Measuring the fit of the model - SSR

Once we've determined our estimated regression line, we'd like to know how well the model fits. How far/close are the observations to the fitted line?

- One way to do this is take some measure of how big the errors/residuals are, where the errors/residuals are given by $e_i = y_i - \hat{y}_i$.
- This measure is called the sum of squares due to error, $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$; it is the quantity which the least squares procedure attempts to minimize.
- Note that this quantity depends on the units in which the dependent variable is measured.

Measuring the fit of the model - R^2

- Another way to measure the fit of the model is to look at the proportion of the total variability in the dependent variable that can be explained by the independent variable.
- We can measure the total variability in the dependent variable using the total sum of squares $SST = \sum_{i=1}^{n}(y_i - \bar{y_i})^2$.
- We can meausure the variability in the dependent variable that can be explained by the independent variable $SSR = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2$
- This means that the proportion of total variability in the dependent variable that can be explained by the independent variable is $\frac{SSR}{SST}$.
- This quantity is called the coefficient of determination and denoted R^2.
- Since $SST = SSR + SSE$, R^2 can also be written $\frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST}$

Sample correlation

- We'd like to have a way to estimate the true correlation, ρ, using the data
- This is the sample correlation r, given by:

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

- This can be re-expressed in terms that we have used before. Remember, that we can write $\hat{\beta}_1$ as $\frac{S_{xy}}{S_{xx}}$. This yields $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \hat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}}$.
- This relationship also means that we can write the regression equation given r, S_{xx}, S_{yy}, and the sample means of x and y. We know $\hat{\beta}_1 = r\sqrt{\frac{S_{yy}}{S_{xx}}}$.
- In the case of simple linear regression (one independent variable), the coefficient of determination $R^2 = r^2$.

Inferences concerning the linear model parameters

- The least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained using our sample are only estimates of β_0 and β_1
- How good are these estimators?
- What are their means, variances, etc.?
- How can we make a confidence interval/hypothesis test for these parameters?

Sampling distribution for slope estimate, $\hat{\beta}_1$

- $E(\hat{\beta}_1) = \beta_1$, so $\hat{\beta}_1$ is unbiased
- $Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$ where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and $\sigma^2 = Var(Y) = Var(\varepsilon)$
- The distribution of $\hat{\beta}_1$ depends on the distribution of the error term ε. It is normally distributed if ε is normally distributed.
- We will generally be looking at models, in which we assume that ε is normally distributed.

Sampling distribution for intercept estimate, $\hat{\beta}_0$

- $E(\hat{\beta}_0) = \beta_0$, so $\hat{\beta}_0$ is unbiased
- $Var(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{n S_{xx}}$
- The distribution of $\hat{\beta}_0$ also depends on the distribution of the error term ε. It is normally distributed if ε is normally distributed.

Estimator for σ^2

- We rarely know σ^2, so we will need to estimate it based on the data
- Since σ^2 represents the variance of the Y_is around the line $\beta_0 + \beta_1 X_i$, it makes since to estimate it using some function of the distances between the data points and the fitted line.
- This (unbiased) estimator for σ^2 is $s^2 = \frac{SSE}{n-2}$, where $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- Given that $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed (given known σ^2), when we substitute our estimate s^2 for σ^2, these estimators have t distributions with $n - 2$ degrees of freedom.
- Knowledge of the sampling distributions of these statistics enables us to conduct hypothesis tests and form confidence intervals.

Hypothesis tests/CIs for coefficients

- After fitting a linear model, we might ask whether there is sufficient evidence to conclude that the x variable is a useful predictor of the y variable.
- This is a hypothesis test with H_0: $\beta_1 = 0$ and H_A: $\beta_1 \neq 0$.
- We can conduct the test as usual, formulating the test statistic as:

$$T^{n-2} = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{s^2}{S_{xx}}}}$$

- Of course, we can also use the same methodology to test hypotheses which involve another value of β_1 (instead of 0) or to test hypotheses involving β_0.
- Using the information about the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$, we can form confidence intervals for these parameters. To find a $(1-\alpha)100\%$ confidence interval:

$$\hat{\beta}_i \pm t_{\frac{\alpha}{2}} SE_{\hat{\beta}_i}$$

Are the assumptions of the model met?

- Suppose we use least squares to obtain an estimated regression line.
- In order to make inferences concerning the parameters β_0 and β_1, we need to make assumptions about the distribution/correlation of the residuals
- One way to examine the truthfulness of the assumptions is to look at a scatter plot of the residuals ($e = y - \hat{y}$) vs. the fitted values ($\hat{y}$)
- They should form a cloud (no patterns), symmetric about 0, with fairly even variation in the the variation of the residuals over the range of fitted values.

Confidence interval for $E(Y)$

- Remember that our regression line is just an estimate for the expected value of the Y variable.
- This means we're estimating $E(Y) = \beta_0 + \beta_1 X^*$ with $\hat{\beta}_0 + \hat{\beta}_1 X^*$, where X^* is just the value of X for which we want to estimate $E(Y)$
- We know that since $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators, the quantity $\hat{\beta}_0 + \hat{\beta}_1 X$ is an unbiased estimator for $E(Y)$.
- The standard error for our estimate is fairly complicated to derive (see pp. 502-4), but it can be shown to be $s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$
- This yields a confidence interval for $E(Y) = \beta_0 + \beta_1 X^*$, of the form

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t^{n-2}_{\frac{\alpha}{2}} s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

where $s^2 = \frac{SSE}{n-2}$.

Prediction interval for Y when $X = x^*$

- Let's say that instead of a confidence interval for the mean $E(Y)$, we want a confidence interval for a prediction of Y when $X = x^*$.
- Before, we were estimating a parameter $E(Y)$. Now we want to estimate the value of a RV, the Y we observe at some specific time when $X = x^*$.
- Intuitively, we would estimate this value somewhere near the middle of the distribution for Y for $X = x^*$. The center of this distribution is $E(Y) = \beta_0 + \beta_1 x^*$, which is estimated by $\hat{\beta}_0 + \hat{\beta}_1 x^*$.
- So we have the same estimate for $E(Y)$ as we do for a prediction of Y, but intuitively, the variance for a prediction must be larger.
- The S.E. is (again) fairly complicated to derive (see pp. 506-8). It can be shown to be $s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$, yielding a prediction interval for Y (when $X = x^*$ and $s^2 = \frac{SSE}{n-2}$)

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t^{n-2}_{\frac{\alpha}{2}} s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$