4 Inference about the regression parameters

In our simple linear regression model of

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where the ε_i are iid $\varepsilon_i \sim N(0, \sigma^2)$

we have found (see section 2.2 above) that the least squares estimates of β_0 and β_1 are given by

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

and

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

4.1 Confidence Interval for β_1

We found earlier that the sampling distribution of $\widehat{\beta_1}$ is

$$\widehat{\beta_1} \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

(Note that even where the y_i are not normally distributed the distribution of $\widehat{\beta_1}$ is approximately normal.)

We can standardise the $\widehat{\beta_1}$, that is find the function of $\widehat{\beta_1}$ that follows a standard normal N(0,1) distribution,

$$\frac{\widehat{\beta_1} - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1)$$

However σ^2 is generally not known, so we will need to replace it with it's estimate s^2. When we do this, the normal distribution becomes a Student t-distribution.

That is because, more generally, if $Z \sim$ N(0,1) and $U \sim \chi^2_v$ then $\frac{Z}{\sqrt{U/v}} \sim t_v$

We already have

$$Z = \frac{\widehat{\beta_1} - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1)$$

and we will see later in the course that

$$U = \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

therefore $T = \frac{\frac{\widehat{\beta_1} - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}}}{\sqrt{\frac{(n-2)S^2}{\sigma^2(n-2)}}} = \frac{\widehat{\beta_1} - \beta_1}{\frac{S}{\sqrt{S_{xx}}}} \sim t_{n-2}$

If we have some parameter Θ a 95% confidence interval for Θ means to find boundaries *a* and *b* such that $P(a < \theta < b) = 0.95$. More generally a 100$(1 - \alpha)$% confidence interval for Θ is to find *a* and *b* such that $P(a < \theta < b) = 1 - \alpha$.

In practice, a confidence interval for β_1 will depend on the data and the estimate $\widehat{\beta_1}$ found from that data. Using the Student-t distribution above, and defining $t_{\frac{\alpha}{2}}$ to be the quantity such that

$$P\left(|t_v| < t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

then

$$P\left(\widehat{\beta_1} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{S_{xx}}} < \beta_1 < \widehat{\beta_1} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{S_{xx}}}\right) = 1 - \alpha$$

So for a particular data set where $\widehat{\beta_1}$ and *S* become values from observed data rather than random variables, we can calculate the 100$(1 - \alpha)$% confidence interval $[a, b]$ for β_1 where

$$[a, b] = \left[\widehat{\beta_1} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{S_{xx}}}, \quad \widehat{\beta_1} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{S_{xx}}}\right]$$

4.2 Testing the significance of β_1

In section 3.3 above we saw that we can test the null hypothesis H_0: $\beta_1 = 0$ using the ANOVA table and the *F* statistic. There is another way to test the same null hypothesis based upon how we have derived the confidence interval for β_1.

Under this null hypothesis, the slope is zero and therefore we have a constant model that can be written $y_i = \beta_0 + \varepsilon_i$

We can test this null hypothesis using the test statistic *T* developed above for confidence intervals. If H_0 is true then β_1 is zero and so

$$T = \frac{\widehat{\beta_1}}{\frac{S}{\sqrt{S_{xx}}}} \sim t_{n-2}$$

For a given data set we can calculate the value of *T*. We then reject the null hypothesis H_0: $\beta_1 = 0$ at significance level α if

$$|T| > t_{n-2,\frac{\alpha}{2}}$$

This methodology is in fact equivalent mathematically to the ANOVA table F-statistic route.

Sometimes you will see equations such as those above for the confidence interval and the test statistic *T* written in terms of the estimated *standard error* of $\widehat{\beta_1}$. The standard error $se(\widehat{\beta_1})$ is the square root of the variance of $\widehat{\beta_1}$. Our estimate of the standard error of $\widehat{\beta_1}$ is

$$\widehat{se(\beta_1)} = \sqrt{\frac{S^2}{S_{xx}}}$$

and using the standard error notation, the confidence interval becomes

$$[a,b] = \left[\widehat{\beta_1} - t_{\frac{\alpha}{2}}\, \widehat{se(\beta_1)}\,, \qquad \widehat{\beta_1} + t_{\frac{\alpha}{2}}\, \widehat{se(\beta_1)}\right]$$

and the *T* test statistic is

$$T = \frac{\widehat{\beta_1}}{\widehat{se(\beta_1)}} \sim t_{n-2}$$

4.3 Inference about β_0

Because in modelling we are generally interested in the relationship between *Y* and *X*, we are usually most interested in parameter β_1. We can however also develop confidence intervals and test hypotheses for β_0. We found earlier that the sampling distribution of β_0 is,

$$\widehat{\beta_0} \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}))$$

Using the same methodology as for $\widehat{\beta_1}$ above, we find that the 100$(1-\alpha)$% confidence interval for β_0 is

$$[a,b] = \left[\widehat{\beta_0} - t_{\frac{\alpha}{2}}\, \widehat{se(\beta_0)}\,, \qquad \widehat{\beta_0} + t_{\frac{\alpha}{2}}\, \widehat{se(\beta_0)}\right]$$

and the test statistic to test the null hypothesis H_0: $\beta_0 = B$ for some value B (which may or may not be zero) is

$$T = \frac{\widehat{\beta_0} - B}{\widehat{se(\beta_0)}} \sim t_{n-2}$$

where $\widehat{se(\beta_0)} = \sqrt{S^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}$

