Multiple Regression Analysis

Chapter 2 (chapter 3 of the book reference)

Reference:
Wooldridge: Introductory Econometrics:
A Modern Approach, 5e

Multiple Regression Analysis: Estimation

- **<u>Definition of the multiple linear regression model</u>**

"Explains variable y in terms of variables $x_1, x_2, \ldots, x_k$"

Intercept

Slope parameters

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

Dependent variable,
explained variable,
response variable,…

Independent variables,
explanatory variables,
regressors,…

Error term,
disturbance,
unobservables,…

Multiple Regression Analysis: Estimation

- **<u>Motivation for multiple regression</u>**
  - Incorporate more explanatory factors into the model
  - Explicitly hold fixed other factors that otherwise would be in u
  - Allow for more flexible functional forms
- **Example: Wage equation**

Now measures effect of education <u>explicitly holding experience fixed</u>

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

All other factors…

Hourly wage

Years of education

Labor market experience

Multiple Regression Analysis: Estimation

- **Example: Average test scores and per student spending**

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

Other factors

Average standardized test score of school

Per student spending at this school

Average family income of students at this school

  - Per student spending is likely to be correlated with average family income at a given high school because of school financing
  - Omitting average family income in regression would lead to biased estimate of the effect of spending on average test scores
  - In a simple regression model, effect of per student spending would partly include the effect of family income on test scores

Multiple Regression Analysis: Estimation

- **Example: Family income and family consumption**

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$$

Family consumption

Family income

Family income squared

Other factors

- Model has two explanatory variables: inome and income squared
- Consumption is explained as a quadratic function of income
- One has to be very careful when interpreting the coefficients:

By how much does consumption increase if income is increased by one unit?

$$\frac{\partial cons}{\partial inc} = \beta_1 + 2\beta_2 inc$$

Depends on how much income is already there

Multiple Regression Analysis: Estimation

- **Example: CEO salary, sales and CEO tenure**

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 ceoten + \beta_3 ceoten^2 + u$$

Log of CEO salary

Log sales

Quadratic function of CEO tenure with firm

  - Model assumes a constant elasticity relationship between CEO salary and the sales of his or her firm
  - Model assumes a quadratic relationship between CEO salary and his or her tenure with the firm

- **Meaning of "linear" regression**
  - The model has to be linear **in the parameters** (not in the variables)

Multiple Regression Analysis: Estimation

- **OLS Estimation of the multiple regression model**

- **Random sample**
  $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : \ i = 1, \ldots n\}$

  $$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$

- **Fitted values:** suppose estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$

  $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$$

- **Regression residuals**

  $$\widehat{u}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \ldots - \widehat{\beta}_k x_{ik}$$

- **Minimize sum of squared residuals**

  $$\min \sum_{i=1}^{n} \widehat{u}_i^2 \quad \rightarrow \quad \widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$$

Minimization will be carried out by computer

Multiple Regression Analysis: Estimation

- **Ordinary Least Squares (OLS) for the Simple Regression Model**

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- **Ordinary Least Squares (OLS) for the Multiple Regression Model with matrix notation**

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i \quad i = 1, 2, \ldots, n$$

$$\Updownarrow$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

$$\boxed{\boldsymbol{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}}$$

Multiple Regression Analysis: Estimation

- **Properties of OLS on any sample of data**
- **Fitted values and residuals**

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \ldots + \widehat{\beta}_k x_{ik} \qquad \widehat{u}_i = y_i - \widehat{y}_i$$

Fitted or predicted values

Residuals

- **Algebraic properties of OLS regression**

$$\sum_{i=1}^{n} \widehat{u}_i = 0 \qquad \sum_{i=1}^{n} x_{ij}\widehat{u}_i = 0 \qquad \bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}_1 + \ldots + \widehat{\beta}_k \bar{x}_k \Leftrightarrow \bar{\hat{y}} = \bar{y}$$

Deviations from regression line sum up to zero

Correlations between deviations and regressors are zero

Sample averages of y and of the regressors lie on regression line

Multiple Regression Analysis: Estimation

- **Interpretation of the multiple regression model**

$$\beta_j = \frac{\partial y}{\partial x_j}$$

By how much does the dependent variable change if the j-th independent variable is increased by one unit, holding all other independent variables and the error term constant

