Unit 3: Ordinary Least Squares Estimation

Readings: Wooldridge Chapter 2, 3 (except Omitted Variable Bias: the Simple case and Omitted Variable Bias: More General Cases)

Introducing the OLS Population Model

Reading: Chapter 2

Ordinary Least Squares Regression

- Remember that we begin an analysis by assuming a population model.
  - We always need some assumptions about the world before we can do anything. We have to ensure that our parameters exist and describe the world in the way we think.
  - These assumptions might be wrong, and we have to assess how realistic they are.
- In the case of simple regression, our population model looks like this:

Intercept

Slope parameter

Dependent variable, outcome variable, response variable,...

$$y = \beta_0 + \beta_1 x + u$$

Error term, Disturbance, includes the effect of all other factors aside from x.

Independent variable, explanatory variable, regressor,...

We'll need more assumptions about the error. That's what statisticians spend most time worrying about.

Interpreting OLS

- The most important parameter here is β_1.
- The intercept is important in certain circumstances, but only if x=0 has some special meaning.
- β_1 represents the expected change in y given a unit change in x, and holding the error constant.

$$y = \beta_0 + \beta_1 x + u$$

Interpreting OLS

Two examples from Wooldridge:

- **Soybean yield and fertilizer**

$$yield = \beta_0 + \beta_1 fertilizer + u$$

Rainfall,
land quality,
presence of parasites, …

Measures the effect of fertilizer on
yield, holding all other factors fixed

- **A simple wage equation**

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

Labor force experience,
tenure with current employer,
work ethic, intelligence …

Measures the change in hourly wage
given another year of education,
holding all other factors fixed

Constraining the Error

- So far, we haven't assumed anything about the error term
  - This is a problem because any line will fit the data for some error distribution.
- What do we need to assume about the error term?
- First, we assume the errors have mean 0, $E(\mu) = 0$

- This isn't a strong assumption, because we could always change β_0 to move our line up or down so that the mean error is zero.

Zero-Conditional Mean

- Our next assumption is more serious, and much more often questioned.
- **Zero conditional mean assumption**

$$E(u|x) = 0$$

Even if we look at a specific value of x, we still expect errors to average to zero.
The explanatory variable must not contain information about the mean of **ANY** unobserved factors
Note that we're talking about the actual parameters here, not our estimates!

- **Example: wage equation**

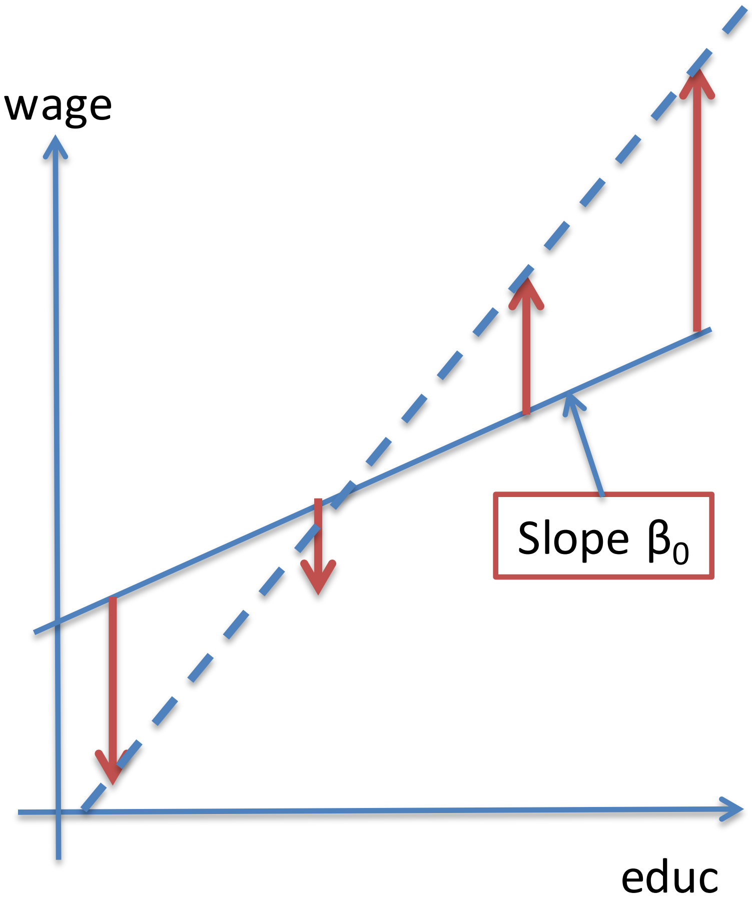
$$wage = \beta_0 + \beta_1 educ + u$$

e.g. intelligence …

- This is the most famous equation in a field called labor economics.
- Here, the error term contains unobserved variables like work experience, ability, etc.
- The conditional mean independence assumption is unlikely to hold because individuals with more education may have more ability, on the average.
  - What does that mean for our interpretation?

A Violation of Zero Conditional Mean

- Here's a graphical depiction of what happens when the zero-conditional mean assumption fails.
- Suppose the solid line, with slope β_0 depicts the causal effect of education. That is, we're assuming that a unit increase in education will increase wages for an individual by β_0.
- The red arrows represent E(u), the average effect of unobserved variables like ability. Here, people with more education also have more ability, which increases their wage.
- The dashed line is the relationship we actually see in data
  - This is the observed relationship between educ and wage in the population
  - But it does not represent what would happen to an individual who gets an extra year of education.
  - The slope we would estimate is not the slope in our population model, and there's no way to recover the real β_0
- we have to assume this doesn't happen in order for OLS to work.



