Questions of Description

Questions of Description

Questions of Description

What is the shape of the relationship between a country's economic output and Internet access?

Questions of Description
What is the shape of the relationship between a country's economic output and Internet access?

1. Here are some questions that you might want to answer with a statistical analysis...

Questions of Description

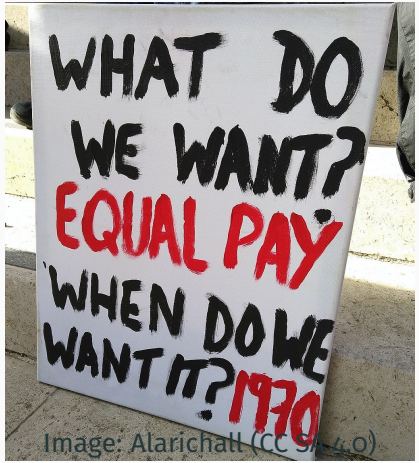


Image: Alarichall (CC BY 4.0)

How big is the pay gap in the United States?

Questions of Description

How big is the pay gap in the United States?

1. That is, what is the difference in pay between workers of different genders?

Questions of Description

How does the pay gap depend on the age of the worker?

Questions of Description
How does the pay gap depend on the age of the worker?

Description

1. One thing these questions have in common: They are questions of description.
2. They are also questions that can be addressed using linear regression.
3. You've learned the mechanics of how linear regression works. You know the nuts and bolts of interpreting coefficients and statistical guarantees.
4. But how do you build a model, when the purpose is description. that's a more strategic level of thinking
5. There's an entire set of concerns that goes into building models for description, and that's going to be out topic in this unit.

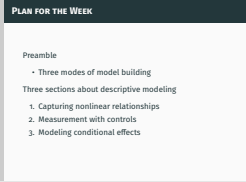
Plan for the Week

Preamble

- Three modes of model building

Three sections about descriptive modeling

1. Capturing nonlinear relationships
2. Measurement with controls
3. Modeling conditional effects



Plan for the Week (cont.)

At the end of this week, you will be able to:

- Understand how three major modes of model building lead to very different models
- Balance design goals when creating a model for description
- Plan a set of model specifications for a regression table

Plan for the Week (cont.)

At the end of this week, you will be able to:

- Understand how three major modes of model building lead to very different models
- Balance design goals when creating a model for description
- Plan a set of model specifications for a regression table

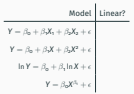
What Is Linear Model Building?

What Is Linear Model Building?

What Is Linear?

Model	Linear?
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$	
$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$	
$\ln Y = \beta_0 + \beta_1 \ln X + \epsilon$	
$Y = \beta_0 X^{\beta_1} + \epsilon$	

What Is Linear?

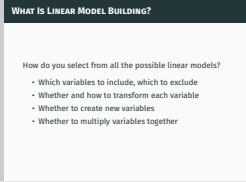


1. Before we talk about model building, let's review a basic definition. What makes a linear model linear?
2. Here's our classic linear model equation, you know this is linear, but why?
3. You might answer, Y is linear in the X's. But that's not quite right. What's important here is Y is linear in the terms.
4. Here's a model, where we've squared X. If you think about it, this is the equation for a parabola, not a line. But actually, you can run an ols regression on this, no problem, because X and X^2 are just like two variables.
5. Another example, where we've taken logs of Y and X.

What Is Linear Model Building?

How do you select from all the possible linear models?

- Which variables to include, which to exclude
- Whether and how to transform each variable
- Whether to create new variables
- Whether to multiply variables together



1. Once you recognize that there are actually a lot of different models that count as linear, you might wonder: how do you choose one?
2. Here are some more specific questions that you face as a data scientist

The Model-Building Process

Propose Model

Evaluate Model

- Modeling goals
- Observations from data
- Background knowledge
- Background theory

- Diagnostic plots
- Measures of fit
- Tests of model assumptions

The Model-Building Process
Propose Model
Evaluate Model
- Modeling goals
- Observations from data
- Background knowledge
- Background theory
- Diagnostic plots
- Measures of fit
- Tests of model assumptions

1. Here's a very very rough schematic, just to emphasize that model building is usually an iteratie prcess. You propose a model, then evaluate it, then make adjustments and propose a new model
2. When you are proposing a model, here are some of the factors that will guide your decision
3. And when you are evaluating
4. In the following segments, we're going to dig into all of these components, and give you an idea of wha this process is like

Modes of Model Building

Modes of Model Building

An Important Question to Keep in Mind:

What are your goals?

An Important Question to Keep in Mind:
What are your goals?

1. If I can give you one piece of advice as you start model building, it's to keep this question in your mind: what are your goals?
2. scribble in "modeling"
3. Depending on what your modeling goals are, even if you have the same data, you may end up with very very different models.
4. That's true of linear regression specifically. A statistician, a political scientist, and a machine learning researcher can all use linear regression on the same data, but the models will look very differerent. Those differences can be traced to the different goals that each of those people

Three Modes of Model Building

1. Predictive modeling: What value can we expect for data we haven't seen?
2. Descriptive modeling: How can we make sense of the patterns in data?
3. Explanatory modeling: How can we measure effects inside a causal theory?

Three Modes of Model Building

1. Predictive modeling: What value can we expect for data we haven't seen?
2. Descriptive modeling: How can we make sense of the patterns in data?
3. Explanatory modeling: How can we measure effects inside a causal theory?

1. So there is a large variety in the way that people in different fields build models.
2. At the same time, we can identify three major modes of model building that account for most of the variety we see: predictive, descriptive, and explanatory. It's important that you understand the differences between them, and the goals that are common in each one.

Modes of Model Building

Predictive Modeling

Modes of Model Building
Predictive Modeling

Predictive Modeling

Prediction: making guesses for unknown values

- For new people, new pictures of cats, or other units
- For future time periods

A key focus of machine learning and time series analysis

Predictive Modeling

Prediction: making guesses for unknown values
- For new people, new pictures of cats, or other units
- For future time periods

A key focus of machine learning and time series analysis

1. Let's start with predictive modeling. When we say prediction, we mean computing a guess for data we haven't seen.
2. It could be new people. Ex: you get data about a patient's symptoms and you want to predict if they have an immune disorder.
3. It could also be future values. Ex. given the last 3 years of data, how many waffle irons can we sell next december?
4. When we think about predictive modeling, the field that immediately jumps to mind is machine learning, especially supervised machine learning. But a lot of time series analysis is also very focused on prediction.

Key Goals of Predictive Modeling

1. Accurately predict values
2. Be interpretable by humans (usually less important)

Implications for linear regression

- Hundreds of variables, or more
- Variable selection by algorithm
  - False discovery: Out of many variables, some will look important by chance.
- Coefficients usually don't have meaning

Key Goals of Predictive Modeling

1. Accurately predict values
2. Be interpretable by humans (usually less important)

Implications for linear regression

- Hundreds of variables, or more
- Variable selection by algorithm
  - False discovery: Out of many variables, some will look important by chance.
- Coefficients usually don't have meaning

1. What are the goals of predictive modeling? Number 1: accuracy - make predictions that are as close to ground truth as possible.
2. In a way, predictive modeling is the most well-defined mode of model building, because it usually has just this one goal. The internals of the model don't matter as long as it predicts well. So you have a lot of flexibility when it comes to choosing the model, you can try a neural network, then switch to a random forest. still, linear regression is still important here.
3. It is true that sometimes there is a second goal in predictive modeling, being interpretable.

Modes of Model Building

Descriptive Modeling

Modes of Model Building
Descriptive Modeling

Descriptive Modeling

Description: summarizing or representing data in a compact, human-understandable way

- Gain understanding by interpreting the model's internal structure
- Popular in statistics, but "not commonly used for theory building and testing in other disciplines"
  —Shmueli, Galit. *To Explain or to Predict?*

Descriptive Modeling

Description: summarizing or representing data in a compact, human-understandable way

- Gain understanding by interpreting the model's internal structure
- Popular in statistics, but "not commonly used for theory building and testing in other disciplines"
  —Shmueli, Galit. *To Explain or to Predict?*

1. Descriptive modeling is really popular among statisticians, but not very popular outside of statistics. Why? Associations are not Causations

Key Goals of Descriptive Modeling

1. Measure complex concepts
2. Capture features of the world
3. Simplify phenomena to make them understandable
4. Highlight associations
5. Generate hypotheses

Implications for linear regression

- Parsimonious models
- Use of logarithms and other transforms with clear interpretation
- Long build process relying on EDA

Key Goals of Descriptive Modeling
1. Measure complex concepts
2. Capture features of the world
3. Simplify phenomena to make them understandable
4. Highlight associations
5. Generate hypotheses
Implications for linear regression
- Parsimonious models
- Use of logarithms and other transforms with clear interpretation
- Long build process relying on EDA

1. We want to measure complex concepts, or especially relationships
2. We want to capture features of the world. This overlaps with making accurate predictions, but it's not really the same goal. we want to capture features that help us make sense of the world.

Modes of Model Building

Explanatory Modeling

Modes of Model Building
Explanatory Modeling

Explanatory Modeling

Explanation: using data to test or estimate parameters in a causal theory

- Explanation lets us reason about actions
- Common mode in economics, political science, epidemiology, psychology, environmental science...
- Causal assumptions are required to make causal claims

Explanatory Modeling

Explanation: using data to test or estimate parameters in a causal theory

- Explanation lets us reason about actions
- Common mode in economics, political science, epidemiology, psychology, environmental science...
- Causal assumptions are required to make causal claims

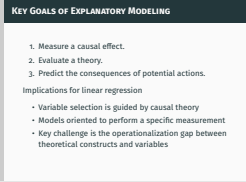
1. Why is this important? Because we may want to learn something from a model and then DO something. Maybe we want to set a policy to improve public health. Maybe we want to change our prices to increase profits.
2. Business problems are also usually causal. You want to know what will happen if the business changes its prices, or adopts a new policy. You don't care if lower prices are associated with more profit, you want to know what will happen if you actually change the prices.
3. I want to point out something in our definition - the causal theory is there first, then we combine it with data. It may not be a big Theory with a capital T. However: You always need to start with some causal background

Key Goals of Explanatory Modeling

1. Measure a causal effect.
2. Evaluate a theory.
3. Predict the consequences of potential actions.

Implications for linear regression

- Variable selection is guided by causal theory
- Models oriented to perform a specific measurement
- Key challenge is the operationalization gap between theoretical constructs and variables



1. very often, our primary goal is measuring a causal effect. that's asking, if we manipulate one variable X, what will be the resulting effect on Y. We want to know the strength of that relationship.