- The multiple linear regression model manages to hold the values of other explanatory variables fixed even if, in reality, they are correlated with the explanatory variable under consideration
- "Ceteris paribus"-interpretation
- It has still to be assumed that unobserved factors do not change if the explanatory variables are changed

Multiple Regression Analysis: Estimation

- **Interpretation of $\hat{\beta}_j$**

$$\hat{\beta}_j = \Delta\hat{y} \text{ if } \Delta x_j = 1 \textit{ ceteris paribus}$$

$$\Delta x_j \textit{ ceteris paribus} \Rightarrow \Delta\hat{y} = \hat{\beta}_j \Delta x_j$$

- **Total variation of y**

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

Multiple Regression Analysis: Estimation

- **<u>Example:</u> variables in Levels (LEVEL – LEVEL)**

$$\hat{wage} = -3.39 + 0.64 educ + 0.07 exper$$

- Interpret the coefficients
- Estimate the variation on wage of a person with 1 more year of education and 2 more years of experience
- Estimate the wage of a person with 15 years of education and 3 years of experience
- Suppose that the observed wage of a person with 15 years of education and 3 years of experience is 7$/h. Calculate the residual and interprert.
- How much does the experience need to vary in order that the wage increases 0.63$/h?

Multiple Regression Analysis: Estimation

- **Example: Determinants of college GPA**

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

Grade point average at college

High school grade point average

Achievement test score

- **Interpretation**
  - Holding ACT fixed, another point on high school grade point average is associated with another .453 points college grade point average
  - Or: If we compare two students with the same ACT, but the hsGPA of student A is one point higher, we predict student A to have a colGPA that is .453 higher than that of student B
  - Holding high school grade point average fixed, another 10 points on ACT are associated with less than one point on college GPA

Multiple Regression Analysis: Functional Form

- **<u>Incorporating nonlinearities: Log-logarithmic form</u> (LOG – LOG)**
- **CEO salary and firm sales**

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u$$

Natural logarithm of CEO salary

Natural logarithm of his/her firm's sales

- **This changes the interpretation of the regression coefficient:**

$$\beta_1 = \frac{\partial \log(salary)}{\partial \log(sales)} = \frac{\frac{\partial salary}{salary}}{\frac{\partial sales}{sales}}$$

<u>Percentage</u> change of salary

… if sales increase <u>by 1 %</u>

$$\boxed{\%\Delta salary = \beta_1}$$

<u>Logarithmic changes are always percentage changes</u>

The Simple Regression Model

- **<u>Incorporating nonlinearities: Log-logarithmic form</u> (LOG – LOG)**
- **CEO salary and firm sales**

$$\widehat{\log}(salary) = 4.822 + 0.257 \log(sales)$$

+ 1 % sales $\Rightarrow$ + 0.257 % salary

- **The log-log form postulates a <u>constant elasticity</u> model**

Multiple Regression Analysis: Functional Form

- <u>**Incorporating nonlinearities: Semi-logarithmic form)**</u> **(LOG – LEVEL)**
- **Regression of log wages on years of eduction**

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

Natural logarithm of wage

- **This changes the interpretation of the regression coefficient:**

$$\beta_1 = \frac{\partial \log(wage)}{\partial educ} = \frac{1}{wage} \cdot \frac{\partial wage}{\partial educ} = \frac{\frac{\partial wage}{wage}}{\partial educ}$$

✖ **100** is the <u>percentage</u> change of wage

... if years of education are increased by one year

$$\boxed{\%\Delta wage = \beta_1 \times 100\%}$$

Multiple Regression Analysis: Functional Form

- **Incorporating nonlinearities: Semi-logarithmic form** **(LOG – LEVEL)**
- **Regression of log wages on years of eduction**

$$\widehat{\log}(wage) = 0.584 + 0.083\ educ$$

The wage increases by 8.3 % for every additional year of education (= return to education)

