Advanced Data Analysis I Multiple Regression

PA 541 Week 4

Michael D. Siciliano

Department of Public Administration

College of Urban Planning and Public Affairs

Admin stuff

- Reminder HW1 due on Feb 8th
- Please be sure to check final data project groupings.
  - Begin to look for interesting data sets and research questions to explore.

QUESTION 3 (8 pts)

Run an ANOVA model to test the predictive capability of the percip.cat variable on total.medals. (a) What do you conclude from the results? (b) What might cause us to find an association between these variables? Do you think precipitation is having a causal effect on total.medals? What else might be going on?

```
aov.1 = aov(total.medals ~ precip.cat, data=oly)
summary(aov.1)
```

```
##                Df  Sum Sq  Mean Sq  F value    Pr(>F)
## precip.cat      2     554   277.11    11.83  8.25e-06 ***
## Residuals    1096   25670    23.42
##  ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 23 observations deleted due to missingness
```

We can conlude from results ($F = 11.83$, $p < .001$) above that knowing the preciptiation category explains a significant amount of variance in the total medal count. Your response to why we find this may vary considerably, from perhaps those with more precipitation have more opportunity to enage in winter sports (higher snowfall) or maybe you think it hurts their ability as it reduces the number of training days. While precipitation might have an effect, there may be endogeneity here, perhaps those in warmer/tropical climates have more precipitation and are also less likely to be successful in winter sports. Again answers may vary...just looking to have you think through potential issues and concerns.

Today's topics

- Multiple regression analysis and interpretation
- Omitted variable bias
- Multiple regression assumptions
- Three in-class exercises; work on HW

- Again, and as noted in Wooldridge, we are holding off on dealing with interpreting log transformations and quadratics until later this semester.
- We will also spend time covering multicollinearity in more detail in week 8 when we assess our regression assumptions.
  - For now, simply having some familiarity with these concepts is enough.

Beta Coefficients

Regression model

Causality

Central Limit Theorem

Let's

Merging Data

Regression Standard Error

Coefficient of Determination

REVIEW

Creating new variables

Statistically Significant/ p-values

Week 3

Factor variables

Missing Data

Predicted Values and Residuals

LET'S REVIEW THE IN-CLASS EXERCISE

Multiple Regression

Analysis and Estimation

- Assume one is only interested in understanding the effect of X on Y. Why is simple regression (i.e. regression with a single predictor) insufficient for understanding the effect of X on Y?

Why Multiple Regression?

- Multiple regression allows us to specifically control for many other factors that may simultaneously affect our dependent variable.

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

- Idea of **ceteris paribus** – which means "other relevant factors being equal".
  - Hence, we can view the effect of a unit change in one IV on the DV while holding the change in our other IVs constant.
  - Our parameter estimates of B_1 and B_2 etc..have partial effect interpretations. We will return to this concept later in the lecture.

- From Wooldridge P. 71
  - A simple model to explain city murder rates (*murdrate*) in terms of the probability of conviction (*pbconv*) and average sentence length (*avgsen*) is:
    - *Murdrate* = B_0 + B_1*prbconv* + B_2*avgsen* + *u*

  - What are some factors contained in *u*? Do you think the key assumption that the error has an expected mean value of zero given any values of the independent variables holds?
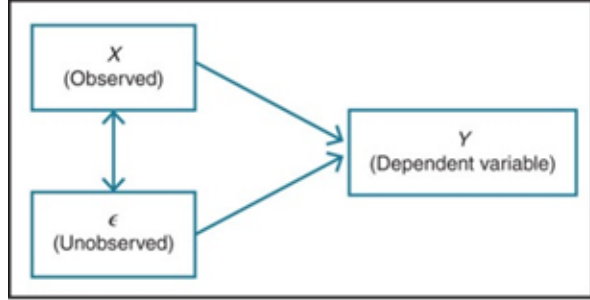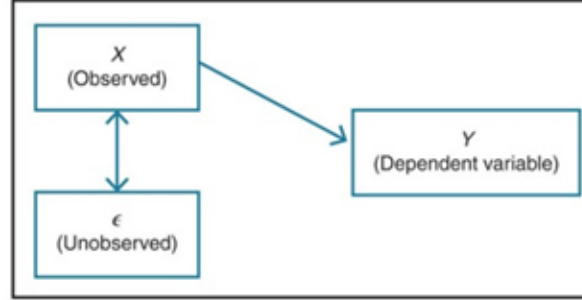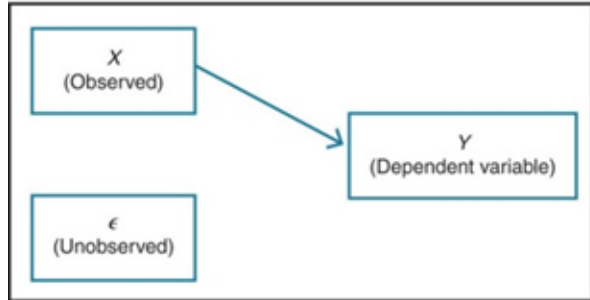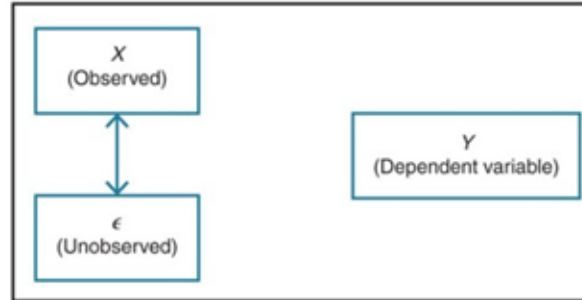
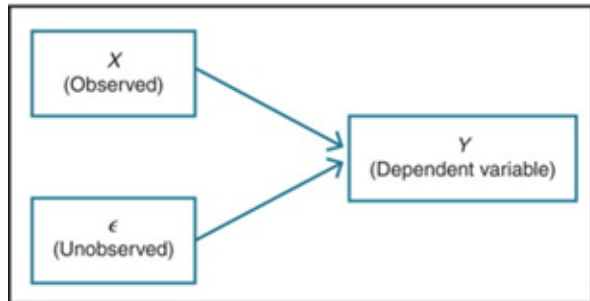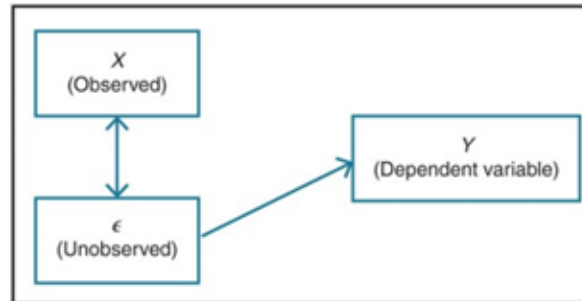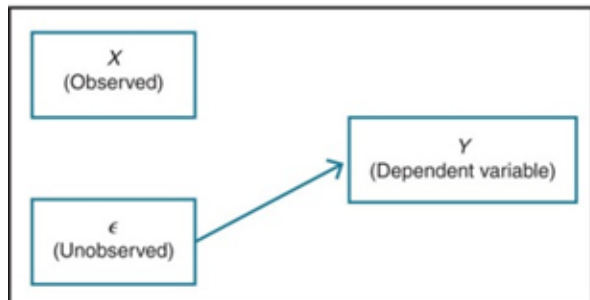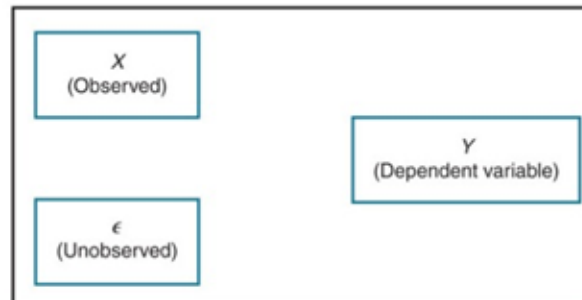For each panel, explain whether endogeneity will cause problems for an analysis of the relationship between X and Y. For concreteness, assume X is grades in college, *E* is IQ, and Y is salary at age 26.

