ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών

Γραμμικά Μοντέλα

Σημειώσεις Εργαστηρίων

Διδάσκουσα: Λουκία Μελιγκοτσίδου

Τμήμα Μαθηματικών

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Contents

1 Εισαγωγή

Οι σημειώσεις αυτές αποτελούν βοηθητικό υλικό για τα εργαστήρια του μαθήματος "Γραμμικά Μοντέλα". Περιλαμβάνουν τη στατιστική ανάλυση του σετ δεδομένων **cars.sav** σε SPSS και σε R. Παράλληλα, οι φοιτητές μπορούν να εξασκούνται στο αντίστοιχο σετ δεδομένων **grades.sav**.

Cars

Αυτό το σετ δεδομένων αφορά χαρακτηριστικά από 391 αυτοκίνητα. Συγκεκριμένα περιέχει τις μεταβλητές:

- Price: Η τιμή του αυτοκινήτου σε ευρώ.
- Engine: Το μέγεθος του κινητήρα σε cc (κυβικά εκατοστά).
- Horsepower: Η ιπποδύναμη του αυτοκινήτου σε ίππους.
- Weight: Η μάζα του αυτοκινήτου σε κιλά.
- Acceleration: Ο χρόνος που απαιτείται για να φτάσει το αυτοκίνητο ταχύτητα 100Km/hr, ξεκινώντας από την ηρεμία.
- Year: Το έτος κυκλοφορίας του αυτοκινήτου.
- Fuel: Η κατανάλωση καυσίμου του οχήματος, κωδικοποιημένη ως 1: Χαμηλή, 2: Μέτρια, 3: Υψηλή και 4: Πολύ υψηλή.
- Brand: Η μάρκα του αυτοκινήτου, κωδικοποιημένη ως 1: BMW, 2: Datsun και 3: Mercedes-Benz.
- Type: Ο τύπος του κινητήρα, κωδικοποιημένος ως 0: std, 1: turbo

Grades

Αυτό το σετ δεδομένων αφορά βαθμούς μαθητών δύο τάξεων γυμνασίου. Συγκεκριμένα περιέχει τις μεταβλητές:

- Grades: Ο γενικός μέσος όρος του μαθητή, στην κλίμακα του 100.
- Maths: Ο βαθμός του μαθητή στα μαθηματικά, στην κλίμακα του 100.
- Physics: Ο βαθμός του μαθητή στη Φυσική, στην κλίμακα του 100.

- Gymnastics: Ο βαθμός του μαθητή στη Γυμναστική, στην κλίμακα του 100.
- Hours: Οι ώρες καθημερινού διαβάσματος του μαθητή.
- GoingOut: Η συχνότητα εξόδων του μαθητή, κωδικοποιημένη ως 0: Καθόλου, 1: Μία φορά το μήνα, 2: 2-3 φορές το μήνα, 3: Μία φορά τη βδομάδα, 4: 2-3 φορές τη βδομάδα, 5: Κάθε μέρα.
- Class: Το τμήμα του μαθητή, κωδικοποιημένο ως 1 ή 2.
- Sex: Το φύλο του μαθητή, κωδικοποιημένο ως 0: κορίτσι, 1: αγόρι.

2 Περιγραφική Στατιστική

Σύνοψη: Το πρώτο βήμα στην ανάλυση δεδομένων είναι η παρουσίαση και σύνοψη των πληροφοριών του δείγματος για τις μεταβλητές που περιλαμβάνονται σε αυτό. Στις ενότητες που ακολουθούν παραθέτουμε τη διαδικασία για τη συνοπτική παρουσίαση ποιοτικών και ποσοτικών δεδομένων.

Είδη Μεταβλητών

Οι μεταβλητές διακρίνονται σε ποσοτικές (scale) και ποιοτικές/κατηγορικές (categorical), οι οποίες με τη σειρά τους διακρίνονται σε διατάξιμες (ordinal) και ονοματικές (nominal). Ποσοτικές ονομάζονται οι μεταβλητές που μπορούν να μετρηθούν (έχουν δηλαδή αριθμητικές τιμές). Παραδείγματα ποσοτικών μεταβλητών είναι η ηλικία, το βάρος, το ύψος, η αξία μίας μετοχής κ.ά. Ποιοτικές ονομάζονται οι μεταβλητές που περιγράφουν χαρακτηριστικά του πληθυσμού που μεταβάλλονται κατά ποιότητα ή είδος. Τέτοιες μεταβλητές είναι το φύλο, το χρώμα των ματιών, η τάξη ενός μαθητή κ.ο.κ. Οι ποιοτικές μεταβλητές που έχουν τη δυνατότητα διάταξης ονομάζονται διατάξιμες (π.χ. η τάξη ενός μαθητή) ενώ οι άλλες ονοματικές (π.χ. το χρώμα των ματιών).

Παράδειγμα: Παρατηρούμε ότι στο αρχείο cars έχουμε 5 ποσοτικές μεταβλητές (Price, Engine, Horsepower, Weight, Acceleration) και 4 κατηγορικές (Year, Fuel, Brand, Type), από τις οποίες 2 είναι διατάξιμες (Year, Fuel) και 2 ονοματικές (Brand, Type).

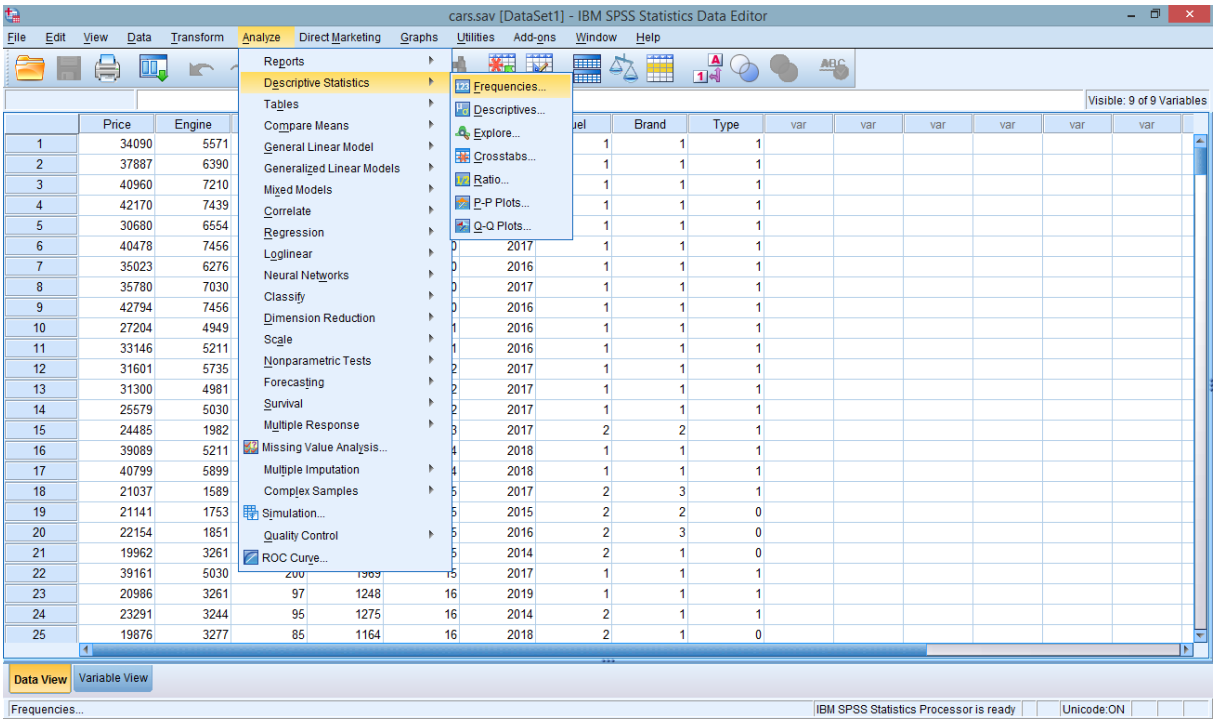
Ποιοτικές Μεταβλητές

Η συνοπτική παρουσίαση των δεδομένων μίας ποιοτικής μεταβλητής με τον πίνακα συχνοτήτων και με τις γραφικές τους παραστάσεις, το ραβδόγραμμα (bar chart) και το κυκλικό διάγραμμα (pie chart). Ο πίνακας συχνοτήτων μιας ποιοτικής μεταβλητής προκύπτει από την απαρίθμηση και καταγραφή των δειγματικών τιμών στην αντίστοιχη κατηγορία. Ένας ολοκληρωμένος πίνακας συχνοτήτων μίας ποιοτικής μεταβλητής περιλαμβάνει τη στήλη των συχνοτήτων και τη στήλη των σχετικών συχνοτήτων. Επιπλέον, αν η μεταβλητή είναι διατάξιμη, μπορεί να συμπεριληφθεί η στήλη των αθροιστικών συχνοτήτων και των αθροιστικών σχετικών συχνοτήτων (ή των αθροιστικών ποσοστών).

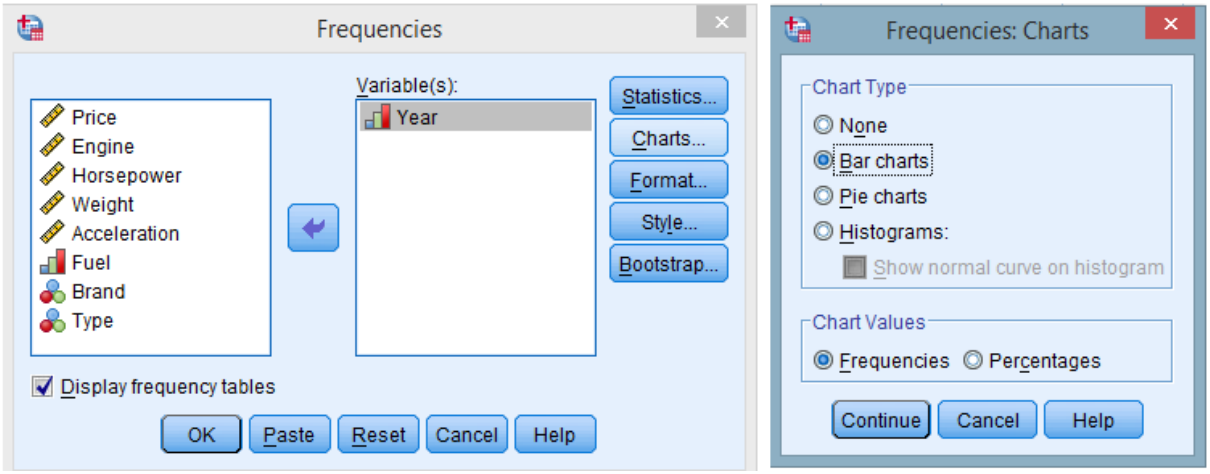
Ένας τρόπος άμεσης κατανόησης των χαρακτηριστικών της κατανομής των συχνοτήτων επιτυγχάνεται με μία ειδική γραφική παράσταση που ονομάζεται ραβδόγραμμα. Στον οριζόντιο άξονα ενός ραβδογράμματος συχνοτήτων σημειώνονται οι κατηγορίες στις οποίες τα μέλη του πληθυσμού κατατάσσονται, ενώ στον κατακόρυφο άξονα οι αντίστοιχες συχνότητες (εναλλακτικά οι αντίστοιχες σχετικές συχνότητες).

Παράδειγμα: Θα δημιουργήσουμε ένα πίνακα συχνοτήτων και ένα ραβδόγραμμα για την (διατάξιμη) μεταβλητή Year.

Στη γραμμή εντολών ακολουθούμε τη διαδρομή Analyze → Descriptive Statistics → Frequencies.



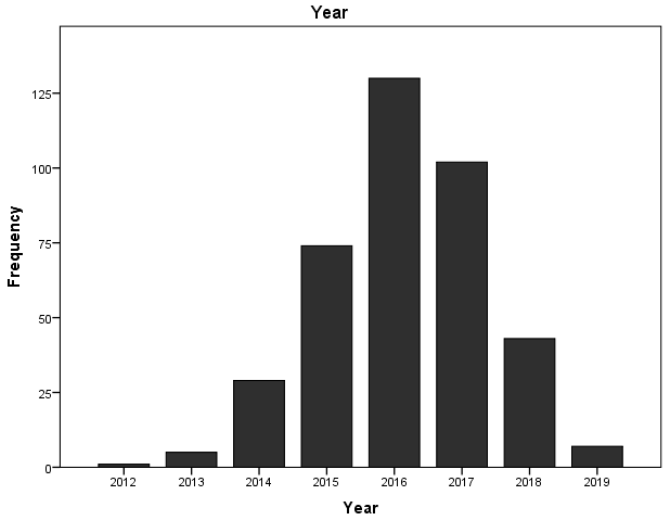
Μετακινούμε δεξιά την μεταβλητή Year, επιλέγουμε το Display Frequency Tables και στο μενού Charts επιλέγουμε το Bar charts για να δημιουργήσουμε το ραβδόγραμμα.



Το SPSS μας δίνει τον πίνακα συχνοτήτων (μαζί με τα αθροιστικά ποσοστά, αφού η Year είναι διατάξιμη μεταβλητή) και το ραβδόγραμμα. Κάνοντας κλικ πάνω στο γράφημα, ένα νέο παράθυρο ανοίγει και μας δίνεται η επιλογή να αλλάξουμε την εμφάνισή του γραφήματος (χρώμα, μέγεθος κ.ά.).

Year

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2012	1	,3	,3	,3
	2013	5	1,3	1,3	1,5
	2014	29	7,4	7,4	9,0
	2015	74	18,9	18,9	27,9
	2016	130	33,2	33,2	61,1
	2017	102	26,1	26,1	87,2
	2018	43	11,0	11,0	98,2
	2019	7	1,8	1,8	100,0
	Total	391	100,0	100,0	



Ποσοτικές Μεταβλητές

Η συνοπτική παρουσίαση των δεδομένων ποσοτικών μεταβλητών γίνεται με τον υπολογισμό των τιμών διάφορων περιγραφικών μέτρων, όπως η μέση τιμή (mean), η τυπική απόκλιση (std deviation) κ.ά., και την κατασκευή του ιστογράμματος (histogram) και του θηκογράμματος (boxplot) της ποσοτικής μεταβλητής. Επιπλέον, γίνεται έλεγχος αν οι διαθέσιμες δειγματικές τιμές περιγράφονται ικανοποιητικά από την κανονική κατανομή.

Τα περιγραφικά μέτρα χωρίζονται σε μέτρα θέσης (που μας δείχνουν πού συγκεντρώνονται οι τιμές) και μέτρα διασποράς (που μας δείχνουν πόσο εξαπλώνονται οι τιμές). Τα σημαντικότερα μέτρα θέσης είναι η μέση τιμή (mean), η διάμεσος (median) και η επικρατούσα τιμή (mode). Σημαντικά μέτρα διασποράς είναι η τυπική απόκλιση (std deviation), το εύρος (range), η ελάχιστη (min) και μέγιστη (max) τιμή και τα ποσοστιαία σημεία (percentile values).

Είναι σημαντικό να μην υπάρχει σύγχυση ανάμεσα στη μέση τιμή (τον μέσο όρο των μετρήσεων), τη διάμεσο (την τιμή που χωρίζει τα δεδομένα σε δύο ίσα μέρη έτσι ώστε το πλήθος των μετρήσεων που βρίσκονται αριστερά της να είναι ίσο με το πλήθος των μετρήσεων που βρίσκεται δεξιά της) και την επικρατούσα τιμή (την τιμή με τη μεγαλύτερη συχνότητα).

Άλλα περιγραφικά μέτρα είναι ο συντελεστής ασυμμετρίας (skewness) και ο συντελεστής κύρτωσης (kyrtosis), που περιγράφουν την ασυμμετρία και την κύρτωση της κατανομής αντίστοιχα. Με βάση την ασυμμετρία, οι κατανομές διακρίνονται σε συμμετρικές όταν $l = 0$ (σε αυτές ανήκει η κανονική κατανομή), σε θετικά ασύμμετρες (ή λοξές δεξιά) όταν $l > 0$ και σε αρνητικά ασύμμετρες (ή λοξές αριστερά) όταν $l < 0$. Με βάση την κύρτωση οι κατανομές διακρίνονται σε λεπτόκυρτες όταν $k > 3$, σε μεσόκυρτες όταν $k = 3$ (σε αυτές ανήκει η κανονική κατανομή) και σε πλατύκυρτες όταν $k < 3$. Το S.P.S.S. υπολογίζει ως κύρτωση την τιμή $k - 3$, έτσι ώστε ο έλεγχος της κανονικότητας των δεδομένων να γίνεται με το μηδέν.

Τέλος, τα τεταρτημόρια έχουν την ιδιότητα να χωρίζουν το σύνολο των μετρήσεων σε τέσσερα ίσα μέρη (0%, 25%, 50%, 75%, 100%).

Για την γραφική αναπαράσταση των ποσοτικών μεταβλητών, χρησιμοποιούμε το ιστόγραμμα συχνοτήτων και το θηκόγραμμα. Το ιστόγραμμα παρουσιάζει ομαδοποιημένα τα δεδομένα με τη μορφή συνεχόμενων

ορθογωνίων, τα οποία έχουν ύψος ανάλογο με τη συχνότητα κάθε ομάδας. Οι τιμές της μεταβλητής τοποθετούνται στον οριζόντιο άξονα, ενώ οι συχνότητες στον κατακόρυφο άξονα.

Το θηκόγραμμα παριστάνει τα σημαντικότερα ποσοστημόρια των δεδομένων και έχει τη μορφή κουτιού με κάτω και άνω άκρα τα 25% και 75% ποσοστημόρια αντίστοιχα. Επιπλέον, διαθέτει δύο προεκτάσεις (whiskers) οι οποίες βοηθούν στον εντοπισμό ακραίων παρατηρήσεων. Η κάτω γραμμή φτάνει μέχρι το (1ο τεταρτημόριο – 1.5 * ενδοτεταρτημοριακό εύρος) ενώ η πάνω μέχρι (3ο τεταρτημόριο + 1.5 * ενδοτεταρτημοριακό εύρος).

Έλεγχος Κανονικότητας

Το μεγαλύτερο μέρος της Στατιστικής Συμπερασματολογίας, προϋποθέτει ότι τα δεδομένα προέρχονται από έναν πληθυσμό ο οποίος περιγράφεται ικανοποιητικά από την κανονική κατανομή. Η κανονικότητα ή μη των δεδομένων θα κρίνει τα μέτρα θέσης και διασποράς που τα περιγράφουν ορθότερα. Τα κανονικά δεδομένα περιγράφονται καλύτερα από την μέση τιμή και την τυπική απόκλιση, ενώ τα μη-κανονικά δεδομένα από τη διάμεσο και τα τεταρτημόρια. Ο έλεγχος της κανονικότητας γίνεται τόσο στατιστικά όσο και γραφικά.

Για τον στατιστικό έλεγχο κανονικότητας χρησιμοποιούμε το τεστ Shapiro-Wilk (ή το τεστ Kolmogorov-Smirnov για μικρά δείγματα, με $n < 30$). Οι έλεγχοι κανονικότητας έχουν τη μορφή:

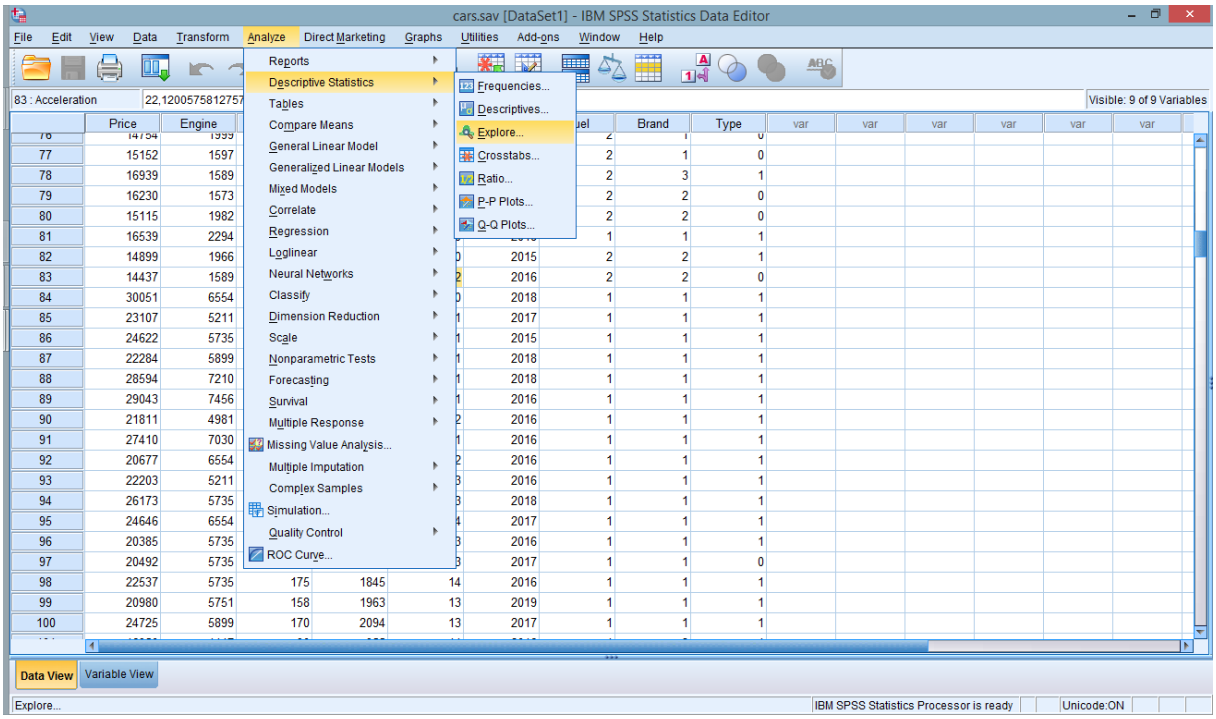
$$H_0 : \text{ τα δεδομένα προέρχονται από κανονική κατανομή}$$
$$H_1 : \text{ η κανονικότητα απορρίπτεται}$$

Επομένως, αν ο έλεγχος δώσει αρκετά μικρό p-value, η υπόθεση της κανονικότητας απορρίπτεται.

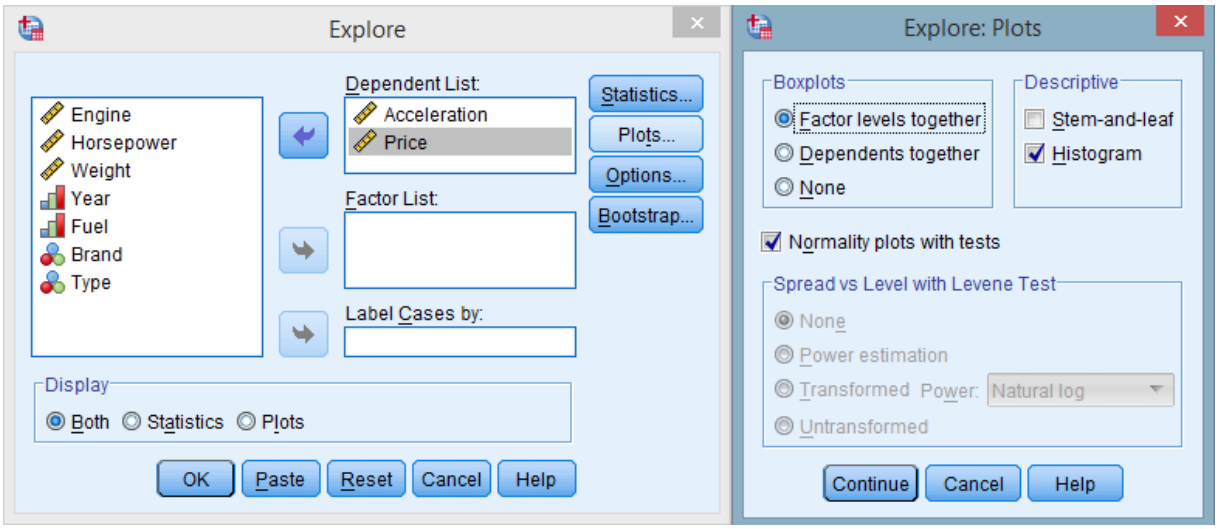
Γραφικά, η κανονικότητα ελέγχεται με το Q-Q (quantile-quantile) γράφημα, το οποίο συγκρίνει τα ποσοστιαία σημεία (quantile) του δείγματος έναντι των πληθυσμιακών ποσοστιαίων σημείων της κανονικής κατανομής. Παρεκκλίσεις από την ευθεία δηλώνουν μη κανονικότητα.

Παράδειγμα: Θα ελέγξουμε την κανονικότητα των (ποσοτικών) μεταβλητών Acceleration και Price.

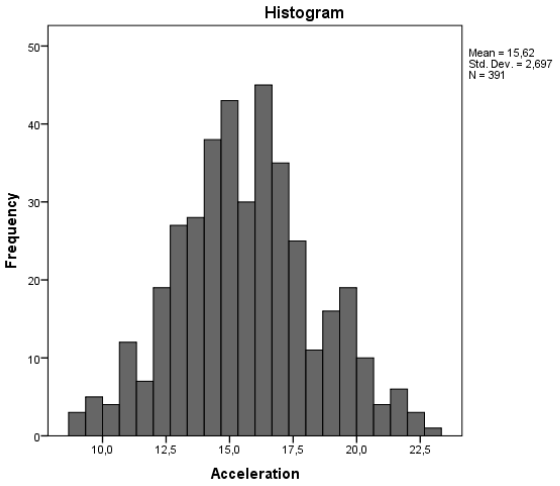
Στη γραμμή εντολών ακολουθούμε τη διαδρομή Analyze → Descriptive Statistics → Explore.

