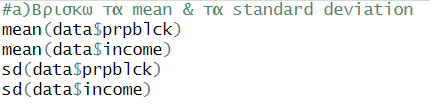
**Εκφώνηση:**

Το συγκεκριμένο project πρόκειται για δεδομένα σε επίπεδο ταχυδρομικού κώδικα για τις τιμές διαφόρων αγαθών σε εστιατόρια fast food μαζί με τα χαρακτηριστικά του πληθυσμού για τον εκάστοτε ταχυδρομικό κώδικα στο Νιου Τζέρσι και την Πενσιλβάνια. Σκοπός μας είναι να δούμε εάν τα εστιατόρια fast food χρεώνουν υψηλότερες τιμές σε περιοχές με μεγαλύτερη συγκέντρωση μαύρων. Το project αυτό αποτελείται από 10 ερωτήματα (a-i) τα οποία καλύπτουν ένα γενικό εύρος γνώσεων και χρησιμοποιώ την γλώσσα R για να απαντήσω σε αυτά.

1. Βρείτε τις μέσες τιμές των *prpblck* και *income* στο δείγμα μαζί με τις τυπικές αποκλίσεις τους. Ποιες είναι οι μονάδες μέτρησης των *prpblck* και *income*;

**Απάντηση:**

Οι μέσες τιμές είναι: prpblck=0.1134864 και income=47053.78 , ενώ οι τυπικες αποκλίσεις είναι: prpblck=0.1821933 και income=13163.16 . To prpblck εχει ποσοστό(%) για μονάδα μέτρησης ενώ το income τα δολάρια($).

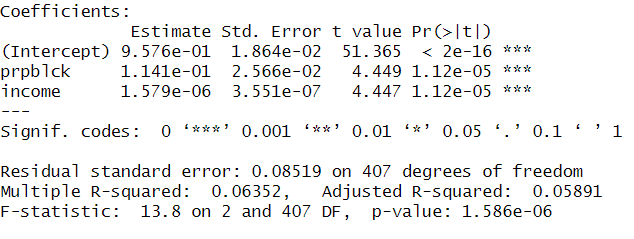


1. Πάρτε ένα υπόδειγμα για να ερμηνεύσετε την τιμή του αναψυκτικού, *psoda*, σε όρους αναλογίας του πληθυσμού που είναι μαύροι και του εισοδήματος:

Εκτιμήστε αυτό το υπόδειγμα με τη μέθοδο OLS και αναφέρετε τα αποτελέσματα με τη μορφή εξίσωσης, συμπεριλαμβάνοντας το μέγεθος του δείγματος και το R-τετράγωνο.

**Απάντηση:**

Στο συγκεκριμένο υπόδειγμα βλεπουμε ότι αν το prpblck και το Income αυξηθούν κατά 1% τότε το psoda αυξάνεται αντίστοιχα 1.141e-01$ και 1.579e-6$ ,το R-squared είναι 0.06352 και το p-value = 1.586e-06 (δηλαδή στατιστικά σημαντικό). Το μέγεθος του δείγματος είναι 410 παρατηρήσεις.



1. Ερμηνεύστε τον συντελεστή του ποσοστού μαύρων (*prpblck*). Είναι στατιστικά διάφορος του μηδενός σε επίπεδο σημαντικότητας 5%; Έχει νόημα από οικονομική άποψη;

**Απάντηση:**

Όπως είδαμε και στο b) ερωτημα μια αύξηση του prpblck κατά 1% αυξάνει το psoda κατά 1.141e-0.1$. Σύμφωνα με το p-value είναι στατιστικά σημαντικό στο 5% . Οικονομικά αυτό σημαίνει ότι όταν υπάρχει μεγαλύτερο prpblck υπάρχουν και υψηλότερες τιμές στα αναψυκτικά των fast food (psoda), συνήθως λόγο οικονομικής εκμετάλλευσης από τις αλυσίδες fast food.