The growth rate of wage is 8.3 % per year of education

- **the semi-log form assumes a semi-elasticity model**

Multiple Regression Analysis: Functional Form

- **<u>Incorporating nonlinearities: Semi-logarithmic form (LEVEL-LOG)</u>**
- **Regression of salary on logarithm of sales**

$$salary = \beta_0 + \beta_1 \log(sales) + u$$

CEO salary in thousands of $

- **This changes the interpretation of the regression coefficient:**

$$\beta_1 = \frac{\partial salary}{\partial \log(sales)} = \frac{\partial salary}{\frac{\partial sales}{sales}}$$

Variation of CEO salary in thousands of $

✖ **100** is the percentage change in sales

$$\boxed{\Delta salary = \frac{\beta_1}{100} \quad \text{if the} \quad \% \Delta sales = 1\%}$$

Multiple Regression Analysis: Functional Form

- **Incorporating nonlinearities: Semi-logarithmic form (LEVEL-LOG)**

- **Regression of salary on logarithm of sales**

$$salary = -898.9 + 262.9 \log(sales)$$

CEO salary in thousands of \$

If sales increase by 1% then the CEO salary increases

by 2.629 thousands of \$

Multiple Regression Analysis: Estimation

- **<u>"Partialling out" interpretation of multiple regression</u>**
- **One can show that the estimated coefficient of an explanatory variable in a multiple regression can be obtained in two steps:**
  - 1) Regress the explanatory variable on all other explanatory variables
  - 2) Regress y on the residuals from this regression
- **Why does this procedure work?**
  - The residuals from the first regression is the part of the explanatory variable that is uncorrelated with the other explanatory variables
  - The slope coefficient of the second regression therefore represents the isolated effect of the explanatory variable on the dep. variable

Multiple Regression Analysis: Estimation

- <u>**Goodness-of-Fit**</u>
- **Decomposition of total variation**

$$SST = SSE + SSR$$

- **R-squared**

$$R^2 = SSE/SST = 1 - SSR/SST$$

Notice that R-squared can only increase if another explanatory variable is added to the regression

- **Alternative expression for R-squared**

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right)}$$

R-squared is equal to the squared correlation coefficient between the actual and the predicted value of the dependent variable

Multiple RegressionAnalysis: Estimation

- **Example: Explaining arrest records**

Number of times arrested 1986

Proportion prior arrests that led to conviction

Months in <u>prison 1986</u>

Quarters employed <u>1986</u>

$$\widehat{narr86} = .712 - .150pcnv - .034ptime86 - .104qemp86$$

$$n = 2,725, \quad R^2 = .0413$$

- **Interpretation:**
  - Proportion prior arrests: Δ=0.5 $\rightarrow$ Δy =-.075 or -7.5 arrests per 100 men
  - Months in prison: Δ=12 $\rightarrow$ Δy =-.034(12) = -0.408 arrests for given man
  - Quarters employed: Δ=1 $\rightarrow$ Δy = -.104 or -10.4 arrests per 100 men

Multiple Regression Analysis: Estimation

- **Example: Explaining arrest records (cont.)**
  - An additional explanatory variable is added:

$$\widehat{narr86} = .707 - .151 pcnv + .0074 avgsen - .037 ptime86 - .103 qemp86$$

$$n = 2,725, \quad R^2 = .0422$$

Average sentence in prior convictions

R-squared increases only slightly

- **Interpretation:**
  - Average prior sentence increases number of arrests (?)
  - Limited additional explanatory power as R-squared increases by little
- **General remark on R-squared**
  - Even if R-squared is small (as in the given example), regression may still provide good estimates of ceteris paribus effects

Multiple Regression Analysis: Expected value of OLS

- **<u>Standard assumptions for the multiple regression model</u>**
- **Assumption MLR.1 (Linear in parameters)**

In the population, the relationship between y and the coefficients is linear

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

The model includes an unobserved error term

- **Assumption MLR.2 (Random sampling)**

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : \ i = 1, \ldots n\}$$

The data is a random sample drawn from the population

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

Each data point therefore follows the population equation

Multiple Regression Analysis: Expected value of OLS

- **<u>Standard assumptions for the multiple regression model (cont.)</u>**
- **Assumption MLR.3 (No perfect collinearity)**

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

- **Remarks on MLR.3**
  - The assumption only rules out <u>perfect</u> collinearity/correlation between explanatory variables; imperfect correlation is allowed
  - If an explanatory variable is a perfect linear combination of other explanatory variables it is superfluous and may be eliminated
  - Constant variables are also ruled out (collinear with intercept)