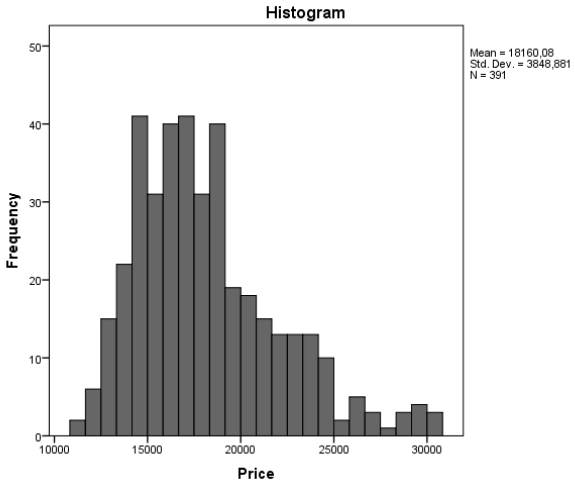


Μετακινούμε δεξιά τις μεταβλητές Acceleration και Price, και στο μενού Charts επιλέγουμε Histogram και Normality plots with tests για να εκτελέσουμε τον έλεγχο κανονικότητας (ο οποίος θα παράγει αυτόματα το boxplot και το Q-Q Plot).

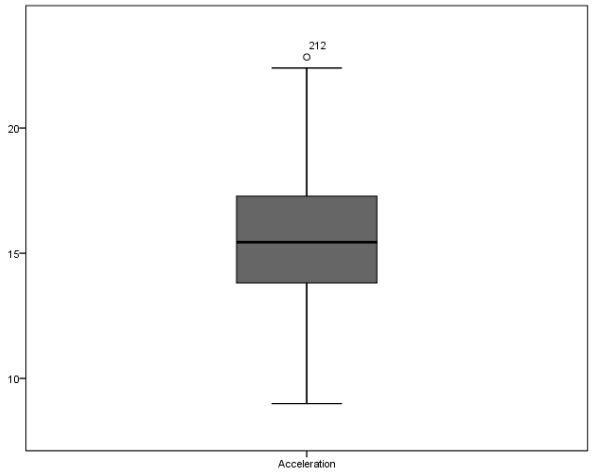


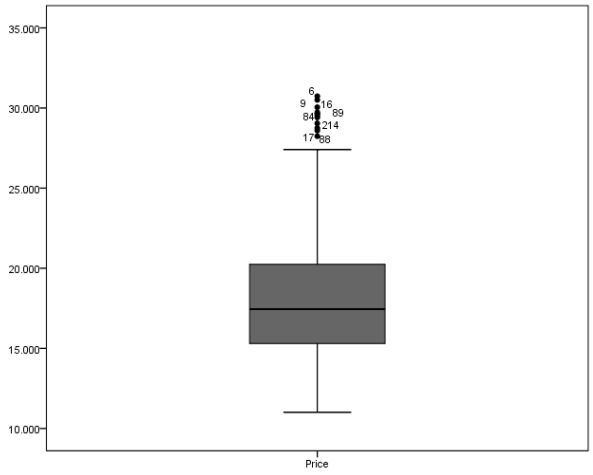
Ξεκινάμε την περιγραφική ανάλυση με τα ιστογράμματα των δύο μεταβλητών. Η Acceleration είναι συμμετρική, αλλά παρουσιάζει μία απότομη πτώση σε κεντρικές τιμές. Η Price εμφανίζει έντονη δεξιά ασυμμετρία (οι τιμές συγκεντρώνονται στα αριστερά).





Συνεχίζουμε με τα boxplot, τα οποία βοηθούν στον εντοπισμό ακραίων τιμών (outliers). Εδώ η Price εμφανίζει μεγάλο πλήθος ακραίων παρατηρήσεων (οι οποίες αντιστοιχούν στις τιμές του ιστογράμματος που προκαλούν την ασυμμετρία, κοντά στο 30000). Αντίθετα, η Acceleration εμφανίζει μόλις μία ακραία παρατήρηση.





Στα γραφήματα Q-Q Plot μπορούμε να δούμε πόσο καλά εφαρμόζουν τα σημεία της μεταβλητής Acceleration πάνω στα ποσοστημόρια της κανονικής κατανομής, σε αντίθεση με αυτά της Price που παρουσιάζουν σοβαρές αποκλίσεις. Υποψιαζόμαστε ότι ο έλεγχος θα απορρίψει την υπόθεση της κανονικότητας για την Price.

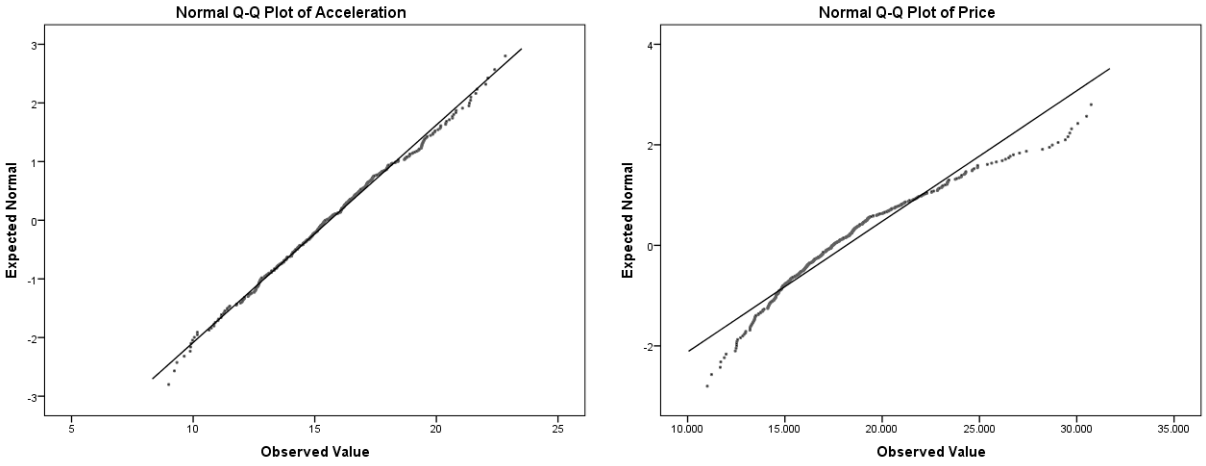
Έχουμε ένα δείγμα 391 αυτοκινήτων, επομένως κοιτάμε τον έλεγχο Shapiro-Wilk. Για την μεταβλητή Acceleration έχουμε p-value$= 0.259 (> 0.05)$ επομένως δεν απορρίπτουμε την (H_0) υπόθεση της κανονικότητας σε επίπεδο σημαντικότητας 5%. Αντίθετα για την μεταβλητή Price έχουμε p-value< 0.05 (η τιμή δεν είναι πραγματικά 0. Κάνοντας κλικ πάνω στον έλεγχο βλέπουμε ότι p-value$= 6,2 \cdot 10^{-11}$). Επομένως απορρίπτουμε την υπόθεση της κανονικότητας για την μεταβλητή Price.

Tests of Normality

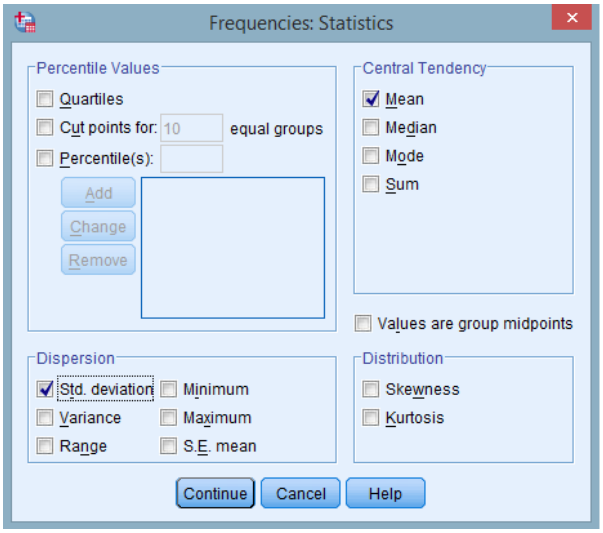
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Acceleration	,028	391	,200*	,995	391	,259
Price	,094	391	,000	,944	391	,000

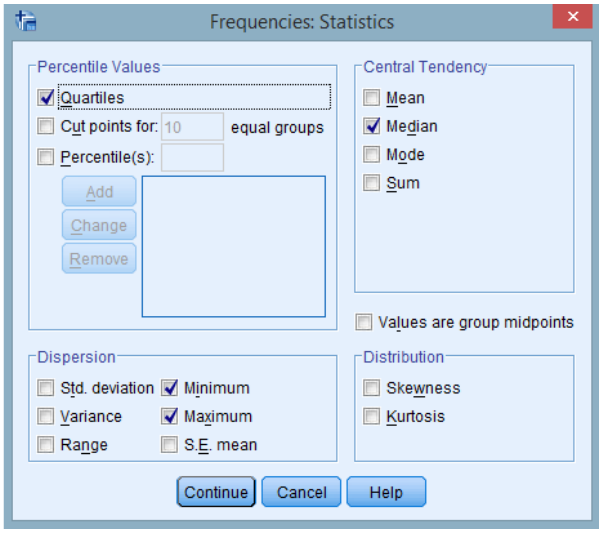
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



Επομένως, τα σωστά περιγραφικά μέτρα για την Acceleration είναι η μέση τιμή και η τυπική απόκλιση, ενώ για την Price η διάμεσος, το ελάχιστο, το μέγιστο και τα τεταρτημόρια. Για να τα επιλέξουμε, στη γραμμή εντολών ακολουθούμε τη διαδρομή Analyze $\rightarrow$ Descriptive Statistics $\rightarrow$ Frequencies. Επιλέγουμε τα επιθυμητά μέτρα για κάθε μεταβλητή (φυσικά, μπορούμε να ζητήσουμε περισσότερα, όπως η κύρτωση και η ασυμμετρία).





Τελικά, τα περιγραφικά μέτρα των δύο μεταβλητών εμφανίζονται στους παρακάτω πίνακες.

Acceleration

Μέτρο	Τιμή
Μέση Τιμή	15.62
Τυπική Απόκλιση	2.70

Price

Μέτρο	Τιμή
Ελάχιστο (0%)	11013
1ο Τεταρτημόριο (25%)	15282.88
Διάμεσος (50%)	17444.97
3ο Τεταρτημόριο (75%)	20243.15
Μέγιστο (100%)	30738

Ασκήσεις

Επαναλάβετε τις παραπάνω περιγραφικές αναλύσεις για τα δεδομένα του αρχείου grades.

1. Αξιολογήστε τις μεταβλητές ως ποιοτικές/ποσοτικές (και τις ποιοτικές σε διατάξιμες/ονοματικές).
2. Δημιουργήστε ένα πίνακα συχνοτήτων και ένα ραβδόγραμμα για τις ποιοτικές μεταβλητές Hours και Class.
3. Εφαρμόστε έλεγχο κανονικότητας για τις ποσοτικές μεταβλητές Grades και Physics. Τι εικόνα δίνουν τα γραφήματα; Γιατί δεν απορρίπτεται η υπόθεση της κανονικότητας;

3 Συσχέτιση

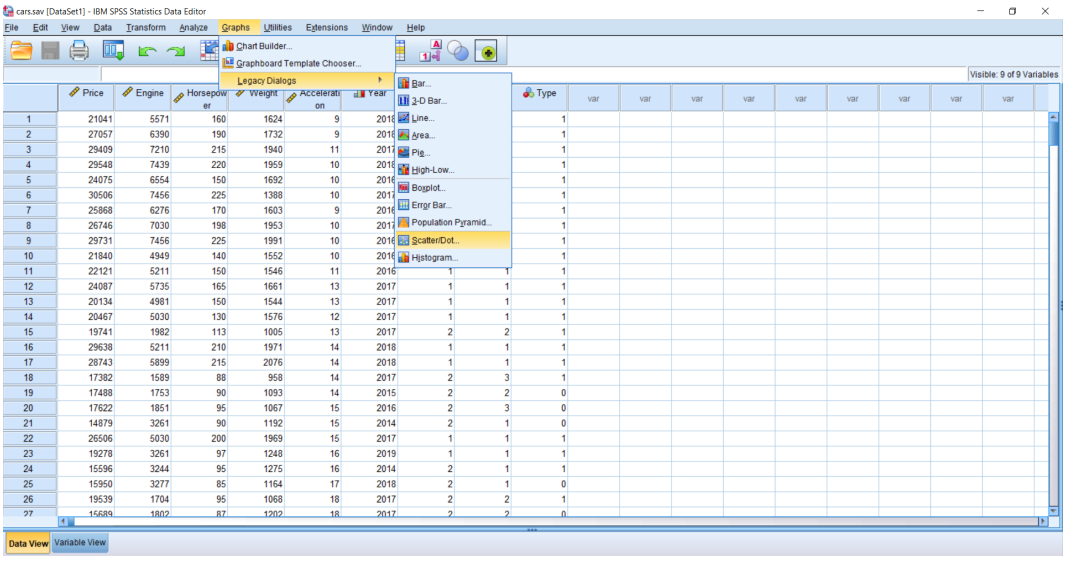
Σύνοψη: Το επόμενο βήμα στην ανάλυση των δεδομένων είναι να εξεταστεί κατά πόσο υπάρχει συσχέτιση μεταξύ των μεταβλητών. Αυτό επιτυγχάνεται τόσο γραφικά, με το διάγραμμα διασποράς, όσο και υπολογιστικά με το συντελεστή γραμμικής συσχέτισης Pearson.

Διάγραμμα διασποράς

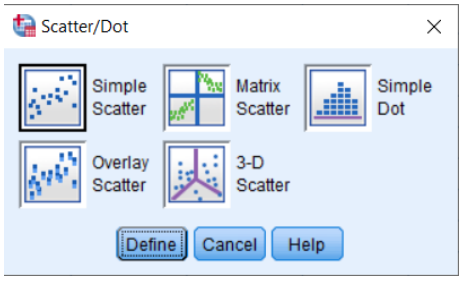
Σε πολλές περιπτώσεις της ανάλυσης δεδομένων το ενδιαφέρον εστιάζεται στη μελέτη δύο ή περισσότερων χαρακτηριστικών του δείγματος. Για το λόγο αυτό, είναι λογικό να αναζητήσουμε μέτρα τα οποία μπορούν να εκφράσουν και να ποσοτικοποιήσουν την πιθανή συμμεταβολή-συσχέτιση των χαρακτηριστικών αυτών. Καθώς η οπτικοποίηση των δεδομένων αποτελεί πρωταρχικό στάδιο της ανάλυσης, το διάγραμμα το οποίο απεικονίζει τη συσχέτιση ματαξύ δύο μεταβλητών είναι το **διάγραμμα διασποράς (Scatter Diagram) ή σημειόγραμμα**

Η εφαρμογή θα γίνει στο σετ δεδομένων cars. Πιο συγκεκριμένα, θα εξετάσουμε γραφικά τη συσχέτιση μεταξύ των μεταβλητών Acceleration και Horsepower. Κατά πόσο, δηλαδή, ο χρόνος που απαιτείται για να φτάσει το αυτοκίνητο ταχύτητα 100Km/h ξεκινώντας από την ηρεμία, εξαρτάται από την ιπποδύναμή του.

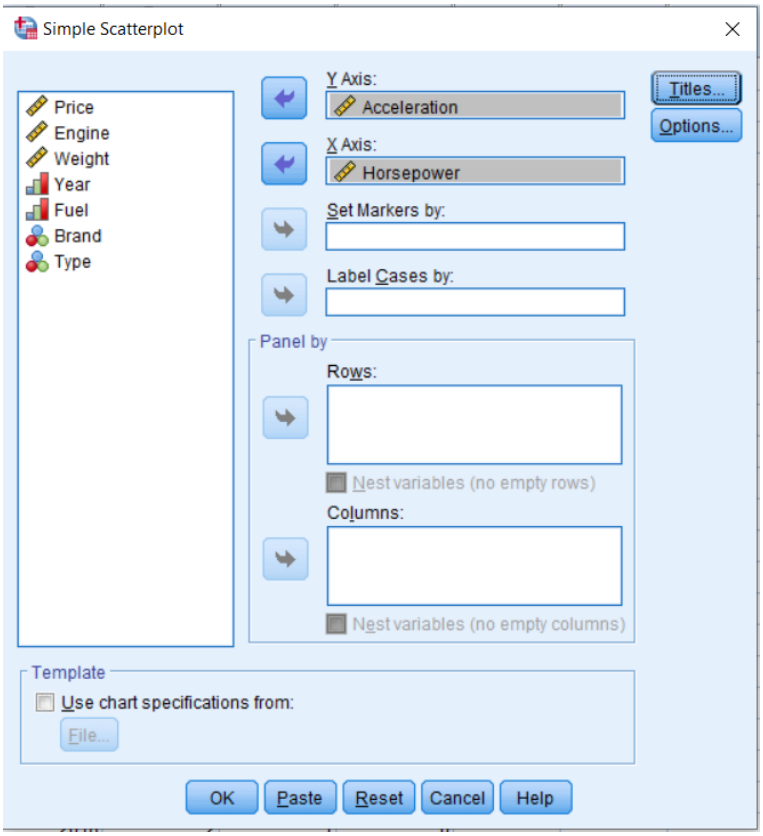
Στη γραμμή εντολών ακολουθούμε τη διαδρομή Graphs → Legacy Dialogs → Scatter/Dot.



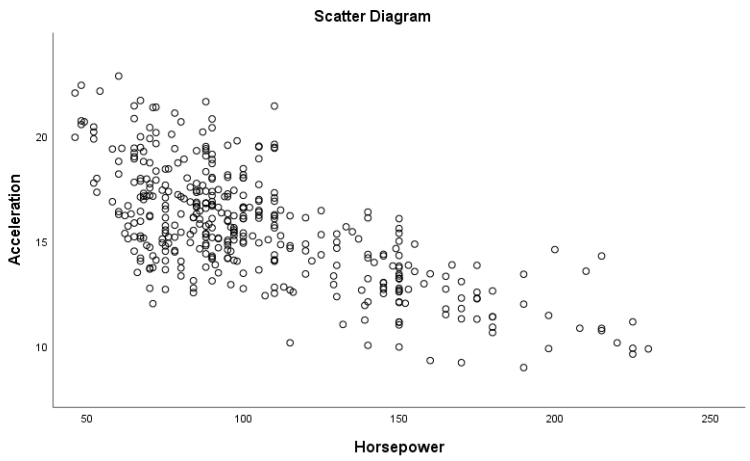
Έπειτα, επιλέγουμε το Simple Scatter. Στο συγκεκριμένο παράδειγμα επιλέγουμε να εξετάσουμε γραφικά τη συσχέτιση μεταξύ δύο μεταβλητών. Αν θέλαμε να εξετάσουμε τη συσχέτιση παραπάνω από δύο ποσοτικών μεταβλητών, θα επιλέγαμε το Matrix Scatter.



Στο αναδυόμενο παράθυρο, μετακινούμε στο κελί Y axis τη μεταβλητή Acceleration και στο κελί X axis τη μεταβλητή Horsepower, όπως φαίνεται παρακάτω.



Το διάγραμμα διασποράς φανερώνει μία γενικά αρνητική συσχέτιση μεταξύ των δύο μεταβλητών, καθώς όσο αυξάνεται η ιπποδύναμη του αυτοκινήτου τόσο ελαττώνεται ο χρόνος της επιτάχυνσής του.



Συντελεστής γραμμικής συσχέτισης Pearson

Από το διάγραμμα διασποράς φαίνεται ότι οι δύο μεταβλητές είναι αρνητικά συσχετισμένες, αλλά πόσο ισχυρή είναι αυτή η συσχέτιση; Δεδομένου του γεγονότος ότι και οι δύο μεταβλητές είναι ποσοτικές, χρησιμοποιείται ο δειγματικός συντελεστής συσχέτισης του Pearson

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

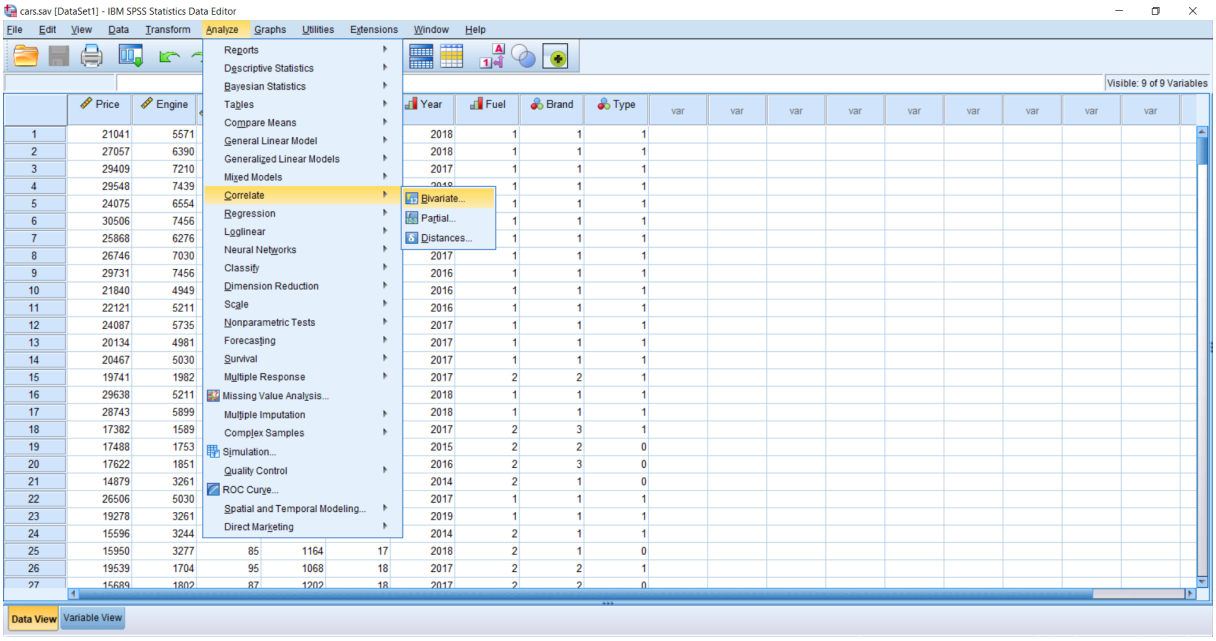
Ο συντελεστής r δίνει ένα μέτρο του μεγέθους της γραμμικής συσχέτισης μεταξύ δύο μεταβλητών. Παίρνει τιμές στο $[-1, 1]$ και το πρόσημό του υποδηλώνει αν η συσχέτιση είναι θετική ή αρνητική. Πιο συγκεκριμένα, αν:

- $r = \pm 1$ τότε η γραμμική συσχέτιση είναι τέλεια.
- $-0.3 \leq 0.3$ δεν υπάρχει γραμμική συσχέτιση, χωρίς αυτό να σημαίνει ότι δεν υπάρχει κάποιου άλλου είδους συσχέτιση (όπως καμπυλόγραμμη συσχέτιση).
- $-0.5 < r \leq -0.3$ ή $0.3 \leq r < 0.5$ υπάρχει ασθενής γραμμική συσχέτιση.
- $-0.7 < r \leq -0.5$ ή $0.5 \leq r < 0.7$ υπάρχει μέση γραμμική συσχέτιση.
- $-0.8 < r \leq -0.7$ ή $0.7 \leq r < 0.8$ υπάρχει ισχυρή γραμμική συσχέτιση.
- $-1 < r \leq -0.8$ ή $0.8 \leq r < 1$ υπάρχει πολύ ισχυρή γραμμική συσχέτιση.

Ο συντελεστής Pearson χρησιμοποιείται κυρίως για ποσοτικές μεταβλητές και έχει πιο ακριβή εφαρμογή, όταν αυτές είναι κανονικά κατανεμημένες.

Η εφαρμογή στο SPSS γίνεται μεταξύ των μεταβλητών Acceleration και Horsepower.

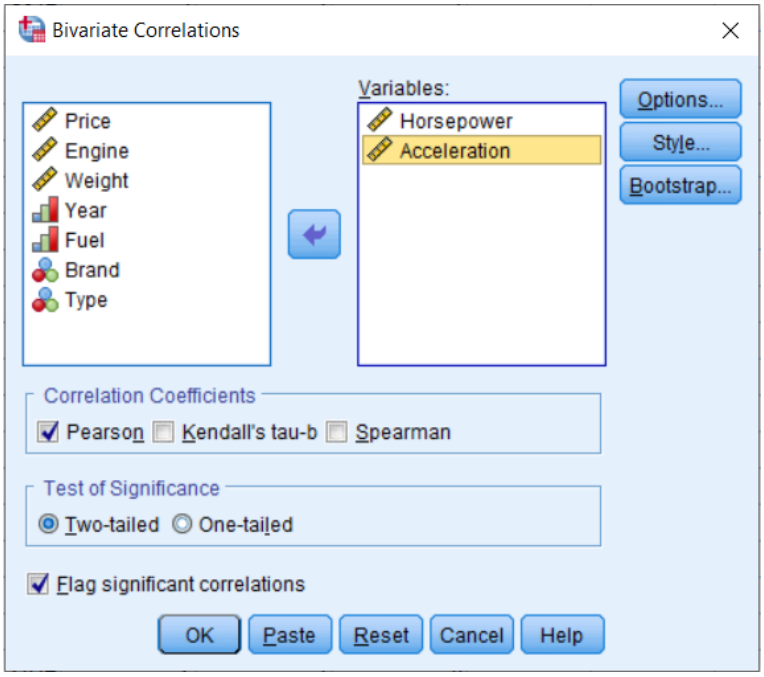
Στη γραμμή εντολών ακολουθούμε τη διαδρομή Analyze → Correlate → Bivariate.



Στο παράθυρο που εμφανίζεται παρακάτω μετακινούμε δεξιά τις μεταβλητές Acceleration και Horsepower, επιλέγουμε τον συντελεστή Pearson και τον αμφίπλευρο έλεγχο (two-tailed) στατιστικής σημντικότητας. Ο έλεγχος αυτός εξετάζει τη μηδενική υπόθεση ότι οι δύο μεταβλητές είναι ασυσχέτιστες έναντι της εναλλακτικής ότι υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ τους.

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$



Ο πίνακας συσχέτισης που εξάγει το SPSS δείχνει μία μέση αρνητική γραμμική συσχέτιση μεταξύ των δύο μεταβλητών, καθώς ο συντελεστής Pearson είναι -0.684. Η συσχέτιση είναι στατιστικά σημαντική σε επίπεδο 1%, όπως φαίνεται και από την τιμή του p-value: Sig.$<< 0.01$. Αυτό σημαίνει, ότι απορρίπτουμε την αρχική υπόθεση, ότι οι δύο μεταβλητές είναι ασυσχέτιστες.