Three Modes of Model Building

1. Predictive modeling: What value can we expect for data we haven't seen?
2. Descriptive modeling: How can we make sense of the patterns in data?
3. Explanatory modeling: How can we measure effects inside a causal theory?

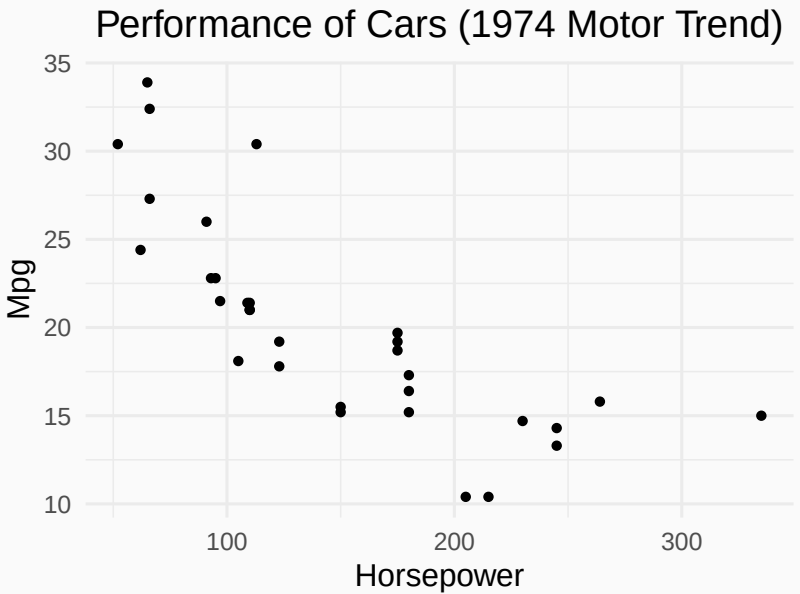
Three Modes of Model Building

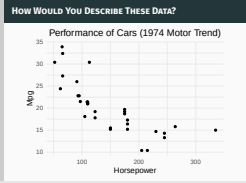
1. Predictive modeling: What value can we expect for data we haven't seen?
2. Descriptive modeling: How can we make sense of the patterns in data?
3. Explanatory modeling: How can we measure effects inside a causal theory?

1. So there you have the three major modes of model building. We've seen that each of them has very different goals, which leads to quite different linear models. Let me also point out that the divisions aren't always that clear. In descriptive modeling, a key goal is capturing features of the joint distribution. But of course, that can also help us to make sense of an explanatory model. In fact, trying to capture features of the joint distribution can sometimes lead you to more accurate machine learning models for prediction.
2. But as you keep learning, it can really help to pay attention to which of the three main modeling modes you're working in.

Introduction to Descriptive Modeling

How Would You Describe These Data?





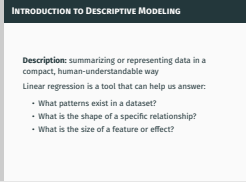
1. If you had to describe this data over the phone, how would you do it? Take 10 seconds, and try to say a few words about it right now.
2. What words are you using? Slope? Decreasing? Maybe hyperbolic or exponential?
3. If you used any of those words, really, what you're doing right now is descriptive model building
4. You may be doing it very informally, or even subconsciously, but you're probably imagining a smooth curve passing through these points and describing that curve.

Introduction to Descriptive Modeling

Description: summarizing or representing data in a compact, human-understandable way

Linear regression is a tool that can help us answer:

- What patterns exist in a dataset?
- What is the shape of a specific relationship?
- What is the size of a feature or effect?



1. Let's review what Description means.
2. There are many tools we can use for describing data. There's clustering algorithms, principle component analysis, and linear regression is also a description tool
3. the sizes of different features are encoded in the coefficients!
4. Unlike a predictive model, in a descriptive model, we really care about our betas, they are measurements of the world.

What Does It Take to Build a Descriptive Model?

Skill, art, and a lot of iteration

- No causal theory to constrain model choice
- Often many ways to operationalize concepts
- Need to balance many competing goals
- Takes time to build intuition in many dimensions

What Does It Take to Build a Descriptive Model?
Skill, art, and a lot of iteration
- No causal theory to constrain model choice
- Often many ways to operationalize concepts
- Need to balance many competing goals
- Takes time to build intuition in many dimensions

1. All together this can be a very creative form of model building.
2. It's our main topic for today - I hope you're excited to dive in.

Logarithms

Logarithms

A Question of Development

How does the economic productivity of a country relate to Internet access?

Data taken from the World Bank

- GDP per capita
- Internet users per 100

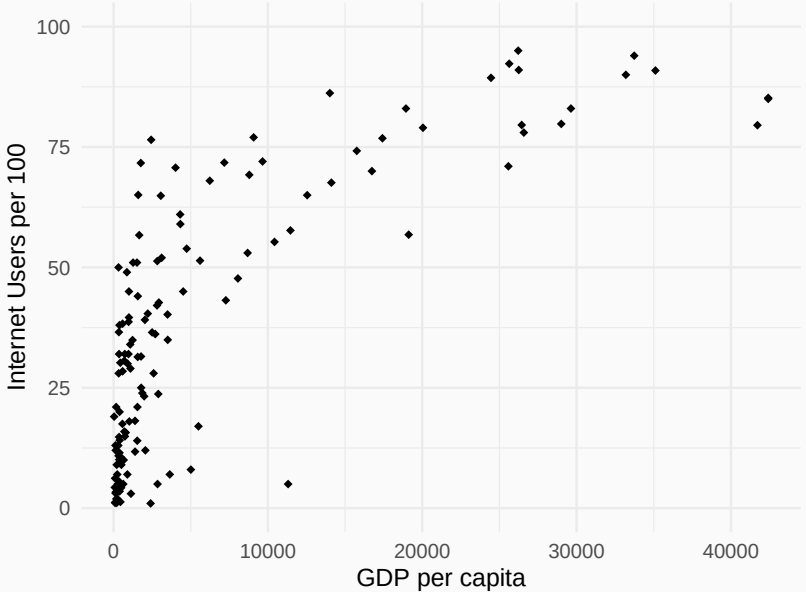
A Question of Development

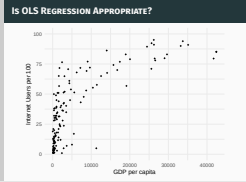
How does the economic productivity of a country relate to Internet access?

Data taken from the World Bank

- GDP per capita
- Internet users per 100

Is OLS Regression Appropriate?





1. Here's a graph of our data.
2. You can see it has this very interesting pattern - a very nonlinear pattern. Try taking your finger, and drawing a line to represent this data. Take a moment to think about this question: is it ok to run a linear regression? here?
3. Alex: turn this into a learnosity check?
4. The short answer is that you can absolutely run an ols regression. We know that OLS coefficients are consistent for the best linear predictor, as long as the best linear predictor exists and is unique. (there is another assumption to worry about, which is iid. Can you really believe that countries are independent from each other? I'm skeptical, but there are a lot of country-by-country

Interpreting the Level-Level Model

$$\widehat{users} = 2.4 + .0021\ GDP$$

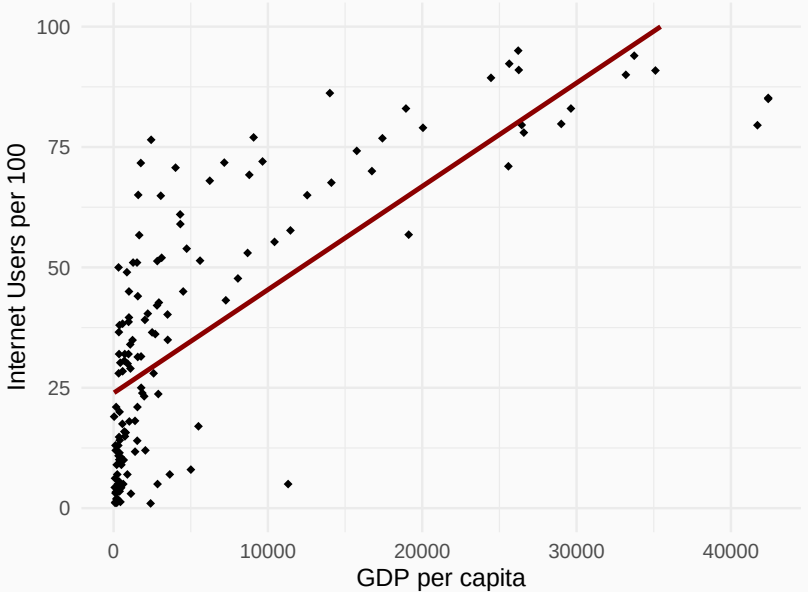
Interpretation 1: A country with an extra $1 of GDP per capita is predicted to have another .0021 Internet users per 100.

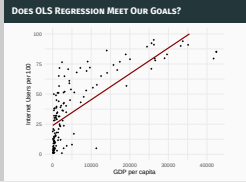
Interpretation 2: A country with an extra $1,000 of GDP per capita is predicted to have another 2.1 Internet users per 100.

Interpreting the Level-Level Model
$\widehat{users} = 2.4 + .0021\ GDP$
Interpretation 1: A country with an extra $1 of GDP per capita is predicted to have another .0021 Internet users per 100.
Interpretation 2: A country with an extra $1,000 of GDP per capita is predicted to have another 2.1 Internet users per 100.

