Week 6 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 6 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.

Matrix Approach to Multiple Linear Regression

Model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$

k = number of regressor variables

p = number of parameters in the model

$p = k + 1$ (for full term model)

Can be written in matrix form: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$

Observations

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Model matrix:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

Model parameter vector:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

Solving the normal equations:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Fitted matrix form: $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

All of the above is best done in Excel using array functions (MINVERSE, TRANSPOSE, and MMULT).

Statistical Properties of Least Squares Estimators $\hat{\beta}$

The standard errors of the coefficients in the model, $se(\ _j)$, can be obtained from the diagonal elements of the C matrix:

$$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$$

$$C = (X'X)^{-1} = \begin{bmatrix} C_{00} & C_{01} & \cdots & C_{0k} \\ C_{10} & C_{11} & \cdots & C_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ C_{k0} & C_{k1} & \cdots & C_{kk} \end{bmatrix}$$

Standard error ($\hat{\sigma}$): $\hat{\sigma}^2 = \frac{SS_E}{n-p}$

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

Hypothesis Tests in Multiple Linear Regression (Parts 1 and 2)

Significance of Regression (ANOVA for Regression)

Partition total sum of squares (SS_T) into two parts, that due to the regression model (SS_R) and that due to random error (SS_E):

$$SS_T = SS_R + SS_E$$

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

(SS_E as computed above.)

Shortcut formula in Excel: SS_T **=VAR.S(y)*(n-1)**

There is a formula for SS_R but it's easiest to compute by subtraction:

$$SS_R = SS_T - SS_E$$

Hypothesis test:

H_0: $\boldsymbol{\beta} = 0$
H_1: $\beta_j \neq 0$ for at least one j

Test statistic: $F_0 = \frac{MS_R}{MS_E} = \frac{SS_R/(p-1)}{SS_E/(n-p)}$

If $f_0 > f_{\alpha,p-1,n-p}$, then accept H_1.

ANOVA table format:

Source of variation	Sum of squares	Degrees of freedom	Mean square	F_0
Regression	SS_R	$k = p - 1$	$MS_R = SS_R/k$	MS_R/MS_E
Error (residual)	SS_E	$n - p$	$MS_E = SS_E/(n-p)$	
Total	SS_T	$n - 1$		

Hypothesis tests on individual regression coefficients:

$$H_0: \beta_j = \beta_{j0}$$

$$H_1: \beta_j \neq \beta_{j0}$$

Test statistic:

$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$ (if $\beta_{j0} = 0$)

Accept H_1 if $|t_0| > t_{\alpha,n-p}$ (for one-tailed test; Excel's Regression tool performs two-tailed test)

Model Performance

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

Adjusted R-squared (more common for engineers/scientists): $R^2_{adj} = 1 - \frac{\frac{SS_E}{(n-p)}}{\frac{SS_T}{(n-1)}}$

Confidence Intervals in Multiple Linear Regression

Confidence interval on model parameters:

$$\hat{\beta}_j - t_{\alpha/2,n-p}\, se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p}\, se(\hat{\beta}_j)$$

$se(\hat{\beta}_j)$ can be obtained from the diagonal elements of the C matrix (see above).

Confidence interval on the mean response at $\boldsymbol{x_0}$ (note that $\boldsymbol{x_0}$ is a vector of inputs):

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \boldsymbol{x_0}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0}} \leq \mu_{Y|x_0} \leq \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \boldsymbol{x_0}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0}}$$

Model output at point $\boldsymbol{x_0}$: $\hat{\mu}_{Y|x_0} = \boldsymbol{x_0}'\hat{\boldsymbol{\beta}}$

Prediction interval on a future observation:

$$\hat{y}_0 \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(\mathbf{1} + \boldsymbol{x_0}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0})}$$

Model output at point $\boldsymbol{x_0}$: $\hat{y}_0 = \boldsymbol{x_0}'\widehat{\boldsymbol{\beta}}$

All of the above are best calculated in Excel (would be difficult to do by hand!).