Correlations

		Acceleration	Horsepower
Acceleration	Pearson Correlation	1	-,684**
	Sig. (2-tailed)		,000
	N	391	391
Horsepower	Pearson Correlation	-,684**	1
	Sig. (2-tailed)	,000	
	N	391	391

**. Correlation is significant at the 0.01 level (2-tailed).

4 Απλή Γραμμική Παλινδρόμηση

Σύνοψη: Αφού εξετάστηκε η συσχέτιση μεταξύ των μεταβλητών, ακολουθεί η εφαρμογή και ανάλυση της απλής γραμμικής παλινδρόμησης.

Εισαγωγή

Ο συντελεστής συσχέτισης μας δείχνει αν και κατά πόσο δύο μεταβλητές σχετίζονται, χωρίς όμως να μας παρέχει πληροφορίες για τον τρόπο με τον οποίο μεταβάλλονται οι τιμές της μίας σε σχέση με της άλλης. Ο στόχος της απλής γραμμικής παλινδρόμησης είναι να περιγράψει με μία μαθηματική σχέση τις παρατηρήσεις των δύο μεταβλητών.

Από το διάγραμμα διασποράς, προκύπτει ότι δεν υπάρχει κάποια καμπύλη να ενώνει όλα τα σημεία και για το λόγο αυτό αναζητείται εκείνη η καμπύλη που προσαρμόζεται καλύτερα στα δεδομένα. Η απλούστερη μορφή συσχέτισης δύο μεταβλητών είναι η γραμμική, η οποία εκφράζεται μαθηματικά από τη σχέση:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, .., n$$

Όπου:

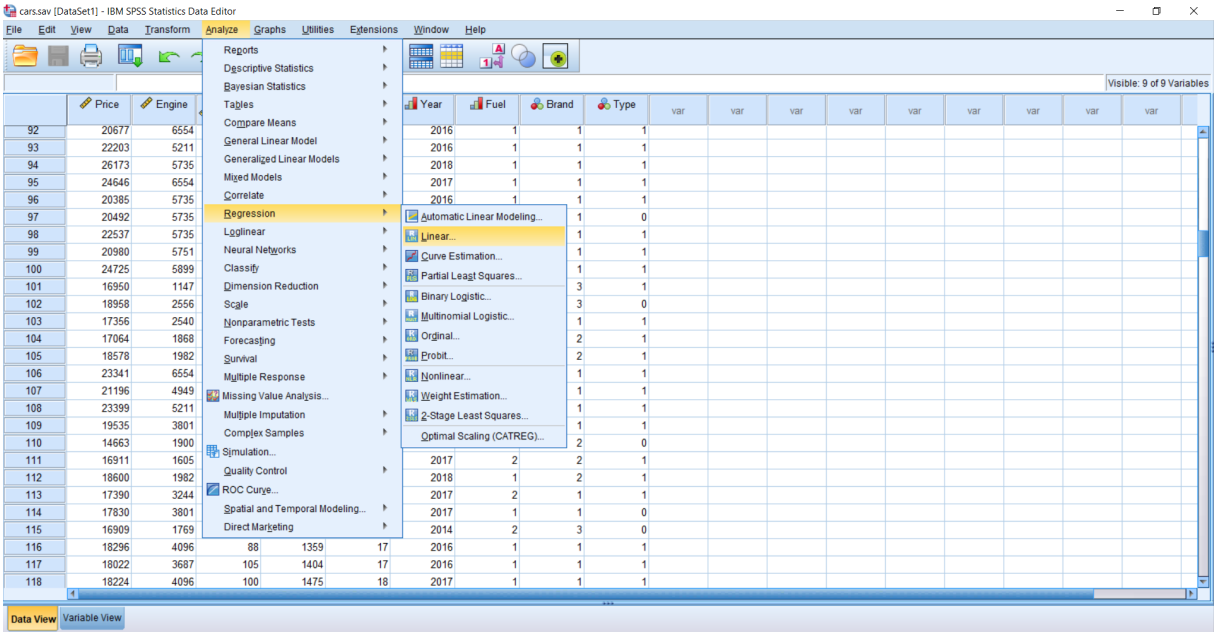
- Y είναι η εξαρτημένη μεταβλητή,
- X είναι η ανεξάρτητη μεταβλητή,
- β_0 είναι η τιμή της εξαρτημένης μεταβλητής για $X = 0$, χωρίς αυτή να έχει πάντα φυσική σημασία στην ανάλυση,
- β_1 είναι η κλίση της ευθείας της γραμμικής παλινδρόμησης και εκφράζει τη μεταβολή της Y για μοναδιαία μεταβολή της X και
- ϵ_i τα τυχαία σφάλματα, τα οποία στο απλό γραμμικό μοντέλο υποτίθενται ομοσκεδαστικά με $\epsilon_i \sim N(0, \sigma^2)$

Για να βρεθεί η ευθεία η οποία προσαρμόζεται καλύτερα στα δεδομένα πρέπει να εκτιμηθούν οι σταθερές β_0 και β_1. Οι εκτιμημένες παράμετροι υπολογίζονται μέσω της μεθόδου ελαχίστων τετραγώνων (Μ.Ε.Τ.) των σφαλμάτων, όπως έχει ειπωθεί αναλυτικά και στη θεωρία.

Εφαρμογή του μοντέλου

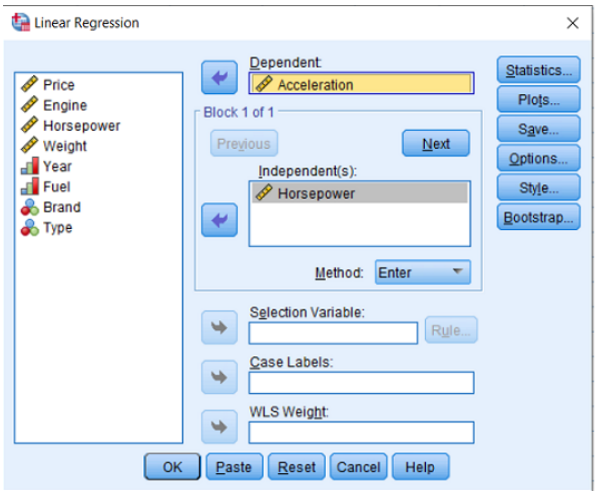
Η εφαρμογή της απλής γραμμικής παλινδρόμησης πραγματοποιείται στο σετ δεδομένων cars, ως συνέχεια της προηγούμενης ενότητας. Θα θεωρήσουμε ως εξαρτημένη μεταβλητή (Y) την Acceleration και ως ανεξάρτητη (X) την Horsepower.

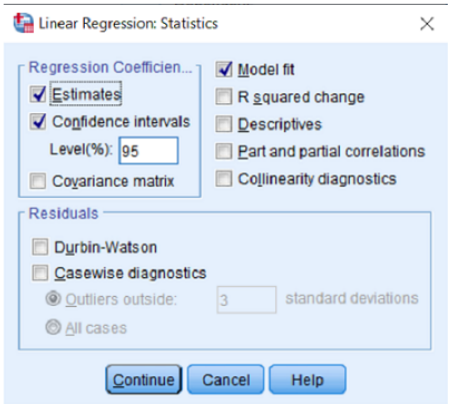
Στη γραμμή εντολών ακολουθούμε τη διαδρομή Analyze $\rightarrow$ Regression $\rightarrow$ Linear.



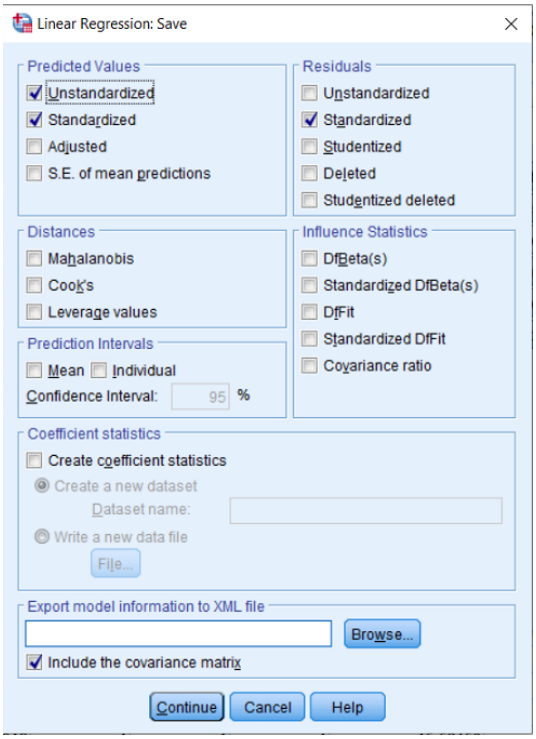
Στο παράθυρο που εμφανίζεται μετακινούμε δεξιά στο κελί Dependent την εξαρτημένη μεταβλητή Acceleration και στο κελί Independent(s) την ανεξάρτητη μεταβλητή Horsepower. Στο πεδίο Method μπορούμε να επιλέξουμε τη μέθοδο επιλογής του βέλτιστου μοντέλου, η οποία έχει νόημα μόνο όταν οι ανεξάρτητες μεταβλητές είναι παραπάνω από μία, οπότε αφήνουμε την προεπιλογή Enter ως έχει.

Στη συνέχεια, στην καρτέλα **Statistics** επιλέγουμε Estimates, Confidence Intervals και Model Fit.





Από την καρτέλα **Save** παρακάτω μας ενδιαφέρουν τα Predicted values και τα Residuals. Από τα Predicted values επιλέγουμε τα Standardized και τα Unstandardized, ενώ από τα Residuals επιλέγουμε τα Standardized. Στη συνέχεια πατάμε Continue και τέλος OK.



Αποτελέσματα

Το ενδιαφέρον εστιάζεται στους πίνακες Model Summary, ANOVA και Coefficients.

Πίνακας Model Summary

Στον παρακάτω πίνακα η πρώτη τιμή αντιστοιχεί στο συντελεστή προσδιορισμού R του Pearson, ο οποίος αναλύθηκε στην προηγούμενη ενότητα. Καλύτερη φυσική ερμηνεία της συσχέτισης δύο μεταβλητών επιτυγχάνεται με το συντελεστή προσδιορισμού R^2, ο οποίος παίρνει τιμές στο $[0, 1]$ και δίνεται από τον τύπο

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

όπου $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, $SSE = \sum_{i=1}^{n}(\hat{Y}_i - \hat{Y}_i)^2$ και $SST = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$.

Η τιμή του συντελεστή προσδιορισμού στο συγκεκριμένο παράδειγμα είναι 0.468 ή περίπου 47% . Αυτό σημαίνει ότι περίπου το 47% της μεταβλητότητας της μεταβλητής Acceleration εξηγείται από την Horse-power και το 53% από τα τυχαία σφάλματα.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,684[a]	,468	,467	1,969

a. Predictors: (Constant), Horsepower

b. Dependent Variable: Acceleration

Πίνακας ANOVA

Η πρώτη στήλη του πίνακα ANOVA δείχνει τα τετραγωνικά αθροίσματα $SSR = 1.328,213$, $SSE = 1.508,377$ και $SST = 2.836,590$. Οι βαθμοί ελευθερίας για το συγκεκριμένο μοντέλο είναι $p = 1$ για την παλινδρόμηση, $n = p - 1 = 391 - 1 - 1 = 389$ για τα κατάλοιπα και $n - 1 = 390$ για το συνολικό τετραγωνικό άθροισμα. Πραγματοποιείται ο παρακάτω έλεγχος (F-test) για την καλή προσαρμογή του μοντέλου στα δεδομένα:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Απορρίπτουμε τη μηδενική υπόθεση αν:

$$F = \frac{MSR}{MSE} > \mathcal{F}_{\nu_1,\nu_2}(\alpha)$$

όπου $\nu_1 = 1, \nu_2 = 389$ και επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$. Ωστόσο η τιμή της ελεγχοσυνάρτησης F δεν είναι πληροφοριακή από μόνη της, καθώς πρέπει να γνωρίζουμε και το άνω α - σημείο της κατανομής F με ν_1 και ν_2 βαθμούς ελευθερίας αντίστοιχα.

Για το λόγο αυτό καταφεύγουμε στην τελευταία στήλη του πίνακα ANOVA, όπου η μηδενική υπόθεση απορρίπτεται αν η τιμή του Sig. -που αντιστοιχεί στο p-value- είναι μικρότερη του 0.05. Στο παράδειγμά μας είναι $Sig. << 0.05$, οπότε συμπεραίνουμε ότι η παράμετρος β_1 είναι στατιστικά σημαντική και επομένως η ύπαρξη της μεταβλητής Horsepower έχει νόημα στο μοτέλο.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1328,213	1	1328,213	342,537	,000[b]
	Residual	1508,377	389	3,878		
	Total	2836,590	390			

a. Dependent Variable: Acceleration

b. Predictors: (Constant), Horsepower

Πίνακας Coefficients

Ο πίνακας Coefficients είναι τελικά αυτός που δίνει δομή στην εξίσωση της απλής γραμμικής παλινδρόμησης, καθώς περίεχει τις εκτιμημένες παραμέτρους, τη στατιστική σημαντικότητά τους αλλά και τα διαστήματα εμπιστοσύνης.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	20,647	,289		71,385	,000	20,078	21,215
	Horsepower	-,048	,003	-,684	-18,508	,000	-,053	-,043

a. Dependent Variable: Acceleration

Η πρώτη στήλη του πίνακα αφορά τις εκτιμήσεις των παραμέτρων.

Ερμηνεία β_0: Η τιμή του β_0 είναι $20,647$ και δηλώνει την τιμή της εξαρτημένης μεταβλητής όταν η ανεξάρτητη παίρνει την τιμή μηδέν. Ωστόσο, στη συγκεκριμένη περίπτωση δεν έχει φυσικό νόημα, αφού ένα αυτοκίνητο είναι αδύνατο να ξεκινήσει όταν δεν έχει ιπποδύναμη (!) και προφανώς δε θα έχει και επιτάχυνση.

Ερμηνεία β_1: Η τιμή του β_1 είναι -0.048 και είναι η αρνητική κλίση της ευθείας της γραμμικής παλινδρόμησης. Υποδηλώνει ότι για μοναδιαία αύξηση της ιπποδύναμης, ο χρόνος που απαιτείται για να φτάσει το αυτοκίνητο το $100km/h$ μειώνεται κατά 0.048 χρονικές μονάδες.

Ο πίνακας Coefficients περιλαμβάνει και τους ελέγχους στατιστικής σημαντικότας του μοντέλου για τις δύο παραμέτρους. Ο έλεγχος αυτός (t-test) είναι ισοδύναμος με τον έλεγχο F-test που αναφέρθηκε παραπάνω, αλλά μόνο στην απλή γραμμική παλινδρόμηση. Ο έλεγχος είναι ο εξής:

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

για $i = 0, 1$. Απορρίπτουμε την αρχική υπόθεση αν $\mid T \mid > t_{n-p-1,1-\frac{\alpha}{2}}$, όπου $T = \frac{\beta_i - 0}{s.e(\beta_i)}$ και $t_{n-p-1,1-\frac{\alpha}{2}}$ η κριτική τιμή της κατανομής Student-t με $n - p - 1$ β.ε. και επίπεδο σημαντικότητας α. Στο παράδειγμά μας έχουμε $n = 391$, $p = 1$ και $\alpha = 0.05$.

Ένας ευκολότερος τρόπος για να εξετάσουμε αν απορρίπτεται η μηδενική υπόθεση είναι να ελέγξουμε αν το p-value είναι μικρότερο από το επίπεδο σημαντικότητας (5%). Παρατηρούμε από την πέμπτη στήλη του πίνακα ότι $Sig. << 0.05$ και για τις δύο μεταβλητές, οπότε έχει νόημα η ύπαρξή τους στο μοντέλο.

Ένας άλλος τρόπος να ελέγξουμε αν οι παράμετροι είναι στατιστικά σημαντικές είναι να δούμε αν το 0 συμπεριλαμβάνεται στο 95% διάστημα επιστοσύνης. Και στις δύο περιπτώσεις το 0 δεν ανήκει στο διάστημα, οπότε καταλήγουμε στο ίδιο συμπέρασμα.
Τελικά, η εκτιμημένη εξίσωση της απλής γραμμικής παλινδρόμησης είναι:

$$(Acceleration) = 20.647 - 0.048 \cdot (Horsepower)$$

Έλεγχος υποθέσεων γραμμικού μοντέλου

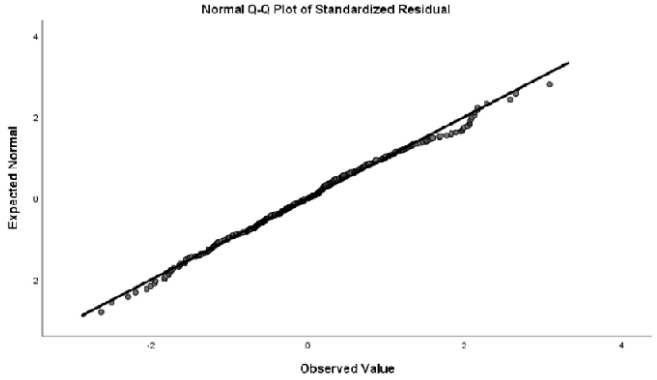
Έλεγχος κανονικότητας: Για να ελέγξουμε αν τα κατάλοιπα του μοντέλου ακολουθούν κανονική κατανομή, ακολουθούμε την παρακάτω διαδικασία:

Στη γραμμή εντολών ακολουθούμε τη διαδρομή Analyze → Descriptive Statistics → Explore.

Στο παράθυρο που εμφανίζεται στο κελί Dependent list μετακινούμε δεξιά τα Standardized Residuals τα οποία είχαμε αποθηκεύσει νωρίτερα. Στην καρτέλα Statistics επιλέγουμε Normality plots with tests.

Προκύπτει από τον παρακάτω πίνακα ότι η αρχική υπόθεση της κανονικότητας δεν μπορεί να απορριφθεί σε επίπεδο στατιστικής σημαντικότητας $\alpha = 5\%$, καθώς σύμφωνα με τους ελέγχους Kolmogorov-Smirnov και Shapiro-Wil $Sig > 0.05$.

Τέλος, από το Q-Q plot των Standardized residuals δεν φαίνεται γενικά να έχουμε πολύ μεγάλες αποκλίσεις από τα ποσοστημόρια της κανονικής κατανομής, αν εξαιρέσουμε κάποιες αποκλίσεις στο άνω δεξί τμήμα του γραφήματος.

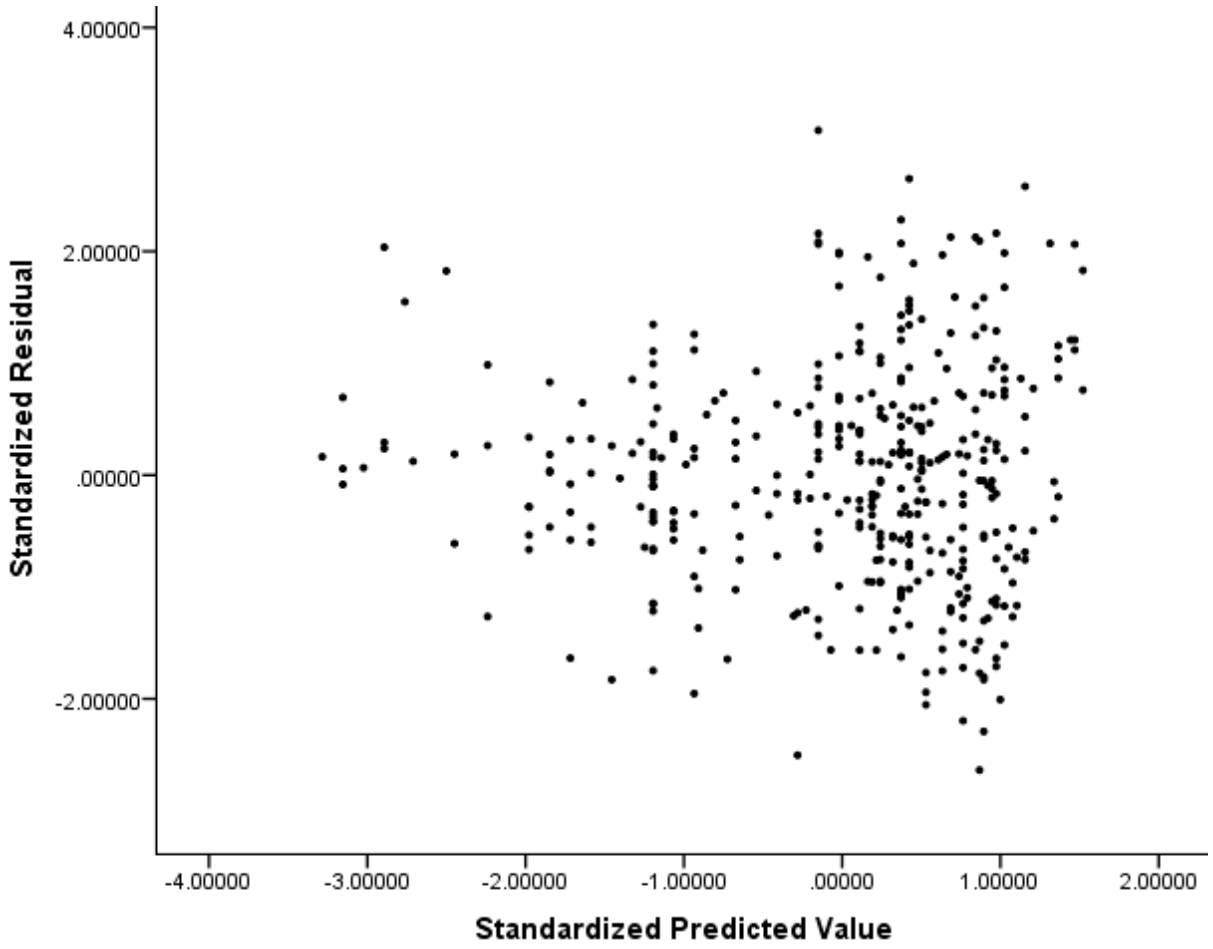


Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	,040	391	,141	,994	391	,143

a. Lilliefors Significance Correction

Ομοσκεδαστικότητα και Ανεξαρτησία: Παρότι υπάρχουν στατιστικοί έλεγχοι για την ομοσκεδαστικότητα και την ανεξαρτησία των καταλοίπων, εδώ θα αρκεστούμε απλά σε ένα γραφικό έλεγχο. Συγκεκριμένα θα δημιουργήσουμε το σημειόγραμμα (scatterplot) των τυποποιημένων καταλοίπων (standardized residuals) ως προς τις τυποποιημένες προβλεπόμενες τιμές της y (standardized predicted values).



Αρχικά, εξετάζουμε την ανεξαρτησία των καταλοίπων. Δεν παρατηρούμε τάσεις (όπως για παράδειγμα πολλά κατάλοιπα συγκεντρωμένα στην θετική μεριά, έπειτα πολλά συγκεντρωμένα στα αρνητικά κ.ο.κ.). Επομένως υποθέτουμε με μία σχετική βεβαιότητα ότι τα κατάλοιπα είναι ανεξάρτητα. Προχωρώντας όμως στην υπόθεση της ομοσκεδαστικότητας (σταθερής διασποράς), παρατηρούμε ότι το γράφημα δεν έχει τη μορφή ενός τυχαίου συννέφου γύρω από την ευθεία $y = 0$. Τα κατάλοιπα παρουσιάζουν ετεροσκεδαστικότητα, δείχνουν δηλαδή να μην κρατάνε σταθερή απόσταση από το 0 καθ' όλη τη διάρκεια του δείγματος, αλλά να αυξάνονται καθώς αυξάνονται και οι τιμές στον άξονα x. Τα παραπάνω υποδεικνύουν ότι πρέπει να δοκιμάσουμε εναλλακτικές μεθόδους μοντελοποίησης ώστε να έχουμε ομοσκεδαστικότητα.

Ασκήσεις

Κατασκευάστε απλά γραμμικά μοντέλα για τα δεδομένα του αρχείου grades.

1. Επιλέξτε ως εξαρτημένη μεταβλητή (y) την Grades και σαν ανεξάρτητη (x) την Maths.
2. Επιλέξτε ως εξαρτημένη μεταβλητή (y) την Grades και σαν ανεξάρτητη (x) την Physics.
3. Ανταλλάξτε την εξαρτημένη μεταβλητή με την ανεξάρτητη. Επιλέξτε δηλαδή ως εξαρτημένη μεταβλητή (y) την Maths και σαν ανεξάρτητη (x) την Grades. Συγκρίνετε τα μοντέλα (1) και (3). Θα μπορούσατε από το ένα να συμπεράνετε το άλλο;

5 Πολλαπλή Γραμμική Παλινδρόμηση

Σύνοψη: Αφού εξετάστηκε η συσχέτιση μεταξύ των μεταβλητών και κατασκευάστηκαν κατάλληλα μοντέλα απλής γραμμικής παλινδρόμησης, ακολουθεί η εφαρμογή και ανάλυση της πολλαπλής γραμμικής παλινδρόμησης.