Bigger images on next slide.
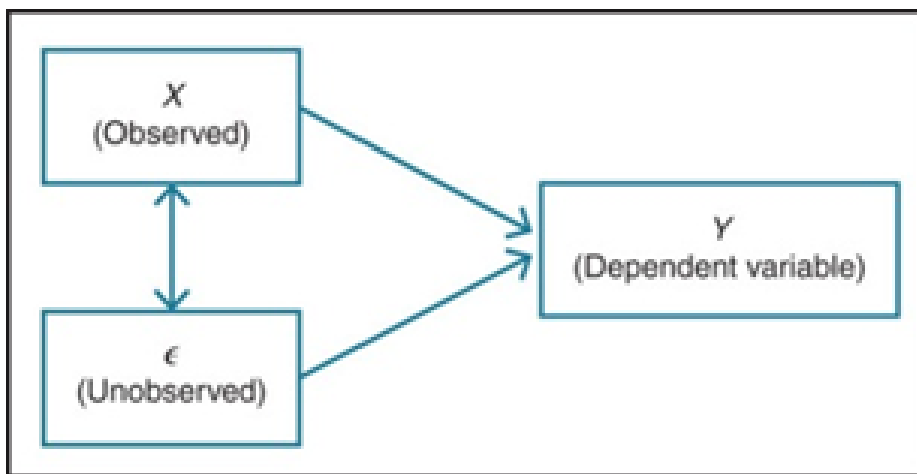
(a)
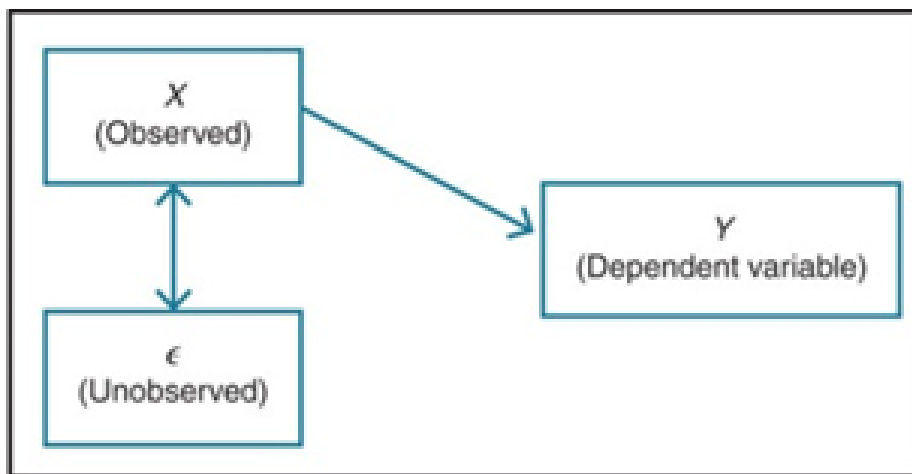
(b)
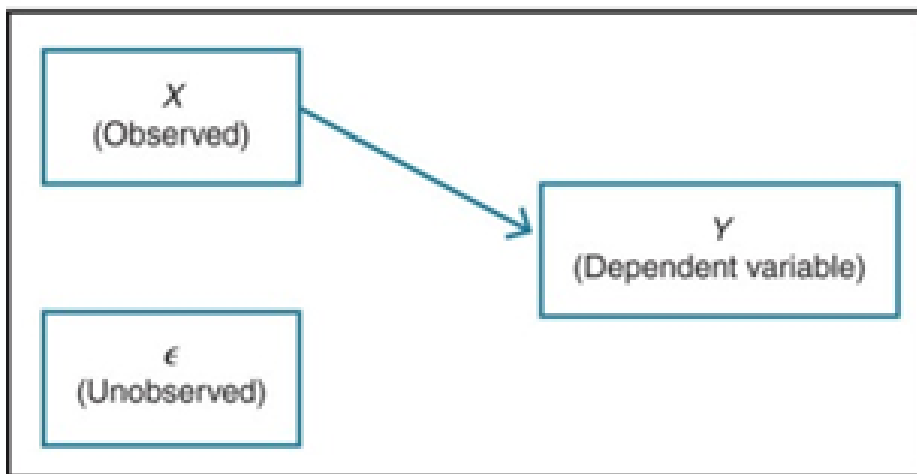
(c)
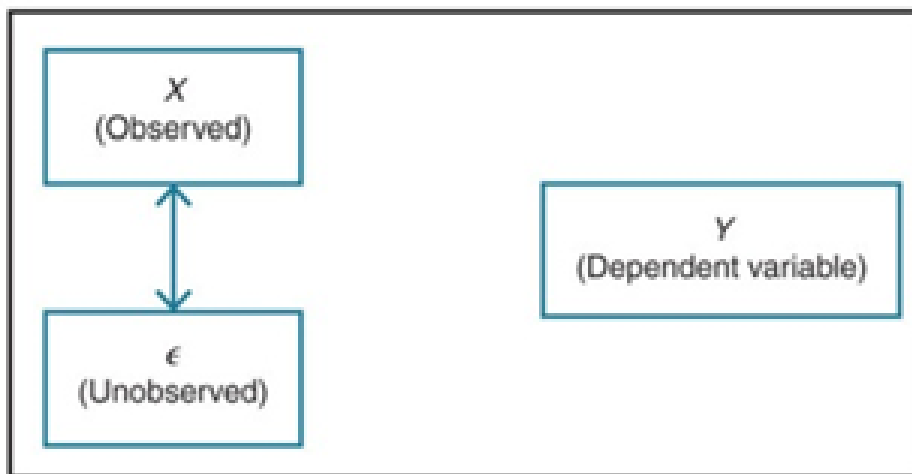