Interpreting OLS

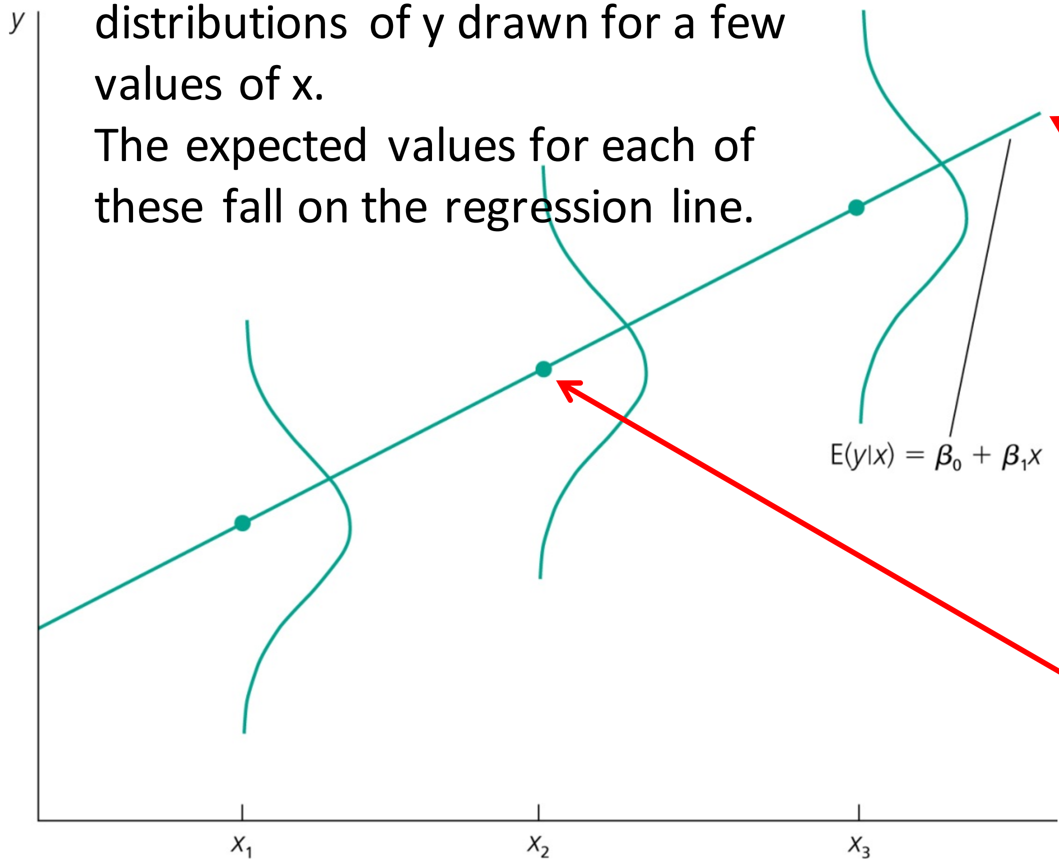
- **Population regression function (PFR)**
  - Given the conditional mean assumption, we can solve for the expected value of the outcome, given a value of x.

$$\begin{aligned} E(y|x) &= E(\beta_0 + \beta_1 x + u|x) \\ &= \beta_0 + \beta_1 x + E(u|x) \\ &= \beta_0 + \beta_1 x \end{aligned}$$

  - This means that the average value of the dependent variable can be expressed as a linear function of the explanatory variable

Conditional Mean Independence

Here, you can see conditional distributions of y drawn for a few values of x.
The expected values for each of these fall on the regression line.



Population regression function

For individuals with $x = x_2$, the average value of y is $\beta_0 + \beta_1 x$

Bivariate OLS Estimation

Introducing Estimation

- So far, we‘re only been talking about the population model, which we assume to be true
- The next step is the estimate the parameters of our model. For that we need data.
  - We‘ll assume we have a random sample of *n* observations

(x_1, y_1) First observation

(x_2, y_2) Second observation

(x_3, y_3) Third observation

$\vdots$

(x_n, y_n) n-th observation

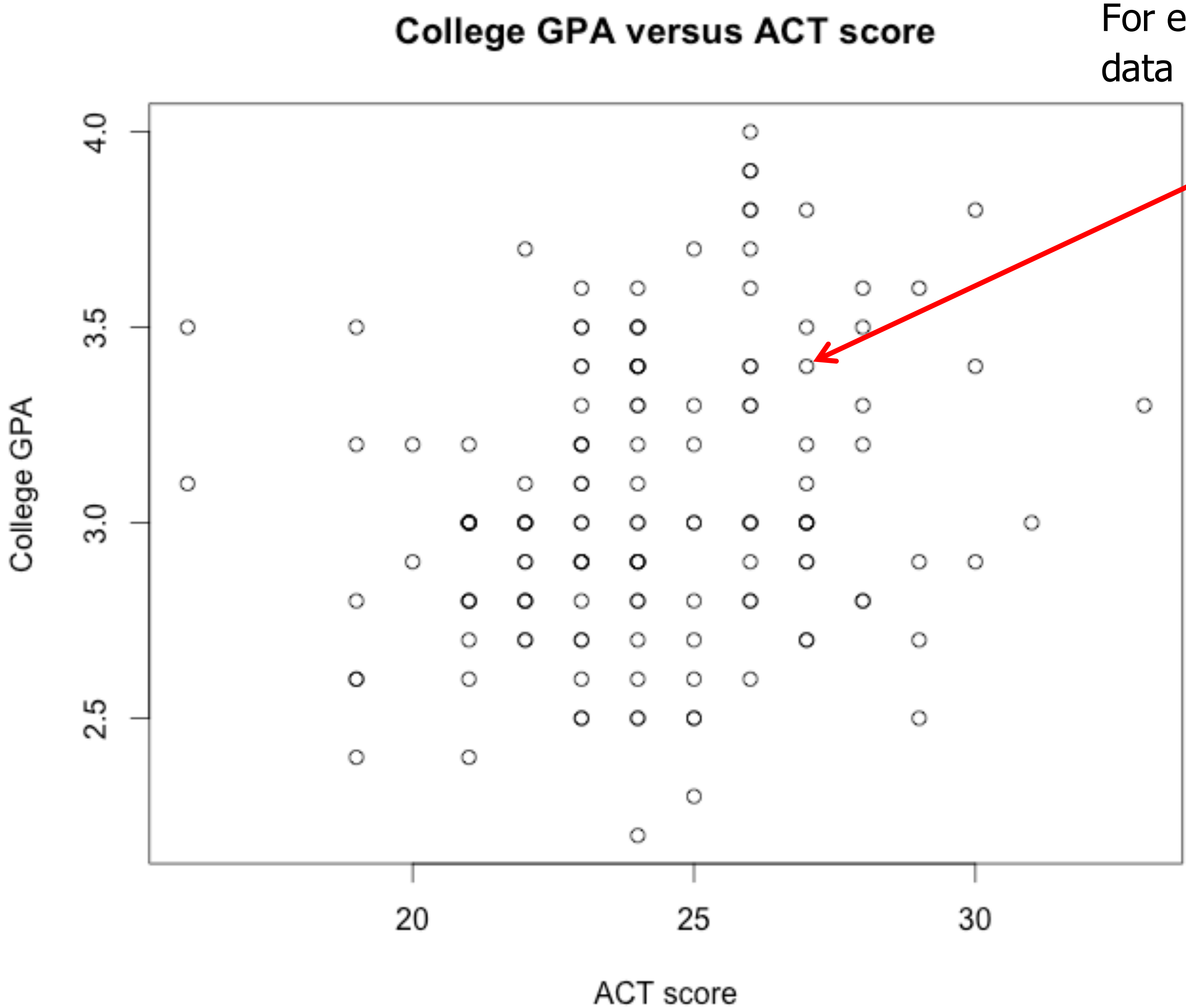
$$\{(x_i, y_i) : \; i = 1, \ldots n\}$$

Value of the <u>explanatory variable</u> of the i-th observation

Value of the <u>dependent</u> variable of the i-th observation

Plotting Bivariate Data

For example, the i-th data point (x_i, y_i)