Εισαγωγή

Συχνά οι τιμές μιας μεταβλητής ενδιαφέροντος εξαρτώνται από τις τιμές περισσότερων από μίας επεξηγηματικών μεταβλητών. Όταν μάλιστα αυτή η σχέση εξάρτησης είναι γραμμική μπορεί να εκφραστεί μέσα από ένα πολλαπλό γραμμικό μοντέλο. Ένα τέτοιο μοντέλο έχει τη μορφή:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + \epsilon_i, \quad i = 1, .., n$$

Όπου:

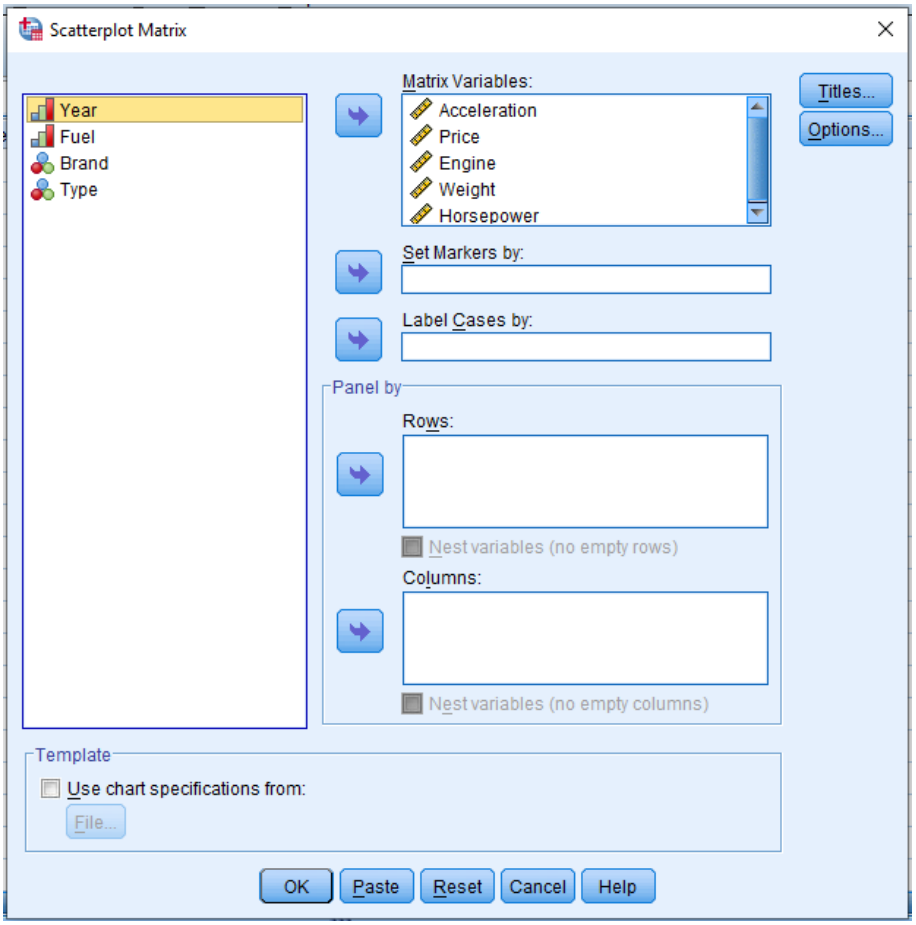
- Y είναι η εξαρτημένη μεταβλητή,
- $X_j,\ j = 1, ..., p$ είναι η j-οστή ανεξάρτητη μεταβλητή,
- β_0 είναι η τιμή της εξαρτημένης μεταβλητής όταν όλες οι επεξηγηματικές μεταβλητές πάρουν την τιμή 0, χωρίς αυτή να έχει πάντα φυσική σημασία στην ανάλυση,
- $\beta_j,\ j = 1, ..., p$ εκφράζει τη μεταβολή της Y για μοναδιαία μεταβολή της X_j με την προϋπόθεση ότι οι τιμές των υπόλοιπων επεξηγηματικών μεταβλητών παραμένουν σταθερές και
- ϵ_i τα τυχαία σφάλματα, τα οποία υποθέτουμε πως είναι κανονικά κατανεμημένα και ομοσκεδαστικά με $\epsilon_i \sim N(0, \sigma^2)$

Σκοπός μας είναι η εκτίμηση των συντελεστών $\beta_j,\ \ j = 0, 1, , ..., p.$ Οι εκτιμήσεις των παραμέτρων υπολογίζονται μέσω της μεθόδου ελαχίστων τετραγώνων (Μ.Ε.Τ.), όπως έχει αναλυθεί στη θεωρία.

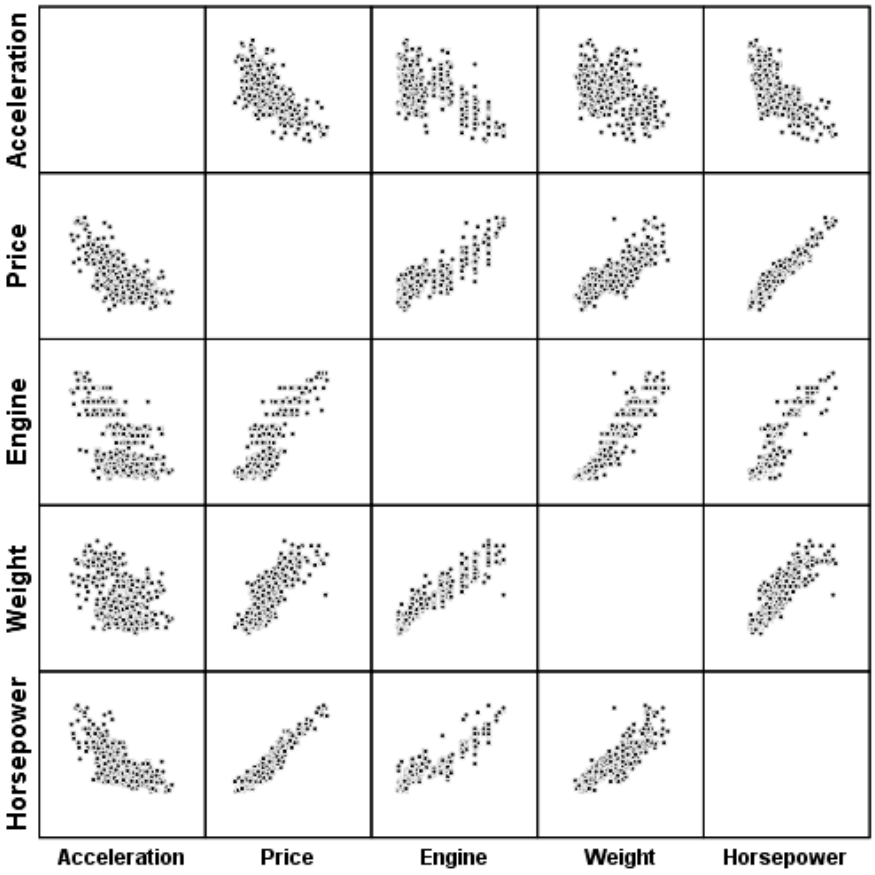
Εφαρμογή του Πλήρους Μοντέλου

Η εφαρμογή της πολλαπλής γραμμικής παλινδρόμησης πραγματοποιείται στο σετ δεδομένων cars, ως συνέχεια της προηγούμενης ενότητας. Θα θεωρήσουμε ως εξαρτημένη μεταβλητή (Y) την Acceleration και θα διερευνήσουμε το αν και πώς αυτή εξαρτάται από τις υπόλοιπες ποσοτικές μεταβλητές που έχουμε στα χέρια μας (Engine-X_1, Price-X_2, Weight-X_3, Horsepower-X_4) συνολικά.

Αρχικά παίρνουμε μια πρώτη ιδέα για τις σχέσεις μεταξύ των μεταβλητών οπτικοποιώντας τα δεδομένα που διαθέτουμε όπως είδαμε στην ενότητα 3. Στη γραμμή εντολών ακολουθούμε τη διαδρομή Graphs $\rightarrow$ Legacy Dialogs $\rightarrow$ Scatter/Dot. Και αυτή τη φορά επιλέγουμε το Matrix Scatter πατώντας το κουμπί Define. Στο αναδυόμενο παράθυρο, μετακινούμε τις μεταβλητές που μας ενδιαφέρουν, εδώ Acceleration, Engine, Price, Weight και Horsepower, στο κελί Matrix Variables.

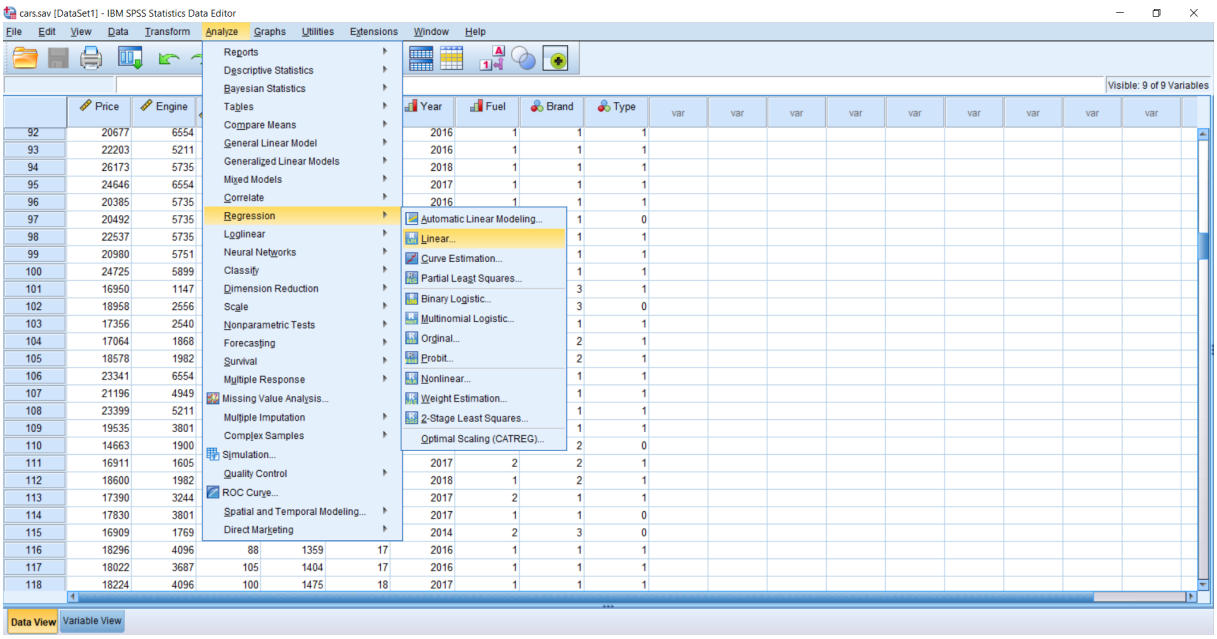


Κάνοντας διπλό κλικ στο γράφημα που προκύπτει μπορούμε να επεξεργαστούμε τα μεγέθη, τα χρώματα και ό,τι άλλο μας ενδιαφέρει. Το γράφημα που παίρνουμε είναι το ακόλουθο:

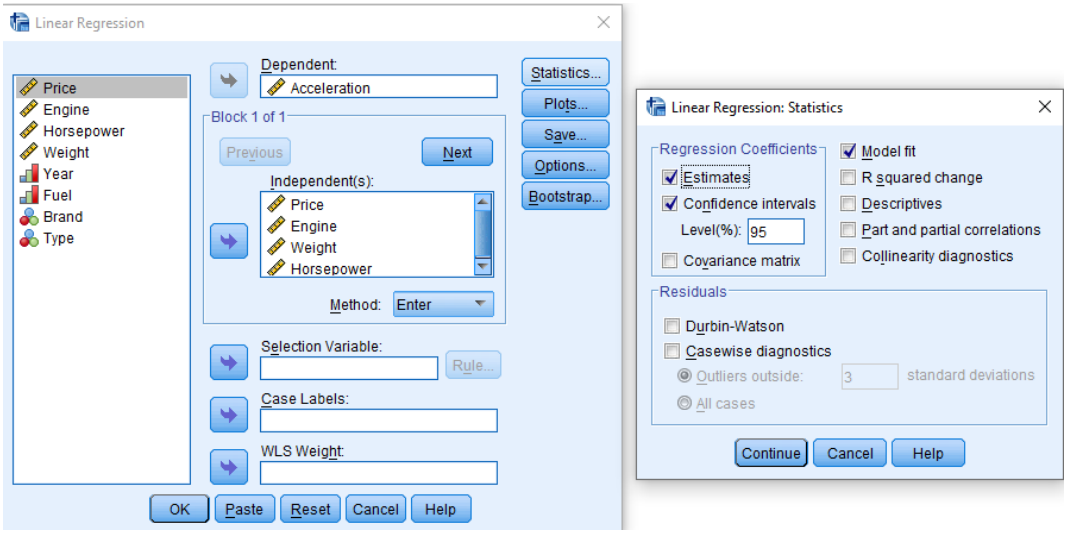


Στη συνέχεια προχωράμε με την εφαρμογή του μοντέλου. Ένα βασικό ζήτημα που καλούμαστε να αντιμετωπίσουμε είναι το ποιες επεξηγηματικές μεταβλητές θα συμπεριλάβουμε στην ανάλυσή μας. Αρχικά τις εισάγουμε όλες και θα συζητήσουμε στη συνέχεια περισσότερα.

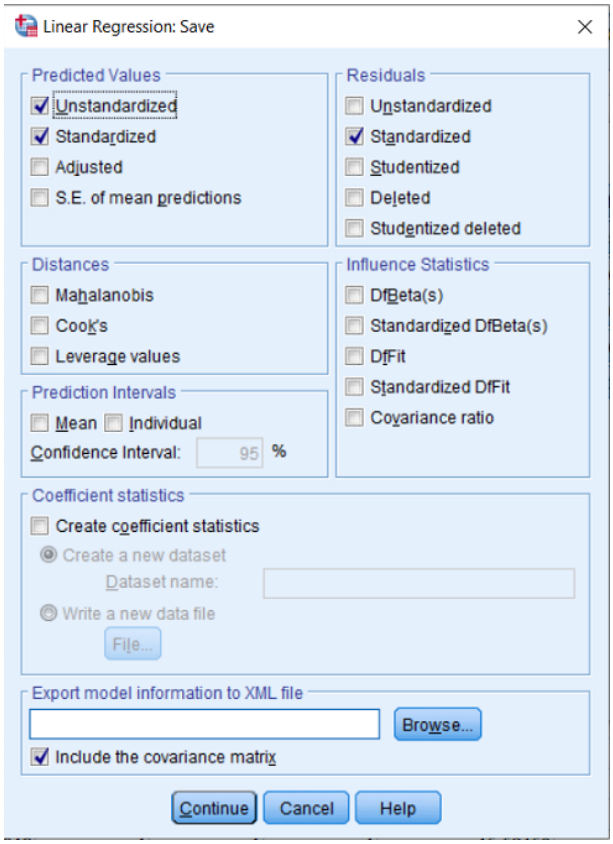
Στη γραμμή εντολών ακολουθούμε τη διαδρομή Analyze $\rightarrow$ Regression $\rightarrow$ Linear όπως και στην απλή γραμμική παλινδρόμηση.



Στο παράθυρο που εμφανίζεται μετακινούμε δεξιά, στο κελί Dependent, την εξαρτημένη μεταβλητή Acceleration και στο κελί Independent(s) τις ανεξάρτητες μεταβλητές Engine, Price, Weight και Horsepower. Στο πεδίο Method μπορούμε να επιλέξουμε τη μέθοδο επιλογής του βέλτιστου μοντέλου στην οποία αναφερθήκαμε παραπάνω. Αφήνοντας την προεπιλογή Enter δεν γίνεται επιλογή των μεταβλητών και συνεπώς εισάγονται όλες στο προς εκτίμηση μοντέλο. Στη συνέχεια, στην καρτέλα **Statistics** επιλέγουμε Estimates, Confidence Intervals και Model Fit.



Από την καρτέλα **Save** παρακάτω μας ενδιαφέρουν τα Predicted values και τα Residuals. Από τα Predicted values επιλέγουμε τα Standardized και τα Unstandardized, ενώ από τα Residuals επιλέγουμε τα Standardized. Στη συνέχεια πατάμε Continue και τέλος OK.



Έτσι ζητάμε την εκτίμηση των συντελεστών, τα αντίστοιχα διαστήματα εμπιστοσύνης και άλλες ποσότητες που μας ενδιαφέρουν για το παρακάτω μοντέλο:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i, \quad i = 1, .., 391$$

Αποτελέσματα για το Πλήρες Μοντέλο

Το ενδιαφέρον, και εδώ, εστιάζεται στους πίνακες Model Summary, ANOVA και Coefficients.

Πίνακας Model Summary

Στον παρακάτω πίνακα η πρώτη τιμή αντιστοιχεί στο συντελεστή προσδιορισμού R του Pearson, η δεύτερη στο συντελεστή προσδιορισμού R^2, οι οποίοι αναλύθηκαν σε προηγούμενη ενότητα. Η τρίτη στήλη αντιστοιχεί στην τιμή του προσαρμοσμένου συντελεστή προσδιορισμού R^2_{adj}, ο οποίος παίρνει τιμές και αυτός στο $[0, 1]$ και λαμβάνει υπόψη του, επιπλέον, τον αριθμό των επεξηγηματικών μεταβλητών του εκάστοτε μοντέλου, δίνοντας μια ποινή για αύξηση του αριθμού αυτού. Αυτό έχει ως αποτέλεσμα σε περίπτωση που εισάγουμε, για παράδειγμα, στο μοντέλο μια μεταβλητή που δεν εξηγεί σημαντικό ποσοστό της εναπομένουσας μεταβλητότητας, να μειωθεί η τιμή του R^2_{adj} υπονοώντας ότι το μοντέλο χωρίς την τελευταία προσθήκη είναι καταλληλότερο. Η τελευταία ιδιότητα έρχεται σε αντίθεση με τη συμπεριφορά του συντελεστή R^2, ο οποίος πάντα αυξάνεται με την εισαγωγή επιπλέον επεξηγηματικών μεταβλητών. Συνεπώς, ο προσαρμοσμένος συντελεστή προσδιορισμού είναι καταλληλότερος για τη σύγκριση μοντέλων με διαφορετικό αριθμό επεξηγηματικών μεταβλητών. Δίνεται από τον τύπο

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}\frac{SSE}{SST}$$

όπου $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,\ SST = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, n το πλήθος των παρατηρήσεων και p το πλήθος των επεξηγηματικών μεταβλητών.

Η τιμή του R^2_{adj} είναι 0.601 ή περίπου 60%, που σε σχέση με το απλό γραμμικό μοντέλο που είχαμε R^2_{adj} ίσο με 0.467 είναι μεγαλύτερη δηλώνοντας καλύτερη προσαρμογή.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,778[a]	,606	,601	1,703

a. Predictors: (Constant), Horsepower, Weight, Price, Engine

Πίνακας ANOVA

Στην πρώτη στήλη του πίνακα ANOVA μπορούμε να δούμε τα τετραγωνικά αθροίσματα SSR, SSE και SST. Οι βαθμοί ελευθερίας για το συγκεκριμένο μοντέλο είναι $p = 4$ για την παλινδρόμηση, $n - p -$

$1 = 391 - 4 - 1 = 386$ για τα κατάλοιπα και $n - 1 = 390$ για το συνολικό τετραγωνικό άθροισμα. Πραγματοποιείται ο παρακάτω έλεγχος στατιστικής σημαντικότητας για το μοντέλο (F-test):

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_1 : \beta_j \neq 0 \text{ για ένα τουλάχιστον } j,\ j = 1, 2, 3, 4$$

Η H_0 αντιστοιχεί στη μη ύπαρξη εξάρτησης της εξαρτημένης μεταβλητής από καμία από τις ανεξάρτητες και επομένως δεν υπάρχει μοντέλο για να εκτιμήσουμε.

Απορρίπτουμε τη μηδενική υπόθεση αν το p-value είναι μικρότερο από το επιθυμητό επίπεδο στατιστικής σημαντικότητας, συνήθως 0.05. Όπως είδαμε και στην περίπτωση της απλής γραμμικής παλινδρόμησης, το p-value θα το βρούμε στον πίνακα ANOVA στην τελευταία στήλη με τίτλο Sig. Στην περίπτωσή μας είναι Sig$<< 0.05$, επομένως απορρίπτεται η μηδενική υπόθεση και έχει νήμα να προχωρήσουμε στην ανάλυσή μας.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1717,662	4	429,416	148,137	,000[b]
	Residual	1118,928	386	2,899		
	Total	2836,590	390			

a. Dependent Variable: Acceleration

b. Predictors: (Constant), Horsepower, Weight, Price, Engine

Πίνακας Coefficients

Ο πίνακας Coefficients είναι τελικά αυτός που δίνει δομή στην εξίσωση της παλινδρόμησης, καθώς περιέχει τις εκτιμημένες παραμέτρους, τη στατιστική σημαντικότητά τους αλλά και τα αντίστοιχα διαστήματα εμπιστοσύνης.

Πριν ασχοληθούμε όμως, με τις εκτιμήσεις των παραμέτρων πρέπει να ελέγξουμε τη στατιστική σημαντικότητα για κάθε έναν από τους συντελεστές του μοντέλου. Αυτό γίνεται με τη πραγματοποίηση τεσσάρων διαφορετικών t-tests (για $i = 1, ..., 4$), ελέγχους της μορφής:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Την τιμή της κάθε ελεγχοσυνάρτησης θα τη βρούμε στη στήλη του πίνακα Coefficients, με τίτλο t, ενώ το αντίστοιχο p-value, που είναι και αυτό που μας ενδιαφέρει για τη λήψη απόφασης, θα το βρούμε στη διπλανή στήλη με τίτλο Sig. Όπως μπορούμε να δούμε όλοι οι συντελεστές προκύπτουν στατιστικά σημαντικοί, εκτός από εκείνον τη μεταβλητής Price.

Coefficients[a]

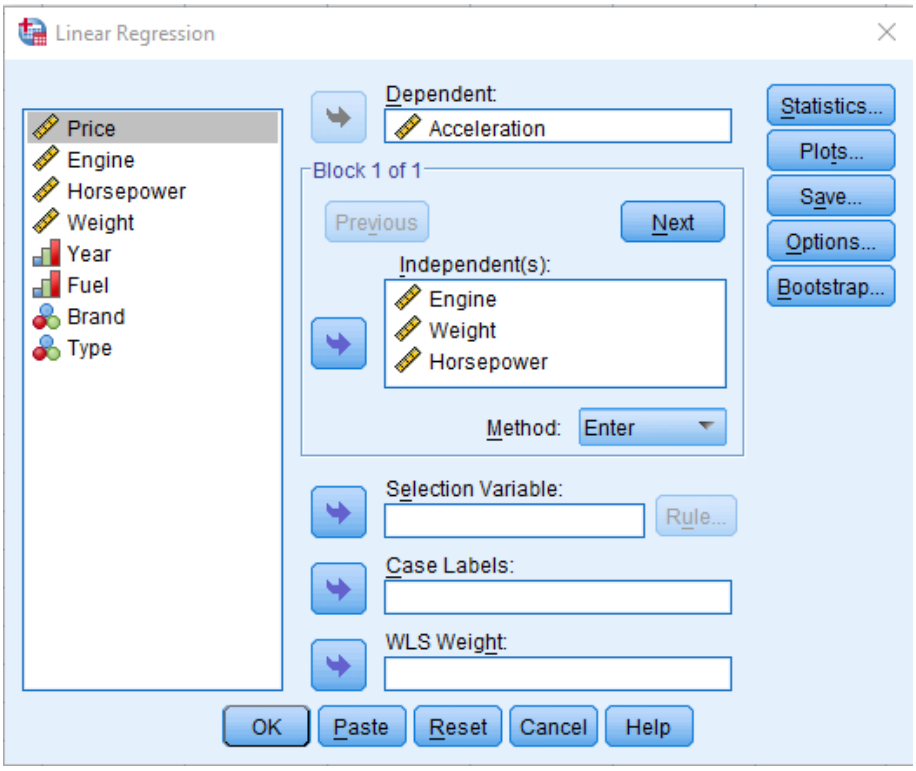
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	17,288	,761		22,707	,000	15,791	18,785
	Price	-7,015E-006	,000	-,010	-,097	,922	,000	,000
	Engine	-,001	,000	-,341	-3,273	,001	-,001	,000
	Weight	,006	,001	,915	10,116	,000	,005	,008
	Horsepower	-,082	,009	-1,158	-9,504	,000	-,098	-,065

a. Dependent Variable: Acceleration

Επιλογή Μοντέλου

Σε αυτό το σημείο έφτασε η ώρα να επανέλθουμε στο θέμα της επιλογής επεξηγηματικών μεταβλητών που αναφέρθηκε στην αρχή της ενότητας. Ένας τρόπος να προχωρήσουμε, είναι να αφαιρέσουμε ”χειροκίνητα” τη μεταβλητή με το μεγαλύτερο p-value και να εκτιμήσουμε από τη αρχή ένα νέο μοντέλο, αυτή τη φορά με τρεις επεξηγηματικές μεταβλητές. Προσοχή, ακόμα και αν έχουμε εικόνα για το ποιες μεταβλητές είναι στατιστικά σημαντικές από το πλήρες μοντέλο, θα πρέπει να ξαναγίνουν εκ νέου οι έλεγχοι για το νέο, μειωμένο μοντέλο.

Συνεχίζουμε, λοιπόν, επιλέγοντας στη γραμμή εντολών τη διαδρομή Analyze → Regression → Linear όπως και στην προηγούμενη περίπτωση, με τη διαφορά πως αυτή τη φορά δεν προσθέτουμε τη μεταβλητή Price στο κελί Independent(s), όπως φαίνεται παρακάτω:



Στο παράθυρο των αποτελεσμάτων, αρχικά εξετάζουμε τον πίνακα Model Summary ο οποίος για την ώρα δεν μας παρέχει κάποια ιδιαίτερα χρήσιμη πληροφορία καθώς διαφέρει ελάχιστα από τον αντίστοιχο για το πλήρες μοντέλο.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,778[a]	,606	,602	1,700

a. Predictors: (Constant), Horsepower, Weight, Engine

Ο πίνακας που μας ενδιαφέρει περισσότερο για τη συνέχεια της ανάλυσής μας είναι ο πίνακας Coefficients από τον οποίο μπορούμε να διαπιστώσουμε, κοιτώντας τη στήλη Sig με τα p-value, πως όλοι οι συντελεστές είναι στατιστικά σημαντικοί. Αυτό αποτελεί ένδειξη πως αν αφαιρέσουμε κάποια από τις υπάρχουσες επεξηγηματικές μεταβλητές θα οδηγηθούμε σε ένα λιγότερο ερμηνευτικό μοντέλο. Καλείστε να το ελέγξετε συγκρίνοντας τους προσαρμοσμένους συντελεστές προσδιορισμού για τα διάφορα μοντέλα.

