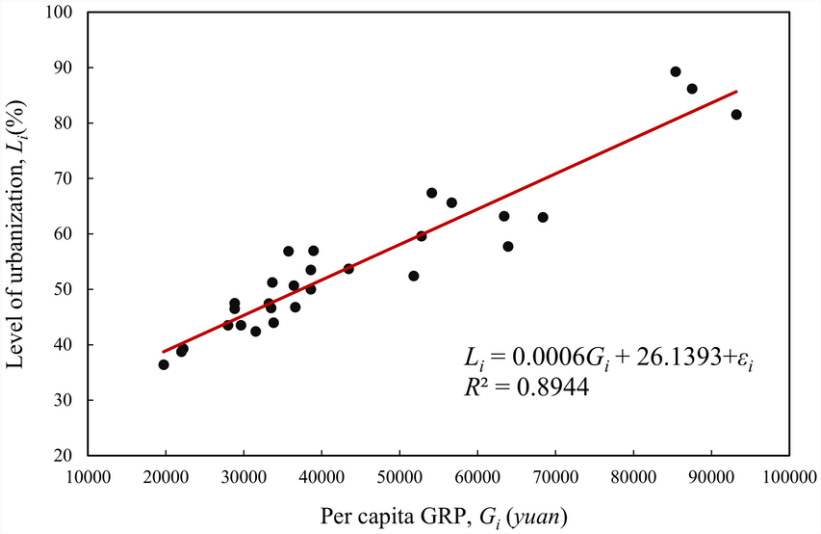
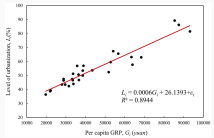
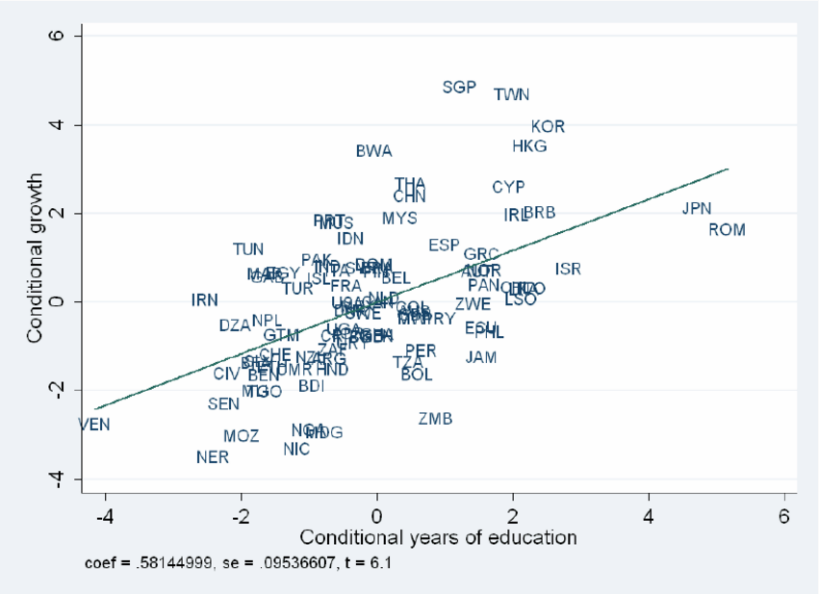
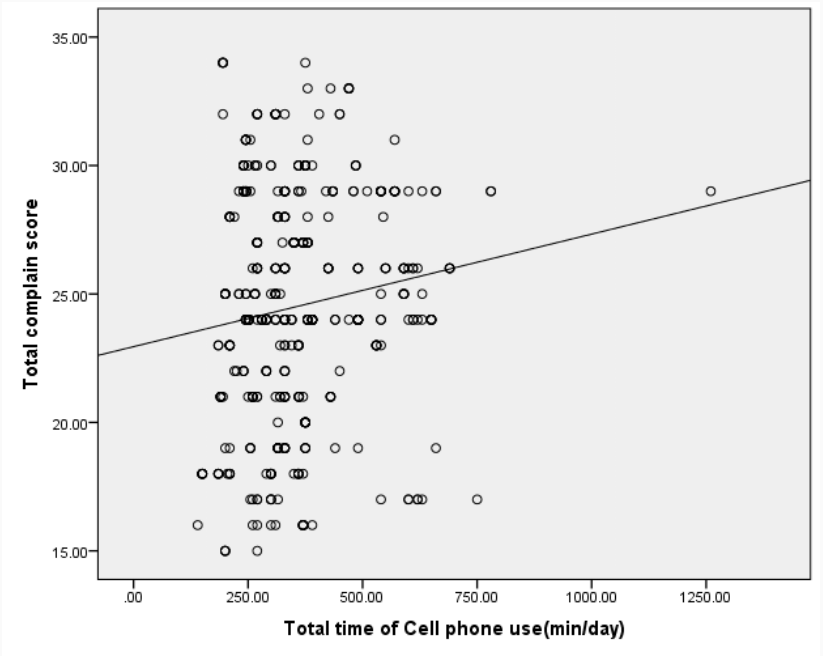
Introduction to Regression





1. If you open any academic paper that uses data today - there's a quite good chance that you'll find a linear regression inside. If you don't find a regression, you might find a technique that's based on linear regression. Even the most cutting edge machine learning techniques - most still contain elements of linear regression.


