Chapter 3 - Multiple Regression Analysis

We will now generalize the SLR model to include multiple explanatory variables.

Ex: Consider the simple wage equation we have discussed. We may want to include an additional factor, experience as

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

Where exper is years of labor market experience

As another example, consider the problem of explaining the effect of per student spending (expend) on the average standardized test score (avgscore) at the high school level. Suppose that the average test score depends on funding, average family income (avginc) and other variables:

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u$$

The Multiple Linear Regression Model (MLR)

MLR can be written in the population as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$$

where

β_0 is the intercept

β_1 is the parameter associated with x_1

β_2 is the parameter associated with x_2

and so on

Obtaining OLS Estimates

Without loss of generality we can examine the case with $k=2$ explanatory variables. The estimated OLS equation is written in a form similar to the SLR:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

where

$\hat{\beta}_0$ = the estimate of β_0

$\hat{\beta}_1$ = the estimate of β_1

$\hat{\beta}_2$ = the estimate of β_2

Just as with the SLR in the MLR model OLS chooses

$\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ to minimize the sum of squared errors:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2$$

Interpreting the OLS Regression Equation

Again begin with the $k=2$ case

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$\hat{\beta}_0$ is the predicted value of y when $x_1 = 0$ and $x_2 = 0$

$\hat{\beta}_1$ and $\hat{\beta}_2$ have partial effect, or ceteris paribus interpretations.

From the above we have

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$$

So that we can obtain the predicted change in y given changes in x_1 and x_2.

NB: Note how the intercept has nothing to do with changes in y

When X_2 is held constant, so that $\Delta X_2 = 0$, then

$$\Delta \hat{y} = \hat{\beta}_1 \Delta X_1$$

Similarly,

$$\Delta \hat{y} = \hat{\beta}_2 \Delta X_2$$

when $\Delta X_1 = 0$ (X_1 held constant)

NB: For the initiated:

if you recall your calculus, these are partial derivatives:

$$\frac{\partial E(y|x_1, x_2)}{\partial x_1} = \hat{\beta}_1$$

and

$$\frac{\partial E(y|x_1, x_2)}{\partial x_2} = \hat{\beta}_2$$

Ex: Determinants of College GPA (p. 75)

Using a sample of data for $n = 141$ students from a large university we obtain the following OLS regression to predict college GPA from high school GPA and achievement test scores:

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT$$

Q: How do we interpret this?

- A 1 point increase in hsGPA leads to (predicted) an increase of .453 in colGPA. Nearly half a point! In other words, holding ACT scores fixed: take two students A and B, both with the same ACT scores. But A has one point higher hsGPA than B. A's colGPA is predicted to be .453 higher than B's!

The sign on ACT implies that, holding hsGPA fixed, a change in ACT of 10 points (very large change; avg = 24), affects colGPA by less than one-tenth of a point. This is not a strong effect. It suggests that, after accounting for hsGPA, ACT is not a strong predictor of colGPA.

If we were to run an SLR relating colGPA to ACT we would obtain

$$\widehat{colGPA} = 2.40 + 0.0271\, ACT$$

But this does not allow us to compare two people with the same hsGPA because are not controlling for that.

Goodness-of-Fit

As with SLR we can define SST (total sum of squares), SSE (the explained sum of squares), and SSR (residual sum of squares) for the MLR model:

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

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

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2$$

As before: $SST = SSE + SSR$

In other words, total variation in $\{y_i\}$ is the sum of the total variation in $\{\hat{y}_i\}$ and $\{\hat{u}_i\}$

Then

$$\frac{SSR}{SST} + \frac{SSE}{SST} = 1$$

We can (again) define the R-squared to be

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

NB: R^2 will only increase (at least will never decrease) when adding an additional variable into the regression equation

The Sampling Distribution of the OLS Estimators

Again we will need some assumptions for the MLR:

Assumption MLR.1 – Linear in Parameters

The model in the population can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the parameters of interes and u is the error term

Assumption MLR.2 – Random Sampling

We have a RS of size n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1

Assumption MLR.3 – No Perfect Colinearity

In the sample (also the population), none of the independent variables is constant, and there are no exact linear relationships among the independent variables

Assumption MLR.4 - Zero Conditional Mean

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u \mid x_1, x_2, \ldots, x_k) = 0$$

Given MLR.1 through MLR.4 we can show that OLS is unbiased for the MLR:

$$E(\hat{\beta}_j) = \beta_j \quad , \text{ for } j = 1, 2, \ldots, k$$

Assumption MLR.5 – Homoskedasticity

The error u has the same variance given any values of the explanatory variables. So

$$Var(u | x_1, x_2, \ldots, x_k) = \sigma^2$$

The Sampling Variances of the OLS Slope Estimators

Under MLR.1 through MLR.5, conditional on the sample values of the independent variables,

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j\,(1 - R_j^2)}$$

for $j = 1, 2, \ldots, k$ where $SST_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j, and R_j^2 is the R-squared from regressing x_j on all the other independent variables (including an intercept)

Estimating σ^2: The Standard Errors of the OLS Estimators

Recall that $E(u^2) = \sigma^2$ and an unbiased estimator is

$\frac{1}{n}\sum_{i=1}^{n} u_i^2$. Unfortunately, we don't observe u_i.

u_i can be written as

$$u_i = Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_k X_{ik}$$

The reason we don't observe u_i is because we don't know

the true population β_j

Replacing the β_j with $\hat{\beta}_j$, we get the OLS residuals

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_K X_{iK}$$

Thus an unbiased estimator is

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2 = \frac{SSR}{n-k-1}$$

NB: when $k=1$ we get the estimator for σ^2 in the SLR with denominator $n-2$

The square root of $\hat{\sigma}^2$ is $\hat{\sigma}$, the standard error of the regression.

For confidence intervals and hypothesis tests we will need the standard deviation of $\hat{\beta}_j$ which is just

$$sd(\hat{\beta}_j) = \frac{\sigma}{[SST_j(1-R_j^2)]^{1/2}}$$

Replacing σ with its estimator $\hat{\sigma}$ gives us the standard error of $\hat{\beta}_j$:

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{[SST_j(1-R_j^2)]^{1/2}}$$

NB: This is exactly what lm outputs in R, together with t stats, p-values etc.
~~[illegible]~~ R FTW!