Multiple Regression Analysis: Expected value of OLS

- **Example for perfect collinearity: small sample**

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

In a small sample, avginc may accidentally be an exact multiple of expend; it will not be possible to disentangle their separate effects because there is exact covariation

- **Example for perfect collinearity: relationships between regressors**

$$voteA = \beta_0 + \beta_1 shareA + \beta_2 shareB + u$$

Either shareA or shareB will have to be dropped from the regression because there is an exact linear relationship between them: shareA + shareB = 1

Multiple Regression Analysis: Expected value of OLS

- **<u>Standard assumptions for the multiple regression model (cont.)</u>**
- **Assumption MLR.4 (Zero conditional mean)**

$$E(u_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$$

The value of the explanatory variables must contain no information about the mean of the unobserved factors

  - In a multiple regression model, the zero conditional mean assumption is much more likely to hold because fewer things end up in the error

- **Example: Average test scores**

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

If avginc was not included in the regression, it would end up in the error term; it would then be hard to defend that expend is uncorrelated with the error

Multiple Regression Analysis: Expected value of OLS

- **Discussion of the zero mean conditional assumption**
  - Explanatory variables that are correlated with the error term are called <u>endogenous</u>; endogeneity is a violation of assumption MLR.4
  - Explanatory variables that are uncorrelated with the error term are called <u>exogenous</u>; MLR.4 holds if all explanat. var. are exogenous
  - Exogeneity is the key assumption for a causal interpretation of the regression, and for unbiasedness of the OLS estimators

- **<u>Theorem 3.1 (Unbiasedness of OLS)</u>**

$$MLR.1-MLR.4 \quad \Rightarrow \quad E(\widehat{\beta}_j) = \beta_j, \quad j = 0, 1, \ldots, k$$

  - **Unbiasedness** is an average property in repeated samples; in a given sample, the estimates may still be far away from the true values

Multiple Regression Analysis: Variance of OLS

- **<u>Standard assumptions for the multiple regression model (cont.)</u>**
- **Assumption MLR.5 (Homoscedasticity)**

$$Var(u_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2$$

The value of the explanatory variables must contain no information about the variance of the unobserved factors

- **Example: Wage equation**

$$Var(u_i|educ_i, exper_i, tenure_i) = \sigma^2$$

This assumption may also be hard to justify in many cases

- **Short hand notation**

All explanatory variables are collected in a random vector

$$Var(u_i|\mathbf{x}_i) = \sigma^2 \quad \text{with} \quad \mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$$

Multiple Regression Analysis: Variance of OLS

- **Theorem 3.2 (Sampling variances of OLS slope estimators)**

Under assumptions MLR.1 – MLR.5:

Variance of the error term

$$Var(\widehat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}, \quad j = 1, \ldots, k$$

Total sample variation in explanatory variable x_j:

$$\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

R-squared from a regression of explanatory variable x_j on all other independent variables (including a constant)

Multiple Regression Analysis: Variance of OLS

- **Components of OLS Variances:**
- **1) The error variance**
  - A high error variance increases the sampling variance because there is more "noise" in the equation
  - A large error variance necessarily makes estimates imprecise
  - The error variance does not decrease with sample size
- **2) The total sample variation in the explanatory variable**
  - More sample variation leads to more precise estimates
  - Total sample variation automatically increases with the sample size
  - Increasing the sample size is thus a way to get more precise estimates

Multiple Regression Analysis: Variance of OLS

- **3) Linear relationships among the independent variables**

Regress x_j on all other independent variables (including a constant)

The R-squared of this regression will be the higher the better x_j can be linearly explained by the other independent variables

- Sampling variance of $\widehat{\beta}_j$ will be the higher the better explanatory variable x_j can be linearly explained by other independent variables
- The problem of almost linearly dependent explanatory variables is called <u>multicollinearity</u> (i.e. $R_j \rightarrow 1$ for some j)