(d)
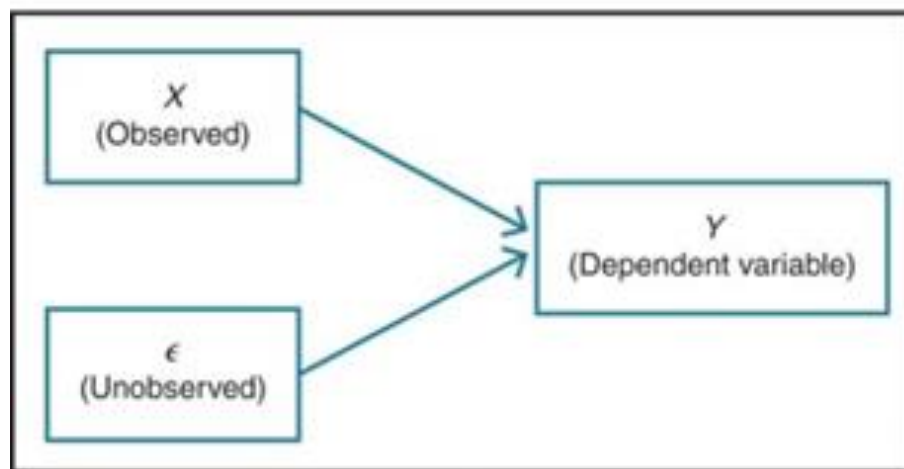
(e)
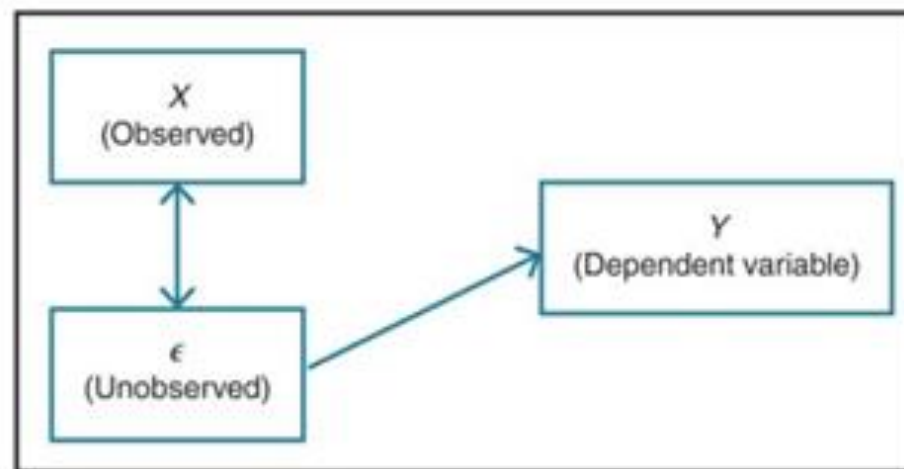
(f)
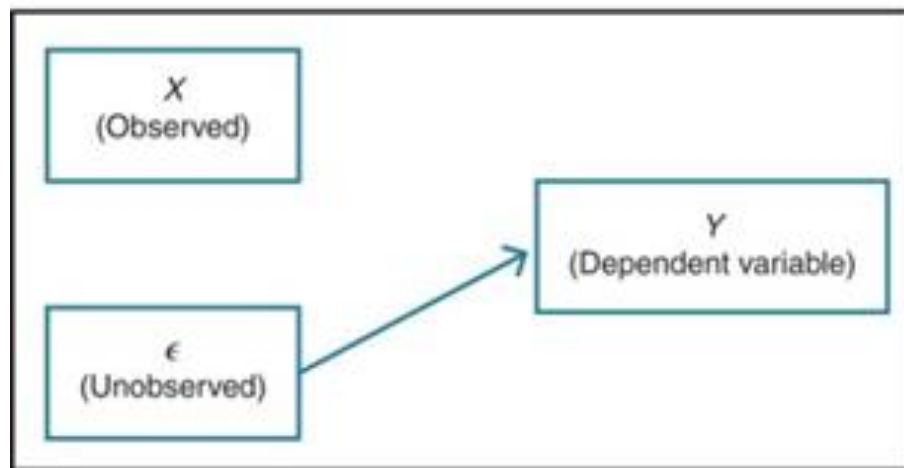
(g)
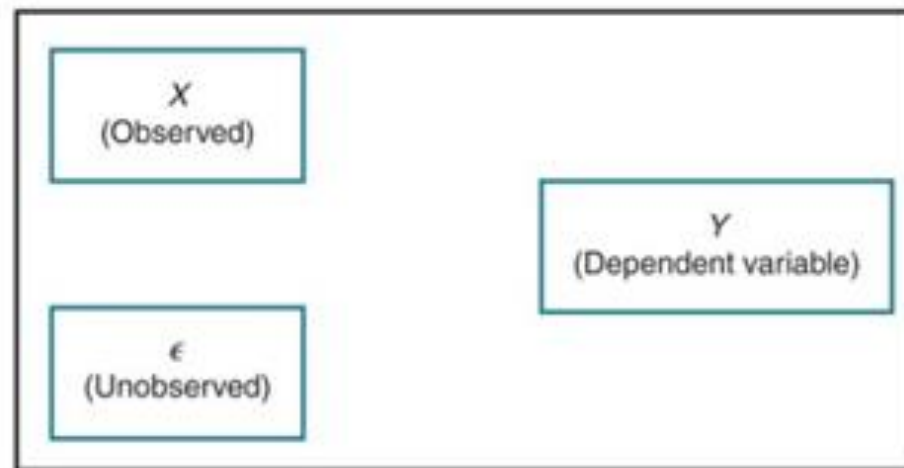
(h)