Καταλήγουμε λοιπόν σε ένα μοντέλο με τρεις επεξηγηματικές μεταβλητές: Engine-X_1, Horsepower-X_2 και Weight-X_3 για την ανεξάρτητη μεταβλητή Acceleration. Το οποίο είναι και αυτό που θα εκτιμήσουμε

στη συνέχεια.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i, \quad i = 1, .., 391$$

Η μέθοδος επιλογής μοντέλου που περιγράφηκε αντιστοιχεί σε μια εκτέλεση της μεθόδου backward, που είναι και μία από τις μεθόδους που μπορεί να εκτελέσει το SPSS αυτόματα όπως θα δούμε στη συνέχεια.

Επειδή συχνά οι διαθέσιμες επεξηγηματικές μεταβλητές είναι πολύ περισσότερες από ότι στο παράδειγμα που εξετάζουμε, όπως μπορείτε να φανταστείτε, η εκτέλεση της παραπάνω διαδικασίας μπορεί να γίνει πολύ πιο χρονοβόρα. Όπως ήδη αναφέρθηκε στο SPSS υπάρχουν διαθέσιμες κάποιες αυτοματοποιημένες μέθοδοι επιλογής. Εν συντομία αυτές είναι οι ακόλουθες:

Forward: Η μέθοδος αυτή ξεκινάει από το κενό/μηδενικό μοντέλο (null model), χωρίς επεξηγηματικές μεταβλητές και εισάγει στο μοντέλο τη μεταβλητή την περισσότερο στατιστικά σημαντική (με το μικρότερο p-value). Η διαδικασία επαναλαμβάνεται έως ότου δεν υπάρχει μεταβλητή που βελτιώνει στατιστικά σημαντικά το μοντέλο.

Backward: Η μέθοδος αυτή ξεκινά με το μοντέλο που περιλαμβάνει όλες τις υποψήφιες επεξηγηματικές μεταβλητές και συνεχίζει αφαιρώντας κάθε φορά τη λιγότερο στατιστικά σημαντική (με το μεγαλύτερο p-value). Η διαδικασία επαναλαμβάνεται μέχρι όλες οι επεξηγηματικές μεταβλητές που περιλαμβάνονται στο μοντέλο να είναι στατιστικά σημαντικές.

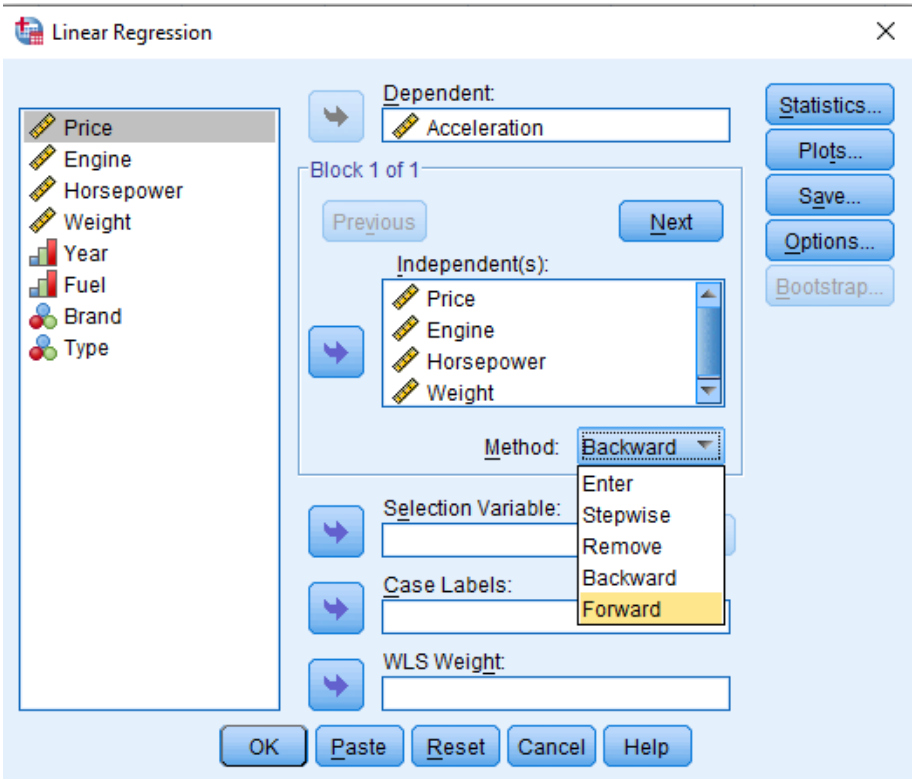
Stepwise: Η παρούσα μέθοδος αποτελεί ένα συνδυασμό των δύο παραπάνω. Είναι μια τροποποίηση της μεθόδου Forward, όπου σε κάθε βήμα, μετά την εισαγωγή της εκάστοτε μεταβλητής, ελέγχεται αν πρέπει να αφαιρεθεί κάποια από τις ήδη υπάρχουσες με βάση κάποιο κριτήριο επιλογής. Η μέθοδος σταματά όταν δεν μπορεί να εισαχθεί ή να αφαιρεθεί κάποια μεταβλητή.

Remove: Η συγκεκριμένη μέθοδος είναι το αντίθετο της Enter. Δηλώνουμε τις μεταβλητές που θέλουμε να μην μπουν στο μοντέλο μας. Σπάνια την χρησιμοποιούμε.

Σημειώνεται ότι διαφορετικές μέθοδοι επιλογής μπορεί να οδηγήσουν σε διαφορετικά μοντέλα.

Ας δούμε τώρα πως μπορούν να εφαρμοστούν τα όσα αναλύσαμε στο SPSS. Εφόσον πήραμε μια εικόνα για

την Bacward μέθοδος επιλογής στη ”χειροκίνητη” εκδοχή της ας δούμε πως μπορούμε να εφαρμόσουμε τη Forward μέθοδο. Στη γραμμή εντολών ακολουθούμε τη διαδρομή Analyze → Regression → Linear ακριβώς όπως πριν. Στο παράθυρο που εμφανίζεται μετακινούμε δεξιά, στο κελί Dependent, την εξαρτημένη μεταβλητή Acceleration και στο κελί Independent(s) όλες τις υποψήφιες επεξηγηματικές μεταβλητές Engine, Price, Weight και Horsepower. Η διαφορά αφορά στο πεδίο Method, από το οποίο, όπως αναφέρθηκε, μπορούμε να επιλέξουμε τη μέθοδο επιλογής του βέλτιστου μοντέλου. Διαλέγουμε λοιπόν την επιλογή Forward. Στη συνέχεια, στην καρτέλα **Statistics** επιλέγουμε Estimates, Confidence Intervals και Model Fit, όπως κάναμε και πριν.



Εφαρμόζοντας τα παραπάνω, στα αποτελέσματα θα δούμε πέντε πίνακες αυτή τη φορά. Αρχικά, εστιάζουμε τους πίνακες Model Summary, ANOVA και Coefficients, που είδαμε και παραπάνω, με τη διαφορά πως αυτή τη φορά μας παρέχουν πληροφορίες για όλα τα μοντέλα που εξετάστηκαν και εκτιμήθηκαν μέχρι να προκύψει το τελικό. Κάθε γραμμή αντιστοιχεί σε διαφορετικό μοντέλο και μπορούμε να την ερμηνεύσουμε όπως αναλύθηκε παραπάνω.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,684[a]	,468	,467	1,969
2	,771[b]	,595	,592	1,722
3	,778[c]	,606	,602	1,700

a. Predictors: (Constant), Horsepower

b. Predictors: (Constant), Horsepower, Weight

c. Predictors: (Constant), Horsepower, Weight, Engine

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1328,213	1	1328,213	342,537	,000[b]
	Residual	1508,377	389	3,878		
	Total	2836,590	390			
2	Regression	1686,587	2	843,294	284,519	,000[c]
	Residual	1150,003	388	2,964		
	Total	2836,590	390			
3	Regression	1717,635	3	572,545	198,019	,000[d]
	Residual	1118,955	387	2,891		
	Total	2836,590	390			

a. Dependent Variable: Acceleration

b. Predictors: (Constant), Horsepower

c. Predictors: (Constant), Horsepower, Weight

d. Predictors: (Constant), Horsepower, Weight, Engine

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	20,647	,289		71,385	,000
	Horsepower	-,048	,003	-,684	-18,508	,000
2	(Constant)	18,439	,323		57,104	,000
	Horsepower	-,091	,005	-1,291	-20,186	,000
	Weight	,005	,000	,703	10,996	,000
3	(Constant)	17,231	,487		35,355	,000
	Horsepower	-,082	,005	-1,168	-15,872	,000
	Weight	,006	,001	,915	10,129	,000
	Engine	-,001	,000	-,341	-3,277	,001

a. Dependent Variable: Acceleration

Οι δύο επιπλέον πίνακες που εμφανίζονται, με τίτλους Variables Entered/Removed και Variables Excluded μας πληροφορούν για το ποιές μεταβλητές εισήχθησαν ή αφαιρέθηκαν από το μοντέλο και με ποιά κριτήρια,

ο πρώτος, και για το ποιές από τις υποψήφιες μεταβλητές εξαιρέθηκαν και δεν μπήκαν στο μοντέλο σε κάθε επανάληψη της διαδικασίας, ο δεύτερος.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Horsepower	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)
2	Weight	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)
3	Engine	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)

a. Dependent Variable: Acceleration

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	Price	,009[b]	,074	,941	,004	,097
	Engine	,412[b]	5,043	,000	,248	,193
	Weight	,703[b]	10,996	,000	,487	,255
2	Price	-,008[c]	-,072	,942	-,004	,097
	Engine	-,341[c]	-3,277	,001	-,164	,094
3	Price	-,010[d]	-,097	,922	-,005	,097

a. Dependent Variable: Acceleration

b. Predictors in the Model: (Constant), Horsepower

c. Predictors in the Model: (Constant), Horsepower, Weight

d. Predictors in the Model: (Constant), Horsepower, Weight, Engine

Μπορούμε, λοιπόν, να δούμε από την πρώτη γραμμή του πίνακα Variables Entered/Removed, πως η πρώτη μεταβλητή που εισήχθη στο μοντέλο είναι η Horsepower, ενώ στην πρώτη γραμμή του πίνακα Variables Excluded μπορούμε να δούμε τις μεταβλητές που εξαιρέθηκαν από την εισαγωγή, συμπληρωματικά. Παρατηρούμε, επομένως, ότι η σειρά εισαγωγής των μεταβλητών είναι Horsepower, Weight και Engine, ενώ η μεταβλητή Price δεν επιλέγεται ποτέ.

Αποτελέσματα για το Τελικό Μοντέλο

Το μοντέλο στο οποίο καταλήγουν και οι δύο μέθοδοι επιλογής που εφαρμόσαμε είναι αυτό με τρεις επεξηγηματικές μεταβλητές, τις Engine-X_1, Horsepower-X_2 και Weight-X_3.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i, \quad i = 1, .., 391$$

Στον πίνακα Coefficients παρακάτω στη στήλη B, μπορούμε να δούμε τις εκτιμήσεις των συντελεστών του γραμμικού μοντέλου, ενώ από τις δύο τελευταίες με τίτλους Lower Bound και Upper Bound παίρνουμε τα αντίστοιχα διαστήματα εμπιστοσύνης. Έτσι, για παράδειγμα η εκτίμηση του συντελεστή της μεταβλητής X_2, Horsepower, είναι $\hat{\beta}_2 = -0.082$, με 95% διάστημα εμπιστοσύνης $(-0.092, -0.07200)$. Τελικά, η εκτιμημένη εξίσωση της πολλαπλής γραμμικής παλινδρόμησης είναι:

$$Y_i = 17.231 - 0.001 X_{1i} - 0.082 X_{2i} + 0.006 X_{3i}, \quad i = 1, .., 391$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	17,231	,487		35,355	,000	16,273	18,189
	Engine	-,001	,000	-,341	-3,277	,001	-,001	,000
	Horsepower	-,082	,005	-1,168	-15,872	,000	-,092	-,072
	Weight	,006	,001	,915	10,129	,000	,005	,008

a. Dependent Variable: Acceleration

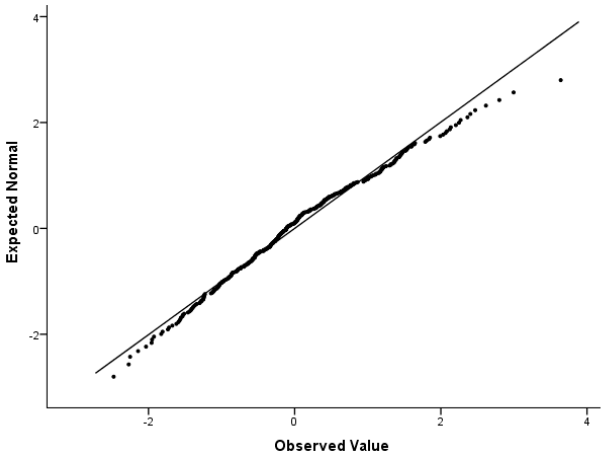
Έλεγχος υποθέσεων γραμμικού μοντέλου

Έλεγχος κανονικότητας: Για να ελέγξουμε αν τα κατάλοιπα του μοντέλου ακολουθούν κανονική κατανομή, ακολουθούμε τη διαδικασία που περιγράφηκε στο τέλος της ενότητας 4.

Στη γραμμή εντολών ακολουθούμε τη διαδρομή Analyze $\rightarrow$ Descriptive Statistics $\rightarrow$ Explore.

Στο παράθυρο που εμφανίζεται στο κελί Dependent list μετακινούμε δεξιά τα Standardized Residuals τα οποία είχαμε αποθηκεύσει νωρίτερα. Στην καρτέλα Plots επιλέγουμε Normality plots with tests.

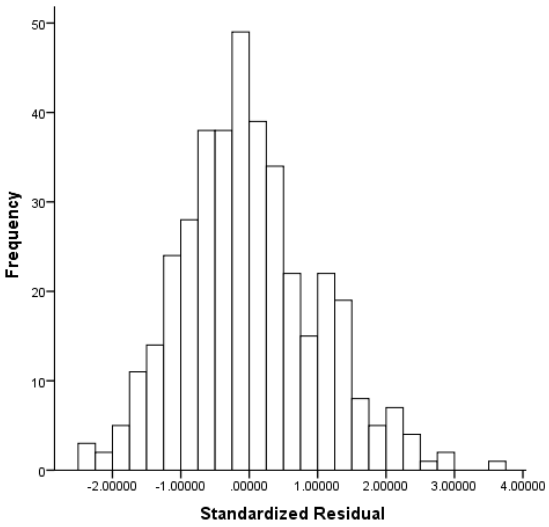
Βλέπουμε πως ο έλεγχος Shapiro-Wilk απορρίπτει την κανονικότητα των καταλοίπων σε επίπεδο στατιστικής σημαντικότητας 5% (p-value=2%). Από τα τρία γραφήματα είναι αρκετά εμφανής η δεξιά ουρά των καταλοίπων, η οποία προκαλεί πολλές ακραίες παρατηρήσεις.

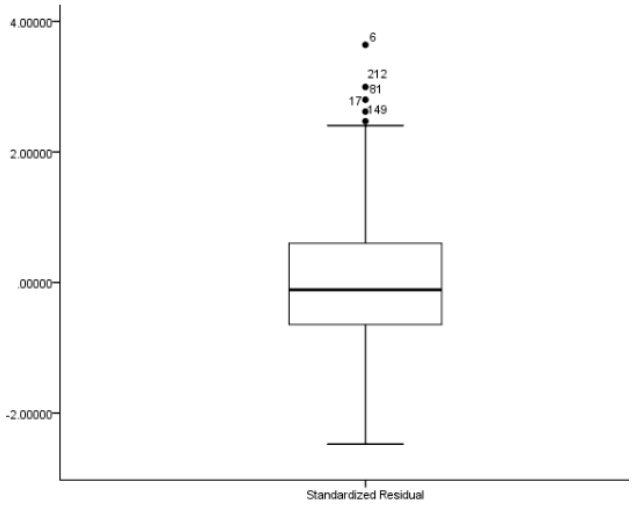


Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.069	391	.000	.988	391	.002

a. Lilliefors Significance Correction

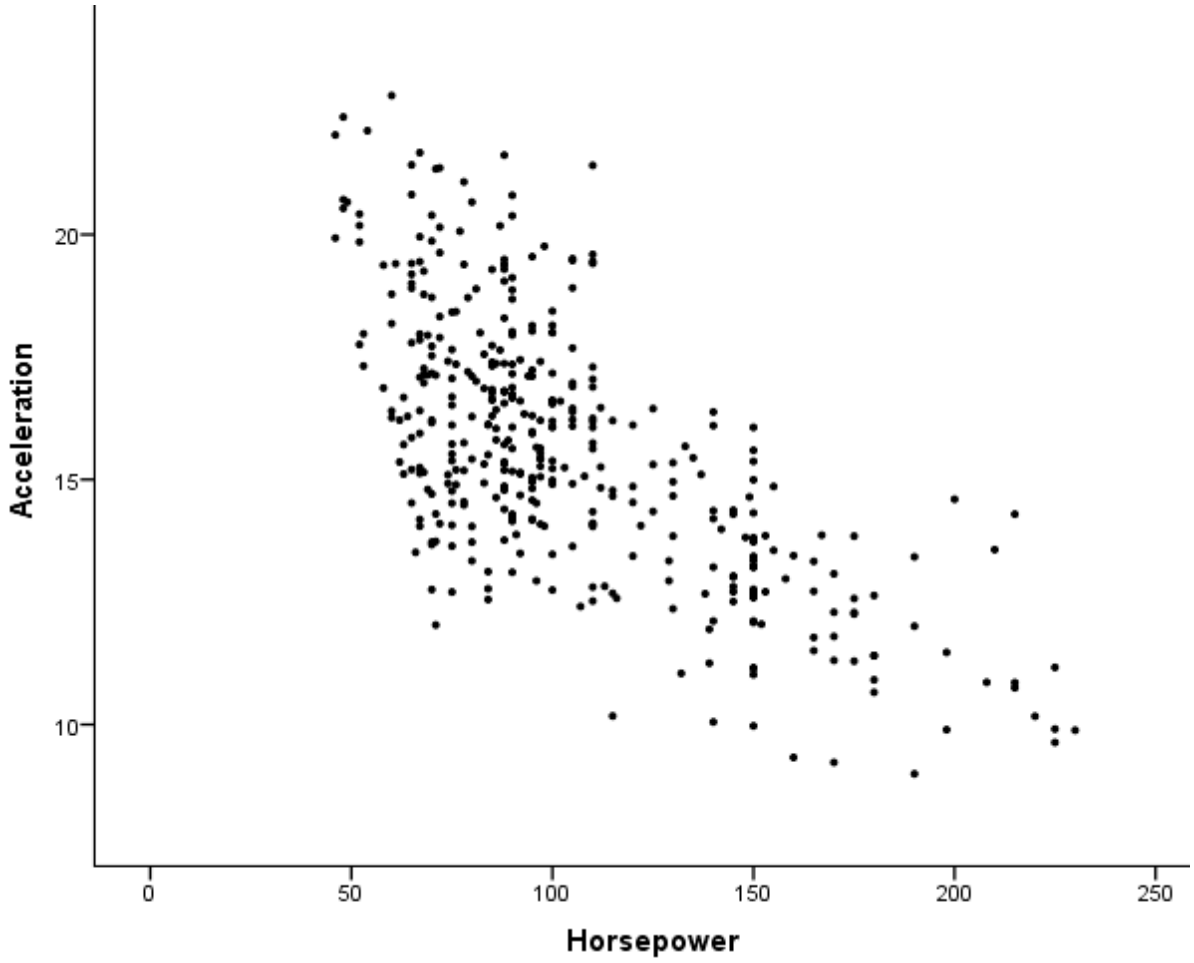




Δεν μπορούμε να σταματήσουμε εδώ την ανάλυσή μας, αφού η υπόθεση της κανονικότητας δεν πληρείται.

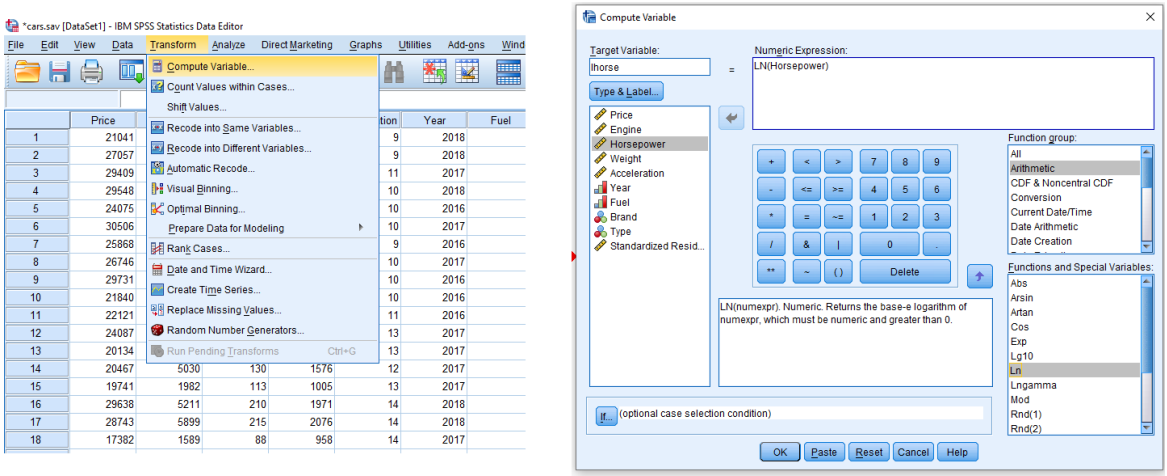
Μετασχηματισμός Μοντέλου

Ας ξανακοιτάξουμε το scatterplot της acceleration με την horsepower. Παρατηρούμε ότι η σχέση τους δεν είναι ακριβώς γραμμική. Το γράφημα σχηματίζει μία γωνία η οποία θυμίζει πιο πολύ κυρτή συνάρτηση παρά γραμμική.

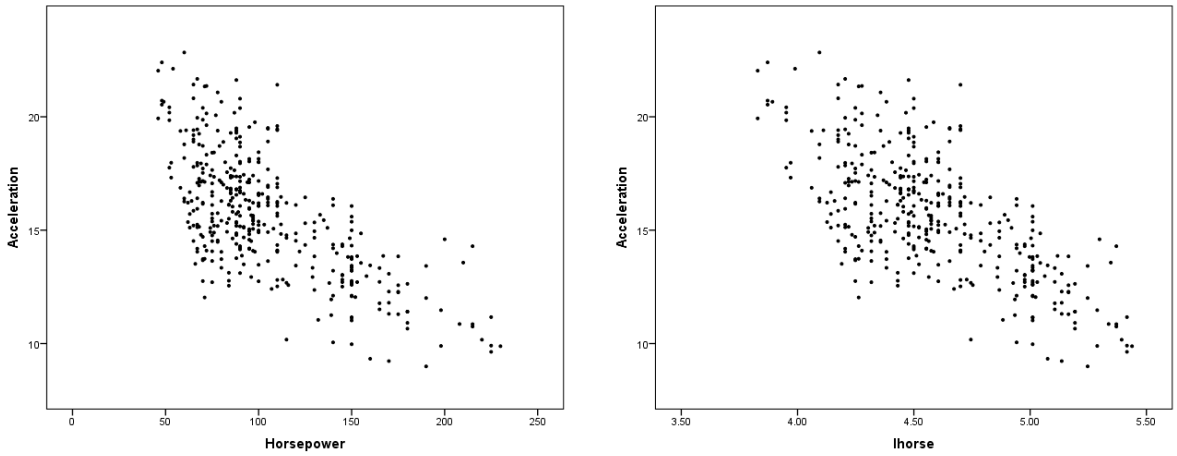


Θα επιχειρήσουμε να μετασχηματίσουμε τη μεταβλητή Horsepower ώστε να αναγκάσουμε τη σχέση να γίνει γραμμική. Αφού το γράφημα εμφανίζει μία κυρτή καμπύλη, επιλέγουμε μία κοίλη συνάρτηση για τον μετασχηματισμό μας. Δημιουργούμε λοιπόν την μεταβλητή lhorse=ln(Horsepower).

Στη γραμμή εντολών ακολουθούμε τη διαδρομή Transform → Compute Variable. Μπορούμε να ορίσουμε τη νέα μεταβλητή είτε χειροκίνητα, είτε από τις επιλογές στο παράθυρο που μας εμφανίζεται.



Συγκρίνοντας τώρα τα scatterplot των δύο μεταβλητών, βλέπουμε βελτίωση στη γραμμικότητα της σχέσης.



Είμαστε σε θέση να κατασκευάζουμε ξανά το γραμμικό μας μοντέλο, αυτή τη φορά με την lhorse στην θέση της Horsepower.

