ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΚΙΝΔΥΝΟΥ ΜΕ ΧΡΗΣΗ ΓΕΝΙΚΕΥΜΕΝΩΝ ΓΡΑΜΜΙΚΩΝ ΜΟΝΤΕΛΩΝ ΣΕ ΜΟΝΟΕΤΗ ΑΣΦΑΛΙΣΤΙΚΑ ΣΥΜΒΟΛΑΙΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΟΥ

ΕΥΑΓΟΡΟΥ ΒΑΣΙΛΗ

GE19712

Επιβλέπων καθηγητής:
Φουσκάκης Δημήτρης
Καθηγητής, Τομέας Μαθηματικών

Αθήνα, Ιούνιος 2024

Τριμελής επιτροπή:
Φουσκάκης Δημήτρης, Καθηγητής, Τομέας Μαθηματικών
Λουλάκης Μιχαήλ, Καθηγητής, Τομέας Μαθηματικών
Παπαπαντολέων Αντώνης, Αναπλ. Καθηγητής, Τομέας Μαθηματικών

Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

.....

Ευαγόρου Βασίλης

στους γονείς μου

Ευχαριστίες

Θα ήθελα αρχικά να ευχαριστήσω θερμά τον επιβλέποντα Καθηγητή της παρούσας διπλωματικής εργασίας κ. Φουσκάκη Δημήτρη για την εμπιστοσύνη που έδειξε σε εμένα και τις δυνατότητες μου αναλαμβάνοντας την επίβλεψη της εργασίας. Είναι τιμή μου που αυτή η διπλωματική εργασία θα έχει τη σφραγίδα ενός καταξιωμένου Καθηγητή στον τομέα της Στατιστικής προσδίδοντας στην εργασία και σε εμένα κύρος. Η ακαδημαϊκή μου εμπειρία μαζί σας ήταν και θα είναι εξαιρετικά πολύτιμη.

Επίσης, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα κ. Ζαμπέλη Σωτήρη για την προσφορά του και τις κατατοπιστικές διαπιστώσεις του σε όλη τη διαδικασία εκπόνησης της εργασίας αυτής. Η παρουσία του ήταν ωφέλιμη και ευεργετική όχι μόνο από τις γνώσεις του στο αντικείμενο για την εκπόνηση της διπλωματικής εργασίας αλλά και γενικότερα στις εμπειρίες και συμβουλές που μοιράστηκε μαζί μου σε όλη αυτή τη διάρκεια.

Τέλος, με αφορμή το πέρας της ακαδημαϊκής μου εμπειρίας στη Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών, θέλω να ευχαριστήσω τους φίλους και συμφοιτητές μου, Χατζηγιάννη Γεωργία, Αντωνίου Σταύρο και Ήλια Δημητρίου για την πολύτιμη βοήθεια τους σε ακαδημαϊκό και προσωπικό επίπεδο. Τέλος, το μεγαλύτερο ευχαριστώ στην οικογένεια και τους φίλους μου για την αμέριστη στήριξη τους όλα αυτά τα χρόνια.

Σας ευχαριστώ όλους.

Περίληψη

Κύριο μέλημα της διπλωματικής εργασίας αυτής είναι να μελετηθεί η συχνότητα των απαιτήσεων σε μονοετή ασφαλιστικά συμβόλαια χαρτοφυλακίου του τομέα ασφάλισης οχημάτων, καθώς και να διερευνηθεί η σχέση της με διάφορους παράγοντες που την επηρεάζουν. Το αντικείμενο της διπλωματικής αρχικά επικεντρώνεται στην εκτενή παρουσίαση όλου του θεωρητικού υπόβαθρου των Γενικευμένων Γραμμικών Μοντέλων και μέσω αυτού στην ανάπτυξη τριών διαφορετικών τύπων μοντέλων για την απάντηση του ερευνητικού ερωτήματος. Σκοπός είναι να κατασκευαστούν μοντέλα τα οποία θα μοντελοποιούν τη συχνότητα απαιτήσεων με τη χρήση πραγματικών δεδομένων που παρέχουν πληροφορία σχετικά με ασφαλιστικά συμβόλαια. Μέσω των μοντέλων αυτών θα εντοπιστούν οι κυριότερες μεταβλητές που επηρεάζουν τη συχνότητα απαιτήσεων και θα γίνουν προβλέψεις. Η διαδικασία κατασκευής των μοντέλων θα γίνει μέσω μίας ολοκληρωμένης στατιστικής ανάλυσης δεδομένων.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο της εργασίας

Το αντικείμενο της διπλωματικής εργασίας αυτής σχετίζεται άμεσα με την Αναλογιστική Επιστήμη. Στον ασφαλιστικό κόσμο, τα Αναλογιστικά μαθηματικά κατέχουν μεγάλη σημασία αφού αποτελούν το επιστημονικό υπόβαθρο με βάση το οποίο μπορούν να αναπτυχθούν μοντέλα τα οποία συμβάλλουν στην καλύτερη αξιολόγηση του ρίσκου οδηγώντας στη βιωσιμότερη οικονομική ανάπτυξη και σταθερότητα κάποιας εταιρίας. Η ανάλυση του ρίσκου αποτελεί πλέον βασική ανάγκη των εταιριών, ειδικότερα σε μία εποχή όπου η αβεβαιότητα είναι έντονη και έχει εισβάλει σε πολλούς τομείς της καθημερινότητας του ανθρώπου όπως για παράδειγμα η υγεία και η κλιματική κρίση. Σκοπός κάθε μοντελοποίησης είναι η πρόβλεψη τιμών μίας μεταβλητής μέσω της χρήσης κάποιων άλλων. Αυτό συμβαίνει και στο μέρος της Αναλογιστικής επιστήμης σε αυτή τη διπλωματική εργασία, όπου αναπτύσσονται μοντέλα τα οποία θα εξυπηρετούν τις ανάγκες κάποιας ασφαλιστικής εταιρίας να προβλέψει ζημιές, αποθέματα, τιμές ασφαλίστρων κ.λ.π.

Τα τελευταία χρόνια, στο πεδίο της Στατιστικής υπάρχει τεράστια εξέλιξη και προτροπή για χρήση όλο και πιο νέων και ευέλικτων τεχνικών μοντελοποίησης οι οποίες μπορούν να διαχειριστούν πληθώρα τύπου δεδομένων και κατανομών. Μία από αυτές τις κλάσεις μοντέλων που έχει αποκτήσει ευρεία χρησιμότητα είναι τα Γενικευμένα Γραμμικά Μοντέλα (ΓΓΜ). Τα μοντέλα αυτά παρέχουν προσαρμοστικότητα στη μοντελοποίηση ποσοτικών και κατηγορικών μεταβλητών, ενώ δεν απαιτούν την κανονικότητα των σφαλμάτων όπως στο Πολλαπλό Γραμμικό Μοντέλο ή αλλιώς τη γνωστή Πολλαπλή Παλινδρόμηση. Η θεωρία των Γενικευμένων γραμμικών μοντέλων έχει θεμελιωθεί από τους Nelder και Wedderburn το 1972, σε μία προσπάθεια να επεκτείνουν την εφαρμογή των παραδοσιακών γραμμικών μοντέλων. Από τότε, τα Γενικευμένα γραμμικά μοντέλα έχουν αποκτήσει τεράστια χρησιμότητα με εφαρμογή σε πλήθος πεδίων όπως η επιδημιολογία, τα χρηματοοικονομικά, οι κοινωνικές επιστήμες και η αναλογιστική επιστήμη.

Σκοπός της παρούσας διπλωματικής εργασίας είναι αρχικά να παρουσιαστεί ολοκληρωμένα η θεωρία των Γενικευμένων γραμμικών μοντέλων και μέσω αυτής να αναπτυχθούν τρεις διαφορετικοί τύποι μοντέλων. Ενδιαφερόμαστε να προσαρμοστούν μοντέλα τα οποία θα εκτιμούν τη συχνότητα απαιτήσεων δοσμένου δεδομένων που παρέχουν πληροφορία σχετικά με ασφαλιστικά συμβόλαια οχημάτων από ένα χαρτοφυλάκιο. Μέσω τον μοντέλων αυτών θα εντοπιστούν οι κυριότερες μεταβλητές που επηρεάζουν τη συχνότητα απαιτήσεων και θα γίνουν προβλέψεις. Η διαδικασία κατασκευής μοντέλων θα γίνει μέσω μίας ολοκληρωμένης στατιστικής ανάλυσης των δεδομένων.

1.2 Ασφάλιση

Η ασφάλιση λειτουργεί μέσω της διαδικασίας μεταφοράς ρίσκου από το άτομο σε κάποιο ασφαλιστικό φορέα με αντάλλαγμα την πληρωμή ενός ασφαλίστρου (premium). Η διαδικασία αυτή επισημοποιείται μέσω μίας σύμβασης μεταξύ του ατόμου και του φορέα η οποία θέτει τους όρους και τις προϋποθέσεις της κάλυψης του ασφαλιζόμενου. Κύριο μέλημα του ασφαλιστικού φορέα είναι η κάλυψη απώλειας ή ζημιάς, όταν αυτή καλύπτεται από το συμβόλαιο στην περίπτωση που υποβληθεί κάποια απαίτηση (claim) από την πλευρά του ασφαλιζόμενου.

Η συχνότητα στην οποία θα γίνουν οι απαιτήσεις αυτές, δηλαδή η απαίτηση προς την εταιρία να καλύψει κάποιο συμβάν, αναφέρεται στον αριθμό των απαιτήσεων που υποβάλλονται εντός μίας καθορισμένης περιόδου (claims frequency). Η εκτίμηση της συχνότητας των απαιτήσεων παρέχει στους ασφαλιστές πολύτιμες πληροφορίες σχετικά με μελλοντικές απαιτήσεις δίνοντας τους τη δυνατότητα να εκτιμήσουν και να εξασφαλίσουν αποθεματικά με σκοπό την αποπληρωμή μελλοντικών απαιτήσεων αλλά και να ορίσουν τα ασφάλιστρα τους. Η συχνότητα κυρίως εξαρτάται από την παράμετρο της έκθεσης (exposure). Στην ασφάλιση η έκθεση αναφέρεται ως ο βαθμός στον οποίο ο ασφαλιστικός φορέας υπόκειται σε ρίσκο ή ο λήπτης της ασφάλισης σε ενδεχόμενη απώλεια ή ζημιά. Η έκθεση μπορεί να μετρηθεί με διάφορους τρόπους. Για παράδειγμα, στη συχνότητα απαιτήσεων θεωρείται ως έκθεση η χρονική περίοδος στην οποία ο ασφαλιζόμενος υπόκειται σε απώλεια ή ζημιά. Αυτή ορίζεται ως:

$$\text{Συχνότητα απαιτήσεων} = \frac{\text{Αριθμός απαιτήσεων}}{\text{Έκθεση}}.$$

Η συχνότητα των απαιτήσεων συχνά αναφέρεται και ως αριθμός απαιτήσεων όμως και οι δύο έννοιες ερμηνεύονται με βάση την διάρκεια έκθεσης.

Στην παρούσα διπλωματική εργασία, η ασφάλιση έγκειται στον χώρο ασφάλισης οχημάτων (auto insurance). Η συχνότητα απαιτήσεων διαφοροποιείται ανάλογα με τον χώρο ασφάλισης και σε κάποιες περιπτώσεις είναι ιδιαίτερα υψηλή. Στο πεδίο αυτό, μία ασφαλιστική κάλυψη υπόκειται σε ρίσκα ατυχήματος, κλοπής, βανδαλισμού και ζημιάς. Επομένως, οι συχνότητες είναι αυξημένες. Παράγοντες που την επηρεάζουν είναι τα χαρακτηριστικά του φυσικού προσώπου, του οχήματος, γεωγραφικά και δημογραφικά χαρακτηριστικά. Την τελευταία δεκαετία η συχνότητα απαιτήσεων στον χώρο ασφάλισης οχημάτων έχει σταθερή αύξηση. Αυτό λόγω της αύξησης της ιδιοκτησίας αυτοκινήτων, της κυκλοφοριακής συμφόρησης και των κακών συνήθειων οδήγησης. Η συχνότητα αυτή τείνει να είναι αυξημένη σε κατοικημένες περιοχές με μεγάλη πυκνότητα του πληθυσμού και άρα μεγαλύτερης πιθανότητας ατυχήματος έναντι αγροτικών περιοχών. Η ηλικία, το φύλο του ατόμου, οι καιρικές συνθήκες και η επιρροή της τεχνολογίας στην οδική εμπειρία επίσης επηρεάζουν τη συχνότητα απαιτήσεων.

Επίσης, βασικό μέλημα της μελέτης των απαιτήσεων είναι και η εκτίμηση της σφοδρότητας τους (severity) η οποία αναφέρεται στο ποσό, μέγεθος κάποιας απαίτησης. Σε αυτή την περίπτωση η έκθεση ορίζεται ως ο αριθμός των απαιτήσεων που θέτουν ως ποσό αποπληρωμής τη συνολική σφοδρότητα (claims severity). Αυτή ορίζεται ως:

$$\text{Σφοδρότητα απαιτήσεων} = \frac{\text{Μέγεθος απαιτήσεων}}{\text{Αριθμός απαιτήσεων}}.$$

Η σφοδρότητα των απαιτήσεων μεταβάλλεται ανάλογα με τον τύπο της ασφάλισης και το μέγεθος της ζημιάς. Απαιτήσεις με μεγάλες ζημιές μπορούν να αποβούν επιδραστικές στην οικονομική

σταθερότητα κάποιας ασφαλιστικής εταιρίας. Η σχέση μεταξύ της συχνότητας και σφοδρότητας των απαιτήσεων είναι αντίστροφη. Είναι γεγονός ότι απαιτήσεις με υψηλές συχνότητες εμφανίζουν χαμηλά επίπεδα σφοδρότητας και αντίστροφα.

1.3 Δεδομένα

Τα δεδομένα που θα χρησιμοποιηθούν στην παρούσα διπλωματική εργασία προέρχονται από μία Ισπανική ασφαλιστική εταιρία που εδράζεται στον τομέα της ασφάλισης οχημάτων. Το αρχείο δεδομένων είναι ένα χαρτοφυλάκιο και περιέχει πληροφορίες σχετικά με ασφαλιστικά συμβόλαια οχημάτων. Πρόκειται για μονοετή ασφαλιστικά συμβόλαια συγκεκριμένων τύπων οχημάτων τα οποία ήταν ενεργά την τριετία μεταξύ την περίοδο Νοεμβρίου του 2015 και Νοεμβρίου του 2018. Το αρχείο δεδομένων `Motor vehicle insurance data` που θα χρησιμοποιηθεί, είναι τύπου `xlsx` και έχει αντληθεί από ιστοσελίδα που λειτουργείται από το Ινστιτούτο Κοινωνικών Ερευνών του Πανεπιστημίου του Michigan (ISR) και αποτελεί βάση δεδομένων για πολλές έρευνες και δεδομένα. Το αρχείο βρίσκεται σε αυτό τον σύνδεσμο όπως επίσης και το αρχείο `Descriptive of the variables.xlsx` στο οποίο περιέχεται η επεξήγηση όλων των διαθέσιμων μεταβλητών. Κάθε γραμμή του αρχείου αποτελεί ένα συμβόλαιο ενώ κάθε στήλη είναι μία μεταβλητή που περιέχει δημογραφικά στοιχεία, πληροφορίες για το συμβόλαιο, τον ασφαλιζόμενο και το όχημα. Για σκοπούς ομοιογένειας, στα αρχικά ονόματα των μεταβλητών έχει αντικατασταθεί η χρήση κάτω παύλας με τελεία, ενώ όλες οι μεταβλητές ξεκινούν με κεφαλαίο γράμμα.

Κεφάλαιο 2

Γενικευμένα Γραμμικά Μοντέλα

2.1 Εισαγωγή στα Γενικευμένα Γραμμικά Μοντέλα

Έστω Y συνεχής μεταβλητή με βάση την οποία δημιουργείται το τυχαίο δείγμα $Y_1, \ldots, Y_n$ με παρατηρήσεις y_i, $i = 1, \ldots, n$. Άρα, οι τυχαίες μεταβλητές Y_i είναι ανεξάρτητες και ισόνομες μεταξύ τους. Ένα απλό γραμμικό μοντέλο δίνεται από τη σχέση

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n,$$

όπου β_0 και β_1 οι άγνωστες παράμετροι του μοντέλου, x_i η τιμή της επεξηγηματικής μεταβλητής X και ϵ_i τυχαία σφάλματα που ακολουθούν την $\mathcal{N}(0, \sigma^2)$. Το πολλαπλό γραμμικό μοντέλο αποτελεί μία επέκταση του πιο πάνω. Στο πολλαπλό γραμμικό μοντέλο η μεταβλητή απόκρισης Y_i μοντελοποιείται μέσω της σχέσης

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \tag{2.1}$$

όπου x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$ η τιμή για την i-οστή παρατήρηση της j-οστής επεξηγηματικής μεταβλητής, β_j, $j = 0, \ldots, p$ οι άγνωστες παραμέτροι του μοντέλου, ϵ_i τυχαίο σφάλμα που ακολουθεί την $\mathcal{N}(0, \sigma^2)$. Δύο βασικές υποθέσεις των πολλαπλών γραμμικών μοντέλων είναι ότι τα σφάλματα έχουν μηδενική μέση τιμή $\mathbb{E}(\epsilon_i) = 0$ για κάθε i και ότι η διασπορά τους είναι σταθερή και δεν αλλάζει μεταξύ των παρατηρήσεων, $V(\epsilon_i) = \sigma^2$. Η ιδιότητα αυτή αναφέρεται ως ομοσκεδαστηκότητα. Υποθέτουμε επομένως ότι τα ϵ_i είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές, κανονικά κατανεμημένες, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Λαμβάνοντας μέσες τιμές στην 2.1, η αναμενόμενη τιμή της εξαρτημένης μεταβλητής Y_i είναι

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \mu_i,$$

ενώ από την ίδια εξίσωση λαμβάνεται ότι

$$V(Y_i) = V(\epsilon_i) = \sigma^2.$$

Η κανονικότητα των σφαλμάτων συνεπάγει και την κανονικότητα της μεταβλητής απόκρισης Y_i δοσμένων των τιμών των επεξηγηματικών μεταβλητών.

Ο σκοπός των Γενικευμένων γραμμικών μοντέλων είναι η επέκταση της έννοιας των πολλαπλών γραμμικών μοντέλων. Η ανάπτυξη της θεωρίας αυτής προήλθε από την ανάγκη μοντελοποίησης σε περιπτώσεις όπου η πολλαπλή παλινδρόμηση δεν ήταν εφαρμόσιμη. Η αδυναμία των πολλαπλών

γραμμικών μοντέλων έγκειται κυρίως στο ότι μπορούν να μοντελοποιήσουν μόνο κανονικές τυχαίες μεταβλητές αλλά και το ότι η διασπορά της μεταβλητής απόκρισης θεωρείται σταθερή μεταξύ των παρατηρήσεων.

Για παράδειγμα, έστω Y ο αριθμός των απαιτήσεων και $\boldsymbol{x}_i$ οι παρατηρήσεις ενός χαρτοφυλακίου με δεδομένα ασφάλισης. Σκοπός είναι να μοντελοποιηθεί ο αναμενόμενος αριθμός απαιτήσεων με βάση τις p επεξηγηματικές μεταβλητές με τιμές $x_{i1}, x_{i2}, \ldots, x_{ip}$. Έστω ότι για τη μοντελοποίηση γίνεται χρήση ενός πολλαπλού γραμμικού μοντέλου. Έστω ότι η Y ακολουθεί κατανομή Poisson, τότε η διασπορά $V(Y_i) = \mathbb{E}(Y_i)$. Αυτό σημαίνει πως η διασπορά δεν είναι σταθερή μεταξύ των παρατηρήσεων και άρα παραβιάζεται η υπόθεση της ομοσκεδαστηκότητας. Επιπλέον, αν μοντελοποιηθεί ο αριθμός απαιτήσεων, το δεξί μέλος της 2.1 θα πρέπει να είναι μη αρνητικό. Αυτό όμως δεν εξασφαλίζεται από το μοντέλο. Είναι πιθανόν κάποιος συνδυασμός των x_{ij} να φέρει αρνητική τιμή της γραμμικής προβλέπουσας $\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ και άρα το μοντέλο θα πρέπει να τροποποιηθεί. Αν θεωρηθεί ότι η ποσότητα $\ln(\mathbb{E}(Y_i))$ ισούται με την η_i, τότε

$$\ln(\mathbb{E}(Y_i)) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

και άρα

$$\mathbb{E}(Y_i) = e^{\beta_0} e^{\beta_1 x_{i1}} \ldots e^{\beta_p x_{ip}}.$$

Με το μοντέλο αυτό, ο αναμενόμενος αριθμός απαιτήσεων θα είναι μη αρνητικός και το μοντέλο γίνεται πολλαπλασιαστικό αντί προσθετικό. Η τροποποίηση αυτή όμως δυσκολεύει στην εκτίμηση των παραμέτρων β_j αφού πλέον η εξίσωση είναι μη γραμμική. Τα Γενικευμένα γραμμικά μοντέλα δίνουν λύσεις στα προβλήματα αυτά.

Τα Γενικευμένα γραμμικά μοντέλα επεκτείνουν την έννοια των πολλαπλών γραμμικών μοντέλων με δύο τρόπους:

- Οι μεταβλητές απόκρισης Y_i συνδέονται με τον γραμμικό συνδυασμό των επεξηγηματικών μεταβλητών x_{ij} μέσο μίας μη γραμμικής συνάρτησης.
- Η διασπορά των Y_i δεν είναι απαραίτητα σταθερή αλλά μπορεί να είναι συνάρτηση της αναμενόμενης τιμής της Y_i.

Η μορφή ενός Γενικευμένου γραμμικού μοντέλου με μεταβλητή απόκρισης Y είναι

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \tag{2.2}$$

με $g(\mathbb{E}(Y_i)) = g(\mu_i)$ η συνάρτηση σύνδεσης (link function) της μεταβλητής απόκρισης Y_i με τη γραμμική προβλέπουσα η_i. Η συνάρτηση σύνδεσης δεν είναι απαραίτητα γραμμική αλλά πρέπει να είναι παραγωγίσιμη και γνησίως μονότονη. Λόγω της ιδιότητας της αυτής, θα είναι και αντιστρέψιμη και άρα η σχέση 2.2 μπορεί να γραφεί ως

$$\mathbb{E}(Y_i) = g^{-1}(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}). \tag{2.3}$$

Ακόμη και αν οι επεξηγηματικές μεταβλητές x_{ij} είναι γραμμικά συνδεδεμένες μεταξύ τους, η μεταβλητή απόκρισης μπορεί να είναι μη γραμμικά συνδεδεμένη με αυτό τον γραμμικό συνδυασμό. Αφού $\mu_i = \mathbb{E}(Y_i)$ λαμβάνεται η πιο σύντομη μορφή της 2.3,

$$\mu_i = g^{-1}(\eta_i). \tag{2.4}$$

Η ποσότητα

$$\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta},$$

καλείτε συστηματικό ή μη στοχαστικό μέρος του μοντέλου αφού οι τιμές $\mathbf{x}_{ij}$ είναι ήδη προκαθορισμένες στο δείγμα. Γράφουμε ως $\mathbf{x}_i$ την i-οστή παρατήρηση με $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$, το i-οστό διάνυσμα γραμμή του πίνακα σχεδιασμού $\mathbf{X}$ με μορφή

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix},$$

με πλήθος γραμμών n και $p+1$ στήλες, όπου p δηλώνει το πλήθος των επεξηγηματικών μεταβλητών.

Μία σημαντική υπόθεση των Γενικευμένων γραμμικών μοντέλων είναι ότι η κατανομή της μεταβλητής απόκρισης Y ανήκει στην Εκθετική οικογένεια κατανομών. Στην πιο κάτω ενότητα παρουσιάζονται οι ιδιότητες της Εκθετικής οικογένειας κατανομών.

2.2 Η Εκθετική Οικογένεια Κατανομών

Η Εκθετική οικογένεια κατανομών (ΕΟΚ) αποτελεί βασική συνιστώσα των Γενικευμένων γραμμικών μοντέλων και αυτό διότι η μεταβλητή απόκρισης θα πρέπει ανήκει στην οικογένεια αυτή. Οποιαδήποτε τυχαία μεταβλητή της οποίας η συνάρτηση πυκνότητας πιθανότητας (σππ) για συνεχή Y ή η συνάρτηση μάζας πιθανότητας (σμπ) για διακριτή Y είναι της μορφής

$$f_Y(y; \theta, \phi) \equiv f(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right), \tag{2.5}$$

με θ και ϕ παραμέτρους, ανήκει στην Εκθετική οικογένεια. Το στήριγμα $S = \{y \in \mathbb{R} : f(y) > 0\}$ είναι ανεξάρτητο του θ και ϕ, ενώ οι συναρτήσεις $b(\theta)$ και $a(\phi)$ είναι γνωστές. Η παράμετρος θ ονομάζεται κανονική παράμετρος και η ϕ παράμετρος μεταβλητότητας (dispersion). Για τυχαία μεταβλητή Y με βάση την οποία δημιουργείται το τυχαίο δείγμα $Y_1, Y_2, \ldots, Y_n$, η $a(\phi)$ συχνά έχει τη μορφή $\frac{\phi}{w_i}$ με την παράμετρο μεταβλητότητας ϕ να θεωρείται σταθερή στο δείγμα και τα βάρη w_i μεταβλητά. Η συνάρτηση $b(\theta)$ καθορίζει τη μέση τιμή της μεταβλητής Y.

2.2.1 Μέση τιμή και διασπορά

Η μέση τιμή μ και διασπορά σ^2 κατανομής μέλους της ΕΟΚ δίνονται αντίστοιχα από τις σχέσεις:

1. $\mathbb{E}(Y) = \mu = b'(\theta)$
2. $V(Y) = \sigma^2 = a(\phi)b''(\theta)$

Απόδειξη: Έστω χωρίς βλάβη της γενικότητας ότι η Y είναι συνεχής τυχαία μεταβλητή με τιμές στον $\mathbb{R}$.

1. Αφού η $f(y; \theta, \phi)$ είναι μία συνάρτηση πυκνότητας πιθανότητας, θα ισχύει ότι

$$\int_{\mathbb{R}} f(y; \theta, \phi)\, dy = 1 \implies \int_{\mathbb{R}} \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right) dy = 1. \tag{2.6}$$

Παραγωγίζοντας και τα δύο μέλη της εξίσωσης ως προς θ προκύπτει ότι

$$\int_{\mathbb{R}} \left(\frac{y - b'(\theta)}{a(\phi)} \right) \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right) dy = 0,$$

άρα

$$\int_{\mathbb{R}} y \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right) dy = b'(\theta) \int_{\mathbb{R}} \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right) dy,$$

όπου το αριστερό μέλος είναι η μέση τιμή μ της τυχαίας μεταβλητής Y και από το δεξί παραμένει η $b'(\theta)$ λόγω της (2.6). Άρα

$$\mathbb{E}(y) = \mu = b'(\theta). \tag{2.7}$$

2. Παραγωγίζοντας την (2.6) δύο φορές ως προς θ προκύπτει ότι

$$-\frac{b''(\theta)}{a(\phi)} \int_{\mathbb{R}} \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right) dy$$

$$+ \int_{\mathbb{R}} \frac{(y - b''(\theta))^2}{(a(\phi))^2} \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right) dy = 0,$$

όπου το ολοκλήρωμα του πρώτου όρου είναι ίσο με 1 και αφού $\mathbb{E}(Y) = \mu = b'(\theta)$ προκύπτει ότι

$$\frac{1}{a^2(\phi)} V(Y) = \frac{b''(\theta)}{a^2(\phi)}.$$

Άρα

$$V(Y) = a(\phi) b''(\theta) = a(\phi) V(\mu), \tag{2.8}$$

με

$$V(\mu) = b''(\theta),$$

η συνάρτηση διασποράς (variance function) που δείχνει τη σχέση μεταξύ της μέσης τιμής και της διασποράς αφού

$$V(\mu) = \frac{V(y)}{a(\phi)}.$$

Συχνά επίσης επιλέγεται

$$a(\phi) = a\phi,$$

με a σταθερά διότι σε πολλές κατανομές η συνάρτηση αυτή είναι πολλαπλάσιο της ϕ όπως φαίνεται και στον Πίνακα 2.2.

2.2.2 Η συνάρτηση διασποράς

Αν $Y_1, \ldots, Y_n$ τυχαίο δείγμα από κατανομή της εκθετικής οικογένειας αποδείχτηκε ότι

$$b''(\theta_i) = \frac{db'(\theta_i)}{d\theta_i} = \frac{d\mu_i}{d\theta_i} = V(\mu_i)$$

και αφού

$$\mu_i = b'(\theta_i)$$

$$\theta_i = b'^{-1}(\mu_i),$$

Κατανομή	Συνάρτηση διασποράς $V(\mu)$
Κανονική	1
Poisson	μ
Διωνυμική	$\mu(1-\mu)$
Γάμμα	μ^2
Αντίστροφη Γάμμα	μ^3
Αρνητική Διωνυμική	$\mu(1+\kappa\mu)$

Πίνακας 2.1: Συναρτήσεις διασποράς

τότε η συνάρτηση διασποράς γράφεται ως

$$V(\mu_i) = b''(b'^{-1}(\mu_i)).$$

Επιπλέον,

$$V(Y_i) = a(\phi)V(\mu_i)$$

και άρα

$$V(Y_i) = a(\phi)b''(b'^{-1}(\mu_i)).$$

Η συνάρτηση διασποράς είναι μεγάλης σημασίας αφού ορίζει τη σχέση μεταξύ της μέσης τιμής και της διασποράς κάποιας κατανομής της εκθετικής οικογένειας. Στα Γενικευμένα γραμμικά μοντέλα η αναμενόμενη τιμή μ_i εξαρτάται από τις επεξηγηματικές μεταβλητές και άρα μεταβάλλεται. Όταν μεταβάλλεται η μέση τιμή θα μεταβάλλεται και η διασπορά μέσω της $V(\mu_i)$. Στα πολλαπλά γραμμικά μοντέλα όπου η μεταβλητή απόκρισης ακολουθεί την κανονική κατανομή ισχύει ότι $V(\mu) = 1$ και επομένως η διασπορά δε μεταβάλλεται με τη μέση τιμή ενώ επίσης η συνάρτηση $a(\phi)$ είναι σταθερή και ίση με σ^2.

2.2.3 Βασικές κατανομές της Εκθετικής οικογένειας

Θα παρουσιαστούν οι κύριες κατανομές της ΕΟΚ. Οποιαδήποτε σππ ή σμπ μπορεί να γραφεί στην πιο κάτω μορφή ανήκει στην Εκθετική οικογένεια:

$$\ln\, f(y) = \ln(c(y,\phi)) + \frac{y\theta - b(\theta)}{a(\phi)}. \tag{2.9}$$

Διωνυμική κατανομή: Υποθέστε ότι $Y \sim B(n,\pi)$, τότε η συνάρτηση μάζας πιθανότητας της Y είναι

$$f(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y = 0,1,\ldots,n.$$

Από τον παραγοντικό όρο προκύπτει η $c(y,\phi)$ που δεν προσφέρει κάποια πληροφορία. Αγνοώντας τον προκύπτει ότι

$$\ln f(y) = \ln(\pi^y(1-\pi)^{n-y}) = y\ln(\frac{\pi}{1-\pi}) + n\ln(1-\pi) = \frac{y\theta - b(\theta)}{a(\phi)}$$

και άρα

$$\theta = \ln(\tfrac{\pi}{1-\pi}),\ b(\theta) = n\ln(1+e^\theta) \text{ και } a(\phi) = 1.$$

Επομένως, η Διωνυμική κατανομή ανήκει στην Εκθετική οικογένεια και προκύπτει ότι η μέση τιμή είναι

$$\mathbb{E}(y) = b'(\theta) = \frac{ne^{\theta}}{1+e^{\theta}} = n\pi,$$

ενώ για τη διασπορά

$$V(y) = \phi b''(\theta) = n\pi(1-\pi).$$

Κατανομή Poisson: Υποθέστε ότι $Y \sim P(\mu)$, τότε η συνάρτηση μάζας πιθανότητας της Y είναι

$$f(y) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

και άρα

$$\ln f(y) = -\mu + y\ln(\mu) - \ln(y!) = -\ln(y!) + \frac{y\theta - b(\theta)}{a(\phi)},$$

με $\theta = \ln(\mu)$, $b(\theta) = \mu$ και $a(\phi) = 1$. Επομένως, η μέση τιμή είναι

$$\mathbb{E}(y) = b'(\theta) = (e^{\theta})' = e^{\theta} = \mu$$

και η διασπορά

$$V(y) = b''(\theta) = e^{\theta} = \mu.$$

Προκύπτει λοιπόν ότι στην Poisson κατανομή η μέση τιμή και η διασπορά ταυτίζονται και άρα μεταβολές στην μ επηρεάζουν άμεσα τη σ^2.

Κανονική κατανομή: Υποθέστε ότι $Y \sim \mathcal{N}(\mu, \sigma^2)$, τότε η συνάρτηση πυκνότητας πιθανότητας της Y είναι

$$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left(\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right) \exp\left(-\frac{y^2 - 2y\mu + \mu^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \left(\frac{y^2}{2\sigma^2} + \ln(\sqrt{2\pi\sigma^2})\right)\right), y \in \mathbb{R} \end{aligned} \tag{2.10}$$

και άρα αγνοώντας τον σταθερό όρο $\ln(\sqrt{2\pi})$ προκύπτει ότι

$$\ln f(y) = -\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{(y-\mu)^2}{2\sigma^2} = -\ln\sigma - \frac{y^2/2}{\sigma^2} + \frac{y\mu - \mu^2/2}{\sigma^2},$$

με $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$, $a(\phi) = \sigma^2$ και $c(y,\phi) = -\ln\sigma - \frac{y^2/2}{\sigma^2}$. Επομένως,

$$\mathbb{E}(y) = b'(\theta) = \left(\frac{\theta^2}{2}\right)' = \theta = \mu$$

και

$$V(y) = a(\phi)b''(\theta) = \sigma^2.$$

Γάμμα κατανομή: Υποθέστε ότι $Y \sim \text{Gamma}(a, \gamma)$, τότε η συνάρτηση πυκνότητας πιθανότητας της Y είναι

$$f(y;a,\gamma) = \frac{\gamma^{a} y^{a-1} e^{-\gamma y}}{\Gamma(a)},\ y \geq 0.$$

Είναι γνωστό ότι η μέση τιμή της κατανομής είναι $\mathbb{E}(Y) = \frac{a}{\gamma}$ και η διασπορά της $V(Y) = \frac{a}{\gamma^2}$. Αλλάζοντας τις παραμέτρους της κατανομής σε $\mu = \frac{a}{\gamma}$, $\nu = a$ και άρα $\gamma = \frac{\nu}{\mu}$ προκύπτει η κάτωθι παραμετροποίηση της σππ:

$$\begin{aligned} f(y;\mu,\nu) &= \frac{\left(\frac{\nu}{\mu}\right)^{\nu}}{\Gamma(\nu)} y^{\nu-1} e^{-\frac{\nu}{\mu}y} \\ &= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu} y\right)^{\nu} y^{-1} e^{-\frac{\nu}{\mu}y} \\ &= \exp\left(-\ln\Gamma(\nu) + \nu\ln\nu - \nu\ln\mu + (\nu-1)\ln y - \frac{ny}{\mu}\right) \\ &= \exp\left(\frac{y(-\frac{1}{\mu}) - \ln\mu}{1/\nu} + (\nu-1)\ln y - \ln\Gamma(\nu) + \nu\ln\nu\right). \end{aligned}$$

Επομένως,

$$\theta = -\frac{1}{\mu},\ b(\theta) = \ln\mu,\ a(\phi) = \frac{1}{\nu}$$

και άρα

$$c(y,\nu) = (\nu-1)\ln y - \ln\Gamma(\nu) + \nu\ln\nu.$$

Επειδή $\theta = -\frac{1}{\mu}$ και η $\mu = \frac{a}{b}$ είναι θετική ποσότητα πρέπει $\theta < 0$. Αντικαθιστώντας τη μέση τιμή στην $b(\theta)$ προκύπτει

$$b(\theta) = \ln\left(-\frac{1}{\theta}\right) - \ln(-\theta)$$

και άρα η μέση τιμή είναι

$$b'(\theta) = \frac{d}{d\theta}(-\ln\theta) = -\frac{1}{\theta} = \mu.$$

Επομένως, η διασπορά είναι

$$\frac{1}{\nu\theta^2},$$

το οποίο ισούται με

$$\frac{a}{\gamma^2},$$

στην αρχική παραμετροποίηση.

Αντίστροφη κανονική κατανομή: Υποθέστε ότι $Y \sim IG(\mu,\lambda)$, τότε η συνάρτηση πυκνότητας πιθανότητας της Y είναι

$$f(y;\mu,\sigma^2) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left(\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right),\ y>0, \lambda>0.$$

Αλλάζοντας τη μεταβλητή $\lambda = \frac{1}{\sigma^2}$ προκύπτει η κάτωθι παραμετροποίηση της σππ:

$$f(y) = \frac{1}{\sqrt{2\pi y^3}\sigma} \exp\left(-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right)$$

και άρα

$$\begin{aligned} \ln f(y) &= -\ln\sqrt{2\pi y^3}\sigma - \frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2 \\ &= -\frac{1}{2}\left(\ln 2\pi y^3\sigma^2 + \frac{1}{y\sigma^2}\right) + \frac{-\frac{y}{2\mu^2} + \frac{1}{\mu}}{\sigma^2}. \end{aligned}$$

Επομένως προκύπτει ότι $\theta = -\frac{1}{2\mu^2}$, $b(\theta) = -\sqrt{-2\theta} = \frac{1}{\mu}$, $c(y,\phi) = -\frac{1}{2}\left[\ln(2\pi\phi y^3) + \frac{1}{\phi y}\right]$ και $a(\phi) = \sigma^2$. Η μέση τιμή είναι

$$b'(\theta) = -\frac{1}{\sqrt{-2\theta}} = \mu$$

και η διασπορά

$$V(Y) = b''(\theta)a(\phi) = \frac{1}{-2\theta\sqrt{-2\theta}}\sigma^2 = \mu^3\sigma^2.$$

Αρνητική διωνυμική κατανομή: Υποθέστε ότι $Y \sim NB(r,\pi)$ με γνωστό r, τότε η συνάρτηση μάζας πιθανότητας της Y είναι

$$f(y;r,\pi) = \pi\binom{r+y-1}{r-1}\pi^{r-1}(1-\pi)^y = \binom{r+y-1}{r-1}\pi^r(1-\pi)^y, \quad y = 0,1,\ldots.$$

Για οποιοδήποτε $r > 0$ η πιο πάνω συνάρτηση μπορεί να γραφεί με τη χρήση συναρτήσεων Gamma ως

$$f(y) = \frac{\Gamma(y+r)}{y!\Gamma(r)\pi^r(1-\pi)^y}, \quad y = 0,1,2,\ldots.$$

Έστω

$$\mu = \frac{r(1-\pi)}{\pi}, \quad \kappa = \frac{1}{r}$$

και προκύπτει η σμπ της Y όταν ακολουθεί την $\mathrm{NB}(\mu,\kappa)$

$$f(y) = \frac{\Gamma(y+\frac{1}{\kappa})}{y!\Gamma(\frac{1}{\kappa})}\left(\frac{1}{1+\kappa\mu}\right)^{\frac{1}{\kappa}}\left(\frac{\kappa\mu}{1+\kappa\mu}\right)^y, \quad y = 0,1,2,\ldots. \tag{2.11}$$

Επομένως,

$$\begin{aligned}\ln f(y;\kappa,\mu) &= \ln\left[\frac{\Gamma\left(y+\frac{1}{\kappa}\right)}{y!\Gamma\left(\frac{1}{\kappa}\right)}\right] + \frac{1}{\kappa}\ln\left(\frac{1}{1+\kappa\mu}\right) + y\ln\left(\frac{\kappa\mu}{1+\kappa\mu}\right)\\ &= \ln\left[\frac{\Gamma\left(y+\frac{1}{\kappa}\right)}{y!\Gamma\left(\frac{1}{\kappa}\right)}\right] - \frac{1}{\kappa}\ln(1+\kappa\mu) + y\ln(\kappa) + y\ln\left(\frac{\mu}{1+\kappa\mu}\right)\end{aligned} \tag{2.12}$$

και άρα

$$\phi = 1,\ \theta = \ln\left(\frac{\mu}{1+\kappa\mu}\right),\ b(\theta) = -\frac{1}{\kappa}\ln(1+\kappa\mu),\ c(y,\phi) = \ln\left[\frac{\Gamma\left(y+\frac{1}{\kappa}\right)}{y!\Gamma\left(\frac{1}{\kappa}\right)}\right] + y\ln(\kappa).$$

Επομένως,

$$\mu = \frac{e^\theta}{1-\kappa e^\theta} \implies b(\theta) = -\frac{1}{\kappa}\ln\left(\frac{1}{1-\kappa e^\theta}\right) = -\frac{1}{\kappa}\ln(1-\kappa e^\theta)$$

και η μέση τιμή είναι ίση με την παράμετρο μ ενώ η διασπορά είναι

$$V(y) = a(\phi)b''(\theta) = \frac{e^\theta}{(1-\kappa e^\theta)^2} = \mu(1+\kappa\mu). \tag{2.13}$$

Στον Πίνακα 2.2 παρουσιάζονται έξι βασικές κατανομές της εκθετικής οικογένειας με τις παραμέτρους τους.

Συνάρτηση μάζας ή πυκνότητας πιθανότητας	θ	$b(\theta)$	ϕ	$a(\phi)$	$c(y,\phi)$
Κανονική: $\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$	μ	$\frac{\theta^2}{2}$	σ^2	ϕ	$-\ln(\sigma)-\frac{y^2/2}{\sigma^2}$
Poisson: $\frac{e^{-\lambda}\lambda^y}{y!}$	$\ln(\lambda)$	e^θ	1	1	$-\ln(y!)$
Διωνυμική: $\binom{n}{y}p^y(1-p)^{n-y}$	$\ln\left(\frac{p}{1-p}\right)$	$n\ln(1+e^\theta)$	1	1	$\ln\binom{n}{y}$
Γάμμα: $\frac{\gamma^a y^{a-1}\exp(-\gamma y)}{\Gamma(a)}$	$-\frac{\gamma}{a}$	$-\ln(-\theta)$	$\frac{1}{a}$	ϕ	$\frac{1}{\phi}\ln\frac{y}{\phi}-\ln y-\Gamma\left(\frac{1}{\phi}\right)$
Αντίστροφη-Κανονική: $\frac{1}{\sqrt{2\pi y^3\sigma}}\exp\left(-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right)$	$\frac{-1}{2\mu^2}$	$-\sqrt{-2\theta}$	σ^2	ϕ	$-\frac{1}{2}\left[\ln(2\pi\phi y^3)+\frac{1}{\phi y}\right]$
Αρνητική Διων.: $\frac{\Gamma(y+\frac{1}{\kappa})}{y!\Gamma(\frac{1}{\kappa})}\left(\frac{1}{1+\kappa\mu}\right)^{\frac{1}{\kappa}}\left(\frac{\kappa\mu}{1+\kappa\mu}\right)^y$	$-\frac{1}{\kappa}\ln(1+\kappa\mu)$	μ	1	1	$\ln\left[\frac{\Gamma(y+\frac{1}{\kappa})}{y!\Gamma(\frac{1}{\kappa})}\right]+y\ln(\kappa)$

Πίνακας 2.2: Βασικές κατανομές της εκθετικής οικογένειας και οι παράμετροι τους

2.3 Προσαρμογή Μοντέλου

Το γενικευμένο γραμμικό μοντέλο γράφεται ως

$$f(y;\theta,\phi)=\exp\left(\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right)$$

με

$$g(\mu)=\boldsymbol{x'\beta},$$

όπου $f(y)$ η συνάρτηση κατανομής της μεταβλητής απόκρισης Y και η $g(\mu)$ η συνάρτηση σύνδεσης μεταξύ της μέσης τιμής μ της Y και των τιμών $\boldsymbol{x}_i$ των επεξηγηματικών μεταβλητών. Η γραμμική προβλέπουσα $\eta_x=\boldsymbol{x'\beta}$ η οποία σχετίζεται ακριβώς με τη συνάρτηση σύνδεσης $g(\mu)$ ως

$$\eta_x=\boldsymbol{x'\beta}=g(\mu_x),$$

γραμμικοποιεί τη σχέση μεταξύ της αναμενόμενης τιμής μ και του γραμμικού εκτιμητή $\boldsymbol{x'\beta}$. Οι κατανομές που παρουσιάστηκαν έχουν μία ή δύο παραμέτρους οι οποίες γενικά είναι άγνωστες και πρέπει να εκτιμηθούν με βάση το διαθέσιμο δείγμα. Είναι βασικό να σημειωθεί ότι οι παρατηρήσεις της μεταβλητής ενδιαφέροντος Y προέρχονται όλες από την ίδια κατανομή με διαφορετικές ωστόσο παραμέτρους. Άρα κάθε παρατήρηση y_i προέρχεται από την $f_Y(y)$ και θεωρούνται ότι είναι ανεξάρτητες μεταξύ τους.

2.3.1 Η συνάρτηση σύνδεσης

Η συνάρτηση σύνδεσης όπως έχει προαναφερθεί θα πρέπει να είναι γνησίως μονότονη, αύξουσα ή φθίνουσα, έτσι ώστε να υπάρχει η αντίστροφη της. Η συνάρτηση σύνδεσης είναι

$$g(\mu_i)=\eta_i,$$

με $\eta_i=\beta_0+\beta_1x_{i1}+\ldots+\beta_px_{ip}$ η γραμμική προβλέπουσα και άρα

$$\mu_i=g^{-1}(\eta_i).$$

Υπάρχουν αρκετές επιλογές για τη συνάρτηση σύνδεσης σε κάθε είδος μεταβλητής απόκρισης αλλά κάποιες μπορεί να μην είναι κατάλληλες. Για παράδειγμα, θα πρέπει να λαμβάνεται υπόψη το σύνολο

τιμών που πρέπει να λαμβάνει η $\mathbb{E}(Y_i)$. Στην εισαγωγή θεωρήθηκε η περίπτωση όπου η Y_i αντιπροσωπεύει τον αριθμό των απαιτήσεων της i παρατήρησης. Ο αναμενόμενος αριθμός απαιτήσεων μ_i είναι θετικός και άρα λαμβάνει τιμές στο $(0, \infty)$. Από την άλλη η γραμμική προβλέπουσα η_i λαμβάνει τιμές στο $(-\infty, \infty)$ και άρα για συγκεκριμένους συνδυασμούς των επεξηγηματικών μεταβλητών ίσως προκύψει ότι $\eta_i < 0$. Αυτό όμως είναι αδύνατο για τον αριθμό των απαιτήσεων. Η λύση σε αυτό το πρόβλημα είναι για παράδειγμα η χρήση της λογαριθμικής συνάρτησης σύνδεσης. Η συνάρτηση είναι η

$$g(\mu) = \ln \mu$$

και είναι τέτοια ώστε

$$\ln \mu : (0, \infty) \to (-\infty, \infty),$$

ενώ η αντίστροφη της λογαριθμικής συνάρτησης σύνδεσης

$$g^{-1}(\eta) = e^{\eta},$$

απεικονίζει το $(-\infty, \infty)$ στο $(0, \infty)$. Με την επιλογή αυτή η αναμενόμενη τιμή είναι πάντα θετική.

