Empirical Microeconomics: Classical Linear Regression

University of Alabama

September 8, 2016

Last Class

- We formally introduced the Grossman (1972) Model of Health Demand.
- In the model, consumers can both consume health today and invest in their health tomorrow.
- Individuals produce health by using both medical care inputs and time inputs, such as going to the gym.
- After a bit of arduous math, Grossman came to some pretty basic results.
- Overall, an individual's level of health stock holdings depends on more than simply the price of medical care, but also on the depreciation of health across time, a consumer's age, an individual's level of education, etc.

What is Econometrics?

"Econometrics is the study of the application of statistical methods to the analysis of economic phenomena."

What distinguishes an econometrician from a statistician is the former's preoccupation with problems caused by violations of statisticians' standard assumptions; owing to the nature of economic relationships and lack of controlled experimentation, these assumptions are seldom met.

Moreover, econometricians are often concerned with inferring causality as opposed to simply correlation within samples of data.

Econometricians are often accused of using sledgehammers to crack open peanuts, i.e. using unnecessarily complicated statistical techniques to answer very basic questions.

Econometrics is Data Driven

Applied Microeconomics research is data driven. So what is data?

When we talk about economic data, we are referring to information obtained typically via survey of either nations, firms, or individuals.

For example:

- If we survey different nations and ask about Gross Domestic Product, then we have aggregated national-level data on GDP.
- If we survey all the Fortune 500 companies and ask about revenues or expenditures, we have firm-level data.
- If we survey everyone in this room, then we have individual-level data on a sample of University of Alabama college students.
- Many popular health economics datasets survey samples of households that are representative of the U.S. population.

The Disturbance Term

There is a distinction between an economic theorist and an econometrician. A theorist will claim that health is a function of income and write:

$$H = f(Y)$$

An econometrician will claim that this relationship must also include a disturbance term, or an error term, and may write the equation as:

$$H = f(Y) + \varepsilon$$

The inclusion of a disturbance term indicates that the relationship is *stochastic*, i.e. occurring with some randomness, as opposed to *deterministic*, or exact.

The Disturbance Term

$$H = f(Y) + \varepsilon$$

In this model, H is the dependent variable, and Y is the independent, or explanatory variable. Why do we include a disturbance term?

1. Omission of some variables, i.e. specification error (the model is misspecified). Sure, income effects health, but it is certainly not the only factor influencing health. What about luck? Guess what, we don't have any data on luck.

2. Measurement Error, i.e. the outcome variable cannot be measured with exactness due to data collection difficulties or because it is inherently unmeasurable.

3. Human indeterminacy, i.e. some believe that because, in economics, we are studying the behavior of humans, we should include a disturbance term to represent the inherent randomness in human behavior.

The Goal of Econometrics

Econometrics is primarily concerned with obtaining the "good", or "preferred" estimator given a particular situation.

An **Estimator** is a method for arriving at an estimate.

Every situation is different, each with its unique set of statistical problems. Hence, there is no single path to obtaining the correct specification. Preferred specifications will vary drastically across different econometric situations. Always keep in mind that nothing is exact, and the disturbance term represents this inexactness.

You might think of an econometrics textbook as a catalog of which estimators are most desirable in different estimating situations. The researcher, facing a certain set of problems, turns to the catalog to determine which estimator is most appropriate.

The Most Common Estimator

The most commonly used estimator in econometrics is known as the Ordinary Least Squares (OLS) Estimator.

The OLS estimator is considered the optimal estimator in estimating the Classical Linear Regression Model.

When someone claims to run a "regression," they are almost always referring to an OLS regression or a Linear regression.

The OLS estimator is considered optimal under a set of certain assumptions. If these assumptions are violated, then OLS may no longer be optimal.

Classical Linear Regression

We want to determine how an independent (explanatory) variable effects a dependent (outcome) variable. Suppose that our dependent variable is given by y and the explanatory variable of interest is given by x. Then