Παρακάτω βλέπουμε τον πίνακα των συντελεστών του μοντέλου. Η εκτιμημένη εξίσωση της πολλαπλής γραμμικής παλινδρόμησης είναι:

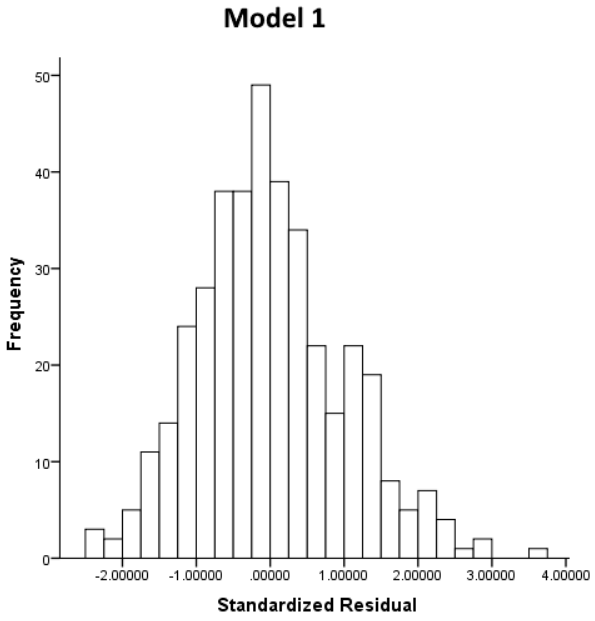
$$Y_i = 52.076 - 0.001X_{1i} - 9.771X_{2i} + 0.008X_{3i}, \quad i = 1, .., 391$$

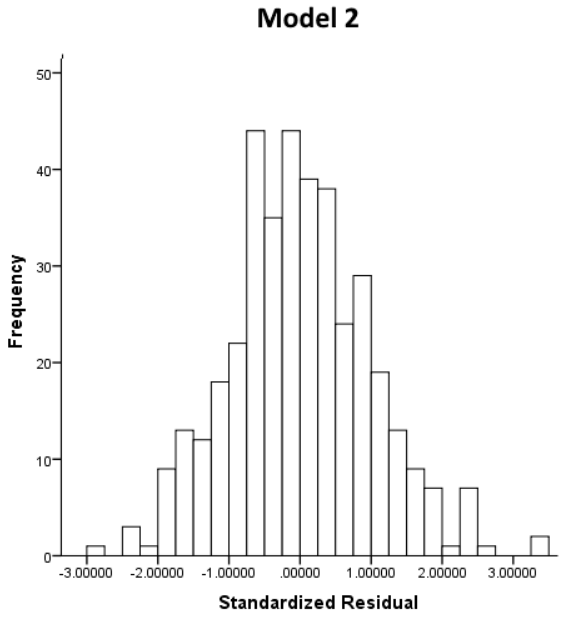
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	52.076	2.080		25.042	.000
	Engine	-.001	.000	-.512	-5.580	.000
	Ihorse	-9.771	.535	-1.240	-18.263	.000
	Weight	.008	.001	1.152	13.141	.000

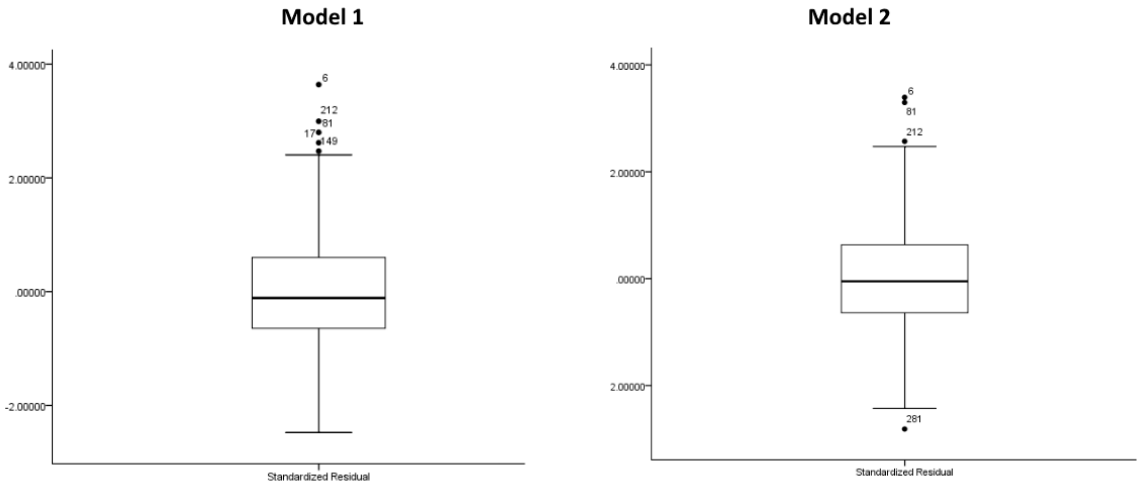
a. Dependent Variable: Acceleration

Ας συγκρίνουμε τα γραφήματα των καταλοίπων για το αρχικό μοντέλο και το μετασχηματισμένο. Στο ιστόγραμμα παρατηρούμε βελτίωση στην συμμετρία, αν και κάποιες ακραίες παρατηρήσεις φαίνεται να υπάρχουν ακόμα.

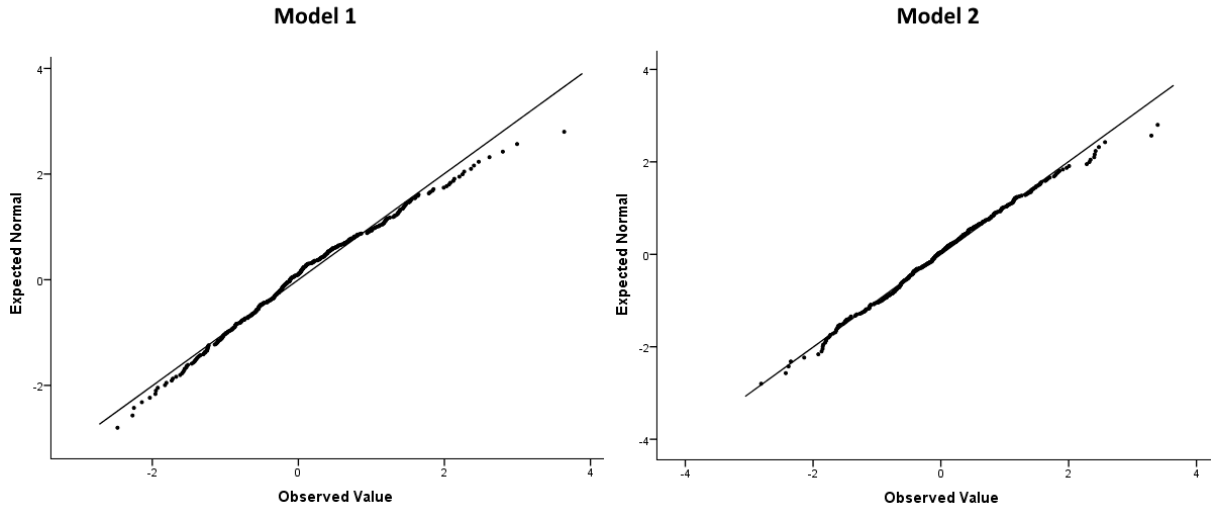




Στο boxplot βλέπουμε πως όντως κάποιες ακραίες παρατηρήσεις επιμένουν, αλλά είναι λιγότερες και επιπλέον εμφανίζονται και στις δύο ουρές (κάτι που υποδηλώνει συμμετρία).



Στο Q-Q Plot βλέπουμε εμφανή βελτίωση στην δεξιά ουρά.



Τέλος, κάνουμε τον έλεγχο κανονικότητας Shapiro-Wilk, ο οποίος μας δίνει p-value=0.18. Επομένως, η κανονικότητα δεν απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας 5%.

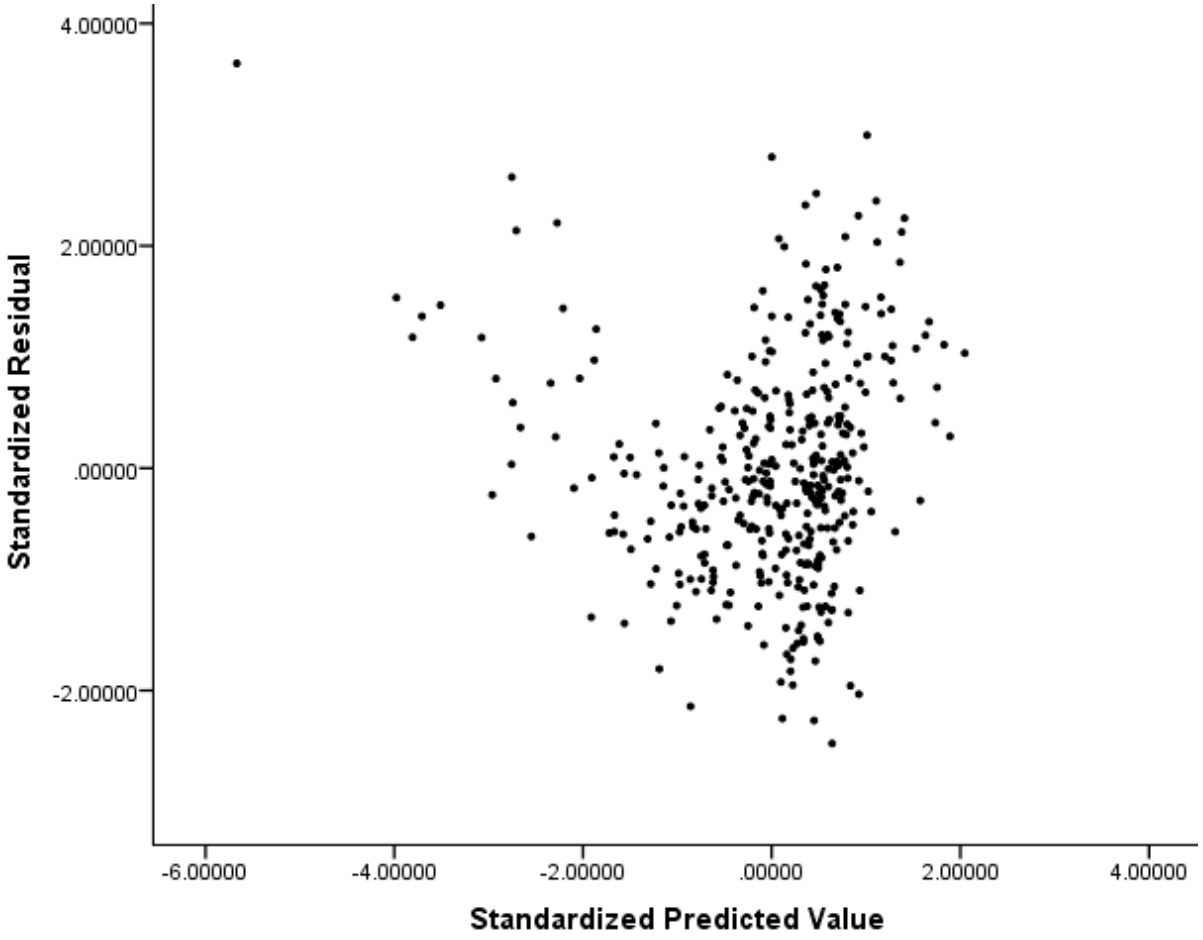
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.033	391	.200*	.995	391	.180

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Ομοσκεδαστικότητα και Ανεξαρτησία: Παρότι υπάρχουν στατιστικοί έλεγχοι για την ομοσκεδαστικότητα και την ανεξαρτησία των καταλοίπων, εδώ θα αρκεστούμε απλά σε ένα γραφικό έλεγχο. Συγκεκριμένα θα δημιουργήσουμε το σημειόγραμμα (scatterplot) των τυποποιημένων καταλοίπων (standardized residuals) ως προς τις τυποποιημένες προβλεπόμενες τιμές της y (standardized predicted values).



Παρατηρούμε πώς για τιμές μικρότερες εκατέρωθεν του x=1, τα κατάλοιπα παρουσιάζουν πολύ μεγάλες διαφορές. Για μικρές τιμές του x, τα κατάλοιπα εμφανίζουν μεγάλη συγκέντρωση στα θετικά και μία πολύ ακραία παρατήρηση. Επομένως τα κατάλοιπα δεν φαίνονται να είναι ανεξάρτητα. Επιπλέον, τα κατάλοιπα μέχρι το σημείο x=1 φαίνεται να έχουν μία φθίνουσα πορεία, επομένως παρουσιάζουν και ετεροσκεδαστικότητα. Τα παραπάνω υποδεικνύουν ότι πρέπει να δοκιμάσουμε εναλλακτικές μεθόδους μοντελοποίησης ώστε να έχουμε τόσο ανεξαρτησία, όσο και ομοσκεδαστικότητα των καταλοίπων.

Ασκήσεις

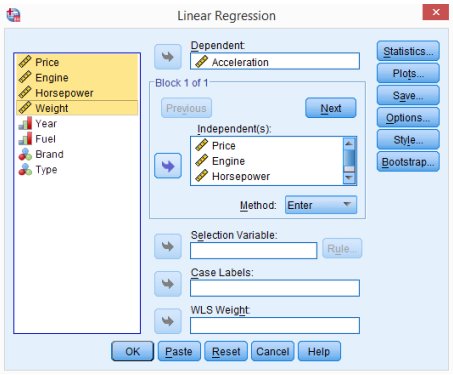
Εξελίξτε τα απλά γραμμικά μοντέλα για τα δεδομένα του αρχείου grades που κατασκευάσατε στην προηγούμενη ενότητα. Επιλέξτε ως εξαρτημένη μεταβλητή την Grades και σαν ανεξάρτητες τις Maths, Physics και Gymnastics.

6 Διαστήματα Εμπιστοσύνης & Έλεγχοι Υποθέσεων

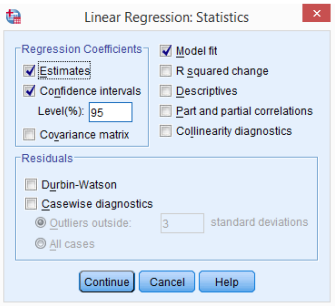
Σύνοψη: Σε αυτή την ενότητα θα εμβαθύνουμε στην κατασκευή διαστημάτων εμπιστοσύνης και τους ελέγχους υποθέσεων που μπορούμε να εκτελέσουμε για τους συντελεστές (β) ενός γραμμικού μοντέλου, καθώς και στις μελλοντικές προβλέψεις (y).

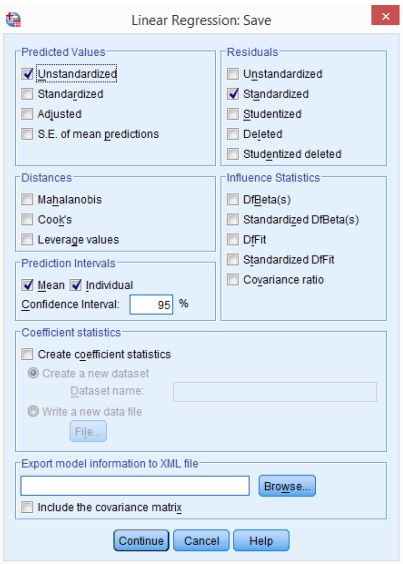
Διαστήματα Εμπιστοσύνης

Αρχικά, θα κατασκευάσουμε ξανά ένα γραμμικό μοντέλο. Ως εξαρτημένη μεταβλητή επιλέγουμε την acceleration, ενώ ως ανεξάρτητες τις price, engine, horsepower και weight.



Στο μενού Statistics επιλέγουμε επιπλέον το confidence intervals που προς το παρόν αφήνουμε στο 95% που έχει ως προεπιλογή το SPSS. Στο μενού Save επιλέγουμε Mean και Individual, ξανά με την προεπιλογή 95%.





Τρέχουμε το γραμμικό μοντέλο και παρατηρούμε τον πίνακα Coefficients. Μπορούμε να δούμε ότι το SPSS κατασκεύασε 95% διαστήματα εμπιστοσύνης για κάθε συντελεστή $\hat{\beta}_i$. Τα διαστήματα που κατασκευάζει το SPSS είναι συμμετρικά, αφήνουν δηλαδή ένα σφάλμα 2.5% προς κάθε κατεύθυνση. Μπορούμε να αλλάξουμε το επίπεδο σημαντικότητας ώστε να πάρουμε τα διαστήματα της αρεσκείας μας. Με τα διαστήματα αυτά μπορούμε να πάρουμε μία καλύτερη εικόνα για το "πού είναι πιθανό να βρίσκεται η πραγματική τιμή β_i".

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		
		B	Std. Error	Beta			Lower Bound	Upper Bound	
1	(Constant)	17,288	,761		22,707	,000	15,791	18,785	
	Price	-7,015E-6	,000	-,010	-,097	,922	,000	,000	(-0.000148 , 0.000134)
	Engine	-,001	,000	-,341	-3,273	,001	-,001	,000	(-0.000858 , -0.000214)
	Horsepower	-,082	,009	-1,158	-9,504	,000	-,098	-,065	
	Weight	,006	,001	,915	10,116	,000	,005	,008	

a. Dependent Variable: Acceleration

Σε αυτό το σημείο πρέπει να δώσουμε ιδιαίτερη προσοχή: Επαναλαμβάνοντας πολλές φορές το πείραμά μας (παίρνοντας διαφορετικά δείγματα και δημιουργώντας τα γραμμικά μοντέλα), θα παίρνουμε ελαφρώς διαφορετικά αποτελέσματα. Αυτό που αλλάζει από επανάληψη σε επανάληψη είναι η εκτίμησή μας $\hat{\beta}_i$, όχι το β_i το οποίο πραγματικό, σταθερό, αλλά άγνωστο σε εμάς. Επομένως είναι λάθος να πούμε "το 95% των φορών που επαναλαμβάνουμε το πείραμα το β_i θα πέφτει μέσα στο διάστημα εμπιστοσύνης" γιατί αυτό υποδηλώνει πώς το διάστημά μας είναι σταθερό ενώ το β_i αλλάζει τιμές, ενώ στην πραγματικότητα συμβαίνει το αντίθετο. Μια σωστή περιγραφή θα ήταν "το 95% των φορών που επαναλαμβάνουμε το πείραμα το διάστημά μας περιλαμβάνει το β_i".

Έλεγχοι Υποθέσεων

Πέρα από τα διαστήματα εμπιστοσύνης θα θέλαμε να εκτελέσουμε και κάποιους ελέγχους υποθέσεων για τα β_i. Αυτοί οι έλεγχοι μπορεί να είναι τριών διαφορετικών μορφών:

$$H_0: \ \beta_i = 0 \qquad H_0: \ \beta_i = 0 \qquad H_0: \ \beta_i = 0$$
$$H_1: \ \beta_i \neq 0 \qquad H_1: \ \beta_i > 0 \qquad H_1: \ \beta_i < 0$$

Στο SPSS δεν μπορούμε να εκτελέσουμε απ'ευθείας τους ελέγχους για τα β_i. Μπορούμε όμως να χρησιμοποιήσουμε την άλλη όψη του ίδιου νομίσματος, τα διαστήματα εμπιστοσύνης.

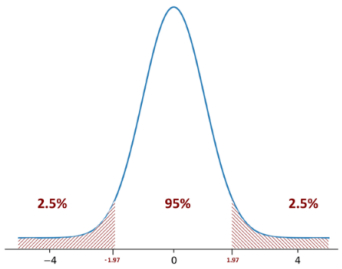
Έστω ότι θέλουμε να εκτελέσουμε έναν έλεγχο της πρώτης μορφής σε επίπεδο σημαντικότητας 5%. Αυτό ισοδυναμεί με το να κατασκευάσουμε ένα διάστημα εμπιστοσύνης όπως παραπάνω, όπου επιτρέπουμε

ένα σφάλμα 2.5% προς την κατεύθυνση $\beta_i > 0$ και ένα σφάλμα 2.5% προς την κατεύθυνση $\beta_i < 0$. Αν το 0 εμπεριέχεται στο διάστημά μας, τότε μένουμε στην H_0, αν όχι, τότε την απορρίπτουμε και δεχόμαστε την H_1.

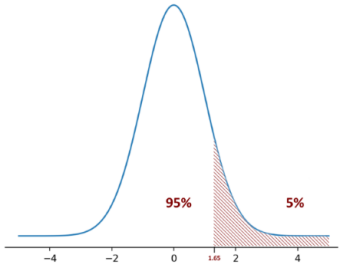
Αν θέλουμε να εκτελέσουμε έναν έλεγχο της δεύτερης μορφής σε επίπεδο σημαντικότητας 5%, τότε όλο το σφάλμα 5% συγκεντρώνεται στην κατεύθυνση $\beta_i > 0$. Για να κατασκευάσουμε το αντίστοιχο διάστημα εμπιστοσύνης, ζητάμε από το SPSS να κατασκευάσει 90% διαστήματα, ώστε το διάστημα (a, b) να αφήνει 5% σφάλμα προς κάθε κατεύθυνση. Το διάστημα που μας ενδιαφέρει τότε είναι το $(a, +\text{inf})$. Αν το 0 εμπεριέχεται στο διάστημά μας, τότε μένουμε στην H_0, αν όχι, τότε την απορρίπτουμε και δεχόμαστε την H_1. Αντίστοιχα για ελέγχους της τρίτης μορφής τα διαστήματά μας έχουν τη μορφή $(-\text{inf}, b)$.

Οι έλεγχοι της μορφής $\beta_i = 0$ μας δείχνουν την σημαντικότητα των μεταβλητών μας, αν δηλαδή έχει νόημα να έχουμε τη αντίστοιχη x_i στο μοντέλο μας. Αν θέλουμε να εκτελέσουμε ελέγχους της μορφής $\beta_i = c$ τότε ενεργούμε όπως παραπάνω και απλά ελέγχουμε αν το c εμπεριέχεται στα διαστήματα (αντί για το 0). Ας δούμε ένα βοηθητικό γράφημα και μερικά παραδείγματα:

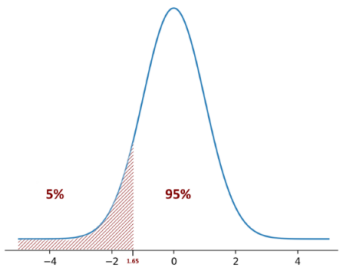
$H_O: \beta_i = 0$
$H_1: \beta_i \neq 0$



$H_O: \beta_i = 0$
$H_1: \beta_i > 0$



$H_O: \beta_i = 0$
$H_1: \beta_i < 0$



Παραδείγματα

Παράδειγμα 1: Για το μοντέλο που κατασκευάσαμε, θέλουμε να εκτελέσουμε τον παρακάτω δίπλευρο έλεγχο σε επίπεδο σημαντικότητας 5%:

$$H_0: \ \beta_{weight} = 0$$

$$H_1: \ \beta_{weight} \neq 0$$

Ζητάμε από το SPSS να μας κατασκευάσει 95% διαστήματα εμπιστοσύνης (όπως παραπάνω):

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		
		B	Std. Error	Beta			Lower Bound	Upper Bound	
1	(Constant)	17,288	,761		22,707	,000	15,791	18,785	
	Price	-7,015E-6	,000	-,010	-,097	,922	,000	,000	(-0.000148 , 0.000134)
	Engine	-,001	,000	-,341	-3,273	,001	-,001	,000	(-0.000858 , -0.000214)
	Horsepower	-,082	,009	-1,158	-9,504	,000	-,098	-,065	
	Weight	,006	,001	,915	10,116	,000	,005	,008	

a. Dependent Variable: Acceleration

Παρατηρούμε πως έχουμε $b_{weight} = 0.006$. Το διάστημα εμπιστοσύνης που κατασκευάσαμε είναι το $(0.005, 0.008)$. Βλέπουμε πως δεν περιέχει το 0, επομένως απορρίπτουμε την μηδενική υπόθεση.

Παράδειγμα 2: Για το μοντέλο που κατασκευάσαμε, θέλουμε να εκτελέσουμε τον παρακάτω μονόπλευρο έλεγχο σε επίπεδο σημαντικότητας 2.5%:

$$H_0: \ \beta_{price} = 0$$

$$H_1: \ \beta_{price} > 0$$

Το διάστημα εμπιστοσύνης που θέλουμε είναι της μορφής $(a, +\text{inf})$. Το SPSS μας κατασκευάζει μόνο συμμετρικά διαστήματα (a, b) , επομένως θα του ζητήσουμε να έχει 2.5% σφάλμα σε κάθε μεριά, δηλαδή να κατασκευάσει ένα 95% διάστημα εμπιστοσύνης (ξανά, όπως παραπάνω).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		
		B	Std. Error	Beta			Lower Bound	Upper Bound	
1	(Constant)	17,288	,761		22,707	,000	15,791	18,785	
	Price	-7,015E-6	,000	-,010	-,097	,922	,000	,000	(-0.000148 , 0.000134)
	Engine	-,001	,000	-,341	-3,273	,001	-,001	,000	(-0.000858 , -0.000214)
	Horsepower	-,082	,009	-1,158	-9,504	,000	-,098	-,065	
	Weight	,006	,001	,915	10,116	,000	,005	,008	

a. Dependent Variable: Acceleration

Παρατηρούμε πως έχουμε $b_{price} = -0.000007015$. Το διάστημα που παίρνουμε είναι $(-0.000148, +\text{inf})$. Βλέπουμε πως περιέχει το 0, επομένως δεν απορρίπτουμε την μηδενική υπόθεση. Αυτό μας δείχνει πως

η μεταβλητή price πρέπει να βγει από το μοντέλο.

Παράδειγμα 3: Για το μοντέλο που κατασκευάσαμε, θέλουμε να εκτελέσουμε τον παρακάτω μονόπλευρο έλεγχο σε επίπεδο σημαντικότητας 5%:

$$H_0: \ \beta_0 = 16$$

$$H_1: \ \beta_0 < 16$$

Το διάστημα εμπιστοσύνης που θέλουμε είναι της μορφής $(-\text{inf}, b)$. Το SPSS μας κατασκευάζει μόνο συμμετρικά διαστήματα (a, b) , επομένως θα του ζητήσουμε να έχει 5% σφάλμα σε κάθε μεριά, δηλαδή να κατασκευάσει ένα 90% διάστημα εμπιστοσύνης.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	90,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	17,288	,761		22,707	,000	16,033	18,543
	Price	-7,015E-6	,000	-,010	-,097	,922	,000	,000
	Engine	-,001	,000	-,341	-3,273	,001	-,001	,000
	Horsepower	-,082	,009	-1,158	-9,504	,000	-,096	-,067
	Weight	,006	,001	,915	10,116	,000	,005	,008

a. Dependent Variable: Acceleration

Παρατηρούμε πως έχουμε $\beta_0 = 17.288$. Το διάστημα που παίρνουμε είναι $(-\text{inf}, 18.543)$. Βλέπουμε πως περιέχει το 16, επομένως δεν απορρίπτουμε την μηδενική υπόθεση. Αν στη θέση του 16 είχαμε το 19, θα απορρίπταμε την μηδενική υπόθεση.

