Chapter 7
Simple linear regression and correlation

Department of Statistics and Operations Research

December 1, 2019

Plan

1. Correlation

2. Simple linear regression

   - Least Squares and the Fitted Model
   - Properties of the Least Squares Estimators
   - Inferences Concerning the Regression Coefficients
   - Hypothesis Testing on the Slope
   - Measuring Goodness-of-Fit: the Coefficient of Determination

Plan

We consider the problem of measuring the relationship between the two variables X and Y. We want to determine whether large values of X are associated with large values of Y, and vice versa. Correlation analysis attempts to measure the strength of such relationships between two variables X and Y by means of a single number called a correlation coefficient.

Definition 1 [Coefficient of corroleation]

The measure of linear association r between two variables X and Y is estimated by the sample correlation coefficient r, where

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

with $S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$, $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$ and

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

Example 2

Let consider the following grades of 6 students selected at random

Mathematics grade	70	92	80	74	65	83
English grade	74	84	63	87	78	90

We have $n = 6$

$$\overline{X} = \frac{1}{6}(70 + 92 + 80 + 74 + 65 + 83) = 77.33$$

$$\overline{Y} = \frac{1}{6}(74 + 84 + 63 + 87 + 78 + 90) = 79.33$$

$$\sum_{i=1}^{6} x_i^2 = 70^2 + 92^2 + 80^2 + 74^2 + 65^2 + 83^2 = 36354$$

$$\sum_{i=1}^{6} y_i^2 = 74^2 + 84^2 + 63^2 + 87^2 + 78^2 + 90^2 = 38254$$

We get

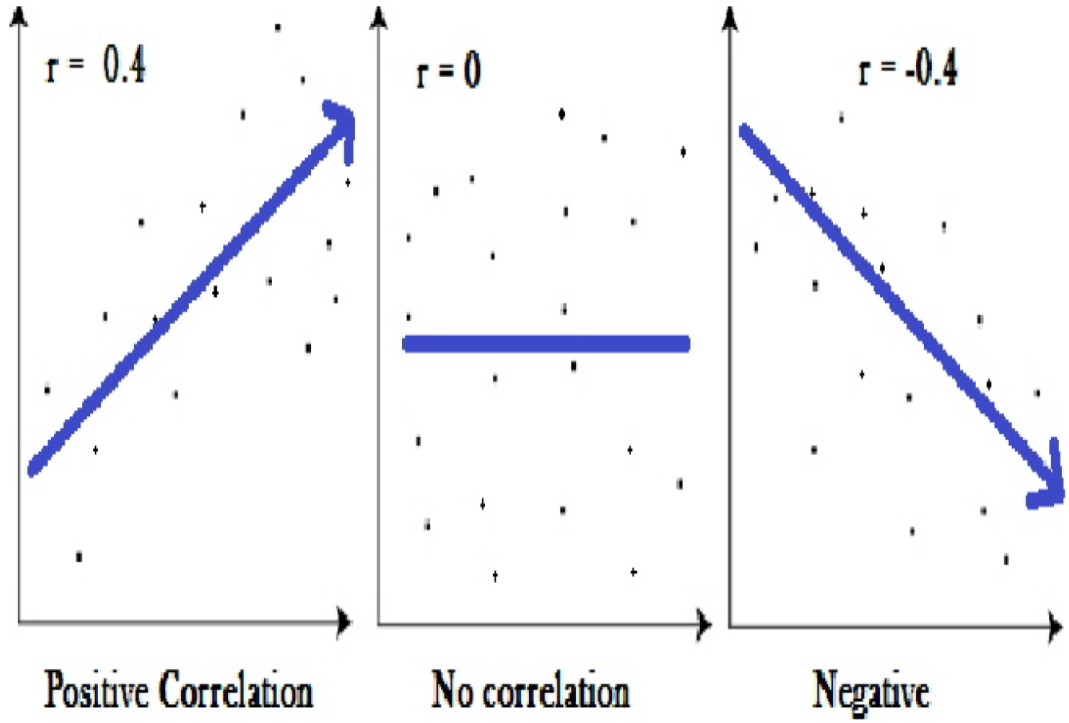
$$S_{xy} = \sum_{i=1}^{6} x_i y_i - n\overline{X}\ \overline{Y} = 36926 - 6 * 77.33 * 79.33 = 115.33$$

$$S_{xx} = \sum_{i=1}^{6} x_i^2 - n\overline{X}^2 = 36354 - 6 * (77.33)^2 = 471.33$$

$$S_{yy} \sum_{i=1}^{6} y_i^2 - n\overline{Y}^2 = 38254 - 6 * (79.33)^2 = 491.33.$$

Hence

$$r = \frac{115.33}{\sqrt{(471.33)(491.33)}} = 0.24.$$



Properties of r

1. $r = 1$ iff all (x_i, y_i) pairs lie on straight line with positive slope,
2. $r = -1$ iff all (x_i, y_i) pairs lie on a straight line with negative slope.

Plan

1. Correlation
2. Simple linear regression
   - Least Squares and the Fitted Model
   - Properties of the Least Squares Estimators
   - Inferences Concerning the Regression Coefficients
   - Hypothesis Testing on the Slope
   - Measuring Goodness-of-Fit: the Coefficient of Determination

The form of a relationship between the response Y (the dependent or the response variable) and the regressor X (the independent variable) is in mathematically the linear relationship

$$Y = \beta_0 + \beta_1 X + \varepsilon_i$$

where, β_0 is the intercept, β_1 the slope and ε_i, the error term in the model, is a random variable with mean 0 and constant variance. An important aspect of regression analysis is to estimate the parameters β_0 and β_1 (i.e., estimate the so-called regression coefficients). The method of estimation will be discussed in the next section. Suppose we denote the estimates b_0 for β_0 and b_1 for β_1. Then the estimated or fitted regression line is given by

$$\hat{y} = b_0 + b_1 x$$

where $\hat{y}$ is the predicted or fitted value.

