Week 5 Cheat Sheet

Statistics and Data Analysis with Excel, Part 2

Charlie Nuttelman

Here, I provide the mathematical equations and some of the important Excel functions required to perform various calculations in Week 4 of the course. The headings represent the screencasts in which you will find those calculations and concepts. Not all screencasts are referenced below – just the ones that have complex mathematical formulas or Excel formulas that are tricky to use.

My equations and analysis follow that of Montgomery and Runger in: *Applied Statistics and Probability for Engineers, 7th edition*, Douglas C. Montgomery and George C. Runger, Wiley (2018). This is an excellent text for applied probability and statistics and highly recommended if you need a supplementary text for the course.

Simple Linear Regression Using Least Squares Estimators

A simple linear regression model, or straight-line linear model, has the form:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Each value of the independent variable, x, is related to the dependent variable, Y, through the relationship shown above, where β_0 is the intercept and β_1 is the slope. ε is a random error term – if samples were obtained that had the same value of x, the response (Y) would be different for each sample due to this random error term.

β_0 and β_1 cannot be known for sure. Instead, we use our sample data to put together estimates for the intercept and slope, $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. We collect data and can calculate these estimators using the following equations:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{\bar{x}}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

A shortcut formula for S_{xx} in Excel, if the x data have been named “x” and the number of observations has been named “n” is:

=VAR.S(x)*(n-1)

Hypothesis Tests for Simple Linear Regression

Oftentimes, the question arises as to whether the intercept and/or slope are statistically significant. In other words, are intercept and slope different from zero. If not different from zero, there's no need to include them in our model (by the way, even if an intercept is not significant, it is typically included in the model regardless).

We can set up hypothesis tests on the intercept and slope, shown here for a test on the slope:

$$H_0: \beta_1 = \beta_{1,0}$$

$$H_1: \beta_1 > \beta_{1,0}$$

$\beta_{1,0}$ is just some constant and is typically 0. To perform the hypothesis test on the slope, we need to calculate a test statistic based on the T distribution:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se\left[\hat{\beta}_1\right]}$$

A test statistic for a hypothesis test on the intercept would be similar to the above equation.

The standard error of slope and intercept are given by the following equations:

$$se\left(\hat{\beta}_0\right) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

$$se\left(\hat{\beta}_1\right) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$\hat{\sigma}$ is known as the standard error and can be calculated from the following equation:

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

Note the 2 that is subtracted from n in the denominator. A straight-line linear model is a 2-parameter model (β_0 and β_1), hence the 2.

SS_E is known as the residual sum of squares and is calculated from:

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Once the test statistic, t_0 is calculated, we compare it to the T distribution with $n - 2$ degrees of freedom. To accept a lower-tailed alternate hypothesis, we must show that $t_0 < -t_{\alpha,n-2}$. To accept an upper-tailed alternate hypothesis, we must show that $t_0 > t_{\alpha,n-2}$. And to accept a two-tailed alternate hypothesis, we must show that $t_0 < -t_{\frac{\alpha}{2},n-2}$ or $t_0 > t_{\frac{\alpha}{2},n-2}$. P-values can also be calculated, as we have done in previous weeks.

Confidence Intervals on the Slope and Intercept

Two-sided confidence intervals on the slope and intercept can be put together easily using the standard errors for slope and intercept calculated above. Based on our estimator for the intercept, $\hat{\beta}_0$, we can be $(1-\alpha)$% sure that the true intercept, β_0, lies in the following interval:

$$P\left[\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2} \cdot se(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t_{\frac{\alpha}{2},n-2} \cdot se(\hat{\beta}_0)\right] = 1-\alpha$$

Similarly, based on our estimator for the slope, $\hat{\beta}_1$, we can be $(1-\alpha)$% sure that the true slope, β_1, lies in the following interval:

$$P\left[\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2} \cdot se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2} \cdot se(\hat{\beta}_1)\right] = 1-\alpha$$

Confidence Interval on the Mean Response

The response of a particular process at a certain value of the independent variable, x_0, will vary due to random error. A confidence interval on the true mean response at $x = x_0$, $\mu_{Y|x_0}$, is centered around our estimate for the mean response, $\hat{\mu}_{Y|x_0}$:

$$\left[\hat{\mu}_{Y|x_0} - t_{\frac{\alpha}{2},n-2} \cdot se\left(\hat{\mu}_{Y|x_0}\right) \le \mu_{Y|x_0} \le \hat{\mu}_{Y|x_0} + t_{\frac{\alpha}{2},n-2} \cdot se\left(\hat{\mu}_{Y|x_0}\right)\right] = 1-\alpha$$

The estimate for the mean response, $\hat{\mu}_{Y|x_0}$, is simply the model prediction at x_0:

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

The standard error of the mean response at $x = x_0$ is:

$$se\left(\hat{\mu}_{Y|x_0}\right) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

Prediction Interval for Straight-Line Regression (Guided Workshop 5)

A prediction interval on a single future observation (Y_0) made at $x = x_0$ is more useful than a confidence interval on the mean response. It tells you, with $(1-\alpha)$% confidence, a range over which you'd expect the response to be at $x = x_0$. The width of a prediction interval is always much larger than that for the confidence interval on the mean response.

$$\left[\hat{y}_0 - t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \le Y_0 \le \hat{y}_0 + t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}\right] = 1-\alpha$$

The estimate for the future observation, $\hat{y}_0$, is simply the model prediction at x_0:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Coefficient of Determination and Adjusted R-Squared

The coefficient of determination, R^2, and adjusted R-squared, R_{adj}^2, are useful in judging adequacy of a model. Ideally, you want a coefficient of determination or adjusted R-squared to be near 1.0. Adjusted R-squared is much better than coefficient of determination in judging model quality and should be used (R^2 should not be used, in general).

To calculate R^2 and R_{adj}^2, we need to obtain the total sum of squares, SS_T:

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

A shortcut formula in Excel for calculating SS_T is:

=VAR.S(y)*(n-1), assuming that "n" and "y" have been named.

We can then obtain the regression sum of squares, SS_R:

$$SS_R = SS_T - SS_E$$

SS_E has been defined previously (see above) and is the residual sum of squares.

Now, we can calculate R^2:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

And the preferred R_{adj}^2:

$$R_{adj}^2 = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}$$

Here, p is the number of parameters in our model, which is 2 for a simple linear regression model.