Lecture 6

Multiple Regression

Outline

- R-square and adjusted R-square
- Multiple regression
  - Interpret $\hat{\beta}$
  - formula for $\hat{\beta}$
- F-test
- Unbiasedness and consistency
- The notion of control

R-square: how well did our line fit the data?

- R^2 is the goodness of fit. It is the most widely used measure of fit.
- $\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i {e_i}^2$
- TSS = ESS + RSS
- $R^2 = \frac{ESS}{TSS} = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i {e_i}^2}{\sum_i (Y_i - \bar{Y})^2}$
- The R^2 is the ratio of explained variation to total variation
  - The proportion of total variation in Y that our model has captured (with independent variables)
  - We can use R^2 to compare models if our objective is to fit data.

Examples of R-square

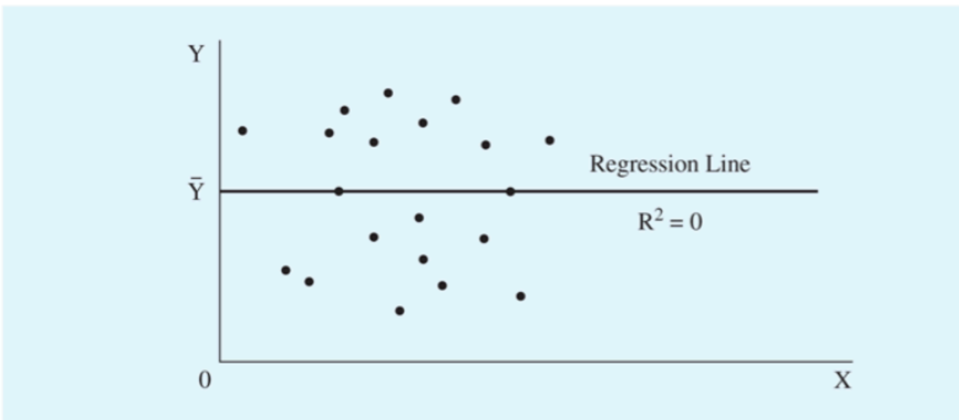


Figure 2.4
X and Y are not related; in such a case, R^2 would be 0.

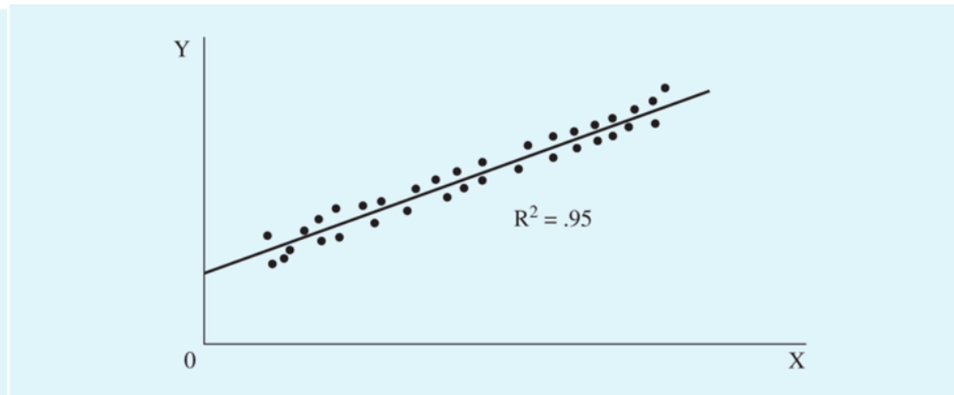


Figure 2.5
A set of data for X and Y that can be "explained" quite well with a regression line (R^2 = .95).

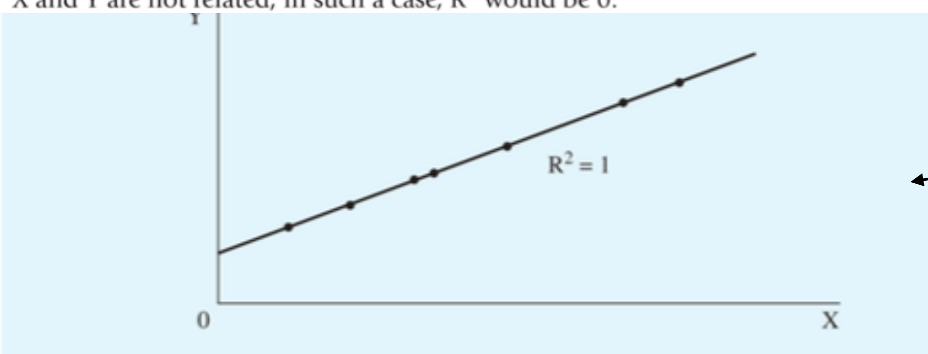


Figure 2.6
A perfect fit: all the data points are on the regression line, and the resulting R^2 is 1.