1. Here's are the results of fitting the ols regression
2. Here's how we would interpret the coefficient...
3. Both of these numbers are really small. So we can improve this interpretation by choosing a bigger change...

Does OLS Regression Meet Our Goals?





1. Here's what the regression line looks like. We know this is valid as far as ols assumptions.
2. But does it meet our goals? In particular, one of the main goals of descriptive modeling is capturing features of the world.
3. There's a clear structure that is not reflected in our model.
4. If you look in the middle section, you can see that the prediction is consistently too low. At the extremes, the prediction is too high.
5. There's one other small problem that's worth mentioning. it's only a minor problem, but notice that most of the

Variable Transformation

Variable transformation

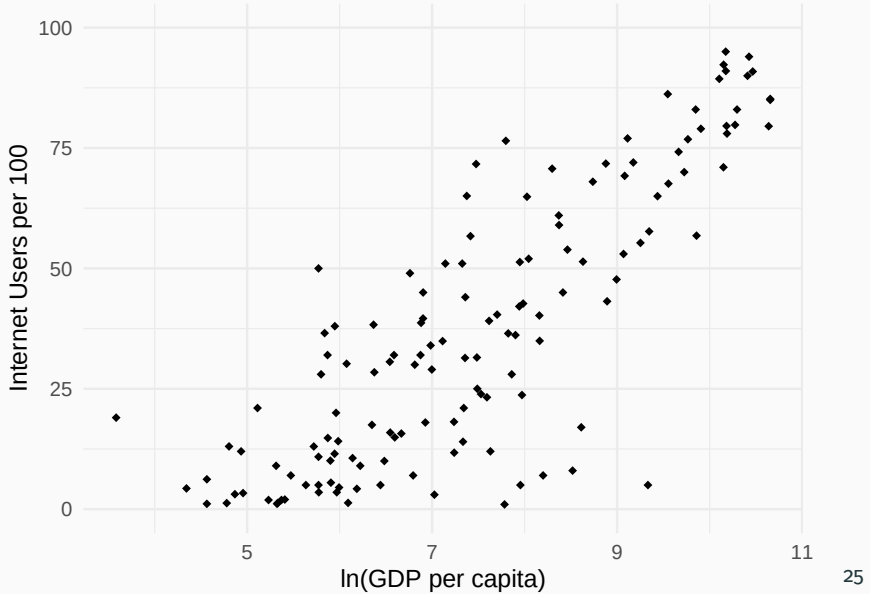
Replace X with $f(X)$ for some $f : \mathbb{R} \to \mathbb{R}$

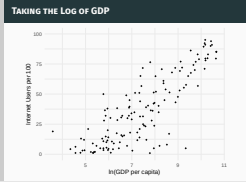
- $\ln(X)$
- $\log_{10}(X)$
- Indicator functions
- X^a
- Polynomials(X)

Variable Transformation
Variable transformation
Replace X with $f(X)$ for some $f : \mathbb{R} \to \mathbb{R}$
- $\ln(X)$
- $\log_{10}(X)$
- Indicator functions
- X^a
- Polynomials(X)

1. Transforming a variable, means that we're replacing some X with a function f(X)
2. What function f should you use? Here are some of the most common choices.
3. Out of all of these, the logarithm is the probably the best choice for our data. Let's see what it does.

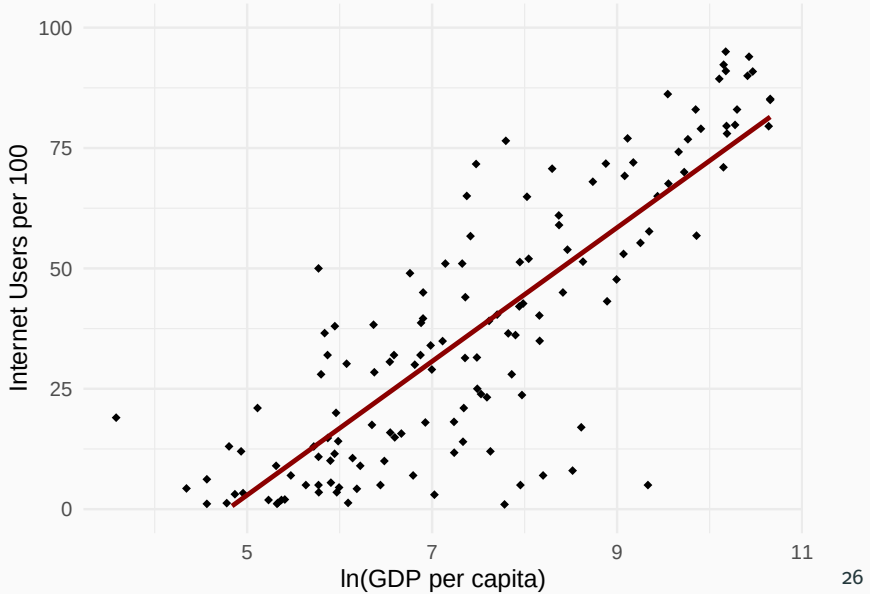
Taking the Log of GDP

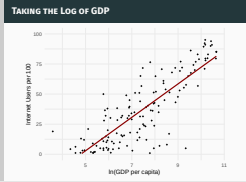




1. Here's what the data looks like after you take a log of GDP per capita. Can you draw a line through it? Much easier!

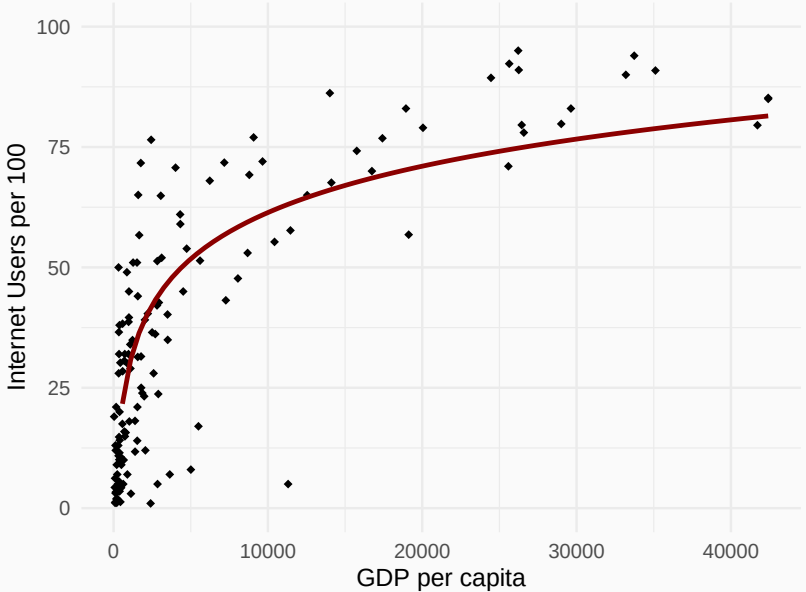
Taking the Log of GDP

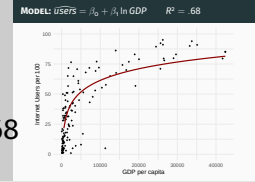




1. And here is the new regression line, which matches the shape of this relationship.

Model: $\widehat{users} = \beta_0 + \beta_1 \ln GDP$ $R^2 = .68$





1. There's another way to view this model. Let's go back to the original x-axis, without the log
2. If we do that, this is what our model looks like
3. This is a logarithmic curve. That makes sense. Our model says that predicted internet users is a linear function of the log of GDP.
4. Think about it this way, we took the log of X and revealed a straight line in the scatterplot. That tells us that we have a logarithmic relationship.

Interpreting and Applying Logarithms

Interpreting Logarithms

Base 10 log

$$\widehat{users} = -66.5 + 32.0 \log_{10} GDP$$

Interpretation: adding a 0 to GDP associated with 32 more Internet users per 100

Base e log

$$\widehat{users} = -66.5 + 13.9 \ln GDP$$

Interpretation: multiplying GDP by a small $1 + \alpha$ associated with 13.9α more Internet users per 100

Interpreting Logarithms
Base 10 log
$\widehat{users} = -66.5 + 32.0 \log_{10} GDP$
Interpretation: adding a 0 to GDP associated with 32 more Internet users per 100
Base e log
$\widehat{users} = -66.5 + 13.9 \ln GDP$
Interpretation: multiplying GDP by a small $1 + \alpha$ associated with 13.9α more Internet users per 100

1. There are actually two common choices for the logarithm, you can take a base 10 log, or a natural log
2. A base 10 log is helpful when we want to think about a variable in powers of 10. We're all used to powers of 10.

Interpreting Logarithms (cont.)

$$\frac{\partial \widehat{users}}{\partial GDP} = \frac{\partial}{\partial GDP}[-66.5 + 13.9 \ln GDP]$$

Interpreting Logarithms (cont.)
$\frac{\partial \widehat{users}}{\partial GDP} = \frac{\partial}{\partial GDP}[-66.5 + 13.9 \ln GDP]$

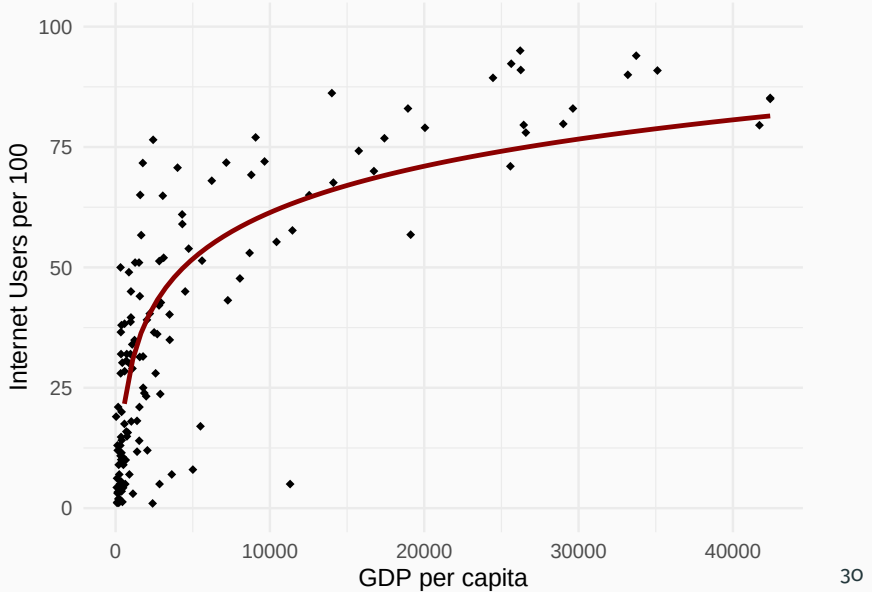
1. $\frac{\partial}{\partial x} \ln x = 1/x$
2.