Έστω ότι η τυχαία μεταβλητή Y είναι δύτιμη και λαμβάνει τιμές 0 ή 1, ανάλογα με το αν συμβαίνει ή όχι κάποιο γεγονός. Η κατανομή της μεταβλητής απόκρισης Y θα είναι Bernulli(p), όπου p η πιθανότητα επιτυχίας. Η μέση τιμή της είναι η πιθανότητα επιτυχίας p η οποία λαμβάνει τιμές στο $(0, 1)$ και άρα $\mu = p$. Όπως εξηγήθηκε πριν, είναι πιθανόν ο γραμμικός συνδυασμός η να επιφέρει και αρνητικές τιμές. Πρέπει να βρεθεί μία συνάρτηση σύνδεσης η οποία θα έχει πεδίο τιμών το $(0, 1)$ και η επιλογή της γίνεται ως εξής. Αφού p είναι η πιθανότητα επιτυχίας, τότε η σχετική πιθανότητα (odds) είναι $\frac{p}{1-p}$ με

$$\frac{p}{1-p} : (0, 1) \to (0, \infty),$$

ενώ ο λογάριθμος της είναι

$$\ln \frac{p}{1-p} : (0, 1) \to (-\infty, \infty).$$

Η συνάρτηση σύνδεσης $g(\mu_i) = \ln \frac{\mu_i}{n_i - \mu_i} = \ln \frac{p}{1-p}$ καλείται συνάρτηση logit και άρα μέσω αυτής μοντελοποιείται η πιθανότητα p_i. Επιλέγεται $n_i = 1$ αφού σε Bernulli κατανομή τα δεδομένα είναι δυαδικά ενώ $n_i > 1$ για διωνυμικά δεδομένα. Η παραπάνω ανάλυση χρησιμοποιείται στη λογιστική παλινδρόμηση κατά την οποία γίνεται η μοντελοποίηση μίας δύτιμης μεταβλητής απόκρισης. Επομένως, το μοντέλο της λογιστικής παλινδρόμησης έχει τη μορφή

$$\mu_i = \mathbb{E}(Y_i \,|\, \boldsymbol{X} = x) = \mathbb{P}(Y_i = 1 \,|\, \boldsymbol{X} = x) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}, \tag{2.14}$$

ενώ η συμπληρωματική πιθανότητα είναι

$$\mathbb{P}(Y_i = 0 \,|\, \boldsymbol{X} = x) = 1 - \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \frac{1}{1 + e^{\eta_i}}. \tag{2.15}$$

Σε αυτή την περίπτωση μπορεί να χρησιμοποιηθεί και η συνάρτηση probit, $g(\mu) = \Phi^{-1}(\mu)$, με $\Phi^{-1}()$ να είναι η αντίστροφη της συνάρτησης κατανομής της κανονικής κατανομής.

Η συνάρτηση σύνδεσης ονομάζεται κανονική όταν $g(\mu) = \theta$ και έτσι ισχύει ότι

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip},$$

άρα $g(\mu_i) = \theta_i = b'^{-1}(\mu_i)$. Στον Πίνακα 2.3 φαίνονται κάποιες συναρτήσεις σύνδεσης, με την αντίστροφη και το σύνολο τιμών που λαμβάνουν.

	$g(\mu)$	$g^{-1}(\eta)$	Σύνολο τιμών της $g^{-1}(\eta)$
Ταυτοτική	μ	η	$(-\infty, \infty)$
Λογαριθμική	$\ln(\mu)$	e^{η}	$(0, \infty)$
Logit	$\ln(\frac{\mu}{1-\mu})$	$\frac{e^{\eta}}{1+e^{\eta}}$	$(0, 1)$
Probit	$\Phi^{-1}(\mu)$	$\Phi(\eta)$	$(0, 1)$

Πίνακας 2.3: Συναρτήσεις σύνδεσης

2.3.2 Μέγιστη πιθανοφάνεια

Η μέθοδος της μέγιστης πιθανοφάνειας αποσκοπεί στην εκτίμηση των παραμέτρων $\boldsymbol{\theta}$ και ϕ έτσι ώστε να μεγιστοποιείται η πιθανοφάνεια ή ομοίως η λογαριθμοποιημένη πιθανοφάνεια της κατανομής, όταν σε αυτή παρατηρούνται οι τιμές $(y_1, y_2, \ldots, y_n)$. Η συνάρτηση πιθανοφάνειας κατανομής μέλους της εκθετικής οικογένειας είναι

$$L(\theta_i, \phi) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi) = \prod_{i=1}^{n} \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right),$$

με τη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας να είναι η

$$\ell(\boldsymbol{\theta}, \phi) \equiv \ln L(\theta_i, \phi) = \sum_{i=1}^{n} \ln f(y_i; \theta_i, \phi),$$

με $\boldsymbol{\theta}$ το διάνυσμα των παραμέτρων θ_i και σταθερό ϕ μεταξύ των παρατηρήσεων. Επομένως, η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας είναι

$$\sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \ln(c(y_i, \phi))\right) = \sum_{i=1}^{n} \ell_i, \tag{2.16}$$

όπου

$$\ell_i(y_i, \theta_i, \phi) = \left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \ln(c(y_i, \phi))\right).$$

2.3.3 Εκτιμήτριες μέγιστης πιθανοφάνειας

Από τη σχέση

$$\eta_x = \boldsymbol{x'\beta} = g(\mu_x),$$

παρατηρούμε ότι υπάρχει σχέση μεταξύ των παραμέτρων β_j του μοντέλου και των θ_i και ϕ των κατανομών των Y_i. Η εκτίμηση των παραμέτρων β_j του μοντέλου γίνεται μέσω της επίλυσης των εξισώσεων

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = 0,\ j = 0, 1, \ldots, p.$$

Αφού στην εκθετική οικογένεια κατανομών ισχύουν οι σχέσεις

$$\mathbb{E}(y_i) = \mu_i = b'(\theta_i), \quad V(y_i) = a_i(\phi)b''(\theta_i) = a_i(\phi)V(\mu_i), \quad \eta_i = \boldsymbol{x_i'\beta},$$

τότε για την Εκτιμήτρια μέγιστης πιθανοφάνειας (ΕΜΠ) του β_j ισχύει ότι

$$\frac{\partial \ell_i(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

και

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - \mu_i}{a_i(\phi)},$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1} = \left(\frac{\partial b'(\theta_i)}{\partial \theta_i}\right)^{-1} = \big(b''(\theta_i)\big)^{-1} = \left(\frac{V(Y_i)}{a_i(\phi)}\right)^{-1},$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1} = (g'(\mu_i))^{-1},$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij},$$

όπου με τόνο ($'$) συμβολίζεται η παράγωγος ως προς την αντίστοιχη μεταβλητή. Επίσης, μόνο τα θ_i είναι συναρτήσει των β_j. Προκύπτει λοιπόν ότι

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{y_i - \mu_i}{V(Y_i)} \cdot \frac{1}{g'(\mu_i)} x_{ij},\ i = 1, 2, \ldots, n$$

και άρα

$$u_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(Y_i)} \cdot \frac{1}{g'(\mu_i)} x_{ij},\ j = 0, 1, \ldots, p. \tag{2.17}$$

Η (2.17) καλείται συνάρτηση score. Θέτοντας $u_j = 0$, $\forall j$ προκύπτει ένα σύστημα p+1 εξισώσεων, μία για κάθε παράμετρο β_j του μοντέλου. Ισοδύναμα η (2.17) μπορεί να γραφεί και σε μορφή πινάκων ως

$$\boldsymbol{X'D(Y - \mu)} = \mathbf{0}, \tag{2.18}$$

όπου $\boldsymbol{Y}$ είναι ο $n \times 1$ πίνακας στήλη των παρατηρήσεων y_i και D ο διαγώνιος πίνακας με στοιχεία $(V(Y_i)g'(\mu_i))^{-1}$. Οι εξισώσεις αυτές καλούνται εξισώσεις score και από τις οποίες προκύπτουν οι εκτιμήτριες μέγιστης πιθανοφάνειας των β_j. Θεωρώντας τους διαγώνιους πίνακες G και W με στοιχεία $g'(\mu_i)$ και $(V(Y_i)(g'(\mu_i))^2)^{-1}$ αντίστοιχα, τότε $D = WG$ και η (2.18) γίνεται

$$\boldsymbol{X'WG(Y - \mu)} = \mathbf{0}. \tag{2.19}$$

Η εκτίμηση των παραμέτρων β_j γίνεται με τη χρήση της επαναληπτικής μεθόδου Newton-Raphson η οποία αναπτύσσεται στην επόμενη ενότητα. Η (2.17) με τη χρήση της (2.8) και της υπόθεσης ότι $a_i(\phi) = \frac{\phi}{w_i}$ γίνεται

$$u_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{w_i(y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)},\ j = 0, 1, \ldots, p.$$

Παρατηρούμε λοιπόν ότι το βάρος που δίνεται στην ποσότητα διαφορών $(y_i - \mu_i)$ κατά τον υπολογισμό των παραμέτρων β_j είναι

$$\frac{w_i x_{ij}}{\phi V(\mu_i)g'(\mu_i)},$$

με τα βάρη w_i γνωστά. Μεγαλύτερο w_i προσδίδει μεγαλύτερο βάρος και στη $(y_i - \mu_i)$. Επομένως,

$$\text{βάρος στην παρατήρηση } i \propto \frac{w_i}{\phi V(\mu_i)} = \frac{1}{V(Y_i)},$$

και αν η διασπορά είναι μεγάλη θα δοθεί μικρό βάρος στη διαφορά για την εκτίμηση των παραμέτρων β_j.

2.3.4 Μέθοδος Newton-Raphson

Οι εξισώσεις πρώτου βαθμού ως προς τα β_j (2.18), είναι αδύνατον να επιλυθούν αναλυτικά εκτός στην περίπτωση μεταβλητής απόκρισης της κανονικής κατανομής όπου η συνάρτηση σύνδεσης είναι η ταυτοτική. Έστω λοιπόν ότι η παράμετρος ϕ είναι γνωστή και η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας $\ell(\boldsymbol{\beta})$ γράφεται ως συνάρτηση της άγνωστης παραμέτρου $\boldsymbol{\beta}$. Έστω ότι $\beta \in \mathbb{R}$, τότε το τετραγωνικού βαθμού ανάπτυγμα Taylor της $\ell(\beta)$ είναι

$$\ell(\beta+\delta) \approx \ell(\beta) + \ell'(\beta)\delta + \frac{\delta^2}{2}\ell''(\beta).$$

Παραγωγίζοντας το δεξί μέλος ως προς δ και εξισώνοντας με το μηδέν προκύπτει ότι

$$\ell'(\beta) + \delta\ell''(\beta) = 0 \implies \delta = -(\ell''(\beta))^{-1}\ell'(\beta).$$

Επομένως, με δοσμένο β και δ, μία καλύτερη εκτίμηση του β θα είναι η

$$\beta - (\ell''(\beta))^{-1}\ell'(\beta).$$

Θεωρώντας ως β^m την τιμή του β στη m-οστή επανάληψη, το επαναληπτικό σχήμα έχει ως εξής:

$$\beta_{m+1} = \beta_m - (\ell''(\beta_m))^{-1}\ell'(\beta_m). \tag{2.20}$$

Η διαδικασία μπορεί να προσαρμοστεί και όταν $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$. Το επαναληπτικό σχήμα γίνεται

$$\boldsymbol{\beta}_{m+1} = \boldsymbol{\beta}_m - \boldsymbol{H}_m^{-1}\boldsymbol{u}_m, \tag{2.21}$$

με $\boldsymbol{\beta}_{m+1}$ τη νέα εκτίμηση του $\boldsymbol{\beta}$ ως συνάρτηση της τιμής $\boldsymbol{\beta}_m$ του προηγούμενου βήματος m της διαδικασίας καθώς και των τιμών των συναρτήσεων score $\boldsymbol{u'} = (u_0, u_1, \ldots, u_p)$ επίσης στο βήμα m, και του $k \times k$ Εσσιανού πίνακα $\boldsymbol{H}$ των δευτέρων μερικών παραγώγων του οποίου το jm-οστό στοιχείο ισούται με

$$\left.\frac{\partial^2 \ell}{\partial\beta_j \partial\beta_m}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_m}.$$

Για να επιτευχθεί το μέγιστο της συνάρτησης $\ell(\boldsymbol{\beta})$ θα πρέπει ο πίνακας $\boldsymbol{H}$ να είναι μη-θετικά ορισμένος, δηλαδή οι ποσότητες $\frac{\partial^2 \ell}{\partial\beta_j\partial\beta_m}$ να είναι αρνητικές. Με την επαναληπτική διαδικασία ενημέρωσης του $\boldsymbol{\beta}$ προκύπτει μία ακολουθία $(\boldsymbol{\beta}_n)_n$ η οποία και τελικά θα συγκλίνει. Η σχέση (2.21) γράφεται και ως

$$\boldsymbol{\beta}_{m+1} = \boldsymbol{\beta}_m - (\boldsymbol{\ell''}(\boldsymbol{\beta}_m))^{-1}\boldsymbol{\ell'}(\boldsymbol{\beta}_m), \tag{2.22}$$

με $\boldsymbol{\ell'}(\boldsymbol{\beta})$ το διάνυσμα στήλη των μερικών παραγωγών $\frac{\partial\ell}{\partial\beta_j}$, $j = 0, 1, \ldots, p$.

2.3.5 Μέθοδος Fisher Scoring

Εναλλακτικά της Newton Raphson, μπορεί να χρησιμοποιηθεί η μέθοδος Fisher Scoring, κατά την οποία ο πίνακας $(-\boldsymbol{H})$ αντικαταστείτε από τον πίνακα πληροφορίας $\boldsymbol{I}(\boldsymbol{\beta})$, των αρνητικά αναμενόμενων τιμών των δευτέρων μερικών παραγώγων της λογαριθμικής πιθανοφάνειας. Στην ουσία, ο πίνακας $\boldsymbol{H}$ αντικαταστείτε από την αναμενόμενη τιμή του $\mathbb{E}(\boldsymbol{H})$,

$$-\boldsymbol{H} \implies \boldsymbol{I}(\boldsymbol{\beta}) = -\mathbb{E}\left(\frac{\partial^2\ell}{\partial\beta_j\partial\beta_m}\right)$$

και

$$
\begin{aligned}
-\mathbb{E}\left(\frac{\partial^2 \ell}{\partial\beta_j \partial\beta_m}\right) &= \mathbb{E}\left(\sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j}\frac{\partial \ell_i}{\partial \beta_m}\right) \\
&= \sum_{i=1}^{n} \frac{\mathbb{E}(y_i-\mu_i)^2}{(V(Y_i))^2}\frac{x_{ij}x_{im}}{(g'(\mu_i))^2} \\
&= \sum_{i=1}^{n} \frac{x_{ij}x_{im}}{V(Y_i)(g'(\mu_i))^2} \\
&= [\boldsymbol{X'WX}]_{jm}, \quad j,m = 0,1,2,\ldots p,
\end{aligned} \tag{2.23}
$$

όπου $\boldsymbol{W} = \text{diag}(w_{ii})$ και

$$
w_{ii} = \frac{1}{V(Y_i)(g'(\mu_i))^2}, \quad i = 1,2,\ldots,n.
$$

Ο πίνακας $\boldsymbol{I(\beta)}$ καλείται πίνακας πληροφορίας Fisher. Αντικαθιστώντας στην (2.21) τον πίνακα $\boldsymbol{H}$ με την εκτίμηση του, $\boldsymbol{I(\beta)}$ στο m-οστό βήμα, προκύπτει ότι

$$
\begin{aligned}
[\boldsymbol{X'W_mX}]\boldsymbol{\beta_{m+1}} &= [\boldsymbol{X'W_mX}]\boldsymbol{\beta_m} + \boldsymbol{u_m} \\
&= [\boldsymbol{X'W_mz_m}],
\end{aligned} \tag{2.24}
$$

με το στοιχείο i του n-διάστατου διανύσματος $\boldsymbol{z_m}$ να ισούται με

$$
z_i = \sum_{j=0}^{p} \beta_{j,m}x_{ij} + (y_i-\mu_i)g'(\mu_i) \tag{2.25}
$$

και τα μ_i και $g'(\mu_i)$ υπολογισμένα στο $\boldsymbol{\beta} = \boldsymbol{\beta_m}$.

Οι μέθοδοι Fisher και Newton-Raphson για την εκτίμηση των παραμέτρων του μοντέλου, συγκλίνουν στην ΕΜΠ $\hat{\boldsymbol{\beta}}$. Οι δύο πίνακες $\boldsymbol{H}$ και $-\boldsymbol{I(\beta)}$ ταυτίζονται όταν $g(\mu_i) = \theta_i$ δηλαδή όταν η συνάρτηση σύνδεσης είναι η κανονική. Αν $g(\mu_i) = \theta_i$ τότε

$$
\frac{d\theta_i}{d\mu_i} = \frac{dg(\mu_i)}{d\mu_i} \implies \frac{1}{g'(\mu_i)} = \frac{d\mu_i}{d\theta_i} = b''(\theta_i) \equiv V(\mu_i)
$$

και ο πίνακας $\boldsymbol{G}$ δεν εξαρτάται από τις παρατηρήσεις y_i.

Σύμφωνα με τις ιδιότητες των ΕΜΠ, η εκτιμήτρια $\hat{\boldsymbol{\beta}}$ ακολουθεί ασυμπτωτικά την κανονική κατανομή και είναι αμερόληπτη με παρατηρούμενο πίνακα διασποράς-συνδιασποράς τον αντίστροφο του πίνακα πληροφορίας $\boldsymbol{I(\beta)}$, δηλαδή

$$
\hat{V}(\hat{\boldsymbol{\beta}}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta}) = (\boldsymbol{X'\hat{W}X})^{-1} \tag{2.26}
$$

και άρα

$$
\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, (\boldsymbol{X'WX})^{-1}).
$$

Ο πίνακας διασποράς-συνδιασποράς γράφεται και ως $\hat{\phi}(\boldsymbol{X'WX})^{-1}$ στη γενικότερη περίπτωση. Για παράδειγμα, στην Poisson κατανομή αποδείχτηκε ότι $\phi = 1$ ενώ στην Κανονική κατανομή $\phi = \sigma^2$.

2.3.6 Σταθμισμένα ελάχιστα τετράγωνα

Στην πολλαπλή γραμμική παλινδρόμηση όπου ισχύουν οι προϋποθέσεις της ομοσκεδαστικότητας και κανονικότητας της κατανομής των παρατηρήσεων, η εκτιμήτρια ελαχίστων τετραγώνων είναι η

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \tag{2.27}$$

συμπίπτει με την εκτιμήτρια μέγιστης πιθανοφάνειας και έχει την ελάχιστη διασπορά από όλες τις αμερόληπτες εκτιμήτριες. Σε αυτή την περίπτωση και όπως προϋποθέτει το γραμμικό μοντέλο, ο πίνακας διασποράς-συνδιασποράς των σφαλμάτων είναι διαγώνιος και ίσος με $V(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. Έστω ότι

$$V(\epsilon_i) = \sigma_i^2 = \sigma_{ii}^2,\ i = 1, 2, \ldots, n,$$

δηλαδή ο πίνακας $\boldsymbol{\epsilon}$ των σφαλμάτων είναι διαγώνιος, τα σφάλματα ασυσχέτιστα και η υπόθεση της ομοσκεδαστικότητας δεν ικανοποιείται. Όταν $\sigma_{ii} \to 0$ υπάρχει ακρίβεια ενώ όταν $\sigma_{ii} \to \infty$ η παρατήρηση είναι δεν ακριβής και λαμβάνει μεγάλη διασπορά. Τότε, στο γραμμικό μοντέλο

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

πολλαπλασιάζοντας με $\boldsymbol{\Sigma}^{-\frac{1}{2}} = \mathrm{diag}(\sigma_{ii}^{-\frac{1}{2}})$ και τα δύο μέλη προκύπτει ότι

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\epsilon} \tag{2.28}$$

και θέτοντας $\mathbf{u} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Y}$, $\mathbf{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}$ και $\boldsymbol{\delta} = \boldsymbol{\epsilon}$ καταλήγουμε στο μοντέλο

$$\mathbf{u} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\delta}. \tag{2.29}$$

Με τον πιο πάνω μετασχηματισμό αποκτήθηκε η επιθυμητή ομοσκεδαστικότητα των σφαλμάτων. Πράγματι, για τη διασπορά των σφαλμάτων $\boldsymbol{\delta}$ έχουμε:

$$\begin{aligned} V(\boldsymbol{\delta}) &= V(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\epsilon}) \\ &= \boldsymbol{\Sigma}^{-\frac{1}{2}}V(\boldsymbol{\epsilon})\boldsymbol{\Sigma}'^{-\frac{1}{2}} \\ &= \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}} \\ &= \boldsymbol{I}. \end{aligned}$$

Η εκτιμήτρια ελαχίστων τετραγώνων (2.27) του $\boldsymbol{\beta}$ γίνεται

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u} \tag{2.30}$$

και άρα επιστρέφοντας στις αρχικές μεταβλητές $\boldsymbol{X}$ και $\boldsymbol{Y}$ ισχύει ότι

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u} \\ &= [(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X})'\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}]^{-1}(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X})'\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Y} \\ &= (\boldsymbol{X}'\boldsymbol{\Sigma}^{\mathbf{1}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{Y}. \end{aligned} \tag{2.31}$$

Η εκτιμήτρια (2.31) καλείται εκτιμήτρια σταθμισμένων ελαχίστων τετραγώνων (weighted least squares estimator), είναι αμερόληπτη και είναι αυτή με την ελάχιστη διασπορά. Να σημειωθεί ότι ο

πίνακας $\mathbf{\Sigma}$ πρέπει να είναι γνωστός. Για την i παρατήρηση λοιπόν, το μοντέλο (2.29) διαμορφώνεται ως εξής

$$u_i = \beta_0 \frac{1}{\sqrt{\sigma_{ii}}} + \beta_1 z_{i1} + \ldots + \beta_k z_{ik} + \delta_i,$$

όπου

$$z_{ij} \equiv (\sqrt{\sigma_{ii}})^{-1} x_{ij},\ \delta_i \equiv (\sqrt{\sigma_{ii}})^{-1} \epsilon_i,$$

και u_i είναι ομοσκεδαστικές παρατηρήσεις. Αντίστοιχα, για τις ετεροσκεδαστικές παρατηρήσεις $(y_1, y_2, \ldots, y_n)$ μπορεί να θεωρηθεί ότι η

$$V(\epsilon_i) = \sigma_{ii}^2 = \frac{\sigma^2}{w_i}, \quad i = 1, 2, \ldots, n.$$

Δηλαδή η ετεροσκεδαστικότητα των σφαλμάτων προσαρμόζεται με ένα βάρος w_i σε κάθε σφάλμα και έτσι δε συμμετέχουν όλες οι παρατηρήσεις με το ίδιο βάρος στην εκτίμηση των παραμέτρων. Με τον μετασχηματισμό (2.28) καταφέραμε να πετύχουμε την επιθυμητή ομοσκεδαστικότητα των σφαλμάτων και έτσι οι εκτιμήσεις γίνονται όπως γίνονται στο πολλαπλό γραμμικό μοντέλο.

Πιο πάνω θεωρήθηκε ότι ο πίνακας διασποράς-συνδιασποράς των σφαλμάτων είναι διαγώνιος και άρα τα τυχαία σφάλματα ήταν ασυσχέτιστα μεταξύ τους. Γενικότερα, τα σφάλματα μπορεί να είναι συσχετισμένα, δηλαδή $\mathbb{E}(\epsilon_i \epsilon_j) \neq 0$. Σε αυτή την περίπτωση και με βάση τα πιο πάνω, η εκτιμήτρια του $\boldsymbol{\beta}$ δίνεται από τη σχέση

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'W^{-1}X})^{-1}\boldsymbol{X'W^{-1}Y}, \tag{2.32}$$

όπου $\mathbf{W}$ είνα ο πίνακας διασποράς-συνδιασποράς των τυχαίων σφαλμάτων, και ονομάζεται εκτιμήτρια γενικευμένων ελαχίστων τετραγώνων (generalized least squares estimator). Όμως σπάνια είναι γνωστός αυτός ο πίνακας και άρα η εφαρμογή της σχέσης (2.32) δεν είναι απλή.

Η σχέση (2.24) της μεθόδου Fisher θυμίζει κατά πολύ την εκτίμηση (2.32) του $\boldsymbol{\beta}$ της μεθόδου των σταθμισμένων ελαχίστων τετραγώνων. Για τον λόγο αυτό η μέθοδος Fisher scoring ονομάζεται και μέθοδος επαναληπτικών σταθμισμένων ελαχίστων τετραγώνων (iteratively reweighted least squares).

Επιπλέον, ας θεωρηθεί το ανάπτυγμα Taylor πρώτου βαθμού της συνάρτησης $g(y_i)$ γύρω από το y_i,

$$g(y_i) \approx g(\mu_i) + g'(\mu_i)(Y_i - \mu_i) \implies \boldsymbol{g(y) \approx g(\mu) + G(Y - \mu)}, \tag{2.33}$$

με $\boldsymbol{g(y)}$ διάνυσμα με στοιχεία $g(y_i)$. Άρα, αφού $\boldsymbol{g(\mu) = X\beta}$ η πιο πάνω σχέση γράφεται ως $\boldsymbol{G(Y - \mu) \approx g(y) - X\beta}$. Αντικαθιστώντας τη σχέση αυτή στη (2.19) προκύπτει η προσεγγιστική εκτίμηση για το $\boldsymbol{\beta}$

$$\boldsymbol{X'Wg(y) - X'WX\beta \approx 0} \implies \boldsymbol{\hat{\beta} \approx (X'WX)^{-1}X'Wg(y)}. \tag{2.34}$$

Επομένως, λύνοντας τη (2.19) προκύπτει και πάλι η εξίσωση (2.32) των σταθμισμένων ελαχίστων τετραγώνων με μετασχηματισμένες παρατηρήσεις $\boldsymbol{g(y_i)}$ με βάρη ανάλογα της διασποράς τους. Αυτό φαίνεται από την (2.33) όπου η διασπορά του $g(y_i)$ είναι ανάλογη του $(g'(\mu_i))^2 V(\mu_i)$. Στην περίπτωση όπου η συνάρτηση σύνδεσης είναι η ταυτοτική, τότε η (2.33) είναι ακριβής και άρα

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'WY},$$

όπου τώρα ο $\boldsymbol{W}$ είναι διαγώνιος με τιμές $\frac{1}{V(\mu_i)}$. Άρα, με την ταυτοτική συνάρτηση και με $V(\mu_i)$ ανεξάρτητο των μ_i, προκύπτει ότι η ΕΜΠ του $\boldsymbol{\beta}$ είναι η εκτιμήτρια σταθμισμένων ελαχίστων τετραγώνων.

2.4 Έλεγχος καλής προσαρμογής και σύγκριση μοντέλων

Η καταλληλότητα του μοντέλου είναι βασικό επακόλουθο κάθε στατιστικής μοντελοποίησης. Μας ενδιαφέρει ειδικότερα να μελετήσουμε κατά πόσο το μοντέλο περιγράφει σωστά τα διαθέσιμα δεδομένα. Προσθέτοντας μεταβλητές, μειώνονται τα σφάλματα $y_i - \hat{y}_i$, όμως αυξάνεται η διασπορά των $\hat{\beta}_j$. Η κεντρική ιδέα εδώ είναι η σχέση μεταξύ μεροληψίας και μεταβλητότητας (bias-variance tradeoff). Η μεροληψία (bias) αναφέρεται στη διαφορά μεταξύ παραμέτρου και της αναμενόμενης εκτίμησης της. Μεγάλη μεροληψία σημαίνει ακαταλληλότητα του μοντέλου ως προς την εκτίμηση των παρατηρούμενων τιμών. Από την άλλη, η μεταβλητότητα (variance) αφορά τη διασπορά των εκτιμήσεων και σχετίζεται με τη χρησιμότητα του μοντέλου σε ένα διαφορετικό δείγμα. Μεγάλη μεταβλητότητα σημαίνει πως το μοντέλο είναι ευαίσθητο σε μικρές αλλαγές των παρατηρήσεων και αδυνατεί να εκτιμήσει σωστά τις παραμέτρους.

2.4.1 Η ελεγχοσυνάρτηση Deviance

Για τη σύγκριση της καταλληλότητας δύο μοντέλων χρησιμοποιείται κυρίως η τεχνική του λόγου των πιθανοφανειών τους από τον οποίο ορίζεται η ελεγχοσυνάρτηση Deviance. Έστω ότι θέλουμε να ελέγξουμε τη μηδενική υπόθεση

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

έναντι της εναλλακτικής

$$H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0,$$

χρησιμοποιόντας τον λόγο των μεγίστων πιθανοφανειών (maximum likehood ratio test)

$$\lambda = \frac{\hat{L}_0}{\hat{L}_1} = \frac{L(\boldsymbol{\theta}_0, \boldsymbol{x})}{L(\hat{\boldsymbol{\theta}}, \boldsymbol{x})},$$

όπου $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{x})$, $L(\boldsymbol{\theta}, \boldsymbol{x})$ η συνάρτηση πιθανοφάνειας και $\boldsymbol{\theta} \in \mathbb{R}^k$ ο παραμετρικός χώρος του κάθε μοντέλου. Παρατηρήστε ότι $0 \leq \lambda \leq 1$. Επίσης, είναι γνωστό ότι

$$-2 \ln \lambda = -2(\ln \hat{L}_0 - \ln \hat{L}_1) = -2(\hat{\ell}_0 - \hat{\ell}_1) \dot{\sim} X^2_d \tag{2.35}$$

ασυμπτωτικά, με d τη διαφορά των παραμετρικών χώρων $\mathbb{R}^{k_i}$ των δύο υποθέσεων και $\hat{\ell}_i$ οι μεγιστοποιημένες λογαριθμικές συναρτήσεις πιθανοφάνειας των μοντέλων.

Ένας τρόπος να αξιολογηθεί η καταλληλότητα ενός μοντέλου είναι να συγκριθεί με το καταλληλότερο. Το καταλληλότερο μοντέλο θα είναι αυτό το οποίο έχει τόσες παραμέτρους όσες και παρατηρήσεις και αυτό ονομάζεται πλήρες ή κορεσμένο μοντέλο M_S (saturated model). Αυτό το μοντέλο έχει n παραμέτρους $\boldsymbol{\psi'} = (\psi_1, \psi_2, \ldots, \psi_n)$ και η συνάρτηση σύνδεσης $g(\mu_i)$ είναι η ταυτοτική αφού δεν εφαρμόζεται κάποιος περιορισμός τύπου $g(\mu_i) = \boldsymbol{x_i'\beta}$ στα δεδομένα. Επομένως, $\psi_i = y_i = \mathbb{E}(\mu_i)$ και έτσι οι προβλεπόμενες τιμές $\tilde{\mu}_i$ ισούνται με τις παρατηρούμενες, $\widetilde{\mu}_i = y_i$. Ο έλεγχος υποθέσεων με μηδενική υπόθεση

$$H_0 : \text{Ισχύει το μοντέλο } M_0 \text{ με } k_0 < n \text{ παραμέτρους,}$$

έναντι της εναλλακτικής

$$H_S : \text{Ισχύει το μοντέλο } M_S \text{ με } k_1 = n \text{ παραμέτρους,}$$

όπου n το μέγεθος του δείγματος και M_0 το μοντέλο που εξετάζεται. Η μεγιστοποιημένη λογαριθμική συνάρτηση πιθανοφάνειας στο υπό μελέτη μοντέλο M_0 είναι

$$\hat{\ell}_0 = \ell(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \left(\frac{y_i \hat{\theta}_i - b(\hat{\theta}_i)}{a_i(\phi)} + \ln\, c(y_i, \phi) \right),$$

ενώ στο κορεσμένο μοντέλο M_S είναι

$$\tilde{\ell}_S = \ell(\boldsymbol{y}, \tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \left(\frac{y_i \tilde{\theta}_i - b(\tilde{\theta}_i)}{a_i(\phi)} + \ln\, c(y_i, \phi) \right),$$

με $\tilde{\theta}$ και $\hat{\theta}$ οι εκτιμήσεις των θ_i υπό τις H_S και H_0 αντίστοιχα. Επομένως, προκύπτει ότι

$$-2(\hat{\ell}_0 - \tilde{\ell}_S) = -2 \sum_{i=1}^{n} (\hat{\ell}_{0i} - \tilde{\ell}_{Si}) = -2 \sum_{i=1}^{n} \frac{y_i(\hat{\theta}_i - \tilde{\theta}_i) - (b(\hat{\theta}_i) - b(\tilde{\theta}_i))}{a_i(\phi)}.$$

Η πιο πάνω ποσότητα είναι θετική αφού $\ell(\boldsymbol{y}, \tilde{\boldsymbol{\theta}}) \geq \ell(\boldsymbol{y}, \hat{\boldsymbol{\theta}})$. Ορίζεται η ελεγχοσυνάρτηση

$$\begin{aligned} D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) &= -2(\hat{\ell}_0 - \tilde{\ell}_S)\phi \\ &= \sum_{i=1}^{n} 2 w_i \left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i)) \right] \\ &= \sum_{i=1}^{n} d_i(y_i, \hat{\mu}_i) = \sum_{i=1}^{n} d_i(\hat{\boldsymbol{\beta}}) = D(\hat{\boldsymbol{\beta}}), \end{aligned} \tag{2.36}$$

η οποία καλείται ελεγχοσυνάρτηση Deviance, ενώ θεωρήθηκε ότι $a_i(\phi) = \frac{\phi}{w_i}$. Η $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ είναι ένα μέτρο απόστασης μεταξύ του προσαρμοσμένου μοντέλου M_0 και του κορεσμένου M_S. Όταν το υποψήφιο μοντέλο M_0 παρέχει καλή προσαρμογή τότε η τιμή της $\hat{\ell}_0$ θα προσεγγίζει τη τιμή της $\tilde{\ell}_S$ και άρα η $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ θα έχει μικρή τιμή. Σε αντίθετη περίπτωση, η προσαρμογή δε θεωρείται ικανοποιητική και το M_0 αποκλίνει από τα δεδομένα. Για να κριθεί αν τιμή της $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ θεωρείται στατιστικά μικρή ή μεγάλη, ορίζεται η τυποποιημένη ελεγχοσυνάρτηση Deviance (scaled Deviance) ως

$$D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \frac{D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})}{\phi} = \sum_{i=1}^{n} 2 \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)]}{a_i(\phi)},$$

για την οποία γνωρίζουμε ότι

$$D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) \quad \dot{\sim} \quad \chi^2_d,$$

ασυμπτωτικά υπό την H_0, δηλαδή όταν το υποψήφιο προσαρμοσμένο μοντέλο M_0 είναι ορθό. Η αναμενόμενη τιμή της τυποποιημένης ελεγχοσυνάρτησης Deviance είναι οι βαθμοί ελευθερίας της, d με

$$\mathbb{E}[D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}})] = d = n - k = n - (p + 1),$$

όπου p το πλήθος των επεξηγηματικών μεταβλητών του M_0. Τιμές της συνάρτησης $D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ που αποκλίνουν από την αναμενόμενη τιμή, συνιστούν κακή προσαρμογή του μοντέλου.

Η ασυμπτωτική συμπεριφορά της Deviance τη θέτει σε ένα αναξιόπιστο μέτρο καταλληλότητας αφού έχει ισχύ υπό κάποιες προϋποθέσεις [1]. Από την άλλη, η προσέγγιση της τυποποιημένης ελεγχοσυνάρτησης Deviance είναι πολύ καλύτερη όταν γίνεται σε μεταβολές των τιμών της, όπως στην επόμενη ενότητα. Παρατηρείστε επίσης ότι στην τυποποιημένη ελεγχοσυνάρτηση Deviance συμμετέχει και η παράμετρος μεταβλητότητας ϕ η οποία γενικά είναι άγνωστη. Η ασυμπτωτική προσέγγιση της Deviance είναι και σε αυτή την περίπτωση καλή.

2.4.2 Σύγκριση μοντέλων

Σκοπός είναι να βρεθεί το καταλληλότερο μοντέλο και όχι απλά να αξιολογηθεί η προσαρμογή ενός μεμονωμένου μοντέλου με τη χρήση της Deviance. Εφαρμόζονται λοιπόν τα πιο κάτω για τη σύγκριση δύο μοντέλων.

Έστω το μοντέλο M_0 που αντιστοιχεί στη μηδενική υπόθεση H_0 στο οποίο συμμετέχουν p_0 επεξηγηματικές μεταβλητές με τιμές $x_1, \ldots, x_{p_0}$. Άρα, καλούμαστε να εκτιμήσουμε $k_0 = p_0 + 1$ παραμέτρους, $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_{p_0})$. Έστω επίσης το μοντέλο M_1 της εναλλακτικής υπόθεσης H_1, με $M_0 \subset M_1$ στο οποίο συμμετέχουν οι p_0 συμμεταβλητές του M_0 και ακόμη d. Άρα, το πλήθος των επεξηγηματικών μεταβλητών του μοντέλου M_1 είναι $p_1 = d + p_0$. Πρόκειται λοιπόν για τη σύγκριση μεταξύ δύο εμφωλευμένων (nested) μοντέλων, όπου το M_0 προκύπτει από το M_1 αν σε αυτό θέσουμε $d = p_1 - p_0$ περιορισμούς της μορφής $\beta_j = 0$ για κάποια j με $j = 1, 2, \ldots, p_1$. Από την (2.35) προκύπτει ότι

$$-2(\hat{\ell}_0 - \hat{\ell}_1) \quad \dot{\sim} \quad \chi^2_d$$

ασυμπτωτικά, με $\hat{\ell}_0, \hat{\ell}_1$ οι μεγιστοποιημένες λογαριθμοποιημένες συναρτήσεις πιθανοφάνειας των δύο υπό μελέτη μοντέλων M_0 και M_1. Ορίζονται επίσης οι ελεγχοσυναρτήσεις deviance των μοντέλων M_0 και M_1 ως

$$D^*_0(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = -2(\hat{\ell}_0 - \tilde{\ell}_S)$$

και

$$D^*_1(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = -2(\hat{\ell}_1 - \tilde{\ell}_S),$$

αντίστοιχα. Ορίζεται η μεταβολή της deviance ως

$$D^*_0 - D^*_1 = -2(\hat{\ell}_0 - \tilde{\ell}_S) + 2(\hat{\ell}_1 - \tilde{\ell}_S) = -2(\hat{\ell}_0 - \hat{\ell}_1) \dot{\sim} \chi^2_d,$$

η οποία είναι πάντα θετική αφού το μοντέλο με τις περισσότερες μεταβλητές M_1 θα διαφέρει λιγότερο από το κορεσμένο σε σχέση με το M_0 και άρα θα έχει μικρότερη deviance, $D_0 > D_1$. Συνεπώς, το κατά πόσο η μεταβολή αυτή είναι μικρή ή μεγάλη εξετάζεται με βάση την κατανομή χ^2_d της οποίας η προσέγγιση θεωρείται καλύτερη από της deviance. Αν η τιμή κριθεί μεγάλη, τότε οι περιορισμοί $\beta_j = 0$ για κάποια j από τους οποίους προκύπτει το M_0, δεν ισχύουν και άρα απορρίπτεται η H_0. Η πιο πάνω ανάλυση είναι χρήσιμη όταν η παράμετρος ϕ είναι γνωστή και άρα η D^* μπορεί να υπολογιστεί. Σε αντίθετη περίπτωση η σύγκριση των μοντέλων μπορεί να γίνει ως εξής. Γνωρίζουμε ότι

$$D^*_0 - D^*_1 \dot{\sim} \chi^2_{p_1 - p_0}$$

και

$$D^*_1 \dot{\sim} \chi^2_{n-k_1}.$$

Άν τα $D^*_0 - D^*_1$ και D^*_1 είναι ανεξάρτητα, τότε

$$F = \frac{(D^*_0 - D^*_1)/(p_1 - p_0)}{D^*_1/(n - k_1)} \dot{\sim} F_{p_1 - p_0, n - k_1}.$$

Επομένως αφού $D^* = \frac{D}{\phi}$ προκύπτει ότι

$$F = \frac{(D_0 - D_1)/(p_1 - p_0)}{D_1/(n - k_1)} \dot{\sim} F_{p_1 - p_0, n - k_1},$$

το οποίο μπορεί να χρησιμοποιηθεί στην περίπτωση όπου η ϕ είναι άγνωστη.

2.4.3 Εκτίμηση της φ

Η αναμενόμενη τιμή τυχαίας μεταβλητής χ^2_{n-k} είναι $n-k$, επομένως αφού

$$D^* = \frac{D}{\phi},$$

προκύπτει μία εκτιμήτρια για το ϕ

$$\hat{\phi}_D = \frac{\hat{D}}{n-k},$$

αφού η αναμενόμενη τιμή $\mathbb{E}\left(\frac{D^*}{n-k}\right) = 1$. Μία δεύτερη εκτιμήτρια της ϕ βασίζεται στο στατιστικό Pearson το οποίο ορίζεται ως

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

και ακολουθεί χ^2_{n-k} κατανομή. Άρα

$$\frac{X^2}{\phi} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\phi V(\hat{\mu}_i)} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{w_i V(\hat{y}_i)},$$

και είναι άθροισμα κανονικών μεταβλητών με μηδενική μέση τιμή και μοναδιαία διασπορά. Θυμηθείτε ότι από το άθροισμα ανεξάρτητων και ισόνομων τυχαίων μεταβλητών Z_i με $Z_i \sim \mathcal{N}(0, \sigma^2)$ προκύπτει ότι $\sum_{i=1}^{n} Z_i^2 \sim X^2_{n-k}$. Επομένως, όπως προηγουμένως προκύπτει ότι

$$\hat{\phi}_{X^2} = \frac{X^2}{n-k},$$

αφού η αναμενόμενη τιμή του πιο πάνω αθροίσματος είναι $n-k$.

2.4.4 Έλεγχος Wald των συντελεστών β

Ο έλεγχος της στατιστικής σημαντικότητας των παραμέτρων β_j έχει μεγάλη σημασία για τον λόγο ότι σχετίζεται με το αν μία επεξηγηματική μεταβλητή χρειάζεται να συμμετάσχει ή όχι στο μοντέλο. Γενικότερα, μπορεί να γραφεί μία υπόθεση, δηλαδή ένας περιορισμός ως

$$\boldsymbol{C\beta} = \boldsymbol{r},$$

όπου $\boldsymbol{C}$ και $\boldsymbol{r}$ γνωστά στοιχεία από τα οποία προκύπτουν και οι περιορισμοί.

Έστω $\tilde{\boldsymbol{\beta}}$ η ΕΜΠ στο μοντέλο χωρίς περιορισμούς και $\hat{\boldsymbol{\beta}}$ η ΕΜΠ υπό τους περιορισμούς. Ο έλεγχος Wald μετρά πόσο αποκλίνει η τιμή $\boldsymbol{C}\tilde{\boldsymbol{\beta}}$ από το $\boldsymbol{r}$. Η μηδενική υπόθεση γράφεται ως

$$H_0 : \text{Μοντέλο με περιορισμούς } \boldsymbol{C\beta} = \boldsymbol{r}$$

και την εναλλακτική

$$H_1 : \text{Μοντέλο χωρίς περιορισμούς } \boldsymbol{C\beta} \neq \boldsymbol{r}.$$

Μικρή διαφορά $\boldsymbol{C}\tilde{\boldsymbol{\beta}} - \boldsymbol{r}$ σημαίνει πως υπάρχουν ενδείξεις εναντίον της μηδενικής υπόθεσης και των περιορισμών $\boldsymbol{C\beta} = \boldsymbol{r}$. Επίσης, είναι γνωστό ότι

$$\tilde{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, (\boldsymbol{X'\tilde{W}X})^{-1}),$$

άρα αφού $\boldsymbol{C}$ και $\boldsymbol{r}$ σταθερά, προκύπτει ότι

$$\boldsymbol{C}\tilde{\boldsymbol{\beta}} - \boldsymbol{r} \sim \mathcal{N}(\mathbf{0}, C(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{C}').$$

Επομένως, από γνωστό θεώρημα προκύπτει το στατιστικό Wald για τον έλεγχο της $\boldsymbol{C\beta} = \boldsymbol{r}$ ως

$$(\boldsymbol{C}\tilde{\boldsymbol{\beta}} - \boldsymbol{r})'(\boldsymbol{C}(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{C}')^{-1}(\boldsymbol{C}\tilde{\boldsymbol{\beta}} - \boldsymbol{r}) \dot{\sim} \chi^2_d. \tag{2.37}$$

Το πιο πάνω στατιστικό είναι η τετραγωνική απόσταση του $\boldsymbol{C}\tilde{\boldsymbol{\beta}}$ από το $\boldsymbol{r}$, χρησιμοποιώντας τον πίνακα διασποράς-συνδιασποράς του $\boldsymbol{C}\hat{\boldsymbol{\beta}}$ και μίας εκτίμησης του πίνακα $\boldsymbol{W}$.

Στην περίπτωση ελέγχου της τιμής μίας παραμέτρου β_j δεδομένου ότι οι υπόλοιπες β_q, $q \neq j, q = 0, \ldots, p$ υπάρχουν στο μοντέλο, ο έλεγχος Wald διαμορφώνεται ως εξής. Έστω ότι θα ελεγχθεί η υπόθεση $\beta_j = r$, τότε ο πίνακας $\boldsymbol{C}$ θα είναι πίνακας γραμμή με μηδενικά, εκτός του j-οστού στοιχείου. Ο όρος διασποράς-συνδιασποράς στην (2.37) γίνεται $(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}')^{-1}$ όπου είναι η διασπορά $\tilde{V}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{I}^{-1}(\tilde{\boldsymbol{\beta}}) = (\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X})^{-1}$. Επομένως, από γνωστό θεώρημα προκύπτει ότι

$$\frac{(\tilde{\beta}_j - r)^2}{\tilde{V}(\tilde{\boldsymbol{\beta}}_j)} \dot{\sim} \chi^2_1$$

και άρα για το στατιστικό Z προκύπτει ότι

$$Z = \frac{\tilde{\beta}_j - r}{(\boldsymbol{I}^{-1}(\tilde{\boldsymbol{\beta}})_{jj})^{1/2}} \dot{\sim} \mathcal{N}(0,1), \tag{2.38}$$

με $\boldsymbol{I}^{-1}(\tilde{\boldsymbol{\beta}})_{jj}$ το j-οστό διαγώνιο στοιχείο του παρατηρημένου πίνακα πληροφορίας $\boldsymbol{I}(\tilde{\boldsymbol{\beta}})$. Ο πιο πάνω έλεγχος μπορεί να χρησιμοποιηθεί προφανώς και για τη μηδενική υπόθεση $\beta_j = 0$ έναντι της εναλλακτικής $\beta_j \neq 0$ αλλά και για να κατασκευαστούν διαστήματα εμπιστοσύνης για τα β_j. Τέλος, να σημειωθεί πως για τον ολικό έλεγχο των παραμέτρων β_j εκτός του σταθερού όρου, ο $\boldsymbol{C}$ είναι ένας $p \times (p+1)$ πίνακας της μορφής

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ 0 & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}.$$

2.4.5 Έλεγχος Score

Ο έλεγχος Score σχετίζεται με την κλίση της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας ℓ στο $\hat{\boldsymbol{\beta}}$, την ΕΜΠ του μοντέλου υπό περιορισμούς. Όσο πιο μεγάλη είναι η κλίση, τόσο πιο πολύ μπορεί να κερδηθεί στην τιμή της ℓ μειώνοντας και άλλο (dropping) τους περιορισμούς. Έτσι, θα υπάρχουν ισχυρές ενδείξεις εναντίον της μηδενικής υπόθεσης $\boldsymbol{C\beta} = \boldsymbol{r}$. Όπως έχει αποδειχθεί, οι συναρτήσεις score σε μορφή πινάκων είναι οι

$$\boldsymbol{X}'\boldsymbol{W}\boldsymbol{G}(\boldsymbol{y} - \boldsymbol{\mu}) = \mathbf{0}.$$

Σημειώστε ότι ο πίνακας $\boldsymbol{X}$ έχει διαστάσεις $n \times q$, όπου q η διάσταση του παραμετρικού χώρου στο μοντέλο χωρίς περιορισμούς που περιέχει μεταβλητές που δεν περιέχονται στο περιορισμένο μοντέλο. Το $\mathbf{0}$ είναι διάνυσμα στήλη $q \times 1$. Έχουμε ότι ισχύει

$$\mathbb{E}[\boldsymbol{\ell}'(\boldsymbol{\beta})] = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{G}\,\mathbb{E}[\boldsymbol{Y} - \boldsymbol{\mu}] = \mathbf{0}$$

και άρα

$$V[\boldsymbol{\ell}'(\boldsymbol{\beta})] = \mathbb{E}[\boldsymbol{\ell}'(\boldsymbol{\beta})(\boldsymbol{\ell}'(\boldsymbol{\beta}))^{\mathrm{T}}] = \boldsymbol{X'WX},$$

με $\boldsymbol{\ell}'(\boldsymbol{\beta})$ το διάνυσμα των μερικών παραγώγων $\frac{\partial \ell}{\partial \beta_j}$, $j = 0, 1, \ldots, p$. Επομένως, κατασκευάζεται το στατιστικό score

$$S = \boldsymbol{\ell}'(\hat{\boldsymbol{\beta}})^{\mathrm{T}} \big[V(\boldsymbol{\ell}'(\hat{\boldsymbol{\beta}}))\big]^{-1} \boldsymbol{\ell}'(\hat{\boldsymbol{\beta}}) \dot{\sim} \chi^2_d, \tag{2.39}$$

με $\boldsymbol{\ell}'(\hat{\boldsymbol{\beta}}) = \boldsymbol{X'\hat{W}G}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$. Στην πράξη, το W και άρα και το ϕ θα αντικατασταθούν από εκτιμήσεις τους και τότε η κατανομή αυτή θα είναι ακόμη πιο προσεγγιστική. Άρα αρχικά το μοντέλο προσαρμόζεται με βάση τους περιορισμούς $\boldsymbol{C\beta} = \boldsymbol{r}$ και προκύπτει το διάνυσμα των υπολοίπων $\boldsymbol{y} - \hat{\boldsymbol{\mu}}$. Τα υπόλοιπα αυτά συνδιάζονται στην (2.39) μαζί με τον πίνακα X ο οποίος περιέχει όλες τις μεταβλητές, ακόμη και αυτές που έχουν μηδενικούς συντελεστές υπό τον περιορισμό $\boldsymbol{C\beta} = \boldsymbol{r}$. Μεγάλη τιμή στο στατιστικό score είναι ένδειξη εναντίον της μηδενικής υπόθεσης με πιθανότητα απόρριψης την $\mathbb{P}(\chi^2_{d,1-a} > S)$, σε επίπεδο σημαντικότητας a.

