Chapter 6 Regression

6.1. Least Squares Estimation

Regression models relate a response variable to one or several predictors. Having observed predictors, we can forecast the response by computing its regression function, given all the available predictors.

Dependent Variable

1. Also known as variable *Y*.
2. It measures the outcome of a study.
3. It is the response variable that is being predicted or estimated.

Independent Variable

1. Also known as variable *X*.
2. It attempts to explain the variation in *Y*.
3. It is the predictor/explanatory variable.

The Regression Model

A (simple) regression model that gives a straight-line relationship between two variables is called a ***linear regression model*** given by

$$y = \beta_0 + \beta_1 x + \varepsilon \,,$$

where y = dependent variable
x = independent variable
β_0 = intercept
β_1 = slope
ε = random error.

Assumptions.

1. The error terms ε are normally and independently distributed with mean zero and constant variance σ^2, namely $\varepsilon \sim NID(0, \sigma^2)$.
2. The errors are uncorrelated with each other. Hence, the responses y are also uncorrelated with each other.
3. $E(y \mid x) = \beta_0 + \beta_1 x$ and $Var(y \mid x) = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2$.

Method of Least Squares

The least squares regression line (or the fitted linear regression model), given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \,,$$

where $\hat{\beta}_1 = \dfrac{S_{XY}}{S_{XX}}$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum y}{n} - \hat{\beta}_1 \left[\frac{\sum x}{n} \right]$$

$$S_{XY} = \sum (x-\bar{x})(y-\bar{y}) = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}$$ * can be positive or negative

$$S_{XX} = \sum (x-\bar{x})^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$ * always positive

is a straight line that best represents the relationship between *x* and *y* so that the ***error sum of squares***, denoted by *SSE* such that

$$SSE = \sum (y - \hat{y})^2$$

is **minimized**. The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ which give the minimum *SSE* are called the *least squares estimates* of β_0 and β_1 respectively.

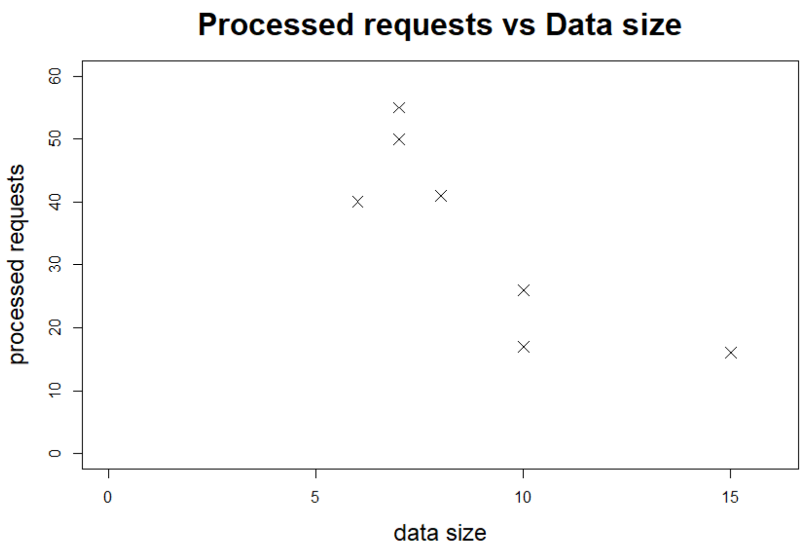
Example 6.1.

A computer manager needs to know how efficiency of her new computer program depends on the size of incoming data. Efficiency will be measured by the number of processed requests per hour. Applying the program to data sets of different sizes, she gets the following results,

Data size (gigabytes), *x*	6	7	7	8	10	10	15
Processed request, *y*	40	55	50	41	17	26	16

In general, larger data sets require more computer time, and therefore, fewer requests are processed within 1 hour. The response variable here is the number of processed requests (*y*), and we attempt to predict it from the size of a data set (*x*).

The corresponding scatterplot is given as



(a) Estimate the least square regression line for the data.
(b) What do the estimated parameters in *part (a)* mean?
(c) What are the estimated number of processed requests per hour when the incoming data is 12 gigabytes?

Solution:

(a) $\sum x = 63,\ \sum x^2 = 623,\ \sum y = 245,\ \sum y^2 = 10027,\ \sum xy = 1973$

(b) We are expected to process 72.287 requests per hour when the incoming data set is 0 gigabyte. However, it is not practical.

Note that $x = 0$ is outside the range of values included in the sample and, therefore, should not be used to estimate the number processed requests. The data size ranged from 6 to 15 gigabytes, so estimates should be limited to that range.

Notice the negative slope. It means that increasing incoming data sets by 1 gigabyte, we expect to process 4.143 fewer requests per hour.