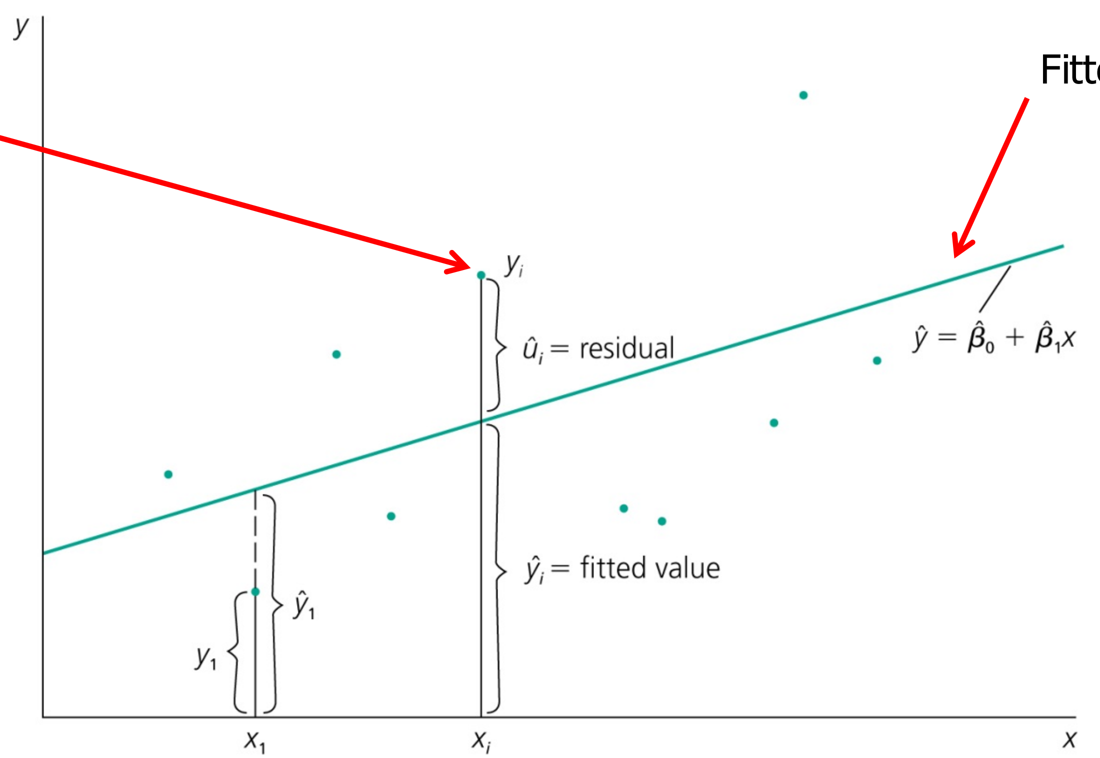
OLS as Best Fit

Artist: I sometimes write a ^ after a variable to save time, but the hat goes over the variable.

- One way of thinking of OLS is as yielding the “best fit” line through our scatter plot

For example, the i-th data point (x_i, y_i)

Fitted regression line



The regression line defines a predicted $y^\wedge_i$ for each x_i. The difference between the observed y_i and $y^\wedge_i$ is the residual, $u^\wedge_i$.

OLS as Error Minimization

- <u>**What does error minimization mean?**</u>
- **The regression residuals are our estimated errors**

$$\widehat{u}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$$

- **We *minimize* the sum of squared regression residuals**

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

- Solving the minimization problem, we arrive at the OLS estimates:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\operatorname{cov}(x_i, y_i)}{\operatorname{var}(x_i)}, \quad \beta_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

The slope is the most important part, and note that it has a very simple form in terms of covariance

Some properties of OLS

Algebraic properties of the OLS estimators

$$\sum_{i=1}^{n} \widehat{u}_i = 0$$

The (estimated) errors sum up to zero

$$\sum_{i=1}^{n} x_i \widehat{u}_i = 0$$

The correlation between residuals and regressors is zero

$$\bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}$$

The sample averages of y and x lie on regression line

Example

- **CEO Salary and return on equity. We assume the following population model**

$$salary = \beta_0 + \beta_1 roe + u$$

Salary in thousands of dollars

Return on equity of the CEO‘s firm (percentage)