Διαστήματα Πρόβλεψης για την Εξαρτημένη Μεταβλητή

Έχοντας δει τα αυτοκίνητα του σετ δεδομένων μας, θα θέλαμε να προβλέψουμε την επιτάχυνση ενός αυτοκινήτου γνωρίζοντας μόνο τα άλλα χαρακτηριστικά του. Θα θέλαμε δηλαδή να κατασκευάσουμε διαστήματα εμπιστοσύνης και για την εξαρτημένη μεταβλητή ενός μοντέλου. Υπάρχουν δύο είδη τέτοιων διαστημάτων. Ας τα γνωρίσουμε μέσα από δύο σενάρια:

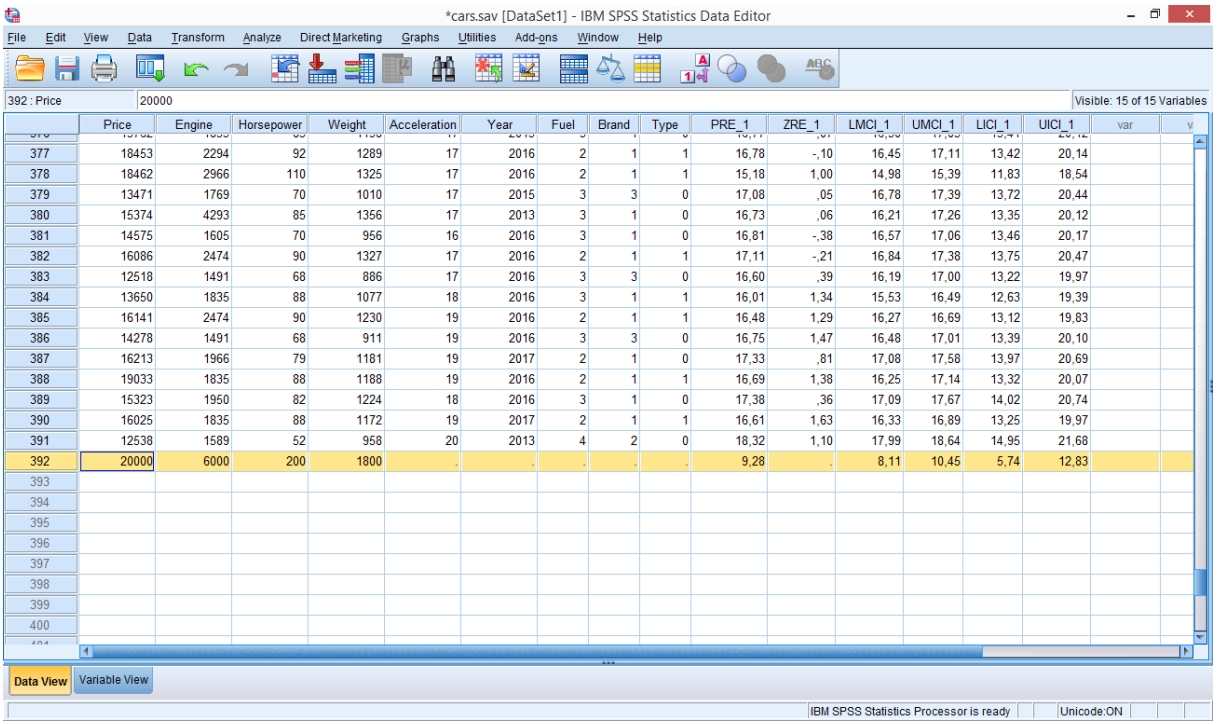
Σενάριο 1: Βρισκόμαστε σε μία έκθεση αυτοκινήτων, όπου μπορούμε να δούμε και να μελετήσουμε τα 391 αυτοκίνητα. Γνωρίζουμε πως σύντομα θα καταφτάσει στην έκθεση ένα ακόμα αυτοκίνητο, το No. 392. Οι οργανωτές της έκθεσης μας πληροφορούν για τα χαρακτηριστικά του ($price = 20000, engine = 6000, horsepower = 200, weight = 1800$) αλλά όχι για την επιτάχυνσή του ($acceleration$). Θα θέλαμε να προβλέψουμε την επιτάχυνση του συγκεκριμένου αυτοκινήτου και να δημιουργήσουμε ένα διάστημα πρόβλεψης για αυτήν.

Σενάριο 2: Βρισκόμαστε σε μία έκθεση αυτοκινήτων, όπου μπορούμε να δούμε και να μελετήσουμε τα 391 αυτοκίνητα. Σύντομα αναχωρούμε για το εργοστάσιο παραγωγής αυτοκινήτων με συγκεκριμένα χαρακτηριστικά τα οποία γνωρίζουμε ($price = 20000, engine = 6000, horsepower = 200, weight = 1800$). Θα θέλαμε να προβλέψουμε τη μέση επιτάχυνσή τους και να δημιουργήσουμε ένα διάστημα πρόβλεψης για αυτήν.

Παρότι τα δύο παραπάνω σενάρια δείχνουν ίδια με μια πρώτη ανάγνωση, αν τα εξετάσουμε αναλυτικά, θα δούμε πως διαφέρουν. Πράγματι, η σημειακή πρόβλεψη που θα κάνουμε για την επιτάχυνση θα ταυτίζεται στα δύο σενάρια, αφού το μόνο που κάνουμε είναι να αντικαταστήσουμε τα συγκεκριμένα x που μας δίνονται στο γραμμικό μοντέλο και να υπολογίσουμε την y. Τα διαστήματα που θα κατασκευάσουμε όμως διαφέρουν. Όταν το αυτοκίνητο No. 392 φτάσει στην έκθεση, είμαστε αρκετά αβέβαιοι για το πόσο κοντά θα πέσουμε στην πρόβλεψή μας, ενώ όταν πάμε στο εργοστάσιο παραγωγής γνωρίζουμε πως η μέση επιτάχυνση όλων των αυτοκινήτων και η πρόβλεψή μας θα είναι σχετικά κοντά. Τα δύο διαστήματα δηλαδή κατασκευάζονται χρησιμοποιώντας διαφορετική διασπορά.

Για να κατασκευάσουμε τα εν λόγω διαστήματα, δημιουργούμε μία ακόμα σειρά στο σετ δεδομένων μας στην οποία βάζουμε τα χαρακτηριστικά που θέλουμε, αφήνοντας την επιτάχυνση κενή. Με αυτόν τον τρόπο, το SPSS δεν χρησιμοποιεί την παρατήρηση αυτή στο γραμμικό μοντέλο. Στην παλινδρόμηση,

επιλέγουμε Mean και Individual Prediction Intervals όπως δείξαμε στην αρχή. Το SPSS δημιουργεί καινούριες μεταβλητές που δείχνουν το κάτω και το άνω άκρο των προβλέψεων μας, τόσο για αυτοκίνητα με χαρακτηριστικά σαν τα 391 που είχαμε στο σετ δεδομένων μας εξ αρχής, όσο και για την τελευταία παρατήρηση που βάλαμε.



Βλέπουμε πώς η σημειακή πρόβλεψη είναι 9.28 για την επιτάχυνση (και για τα δύο σενάρια). Το Mean Confidence Interval (MCI) που προβλέπουμε για το εργοστάσιο είναι το $(8.11, 10.45)$ ενώ το Individual Confidence Interval (ICI) που προβλέπουμε για το αυτοκίνητο που θα φτάσει στην έκθεση είναι $(5.74, 12.83)$, αρκετά μεγαλύτερο του MCI.

7 ANOVA & t-test

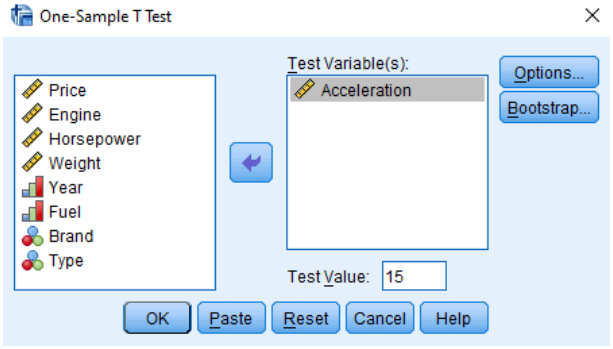
Σύνοψη: Σε αυτή την ενότητα θα εμβαθύνουμε σε ελέγχους για τον μέσο μίας μεταβλητής και τους μέσους διαφορετικών υποκατηγοριών μιας μεταβλητής. Συγκεκριμένα θα ασχοληθούμε με τα t-test και τα μοντέλα ANOVA.

One Sample t-test

Ας υποθέσουμε πως θέλουμε να κάνουμε έναν έλεγχο υποθέσεων για την μέση επιτάχυνση των αυτοκινήτων. Αν τρέξουμε κάποια περιγραφικά στατιστικά, θα δούμε πώς η μέση επιτάχυνση των αυτοκινήτων των δεδομένων μας είναι 15.62 δευτερόλεπτα. Θέλουμε λοιπόν να εφαρμόσουμε τον παρακάτω έλεγχο:

$$H_0 : \ \mu_{accel} = 15$$
$$H_1 : \ \mu_{accel} \neq 15$$

Ο κατάλληλος έλεγχος για να το κάνουμε αυτό ονομάζεται t-test και συγκεκριμένα one-sample t-test (θα δούμε και άλλα είδη t-test παρακάτω). Απαραίτητη προϋπόθεση για το one sample t-test είναι η μεταβλητή να προέρχεται από κανονική κατανομή, κάτι που έχουμε ήδη ελέγξει για την Acceleration σε προηγούμενη ενότητα. Για να εκτελέσουμε τον έλεγχο στο SPSS, ακολουθούμε τη διαδρομή Analyze → Compare Means → One-Sample T Test και συμπληρώνουμε το παράθυρο που εμφανίζεται όπως παρακάτω:



Τρέχουμε το t-test και το SPSS μας εμφανίζει δύο πίνακες, έναν με περιγραφικά στατιστικά και έναν όπου εκτελείται ο έλεγχος. Παίρνουμε $p - value = 0.000007 < 0.05$ επομένως απορρίπτουμε την μηδενική υπόθεση σε επίπεδο στατιστικής σημαντικότητας 5%. Ακόμα, μας δίνεται η διαφορά ανάμεσα στο δειγματικό μέσο και την τιμή που θέσαμε ως πραγματικό μέσο (0.62) και ένα 95% διάστημα εμπιστοσύνης για αυτή τη διαφορά $(0.35, 0.89)$.

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Acceleration	391	15.62	2.697	.136

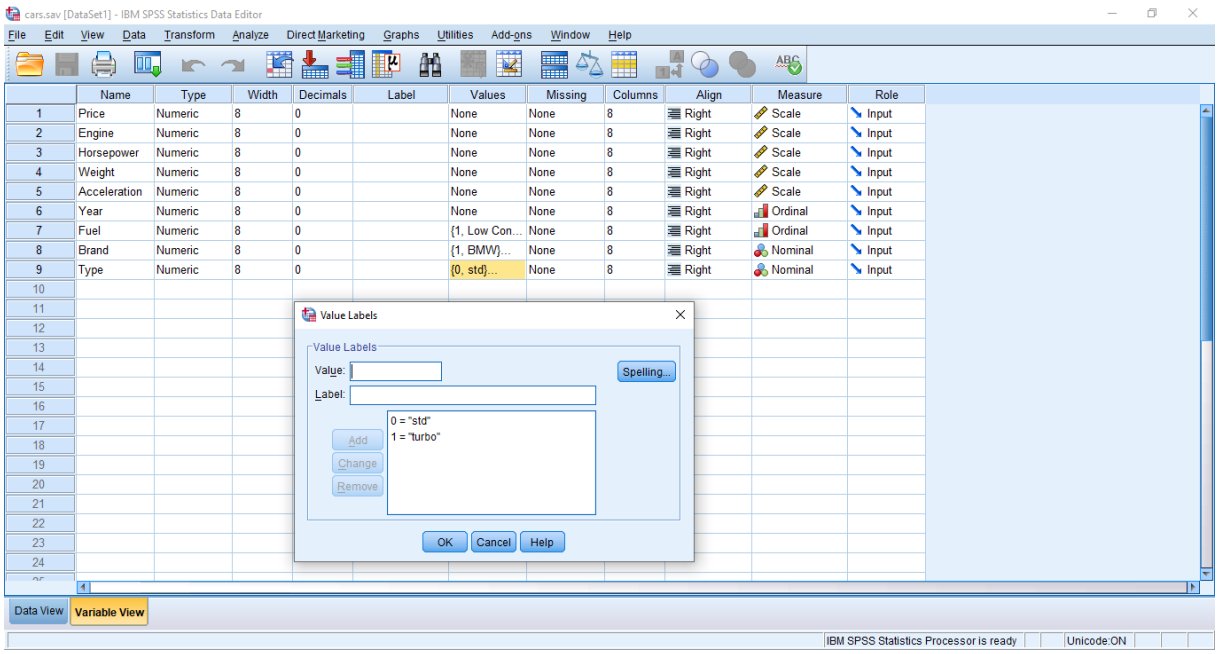
One-Sample Test

	Test Value = 15					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Acceleration	4.553	390	.000	.621	.35	.89

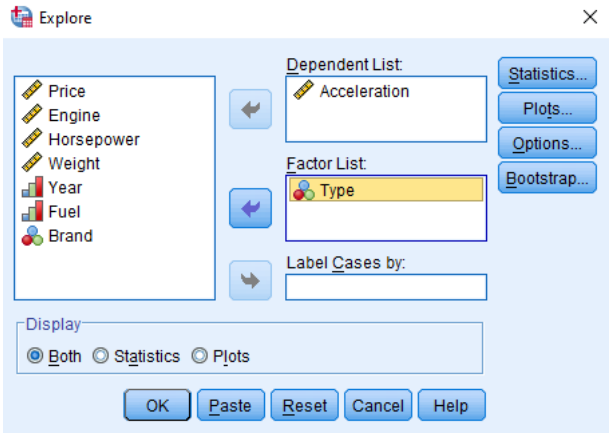
Με το παραπάνω t-test κάνουμε έλεγχο για τον μέσο μιας μεταβλητής ως προς μια συγκεκριμένη τιμή που θέτουμε εμείς. Προχωράμε σε ένα t-test που αφορά την σύγκριση των μέσων δύο υποκατηγοριών της μεταβλητής.

Independent Sample t-test

Τα αυτοκίνητα στο σετ δεδομένων μας έχουν δύο είδη κινητήρα, std και turbo. Αυτό μπορούμε να το δούμε από την μεταβλητή Type, η οποία έχει κωδικοποιημένα τα δύο είδη με 0 και 1 για std και turbo αντίστοιχα. Στο SPSS, πάμε στην καρτέλα Variable View, στην στήλη Values, όπως φαίνεται παρακάτω:

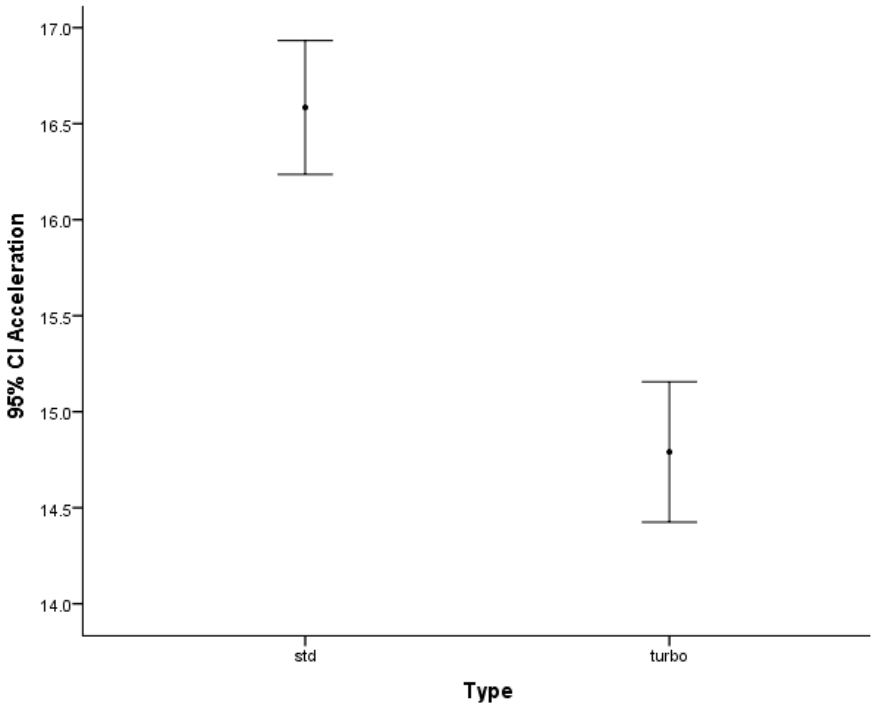


Ας υποθέσουμε λοιπόν ότι θέλουμε να συγκρίνουμε τα δύο είδη αυτοκινήτων και συγκεκριμένα την επιτάχυνσή τους. Σημαντική προϋπόθεση για αυτό το t-test είναι οι δύο υποκατηγορίες να προέρχονται από κανονική κατανομή. Πρέπει λοιπόν να κάνουμε έλεγχο κανονικότητας για τα std και turbo ξεχωριστά. Για να "σπάσουμε" το δείγμα μας σε κομμάτια σύμφωνα με την μεταβλητή Type, στο μενού Explore που ανοίγουμε για τους ελέγχους κανονικότητας, βάζουμε την Type στο κουτί Factor List, όπως φαίνεται παρακάτω:



Από τον έλεγχο προκύπτει πως δεχόμαστε την κανονικότητα για την κατηγορία turbo, αλλά όχι για την std. Αυτό σημαίνει πώς το t-test που θα κάνουμε θα ισχύει προσεγγιστικά (και δεν θα μπορούσα να το εμπιστευτώ για μικρά δείγματα).

Αρχικά, θα κατασκευάσουμε ένα error bar για να έχουμε μια καλύτερη εικόνα των δύο κατηγοριών. Στη γραμμή εντολών ακολουθούμε τη διαδρομή Graphs → Legacy Dialogs → Error Bar και επιλέγουμε το Simple. Στο μενού που εμφανίζεται επιλέγουμε ως Variable την Acceleration και ως Category Axis την Type.
Το SPSS κατασκευάζει το γράφημα που του ζητήσαμε. Παρατηρούμε πως ο άξονας x έχει τις δύο κατηγορίες της μεταβλητής Type. Για κάθε μία από τις δύο κατηγορίες, μπορούμε να δούμε τον δειγματικό μέσο του acceleration (μαύρη κουκκίδα) και ένα 95% διάστημα γύρω από αυτόν με πλάτος ανάλογο με τη δειγματική τυπική απόκλιση (γραμμές και φράχτες). Ο μέσος για τα αυτοκίνητα τύπου std βρίσκεται πολύ υψηλότερα από τον μέσο των αυτοκινήτων turbo. Αυτό σημαίνει πώς τα std χρειάζονται περισσότερο χρόνο για να επιταχύνουν (είναι δηλαδή πιο αργά). Αν έπρεπε να βγάλουμε ένα συμπέρασμα μόνο από το γράφημα, θα λέγαμε πως αυτή διαφορά δεν είναι ούτε μικρή ούτε τυχαία, αφού πέρα από τους μέσους έχουμε κατασκευάσει και τα διαστήματα εμπιστοσύνης τα οποία δεν τέμνονται καν.



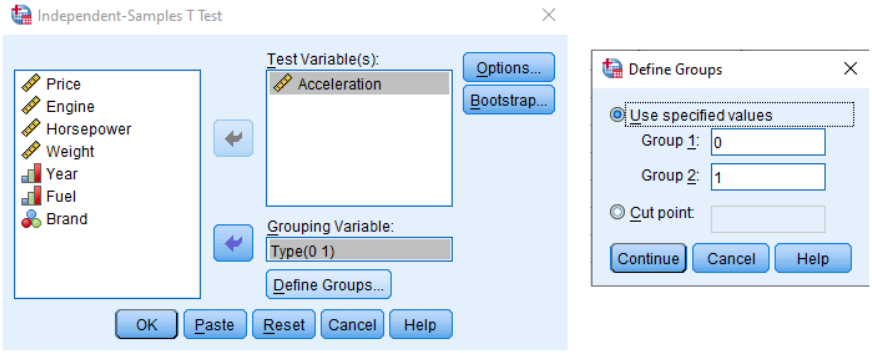
Προσοχή: Το box plot (που κατασκευάσαμε σε προηγούμενα εργαστήρια) και το error bar είναι διαφορετικά γραφήματα. Το box plot έχει σαν κέντρο την διάμεσο και διάστημα βασισμένο σε ποσοστημόρια, ενώ το error bar έχει σαν κέντρο τον δειγματικό μέσο και διάστημα βασισμένο στην τυπική απόκλιση. Το box plot δηλαδή είναι ανεξάρτητο κατανομής, ενώ το error bar στηρίζεται στην κανονική κατανομή της μεταβλητής y (εδώ η Acceleration).

Θέλουμε να ελέγξουμε όσα παρατηρήσαμε στο γράφημα με έναν στατιστικό έλεγχο. Θέλουμε δηλαδή να ελέγξουμε αν οι δειγματικοί μέσοι των δύο κατηγοριών είναι ίσοι ή οχι. Αυτός ο έλεγχος θα μπορούσε να περιγραφεί ως:

$$H_0 : \ \mu_{std} = \mu_{turbo}$$
$$H_1 : \ \mu_{std} \neq \mu_{turbo}$$

Ο έλεγχος αυτός ονομάζεται independent sample t-test. Το independent sample δηλώνει πως τα δύο υποσετ αυτοκινήτων είναι ανεξάρτητα μεταξύ τους, αφού δεν έχουμε λόγο να υποστηρίξουμε πως ένα αυτοκίνητο (std) επηρεάζει κάπως ένα άλλο αυτοκίνητο (turbo). Αν για παράδειγμα εξετάζαμε τους βαθμούς ενός μαθητή στα μαθηματικά και τη φυσική, δε θα μπορούσαμε να υποστηρίξουμε πώς είναι ανεξάρτητοι, αφού καλή γνώση μαθηματικών γενικά ενισχύει την επίδοση στη φυσική.

Για να εκτελέσουμε τον έλεγχο στο SPSS, ακολουθούμε τη διαδρομή Analyze → Compare Means → Independent-Samples T Test. Στο μενού που εμφανίζεται τοποθετούμε τις μεταβλητές Acceleration και Type όπως παρακάτω. Για την μεταβλητή Type, πρέπει να δώσουμε στο SPSS τις τιμές 0 και 1 τις οποίες θέλουμε να συγκρίνουμε. Αυτό πρέπει να το κάνουμε γιατί θα μπορούσε η Type να είχε πολλές κατηγορίες, ενώ το t-test μπορεί να συγκρίνει μόλις δύο.



Τρέχουμε το t-test και το SPSS μας εμφανίζει δύο πίνακες. Ο πρώτος έχει περιγραφικά στατιστικά για τις δύο κατηγορίες. Μπορούμε να δούμε ότι έχουμε 181 αυτοκίνητα με κινητήρα std και 210 με turbo. Ο δεύτερος πίνακας εκτελεί το t-test. Βασική προϋπόθεση είναι να γνωρίζουμε αν οι δύο κατηγορίες δείχνουν να έχουν ίδια διασπορά ή όχι. Για να το ελέγξουμε αυτό, το SPSS τρέχει το Levene's Test (πρώτες δύο στήλες του πίνακα). Μπορούμε να δούμε ότι έχουμε $p - value = 0.137 > 0.05$ επομένως μένουμε στην αρχική υπόθεση ότι οι διασπορές είναι ίσες. Επομένως θα κοιτάξουμε την πρώτη γραμμή του πίνακα (Equal Variances Assumed). Αν είχαμε απορρίψει την ισότητα, θα κοιτούσαμε την δεύτερη γραμμή (αν συγκρίνουμε τις δύο γραμμές, μπορούμε να δούμε πως τα δύο t-test είναι ελαφρώς διαφορετικά).

Βλέπουμε ότι το t-test δίνει $p - value = 1.6 \cdot 10^{-11}$ επομένως απορρίπτουμε την ισότητα των δύο μέσων, όπως αναμέναμε. Επιπλέον, μας δίνει την διαφορά των μέσων (1.794), καθώς και ένα 95% διάστημα εμπιστοσύνης για αυτή την διαφορά (από 1.286 έως 2.302).