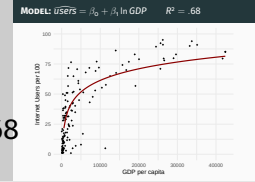
$$= 13.9 \frac{1}{GDP}$$

$$\partial \widehat{users} = 13.9 \frac{\partial GDP}{GDP}$$

$$\Delta \widehat{users} \approx 13.9 \frac{\Delta GDP}{GDP}$$

Model: $\widehat{users} = \beta_0 + \beta_1 \ln GDP$ $R^2 = .68$





1. Let's see what that rule of thumb looks like on the graph. Let's choose some point on the curve, GDP = 10,000.
2. I'm going to draw the tangent line at that point - it's a good approximation for small changes in GDP.
3. I know the slope of the tangent is 13.9/GDP
4. Let's say that I increase GDP by some proportion, αGDP so I can mark that as the change in X.
5. What's the change in Y? I multiple the change in X by the slope and I get 13.9α
6. If alpha is small, then the tangent line is close to the curve, so this is a good approximation.
7. For example, if we increase GDP by 1%. Then $\alpha = .01$ then

Other Uses of Logarithms

Log-linear form: $\ln Y = \beta_0 + \beta_1 X$

- Interpretation: $\frac{\Delta Y}{Y} \approx \beta_1 \Delta X$
- Example: add 0.1 to $X \rightarrow$ add $0.1\beta_1 \cdot Y$ to Y

Log-log form: $\ln Y = \beta_0 + \beta_1 \ln X$

- Constant elasticity model
- Interpretation: $\frac{\Delta Y}{Y} \approx \beta_1 \frac{\Delta X}{X}$
- Example: increase X by 1% $\rightarrow$ increase Y by β_1%

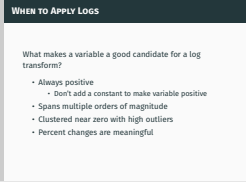
Other Uses of Logarithms
Log-linear form: $\ln Y = \beta_0 + \beta_1 X$
- Interpretation: $\frac{\Delta Y}{Y} \approx \beta_1 \Delta X$
- Example: add 0.1 to $X \rightarrow$ add $0.1\beta_1 \cdot Y$ to Y

Log-log form: $\ln Y = \beta_0 + \beta_1 \ln X$
- Constant elasticity model
- Interpretation: $\frac{\Delta Y}{Y} \approx \beta_1 \frac{\Delta X}{X}$
- Example: increase X by 1% $\rightarrow$ increase Y by β_1%

When to Apply Logs

What makes a variable a good candidate for a log transform?

- Always positive
  - Don’t add a constant to make variable positive
- Spans multiple orders of magnitude
- Clustered near zero with high outliers
- Percent changes are meaningful



1. This is especially true for monetary variable. We all understand what a 20% increase in salary is. A dollar increase can feel very different depending on how much money you have, but 10% more is meaningful for almost everyone.

Power Transformations

Power transformation: Replace X with X^α for some $\alpha \in \mathbb{R}$.

- $\sqrt{X}$ behaves similarly to $\ln X$.
- X^2, X^3,... spreads values far from zero further apart.

Interpretation is more difficult than with logs.

Power Transformations
Power transformation: Replace X with X^α for some $\alpha \in \mathbb{R}$.
- $\sqrt{X}$ behaves similarly to $\ln X$.
- X^2, X^3,... spreads values far from zero further apart.

Interpretation is more difficult than with logs.

1. Alex, tried to trim this down, but also happy to skip the slide.
2. I want to very quickly mention a broader class of transformations, called the power transformations
3. Here, we're raising X to a power.
4. For example, you might take the square root of X in situations where you can't take the log, because a variable sometimes equals zero.
5. However, these transformations make it much harder to interpret the results of a model, so they're not used very often in descriptive modeling.

When to Apply Transformations

Question: Should you always transform your variables to make them normal?

- Normal variables are **never** a requirement for OLS regression.
  - Even in the classical linear model, *errors* are normal, not variables.

$\Longrightarrow$ Don't focus too much on normality. Capturing relationships is more important.

When to Apply Transformations

Question: Should you always transform your variables to make them normal?

- Normal variables are **never** a requirement for OLS regression.
  - Even in the classical linear model, *errors* are normal, not variables.

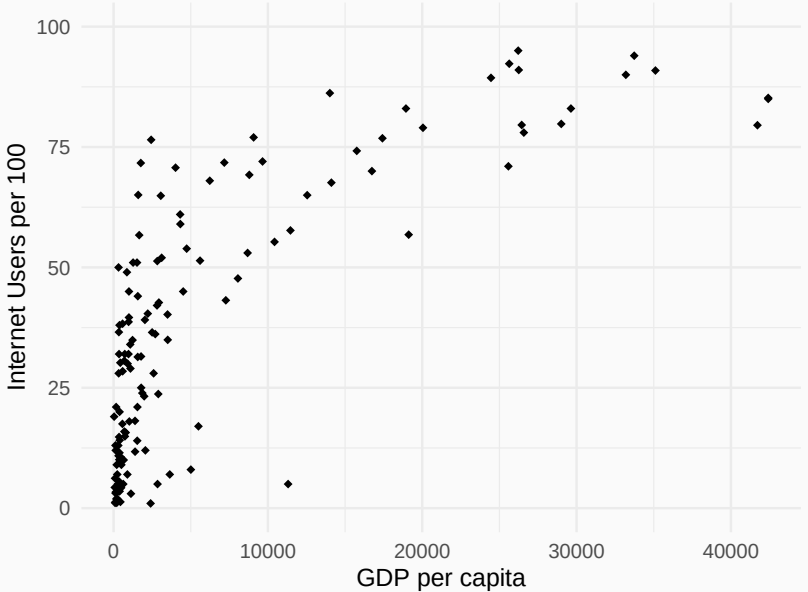
$\Longrightarrow$ Don't focus too much on normality. Capturing relationships is more important.

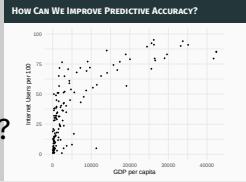
1. Before we move on, I want to address one issue, a lot of websites will tell you to always transform your variables until they look normal. Is that a good idea?
2. Well, first of all, normal variables are never a requirement for ols. Even in the classical linear model, the errors must be normal, not the variables. Of course if you're in a large sample, there's no assumption of normality at all.
3. But I will say that normal variables are generally a good thing. Normal means that most of your data isn't so clustered that it doesn't affect the model. So it's good to try for normal variables, but bad if that's the only thing you're thinking about

Polynomials

Polynomials

How Can We Improve Predictive Accuracy?





1. Here's another look at our GDP and Internet User data. Previously, we used a log transform. That was good for capturing the shape and great for model interpretability.
2. What if I tell you that your goals are little different. You still care about interpretation, but you really want more accuracy. You'd be willing to give up a little bit of interpretability to get a higher R^2
3. In that case, you might want to consider a polynomial regression.

Polynomial Regression

Motivation: Polynomials can approximate any continuous function (e.g. Weierstrass theorem)

Quadratic: $\widehat{Users} = \beta_0 + \beta_1 GDP + \beta_2 GDP^2$

- OLS finds the parabola that minimizes MSE.

Cubic: $\widehat{Users} = \beta_0 + \beta_1 GDP + \beta_2 GDP^2 + \beta_3 GDP^3$

- OLS finds the cubic function that minimizes MSE.

Polynomial Regression

Motivation: Polynomials can approximate any continuous function (e.g. Weierstrass theorem)

Quadratic: $\widehat{Users} = \beta_0 + \beta_1 GDP + \beta_2 GDP^2$

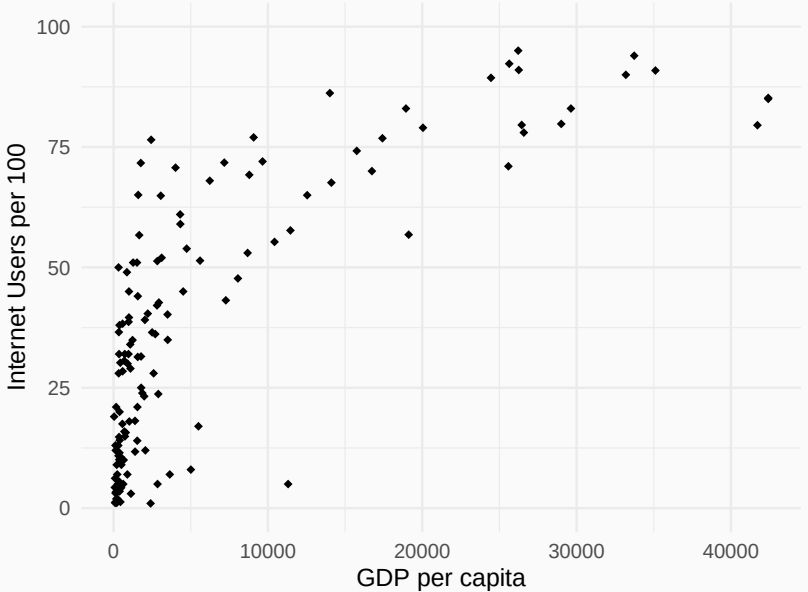
- OLS finds the parabola that minimizes MSE.