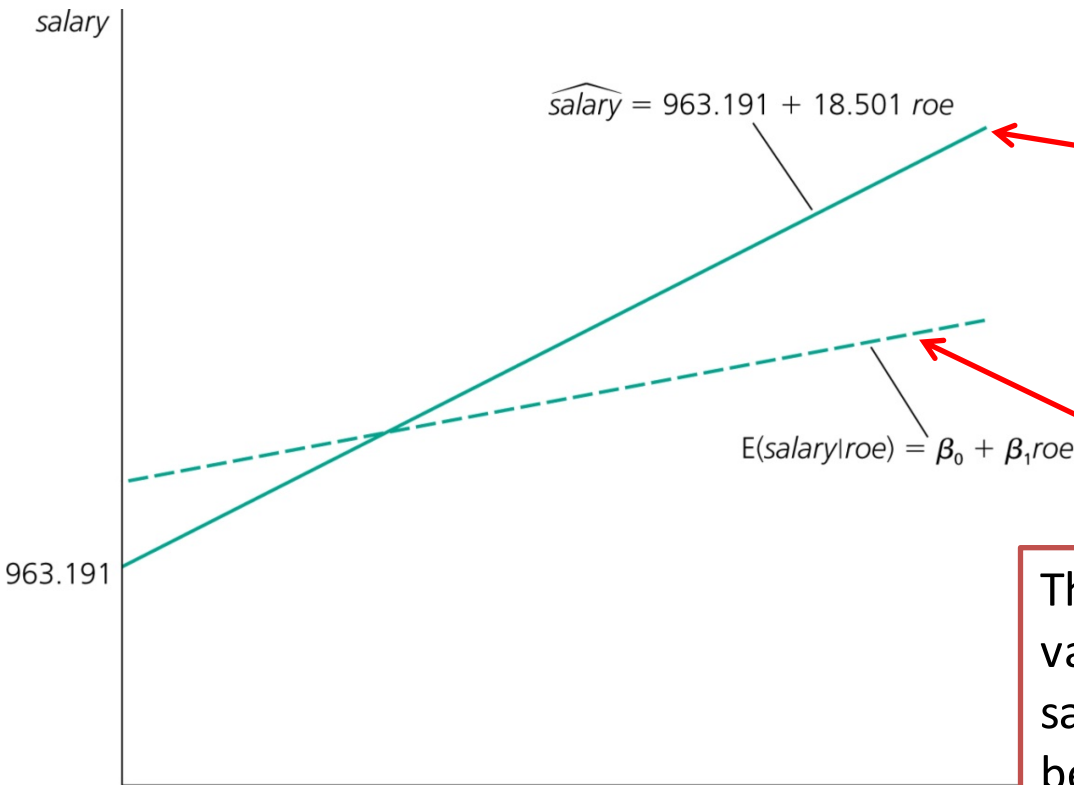
- **Here‘s an example of a fitted regression, from Wooldridge.**

$$\widehat{salary} = 963.191 + 18.501\ roe$$

Intercept

If the return on equity increases by 1 percent, then salary is predicted to change by $18,501

OLS Coefficients as Random Variables



Fitted regression line (depends on sample)

Unknown population regression line

The coefficients we get are random variables. They will differ from sample to sample. So the line we fit varies and will be different from the true population regression line.

Deriving the Bivariate OLS Estimators

10 minute Lightboard

Moments derivation

Optional lightboard (5 min)

Goodness of Fit

How do we know how much of Y our variable X explains?

- **Goodness-of-Fit**

  By this, we mean how well does the explanatory variable explain the dependent variable?

- **Three Important Measures of Variation**

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad SSR = \sum_{i=1}^{n} \hat{u}_i^2$$

Total sum of squares, represents total variation in dependent variable. It‘s an unscaled version of variance

Explained sum of squares, represents variation explained by regression. Notice we put in the predicted values

Residual sum of squares, represents variation not explained by regression

How do we know how much of Y our variable X explains?

- **Decomposition of total variation**

$$SST = SSE + SSR$$

Total variation | Explained part | Unexplained part

- **Goodness-of-fit measure (R-squared)**

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

R-squared measures the fraction of the total variation that is explained by the regression
R-squared requires all OLS assumptions to be interpreted correctly

How do we know how much of Y our variable X explains?

- CEO salary and return on equity

$$\widehat{salary} = 963.191 + 18.501\ roe$$

$$n = 209, \quad R^2 = 0.0132$$

The regression explains only 1.3 % of the total variation in salaries

- R-squared is an important measure – but often used inappropriately
  - When asked to assess the practical significance of a model, students (and researchers) often report R-squared without considering if it's the right metric.
- A high R-squared only tells us that a lot of the variation in our Y variable is explained by our model
  - This is important in certain circumstances: if our primary objective is prediction, this tells us that our predicted values are close to the true values.
  - R-squared can be considered a measure of predictive accuracy
- In most applications we'll talk about, you care about inference, understanding an effect, or testing a theory.
  - In these cases, R-squared is not the measure you want and may mislead you.
  - Example: If we regress hospital admissions on whether a person was recently shot, the R-squared is low since there are a lot of other reasons why people go to the hospital.
    - But doesn't mean the effect of getting shot is unimportant!

Assessing Practical Significance

- You should get into the habit of commenting on the practical significance of your results.
  - Statistical significance is about whether our results are unlikely to occur by chance
  - Practical significance is about whether we should care
    - what is the effect size?
    - What number would you put into a newspaper headline to inform readers about the relevance of your fitted model?
- Guidelines:
  - In linear regression, your slope coefficients are usually much more relevant than your R-squared.
    - In the hospital example, the coefficient for getting shot shows how much more likely a person is to be admitted to the hospital if they were just shot.
  - But context matters.
    - If you're deciding whether to fund a program to reduce firearm violence, the number of shooting victims matters too. You may want to report an estimated decrease in hospital admissions.
  - Consider the units: effect size measures should be understandable.
    - Understandable: every minute waiting for the bill decreases a restaurant rating by .12 stars out of 5.
    - Not understandable: A 1% increase in sugar intake per pound of hay per height of race horse results in 5% less heart rate increase per 100 meters of track.
      - When the unit are hard to understand, you have the option of standardizing the variable first
        - » Subtract mean and divide by standard deviation.
        - » Then instead of obscure units, you can talk about a standard deviation increase in a variable, which likely provides more sense of scale.

Multivariate OLS

Expanding OLS to multiple dimensions

- Bivariate OLS can be very useful
  - You can learn a lot just by comparing two variables
- More often, we have a larger number of variables, and want to understand their relationship, or use the information they contain to make better predictions
- Fortunately, the mechanics of multiple OLS regression are similar to simple regression.
  - Multiple regression is a workhorse of statistical analysis in a wide variety of fields

The Multiple Regression Population Model

- As before, we have to start with a population model.
- This is similar to the population model for simple regression, but we have several x variables and a coefficient for each one.

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,…

Interpreting Coefficients in Multiple Regression

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

Consider the meaning of each coefficient. β_j now represents the expected change in y from a unit change in x_j, **holding all the other x's and u constant**.

- Our interpretation is ceteris paribus.
- This is true, even if the other variables are correlated with x_j.

- Here‘s an example, from a study of test scores.
- We model scores as a function of school spending and average family income.
- schools that spend a lot on each student are also likely to be in areas with high family income – these variables are correlated.
- Omitting average family income in regression would lead to a biased estimate of the effect of spending on average test scores
- If we want to assess a spending plan, we should hold family income fixed since this is unlikely to change, at least in the short term.