T-Test

Group Statistics

	Type	N	Mean	Std. Deviation	Std. Error Mean
Acceleration	std	181	16.58	2.375	.177
	turbo	210	14.79	2.687	.185

Independent Samples Test

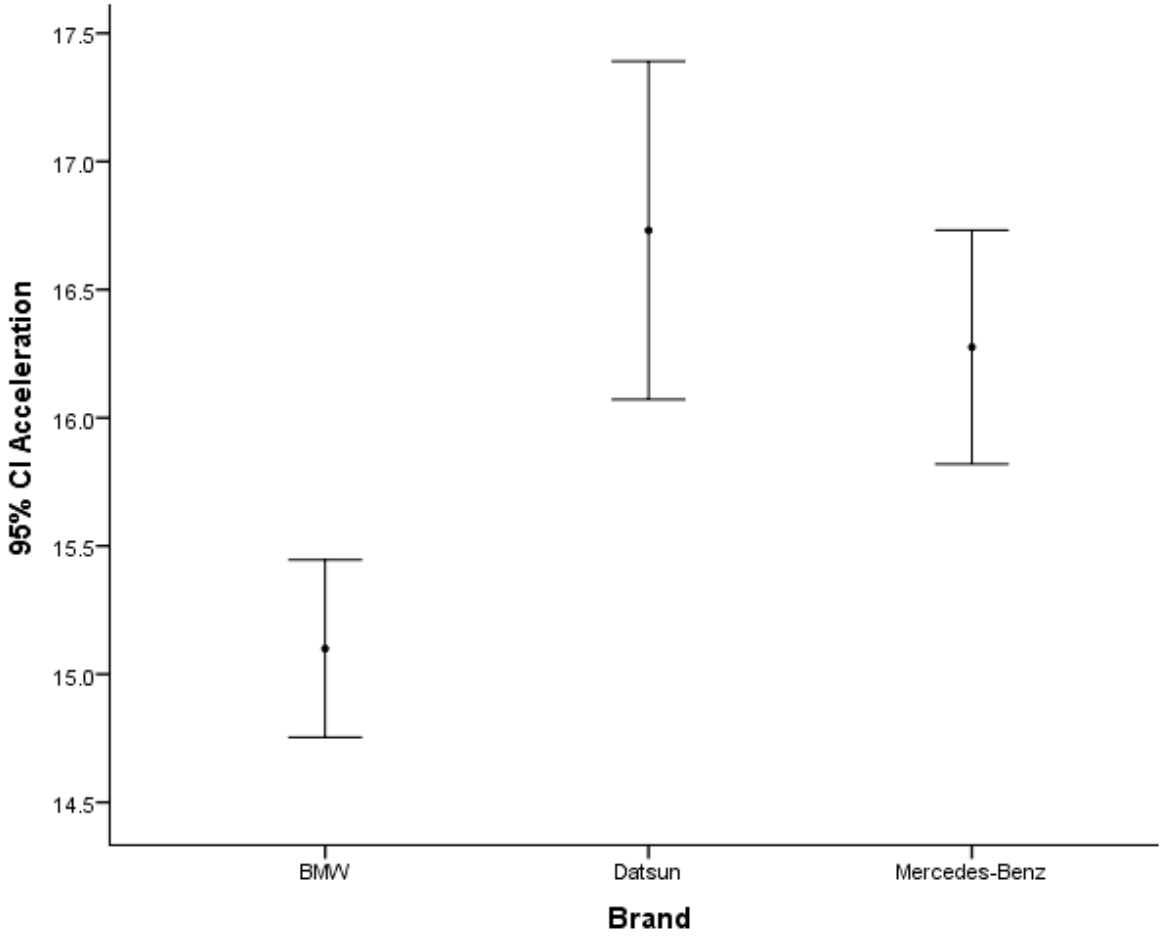
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Acceleration	Equal variances assumed	2.216	.137	6.945	389	.000	1.794	.258	1.286	2.302
	Equal variances not assumed			7.008	388.745	.000	1.794	.256	1.291	2.297

Προσοχή: Είναι σημαντικό να καταλάβουμε πως για τον παραπάνω έλεγχο χρησιμοποιήσαμε την ανεξαρτησία των δύο κατηγοριών. Αν είχαμε εξαρτημένες παρατηρήσεις, ο κατάλληλος έλεγχος θα ήταν το Paired-Samples T test.

One-Way ANOVA

Το t-test μπορεί να συγκρίνει μόνο δύο υποκατηγορίες μίας μεταβλητής. Αν θέλουμε να συγκρίνουμε την επιτάχυνση με βάση την μάρκα των αυτοκίνητων, για την οποία έχουμε 3 κατηγορίες (BMW, Datsun, Mercedes-Benz) πρέπει να χρησιμοποιήσουμε μία γενίκευση του t-test, την One-Way ANOVA (κατά ένα παράγοντα). Όπως και με το t-test, έτσι και με την ANOVA, προϋπόθεση αποτελεί η κανονικότητα των υποκατηγοριών. Στην πραγματικότητα, η ANOVA είναι ένα γραμμικό μοντέλο με τις ποιοτικές ανεξάρτητες μεταβλητές. Επομένως, μπορούμε να εκτελέσουμε έλεγχο κανονικότητας στα κατάλοιπα.

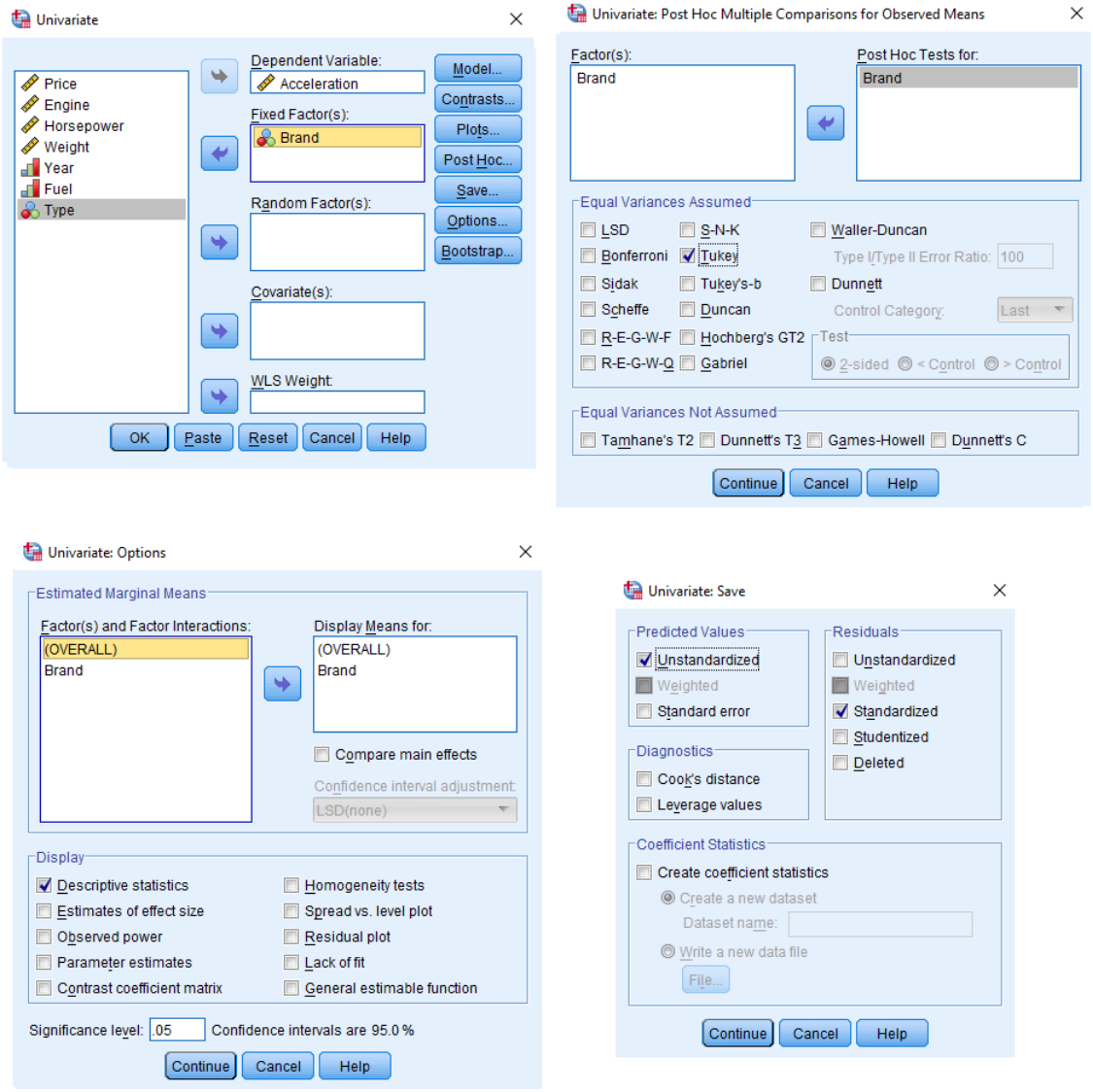
Αρχικά, θα κατασκευάσουμε ένα error bar όπως παραπάνω.



Μπορούμε να δούμε πώς τα αυτοκίνητα μάρκας BMW επιταχύνουν πολύ πιο γρήγορα από τις άλλες δύο μάρκες. Επομένως αναμένουμε πως ο μέσος των BMW θα διαφέρει στατιστικά σημαντικά από τους άλλους δύο μέσους. Για την σύγκριση Datsun και Mercedes-Benz, δεν μπορούμε να είμαστε σίγουροι,

αφού τα δύο διαστήματα εμφανίζουν μεγάλη αλληλεπικάλυψη, με τον μέσο της μίας μάρκας να βρίσκεται μέσα στο διάστημα της άλλης. Επιθυμούμε λοιπόν να κατασκευάσουμε το μοντέλο ANOVA για να δούμε αν ισχύουν όσα παρατηρήσαμε.

Για να τρέξουμε ένα μοντέλο ANOVA στο SPSS, ακολουθούμε τη διαδρομή Analyze → General Linear Model → Univariate και μιμούμαστε τις επιλογές που φαίνονται παρακάτω:



Παρατήρηση: Για την One-Way ANOVA θα μπορούσαμε να ακολουθήσουμε την διαδρομή Analyze → Compare Means → One-Way ANOVA. Αυτό το μενού όμως δεν μας επιτρέπει να αποθηκεύσουμε τα κατάλοιπα, τα οποία χρειαζόμαστε για τον έλεγχο κανονικότητας.

Το SPSS μας εμφανίζει δύο πινακάκια με γενικά περιγραφικά στατιστικά για την επιτάχυνση κάθε μάρκας ξεχωριστά. Ο πίνακας που μας ενδιαφέρει είναι ο τρίτος, (Tests of Between-Subjects Effects):

Tests of Between-Subjects Effects

Dependent Variable: Acceleration

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	183.984[a]	2	91.992	13.456	.000
Intercept	73555.299	1	73555.299	10759.026	.000
Brand	183.984	2	91.992	13.456	.000
Error	2652.606	388	6.837		
Total	98246.722	391			
Corrected Total	2836.590	390			

a. R Squared = .065 (Adjusted R Squared = .060)

Από την πρώτη γραμμή του πίνακα βλέπουμε πως το μοντέλο είναι στατιστικά σημαντικό ($p - value = 2 \cdot 10^{-7}$), δηλαδή το μοντέλο είναι στατιστικά σημαντικό. Κοιτώντας την γραμμή Brand μπορούμε να δούμε ότι έχει το ίδιο p-value και επομένως η μέση επιτάχυνση αυτοκινήτων διαφορετικής μάρκας διαφέρουν (χωρίς να ξέρουμε λεπτομέρειες). Στην υποσημείωση (a) βλέπουμε πως το μοντέλο έχει $R^2 = 0.065, R^2_{adj} = 0.060$, δηλαδή αρκετά ασθενές.

Το SPSS έπειται εμφανίζει δύο ακόμα πίνακες με τον δειγματικό μέσο και διαστήματα εμπιστοσύνης για την επιτάχυνση κάθε κατηγορίας. Αυτοί οι πίνακες έχουν την ίδια λειτουργία με τα error bar. Ο δεύτερος πίνακας που μας ενδιαφέρει είναι ο ”Multiple Comparisons”, που συγκρίνει τις κατηγορίες ανά δύο, ώστε να εντοπίσουμε αυτές που διαφέρουν μεταξύ τους. Από τους αστερίσκους του πίνακα μπορούμε να βρούμε τις στατιστικά σημαντικές διαφορές. Εδώ παρατηρούμε ότι η BMW διαφέρει και από τις δύο άλλες μάρκες, οι οποίες μεταξύ τους δεν διαφέρουν, όπως ακριβώς αναμέναμε.

Multiple Comparisons

Dependent Variable: Acceleration

Tukey HSD

(I) Brand	(J) Brand	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
BMW	Datsun	-1.632*	.359	.000	-2.48	-.79
	Mercedes-Benz	-1.176*	.338	.002	-1.97	-.38
Datsun	BMW	1.632*	.359	.000	.79	2.48
	Mercedes-Benz	.456	.433	.543	-.56	1.47
Mercedes-Benz	BMW	1.176*	.338	.002	.38	1.97
	Datsun	-.456	.433	.543	-1.47	.56

*. The mean difference is significant at the 0.05 level.

Ο πίνακας Homogeneous Subsets είναι προέκταση του προηγούμενου πίνακα και χωρίζει τις κατηγορίες με διαφορετικούς μέσους. Εδώ βλέπουμε πως έχουμε σε μία ομάδα την BMW (μόνη της) και σε μία άλλη τις Mercedes-Benz και Datsun.

Homogeneous Subsets

Acceleration

Tukey HSD[a,b]

Brand	N	Subset for alpha = 0.05	
		1	2
BMW	244	15.10	
Mercedes-Benz	79		16.28
Datsun	68		16.73
Sig.		1.000	.452

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 95.352.

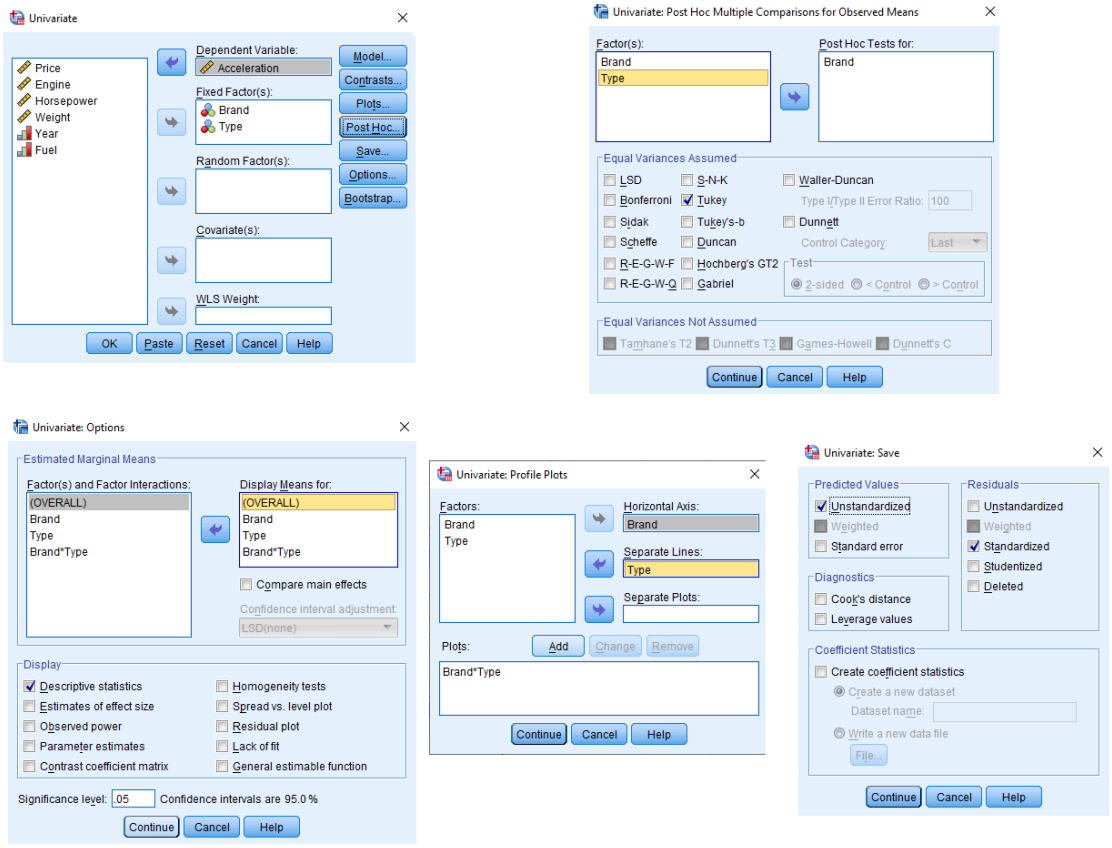
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Τέλος, πρέπει να κάνουμε τον έλεγχο κανονικότητας των καταλοίπων (κατά τα γνωστά). Ο έλεγχος Shapiro-Wilk βγάζει $p - value = 0.04$. Επομένως, η κανονικότητα των καταλοίπων απορρίπτεται σε επίπεδο σημαντικότητας 5%, αλλά όχι σε επίπεδο 1%. Σε κάθε περίπτωση, τα αποτελέσματά μας ισχύουν προσεγγιστικά, λόγω του μεγάλου μέγεθους του δείγματος.

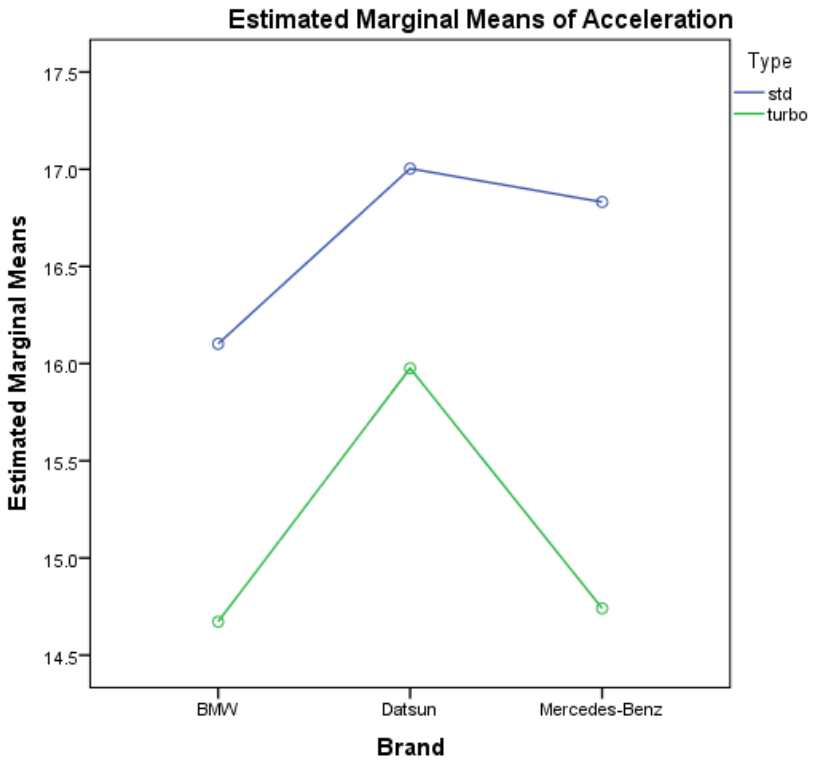
Two-Way ANOVA

Ας υποθέσουμε τώρα ότι θέλουμε να εξετάσουμε την διαφορά της μέσης επιτάχυνσης για τις διάφορες μάρκες και τους διάφορους τύπους κινητήρων ταυτόχρονα (δηλαδή για τις $3 \cdot 2 = 6$ κατηγορίες). Θα κατασκευάσουμε μία Two-Way ANOVA (κατά δύο παράγοντες). Βασική προϋπόθεση είναι η κανονικότητα, την οποία θα ελέγξουμε μέσα από τα κατάλοιπα.

Για να την τρέξουμε στο SPSS, ακολουθούμε τη διαδρομή Analyze → General Linear Model → Univariate και μιμούμαστε τις επιλογές που φαίνονται παρακάτω:



Αρχικά, κοιτάμε το γράφημα, όπου μπορούμε να δούμε τους μέσους κάθε μίας από τις 6 κατηγορίες. Με μπλε χρώμα συμβολίζονται τα αυτοκίνητα με κινητήρα std και με πράσινο τα turbo. Μπορούμε να δούμε σημαντικές διαφορές τόσο στις τρεις μάρκες, όσο και στα δύο χρώματα. Το ερώτημα που θέλουμε να θέσουμε σε αυτό το σημείο είναι αν αυτοκίνητα ίδιας μάρκας και διαφορετικού κινητήρα έχουν σταθερή διαφορά για τα όλα τα είδη κινητήρα, καθώς και αν αυτοκίνητα ίδιου κινητήρα και διαφορετικής μάρκας έχουν σταθερή διαφορά για όλες τις μάρκες.



Ο πίνακας που εκτελεί τους ελέγχους που μας ενδιαφέρουν είναι ο Tests of Between-Subjects Effects:

Tests of Between-Subjects Effects

Dependent Variable: Acceleration

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	369.959[a]	5	73.992	11.549	.000
Intercept	56805.782	1	56805.782	8866.433	.000
Brand	52.627	2	26.313	4.107	.017
Type	129.303	1	129.303	20.182	.000
Brand * Type	8.579	2	4.289	.670	.513
Error	2466.632	385	6.407		
Total	98246.722	391			
Corrected Total	2836.590	390			

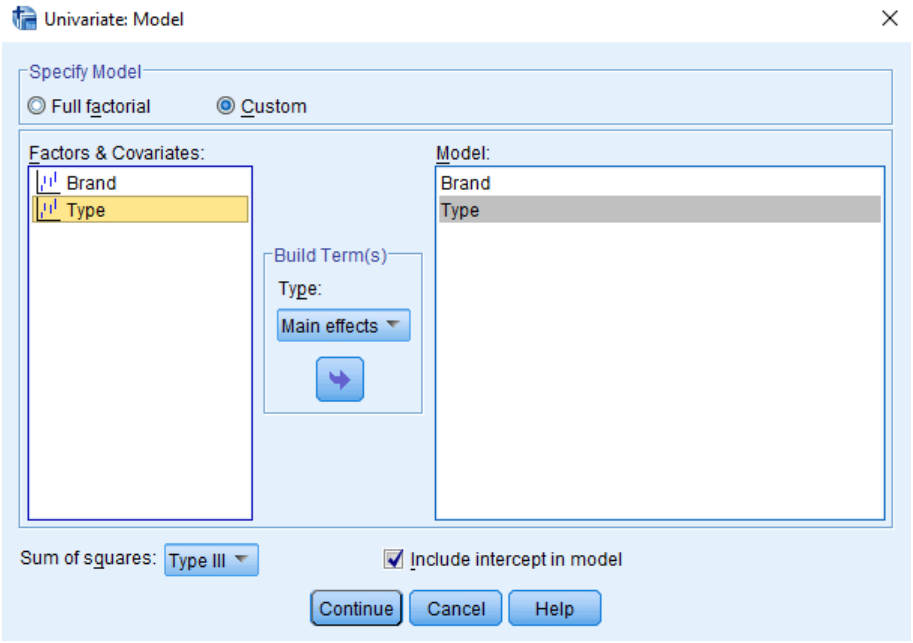
a. R Squared = .130 (Adjusted R Squared = .119)

Από την πρώτη γραμμή του πίνακα βλέπουμε πως το μοντέλο είναι στατιστικά σημαντικό ($p - value = 2 \cdot 10^{-10}$), δηλαδή τουλάχιστον δύο κατηγορίες διαφέρουν μεταξύ τους. Κοιτώντας την γραμμή Brand μπορούμε να δούμε ότι και αυτή έχει πολύ μικρό p-value και επομένως η μέση επιτάχυνση αυτοκινήτων διαφορετικής μάρκας διαφέρουν (κάτι που είδαμε και στην One-Way ANOVA). Ομοίως για την γραμμή Type. Στην υποσημείωση (a) βλέπουμε πως το μοντέλο έχει $R^2 = 0.130, R^2_{adj} = 0.119$, που είναι καλύτερο από το One-Way ANOVA, αλλά παραμένει αρκετά ασθενές.

Η γραμμή που μας ενδιαφέρει όμως είναι η Brand*Type, που μας δείχνει την αλληλεπίδραση των δύο μεταβλητών. Παρατηρούμε ότι έχουμε $p - value = 0.513$ επομένως ο όρος αυτός δεν είναι στατιστικά σημαντικός και μπορεί να αφαιρεθεί από το μοντέλο. Αυτό σημαίνει πώς η διαφορά στην επιτάχυνση δύο αυτοκινήτων ίδιας μάρκας με διαφορετικό κινητήρα παραμένει σταθερή, από όποια μάρκα κι αν προέρχονται τα δύο αυτοκίνητα. Ομοίως, η διαφορά στην επιτάχυνση δύο αυτοκινήτων ίδιου κινητήρα και διαφορετικής μάρκας παραμένει σταθερή, όποιο κι αν είναι το είδους του κινητήρα.

Οι επιλογές που κάναμε στο μενού Post Hoc μας εμφανίζουν ακριβώς τον ίδιο πίνακα Multiple Comparisons όπως στην One-Way ANOVA. Οι υπόλοιποι πίνακες αφορούν περιγραφικά στατιστικά και διαστήματα εμπιστοσύνης για τους μέσους.

Σε αυτό το σημείο, επαναλαμβάνουμε το Two-Way ANOVA, αυτή τη φορά χωρίς την αλληλεπίδραση των δύο παραγόντων (interaction term). Ανοίγουμε ξανά το παράθυρο για την Two-Way ANOVA, αλλά πέραν των άλλων επιλογών, ρυθμίζουμε και την επιλογή Model όπως φαίνεται παρακάτω:



Τρέχουμε το μοντέλο και παρατηρούμε πως και η Type και η Brand είναι στατιστικά σημαντικές, ενώ το R^2 έχει μια πολύ μικρή μείωση. Αυτό είναι το τελικό μας μοντέλο.

Tests of Between-Subjects Effects

Dependent Variable: Acceleration

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	361.380[a]	3	120.460	18.834	.000
Intercept	70314.632	1	70314.632	10993.716	.000
Brand	48.479	2	24.239	3.790	.023
Type	177.395	1	177.395	27.736	.000
Error	2475.211	387	6.396		
Total	98246.722	391			
Corrected Total	2836.590	390			

a. R Squared = .127 (Adjusted R Squared = .121)

Τέλος, κάνουμε τον έλεγχο κανονικότητας των καταλοίπων. Ο έλεγχος Shapiro-Wilk δίνει $p - value = 0.005$, επομένως η κανονικότητα απορρίπτεται σε επίπεδο σημαντικότητας 5% ή και 1%. Τα αποτελέσματά μας ισχύουν προσεγγιστικά λόγω του μεγάλου μέγεθους του δείγματος.

8 Παράρτημα

Μέτρα Θέσης και Διασποράς

Παρακάτω δίνονται τα σημαντικότερα περιγραφικά μέτρα, τα δειγματικά τους ανάλογα και οι τύποι υπολογισμού τους. Ιδιαίτερη προσοχή στη σύγχυση παραμέτρου (π.χ. μέση τιμή μ) και εκτιμήτριας (π.χ. δειγματικός μέσος $\bar{X}$).

Table 1: Μέτρα Θέσης και Διασποράς

Μέτρο	Εκτιμήτρια	Τύπος
Μέση Τιμή	Δειγματικός Μέσος	$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
Διασπορά	Δειγματική Διασπορά	$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$
Τυπική Απόκλιση	Τυπικό Σφάλμα	$S = \sqrt{S^2}$
Ασυμμετρία	Δειγματική Ασυμμετρία	$l = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{nS^3}$
Κύρτωση	Δειγματική Κύρτωση	$k = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4}{nS^4}$