LABORATORY OF STATISTICS AND DATA ANALYSIS

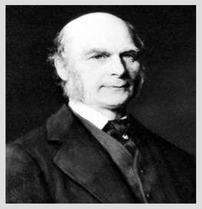
Reversion and Regression: two peas in a pod – Part 1

e-ΠΕΡΙΣΚΟΠΙΟ

No.2/2020

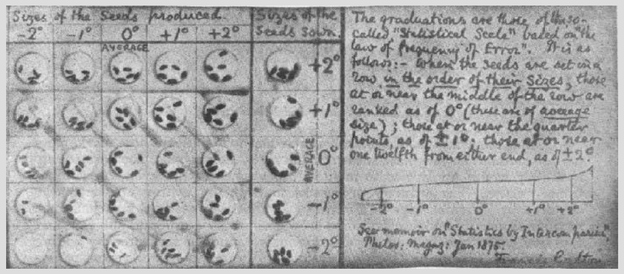
Το e-Περισκόπιο του Εργαστηρίου Στατιστικής και Ανάλυσης Δεδομένων του Πανεπιστημίου αποτελεί μια πρωτοβουλία των φοιτητών-ερευνητών που το πλαισιώνουν και δεν απευθύνεται αποκλειστικά σε άτομα με στατιστικό υπόβαθρο.
Το κάθε τεύχος είναι ανεξάρτητο των υπολοίπων και ακολουθεί συγκεκριμένη θεματολογία.
Συγκεκριμένα απαρτίζεται από: (1) Βιογραφικό σημείωμα ενός ατόμου που έχει συνδεθεί με το υπό ανάλυση θέμα και η συμβολή του στην στατιστική ήταν καθοριστική και θεμελιώδης; (2) Εισαγωγική συζήτηση του θέματος συνοδευόμενη από κάποια funny corners καθώς και ένα quiz/paradox. Στόχος του e-Περισκοπίου είναι η ενημέρωση, η ψυχαγωγία και ο προβληματισμός των αναγνωστών σε θέματα που έχουν ως κεντρικό άξονα την στατιστική. Αν επιθυμείτε να συμβάλετε στο περισκόπιο θέτοντας κάποιο θέμα προς ανάλυση, επισκεφτείτε την σελίδα του εργαστηρίου ή/και ελάτε σε επικοινωνία μαζί μας μέσω των πληροφοριών που βρίσκονται στην καρτέλα *"ΣΤΟΙΧΕΙΑ ΕΠΙΚΟΙΝΩΝΙΑΣ"*.

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ



Ο Sir Francis Galton (*1822-1911*) ήταν Βρετανικής καταγωγής ανθρωπολόγος με σημαντική συνεισφορά στη Στατιστική, την Κοινωνιολογία αλλά και την Ψυχολογία. Για το ερευνητικό του έργο (*το οποίο υπολογίζεται σε πάνω από 340 επιστημονικές δημοσιεύσεις*) έλαβε τον τίτλο Sir το 1909. Ήταν ο πρώτος επιστήμονας που εφάρμοσε τεχνικές Στατιστικής για τη μελέτη ανθρωπολογικών διαφορών. Σε μία από τις σημαντικότερες μελέτες του με τίτλο "Παλινδρόμηση περί του μετρίου κληρονομικού αναστήματος" μελέτησε τη σχέση μεταξύ της διαμέτρου του σπόρου του μπιζελιού και της διαμέτρου του παραγόμενου καρπού και πρότεινε τη τεχνική της "Επαναστροφής" (*Reversion*) που μεταγενέστερα επαναπροσδιόρισε και επανεισήγαγε με τον όρο "Παλινδρόμηση" (*Regression*). Ανάμεσα στα σημαντικότερα ερευνητικά του επιτεύγματα συγκαταλέγονται, μεταξύ άλλων, τα ακόλουθα:

1. **Συσχέτιση:** στατιστικό μέτρο το οποίο ποσοτικοποιεί τη "σχέση" μεταξύ δύο τυχαίων μεταβλητών.
2. **Παλινδρόμηση:** τεχνική μοντελοποίησης μιας τυχαίας μεταβλητής βάσει άλλων "σχετικών" μεταβλητών.
3. **Διάμεσος:** πρότεινε τη χρήση της διαμέσου ως μέτρου κεντρικής τάσης των κατανομών.