e.g. $WAGE_i = \beta_0 + \beta_1 WAGE_i + \varepsilon_i$

Exercise: compare models using R-square

Source	SS	df	MS
Model	611.927847	1	611.927847
Residual	60354.2942	46948	1.28555624
Total	60966.222	46949	1.29856274

Source	SS	df	MS
Model	4293.73993	1	4293.73993
Residual	56672.4821	46948	1.20713304
Total	60966.222	46949	1.29856274

- Model 1: age and health status
  - reg phstat age_yrs
- Model 2: education and health status
  - reg phstat educ_r1
- Manually calculate the R-squares from ESS and RSS
- Interpret what R-square means in each model in words
- Which model is better in terms of making predictions (fitting data)? Why?
- Is it necessary that the bigger R-square the better our model is?
- Does big R-square necessarily imply big β? Why?
- Note the total SS are the same in both models, why?

Adjusted R-square

- Let's compare the R-square of these two models
  - Model 1: $Y=\beta_1 X_1+\varepsilon$
  - Model 2: $Y=\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\varepsilon$

- Which model have a bigger R-square? (R-square is the explained variation divided by the total variation-which model can explain more of the variations in Y?)
  - As we add more X in the model, R-square will ALWAYS increase
  - We need another measure to penalize for adding more irrelevant variables

Adjusted R-square

- Adjusted R^2 denoted $\overline{R^2}$

$$\overline{R}^2 = 1 - \frac{\sum e_i^2/(N-K-1)}{\sum(Y_i - \overline{Y})^2/(N-1)}$$

Where N – K – 1 = degrees of freedom for residual, N is sample size, K is number of coefficients excluding the constant

- As K increases (number of independent variables increases), what happens to $\overline{R^2}$?
- Don't have to memorize the exact formula, but need to understand the intuition: **adjusted R-square is an improvement over R-square. It penalizes for adding more independent variables (bigger K)**.
- People usually look at the adjusted R-square to evaluate how well they fit the regression line

STATA output of R-square and adjusted R-square

Source	SS	df	MS
Model	611.927847	1	611.927847
Residual	60354.2942	46948	1.28555624
Total	60966.222	46949	1.29856274

Number of obs =	46950
F(1, 46948) =	476.00
Prob > F =	0.0000
R-squared =	0.0100
Adj R-squared =	0.0100
Root MSE =	1.1338

phstat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age_yrs	.0198173	.0009083	21.82	0.000	.018037	.0215977
_cons	1.426175	.0579778	24.60	0.000	1.312538	1.539813

T or F: Adjusted R-square is never bigger than R-square

Our Regression

```
. reg wage female

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =   68.54
       Model |  828.220467     1  828.220467           Prob > F      =  0.0000
    Residual |  6332.19382   524  12.0843394           R-squared     =  0.1157
-------------+------------------------------           Adj R-squared =  0.1140
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.4763

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -2.51183   .3034092    -8.28   0.000    -3.107878   -1.915782
       _cons |   7.099489   .2100082    33.81   0.000     6.686928     7.51205
------------------------------------------------------------------------------
```

Comparing wages for women and men with same education

```
. reg wage female educ

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   91.32
       Model |  1853.25304     2  926.626518           Prob > F      =  0.0000
    Residual |  5307.16125   523  10.1475359           R-squared     =  0.2588
-------------+------------------------------           Adj R-squared =  0.2560
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.1855

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.273362   .2790444    -8.15   0.000    -2.821547   -1.725176
        educ |   .5064521   .0503906    10.05   0.000     .4074592     .605445
       _cons |   .6228168   .6725334     0.93   0.355     -.698382    1.944016
------------------------------------------------------------------------------
```