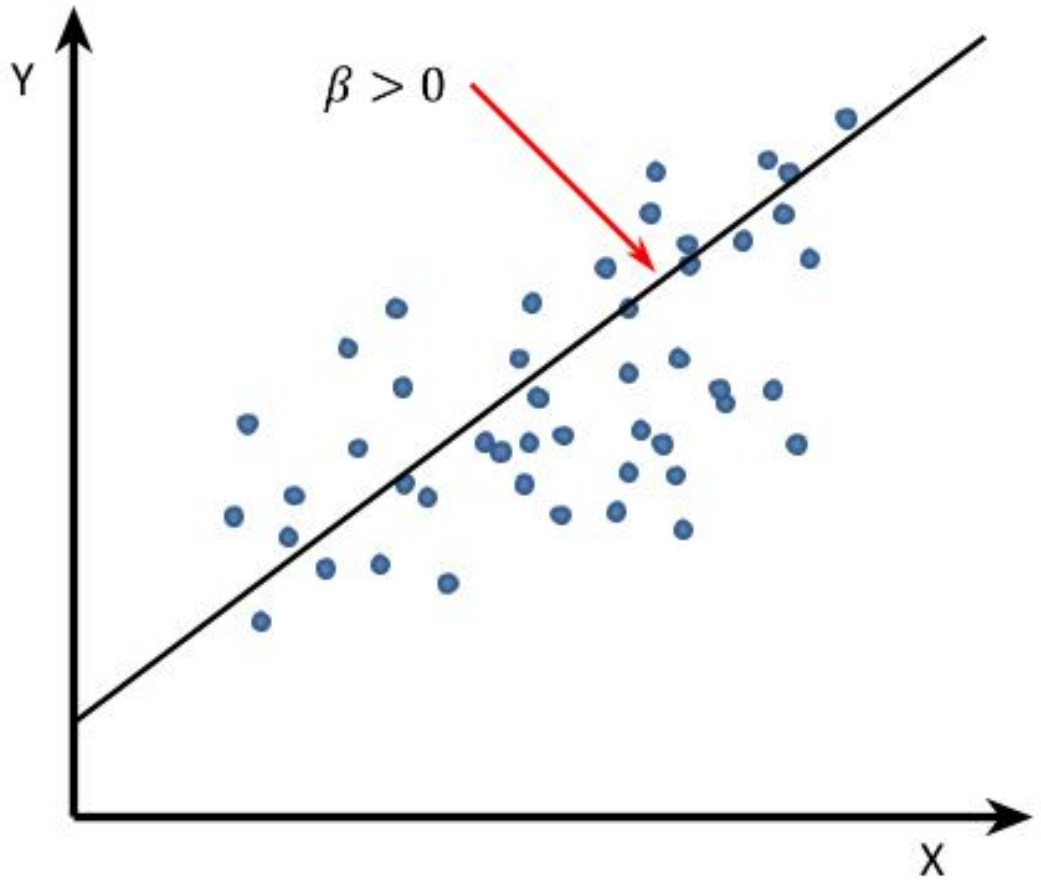
$$y = \alpha + \beta x + \varepsilon$$

represents the standard notation for a linear regression. Recall that ε represents the disturbance (error) term. Here, y and x are data, α, β, and ε are estimated.

The direction and magnitude of the estimated β tells us about the relationship between y and x.