1. Ένα υπόδειγμα με μια σταθερή ελαστικότητα τιμής αναφορικά με το εισόδημα μπορεί να είναι καταλληλότερο. Αναφέρετε τις εκτιμήσεις του υποδείγματος:

Ποιο κριτήριο θα επιλέγατε για να συγκρίνετε το συγκεκριμένο υπόδειγμα με το προηγούμενο; Ποιο υπόδειγμα είναι καλύτερο σύμφωνα με το κριτήριο που επιλέξατε;

**Απάντηση:**

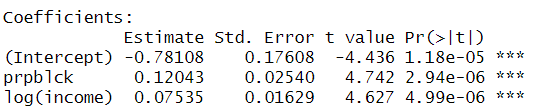
Τώρα παρατηρούμε ότι αν το prpblck και το log(Income) αυξηθούν κατά 1% τότε το log(psoda) αυξάνεται αντίστοιχα 0.12043$ και 0.07535%, το R-squared είναι 0.06723 και το p-value = 7.06e-07. Το μέγεθος του δείγματος είναι 410 παρατηρήσεις. Με σκοπό να συγκρίνουμε τα 2 υποδείγματα χρησιμοποιούμε το AIC και επιλέγουμε το μικρότερο (δηλαδή το ols2).



1. Ο συντελεστής του εισοδήματος (*)* είναι στατιστικά σημαντικός για επίπεδο 5%; Εάν το εισόδημα αυξηθεί κατά 1%, ποια θα είναι η μεταβολή στην τιμή του αναψυκτικού (*psoda)*; Πως εξηγείται αυτό με οικονομικούς όρους;

**Απάντηση:**

Είναι στατιστικά σημαντικός γιατί p-value = 4.99e-06 < 0.05 , αν το εισόδημα αυξηθεί κατά 1% τότε το log(psoda) αυξάνεται κατά 0.07535%. Αυτό σημαίνει ότι όσο αυξάνεται το εισόδημα αυξάνεται και η τιμή του αναψυκτικού στα fast food μαγαζιά δηλαδή από οικονομικής άποψης το συγκεκριμένο γεγονός δηλώνει ότι όσο αυξάνεται το ποσό χρημάτων κάποιου, λόγο οικονομικής άνεσης θα αυξήσει και την κατανάλωσή του.

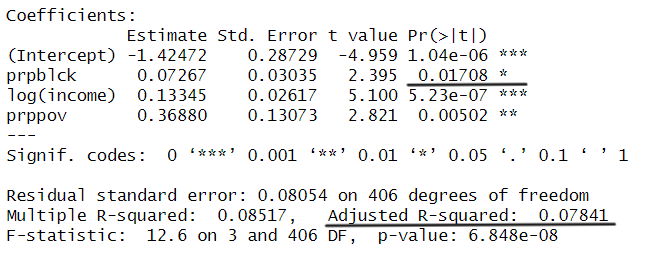


1. Προσθέστε τώρα τη μεταβλητή ποσοστού φτώχειας (*prppov)* στην παλινδρόμησης του ερωτήματος d). Πως επηρεάζεται ο συντελεστής του prpblck () σε αυτή την

περίπτωση; Η ερμηνευτική ικανότητα του υποδείγματος αυξάνεται ή μειώνεται;

**Απάντηση:**

Βλεπουμε ότι σε αντίθεση με τα προηγούμενα υποδείγματα το prpblck μειώνεται. Η ερμηνευτική ικανότητα αυξάνεται διότι αυξάνεται και το adjust R-squared από ότι πριν.



1. Βρείτε τη συσχέτιση μεταξύ log(income) και *prppov*. Είναι χονδρικά αυτή που αναμένατε;

**Απάντηση**:

Η συσχέτιση είναι: -0.8384476. Ναι είναι χονδρικά αυτό που αναμέναμε γιατί είναι λογικό όσο αυξάνεται το εισόδημα να μειώνεται η φτώχια.



1. Αξιολογήστε την ακόλουθη πρόταση: «Επειδή η log(*income*) και η *prpov* έχουν τόσο υψηλή συσχέτιση, δεν έχουν θέση στην ίδια παλινδρόμηση».