Extrapolation

Using the regression line to predict the value of a response corresponding to an x value that is outside the range of the data used to determine the regression line. *Extrapolation* can lead to unreliable predictions.

Example 6.2.

The growth of children from early childhood through adolescence generally follows a linear pattern. Data on the heights of female during childhood, from four to nine years old, were compiled and the least squares regression line was obtained as $\hat{y} = 31.496 + 2.3622x$ where Y denotes height in inches and X denotes age in years.

(a) Interpret the value of the estimated slope $\hat{\beta}_1 = 2.3622$.

(b) Would interpretation of the value of the estimated intercept $\hat{\beta}_0 = 31.496$ make sense here? If yes, interpret it. If no, explain why not.

(c) What would you predict the height to be for a female at 8 years old?

(d) What would you predict the height to be for a female at 25 years old?

(e) Why do you think your answer to *part (d)* was so inaccurate?

Solution:

6.2. Regression and Correlation

Linear Correlation Coefficients

The correlation coefficient calculated for the population data denoted by ρ (Greek letter *rho*) and for sample data denoted by r, measure the strength of the **linear relationship** between two variables. The value of the correlation coefficient always lies in the range from -1 to 1, that is,

$$-1 \le \rho \le 1 \quad \text{and} \quad -1 \le r \le 1.$$

Pearson Product Moment Correlation Coefficient, r

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}}$$

where $S_{XY} = \sum xy - \dfrac{\sum x \sum y}{n}$ *can be positive or negative

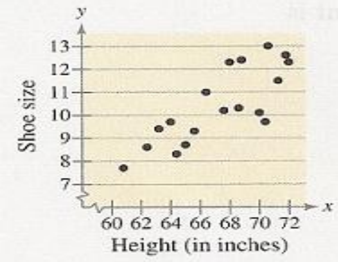
$S_{XX} = \sum x^2 - \dfrac{\left(\sum x\right)^2}{n}$ *always positive

$S_{YY} = \sum y^2 - \dfrac{\left(\sum y\right)^2}{n}$. *always positive

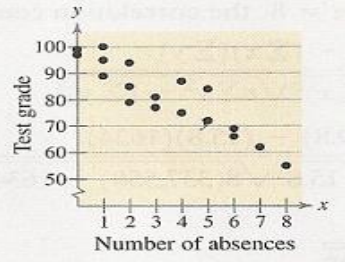
Properties of correlation coefficient.

1. A correlation of 1.0 shows that all of the variance can be explained.
2. A correlation of zero shows that none of the variance can be explained or there is **no linear relationship** between the two variables.
3. The closer to 1 (or -1) the stronger the relationship; the closer to 0 the weaker the relationship.
4. Positive values indicate a positive relationship.
5. Negative values indicate a negative relationship.

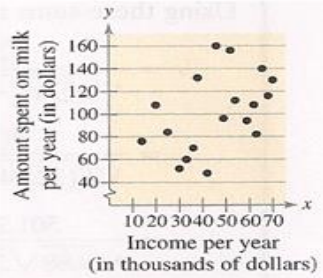
Examples:



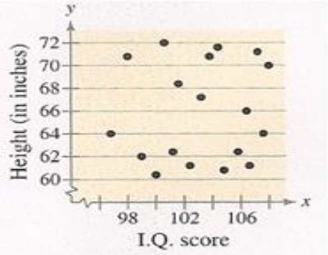
Strong positive correlation
$r = 0.81$



Strong negative correlation
$r = -0.92$



Weak positive correlation
$r = 0.45$



No correlation
$r = 0.04$

Degree of Correlation

Degree of correlation	**Positive correlation**	**Negative correlation**
Perfect	+1	- 1
Strong	$0.8 \le r < 1.0$	$-1.0 < r \le -0.8$
Moderate	$0.4 \le r < 0.8$	$-0.8 < r \le -0.4$
Weak	$0 < r < 0.4$	$-0.4 < r < 0$
Absent	0	0

<u>*Example 6.3.*</u>
Reconsider *Example 6.1.*. Compute the correlation coefficient *r*. Comment on your result.
Solution:
$\sum x = 63,\ \sum x^2 = 623,\ \sum y = 245,\ \sum y^2 = 10027,\ \sum xy = 1973$

Estimated Regression Slope

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = r\sqrt{\frac{S_{YY}}{S_{XX}}}$$

6.3. Analysis of Variance Approach to Regression Analysis

The total variation among observed responses is measured by the ***total sum of squares***

$$\begin{aligned} SST &= \sum (y-\bar{y})^2 \\ &= S_{YY} \\ &= \sum y^2 - \frac{\left(\sum y\right)^2}{n} \end{aligned}$$

A portion of this total variation is attributed to predictor X and the regression model connecting predictor and response. This portion is measured by the ***regression sum of squares***

$$\begin{aligned} SSR &= \sum (\hat{y}-\bar{y})^2 \\ &= \sum \hat{\beta}_1^2 (x-\bar{x})^2 \\ &= \hat{\beta}_1^2 S_{XX} \\ &= \hat{\beta}_1 S_{XY} \end{aligned}$$

This is the portion of total variation *explained* by the model.