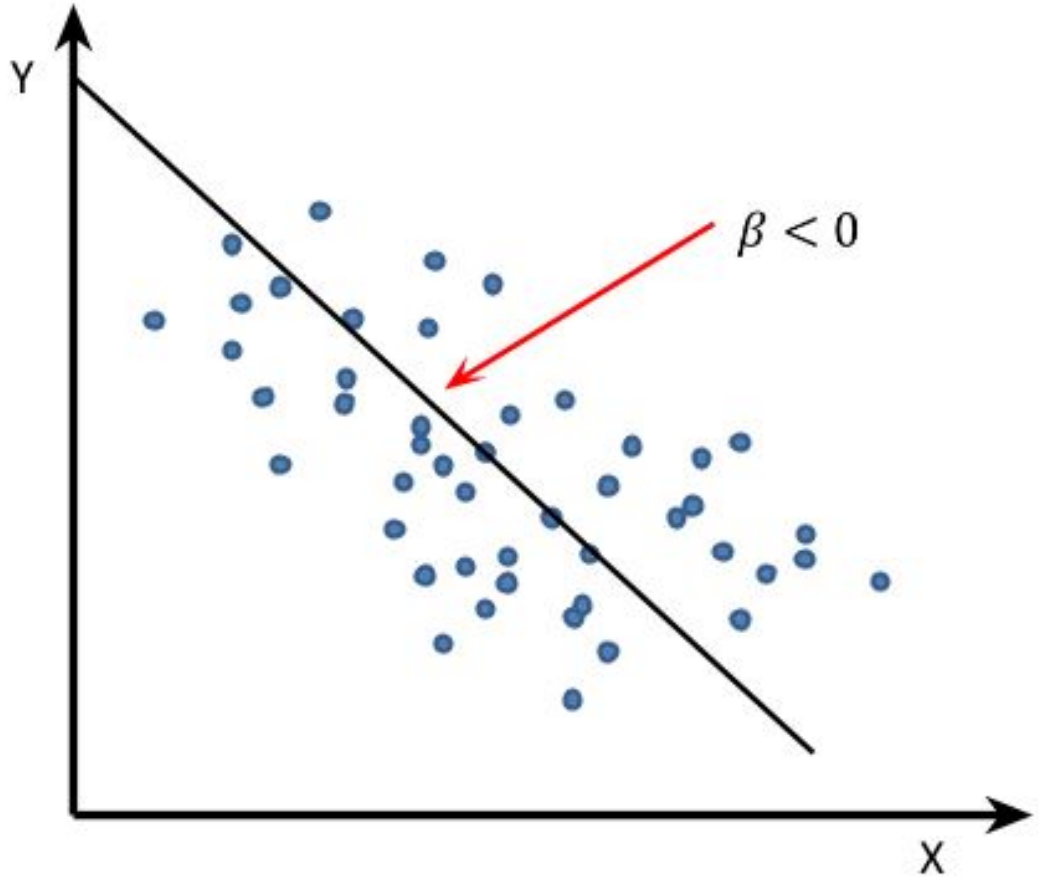
Positive Relationship

- Many different relationships
  - Positive linear



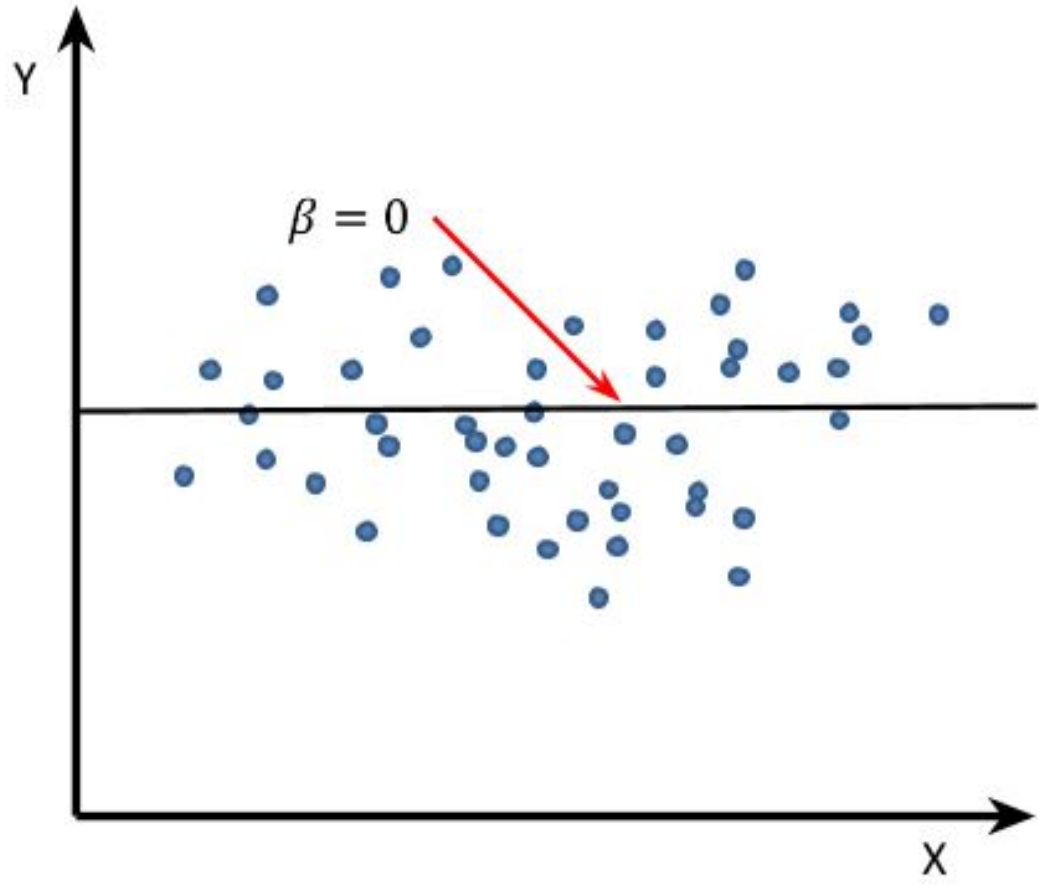
Negative Relationship

- Many different relationships
  - Negative linear



No Relationship

- Many different relationships
  - No relationship



Assumptions of Classical Linear Regression

1. The model is linear in parameters and is correctly specified.
2. The matrix of explanatory variables X must have full rank.
3. Explanatory variables must be exogeneous.
4. The error term must be independently and identically distributed.
5. The error term must be normally distributed in the population.

Assumptions of Classical Linear Regression

1. The model is linear in parameters and correctly specified.

The dependent variable can be calculated as a linear function of a specific set of independent variables plus a disturbance term.

Specification errors will violate this assumption. Specification errors might include:

- using the wrong regressors, i.e. omitting relevant variables or including irrelevant variables.
- Nonlinearity- the relationship between the dependent and independent variables is nonlinear

Assumptions of Classical Linear Regression

2. The matrix of explanatory variables X must have full rank.

   - This simply means that the number of observations must be greater than or equal to the number of explanatory variables.
   - Additionally, no two explanatory variables can have an exact linear relationship. When explanatory variables are approximately linearly related to one another, this is a problem called multicollinearity.

Assumptions of Classical Linear Regression

3. Explanatory Variables must be Exogeneous.

The textbook definition of endogeneity is that the explanatory variable and the error term are influenced by common factors. To avoid endogeneity, we need the explanatory variable X to be independent of the error term ε, i.e.

$$E[\varepsilon \mid X] = 0$$

This assumption is often violated through omitted variable bias or simultaneity, and endogeneity remains the biggest problem in applied microeconomics research.

Assumptions of Classical Linear Regression

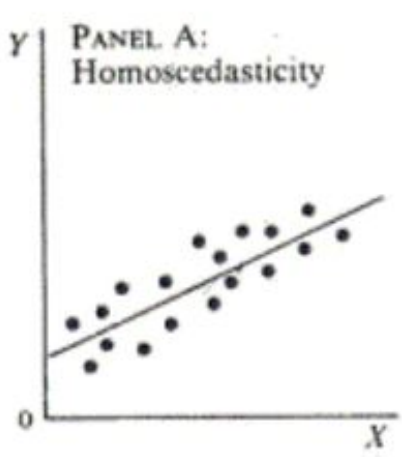
4. The error term must be independently and identically distributed.

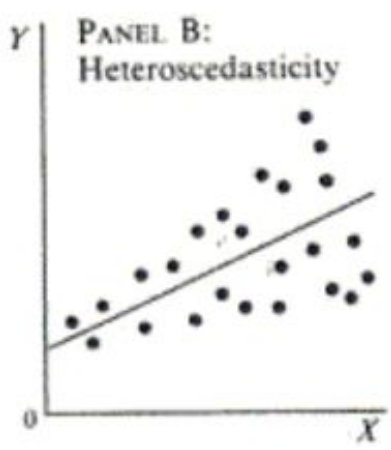
$$\varepsilon_i \sim iid(0, \sigma^2)$$

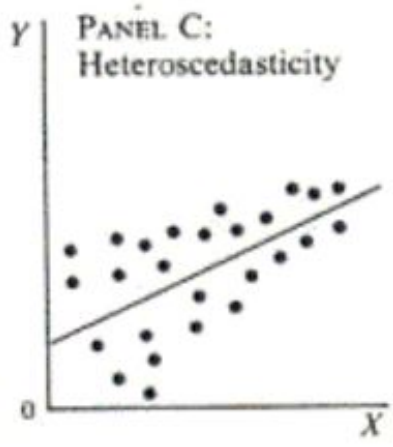
This implies that the expected value, or the mean, of the error term in the population is zero. In other words, the mean of the distribution from which the disturbance term is drawn is zero. A non-zero expectation of the error term may bias the intercept term.

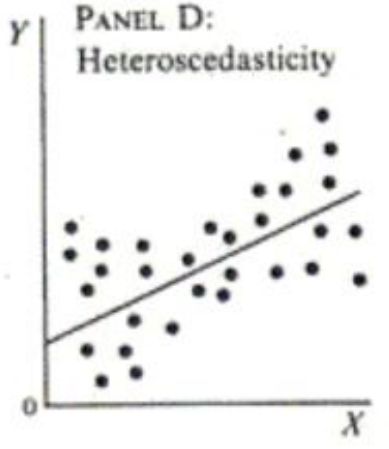
We also assume that the error terms, or residuals, have a constant variance across the sample, i.e. we assume homoskedasticity. A violation of this assumption is occurs when the residuals do not a have a constant variance, i.e. heteroskedasticity. This problem can bias estimates, and this is typically corrected by using "clustered" standard errors.

Heteroskedasticity









Assumptions of Classical Linear Regression

If all of the above assumptions (the Gauss-Markov Assumptions) are met, then the OLS estimator is said to be the Best Linear Unbiased Estimator (BLUE).

Where here "Best" refers to smallest variance. Hence, if all of the Gauss-Markov assumptions hold, then the OLS estimator is the preferred estimator.

Criteria for Estimators

What do we mean by "preferred" estimator?