Multiple Regression Analysis: Variance of OLS

- **Discussion of the multicollinearity problem**
  - In the above example, it would probably be better to lump all expenditure categories together because effects cannot be disentangled
  - In other cases, dropping some independent variables may reduce multicollinearity (but this may lead to omitted variable bias)
  - Only the sampling variance of the variables involved in multicollinearity will be inflated; the estimates of other effects may be very precise
  - Note that multicollinearity is not a violation of MLR.3 in the strict sense
  - Multicollinearity may be detected through "variance inflation factors"

$$VIF_j = 1/(1 - R_j^2)$$

As an (arbitrary) rule of thumb, the variance inflation factor should not be larger than 10

Multiple Regression Analysis: Variance of OLS

- **An example for multicollinearity**

Average standardized test score of school

Expenditures for teachers

Expenditures for instructional materials

Other expenditures

$$avgscore = \beta_0 + \beta_1 teachexp + \beta_2 matexp + \beta_3 othexp + \ldots$$

The different expenditure categories will be strongly correlated because if a school has a lot of resources it will spend a lot on everything.

It will be hard to estimate the differential effects of different expenditure categories because all expenditures are either high or low. For precise estimates of the differential effects, one would need information about situations where expenditure categories change differentially.

As a consequence, sampling variance of the estimated effects will be large.

Multiple Regression Analysis: Variance of OLS

- **Theorem 3.4 (Gauss-Markov Theorem)**
  - Under assumptions MLR.1 - MLR.5, the OLS estimators are the best linear unbiased estimators (BLUEs) of the regression coefficients, i.e.

$$Var(\widehat{\beta}_j) \leq Var(\tilde{\beta}_j) \quad j = 0, 1, \ldots, k$$

for all $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$ for which $E(\tilde{\beta}_j) = \beta_j, j = 0, \ldots, k.$

- **OLS is only the best estimator if MLR.1 – MLR.5 hold; if there is heteroscedasticity for example, there are better estimators.**

Multiple Regression Analysis: Variance of OLS

- **Efficiency of OLS: The Gauss-Markov Theorem**
  - Under assumptions MLR.1 - MLR.5, OLS is unbiased
  - However, under these assumptions there may be many other estimators that are unbiased
  - Which one is the unbiased estimator with the smallest variance?
  - In order to answer this question one usually limits oneself to linear estimators, i.e. estimators linear in the dependent variable

$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$$

May be an arbitrary function of the sample values of all the explanatory variables; the OLS estimator can be shown to be of this form

Multiple Regression Analysis: Variance of OLS

- **Estimating the error variance**

$$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} \hat{u}_i^2 \right) / [n - k - 1]$$

An unbiased estimate of the error variance can be obtained by substracting the number of estimated regression coefficients from the number of observations. The number of observations minus the number of estimated parameters is also called the degrees of freedom. The n estimated squared residuals in the sum are not completely independent but related through the k+1 equations that define the first order conditions of the minimization problem.

- **Theorem 3.3 (Unbiased estimator of the error variance)**

$$MLR.1 - MLR.5 \quad \Rightarrow \quad E(\hat{\sigma}^2) = \sigma^2$$

Multiple Regression Analysis: Variance of OLS

- **Estimation of the sampling variances of the OLS estimators**

The true sampling variation of the estimated β_j

$$sd(\widehat{\beta}_j) = \sqrt{Var(\widehat{\beta}_j)} = \sqrt{\sigma^2 / \left[SST_j (1 - R_j^2) \right]}$$

Plug in $\widehat{\sigma}^2$ for the unknown σ^2

The estimated sampling variation of the estimated β_j

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

- **Note that these formulas are only valid under assumptions MLR.1-MLR.5 (in particular, there has to be homoscedasticity)**

Multiple Regression Analysis: Additional Topics

- **Including irrelevant variables in a regression model**

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

= 0 in the population

No problem because $E(\widehat{\beta}_3) = \beta_3 = 0.$

However, including irrelevant variables may increase sampling variance.