Conditional growth
Conditional years of education
coef = .58144999, se = .09536607, t = 6.1
SGP
TWN
KOR
HKG
BWA
THA
CHN
CYP
MYS
IRL
BRB
JPN
ROM
IDN
ESP
GRC
TUN
ISR
BEL
FRA
TUR
IRN
ZWE
LSO
DZA
NPL
PER
JAM
TZA
BOL
CIV
BEN
BDI
SEN
ZMB
VEN
MOZ
NER
NIC


35.00
30.00
25.00
20.00
15.00
Total complain score
.00
250.00
500.00
750.00
1000.00
1250.00
Total time of Cell phone use(min/day)

The Regression Algorithm

Idea: Draw a line given a sample of data

The Regression Algorithm
Idea: Draw a line given a sample of data

1. So regression is absolutely foundational to modern data science.
2. And, in one sense it's easy.
3. The basic idea is quite simple
4. The algorithm is quite simple - I can show you how to calculate the line in a few minutes. But just drawing a line isn't statistics - anyone can draw a line, but a line is just a line.
5. After you draw a line, there are two questions you need to ask...

What does this line *mean*?

Under what assumptions?

What does this line *mean*?
Under what assumptions?

1. **What's interesting is that even though everyone agrees on how to compute a regression line, people have different answers to these questions.**
2. **There's a statistical perspective, a structural perspective, a machine learning perspective...**
3. **Which do you need to learn? All of them. As a data scientist, you'll need to converse with many types of people...**
4. **But you have to start somewhere. We have a very strong opinion about this - we're going to teach you what we think is the most fundamental perspective on regression.**

The Mechanics of OLS Regression

- The OLS algorithm
- Statistical assumptions
- Statistical guarantees

The Mechanics of OLS Regression
- The OLS algorithm
- Statistical assumptions
- Statistical guarantees

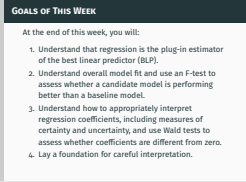
1. You can think of this unit as covering the mechanics of linear regression. The algorithm itself, statistical assumptions, and statistical guarantees. This is a foundation, and later on, we'll move on to applications. let's get started.

Unit Plan

Goals of This Week

At the end of this week, you will:

1. Understand that regression is the plug-in estimator of the best linear predictor (BLP).
2. Understand overall model fit and use an F-test to assess whether a candidate model is performing better than a baseline model.
3. Understand how to appropriately interpret regression coefficients, including measures of certainty and uncertainty, and use Wald tests to assess whether coefficients are different from zero.
4. Lay a foundation for careful interpretation.



1. When we discuss fitting, we're going to reason from the closed-form solutions. But, we will also nod to how we might produce estimates through a brute-force, numeric optimization route.

Reading: The Golem of Prague

Reading: The Golem of Prague

Reading: The Golem of Prague

This is a placeholder for a reading call. We're just placing it here for organization.

- Read Sections 1.0 and 1.1 of *Statistical Rethinking*, which we have provided a copy of in PDF form from the publisher.
- *Statistical Rethinking* is a great book and reference that you should consider later in your data science and statistics path.

Reading: The Golem of Prague

This is a placeholder for a reading call. We're just placing it here for organization.

- Read Sections 1.0 and 1.1 of *Statistical Rethinking*, which we have provided a copy of in PDF form from the publisher.
- *Statistical Rethinking* is a great book and reference that you should consider later in your data science and statistics path.

Regression, a Statistical Golem

Regression, a Statistical Golem

Regression, a Statistical Golem

- Regression—like all models—is a tool.
- We put tools to use toward a data scientific purpose; however, tools are only tools.
- Use of a straight-edge, scale, and T-square doesn't make one an architect any more than use of {insert language} or {insert technique} makes one a data scientist.

Regression, a Statistical Golem
- Regression—like all models—is a tool.
- We put tools to use toward a data scientific purpose; however, tools are only tools.
- Use of a straight-edge, scale, and T-square doesn't make one an architect any more than use of {insert language} or {insert technique} makes one a data scientist.

1. A major point in 201, and again here in 203 is that we cannot, under any circumstance – whether question formation, ethics, data viz, model architecture – be unthinking as data scientists. Because our models *are* unthinking.
2. In the case of regression, it is *very* literally a measure of linear dependence between two dimensions of data.
3. The tasks that we pursue require *careful* thought. This is the task that makes a pursuit a science, rather than a psuedo-science; an art, rather than a craft.
4. But, combined with how we *know* human process – we're limited processors of information; we need to develop skills for managing where we are working.

The Machinery of OLS Regression

- OLS regression is a plug-in estimator for the best linear predictor (BLP).
- The BLP is the lowest mean squared error (MSE) estimator, out of all linear functions.

The Machinery of OLS Regression

- OLS regression is a plug-in estimator for the best linear predictor (BLP).
- The BLP is the lowest mean squared error (MSE) estimator, out of all linear functions.

1. I removed CEF, because it's an intermediary step - the target variable is the ultimate estimation goal
2. On the one hand, there is little special about regression.
3. It is a simple formula, with a closed form solution that is both easy to fit (because of squared loss). It just so happens that minimizing MSE happens to be desirable in many cases.
4. But this is quite literally the beginning and end of what it does. If we wanted to minimize quadratic squared error; or absolute deviations, we would derive another estimator.
5. What if we wanted to make a model that was usefully understandable by a stakeholder for descriptive purposes? What if we wanted to identify a causal effect? There is no guarantee that minimizing MSE will accomplish these goals. We as scientists have to carefully define the question and task

Applications of OLS Regression

Regression is fantastically versatile

- Under some circumstances, regression has explainable internal weights (coefficients) that are of interest.
- Under other circumstances, regression identifies causal effects.
- Under many circumstances, regression is the *de facto* baseline estimator.