4.4 Inference about the mean response

We may also develop confidence intervals and test hypotheses for the mean of the response variable given some value of the explanatory variable, that is $E[Y_i|X_i = x_i]$ which is also often written as μ_i.

Under the simple linear regression model,

$$\mu_i = E[Y_i|X_i = x_i] = \beta_0 + \beta_1 x_i$$

and the least squares estimator of μ_i is given by

$$\widehat{\mu_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

so that for any value of the explanatory variable x_i we can estimate the mean response.

Under the simple linear regression model, the sampling distribution of μ_i is also Normal,

$$\widehat{\mu_i} \sim N(\mu_i,\ \sigma^2(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}))$$

which allows us to obtain a 100$(1-\alpha)$% confidence interval for $\widehat{\mu_i}$ which is

$$[a,b] = \left[\widehat{\mu_i} - t_{\frac{\alpha}{2}}\, \widehat{se(\mu_i)}\,,\qquad \widehat{\mu_i} + t_{\frac{\alpha}{2}}\, \widehat{se(\mu_i)}\right]$$

and we can test the null hypothesis, H_0: $\mu_i = M$ for some value M, with the test statistic

$$T = \frac{\widehat{\mu_i} - M}{\widehat{se(\mu_i)}} \sim t_{n-2}$$

where $\widehat{se(\mu_i)} = \sqrt{S^2(\frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}})}$

We should note here though that the value of x_i should be within the range of observed data values for *X* for this estimation of the mean response to be valid. The model has said nothing about the applicability of the linear regression beyond that data range and this should not be used as a method of extrapolation. What we can do though, and will consider next, is to use the model to predict the value of the response variable when presented with some new value for x_i for which y_i has not yet been observed.

4.5 A Prediction Interval for a new observation

More precisely we can develop what is known as a Prediction Interval (sometimes just PI) for some new observation. Let us say that we have a new value for x_i which we will label x_0. We have yet to observe the response for x_0 but we wish to predict it, which we will do by way of an interval rather than a single value given the stochastic nature of our model.

We seek y_0 where $y_0 = \mu_0 + \varepsilon_0$ and the point "prediction" for this would be

$$\widehat{y_0} = \widehat{\mu_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

We know that

$$\widehat{\mu_0} \sim N(\mu_0,\ \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}))$$

and therefore the distribution of $\widehat{\mu_0} - \mu_0$ is

$$\widehat{\mu_0} - \mu_0 \sim N(0,\ \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}))$$

To gain a prediction interval we would like to have the distribution for $\widehat{y_0} - y_0$ rather than $\widehat{\mu_0} - \mu_0$

So taking our previous equation and then adding and subtracting ε_0 to the left-hand side

$$\widehat{\mu_0} - (\mu_0 + \varepsilon_0) + \varepsilon_0 \sim N(0,\ \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}))$$

but the term in the brackets $(\mu_0 + \varepsilon_0)$ is y_0 and $\widehat{y_0} = \widehat{\mu_0}$ so we can re-write this equation as

$$\widehat{y_0} - y_0 + \varepsilon_0 \sim N(0, \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}))$$

and because we know $\varepsilon_0 \sim N(0, \sigma^2)$ from the original specification of the simple linear model we have

$$\widehat{y_0} - y_0 \sim N(0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) + \sigma^2)$$

or

$$\widehat{y_0} - y_0 \sim N(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right))$$

To find a formula for the prediction interval we need to standardise the normal distribution, that is find the function of $\widehat{y_0} - y_0$ that follows N(0,1).

$$\frac{\widehat{y_0} - y_0}{\sqrt{\sigma^2\left(1 + \frac{1}{n} + \frac{x_0 - \bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$$

and if we replace σ^2 with its estimator S^2 we have

$$\frac{\widehat{y_0} - y_0}{\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$$

which allows us to find the 100$(1-\alpha)$% prediction interval for y_0 which is

$$\widehat{y_0} \pm t_{\frac{\alpha}{2}} \sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

The prediction interval for y_0 is usually much wider than the confidence interval for μ_0 at the same value of α. This is because the random variability term ε_0 impacts the prediction interval.