How do we judge whether an estimator is a good one?

1. Computational Cost
2. Least Squares
3. Highest R^2
4. Unbiasedness
5. Efficiency
6. Asymptotic Properties

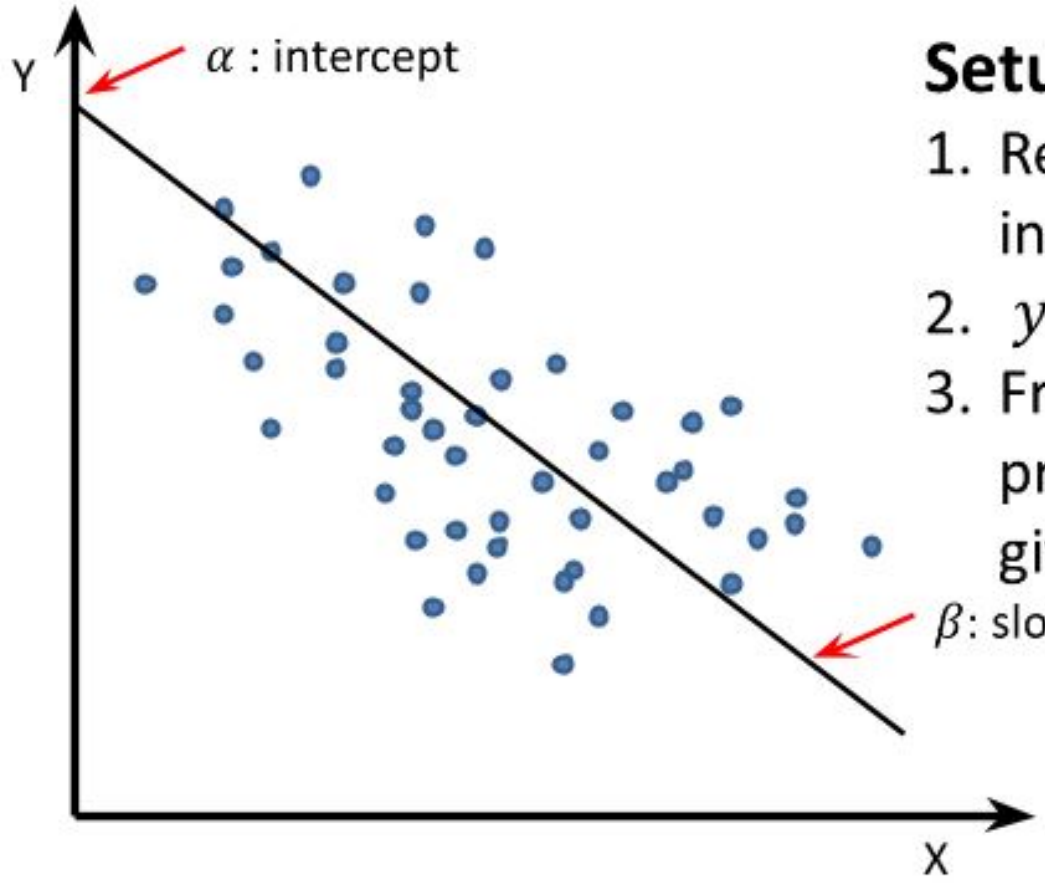
Computational Cost

As economists, we must always compare benefits to costs. Thus, the computational ease and cost of using one estimator rather than another must be taken into account whenever selecting an estimator.

Thanks to modern computing power, computational cost is not as much of a concern as in the old days. Still coding an estimator may prove too cumbersome and not worth the trouble.

Least Squares

Suppose we want to estimate a linear relationship between the dependent variable some measure of health status, y, and an explanatory variable number of cigarettes smoked per day, x.



Setup:

1. Remember slope-intercept formula?
2. $y = \alpha + \beta x$
3. From α and β, can predict value of Y for a given value of X

Least Squares

If we estimate a linear model given by

$$y = \alpha + \beta x + \varepsilon$$

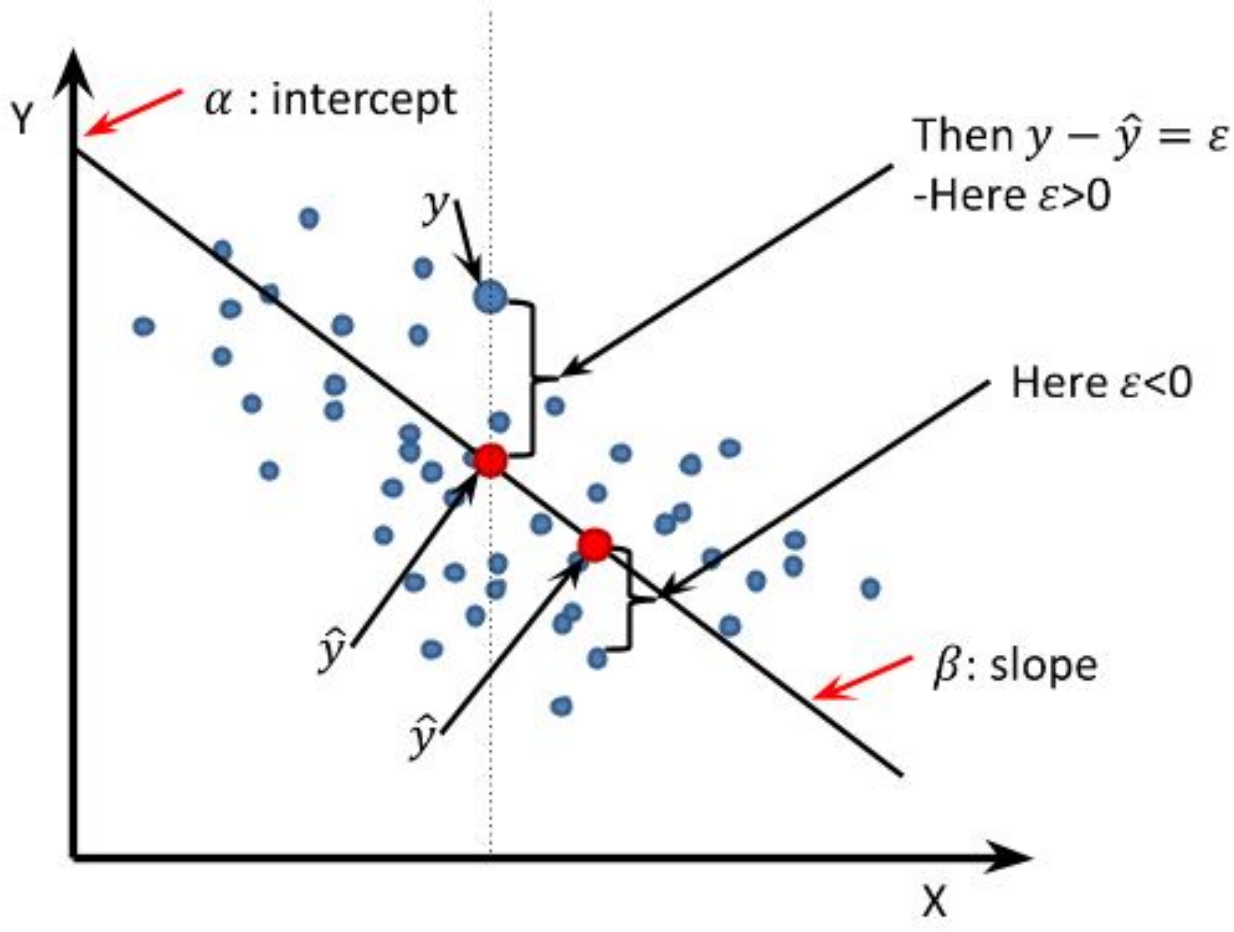
For every x, we can come up with a predicted value for y, call it y-hat. Given some amount of cigarettes smoked per day, we can estimate a health status.