Applications of OLS Regression

Regression is fantastically versatile

- Under some circumstances, regression has explainable internal weights (coefficients) that are of interest.
- Under other circumstances, regression identifies causal effects.
- Under many circumstances, regression is the *de facto* baseline estimator.

1. But, on the other hand...

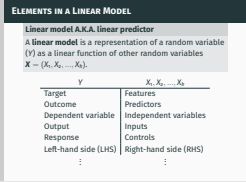
Elements of a Linear Model

Elements in a Linear Model

Linear model A.K.A. linear predictor

A **linear model** is a representation of a random variable (Y) as a linear function of other random variables $\boldsymbol{X} = (X_1, X_2, ..., X_k)$.

Y	$X_1, X_2, ..., X_k$
Target	Features
Outcome	Predictors
Dependent variable	Independent variables
Output	Inputs
Response	Controls
Left-hand side (LHS)	Right-hand side (RHS)
⋮	⋮



1. Before we dive into OLS regression, let's discuss linear models more broadly.
2. By a linear model, we mean that we have one random variable Y and we're going to represent or summarize it as a linear function of other random variables $X_1, X_2, ..., X_k$
3. These variables go by many different names.
4. In ML, Y is usually the target, X's are features
5. In classical stats, dependent and independent variables, or sometimes RHS and LHS are more popular
6. Those are ok, but we don't like that dependent variable sounds like a causal effect. As we'll discuss, we can't just assume that a linear model represents any kind of causal relationship. For that reason, we'll try to stick with the more prediction-focused terms, like target and features

The Linear Model Formula

The linear model formula

$$\begin{aligned}\hat{Y} &= g(X_1, X_2, ..., X_k) \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k\end{aligned}$$

The Linear Model Formula
The linear model formula
$\hat{Y} = g(X_1, X_2, ..., X_k)$
$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$

1. Here's what the linear model looks like
2. We may write g of the X's to emphasize that this is a function
3. The number that comes out of the function is called a prediction
4. Since we're putting random variables into the function, the output is also a random variable
5. We'll label the prediction $\hat{Y}$. The hat means that this is a prediction for Y, it's not the same random variable, but it is related.
6. β_k are the coefficients. we can also call them weights
7. Depending on the type of linear model building that you're doing, either β or $\hat{Y}$ might be the target of your inquiry.

A Linear Model Example

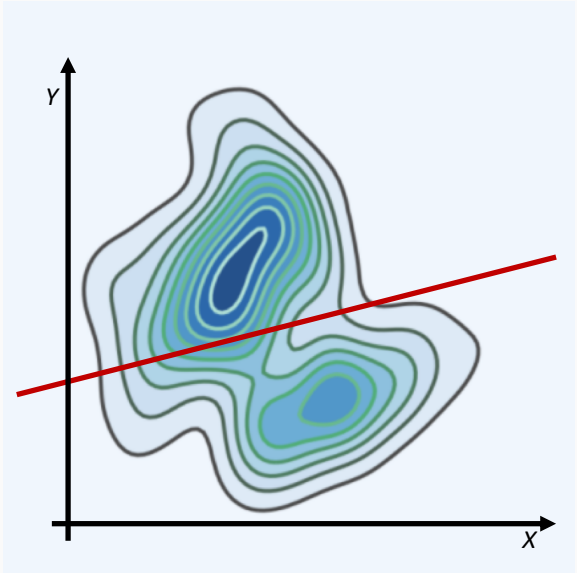
Brunch in Berkeley

$$\widehat{\text{Avocados}} = 2 + 1 \cdot \text{Lemons} + 2 \cdot \text{Loaves\_Bread}$$

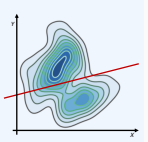
A Linear Model Example
Brunch in Berkeley
$\widehat{\text{Avocados}} = 2 + 1 \cdot \text{Lemons} + 2 \cdot \text{Loaves\_Bread}$

Review: Outcome, Prediction, and Error

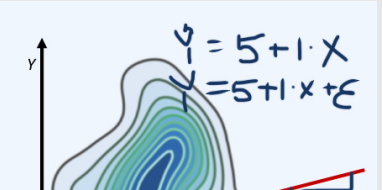
Review: Outcome, Prediction, and Error



Review: Outcome, Prediction, and Error



1. Here's a picture of a joint distribution, with just one X and Y.
2. And here's a line representing a linear model.
3. Is this a good model? Doesn't look great to my eyes, but that's actually intentional. We haven't said anything about what makes a good model yet!
4. If I imagine drawing a point (x, y) from this distribution, I can draw the distance from the point to the line. that distance is a random variable, with a special name: the error. This point on the y axis is the predicted y, $\hat{y}$



Concept Check: Making Predictions with a Linear Model

Concept Check: Making Predictions with a Linear Model

Concept Check: Making Predictions with a Linear Model

- Students will be given a model and (x,y) and compute prediction and error.