Same education and experience

```
. reg wage female educ exper

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =   77.92
       Model |  2214.74206     3  738.247353           Prob > F      =  0.0000
    Residual |  4945.67223   522  9.47446788           R-squared     =  0.3093
-------------+------------------------------           Adj R-squared =  0.3053
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.0781

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.155517   .2703055    -7.97   0.000    -2.686537   -1.624497
        educ |   .6025802   .0511174    11.79   0.000     .5021591    .7030012
       exper |   .0642417   .0104003     6.18   0.000     .0438101    .0846734
       _cons |  -1.734481   .7536203    -2.30   0.022    -3.214982   -.2539797
------------------------------------------------------------------------------
```

Takeaways about R-square and adjusted R-square

- Adjusted R-square is a standardized measure that can be used to compare models. It is useful if our objective is to making prediction
- In micro-econometrics, R-square is not that useful. Because our objective is NOT fitting the data, but to find whether one specific X has causal impact on Y.
  - e.g. How many factors can explain variation in one's health status?
  - Perhaps many more than we can include in the regression
  - It's natural to get a small R-square if we only include education as X
  - Our interest lies in whether education has caused variation in health, not in predicting health status based only on one variable

Multiple regression

- Definition: a regression with more than one independent variable
- Back to the health status example. You might want to include both age and education in the regression. Now we have two independent variables, it is a multiple regression.
- The general multiple regression model with K independent variables is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + \varepsilon_i$$

How to interpret β?

- A big difference between multiple and single regression model is in the interpretation of the slope coefficients
- Now a slope coefficient indicates the change in the average of the dependent variable associated with a one-unit increase in the explanatory variable *holding the other explanatory variables constant or fixed*
- Example: $Health_i = \beta_0 + \beta_1 Educ_i + \beta_2 Age_i + \varepsilon_i$
- $\widehat{\beta}_1$=-0.15: Holding other explanatory variables constant, one unit increase in education level is associated with 0.15 unit decrease in average health index. (note here 1=excellent, so it actually increase health)
- Exercise: $\hat{\beta}_2$=0.07, explain what it means in the context.

Mathematical formula for β (difficult)

- We run a multiple regression
  $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_N X_{Ni} + \varepsilon_i$
- We could run a regression of X_{1i} on all of the other X variables, and then compute the residuals- call it $\widetilde{X_{1i}}$(partial out the effect of all the other X variables)
- Then $\hat{\beta}_1 = \frac{Cov(\widetilde{X_{1i}}, Y_i)}{Var(\widetilde{X_{1i}})}$
- The relationship between $X1$ and Y holding all other X variables fixed

F-test

- With a multiple regression $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_N X_{Ni} + \varepsilon_i$, we can still perform statistical inference using p-value or confidence interval to determine if a specific β is **statistically different** from 0 (or other number)
- Now if we want to test whether a group of variables jointly have any effect on Y, we will use F-test.
- H_0: $\beta_1 = \beta_2 = .. = \beta_N = 0$
- H_a: at least one of the βs is not zero

Example of F-test

- $Health_i = \beta_0 + \beta_1 Educ_i + \beta_2 Age_i + \beta_3 male + \varepsilon_i$
- I want to know whether education, age and gender are jointly affecting health
- $H_0: \beta_1 = \beta_2 = \beta_3 = 0$
- What is H_a?
- Where is F-test in STATA output?
  - Top right panel
  - Like t-stat, F-stat follows a specific distribution, a bigger F-stat means higher power to reject the H_0
  - Prob>F is like the p-value in t-test. We can look at "Prob>F" and compare it with α to decide whether rejecting H_0
- Why is this statistics in the section along with R-square? What do you think this F- stat is mostly used for?

```
Number of obs  =    46950
F(  3, 46946)  =  1272.88
Prob > F       =   0.0000
R-squared      =   0.0752
Adj R-squared  =   0.0752
Root MSE       =   1.0959
```