2.5 Διαγνωστικές Μεθόδοι

Στο πολλαπλό γραμμικό μοντέλο οι διαγνωστικές μέθοδοι επικεντρώνονται στον έλεγχο της κανονικότητας και της ομοσκεδαστικότητας των υπολοίπων. Αν και οι υποθέσεις των Γενικευμένων γραμμικών μοντέλων είναι διαφορετικές από των πολλαπλών γραμμικών μοντέλων, ο σκοπός των διαγνωστικών ελέγχων είναι ο ίδιος και είναι η ανίχνευση ενδείξεων εναντίον των υποθέσεων. Οι έλεγχοι αυτοί βασίζονται στα υπόλοιπα, τα οποία εκφράζουν την καταλληλότητα του μοντέλου ως προς τη συμφωνία παρατηρούμενων και προσαρμοσμένων τιμών. Τα υπόλοιπα $\epsilon_i = y_i - \hat{y}_i$ δεν έχουν σταθερή διασπορά και έτσι προτιμάται να τυποποιούνται έτσι ώστε αν οι υποθέσεις του μοντέλου είναι σωστές αυτά να έχουν κοινή διασπορά και να συμπεριφέρονται όσο γίνεται όπως τα υπόλοιπα του πολλαπλού γραμμικού μοντέλου. Υπάρχουν τρία είδη υπολοίπων.

2.5.1 Υπόλοιπα

1. **Υπόλοιπα Pearson:** Έστω τα συνήθη υπόλοιπα

$$\epsilon_i = y_i - \hat{y}_i = y_i - \hat{\mu}_i,\ i = 1, \ldots, n.$$

Η διασπορά των Y_i είναι

$$V(Y_i) = a_i(\phi) b''(\theta_i),$$

με

$$b'(\theta_i) = \mu_i$$

και άρα

$$\hat{V}(Y_i) = a_i(\hat{\phi}) b'(\hat{\mu}_i).$$

Επομένως, τα υπόλοιπα δεν είναι συγκρίσιμα αφού δεν έχουν κοινή διασπορά ακόμα και αν έχει επιλεγεί σταθερή συνάρτηση

$$a_i(\phi) = a(\phi),\ \forall\, i = 1, \ldots, n.$$

Ίση διασπορά όλων των τυχαίων μεταβλητών απόκρισης συμβαίνει μονό στην περίπτωση της κανονικής κατανομής στην οποία

$$b''(\theta_i) = \left(\frac{\theta^2}{2}\right)' = 1$$

και άρα

$$V(Y_i) = \sigma^2.$$

Εισάγεται λοιπόν η έννοια των υπολοίπων Pearson, κατά την οποία τα συνήθη υπόλοιπα διαιρούνται με τις εκτιμήσεις των συναρτήσεων διασπορών $V(\hat{\mu}_i)$ και έτσι προκύπτουν ως

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}},\ i = 1, \ldots, n.$$

Η ονομασία των υπολοίπων προέρχεται από το γεγονός ότι το άθροισμα των τετραγώνων τους ισούται με την ελεγχοσυνάρτηση Pearson για τον έλεγχο χ^2 καλής προσαρμογής

$$\boldsymbol{X}^2 = \sum_{i=1}^{n} \left(r_i^P\right)^2.$$

Τα υπόλοιπα Pearson αν και διαιρούνται με $\sqrt{V(\hat{\mu}_i)}$, η διασπορά τους παραμένει μη σταθερή και είναι ανάλογη της $a_i(\hat{\phi})$. Θα ήταν πιο σωστό να διαιρεθούν με το τυπικό σφάλμα των συνήθων υπολοίπων $se(y_i - \hat{\mu}_i)$ το οποίο αποδεικνύεται πως ισούται με

$$\sqrt{V(\hat{\mu}_i)(1 - \hat{h}_{ii})},$$

[1], με $\hat{h}_{ii}$ να είναι το i-οστό στοιχείο του πίνακα προβολής

$$\hat{\boldsymbol{H}} = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}^{1/2},$$

ο οποίος απεικονίζει τα y_i στα $\hat{\mu}_i$ και έχει προκύψει από την αντικατάσταση του πίνακα $\boldsymbol{X}$ με $\boldsymbol{W}^{1/2}\boldsymbol{X}$, όπου $\boldsymbol{W}$ όπως αυτός ορίστηκε στην (2.19) στη μορφή του πίνακα προβολής πολλαπλών γραμμικών μοντέλων. Ορίζονται λοιπόν τα τυποποιημένα (standaralized) υπόλοιπα Pearson ως

$$\begin{aligned} r_i^{PS} &= \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1 - h_{ii})}} \\ &= \frac{y_i - \hat{\mu}_i}{\sqrt{a_i(\hat{\phi})V(\hat{y}_i)(1 - \hat{h}_{ii})}},\ i = 1, \ldots, n, \end{aligned} \tag{2.40}$$

με ασυμπτωτική διασπορά $V(r_i^{PS}) = 1$.

2. **Υπόλοιπα Deviance:** Στην πράξη τα υπόλοιπα Pearson είναι αρκετά ασύμμετρα γύρω από το μηδέν και άρα η συμπεριφορά τους δεν είναι ανάλογη των υπολοίπων του πολλαπλού γραμμικού μοντέλου. Σε αυτή την περίπτωση, τα υπόλοιπα Deviance είναι προτιμότερα. Θεωρήστε την ελεγχοσυνάρτηση scaled Deviance όπως αυτή ορίστηκε στην ενότητα 2.4.1 με

$$D^*(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} d_i(y_i, \hat{\mu}_i) = \sum_{i=1}^{n} 2\frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))}{a_i(\phi)}.$$

Τα υπόλοιπα deviance ορίζονται ως

$$r_i^D = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{d_i(y_i, \hat{\mu}_i)} = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{d_i(\hat{\boldsymbol{\beta}})},$$

μία ποσότητα που αυξάνεται με το $(y_i - \hat{\mu}_i)$, όπου $d_i(\hat{\boldsymbol{\beta}})$ η προσφορά της παρατήρησης i στην ελεγχοσυνάρτηση και $\text{sgn}(y_i - \hat{\mu}_i)$ το πρόσημο της διαφοράς $(y_i - \hat{\mu}_i)$. Προφανώς, η Deviance $D^*(\hat{\boldsymbol{\beta}})$ γράφεται και ως

$$D^*(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} d_i(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} (r_i^D)^2,$$

το άθροισμα των τετραγώνων των υπολοίπων όπως στην περίπτωση των υπολοίπων Pearson. Αν το υπόλοιπο της παρατήρησης i είναι μεγάλο, θα έχει και μεγάλη προσφορά στη Deviance, ένδειξη ότι υπάρχει πρόβλημα με τη συγκεκριμένη παρατήρηση. Αν η Deviance έχει υπολογιστεί σε ένα μοντέλο όπου όλες οι παράμετροι είναι γνωστές, τότε

$$D^*(\hat{\boldsymbol{\beta}}) \sim \chi_n^2$$

και αυτό υποδεικνύει ότι για μία παρατήρηση i η

$$d_i(\hat{\boldsymbol{\beta}}) \sim \chi_1^2$$

και άρα το υπόλοιπο

$$r_i^D \sim \mathcal{N}(0, 1).$$

Επομένως, η πιο πάνω προσέγγιση υποδεικνύει ότι αναμένεται τα υπόλοιπα να συμπεριφέρονται σαν κανονικές μεταβλητές. Ορίζονται ομοίως τα τυποποιημένα υπόλοιπα Deviance ως

$$r_i^{DS} = \frac{r_i^D}{\sqrt{1 - \hat{h}_{ii}}}.$$

3. **Υπόλοιπα Anscombe:** Η ανάγκη των υπολοίπων Anscombe φαίνεται στα Γενικευμένα γραμμικά μοντέλα όπου η μεταβλητή απόκρισης δεν ακολουθεί συχνά την κανονική κατανομή. Η κατανομή των υπολοίπων Pearson έχει συνήθως μεγάλες ουρές (skewed) και άρα δεν έχει τις επιθυμητές ιδιότητες της κανονικότητας. Εισάγεται λοιπόν η έννοια των υπολοίπων Anscombe που βασίζονται στη συνάρτηση h η οποία επιλέγεται έτσι ώστε η κατανομή της $h(y)$ να είναι προσεγγιστικά κανονική. Η συνάρτηση αυτή είναι η

$$h(u) = \int \frac{1}{V^{1/3}(u)}\, du,$$

[9] και επομένως τα υπόλοιπα βασίζονται στις διαφορές

$$h(y_i) - h(\hat{\mu}_i).$$

Παρολ' αυτά, ο μετασχηματισμός αυτός δε σταθεροποιεί τη διασπορά και άρα τα υπόλοιπα θα πρέπει να διαιρεθούν με την τυπική απόκλιση της $h(y)$ η οποία σε πρώτο βαθμό δίνεται από την

$$h'(y)\sqrt{V(y)}.$$

Άρα, τα υπόλοιπα Anscombe ορίζονται ως

$$\frac{h(y_i) - h(\hat{\mu}_i)}{h'(\hat{\mu}_i)\sqrt{V(\hat{\mu}_i)}},$$

ακολουθώντας προσεγγιστικά κανονική κατανομή, αν έχει επιλεγεί σωστή συνάρτηση κατανομής της μεταβλητής απόκρισης. Αν και τα υπόλοιπα Anscombe και Deviance είναι τυπικά διαφορετικά, αριθμητικά είναι σχεδόν ίδια. Επομένως, τα υπόλοιπα Deviance ακολουθούν προσεγγιστικά την κανονική κατανομή όπως και τα Anscombe [1].

Χρήση των υπολοίπων

Τα διάφορα είδη υπολοίπων χρησιμοποιούνται για να ελεγχθεί με τη βοήθεια γραφημάτων η καταλληλότητα του μοντέλου. Ο πιο απλός τρόπος είναι να κατασκευαστεί ένα γράφημα δείκτη (index plot) των υπολοίπων σε σχέση με τη σειρά των παρατηρήσεων στο δείγμα. Ασυνήθιστα μεγάλα υπόλοιπα υποδεικνύουν ότι το μοντέλο δεν είναι ικανοποιητικό. Το ίδιο γράφημα μπορεί να χρησιμοποιηθεί για την εξέταση συσχέτισης μεταξύ των υπολοίπων όταν αυτά παρατηρούνται σε χρονική σειρά. Αυξητική τάση των υπολοίπων δείχνει συσχέτιση μεταξύ τους και συσσώρευση των σφαλμάτων. Επίσης, χρήσιμο είναι το γράφημα των υπολοίπων έναντι των προσαρμοσμένων τιμών μ_i. Σε αυτό το γράφημα αναμένονται τα υπόλοιπα να είναι ομοιόμορφα κατανεμημένα γύρω από το μηδέν. Αν υπάρχει κάποια τάση στη μέση τιμή των υπολοίπων τότε παραβιάζεται η υπόθεση της ανεξαρτησίας των παρατηρήσεων και αυτό προκύπτει από την κατασκευή του μοντέλου. Η τάση αυτή ίσως είναι αποτέλεσμα λάθος μετασχηματισμού ή μη συμπερίληψης κάποιας μεταβλητής. Τα δύο πιο πάνω γραφήματα χρησιμεύουν και στον εντοπισμό έκτροπων (outliers) τιμών στα δεδομένα. Τέλος, ένας άλλος γραφικός έλεγχος για τα υπόλοιπα είναι η κατασκευή Q-Q διαγραμμάτων κατά την οποία θα εξετάζεται η προσέγγιση της Κανονικής κατανομής των υπολοίπων. Αν και τα υπόλοιπα δεν ακολουθούν κανονική κατανομή, τα γραφήματα αυτά χρησιμοποιούνται σαν ένδειξη καλής προσαρμογής του μοντέλου. Θεωρητικά, έχουμε κατασκευάσει υπόλοιπα που ακολουθούν Κανονική κατανομή όπως τα υπόλοιπα Deviance και άρα ελέγχουμε τη θεώρηση αυτή. Τιμές που ξεφεύγουν από την ευθεία της Κανονικής κατανομής υποδεικνύουν πιθανή κακή προσαρμογή και ακατάλληλη επιλογή κατανομής απόκρισης.

2.5.2 Μερικά υπόλοιπα

Όπως και στο πολλαπλό γραμμικό μοντέλο, η χρήση των μερικών υπολοίπων είναι σημαντική για να εξεταστεί αν μία επεξηγηματική μεταβλητή χρειάζεται στο μοντέλο όπως είναι ή αν πρέπει να μετασχηματιστεί. Τα μερικά υπόλοιπα για τη μεταβλητή X ορίζονται ως

$$r_{ji}^T = (y_i - \hat{\mu}_i)g'(\mu_i) + \hat{\beta}_j x_{ij},\ i = 1, \ldots, n,$$

με $\hat{\beta}_j$ ο εκτιμώμενος συντελεστής της μεταβλητής X_j. Ο πρώτος όρος αποτελεί την προσέγγιση Taylor πρώτου βαθμού όπως φαίνεται στην εξίσωση 2.33. Στην περίπτωση όπου έχει επιλεγεί κανονική συνάρτηση σύνδεσης ισχύει ότι $g(\mu_i) = \theta_i \implies g'(\mu_i) \sim \frac{1}{V(\mu_i)}$ και άρα ο πρώτος όρος συμπεριφέρεται κανονικά. Η σχέση των μερικών υπολοίπων είναι αντίστοιχη της (2.25) και είναι το υπόλοιπο προσθέτοντας την προσφορά της επεξηγηματικής μεταβλητής X_j. Το διάγραμμα των πιο πάνω ποσοτήτων έναντι των x_{ij}, $i = 1, \ldots, n$ αποτελεί το διάγραμμα των μερικών υπολοίπων και χρησιμοποιείται για την εξέταση της γραμμικότητας μεταξύ της X_j και της μεταβλητής απόκρισης

Y. Αν το γράφημα δεν παρουσιάζει ευθεία τάση τότε η μεταβλητή X_j θα πρέπει να τεθεί εκτός μοντέλου ή να μετασχηματιστεί. Σε διαφορετική περίπτωση, το γράφημα θα έχει ευθεία τάση με κλίση $\hat{\beta}_j$ λόγω της κανονικότητας του πρώτου όρου και άρα η μεταβλητή X_j παραμένει στο μοντέλο χωρίς να χρειάζεται κάποιο μετασχηματισμό. Επομένως, το διάγραμμα των μερικών υπολοίπων είναι επίσης σημαντικό και κατασκευάζεται για κάθε επεξηγηματική μεταβλητή του μοντέλου.

2.5.3 Πρόσθετες μεταβλητές

Το διάγραμμα των πρόσθετων μεταβλητών χρειάζεται επίσης για να κριθεί αν μία μεταβλητή $\boldsymbol{Z}$ πρέπει να προστεθεί στο μοντέλο ή όχι. Αρχικά, προσαρμόζεται το μοντέλο με τις συμμεταβλητές $X_1, X_2, \ldots, X_p$ και προσδιορίζονται τα υπόλοιπα Pearson όπως αυτά έχουν ορισθεί. Προσδιορίζονται λοιπόν τα υπόλοιπα του μοντέλου με μεταβλητή απόκρισης την $\boldsymbol{Y}$ και επεξηγηματικές μεταβλητές τις $X_1, X_2, \ldots, X_p$. Στη συνέχεια υπολογίζονται τα υπόλοιπα Pearson του μοντέλου με μεταβλητή απόκρισης την $\boldsymbol{Z}$ και επεξηγηματικές μεταβλητές τις $X_1, X_2, \ldots, X_p$. Αυτά θα έχουν τη μορφή

$$(\boldsymbol{I} - \hat{\boldsymbol{H}})\hat{\boldsymbol{W}}^{1/2}\boldsymbol{Z},$$

[11], όπου $\hat{\boldsymbol{H}}$ ο εκτιμημένος πίνακας προβολής $\boldsymbol{H}$

$$\hat{\boldsymbol{H}} = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}^{1/2}$$

και $\boldsymbol{W}$ ο πίνακας με στοιχεία $\frac{1}{V(Y_i)(g'(\mu_i)^2}$ και $\mathbf{Z}$ ο $n \times 1$ πίνακας γραμμή των τιμών της υποψήφιας μεταβλητής Z. Αν το γράφημα δεν παρουσιάζει κάποια τάση αλλά τυχαίες τιμές, τότε η μεταβλητή $\boldsymbol{Z}$ δε χρειάζεται στο μοντέλο. Αν υπάρχει ευθεία τάση η μεταβλητή χρειάζεται, ενώ αν η τάση δεν είναι ευθεία τότε χρειάζεται μετασχηματισμός.

2.5.4 Επιρροή

Η επιρροή αναφέρεται σε συγκεκριμένες παρατηρήσεις του δείγματος οι οποίες ασκούν μεγάλη επιρροή στην εκτίμηση των παραμέτρων του μοντέλου. Οι παρατηρήσεις αυτές ονομάζονται σημεία ή παρατηρήσεις επιρροής (influential observations) και η παρουσία τους ή μη επηρεάζει αισθητά τα αποτελέσματα. Η επιρροή είναι ένα μέτρο το οποίο σχετίζεται με την απόσταση των

$$\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}},$$

με $\hat{\boldsymbol{\beta}}_i$ η εκτίμηση του $\boldsymbol{\beta}$ αφού έχει παραλειφθεί η παρατήρηση i και $\hat{\boldsymbol{\beta}}$ όταν χρησιμοποιείται όλο το δείγμα. Ένα πολύ χρήσιμο μέτρο για τον εντοπισμό σημείων των οποίων η αφαίρεση θα επηρεάσει τις εκτιμήσεις των παραμέτρων είναι η απόσταση Cook. Χρησιμοποιείται λοιπόν η στατιστική συνάρτηση του Cook

$$CD_i = \frac{1}{k\phi}\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}\right)\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}\right), \; i = 1, \ldots, n,$$

με $\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}$ ο παρατηρούμενος πίνακας πληροφορίας $\boldsymbol{I}(\hat{\boldsymbol{\beta}})$, $\hat{\phi}$ η εκτιμώμενη παράμετρος μεταβλητότητας και k το πλήθος των παραμέτρων στο μοντέλο [1]. Το πιο πάνω στατιστικό μπορεί να γραφεί και ως

$$CD_i = \frac{\hat{h}_{ii}(r_i^{PS})^2}{k(1-\hat{h}_{ii})},$$

[10] με $\hat{h}_{ii}$ το i-οστό διαγώνιο στοιχείο του πίνακα προβολής $\hat{\boldsymbol{H}}$ που ορίστηκε στην ενότητα 2.5.1.

Η στατιστική συνάρτηση Cook σχετίζεται με τη διαφορά των $\hat{\boldsymbol{\beta}}_i$ και $\hat{\boldsymbol{\beta}}$ όπως αυτή ορίζεται στον παραμετρικό χώρο όπου ανήκουν. Επομένως, είναι μία συνάρτηση που δεν μπορεί εύκολα να ανιχνεύσει τη μεταβολή ενός μόνο στοιχείου β_j του διανύσματος $\boldsymbol{\beta}$, από την παράλειψη κάποιας παρατήρησης i ειδικά σε μεγάλες διαστάσεις. Παρουσιάζει λοιπόν ενδιαφέρον να μελετηθεί η επιρροή που ασκεί η παρατήρηση i στην εκτίμηση καθεμίας παραμέτρου β_j και αυτή η μεταβολή μπορεί να υπολογιστεί από τη σχέση

$$\Delta_i\hat{\beta}_j = \frac{\left(\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}_{j+1}\boldsymbol{x}_i(y_i-\mu_i)}{(1-\hat{h}_{ii})\,\mathrm{se}(\hat{\beta}_j)},\ i=1,\ldots,n,\ j=0,\ldots,p,$$

με $\left(\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}_{j+1}$ είναι η $(j+1)$-οστή γραμμή του πίνακα διασποράς-συνδιασποράς του $\hat{\beta}$ και $\boldsymbol{x}_i$ το διάνυσμα τιμών της i-οστής παρατήρησης. Η πιο πάνω συνάρτηση ονομάζεται δέλτα-βήτα (delta-beta) [11]. Με τη χρήση της σχέσης αυτής μπορούν να κατασκευαστούν $p+1$ γραφήματα δείκτη, ένα για κάθε παράμετρο, στα οποία θα απεικονίζεται η τιμή της συνάρτησης για κάθε παρατήρηση i. Μεγάλες τιμές δείχνουν ποιες παρατηρήσεις ασκούν επιρροή στην εκτίμηση του υπό εξέταση συντελεστή β_j.

2.6 Επιλογή μοντέλου

Όπως έχει αναφερθεί στην Ενότητα 2.4.1 η ελεγχοσυνάρτηση Deviance είναι ένα χρήσιμο μέτρο για να συγκριθούν εμφωλευμένα μοντέλα. Όταν όμως τα μοντέλα δεν είναι εμφωλευμένα, ο έλεγχος χ^2 δεν εφαρμόζεται. Υπάρχει λοιπόν η ανάγκη να βρεθεί ένα μέτρο σύγκρισης μεταξύ τέτοιων μοντέλων. Η συνάρτηση μεγιστοποιημένης λογαριθμικής πιθανοφάνειας $\ell(\boldsymbol{y},\hat{\boldsymbol{\theta}})$ από μόνη της δεν μπορεί να παρέχει ένα μέτρο σύγκρισης μεταξύ δύο μοντέλων, αλλά γενικότερα όσο μεγαλύτερη είναι τόσο το καλύτερο. Αυτό διότι μπορεί τα μοντέλα να μην έχουν τον ίδιο αριθμό επεξηγηματικών μεταβλητών και κατά πάσα πιθανότητα το μοντέλο με τις περισσότερες θα έχει την υψηλότερη πιθανοφάνεια. Για να επιλυθεί το πρόβλημα αυτό, χρησιμοποιούνται οι δύο πιο κάτω δείκτες.

2.6.1 Δείκτες καλής προσαρμογής AIC και BIC

Όπως και στο πολλαπλό γραμμικό μοντέλο, χρησιμοποιείται το κριτήριο επιλογής AIC (Akaike's information criterion) το οποίο και ποινικοποιεί την εισαγωγή πολλών παραμέτρων στο μοντέλο. Στη γενική περίπτωση ορίζεται από τη σχέση

$$\mathrm{AIC} = 2k - 2\ln\ell(\boldsymbol{y},\hat{\boldsymbol{\theta}}),$$

με $\ell(\boldsymbol{y},\hat{\boldsymbol{\theta}})$ η μεγιστοποιημένη λογαριθμική συνάρτηση πιθανοφάνειας και k ο αριθμός των παραμέτρων του υπό εξέταση μοντέλου. Προτιμότερο είναι το μοντέλο με το μικρότερο AIC. Αν και η εισαγωγή επιπλέον μεταβλητών στο μοντέλο αυξάνει την πιθανοφάνεια, η τιμή του AIC μειώνεται όταν η εισαγωγή τους βελτιώνει την προσαρμογή του μοντέλου σε βαθμό που να υπερβαίνει το βάρος του πρώτου όρου ποινής. Επομένως, επιλέγοντας το χαμηλότερο AIC πετυχαίνεται η εξισορρόπηση της πολυπλοκότητας και προσαρμογής του μοντέλου. Ορίζεται επίσης ο δείκτη BIC (Bayesian information criterion) ο οποίος αποθαρρύνει περισσότερο την εισαγωγή επιπρόσθετων παραμέτρων στο μοντέλο. Στη γενική περίπτωση ορίζεται από τη σχέση

$$\mathrm{BIC} = k\ln n - 2\ln\ell(\boldsymbol{y},\hat{\boldsymbol{\theta}}).$$

Κεφάλαιο 3

Μοντελοποίηση Δεδομένων Μέτρησης

Όταν η μεταβλητή απόκρισης αφορά τον αριθμό εμφάνισης κάποιου γεγονότος στον χώρο ή τον χρόνο τότε οι παρατηρήσεις είναι δεδομένα μέτρησης. Αυτά μπορούν να μοντελοποιηθούν μέσω διακριτών κατανομών. Σε αυτή την ενότητα θα παρουσιαστούν τρία μοντέλα με τη χρήση διαφορετικών κατανομών που αποσκοπούν στη μοντελοποίηση τέτοιων μεταβλητών.

Έστω η τυχαία μεταβλητή Y με βάση την οποία κατασκευάζεται το τυχαίο δείγμα $Y_1, Y_2, \ldots, Y_n$. Η τυχαία μεταβλητή Y θα αφορά δεδομένα μέτρησης και συγκεκριμένα θα είναι ο αριθμός απαιτήσεων στον οποίο έχει προβεί ένας ασφαλιζόμενος κατά τη διάρκειας κάποιας περιόδου w. Η διάρκεια w_i ονομάζεται έκθεση και αντιπροσωπεύει τον χρόνο όπου ο i-ασφαλιζόμενος βρίσκεται σε έκθεση στον κίνδυνο. Σε αυτή τη διπλωματική εργασία γίνεται μελέτη μονοετών συμβολαίων. Σε ένα δείγμα όμως δεν είναι πάντα δυνατόν οι παρατηρήσεις να είναι μονοετούς έκθεσης αφού αρκετοί ασφαλιζόμενοι διακόπτουν το συμβόλαιο τους πρόωρα. Επομένως, θα πρέπει στο μοντέλο η προσαρμογή να γίνει με βάση τη διάρκεια w που ήταν τελικά εκτεθειμένος ο ασφαλιζόμενος, ώστε η εκτιμήσεις να γίνουν υπό την ίδια αναφορά.

3.1 Poisson Μοντέλο

Το δημοφιλέστερο μοντέλο για τη μοντελοποίηση δεδομένων μέτρησης είναι το Poisson. Σε αυτό, η μεταβλητή απόκρισης ακολουθεί την κατανομή Poisson. Το μοντέλο γράφεται ως

$$Y_i \sim \text{Poisson}(\mu_i)\,, g(\mu_i) = \boldsymbol{x_i'\beta},$$

με συγκεκριμένη επιλογή της συνάρτησης σύνδεσης $g(\mu_i)$. Οι επεξηγηματικές μεταβλητές X_j με τιμές x_{ij} χαρακτηρίζουν το προφίλ ρίσκου του i ασφαλιζόμενου. Η αναμενόμενη τιμή μ_i αυτής της κατανομής είναι υποχρεωτικά θετική και άρα όπως εξηγήθηκε στην Ενότητα 2.3.1 χρησιμοποιείται η λογαριθμική συνάρτηση σύνδεσης

$$g(\mu_i) = \ln \mu_i,$$

που είναι τέτοια ώστε

$$\ln \mu_i : (0, \infty) \to (-\infty, \infty),$$

και έτσι η αναμενόμενη τιμή

$$\mu_i = e^{\boldsymbol{x_i'\beta}},$$

λαμβάνει μόνο θετικές τιμές.

Έστω Y_i ο αριθμός απαιτήσεων που προέβει ο i ασφαλιζόμενος σε διάρκεια w_i. Η διάρκεια w_i είναι ο συνολικός χρόνος (exposure) που παρέμεινε ενεργό το i-οστό συμβόλαιο κατά τη διάρκεια ενός έτους. Θεωρείται ως $w_i = \frac{\text{διάρκεια σε ημέρες}}{365 \text{ μέρες}}$ και άρα τιμή ίση με 1 θα αφορά ένα ολόκληρο έτος. Θεωρούμαι ότι η τυχαία μεταβλητή Y_i ακολουθεί Poisson κατανομή με παράμετρο $w_i\mu$, δηλαδή $Y_i \sim P(w_i\mu)$. Παρατηρείστε ότι για έκθεση ενός έτους $w_i = 1$ ο αναμενόμενος αριθμός απαιτήσεων είναι μ και άρα σταθερός στο τυχαίο δείγμα. Η εκτίμηση της μ γίνεται με τη μέθοδο της μέγιστης πιθανοφάνειας. Η συνάρτηση πιθανοφάνειας έχει τη μορφή

$$L(y_i, \mu, w_i) = \prod_{i=1}^{n} \frac{e^{-w_i\mu}(w_i\mu)^{y_i}}{y_i!},$$

και άρα ο λογάριθμος της γράφεται ως

$$\ell(y_i, \mu, w_i) = -\mu \sum_{i=1}^{n} w_i + \sum_{i=1}^{n} y_i \ln(\mu w_i) - \ln \prod_{i=1}^{n} y_i!,$$

με παράγωγο

$$\frac{\partial \ell(y_i, \mu, w_i)}{\partial \mu} = 0 \implies -\sum_{i=1}^{n} w_i + \frac{1}{\mu} \sum_{i=1}^{n} y_i = 0$$

και άρα

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} w_i} = \sum_{i=1}^{n} z_i \frac{y_i}{w_i},$$

ο δειγματικός μέσος της συχνότητας απαιτήσεων με

$$z_i = \frac{w_i}{\sum_{i=1}^{n} w_i}.$$

Σκοπός όμως είναι να κατασκευαστεί ένα μοντέλο το οποίο θα εκτιμά τον αναμενόμενο αριθμό απαιτήσεων ξεχωριστά για κάθε ασφαλιζόμενο ο οποίος φέρει διαφορετικό ρίσκο. Επομένως, η πιο πάνω ανάλυση δεν είναι κατάλληλη. Έστω πάλι Y_i ο αριθμός απαιτήσεων σε διάρκεια w_i και έστω μ_i η αναμενόμενη τιμή της Y_i όταν $w_i = 1$. Άρα, $\mathbb{E}(Y_i) = w_i\mu_i$ και η Y_i ακολουθεί κατανομή Poisson με παράμετρο $w_i\mu_i$. Η συνάρτηση μάζας πιθανότητας της Y_i είναι

$$f_{Y_i}(y_i, \mu_i) = \frac{(w_i\mu_i)^{y_i}}{y_i!} e^{-w_i\mu_i}, \quad y_i = 0, 1, 2, \ldots.$$

Επομένως, με τη μέθοδο μέγιστης πιθανοφάνειας γίνεται η εκτίμηση των συντελεστών β_j του μοντέλου. Η λογαριθμική πιθανοφάνεια θα έχει τη μορφή

$$\ell(y_i, \mu_i) = \sum_{i=1}^{n} y_i \ln w_i\mu_i - w_i\mu_i - \ln(y_i!)$$

και αφού χρησιμοποιείται η λογαριθμική συνάρτηση σύνδεσης θα ισχύει ότι

$$\mu_i = \exp\left(\sum_{i=0}^{p} x'_{ij}\beta_j\right),$$

το οποίο αν αντικατασταθεί στη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας προκύπτει ότι

$$\ell(y_i, \mu_i) = \sum_{i=1}^{n} y_i \left(\ln w_i + \sum_{i=0}^{p} x'_{ij}\beta_j\right) - w_i \exp\left(\sum_{i=0}^{p} x'_{ij}\beta_j\right) - \ln(y_i!).$$

Παραγωγίζοντας ως προς κάθε β_j προκύπτει ότι

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} y_i x'_{ij} - w_i x'_{ij} \exp\left(\sum_{i=0}^{p} x'_{ij}\beta_j\right) = \sum_{i=1}^{n} (y_i w_i - \mu_i) x'_{ij}.$$

Θέτοντας τις $p+1$ μερικές παραγώγους μηδέν, λαμβάνεται ότι οι εκτιμήτριες μέγιστης πιθανοφάνειας είναι οι εξισώσεις

$$\sum_{i=1}^{n} (y_i w_i - \mu_i) x'_{ij} = 0, \quad j = 0, 1, 2, \ldots, p,$$

οι οποίες λύνονται αριθμητικά. Επομένως, το Poisson μοντέλο γράφεται ως

$$\ln \mathbb{E}(Y_i) = \boldsymbol{x'_i \beta} + \ln w_i,$$

με τον όρο $\ln w_i$ να ονομάζεται offset. Άρα, ο αναμενόμενος αριθμός απαιτήσεων σε διάστημα w_i είναι

$$\mathbb{E}(Y_i) = w_i e^{\beta_0} e^{\beta_1 x_{i1}} \ldots e^{\beta_p x_{ip}},$$

και αν $w_i = 1$ τότε γίνεται κάλυψη ενός ολόκληρου έτους.

Κατά τη μοντελοποίηση με τη χρήση της κατανομής Poisson, συχνά εμφανίζεται το φαινόμενο της ύπαρξης μεγαλύτερης διασποράς στα δεδομένα από τις θεωρητικές τιμές $V(Y_i) = \mathbb{E}(Y_i)$ της κατανομής. Αυτό ορίζεται ως υπερμεταβλητότητα (overdispersion) και μπορεί να επιλυθεί με χρήση πιο γενικών μοντέλων. Η κατανομή Poisson χαρακτηρίζεται από υπομεταβλητότητα και έτσι η χρήση της πολλές φορές δεν είναι κατάλληλη [4]. Με τη χρήση των ΕΜΠ των β_j, είναι δυνατό να διορθωθεί η διασπορά της κατανομής Poisson πολλαπλασιάζοντας τη με την παράμετρο ϕ έτσι ώστε

$$V_{\text{over.}}(Y_i) = \phi w_i \mu_i$$

και να λαμβάνονται αυξημένες διασπορές. Η παράμετρος ϕ μπορεί να εκτιμηθεί μέσω της σχέσης

$$\hat{\phi} = \frac{\sum_{i=1}^{n} (y_i - w_i \hat{\mu}_i)^2}{\sum_{i=1}^{n} w_i \hat{\mu}_i}.$$

Με αυτή την τροποποίηση και τη χρήση της συνάρτησης διασποράς των παραμέτρων β_j, 2.26, αποδεικνύεται ότι

$$V_{\text{over.}}(\hat{\boldsymbol{\beta}}) = \phi V(\hat{\boldsymbol{\beta}}),$$

και προκύπτει ότι

$$V_{\text{over.}}(\hat{\boldsymbol{\beta}}) = \phi(\boldsymbol{X'\hat{W}X})^{-1} = \phi \left[\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}'_i w_i \hat{\mu}_i\right]^{-1},$$

αφού $g'(\hat{\mu}_i) = \frac{1}{\hat{\mu}_i}$ και άρα τα στοιχεία του πίνακα $\hat{\mathbf{W}}$ είναι $w_i \hat{\mu}_i$. Η ποσότητα αυτή εμφανίζεται σε αγκύλες αφού αποτελεί έναν τετραγωνικό $k \times k$ διαγώνιο πίνακα. Ο πίνακας $\boldsymbol{X}$ έχει διαστάσεις $n \times k$ και ο $\boldsymbol{W}$ $n \times n$.

3.2 Αρνητικό Διωνυμικό Μοντέλο

Αντί να γενικευθεί η κατανομή Poisson προσθέτοντας τον όρο μεταβλητότητας φ, είναι δυνατό να κατασκευαστούν νέες κατανομές οι οποίες θα ανταποκρίνονται στη μεταβλητότητα των δεδομένων.

Θεωρούμαι ότι η επιπρόσθετη μεταβλητότητα προκαλείται από την αγνόηση κάποιων σοβαρών μεταβλητών από το μοντέλο. Στην περίπτωση της μοντελοποίησης της συχνότητας απαιτήσεων, η κατανάλωση ουσιών από τον οδηγό θεωρείτε ένας τέτοιος παράγοντας. Επίσης, η υπερβολική αυτοπεποίθηση ή η αποδυνάμωση των οδικών ανταναακλαστικών είναι παράγοντες οι οποίοι δε λαμβάνονται υπόψη, αλλά σχετίζονται με τη συμπεριφορά του ασφαλιζόμενου ενώ δεν ποσοτικοποιούνται κατά τη μοντελοποίηση. Είναι παράγοντες των οποίων η επίδραση φαίνεται έμμεσα από τις επεξηγηματικές μεταβλητές του μοντέλου. Κατασκευάζονται επομένως σύνθετες κατανομές Poisson [2].

Αν Y είναι ο αριθμός των απαιτήσεων, τότε η Y μπορεί να μοντελοποιηθεί μέσω κατανομής Poisson, όπως πριν. Επιλέγουμε παράμετρο λ για την κατανομή αυτή. Έστω όμως ότι η Λ είναι μία συνεχής τυχαία θετική μεταβλητή η οποία επηρεάζει τη μεταβλητότητα, με συνάρτηση πυκνότητας πιθανότητας $f_\Lambda(\lambda)$. Τότε, δεδομένου του λ, η $Y \,|\, \lambda \sim \text{Poisson}(\lambda)$ και η συνάρτηση μάζας πιθανότητας της Y είναι

$$\begin{aligned} f_Y(y) &= \int_0^\infty f(y \,|\, \lambda)\, f_\Lambda(\lambda)\, d\lambda \\ &= \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!}\, f_\Lambda(\lambda)\, d\lambda, \quad y = 0, 1, 2, \ldots . \end{aligned} \tag{3.1}$$

Κατανομές της μορφής 3.1 ονομάζονται σύνθετες κατανομές Poisson (compound poisson distributions) και έχουν την ιδιότητα της ανομοιογένειας με την έννοια ότι η δεσμευμένη μέση τιμή της Y είναι και αυτή μία τυχαία μεταβλητή. Η κατανομή $f_\Lambda(\lambda)$ ονομάζεται κατανομή μίξης και μία συνήθης επιλογή για αυτή είναι η Γάμμα. Αν η μεταβλητή $\Lambda \sim \text{Gamma}(\mu, \nu)$, με μέση τιμή μ και παράμετρο ρυθμού ν, τότε η συνάρτηση πυκνότητας πιθανότητας της είναι

$$f_\Lambda(\lambda, \mu, \nu) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\lambda\right)^\nu \lambda^{-1} e^{-\frac{\nu}{\mu}\lambda},$$

οπότε προκύπτει ότι

$$\begin{aligned} f_Y(y) &= \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!} \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\lambda\right)^\nu \lambda^{-1} e^{-\frac{\nu}{\mu}\lambda}\, d\lambda \\ &= \frac{1}{y!\,\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^\nu \int_0^\infty \lambda^{y+\nu-1} e^{-\lambda\left(1+\frac{\nu}{\mu}\right)}\, d\lambda \\ &= \frac{1}{y!\,\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^\nu \int_0^\infty \left(\frac{u}{1+\frac{\nu}{\mu}}\right)^{y+\nu-1} e^{-u} \cdot \frac{\mu}{\mu+\nu}\, du \\ &= \frac{1}{y!\,\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^\nu \frac{\mu}{\mu+\nu} \int_0^\infty \left(\frac{\mu u}{\mu+\nu}\right)^{y+\nu-1} e^{-u}\, du \\ &= \frac{1}{y!\,\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^\nu \left(\frac{\mu}{\mu+\nu}\right)^{y+\nu} \Gamma(y+\nu) \\ &= \frac{\Gamma(y+\nu)}{y!\,\Gamma(\nu)} \left(\frac{\nu}{\nu+\mu}\right)^\nu \left(\frac{\mu}{\mu+\nu}\right)^y, \quad y = 0, 1, 2, \ldots, \end{aligned} \tag{3.2}$$

με τη χρήση του μετασχηματισμού

$$u = \lambda\left(1 + \frac{\nu}{\mu}\right).$$

Εάν θεωρηθεί $k = \frac{1}{\nu}$ τότε η 3.2 είναι η συνάρτηση πυκνότητας πιθανότητας Αρνητικής διωνυμικής κατανομής με παραμέτρους (μ, κ), όπως στη 2.11 με

$$\mathbb{E}(Y) = \mu \quad \text{και} \quad V(Y) = \mu(1 + \kappa\mu).$$

Επομένως, η Αρνητική διωνυμική κατανομή προκύπτει όταν διαφορετικά προφίλ ρίσκου δηλαδή παρατηρήσεις, χαρακτηρίζονται από ξεχωριστές Poisson κατανομές των οποίων η μέσες τιμές είναι κατανεμημένες με βάση τη Γάμμα κατανομή. Η παράμετρος κ ονομάζεται παράμετρος υπερμεταβλητότητας διότι μέσω αυτής προστίθεται μεταβλητότητα στη θεωρητική κατανομή. Όταν $\kappa \to 0$ τότε η $\mathrm{NB}(\mu, \kappa)$ προσεγγίζει την $\mathrm{P}(\mu)$ και όταν $\kappa = 0$ τότε $\mathbb{E}(Y) = \mathrm{V}(Y)$, άρα δεν υπάρχει επιπρόσθετη μεταβλητότητα. Η διασπορά είναι $V(Y) = \mu(1 + \kappa\mu) = \mu + \kappa\mu^2$ (2.13), με τον τετραγωνικό όρο $\kappa\mu^2$ η κατανομή ονομάζεται NB2 και μπορεί να ανταποκριθεί στην υπερμεταβλητότα.

Γενικότερα, μπορεί να κατασκευαστεί η NBp μορφή της Αρνητικής διωνυμικής κατανομής ανάλογα με την επιλογή της κατανομής μίξης. Η διασπορά τότε είναι

$$V(Y) = \mu + \kappa\mu^p,$$

όπου μ η μέση τιμή της κατανομής μίξης και p που προκύπτει από την επιλογή. Θέτοντας $g(\mu) = \mu^p$, η κατανομή Poisson μπορεί να συγκριθεί με κάποια σύνθετη κατανομή μέσω της συνάρτησης διασποράς της, $\mu + \kappa\mu^p$. Επομένως, η μηδενική υπόθεση $H_0 : \kappa = 0$ ελέγχεται εναντίον της $H_1 : \kappa > 0$ με τη χρήση του στατιστικού T το οποίο χρησιμοποιείται για τον έλεγχο οποιασδήποτε σύνθετης κατανομής Poisson έναντι της Poisson. Το στατιστικό αυτό δίνεται από τη σχέση

$$T = \left(\sum_{i=1}^{n} \frac{1}{2} \hat{\mu}_i^{-2} g^2(\hat{\mu}_i) \right)^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{1}{2} \hat{\mu}_i^{-2} g(\hat{\mu}_i)[(y_i - \hat{\mu}_i)^2 - y_i]$$

και ακολουθεί τυποποιημένη κανονική κατανομή κάτω από τη μηδενική υπόθεση. Η υλοποίηση του ελέγχου γίνεται μέσω της εντολής `dispersiontest()` της βιβλιοθήκης `AER` [4].

Επομένως, στην περίπτωση όπου γίνεται μοντελοποίηση του αριθμού των απαιτήσεων, τότε απαιτείται να γίνει η διόρθωση με τη χρήση offset, όπως έγινε και στο Poisson μοντέλο. Θεωρώντας ότι η μέση τιμή μ_i είναι ο αναμενόμενος αριθμός απαιτήσεων που προέβει ο i ασφαλιζόμενος σε διάρκεια ενός έτους, τότε ο αναμενόμενος αριθμός απαιτήσεων σε διάρκεια w_i θα είναι $\mu_i w_i$. Αν στην παραπάνω ανάλυση θεωρηθεί $\mu_i w_i$ ως η μέση τιμή της Γάμμα κατανομής, τότε προκύπτει ότι

$$\mathbb{E}(Y_i) = \mu_i w_i \quad \text{και} \quad V(Y_i) = \mu_i w_i (1 + \kappa \mu_i w_i),$$

με το κ να θεωρείται σταθερό στο δείγμα. Το Αρνητικό διωνυμικό μοντέλο γράφεται όπως το Poisson, με χρήση λογαριθμικής συνάρτησης σύνδεσης και του όρου offset ως

$$\ln \mathbb{E}(Y_i) = \boldsymbol{x_i'\beta} + \ln w_i.$$

3.3 Μοντέλο Μηδενικής Διόγκωσης

Η παραβίαση της ισότητας μεταξύ μέσης τιμής και διασποράς σε ένα Poisson μοντέλο είναι σχεδόν βέβαιη σε πραγματικά δεδομένα θέτοντας το μοντέλο ακατάλληλο. Επίσης, η υπερμεταβλητότητα των πραγματικών δεδομένων απαιτεί την εύρεση ενός εναλλακτικού μοντέλου. Αν η μεταβλητή απόκρισης λαμβάνει πολλές μηδενικές τιμές, ένα Poisson μοντέλο δε θα ήταν το καταλληλότερο. Επομένως, είναι αναγκαίο να βρεθεί η κατάλληλη κατανομή η οποία θα ανταποκρίνεται στην υπερμεταβλητότητα των δεδομένων και στις μηδενικές τιμές και αυτή είναι η κατανομή Μηδενικής διόγκωσης.

Πρόκειται για τον συνδυασμό δύο διαφορετικών κατανομών. Υποθέτετε ότι οι παρατηρήσεις της μεταβλητής απόκρισης Y ανήκουν σε δύο διαφορετικές ομάδες. Στην πρώτη, ανήκουν οι παρατηρήσεις που με πιθανότητα ένα έχουν μηδενική μέτρηση και στη δεύτερη οι παρατηρήσεις με μέγεθος

που περιγράφεται από μία κατανομή μέτρησης, Poisson ή Αρνητική διωνυμική η οποία μπορεί επίσης να μετρήσει μηδενικά. Επομένως, μηδενική παρατήρηση μπορεί να προέλθει και από τις δύο ομάδες ενώ αν είναι από την πρώτη, τότε είναι ανεξάρτητη της πιθανότητας που προσδίδει η κατανομή μέτρησης. Με αυτό τον τρόπο δεν προέρχονται όλα τα μηδενικά από την κατανομή μέτρησης αλλά και από την πρώτη ομάδα. Αυτό σημαίνει ότι κατανέμεται μεγαλύτερη μάζα πιθανότητας στο μηδέν σε σχέση με τη χρήση μόνο της κατανομής μέτρησης.