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

Other factors

Average standardized test score of school

Per student spending at this school

Average family income of students at this school

Another Multiple Regression Example

- **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**
  - <u>Holding ACT fixed</u>, another point on high school grade point average is associated with another .453 points college grade point average.
  - Or: If we compare two students <u>with the same ACT</u>, 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
  - <u>Holding high school grade point average fixed</u>, another 10 points on ACT are associated with less than 0.1 points on college GPA

Partialling out

- OLS estimates all coefficients simultaneously, but it turns out it can also be done in two steps:

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

1. First, write down the regression of x_1 on all the other x's.

   $x_1 = \delta_0 + \delta_2 x_2 + \ldots + \delta_k x_k + r_1$.

   Let r_1 be the error term in this regression.

   r_1 represents the unique variation in x_1 – the part that's not collinear with other variables

   The other variables have been "partialled out"

2. Now, regress y on just r_1:

   $y = \gamma_0 + \gamma_1 r_1 + v$

- β_1 is the same as the coefficient on r_1 in this new regression.
- $\beta_1 = \mathrm{cov}(r_1,y)/\mathrm{var}(r_1)$
  - This may be called the ***regression anatomy formula***
- So instead of running the full regression, we can just look at the unique variation in each variable, and see how it relates to y.
- Intuitively, it's the unique variation that lets us estimate a coefficient.
  - Any variation that's collinear with other variables doesn't help us because we don't know what variable to ascribe the effect to.

Flexible Functional Forms

- Another major motivation for multiple regression is that it allows more flexible functional forms.
- As an example, here's a population model of consumption as a function of both income and incomes squared.

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

Other factors

Family consumption

Family income

Family income <u>squared</u>

  - Consumption is explained as a quadratic function of income, not linear.
  - We have to be 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

  - Remember that linearity only restricts how our variables interact with each other
    - As long as we combine our terms linearly, we have a lot of flexibility in designing our model

Regression Anatomy

15 minute lightboard (possibly optional for students)

BLUE

OLS Assumptions

- What assumptions do we need for OLS regression to work?
- Not a simple question, it depends on what we mean by “work.”
  - Depending on what guarantees we want, we need to meet different sets of assumptions.
- On one hand, we’ll see that our population model may meet only a weak set of assumptions
  - Assumptions that are realistic for almost any real dataset
  - We can still run an OLS regression, but our ability to draw meaning from the results will be severely limited.
- On the other hand, there is a famous set of fairly strict assumptions called the Gauss-Markov assumptions
  - These are tougher to justify and often unrealistic for real datasets
  - If they hold, we get much stronger guarantees about OLS estimates
  - Specifically, the Gauss-Markov theorem says that under certain assumptions, OLS is BLUE…

BLUE

- BLUE stands for Best Linear Unbiased Estimator.
  - This is often what people mean by OLS "working."
  - We already know what an estimator is
    - OLS coefficients are estimators of the population parameters
  - let's look at the other terms
- Best – Here, we're talking relative efficiency. The OLS coefficients are random variables, and we want them to be as precise as possible
  - OLS coefficients have the smallest variance of all linear unbiased estimators.
- Linear – OLS estimates are a linear function of the y_i's.
  - You can see in the matrix representation that the vector y is multiplied by a matrix, $(X'X)^{-1}X'$, and matrix multiplication is a linear operation.
- Unbiased – each $\hat{\beta}_j$ is an unbiased estimator for the true parameter β_j.

  $$E(\hat{\beta}_j) = \beta_j$$
- BLUE is the most well-known benchmark for OLS performance
- Next, let's look at the actual assumptions that underlie the theorem.
  - We'll start this week by establishing Unbiased, we'll tackle Best next week.

Getting to Unbiased

First Assumptions

- We’ll begin in this segment with a fairly weak set of assumptions about our population model
  - The kind of assumptions that are often quite realistic
  - Safer assumptions to make
- These assumptions are the first 4 Gauss-Markov Assumptions
  - but that’s not enough for the Gauss-Markov Theorem.
- Even so, with just 4 assumptions, we’ll manage to show that OLS estimators are unbiased
  - This is the U in BLUE

Linearity and Random Sampling

- **Assumption MLR.1 (Linear in parameters)**

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

The first assumption is the basic population model – y is linear in the x's. At this point, we don't have to worry about this assumption because we haven't said anything about u, so the assumption isn't really a restriction. Any population distribution could be represented as a linear model plus some error. The error might be very poorly behaved.

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

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

The second assumption states that the data is a random sample drawn from the population

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

All data points follow the population distribution, and they must be independent draws from the distribution. The data points are iid – indendently and identically distributed.

Multicollinearity

- **Assumption MLR.3 (No perfect collinearity)**

  In the sample (and therefore in the population), none of the independent variables are 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
    - In practice high correlation can greatly increase errors
  - 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 (these are collinear with the intercept term)

Multicollinearity Example

- As an example of perfect multicollinearity, imagine a model that predicts the share of the vote earned by Candidate A as a function of how much A spends, how much B spends, and total campaign spending:

  $VoteA = \beta_0 + \beta_1 expendA + \beta_2 expendB + \beta_3 totexpend$

- Here, totexpend is a linear combination of the other variables, so it has no unique variation for OLS to work with.
  - Whatever coefficients we choose, we could subtract 1 from β_1 and β_2 and add one to β_3 and the model stays exactly the same. There's no unique set of coefficients for us to estimate.
  - Conceptually, multicollinearity is equivalent to asking "So, did you buy 12 eggs or a dozen?" and demanding one answer or the other
- To solve this problem, one variable has to be dropped from the model.

Zero-Conditional Mean

- **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

- This is the strongest assumption so far.
- You can think of it as ensuring linearity. MLR.1 establishes a linear population model, but MLR4 ensures that the population actually follows that linear model.

Unbiased Coefficients

- **<u>Theorem3.1 (Unbiasedness of OLS)</u>**
  - Under MLR.1-4, OLS estimates are unbiased.

$$E(\hat{\beta}_j) = \beta_j$$

  - Remember that unbiasedness is an average property in repeated samples; in a given sample, the estimates may still be far away from the true values
  - But at least we know that in expectation, we‘re measuring the right thing.

Troubleshooting the Bias Assumptions

Linearity

- We've listed the assumptions needed for OLS to be unbiased. Now let's see how to test them, and what to do if our data seems to violate them.
- Linear model
- Our linear model assumption just expresses y as a linear function of our x's.