Then

$$y = \hat{y} + \varepsilon$$

Linear regression estimates β such that the sum of squared residuals are minimized.

Least Squares



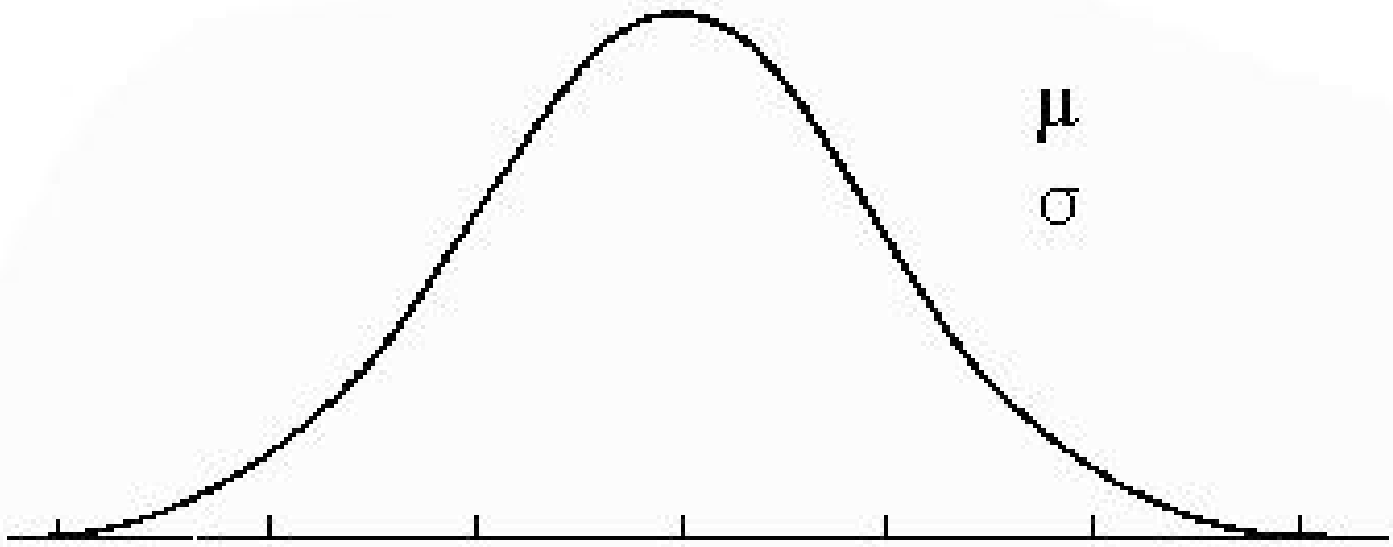
Highest R^2

R-squared is known as the coefficient of determination. It represents the proportion of the variation in the dependent variable "explained" by variation in the independent variables.

Though in the old days people cared a lot about R^2, We are not all that concerned with R-squared in modern times. Typically, within Panel Econometric Models R^2 tends to be pretty low, while in Time-Series R^2 tends to be pretty high.

In econometrics we are primarily concerned with "good" parameter estimates, where "good" is not determined by R^2. We care much more about correctly specifying the model.

Sampling Distribution



Each representative sample comes from some population. When we resample over and over again, each sample will have a mean and a variance. Each mean will be an "unbiased" estimator of the population mean. "Efficiency" refers to minimizing the variance associated with draws from the sampling distribution.

Unbiasedness

Unbiasedness essentially means accurate "on average"

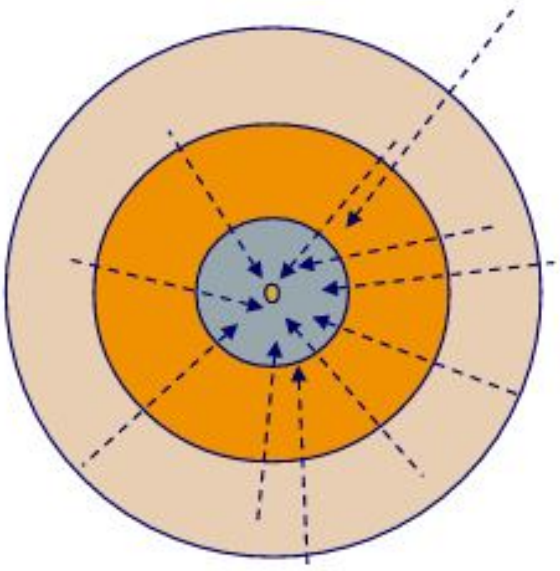
A statistic is said to be an unbiased estimate of a given parameter when the mean of the sampling distribution of that statistic can be shown to be equal to the parameter being estimated.