Unbiasedness of OLS

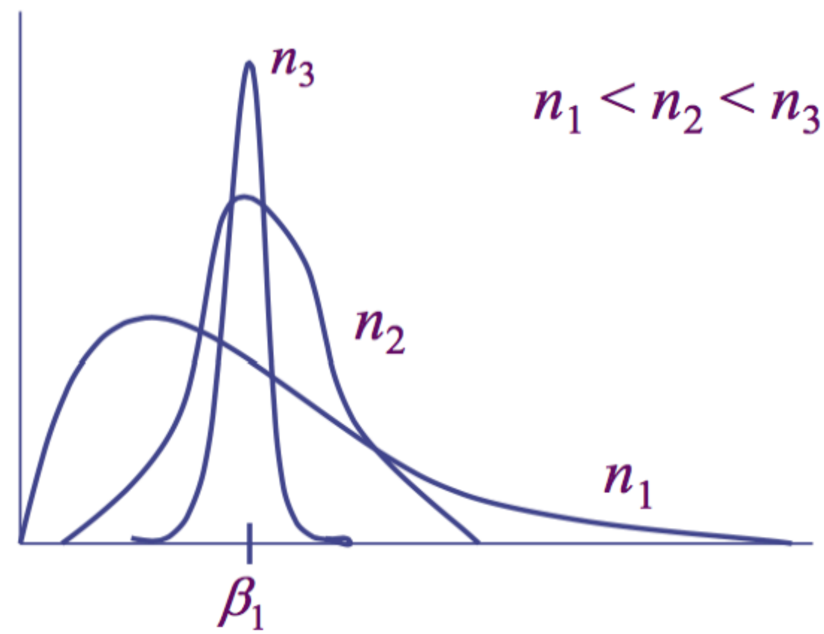
- In the econometric model we have population parameter β, and what we get from data is a sample estimate $\hat{\beta}$. If we draw many samples, we will get multiple $\hat{\beta}$. Those $\hat{\beta}$ form a sampling distribution with a certain mean and standard deviation
- Unbiased estimator:
  - If $E(\hat{\beta}) = \beta$, then $\hat{\beta}$ is unbiased
  - If the mean of all those $\hat{\beta}s$ equal to the true parameter β, then it is unbiased
- If the estimate is unbiased, it is likely we get the true causal effect.
  - It doesn't mean the specific $\hat{\beta}$ is the true β, it means if the sample is "typical", we are near the true effect

When is OLS estimate unbiased?

- There are four conditions
- The most important one is "zero conditional mean", $E(\varepsilon|X)=0$
  - It is an assumption made in the population, it's not about a specific sample
  - The expected value of error term is zero for every given X.
  - In other words, the error is randomly distributed for a given X.
- It is an untestable assumption
- Likely to be true or not? Can you think of an example when OLS is biased?

Consistency of OLS

- The conditions for unbiased estimator are hard to meet, so we settle for estimators that are consistent.
- Consistency: the distribution of the estimate becomes more tightly distributed around β as the sample size grows
- As n→ ∞, the distribution of the estimate $\hat{\beta}$ collapse to the parameter value β



Unbiasedness and consistency

- For unbiasedness, we have "zero conditional mean":

$$E(\varepsilon|X)=0 \quad (1)$$

- For consistency, we have a weaker assumption-"zero mean and zero correlation":

$$E(\varepsilon)=0 \ \& \ Cov(X, \varepsilon)=0 \quad (2)$$

- E(ε)=0 is always true for OLS estimator, so Cov(X, ε)=0 is the key condition
- Why is (2) a weaker condition of (1)?
- Assumption (1) requires the error term to be completely random given each X
  - e.g. ε cannot be related to X^2
- Assumption (2) says that error is not correlate with X, but can be correlated with other stuff
  - e.g. ε can be correlated with X^2

Unbiasedness and consistency

- Unbiasedness is a **finite** sample property
- Consistency is a **large** sample property
- Unbiased means on average you will get the true β
- Consistent means on average you may not get the true β, but as sample size increase, you will be more likely to get true β

- Consistency requires a weaker assumption, therefore the condition to get correct causal estimate is relaxed to **Cov(X, ε)=0**.

OLS with control

- Look at multiple regression from another perspective:
- $Y = \alpha + \beta D + \gamma_1 X_1 + \gamma_2 X_2 + \cdots + \gamma_n X_n + \varepsilon$
- We have treated all independent variables equally. Now we give special attention to our variable of interest, ***the treatment variable D***, and consider other variables $\boldsymbol{X_1}$-$\boldsymbol{X_n}$ as ***controls***
- We put $\mathrm{X_1}$-$\mathrm{X_n}$ in the regression, because we can observe them and the data allow us to measure them. Therefore they are also called ***the observables***
- There can be factors that we want to control but are not observable to us (or we cannot measure them in the data). These factors are called ***unobserved variables (unobservables)***. They are represented by the ***error term ε***

Why do we have biased $\hat{\beta}$?

