Regression

In some situations, a response (dependent) variable Y depends on one or more independent variables (or input variables or covariates) denoted by $X_1, X_2, \cdots, X_p$. The simplest type of relationship is a linear relationship as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

where ϵ is a random "*error*" term that has mean zero. The above model is called a *linear regression model*. When there is only one covariate, we call it a *simple linear regression model*. When there are two or more covariates, we call it a *multiple linear regression model*. The quantities $\beta_0, \beta_1, \cdots, \beta_p$ are called regression coefficients, and they are estimated from the data.

We can write the model as follows:

$$E(Y \mid \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Least squares estimators of regression coefficients

Let $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ be estimators of the regression coefficients mentioned above. These estimators should be such that the following quantity is minimized:

$$E = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$$

The fitted model

$$Y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p + e$$

where e is the "*residual*".

Simple linear regression

When $p = 1$, i.e., there is only one covariate x, the model is often written as

$$Y = \alpha + \beta x + \epsilon$$

Least Squares (LS) estimators of the regression coefficients are:

$$\hat{\beta} = \frac{\sum(x - \bar{x})(Y - \bar{Y})}{\sum(x - \bar{x})^2}$$

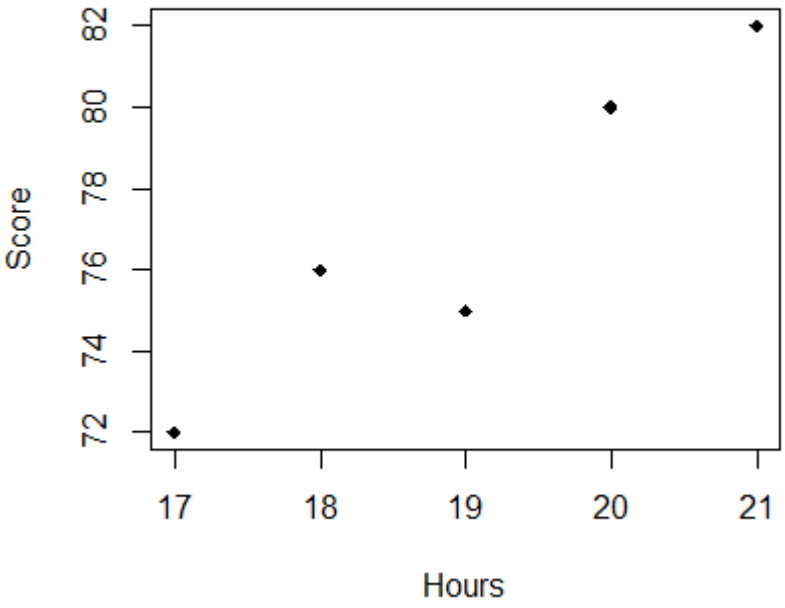
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

Example

You want to study the relationship between hours of study and exam score. Listed below are data from a random sample of 5 students.

Hours (X):	17	18	19	20	21
Score (Y):	72	76	75	80	82

First, we draw a 'scatter plot' as follows:



The LS estimates of regression parameters are:

$\hat{\beta} = 2.4, \;\; \hat{\alpha} = 31.4$

Fitted regression line is as follows:

Fitted Average Exam Score = 31.4 + 2.4 Study Time

For students who study for 10 hours per week, the average exam score is:

$31.4 + 2.4 \times 10 = 55.4$

Interpretation of $\hat{\beta}$:

For 1 unit increase in 'study time', the average exam score increases $\hat{\beta}$ units.

Distribution of the estimators

We assume, $Y_i \sim N(\alpha + \beta x,\ \sigma^2)$. It can be shown that, $E(\hat{\beta}) = \beta$. That is, $\hat{\beta}$ is an unbiased estimator of β. The variance of $\hat{\beta}$ is given by

$$\frac{\sigma^2}{\sum(x - \bar{x})^2}$$

Thus, $\hat{\beta} \sim N(\beta,\ \ \sigma^2 / \sum(x - \bar{x})^2)$.

Again, it can be shown that, $E(\hat{\alpha}) = \alpha$. That is, $\hat{\alpha}$ is an unbiased estimator of α. The variance of $\hat{\alpha}$ is given by

$$\frac{\sigma^2 \sum x^2}{n\ \sum(x - \bar{x})^2}$$

Thus, $\hat{\alpha} \sim N(\alpha,\ \ \sigma^2 \sum x^2 / (n\ \sum(x - \bar{x})^2)\ )$.

The sum of squares of residuals

$$SS_e = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

is divided by $n - 2$ to estimate the error variance σ^2. It can be shown that

$$\frac{SS_e}{\sigma^2} \sim \chi^2_{n-2}$$

That is, SS_e / σ^2 has a chi-square distribution with $n - 2$ d.f.

Inference about $\boldsymbol{\beta}$

For a simple linear regression model, we want to test the hypothesis:

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

Test statistic:

$$Z = \frac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / \sum(x - \bar{x})^2}}$$

Since σ^2 is unknown, we use $SS_e/(n - 2)$ as its estimate. Then the statistic is

$$T = \frac{\hat{\beta} - \beta}{\sqrt{SS_e/((n-2)\sum(x-\bar{x})^2)}}$$

which follows t distribution with $n - 2$ d.f.

Exercise

An individual claims that the fuel consumption of his automobile does not depend on how fast the car is driven. To test this hypothesis, the car was driven at various speeds between 45 and 70 miles per hour. The miles per gallon attained at each of these speeds was determined, with the following data resulting:

Speed	Miles per Gallon
45	24.2
50	25.0
55	23.3
60	22.0
65	21.5
70	20.6
75	19.8

What is your conclusion?

Solution (main results)

$\hat{\beta} = -0.17$

H_0: $\beta = 0$

H_1: $\beta \neq 0$

$T = -8.14$

$t_{0.025,5} = -2.57$

(Note that this is a two-sided test.)

H_0 can be rejected. The fuel consumption of this automobile depends on how fast the car is driven.

Coefficient of determination

The coefficient of determination is given by

SSy=sum{Xi-ave(x))2

$$R^2 = \frac{SS_Y - SS_e}{SS_Y} = 1 - \frac{SS_e}{SS_Y}$$

Here, SS_e/SS_Y is the proportion of variation in Y unexplained by the covariates. Thus, R^2 is the proportion of variation in Y explained by the covariates. The value of R^2 is often used as an indicator of how well the regression model fits the data, with a value near 1 indicating a good fit, and a value near 0 indicating a poor fit.

Sample correlation coefficient

The sample correlation coefficient r measures the degree of linear relationship between x and Y. It is defined as

$$r = \frac{\sum(x - \bar{x})(Y - \bar{Y})}{\sqrt{\sum(x - \bar{x})^2}\sqrt{\sum(Y - \bar{Y})^2}}$$

Properties of r

1. $-1 \leq r \leq 1$.
2. When r is positive: If x increases, then Y increases.
   When r is negative: If x increases, then Y decreases.
3. When r is close to -1 or $+1$, the linear relationship is strong.
   When r is close to zero, the linear relationship is weak.

For simple linear regression, it can be shown that $r^2 = R^2$