- We will be looking at examples from a dataset called 'prestige'. This data concerns the prestige of 102 different occupations in Canada.
- Variables include:
  - *education* - Average education of occupational incumbents, years, in 1971.
  - *income* - Average income of incumbents, dollars, in 1971.
  - *women* - Percentage of incumbents who are women.
  - *prestige* - Pineo-Porter prestige score for occupation. A scale that can run between 0 and 100.
  - *census* - Canadian Census occupational code.
  - *type* - Type of occupation. A factor with levels: bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar.

Let's look at a model

```
> lm1 = lm (prestige ~ education + income, data=prest)
> summary(lm1)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.8477787  3.2189771  -2.127   0.0359 *
education    4.1374444  0.3489120  11.858  < 2e-16 ***
income       0.0013612  0.0002242   6.071 2.36e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.81 on 99 degrees of freedom
Multiple R-squared: 0.798,     Adjusted R-squared: 0.7939
F-statistic: 195.6 on 2 and 99 DF,  p-value: < 2.2e-16

>
> anova(lm1)
Analysis of Variance Table

Response: prestige
          Df  Sum Sq Mean Sq F value     Pr(>F)
education  1 21608.4 21608.4 354.245  < 2.2e-16 ***
income     1  2248.1  2248.1  36.856  2.355e-08 ***
Residuals 99  6038.9    61.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Major component of the regression output are identical to SLR. Let's first talk about **the overall predictive capacity** of the model by looking at the R-squared and then we will discuss interpreting the coefficients.

From Last Week

- Three measures of variation in our simple regression model:
  - SST: Total sum of squares
  - SSR: Residual Sum of Squares (Error Sum of Squares)
  - SSM: Model Sum of Squares (Regression Sum of Square)
  - SST = SSM + SSR

Calculating Each Value

Total Sum of Squares

Sum of Squared Residual
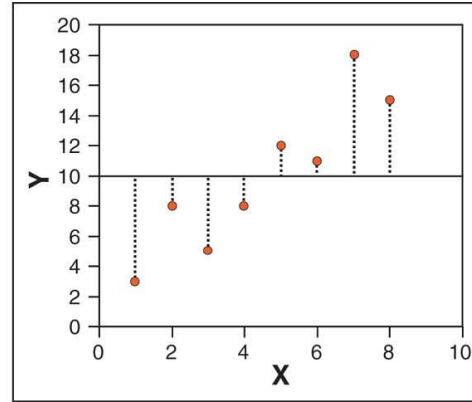
SS_T uses the differences between the observed data and the mean value of Y
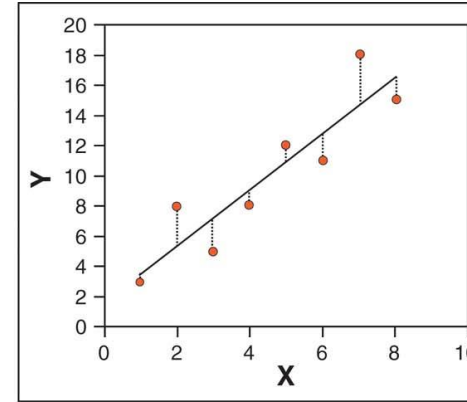
SS_R uses the differences between the observed data and the regression line

NOTE: Be very careful when using the terms SSE, SSR, etc...many textbooks use different forms. For example, SSR could be for "Residual" which is error or "Regression" which is explained from the model.

Sum of Squares for the Model
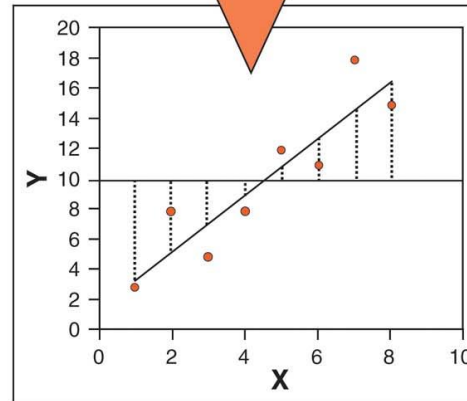
SS_M uses the differences between the mean value of Y and the regression line

Another View of the Measures of Variation

The R here is for residual.

$$SSR = \Sigma(y_i - \hat{y}_i)^2$$

$$SST = \Sigma(y_i - \bar{y})^2$$

$$SSM = \Sigma(\hat{y}_i - \bar{y})^2$$

Note, that Wooldridge labels this the SSE, for explained sum of squares
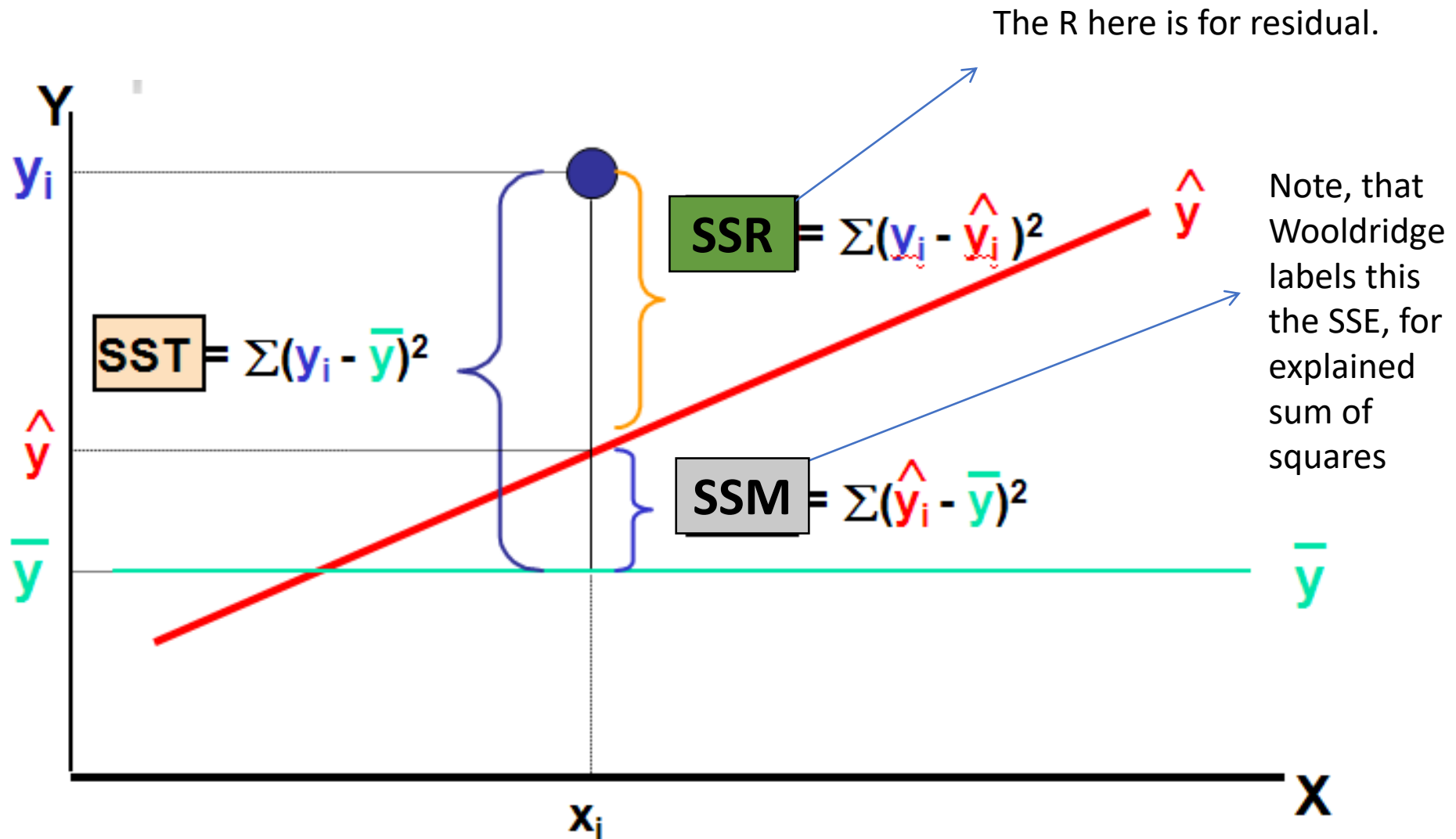
A Note about Adjusted R-Squared

- R^2 = SSM/SST or

- R^2 = 1 − SSR/SST

- R^2 does not decreases when a new independent variable is added to the model, even if the new variable is not an important predictor.

Adjusted Coefficient of Determination, $\overline{R}^2$

- Used to correct for the fact that adding non-relevant independent variables will still reduce the residual sum of squares

$$\overline{R}^2 = 1 - \frac{SSR/(n-K-1)}{SST/(n-1)}$$

Again, SSR is the residual sum of squares. What is left over that still needs explained.

(where n = sample size, K = number of independent variables)

- Adjusted R^2 provides a better comparison between multiple regression models when different numbers of independent variables are used.
- Adjusted R^2 penalizes excessive use of unimportant independent variables
- Smaller than R^2

Let's go back to our output