- Assume the true model for whether insurance affect health is
  - $Health = \alpha + \beta Insured + \gamma_1 Age + \gamma_2 Educ + \gamma_3 Income + \varepsilon$
- We are now running this model:
  - $Health = \alpha' + \beta' Insured + \gamma_1' Age + \gamma_2' Educ + \varepsilon'$
- Is β' the true causal impact of insured on health (β)?
- Recall the condition for β' to be causal is Cov(ε',Insured)=0
- Which variable is now left in ε'?
  - Income
- Is Cov (income, Insured) likely to be 0?
  - No. Therefore $\widehat{\beta'}$ is biased
- Failure to include proper controls result in biased estimates

The notion of control

- Meaning of control #1:
  - When you control for variables you hold that variable constant or fixed
  - Similar notion to a lab experiment –keep other factors held fixed (ceteris paribus)
- Meaning of control #2:
  - e.g. $Health = \alpha + \beta Insured + \gamma_1 Age + \gamma_2 Educ + \gamma_3\ Income + \varepsilon$
  - Within group comparison: compare mean health status of the insured with the uninsured for those with same age, education and income. It's more proper than across group comparison
- Meaning of control #3:
  - "Partialling Out"
  - β tells us the causal effect of whether insured on health after the confounding effect of age, education and income has been partialled (or netted) out

Omitted variable bias

- The bias in the OLS estimator that occurs as a result of an omitted factor is called ***omitted variable*** bias. For omitted variable bias to occur, the omitted factor "*Z*" must be:

   1. A determinant of Y (i.e. Z is part of ε); **and**

   2. Correlated with the regressor X (*i.e.* $\text{corr}(Z,X) \neq 0$)

- ***Both*** *conditions must hold for the omission of Z to result in omitted variable bias*.

Omitted variable bias direction

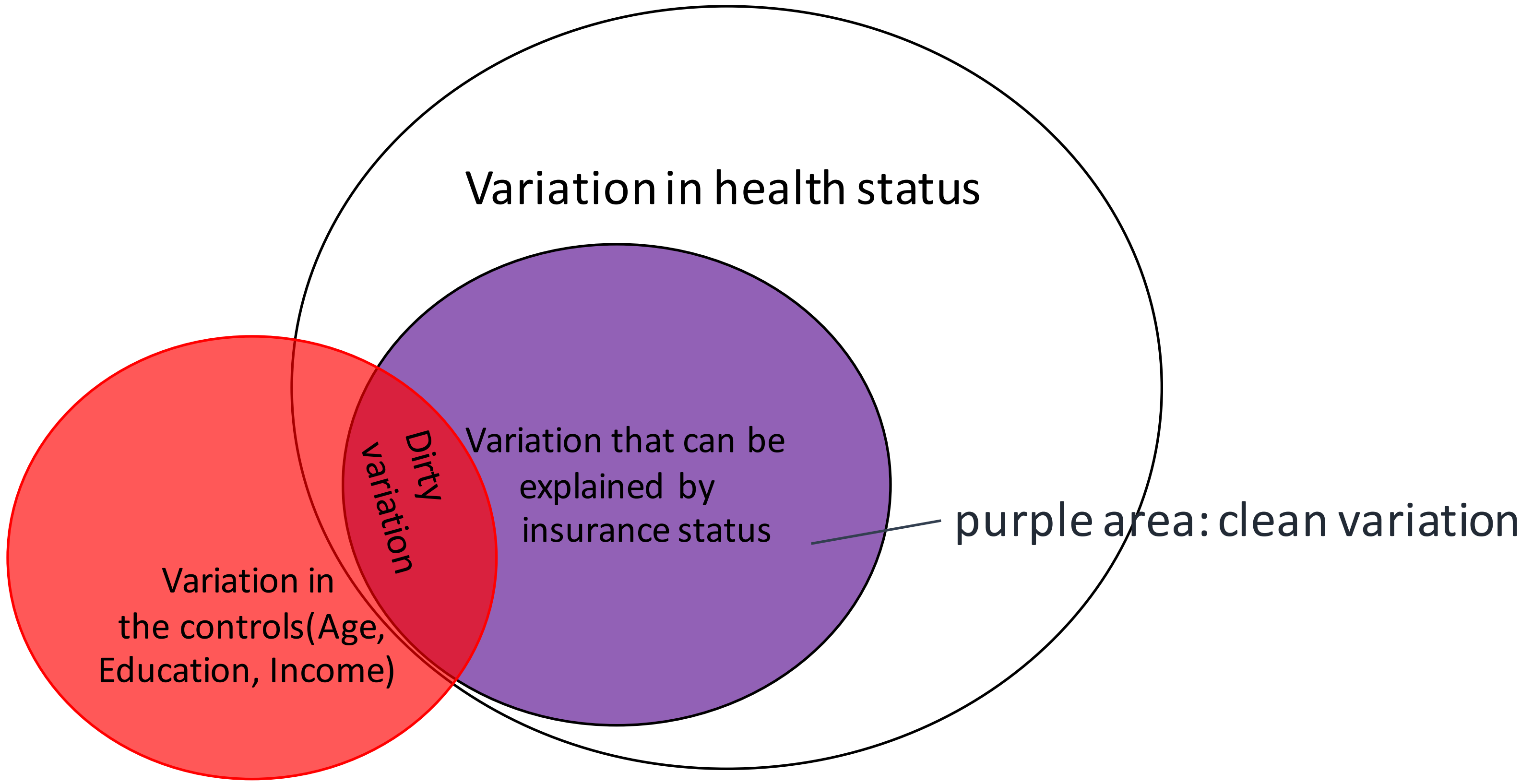
		Corr(omitted variable,x)	
		positive	negative
Corr(omitted variable,y)	positive	upward bias	downward bias
	negative	downward bias	upward bias