Χειρόγραφη δημοσίευση του Sir F. Galton αναφορικά με την Παλινδρόμηση

ΜΠΙΖΕΛΙΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

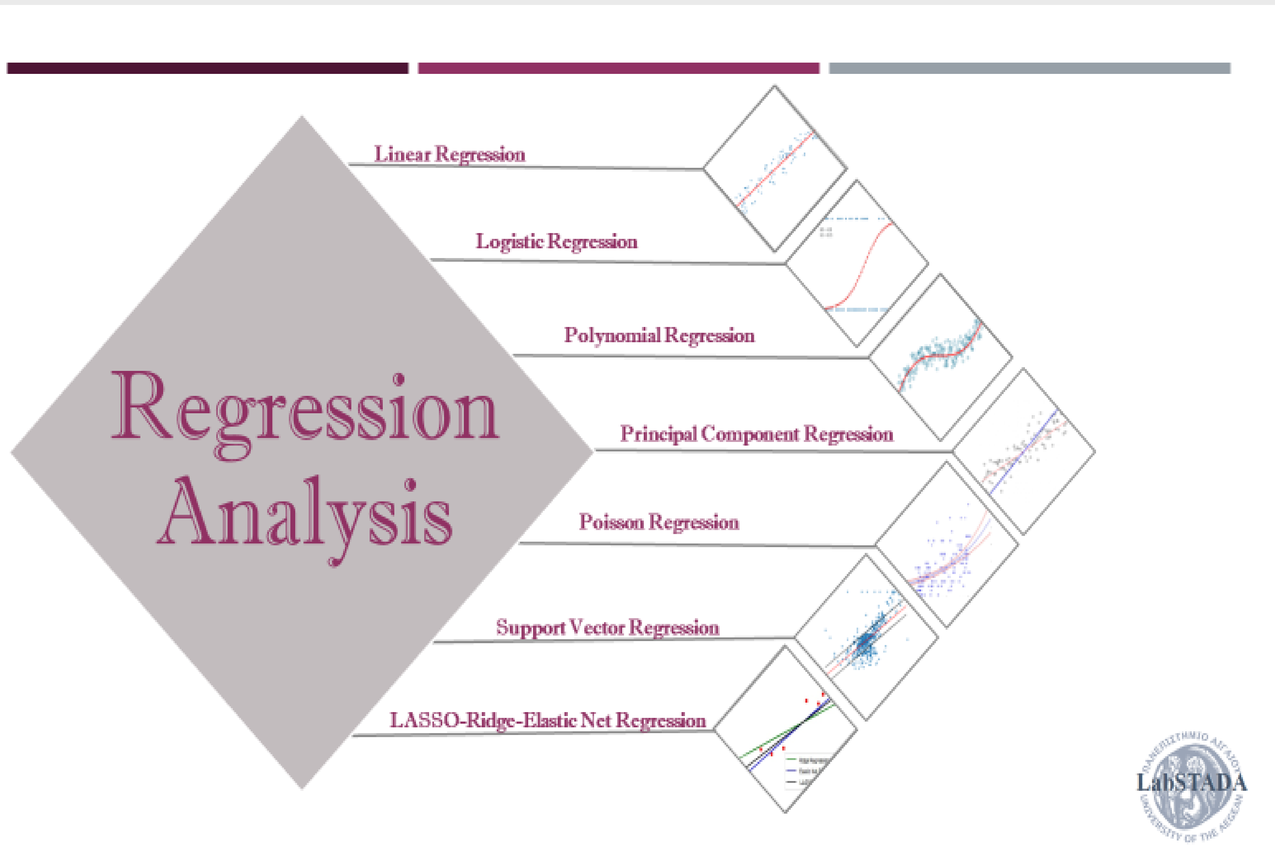
Πως από ένα μπιζέλι "φύτρωσε" ο όρος παλινδρόμηση

