1 Inferences about regression parameters

For our linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, we have not made any assumptions about the data to calculate the parameter estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$, since we are simply applying the method of least squares. To conduct inference about regression parameters, however, we need to make the following assumptions about the errors ε_i :

1. **Linearity**: the above model is actually correct, we are not neglecting any terms,
2. **Independence**: the errors ε_i are independent,
3. **Normality**: the errors ε_i follow a normal distribution, and
4. **Homogeneity of variance**: $\text{Var}(\varepsilon_i) = \sigma_\varepsilon^2$ for all observations.

2 Inference for the regression coefficients β_0 and β_1

With the assumptions above, we can show that the least squares estimators are unbiased and have the following variances:

$$Var(\widehat{\beta}_0) = \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right) \text{ and } Var(\widehat{\beta}_1) = \frac{\sigma_\varepsilon^2}{S_{xx}}.$$

We can estimate σ_ε^2 by $\widehat{\sigma}_\varepsilon^2 = s_\varepsilon^2 = \frac{1}{n-2} \sum (y_i - \widehat{y}_i)^2$. Thus we can construct confidence intervals for β_0 and β_1 with:

$$\widehat{\beta}_0 \pm t_{\alpha/2}\ s_\varepsilon \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}} \text{ and } \widehat{\beta}_1 \pm t_{\alpha/2}\ \frac{s_\varepsilon}{\sqrt{S_{xx}}}.$$

In a similar way, we can conduct tests of hypotheses as illustrated in the text on page 591.

3 Inference for $E(y_{n+1})$ and y_{n+1} (Confidence intervals and prediction intervals)

Two common uses of regression analyses are to predict the mean value of y at a given x_{n+1} value ($E(y_{n+1})$), and to predict the value of an individual observation at a given x_{n+1} value (y_{n+1}). In both cases, the estimated value is obtained by evaluating the least-squares prediction equation at x_{n+1}. The difference is that an interval estimate for an individual value should be much wider than for a mean. Thus we have:

$$\widehat{y}_{n+1} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{n+1} = \overline{y} + \widehat{\beta}_1 (x_{n+1} - \overline{x}) \ ,$$

as the estimate for either $E(y_{n+1})$ or y_{n+1}. For the confidence interval for $E(y_{n+1})$ we have

$$\widehat{y}_{n+1} \pm t_{\alpha/2}\, s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{S_{xx}}}.$$

We can compute these intervals for individual values of x_{n+1}, or we can compute them for many values of x_{n+1} to create confidence bands. Interval estimates for an individual observation at a given x_{n+1} value are called prediction intervals, and are given by:

$$\widehat{y}_{n+1} \pm \ t_{\alpha/2}\, s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{S_{xx}}}.$$

The extra '1' under the square root sign arises from the extra variability for an individual observation. As with confidence interval calculations, often several prediction intervals are calculated to create prediction bands. Check the SAS code for this lecture to see how to generate confidence bands and prediction bands.

4 The ANOVA table

A summary of the regression results that is used for many linear model analyses is called the analysis of variance (ANOVA) table. It is based on the idea that the variability of the y values about their mean can be partitioned into two sources: one part describing variability in y 'explained' by the regression, and a second part due to error in predicting y from the regression model. This relationship can be seen by noting that:

$y_i - \overline{y} = (\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i)$ (also see Figure 11-10 of page 582 of the Ott text). When each side is squared and summed (and the crossproduct term vanishes) we get the result:

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \ ,$$

which is also expressed as $\text{SSY} = \text{SS(Regression)} + \text{SS(Residual)}$. The term SS(Regression) is called the sum of squares due to regression. These sums of squares are presented in the ANOVA table, and then SS(Regression) and SS(Residual) are divided by their degrees of freedom to obtain mean squares, denoted MS(Regression) and MS(Residual), respectively. MS(Regression) and MS(Residual) are statistically independent, and if the null hypothesis

of $H_0 : \beta_1 = 0$ is true, then they both estimate the population variance σ^2. Thus we can take the ratio F =MS(Regression)/MS(Residual) and use it as a test of $H_0 : \beta_1 = 0$. F follows an F distribution with numerator degrees of freedom = df_1=1 and denominator degrees of freedom = df_2= n-2. This test is equivalent to the t-test of $H_0 : \beta_1 = 0$ presented earlier, and in fact $F_{\alpha,1,n-2} = t^2_{\alpha/2,n-2}$.