Concept Check: Making Predictions with a Linear Model
- Students will be given a model and (x,y) and compute prediction and error.

Metric Inputs

Interpreting Model Weights (Coefficients)

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Interpretation of coefficients

- If X_i changes by ΔX_i units, the predicted value of the target, $\hat{Y}$ changes by $\beta_i \cdot \Delta X_i$ units.
- If X_i and X_j change by ΔX_i and ΔX_j respectively, then the predicted value of the target, $\hat{Y}$ changes by $\beta_i \cdot \Delta X_i + \beta_j + \Delta X_j$.

Ceteris paribus: all else equal

1. How do we interpret the betas in a linear model?
2. A linear model is a *sharp* abstraction of the world. **But**, because it is linear, the *model* is easily interpretable. However, it is important to note that we frequently want our *inference* to be about the world, not our data. Only when the data meets certain assumptions does what we learn from model tell us about the world. (Recall properties of estimators.)
3. To get an idea, try taking the partial derivative of the predicted outcome, with respect to some X_i
4. This means that you can think of β_i as the change in the outcome when X_i changes by 1 unit, but the partial is very important. This is ONLY true if the other X's are held constant.

Interpreting Model Coefficients: Example

Does this model say peacocks with longer tails fly slower?

$$\text{air\_speed} = 4.3 - 1.2 \cdot \text{tail\_length} + 0.8 \cdot \text{muscle\_mass}$$

Photo by Thimindu Goonatillake CC BY-SA 2.0

Interpreting Model Coefficients: Example
Does this model say peacocks with longer tails fly slower?
air_speed = 4.3 − 1.2 · tail_length + 0.8 · muscle_mass

Photo by Thimindu Goonatillake CC BY-SA 2.0

1. Here's a model that we developed to describe the flying speed of male peacocks.
2. Notice the coefficient on the tail length variable: -1.2
3. Does this mean that peacocks with longer tails are slower flyers?
4. The answer is no - we have to remember ceteris paribus
5. The correct interpretation is that peacocks with1 unit longer tails, but the same muscle mass, are predicted to fly 1.2 units slower. But what about peacocks with longer tails as a whole? We just don't know. You can imagine that peacocks with longer tails also tend to have more muscle mass. Maybe they fly faster.

Categorical Inputs

Alex fills in this section

Alex fills in this section

Learnosity: Interpreting Model Weights

This is a placeholder for a Learnosity activity.

- have to change this, so it's about interpretation, not prediction
- In this activity, students will be given a fitted model that conforms with the data that they have for the peacock
- They will make predictions *first* from the data, and then
- Second, from newly created data to see how the predictions change

This is a placeholder for a Learnosity activity.

- have to change this, so it's about interpretation, not prediction
- In this activity, students will be given a fitted model that conforms with the data that they have for the peacock
- They will make predictions *first* from the data, and then
- Second, from newly created data to see how the predictions change

Part 2: Selecting a Linear Model with OLS

OLS is Regression for Estimating the BLP

OLS is Regression for Estimating the BLP

Fitting Linear Models

Linear regression: an algorithm for fitting a linear model given a sample of data

- Ordinary least squares (OLS) regression
- Quantile regression
- Regularized regression
  - Lasso
  - Ridge regression

Fitting Linear Models

Linear regression: an algorithm for fitting a linear model given a sample of data

- Ordinary least squares (OLS) regression
- Quantile regression
- Regularized regression
  - Lasso
  - Ridge regression

1. Where do linear models come from? Up till now, we've just made up numbers.
2. But what we really want to do is choose the model using data.
3. There are a lot of different algorithms that you can use to choose a linear model. We've listed a few famous ones here, but there are many options.
4. If you start with a different objective, you will design a different algorith..
5. Out of these, the most foundational type of regression is OLS, so what is it designed to do?

Understanding OLS

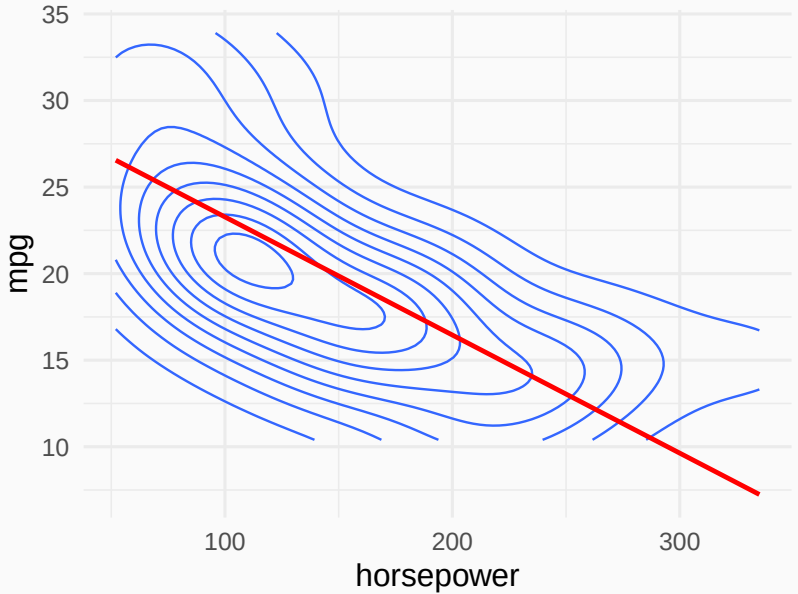
Ordinary least squares (OLS) regression

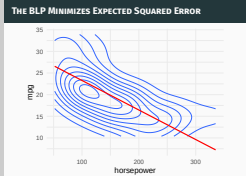
- The most well-known type of linear regression
- A foundation for many other types of regression
- Key goal: estimating the best linear predictor (BLP)

Understanding OLS
Ordinary least squares (OLS) regression
• The most well-known type of linear regression
• A foundation for many other types of regression
• Key goal: estimating the best linear predictor (BLP)

1. In this unit, our main topic is a type of regression called OLS
2. A key goal is estimating the BLP. Let's review what the BLP is and why it's so great.