Click to learn more

Ένα από τα πιο σημαντικά προβλήματα στα οποία εστιάζεται η Στατιστική είναι η ταυτόχρονη μελέτη δύο ή περισσότερων μεταβλητών ώστε να προσδιοριστεί η μεταξύ τους σχέση. Κύριος στόχος είναι η πρόβλεψη των τιμών της μίας μέσω των τιμών της άλλης, για παράδειγμα η ηλικία και το ύψος ενός παιδιού έχουν μια κάποια θετική εξάρτηση (όσο μεγαλύτερη η ηλικία τόσο μεγαλύτερο και το ύψος). Αυτός ο τομέας της Στατιστικής ονομάζεται *Ανάλυση Παλινδρόμησης*. Στην Παλινδρόμηση υπάρχουν τα εξής δύο είδη μεταβλητών:

- **Επεξηγηματικές (ή ανεξάρτητες)** είναι εκείνες τις οποίες μπορούμε να ελέγξουμε και να καθορίσουμε τις τιμές τους.
- **Απόκρισης (ή εξαρτημένες)** είναι εκείνες στις οποίες αποτυπώνεται το αποτέλεσμα των μεταβολών των ανεξάρτητων μεταβλητών.

Με βάση το πλήθος των επεξηγηματικών μεταβλητών, η Παλινδρόμηση μπορεί να διακριθεί σε δύο κατηγορίες. Η παλινδρόμηση στην οποία υπάρχει μόνο μία επεξηγηματική μεταβλητή καλείται **Απλή Παλινδρόμηση** ενώ στην περίπτωση που υπάρχουν περισσότερες από μία επεξηγηματικές μεταβλητές καλείται **Πολλαπλή Παλινδρόμηση**. Για παράδειγμα, η εύρεση της σχέσης μεταξύ των συνολικών κρουσμάτων ενός ιού σε μία χώρα και της θερμοκρασίας αυτής της χώρας την ίδια χρονική περίοδο είναι πρόβλημα Απλής Παλινδρόμησης, ενώ η εύρεση της σχέσης μεταξύ των συνολικών κρουσμάτων του ιού στη χώρα, της θερμοκρασίας αυτής αλλά και της υγρασίας αποτελεί πρόβλημα Πολλαπλής Παλινδρόμησης.



Regression analysis: If everything else fails, ignore it!

Στη Στατιστική υπάρχουν διάφορα είδη Παλινδρόμησης με πιο σύνηθες αυτό της Γραμμικής, η οποία διακρίνεται σε Απλή και Πολλαπλή. Με αυτό το είδος Παλινδρόμησης μελετάται η γραμμική σχέση μεταξύ της μεταβλητής απόκρισης και επεξηγηματικής(ων) μεταβλητής(ων) με βάση τις ακόλουθες εξισώσεις.

Απλή Γραμμική Παλινδρόμηση

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

όπου Y η μεταβλητή απόκρισης, X η επεξηγηματική, τα β_0 και β_1 οι (*άγνωστες*) σταθερές παράμετροι του μοντέλου και το ε ο τυχαίος παράγοντας (*σφάλμα*). Η παράμετρος β_0 είναι το σημείο που η εξίσωση (*ευθεία*) παλινδρόμησης τέμνει τον άξονα των y, δηλαδή αντιστοιχεί στην αναμενόμενη τιμή της Y για $X = 0$ και για αυτό ονομάζεται διαφορά ύψους (*intercept*). Η παράμετρος β_1 είναι η κλίση (*slope*) της ευθείας παλινδρόμησης και αντιπροσωπεύει τη μεταβολή (*αύξηση ή μείωση*) στην αναμενομένη τιμή της Y που αντιστοιχεί σε μεταβολή της X κατά μία μονάδα.