For example, the mean of a sample is an unbiased estimate of the mean of the population from which the sample was drawn.

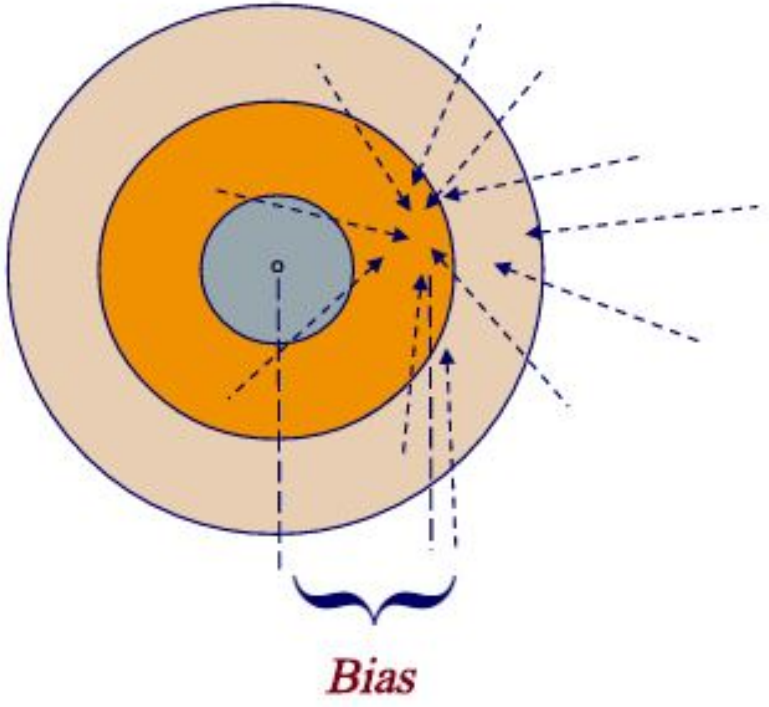
Suppose we want to estimate the average height of a student in this class. In this scenario, the entire group of people in the class makes up the population. We randomly sample ONE and record his or her height. This is an unbiased estimate of the population mean height.

Suppose a crazy statistician samples fifteen students and records their heights, and then adds 0.03% to this average. Though this is likely a good estimate of the average height, this is a biased estimate.

Unbiasedness vs Biasedness



An **unbiased** estimator is on target on average.



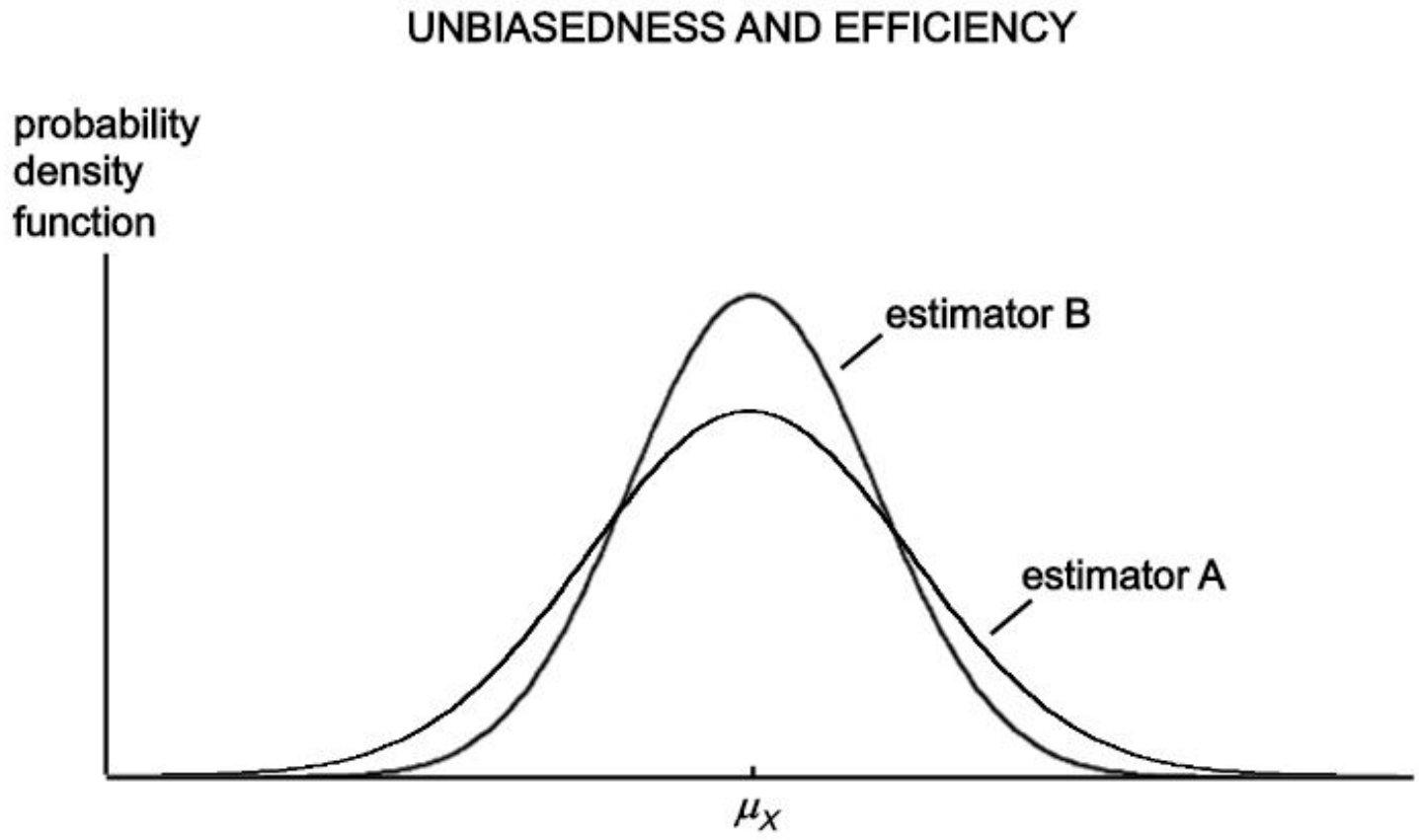
A **biased** estimator is off target on average.