```
> lm1 = lm (prestige ~ education + income, data=prest)
> summary(lm1)

Call:
lm(formula = prestige ~ education + income, data = prest)

Residuals:
     Min       1Q    Median       3Q      Max
-19.4040  -5.3308    0.0154   4.9803  17.6889

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.8477787  3.2189771  -2.127   0.0359 *
education    4.1374444  0.3489120  11.858  < 2e-16 ***
income       0.0013612  0.0002242   6.071 2.36e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.81 on 99 degrees of freedom
Multiple R-squared: 0.798,     Adjusted R-squared: 0.7939
F-statistic: 195.6 on 2 and 99 DF,  p-value: < 2.2e-16
```

- How do we interpret the coefficients?

Interpreting OLS Regression Equation cont…

- We can obtain the predicted change in y given changes in x_1 and x_2. Note, that the intercept has nothing to do with changes in y.

$$\Delta y = B_1 \Delta x_1 + B_2 \Delta x_2$$

- In particular when x_2 is held constant, so that the change in $x_2 = 0$ then:

$$\Delta y = B_1 \Delta x_1 \text{ , } \quad \text{holding } x_2 \text{ constant}$$

- The key point is that by including x_2 in our model, we obtain a coefficient on x_1 that controls for the effect of x_2 on our dependent variable.

- Similarly, $\quad \Delta y = B_2 \Delta x_2$, holding x_1 constant.

Thinking more deeply about the 'partialling out' interpretation

- **<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 dependent variable

Visual representation of partialling out

- Consider the Venn Diagram below:
  - R-squared for this situation is A+B+C+D+E
  - Area A comes unequivocally from IV1; area C unequivocally from IV2; and area E unequivocally from IV3.
  - However, there is ambiguity regarding B and D. Both areas could be predicted from either of two IVs. To which IV should the contested areas be assigned?
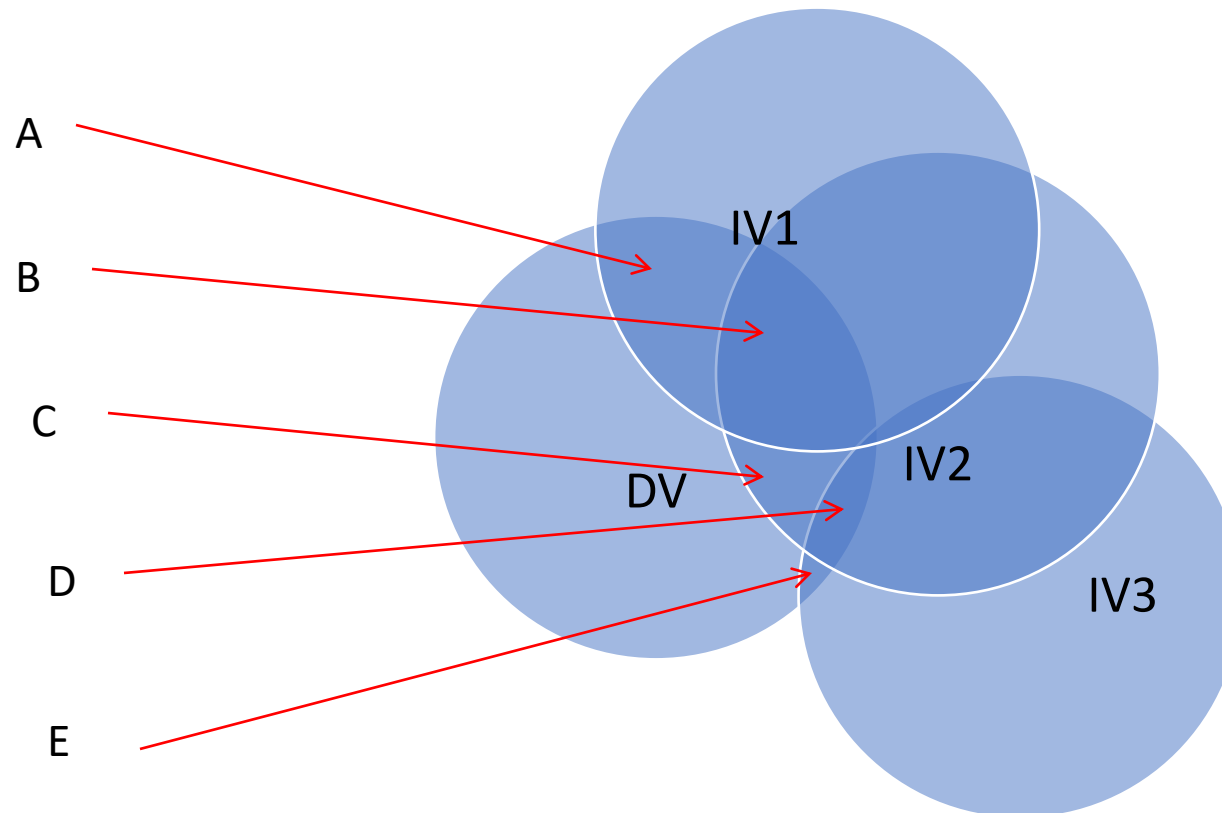
Standard Multiple Regression

- In the standard model all IVs are entered into the regression equation simultaneously; each one is thus assessed as if it had entered the regression after all other IVs had been entered.

- **Each IV is evaluated in terms of what it adds to prediction of the DV that is different from the predictability afforded by all other IVs.**

Using standard multiple regression IV1 "gets credit" for area A, IV2 gets credit for area C, and IV3 gets credit for area E. That is, each IV is assigned only the area of its unique contribution. The overlapping areas, B and D, contribute to the r-squared, but are not assigned to any of the individual IVs. The effects of the other variables have been partialled out.
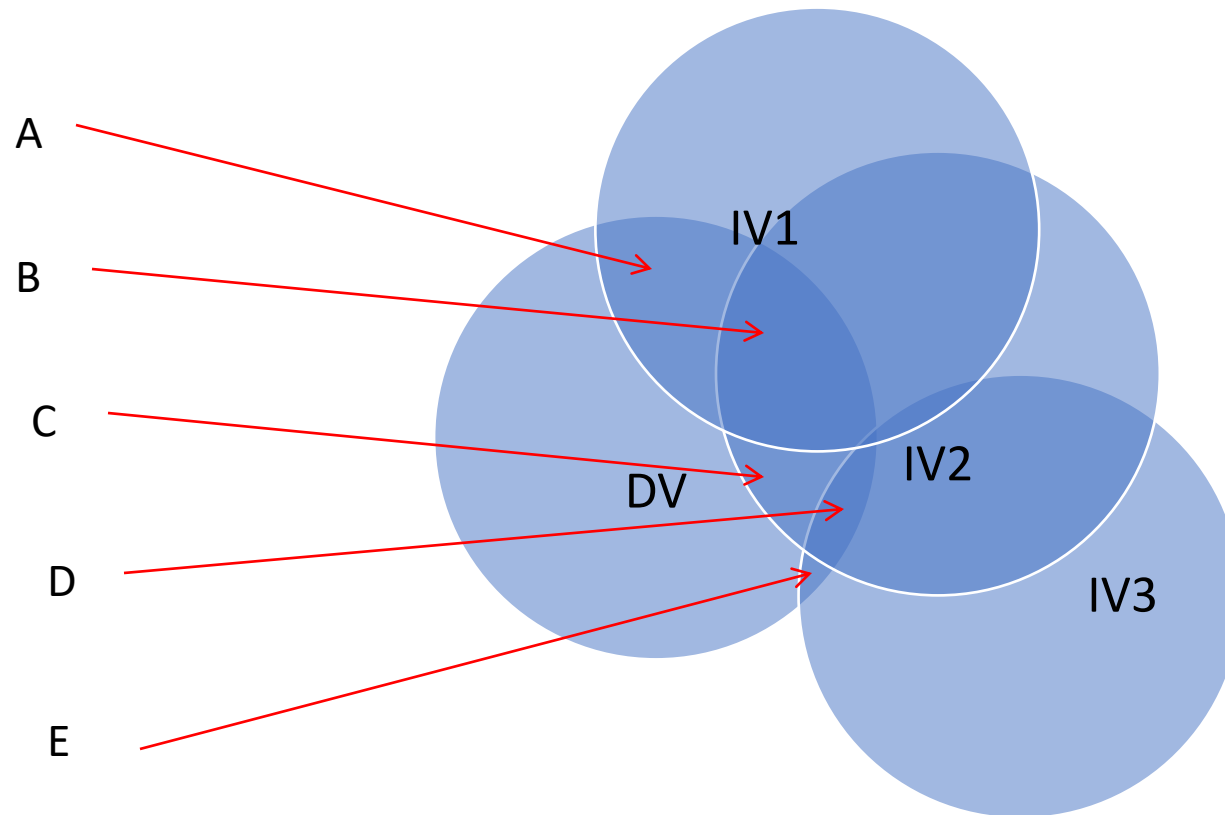
Precision of the slope estimates

- The variance of the slope estimate, $\widehat{\beta_1}$, is the width of the $\hat{\beta}_1$ distribution. When certain regression assumptions are met the variance is calculated as follows for the **bivariate case**:

$$var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{N \; X \; var(X)}$$

- Where:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N-k}$$

