Multiple Regression Analysis

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$

1. Estimation

Introduction

- Main drawback of simple regression is that with 1 RHS variable it is unlikely that *u* is uncorrelated with *x*
- Multiple regression allows us to control for those "other" factors
- The more variables we have the more of *y* we will be able to explain (better predictions)
- e.g. Earn=β_0+ β_1Educ+ β_2Exper+u
- Allows us to measure the effect of education on earnings holding experience fixed

Parallels with Simple Regression

- β_0 is still the intercept
- β_1 to β_k all called slope parameters
- u is still the error term (or disturbance)
- Still need to make a zero conditional mean assumption, so now assume that
- $E(u|x_1,x_2, \ldots,x_k) = 0$
- other factors affecting y are not related on average to $x_1,x_2, \ldots, x_k$
- Still minimizing the sum of squared residuals

Estimating Multiple Regression

- The estimated equation can be written:

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

- And we want to minimize:

$$\sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)^2$$

- We will now have k+1 first order conditions

$$\sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right) = 0$$
$$\sum x_{i1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right) = 0$$
$$\sum x_{i2} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right) = 0$$

Etc.

Interpreting The Coefficients

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

- Then we can obtain the predicted change in y given changes in the *x* variables

A "Partialling Out" Interpretation

- Consider the case where k=2, then

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

- The estimated coefficient of β_1 is:

$$\hat{\beta}_1 = \left(\sum x_{i1} - \overline{x}_1\right) y_i \Big/ \sum \left(x_{i1} - \overline{x}\right)^2$$

- Another way to express this same estimate is:

"Partialling Out" continued

- This implies that we estimate the same effect of x_1

1. Regressing y on x_1 and x_2, and
2. Regressing y on residuals from a regression of x_1 on x_2

Simple vs Multiple Reg Estimate

- Compare the simple regression $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$

with the multiple regression $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

Generally, $\tilde{\beta}_1 \neq \hat{\beta}_1$ unless :

Assumptions for Unbiasedness

- The assumptions needed to get unbiased estimates in simple regression can be restated

1. Population model is linear in parameters:
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$

2. We can use a random sample of size n from the population model, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i=1, 2, .., n\}$

- i.e. no sample selection bias

Then the sample model is
$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$

Assumptions for Unbiasedness

3. $E(u|x_1, x_2, \ldots x_k) = 0$

4. None of the x's is constant, and there are no <u>exact linear</u> relationships among them

Examples of Perfect Collinearity

1. $x_1 = 2(x_2)$ - One of these is redundant

2. The linear combinations can be more complicated

3. Including income and income^2 does not violate this assumption (not an exact linear function)

Too Many or Too Few Variables

- Assumptions 1-4 can be used to show that:

  $E\left(\hat{\beta}_j\right) = \beta_j, \ j = \{0, 1, ..., \text{k}\}$

  i.e. that the estimated parameters are unbiased
- What happens if we include variables in our specification that don't belong (over-specify)?
- What if we exclude a variable from our specification that does belong?

Omitted Variable Bias

- Suppose the true model is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- But we estimate

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u \text{, then}$$

$$\tilde{\beta}_1 = \frac{\sum (x_{i1} - \bar{x}_1) y_i}{\sum (x_{i1} - \bar{x}_1)^2}$$

- Recall the true model, so that

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

Omitted Variable Bias (cont)

- So the numerator becomes

$$\sum (x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i)$$

$$\text{recall}: \sum (x_i - \bar{x}) = 0, \text{ and } \sum (x_i - \bar{x}) x_i = \sum (x_i - \bar{x})^2$$

- Then, the numerator becomes

Omitted Variable Bias (cont)

$$\text{so, } \tilde{\beta} = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum \left((x_{i1} - \bar{x}_1)^2 \right)} + \frac{\sum (x_{i1} - \bar{x}_1) u_i}{\sum \left((x_{i1} - \bar{x}_1)^2 \right)}$$

- Taking expectations (conditional on *x*'s) and recalling that E(u_i)=0, we have

Omitted Variable Bias (cont)

- How might we interpret the bias?
- Consider the regression of x_2 on x_1:

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1$$

$$\text{then} \quad \tilde{\delta}_1 = \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum \left((x_{i1} - \bar{x}_1)^2 \right)}$$

- The omitted variable bias equals zero if:

Summary of Direction of Bias

- The direction of the bias depends on:

1. The sign of δ_1

2. The sign of β_2

Summary Table

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$		
$\beta_2 < 0$		

Omitted Variable Bias Summary

- The size of the bias is also important
- Typically we don't know β_2 or the correlation but we can usually make an educated guess whether these are positive or negative

What about the more general case (k+1 variables)?

- Technically we can only sign the bias of the more general case if all the *x*'s are uncorrelated

The More General Case

- For example, suppose that the true model is:

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

and we estimate :

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

Variance of the OLS Estimators

- We now know that the sampling distribution of our estimate is centered around the true parameter
- Want to think about how spread out this distribution is
- Much easier to think about this variance under an additional assumption, so

Assumption 5:

- Assume $\mathrm{Var}(u|x_1, x_2,\ldots, x_k) = \sigma^2$ (Homoskedasticity)
- This implies that $\mathrm{Var}(y|x_1, x_2,\ldots, x_k) = \sigma^2$

Variance of OLS (cont)

- The 4 assumptions for unbiasedness, plus this homoskedasticity assumption are known as the Gauss-Markov assumptions
- These can be used to show:

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

Components of OLS Variances

- Thus, how precisely we can estimate the coefficients (variance) depends on:

1. The error variance (σ^2):

2. The total sample variation in xj (SST):

Components of OLS Variances

3. The linear relationships among the independent variables (R_j^2):

- The problem of highly correlated x variables is called "Multi-Collinearity"

Example of Multi-Collinearity

- Suppose that you want to try to explain differences in mortality rates across different states
- Might use variation in medical expenditures across the states

- *Problem?*

Misspecified Models

- Consider again the misspecified model:
  $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$
- We showed that the estimate of β_1 is biased if β_2 does not equal zero and that including x_2 in this model did not bias the estimate even if $\beta_2=0$
- Therefore, why not just include β_2?

Misspecified Models (cont)

Thus, $Var\left(\tilde{\beta}_1\right) < Var\left(\hat{\beta}_1\right)$

- Unless x_1 and x_2 are uncorrelated ($R_j^2=0$) then they are the same
- Thus, assuming that x_1 and x_2 are **not** uncorrelated

β_2=0:

β_2 not equal to 0:

Estimating the Error Variance

- We don't know what the error variance, σ^2, is, because we don't observe the errors, u_i
- What we observe are the residuals, $\hat{u}_i$
- We can use the residuals to form an estimate of the error variance as we did for simple regression

$$\hat{\sigma}^2 = \left(\sum \hat{u}_i^2\right) / (n - k - 1)$$
$$= SSR / df$$

Error Variance Estimate (cont)

- Thus, an estimate of the standard error of the coefficients is:

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