The rest of total variation is attributed to "error". It is measured by the ***error sum of squares***

$$\begin{aligned} SSE &= \sum (y-\hat{y})^2 \\ &= \sum e^2 \end{aligned}$$

This is the portion of total variation *not explained* by the model.

Regression and error sums of squares partition SST in two parts, $SST = SSR + SSE$.

A standard way to present analysis of variance is the *ANOVA table*.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	F
Regression	SSR	1	MSR	$F = \frac{MSR}{MSE}$
Error	SSE	$n-2$	MSE	
Total	SST	$n-1$		

where

$$MSR = \frac{SSR}{1} = SSR$$

and

$$MSE = \frac{SSE}{n-2} = s^2 .$$ ***Sample Regression Variance***

The $F-$ratio, $F = \frac{MSR}{MSE}$ is used to test **significance** of the entire regression model.

Note. The ANOVA test is always a **right-tailed test**.

The F Distribution

1. The *F* distribution is continuous and skewed to the right.
2. The value of *F* is nonnegative (ie: always positive).
3. The *F* distribution depends on two number of degrees of freedom, (ie: ν_1 and ν_2), and is denoted by $F(\nu_1, \nu_2)$.
4. Reject H_0 if $F > F_{\alpha,\nu_1,\nu_2}$.

ANOVA F-test

A popular method of testing significance of a model is the *ANOVA F-test*. It compares the portion of variation explained by regression with the portion that remains unexplained. Significant models explain a relatively large portion.

For a given significance level α, F-test of

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

is equivalent algebraically to the two-tailed t-test.

$$F = \frac{MSR}{MSE} = \frac{\hat{\beta}_1^2 S_{XX}}{MSE} = \frac{\hat{\beta}_1^2}{MSE/S_{XX}} = t^2 .$$

Hence, both tests give us the same result.

<u>*Example 6.4.*</u>

Reconsider *Example 6.1.*. Construct the ANOVA table and test for significance of regression by using the significance level, $\alpha = 0.05$.

Solution:

$\sum x = 63,\ \sum x^2 = 623,\ \sum y = 245,\ \sum y^2 = 10027,\ \sum xy = 1973,\ n = 7$

$S_{XX} = 56$, $S_{YY} = 1452$, $S_{XY} = -232$, $\hat{\beta}_1 = -4.143$, $\hat{\beta}_0 = 72.287$

Coefficient of Determination

The *goodness of fit*, appropriateness of the predictor and the chosen regression model can be judged by the proportion of SST that the model can explain.

The *coefficient of determination*, denoted by R^2, represents the proportion of SST that is explained by the use of the linear regression model such that

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

It is always between 0 and 1, with high values generally suggesting a good fit.

In univariate regression,

$$R^2 = r^2.$$

The computational formula for R^2 is

$$R^2 = \frac{S_{XY}^2}{S_{XX}\ S_{YY}} \qquad \text{and} \qquad 0 \le R^2 \le 1.$$

Note.

1. R^2 measures the proportion of variation in *Y* that explained by the regressor variable *X*. i.e. $R^2 \times 100\%$ of the variation in *Y* can be explained by using *X* to predict *Y*.
2. $F = \dfrac{R^2/1}{\left(1-R^2\right)/\left(n-2\right)} = \dfrac{\left(n-2\right)R^2}{1-R^2}.$
3. Use R^2 as a measure of fit when the sample size is substantially larger than the number of variables in the model; otherwise, R^2 may be artificially high.
4. R^2 is only measuring the linear relationship.
5. R^2 is a measure of how the estimated regression line fits in the sample only.

Example 6.5.

Reconsider *Example 6.4.*. Find the coefficient of determination and explain what it means.

Solution:

6.4. Inference About The Regression Slope

The Sampling Distribution for β₁

Population regression model:

$$y = \beta_0 + \beta_1 x + \varepsilon \,, \qquad \varepsilon \sim NID(0, \sigma^2)\,.$$

We have

$$y \sim NID(\beta_0 + \beta_1 x, \sigma^2)$$

and

$$\hat{\beta}_1 \sim NID\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right).$$

Interval Estimation of β₁

If σ^2 is not known, we estimate the standard error of $\hat{\beta}_1$ by

$$se\left(\hat{\beta}_1\right) = \sqrt{\frac{MSE}{S_{XX}}}\,, \qquad MSE = \frac{SSE}{n-2}\,.$$

The $(1-\alpha)100\%$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2,\, n-2} \times se\left(\hat{\beta}_1\right).$$

Hypothesis Testing About β₁

To test the hypothesis that the slope equals a constant, we have

$$H_0 : \beta_1 = B$$

and the test statistic is

$$t = \frac{\hat{\beta}_1 - B}{se\left(\hat{\beta}_1\right)} \sim t_{\alpha, \nu = n-2}\,.$$

A non-zero slope indicates **significance** of the model, relevance of predictor *X* in the inference about response *Y*, and the existence of a linear relation among them. It means that a change in *X* causes changes in *Y*. In the absence of such relation, $E\left(y\right) = \beta_0$ remains constant.