Precision of the slope estimates

- The variance of the slope estimate, $\widehat{\beta_1}$, is the width of the $\hat{\beta}_1$ distribution. When certain regression assumptions are met the variance is calculated as follows for the **multivariate case**:

$$var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{N \; X \; var(X)(1 - R_j^2)}$$

- Where:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N - k}$$

Omitted variable bias

What happens when we fail to include a variable in our model that should be there?

Also, what happens when we add a variable into our model that shouldn't be there?

Model misspecification

- Recall, that we never know the **true model**.

- When interpreting our parameter estimates and making inferences based on our model, we assume that it has been properly specified.
  - This is noted by the regression assumption that each independent variable is uncorrelated with the error term.

Model Specification cont…

- For instance, if the true model is

$$\hat{salary} = \beta_0 + \beta_1 ceoten + \beta_2 profits + \varepsilon$$

- And we run

$$\hat{salary} = \beta_0 + \beta_1 ceoten + \varepsilon$$

- Then the error term is now composed of both the true error in the model (or what Berry (1993) terms free will or the intrinsic randomness in human behavior) plus the omitted variable. If the omitted variable is theoretically correlated with the included variable, then there is concern over our assumption of independence between predictors and error term (SLR.4/MLR.4). We will see why this is a concern in a few slides.

Model Specification cont…

- Two types of model specification issues:
  - We include an irrelevant variable in our model (this is often termed **overspecifying** the model).

  - We exclude a relevant variable from our model (this is often termed **underspecifying** the model).

Including an **Irrelevant** Variable

- For instance, suppose the true model is:

$$y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \mu_i$$

- And we estimate:

$$y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

- Hence, we developed a model that included a variable that has no partial effect on our dependent variable.

- More specifically, $\beta_3 = 0$, indicating that x_3 has no effect on y after x_1 and x_2 have been controlled for. (Note: x_3 may or may not be correlated with x_1 and x_2; all that matters is that once x_1 and x_2 are controlled for, x_3 has no effect on y).

- What are the consequences?

Including an Irrelevant Variable

- Including one or more irrelevant variables in a multiple regression model, or overspecifying the model, does not affect the unbiasedness of the OLS estimators. Does this mean that it is harmless to include irrelevant variables? No. Including irrelevant variables can have undesirable effects on the variances of OLS estimators (Wooldridge, 2013, p. 88).

Example of including an irrelevant variable

- Let's create a variable that is drawn completely at random...thus it should not be in our model because it should not be related to our dependent variable.

