# Απλή γραμμική παλινδρόμηση (Simple linear Regression)

Data_set-12

Questions: Given a person's height what would be their predicted weight? How can we best define the relationship between height and weight? By studying the graph (scatterplot) below we see that the relationship is approximately linear.

We can imagine drawing a straight line on the graph with most of the data points being only a short distance from the line. The vertical distance from each data point to this line is called a residual (υπόλοιπο - κατάλοιπο).

Question: Πως θα καθορίσουμε την "βέλτιστη" ευθεία γραμμή που πρέπει/μπορεί να προσαρμοστεί στα δεδομένα height/weight;

Η μέθοδος των ελαχίστων τετραγώνων (Method of least squares) συνήθως χρησιμοποιείται για να καθορίσει η γραμμή παλινδρόμησης (regression line) που ελαχιστοποιεί το άθροισμα των τετραγώνων των residuals (υπόλοιπα, κατάλοιπα ή σφάλματα).

Η απλή γραμμική παλινδρόμηση (simple linear regression) ορίζει τη σχέση δύο ποσοτικών μεταβλητών $X$ και $Y$, μέσω ενός υποδείγματος της μορφής:

$$Y = b_0 + b_1 X + \varepsilon, \quad \varepsilon: \text{σφάλμα}$$

$b_0, b_1$: παράμετροι

$Y$: εξαρτημένη μεταβλητή (dependent or response variable)

$X$: ανεξάρτητη μεταβλητή (independent or predictor variable).

Η διαδικασία προσδιορισμού της ευθείας των ελαχίστων τετραγώνων ορίζεται από την εξίσωση:

$\hat{Y} = b_0 + b_1 X$ και απαιτεί τον προσδιορισμό των ποσοτήτων $b_0$ και $b_1$. Από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων

$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$ προκύπτει ότι

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$ και $b_0 = \bar{Y} - b_1 \bar{X}$

διότι $\bar{Y} = b_0 + b_1 \bar{X}$.

Συνήθως, υποθέτουμε ότι $\varepsilon_i \sim N(0, \sigma^2)$, και θεωρούμε ότι

$Cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$

δηλ. $\varepsilon_i, \varepsilon_j$ ασυσχέτιστα.

$\sigma^2$: σταθερή $\forall i = 1, \ldots, n$.

Analyze → Regression → Linear

Dependent → Weight

Independent → Height

Statistics
- Estimates, Confidence intervals, 95%
- Model Fit
- Descriptives
- Casewise diagnostics (All cases)

Plots
- ✓ Histogram
- ✓ Normal Probability Plot

Options
- ✓ Include constant in equation
- • Exclude cases listwise

We can write the equation for the "best" straight line defined by our height/weight data:

$$WEIGHT = -592.64 + 11.19 \cdot HEIGHT$$

Given any height, we can now predict the weight. For example, the predicted weight of a 70-inch-tall (inches) person is: (pounds)

$$WEIGHT = -592.64 + 11.19 \times 70 = 190.66 \text{ lb}$$

T-tests $H_0: b_0 = 0$ vs $H_1: b_0 \neq 0$

$$T_{b_0} = \frac{|\hat{b}_0|}{s_{b_0}} > t_{n-2;\alpha/2}$$

$$s_{b_0}^2 = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right), \quad s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\hat{b}_0 \sim N\left(b_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)\right)$$

$H_0: b_1 = 0$ vs $H_1: b_1 \neq 0$

$$T_{b_1} = \frac{|\hat{b}_1|}{s_{b_1}}, \quad s_{b_1}^2 = \frac{s^2}{(n-1)s_x^2}, \quad \hat{b}_1 \sim N\left(b_1, \frac{\sigma^2}{(n-1)s_x^2}\right)$$

$b_1 = 11.19$ $\quad b_0 = -592.64$

$s_{b_1} = 1.21$ $\quad s_{b_0} = 81.54$

The T-values test the hypotheses that the parameters $b_0, b_1$ are actually zero.

The standard errors $(s_{b_0}, s_{b_1})$ reflects the accuracy with which we know the true slope $(b_1)$ and intercept $(b_0)$.

ΠΙΝΑΚΑΣ ΑΝΑΛΥΣΗΣ ΔΙΑΣΠΟΡΑΣ (Analysis of Variance Table)

Η μεταβλητότητα της εξαρτημένης μεταβλητής αναλύεται στις διάφορες πηγές της.

$$\text{Total Sum of Squares} = SST = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

$$SSR\,(\text{Regression}) = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$$SSE\,(\text{Error}) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$SST = SSR + SSE$$

$H_0: b_1 = b_2 = \dots = b_K = 0$

$H_1: b_j \neq 0$ for at least one $j$

$$\frac{SSR}{\sigma^2} \overset{H_0}{\sim} \chi^2_K, \quad K := \#\text{ of regressor variables in the model}$$

$$\frac{SSE}{\sigma^2} \overset{H_0}{\sim} \chi^2_{n-K-1}, \quad SSE, SSR \text{ independent r.v.}$$

$$F = \frac{SSR/K}{SSE/(n-K-1)} = \frac{MSR}{MSE} \sim f_{K, n-K-1}$$

$H_0$ σε ε.σ. $\alpha$ αν $F > f_{\alpha; K, n-K-1}$

If the deviations about the regression line are small (small error mean square) compared to the deviation between the predicted values and the mean squared regression then we have a good regression line.