Έστω π η πιθανότητα η παρατήρηση i να προέρχεται από την πρώτη ομάδα. Τότε η πιθανότητα η παρατήρηση i να προέρχεται από τη δεύτερή ομάδα είναι $1-\pi$. Η συνάρτηση μάζας πιθανότητας για τη σύνθετη κατανομή δίνεται από την

$$\mathbb{P}(Y_i = k) = \begin{cases} \pi + (1-\pi)\mathbb{P}(V_i = 0), & \text{εάν } k = 0 \\ (1-\pi)\mathbb{P}(V_i = k), & \text{εάν } k = 1, 2, \ldots, \end{cases} \tag{3.3}$$

όπου η τυχαία μεταβλητή V_i ακολουθεί μία κατανομή μέτρησης. Το πρώτο σκέλος της συνάρτησης μάζας πιθανότητας είναι η πιθανότητα μηδενικής παρατήρησης και άρα προέρχεται και από τις δύο ομάδες. Το δεύτερο σκέλος αφορά θετικές μετρήσεις η οποίες προέρχονται μόνο από τη δεύτερη ομάδα. Παρατηρείστε ότι στην οριακή περίπτωση όπου $\pi \to 0$ τότε η κατανομή Μηδενικής διόγκωσης εκφυλίζεται στην κατανομή μέτρησης. Μεγάλη τιμή της π σημαίνει πως πολλές παρατηρήσεις ανήκουν στην πρώτη ομάδα. Η φύση του μοντέλου είναι τέτοια ώστε να κατανέμει λιγότερη από την κατανομή μέτρησης μάζα σε θετικές μετρήσεις και αυτό λόγο του όρου $(1-\pi)\mathbb{P}(V_i = k)$ με $0 \leq \pi \leq 1$.

Μπορεί επίσης η πιθανότητα μία παρατήρηση να ανήκει στην πρώτη ομάδα, να μην είναι σταθερή μεταξύ των παρατηρήσεων. Έτσι, για κάθε παρατήρηση i η συνάρτηση μάζας πιθανότητας είναι η

$$\mathbb{P}(Y_i = k) = \begin{cases} \pi_i + (1-\pi_i)\mathbb{P}(V_i = 0), & \text{εάν } k = 0 \\ (1-\pi_i)\mathbb{P}(V_i = k), & \text{εάν } k = 1, 2, \ldots, \end{cases}$$

με π_i η πιθανότητα η παρατήρηση i να προέρχεται από την πρώτη ομάδα. Η πιθανότητα π_i μπορεί να μοντελοποιηθεί μέσω Λογιστικής παλινδρόμησης και της logit συνάρτησης σύνδεσης

$$\ln \frac{\pi_i}{1-\pi_i} = \mathbf{q_i}'\boldsymbol{\gamma},$$

με $\mathbf{q_i}$ οι τιμές των επεξηγηματικών μεταβλητών που θα χρησιμοποιηθούν και $\boldsymbol{\gamma}$ οι αντίστοιχοι συντελεστές τους. Οι επεξηγηματικές μεταβλητές αυτές μπορούν να είναι διαφορετικές από αυτές που θα επιλεγούν για τη μοντελοποίηση της κατανομής μέτρησης.

Αν χρησιμοποιηθεί ως κατανομή μέτρησης η Poisson με παράμετρο μ_i, τότε προκύπτει το Poisson μοντέλο Μηδενικής διόγκωσης (Zero-infated Poisson) το οποίο έχει τη μορφή

$$\mathbb{P}(Y_i = k) = \begin{cases} \pi_i + (1-\pi_i)e^{-\mu_i}, & \text{εάν } k = 0 \\ (1-\pi_i)\frac{e^{-\mu_i}\mu_i^k}{k!}, & \text{εάν } k = 1, 2, \ldots. \end{cases}$$

Από το πιο πάνω μπορούν να υπολογιστούν οι δύο πρώτες ροπές της κατανομής Μηδενικής διόγκωσης της μεταβλητής Y. Επομένως,

$$\mathbb{E}(Y_i) = 0 + (1-\pi_i)\mathbb{E}\left(\frac{e^{-\mu_i}\mu_i^k}{k!}\right) = (1-\pi_i)\mu_i$$

και

$$V(Y_i) = \pi_i\mu_i + \pi_i\mu_i^2(1-\pi_i) = (1-\pi_i)\mu_i(1+\mu_i\pi_i).$$

Επομένως, η διασπορά της κατανομής είναι μεγαλύτερη της μέσης τιμής και άρα ανταποκρίνεται στην υπερμεταβλητότητα.

Το μοντέλο αυτό, μπορεί επομένως να χρησιμοποιηθεί για τη μοντελοποίηση της συχνότητας των απαιτήσεων. Ως πιθανότητα π θεωρείται η πιθανότητα ένας ασφαλιζόμενος να ανήκει στην πρώτη ομάδα που σίγουρα θα έχει μηδενικές απαιτήσεις. Όπως και στο Poisson μοντέλο μπορεί να γίνει η διόρθωση με τη χρήση της έκθεσης w. Αυτή μπορεί να γίνει και στην κατανομή μέτρησης και στη λογιστική παλινδρόμηση. Για την πρώτη περίπτωση η διόρθωση γίνεται στην παράμετρο της κατανομής μέτρησης ως $w_i\mu_i$ αν θεωρήσουμε ότι η μ είναι η μέση τιμή στη διάρκεια ενός έτους. Στο λογιστικό μοντέλο η διόρθωση γίνεται στην πιθανότητα π. Συγκεκριμένα, με βάση τη λογιστική παλινδρόμηση υπάρχει τυχαία μεταβλητή έστω Z η οποία λαμβάνει τιμή 1 σε ενδεχόμενο επιτυχίας και 0 αλλιώς, με πιθανότητα π και $1-\pi$ αντίστοιχα. Η μεταβλητή Z είναι η μεταβλητή απόκρισης του λογιστικού μοντέλου και θα ακολουθεί Bernoulli κατανομή με παράμετρο π. Το μοντέλο γράφεται ως

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{q_i'}\boldsymbol{\gamma}$$

και άρα η διόρθωση γίνεται στο μοντέλο ως εξής:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{q_i'}\boldsymbol{\gamma} + \ln\,\mathbf{w_i}.$$

Η πιθανότητα π_i είναι επομένως ίση με

$$\frac{w_i e^{\mathbf{q_i'}\boldsymbol{\gamma}}}{1+w_i e^{\mathbf{q_i'}\boldsymbol{\gamma}}},$$

ενώ για ένα ολόκληρο έτος είναι

$$\frac{e^{\mathbf{q_i'}\boldsymbol{\gamma}}}{1+e^{\mathbf{q_i'}\boldsymbol{\gamma}}}.$$

Κεφάλαιο 4

Ανάλυση των Δεδομένων

Στο κεφάλαιο αυτό γίνεται η προβολή και η ανάλυση των δεδομένων του αρχείου `Motor vehicle insurance data`. Η ανάλυση θα γίνει με τη χρήση του στατιστικού πακέτου `R` όπως και η κατασκευή όλων των διαγραμμάτων. Ξεκινώντας με την εισαγωγή των δεδομένων στην `R`, θα παρουσιαστούν τα κυριότερα χαρακτηριστικά του δείγματος μέσω της περιγραφικής στατιστικής. Η διαδικασία αυτή θα δώσει το έναυσμα για την έναρξη της μοντελοποίησης της συχνότητας των απαιτήσεων. Αυτή θα γίνει με τη χρήση διαφορετικών κατανομών και της κατασκευής τριών τύπων μοντέλων τα οποία θα εξεταστούν ως προς την προσαρμογή και την ακρίβεια τους.

4.1 Μεταφορά δεδομένων

Με τη χρήση της βιβλιοθήκης `readxl` διαβάζεται το αρχείο δεδομένων `Motor vehicle insurance data` και αποθηκεύεται ως πλαίσιο δεδομένων στη μεταβλητή `data`.

```
install.packages("readxl")
library(readxl)
data <- read_excel("C:/Downloads/Car_Insurance_Dataset.xlsx")
data <- as.data.frame(data)
```

4.2 Περιγραφή των μεταβλητών

Υπάρχουν 105,555 εγγραφές και 27 μεταβλητές. Οι μεταβλητές μπορούν να χωριστούν ανάλογα με το πού αναφέρονται σε 5 κατηγορίες: στα χαρακτηριστικά του οδηγού, του οχήματος, του συμβολαίου, σε γεωγραφικές μεταβλητές και στις μεταβλητές απόκρισης όπως φαίνεται στον Πίνακα 4.2. Στον Πίνακα 4.1 γίνεται η περιγραφή των διαθέσιμων και επιπρόσθετων μεταβλητών.

	Μεταβλητή	Τύπος	Περιγραφή
1	`ID`	Ποσοτική	Ο κωδικός ταυτοποίησης του κάθε ασφαλιζόμενου. Η μεταβλητή αυτή μπορεί να επαναλαμβάνετε αρκετές φορές υποδεικνύοντας ότι ο ασφαλιζόμενος έχει συνάψει ένα νέο μονοετές συμβόλαιο, δηλαδή έχει ανανεώσει.

2	Date.start.contract	Ημερομηνία	Ημερομηνία έναρξης κάποιου συμβολαίου χωρίς αυτή να είναι απαραίτητα στο διάστημα 2015 με 2018
3	Date.birth	Ημερομηνία	Ημερομηνία γέννησης ασφαλιζόμενου
4	Date.driving.licence	Ημερομηνία	Ημερομηνία απόκτησης άδειας οδηγού
5	Year.matriculation	Ποσοτική	Χρονιά εγγραφής του οχήματος
6	Power	Ποσοτική	Δύναμη του οχήματος σε άλογα
7	Cylinder.capacity	Ποσοτική	Χωρητικότητα του κυλίνδρου
8	Value.vehicle	Ποσοτική	Αξία οχήματος όπως αυτή ορίζεται την 31 Δεκεμβρίου 2019 σε απροσδιόριστη χρηματική μονάδα
9	N.doors	Κατηγορική	Αριθμός των πορτών του οχήματος
10	Length	Ποσοτική	Μήκος του οχήματος σε μέτρα
11	Weight	Ποσοτική	Βάρος του οχήματος σε χιλιόγραμμα
12	Type.fuel	Κατηγορική	P για βενζινοκίνητο όχημα, D για πετρελαιοκίνητο
13	Date.last.renewal	Ημερομηνία	Ημερομηνία έναρξης συμβολαίου
14	Date.next.renewal	Ημερομηνία	Ημερομηνία λήξης συμβολαίου
15	Distribution.channel	Κατηγορική	0, αν το συμβόλαιο έγινε μέσω πράκτορα, 1 αν έγινε μέσω μεσίτη ασφαλίσεων ή 00/01/1900
16	Seniority	Ποσοτική	Χρόνια τα οποία ο ασφαλιζόμενος είναι συνδεδεμένος με την εταιρία
17	Policies.in.force	Ποσοτική	Αριθμός υπηρεσιών ανά συμβόλαιο
18	Lapse	Ποσοτική	Προσμετρά τον αριθμό των συμβολαίων που έχουν διακοπεί στην περίοδο κάλυψης
19	Date.lapse	Ημερομηνία	Ημερομηνία διακοπής του συμβολαίου
20	Payment	Κατηγορική, δύτιμη	0, αν η αποπληρωμή του ασφαλίστρου γίνεται ετήσια και 1, για εκ του μισού πληρωμή
21	Premium	Ποσοτική	Ασφάλιστρο σε άγνωστες χρηματικές μονάδες
22	Type.risk	Κατηγορική	1, αν ο τύπος οχήματος είναι μοτοσικλέτα, 2 για φορτηγό, 3 για επιβατικό όχημα και 4 για αγροτικό όχημα
23	Second.Driver	Κατηγορική, δύτιμη	1, αν υπάρχει και άλλος ασφαλιζόμενος στο ίδιο όχημα και 0, αλλιώς
24	Area	Κατηγορική, δύτιμη	0, αν το συμβάν που χρήζει αποζημίωσης έχει γίνει σε αγροτική περιοχή και 1, σε κατοικημένη (άνω των 30χιλ. κατοίκων)
25	Cost.claims.year	Ποσοτική	Συνολικό ποσό των απαιτήσεων σε απροσδιόριστη χρηματική μονάδα
26	N.claims.year	Ποσοτική	Αριθμός των απαιτήσεων

27	N.claims.history	Ποσοτική	Αριθμός των απαιτήσεων πριν τον Νοέμβριο του 2015
28	R.claims.history	Ποσοτική	Λόγος του αριθμού των απαιτήσεων προς το χρόνο κάλυψης
29	Exposure	Ποσοτική	Χρόνος κάλυψης, έκθεσης στον κίνδυνο, σε χρόνια
30	Driver.age	Ποσοτική	Ηλικία ασφαλιζόμενου, ακέραιος αριθμός
31	Drv.age	Κατηγορική, διάταξης	Ηλικία ασφαλιζόμενου σε κατηγορίες
32	Driver.age.cut	Κατηγορική	Ομαδοποιημένες ηλικίες
33	Years.licenced	Ποσοτική	Χρόνια οδήγησης, ακέραιος αριθμός
34	Years.lic	Κατηγορική, διάταξης	Χρόνια οδήγησης σε κατηγορίες
35	Vehicle.age	Ποσοτική	Ηλικία οχήματος, ακέραιος αριθμός
36	Veh.age	Κατηγορική, διάταξης	Ηλικία οχήματος σε κατηγορίες
37	Vehicle.age.cut	Κατηγορική	Ομαδοποιημένες ηλικίες
38	Year	Κατηγορική, διάταξης	Έτος σύναψης συμβολαίου
39	Frequency	Ποσοτική	Συχνότητα απαιτήσεων

Πίνακας 4.1: Περιγραφή μεταβλητών

Οδηγός	Όχημα	Συμβόλαιο	Γεωγραφικές	Απόκρισης
ID	Year.matriculation	Date.last.renewal	Area	Cost.claims.year
Date.start.contract	Power	Date.next.renewal		N.claims.year
Date.birth	Cylinder.capacity	Distribution.channel		N.claims.history
Date.driving.licence	Value.vehicle	Seniority		R.claims.history
	N.doors	Policies.in.force		
	Type.fuel	Max.policies		
	Length	Max.products		
	Weight	Lapse		
		Date.lapse		
		Payment		
		Premium		
		Type.risk		
		Second.Driver		

Πίνακας 4.2: Κατηγοριοποίηση διαθέσιμων μεταβλητών

4.3 Έλεγχος Ορθότητας των Δεδομένων

Στο παρόν στάδιο, γίνετε ο έλεγχος της ορθότητας των δεδομένων (`data validation`) για λάθη και άγνωστες τιμές. Οι εντολές στο Παράρτημα εκτελούνται με σκοπό να διασφαλιστεί ο επιθυμητός τύπος των μεταβλητών αλλά επίσης για να γίνουν οι απαραίτητοι έλεγχοι στα δεδομένα. Με βάση τους ελέγχους αυτούς παρατηρούνται και κάποια λάθη.

Αρχικά, υπάρχουν 225 εγγραφές στις οποίες η ημερομηνία `Date.start.contract` είναι παλαιότερη της `Date.driving.licence`. Αυτό προφανώς είναι αδύνατο και άρα πρόκειται για εσφαλμένες τιμές. Επίσης, υπάρχουν 404 συμβόλαια-εγγραφές των οποίων η ημερομηνία διακοπής `Date.lapse` έγινε νωρίτερα της ημερομηνίας έναρξης κάποιου συμβολαίου `Date.next.renewal`. Αυτό, ίσως να σημαίνει πως ο ασφαλιζόμενος έχει διακόψει κάποιο συμβόλαιο του πριν από την περίοδο στην οποία έχουν συλλεχθεί τα δεδομένα ή έχει διακόψει συμβόλαιο και ξεκινήσει ένα νέο. Υπάρχουν επίσης 31 συμβόλαια στα οποία η `Date.driving.licence` είναι αργότερα της `Date.last.renewal`, κάτι το οποίο είναι επίσης αδύνατο. Με την αφαίρεση των εγγραφών που περιέχουν τέτοιες τιμές, το μέγεθος του δείγματος είναι 105,328.

Όσο αφορά τις ελλιπείς τιμές, υπάρχουν συνολικά 12093 άγνωστες τιμές από τις οποίες οι 10329 αφορούν τη μεταβλητή `Length` και οι υπόλοιπες τη `Type.fuel`. Επιλέγεται να αφαιρεθούν οι εγγραφές αυτές καθώς υπάρχει αυτή η πολυτέλεια λόγω του μεγέθους του δείγματος. Επίσης, και η μεταβλητή `Date.lapse` περιέχει ελλιπείς τιμές αλλά δεν προσμετρώνται αφού ελλιπής τιμή σημαίνει πως το συμβόλαιο δεν έχει διακοπεί σε διάστημα ενός χρόνου. Με την αφαίρεση και αυτών των τιμών το νέο μέγεθος του δείγματος είναι 95,040.

Τέλος, κατά τη διαδικασία έχει διασφαλιστεί ο επιθυμητός τύπος των μεταβλητών ειδικότερα για τις μεταβλητές που αφορούν ημερομηνίες και επίσης η επιλογή των κατηγοριών βάσης στις κατηγορικές μεταβλητές.

4.4 Περιγραφική Στατιστική

Σε αυτή την ενότητα παρουσιάζονται τα χαρακτηριστικά των δεδομένων του δείγματος `Motor vehicle insurance data` ενώ κατασκευάζονται μεταβλητές που θα φανούν χρήσιμες στη μετέπειτα μοντελοποίηση. Στον Πίνακα 4.3 φαίνονται τα αριθμητικά μέτρα των ποσοτικών μεταβλητών ενώ στον Πίνακα 4.4 οι συχνότητες (ή οι υψηλότερες συχνότητες) των κατηγορικών μεταβλητών. Η περιγραφή γίνεται με βάση ολόκληρο το δείγμα.

Αρχικά, κατασκευάζεται η ποσοτική μεταβλητή έκθεσης στον κίνδυνο, `Exposure`, η οποία θα δηλώνει την πραγματική διάρκεια ενός συμβολαίου σε χρόνια. Η κατασκευή της αποσκοπεί στο να μπορούν να γίνονται συγκρίσεις απαιτήσεων που έχουν γίνει σε διαφορετική διάρκεια. Από την κατασκευή της, η μεταβλητή αυτή είναι θετική και μικρότερη της μονάδας. Τιμή ίση με ένα σημαίνει πως η τελική διάρκεια του συμβολαίου αφορούσε ένα ολόκληρο έτος κάλυψης ενώ η κάθε τιμή της είναι το μέρος του έτους στο οποίο υπήρχε έκθεση. Υπενθυμίζεται πως η συμφωνημένη διάρκεια συμβολαίων είναι μονοετής και άρα δε δύναται να υπάρχουν τιμές μεγαλύτερες της μονάδας. Εγγραφές με μηδενικό `Exposure` δεν έχουν σημασία και για αυτό αφαιρούνται με αποτέλεσμα το δείγμα να φτάνει τις 81,764 εγγραφές. Από τον υπολογισμό του πρώτου και τρίτου τεταρτημορίου της μεταβλητής αυτής, είναι αντιληπτό πως πάνω από το 75% των ασφαλιζόμενων δεν τερματίζει το συμβόλαιο του πρόωρα. Η μέση περίοδος έκθεσης είναι 0.98 χρόνια (δηλαδή σχεδόν 358 μέρες,

Μεταβλητή	Ελάχιστη τιμή	Διάμεσος	Τυπική Απόκλιση	Δειγματικός Μέσος	Μέγιστη τιμή
`Cost.claims.year`	0.00	0.00	151,675.90	14,342.00	26,085,324.00
`N.claims.year`	0.00	0.00	1.13	0.46	25.00
`N.claims.history`	0.00	1.00	3.93	2.76	52.00
`R.claims.history`	0.00	0.00	61.42	30.34	2,607.00
`Power`	0.00	90.00	30.10	92.86	560.00
`Cylinder.capacity`	49.00	1,598.00	409.41	1,617.00	7,480.00
`Value.vehicle`	656.00	21,900.00	906,161.20	457,539.00	14,019,999.00
`Length`	1.98	4.23	0.39	4.25	8.22
`Weight`	43.00	1,210.00	264.04	1,195.00	7,300.00
`Driver.age`	18.00	47.00	12.84	47.33	99.00
`Years.licensed`	0.00	24.00	12.49	24.77	74.00
`Vehicle.age`	0.00	13.00	6.78	12.74	69.00
`Exposure`	0.01	1.00	0.10	0.98	1.00
`Frequency`	0.00	0.00	1.50	0.51	166.67

Πίνακας 4.3: Κύρια μέτρα θέσης και μεταβλητότητας ποσοτικών μεταβλητών

1 χρόνος) ενώ υπάρχουν και εγγραφές με έκθεση μόλις 4 μέρες. Το 96% των ασφαλιζόμενων του δείγματος δε διακόπτει το συμβόλαιο του, ενώ λιγότερο του 2.5% διακόπτει στο μισό. Ο υπολογισμός της διάρκειας έκθεσης γίνεται βάση την ημερομηνίας ανανέωσης και την ημερομηνίας λήξης ή διακοπής, αν έχει γίνει.

Είναι επίσης απαραίτητη η κατασκευή των μεταβλητών `Driver.age`, `Years.licensed`, `Vehicle.age`, `Drv.age`, `Years.lic`, `Veh.age`, `Year` και `Frequency` των οποίων η περιγραφή φαίνεται στον Πίνακα 4.1. Η κατασκευή επιπρόσθετων μεταβλητών φαίνεται στο Παράρτημα. Στο Διάγραμμα 4.1 φαίνεται η κατανομή των μεταβλητών `Driver.age`, `Years.licenced` και `Vehicle.age` ενώ στον Πίνακα 4.5 τα αριθμητικά τους μέτρα. Η κατανομή των ηλικιών δείχνει ότι το μεγαλύτερο μέρος του δείγματος αφορά μεσήλικες 40 με 60 ετών και η μέση ηλικία είναι 47 ετών. Επίσης, από τις τιμές της `Years.licenced` γίνεται αντιληπτό ότι δεν ασφαλίζονται κατά κόρον νέοι οδηγοί αφού η μέση τιμή ετών οδήγησης φτάνει τα 24 έτη. Το ποσοστό των νέων οδηγών στο δείγμα είναι το 32%. Κατά μέσο όρο ασφαλίζονται οχήματα ηλικίας πάνω από 12 έτη ενώ υπάρχει ασφάλιση που αφορά όχημα μέχρι και 66 χρονών.

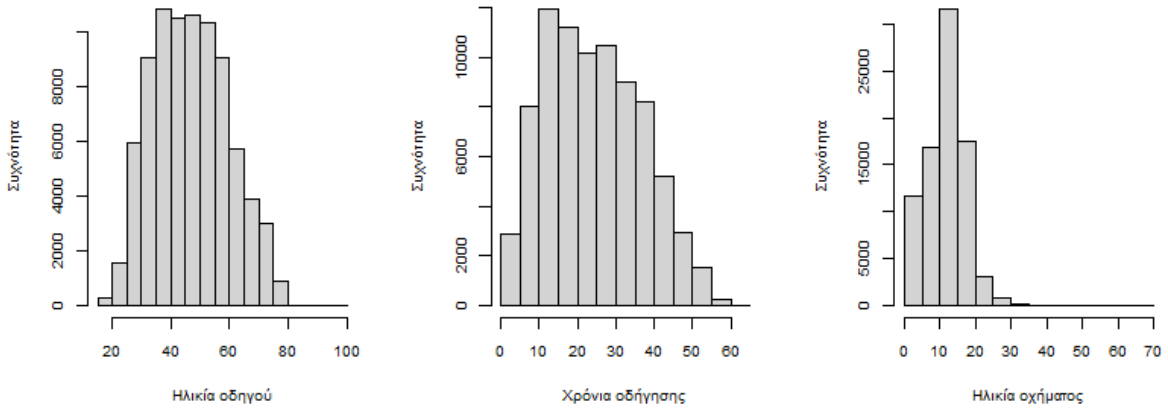
Στον Πίνακα 4.6 γίνεται ανάλυση της συνολικής έκθεσης, αριθμού και μεγέθους απαιτήσεων καθώς και της μέσης συχνότητας και σφοδρότητας για κάθε ένα από τα 4 έτη παρατήρησης σε ολόκληρο το δείγμα. Τα έτη 2015 (2 μήνες) και 2018 (11 μήνες) δεν είναι ολοκληρωμένα. Παρατηρείται ότι με το πέρασμα των ετών η συχνότητα των απαιτήσεων μειώνεται. Για το 2015 η μέση συχνότητα φτάνει το 91% ενώ στο 2017 μειώνεται σχεδόν στο μισό φτάνοντας το 47%. Η τιμή της μέσης συχνότητας υπολογίζεται από τον λόγο του συνολικού αριθμού απαιτήσεων που αναφέρθηκαν σε κάθε έτος με τη συνολική έκθεση στον κίνδυνο για όλους τους ασφαλιζόμενους. Για τα δύο ολοκληρωμένα έτη 2016 και 2017, ο όγκος εργασίας όσο αφορά τη συνολική έκθεση έχει αυξηθεί κατά $\frac{25258}{23226} - 1 = 8.74\%$ ενώ η συχνότητα έχει μειωθεί σημαντικά κατά $1 - \frac{0.47}{0.86} = 45.34\%$ στο επόμενο έτος. Από το πρώτο έτος παρατήρησης, 2015, μέχρι το τελευταίο, 2018, η συχνότητα έχει μειωθεί κατά $1 - \frac{0.17}{0.91} = 81.31\%$.

Μεταβλητή	Κατηγορίες	Συχνότητα	Κατηγορία αναφοράς
Distribution.channel	0	42751	1
	1	36339	
	00/01/1900	2674	
Type.risk	1	2	3
	2	11309	
	3	70433	
	4	20	
Area	0	60268	0
	1	21496	
N.doors	0	22	5
	2	2574	
	3	11945	
	4	11850	
	5	54859	
	6	514	
Type.fuel	D	54676	D
	P	27088	
Year	2015	3537	2015
	2016	23854	
	2017	25873	
	2018	28500	
Drv.age	18...79 και 93...99	40	2275
		41	2233
		38	2216
Driver.age.cut	[18, 24]	1367	[18, 24]
	[25, 44]	34801	
	[45, 59]	30622	
	[60, 69]	10387	
	[70, Inf)	4587	
Veh.age	0...69	12	6696
		13	6656
		14	6389
Vehicle.age.cut	5-	9640	5-
	6+	72124	
Years.lic	0...64, 74	12	2525
		11	
		13	2461
		10	2421
		14	
		20	

Πίνακας 4.4: Πίνακας συχνοτήτων

	Ελ. τιμή	1ο τετ.	2ο τετ.	Μέση τιμή	3ο τετ.	Μέγ. τιμή
`Driver.age`	18.00	37.00	47.00	47.32	57.00	98.00
`Years.licenced`	0.00	14.00	24.00	24.73	34.00	64.00
`Vehicle.age`	0.00	9.00	12.00	12.13	16.00	66.00

Πίνακας 4.5: Αριθμητικά μέτρα των `Driver.age`, `Years.licences` και `Vehicle.age`



Διάγραμμα 4.1: Ιστογράμματα των `Driver.age` (αριστερό γράφημα), `Years.licences` (μεσαίο γράφημα) και `Vehicle.age` (δεξί γράφημα) για το δείγμα μάθησης

Η μέση συχνότητα για όλο το δείγμα είναι στο 50% και αυτή αποκλίνει σημαντικά από όλα τα έτη εκτός το 2017. Η συχνότητα των απαιτήσεων λαμβάνει αρκετές ερμηνείες. Αρχικά, κατά μέσο όρο το 50% των ασφαλισμένων μονάδων έχει αναφέρει τουλάχιστον μία απαίτηση στο διάστημα των ετών που μελετώνται. Μία άλλη ερμηνεία είναι πως ο κάθε ασφαλιζόμενος θα προβεί κατά μέσο όρο σε λιγότερο από μία απαίτηση στο διάστημα ενός έτους. Κατά την πιθανοθεωρητική προσέγγιση υπάρχει 50% πιθανότητα ένας ασφαλιζόμενος να αναφέρει τουλάχιστον μία απαίτηση σε διάρκεια ενός έτους. Παράλληλα, το μέσο μέγεθος της κάθε απαίτησης ανά χρονιά δε φαίνεται να αποκλίνει από τον ολικό μέσο όρο 31,725. Συμπερασματικά, η συχνότητα των απαιτήσεων φαίνεται να έχει συνεχή μείωση με το πέρασμα των ετών ειδικότερα στα τρία τελευταία έτη. Επομένως, φαίνεται αρχικά πως το έτος σύναψης κάποιου συμβολαίου να έχει σχέση με τη συχνότητα των απαιτήσεων και αυτό είναι λογικό μέχρι σε ένα σημείο. Η μείωση αυτή χαρακτηρίζεται αρχικά σημαντική και θα διερευνηθεί αργότερα.

Η μεταβλητή `N.claims.year` αντιπροσωπεύει τον αριθμό των απαιτήσεων σε κάθε συμβόλαιο. Για ολόκληρο το δείγμα, η κατανομή του αριθμού των απαιτήσεων φαίνεται στον Πίνακα 4.7. Το 77% των τιμών είναι μηδέν, ενώ για τις υπόλοιπες μη μηδενικές τιμές ο μέσος αριθμός είναι 2.12 απαιτήσεις. Στο Διάγραμμα 4.2 φαίνεται η κατανομή της `N.claims.year` για διάφορες ομάδες ηλικιών. Μπορούμε να συμπεράνουμε ότι η χρήση κατανομής Poisson για τον αριθμό απαιτήσεων φαίνεται να είναι αρκετά λογική. Χρησιμοποιείται η μεταβλητή ηλικίας αφού είναι βασική για το ερευνητικό ερώτημα. Παρόλαυτα, η αρκετά υψηλή συχνότητα των μηδενικών απαιτήσεων δημιουργεί ερωτήματα του κατά πόσο η κατανομή Poisson μπορεί να ανταποκριθεί στη συμπεριφορά αυτή.

Έτος	Έκθεση	Αριθμός απαιτήσεων	Μέγεθος απαιτήσεων	Μέση συχνότητα	Μέση σφοδρότητα
2015	3438.34	3,144	95,163,794	0.91	30,268.38
2016	23,225.73	19,982	619,873,689	0.86	31,021.60
2017	25,257.90	11,762	377,443,984	0.47	32,090.12
2018	28,206.46	4,881	169,212,429	0.17	34,667.57
Σύνολο	80,128.43	39,769	1,261,693,896	0.50	31,725.56

Πίνακας 4.6: Ανάλυση έκθεσης και απαιτήσεων ανά έτος

N.claims.year	0	1	2	3	4	5	6	7	8	9	10	11	12
Συχνότητα	63,062	9,023	4,744	2,349	1,139	579	302	216	128	77	60	24	17

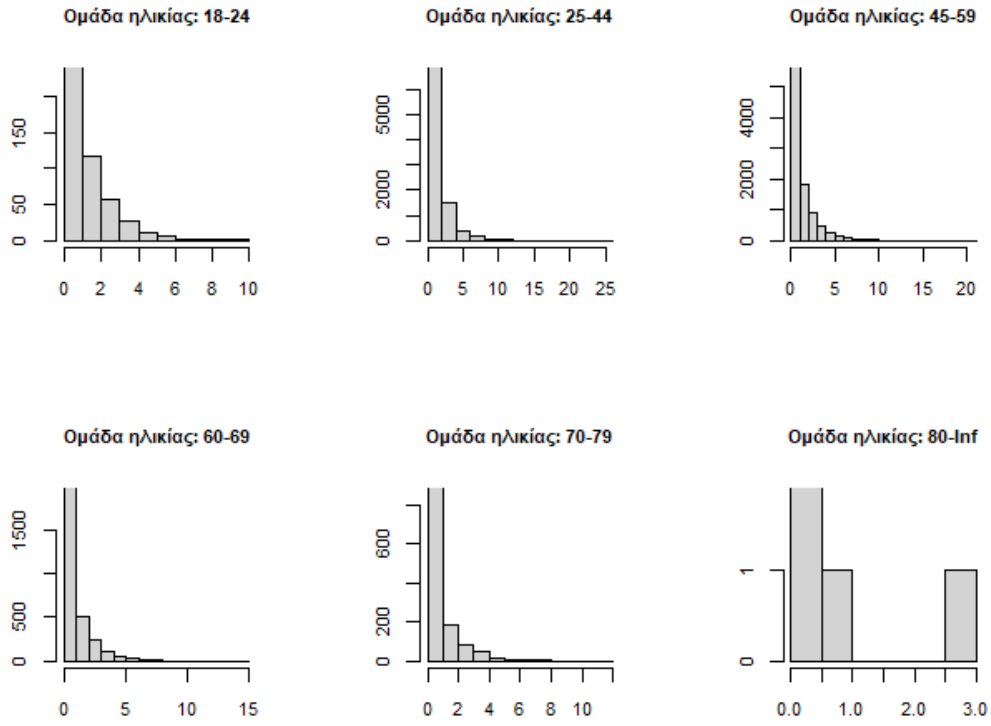
N.claims.year	13	14	15	16	17	18	19	20	21	22	23	24	25
Συχνότητα	17	9	8	2	3	2	1	0	1	0	0	0	1

Πίνακας 4.7: Πλήθος εγγραφών ανά αριθμό απαιτήσεων

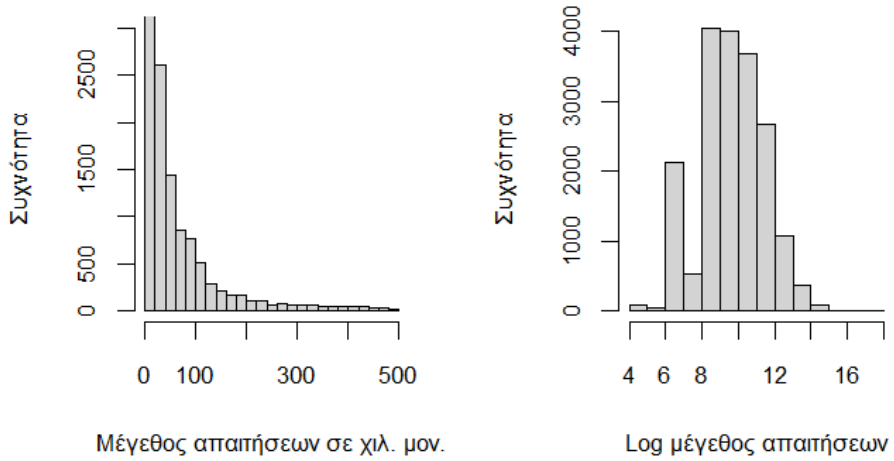
Η σχέση μεταξύ του αριθμού των απαιτήσεων και της έκθεσης είναι ένας τρόπος για τον εντοπισμό έκτροπων τιμών (outliers). Για παράδειγμα, ο ασφαλιζόμενος με `ID` 26435 έχει αναφέρει 10 απαιτήσεις σε διάστημα μόλις 21 ημερών ενώ ο μέσος όρος απαιτήσεων στο δείγμα είναι το πολλή μία απαίτηση σε διάστημα ενός χρόνου. Στην περίπτωση αυτή πετυχαίνεται η μέγιστη τιμή της μεταβλητής `Frequency` και είναι 166.67 απαιτήσεις. Υπάρχει δηλαδή ασφαλιζόμενος ο οποίος θα υποβάλει σχεδόν 167 απαιτήσεις σε διάστημα ενός χρόνου. Ταυτόχρονα, η μέση τιμή της `Frequency` είναι μόλις 0.5 απαιτήσεις. Στον Πίνακα 4.8 φαίνονται τα αριθμητικά μέτρα της μεταβλητής `Frequency`. Πάνω από το 75% των ασφαλιζόμενων έχει μηδενική συχνότητα. Προσέξτε ότι γίνεται αναφορά σε συχνότητα απαιτήσεων και αυτή αφορά τη μέση συχνότητα ανάλογα με το προφίλ του ρίσκου/ασφαλιζόμενου που μελετάται. Η μεταβλητή `Frequency` αντιπροσωπεύει την πραγματική συχνότητα απαιτήσεων σε μεμονωμένο συμβόλαιο.

Για τη μεταβλητή μεγέθους απαιτήσεων, `Cost.claims.year`, πάνω από το 75% των συμβολαίων έχουν μηδενικό μέγεθος απαιτήσεων και αυτό αντικατοπτρίζει τον μηδενικό αριθμό απαιτήσεων. Συγκεκριμένα 63,062 εγγραφές έχουν μηδενική τιμή μεγέθους απαιτήσεων και αντιπροσωπεύουν το 77.13% του δείγματος. Στο Διάγραμμα 4.3 φαίνεται η κατανομή της μεταβλητής για τιμές μέχρι και 500 χιλ. χρηματικές μονάδες. Τιμές αυτές αντιπροσωπεύουν λιγότερο του 1% ολόκληρου του δείγματος. Παρατηρείται ότι απαιτήσεις μεγάλου μεγέθους είναι όλο και πιο σπάνιες.

Στον Πίνακα 4.10 παρουσιάζονται τα αριθμητικά μέτρα της μεταβλητής `Value.vehicle`. Το μισό δείγμα αποτελείται από οχήματα αξίας μικρότερης των 22,260 χρηματικών μονάδων, ενώ ασφαλίζονται οχήματα αξίας μέχρι 14 εκατ. χρηματικών μονάδων. Οχήματα αξίας εκατ. μονάδων και άνω αποτελούν λιγότερο του 23% του δείγματος ενώ αξίας λιγότερων των 50 χιλ. μονάδων είναι το 71%. Η μέση αξία οχημάτων λιγότερης των 50 χιλ. μονάδων είναι 19,766 χρηματικές μονάδες. Η εταιρία εδράζεται στην Ισπανία επομένως εικάζεται ότι οι χρηματικές μονάδες είναι ευρώ. Επομένως, διαπιστώνεται πως σε μεγάλο βαθμό η ασφαλιστική εταιρία συνάπτει συμβόλαια τα οποία αφορούν αυξημένα ρίσκα σε μεγάλες αποζημιώσεις λόγω της μεγάλης αξίας των αυτοκινήτων.



Διάγραμμα 4.2: Ιστογράμματα της `N.claims.year` ανά ομάδα ηλικίας



Διάγραμμα 4.3: Ιστόγραμμα των `Cost.claims.year` (αριστερό γράφημα) και `log(Cost.claims.year)` (δεξί γράφημα). Στο αριστερό γράφημα ο y-άξονας έχει περιοριστεί στο $[0, 3000]$ για τη βελτίωση του γραφήματος

	Ελ. τιμή	1ο τετ.	2ο τετ.	3ο τετ.	Μέγ. τιμή
`Frequency`	0.00	0.00	0.00	0.00	166.67

Πίνακας 4.8: Αριθμητικά μέτρα της `Frequency`

	Ελ. τιμή	1ο τετ.	2ο τετ.	3ο τετ.	Μέγ. τιμή
`Cost.claims.year`	57	5685	14978	52114	26085324

Πίνακας 4.9: Αριθμητικά μέτρα της `Cost.claims.year` για θετικά μεγέθη

Για τη μεταβλητή `premium`, το 75% των ασφαλιζόμενων πληρώνει ασφάλιστρο κάτω των 35,668 χρηματικών μονάδων ετησίως όπως φαίνεται και στον Πίνακα 4.11.

Στη συνέχεια γίνεται η ανάλυση των κατηγορικών μεταβλητών και αρχικά για τη μεταβλητή `Area` (0 για αγροτική, 1 για κατοικημένη περιοχή). Το 74% του δείγματος αφορά αγροτικές περιοχές και το υπόλοιπο κατοικημένες. Στον Πίνακα 4.12 φαίνεται η συνολική έκθεση, αριθμός και μέγεθος απαιτήσεων για κάθε μία από τις κατηγορίες των μεταβλητών. Από αυτά προκύπτουν επίσης η μέση συχνότητα, διασπορά συχνότητας και μέση σφοδρότητα των απαιτήσεων. Σε κατοικημένες περιοχές εμφανίζεται αυξημένη συχνότητα σε σχέση με αγροτικές περιοχές όπως επίσης και σφοδρότητα. Για τη μεταβλητή `Type.risk` δυστυχώς το παρόν δείγμα δεν εμφανίζει απαιτήσεις για οχήματα τύπου 1 (μοτοσικλέτες) και 4 (αγροτικά οχήματα). Παρατηρείται όμως κατά μέσο όρο μεγαλύτερη συχνότητα σε οχήματα τύπου 2 (φορτηγά) παρά σε επιβατικά οχήματα (τύπου 3). Από την άλλη η μέση σφοδρότητα των απαιτήσεων είναι μεγαλύτερη σε επιβατικά αυτοκίνητα. Για τη μεταβλητή `Type.fuel`, εμφανίζεται υψηλότερη μέση συχνότητα σε πετρελαιοκίνητα οχήματα ενώ η σφοδρότητα σε αυτή την περίπτωση είναι χαμηλότερη της χρήσης βενζίνης. Στο Διάγραμμα 4.4 παρουσιάζονται κάποια στοιβαγμένα ραβδογράμματα μεταξύ των μεταβλητών αυτών. Στο αριστερό ραβδόγραμμα, οι στοίβες ορίζονται ως οι δεσμευμένες πιθανότητες της `Type.risk` ως προς την `Area`. Όπως αναμενόταν, η χρήση αγροτικών οχημάτων είναι συχνότερη σε αγροτικές περιοχές (σκούρο χρώμα) παρά σε κατοικημένες (ανοικτό χρώμα). Στο μεσαίο ραβδόγραμμα φαίνονται οι δεσμευμένες πιθανότητες της `Type.risk` ως προς την `Type.fuel`. Παρατηρείται όπως είναι λογικό η εντονότερη χρήση πετρελαίου (σκούρο χρώμα) σε αγροτικά οχήματα και φορτηγά ενώ η χρήση του είναι μειωμένη σε αυτοκίνητα και σχεδόν μηδενική σε μοτοσικλέτες. Στο δεξί ραβδόγραμμα φαίνονται οι συχνότητες εγγραφών της `Type.fuel` με βάση την `Area`. Η χρήση πετρελαίου είναι συχνότερη της χρήσης βενζίνης ενώ είναι συχνότερη σε αγροτικές (σκούρο χρώμα) περιοχές παρά σε κατοικημένες (ανοικτό χρώμα).

Αυτά ήταν τα κυριότερα χαρακτηριστικά του δείγματος. Μέσα από την ανάλυση αυτή έγινε αντιληπτό το πεδίο στο οποίο κινείται ο ασφαλιστικός φορέας από τον οποίο προήλθαν τα δεδομένα αλλά και η κλίμακα ποσοτήτων στην οποία εργάζεται η εταιρία. Επίσης, έγινε η κατασκευή εκείνων των μεταβλητών που είναι αναγκαίες για τη μοντελοποίηση και θέτουν τη βάση της διερευνητικής ανάλυσης που ακολουθεί.

	Ελ. τιμή	1ο τετ.	2ο τετ.	3ο τετ.	Μέγ. τιμή
Value.vehicle	656	16595	22260	611367	14019999

Πίνακας 4.10: Αριθμητικά μέτρα της `Value.vehicle`

	Ελ. τιμή	1ο τετ.	2ο τετ.	3ο τετ.	Μέγ. τιμή
Premium	44	23491.5	28711	35668	299334

Πίνακας 4.11: Αριθμητικά μέτρα της `Premium`

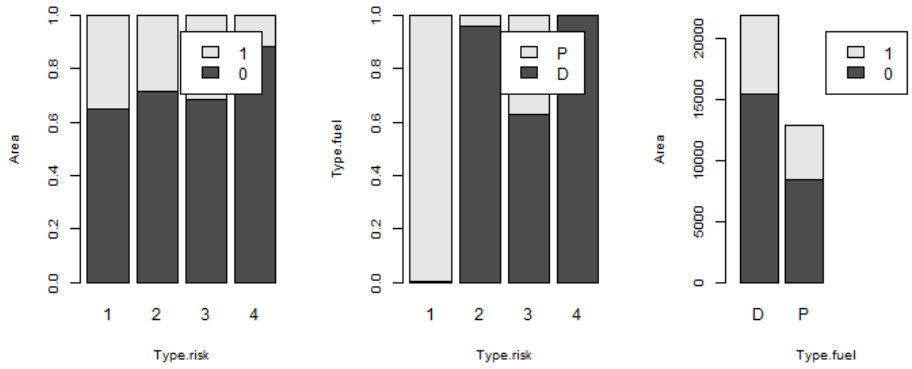
4.5 Μοντελοποίηση Συχνότητας Απαιτήσεων

4.5.1 Μεθοδολογία

Αφού έχει γίνει η περιγραφή των κύριων χαρακτηριστικών των διαθέσιμων και επιπρόσθετων μεταβλητών θα ακολουθήσει η μοντελοποίηση της συχνότητας των απαιτήσεων. Γίνεται αρχή με τη διερευνητική ανάλυση των δεδομένων όπου σκοπός είναι να εξεταστούν οι σχέσεις μεταξύ διαφόρων μεταβλητών και της συχνότητας. Με αυτό τον τρόπο θα βρεθούν εκείνες οι μεταβλητές οι οποίες σχετίζονται ή και όχι με τη συχνότητα των απαιτήσεων και αυτές που είναι πιο πιθανόν να συμπεριληφθούν κατά την προσαρμογή του μοντέλου. Στη συνέχεια θα γίνει η κατασκευή των μοντέλων με βάση τις μεταβλητές που μελετήθηκαν κατά τη διερευνητική ανάλυση.

Υπάρχουν συνολικά 81,764 εγγραφές και αυτό επιτρέπει στην ανάλυση να γίνει η χρήση της τεχνικής ενδοεπικύρωσης με χρήση κ ομάδων (k-fold cross-validation). Στην παρούσα ανάλυση το δείγμα χωρίζεται αρχικά σε δύο ομάδες (2-fold cross-validation), το δείγμα μάθησης (training data) το οποίο επιλέγεται να είναι το 60% του δείγματος και το δείγμα ελέγχου (testing data), που είναι το υπόλοιπο. Με αυτό τον τρόπο τα μοντέλα εκπαιδεύονται στα δεδομένα του δείγματος μάθησης ενώ θα γίνει χρήση του δείγματος ελέγχου για την επικύρωση τους. Έχοντας τις τιμές της μεταβλητής απόκρισης στο δείγμα ελέγχου θα συγκριθούν με τις εκτιμήσεις του μοντέλου. Για την υλοποίηση αυτού, ορίζεται μία ομοιόμορφη μεταβλητή $u \in (0, 1)$, για κάθε εγγραφή ολόκληρου του δείγματος. Το δείγμα μάθησης θα είναι αυτό με $u < 0.6$ και το δείγμα ελέγχου αυτό με $u > 0.6$. Το δείγμα μάθησης έχει μέγεθος 49,292 και το δείγμα ελέγχου 32,472. Η διερευνητική ανάλυση γίνεται στο δείγμα μάθησης ενώ η περιγραφική στατιστική έχει γίνει σε ολόκληρο το δείγμα.