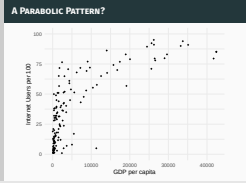
Cubic: $\widehat{Users} = \beta_0 + \beta_1 GDP + \beta_2 GDP^2 + \beta_3 GDP^3$

- OLS finds the cubic function that minimizes MSE.

1. First, why polynomials? well, polynomials are really flexible. They can approximate any continuous function.
2. The idea here is we're going to run a regression with different powers of the same variable.
3. Most of the time, we'll use just a quadratic. So we'll put in GDP and GDP^2
4. What will ols regression do? ols finds the beta's that minimize MSE. If you change these betas, you get all the different parabolas in the X-Y plane.
5. You can also put in the cubic term, though that's much less common.

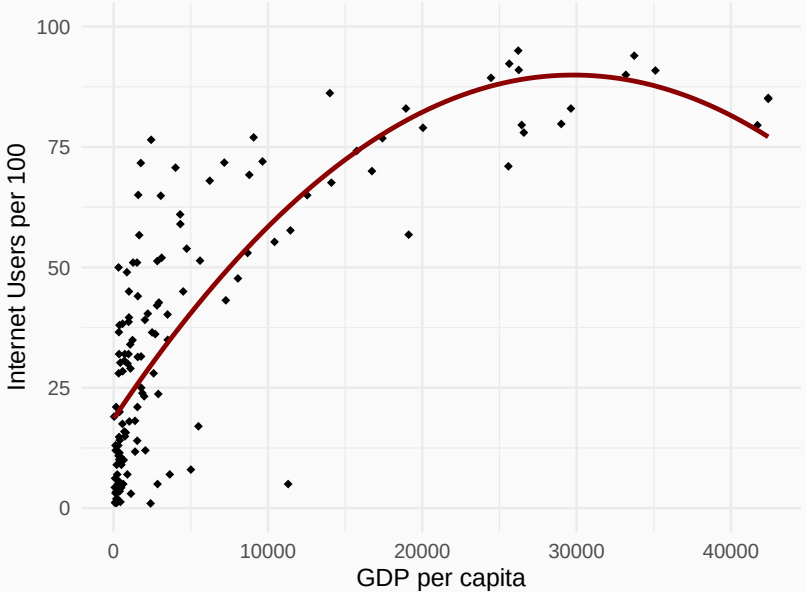
A Parabolic Pattern?

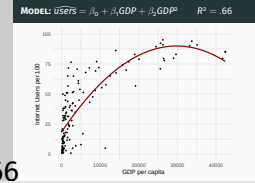




1. Let's try the quadratic form.
2. Here's the data again. I want you to imagine what you think the best fitting parabola looks like. try tracing it with your finger.

Model: $\widehat{users} = \beta_0 + \beta_1 GDP + \beta_2 GDP^2$ $R^2 = .66$





1. There you have the best fitting parabola
2. You can see that it definitely looks closer to the data than the line. The fit is actually slightly worse

Interpreting the Quadratic Specification

$$\widehat{Users} = 18.5 + .0048\ GDP - 8.0 \cdot 10^{-8}\ GDP^2$$

$$\frac{\partial \widehat{Users}}{\partial GDP} = .0048 - 1.6 \cdot 10^{-7}\ GDP$$

Example 1: If $GDP = 10,000$, then $\frac{\partial \widehat{Users}}{\partial GDP} = .0032$.

(Extra \$1,000 in GDP associated with 3.2 extra users)

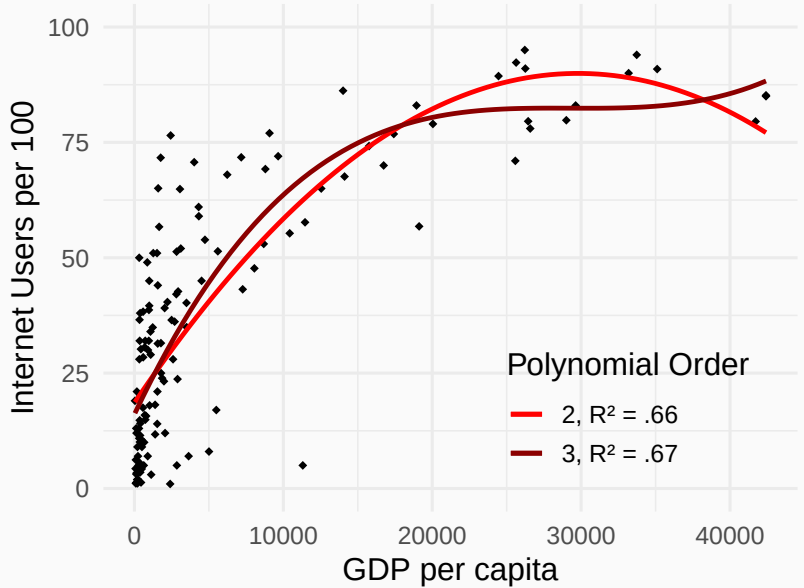
Example 2: If $GDP = 20,000$, then $\frac{\partial \widehat{Users}}{\partial GDP} = .0016$.

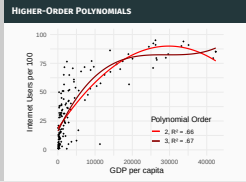
(Extra \$1,000 in GDP associated with 1.6 extra users)

Interpreting the Quadratic Specification
$\widehat{Users} = 18.5 + .0048\ GDP - 8.0 \cdot 10^{-8}\ GDP^2$
$\frac{\partial \widehat{Users}}{\partial GDP} = .0048 - 1.6 \cdot 10^{-7}\ GDP$
Example 1: If $GDP = 10,000$, then $\frac{\partial \widehat{Users}}{\partial GDP} = .0032$.
(Extra \$1,000 in GDP associated with 3.2 extra users)
Example 2: If $GDP = 20,000$, then $\frac{\partial \widehat{Users}}{\partial GDP} = .0016$.
(Extra \$1,000 in GDP associated with 1.6 extra users)

1. Here's the equation with the fitted coefficients
2. What's the interpretation of this equation?
3. Let's look at the slope... we get a function - the partial depends on GDP

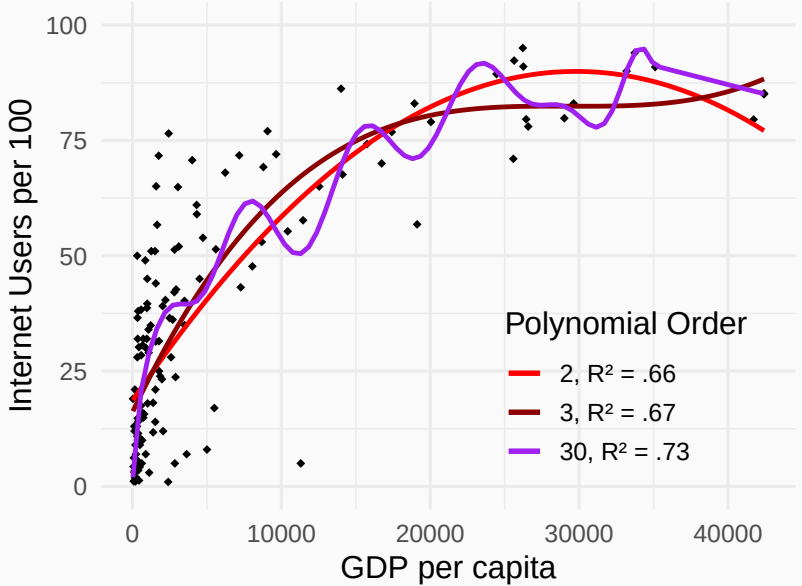
Higher-Order Polynomials

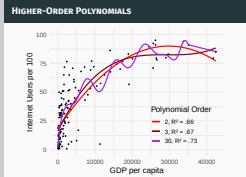




1. what happens if we increase the order of our polynomial?
2. Here you can see the best fitting cubic function. Notice that the R^2 went up a bit - of course, we knew that would happen. R^2 can only go up from adding variables.
3. The curve looks a little better at the end, it doesn't turn downwards.
4. How far can we push this? can we keep increasing the order of the polynomial and keep getting better and better predictions?

Higher-Order Polynomials





1. Here, I tried to fit an order 30 polynomial. Indeed, the R^2 goes up even more.
2. But the shape doesn't make sense. Why would internet users go up and down with GDP?
3. Especially on the right hand side, you can see that the curve traces from one point to the next.
4. This is a situation called overfitting. We have a high R^2, but we're really just modeling the noise in this one dataset. If we tried to use this model to predict internet users for a new set of countries, it would probably perform worse than the quadratic and cubic models.

Measurement With Controls

Measurement With Controls

Measurement With Controls

Note: This is a placeholder slide for an introduction that we will provide to the section. We're just placing it here for organization.

Measurement With Controls

Note: This is a placeholder slide for an introduction that we will provide to the section. We're just placing it here for organization.

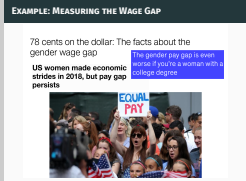
Interpreting Indicator Variables

Example: Measuring the Wage Gap

78 cents on the dollar: The facts about the gender wage gap

The gender pay gap is even worse if you're a woman with a college degree

US women made economic strides in 2018, but pay gap persists



1. Let's take a look at an example measurement goal. The example we'll use is the wage gap.
2. You probably already know about the wage gap - newspapers in the US talk about it from time to time.
3. In particular, you may have heard this figure: that a woman in the US earns 78 cents for every dollar that a man earns
4. I hope you agree that this is an important topic - we would like to live in a country where people of all genders are paid equally
5. But how can you actually measure the wage gap? Should we compare all men and all women? Or men and women

Example: Measuring the Wage Gap (cont.)