Consequently the larger the ratio F, the better the fit. If there is one dependent variable only, the probability of the F-statistic is the same as the probability associated with testing for the significance of the correlation coefficient. If this probability is "large" (p-value is large) our linear model is not doing a good job of describing the relationship between the variables.

Προσαρμογή ευθείας Analyze → Regression
→ Curve Estimation
Dependent → Weight
Independent → Height
√ Linear (+ άλλες προσαρμογές)

ANOVA TABLE

| | Sum of Squares | df | Mean Square | F | $\hat{\alpha}$=Sig |
|---|---|---|---|---|---|
| Regression | SSR | K | MSR | MSR/MSE | $\hat{\alpha}$ |
| Residual | SSE | n-K-1 | MSE | | |
| Total | SST | n-1 | | | |

R-square is the square of the multiple correlation coefficient. $0 \leq R^2 \leq 1$ Επειδή έχουμε μόνο μία ανεξάρτητη μεταβλητή την HEIGHT, R-square is the square of the Pearson correlation coefficient between HEIGHT and WEIGHT. The square of the correlation coefficient tells us how much variation in the dependent variable can be accounted for by variation of the independent variable. Όταν υπάρχουν περισσότερες από μία ανεξάρτητες μεταβλητές, το R-square αναπαριστά τη μεταβλητότητα στην εξαρτημένη μεταβλητή που

οφείλεται στο γραμμικό συνδυασμό όλων των ανεξάρτητων μεταβλητών.

Adj $R^2$-Sq: the squared CC. corrected for the number of independent variables in the equation. This adjustment has the effect of decreasing the value of R-squared. Η διαφορά είναι μικρή αλλά γίνεται μεγαλύτερη και περισσότερο σημαντική όταν έχουμε πολλές ανεξάρτητες μεταβλητές.

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

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2), \quad k := \#\text{ ανεξ. μεταβλητών}$$

When $R^2$ και $R^2_{adj}$ differ dramatically there is a good chance that non-significant terms have been included in the model.

Std. Error of Estimate = Root MSE: is the square root of the error variance. It is the standard deviation of the residuals

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} = \sqrt{MSSE}$$

Confidence Intervals for the Coefficients

$b_0$: $\left(\hat{b}_0 \pm s_{b_0} \cdot t_{n-2;\alpha/2}\right)$ n: μικρό we should go to a t-table to find a 95% CI for $b_0$ since $n<30$

$t_{n-2;\frac{0.05}{2}} = 2.57$

$df = n-2 = 5, \alpha = 0.05$

Για n μεγάλο: $\hat{b}_0 \pm s_{b_0} z_{\alpha/2}$

Όρια $\rightarrow (11.19 \pm 2.57 \cdot 1.218) = (8.064, 14.322)$

$b_1$: $\hat{b}_1 \pm s_{b_1} \cdot t_{n-2;\alpha/2}$ για n μικρό, $n<30$ ; n μεγάλο $\hat{b}_1 \pm s_{b_1} z_{\alpha/2}$

## Coefficient of Variation (C.V.)

C.V. = 7.62 (×100) (Συντελεστής Μεταβλητότητας)

Two different settings → (α) Ανάλυση μίας μεταβλητής
→ (β) Ερμηνεία μοντέλου

(α) $C.V. = \frac{s}{\bar{x}} (\times 100)$

Προσπαθεί να περιγράψει τη διασπορά (τη μεταβλητότητα) της μεταβλητής με έναν τρόπο ώστε να μην εξαρτάται από τη μονάδα μέτρησης της μεταβλητής.

(β) In the model setting $C.V. = \frac{\text{Root MSE}}{\text{Dep Mean}} \times 100$

$= \frac{11.86}{155.57} \times 100 \cong 7.623$

(Dep Mean = mean of the dependent variable)

(α): The higher the C.V., the greater the dispersion in the variable.

(β): The C.V. for a model, aims to describe the model fit (προσπαθεί να περιγράψει την προσαρμογή του μοντέλου με βάση τα σχετικά μεγέθη των τετραγώνων των καταλοίπων και των τιμών της εξαρτημένης μεταβλητής. Χαμηλές τιμές για το C.V. είναι ένδειξη για "καλή" προσαρμογή του μοντέλου. Το μοντέλο με το μικρότερο C.V. έχει predicted values κοντινότερες στις πραγματικές τιμές. Both R-squared and C.V. values are unitless (ανεξάρτητες μονάδων μέτρησης) ενδείξεις για τη προσαρμογή ενός μοντέλου αλλά ορίζουν την προσαρμογή του με δύο διαφορετικούς τρόπους.

C.V.: Αξιολογεί τη σχετική εγγύτητα των προβλεπόμενων τιμών με τις πραγματικές τιμές.

R-squared: Αξιολογεί κατά πόσο η μεταβλητότητα των πραγματικών τιμών μπορεί να εξηγηθεί μέσω του μοντέλου.

Πλεονέκτημα και των δύο: Ανεξαρτησία από τις μονάδες μέτρησης των μεταβλητών.

Μειονέκτημα του C.V. είτε στο (α) είτε στο (β).

It can be considered as a reasonable measure if the variable contains only positive values. This is a disadvantage of C.V. (δηλ. αν οι τιμές είναι για την μεταβλητή άλλες θετικές, άλλες αρνητικές και C.V. → 0, τότε η τιμή του μπορεί να είναι παραπλανητική για το μοντέλο. Δεν μπορεί να το C.V. να οριστεί για τέτοιες περιπτώσεις και να δώσει αξιόπιστες ερμηνείες για το υπό-μελέτη μοντέλο.