Πολλαπλή Γραμμική Παλινδρόμηση

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},$$

όπου $\boldsymbol{Y}$ το διάνυσμα στήλη των τιμών απόκρισης, $\boldsymbol{X}$ ο πίνακας τιμών των επεξηγηματικών μεταβλητών (*πίνακας σχεδιασμού*), $\boldsymbol{\beta}$ το διάνυσμα στήλη των (*άγνωστων*) σταθερών παραμέτρων του μοντέλου και $\boldsymbol{\varepsilon}$ το διάνυσμα στήλη των τυχαίων σφαλμάτων.
Στο γραμμικό μοντέλο πολλαπλής παλινδρόμησης ισχύει, όπως και στο απλό γραμμικό μοντέλο, η βασική ισότητα *Ανάλυσης Διακύμανσης*,

SST=SSR+SSE,

Η οποία υποδεικνύει ότι η ολική μεταβλητότητα (*Total Sum of Squares, SST*) των παρατηρήσεων της μεταβλητής απόκρισης αναλύεται σε δύο μέρη:

- ένα μέρος που ερμηνεύεται από την παλινδρόμηση (*Regression Sum of Squares, SSR*)
- ένα μέρος που παραμένει ανερμήνευτο (*Error Sum of Squares, SSE*).

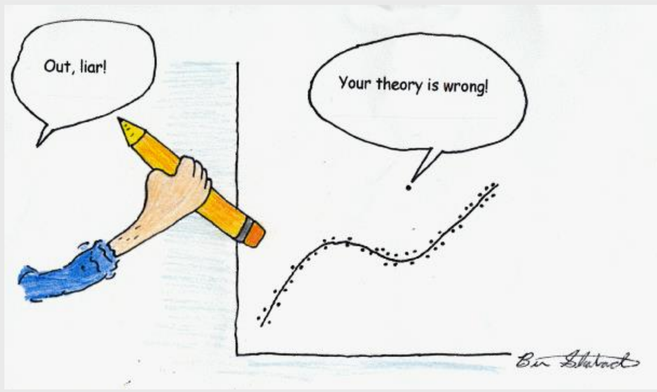
Στα ανωτέρω μοντέλα, ο όρος "γραμμικό" αναφέρεται στις παραμέτρους του μοντέλου και όχι στις επεξηγηματικές μεταβλητές οι οποίες θα μπορούσαν να υπεισέλθουν στο μοντέλο και με μεγαλύτερες δυνάμεις ή και ως πιο πολύπλοκες συναρτήσεις.

Στο γραμμικό μοντέλο μία σημαντική υπόθεση είναι η κανονικότητα του σφάλματος (*και κατ' επέκταση και της μεταβλητής απόκρισης*). Μια δημοφιλής προσέγγιση στις περιπτώσεις που η υπόθεση αυτή παραβιάζεται είναι η υιοθέτηση των Γενικευμένων Γραμμικών Μοντέλων *(Generalized Linear Models, GLMs*). Τα GLMs είναι μια γενίκευση των κλασικών γραμμικών μοντέλων τα οποία περιλαμβάνουν, μεταξύ άλλων, την ανάλυση διασποράς, τα logit και probit μοντέλα, τα log-linear και τα πολυωνυμικά μοντέλα, τα μοντέλα Ελαστικής Παλινδρόμησης, κ.ά. Το 1972, οι Nelder και Wedderburn, παρουσίασαν μια ενοποιημένη θεωρία για γραμμικά μοντέλα σύμφωνα με την οποία τα μοντέλα αυτά μπορούν να μελετηθούν ενιαία κάτω από την υπόθεση ότι η κατανομή της μεταβλητής απόκρισης ανήκει στην εκθετική οικογένεια κατανομών.
Να σημειωθεί ότι για την εκτίμηση των άγνωστων παραμέτρων του μοντέλου, έχουν προταθεί διάφορες τεχνικές εκτίμησης, με τη μέθοδο των Ελαχίστων Τετραγώνων να αποτελεί την πλέον δημοφιλέστερη κατά την κλασική γραμμική παλινδρόμηση και τη μέθοδο της Μέγιστης Πιθανοφάνειας να είναι συνήθως εφαρμοζόμενη στα γενικευμένα γραμμικά υποδείγματα.

Regression analysis is the hydrogen bomb of the statistics arsenal.

-Charles Wheelan

And that's all for the day fellow statistician folks.
Until next time take care and remember…
The probability to be killed by a cow is low
but never zero!



ΣΤΟΙΧΕΙΑ ΕΠΙΚΟΙΝΩΝΙΑΣ

Laboratory of
Statistics and Data Analysis