Unbiasedness

Though we have become somewhat obsessed with the unbiasedness criterion in econometrics, many have downplayed the importance of the concept. Recall that unbiasedness means that we would get the correct estimate "on average."

Efficiency

Efficiency refers to the idea of drawing an estimate from a sampling distribution that has as small a variance as possible.



In the diagram, A and B are both unbiased estimators but B is superior because it is more efficient.

Asymptotic Properties

Asymptotics refers to what would happen as the number of observations tends to infinite. Asymptotic theory is relied upon to generalize many finite-sample results to that of a population.

Consistency is a concept associated with the asymptotic behavior of an estimator. An estimator is consistent if as the sample size grows to infinite, we obtain the true population parameter β.

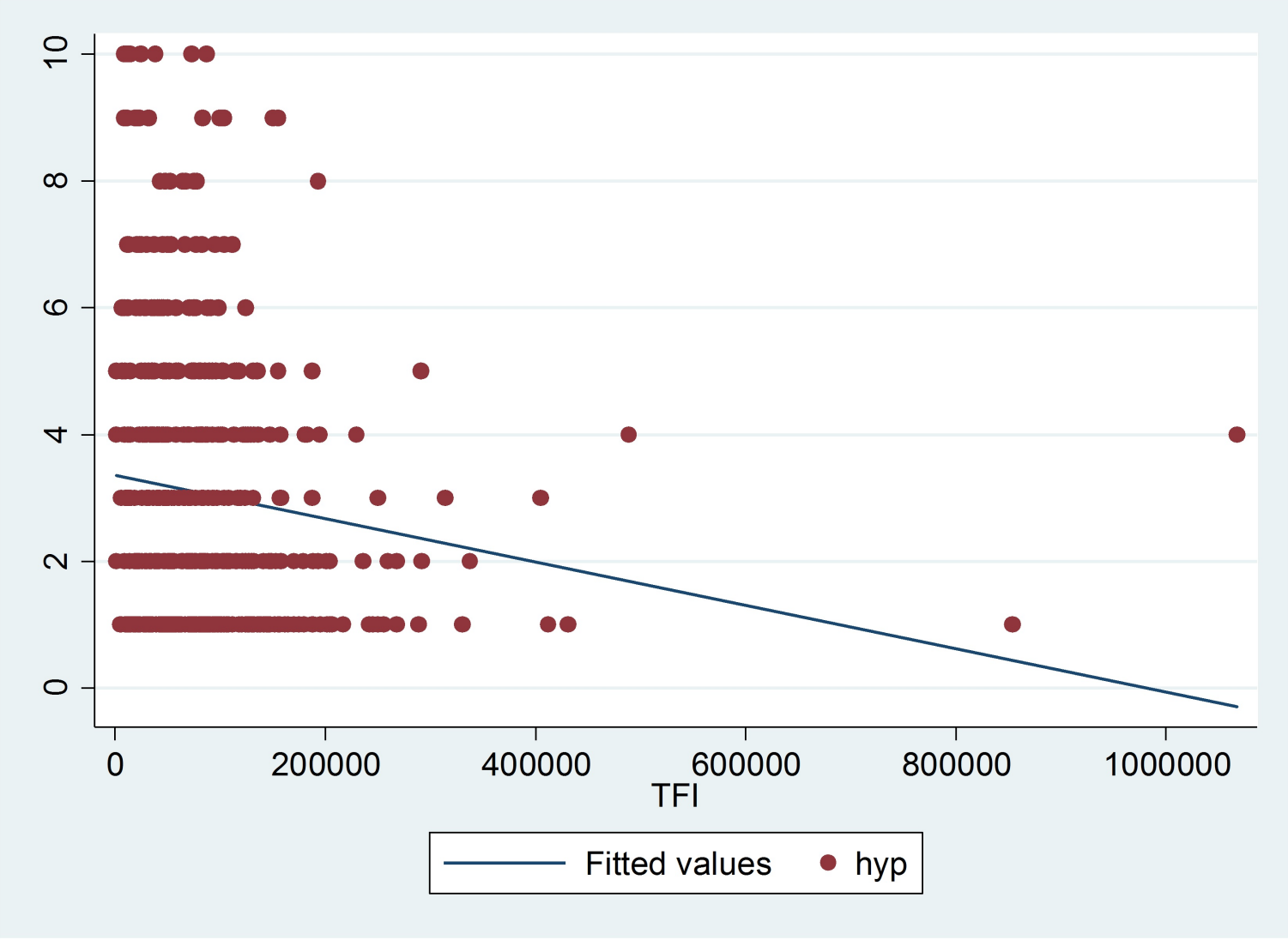
Linear Regression Example

Suppose we want to study the effect of Income on Mother-Reported Hyperactivity, a proxy for ADHD diagnosis. Then our model might be:

$$hyp = \alpha + \beta Income + \varepsilon$$

where our dependent variable on the left-hand side is a report of child behavioral problems ranging from 0 (no behavioral problems) to 10 (extreme hyperactivity) and our independent variable on the right-hand side is Total Family Income. Would you expect the relationship to be positive or negative?

Hyperactivity vs Total Family Income



y-axis is mother-reported hyperactivity ranging from 1 to 10. This plots a fitted, linear relationship for 500 children in 2007

Regression Output

This is keeping only 500 children from the year 2007 and only non-zero reports.

Number of obs	=	500
F(1, 498)	=	9.03
Prob > F	=	0.0028
R-squared	=	0.0178
Adj R-squared	=	0.0158
Root MSE	=	2.2027

hyp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
TFI	-3.42e-06	1.14e-06	-3.01	0.003	-5.65e-06	-1.18e-06
_cons	3.361556	.1396706	24.07	0.000	3.08714	3.635973

Does this model suffer from omitted variable bias?
If Total Family Income goes up by $10,000, then mothers report 0.342 lower levels of hyperactivity. If income goes up by $100,000, mothers report 3.42 lower levels.