Κατασκευάζονται αρχικά μοντέλα μίας μεταβλητής ενώ στη συνέχεια αυτά αναπτύσσονται σε μοντέλα πολλαπλών μεταβλητών. Η μοντελοποίηση αποτελείται από τρία στάδια. Τη μοντελοποίηση με χρήση κατανομής Poisson, με χρήση Αρνητικής διωνυμικής κατανομής και τέλος με την κατανομή Μηδενικής διόγκωσης. Το μοντέλο Poisson ως το κλασσικότερο στη μοντελοποίηση δεδομένων μέτρησης θα αποτελέσει τη βάση για την κατασκευή και των υπόλοιπων μοντέλων αφού σε αυτό θα γίνει η διαδικασία της βήμα προς βήμα προσθήκης μεταβλητών. Από αυτή τη διαδικασία θα προκύψουν 12 διαφορετικά μοντέλα όπου στο κάθε ένα το σύνολο των μεταβλητών που συμμετέχουν θα είναι διαφορετικό. Η συμπερίληψη ή όχι κάποιας μεταβλητής θα σημαίνει και την επιλογή ή όχι κάποιου μοντέλου. Στο Παράρτημα βρίσκεται ο κώδικας υλοποίησης όλων των κατασκευασμένων μοντέλων που παρουσιάζονται σε αυτή την ενότητα. Στη συνέχεια, γίνεται ο διαγνωστικός έλεγχος των μοντέλων με χρήση γραφημάτων και η επικύρωση τους με χρήση της μεθόδου ενδοεπικύρωσης για $\kappa = 2$ και $\kappa = 5$. Για λόγους συντομίας, η αναφορά σε κάθε ένα από τα τρία μοντέλα θα γίνεται



Διάγραμμα 4.4: Στοιβαγμένα ραβδοδράμματα της `Type.risk` έναντι της `Area` (αριστερό γράφημα), της `Type.risk` έναντι της `Type.fuel` (μεσαίο γράφημα) και της `Type.fuel` έναντι της `Area` (δεξί γράφημα)

		Συνολική έκθεση	Αριθμός απαιτήσεων	Συνολικό μέγεθος	Μέση συχνότητα	Διασπορά συχνότητας	Μέση σφοδρότητα
`Area`	0	59,110.50	27,412.00	869,805,087.00	0.46	1.38	66,069.51
	1	21,017.93	12,357.00	391,888,809.00	0.59	1.82	70,776.38
`Type.risk`	1	2.00	0.00	0.00	0.00	0.00	0.00
	2	11,068.34	7,331.00	166,306,059.00	0.66	2.54	22,685.32
	3	69,038.09	32,438.00	1,095,387,837.00	0.47	1.32	33,768.66
	4	20.00	0.00	0.00	0.00	0.00	0.00
`Type.fuel`	P	26,542.21	10,613.00	377,294,666.00	0.40	1.06	35,550.24
	D	53,586.22	29,156.00	884,399,230.00	0.54	1.74	30,333.35

Πίνακας 4.12: Ανάλυση έκθεσης, συχνότητας και σφοδρότητας απαιτήσεων για τις κατηγορίες των `Area`, `Type.risk` και `Type.fuel`

με τη χρήση των συντομογραφιών `poisson`, `negbin` και `zip` για το μοντέλο Poisson, Αρνητικό διωνυμικό μοντέλο και μοντέλο Μηδενικής διόγκωσης, αντίστοιχα.

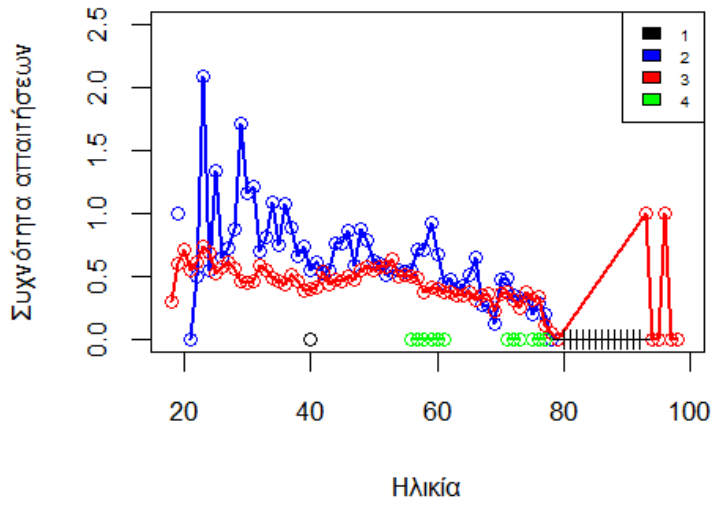
4.5.2 Διερευνητική ανάλυση δεδομένων

Σκοπός της ανάλυσης που ακολουθεί είναι να διερευνηθεί η σχέση μεταξύ της μέσης συχνότητας απαιτήσεων και διαφόρων μεταβλητών. Σε κάθε περίπτωση θα εκτιμάται η μέση συχνότητα ως προς τις κατηγορίες (για κατηγορικές) ή ως προς τις τιμές (για συνεχής) της μεταβλητής που εξετάζεται. Η διαδικασία αυτή ευκολύνει την αναγνώριση τάσεων και αγνοεί τον θόρυβο και τις έκτροπες τιμές αφού υπολογίζει μέσους όρους. Η ανάλυση θα γίνει στο δείγμα μάθησης στο οποίο και θα εκπαιδευτούν τα μοντέλα. Επιθυμούμε να βρούμε γραμμικές τάσεις οι οποίες θα ανταποκρίνονται στη δομή ενός γενικευμένου μοντέλου.

Από τον Πίνακα 4.6, η ολική μέση συχνότητα στο δείγμα είναι 50% ενώ όπως έχει παρατηρηθεί σε ετήσια ανάλυση η μέση συχνότητα ανά έτος μειώνεται σημαντικά με το πέρασμα των ετών. Επομένως, η μεταβλητή `Year` μπορεί να θεωρηθεί καλή υποψήφια για το μοντέλο συχνότητας απαιτήσεων.

Από τον Πίνακα 4.12 οι μεταβλητές `Type.risk` και `Type.fuel` φαίνεται να αποτελούν καλές υποψήφιες για να συμπεριληφθούν στο μοντέλο της συχνότητας απαιτήσεων. Από τον Πίνακα 4.12 έχει παρατηρηθεί μεγαλύτερη μέση συχνότητα απαιτήσεων σε φορτηγά παρά σε επιβατικά αυτοκίνητα. Το αντίστοιχο παρατηρείται και στο Διάγραμμα 4.5 στο οποίο παρουσιάζεται η μέση συχνότητα απαιτήσεων ανά ηλικία. Επιλέγεται η ηλικία για τον λόγο ότι αποτελεί μία μεταβλητή που υπάρχει επιθυμία να συμπεριλαμβάνεται μέσα στο μοντέλο και έτσι μελετώνται οι σχέσεις με βάση και αυτή. Η μέση συχνότητα απαιτήσεων σε φορτηγά (μπλε απεικόνιση) είναι σχεδόν συνέχεια μεγαλύτερη από τη μέση συχνότητα σε επιβατικά αυτοκίνητα (κόκκινη απεικόνιση). Επίσης, υπάρχει έντονη διαφοροποίηση της μέσης συχνότητας κυρίως σε νεαρές ηλικίες, από 20 μέχρι 30 ετών, κάτι που ενδεχομένως να δείχνει την αδυναμία νέων να χειριστούν οχήματα όπως φορτηγά ή ακόμη και την έλλειψη εμπειρίας. Στο Διάγραμμα 4.5 εμφανίζονται με σταυρό ηλικίες που δεν υπάρχουν στο δείγμα και είναι οι 80 με 92 ετών, ενώ επίσης υπάρχουν ηλικίες στις οποίες δεν έχει γίνει κάποια απαίτηση.

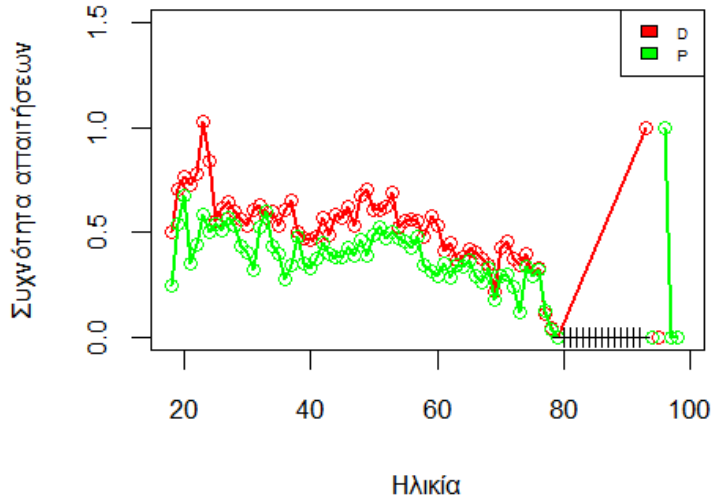
Όσο αφορά τη μεταβλητή `Type.fuel`, στο Διάγραμμα 4.6 φαίνεται πάλι η μέση συχνότητα απαιτήσεων για κάθε ηλικία. Από αυτό, η χρήση καυσίμου πετρελαίου (κόκκινη απεικόνιση) συνιστά αυξημένη μέση συχνότητα απαιτήσεων σε σχέση με τη χρήση βενζίνης και αυτό είναι κάτι το οποίο διαπιστώθηκε και κατά την περιγραφή των μεταβλητών.



Διάγραμμα 4.5: Μέση συχνότητα απαιτήσεων σε σχέση με ηλικία και τύπο οχήματος

Για τη μεταβλητή `Area`, παρατηρείται από τον Πίνακα 4.13 ότι σε αγροτικές περιοχές η μέση εμπειρική συχνότητα είναι μεγαλύτερη της αντίστοιχης για κατοικημένες περιοχές. Η διαφοροποίηση αυτή αποδυναμώνεται με την πάροδο των ετών. Επομένως, η μεταβλητή `Area` αποτελεί επίσης υποψήφια μεταβλητή για το μοντέλο.

Συνεχίζοντας, μελετάτε η σχέση της συχνότητας απαιτήσεων με τη μεταβλητή `Driver.age` περισσότερο. Στο δείγμα μάθησης, υπάρχουν ηλικίες από 18 μέχρι 79 και από 93 μέχρι 99, με σύνολο 68 μοναδικών ηλικιών όπως ακριβώς και σε ολόκληρο το δείγμα. Από τις ίδιες τιμές αποτελείται και το δείγμα ελέγχου. Από τον Πίνακα 4.14, το 75% των ηλικιών έχουν μέση συχνότητα απαιτήσεων



Διάγραμμα 4.6: Μέση συχνότητα απαιτήσεων σε σχέση με ηλικία και τύπο καυσίμου

	Έκθεση		Αριθμός απαιτήσεων		Συχνότητα απαιτήσεων	
Έτος	Αγροτική	Κατοικημένη	Αγροτική	Κατοικημένη	Αγροτική	Κατοικημένη
2015	1,457.70	610.62	1,279	662	0.88	1.08
2016	9,998.65	3,980.97	8,206	3,824	0.82	0.96
2017	11,243.62	3,928.57	4,991	2,089	0.44	0.53
2018	13,000.99	4,085.27	2,167	804	0.17	0.20
Σύνολο	35,700.96	12,605.43	16,643	7,379	0.47	0.59

Πίνακας 4.13: Μέση συχνότητα απαιτήσεων ανά έτος και τύπο περιοχής

κάτω του 57% και η μέγιστη συχνότητα είναι 100%. Αυτή προέρχεται από τις ηλικίες 98 και 96. Η μέγιστη συχνότητα καταδεικνύει ότι σε εκείνες τις ηλικίες σε διάρκεια ενός έτους θα υπάρξει κατά μέσο όρο μία αναφορά απαίτησης. Η συχνότητα των απαιτήσεων έχει κατασκευαστεί υπολογίζοντας τον συνολικό αριθμό απαιτήσεων και διαιρώντας τον με τη συνολική έκθεση για κάθε μία από τις 68 διαφορετικές ηλικίες.

Στον Πίνακα 4.15 φαίνεται η μέση συχνότητα των απαιτήσεων, η διασπορά της και το πλήθος των εγγραφών για διαφορετικές ηλικιακές ομάδες. Η χρήση της Poisson κατανομής υποθέτει πως θα πρέπει η διασπορά να είναι ίση με τη μέση τιμή σε κάθε ηλικιακό εύρος. Παρατηρήστε ότι με την αύξηση της ηλικίας υπάρχει τάση μείωσης της μέσης τιμής και της διασποράς. Όμως, οι διασπορές είναι μεγαλύτερες των μέσων τιμών και μειώνονται όσο αυξάνεται η ηλικία. Επομένως, πιθανότατα υπάρχει παραβίαση της υπόθεσης ισότητας της μέσης τιμής και διασποράς σε περίπτωση που θεωρηθεί ότι η συχνότητα απαιτήσεων ακολουθεί Poisson κατανομή. Αυτό συνεπάγει την πιθανή υπερμεταβλητότητα των δεδομένων, δηλαδή την πολύ μεγαλύτερη διασπορά σε σχέση με τη μέση

Ελ. τιμή	1ο τετ.	Διάμεσος	Μέση τιμή	3ο τετ.	Μέγ. τιμή
0.00	0.37	0.49	0.46	0.57	1.00

Πίνακας 4.14: Στατιστικά της μέσης συχνότητας των απαιτήσεων ανά ηλικία.

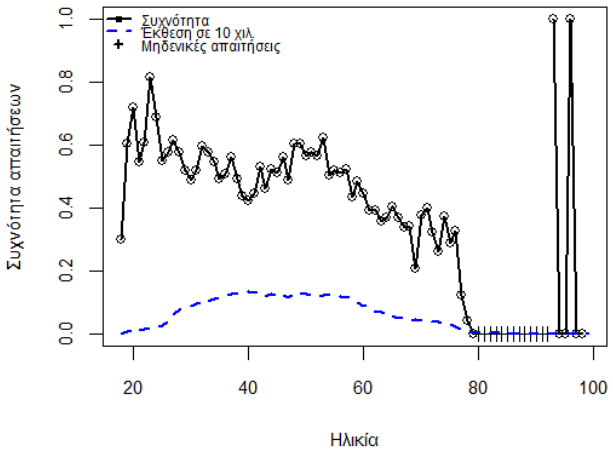
Ηλικιακή ομάδα	Μέση Συχνότητα	Διασπορά	Πλήθος εγγραφών
18-24	0.67	1.57	1,367
25-44	0.51	1.53	34,801
45-59	0.54	1.77	30,622
60-69	0.37	1.00	10,387
70-79	0.31	0.82	4,576
80+	0.50	1.00	11

Πίνακας 4.15: Μέση συχνότητα και διασπορά απαιτήσεων ανά ηλικιακή ομάδα

τιμή η οποία προβλέπεται από ένα μοντέλο Poisson. Η μείωση της διασποράς με την αύξηση της ηλικίας δείχνει ότι σε μεγαλύτερες ηλικίες η μεταβλητότητα των απαιτήσεων είναι πολύ μικρότερη και άρα νεότεροι οδηγοί έχουν την τάση να αναφέρουν απαιτήσεις πολύ πιο ακραίες μεταξύ τους.

Στο Διάγραμμα 4.7 φαίνεται η μέση συχνότητα απαιτήσεων ανά ηλικία και η συνολική έκθεση για κάθε ηλικία σε κλίμακα 10 χιλιάδων ετών. Είναι εμφανές ότι υπάρχει τάση η συχνότητα να μειώνεται με την αύξηση της ηλικίας και μετέπειτα αύξηση σε μεγάλες ηλικίες, όμως υπάρχει αρκετή μεταβλητότητα στη συχνότητα και αρκετές αυξομειώσεις (zigzag patterns). Η υψηλή συχνότητα στις μεγαλύτερες ηλικίες είναι κάτι το οποίο δεν μπορεί να γίνει αντιληπτό για τον λόγο ότι λείπουν ηλικίες στο διάστημα 80 με 92 και άρα η αύξηση δεν είναι ξεκάθαρη. Επιπλέον, φαίνεται ότι κοντά στην ηλικία των 50 η συχνότητα παρουσιάζει τοπικά μέγιστό. Αυτό μπορεί να συμβαίνει λόγω νεαρών οδηγών που ξεκινούν να οδηγούν τα οχήματα των γονιών τους και άρα το ρίσκο να είναι αυξημένο. Αν και μοντελοποιείται η λογαριθμική μέση συχνότητα, στο Διάγραμμα 4.7 φαίνεται κάποια γραμμική σχέση μεταξύ της μέσης συχνότητας και της ηλικίας. Το διάγραμμα της λογαριθμικής συχνότητας έχει παρόμοια συμπεριφορά και γιάυτό δε φαίνεται στην ανάλυση αυτή. Αναγνωρίζονται επίσης κάποιες πιο συγκεκριμένες τάσεις σε ηλικιακές ομάδες. Η συχνότητα αυξάνεται στις ηλικιακές ομάδες 18 με 24 και 41 με 53, ενώ έχει μειωτική τάση στις ηλικίες 25 με 40 και 55 μέχρι 78. Στο ίδιο διάγραμμα φαίνεται η έκθεση στο κίνδυνο κάθε ηλικίας σε κλίμακα 10 χιλιάδων ετών. Αυτό χρησιμοποιείται ποιοτικά. Παρατηρείστε την αυξημένη έκθεση στις ηλικίες 35 με 60 και την αντίστοιχη μεταβολή της μέσης συχνότητας. Παρ΄ όλη την αυξημένη έκθεση, η συχνότητα δεν αυξάνεται στις ηλικίες 35 με 40. Η ανάλυση της `Driver.age` έχει γίνει για τα 4 έτη συνδυαστικά. Φυσικά, αυτή η συμπεριφορά θα πρέπει να συμβαίνει και σε κάθε έτος ξεχωριστά. Επαναλαμβάνεται το Διάγραμμα 4.7 για τα 4 έτη ξεχωριστά στο Διάγραμμα 4.8. Η γενικότερη τάση μείωσης της συχνότητας φαίνεται να επαναλαμβάνεται στα 3 τελευταία έτη με το 2015 να είναι ουδέτερο.

Ομοίως με πιο πάνω, εξετάζεται και η σχέση της συχνότητας έναντι της `Vehicle.age`. Στα Διάγραμματα 4.9 και 4.10 παρουσιάζεται η μέση συχνότητα ανά ηλικία του οχήματος για τα τέσσερα έτη συνδυαστικά αλλά και ξεχωριστά. Γενικότερα φαίνεται να υπάρχει ανοδική τάση της συχνότητας σε όλες τις περιπτώσεις. Ειδικότερα, τα δέκα πρώτα χρόνια η μέση συχνότητα είναι χαμηλότερη



Διάγραμμα 4.7: Μέση συχνότητα απαιτήσεων (μαύρη απεικόνιση) και έκθεση (μπλε απεικόνιση σε κλίμακα 10 χιλιάδων ετών) σε σχέση με ηλικία

των υπόλοιπων ετών. Μετά την περίοδο αυτή η συχνότητα έχει ανοδική πορεία ενώ μετά τα 20 χρόνια η συχνότητα παρουσιάζει έντονες αυξομειώσεις με τάση κυρίως καθοδική. Οι έντονες αυτές διακυμάνσεις ίσως να οδηγήσουν σε ουδέτερη σχέση της ηλικίας του οχήματος με τη συχνότητα απαιτήσεων αφού η μεταβλητότητα είναι μεγάλη. Δυστυχώς, για σκοπούς μίας ολοκληρωμένης μελέτης, σε κανένα έτος δεν υπάρχουν απαιτήσεις για οχήματα ηλικίας 43 και άνω ενώ ηλικίες οχήματος 56 με 62 δεν υπάρχουν στο δείγμα.

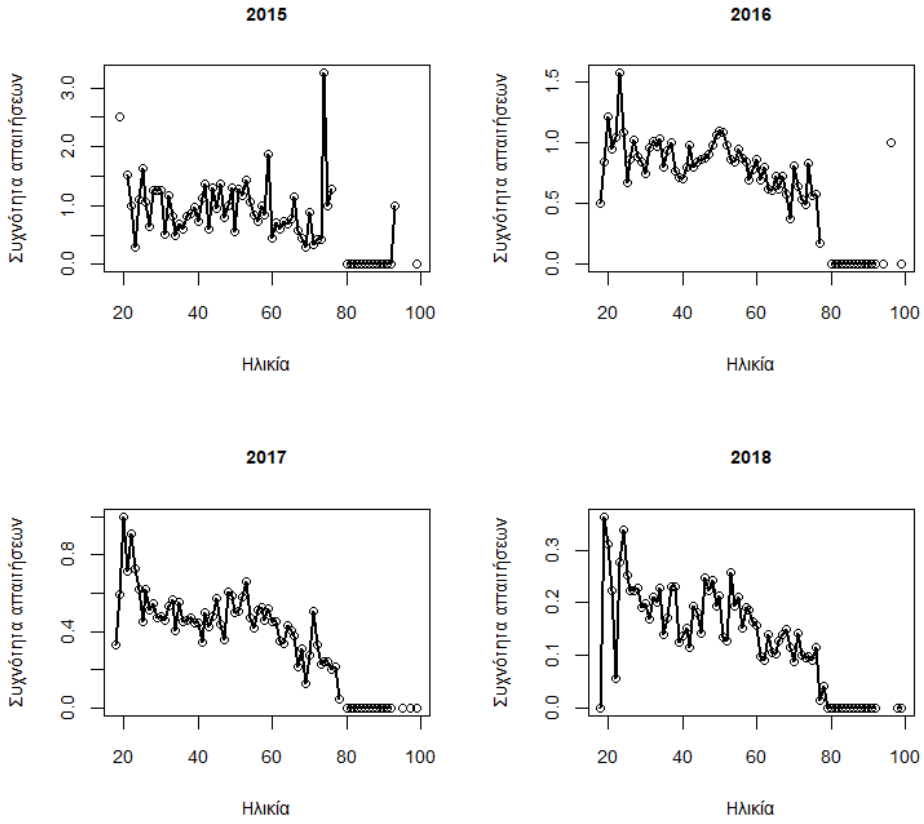
Παρόμοια μεταχείριση επιδέχεται και η μεταβλητή `Years.licenced` η οποία αντιπροσωπεύει τα χρόνια που ένας οδηγός κατέχει άδεια οδήγησης. Υπάρχουν συνολικά 63 διαφορετικές τιμές με εύρος 0 μέχρι 64 ενώ οι τιμές 61 και 63 δεν υπάρχουν στο δείγμα μάθησης. Στη γενική περίπτωση υπάρχει μειωτική τάση της μέσης συχνότητας απαιτήσεων με την αύξηση των ετών οδήγησης και αυτό διότι αποκτάται περισσότερη οδική εμπειρία.

Εστιάζεται τώρα η προσοχή στη μεταβλητή `Power`. Στο δείγμα μάθησης υπάρχουν 215 διαφορετικές τιμές, ξεκινώντας από τη χαμηλότερη 12 και φτάνοντας μέχρι την τιμή 560 χωρίς να έχουν την ίδια συχνότητα. Από το Διάγραμμα 4.13 δεν παρατηρείται κάποια συστηματική σχέση μεταξύ της δύναμης και της συχνότητας αφού οι τιμές κατανέμονται τυχαία. Άρα, η μεταβλητή `Power` δεν αποτελεί καλή υποψήφια για να συμμετάσχει στο μοντέλο που θα κατασκευαστεί για την εκτίμηση της συχνότητας αφού δεν παρέχει κάποια πληροφορία. Οι πέντε υψηλότερες μέσες συχνότητες είναι οι 4.00, 3.00, 2.25, 2.00 και 1.74 και αντιστοιχούν στις δυνάμεις 226, 234 και 315, 146, 235, 186. Σημειώστε επίσης τις τιμές με τη μεγαλύτερη έκθεση που φαίνονται στον πίνακα 4.16.

Δύναμη	105	100	75	110	90
Έκθεση	2764.36	3400.01	3520.15	4038.21	6328.11

Πίνακας 4.16: Τιμές δύναμης με την υψηλότερη έκθεση

Συνεχίζοντας, μελετάται η μεταβλητή `Length` και η σχέση της με τη συχνότητα των απαιτήσεων. Από το μεσαίο γράφημα του Διαγράμματος 4.13, στο οποίο φαίνεται η μέση συχνότητα απαιτήσεων

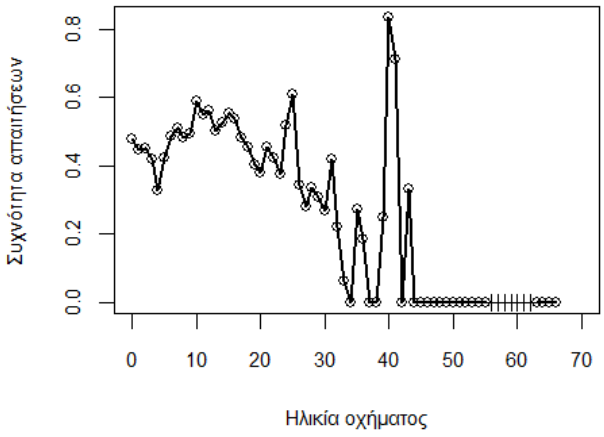


Διάγραμμα 4.8: Μέση συχνότητα απαιτήσεων σε σχέση με ηλικία και έτος

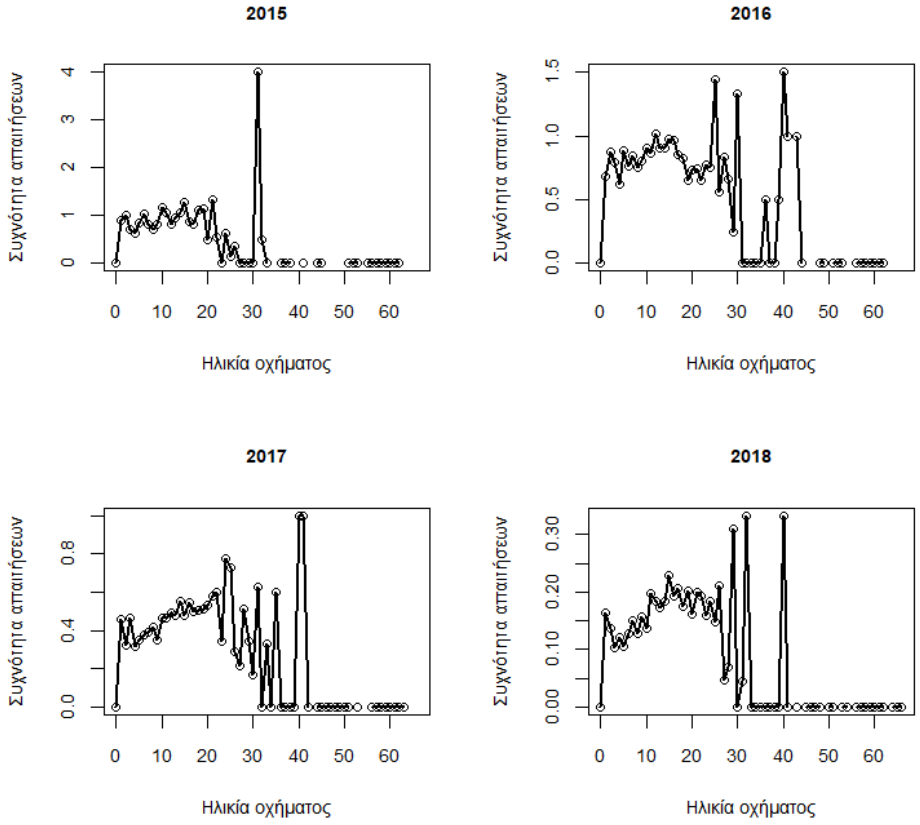
για κάθε τιμή μήκους, υπάρχει συστηματική σχέση μεταξύ του μήκους των οχημάτων και της συχνότητας απαιτήσεων. Σε πρώτη ανάγνωση λοιπόν, η μεταβλητή `Length` είναι καλή υποψήφια για το μοντέλο συχνότητας. Για την κατασκευή του, το μήκος έχει προσεγγιστεί προς το κοντινότερο μισό. Παρατηρείται ικανοποιητική γραμμική σχέση μεταξύ της μέσης συχνότητας απαιτήσεων και του μήκους και άρα αύξηση της αναμενόμενης συχνότητας με την αύξηση του μήκους. Επομένως, η μεταβλητή `Length` είναι μία καλή μεταβλητή για να συμμετάσχει στο μοντέλο.

Τέλος, γίνεται διερεύνηση της μεταβλητής `Weight` σε σχέση με τη συχνότητα των απαιτήσεων. Η συμπεριφορά της `Weight` φαίνεται στο τρίτο γράφημα του Διαγράμματος 4.13. Η αδύναμη γραμμική σχέση μεταξύ της αναμενόμενης συχνότητας και του βάρους του οχήματος δείχνει ότι δεν είναι καλή υποψήφια για το μοντέλο. Αύξηση του βάρους δε σημαίνει αύξηση της αναμενόμενης συχνότητας. Είναι σημαντικό το ότι η μεταβλητές `Length` και `Weight` είναι υψηλά θετικά συσχετισμένες όπως και αναμενόταν, με συντελεστή συσχέτισης 0.82. Επομένως, κατά τη μοντελοποίησή δε συνίσταται να συμπεριληφθούν και οι δύο.

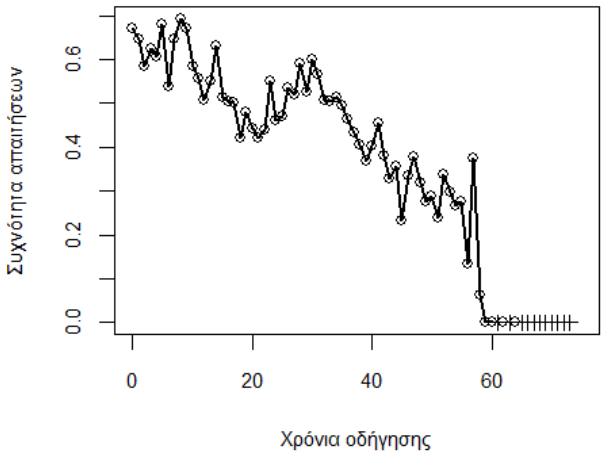
Η διερευνητική ανάλυση έγινε με σκοπό να εντοπιστούν μεταβλητές οι οποίες σχετίζονται με τη συχνότητα των απαιτήσεων. Αυτό έγινε με τη χρήση διαγραμμάτων, απεικονίσεων και πινάκων. Τα κύρια ευρήματα της διερευνητικής ανάλυσης αφορούν κυρίως τις μεταβλητές ηλικίας. Αρχικά, παρατηρήθηκε η διαφοροποίηση της συχνότητας απαιτήσεων σε σχέση με την ηλικία του ασφαλιζόμενου. Η διερευνητική ανάλυση έδειξε ότι νεότεροι και γηραιότεροι ασφαλιζόμενοι έχουν την τάση να αναφέρουν περισσότερες απαιτήσεις σε σχέση με μεσήλικες. Από την άλλη φάνηκε να υπάρχει τοπική κορύφωση της συχνότητας στους ασφαλιζόμενους αυτούς. Επίσης, εξετάστηκε και



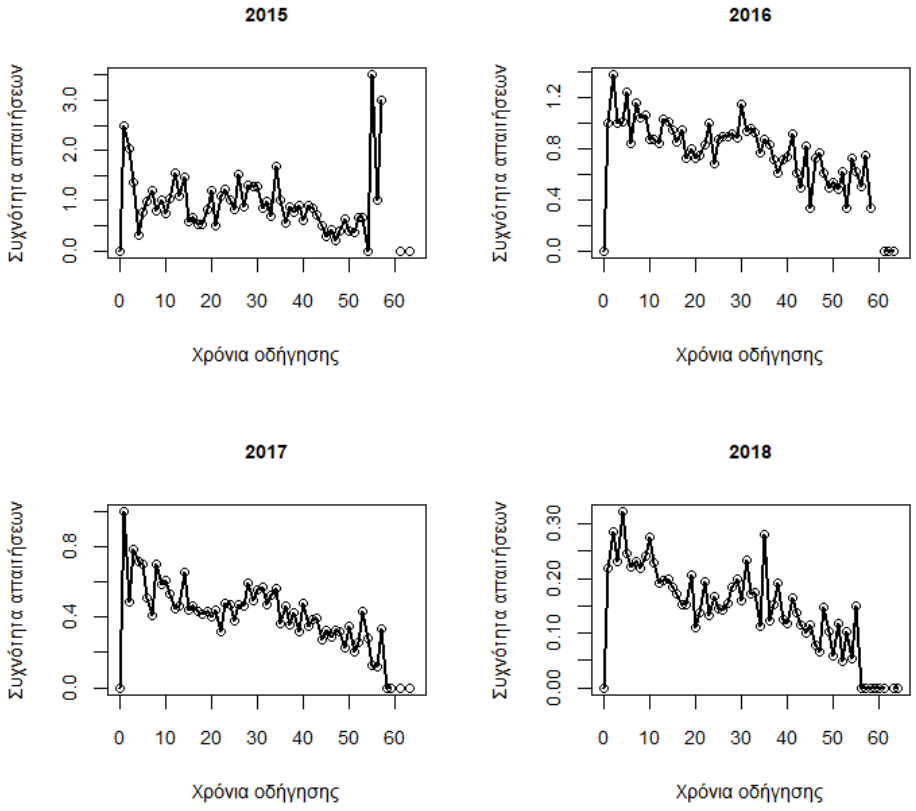
Διάγραμμα 4.9: Μέση συχνότητα απαιτήσεων σε σχέση με ηλικία οχήματος



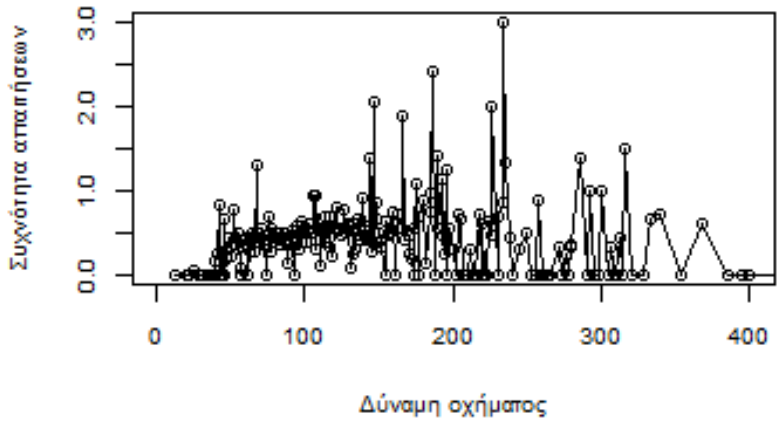
Διάγραμμα 4.10: Μέση συχνότητα απαιτήσεων σε σχέση με ηλικία οχήματος και έτος

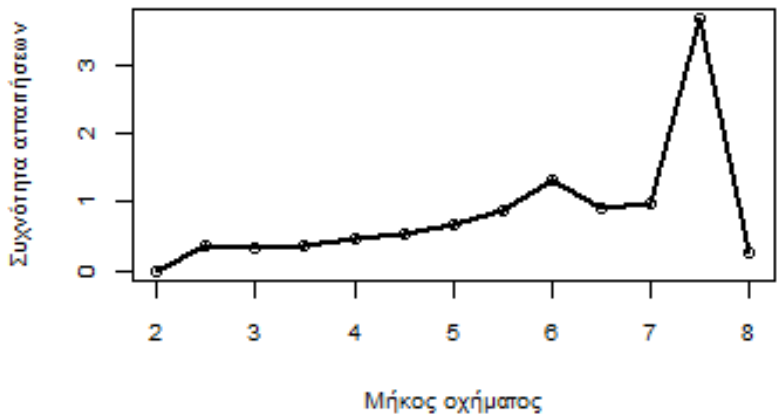


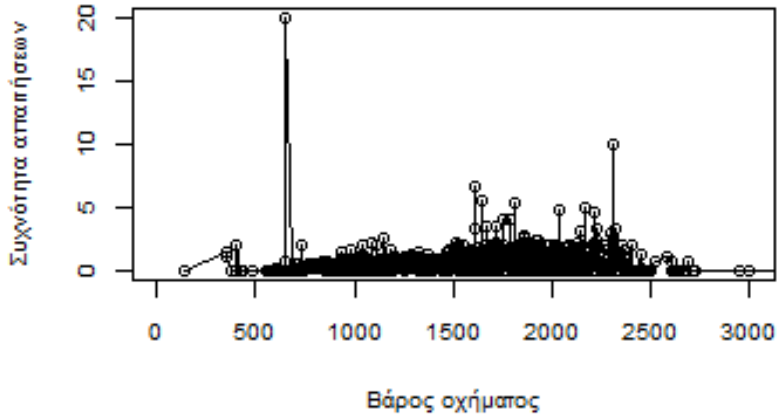
Διάγραμμα 4.11: Μέση συχνότητα απαιτήσεων σε σχέση με έτη οδήγησης



Διάγραμμα 4.12: Μέση συχνότητα απαιτήσεων σε σχέση με έτη οδήγησης και έτος







Διάγραμμα 4.13: Μέση συχνότητα απαιτήσεων σε σχέση με τη δύναμη (πάνω γράφημα), το μήκος (μεσαίο γράφημα) και το βάρος οχήματος (κάτω γράφημα)

κρίθηκε σημαντική η ηλικία του αυτοκινήτου και τα χρόνια οδήγησης. Νεότερα οχήματα φαίνεται να έχουν χαμηλότερη συχνότητα απαιτήσεων σε σχέση με παλαιοτέρα ενώ επίσης νεότεροι οδηγοί έχουν αυξημένη συχνότητα. Σημαντική επίσης κρίθηκε και η σχέση μεταξύ του μήκους και της συχνότητας όπως και εξηγήθηκε. Τέλος, φάνηκε και η σημαντικότητα των τριών κατηγορικών μεταβλητών τύπου οχήματος, καυσίμου και περιοχής με την κάθε μία να σχετίζεται διαφορετικά με τη συχνότητα των απαιτήσεων.

4.5.3 Poisson μοντέλο

Σε αυτό το στάδιο, θα κατασκευαστούν μοντέλα για τη μοντελοποίηση της συχνότητας των απαιτήσεων με χρήση της κατανομής Poisson. Η παρούσα ανάλυση θα αποτελέσει τη βάση για εντοπισμό των μεταβλητών που φέρουν την καλύτερη προσαρμογή στα δεδομένα του δείγματος μάθησης. Οι μεταβλητές που θα κριθούν καταλληλότερες θα χρησιμοποιηθούν στην κατασκευή πολλαπλών γενικευμένων μοντέλων αλλά και στη μετέπειτα μοντελοποίηση με τη χρήση των άλλων δύο κατανομών. Η επιλογή των μοντέλων γίνεται με βάση τις μεταβλητές που εξυπηρετούν τον σκοπό της εργασίας αλλά και την προεργασία που έχει γίνει. Η προσθήκη των μεταβλητών γίνεται διαδοχικά με βάση την τιμή της Deviance και του δείκτη AIC. Κάποιες καλές υποψήφιες για το μοντέλο, με βάση τη διερευνητική ανάλυση είναι για παράδειγμα οι `Year, Driver.age, Vehicle.age, Years.licenced, Area, Type.fuel` και `Type.risk`. Αρχικά θα κατασκευαστούν μοντέλα μίας μεταβλητής και αργότερα θα προστεθούν και άλλες μεταβλητές. Σε αυτή τη φάση θα πρέπει να γίνουν οι επιλογές της συνάρτησης σύνδεσης και της κατανομής της μεταβλητής απόκρισης. Επιλέγεται η χρήση της κατανομής Poisson με τη λογαριθμική συνάρτηση σύνδεσης. Για να γίνει η επιλογή της κατανομής αυτής θα πρέπει η μέση συχνότητα να είναι ίση με τη διασπορά της. Από τον Πίνακα 4.6, η μέση συχνότητα σε ολόκληρο το δείγμα είναι 49.63% ενώ στο δείγμα μάθησης είναι 49.72%. Επίσης, η διασπορά ολόκληρου του δείγματος είναι 1.49 και στο δείγμα μάθησης ίση με 1.52. Οι τιμές αυτές δείχνουν πως η υπόθεση της ισότητας μεταξύ μέσης τιμής και διασποράς μάλλον παραβιάζεται. Υπάρχει υπερμεταβλητότητα των δεδομένων και ίσως το μοντέλο Poisson να μην είναι το καταλληλότερο. Η συμπεριφορά όμως αυτή είναι θεμιτή σε πραγματικά δεδομένα ενώ αργότερα θα γίνει προσπάθεια αντιμετώπισης της. Προς το παρόν θεωρείται το Poisson μοντέλο. Η εμπειρική μέση συχνότητα και διασπορά υπολογίζονται αντίστοιχα από τις σχέσεις

$$m_Y = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} w_i} = 0.49 \quad \text{και} \quad S_Y^2 = \frac{\sum_{i=1}^{n} (Y_i - m_Y\, w_i)^2}{\sum_{i=1}^{n} w_i} = 1.52$$

με Y_i τον παρατηρούμενο αριθμό απαιτήσεων σε διάρκεια w_i. Επίσης, εκτιμάται η παράμετρος μεταβλητότητας ϕ η οποία είναι τέτοια ώστε $S_Y^2 = \phi\, m_Y$, με τιμή 3.06. Θυμηθείτε ότι στην κατανομή Poisson η $\phi = 1$.

Το απλούστερο Poisson μοντέλο που μπορεί να προσαρμοστεί στα δεδομένα είναι το μοντέλο στο οποίο συμμετέχει ο σταθερός όρος και ο όρος διόρθωσης με τη χρήση της `Exposure`. Προσαρμόζεται με τη χρήση της εντολής `glm()` και αποθηκεύεται στο αντικείμενο `poisson1`. Η διόρθωση μπορεί να γίνει στο μοντέλο μέσω της παραμέτρου `offset=log(Exposure)`, αφού γίνεται χρήση της λογαριθμικής συνάρτησης σύνδεσης. Ο όρος `offset=log(Exposure)` δε λαμβάνει συντελεστή στο μοντέλο αλλά επηρεάζει έμμεσα την ερμηνεία των αποτελεσμάτων. Θυμηθείτε ότι ο αναμενόμενος αριθμός απαιτήσεων σε διάρκεια w_i είναι $w_i\mu_i$ με μ_i να λαμβάνεται όταν το συμβόλαιο διήρκησε ένα ολόκληρο έτος. Η εκτίμηση του σταθερού όρου, -0.698, είναι στην κλίμακα του γραμμικού εκτιμητή

(λογαριθμική). Επομένως, στην κλίμακα της μεταβλητής απόκρισης η εκτιμώμενη συχνότητα είναι

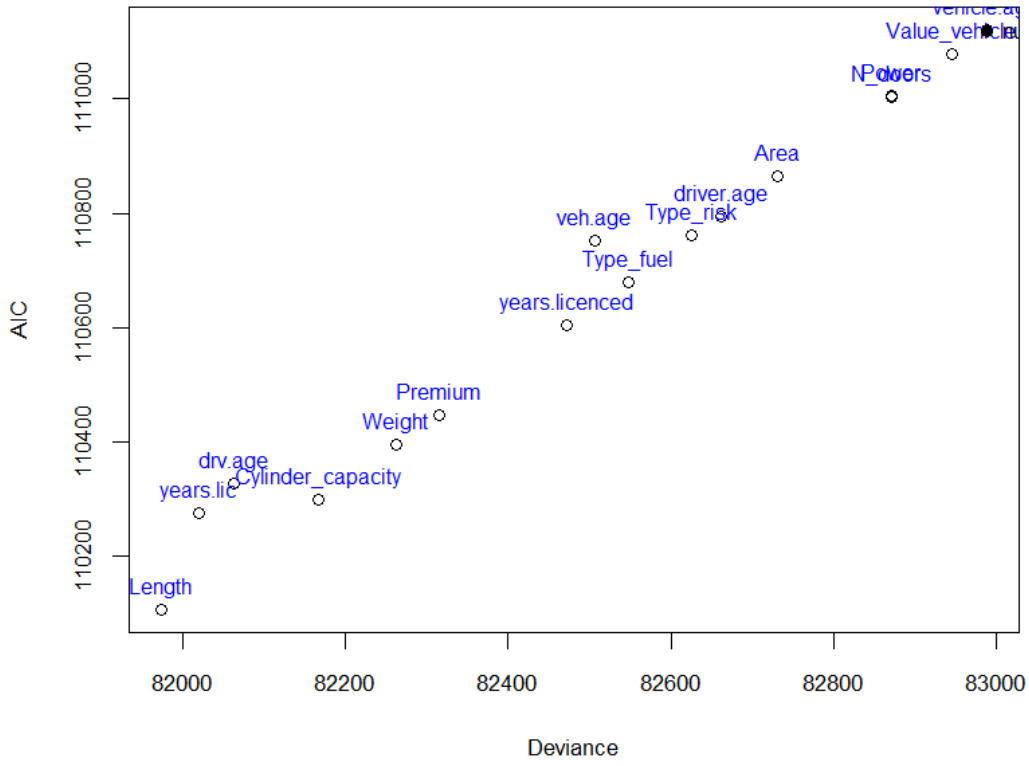
$$e^{-0.698} = 0.497$$

η οποία και αντιστοιχεί στη μέση δειγματική συχνότητα (0.497) όπως και θα έπρεπε.

```
> summary(poisson1)

Call:
glm(formula = N.claims.year ~ 1, family = poisson(link = log),
    data = train.data, offset = log(Exposure))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9973  -0.9973  -0.9973  -0.7052  12.1189

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.698594   0.006452  -108.3   <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *
0.05    .    0.1         1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 82988  on 49291  degrees of freedom
Residual deviance: 82988  on 49291  degrees of freedom
AIC: 111119

Number of Fisher Scoring iterations: 6
```

Σε αυτό το σημείο θα θεωρηθεί ότι το μοντέλο `poisson1` είναι ένα καλό μοντέλο για τα δεδομένα ενώ επιθυμία είναι να βρεθεί ποιό είναι εκείνο το μοντέλο μίας μεταβλητής το οποίο παρέχει καλύτερη προσαρμογή στο διαθέσιμο δείγμα μάθησης. Η επιλογή του κατάλληλου μοντέλου θα γίνει με τη χρήση του δείκτη AIC και της ελεγχοσυνάρτησης Deviance. Η χρήση της δεύτερης έχει ουσία όταν η σύγκριση γίνεται μεταξύ μοντέλων ιδίων βαθμών ελευθερίας και έτσι η χρήση του δείκτη AIC είναι προτιμότερη. Στο Διάγραμμα 4.14 φαίνεται ο δείκτης AIC στον y-άξονα και η Deviance στον x-άξονα για διάφορα μοντέλα που περιέχουν μόνο τις μεταβλητές που αναγράφονται και τον σταθερό όρο. Οι καταλληλότερες μεταβλητές θα είναι αυτές που εμφανίζονται στο κάτω αριστερό μέρος του διαγράμματος όπου ο AIC και η Deviance λαμβάνουν τις χαμηλότερες τιμές. Σε αυτό το μέρος βρίσκονται οι `Year`, `Drv.age` και `Years.lic`. Από την άλλη σε μαύρο σημείο εμφανίζεται το null μοντέλο. Αυτό είναι το πιο απλό μοντέλο και το πιο ακατάλληλο τελικά αφού περιέχει μόνο τον σταθερό όρο. Από το Διάγραμμα 4.14 παρατηρείτε πως υπάρχουν και μεταβλητές που τοποθετούνται κοντά στο κενό μοντέλο όπως η `Value.vehicle, N.doors, Vehicle.age` και `Power`.

Αρχικά, επιλέγονται οι μεταβλητές `Driver.age` και `Drv.age` που βρίσκονται στο κάτω αριστερό μέρος με τιμές (82662, 110794) και (82064, 110328). Με βάση τη Deviance, η κατηγορική μεταβλητή `Drv.age` φαίνεται να έχει καλύτερη προσαρμογή στα δεδομένα. Με βάση τον δείκτη AIC η κατηγορική μεταβλητή `Drv.age` φαίνεται να έχει πάλι καλύτερη προσαρμογή. Όμως η Deviance δε



Διάγραμμα 4.14: AIC και η Deviance για μοντέλα μίας μεταβλητής. Στο Διάγραμμα δεν εμφανίζεται η μεταβλητή `Year` με τιμές (74473, 102609) για τη βελτίωση του διαγράμματος.