Data: Current Population Survey, 2019 Annual Social and Economic Supplement (ASEC)

Key variables:

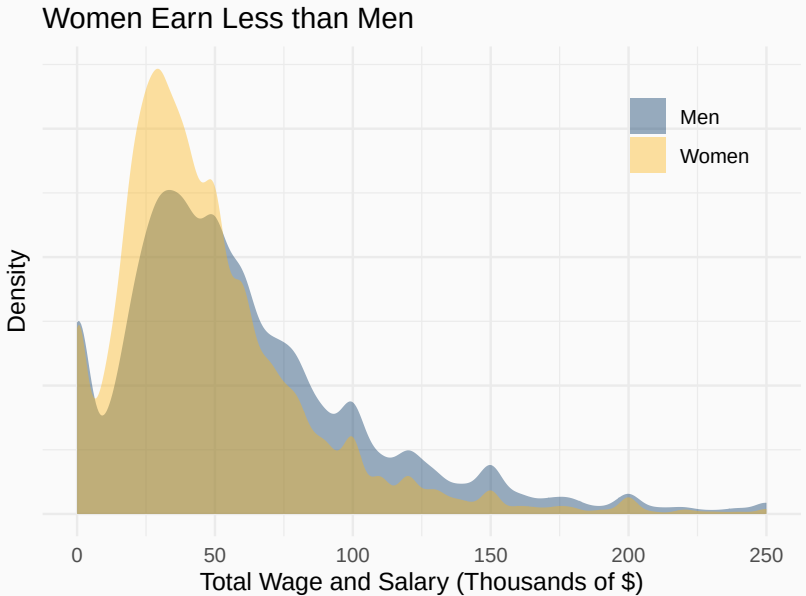
- Status: full-/part-time work status
  - 1: not in labor force
  - 2: full-time hours (35 or more), usually full-time

    ⋮
- Pay: total wage and salary earnings
- Sex
  - 1: male
  - 2: female

Example: Measuring the Wage Gap (cont.)
Data: Current Population Survey, 2019 Annual Social and Economic Supplement (ASEC)
Key variables:
- Status: full-/part-time work status
  - 1: not in labor force
  - 2: full-time hours (35 or more), usually full-time
    ⋮
- Pay: total wage and salary earnings
- Sex
  - 1: male
  - 2: female

1. In the survey, there is an income variable, and unlike most surveys, it's not binned into ranges, it's any integer.
2. The survey has no variable for gender, and it has one variable labeled sex. You can see It only has two levels, male and female. This is not a best practice, because it means that transgender and non-binary people may feel like they don't have an option. Unfortunately, I have to say that we're filming this in 2020, and this is still what most big surveys are like.

Overall Comparison



1. First, let's try the simplest thing we can do: comparing all men against all women.
2. Actually, we've seen this problem before. When we talked about t-tests, we had two groups, and we wanted to know if the means were the same.

Direct Application of t-Test

Mean for men: $68,700

Mean for women: $52,100

Mean difference: $16,600

(76 cents to the dollar)

$t = 27.938, df = 61,733, p < 2.2e - 16$

Direct Application of t-Test

Mean for men: $68,700
Mean for women: $52,100
Mean difference: $16,600
(76 cents to the dollar)
$t = 27.938, df = 61,733, p < 2.2e - 16$

1. Here are the results of that t-test analysis.
2. You can see man pay for men about 69K, 52K for women,
3. That corresponds to about 76 cents to the dollar
4. The t statistic is giant - almost 30, so the null is overwhelmingly rejected
5. Now, what about linear regression? Well, it's important to realize that this exact analysis can be done in a regression framework.

The Same Analysis in a Regression Framework, Part I

$$\widehat{Pay} = \beta_0 + \beta_1 Female$$

For men, $\widehat{Pay} = \beta_0 + \beta_1(0) = \beta_0$.

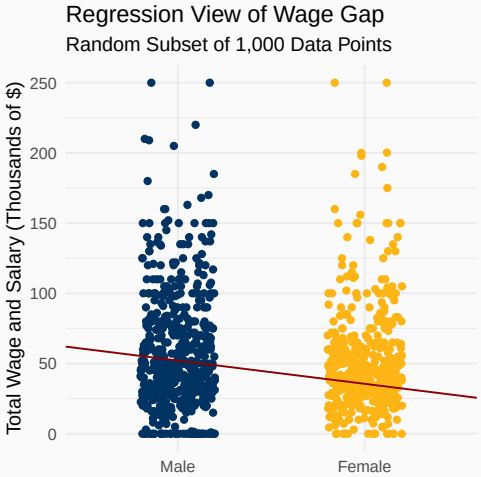
For women, $\widehat{Pay} = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$.

$\implies$ β_1 represents the difference between groups.

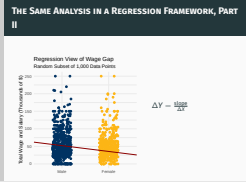
The Same Analysis in a Regression Framework, Part I
$\widehat{Pay} = \beta_0 + \beta_1 Female$
For men, $\widehat{Pay} = \beta_0 + \beta_1(0) = \beta_0$.
For women, $\widehat{Pay} = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$.
$\implies$ β_1 represents the difference between groups.

1. This may seem like a strange view of regression - aren't we supposed to be fitting a line?

The Same Analysis in a Regression Framework, Part II



$$\Delta Y = \frac{\text{slope}}{\Delta X}$$



1. Well, there is a line. Here's what the data looks like if we plot it, and here's the regression line.
2. (I took a small subset and added a little jitter so you can see things better.)
3. Since we have an indicator variable, the only x values are 0 and 1
4. What's the difference in in the predicted Y?...
5. I can write: $\Delta Y = \frac{\text{slope}}{\Delta X} = \frac{\beta_1}{1} = \beta_1$

The Same Analysis in a Regression Framework, Part III

	Dependent Variable: Pay
Female	−16,561***
	(618)
Intercept	68,667***
	(408)
Observations	62,110
Note:	*p<0.05; **p<0.01; ***p<0.001

The Same Analysis in a Regression Framework, Part III

	Dependent Variable: Pay
Female	−16,561***
	(618)
Intercept	68,667***
	(408)
Observations	62,110
Note:	*p<0.05; **p<0.01; ***p<0.001

1. When we look at our regression output, the coefficient for Female will have exactly the same difference in means that we say earlier.
2. One more thing to note: we usually show a standard t-test for each coefficient. Here, we have 3 stars for Female. The null of that test is that the slope is zero, which we know means that men and women are the same. So this is exactly the same as the classic t-test for two groups.
3. Here's another question to think about: what do the stars on the intercept mean?
4. Remember that the intercept represents the mean pay

Indicator Variables as Controls

Controlling for Occupation, Part I

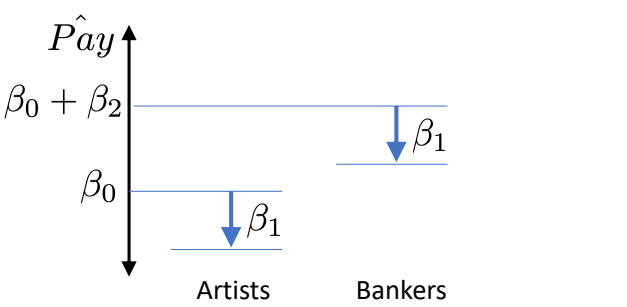
How can we compare men and women in the same occupation?

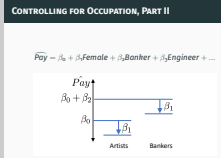
Controlling for Occupation, Part I
How can we compare men and women in the same occupation?

1. So far, we've just been comparing men as a whole to women as a whole
2. But part of the difference we see could be that men and women tend to be in different professions
3. Now, you might argue, so what? If the main problem is that women are being locked out of high-paying professions, it's the overall pay difference that we should be worried about.
4. But at least from certain perspectives, it might be more fair to compare men and women in the same profession.
5. I'm going to try it both ways, because I think we can learn from the results

Controlling for Occupation, Part II

$$\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Banker + \beta_3 Engineer + \ldots$$





1. What we need to do, is add an indicator variable for every possible occupation (we need a base category, let's make it artist).
2. What does this do? Here's a picture of what the model looks like.
3. If you want to know the predicted pay for a female Banker, you start with β_0, you add β_2 since $banker = 1$, then you add β_1 since $female = 1$.
4. The thing to notice is that β_1 now represents a common difference inside each profession. So it doesn't matter if there are a lot of male bankers. If men and women inside each profession earn the same, β_1 will be 0

Controlling for Occupation, Part III

	Dependent Variable: Pay	
	(1)	(2)
Female	−16,561***	−15,035***
	(618)	(679)
Intercept		183,292***
		(3,440)
occup20		84,255***
		(3,714)
occup40		76,881***
		(13,132)
occup50		108,112***
		(3,753)
⋮		⋮

Controlling for Occupation, Part III
Dependent Variable: Pay
(1) (2)
Female −16,561*** −15,035***
(618) (679)
Intercept 183,292***
(3,440)
occup20 84,255***
(3,714)
occup40 76,881***
(13,132)
occup50 108,112***
(3,753)
⋮ ⋮

1. Inside the survey data, we have an occupation code variable with 484 different values. If you just add it to the table directly, you have a problem.
2. Of course there's not enough room to list every occupation

Controlling for Occupation, Part IV

	Dependent Variable: Pay	
	(1)	(2)
Female	−16,561***	−15,035***
	(618)	(679)
Intercept	68,667***	7,110**
	(408)	(2,321)
Occupation FE	No	Yes
Observations	62,110	62,110
Note:	*p<0.05; **p<0.01; ***p<0.001	

Controlling for Occupation, Part IV

	Dependent Variable: Pay	
	(1)	(2)
Female	−16,561***	−15,035***
	(618)	(679)
Intercept	68,667***	7,110**
	(408)	(2,321)
Occupation FE	No	Yes
Observations	62,110	62,110
Note:	*p<0.05; **p<0.01; ***p<0.001	

1. So instead we can put in a row to summarize all these indicators.
2. You can see that actually, the effect for Female has shrunk by about 10 percent in this specification. That corresponds to about 78 cents to the dollar.
3. Why is that? This suggests, that in part, there are more men in occupations with higher pay. But even within each occupation, men are earning more than women.