```
> prest$rand.var = rnorm(102)
>
> lm2 = lm(prestige ~ education + income + rand.var, data=prest)
> summary(lm2)

Call:

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.8883962  3.2397127  -2.126    0.036 *
education    4.1442014  0.3519175  11.776  < 2e-16 ***
income       0.0013538  0.0002277   5.944 4.27e-08 ***
rand.var    -0.1797843  0.8096146  -0.222    0.825
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.848 on 98 degrees of freedom
Multiple R-squared: 0.7981,    Adjusted R-squared: 0.7919
F-statistic: 129.1 on 3 and 98 DF,  p-value: < 2.2e-16
```

Excluding a **<u>Relevant</u>** Variable

- Assume the true population model is:

$$wage = \beta_0 + \beta_1 educ_i + \beta_2 income_i + \mu_i$$

- And due to some reason we specify and run the model:

$$wage = \beta_0 + \beta_1 educ_i + \varepsilon_i.$$

- So now our residual term, ε, is made up of μ (the true model error) and $\beta_2 income$.

- What are the consequences?

Excluding a Relevant Variable

- What are the consequences?
  - When we exclude a variable that should be in our model we potentially bias our parameter estimates, something referred to as **omitted variable bias**.

- Under what conditions do we face these consequences?
  - As long as the omitted variable and the variable included in the model are correlated & and omitted variable affects the dependent variable, omitted variable bias is present. If the two variables are uncorrelated, there is no bias.

- Think of our Venn diagrams.

Take a second to draw the Venn Diagram to represent what happens when only 1 of the 2 conditions for omitted variable bias are present.

What to Do about Omitted Variable Bias

- The main thing we can do is attempt to understand the direction of the bias. On previous slide. From before, the true model was:

  $$\text{wage} = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{income}_i + \mu_i$$

- And we estimated:

  $$\text{wage} = \beta_0 + \beta_1 \text{educ}_i + \varepsilon_i.$$

- If we believe that income has a positive correlation with wage and also that it is positively correlated with education, then the omitted variable bias is positive.

Direction of Omitted Variable Bias

If the true population model is:

wage = β_0 + $\beta_1$educ$_i$ + $\beta_2$income$_i$ + μ_i

And due to some reason we specify and run the model:

wage = β_0 + $\beta_1$educ$_i$ + ε_i.

The direction of the **omitted variable bias in β_1** can be determined from the table below.

	Corr(x_1,x_2) >0	**Corr(x_1,x_2) <0**
B_2>0	Positive bias	Negative bias
B_2<0	Negative bias	Positive bias

Some Econometric Terminology

- Upward bias occurs when $\hat{\beta}_1 > \beta_1$

- Downward bias occurs when $\hat{\beta}_1 < \beta_1$

- Biased toward zero refers to instances when $\hat{\beta}_1$ is closer to zero than β_1
    - Therefore, if the true beta is positive, then the estimated beta is biased toward zero if it has a downward bias.
    - If the true beta is negative, then the estimated beta is biased toward zero if it has an upward bias.

Example of Omitted Variable Bias

