Inference in regression

$$Y_i = a + bX_i + e_i$$

- a and b are point estimates of the true (but unknown) values α and β
- We can therefore calculate confidence interval estimates
- For these we need the standard errors

Standard errors for *a* and *b*

- The variance of *b* is given by

$$s_b^2 = \frac{s_e^2}{\sum (X_i - \bar{X})^2}$$

- where

$$s_e^2 = \frac{\sum e_i^2}{n-2} = \frac{ESS}{n-2}$$

is the estimated error variance

Calculation of the standard error

$$s_e^2 = \frac{\sum e_i^2}{n-2} = \frac{ESS}{n-2} = \frac{170.75}{10} = 17.075$$

- Hence

$$s_b^2 = \frac{s_e^2}{\sum (X_i - \bar{X})^2} = \frac{17.075}{49.37} = 0.346$$

- and the standard error of *b* is

$$s_b = \sqrt{0.346} = 0.588$$

Confidence interval for β

- The 95% CI is given by

$$-2.7 \pm 2.228 \times 0.588 = [-4.01, -1.39]$$

- where 2.228 is the appropriate *t* value (df = 10, α = 5%)

Confidence interval for α

- The variance of *a* is given by

$$s_a^2 = s_e^2 \times \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right) = 17.0754 \times \left(\frac{1}{12} + \frac{3.35^2}{49.37} \right) = 5.304$$

- Hence the standard error is 2.303. The 95% CI is given by

$$40.71 \pm 2.228 \times 2.303 = [35.57, 45.84]$$

Hypothesis testing

- We can test hypotheses in the usual manner. To test for a zero slope coefficient (i.e. *X* does not influence *Y*):

$H_0: \beta = 0$
$H_1: \beta \neq 0$

- The test statistic is: $t = \dfrac{b - \beta}{s_b} \sim t_{n-2}$

Hypothesis testing (continued)

- The test statistic is

$$t = \frac{-2.7 - 0}{0.588} = -4.59$$

which exceeds the critical value $t^*_{10} = 2.228$, so the null hypothesis is rejected. We conclude X does indeed affect Y.

A test of the goodness of fit

- We can also test for the significance of the R^2 statistic:

H_0: $R^2 = 0$
H_1: $R^2 > 0$

- The test statistic is: $F = \dfrac{RSS/1}{ESS/(n-2)} \sim F_{1,n-2}$

Evaluating the test

$$F = \frac{359.19/1}{170.75/10} = 21.078$$

- This exceeds the critical value F^* = 4.96 (df = 1,10) so H_0 is rejected.

- Note: an equivalent formula is

$$F = \frac{R^2/1}{\left(1-R^2\right)/\left(n-2\right)} \sim F_{1,n-2}$$

Prediction

- The regression line may be used for prediction: to predict the birth rate for a country growing at 3% p.a. we insert this value into the regression equation.

$$\hat{Y} = 40.71 - 2.7 \times 3 = 32.6$$

- The predicted birth rate is 32.6.

Confidence interval for the prediction

- The 95% CI for the prediction is given by

$$\left[\hat{Y} - t_{n-2} \times s_e \sqrt{\frac{1}{n} + \frac{(X_P - \bar{X})^2}{\sum (X - \bar{X})^2}}, \hat{Y} + t_{n-2} \times s_e \sqrt{\frac{1}{n} + \frac{(X_P - \bar{X})^2}{\sum (X - \bar{X})^2}}\right]$$

- Evaluating this gives

$$\left[32.6 \pm 2.228 \times 4.132 \sqrt{\frac{1}{12} + \frac{(3-3.35)^2}{49.37}}\right]$$

$$= [29.9, 35.3]$$

Summary

- The regression coefficients *a* and *b* are estimates (as are R^2, etc) so can be the subject of inference
- To obtain CIs and to conduct hypotheses, tests we need to calculate standard errors
- Predictions can be made and CIs of the predictions can be made