Chapter 7

Linear regression

Consider the following linear relationship, or *population regression equation*:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $i = 1, 2...n$ are individual observations from a sample, y_i and x_i are, respectively, dependent and explanatory variables, and where ϵ_i is an iid random error term. Finally the parameters β_0 and β_1 are unknown and have to be estimated.

7.1. OLS estimators

The Ordinary Least Squares (OLS) estimators of β_0 and β_1, $\hat{\beta}_0$ and $\hat{\beta}_1$, are obtained by fitting a line through the data minimising the sum of squared residuals, i.e.

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \quad S(\hat{\beta}_0, \hat{\beta}_1) = \min_{\hat{\beta}_0, \hat{\beta}_1} \quad \sum_{i=1}^{n} {\hat{\epsilon}_i}^2$$

where

$$\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

To derive the OLS estimator, we need to solve the following First Order Conditions (FOC), i.e.

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = 0 \implies -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = 0 \implies -2 \sum_{i=1}^{n} [x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)] = 0$$

From the first FOC, we get:

$$\begin{aligned}
-2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \implies \\
\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \implies \\
\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{\beta}_0 - \sum_{i=1}^{n}\hat{\beta}_1 x_i = 0 \implies \\
\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \implies \\
\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i \implies \\
\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}
\end{aligned}$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Substituting the solution for $\hat{\beta}_0$ into the second FOC, we get:

$$\begin{aligned}
-2\sum_{i=1}^{n}[x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)] = 0 \implies \\
\sum_{i=1}^{n}[x_i(y_i - (\overline{y} - \hat{\beta}_1\overline{x}) - \hat{\beta}_1 x_i)] = 0 \implies \\
\sum_{i=1}^{n}[x_i(y_i - \overline{y}) - \hat{\beta}_1 x_i(x_i - \overline{x})] = 0 \implies \\
\sum_{i=1}^{n}\hat{\beta}_1 x_i(x_i - \overline{x}) = \sum_{i=1}^{n} x_i(y_i - \overline{y}) \implies \\
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n} x_i(x_i - \overline{x})}
\end{aligned}$$

Notice that:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x}) = \sum_{i=1}^{n}(x_i x_i - 2\overline{x}x_i + \overline{x}\overline{x}) = \\
\sum_{i=1}^{n} x_i x_i - 2\sum_{i=1}^{n}\overline{x}x_i + \sum_{i=1}^{n}\overline{x}\overline{x} = \\
\sum_{i=1}^{n} x_i x_i - \overline{x}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} x_i + n\overline{x}\overline{x} = \\
\sum_{i=1}^{n} x_i x_i - \overline{x}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} x_i + \overline{x}\sum_{i=1}^{n} x_i = \\
\sum_{i=1}^{n} x_i x_i - \overline{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i(x_i - \overline{x})
\end{aligned}$$

By the same argument,

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i(y_i - \overline{y})$$

Substituting back into the formula for $\hat{\beta}_1$, we get:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n} x_i(x_i - \overline{x})} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})} = \frac{\text{cov}[x, y]}{\text{var}[x]}$$

i.e. $\hat{\beta}_1$ is the ratio of the sample covariance between x and y to the sample variance of x. Since the variance is always positive, $\hat{\beta}_1$ will have the same sign as the covariance (and the correlation coefficient). Plugging $\hat{\beta}_1$ into $\hat{\beta}_0$, we get

$$\hat{\beta}_0 = \overline{y} - \frac{\text{cov}[x, y]}{\text{var}[x]}\overline{x}$$

7.1.1 Properties

Unbiasedness The OLS estimators are unbiased, i.e.

$$\text{E}[\hat{\beta}_0] = \beta_0$$
$$\text{E}[\hat{\beta}_1] = \beta_1$$

The Gauss-Markov Theorem. Among all linear unbiased estimators, OLS estimators have the smallest variance, i.e. OLS estiamators are efficient estimators. The variance of any other linear unbiased estimator must be larger than that of the OLS estimator. We say that OLS estimators are BLUE (best linear unbiased estimator)

Consistency The OLS estimators are consistent. Since they are also unbiased, this implies that:

$$\lim_{n\to\infty} \text{VAR}[\hat{\beta}_0] = 0$$
$$\lim_{n\to\infty} \text{VAR}[\hat{\beta}_1] = 1$$

7.1.2 Goodness of Fit

We can measure the goodness of fit of a regression model by comparing the variation that has been explained by the model to the total variation in the data. We define the Total Sum of Squares (TSS) to be the total (squared) variation of the y_i values about their mean $\bar{y}$, such that

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

and we define the Explained Sum of Squares (ESS) to be the total (squared) variation of the fitted values $\hat{y}_i$ about their mean $\bar{y}$, such that

$$\text{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Finally, we define the Residual Sum of Squares (RSS) to be the total (squared) unexplained variation in our model, i.e. the total squared difference between the y_i values and the fitted $\hat{y}_i$ values,