λαμβάνει υπόψη της την πολυπλοκότητα του μοντέλου. Το μοντέλο με την `Driver.age` περιέχει δύο παραμέτρους, ενώ το `Drv.age` 68. Προφανώς, δεν είναι γνωστό αν είναι και οι 68 σημαντικές. Ενδεχομένως κάποιες να είναι στατιστικά μη σημαντικές με αποτέλεσμα η αφαίρεση τους να μειώσει περισσότερο τον δείκτη AIC αλλά να αυξήσει τη Deviance χωρίς όμως να θέτει την `Driver.age` καταλληλότερη της αντίστοιχης κατηγορικής. Προσαρμόζοντας ένα μοντέλο στο δείγμα μάθησης με τη χρήση της `Drv.age` αποφέρει 69 παραμέτρους, 1 σταθερό όρο και 68 παραμέτρους ηλικίας. Η ηλικία βάσης είναι τα 40 έτη. Επειδή υπάρχει μόνο μία μεταβλητή στο μοντέλο κάθε παράμετρος αποτελεί τη συχνότητα για την αντίστοιχη ηλικία (ο σταθερός όρος αποτελεί τη συχνότητα της ηλικίας βάσης). Για παράδειγμα, από τη χρήση των δεδομένων οι οδηγοί ηλικίας 30 ετών έχουν συνολικό αριθμό απαιτήσεων 427 με συνολική έκθεση σε ρίσκο 870.25 χρόνια στο δείγμα μάθησης. Επομένως, η εμπειρική μέση συχνότητα θα είναι

$$\frac{\text{Αριθμός απαιτήσεων ατομών 30 χρονών}}{\text{Συνολική έκθεση ατομών 30 χρονών}} = \frac{427}{870.25} = 0.4906636,$$

ενώ από το μοντέλο η συχνότητα προκύπτει από:

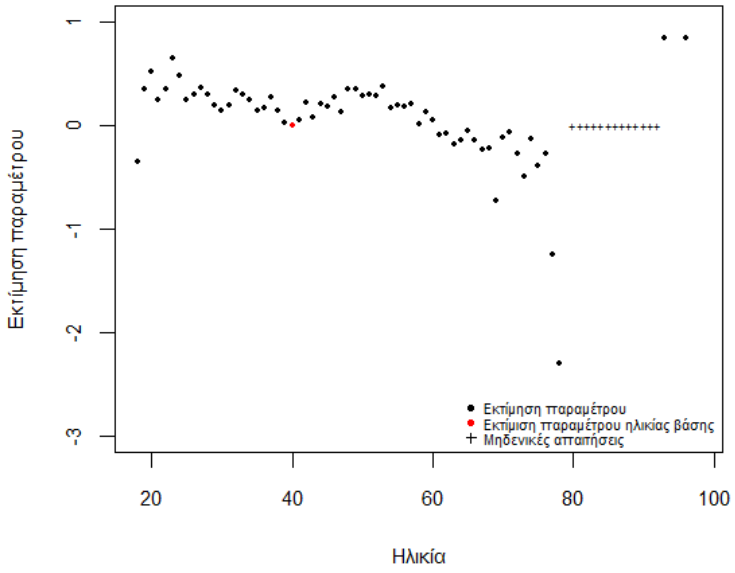
$$(\texttt{Intercept}) = -0.85460,$$

$$\texttt{Drv.age30} = 0.14260$$

και άρα η συχνότητα είναι

$$\exp(-0.85460 + 0.14260) = 0.4906619,$$

η ίδια εκτίμηση με την εμπειρική. Στο Διάγραμμα 4.15 φαίνεται η εκτίμηση της παραμέτρου για κάθε ηλικία. Ηλικίες που λείπουν από το δείγμα φαίνονται με σταυρό ενώ ο η συχνότητα στην ηλικία βάσης εμφανίζεται με κόκκινο χρώμα. Από το Διάγραμμα 4.15 λείπουν 5 τιμές που αντιστοιχούν στις ηλικίες 79, 94, 95, 97, 98 με τις εκτιμήσεις να είναι όλες ίσες με -11.45 και άρα οι συχνότητες απαιτήσεων θα είναι σχεδόν μηδενικές.



Διάγραμμα 4.15: Παράμετροι της `Drv.age` σε λογαριθμική κλίμακα

Σε σχέση με τη μεταβλητή `Years.lic`, από το Διάγραμμα 4.14 γίνεται αντιληπτό πως πάλι η αντίστοιχη κατηγορική μεταβλητή της `Years.licenced`, η `Years.lic`, είναι καταλληλότερη ακόμη και αν σε αυτή την περίπτωση το ένα μοντέλο αποτελείται από 65 παραμέτρους (1 σταθερό όρο και 64 για τα χρόνια) ενώ στην άλλη από 2. Τα αποτελέσματα του μοντέλου με τη χρήση της δεύτερης δείχνουν πως όχι όλες οι μεταβλητές είναι στατιστικά σημαντικές. Αυτό οδηγεί στο να επιλεγεί τελικά η ποσοτική μεταβλητή `Years.licenced` ή να κατασκευαστεί μία νέα κατηγορική μεταβλητή με σκοπό να διερευνηθεί αν μπορεί να αναχθεί η επιθυμητή σημαντικότητα. Για παράδειγμα, κατασκευάζεται η `Years.lic.cut` με τον πιο κάτω τρόπο. Η μεταβλητή αυτή κατηγοριοποιεί τα χρόνια οδήγησης σε 9 κατηγορίες. Με βάση αυτήν, οι μεταβλητές αυτές συνεχίζουν να είναι στατιστικά μη σημαντικές εκτός της `Years.lic.cut8+`. Αυτό δίνει ένα έναυσμα για επιπρόσθετη διερεύνηση με τη δημιουργία μίας δύτιμης μεταβλητής που θα δείχνει αν τα χρόνια οδήγησης είναι λιγότερα ή όχι των 8 ετών. Μοντελοποιώντας με αυτή τη μεταβλητή ανακτάται η σημαντικότητα.

```
years.intervals <- c(0, 1, 2, 3, 4, 5, 6, 7, 8, Inf)
data$Years.lic.cut <- cut(data$Years.licenced, Years.intervals,
labels = c('0', '1', '2', '3', '4', '5', '6', '7', '8+'),
right = FALSE)
```

Μέχρι τώρα προσαρμόστηκαν μοντέλα μίας μεταβλητής στα διαθέσιμα δεδομένα χωρίς όμως να λαμβάνεται υπόψη η σημαντικότητα των συντελεστών των μεταβλητών και άλλες πληροφορίες που παρέχονται από τα μοντέλα. Έχοντας μία αρχική ιδέα για το ποιες μεταβλητές είναι πιο κατάλληλες με βάση την κατασκευή απλών μοντέλων αλλά και από τη διερευνητική ανάλυση, μπορούν να αναπτυχθούν γενικευμένα μοντέλα πολλαπλών μεταβλητών. Από το Διάγραμμα 4.14 οι καλύτερες μεταβλητές είναι οι `Year`, `Length`, `Years.lic`, `Drv.age`, `Cylinder.capacity`, `Weight`, `Premium`, `Years.licenced`, `Type.fuel`, `Veh.age` και `Type.risk`. Η συμπερίληψη άλλων μεταβλητών εκτός αυτών δεν απορρίπτεται και θα διερευνηθεί.

Για τη μεταβλητή `Year` η πρώτη εντύπωση δείχνει πως θα πρέπει να συμπεριληφθεί στο μοντέλο αφού φέρει την καλύτερη προσαρμοστικότητα στα δεδομένα κατά την κατασκευή απλών `Poisson` μοντέλων. Με βάση αυτή τη μεταβλητή μπορεί να γίνει διαχείριση της έννοιας του χρόνου αφού αντιπροσωπεύει το ημερολογιακό έτος στο οποίο έχει συναφθεί ένα συμβόλαιο. Από την περιγραφή των μεταβλητών διαπιστώθηκε η έντονη μεταβολή της συχνότητας των απαιτήσεων με το πέρασμα των ετών και άρα εκ πρώτης άποψης το έτος σχετίζεται σημαντικά με αυτή. Από την άλλη, η μεγάλη αυτή μεταβολή στις μέσες συχνότητες των απαιτήσεων ανά έτος δημιουργεί ερωτήματα σχετικά με το πως μπορεί να ερμηνευθεί μία τόσο σημαντική μεταβολή. Αν για παράδειγμα εξεταστεί η περίπτωση από το 2016 μέχρι το 2017 όπου η συχνότητα μειώθηκε κατά $1 - \frac{0.47}{0.86} = 45.35\%$ αυτό μπορεί να οφείλετε στην αλλαγή πολιτικής της ασφαλιστικής εταιρίας να ασφαλίζει πολύ πιο δύσκολα και να επιλέγει χαμηλότερου ρίσκου ασφαλιζόμενους. Στη συνέχεια αυτού, το 2018 η συχνότητα μειώνεται επιπλέον κατά $1 - \frac{0.17}{0.47} = 63.83\%$ σε σχέση με το προηγούμενο έτος και αυτό δημιουργεί ξανά ερωτήματα. Από το 2016 στο 2018 η συχνότητα απαιτήσεων μειώνεται κατά $1 - \frac{0.17}{0.86} = 80.23\%$ και αυτό θεωρείται επίσης προβληματικό. Δεν είναι ρεαλιστική η υπόθεση αλλαγής πολιτικής της εταιρίας επί δύο έτη αλλά ούτε και της καθυστερημένης αναφοράς απαιτήσεων σε σχέση με τον χρόνο του γεγονότος αποζημίωσης. Επομένως, φαίνεται πως η χρήση της `Year` φέρει ερωτήματα σχετικά με την ερμηνεία και έτσι θεωρείται πως δημιουργεί έναν αδικαιολόγητο θόρυβο ο οποίος θα θέσει τα μοντέλα εκτός πραγματικότητας. Επιλέγεται λοιπόν να μείνει εκτός παρά την πολύ καλή προσαρμοστικότητα που φέρει.

Ξεκινώντας την κατασκευή μοντέλων πολλαπλών μεταβλητών, η `Year` είναι η καλύτερη επιλογή για αρχή. Η επιλογή αυτή εξυπηρετεί επίσης στην άμεση εκτίμηση απαιτήσεων με βάση το έτος. Επιλέγεται όμως να μείνει εκτός όπως συζητήθηκε προηγουμένος. Επίσης, λόγω του πολύ μικρού δείγματος σε οχήματα τύπου 1 και 4, αυτά αφαιρούνται με σκοπό να μην επηρεάσουν την ανάλυση μας. Όλα τα πιο κάτω αποτελέσματα αφορούν το νέο δείγμα μάθησης και νέο δείγμα ελέγχου, `train.data` και `test.data` αντίστοιχα. Το νέο δείγμα μάθησης αποτελείτε από 49,277 εγγραφές και το δείγμα ελέγχου από 32,465.

Η επόμενη υποψήφια με τη βέλτιστη προσαρμογή είναι η `Drv.age` και θα εξεταστεί η προσθήκη της στο `poisson1`, δημιουργώντας το `poisson2` μοντέλο. Κατασκευάζεται λοιπόν το μοντέλο `poisson2` με τη χρήση της `Drv.age`, του σταθερό όρου και του όρου `offset`. Τα αποτελέσματα δείχνουν πως δεν είναι όλες οι εκτιμήσεις των συντελεστών στατιστικά σημαντικές. Η ανάλυση διασποράς για την προσθήκη της `Drv.age` στο `poisson1` φαίνεται στον Πίνακα 4.17 και κρίνεται ως σημαντική. Στο Διάγραμμα 4.15 φαίνονται οι εκτιμήσεις των συντελεστών του μοντέλου και με κόκκινο χρώμα η εκτίμηση του συντελεστή ηλικίας βάσης στο μηδέν. Παρατηρώντας το αναγνωρίζονται τάσεις των εκτιμήσεων σε διάφορες ηλικιακές περιόδους. Παρατηρείτε αύξηση στις ηλικίες 18 με 25, μείωση για 26 με 44, αύξηση για 46 με 59 και ξανά μείωση για ηλικίες 60 και άνω. Κατασκευάζεται λοιπόν η

κατηγορική μεταβλητή ηλικίας `Driver.age.cut` με κατηγορίες [18,24], [25,44], [45,59], [60,69] και [70,`Inf`) και δοκιμάζεται η προσαρμογή του μοντέλου `poisson3`. Κατηγορία βάσης είναι η ηλικιακή περίοδος [18, 24]. Θα μπορούσε επίσης η τελευταία κατηγορία να διαχωριστεί σε [70, 79] και [80, `Inf`) λόγο της διαφοράς στη συχνότητα στις πιο μεγάλες ηλικίες. Αυτό δεν επιλέγεται καθώς θα σήμαινε υπερπροσαρμογή στο δείγμα μάθησης όπως φέρει και η χρήση της `Drv.age`. Με την εντολή `anova(poisson1, poisson3)` λαμβάνονται τα αποτελέσματα του χ^2 ελέγχου ο οποίος συγκρίνει τα δύο μοντέλα. Οι βαθμοί ελευθερίας της κατανομής είναι η διαφορά στο πλήθος των επεξηγηματικών μεταβλητών που χρησιμοποιούνται. Ο έλεγχος αυτός είναι ο ίδιος έλεγχος που παρουσιάστηκε στην Ενότητα 2.4.2. Στον Πίνακα 4.18 φαίνονται τα αποτελέσματα της ανάλυσης διασποράς για την προσθήκη της `Driver.age.cut` στο μοντέλο `poisson1`. Η προσθήκη της `Driver.age.cut` είναι απαραίτητη αφού μειώνει την Deviance κατά 548.30 μονάδες ενώ η μείωση αυτή θεωρείται σημαντική με βάση τον X^2 έλεγχο ο οποίος λαμβάνει υπόψη του τους βαθμούς ελευθερίας των δύο μοντέλων. Στον Πίνακα 4.17 φαίνεται η ανάλυση διασποράς για την προσθήκη της κατηγορικής μεταβλητής `Drv.age` στον οποίο φαίνεται πως και αυτή η μεταβλητή κρίνεται σημαντική αφού μειώνει την Deviance περισσότερο από την `Driver.age.cut` και συγκεκριμένα κατά 921.04 μονάδες. Από την άλλη ο δείκτης AIC δε φαίνεται να αλλάζει σημαντικά. Τελικά, επιλέγεται η χρήση της `Driver.age.cut` λόγω της απλότητας της και άρα επιλέγεται το μοντέλο `poisson3`. Η χρήση της συνεχούς μεταβλητής `Driver.age` φέρει τη φτωχότερη προσαρμοστικότητα και δεν παρουσιάζεται.

```
> summary(poisson3)

Call:
glm(formula = N.claims.year ~ Driver.age.cut,
family = poisson(link = log),
data = train.data, offset = log(Exposure))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1626  -1.0410  -1.0138  -0.7361  12.0523

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.39187    0.04256  -9.207  < 2e-16 ***
Driver.age.cut[25, 44]  -0.27386    0.04366  -6.273 3.55e-10 ***
Driver.age.cut[45, 59]  -0.22100    0.04375  -5.051 4.39e-07 ***
Driver.age.cut[60, 69]  -0.59732    0.04737 -12.609  < 2e-16 ***
Driver.age.cut[70, Inf) -0.76752    0.05460 -14.057  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *
0.05    .    0.1         1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 82973  on 49276  degrees of freedom
Residual deviance: 82417  on 49272  degrees of freedom
AIC: 110556
```

```
Number of Fisher Scoring iterations: 6
```

Μοντέλο	Resid. Df	Resid. Dev	Df	Deviance	$\mathbb{P}(\chi^2_{67} > 924.56)$
1	49276	82973			
2	49209	82052	67	921.04	0

Πίνακας 4.17: Ανάλυση διασποράς για την προσθήκη της `Drv.age`

Μοντέλο	Resid. Df	Resid. Dev	Df	Deviance	$\mathbb{P}(\chi^2_4 > 558.91)$
1	49276	82973			
3	49272	82417	4	556.30	0

Πίνακας 4.18: Ανάλυση διασποράς για την προσθήκη της `Driver.age.cut`

Επόμενη υποψήφια για την προσθήκη στο μοντέλο `poisson3` είναι η μεταβλητή `Years.licenced`, δημιουργώντας το μοντέλο `poisson4`. Ενδιαφέρον είναι το ότι οι δύο μεταβλητές `Years.licenced` και `Driver.age` είναι υψηλά θετικά συσχετισμένες, κατά 88%. Αυτό φαίνεται να επηρεάζει τα αποτελέσματα αφού η προσθήκη της `Years.licenced` στο μοντέλο αυξάνει τις p-τιμές των ελέγχων των συντελεστών των επεξηγηματικών μεταβλητών `Driver.age.cut[25, 44]`, `Driver.age.cut[60, 69]` και `Driver.age.cut[70, Inf)` με αποτέλεσμα να μην υπάρχουν σοβαρές ενδείξεις για να απορρίψουμε τις μηδενικές υποθέσεις ότι οι τρεις συντελεστές είναι μηδέν. Επιλέγεται τελικά να μη συμπεριληφθεί η μεταβλητή αυτή αφού η επίδραση της μπορεί να επεξηγηθεί από τη μεταβλητή ηλικίας και άρα δεν επιλέγεται το μοντέλο `poisson4`.

Μέχρι σε αυτό το στάδιο της ανάλυσης έχουν προσαρμοστεί 4 διαφορετικά μοντέλα και απορρίφθηκε η συμπερίληψη των `Drv.age` και `Years.licenced`. Καταλληλότερο μοντέλο είναι μέχρι στιγμής το `poisson3` με τη μεταβλητή `Driver.age.cut`, τον σταθερό όρο και τον όρο `offset`. Θα εξεταστεί η προσθήκη της ποσοτικής μεταβλητής `Vehicle.age` προσαρμόζοντας το μοντέλο `poisson5`. Η προσθήκη της δεν κρίνεται σημαντική με βάση την τιμή 0.61 του στατιστικού X^2 της ανάλυσης διασποράς που φαίνεται στον Πίνακα 4.19 και άρα παραμένει ως καταλληλότερο το `poisson3`.

Όπως έχει παρατηρηθεί από πριν, η συχνότητα απαιτήσεων για οχήματα ηλικίας 15 ετών και άνω παρουσιάζει πολύ έντονες αυξομειώσεις σε σχέση με οχήματα ηλικίας μέχρι 15 ετών. Έτσι, δοκιμάζεται η χρήση της δύτιμης κατηγορικής μεταβλητής `Vehicle.age.cut` η οποία χωρίζει τις εγγραφές ανάλογα με το αν η χρήση του οχήματος είναι μεγαλύτερη ή όχι των 15 ετών. Κατά την εκτέλεση της ανάλυσης διασποράς, κρίνεται πως η προσθήκη της μεταβλητής αυτής είναι επίσης ασήμαντη και έτσι επιλέγεται ο διαχωρισμός να γίνει με βάση τα 5 πρώτα έτη. Κατασκευάζεται λοιπόν το μοντέλο `poisson6` το οποίο εξετάζεται αν είναι καταλληλότερο του `poisson3`. Η χρήση της νέας κατηγοριοποίησης φέρει σημαντική μεταβολή στην Deviance κατά 100.43 μονάδες και αυτή κρίνεται στατιστικά σημαντική με βάση τα αποτελέσματα στον Πίνακα 4.20, θεωρώντας το μοντέλο `poisson6` ως το μέχρι στιγμής καταλληλότερο. Υπάρχει δηλαδή μεταβολή στην αναμενόμενη συχνότητα απαιτήσεων όταν το όχημα είναι ηλικίας 6 ετών και άνω.

```
Veh.age.intervals <- c(0, 5, Inf)
```

```
data$Vehicle.age.cut <- cut(data$Vehicle.age, Veh.age.intervals,
                            labels = c('5-','6+'), right = FALSE)
```

Μοντέλο	Resid. Df	Resid. Dev	Df	Deviance	$\mathbb{P}(\chi^2_1 > 0.5415)$
3	49272	82417			
5	49271	82316	1	0.61	0.44

Πίνακας 4.19: Ανάλυση διασποράς για την προσθήκη της `Vehicle.age`

Μοντέλο	Resid. Df	Resid. Dev	Df	Deviance	$\mathbb{P}(\chi^2_1 > 102.38)$
3	49272	82417			
6	49271	82315	1	102.43	0

Πίνακας 4.20: Ανάλυση διασποράς για την προσθήκη της `Vehicle.age.cut`

Αν και υπάρχουν μεταβλητές οι οποίες από το Διάγραμμα 4.14 είναι σημαντικότερες των ήδη επιλεγέντων, όπως οι `Weight` και `Length`, οι επιλογές των μεταβλητών είναι τέτοιες έτσι ώστε πρώτα να εξυπηρετούν το ερευνητικό ερώτημα το οποίο και αποσκοπεί στην εκτίμηση μελλοντικών απαιτήσεων. Αυτό απαιτεί μεταβλητές οι οποίες είναι δυναμικές ως προς τον ίδιο ασφαλιζόμενο όπως οι μεταβλητές ηλικίας και όχι οι μεταβλητές που αφορούν το όχημα που θεωρούνται στατικές για κάθε ασφαλιζόμενο. Θα εξεταστεί η προσθήκη της ποσοτικής μεταβλητής `Length` στο μέχρι τώρα καταλληλότερο μοντέλο, `poisson6`, διαμορφώνοντας το νέο μοντέλο `poisson7`. Νωρίτερα, είχε παρατηρηθεί ότι η μεταβλητή αυτή είναι μία αρκετά καλή επιλογή να συμπεριληφθεί στο μοντέλο συχνότητας με βάση και το τρίτο γράφημα του Διαγράμματος 4.13. Η ανάλυση διασποράς στον Πίνακα 4.21, δείχνει ότι η προσθήκη της `Length` κρίνεται στατιστικά σημαντική. Ο δείκτης AIC μειώνεται κατά 1004.53 μονάδες και αυτή η μεταβολή είναι η μεγαλύτερη μέχρι στιγμής. Επομένως, επιλέγεται να προστεθεί η μεταβλητή `Length`, καταλήγοντας στο μοντέλο `poisson7`.

Μοντέλο	Resid. Df	Resid. Dev	Df	Deviance	$\mathbb{P}(\chi^2_1 > 1003.1)$
6	49271	82315			
7	49270	81308	1	1006.5	0

Πίνακας 4.21: Ανάλυση διασποράς για την προσθήκη της `Length`.

Παρόλο που οι μεταβλητές `Weight` και `Length` είναι κατά 84.14% συσχετισμένες, η προσθήκη της `Weight` δεν αποδυναμώνει την παρουσία της `Length` στο νέο μοντέλο `poisson8`. Επιπλέον, με βάση την ανάλυση διασποράς στον Πίνακα 4.22, φαίνεται πως η προσθήκη της `Weight` μειώνει τη Deviance. Από την άλλη ο δείκτης AIC μειώνεται μόλις κατά 31.65 μονάδες στο μοντέλο `poisson8` σε σχέση με το `poisson7` και έτσι σε αυτή τη φάση επιλέγεται η μεταβλητή `Weight` να μείνει εκτός του μοντέλου. Ως καταλληλότερο μοντέλο παραμένει το `poisson7`.

Υπάρχουν ακόμη κάποιες ποσοτικές μεταβλητές οι οποίες δεν έχουν εξεταστεί, όπως η `Premium` και `Value.vehicle`. Μέχρι σε αυτό το στάδιο έχει κατασκευαστεί το μοντέλο `poisson7` με τις μεταβλητές `Driver.age.cut, Vehicle.age.cut` και `Length`. Η μεταβλητή `Premium` δε χρησιμοποιείται σε μοντέλα συχνότητας των απαιτήσεων αφού ένα τέτοιο μοντέλο αποτελεί τη βάση

Μοντέλο	Resid. Df	Resid. Dev	Df	Deviance	$\mathbb{P}(\chi^2_1 > 18.75)$
7	49270	81308			
8	49269	81274	1	33.65	0

Πίνακας 4.22: Ανάλυση διασποράς για την προσθήκη της `Weight`.

για την ανάπτυξη μοντέλων με μεταβλητή απόκρισης το ασφάλιστρο `Premium`. Τελευταία ποσοτική μεταβλητή που εξετάζεται, η `Value.vehicle`. Γίνεται η κατασκευή του μοντέλου `poisson9`. Η σημαντικότητα της `Value.vehicle` στο `poisson9` φαίνεται από τη p-τιμή του ελέγχου στα αποτελέσματα του πίνακα 4.23. Επιλέγεται όμως να μη συμπεριληφθεί στο μοντέλο καθώς ο δείκτης AIC μειώνεται μόλις 4.19 μονάδες και άρα είναι αμελητέος. Επομένως, καταλήγουμε ξανά στο μοντέλο `poisson7`.

Μοντέλο	Resid. Df	Resid. Dev	Df	Deviance	$\mathbb{P}(\chi^2_1 > 4.86)$
7	49270	81308			
9	49269	81302	1	6.20	0.012

Πίνακας 4.23: Ανάλυση διασποράς για την προσθήκη της `Value.vehicle`.

Μέχρι στιγμής έχει δοθεί η προσοχή στην εξέταση ποσοτικών μεταβλητών. Το μοντέλο `poisson7` που έχει διαμορφωθεί μέχρι τώρα, περιέχει τις `Driver.age.cut, Vehicle.age.cut` και `Length`. Τώρα το ενδιαφέρον στρέφεται στις τρεις υποψήφιες κατηγορικές μεταβλητές `Type.risk, Type.fuel` και `Area`. Όσο αφορά τη μεταβλητή `Type.risk` αυτή έχει θεωρηθεί πως είναι σημαντική κατά τη διερευνητική ανάλυση. Προσθέτουμε τώρα την `Type.risk` στο μοντέλο `poisson7` δημιουργόντας το `poisson10`. Όπως κρίνεται από την ανάλυση διασποράς, Πίνακας 4.24, η προσθήκη της είναι αναγκαία ενώ οι p-τιμές των ελέγχων για τις κατηγορίες `Type.risk1` και `Type.risk4` είναι πολύ υψηλές και άρα η μηδενική υπόθεση για τους συντελεστές των δύο αυτών εικονικών μεταβλητών δεν απορρίπτεται. Αυτό συμβαίνει λόγο του ότι υπάρχει μόνο μία εγγραφή με μη μηδενικό αριθμό απαιτήσεων που να ανήκει σε αυτές τις κατηγορίες ενώ υπάρχουν συνολικά μόλις 22 εγγραφές στο δείγμα μάθησης. Επομένως, με τη χρήση του παρόν δείγματος είναι αδύνατο να εκτιμηθεί η συχνότητα απαιτήσεων για οχήματα τύπου 1 και 4. Αυτός είναι και ο λόγος που επιλέχθηκε εξάρχής να αφαιρεθούν οι εγγραφές αυτές. Άρα η ανάλυση θα αφορά μόνο οχήματα τύπου 2 και 3, φορτηγά και επιβατικά οχήματα αντίστοιχα. Η ανάλυση διασποράς έδειξε ότι επιλέγουμε το μοντέλο `poisson10`.

Μοντέλο	Resid. Df	Resid. Dev	Df	Deviance	$\mathbb{P}(\chi^2_1 > 56.79)$
7	49270	81308			
10	49269	81267	1	40.98	0

Πίνακας 4.24: Ανάλυση διασποράς για την προσθήκη της `Type.risk`.

Στη συνέχεια εξετάζεται η προσθήκη της `Type.fuel` στο μοντέλο `poisson10` δημιουργώντας το `poisson11`. Από τον Πίνακα 4.12 παρατηρήθηκε ήδη αυξημένη συχνότητα απαιτήσεων όταν γίνεται χρήση καυσίμου πετρελαίου. Η προσθήκη της μεταβλητής `Type.fuel` φαίνεται από τον Πίνακα 4.25 να είναι σημαντική αφού με έναν επιπλέον βαθμό ελευθερίας κερδίζονται 120.14 μονάδες στην

Deviance του μοντέλου.

Μοντέλο	Resid. Df	Resid. Dev	Df	Deviance	$\mathbb{P}(\chi^2_1 > 120.14)$
10	49269	81267			
11	49268	81147	1	120.14	0

Πίνακας 4.25: Ανάλυση διασποράς για την προσθήκη της `Type.fuel`.

Τέλος, η προσθήκη της `Area` στο `poisson11` θεωρείται επίσης σημαντική όπως φαίνεται στον Πίνακα 4.26 ανάλυσης διασποράς, ενώ μειώνει την τιμή του δείκτη AIC κατά 278.98 μονάδες.

Μοντέλο	Resid. Df	Resid. Dev	Df	Deviance	$\mathbb{P}(\chi^2_1 > 280.99)$
11	49268	81147			
12	49267	80866	1	280.99	0

Πίνακας 4.26: Ανάλυση διασποράς για την προσθήκη της `Area`.

Με την πιο πάνω διαδικασία δημιουργήθηκε μία ακολουθία 12 μοντέλων, από το `poisson1` μέχρι το `poisson12`, όπου σε όλα χρησιμοποιήθηκε ο σταθερός όρος και το `offset`. Χρησιμοποιήθηκαν σε μεγάλο βαθμό οι δείκτες προσαρμογής AIC και Deviance ενώ επίσης έγινε η προσπάθεια κάθε συντελεστής να είναι σημαντικός με σκοπό να μη λείψει καμία μεταβλητή η οποία εξυπηρετεί τον σκοπό της εργασίας. Το τελικό μοντέλο, `poisson12`, φέρει δείκτη AIC ίσο με 109014 μονάδες και είναι ο χαμηλότερος που επιτεύχθηκε μέσα από την ανάλυση αυτή. Σε αυτό συμμετέχουν οι επεξηγηματικές μεταβλητές `Driver.age.cut, Vehicle.age.cut, Length, Type.risk, Type.fuel` και `Area`. Τα αποτελέσματα του τελικού μοντέλου `poisson12` παρουσιάζονται στην Ενότητα 5.1. Στον Πίνακα 4.27 φαίνονται όλα τα μοντέλα που έχουν κατασκευαστεί με τις αντίστοιχες μεταβλητές ενώ σε όλα υπάρχει ο σταθερός όρος.

4.5.4 Αρνητικό διωνυμικό μοντέλο

Η ανάγκη για την εξεύρεση κάποιου εναλλακτικού μοντέλου είναι εμφανής όταν παραβιάζεται η υπόθεση της ισότητας μεταξύ μέσης τιμής και διασποράς. Η μέση συχνότητα των απαιτήσεων είναι 0.497 ενώ η εμπειρική διασπορά είναι 1.52 γεγονός που θέτει την επιλογή της κατανομής Poisson ακατάλληλη αφού θα έπρεπε τα μεγέθη αυτά να είναι ίσα και εμφανίζεται έτσι υπερμεταβλητότητα

	Μεταβλητές											
Μοντέλο	Drv.age	Driver.age.cut	Years.licenced	Vehicle.age	Vehicle.age.cut	Length	Weight	Value.vehicle	Type.risk	Type.fuel	Area	Driver.age*
1												*
2	*											*
3		*										*
4		*	*									*
5		*		*								*
6		*			*							*
7		*			*	*						*
8		*			*	*	*					*
9		*			*	*		*				*
10		*			*	*			*			*
11		*			*	*			*	*		*
12		*			*	*			*	*	*	*

Πίνακας 4.27: Μεταβλητές των τριών τύπων μοντέλων

των πραγματικών δεδομένων. Μία ένδειξη της παρουσίας υπερμεταβλητότητας στα δεδομένα είναι όταν η ελεγχοσυνάρτηση Deviance είναι μεγαλύτερη των βαθμών ελευθερίας $n - p$, οι οποίοι είναι και η αναμενόμενη της τιμή αφού αυτή ακολουθεί X^2 κατανομή [11]. Επιλέγεται λοιπόν να υπολογιστεί ο λόγος της τιμής της ελεγχοσυνάρτησης Deviance προς τους βαθμούς ελευθερίας της, με το αποτέλεσμα να είναι $\frac{80866.01}{49267} = 1.64$. Η τιμή αυτή δείχνει την έντονη παρουσία υπερμεταβλητότητας που εμφανίζεται από το παρόν Poisson μοντέλο, `poisson12`. Τιμή άνω του ενός σημαίνει πως η μεταβλητότητα της μεταβλητής απόκρισης είναι μεγαλύτερη από αυτή που το μοντέλο εκτιμά, θέτοντας το ακατάλληλο. Θυμηθείτε ότι στο Poisson μοντέλο η παράμετρος μεταβλητότητας ϕ είναι ίση με τη μονάδα εξ ορισμού (2.2.3). Όπως έχει αναφερθεί στην ενότητα 3.2 η χρήση της Αρνητικής διωνυμικής κατανομής είναι καταλληλότερη της Poisson για τον λόγο ότι μπορεί να ανταποκριθεί καλύτερα στη μεταβλητότητα των πραγματικών δεδομένων λόγω της συνάρτησης διασποράς της. Πιο κάτω γίνεται ο στατιστικός έλεγχος `dispersiontest` της βιβλιοθήκης `AER`, για ύπαρξη υπερμεταβλητότητας [4]. Με τη χρήση της εντολής αυτής στο τελευταίο μοντέλο `poisson12` λαμβάνονται τα πιο κάτω αποτελέσματα.

```
> dispersiontest(poisson12, trafo = 2, alternative = 'greater')

        Overdispersion test

data:  poisson12
z = 31.262, p-value < 2.2e-16
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
3.725655
```

Η παράμετρος `trafo` αναφέρεται στη διασπορά της εναλλακτικής κατανομής και μπορεί να είναι αριθμός ή θετική συνάρτηση. Η χρήση της τιμής 2 υποδηλώνει τη χρήση της κατανομής NB2 ως εναλλακτικής αφού σε αυτή την περίπτωση η διασπορά αποτελείτε από έναν επιπρόσθετο τετραγωνικό μετασχηματισμό της μέσης τιμής της κατανομής. Η επιλογή `greater` υποδηλώνει ότι στην εναλλακτική υπόθεση $H_1 : \kappa > 0$, το οποίο εννοείται λόγω της επιλογής της NB2 και με μηδενική υπόθεση $H_0 : \kappa = 0$. Η τιμή της ελεγχοσυνάρτησης T είναι 31.26 ενώ η p-τιμή του ελέγχου είναι σχεδόν μηδέν, υποδηλώνοντας ισχυρές ενδείξεις κατά της μηδενικής υπόθεσης και άρα $\kappa \neq 0$, με εκτίμηση 3.72. Η τιμή αυτή μοιάζει να είναι η τιμή της ϕ που εκτιμήθηκε ίση με 3.06 κατά την έναρξη της μοντελοποίησης με τη χρήση της Poisson κατανομής. Επομένως, απορρίπτεται η μηδενική υπόθεση υπέρ της χρήσης Poisson και προσαρμόζονται μοντέλα με τη χρήση Αρνητικής διωνυμικής κατανομής. Αυτό γίνεται μέσω της εντολής `glm.nb` που παρέχεται από τη βιβλιοθήκη `MASS`. Προσαρμόζονται τα μοντέλα που προσαρμόστηκαν και στην περίπτωση της χρήσης κατανομής Poisson και δημιουργείται μία σειρά 12 μοντέλων καταλήγοντας στο `negbin12`. Σημειώστε ότι η προσαρμογή γίνεται στο δείγμα μάθησης το οποίο περιέχει εγγραφές στις οποίες η μεταβλητή `Type.risk` λαμβάνει τιμές 2 και 3 μόνο.

Η προσαρμογή επομένως με τη χρήση Αρνητικής διωνυμικής κατανομής και το μοντέλο `negbin12`, μπορεί να επιφέρει πολύ καλύτερα αποτελέσματα αφού επιτρέπει τη διαφορά μεταξύ μέσης τιμής και διασποράς και άρα λαμβάνει υπόψη της την υπερμεταβλητότητα των δεδομένων. Η χρήση ενός τέτοιου μοντέλου έχει φέρει δείκτη AIC ίσο με 85,934.84 μονάδες και υπόλοιπο Deviance ίσο με 28,566.24 μονάδες σε 49,267 βαθμούς ελευθερίας.

4.5.5 Μοντέλο μηδενικής διόγκωσης

Το 77% του παρόν δείγματος μάθησης αποτελείται από εγγραφές μηδενικών απαιτήσεων. Επομένως, ενδεχομένως οι κατανομές Poisson και Αρνητική Διωνυμική να αδυνατούν να ανταποκριθούν σε αυτά τα μεγέθη. Για αυτό τον λόγο επιλέγεται να χρησιμοποιηθούν κατανομές οι οποίες θα μπορούν να ανταποκριθούν καλύτερα σε δεδομένα όπου ο αριθμός των μηδενικών είναι τεράστιος. Έτσι, σε αυτή τη μοντελοποίηση θα προσαρμοστούν τα μοντέλα Μηδενικής Διόγκωσης. Αυτά προσαρμόζονται μέσω της εντολής `zeroinfl()` που ανήκει στη βιβλιοθήκη `pscl`. Για να προσαρμοστεί το τελευταίο μοντέλο, `zip12`, εκτελούμε την εντολή:

```
zip12 <- zeroinfl( N.claims.year ~  Driver.age.cut + Vehicle.age.cut
                  + Length + Type.risk + Type.fuel + Area
                  + offset(log(Exposure))
                  | Driver.age + offset(log(Exposure))
                  , dist='poisson', link="logit"
                  , data = train.data)
```

Το πιο πάνω αποτελεί την κλασσική αναπαράσταση της προσαρμογής ενός μοντέλου μηδενικής διόγκωσης. Ως κατανομή μέτρησης επιλέγεται η Poisson με τη χρήση της λογαριθμικής συνάρτησης σύνδεσης. Για τη μοντελοποίηση της πιθανότητας π πραγματοποιείται λογιστική παλινδρόμηση με χρήση logit συνάρτησης σύνδεσης, με χρήση σταθερού όρου, της επεξηγηματικής μεταβλητής `Driver.age` αλλά και τον όρο διόρθωσης `offset`. Η επιλογή αυτή φέρει καλύτερη προσαρμοστικότητα στο μοντέλο, παρά τη χρήση σταθερής πιθανότητας π. Το τελικό μοντέλο μηδενικής διόγκωσης που προσαρμόζεται, `zip12`, φέρει δείκτη AIC ίσο με 88,610.26 μονάδες ενώ η χρήση σταθερής πιθανότητας για τη λογιστική παλινδρόμηση φέρει δείκτη ίσο με 88826.69.

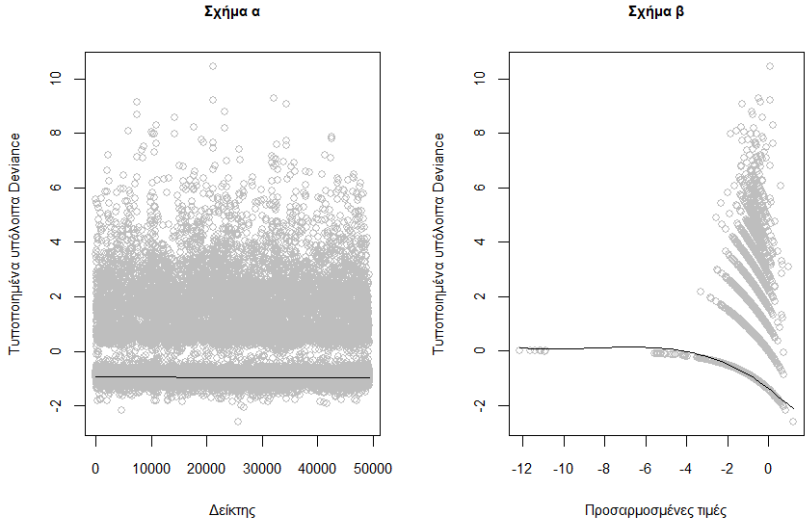
4.6 Διαγνωστικοί έλεγχοι

Μέχρι τώρα έχουν κατασκευαστεί μοντέλα στα οποία η επιλογή των μεταβλητών και η σύγκριση της προσαρμοστικότητας γινόταν μόνο μέσω της χρήση της Deviance και του δείκτη AIC. Η κατασκευή τους όμως είναι βασισμένη σε κάποιες υποθέσεις και αυτές θα πρέπει να εξεταστούν.

Εξετάζεται αρχικά η ανεξαρτησία μεταξύ των παρατηρήσεων και αυτό αφορά και τους τρεις τύπους μοντέλων. Κατασκευάζεται ένα γράφημα δείκτη των υπολοίπων Deviance, αριστερό γράφημα του Διαγράμματος 4.16. Σε αυτό δεν παρατηρείται κάποια τάση στα υπόλοιπα αφού η κατανομή τους είναι τυχαία και άρα δεν υπάρχουν ενδείξεις ότι οι παρατηρήσεις του δείγματος δεν είναι ανεξάρτητες.

Στη συνέχεια, κατασκευάζεται το διάγραμμα των υπολοίπων Deviance έναντι του εκτιμημένου γραμμικού συνδυασμού $\eta_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$, δηλαδή το διάγραμμα διασποράς (η_i, r_i^D) με r_i^D τα Deviance υπόλοιπα. Υπό την υπόθεση της γραμμικότητας, δεν αναμένεται κάποιο συστηματικό σχήμα ή εικόνα στα r_i^D σε σχέση με τη γραμμική προβλέπουσα. Από το δεξί γράφημα του Διαγράμματος 4.16, παρατηρείται κάποια τάση γραμμικής μείωσης των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές. Οι διάφορες γραμμικές κατανομές σημείων αναπαριστούν τα υπόλοιπα για διάφορες τιμές της `N.claims.year`, με συνολική τάση μείωσης. Τα γραφήματα αυτά αφορούν το μοντέλο `poisson12` ενώ η συμπεριφορά τους δεν αλλάζει σημαντικά στην περίπτωση των `negbin12` και `zip12`.

Για την καλύτερη εξέταση των υπολοίπων θα κατασκευαστούν τα διαγράμματα των μερικών υπολοίπων κάθε μεταβλητής έναντι κάθε μεταβλητής. Από αυτά τα γραφήματα μπορεί να διαπιστωθεί

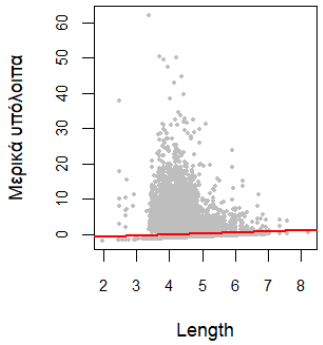


Διάγραμμα 4.16: Γραφικός έλεγχος για την υπόθεση ανεξαρτησίας σφαλμάτων (αριστερό γράφημα) και γραμμικότητας μεταξύ προβλέπουσας και μεταβλητής απόκρισης (δεξί γράφημα) για το `poisson12`

καλύτερα ποια είναι η συναρτησιακή σχέση της κάθε επεξηγηματικής μεταβλητής που χρησιμοποιείται στο μοντέλο σε σχέση με τη μεταβλητή απόκρισης. Αναμένεται από τις ποσοτικές μεταβλητές, γραμμικές συσχετίσεις αφού αυτή ήταν και η αρχική υπόθεση. Στο Διάγραμμα 4.17 παρουσιάζεται η διασπορά των μερικών υπολοίπων για την ποσοτική μεταβλητή `Length` σε σχέση με τις τιμές που λαμβάνει. Δεν παρατηρείται γραμμική κατανομή των σημείων όμως θεωρούμαι ότι η γραμμική τάση που εμφανίζεται με κόκκινη γραμμή είναι ικανοποιητική σε αυτό το στάδιο. Αν και η εξέταση αυτή είναι αρκετά αμφιλεγόμενη, υπάρχουν ενδείξεις πιθανού προβλήματος στο μοντέλο. Έχουν δοκιμαστεί διάφοροι μετασχηματισμοί στη μεταβλητή χωρίς όμως να πετυχαίνεται η επιθυμητή γραμμική συσχέτιση στο γράφημα. Αυτό δείχνει από τη μία ότι χρειάζεται να συμπεριληφθούν νέες μεταβλητές στο μοντέλο ή ότι το πρόβλημα προέρχεται από άλλους παράγοντες όπως για παράδειγμα η υπόθεση λανθασμένης κατανομής στα δεδομένα. Τα πιο πάνω γραφήματα είναι τα ίδια και στην περίπτωση του μοντέλου `negbin12` ενώ για το `zip12` η ερμηνεία τους δεν είναι ικανοποιητική.

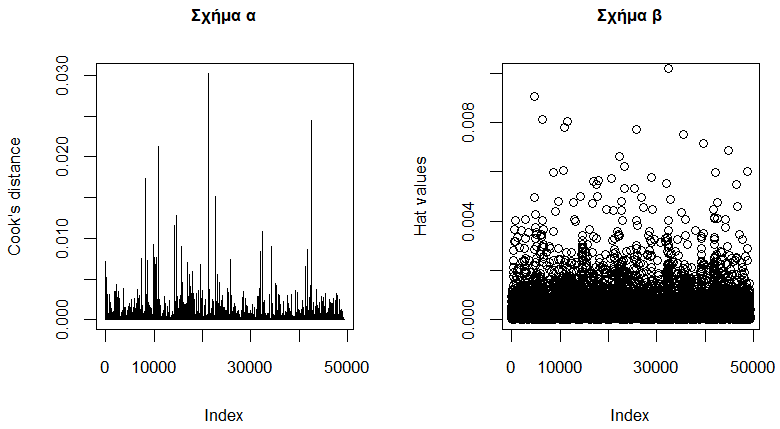
Στη συνέχεια, μπορεί να γίνει ένας άλλος χρήσιμος έλεγχος για τα υπόλοιπα, κατασκευάζοντας ένα Q-Q διάγραμμα. Όπως έχει αναφερθεί αν και τα υπόλοιπα δεν ακολουθούν κανονική κατανομή, γίνεται ο έλεγχος αυτός σαν ένδειξη της καλής προσαρμογής του μοντέλου. Γίνεται η σύγκριση μεταξύ των θεωρητικών και παρατηρούμενων τιμών των score κάθε κατανομής. Με τον όρο score νοείται το σημείο της κατανομής στο οποίο πετυχαίνεται κάποιο x τεταρτημόριο. Παρόλα΄ αυτά, από το προηγούμενο διάγραμμα 4.17 αντιλαμβανόμαστε ότι υπάρχει πρόβλημα με το κατασκευασμένο μοντέλο `poisson12` όσο αφορά τα υπόλοιπα. Επομένως, η θεώρηση κανονικής κατανομής των υπολοίπων και η σύγκριση με αυτή είναι ακατάλληλη και θα επέφερε λανθασμένα συμπεράσματα για την καταλληλότητα του μοντέλου, κάτι το οποίο δε φαίνεται κατά την επικύρωση στην Ενότητα 4.7 και έτσι δε γίνεται αυτός ο έλεγχος.