  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
- At this point, there's nothing to test, because we haven't constrained our error in any way.
  - This formula is always true for some definition of u.
  - Given any set of coefficients, we can just define $u = y - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \ldots - \beta_k x_k$.
- So our work really starts with the next assumptions.

Random Sampling

- Random sampling:
- This assumption says that all data points are independent random draws from our population distribution.
- In part, we must use our knowledge of where the data came from to assess this assumption.
  - What was the procedure for collecting data points?
    - For a study of people, how were subjects found?
- There are two common ways that this assumption can fail.
  - The first is clustering.
    - Clustering occurs when individuals are collected into groups, and researchers can only access a limited number of these groups, known as clusters.
      - As an example, a study might randomly select n schools from a school district and then m_i students from school i.
      - The problem is that students from a particular school are likely to be similar to each other, so we're observing less variation than actually exists in the population.
    - Even with clustering, OLS coefficients are unbiased.
    - However, our estimates become much less precise under clustering.
      - In response, we'll need to use clustered standard errors, or other techniques to account for this.
      - We'll discuss the precision of our estimates next week.

Random Sampling

- Another way that the random sampling assumption may fail is with autocorrelation or serial correlation.
- This occurs when the error for one datapoint is correlated with the error for the next datapoint.
- This is common for time series data.
  - A variable that's unusually high at time t will tend to be high at time t + 1.
- There are tests for autocorrelation. The most common is the Durbin-Watson test.
  - The Durbin-Watson statistic compares the differences between successive data points to the magnitude of the data points.

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2},$$

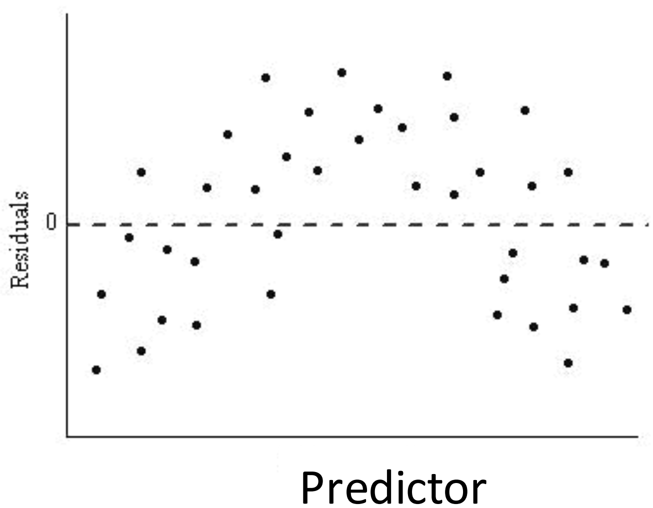
  - The distribution of this statistic is complex, but R handles the details for you.
  - In R, durbinWatsonTest() computes the statistic under the null hypothesis that there is no serial correlation. If it's significant, you have evidence that there is correlation.
- There's no simple fix for serial correlation.
- The latter half of the course is devoted to specialized methods for studying time series data.

Multicollinearity

- Remember that our multicollinearity assumption only rules out perfect multicollinearity.
  - Now that you've seen the regression anatomy formula, this should be more intuitive.
    - OLS operates on the unique variation in each variable. Under multicollinearity, there is no unique variation, so the formula is 0 / 0 – undefined.
- The response to multicollinearity is simple: drop redundant variables
- When variables are highly correlated, but not perfectly collinear, OLS will still work, but as we'll discuss next week, estimates will be much less precise.
  - This means that we often have to make tough choices.
  - Do we put in a variable, and suffer a lot of precision, or leave it out, even though we think it has an important effect on the outcome?

Zero-Conditional Mean

- Zero-conditional mean
  - This is the strongest assumption we've seen.
  - It says that for any possible value of our predictors, our error is zero in expectation.
  - To examine this assumption when there's just one predictor, we could create a residuals versus predictor plot.
    - We have our x on the x axis, and our residuals on the y.
    - Remember that our residuals are our estimates of the error, so we can see how they change for different values of x.
  - On this plot, we can eyeball where the mean of the residual changes from left to right.



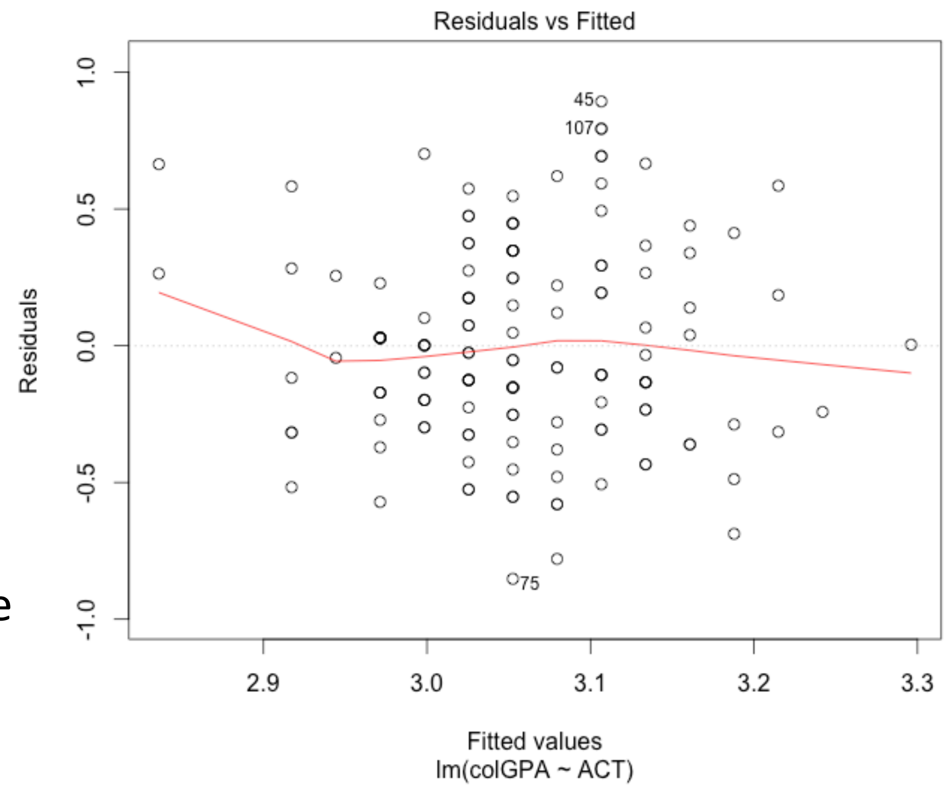
- In this example, you can see that the mean of the residuals seems to go up and then down.
- For zero-conditional mean, we'd want this to be a flat band.