```
> lm.true = lm(prestige ~ education + income, data=prest)
> summary(lm.true)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.8477787  3.2189771  -2.127   0.0359 *
education    4.1374444  0.3489120  11.858  < 2e-16 ***      True Model
income       0.0013612  0.0002242   6.071 2.36e-08 ***
---
```

```
> lm.omitted = lm(prestige ~ education, data=prest)
> summary(lm.omitted)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -10.732      3.677   -2.919  0.00434 **
education      5.361       0.332   16.148  < 2e-16 ***      Omitted Variable Model
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> cor.test(prest$education, prest$income)

        Pearson's product-moment correlation

data:  prest$education and prest$income
t = 7.0753, df = 100, p-value = 2.079e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4315828 0.6940914
sample estimates:
      cor
0.5775802
```

How do we explain what is happening to the coefficient on education?

Let's look at an example from the text

- **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)

Example from the text…

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

Again, here is the 'true' model

- **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

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

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

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

- **Conclusion:** $\boldsymbol{\beta^{OmitX_2} = \beta_1 + \beta_2\delta_1}$

Example from the text cont…

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

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

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

Will both be positive

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

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

Another example of omitted variable Bias
(Bailey 2016)

	Bivariate	Multivariate
Adult height	0.412*	
	(0.0975)	
	[t = 4.23]	
Adolescent height		
Athletics		
Clubs		
Constant	−13.09	
	(6.90)	
	[t = 1.90]	
N	1,910	
σ̂	11.9	
R²	0.01	

Standard errors in parentheses.

* indicates significance at p < 0.05, two-tailed.

Another example of omitted variable Bias
(Bailey 2016)

	Bivariate	Multivariate (a)
Adult height	0.412* (0.0975) [$t = 4.23$]	0.003 (0.20) [$t = 0.02$]
Adolescent height		0.48* (0.19) [$t = 2.49$]
Athletics		
Clubs		
Constant	−13.09 (6.90) [$t = 1.90$]	−18.14* (7.14) [$t = 2.54$]
N	1,910	1,870
$\hat{\sigma}$	11.9	12.0
R^2	0.01	0.01

Standard errors in parentheses.

* indicates significance at $p < 0.05$, two-tailed.

Another example of omitted variable Bias
(Bailey 2016)

	Bivariate	Multivariate	
		(a)	(b)
Adult height	0.412*	0.003	0.03
	(0.0975)	(0.20)	(0.20)
	[t = 4.23]	[t = 0.02]	[t = 0.17]
Adolescent height		0.48*	0.35
		(0.19)	(0.19)
		[t = 2.49]	[t = 1.82]
Athletics			3.02*
			(0.56)
			[t = 5.36]
Clubs			1.88*
			(0.28)
			[t = 6.69]
Constant	−13.09	−18.14*	−13.57
	(6.90)	(7.14)	(7.05)
	[t = 1.90]	[t = 2.54]	[t = 1.92]
N	1,910	1,870	1,851
$\hat{\sigma}$	11.9	12.0	11.7
R^2	0.01	0.01	0.06

Standard errors in parentheses.

* indicates significance at $p < 0.05$, two-tailed.

Multiple regression assumptions

From last week

- Assumptions:
  - SLR.1 – Linear in parameters
  - SLR.2 – Data drawn from a random sample (i.e., the errors are independent - no autocorrelation in the data)
  - SLR.3 – Sample variation in the explanatory variables
  - SLR.4 – Zero conditional mean for the error term
  - SLR.5 – Homoskedasticity (i.e., the errors have equal variance)
  - Normality of the error term
  - Validity (i.e. data maps to research question)

MLR Assumptions

- MLR. 1 - Linear in parameters
- MLR. 2 - Data drawn from a random sample (i.e., the errors are independent - no autocorrelation in the data)
- MLR.3 - No perfect collinearity (no exact linear relationship among the independent variables)
- MLR.4 - Zero conditional mean for the error term
- MLR.5 - Homoskedasticity (i.e., the errors have equal variance)
- Normality of the error term
- Validity (i.e. data maps to research question)

In Class Exercises

- 1. Using the prestige dataset regress *prestige* on *education*, *income,* and *women*.
  - Prove to yourself that the coefficient on *education* is the partialled out effect of *education* on *prestige*.
  - Two steps needed to do this:
    - 1) Regress the explanatory variable on all other explanatory variables
    - 2) Regress *Y* on the residuals from this regression

- Complete computer exercises C1 and C2 in Wooldridge. I have them posted on the following slide if you don't have your text.

C1 A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth weight that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother. An equation that recognizes this is

$$bwght = \beta_0 + \beta_1 cigs + \beta_2\, faminc + u.$$

(i) What is the most likely sign for β_2?

(ii) Do you think *cigs* and *faminc* are likely to be correlated? Explain why the correlation might be positive or negative.

(iii) Now, estimate the equation with and without *faminc*, using the data in BWGHT .RAW. Report the results in equation form, including the sample size and *R*-squared. Discuss your results, focusing on whether adding *faminc* substantially changes the estimated effect of *cigs* on *bwght*.

C2 Use the data in HPRICE1.RAW to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

where *price* is the house price measured in thousands of dollars.

(i) Write out the results in equation form.

(ii) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

(iii) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).

(iv) What percentage of the variation in price is explained by square footage and number of bedrooms?

(v) The first house in the sample has *sqrft* = 2,438 and *bdrms* = 4. Find the predicted selling price for this house from the OLS regression line.

(vi) The actual selling price of the first house in the sample was $300,000 (so *price* = 300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

Note, the required datasets are on Blackboard as R datasets