**Απάντηση:**

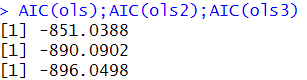
Πράγματι η υψηλή συσχέτιση πολλές φορές είναι σημάδι πιθανού προβλήματος πολυσυγγραμικοτητας. Αλλά αν εφαρμόσουμε τον έλεγχο VIF παρατηρούμε ότι στην δικιά μας περίπτωση δεν ισχύει κάτι τέτοιο (διότι η τιμές είναι κάτω από το 10).



1. Με δεδομένα τα αποτελέσματα των προηγούμενων παλινδρομήσεων, ποια θα θεωρούσατε την πλέον αξιόπιστη παλινδρόμηση; Σύμφωνα με αυτό κατά πόσον η φυλετική σύνθεση μιας περιοχής επηρεάζει τις τιμές αναψυκτικών των τοπικών εστιατορίων fast food; Γιατί πιστεύετε ισχύει αυτό;

**Απάντηση:**

Η πλέον αξιόπιστη παλινδρόμηση είναι η 3η  (ols3) διότι εχει το χαμηλότερο AIC σε σχέση με τις άλλες. Σε όλες τις περιπτώσεις μέχρι τώρα ο συντελεστής του prpblck ήταν πάντα θετικός και στατιστικά σημαντικός, δηλαδή αυτό σημαίνει ότι σε κάποια περιοχή όσο μεγαλύτερο είναι το prpblck τόσο υψηλότερες θα είναι και οι τιμές των αναψυκτικών στα fast food. Αυτό πιστεύω γίνεται διότι όπως ανέφερα πριν υπάρχει οικονομική εκμετάλλευση ή μπορεί να οφείλεται επίσης στο γεγονός οτι το ποσοστό των έγχρωμων ανθρώπων κάνει μεγαλύτερη κατανάλωση στα αναψυκτικά (λόγο καθημερινής συνήθειας).



1. Μπορείτε να ελέγξετε αν παραβιάζονται κάποιες από τις βασικές υποθέσεις της γραμμικής παλινδρόμησης στο τελευταίο υπόδειγμα; Που μπορεί να οφείλεται (αν υπάρχει) η παραβίαση κάποιων υποθέσεων; Τι μέτρα θα λαμβάνατε για την βελτίωση της ερμηνευτικής ικανότητας του υποδείγματός σας;

**Απάντηση:**

Έλεγχος παραβιάσεις βασικών υποθέσεων:

1. Μη κανονικότητα καταλοίπων. Χρησιμοποιω τον έλεγχο Jerque-Bera και εντοπιζω ότι υπάρχει προβλημα.
2. Πολυσυγγραμικοτητα. Χρησιμοποιω τον συντελεστή VIF και δεν εντοπιζω καποιο προβλημα.
3. Αυτοσυσχετιση. Χρησιμοποιω τον έλεγχο Durbin Watchon και δεν εντοπιζω καποιο προβλημα.
4. Ετεροσκεδαστικοτητα. Χρησιμοποιω τον έλεγχο White και δεν εντοπιζω καποιο προβλημα.

Τα προβλήματα αυτά συνήθως οφείλονται σε μεταβλητές που δεν χρειάζονται στο υπόδειγμά μας και τις βάλαμε (π.χ. κανονικότητα καταλοίπων) ή ακόμα και το αντίθετο δηλαδή μια μεταβλητή που θα έπρεπε να εισάγουμε και δεν την εισήγαμε (π.χ. αυτοσυσχετιση). Από τα πιο συνηθισμένα μέτρα που μπορούμε να πάρουμε για την αντιμετώπιση αυτών των προβλημάτων είναι να ελέγχουμε όσο το δυνατόν καλυτέρα τα δεδομένα μας και να χρησιμοποιούμε τις σωστές μεταβλητές, ενώ επίσης να αποφεύγουμε αυτές που δεν βοηθάνε την ερμηνευτική ικανότητα του υποδείγματός μας (παρακάτω ακολουθούν οι κώδικες που χρησιμοποίησα).