The BLP Minimizes Expected Squared Error





The Best Linear Predictor

The BLP (population regression function)

The best linear predictor is defined by the function

$$g(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ are chosen to minimize the expected squared error.

$$\min_{(b_0,...,b_k)} \mathrm{E}\left[(Y - (b_0 + b_1 X_1 + ... + b_k X_k))^2\right]$$

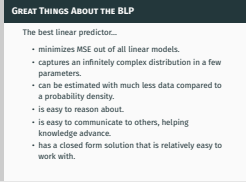
The Best Linear Predictor
The BLP (population regression function)
The best linear predictor is defined by the function
$g(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$
where $\beta = (\beta_0, \beta_1, ..., \beta_k)$ are chosen to minimize the expected squared error.
$\min_{(b_0,...,b_k)} \mathrm{E}\left[(Y - (b_0 + b_1 X_1 + ... + b_k X_k))^2\right]$

1. If we are working with a best linear predictor, then we know we have a function that must take the following form: [read the formula]
2. Remember the formula for the BLP in the single variable case; it is a more complex now – complex enough that we can't work through the statement, but it is just a bit more than $\beta = \frac{Cov[X,Y]}{Var[X]}$

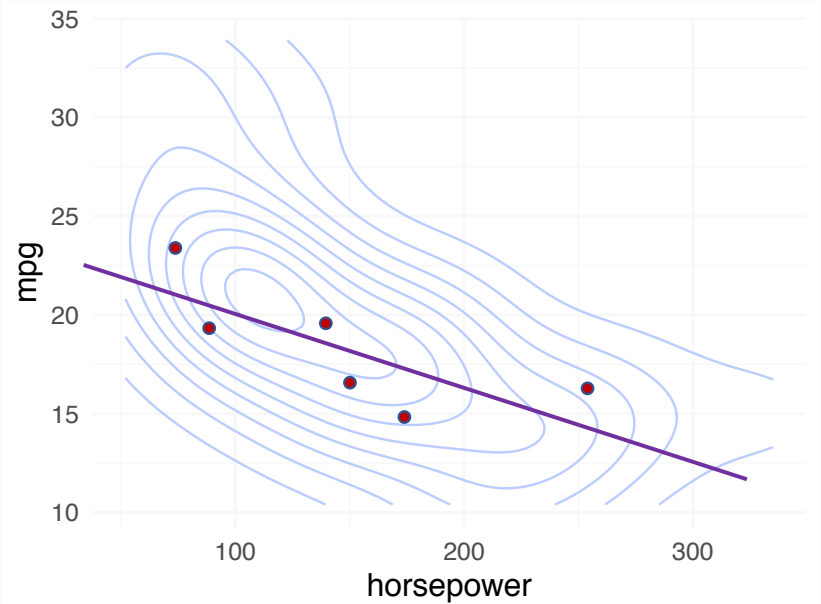
Great Things About the BLP

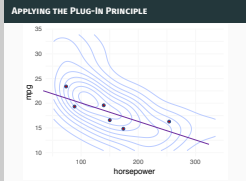
The best linear predictor...

- minimizes MSE out of all linear models.
- captures an infinitely complex distribution in a few parameters.
- can be estimated with much less data compared to a probability density.
- is easy to reason about.
- is easy to communicate to others, helping knowledge advance.
- has a closed form solution that is relatively easy to work with.



Applying the Plug-In Principle





1. We can't see the joint distribution, so we can't analyze the true error
2. But we do have this data, so can use a plug-in estimate
3. Instead of an expectation, we'll look at the distance to each data point and take an average
4. These distances are called residuals

OLS Regression Is the BLP Plug-In Estimator

Plug-in strategy

The OLS regression line is given by

$$\hat{g}(\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ are given by

$$\min_{(b_0,...,b_k)} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - (b_0 + b_1 X_{[1]i} + ... + b_k X_{[k]i})\right)^2.$$

OLS Regression Is the BLP Plug-In Estimator
Plug-in strategy
The OLS regression line is given by
$\hat{g}(\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$
where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ are given by
$\min_{(b_0,...,b_k)} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - (b_0 + b_1 X_{[1]i} + ... + b_k X_{[k]i})\right)^2.$

1. The general plug-in principle says that we write down the sample analogue for the population quantity that we're interested in. Lets do just that.
2. Basically, we had an expectation in the formula before, but now we replace that by taking an average over the n datapoints.
3. Instead of minimizing squared error, we are now minimizing squared residuals
4. The hope is that by doing this, we'll get a good estimator. But we still have to prove it. we'll do that soon.