Definition 3

Given a set of regression data $\{(x_i, y_i); i = 1, 2, ..., n\}$ and a fitted model, $\hat{y}_i = b_0 + b_1 x_i$, the ith residual e_i is given by

$$e_i = y_i - \hat{y}_i,\ i = 1, 2, ..., n.$$

We shall find b_0 and b_1, the estimates of β_0 and β_1, so that the sum of the squares of the residuals is a minimum. This minimization procedure for estimating the parameters is called the method of least squares. Hence, we shall find b_0 and b_1 so as to minimize

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y - \hat{y}_i)^2 = \sum_{i=1}^{n} (y - b_0 - b_1 x_i)^2$$

SSE is called the error sum of squares.

Differentiating SSE with respect to b_0 and b_1, we get:

Theorem 4

Given the sample $\{(x_i, y_i); i = 1, 2, ..., n\}$, the least squares estimates b_0 and b_1 of the regression coefficients β_0 and β_1 are computed from the formulas

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\overline{x}\,\overline{y}}{\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2}$$

$$b_0 = \overline{y} - b_1\overline{x}$$

Example 5

Consider the experimental data in Table the following table (Table 7.1), which were obtained from 33 samples of chemically treated waste in a study conducted at Virginia Tech. Readings on x, the percent reduction in total solids, and y, the percent reduction in chemical oxygen demand, were recorded. We denote by

x: Solids Reduction

y: Oxygen Demand

x (%),	y(%)	x (%),	y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

The estimated regression line is given by

$$\hat{y} = 4.0315 + 0.8895x.$$

Using the regression line, we would predict a 31% reduction in the chemical oxygen demand when the reduction in the total solids is 30%. The 31% reduction in the chemical oxygen demand may be interpreted as an estimate of the population mean $\mu_{Y|30}$ or as an estimate of a new observation when the reduction in total solids is 30%.

We are interested in the expectation and variance the estimator b_1 of β_1 and the expectation of b_0 the estimator of β_0.

Theorem 6

We have

1. $E(b_0) = \beta_0$, $E(b_1) = \beta_1$,
2. $var(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$.

Theorem 7

An unbiased estimate of σ^2, named the mean squared error, is

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

We assume now that the errors ε_i are normally distributed.

Theorem 8

A $100(1-\alpha)\%$ confidence interval for the parameter β_1 in the regression line

$$b_1 - t_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{S_{xx}}}$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-2$ degrees of freedom.

Example 9

Find a 95% confidence interval for β_1 in the regression line , based on the pollution data of Table 7.1.

Solution

We show that

$$\widehat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{418.5935}{33} = 13.503.$$

Therefore, taking the square root, we obtain $\widehat{\sigma} = 3.6746$. Also,

$$S_{xx} = \sum_{i=1}^{n=33} x_i^2 - n\overline{x}^2 = 41086 - 33 * 33.4545 = 4152.28,$$

where

$$\overline{X} = \frac{1}{33}\sum_{i=1}^{33} x_i = 33.4545 \quad \textit{and} \quad \sum_{i=1}^{33} x_i^2 = 41086.$$

Using Table of the t-distribution, we find that $t_{0.025} \approx 2.042$ for 31 degrees of freedom. Therefore, a 95% confidence interval for β_1 is

$$0.8895 - (2.042) * \frac{\sqrt{(13.503)}}{\sqrt{(4152.28)}} < \beta_1 < 0.8895 - (2.042) * \frac{\sqrt{(13.503)}}{\sqrt{(4152.28)}}$$

which simplifies to

$$0.77305 < \beta_1 < 1.0059.$$

To test the null hypothesis H_0 that $\beta_1 = \gamma$ (value) , we again use the t-distribution with $n-2$ degrees of freedom to establish a critical region and then base our decision on the value of

$$t = \frac{b_1 - \gamma}{\widehat{\sigma}/\sqrt{S_{xx}}},$$

which is t-distribution with $n-2$ degrees of freedom.

Example 10

Using the estimated value $b_1 = 0.8895$ of Example 9, test the hypothesis that $\beta_1 = 1.0$ against the alternative that $\beta_1 < 1.0$.

Solution

The hypotheses are $H_0 : \beta_1 = 1.0$ vs $H_1 : \beta_1 < 1.0$. So

$$t = \frac{0.8895 - 1.0}{\sqrt{(\frac{13.503}{4152.18})}} = -1.94,$$

with $n - 2 = 31$ degrees of freedom.
Decision: We accept H_0 because $t = -1.94 < 1.697 = t(0.95, 31)$.

One important t-test on the slope is the test of the hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. When the null hypothesis is not rejected, the conclusion is that there is no significant linear relationship between $E(y)$ and the independent variable x.
Rejection of H_0 above implies that a significant linear regression exists.
A goodness-of-fit statistic is a quantity that measures how well a model explains a given set of data. A linear model fits well if there is a strong linear relationship between x and y.

Definition 11

The coefficient of determination, R^2, is given by

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

where $SSE = \sum_{i=1}^{n}(y - \hat{y}_i)^2$ and $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

Note that if the fit is perfect, all residuals $y - \hat{y}_i$ are zero, and thus $R^2 = 1.0$. But if SSE is only slightly smaller than SST, $R^2 \approx 0$. In the example of table 7.1, the coefficient of determination $R^2 = 0.913$, suggests that the model fit to the data explains 91.3% of the variability observed in the response, the reduction in chemical oxygen demand.

Theorem 12

The square r^2 of the sample correlation coefficient gives the value of the coefficient of determination R^2 that would result from fitting the simple linear regression model.