S6 Linear Regression

S6.1 Background

In this chapter we assume that we have a collection of sampling units ($i = 1, \ldots, n$) and that, associated with sampling unit i, there are two measurements X_i and Y_i.

We need to investigate the extent to which the variations in the Y_i values can be explained by the X_i.

Notation

In this chapter we will use the following notation:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2 \\ S_{yy} &= \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum x_i^2 - n\overline{y}^2 \\ S_{xy} &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - n\overline{xy} \end{aligned}$$

Note the relationship with sample variances: $S_x^2 = S_{xx}/(n-1)$ and $S_y^2 = S_{yy}/(n-1)$.

S6.2 Covariance and Correlation

The *sample covariance* between the samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ is

$$\frac{1}{n-1}\left(\sum_i x_i y_i - n\overline{xy}\right) = \frac{S_{xy}}{n-1}.$$

The *sample correlation* is

$$r_{XY} = \frac{\text{sample covariance}}{S_x S_y}.$$

(Compare these with the definitions of covariance and correlation between two random variables: see chapter P5.)

Note that $-1 \leq r_{XY} \leq 1$ in all cases.

If r_{XY} is sufficiently far from 0, we conclude that there is a significant linear relationship between X and Y. To test significance we can use the fact that

$$T = \sqrt{n-2}\,\frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sim t_{n-2}$$

if the X and Y samples are in fact independent.

S6.3 The model

The model used for linear regression is

$$Y_i = \alpha + \beta x_i + E_i,$$

where α and β are unknown parameters and the E_i are assumed to have mean 0, variance σ^2 and no dependence on each other. The x_i are assumed deterministic (non-random).
Y is the *dependent variable*, x the *independent* or *explanatory* variable.
Sometimes we select the x values in designing the experiment; sometimes we can only observe them.

S6.4 The scatterplot

The observations (x_i, y_i) are plotted on a two-dimensional graph. If there is a strong linear relationship, it will be apparent.

S6.5 The line of best fit

How well does the line $y = a + bx$ fit the data? "Badness of fit" may be measured by the sum of squares of the residuals:

$$B(a, b) = \sum_{i=1}^{n} (y_i - [a + bx_i])^2.$$

The *line of best fit* is the one with smallest "badness". This is a minimisation problem, solved by calculus.

The normal equations

Differentiating $B(a, b)$ w.r.t. a gives

$$-2 \sum_{i=1}^{n} (y_i - [a + bx_i])$$

and differentiating w.r.t b gives

$$-2 \sum_{i=1}^{n} (x_i y_i - [ax_i + bx_i^2]).$$

Setting these equal to 0, we get

$$a + b\overline{x} = \overline{y} \quad \text{and} \quad a \sum x_i + b \sum x_i^2 = \sum x_i y_i.$$

These are two equations in two unknowns, called the *normal equations*.

The estimators

Use a standard method to solve them. The solution is

$$b = \frac{S_{xy}}{S_{xx}} \text{ and } a = \overline{y} - b\overline{x},$$

where S_{xx} and S_{xy} are defined above.
a and b are known as the *Least Squares Estimators* of α and β.

S6.6 Residuals

A *residual* is the vertical distance from a point to the line of best fit, $e_i = y_i - [a + bx_i] = y_i - \hat{y}_i$, where $\hat{y}_i = a + bx_i$ is the *fitted value*.
The residuals can be used to test how well the model fits the data:

- the e_i should be independent of the x_i: draw a scatterplot and check there is no pattern;
- the e_i should be independent of the fitted values $a + bx_i$: scatterplot again;
- the e_i should be roughly normally distributed: use a histogram to check;
- the variance of the e_i should be independent of x_i: if there is a problem, this will typically show up in the form of a cone shape in the scatterplot of e_i against x_i.

S6.7 Transformation

If residual plots indicate a problem, it is sometimes possible to solve it by transforming the data.
A common transformation is to take the log of the x values; other possibilities are a square root transformation, or taking the log of both x and y values.
This kind of transformation is particularly useful if the variance of the e_i appears dependent on the x values (this phenomenon is called *heteroscedasticity*).