Για τον εντοπισμό των σημείων επιρροής, χρησιμοποιείται αρχικά η απόσταση Cook με σκοπό να εξεταστεί κατά πόσο η αφαίρεση κάποιων παρατηρήσεων θα επηρεάσει τις εκτιμήσεις των παραμέτρων του μοντέλου. Η ανάλυση αυτή προφανώς και εξαρτάται από το μοντέλο που έχει προσαρμοστεί

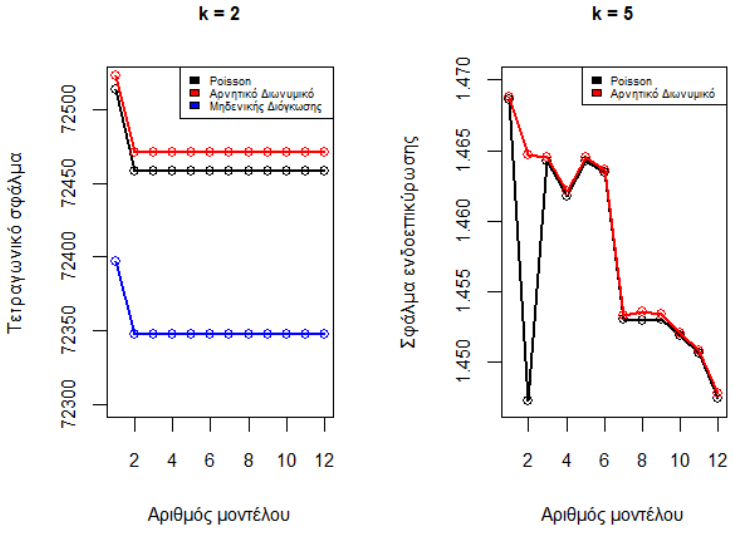


Διάγραμμα 4.17: Διαγράμματα διασποράς των μερικών υπολοίπων της `Length` για τον έλεγχο της υπόθεσης γραμμικότητας στο μοντέλο poisson12

αφού λαμβάνει υπόψη της τη μεταβολή στην εκτίμηση των παραμέτρων με την αφαίρεση κάποιας παρατήρησης. Με τη χρήση του τελευταίου μοντέλου, υπάρχουν 3,523 σημεία του δείγματος μάθησης (το 13.98%), των οποίων η απόσταση Cook, αριστερό γράφημα του Διαγράμματος 4.18, είναι δύο φορές μεγαλύτερη της μέσης απόστασης και θεωρούνται πως επηρεάζουν αρκετά το μοντέλο. Να σημειωθεί πως 3,324 παρατηρήσεις από τις 3,523 είναι παρατηρήσεις οι οποίες έχουν αριθμό απαιτήσεων μεγαλύτερο ή ίσο του 2 ενώ δεν υπάρχει παρατήρηση χωρίς απαιτήσεις (`N.claims.year`). Επομένως, παρατηρήσεις με ψηλό αριθμό απαιτήσεων φαίνεται να έχουν μεγάλη επιρροή στο μοντέλο. Επιπλέον, για τον εντοπισμό των σημείων αυτών κατασκευάζεται το γράφημα δείκτη για την ποσότητα μόχλευσης h_{ii} η οποία αποτελεί επίσης ένα μέτρο επιρροής. Στο δεξί γράφημα του Διαγράμματος 4.18 φαίνονται οι ποσότητες αυτές. Εάν χρησιμοποιηθεί τυχαία το 0.02 ως ένα μέτρο σύγκρισης, υπάρχουν 62 παρατηρήσεις με επιρροή μεγαλύτερη αυτού, όπου οι 49 ανήκουν στο πιο πάνω σύνολο δεικτών.



Διάγραμμα 4.18: Γραφήματα δεικτών για την απόσταση Cook (αριστερό γράφημα) και τη μόχλευση (δεξί γράφημα)



Διάγραμμα 4.19: Συνολικό τετραγωνικό σφάλμα (αριστερό γράφημα) και σφάλμα ενδοεπικύρωσης (δεξί γράφημα) για κάθε τύπο μοντέλου

4.7 Επικύρωση μοντέλων

Η επικύρωση των μοντέλων γίνεται με τη μέθοδο της ενδοεπικύρωσης και με βάση την ακολουθία των 12 μοντέλων που έχουν κατασκευαστεί σε κάθε τύπο. Τα μοντέλα έχουν προσαρμοστεί στο δείγμα μάθησης, επομένως θα πρέπει να εξεταστεί η προσαρμοστικότητα τους και στο δείγμα ελέγχου. Έτσι, υπολογίζεται η συνολική τετραγωνική διαφορά των παρατηρήσεων όταν τα μοντέλα αυτά προσαρμοστούν στο δείγμα ελέγχου. Υπολογίζεται δηλαδή η τιμή $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, που είναι το συνολικό τετραγωνικό σφάλμα (ΣΤΣ), με y_i και $\hat{y}_i$ η πραγματική τιμή και εκτίμηση της κάθε παρατήρησης με βάση το δείγμα ελέγχου, αντίστοιχα.

Αρχικά, κατασκευάζεται το διάγραμμα των τετραγωνικών αποκλίσεων για κάθε ένα από τα 12 μοντέλα και τους τρεις τύπους που φαίνεται στο αριστερό γράφημα του Διαγράμματος 4.19. Τα συνολικά τετραγωνικά σφάλματα για τα μοντέλα poisson_i, negbin_i και zip_i, $i = 1, \cdots, 12$ απεικονίζονται αντίστοιχα με μαύρο, κόκκινο και μπλε χρώμα. Παρατηρήστε τη μείωση του συνολικού σφάλματος με την προσθήκη της πρώτης μεταβλητής `Drv.age`. Οι υπόλοιπες διαφορές στο συνολικό τετραγωνικό σφάλμα συμβαίνουν σε μέγεθος δεκαδικού ψηφίου και άρα θεωρούνται αμελητέες κατά την προσθήκη ή όχι κάποιας μεταβλητής. Είναι εμφανές ότι η ακολουθία των μοντέλων Μηδενικής διόγκωσης παράγει μικρότερα σφάλματα των υπόλοιπων σε σχέση με τα μοντέλα των άλλων δύο τύπων μοντέλων και άρα είναι το καταλληλότερο. Επιπλέον, φαίνεται πως τα μοντέλα με τη χρήση της Αρνητικής διωνυμικής κατανομής έχουν μικρότερη ακρίβεια σε σχέση με τα Poisson.

Ο υπολογισμός του συνολικού τετραγωνικού σφάλματος έχει γίνει με βάση τον τυχαίο τρόπο που κατασκευάστηκαν τα δείγματα μάθησης και ελέγχου και αυτός δεν αντικατοπτρίζει πλήρως την ακρίβεια του μοντέλου σε όλα τα δεδομένα. Η μέθοδος ενδοεπικύρωσης εξυπηρετεί σε αυτόν ακριβώς τον σκοπό. Ολόκληρο το δείγμα αντί να χωρίζεται σε 2 δείγματα, θα χωριστεί σε 5. Σε κάθε επανάληψη ένα από αυτά τα 5 υποδείγματα θα χρησιμοποιείται ως το δείγμα ελέγχου (20%) και το υπόλοιπο (80%) ως δείγμα μάθησης στο οποίο και θα εκπαιδεύεται το μοντέλο. Επιλέγεται λοιπόν $\kappa = 5$ και υπολογίζεται για κάθε ένα από αυτά τα 12 μοντέλα το σφάλμα ενδοεπικύρωσης (ΣΕ) για

	Poisson		Αρνητικό Διωνυμικό		Μηδενικής Διόγκωσης	
Απαιτήσεις	Μάθησης	Ελέγχου	Μάθησης	Ελέγχου	Μάθησης	Ελέγχου
Εκτιμώμενες	24,022.00	15,849.248	24,356.94	16,089.55	23,743.18	15,658.91
Πραγματικές	24,022.00	15,747.00	24,022.00	15,747.00	24,022.00	15,747.00
Λόγος	1.00	1.00	1.01	1.02	0.99	0.99

Πίνακας 4.28: Εκτιμώμενες και παρατηρούμενες αριθμός απαιτήσεις ανά δείγμα

τους δύο πρώτους τύπους μοντέλων καθώς δεν υπάρχει αντίστοιχη συνάρτηση για μοντέλα Μηδενικής διόγκωσης. Κάθε ένα μοντέλο προσαρμόζεται 5 φορές σε 5 διαφορετικά δείγματα μάθησης και υπολογίζεται κάθε φορά το σφάλμα ενδοεπικύρωσης το οποίο και ισούται με το πηλίκο του συνολικού τετραγωνικού σφάλματος με τους βαθμούς ελευθερίας σε κάθε μοντέλο. Η ποσότητα που λαμβάνεται μέσω της εντολής `cv.glm( )` είναι η $\frac{1}{5}\sum_{i=1}^{5}\Sigma\mathrm{E}_i$, και είναι το μέσο σφάλμα ενδοεπικύρωσης από τις 5 επαναλήψεις που έγιναν. Αυτό ορίζεται ως το σφάλμα ενδοεπικύρωσης. Κατασκευάζεται επομένως το δεξί γράφημα του Διάγραμματος 4.19. Από αυτό φαίνεται πως η διαφορά των δύο πρώτων τύπων μοντέλων είναι αδιάφορη ως προς το σφάλμα ενδοεπικύρωσης και έτσι θεωρείται πως είναι όμοιας ακρίβειας. Για το `zip12` μοντέλο, η εκτίμηση του μέσου τετραγωνικού σφάλματος στο δείγμα μάθησης είναι $\frac{\text{συνολικό τετραγωνικό σφάλμα}}{n} = \frac{72229.44}{49277} = 1.4657$. Σημειώστε πως η εμπειρική διασπορά του δείγματος μάθησης είναι 1.52. Επομένως, από το μέσο τετραγωνικό σφάλμα αντιλαμβανόμαστε ότι το μοντέλο `zip12` ανταποκρίνεται σωστά στη μεταβλητότητα των δεδομένων. Από την πιο πάνω ανάλυση, το μοντέλο μηδενικής διόγκωσης θεωρείται το καταλληλότερο.

Στη συνέχεια γίνεται μία ανάλυση στην οποία συγκρίνεται το πραγματικό άθροισμα των απαιτήσεων σε σχέση με το εκτιμώμενο σε διάφορες περιπτώσεις. Αρχικά, πραγματοποιείται για κάθε ένα από τα δείγματα ελέγχου και μάθησης [8]. Η ανάλυση αυτή βοηθά στο να αξιολογηθεί η ακρίβεια των τύπων μοντέλων `poisson12`, `negbin12` και `zip12`. Στον Πίνακα 4.28 φαίνονται οι συνολικές εκτιμώμενες και παρατηρούμενες απαιτήσεις για τα δύο τυχαία κατασκευασμένα δείγματα μάθησης και ελέγχου. Επίσης, παρουσιάζεται ο λόγος των εκτιμημένων απαιτήσεων προς τις πραγματικές για κάθε ένα από τα δείγματα και τους τρεις τύπους μοντέλων. Λόγος κοντά στη μονάδα σημαίνει πως το εκτιμώμενο άθροισμα προσεγγίζει το πραγματικό και αυτό είναι το επιθυμητό. Αρχικά, για το `poisson12` μοντέλο φαίνεται πως προσαρμόζεται πάρα πολύ καλά και στα δύο δείγματα. Για το `negbin12` μοντέλο φαίνεται πως και στις δύο περιπτώσεις γίνεται ελάχιστη υπερεκτίμηση του συνολικού αριθμού απαιτήσεων με τις εκτιμώμενες απαιτήσεις να είναι στην πρώτη περίπτωση κατά 1% και στη δεύτερη κατά 2% αυξημένες σε σχέση με τον πραγματικό αριθμό. Τέλος, από το μοντέλο `zip12` φαίνεται ότι και στο δείγμα μάθησης και στο δείγμα ελέγχου οι εκτιμήσεις είναι πάρα πολύ ακριβής. Συμπερασματικά, φαίνεται ότι ως προς τον συνολικό αριθμό απαιτήσεων τα μοντέλα έχουν πάρα πολύ καλές εκτιμήσεις.

Ενδιαφέρον παρέχει η πιο πάνω ανάλυση να γίνει ξεχωριστά για δύο σημαντικές μεταβλητές, τη `Year` και τη `Driver.age.cut` στο δείγμα ελέγχου. Όπως έχει προαναφερθεί η μεταβλητή `Year` εισάγει κάποιον θόρυβο ο οποίος δεν μπορεί να ερμηνευθεί. Από τον Πίνακα 4.29 παρατηρείται ότι οι εκτιμήσεις των μοντέλων δεν είναι τόσο ικανοποιητικές με βάση τα έτη αφού υπάρχουν έντονες αποκλίσεις. Τα τρία μοντέλα φαίνεται πως στα πρώτα έτη εμφανίζουν έντονη υποεκτίμηση του συνολικού αριθμού απαιτήσεων ενώ με το πέρασμα των ετών η συμπεριφορά αυτή εξελίσσεται σε υπερεκτίμηση των απαιτήσεων. Η συμπεριφορά αυτή θα μπορούσε να είναι καλύτερη αν στο μοντέλο

είχε συμπεριληφθεί η μεταβλητή `Year`. Κατασκευάζεται επίσης ο Πίνακας 4.30 στον οποίο φαίνονται οι λόγοι για κάθε ηλικιακό εύρος στο δείγμα ελέγχου και οι εκτιμήσεις είναι ικανοποιητικές. Με βάση τους δύο Πίνακες 4.29 και 4.30, οι μικρότερες αποκλίσεις γίνονται στο μοντέλο Μηδενικής διόγκωσης, `zip12`.

	Poisson				Αρνητικό Διωνυμικό				Μηδενικής Διόγκωσης			
Απαιτήσεις	2015	2016	2017	2018	2015	2016	2017	2018	2015	2016	2017	2018
Εκτιμώμενες	691	4,640	4,996	5,521	701	4,706	5,067	5,598	685	4,599	4,952	5,423
Πραγματικές	1,203	7,952	4,682	1,910	1,203	7,952	4,682	1,910	1,203	7,952	4,682	1,910
Λόγος	0.57	0.58	1.07	2.89	0.58	0.59	1.08	2.93	0.57	0.58	1.06	2.84

Πίνακας 4.29: Εκτιμώμενες και παρατηρούμενες απαιτήσεις ανά έτος

Ηλικιακό εύρος	[18, 24]	[25, 44]	[45, 59]	[60, 69]	70+
Poisson μοντέλο					
Εκτιμώμενες	353	6,916	6,522	1,503	557
Πραγματικές	306	7,028	6,249	1,603	561
Λόγος	1.15	0.98	1.04	0.94	0.99
Αρνητικό Διωνυμικό Μοντέλο					
Εκτιμώμενες	352	6,977	6,650	1,526	568
Πραγματικές	306	7,028	6,249	1,603	561
Λόγος	1.15	0.99	1.06	0.95	1.01
Μοντέλο Μηδενικής Διόγκωσης					
Εκτιμώμενες	314	6,996	6,263	1,499	587
Πραγματικές	306	7,028	6,249	1,603	561
Λόγος	1.03	0.99	1.00	0.94	1.05

Πίνακας 4.30: Εκτιμώμενες και παρατηρούμενες απαιτήσεις ανά ηλικιακή περίοδο

Συμπερασματικά, αρχικά με βάση το συνολικό τετραγωνικό σφάλμα, φαίνεται πως το Poisson και το μοντέλο Αρνητικής διωνυμικής κατανομής δεν είναι τόσο ακριβή όσο το μοντέλο Μηδενικής διόγκωσης το οποίο φαίνεται να είναι το καταλληλότερο για τη μοντελοποίηση της συχνότητας απαιτήσεων αφού παρέχει τις ακριβέστερες απαιτήσεις. Επίσης, η ανάλυση παρατηρούμενων και εκτιμώμενων ποσοτήτων δε φαίνεται να δείχνει σοβαρές ενδείξεις για την ακαταλληλότητα του μοντέλου `zip12`, οπότε και μπορεί να θεωρηθεί ως η καταλληλότερη επιλογή για τις εκτιμήσεις που θα γίνουν.

Κεφάλαιο 5

Αποτελέσματα

Μετά την κατασκευή των μοντέλων, θα παρουσιαστούν τα αποτελέσματα που προκύπτουν από αυτά μέσω της στατιστικής συμπερασματολογίας και των προβλέψεων που γίνονται. Ακολουθούν γενικά συμπεράσματα και συζήτηση μέσα από όλη τη διαδικασία της μοντελοποίησης, της ανάλυσης των δεδομένων και των περιορισμών που προέκυψαν.

5.1 Στατιστική Συμπερασματολογία

Αφού έχουν προσαρμοστεί τα μοντέλα, μπορεί να γίνει η ερμηνεία των αποτελεσμάτων τους. Μέσα από αυτή μπορούν να αναγνωριστούν οι κύριοι παράγοντες που επηρεάζουν τελικά την εκτιμημένη ποσότητα και η σημαντικότητα τους. Υπενθυμίζεται πως το προσαρμοσμένο μοντέλο έχει τη μορφή

$$\mathbb{E}(Y_i) = w_i e^{\beta_0} e^{\beta_1 x_{i1}} \cdots e^{\beta_p x_{ip}},$$

όπου Y_i ο αναμενόμενος αριθμός απαιτήσεων σε διάρκεια έκθεσης στον κίνδυνο w_i. Η έκθεση έχει συμπεριληφθεί στο μοντέλο μέσω της μεταβλητής `Exposure` και του όρου `offset` που είναι ίσος με `log(Exposure)`. Όταν η διάρκεια έκθεσης w_i είναι ίση με ένα, τότε γίνεται αναφορά σε αριθμό απαιτήσεων σε διάρκεια ενός ολόκληρου έτους.

5.1.1 Μοντέλο `poisson`

Αρχικά, παρουσιάζονται τα αποτελέσματα του μοντέλου `poisson12` που κατασκευάστηκε. Με την εντολή `summary()` λαμβάνονται διάφορα αποτελέσματα της στατιστικής συμπερασματολογίας του μοντέλου. Ως όρισμα η εν λόγω εντολή δέχεται το αντικείμενο `poisson12` το οποίο είναι και το τελικό Poisson μοντέλο που έχει κατασκευαστεί.

```
> summary(poisson12)

Call:
glm(formula = N.claims.year ~ Driver.age.cut + Vehicle.age.cut +
    Length + Type.risk + Type.fuel + Area, family = poisson(link = log),
    data = train.data, offset = log(Exposure))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-2.5195  -1.0273  -0.9200  -0.6741  10.5528

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.12526    0.08366 -25.404  < 2e-16 ***
Driver.age.cut[25, 44]  -0.39693    0.04383  -9.056  < 2e-16 ***
Driver.age.cut[45, 59]  -0.34881    0.04395  -7.937 2.07e-15 ***
Driver.age.cut[60, 69]  -0.72081    0.04757 -15.153  < 2e-16 ***
Driver.age.cut[70, Inf) -0.84707    0.05476 -15.469  < 2e-16 ***
Vehicle.age.cut6+        0.21229    0.02149   9.879  < 2e-16 ***
Length                   0.38091    0.01628  23.404  < 2e-16 ***
Type.risk2               0.09588    0.01936   4.953 7.32e-07 ***
Type.fuelP              -0.17949    0.01537 -11.676  < 2e-16 ***
Area1                    0.23865    0.01401  17.034  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **
0.01    *    0.05    .    0.1         1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 82973  on 49276  degrees of freedom
Residual deviance: 80866  on 49267  degrees of freedom
AIC: 109014

Number of Fisher Scoring iterations: 6
```

Με βάση τα πιο πάνω αρχικά μπορεί να διαπιστωθεί η μη ικανοποιητική κατανομή των υπολοίπων Deviance. Θεωρητικά τα υπόλοιπα αυτά κατανέμονται κανονικά και επομένως οι περιγραφικοί δείκτες των υπολοίπων που παρουσιάζονται (ελάχιστη τιμή, 1ο τεταρτημόριο, διάμεσος, 3ο τεταρτημόριο και μέγιστη τιμή) αναμένεται να δείχνουν μία τέτοια συμπεριφορά. Τα αποτελέσματα δε δείχνουν κάτι τέτοιο αφού υπάρχει απόκλιση από την κανονικότητα. Υπάρχει έντονη συγκέντρωση των υπολοίπων στον αρνητικό ημιάξονα ενώ τα θετικά υπόλοιπα είναι σπανιότερα και μεγαλύτερα. Αυτό σημαίνει πως στο παρόν μοντέλο υπάρχει έντονη τάση αρνητικών υπολοίπων και έτσι οι εκτιμήσεις να είναι μεγαλύτερες των πραγματικών. Από την άλλη η πιθανότητα θετικών σφαλμάτων είναι μειωμένη αλλά η ουρά φαίνεται να είναι πιο απλωμένη στον θετικό ημιάξονα και άρα τα θετικά υπόλοιπα είναι μεγαλύτερα κατά μέτρο από τα αρνητικά αλλά σπανιότερα. Στη συνέχεια, παρουσιάζονται τα αποτελέσματα για τους συντελεστές των 9 μεταβλητών και του σταθερό όρου που συμμετέχουν στο μοντέλο.

Όπως είπαμε, ο όρος `offset = log(Exposure)` δε λαμβάνει κάποιο συντελεστή αλλά επηρεάζει την ερμηνεία των αποτελεσμάτων. Για την ερμηνεία των συντελεστών του μοντέλου θεωρείται πως η έκθεση παραμένει σταθερή κατά τη μεταβολή κάποιας επεξηγηματικής μεταβλητής. Η ίδια ερμηνεία δίνεται και αν θεωρηθεί ότι η έκθεση είναι πάντα ένα ολόκληρο έτος. Θα ακολουθηθεί η τελευταία θεώρηση ενώ για διάρκεια μικρότερη του ενός η εκτίμηση της αναμενόμενης συχνότητας θα πολλαπλασιάζεται με την τιμή έκθεσης αντίστοιχα.

Παρατηρήστε ότι η τιμή του σημειακού εκτιμητή του σταθερού όρου β_0 είναι -2.12 με τυπικό σφάλμα 0.08, ενώ το στατιστικό ελέγχου T_1 για τον έλεγχο Wald του σταθερού όρου με υποθέσεις

$H_0 : \beta_0 = 0$ έναντι της $H_1 : \beta_0 \neq 0$ έδωσε τιμή -25.40. Η p-τιμή του ελέγχου είναι πολύ μικρή οπότε υπάρχουν σοβαρές ενδείξεις για την απόρριψη της μηδενικής υπόθεσης. Συμπερασματικά, η συχνότητα των απαιτήσεων σε ετήσια έκθεση σε κίνδυνο $w_i = 1$, στο ηλικιακό εύρος [18, 24], με πετρελαιοκίνητο επιβατικό όχημα ηλικίας 5 ετών και κάτω, μηδενικού μήκους σε αγροτική περιοχή είναι $e^{\beta_0} = e^{-2.12} = 0.12$ το οποίο φυσικά δεν είναι ρεαλιστικό αφού δεν υπάρχει τέτοια παρατήρηση. Παρόλ' αυτά, η τιμή e^{β_0} θεωρείτε από τους αναλογιστές η τιμή του βασικού ασφαλίστρου σε όποιες τιμές των επεξηγηματικών μεταβλητών.

Για τη μεταβλητή ηλικίας η ανάλυση είναι ως εξής. Υπάρχουν τέσσερεις εικονικές μεταβλητές ενώ το ηλικιακό εύρος βάσης είναι το [18, 24]. Από τις τιμές των συντελεστών των εικονικών μεταβλητών, η μεγαλύτερη συχνότητα απαιτήσεων λαμβάνεται φυσικά στην κατηγορία βάσης [18, 24] ενώ ακολουθά το ηλικιακό εύρος [45, 59], [25, 44], [60, 69] και τέλος το 70+. Εάν ένας οδηγός εμπίπτει στην ηλικία [45, 59] τότε η αναμενόμενη συχνότητα απαιτήσεων πολλαπλασιάζεται με $e^{-0.39} = 0.67$, δηλαδή μειώνεται κατά $\frac{1-0.67}{1}100\% = 32.29\%$ σε σχέση με την αναμενόμενη συχνότητα στην ηλικία βάσης, εάν η έκθεση, κατηγορία ηλικίας οχήματος, μήκος, τύπος οχήματος, καυσίμου και περιοχή δε μεταβληθούν. Αντίστοιχα στο εύρος [25, 44] η αναμενόμενη συχνότητα μειώνεται κατά 30% σε σχέση με το ηλικιακό εύρος [18, 24], δεδομένου ότι η έκθεση, κατηγορία ηλικίας οχήματος, μήκος, τύπος οχήματος, καυσίμου και περιοχή δε μεταβληθούν. Είναι ενδιαφέρον το ότι σε μεγαλύτερες ηλικίες, όπως στο ηλικιακό εύρος [45, 59] η αναμενόμενη συχνότητα εμφανίζει μικρότερη μείωση απ' ότι στις ηλικίες [25, 44], σε σχέση με την ηλικία βάσης [18, 24]. Η αυξημένη συχνότητα σε ηλικίες [45, 59], σε σχέση με την ηλικία βάσης [18, 24], μπορεί να δικαιολογηθεί από το γεγονός ότι νεότεροι οδηγοί έχουν πρόσβαση στο όχημα των γονιών τους οι οποίοι βρίσκονται στις ηλικίες [45, 59] και άρα ο κίνδυνος είναι μεγαλύτερος. Από την άλλη, σε ηλικίες [60, 69] η συχνότητα των απαιτήσεων είναι κατά 0.52% μικρότερη σε σχέση με την αναμενόμενη συχνότητα των ηλικιών βάσης, ενώ είναι κατά $(1 - e^{-0.7208}e^{0.3488})100\% = (1 - e^{-0.3782})100\% = 31.49\%$ μικρότερη από το προηγούμενο ηλικιακό εύρος, [45, 59], δεδομένου ότι η έκθεση, κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος οχήματος, καυσίμου και περιοχή δε μεταβληθούν. Τη μεγαλύτερη μείωση στη συχνότητα σε σχέση με το ηλικιακό εύρος βάσης, σημειώνουν οι ηλικίες 70+ με μείωση κατά 57.13%, δεδομένου ότι η έκθεση, κατηγορία ηλικίας οχήματος, μήκος, τύπος οχήματος, καυσίμου και περιοχή δε μεταβληθούν και αυτό δείχνει την αδυναμία του μοντέλου να εκτιμήσει την αυξημένη συχνότητα σε αυτό το εύρος, όπως είχε διαπιστωθεί κατά τη διερευνητική ανάλυση.

Η μεταβλητή ηλικίας του οχήματος, `Vehicle.age.cut` θεωρείται επίσης σημαντική αφού με βάση τον έλεγχο Wald του συντελεστή η p-τιμή είναι πάρα πολύ μικρή και άρα η ηλικία του οχήματος σχετίζεται με τη συχνότητα των απαιτήσεων. Συγκεκριμένα, η εκτίμηση του συντελεστή της `Vehicle.age.cut6+` είναι 0.21 με τυπική απόκλιση ίση με 0.021 και p-τιμή σχεδόν μηδέν. Επομένως, όταν η ηλικία του οχήματος είναι 6 χρόνια και άνω υπάρχει αύξηση της αναμενόμενης συχνότητας κατά $\frac{e^{0.2123}-1}{1}100\% = 23.65\%$ σε σχέση με οχήματα ηλικίας 5 ετών και κάτω, δεδομένου ότι η έκθεση, το ηλικιακό εύρος, το μήκος, ο τύπος του οχήματος, ο τύπος καυσίμου και η περιοχή παραμείνουν σταθερά.

Εν συνεχεία γίνεται ο σχολιασμός των αποτελεσμάτων για τη μεταβλητή του μήκους του οχήματος, `Length`. Αφού αυτή είναι μία συνεχής μεταβλητή, η αύξηση του μήκους ενός οχήματος κατά ένα μέτρο θα φέρει αύξηση στην αναμενόμενη συχνότητα των απαιτήσεων κατά $\frac{e^{0.3809}-1}{1}100\% = 46.36\%$, δεδομένου ότι οι μεταβλητές έκθεσης, εύρους ηλικίας, κατηγορίες ηλικίας οχήματος, τύπος οχήματος, καυσίμου και περιοχής παραμείνουν σταθερές.

Η μεταβλητή τύπου ρίσκου, `Type.risk`, είναι επίσης σημαντική για όλες τις κατηγορίες. Βασική κατηγορία είναι η τύπου 3, δηλαδή επιβατικό όχημα. Με βάση την εκτίμηση του συντελεστή της εικονικής μεταβλητής `Type.risk2` που ισούται με 0.0959 προκύπτει ότι σε σχέση με ένα επιβατικό όχημα, η αναμενόμενη συχνότητα απαιτήσεων στην περίπτωση φορτηγού είναι κατά $\frac{e^{0.0959}-1}{1}100\% = 10\%$ μεγαλύτερη, δεδομένου ότι η έκθεση, το ηλικιακό εύρος, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος καυσίμου και περιοχή παραμένουν σταθερά.

Για τη μεταβλητή καυσίμου, `Type.fuel`, η συμπερίληψη της στο μοντέλο κρίνεται και αυτή με βάση τον έλεγχο Wald ως σημαντική. Για τον λόγο ότι υπάρχουν δύο κατηγορίες καυσίμου έχει κατασκευαστεί μία εικονική μεταβλητή, η `Type.fuelP` και η οποία αντιπροσωπεύει τη χρήση καυσίμου βενζίνης, με κατηγορία βάσης το πετρέλαιο. Η εκτίμηση του συντελεστή είναι ίση με -0.18. Αυτό σημαίνει πως η χρήση βενζίνης σε ένα όχημα φέρει μείωση στον αναμενόμενο αριθμό απαιτήσεων κατά $\frac{1-e^{-0.1795}}{1}100\% = 16.43\%$, σε σχέση με τη χρήση πετρελαίου, δεδομένου ότι η έκθεση, το ηλικιακό εύρος, η κατηγορίας ηλικίας οχήματος, το μήκος, ο τύπος οχήματος και η περιοχή παραμένουν σταθερά.

Τέλος, για τη μεταβλητή `Area` έχει κατασκευαστεί μία εικονική μεταβλητή η οποία αναγνωρίζει περιοχές τύπου 1, δηλαδή κατοικημένες περιοχές. Κατηγορία βάσης είναι οι αγροτικές περιοχές. Η προσθήκη της `Area` κρίνεται στατιστικά σημαντική και έτσι υπάρχουν ενδείξεις για την απόρριψη της μηδενικής υπόθεσης. Επομένως, με βάση την τιμή του συντελεστή της εικονικής μεταβλητής `Area1` η αναμενόμενη συχνότητα απαιτήσεων αυξάνεται κατά $\frac{e^{0.2386}-1}{1}100\% = 26.95\%$ σε κατοικημένες περιοχές σε σχέση με αγροτικές περιοχές, δεδομένου ότι έκθεση, το ηλικιακό εύρος, η κατηγορίας ηλικίας οχήματος, το μήκος, ο τύπος οχήματος και καυσίμου παραμείνουν σταθερά.

Το `poisson12` μοντέλο που κατασκευάστηκε φέρει ελεγχοσυνάρτηση Deviance με τιμή 80866 σε 49267 βαθμούς ελευθερίας. Επομένως, η ελεγχοσυνάρτηση απέχει από το πλήρες μοντέλο κατά 80866 μονάδες ενώ το μοντέλο με τη χρήση μόνο του σταθερού όρου απέχει από το πλήρες κατά 82973 μονάδες. Ο δείκτης AIC έχει τιμή ίση με 109014 μονάδες και είναι ο χαμηλότερος όλων των Poisson μοντέλων που έχουν κατασκευαστεί. Τέλος, έχουν γίνει 6 επαναλήψεις της μεθόδου Newton-Raphson για την εκτίμηση των παραμέτρων του μοντέλου.

5.1.2 Μοντέλο `negbin`

Στη συνέχεια παρουσιάζονται τα αποτελέσματα του Αρνητικού διωνυμικού μοντέλου `negbin12` και γίνονται οι ερμηνείες με βάση αυτό.

Η τιμή του σημειακού εκτιμητή του σταθερού όρου β_0 είναι -2.05 με τυπικό σφάλμα 0.16, ενώ το στατιστικό ελέγχου T_1 για τον έλεγχο Wald του σταθερού όρου με υποθέσεις $H_0 : \beta_0 = 0$ έναντι της $H_1 : \beta_0 \neq 0$ έδωσε τιμή -12.79. Η p-τιμή του ελέγχου είναι πολύ μικρή οπότε υπάρχουν σοβαρές ενδείξεις για την απόρριψη της μηδενικής υπόθεσης. Με βάση το μοντέλο `negbin12`, η συχνότητα των απαιτήσεων σε ετήσια έκθεση σε κίνδυνο $w_i = 1$, στο ηλικιακό εύρος [18, 24], με πετρελαιοκίνητο επιβατικό όχημα ηλικίας 5 ετών και κάτω, μηδενικού μήκους σε αγροτική περιοχή είναι $e^{\beta_0} = e^{-2.05} = 0.13$ το οποίο φυσικά δεν είναι ρεαλιστικό αφού δεν υπάρχει τέτοια παρατήρηση.

Για τη μεταβλητή `Driver.age.cut` η ανάλυση γίνεται ως εξής. Στο εύρος [25, 44] η αναμενόμενη συχνότητα μειώνεται κατά $\frac{1-e^{-0.38}}{1} = 31.61\%$ σε σχέση με το ηλικιακό εύρος βάσης [18, 24], εάν η έκθεση, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος οχήματος, τύπος καυσίμου και η περιοχή δε μεταβληθούν. Από την τιμή του συντελεστή της εικονικής μεταβλητής `Driver.age.cut[45, 59]`, η αναμενόμενη συχνότητα είναι κατά $\frac{1-e^{-0.32}}{1} = 27.38\%$ μειωμένη σε σχέση με το ηλικιακό εύρος βάσης

Μοντέλο	Poisson		Αρνητικό Διωνυμικό		Μηδενικής Διόγκωσης	
Μεταβλητές	β_i	e^{β_i}	β_i	e^{β_i}	β_i	e^{β_i}
`Intercept`	-2.1253	0.1194	-2.0532	0.1283	-0.6981	0.4975
`Driver.age.cut[25, 44]`	-0.3969	0.6724	-0.3821	0.6824	-0.1180	0.8887
`Driver.age.cut[45, 59]`	-0.3488	0.7055	-0.3228	0.7241	0.0298	1.0303
`Driver.age.cut[60, 69]`	-0.7208	0.4864	-0.6988	0.4972	-0.1804	0.8350
`Driver.age.cut[70, Inf)`	-0.8471	0.4287	-0.8237	0.4388	-0.1709	0.8429
`Vehicle.age.cut+6`	0.2123	1.2365	0.2147	1.2395	0.1985	1.2195
`Length`	0.3809	1.4636	0.3628	1.4374	0.2615	1.2989
`Type.risk2`	0.0959	1.1006	0.0882	1.0922	0.1096	1.1158
`Type.fuelP`	-0.1795	0.8357	-0.1810	0.8344	-0.1280	0.8798
`Area1`	0.2386	1.2695	0.2381	1.2688	0.1372	1.1471
`Intercept`	-	-	-	-	0.2904	1.3370
`Driver.age`	-	-	-	-	0.0137	1.0138

Πίνακας 5.1: Τιμές των συντελεστών των μοντέλων

[18, 24], δεδομένου ότι η έκθεση, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος οχήματος, ο τύπος καυσίμου και η περιοχή παραμείνουν σταθερά. Σε ηλικίες [60, 69] η συχνότητα των απαιτήσεων είναι κατά $\frac{1-e^{0.70}}{1} = 50.28\%$ μικρότερη σε σχέση με την αναμενόμενη συχνότητα των ηλικιών βάσης. Τη μεγαλύτερη μείωση στη συχνότητα σε σχέση με το ηλικιακό εύρος βάσης, σημειώνουν οι ηλικίες 70+ με μείωση κατά $1 - e{-}0.82 = 55.96\%$, δεδομένου ότι η έκθεση, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος οχήματος, ο τύπος καυσίμου και η περιοχή παραμείνουν σταθερά.

Η μεταβλητή `Vehicle.age.cut` είναι επίσης σημαντική αφού με βάση τον έλεγχο Wald του συντελεστή η p-τιμή είναι πάρα πολύ μικρή και άρα η ηλικία του οχήματος σχετίζεται με τη συχνότητα των απαιτήσεων. Συγκεκριμένα, η εκτίμηση του συντελεστή της `Vehicle.age.cut6+` είναι 0.21 με τυπική απόκλιση ίση με 0.036 και p-τιμή σχεδόν μηδέν. Επομένως, όταν η ηλικία του οχήματος είναι 6 χρόνια και άνω υπάρχει αύξηση της αναμενόμενης συχνότητας κατά $\frac{e^{0.2147}-1}{1}100\% = 23.94\%$ σε σχέση με οχήματα ηλικίας 5 ετών και κάτω, δεδομένου ότι η έκθεση, το ηλικιακό εύρος, το μήκος, ο τύπος του οχήματος, ο τύπος καυσίμου και η περιοχή παραμείνουν σταθερά.

Για τη μεταβλητή `Length` έχει εκτιμηθεί συντελεστής με τιμή 0.3628. Αφού αυτή είναι μία συνεχής μεταβλητή, η αύξηση του μήκους ενός οχήματος κατά ένα μέτρο θα φέρει αύξηση στην αναμενόμενη συχνότητα των απαιτήσεων κατά $\frac{e^{0.3628}-1}{1}100\% = 43.73\%$, δεδομένου ότι οι μεταβλητές έκθεσης, εύρους ηλικίας, κατηγορίας ηλικίας οχήματος, τύπου οχήματος, καυσίμου και περιοχής παραμείνουν σταθερές.

Η μεταβλητή τύπου ρίσκου, `Type.risk`, φαίνεται να είναι στατιστικά σημαντική σε όλες τις κατηγορίες. Βασική κατηγορία είναι η τύπου 3, δηλαδή επιβατικό όχημα. Με βάση την εκτίμηση του συντελεστή της εικονικής μεταβλητής `Type.risk2` που ισούται με 0.0882 προκύπτει ότι σε σχέση με ένα επιβατικό όχημα, η αναμενόμενη συχνότητα απαιτήσεων στην περίπτωση φορτηγού είναι κατά $\frac{e^{0.0882}-1}{1}100\% = 9.22\%$ μεγαλύτερη, δεδομένου ότι η έκθεση, το ηλικιακό εύρος, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος καυσίμου και περιοχή παραμένουν σταθερά.

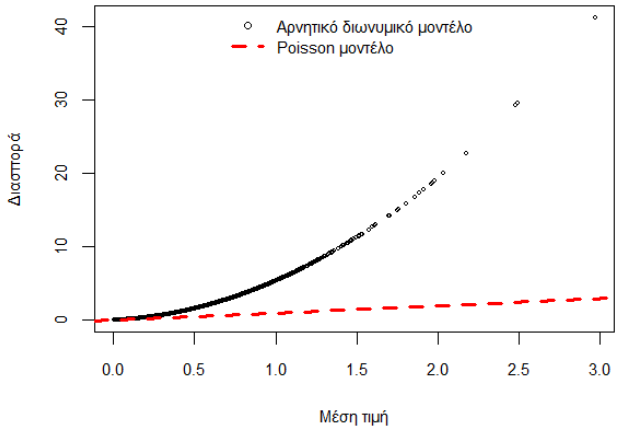
Για τη μεταβλητή `Type.fuel`, η συμπερίληψη της στο μοντέλο κρίνεται και αυτή με βάση τον έλεγχο Wald ως σημαντική. Κατηγορία βάσης είναι το πετρέλαιο. Η εκτίμηση του συντελεστή είναι

ίση με -0.18. Αυτό σημαίνει πως η χρήση βενζίνης σε ένα όχημα φέρει μείωση στον αναμενόμενο αριθμό απαιτήσεων κατά $\frac{1-e^{-0.1795}}{1}100\% = 16.43\%$, σε σχέση με τη χρήση πετρελαίου, δεδομένου ότι η έκθεση, το ηλικιακό εύρος, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος οχήματος και η περιοχή παραμένουν σταθερά.

Τέλος, για τη μεταβλητή `Area` έχει κατασκευαστεί μία εικονική μεταβλητή η οποία αναγνωρίζει περιοχές τύπου 1, δηλαδή κατοικημένες περιοχές. Κατηγορία βάσης είναι οι αγροτικές περιοχές. Η προσθήκη της `Area` κρίνεται στατιστικά σημαντική και έτσι υπάρχουν ενδείξεις για την απόρριψη της μηδενικής υπόθεσης. Επομένως, με βάση την τιμή του συντελεστή της εικονικής μεταβλητής `Area1` η αναμενόμενη συχνότητα απαιτήσεων αυξάνεται κατά $\frac{e^{0.2381}-1}{1}100\% = 26.89\%$ σε κατοικημένες περιοχές σε σχέση με αγροτικές, δεδομένου ότι έκθεση, το ηλικιακό εύρος, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος οχήματος και καυσίμου παραμείνουν σταθερά.

Η διαφορά ανάμεσα στη μοντελοποίηση με χρήση της Poisson και της Αρνητικής διωνυμικής κατανομής βασίζεται στη συνάρτηση διασποράς [4]. Στο Διάγραμμα 5.1 φαίνεται η σχέση μεταξύ της μέσης τιμής και της διασποράς σε κάθε ένα από τα δύο μοντέλα `poisson12` (κόκκινη απεικόνιση) και `negbin12` (μάυρη απεικόνιση). Προσέξτε ότι και οι δύο κατανομές έχουν αύξηση της διασποράς με την αύξηση της μέσης τιμής όμως η Αρνητική διωνυμική (με μαύρο χρώμα) αυξάνεται εκθετικά και αυτό συμβαίνει λόγο της συνάρτησης διασποράς της $V(y_i) = \mu_i + \kappa\mu_i^2$. Για την κατασκευή του γραφήματος έχει χρησιμοποιηθεί η παράμετρος μεταβλητότητας που παρέχει το `negbin12`. Μέσω της `R` λαμβάνεται η τιμή `theta=0.2399` η οποία είναι η αντίστροφη της παραμέτρου υπερμεταβλητότητας κ. Άρα, η τιμή της παραμέτρου αυτής είναι ίση με $\frac{1}{0.24} = 4.17$ και έτσι η συνάρτηση διασποράς είναι η $V(y_i) = \mu_i + 4.17\mu_i^2$. Από τον έλεγχο υπερμεταβλητότητας που έγινε στο μοντέλο `poisson12` είχε προκύψει τιμή 3.726 και άρα η $\theta = \frac{1}{3.726} = 0.27$ προσεγγίζει την εκτιμώμενη `theta=0.24` που προκύπτει από το μοντέλο `negbin12`. Οι δύο αυτές εκτιμήσεις αφορούν την ίδια ποσότητα και παρέχουν μία ένδειξη για την υπερμεταβλητότητα. Παρολ' αυτά, διαφέρουν λόγω της διαφορετικής μεθόδου εκτίμησης τους. Επομένως, ασφαλιζόμενοι με μεγαλύτερη αναμενόμενη συχνότητα θα έχουν και αυξημένη διασπορά. Ασφαλιζόμενοι με υψηλή αναμενόμενη συχνότητα θεωρούνται ως επιρρεπής στο ρίσκο και άρα θα έχουν τη μεγαλύτερη μεταβλητότητα στη συχνότητα των απαιτήσεων που θα αναφέρουν. Τέλος, προσέξτε τον πολύ χαμηλότερο δείκτη AIC του μοντέλου `negbin12` σε σχέση με το poisson12, που ισούται με 85935 μονάδες και άρα προσφέρει καλύτερη προσαρμοστικότητα στα δεδομένα.

```
> summary(negbin12)

Call:
glm.nb(formula = N.claims.year ~ Driver.age.cut + Vehicle.age.cut +
    Length + Type.risk + Type.fuel + Area + offset(log(Exposure)),
    data = train.data, init.theta = 0.2399593297, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1154  -0.7505  -0.7020  -0.5707   4.0022

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.05323    0.16049 -12.793  < 2e-16 ***
```



Διάγραμμα 5.1: Διασπορά απαιτήσεων σε σχέση με τη μέση τιμή

```
Driver.age.cut[25, 44]   -0.38211     0.08500   -4.495 6.94e-06 ***
Driver.age.cut[45, 59]   -0.32280     0.08524   -3.787 0.000153 ***
Driver.age.cut[60, 69]   -0.69877     0.08970   -7.790 6.68e-15 ***
Driver.age.cut[70, Inf)  -0.82368     0.09818   -8.390  < 2e-16 ***
Vehicle.age.cut6+         0.21471     0.03603    5.958 2.55e-09 ***
Length                    0.36283     0.03145   11.536  < 2e-16 ***
Type.risk2                0.08823     0.03540    2.493 0.012681 *
Type.fuelP               -0.18100     0.02602   -6.957 3.47e-12 ***
Area1                     0.23811     0.02541    9.372  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01
   *    0.05    .    0.1         1

(Dispersion parameter for Negative Binomial(0.24) family taken to be 1)

    Null deviance: 29226  on 49276  degrees of freedom
Residual deviance: 28566  on 49267  degrees of freedom
AIC: 85935

Number of Fisher Scoring iterations: 1


              Theta:  0.23996
          Std. Err.:  0.00393

 2 x log-likelihood:  -85912.84000
```

5.1.3 Μοντέλο zip

Θα γίνει η παρουσίαση των αποτελεσμάτων του τελικού μοντέλου Μηδενικής διόγκωσης που επιλέχθηκε, το οποίο έχει αποθηκευθεί στο αντικείμενο zip12.

Η τιμή του σταθερού όρου β_0 είναι -0.70 με τυπικό σφάλμα 0.09, ενώ το στατιστικό ελέγχου T_1 για τον έλεγχο Wald του σταθερού όρου με υποθέσεις $H_0 : \beta_0 = 0$ έναντι της $H_1 : \beta_0 \neq 0$ έδωσε τιμή -7.34. Η p-τιμή του ελέγχου είναι πολύ μικρή οπότε υπάρχουν σοβαρές ενδείξεις για την απόρριψη της μηδενικής υπόθεσης. Με βάση το μοντέλο `zip12`, η συχνότητα των απαιτήσεων σε ετήσια έκθεση σε κίνδυνο $w_i = 1$, στο ηλικιακό εύρος [18, 24], με πετρελαιοκίνητο επιβατικό όχημα ηλικίας 5 ετών και κάτω, μηδενικού μήκους σε αγροτική περιοχή είναι $e^{\beta_0} = e^{-0.70} = 0.50$ το οποίο δεν είναι ρεαλιστικό αφού δεν υπάρχει τέτοια παρατήρηση.

Για τη μεταβλητή `Driver.age.cut` έχουμε ότι στο ηλικιακό εύρος [25, 44] η αναμενόμενη συχνότητα μειώνεται κατά $\frac{1-e^{-0.12}}{1} = 10.41\%$ σε σχέση με το ηλικιακό εύρος βάσης [18, 24], εάν η έκθεση, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος οχήματος, τύπος καυσίμου και η περιοχή δε μεταβληθούν. Από την p-τιμή του ελέγχου του συντελεστή της εικονικής μεταβλητής `Driver.age.cut[45, 59]`, δεν υπάρχουν σοβαρές ενδείξεις για την απόρριψη της μηδενικής υπόθεσης ότι η τιμή του είναι μηδέν και άρα η αναμενόμενη συχνότητα σε αυτό το εύρος δε μεταβάλλεται σε σχέση με το ηλικιακό εύρος βάσης [18, 24], δεδομένου ότι η έκθεση, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος οχήματος, τύπος καυσίμου και η περιοχή δε μεταβληθούν. Σε ηλικίες [60, 69] η συχνότητα των απαιτήσεων είναι κατά $\frac{1-e^{-0.18}}{1} = 16.47\%$ μικρότερη σε σχέση με την αναμενόμενη συχνότητα των ηλικιών βάσης, δεδομένου ότι η έκθεση, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος οχήματος, τύπος καυσίμου και η περιοχή δε μεταβληθούν. Οι ηλικίες 70+ σημειώνουν μείωση κατά $1 - e^{-0.17} = 15.63\%$, σε σχέση με τις ηλικίες βάσης εάν η έκθεση, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος οχήματος, τύπος καυσίμου και η περιοχή δε μεταβληθούν.

Η μεταβλητή `Vehicle.age.cut` είναι επίσης σημαντική αφού με βάση τον έλεγχο Wald του συντελεστή η p-τιμή είναι πάρα πολύ μικρή και άρα η ηλικία του οχήματος σχετίζεται με τη συχνότητα των απαιτήσεων. Συγκεκριμένα, η εκτίμηση του συντελεστή της `Vehicle.age.cut6+` είναι 0.20 με τυπική απόκλιση ίση με 0.025 και p-τιμή σχεδόν μηδέν. Επομένως, όταν η ηλικία του οχήματος είναι 6 χρόνια και άνω υπάρχει αύξηση της αναμενόμενης συχνότητας κατά $\frac{e^{0.1985}-1}{1}100\% = 21.95\%$ σε σχέση με οχήματα ηλικίας 5 ετών και κάτω, δεδομένου ότι η έκθεση, το ηλικιακό εύρος, το μήκος, ο τύπος του οχήματος, ο τύπος καυσίμου και η περιοχή παραμείνουν σταθερά.