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (\epsilon_i)^2$$

It can be shown that $\text{TSS} = \text{ESS} + \text{RSS}$. Intuitively, we can think of the total variation of the data (TSS) being decomposed into a part that is explained by the fitted model (ESS) and a part that is unexplained by the fitted model (RSS).
Therefore, a measure of fitness is

$$\text{R}^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

R^2 is called coefficient of determination. Notice that

$$0 \leq \text{ESS} \leq \text{TSS} \quad \text{and} \quad 0 \leq \text{RSS} \leq \text{TSS} \implies 0 \leq \text{R}^2 \leq 1$$

It can be shown that R^2 is equivalent to the squared sample correlation coefficient between y and x, i.e.

$$\text{R}^2 = \left(\frac{\text{COV}[x, y]}{s_x s_y} \right)^2$$

7.2. Hypothesis Testing of Regression Parameters

Let the regression model be:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Suppose we want to test claims made about the parameters β_1[1]. In particular, consider testing the claim that there is no relationship between x and y, i.e.

$$\text{H}_0 : \beta_1 = 0 \quad \text{versus} \quad \text{H}_0 : \beta_1 \neq 0$$

Assumption: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

This assumption ensure that y_i is normally distributed as well. Since estimator $\hat{\beta}_i$ is

[1] The same argument applies for β_0

a linear function of y_i, it must also follow a normal distribution. Therefore,

$$\hat{\beta}_i \sim \mathcal{N}(\beta_1, \text{VAR}[\beta_1])$$

and we can standardise $\hat{\beta}_i$ in the usual way (by subtracting its mean and dividing by its standard deviation), such that:

$$\frac{\hat{\beta}_i - \beta_1}{\sqrt{\text{VAR}[\beta_1]}} \sim \mathcal{N}(0, 1)$$

It turns out that $\text{VAR}[\beta_1]$, is a function of σ^2 which is unknown. Replacing σ^2 with a sample estimator changes the distribution of the above statistic from standard normal to Student's, i.e.

$$\frac{\hat{\beta}_i - \beta_1}{\sqrt{\widehat{\text{VAR}}[\beta_1]}} \sim t_{n-2}$$

where $\widehat{\text{VAR}}[\beta_1]$ is an estimator of $\text{VAR}[\beta_1]$. This statistic follows a t-distribution with $n-2$ degrees of freedom (instead of $n-1$ in the population mean case) This is because we now estimate two parameters instead of one. Therefore, to test our null hypothesis against the alternative, we have to construct the test statistic, i.e.

$$\frac{\hat{\beta}_i}{\sqrt{\widehat{\text{VAR}}[\beta_1]}}$$

and compare it to critical values from the $t_{\frac{\alpha}{2},t-2}$ distribution. Notice that we can also consider one-sided alternatives of the form

$$\text{H}_0 : \beta_1 = 0 \quad \text{versus} \quad \text{H}_1 : \beta_1 \geq 0$$

and test the hull hypothesis against this alternative using critical values from $t_{\alpha,t-2}$.

7.3. Exercises

Exercise 1 The weight (in kg) and height (in cm) of a sample of 5 individuals is reported below:

Individual	Weight	Height
1	83	181
2	70	175
3	63	165
4	82	193
5	75	178

Calculate the OLS estimates of the following linear regression model:

$$\text{weight}_i = \beta_0 + \beta_1 \text{height}_i + \epsilon_i$$

Exercise 2 The campaign manager for a local politician wants to know if the number of leaflets delivered during an election campaign affects the total number of votes received by a candidate standing for election. She collects data from the last 5 elections. She denotes the number of leaflets delivered as x and the number of votes the candidate received as y. She finds the following results:

$$\begin{aligned} \bar{x} &= 1280 \\ \bar{y} &= 1980 \\ VAR[x] &= 44.5 \\ COV[x, y] &= 25.2 \end{aligned}$$

Calculate the Ordinary Least Squares estimates of the parameters β_0 and β_1 in the following linear regression:

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$$

Exercise 3 Consider the sample data and fitted regression considered in Exercise 2. You are now told that $VAR[y] = 29.7$. Compute R^2 for this regression.

Exercise 4 An economist collects data on GDP (x) and literacy rates (y) for a sample of 30 countries and estimates the following fitted regression model

$$y_i = 12 + 0.2x_i + \hat{\epsilon}_i$$

where the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are 2.3 and 0.08 respectively.

- Test the claim that there is no relationship between GDP and literacy rates at a 5% significance level, against a two-sided alternative.
- Test the claim that β_1 is equal to 0.3 against the one-sided alternative that β_1 is less than 0.3 at a 5% level of significance