A simple way to decide whether a transformation is useful is to calculate R^2 (see next section) for both untransformed and transformed data and see which is larger.

S6.8 Explanatory power of the model

The Total Sum of Squares is

$$SST = S_{yy} = \sum_i (y_i - \overline{y})^2.$$

This can be broken down into

* the *Error Sum of Squares* (SSE), which is $\sum e_i^2$, the sum of squares of the residuals, or can also be expressed as $SSE = S_{yy} - bS_{xy}$, and
* the *Regression Sum of Squares* (SSR), equal to $SST - SSE$.

The *coefficient of determination*, R^2 is SSR/SST, the proportion of the total variation which is explained by the model. A value of R^2 close to 1 means the fit is good.

In fact R^2 is the square of r_{XY}, the sample correlation between the x and y samples; it can also be written as

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}.$$

S6.9 Estimating σ^2

The error sum of squares (SSE) is the only information available relating to σ^2. We use the estimator $\hat{\sigma}^2 = \frac{1}{n-2} \times SSE$, which has expectation equal to σ^2 and, assuming normality, also has the property that

$$(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

(Note: estimating α and β has cost 2 d.f.)

S6.10 Properties of a and b

The LSEs are functions of random variables, so are themselves random. It can be shown that $\mathbb{E}(a) = \alpha$ and $\mathbb{E}(b) = \beta$.

It is also possible to calculate their variances (note that these do not depend on α and β):

$$\mathrm{Var}\,(b) = \frac{\sigma^2}{S_{xx}}, \quad \mathrm{Var}\,(a) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right).$$

a and b are normally distributed as long as the observations are normally distributed.

S6.11 Significance of the slope

To test whether there is a significant linear relationship between y and x, test $\mathrm{H}_0 : \beta = 0$ against $\mathrm{H}_1 : \beta \neq 0$. (One-sided tests are also possible.) If H_0 is true, then

$$T = \frac{b}{\sqrt{\mathrm{EstVar}\,(b)}} = \frac{b}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_{n-2},$$

so that a simple one-sample t test can be applied.
If H_0 is rejected, we conclude that there is a significant linear association between y and x.

Standard t-distribution techniques also enable us to calculate a confidence interval for β, using the pivotal quantity $(b-\beta)/\sqrt{\hat{\sigma}^2/S_{xx}}$.

We can apply the same techniques to test whether α is equal to 0, or to find a CI for α, but usually α is of less interest than β.

S6.12 Prediction

One of the aims of Linear Regression is to be able to predict what y-value would be observed if the x-value were equal to x_0. Denote this r.v. by $Y(x_0)$.

Cautionary note: it is only safe to use predictions for values of x between $\min x_i$ and $\max x_i$. The relationship between x and y may not be linear at all outside this range: we have no evidence.

First note that $\mathbb{E}[Y(x_0)] = \alpha + \beta x_0$.
Since α and β are unknown, we must use estimates for our predictions: our prediction for $Y(x_0)$ is

$$\hat{Y}(x_0) = a + bx_0.$$

The expectation is right. What is the variance?

If we only want to predict $\mathbb{E}[Y]$, the variance term we need is $\mathbb{E}\left[\left(\hat{Y}(x_0) - \alpha - \beta x_0\right)^2\right]$, which is equal to

$$\sigma^2\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right).$$

But to forecast the value of $Y(x_0)$ which will actually be observed we must take account of the random variation as well:

$$\mathbb{E}\left[\left(\hat{Y}(x_0) - Y(x_0)\right)^2\right] = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right).$$

Since σ^2 is unknown we have to use estimates of the variance in both cases, replacing σ^2 by $\hat{\sigma}^2$ and using the t distribution with $n-2$ d.f.

S6.13 Summary

In this chapter we have

* defined a linear model for bivariate data
* derived estimates of the parameters of the model
* obtained a method for testing whether the parameters are different from 0
* specified graphical procedures for evaluating goodness of fit
* determined a method for predicting the values of future observations