An Analogy with the Mean

Given only Y	Given Y and X
E[Y] minimizes MSE out of all numbers.	the BLP minimizes MSE out of all linear models.
We can't compute E[Y] without knowing the distribution.	We can't compute the BLP without knowing the distribution.
$\bar{X}$ is the plug-in estimator for E[Y].	OLS is the plug-in estimator for the BLP.

An Analogy with the Mean

Given only Y	Given Y and X
E[Y] minimizes MSE out of all numbers.	the BLP minimizes MSE out of all linear models.
We can't compute E[Y] without knowing the distribution.	We can't compute the BLP without knowing the distribution.
$\bar{X}$ is the plug-in estimator for E[Y].	OLS is the plug-in estimator for the BLP.

Coming Up Soon...

We still have to:

- solve the minimization problem
- show that OLS is *consistent* for the BLP

Coming Up Soon...
We still have to:
- solve the minimization problem
- show that OLS is *consistent* for the BLP

1. Here's what we have to do.
2. First, we have to get the equation into a more useful form. We want to minimize the sum squared residuals. we have to actually solve that minimization problem and write down the solution.
3. Second, we have to show that it really has good statistical properties
4. Once we do all that, we'll have a great justification for the OLS algorithm. it is a consistent estimator for the model that minimizes MSE.

Learnosity: You Minimize It!

Learnosity: You Minimize It!

Learnosity: You Minimize It!

Note: This is a Learnosity Activity. We're just placing it here for organization.

This is the activity that is currently coded

`regression_fit_2d_exercise/`. **Note that we would like to expand this to ask students to work through several examples in a row. Presently we have a single example made; expanding this to a broader set is relatively easy. In the expanded set, we should ensure that we have some complexity in the data—e.g., a sine curve.**

Learnosity: You Minimize It!

Note: This is a Learnosity Activity. We're just placing it here for organization.

This is the activity that is currently coded

regression_fit_2d_exercise/. Note that we would like to expand this to ask students to work through several examples in a row. Presently we have a single example made; expanding this to a broader set is relatively easy. In the expanded set, we should ensure that we have some complexity in the data—e.g., a sine curve.

Reading: OLS Regression Estimates the BLP

Reading: Linear Regression is a Plug-in Estimator for the BLP

Note: this is a reading call, we're just placing it here for organization.

Read pages 143–147 of *Foundations of Agnostic Statistics.*

Reading: Linear Regression is a Plug-in Estimator for the BLP
Note: this is a reading call, we're just placing it here for organization.
Read pages 143–147 of *Foundations of Agnostic Statistics.*

Choosing Assumptions for OLS Regression

Choosing Assumptions for OLS Regression

When Does OLS "Work"?

The OLS regression line is

$$\hat{g}(\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

where $\hat{\beta}$ are chosen to minimize squared residuals.

When Does OLS "Work"?
The OLS regression line is
$\hat{g}(\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$
where $\hat{\beta}$ are chosen to minimize squared residuals.

1. You've already seen one definition of OLS regression.
2. OLS is just an algorithm. No matter what the situation is, you can always run it and get some numbers out. (It is a Golem!)
3. So you might want to know, "What assumptions are needed for OLS regression to work?" It's not so easy - there are different sets of assumptions out there.
4. Depending on which assumptions you choose, you get different guarrantees.
5. We're going to teach you two main sets of assumptions.

When Does OLS "Work"?

The OLS regression line is

$$\hat{g}(\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

where $\hat{\beta}$ are chosen to minimize squared residuals.

Different assumptions

$\Downarrow$

Different statistical guarantees

When Does OLS "Work"?
The OLS regression line is
$\hat{g}(\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$
where $\hat{\beta}$ are chosen to minimize squared residuals.
Different assumptions
$\Downarrow$
Different statistical guarantees

1. You've already seen one definition of OLS regression.
2. OLS is just an algorithm. No matter what the situation is, you can always run it and get some numbers out. (It is a Golem!)
3. So you might want to know, "What assumptions are needed for OLS regression to work?" It's not so easy - there are different sets of assumptions out there.
4. Depending on which assumptions you choose, you get different guarrantees.
5. We're going to teach you two main sets of assumptions.

More data $\implies$ less restrictive assumptions

More data $\implies$ easier to assess assumptions

1. **First, a very general point: the more data you have, the better off you are.**
2. **More data helps in two major ways.**
3. **First, we don't need as many assumptions with more data. That's because of asymptotics - statistics behave in predictable ways when $n \to \infty$**
4. **Second, having more data helps you to assess your assumptions. It's hard to defend an assumption of normality when you have 5 data points. if you have 100, you have an argument.**

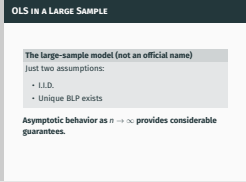
OLS in a Large Sample

The large-sample model (not an official name)

Just two assumptions:

- I.I.D.
- Unique BLP exists

Asymptotic behavior as $n \to \infty$ provides considerable guarantees.



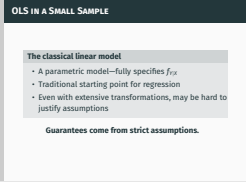
1. With a lot of data, $n \to \infty$, we need to believe very few assumptions for OLS to have good properties.
2. All we need is that the data is distributed i.i.d., and that there is a unique solution to the BLP.
3. If there isn't a unique solution, then we just can't solve for it!
4. When you package these two assumptions together, we'll call them the large-sample model.
5. That's not an official name, so don't use it outside of this class. But it will help us to give it a name.