Για τη μεταβλητή `Length` έχει εκτιμηθεί συντελεστής με τιμή 0.2615. Αφού αυτή είναι μία συνεχής μεταβλητή, η αύξηση του μήκους ενός οχήματος κατά ένα μέτρο θα φέρει αύξηση στην αναμενόμενη συχνότητα των απαιτήσεων κατά $\frac{e^{0.2615}-1}{1}100\% = 29.89\%$, δεδομένου ότι οι μεταβλητές έκθεσης, εύρους ηλικίας, κατηγορίας ηλικίας οχήματος, τύπος οχήματος, καυσίμου και περιοχής παραμείνουν σταθερές.

Η μεταβλητή `Type.risk`, φαίνεται να είναι στατιστικά σημαντική σε όλες τις κατηγορίες. Βασική κατηγορία είναι η τύπου 3, δηλαδή επιβατικό όχημα. Με βάση την εκτίμηση του συντελεστή της εικονικής μεταβλητής `Type.risk2` που ισούται με 0.1095 προκύπτει ότι σε σχέση με ένα επιβατικό όχημα, η αναμενόμενη συχνότητα απαιτήσεων στην περίπτωση φορτηγού είναι κατά $\frac{e^{0.1095}-1}{1}100\% = 11.57\%$ μεγαλύτερη, δεδομένου ότι η έκθεση, το ηλικιακό εύρος, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος καυσίμου και περιοχή παραμένουν σταθερά.

Για τη μεταβλητή `Type.fuel`, η συμπερίληψη της στο μοντέλο κρίνεται και αυτή με βάση τον έλεγχο Wald ως σημαντική. Κατηγορία βάσης είναι το πετρέλαιο. Η εκτίμηση του συντελεστή της `Type.fuelP` είναι ίση με -0.12. Αυτό σημαίνει πως η χρήση βενζίνης σε ένα όχημα φέρει μείωση στον αναμενόμενο αριθμό απαιτήσεων κατά $\frac{1-e^{-0.1280}}{1}100\% = 13.65\%$, σε σχέση με τη χρήση πετρελαίου, δεδομένου ότι η έκθεση, το ηλικιακό εύρος, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος

οχήματος και η περιοχή παραμένουν σταθερά.

Τέλος, για τη μεταβλητή `Area` έχει κατασκευαστεί η εικονική μεταβλητή `Area1`, η οποία αναγνωρίζει περιοχές τύπου 1, δηλαδή κατοικημένες. Κατηγορία βάσης είναι οι αγροτικές περιοχές. Η προσθήκη της `Area` κρίνεται στατιστικά σημαντική και έτσι υπάρχουν ενδείξεις για την απόρριψη της μηδενικής υπόθεσης. Επομένως, με βάση την τιμή του συντελεστή της εικονικής μεταβλητής `Area1` η αναμενόμενη συχνότητα απαιτήσεων αυξάνεται κατά $\frac{e^{0.1372}-1}{1}100\% = 14.70\%$ σε κατοικημένες περιοχές σε σχέση με αγροτικές, δεδομένου ότι η έκθεση, το ηλικιακό εύρος, η κατηγορία ηλικίας οχήματος, το μήκος, ο τύπος οχήματος και καυσίμου παραμείνουν σταθερά.

Για τη λογιστική παλινδρόμηση έχει χρησιμοποιηθεί η μεταβλητή `Driver.age`, ο σταθερός όρος και ο όρος `offset=log(Exposure)`. Η παλινδρόμηση πραγματοποιείται με σκοπό να μοντελοποιηθεί η πιθανότητα π_i ένας ασφαλιζόμενος να ανήκει στην πρώτη ομάδα, όπου δεν αναφέρονται καθόλου απαιτήσεις. Η εκτίμηση του σταθερού όρου είναι 0.29, και του συντελεστή της `Driver.age` είναι 0.01, ενώ θεωρουνται και οι δύο στατιστικά σημαντικοί. Από τις σχέσεις 2.15 και 2.14 προκύπτει ότι ο λόγος των πιθανοτήτων π_i και $1-\pi_i$ είναι

$$\frac{\mathbb{P}(Z_i = 1 \,|\, q_i)}{\mathbb{P}(Z_i = 0 \,|\, q_i)} = e^{\eta_i}$$

και άρα

$$\ln \frac{\mathbb{P}(Z_i = 1 \,|\, q_i)}{\mathbb{P}(Y_i = 0 \,|\, q_i)} = \eta_i,$$

με $Z_i \in \{0, 1\}$, $i = 1, \cdots, n$ και $Z_i = 1$, αν ο i ασφαλιζόμενος ανήκει στην πρώτη ομάδα και 0 αλλιώς και q_i η τιμή της επεξηγηματικής μεταβλητής `Driver.age`. Το μοντέλο της λογιστικής παλινδρόμησης γράφεται ως

$$\ln \frac{\pi_i}{1-\pi_i} = \eta_i = \gamma_0 + \gamma_1 q_i + \ln w_i,$$

με γ_j, $j = 0, 1$ οι τιμές των συντελεστών του μοντέλου, w_i η έκθεση σε χρόνια και q_i η τιμή της επεξηγηματικής μεταβλητής `Driver.age`. Επομένως, η τιμή $e^{\gamma_0} = e^{0.29} = 1.33$, είναι η σχετική πιθανότητα (`odd`) επιτυχίας για ασφαλιζόμενο ηλικίας μηδέν και με έκθεση ενός έτους. Προφανώς, η ερμηνεία αυτή δεν είναι ρεαλιστική αφού δεν υπάρχει μηδενική ηλικία ασφαλιζόμενου. Η τιμή e^{z_1} εκφράζει τη μεταβολή της σχετικής πιθανότητας όταν η ηλικία αυξηθεί κατά ένα έτος και η έκθεση παραμείνει σταθερή. Άρα, αφού $\gamma_1 = 0.01372$ και $e^{0.01372} = 1.0138$ η σχετική πιθανότητα αυξάνεται κατά 1.38% για κάθε έτος που περνά. Ως επιτυχία ορίζεται η $Z_i = 1$ δηλαδή η i-οστή παρατήρηση να ανήκει στην πρώτη ομάδα. Άρα, σε μεγαλύτερους κατά ένα έτος σε ηλικία οδηγούς η πιθανότητα μηδενικής μέτρησης απαιτήσεων αυξάνεται κατά 1.38%. Η αύξηση της πιθανότητας π δείχνει ότι με την αύξηση της ηλικίας είναι πιο πιθανόν να μην αναφερθούν καθόλου απαιτήσεις σε σχέση με το να αναφερθεί τουλάχιστον μία. Αυτό γενικότερα δείχνει την τάση μεγαλύτερων σε ηλικία οδηγών να έχουν μηδενικές απαιτήσεις είτε λόγω μειωμένου χρόνου οδήγησης, περισσότερης εμπειρίας και άλλων λόγων.

```
> summary(zip12)

Call:
zeroinfl(formula = N.claims.year ~ Driver.age.cut + Vehicle.age.cut
+ Length + Type.risk + Type.fuel + Area + offset(log(Exposure))
| Driver.age + offset(log(Exposure)),
```

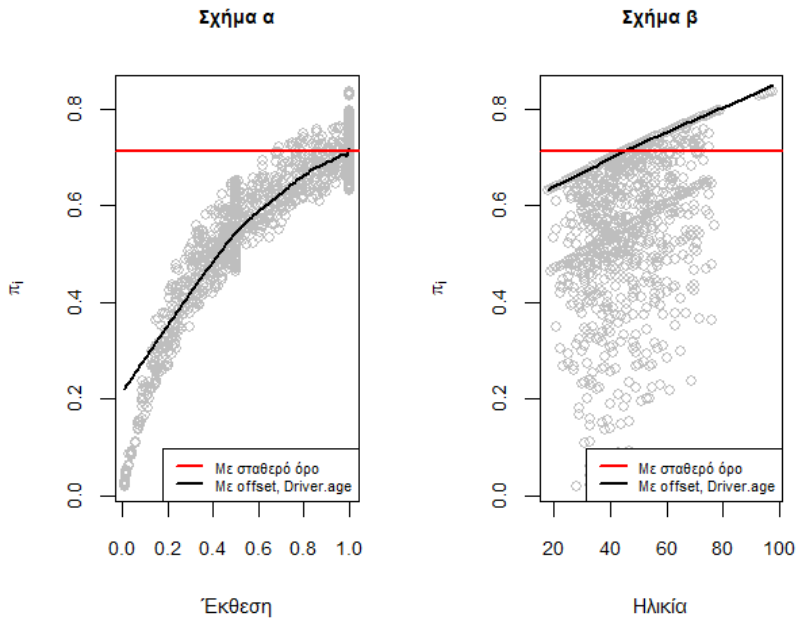
```
data = train.data, dist = "poisson", link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.6922 -0.4828 -0.4528 -0.3761 17.0020

Count model coefficients (poisson with log link):
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.69816    0.09515  -7.337 2.18e-13 ***
Driver.age.cut[25, 44]  -0.11796    0.05013  -2.353  0.01862 *
Driver.age.cut[45, 59]   0.02982    0.05043   0.591  0.55432
Driver.age.cut[60, 69]  -0.18037    0.05554  -3.247  0.00117 **
Driver.age.cut[70, Inf) -0.17086    0.06546  -2.610  0.00905 **
Vehicle.age.cut6+        0.19847    0.02564   7.740 9.95e-15 ***
Length                   0.26150    0.01822  14.352  < 2e-16 ***
Type.risk2               0.10955    0.02232   4.909 9.16e-07 ***
Type.fuelP              -0.12801    0.01840  -6.959 3.43e-12 ***
Area1                    0.13722    0.01623   8.454  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.29044    0.04840   6.001 1.96e-09 ***
Driver.age   0.01372    0.00100  13.713  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 43
Log-likelihood: -4.429e+04 on 12 Df
```

Η χρήση μοντέλων Μηδενικής διόγκωσης είναι επίσης χρήσιμη στη μοντελοποίηση της πιθανότητας αναφοράς κάποιας απαίτησης. Υπάρχουν δύο ερμηνείες που σχετίζονται με τα μοντέλα αυτά [4].

Η πρώτη ερμηνεία έγκειται στην υπόθεση ότι κάποιοι ασφαλιζόμενοι δεν πρόκειται να αναφέρουν καμία απαίτηση στη διάρκεια του χρόνου, ανεξαρτήτως του ρίσκου που έχουν και αυτό φαίνεται από το πρώτο σκέλος της συνάρτησης μάζας πιθανότητας 3.3. Παρ΄ όλη τη συμπεριφορά αυτή, η επιθυμία ασφαλιστικής κάλυψης υπάρχει ως τυπική ασφάλιση απέναντι στον κίνδυνο ακόμη και αν η πιθανότητα ατυχήματος είναι σχεδόν μηδενική.

Η δεύτερη ερμηνεία σχετίζεται με την υπόθεση ότι ο αριθμός των ατυχημάτων είναι κατανεμημένος με βάση την κατανομή Poisson. Η πιθανότητα αναφοράς ενός ατυχήματος θεωρείται καθοριστική. Το πρώτο ατύχημα κάθε ασφαλιζόμενου είναι αυτό που θα ορίσει τη συμπεριφορά του στο υπόλοιπο του χρόνου κάλυψης. Συγκεκριμένα, αν το πρώτο ατύχημα αναφερθεί τότε θα ακολουθήσουν και τα υπόλοιπα. Αν το πρώτο ατύχημα δεν αναφερθεί τότε δε θα ακολουθήσουν άλλα. Ασφαλιζόμενοι οι οποίοι δεν αναφέρουν το πρώτο ατύχημα, επιμένουν στην απόφαση τους αυτή και σε επόμενα πιθανά ατυχήματα, υποστηρίζοντας οικονομικά την απόφαση αυτή μόνοι τους. Η συμπεριφορά αυτή φαίνεται από τη συνάρτηση μάζας πιθανότητας 3.3, η οποία ορίζει τη συνάρτηση μάζας πιθανότητας σε ένα από τα δύο σκέλη.

Κατά την πρώτη ερμηνεία αφού ισχύει η υπόθεση ότι πάντα θα υπάρχουν πελάτες με μηδενι-



Διάγραμμα 5.2: Πιθανότητα μηδενικών απαιτήσεων σε σχέση με την έκθεση (αριστερό γράφημα) και την ηλικία του οδηγού (δεξί γράφημα)

κές απαιτήσεις (πρώτη ομάδα), στη μοντελοποίηση της πιθανότητας π δε χρειάζεται ο όρος `offset=log(Exposure)` αλλά ούτε και η `Driver.age` στο λογιστικό μοντέλο, αφού η πιθανότητα είναι ανεξάρτητη του χρόνου κάλυψης και της ηλικίας. Αντιθέτως, κατά τη δεύτερη ερμηνεία θα πρέπει να συμπεριληφθεί ο όρος `offset=log(Exposure)` στο λογιστικό μοντέλο, αφού επηρεάζει την πιθανότητα της αναφοράς του πρώτου ατυχήματος η οποία είναι και η πιθανότητα ύπαρξης τουλάχιστον μίας απαίτησης, $(1 - \pi_i)$. Παρατηρείστε ότι κατά την κατασκευή του τελικού μοντέλου `zip12`, έχει συμπεριληφθεί ο όρος `offset=log(Exposure)` και στις δύο κατανομές. Στο Διάγραμμα 5.2 φαίνεται πως μεταβάλλεται η π_i σε σχέση με την έκθεση στο αριστερό γράφημα και την ηλικία στο δεξί γράφημα. Η οριζόντια ευθεία στο αριστερό γράφημα είναι η πιθανότητα π, με βάση την πρώτη υπόθεση και το λογιστικό μοντέλο με τη χρήση μόνο σταθερού όρου. Κατά τη δεύτερη υπόθεση, η σχέση μεταξύ έκθεσης και πιθανότητας φαίνεται αυξητική. Με τη χρήση της συνάρτησης `lowess` προσαρμόζεται η καμπύλη με μαύρο χρώμα που είναι η πιθανότητα π_i. Η αυξητική τάση επαληθεύει το συμπέρασμα προηγούμενης παραγράφου σχετικά με την αύξηση της π με την αύξηση της έκθεσης w. Αντίστροφα, η πιθανότητα ύπαρξης τουλάχιστον μίας απαίτησης μειώνεται. Ενδιαφέρον είναι ότι τα δύο μοντέλα ταυτίζονται, όπως φαίνεται στο αριστερό σχήμα του Διαγράμματος 5.2, στην περίπτωση που η έκθεση είναι ίση με ένα έτος. Αυτό συμβαίνει διότι όταν η έκθεση αφορά ένα ολόκληρο έτος τότε η χρήση του όρου `offset` δεν είναι αναγκαία για τη διόρθωση των αποτελεσμάτων και έτσι τα δύο μοντέλα ταυτίζονται. Ομοίως, στο δεξί γράφημα του Διαγράμματος 5.2 φαίνεται η σχέση μεταξύ της ηλικίας του οδηγού και της πιθανότητας π_i. Η οριζόντια ευθεία είναι η πιθανότητα με τη χρήση σταθερού όρου στη λογιστική παλινδρόμηση ενώ η προσαρμοσμένη καμπύλη αφορά τη δεύτερη υπόθεση. Παρατηρείστε την αύξηση της πιθανότητας π_i με την αύξηση της ηλικίας.

Ένας άλλος τρόπος κατασκευής μοντέλων Μηδενικής διόγκωσης στην `R`, είναι μέσω της εντολής `glmmTMB` της βιβλιοθήκης `glmmTMB`. Αυτή έχει τη δυνατότητα να εκτιμήσει τυπικά σφάλματα μέσω

της εντολής `predict(, se.fit = TRUE)` ενώ η κατασκευή αντικειμένων μέσω της `zeroinfl` δε δίνει αυτή τη λειτουργία η οποία θα χρειαστεί για τον υπολογισμό διαστημάτων εμπιστοσύνης για τις προβλέψεις σε επόμενο έτος.

5.2 Προβλέψεις

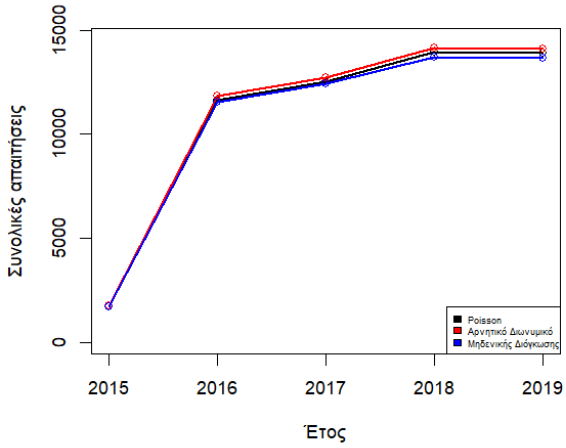
Η εφαρμογή των κατασκευασμένων μοντέλων κατανομών μέτρησης στον ασφαλιστικό τομέα έγκειται στο να γίνουν οι απαραίτητες εκτιμήσεις για τις απαιτήσεις επόμενων ετών με σκοπό ένας ασφαλιστικός φορέας να μπορεί να εξασφαλίσει το απαιτούμενο αποθεματικό για να μπορεί να ανταποκριθεί σε μελλοντικές αποζημιώσεις. Αν θεωρηθεί ως τρέχον έτος το τελευταίο έτος στο οποίο συνήφθει ένα συμβόλαιο, τότε αυτό είναι το 2018. Θεωρούνται ως ενεργοί ασφαλιζόμενοι όσοι έχουν ανανεώσει το συμβόλαιο τους μέσα σε αυτό το έτος. Θα υπολογιστεί ο συνολικός αριθμός απαιτήσεων για τους ασφαλιζόμενους αυτούς για το επόμενο έτος, 2019, δεδομένου ότι όλοι έχουν ανανεώσει ξανά μέσα σε αυτό το έτος. Στην πράξη, σε κάθε επόμενο έτος εισέρχονται και νέοι ασφαλιζόμενοι με διαφορετικά ρίσκα μέσα στο χαρτοφυλάκιο, όμως δεν εξετάζεται η περίπτωση αυτή αφού απαιτεί νέα δεδομένα, αν και είναι η πραγματική. Επίσης, θα γίνουν οι εκτιμήσεις για τα έτη 2015 με 2018 από τους τρεις τύπους μοντέλων.

Επομένως, για την εκτίμηση των απαιτήσεων στο έτος 2019, θα πρέπει να μεταβληθούν οι τιμές των επεξηγηματικών μεταβλητών με σκοπό να αντιστοιχούν σε ένα έτος μεγαλύτερο. Στο νέο δείγμα συμμετέχουν μόνο οι ενεργοί ασφαλιζόμενοι το 2018. Υπάρχουν συνολικά 28,493 ενεργές εγγραφές σε αυτό το έτος. Θα χρησιμοποιήσουμε αυτό το δείγμα για την εκτίμηση των απαιτήσεων του επόμενου έτους. Στο νέο δείγμα, αυξάνουμε τις μεταβλητές `Driver.age` και `Vehicle.age` κατά ένα έτος, ενώ οι μεταβλητές `Driver.age.cut` και `Vehicle.age.cut` θα μεταβληθούν ανάλογα με τις ανανεωμένες τιμές των `Driver.age` και `Vehicle.age`, αντίστοιχα. Επιλέγεται επίσης η μεταβλητή έκθεσης, `Exposure`, να παραμείνει ως έχει. Στον Πίνακα 5.2 παρουσιάζεται ο εκτιμώμενος αριθμός απαιτήσεων για τα έτη 2015 με 2019 και για κάθε έναν από τους τρεις τύπους μοντέλων ενώ έχει εκτιμηθεί και αντίστοιχο 95% διάστημα εμπιστοσύνης με βάση το τυπικό σφάλμα της εκτίμησης για τον αριθμό απαιτήσεων του έτους 2019. Στο Διάγραμμα 5.3 φαίνεται η γραφική αναπαράσταση των αναμενόμενων εκτιμήσεων για τα έτη 2015 με 2019 από τα τρία κατασκευασμένα μοντέλα `poisson12` (μαύρο χρώμα), `negbin12` (κόκκινο χρώμα), `zip12` (μπλέ χρώμα). Από αυτό παρατηρείται ότι οι εκτιμήσεις των τριών μοντέλων είναι πάρα πολύ κοντά, αυτές όμως αφορούν μόνο το δοσμένο δείγμα δεδομένων.

	2015	2016	2017	2018	2019	95% δ.ε	
`poisson12`	1733	11656	12528	13954	13639	13895	13902
`negbin12`	1757	11819	12703	14150	14098	14092	14104
`zip12`	1720.98	11553.26	12409.86	13718.07	13639.44	13635.41	13643.52

Πίνακας 5.2: Αναμενόμενος αριθμός απαιτήσεων

Για το τελευταίο μοντέλο, `zip12` που θεωρείται πως είναι το καταλληλότερο, έχει υπολογιστεί ένα 95% διάστημα εμπιστοσύνης ίσο με [13635.41, 13643.52]. Επομένως, το συγκεκριμένο διάστημα εμπιστοσύνης θα περιέχει την πραγματική τιμή του αριθμού απαιτήσεων στο έτος 2019 με πιθανότητα 0.95.



Διάγραμμα 5.3: Αναμενόμενος αριθμός απαιτήσεων για κάθε έτος από το `poisson12` (μαύρο χρώμα), `negbin12` (κόκκινο χρώμα) και `zip12` (μπλε χρώμα)

5.3 Συμπεράσματα και Συζήτηση

Σκοπός της διπλωματικής εργασίας αυτής είναι να μελετηθεί η σχέση μεταξύ διαφόρων μεταβλητών και της συχνότητας των απαιτήσεων. Μέσα από την ανάλυση των αποτελεσμάτων των μοντέλων και τη συμπερασματολογία έχουν εντοπιστεί οι μεταβλητές εκείνες οι οποίες σχετίζονται περισσότερο και λιγότερο με τη συχνότητα των απαιτήσεων.

Με βάση τους συντελεστές των μεταβλητών έχει διαπιστωθεί πως η σημαντικότερη μεταβλητή είναι η μεταβλητή μήκους του οχήματος (`Length`). Συγκεκριμένα, η αύξηση της τιμής του μήκους κατά ένα μέτρο, προκαλεί αύξηση της αναμενόμενης συχνότητας κατά 29.89%, δεδομένου ότι οι υπόλοιπες μεταβλητές παραμείνουν σταθερές. Η μεταβολή αυτή είναι η εντονότερη και έτσι δημιουργεί ερωτήματα σχετικά με τη ρεαλιστική της ερμηνεία. Θα ήταν πιο λογικό η μεταβολή αυτή να αφορούσε κυρίως οχήματα τύπου 2, δηλαδή φορτηγά. Αν αντί της μεταβλητής `Length` χρησιμοποιηθεί η αλληλεπίδραση `Length:Typerisk` θα δημιουργηθούν δύο εικονικές μεταβλητές, ανάλογα με τη μεταβλητή `Typerisk`. Σε αυτή την περίπτωση η ερμηνεία ίσως να είναι λογικότερη. Στην προσαρμογή του μοντέλου με χρήση αλληλεπίδρασης, φαίνεται πως οι συντελεστές των δύο εικονικών μεταβλητών `Length:Typerisk1` και `Length:Typerisk2` δε διαφέρουν αρκετά και έτσι η ερμηνεία δε βελτιώνεται. Από την ανάλυση διασποράς της προσθήκης της `Length` φαίνεται πως η μεταβλητή αυτή είναι από τις σημαντικότερες λόγω του συντελεστή της και της καλύτερης προσαρμοστικότητας που προσφέρει. Είναι λοιπόν κάποια ένδειξη ότι σε αυτή υπάρχει και άλλη πληροφορία εκτός του μήκους. Είναι μία μεταβλητή η οποία εμμέσως μπορεί να ορίσει και τον τύπο του οχήματος, δεχόμενοι ότι οχήματα με αυξημένο μήκος προσεγγίζουν το φορτηγό ενώ τα υπόλοιπα ένα επιβατικό αυτοκίνητο καθώς παράλληλα ορίζει και την περιοχή αν θεωρηθεί πως οχήματα μεγάλου μήκους κινούνται κυρίως σε αγροτικές περιοχές. Η θεώρηση αυτή έρχεται συμφωνεί με τη διαπίστωση κατά την οποία η συχνότητα είναι αυξημένη σε φορτηγά παρά σε επιβατικά οχήματα. Είναι επομένως μία επεξηγηματική μεταβλητή της οποίας η επίδραση είναι σημαντική. Εν μέρει, η συμπεριφορά αυτή μπορεί να ερμηνευθεί λόγω της οδικής δυσκολίας που αντιμετωπίζουν μακρυά οχήματα.

Επιπλέον, σημαντικές είναι οι μεταβλητές ηλικίας. Για την ηλικία του αυτοκινήτου χρησιμο-

ποιήθηκε ένας διαχωρισμός σε δύο κατηγορίες (`Vehicle.age.cut`) ανάλογα με το αν αυτή είναι μεγαλύτερη ή όχι των 5 ετών. Διαπιστώθηκε αύξηση της αναμενόμενης συχνότητας σε ηλικίες μεγαλύτερες των 5 ετών σε σχέση με οχήματα μικρότερης ηλικίας. Για παράδειγμα, αυτό μπορεί να οφείλεται στο ότι νεότεροι οδηγοί συνήθως οδηγούν παλαιότερα οχήματα λόγω της αδυναμίας τους να χειριστούν νεότερα ή λόγω του χαμηλότερου κόστους επισκευής τους και άρα τα παλαιότερα οχήματα βρίσκονται σε υψηλότερο ρίσκο. Παράλληλα, με την πάροδο του χρόνου τα οχήματα φέρουν περισσότερη φθορά και άρα η πιθανότητα μηχανολογικού ή οποιουδήποτε άλλου προβλήματος είναι μεγαλύτερη, αυξάνοντας έτσι και την πιθανότητα ατυχήματος.

Για τη μεταβλητή ηλικίας (`Driver.age.cut`) έγινε διαχωρισμός σε πέντε ηλικιακές ομάδες. Το μοντέλο έδειξε πως σε όλες τις κατηγορίες υπάρχει μείωση της αναμενόμενης συχνότητας απαιτήσεων σε σχέση με το ηλικιακό εύρος βάσης (18 με 24 ετών), με το ηλικιακό εύρος 45 με 59 να είναι αυτό με τη μικρότερη μείωση. Αυτό σημαίνει πως σε αυτό το εύρος οι συχνότητες είναι αυξημένες. Η ανάλυση των δεδομένων θα μπορούσε να αποφέρει ακόμη πιο επιθυμητά αποτελέσματα. Υπήρξαν περιορισμοί και αυτό οφείλετε στο ότι το δείγμα δεν ήταν πλήρης. Η απουσία ηλικιών από 80 μέχρι 92 ετών δεν έχει συμβάλει στην αναγνώριση της τάσης αύξησης της αναμενόμενης συχνότητας απαιτήσεων σε αυτές τις ηλικίες που παρατηρήθηκε στη διερευνητική ανάλυση. Πιο συγκεκριμένα, αν και αρχικά ο διαχωρισμός των δύο τελευταίων κατηγοριών ηλικίας ήταν 70 μέχρι 79 και 80 και άνω, στη μοντελοποίηση επιλέχθηκε η συγχώνευση των δύο αυτών κατηγοριών με αποτέλεσμα λόγω των αυξημένων παρατηρήσεων του πρώτου εύρους ηλικιών να δοθεί περισσότερη σημασία στην καθοδική τάση της συχνότητας σε αυτό το διάστημα ηλικιών. Επομένως, οι περιορισμένες παρατηρήσεις του δείγματος για ηλικίες 80 και άνω οδήγησαν στο να μην μπορεί να αναγνωριστεί η αυξητική τάση στο τελευταίο εύρος. Ακόμη και αν δεν επιλεγόταν η συγχώνευση των κατηγοριών, η προσαρμογή του νέου μοντέλου δεν άλλαζε σχεδόν καθόλου.

Η ανάλυση έδειξε πως ο τύπος καυσίμου (`Type.fuel`) που δέχεται ένα όχημα επηρεάζει την αναμενόμενη συχνότητα απαιτήσεων και αυτό είναι κάτι που συμβαίνει στον ασφαλιστικό κόσμο. Συγκεκριμένα, φαίνεται πως με τη χρήση βενζίνης η αναμενόμενη συχνότητα είναι μειωμένη κατά 12% σε σχέση με τη χρήση πετρελαίου. Αυτό συμβαίνει γενικότερα λόγω του ότι πετρελαιοκίνητα οχήματα χρησιμοποιούνται κυρίως για εμπορικούς σκοπούς και άρα είναι εκτεθειμένα στον χώρο και χρόνο περισσότερο σε σχέση με βενζινοκίνητα οχήματα τα οποία χρησιμοποιούνται κυρίως στην πόλη. Θυμηθείτε από την περιγραφική στατιστική, ότι το πετρέλαιο χρησιμοποιείται κατά κόρον σε φορτηγά και αγροτικά οχήματα. Συγκεκριμένα, το 62% των φορτηγών χρησιμοποιεί καύσιμο πετρελαίου, ενώ το 100% των αγροτικών οχημάτων χρησιμοποιούν επίσης πετρέλαιο. Επομένως, ο τύπος καυσίμου επηρεάζει σημαντικά τη συχνότητα απαιτήσεων και άρα είναι ένας παράγοντας ο οποίος λαμβάνεται σοβαρά υπόψη κατά την ανάπτυξη μοντέλων.

Διαπιστώθηκε επίσης πως υπάρχει μεταβολή της αναμενόμενης συχνότητας ανάλογα με τον τύπο του οχήματος (`Type.risk`). Δυστυχώς, το πολύ μικρό δείγμα μοτοσικλετών (τύπος 1) και αγροτικών οχημάτων (τύπος 4) δεν επέτρεψε στην ανάλυση μας να ενσωματώσει τους τύπους αυτούς. Σε ολόκληρο το δείγμα υπάρχουν μόνο 2 εγγραφές μοτοσικλέτας και 20 αγροτικού οχήματος όπου σε καμία από τις δύο περιπτώσεις δεν αναφέρθηκε κάποια απαίτηση. Η αφαίρεση λοιπόν του υποδείγματος εγγραφών τύπου 1 και 4 από τα δείγματα μάθησης και ελέγχου ήταν αναγκαία, καθώς κατά την προσαρμογή του μοντέλου με όλους τους τύπους παράγονταν τεράστια τυπικά σφάλματα στις εικονικές μεταβλητές `Type.risk1` και `Type.risk4`, με αποτέλεσμα να επηρέαζαν την ακρίβεια της ανάλυσης. Επομένως, η ανάλυση αφορούσε μόνο φορτηγά (τύπος 2) και επιβατικά οχήματα (τύπος

3). Τα αποτελέσματα έδειξαν πως η αναμενόμενη συχνότητα είναι αυξημένη σε φορτηγά κατά 11% σε σχέση με επιβατικά οχήματα. Ομοίως με πριν, η εντονότερη χρήση πετρελαίου σε φορτηγά φέρει αύξηση της συχνότητας των απαιτήσεων.

Όπως είναι λογικό, τα αποτελέσματα έδειξαν πως η περιοχή (`Area`) επηρεάζει επίσης τη συχνότητα των απαιτήσεων. Όπως αναμέναμε, η συχνότητα είναι υψηλότερη σε κατοικημένες περιοχές λόγω της έντονης κινητικότητας και άρα αυξημένου κινδύνου. Σε κατοικημένες περιοχές, η συχνότητα είναι αυξημένη κατά 14.71% σε σχέση με τις αγροτικές. Το συμπέρασμα αυτό έρχεται σε αντίθεση με τη θεώρηση ότι επιβατικά οχήματα που κινούνται κατά κύριο λόγο σε κατοικημένες περιοχές έχουν μειωμένη συχνότητα σε σχέση με φορτηγά που κινούνται συνήθως σε αγροτικές περιοχές. Παρόλ' αυτά η γενικότερη αυτή σκοπιά δεν ισχύει στο δείγμα μας καθώς δε φαίνεται να υπάρχει μεγάλη διαφορά μεταξύ του ποσοστού των τύπων οχημάτων που κινούνται στις δύο περιοχές. Για τα φορτηγά, το 75% και το 25% κινούνται σε αγροτικές και κατοικημένες περιοχές, αντίστοιχα, ενώ για επιβατικά οχήματα είναι 73% και 25%, αντίστοιχα.

Στην ανάλυση θα μπορούσε να συμμετάσχει όπως προηγήθηκε και η μεταβλητή έτους σύναψης του συμβολαίου (`Year`), όμως αυτό δε γίνεται λόγω του αδικαιολόγητου θορύβου που παράγει. Επίσης, χρήσιμη θα ήταν η ύπαρξη μεταβλητής γένους, `gender` κατά την οποία γενικότερα υπάρχει διαφοροποίηση της συχνότητας από άντρες σε γυναίκες.

Τέλος, η ανάλυση που έγινε αφορούσε τη συχνότητα των απαιτήσεων. Η σφοδρότητα των απαιτήσεων είναι επίσης κάτι που ενδιαφέρει τους μελετητές και αφήνεται ως θέμα για περισσότερη εντρύφηση. Γενικότερα, οι δύο αυτές ποσότητες συμπεριφέρονται αντιστρόφως ανάλογα. Ένα προφίλ ρίσκου στο οποίο οι συχνότητα των απαιτήσεων θα είναι μεγάλη, θα φέρει μικρή σφοδρότητα απαιτήσεων. Αυτό παρατηρείται στον Πίνακα 4.12 στην περίπτωση των μεταβλητών `Type.fuel` και `Type.risk`. Η συμπεριφορά αυτή παρατηρείται γενικότερα σε δείγματα που αφορούν χαρτοφυλάκια δεδομένων ασφάλισης.

Στη διπλωματική εργασία αυτή έγινε εντρύφηση στο αντικείμενο των Γενικευμένων Γραμμικών μοντέλων. Μέσα από τη διαδικασία μοντελοποίησης με τη χρήση αυτών των μοντέλων έχει διαπιστωθεί η τεράστια εφαρμογή τους και δυνατότητες που προσφέρουν στη μοντελοποίηση ποικίλων τύπων μεταβλητών και περιπτώσεων. Επίσης, μέσα από την ανάλυση των δεδομένων έγινε εμβάθυνση στο στατιστικό πακέτο της `R` και σε βιβλιοθήκες, όπου έγινε δουλεία μέσα από την οποία ξεπεράστηκαν οι προσδοκίες μου σχετικά με τις ικανότητες μου στον προγραμματισμό. Η εργασία αυτή ήταν μία νέα εμπειρία και πρόκληση, ειδικότερα όταν το αντικείμενο της εργασίας σχετίζεται με τις επαγγελματικές μου φιλοδοξίες.

Παράρτημα

```
install.packages("readxl")
library(readxl)
data <- read_excel("C:/Users/User/Downloads/Car_Insurance_Dataset.xlsx")
class(data)
data <- as.data.frame(data)

# CHECKS

# Desirable variables type

data$Date_lapse <- as.Date(data$Date_lapse, origin = "1970-01-01")
data$Date_start_contract <- as.Date(data$Date_start_contract
, origin ="1970-01-01")
data$Date_next_renewal <- as.Date(data$Date_next_renewal
, origin ="1970-01-01")
data$Date_last_renewal <- as.Date(data$Date_last_renewal
, origin = "1970-01-01")
data$Date_driving_licence <- as.Date(data$Date_driving_licence
, origin = "1970-01-01")
data$Date_birth <- as.Date(data$Date_birth
, origin = "1970-01-01")

# The desirable variables' names, replace _ with .
names(data)
names(data) <- gsub("_", ".", names(data))

# Desirable variables' type

data$Distribution.channel <- factor(data$Distribution.channel)
levels(data$Distribution.channel)

data$Type.fuel <- factor(data$Type.fuel)
levels(data$Type.fuel)

data$Area <-factor(data$Area)
levels(data$Area)

data$Type.risk <- factor(data$Type.risk)
```

```
levels(data$Type.risk)

data$N.door <- factor(data$N.door)

table(data$Distribution.channel)
data$Distribution.channel <- relevel(data$Distribution.channel
, ref = '0')

table(data$Type.risk)
data$Type.risk <- relevel(data$Type.risk, ref = '3')

table(data$Area)
data$Area <- relevel(data$Area, ref = '0')

table(data$Type.fuel)
data$Type.fuel <- relevel(data$Type.fuel, ref = 'D')

data$Length <- as.numeric(data$Length)

data$Weight <- as.numeric(data$Weight)

# Convert 'matriculation_year' to Date format
# (considering matriculation date as 1st January of that Year)
data$Matriculation.date <- as.Date(paste0(data$Year.matriculation
, "-01-01"))

# Data checks

data <- data[data$Date.next.renewal >= data$Date.last.renewal, ]
data <- data[data$Date.driving.licence <= data$Date.start.contract, ]
data <- data[data$Date.birth <= data$Date.start.contract, ]
data <- data[data$Date.last.renewal >= data$Matriculation.date, ]
data <- data[data$Date.last.renewal >= data$Date.driving.licence, ]
which(data$Date.lapse < data$Date.last.renewal)

# Construction of new variables

# Construction of variable 'Year'

install.packages("lubridate")
library(lubridate)
data$Year <- year(data$Date.last.renewal)
data$Year <- as.numeric(data$Year)

# Construction of variable 'Years.licenced' and 'Years.lic'

data$Years.licenced <- as.numeric(difftime(data$Date.last.renewal
, data$Date.driving.licence, units = "days")/365.25)
data$Years.licenced <- round(data$Years.licenced)
```

```
data$Years.lic <- factor(data$Years.licenced)
data$Years.lic <- relevel(data$Years.lic, ref ='12')
which(data$Years.licenced < 0)

# Construction of variable 'Driver.age' and 'Drv.age'

data$Driver.age <- as.numeric(difftime(data$Date.last.renewal
, data$Date.birth, units = "days")/365.25)
data$Driver.age <- round(data$Driver.age)
data$Drv.age <- factor(data$Driver.age)
data$Drv.age <- relevel(data$Drv.age, ref ='40')

# Construction of variable 'Vehicle.age' and 'Veh.age'

data$Vehicle.age <- as.numeric(difftime(data$Date.last.renewal
, data$Matriculation.date, units = "days")) / 365.25
data$Vehicle.age <- round(data$Vehicle.age)
data$Veh.age <- factor(data$Vehicle.age)

# Construction of variable 'Exposure'

data$Exposure <- ifelse(is.na(data$Date.lapse),
                        as.numeric(difftime(data$Date.next.renewal
                        , data$Date.last.renewal, units = "days"))
                        / 365.25,
                        ifelse(data$Date.lapse > data$Date.last.renewal,
                        as.numeric(difftime(data$Date.lapse
                        , data$Date.last.renewal, units = "days"))
                        / 365.25, 0))
data$Exposure <- round(data$Exposure, 2)
data$Exposure <- ifelse(data$Exposure > 1, 1, data$Exposure)

# Construction of range variables

age.intervals <- c(18, 25, 45, 60, 70, 80, Inf)
data$Driver.age.cut <- cut(data$Driver.age, age.intervals,
                           labels = c('[18,␣24]', '[25,␣44]', '[45,␣59]'
                           , '[60,␣69]', '[70,␣79]', '[80,␣Inf)')
                           , right = FALSE)

age.intervals <- c(18, 25, 45, 60, 70, Inf)
data$Driver.age.cut <- cut(data$Driver.age, age.intervals,
                           labels = c('[18,␣24]', '[25,␣44]', '[45,␣59]'
                           , '[60,␣69]', '[70,␣Inf)'), right = FALSE)

Veh.age.intervals <- c(0, 14, Inf)
data$Vehicle.age.cut <- cut(data$Vehicle.age, Veh.age.intervals,
                            labels = c('14-','15+'), right = FALSE)
```

```
Veh.age.intervals <- c(0, 5, Inf)
data$Vehicle.age.cut <- cut(data$Vehicle.age, Veh.age.intervals,
                            labels = c('5-','6+'), right = FALSE)


# Construction of Semester variable for each contract

# Define the breaks for each semester (including the start,end dates)
semester_breaks <- as.Date(c("2015-11-02", "2016-01-01", "2016-07-01",
"2017-01-01" , "2017-07-01", "2018-01-01", "2018-07-01", "2019-01-01"))

# Define the labels for each semester
semester_labels <- c("Semester␣1", "Semester␣2", "Semester␣3"
, "Semester␣4", "Semester␣5", "Semester␣6", "Semester␣7")

# Categorize Date.last.renewal into semesters
data$Semester <- cut(data$Date.last.renewal, breaks = semester_breaks
, labels = semester_labels, right = FALSE)

# Convert Semester to factor (if needed)
data$Semester <- factor(data$Semester, levels = semester_labels)

data$Semester.number <- as.numeric(gsub("Semester␣", ""
, data$Semester))


# REMOVE UNWANTED DATA

# Remove zero Exposures
data <- data[data$Exposure != 0.00, ]

# Remove NA values

data <- subset(data, Type.fuel %in% c("D", "P"))
data <- data[!is.na(data$Length), ]

# 2 FOLD CROSS VALIDATION

# Generate uniform random variables

set.seed(123)
n <- nrow(data)
u <- runif(n)

# Subset the dataset based on random numbers
# Training dataset: records with u <= 0.6
train.data <- data[u <= 0.6, ]
```

```
# Validation dataset: records with u > 0.6
test.data <- data[u > 0.6, ]

# AIC and Deviance Plot

predictors <- c('Length', 'Power', 'Type.fuel', 'Type.risk', 'Area'
, 'Weight', 'Years.licenced', 'Years.lic', 'Driver.age', 'drv.age'
, 'Vehicle.age', 'Veh.age', 'Value.vehicle', 'Premium'
, 'Cylinder.capacity', 'N.doors')
models <- lapply(predictors, function(x) {
  glm(formula(paste("N.claims.year ~", x))
  , family = poisson(link=log), data = train.data
  , offset=log(Exposure))})

deviance <- sapply(models, function(model) {
  deviance(model)
})
aic <- sapply(models, AIC)

plot(deviance, aic, xlab = "Deviance", ylab = "AIC", main = " "
, cex.lab = 0.85, cex.axis = 0.85)

poisson12 <- glm( N.claims.year ~  Driver.age.cut + Vehicle.age.cut
                  + Length + Type.risk + Type.fuel + Area,
                   family = poisson(link=log)
                   , data = train.data
                   , offset=log(Exposure))

library(MASS)
negbin12 <- glm.nb( N.claims.year ~  Driver.age.cut + Vehicle.age.cut
+ Length + Type.risk + Type.fuel + Area + offset(log(Exposure))
, data = train.data)

library(AER)
dispersiontest(poisson12, trafo = 2, alternative = 'greater')

install.packages("pscl")
library(pscl)
zip12 <- zeroinfl( N.claims.year ~  Driver.age.cut + Vehicle.age.cut +
Length + Type.risk + Type.fuel + Area + offset(log(Exposure))
| Driver.age + offset(log(Exposure))
, dist='poisson', link="logit", data = train.data)

library(glmmTMB)
zip_model <- glmmTMB(N.claims.year ~ Driver.age.cut + Vehicle.age.cut
+ Length + Type.risk + Type.fuel + Area + offset(log(Exposure))
, ziformula = ~ Driver.age + offset(log(Exposure))
, family = poisson, data = train.data)
```

```
library(mgcv)
zip_model2 <- gam(list(N.claims.year ~ Driver.age.cut +
Vehicle.age.cut + Length + Type.risk + Type.fuel + Area +
offset(log(Exposure)), ~ Driver.age + offset(log(Exposure)))
, data = train.data, family = ziplss, method = "REML")


# CROSS VALIDATION k=2

# SSE for Poisson
predictions <- list()
mse_errors <- numeric(12)

for (i in 13:24) {
  # Make predictions on test data
  predictions[i-12] <- predict(get(paste0("model", i))
  , newdata = test.data, type = "response")

  mse_errors[i-12] <- sum((residuals.glm(get(paste0("model", i))
  , type='response'))^2)}

nb_predictions <- list()
nb_sse_errors <- numeric(12)

for (i in 13:24) {
  # Make predictions on test data
  nb_predictions[i-12] <- predict(get(paste0("negbin", i))
  , newdata = test.data, type = "response")

  nb_sse_errors[i-12] <- sum((residuals.glm(get(paste0("negbin", i))
  , type='response'))^2)}

zip_predictions <- list()
zip_mse_errors <- numeric(12)
for (i in 13:24) {
  # Make predictions on test data
  zip_predictions[i-12] <- predict(get(paste0("zip", i))
  , newdata = test.data, type = "response")

  zip_mse_errors[i-12] <- sum((residuals.glm(get(paste0("zip", i))
  , type='response'))^2)}

# CROSS VALIDATION k=5
data_risks23 <- data[data$Type.risk =='2'|data$Type.risk == '3',]

library(boot)
cv_errors <- numeric(12)
cv_errors_cor <- numeric(12)
for (i in 13:24) {
```

```
  if (i != 14) {
    cv_errors[i-12] <- cv.glm(data_risks23, get(paste0("model", i))
    , K = 5)$delta[1]
    cv_errors_cor[i-12] <-
    cv.glm(data_risks23, get(paste0("model", i)), K = 5)$delta[2]}}

x <- cv.glm(data[!data$Drv.age %in%
c("93", "94", "95", "96", "97", "98", "99"), ]
, poisson24, K = 5)$delta[2]

library(boot)
nb_cv_errors <- numeric(12)
nb_cv_errors_cor <- numeric(12)
for (i in 13:24) {
  if (i != 14) {
    nb_cv_errors_cor[i-12] <-
    cv.glm(data_risks23, get(paste0("negbin", i)), K = 5)$delta[2]}}

y <- cv.glm(data_risks23[!data_risks23$Drv.age %in%
c("93", "94", "95", "96", "97", "98", "99"), ]
, negbin15, K = 5)$delta[2]
```

Βιβλιογραφικές αναφορές

1. P. McCullagh, J.A Nelder (1983), Generalized Linear Models, Chapman and Hall.
2. Piet De Jong and Gillian Z. Heller (2008), Generalized Linear Models for Insurance Data, CAMBRIDGE UNIVERSITY PRESS.
3. Roger J. Gray, Susan M Pitts (2012), Risk Modelling in General Insurance, CAMBRIDGE UNIVERSITY PRESS.
4. Edward W. Frees, Richard A. Derrig, Glenn Meyers (2014), Predictive modeling applications in Actuarial Science, *Volume 1: Predictive Modeling Techniques*, CAMBRIDGE UNIVERSITY PRESS.
5. Edward W. Frees, Richard A. Derrig, Glenn Meyers (2014), Predictive modeling applications in Actuarial Science, *Volume 2: Case studies in Insurance*, CAMBRIDGE UNIVERSITY PRESS.
6. Athur Charpentier (2015), Computational Actuarial Science with R, *The R Series*, Chapman and Hall.
7. Wagner Hugo Bonat, Walmes Marques Zeviani, Eduardo Elias Ribeiro Jr (2017), Regression Models for Count Data: beyond Poisson model, XV EMR - Brazilian Regression Model School.
8. Simon N. Wood (2017), Generalized Additive Models: an introduction with R, Chapman and Hall.
9. O. Barndorff Nielsen (1977), Information and Exponential Families, John Wiley and Sons.
10. A. C. Atkinson (1985), Plots Transformations and Regression, Oxford Science Publications.
11. Χ. Καρώνη, Π. Οικονόμου (2020), Στατιστικά Μοντέλα Παλινδρόμισης, Εκδόσεις ΣΥΜΕΩΝ.