To test the hypothesis that *X* does not determine *Y* linearly, we will test the null hypothesis that the slope of the regression line is zero, that is,

$$H_0 : \beta_1 = 0\,.$$

The alternative hypothesis can be

1. $H_1 : \beta_1 \neq 0$, *X* determines *Y*
2. $H_1 : \beta_1 > 0$, *X* determines *Y* positively
3. $H_1 : \beta_1 < 0$, *X* determines *Y* negatively

Note. $SSE = S_{YY} - \hat{\beta}_1 S_{XY}$

Example 6.6.

Reconsider *Example 6.1.*. Test at the 5% significance level whether the slope of the regression line is significant. Does the number of processed requests really depend on the size of data sets? Justify your answer.

Solution:

$\sum x = 63,\ \sum x^2 = 623,\ \sum y = 245,\ \sum y^2 = 10027,\ \sum xy = 1973,\ n = 7$

$S_{XX} = 56,\quad S_{YY} = 1452,\quad S_{XY} = -232,\quad \hat{\beta}_1 = -4.143,\quad \hat{\beta}_0 = 72.287,\quad MSE = 98.16$

6.5. Prediction

One of the main applications of regression analysis is making forecasts, predictions of the response variable *Y* based on the known or controlled predictors *X*.

Confidence Interval for the Mean of Response

Population regression model:

$$y = \beta_0 + \beta_1 x + \varepsilon , \qquad \varepsilon \sim NID(0, \sigma^2) .$$

Let x_h be the value of the predictor X . The corresponding value of the response Y is computed by evaluating the estimated regression line at x_h ,

$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 \left(x_h \right) .$$

Then the expectation is

$$\mu_{Y|x_h} = E\left(\hat{y}_h\right) = E\left[\hat{\beta}_0 + \hat{\beta}_1\left(x_h\right)\right] = \beta_0 + \beta_1\left(x_h\right)$$

and variance

$$\sigma^2_{\hat{Y}_h} = Var\left(\hat{y}_h\right) = \sigma^2\left(\frac{1}{n} + \frac{\left(x_h - \overline{x}\right)^2}{S_{XX}}\right).$$

We estimate the regression variance σ^2 by s^2 ,

$$MSE = \frac{SSE}{n-2} = s^2 .$$

The $(1-\alpha)100\%$ confidence interval for $\mu_{Y|x_h}$ is

$$E\left(\hat{y}_h\right) \pm t_{\alpha/2,\, n-2} \times \sqrt{MSE\left(\frac{1}{n} + \frac{\left(x_h - \overline{x}\right)^2}{S_{XX}}\right)} .$$

Prediction Interval for the Individual Response

Often we are more interested in predicting the actual response rather than the mean of all possible responses. Instead of estimating a *population parameter*, we are now predicting the *actual value* of a random variable.

If x_h is the value of the regressor of interest, the point estimate of the new value of the response, y_h is

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 \left(x_h \right) .$$

Note that the random variable

$$Y_h - \hat{Y}_h \sim N\left(0,\ \sigma^2\left[1 + \frac{1}{n} + \frac{\left(x_h - \overline{x}\right)^2}{S_{XX}}\right]\right).$$

The $(1-\alpha)100\%$ prediction interval for Y_h at a specified value of $x = x_h$ is

$$\hat{Y}_h \pm t_{\alpha/2,\, n-2} \times \sqrt{MSE\left(1 + \frac{1}{n} + \frac{\left(x_h - \overline{x}\right)^2}{S_{XX}}\right)} .$$

Example 6.7.

Reconsider *Example 6.1*..

(a) Compute a 95% confidence interval for the mean number of processed requests per hour for 16 gigabytes of data.

(b) Compute a 95% prediction interval for the number of processed requests within 1 hour for a randomly selected data size of 16 gigabytes.

Solution:

$\sum x = 63, \sum x^2 = 623, \sum y = 245, \sum y^2 = 10027, \sum xy = 1973, n = 7$

$S_{XX} = 56$, $S_{YY} = 1452$, $S_{XY} = -232$, $\hat{\beta}_1 = -4.143$, $\hat{\beta}_0 = 72.287$, $MSE = 98.16$