Metric Variables as Controls

Metric Variables as Controls

Age as a Covariate

Ways to add age to a specification

Linear effect: $\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Age$

Polynomial effect: $\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Age + \beta_3 Age^2$

AGE AS A COVARIATE
Ways to add age to a specification
Linear effect: $\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Age$
Polynomial effect: $\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Age + \beta_3 Age^2$

1. We're trying to measure the pay gap, and another variable that we might want to control for is age
2. People tend to earn more as they get older, and you might wonder if this can explain part of the pay gap that we see.
3. How should we enter age into the regression?
4. You could put it in directly. That would make sense if you wanted to measure the age effect and wanted a single number to summarize the effect.
5. If, on the other hand, age is really just a control variable, you might want a more flexible specification, just to soak up as much variation as you can. For example, a

Understanding the Linear Age Effect

For men: $\widehat{Pay} = \beta_0 + \beta_1(0) + \beta_2 Age = \beta_0 + \beta_2 Age$

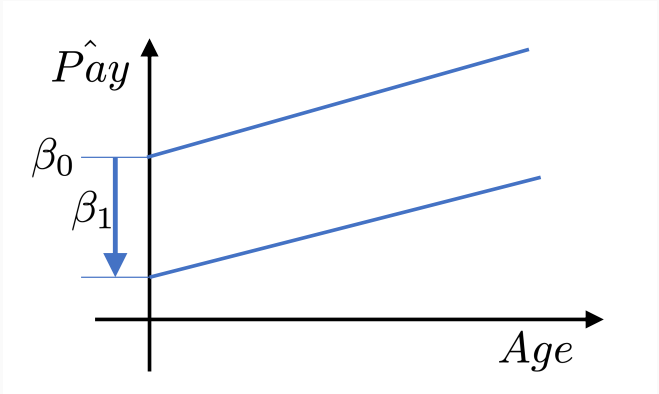
For women: $\widehat{Pay} = \beta_0 + \beta_1(1) + \beta_2 Age = \beta_0 + \beta_1 + \beta_2 Age$

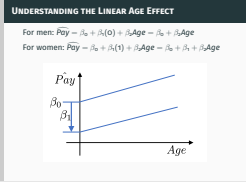
Understanding the Linear Age Effect
For men: $\widehat{Pay} = \beta_0 + \beta_1(0) + \beta_2 Age = \beta_0 + \beta_2 Age$
For women: $\widehat{Pay} = \beta_0 + \beta_1(1) + \beta_2 Age = \beta_0 + \beta_1 + \beta_2 Age$

Understanding the Linear Age Effect

For men: $\widehat{Pay} = \beta_0 + \beta_1(0) + \beta_2 Age = \beta_0 + \beta_2 Age$

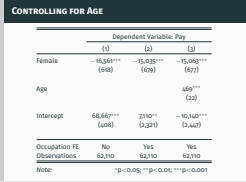
For women: $\widehat{Pay} = \beta_0 + \beta_1(1) + \beta_2 Age = \beta_0 + \beta_1 + \beta_2 Age$





Controlling for Age

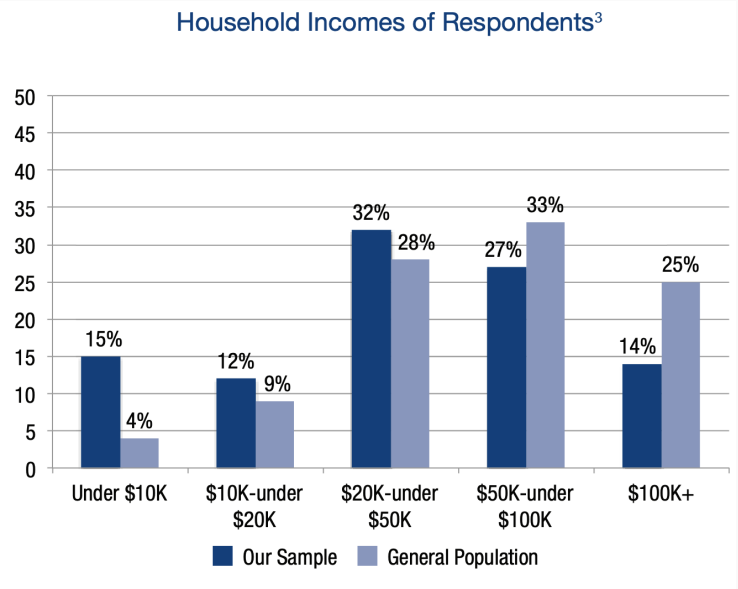
	Dependent Variable: Pay		
	(1)	(2)	(3)
Female	−16,561***	−15,035***	−15,063***
	(618)	(679)	(677)
Age			469***
			(22)
Intercept	68,667***	7,110**	−10,140***
	(408)	(2,321)	(2,447)
Occupation FE	No	Yes	Yes
Observations	62,110	62,110	62,110
Note:			*p<0.05; **p<0.01; ***p<0.001

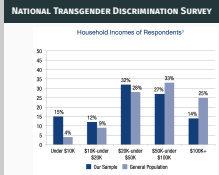


1. Here's the regression table, with the new specification added in. You can see that each year of age is associated with $469 in salary.
2. Also, adding age didn't change the gap between men and women at all.
3. Does that mean we should take this specification out of the table?
4. Well, no. This is all interesting information. We're building up a table that shows how our estimate changes or doesn't change depending on our choices.
5. If we keep trying different specifications and the estimate for Female never changes much, we'll say that it's robust.

Measuring With More Categories

National Transgender Discrimination Survey





1. Because of our data source, we had two categories for gender - male and female.
2. What if you could write your own survey - could you make it more inclusive by listing more options?
3. How would you analyze the data?
4. That could be an important question. I'm showing a graph from the National transgender discrimination survey. This is not a nationally representative survey, but out of the trans and nonconforming people they were able to reach, 15% had a household income under $10,000. At the least, that's a hint that this could be an important issue.

Example of a More Inclusive Question

How would you describe yourself?

- Female
- Male
- Transgender female
- Transgender male
- Non-binary/nonconforming
- Prefer not to answer

1. Here's an example of a question with more options. We're not claiming this is the best possible set of choices - but the intention is to provide options that more people identify with. We would like to see more surveys try to do this.
2. Alex, I realize I'm not sure if the question should be gender or sex. guess I've been mixing the two up...

Analyzing More Gender Categories

$$\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Transgender\_Male + \beta_3 Transgender\_Female + \beta_4 Nonbinary$$

$\implies$ Report each β to describe the pay gap for a different gender identity.

Analyzing More Gender Categories
$\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Transgender\_Male + \beta_3 Transgender\_Female + \beta_4 Nonbinary$
$\implies$ Report each β to describe the pay gap for a different gender identity.

1. Now we need to include each option as an indicator variable. In this case, I'm using male as a base category, because then each beta compares a gender against male, which is something we're very interested in.
2. Now what could happen here? It might be that you just don't get enough data in some categories. That would make your standard error for those categories too high. So that means that you need to be careful about adding too many options.
3. If you face this problem, you may be able to combine some categories together in order to gain accuracy.

Reporting Overall Effect

Law of total variance:

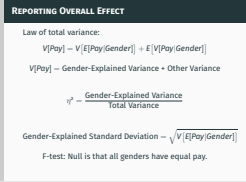
$$V[Pay] = V\big[E[Pay|Gender]\big] + E\big[V[Pay|Gender]\big]$$

$$V[Pay] = \text{Gender-Explained Variance} + \text{Other Variance}$$

$$\eta^2 = \frac{\text{Gender-Explained Variance}}{\text{Total Variance}}$$

$$\text{Gender-Explained Standard Deviation} = \sqrt{V\big[E[Pay|Gender]\big]}$$

F-test: Null is that all genders have equal pay.



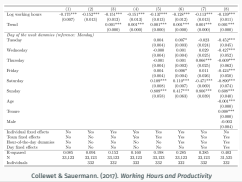
1. On the one hand, you may really be interested in each individual β. but you may also want a summary measure - one number to sum up how big the pay gap is.
2. One idea is to break up the variation in pay and see how much of it can be tied to gender. remember the law of total variance from earlier...
3. eta squared is defined as a fraction, of the gender-explained variance over the total variance. You can think of it as how much of the total variation can be explained through gender.
4. Another idea is to take the gender-explained variance and take the square root, to give gender-explained

Planning Multiple Specifications

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log working hours	-0.175***	-0.152***	-0.151***	-0.151***	-0.137***	-0.128***	-0.113***	-0.139***
	(0.007)	(0.013)	(0.013)	(0.013)	(0.013)	(0.012)	(0.013)	(0.011)
Trend			0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Day of the week dummies (reference: Monday)								
Tuesday					0.004	0.007*	-0.023	-0.452***
					(0.004)	(0.003)	(0.024)	(0.045)
Wednesday					-0.000	0.001	0.029	-0.427***
					(0.004)	(0.004)	(0.025)	(0.052)
Thursday					-0.001	0.001	0.066***	-0.609***
					(0.004)	(0.003)	(0.025)	(0.063)
Friday					0.004	0.006*	0.011	-0.424***
					(0.004)	(0.004)	(0.026)	(0.050)
Saturday					0.109***	0.110***	-0.471***	-0.890***
					(0.008)	(0.007)	(0.069)	(0.074)
Sunday					0.809***	0.417***	0.806***	0.608***
					(0.058)	(0.063)	(0.039)	(0.040)
Age								-0.001***
								(0.000)
Tenure								0.008***
								(0.000)
Male								-0.003
								(0.004)
Individual fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes	No
Team fixed effects	No	No	No	Yes	Yes	Yes	Yes	Yes
Hour-of-the-day dummies	No	No	No	No	No	Yes	Yes	Yes
Day fixed effects	No	No	No	No	No	No	Yes	Yes
R-squared	0.085	0.094	0.152	0.160	0.198	0.285	0.385	0.403
N	33,123	33,123	33,123	33,123	33,123	33,123	33,123	31,525
Individuals		332	332	332	332	332	332	332

Collewet & Sauermann. (2017). *Working Hours and Productivity*

Planning Multiple Specifications



Collewet & Sauermann. (2017). *Working Hours and Productivity*

1. We've been building up a regression table for our study, and so far it has three models in it - we might say three specifications.
2. That's not at all unusual. Actually, if you write articles in some fields, it's common to include many more specifications. Here's a table from a paper on working hours and productivity..
3. You can see that there are 8 specifications. Model 1 only has a single variable - that's the effect that they want to measure. Then for the most part the models on the right have more and more variables.
4. What's the purpose of this?