The notion of control

- $Health = \alpha + \beta Insured + \gamma_1 Age + \gamma_2 Educ + \gamma_3\ Income + \varepsilon$
- More on "partialling out":
  - If insurance status is randomly assigned, we won't have to control for other factors. In that case, all the variation in insured is "clean".
  - Since insurance status is not randomly assigned, it contains both "clean" variation and "dirty" variation
  - The "dirty" variation is the part that correlates with age, educ or income.
  - By controlling for age, educ and income, we are partialling out the "dirty" variation, leaving the "clean" variation for identification. What does it exactly mean?

Diagram showing clean and dirty variation

$$Health = \alpha + \beta Insured + \gamma_1 Age + \gamma_2 Educ + \gamma_3\, Income + \varepsilon$$



Good, bad and the useless

- Good controls are variables that you include in the regression in order to kill dirty identifying variation
- Useless controls are variables that you include in the regression that aren't correlated with treatment variable D.
  - Useless controls can still explain Y
  - Often it's worthwhile to include these controls to reduce your standard errors, but fundamentally, these controls don't do that much
- Bad controls are variables that are caused by D. Controlling for these variables is called "over controlling"

Example: Immigrantion

- Suppose immigration positively impacts crime rate
- If I regress crime on immigration, I may find a positive effect
- $Crime = \alpha + \beta\ Number\ of\ Immigrants + \varepsilon, \hat{\beta}$>0
- Is this relationship biased?
  - Possibly, because some other factors (e.g. big city indicator) is left in the error term. Once controlling for big city, the effect may change
- Note, when we include a control, we are more concerned whether it affects treatment **variable D**, not whether it affects Y.

Example: good, bad and useless controls

- Variable “big city” affects both immigration and crime rate, therefore it is a **good** control.
- Now you include another variable number of rainy days. You think number of rainy days will affect crime. But it has nothing to do with immigration. Therefore it is a **useless** control. (Although the adjusted R-square will increase, it doesn’t eliminate any “dirty” variation)
- Now consider the variable-local unemployment rate. This variable is likely to be caused by immigrants and therefore may be a **bad** control. Because it will take away the “clean” variation too. You will over-control by including it
  - Note: when you try to think of a bad control, think about the mechanisms why immigrants affect crime rate, controlling for these mechanisms are usually over-controlling

Criteria for good, bad and useless controls

- Good controls affect Y and D (variable of interest)
- Useless controls affect Y but not D
- Bad controls affects Y, and are caused by D (Bad controls are often mechanisms through which D affect Y)

Review for quizzes

- Be able to interpret β in multiple regressions
- Know the meaning, null and alternative hypothesis of F-test
- Understand the conditions for unbiasedness and consistency and the relationship between them
- Understand why we need controls-Because $cov(D,\varepsilon)=0$ is violated, and the estimate is biased
- Be able to come up with examples of three types of controls, and explain why they are good, bad and useless