Residuals vs. Fitted Values

- For multiple regression, we can't plot all possible x values in two dimensions.
- We could create a separate residual versus predictor plot for every x – but that would be a lot of graphs and still wouldn't tell us definitively if we have zero-conditional mean.
- More commonly we would create a residuals vs. fitted values plot.
  - Here, the y-axis has residuals, as before.
  - The x-axis has our predicted values of y.
    - These are a linear function of x, so if there's a non-zero mean for some values of some x, it's likely to show up in this plot
    - Notice that if there's just one x, the fitted value of y is just a linear scaling of x, so the plot is the same as the residual vs. predictor plot.
  - So once again, we're looking to see if the plot looks like a nice flat band from left to right.
- Most software including R will easily create a residual versus fitted value plot.

Residuals vs. Fitted Values

- Here's an example of a residuals vs. fitted values plot in R.
- As you can see, this one looks better than the example we had before. There's more of a flat band from left to right.
- R helps us tell if the conditional mean is zero, by including a spline curve in red.
- Ideally, this curve is completely flat.
- Here, there's a tiny bit of curvature, but it's minor.
    - In fact, it might just be that there are few datapoints on the left of the graph, so the mean could be high randomly.



Responding to Violations of Zero-Conditional Mean

- If the conditional mean of the error is not constant, we may be able to change functional form.
  - Sometimes, if you see curvature in the residual vs. fitted value plot, there may be a linear relationship between x and the log of y.
    - Or perhaps the log or x and y, etc..
  - We may also allow a more flexible functional form by regressing y on x and x^2. This fits a parabola to the data and often corrects violations of zero-conditional mean
  - These methods have trade-offs, and we'll discuss them in detail later in the course.
- Sometimes, adding new variables can fix the zero-conditional mean assumption.
  - There may be a better variable out there that has a linear relationship with our outcome
- If all of that fails, we may decide that we can't meet zero-conditional mean.
- In that case, we may be able to meet a weaker assumption: exogeneity.

Exogeneity

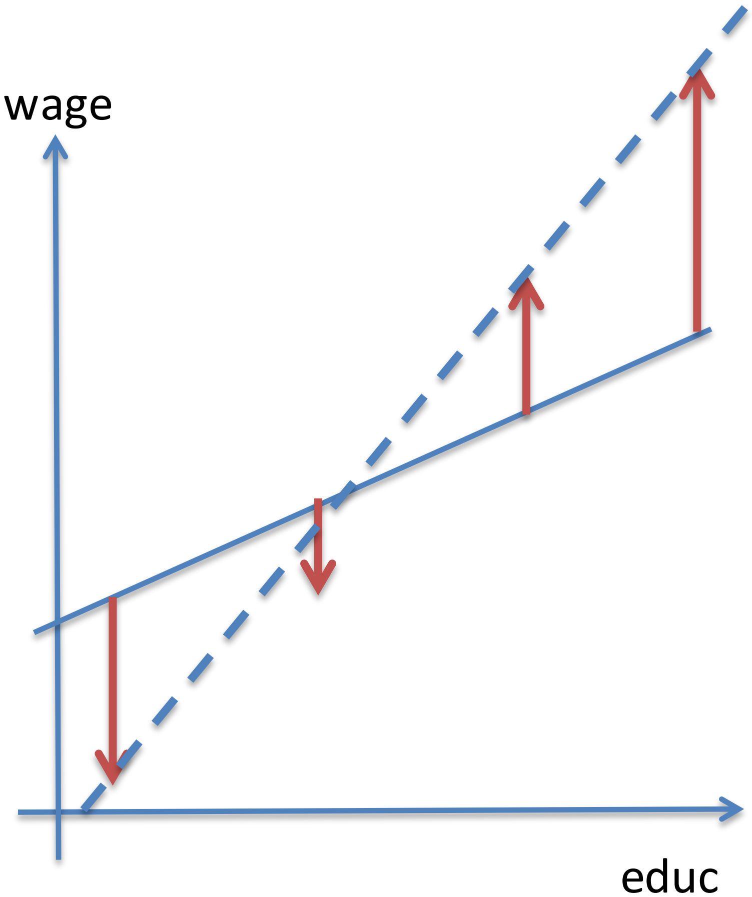
- Explanatory variables that are correlated with the error term are called **endogenous**
    - “originates within the system”
    - However, endogeneity is not a direct statement about causality – it’s about correlation, and that correlation could be present for all sorts of reasons
  - Endogeneity is a violation of zero-conditional mean, and the presence of endogenity implies that OLS coefficients are biased and inconsistent.
- Explanatory variables that are uncorrelated with the error term are called **exogenous.** If x_j is exogenous, $Cov(x_j,u)=0$
- **Assumption MLR.4‘ (Exogeneity)**
  - $Cov(x_j,u)=0$ for all j.
- **Theorem**: Under MLR.1-3 and MLR.4’, the OLS estimators are consistent.

$$\text{plim}_{n\to\infty}(\hat{\beta}_j) = \beta_j$$

- Our estimators are no longer unbiased, but consistent means the bias goes to zero for large sample size.
- As long as we have a data set of a few hundred or thousand observations, researchers generally focus on achieving consistency.
- If you have such a large dataset, exogeneity is the critical assumption.