Why Report Multiple Specifications?

Modeling decisions are tough.

- Is it worthwhile to control for education, even if standard errors increase?
- Is it worth switching from a log to a square root to get a much better fit?
- Should we include occupation, even if that absorbs some of the concept we want to measure?

Would your results be different with different choices?

$\Longrightarrow$ Try different paths to see if your results are **robust**.

Why Report Multiple Specifications?

Modeling decisions are tough.

- Is it worthwhile to control for education, even if standard errors increase?
- Is it worth switching from a log to a square root to get a much better fit?
- Should we include occupation, even if that absorbs some of the concept we want to measure?

Would your results be different with different choices?

⟹ Try different paths to see if your results are **robust**.

Why Report Multiple Specifications? (cont.)

Error rate inflation: a tendency for effects to appear large (and p-values small), especially when a researcher uses the same data for exploration and testing

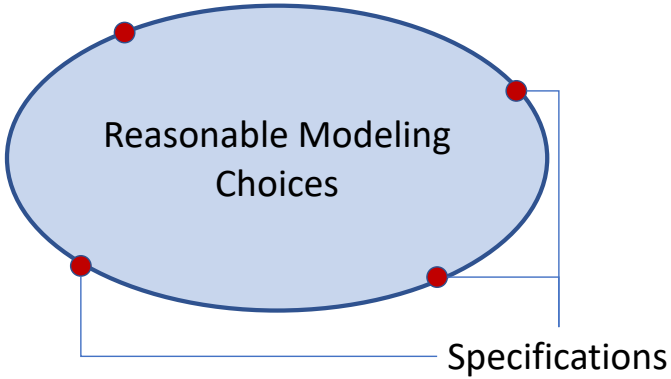
p-Hacking: a deliberate attempt to generate significant results by altering the model specification and other researcher degrees of freedom

$\Longrightarrow$ Report multiple specifications to guard against error rate inflation.

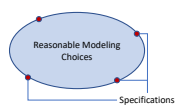
Why Report Multiple Specifications? (cont.)
Error rate inflation: a tendency for effects to appear large (and p-values small), especially when a researcher uses the same data for exploration and testing
p-Hacking: a deliberate attempt to generate significant results by altering the model specification and other researcher degrees of freedom
⟹ Report multiple specifications to guard against error rate inflation.

1. There's a deeper reason that we report multiple specifications, and it has to do with reproducibility.
2. Let's review two very important terms...
3. If you have a lot of data, you may be able to hold out some of it, so you can get your final p-values from fresh data. But even if you do that, your audience may just expect to see multiple specifications, and may be more suspicious if they only see one.

How Should You Think About Specifications?



How Should You Think About Specifications?



1. With all that in mind, how should you think about your model specifications?
2. Imagine a space that represents reasonable modeling choices
3. They still have to reasonable - you shouldn't consider models that can't be defended
4. Try to select your specifications to encircle that space, so your results can represent the larger set of choices.

An Example Specification Table

	Dependent Variable: Pay		
	(1)	(2)	(3)
Female	−16,561***	−15,035***	−15,063***
	(618)	(679)	(677)
Age			469***
			(22)
Intercept	68,667***	7,110**	−10,140***
	(408)	(2,321)	(2,447)
Occupation FE	No	Yes	Yes
Observations	62,110	62,110	62,110
Note:		*p<0.05; **p<0.01; ***p<0.001	

An Example Specification Table
Dependent Variable: Pay
(1) (2) (3)
Female −16,561*** (618) −15,035*** (679) −15,063*** (677)
Age 469*** (22)
Intercept 68,667*** (408) 7,110** (2,321) −10,140*** (2,447)
Occupation FE No Yes Yes
Observations 62,110 62,110 62,110
Note: *p<0.05; **p<0.01; ***p<0.001

1. Let's take another look at our own specification table.
2. A few things to remember. First, this is a measurement study, and our key variable is Female. So we place it on the first row, and of course it has to appear in every single specification.
3. One thing you almost always want to do is start with a base model, that only has your key variable and nothing else. If you have a very strong reason to put in another variable, that's ok, but you want this to be a very minimal model.
4. It's very common to add variables in blocks as you go left to right. If you do that, then your models will be nested,

Modeling Conditional Effects

Conditional Effects

Conditional Effects

Idea: The relationships we measure may be different for different people or units.

- A vaccine may reduce infection rates more in adults than in children.
- Network access may be more associated with loan availability in poor countries.
- The pay gap may be different for people of different ages.

Conditional Effects

Idea: The relationships we measure may be different for different people or units.

- A vaccine may reduce infection rates more in adults than in children.
- Network access may be more associated with loan availability in poor countries.
- The pay gap may be different for people of different ages.

1. Let's take a closer look at that last example...

Conditional Effects of Age, Part I

Interaction term: a variable in a regression formed by multiplying two other variables together

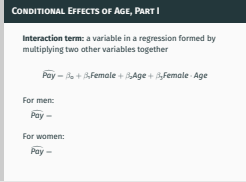
$$\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Age + \beta_3 Female \cdot Age$$

For men:

$$\widehat{Pay} =$$

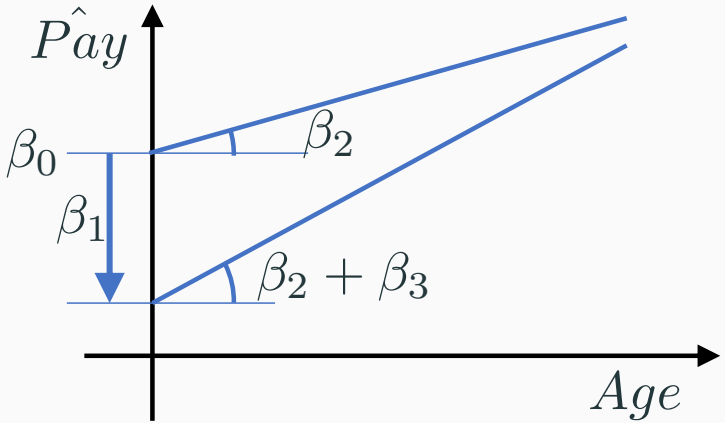
For women:

$$\widehat{Pay} =$$

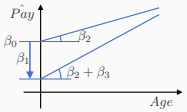


1. There's a simple technique for modeling heterogeneous effects. it's called an interaction term. An interaction term is just two of your variables multiplied together.
2. Let's see what this equation looks like for men and for women.
3. $beta_0 + \beta_1(0) + \beta_2 Age + \beta_3(0) Age = \beta_0 + \beta_2 Age$
4. $\beta_0 + \beta_1(1) + \beta_2 Age + \beta_3(1) Age = \beta_0 + \beta_1 + (\beta_2 + \beta_3) Age$
5. Notice that these are again two different lines. once again, they have different intercepts. but this time, they also have different slopes.

Conditional Effects of Age, Part II



Conditional Effects of Age, Part II



1. Here's what that looks like. we're fitting a very flexible model, with two different slopes and two different intercepts.
2. by the way, how would you test whether the effect really is heterogeneous? that is, how can you test whether the slopes really are different?
3. The answer is that the slopes are the same exactly if $\beta_3 = 0$. so you just use the regular old t-test that's standard in regression output.

Conditional Effects of Age, Part III

	Dependent Variable: Pay
Female	−6,230**
	(1,967)
Age	559***
	(29)
Female:Age	−207***
	(43)
Intercept	−13,971***
	(2,575)
Occupation FE	No
Observations	62,110
Note:	*p<0.05; **p<0.01; ***p<0.001

Conditional Effects of Age, Part III
Dependent Variable: Pay
Female −6,230** (1,967)
Age 559*** (29)
Female:Age −207*** (43)
Intercept −13,971*** (2,575)
Occupation FE No
Observations 62,110
Note: *p<0.05; **p<0.01; ***p<0.001

1. Here's our result, and you can see it's pretty interesting. The interaction term is negative and statistically significant. And the wage gap appears larger for older workers. With each year of age, the gap increases by about $200

Interaction Terms

Interaction Terms

Interactions for Indicator Variables

Use old content: 12.9 Interaction Terms for Indicator Variables, Part 1

If there's extra time, I would redo this using the wage gap example.

Interactions for Indicator Variables

Use old content: 12.9 Interaction Terms for Indicator Variables, Part 1

If there's extra time, I would redo this using the wage gap example.

Guidelines for Polynomial and Interaction Terms

Use old content: 12.12 Guidelines for Polynomial and Interaction Terms

Could redo some of the discussion of goals if there is time.

Guidelines for Polynomial and Interaction Terms
Use old content: 12.12 Guidelines for Polynomial and Interaction Terms
Could redo some of the discussion of goals if there is time.