- **Omitting relevant variables: the simple case**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$ True model (contains x_1 and x_2)

$$y = \alpha_0 + \alpha_1 x_1 + w$$ Estimated model (x_2 is omitted)

Multiple Regression Analysis: Additional Topics

- **Omitted variable bias**

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

If x_1 and x_2 are correlated, assume a linear regression relationship between them

$$\Rightarrow \quad y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u$$

$$= \boxed{(\beta_0 + \beta_2\delta_0)} + \boxed{(\beta_1 + \beta_2\delta_1)}x_1 + \boxed{(\beta_2 v + u)}$$

If y is only regressed on x_1 this will be the estimated intercept

If y is only regressed on x_1, this will be the estimated slope on x_1

error term

- **<u>Conclusion:</u> All estimated coefficients will be biased**

Multiple Regression Analysis: Additional Topics

- **Omitted variable bias: more general cases**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$ True model (contains x_1, x_2 and x_3)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + w$$ Estimated model (x_3 is omitted)

  - No general statements possible about direction of bias
  - Analysis as in simple case if one regressor uncorrelated with others

- **Example: Omitting ability in a wage equation**

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u$$

If exper is approximately uncorrelated with educ and abil, then the direction of the omitted variable bias can be as analyzed in the simple two variable case.

Multiple Regression Analysis: Additional Topics

- **Example: Omitting ability in a wage equation**

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

Will both be positive

$$abil = \delta_0 + \delta_1 educ + v$$

$$wage = (\beta_0 + \beta_2\delta_0) + \boxed{(\beta_1 + \beta_2\delta_1)} educ + (\beta_2 v + u)$$

The return to education β_1 will be <u>overestimated</u> because $\beta_2\delta_1 > 0$. It will look as if people with many years of education earn very high wages, but this is partly due to the fact that people with more education are also more able on average.

- **When is there no omitted variable bias?**
  - If the omitted variable is irrelevant or uncorrelated

Multiple Regression Analysis: Further Issues

- **More on goodness-of-fit and selection of regressors**
- **General remarks on R-squared**
  - A high R-squared does not imply that there is a causal interpretation
  - A low R-squared does not preclude precise estimation of partial effects
- **Adjusted R-squared**
  - What is the ordinary R-squared supposed to measure?

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{SSR/(n-1)}{SST/(n-1)}$$ is an estimate for $$\boxed{1 - \frac{\sigma_u^2}{\sigma_y^2}}$$

Population R-squared

Multiple Regression Analysis: Further Issues

- **Adjusted R-squared (cont.)**
  - A better estimate taking into account degrees of freedom would be

Correct degrees of freedom of nominator and denominator

$$\bar{R}^2 = 1 - \frac{(SSR/\boxed{(n-k-1)})}{(SST/\boxed{(n-1)})} = adjusted\ R^2$$

  - <u>The adjusted R-squared imposes a penalty for adding new regressors</u>
  - The adjusted R-squared increases if, and only if, the ratio between the coef. estimate and the standard error of a newly added regressor is greater than one in absolute value
- **Relationship between R-squared and adjusted R-squared**

$$\bar{R}^2 = 1 - (1 - R^2)(n-1)/(n-k-1)$$

The adjusted R-squared may even get negative

Multiple Regression Analysis: Further Issues

- **Using adjusted R-squared to choose between nonnested models**
  - Models are nonnested if neither model is a special case of the other

$$rdintens = \beta_0 + \beta_1 \log(sales) + u \qquad \boxed{R^2 = .061, \bar{R}^2 = .030}$$

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u \qquad \boxed{R^2 = .148, \bar{R}^2 = .090}$$

  - A comparison between the R-squared of both models would be unfair to the first model because the first model contains fewer parameters
  - In the given example, even after adjusting for the difference in degrees of freedom, the quadratic model is preferred

Multiple Regression Analysis: Further Issues

- **Comparing models with different dependent variables**
  - R-squared or adjusted R-squared must not be used to compare models which differ in their definition of the dependent variable
- **Example: CEO compensation and firm performance**

There is much less variation in log(salary) that needs to be explained than in salary

$$\boxed{\widehat{salary}} = \underset{(223.63)}{223.90} + \underset{(.0163)}{.0089}\, sales + \underset{(11.08)}{19.63}\, roe$$

$$n = 209, R^2 = .029, \bar{R}^2 = .020, \boxed{TSS = 391{,}732{,}982}$$

$$\boxed{\widehat{lsalary}} = \underset{(0.29)}{4.36} + \underset{(.033)}{.275}\, lsales + \underset{(.0040)}{.0179}\, roe$$

$$n = 209, R^2 = .282, \bar{R}^2 = .275, \boxed{TSS = 66.72}$$