Multiple Regression

Very rarely do we ever want to only control for a single explanatory variable. In most cases, we want to include a number of different controls. This situation is called multiple regression. Suppose

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon,$$

where each x represents a different control. For example, maybe we want to control for total family income, sex, race, etc.

When we run a regression, we are essentially determining what our expected outcome would be, i.e. our expected y value, given some value for x. Mathematically speaking, $E[y \mid x]$ So we might imagine what the expected hyperactivity level would be *given* all of the characteristics of a mother or a child.

A Model of Hyperactivity

Suppose we include a host of different controls on the right-hand side. Remember, our outcome variable is mother-reported hyperactivity. Our model might look like:

$$hyp = \alpha + \beta_1 Income + \beta_2 Male + \beta_3 Age + \beta_4 Black +$$

$$\beta_5 White + \beta_6 Female\ Head + \beta_7 Mother\ College +$$

$$\beta_8 Birth\ Order + \beta_9 LowbirthWgt + \varepsilon$$

Note that most of these variables on the right-hand side are dummy variables, i.e. categorical variables taking on values of either 0 or 1.

A Model of Hyperactivity

Number of obs	=	7,920
F(10, 7909)	=	44.00
Prob > F	=	0.0000
R-squared	=	0.0527
Adj R-squared	=	0.0515
Root MSE	=	2.1054

hyp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
TFI	-2.37e-06	4.02e-07	-5.90	0.000	-3.16e-06	-1.58e-06
male	.5615283	.0473606	11.86	0.000	.4686891	.6543675
age	.0667921	.0051836	12.89	0.000	.0566309	.0769533
black	-.1131999	.0790632	-1.43	0.152	-.2681847	.041785
white	.0868678	.0768435	1.13	0.258	-.0637656	.2375013
femalehead	.3017734	.0585459	5.15	0.000	.187008	.4165387
mothercollege	-.3245101	.050112	-6.48	0.000	-.4227429	-.2262773
firstborn	.0357	.05966	0.60	0.550	-.0812494	.1526494
secondborn	-.0094597	.0618844	-0.15	0.879	-.1307694	.1118501
lowb	.097281	.1136938	0.86	0.392	-.125589	.3201509
_cons	1.066129	.0980259	10.88	0.000	.8739725	1.258286

A p-value below 0.05 gives us a significant coefficient at the 95% level. It seems that the male sex, age, and having a female head of household are all positively and significantly related to the mother's reporting of hyperactivity. Alternatively, more educated and higher income mothers tend to report lower levels.

ADHD Diagnosis as a Dependent Variable

Number of obs	=	7,989	
F(10, 7978)	=	23.43	
Prob > F	=	0.0000	
R-squared	=	0.0285	
Adj R-squared	=	0.0273	
Root MSE	=	.24525	

adhd	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
TFI	-9.08e-08	4.66e-08	-1.95	0.051	-1.82e-07	4.88e-10
male	.0618631	.0054933	11.26	0.000	.0510948	.0726314
age	.0052984	.0006006	8.82	0.000	.0041211	.0064758
black	.0135677	.009154	1.48	0.138	-.0043765	.031512
white	.0310646	.0088986	3.49	0.000	.013621	.0485082
femalehead	.0187663	.0067846	2.77	0.006	.0054666	.0320659
mothercollege	-.0153322	.0058088	-2.64	0.008	-.0267189	-.0039454
firstborn	.0071698	.0069223	1.04	0.300	-.0063998	.0207394
secondborn	.0047994	.0071783	0.67	0.504	-.009272	.0188707
lowb	.0103828	.0130603	0.79	0.427	-.0152187	.0359844
_cons	-.0380752	.0113485	-3.36	0.001	-.0603213	-.0158291

This is using hispanics as the reference group. Within this sample, blacks and whites are more likely to be diagnosed with ADHD than hispanics, but the coefficient associated with blacks is insignificant.

Changing the Reference Group

Number of obs	=	7,989	
F(10, 7978)	=	24.28	
Prob > F	=	0.0000	
R-squared	=	0.0295	
Adj R-squared	=	0.0283	
Root MSE	=	.24513	

adhd	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
TFI	-9.55e-08	4.66e-08	-2.05	0.040	-1.87e-07	-4.28e-09
male	.0618075	.0054899	11.26	0.000	.0510459	.0725692
age	.0052948	.0006001	8.82	0.000	.0041185	.0064711
white	.0152787	.006465	2.36	0.018	.0026055	.0279518
hispanic	-.0359992	.0111605	-3.23	0.001	-.0578767	-.0141217
femalehead	.0163837	.0067443	2.43	0.015	.0031632	.0296042
mothercollege	-.0168056	.0058281	-2.88	0.004	-.0282303	-.005381
firstborn	.0074363	.0069185	1.07	0.282	-.0061257	.0209983
secondborn	.0052281	.0071753	0.73	0.466	-.0088374	.0192937
lowb	.009405	.0130519	0.72	0.471	-.0161801	.0349902
_cons	-.0209372	.0100901	-2.08	0.038	-.0407164	-.001158

This is using blacks as the reference group. The positive and significant coefficient with the white variable tells us that whites are 1.5% more likely to be diagnosed with ADHD than blacks.

Limited Dependent Variables

Note that in the three regressions above, the dependent variable was mother-reported hyperactivity in the first, and ADHD diagnosis in the latter two. Hyperactivity takes on discrete values from 0 all the way up to 10, while ADHD diagnosis takes on values of either 0 or 1.

These discrete dependent variables are called limited dependent variables. In many cases OLS may not be preferred to other forms of regression including logit or probit. In most cases, however, OLS will give qualitatively similar results.

Also, note that in the above regressions, we have not constructed any "idenfication strategy" to infer causality of any kind. We have simply performed a discriptive analysis and showed that male, black or white children coming from low-income, female-headed households are more likely to be diagnosed with ADHD. Due to the lack of an identification strategy, these are correlations and not causations.

Next Class

- What are common identification strategies within Health Economics?
- Insurance Theory (Ch. 8 FGS)