OLS in a Small Sample

The classical linear model

- A parametric model—fully specifies $f_{Y|X}$
- Traditional starting point for regression
- Even with extensive transformations, may be hard to justify assumptions

Guarantees come from strict assumptions.



1. We want to draw your attention to the fact that we're presenting the large-sample (asymptotic) version of regression this week. In the future, we will talk about the smaller sample variant known as the *Classic Linear Model*.
2. Whereas we get *consistent* estimates with very minimal assumptions in the infinite data-case, we get a lot less in the more limited CLM case. It is kind of like buying a used car with no warranty.

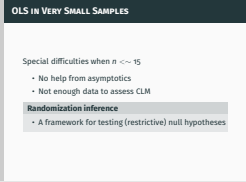
OLS in Very Small Samples

Special difficulties when $n <\sim 15$

- No help from asymptotics
- Not enough data to assess CLM

Randomization inference

- A framework for testing (restrictive) null hypotheses



1. I want to say a few words about very small samples.
2. very small isn't a technical phrase, but think $n \leq 15$.
3. First, should you *really* be doing data quantitative work with this much data? Is what you give up in richness worth it?
4. The problems really compound in this range.
5. First, you get no help from asymptotics, which means that you need to meet the CLM exactly.
6. Second, you don't have enough data to assess the CLM. even if things look ok, how can you convince someone when you only have 10 datapoints?
7. One suggestion: if you need to work with very small data, especially from experiments, you should read up on a field called randomization inference. It can give you a principled way to test hypotheses (at least special types of hypotheses)

Rules of Thumb for OLS Assumptions

Rules of thumb

In general, you might reason about data and regression models in the following way.

Sample size	Required assumptions
$100 \leq n$	Large-sample linear model
$15 \leq n < 100$	Classical linear model
$5 \leq n < 15$	Randomization inference

Rules of Thumb for OLS Assumptions
Rules of thumb
In general, you might reason about data and regression models in the following way.
Sample size | Required assumptions
$100 \leq n$ | Large-sample linear model
$15 \leq n < 100$ | Classical linear model
$5 \leq n < 15$ | Randomization inference

1. These are our 3 main frameworks. I want to put down some numbers to help you choose a framework.
2. These are ONLY rules of thumb.
3. First, you can consider the large-sample model if $n > 60$. That's really just a starting number.
4. It depends on how extreme the violations of the CLM are. In particular, if the error distribution has a really large skew, you may need $n > 100$ or $n > 1000$ or even more. So it's important for you to learn how to do diagnostics for the CLM, even if you think you have a large sample
5. from 15 to 60, we're recommending the CLM as your framework.
6. below 15, the CLM is no longer very credible. you can report your coefficients, and possibly use randomization inference to get p-values.

The Bivariate OLS Solution

The Bivariate OLS Solution

Bring in content from old version of course:

9.5 Deriving the Bivariate OLS Estimators

Alex, note that the notation isn't perfectly aligned. but I think on the balance, this is probably still better than the iPad scribble that would replace it.

Bring in content from old version of course:
9.5 Deriving the Bivariate OLS Estimators
Alex, note that the notation isn't perfectly aligned. but I think on the balance, this is probably still better than the iPad scribble that would replace it.

Consistency of Bivariate OLS Under the Large-Sample Model

Consistency of Bivariate OLS Under the Large-Sample Model

Continuous mapping theorem: Let $(S^{(1)}, S^{(2)}, S^{(3)}, \ldots)$ and $(T^{(1)}, T^{(2)}, T^{(3)}, \ldots)$ be sequences of random variables. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function that is continuous at $(a, b) \in \mathbb{R}^2$. If $S^{(n)} \xrightarrow{p} a$ and $T^{(n)} \xrightarrow{p} b$, then $g(S^{(n)}, T^{(n)}) \xrightarrow{p} g(a, b)$

Assumptions: 1) I.I.D. 2) Unique BLP exists ($V(X) > 0$)

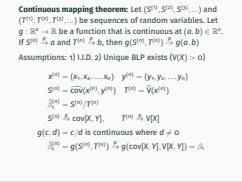
Continuous mapping theorem: Let $(S^{(1)}, S^{(2)}, S^{(3)}, \ldots)$ and $(T^{(1)}, T^{(2)}, T^{(3)}, \ldots)$ be sequences of random variables. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function that is continuous at $(a, b) \in \mathbb{R}^2$. If $S^{(n)} \xrightarrow{p} a$ and $T^{(n)} \xrightarrow{p} b$, then $g(S^{(n)}, T^{(n)}) \xrightarrow{p} g(a, b)$
Assumptions: 1) I.I.D. 2) Unique BLP exists ($V(X) > 0$)

1. In the last lightboard we derived that $\hat{\beta}_1 = \frac{\widehat{cov}(x,y)}{\widehat{V}(x)}$. This *looks* just like the formula that we use for the BLP!
2. Here, we'll prove that these estimates are consistent estimates in the large sample case.