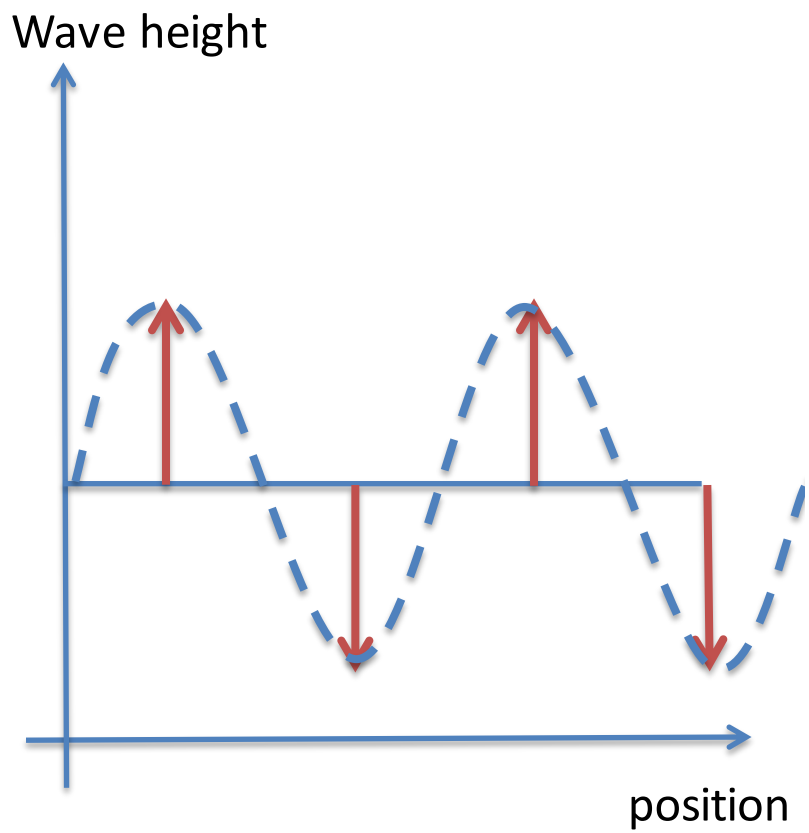
The Problem with Endogeneity

- Let’s see why exogeneity is important.
- Remember this example from earlier, graphing wage as a function of education
- Here, educ is an endogenous variable, since it’s correlated with the errors, represented by the red arrows.
  - These include factors like ability. We’re assuming that people with more education also have more ability, on the average.
- We can see that Cov(educ, u) > 0.
- That means that OLS can’t find the real slope of the population model, which is represented by the solid line.
- Instead, it would pick up a totally different slope, represented by the dashed line.
  - This is called Endogeneity Bias.



The Problem with Endogeneity

- Exogeneity is a weaker assumption than zero-conditional mean.
- Here's another population relationship, wave height as a function of position (for a fixed time)
- The population model is the flat line.
- The arrows represent the expected error E(u|x),which goes up and down sinusoidally
  - So the relationship isn't really a linear one, E(u|x) ≠ 0.
- However, we still have Cov(x,u) = 0. the variable is endogenous
- That means that OLS will correctly identify the slope of the population as zero.
- This shows that exogeneity is weaker assumption than zero conditional mean
  - It's easier to meet.
  - Often more realistic.



Causality

Schools of Thought

$$y = \beta_0 + \beta_1 x + u$$

- When can we interpret β_1 as the causal effect of x on y?
- There are many competing theories of causality.
  - Researchers have deep philosophical debates about this.
- According to one popular school of thought, we have to consider a ***counterfactual***: what if x were some other value?
  - Would *y* change in the way our population model predicts?
  - Note that we're talking about changing *x* for one individual or unit of analysis.
    - If you were to take an individual and give them an extra year of school, how would their wage change?
  - We want to leave other factors equal – this is known as the ceteris paribus assumption.
    - It may be problematic to imagine leaving other factors equal in some cases.
    - Can you convince somebody to stay in school for an extra year without changing something else about who they are?
      - But in your own life, you might imagine making different choices, and wonder what would have happened.
- This counterfactual idea is related to the idea of ***manipulation***.
  - We often investigate data because we want to make a decision or change something.
    - What if we instituted a policy on health insurance?
    - What if we increased vacation time to 4 weeks for new employees?
  - Intuitively, we can imagine making different choices and try to imagine what the results are.

Causality and the Error Term

$$y = \beta_0 + \beta_1 x + u$$

$$\frac{\partial y}{\partial x} = \beta_1 \quad \text{as long as} \quad \frac{\partial u}{\partial x} = 0$$

- Mathematically, the idea of a manipulation can be represented by a change in x.
  - Our equation tells us that β_1 is the rate of change of y with respect to x, but only if the the rate of change of u with respect to x is zero
  - This is the ceteris paribus assumption – everything else remaining equal.
- So our population model is causal if manipulations to x do not affect u.
  - This is an extra assumption on top of our population model
  - And you can't prove it with math, this is something that you usually have to argue intuitively or philosophically.

Causality vs. Exogeneity

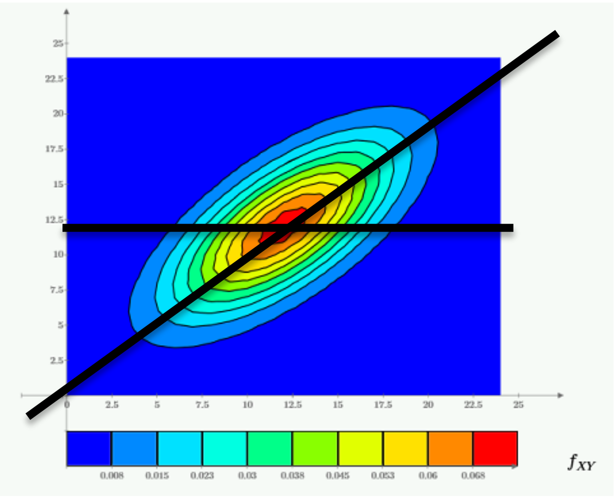
- Here's one clarification that you might find useful.
- In a population model, being causal is not the same as exogeneity.
  - Causality is about whether manipulations to x do not influence the error term.
  - Exogeneity is about whether OLS can correctly estimate (identity) β_0.

Example

- Suppose Y represents rainfall and X is umbrella sales.
  - Our joint distribution might look like the heatmap to the right
- We could represent this naively as a population model,

$$y = \beta_0 + \beta_1 x + u$$

With $\beta_1 > 0$ and zero conditional mean, $E(u|x)=0$
  - This model is not causal. If I buy another umbrella, I move to the right of the plot, but not up. So the error goes down.
  - On the other hand, zero-conditional mean implies that x is exogenous, and OLS will correctly estimate β_1
- We could also represent this as a causal population model, in which $\beta_1 = 0$. But then the error is no longer exogenous.
  - This means that OLS cannot estimate β_1.



In either case, the coefficients you compute from OLS do not have a causal interpretation.

- We’ll come back to the topic of causality later in the course when we discuss identification strategies.
- For the next couple of weeks, we’ll go further into the nuts of bolts of OLS regression.
- By the end, you should have a firm grasp of all the assumptions required and know best practices of what to do when assumptions are not met.