Continuous mapping theorem: Let $(S^{(1)}, S^{(2)}, S^{(3)}, ...)$ and $(T^{(1)}, T^{(2)}, T^{(3)}, ...)$ be sequences of random variables. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function that is continuous at $(a, b) \in \mathbb{R}^2$. If $S^{(n)} \xrightarrow{p} a$ and $T^{(n)} \xrightarrow{p} b$, then $g(S^{(n)}, T^{(n)}) \xrightarrow{p} g(a, b)$

Assumptions: 1) I.I.D. 2) Unique BLP exists ($V(X) > 0$)

$$x^{(n)} = (x_1, x_2, ..., x_n) \quad y^{(n)} = (y_1, y_2, ..., y_n)$$
$$S^{(n)} = \widehat{\text{cov}}(x^{(n)}, y^{(n)}) \quad T^{(n)} = \widehat{V}(x^{(n)})$$
$$\hat{\beta}_1^{(n)} = S^{(n)}/T^{(n)}$$
$$S^{(n)} \xrightarrow{p} \text{cov}[X, Y], \qquad T^{(n)} \xrightarrow{p} V[X]$$
$$g(c, d) = c/d \text{ is continuous where } d \neq 0$$
$$\hat{\beta}_1^{(n)} = g(S^{(n)}, T^{(n)}) \xrightarrow{p} g(\text{cov}[X, Y], V[X, Y]) = \beta_1$$



1. In the last lightboard we derived that $\hat{\beta}_1 = \frac{\widehat{cov}(x,y)}{\widehat{V}(x)}$. This *looks* just like the formula that we use for the BLP!
2. Here, we'll prove that these estimates are consistent estimates in the large sample case.

The Matrix Formulation of a Linear Model

The Matrix Formulation of a Linear Model

- Insert content from previous version of course: 10.7 Matrix Form of the Linear Model
- This content leads into the next lightboard of the derivation of the OLS normal equations

The Matrix Formulation of a Linear Model

- Insert content from previous version of course: 10.7 Matrix Form of the Linear Model
- This content leads into the next lightboard of the derivation of the OLS normal equations

Reading: The Matrix Solution For OLS Regression

Reading: The Matrix Solution For OLS Regression

Reading: The Matrix Solution For OLS Regression

Read section 4.1.3, which is on pages 147 - 151.

Reading: The Matrix Solution For OLS Regression
Read section 4.1.3, which is on pages 147 - 151.

The Multiple OLS Solution

The Multiple OLS Solution

The Multiple OLS Solution

- Pull in the lightboard called *Matrix Derivation of the OLS Estimator*.
- In the next iteration of the course, pull in a geometric derivation of the OLS coefficients.

The Multiple OLS Solution

- Pull in the lightboard called *Matrix Derivation of the OLS Estimator*.
- In the next iteration of the course, pull in a geometric derivation of the OLS coefficients.

Sample Moment Conditions

Sample Moment Conditions

Review: Population Moment Conditions

Population: Let ϵ represent error from the BLP.

Version 1: $E[\epsilon] = 0, E[X_j \epsilon] = 0$ for all j.

Version 2: $E[\epsilon] = 0, \text{cov}[X_j, \epsilon] = 0$ for all j.

Review: Population Moment Conditions
Population: Let ϵ represent error from the BLP.
Version 1: $E[\epsilon] = 0, E[X_j \epsilon] = 0$ for all j.
Version 2: $E[\epsilon] = 0, \text{cov}[X_j, \epsilon] = 0$ for all j.

Sample Moment Conditions

Sample: Let $\boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$ represent OLS residuals.

Sample Moment Conditions
Sample: Let $\boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$ represent OLS residuals.

Sample Moment Conditions

Sample: Let $\boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$ represent OLS residuals.

$$\mathbb{X}^T\boldsymbol{Y} = \mathbb{X}^T\mathbb{X}\beta, \qquad 0 = \mathbb{X}^T(\boldsymbol{Y} - \mathbb{X}\beta) = \mathbb{X}^T\boldsymbol{e}$$

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{[1]1} & X_{[1]2} & \dots & X_{[1]n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{[k]1} & X_{[k]2} & \dots & X_{[k]n} \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$$\sum e_i = 0, \sum X_{[j]i}e_i = 0. \text{ or } \widehat{\text{cov}}(\boldsymbol{X}_{[j]}, \boldsymbol{e}) = 0$$

Sample Moment Conditions
Sample: Let $\boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$ represent OLS residuals.
$\mathbb{X}^T\boldsymbol{Y} = \mathbb{X}^T\mathbb{X}\beta, \quad 0 = \mathbb{X}^T(\boldsymbol{Y} - \mathbb{X}\beta) = \mathbb{X}^T\boldsymbol{e}$
$\begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{[1]1} & X_{[1]2} & \dots & X_{[1]n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{[k]1} & X_{[k]2} & \dots & X_{[k]n} \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$
$\sum e_i = 0, \sum X_{[j]i}e_i = 0.$ or $\widehat{\text{cov}}(\boldsymbol{X}_{[j]}, \boldsymbol{e}) = 0$

Consistency of Multiple OLS

Consistency of Multiple OLS

Consistency of Multiple OLS

Assumptions: 1) I.I.D. 2) Unique BLP exists

In population: $\beta = \mathrm{E}[X^T X]^{-1}\mathrm{E}[X^T Y]$

Consistency of Multiple OLS
Assumptions: 1) I.I.D. 2) Unique BLP exists
In population: $\beta = \mathrm{E}[X^T X]^{-1}\mathrm{E}[X^T Y]$

1. To apply the continuous mapping theorem, we need to know that $f(A, B) = A^{-1}B$ is continuous. unique BLP tells us that $E[\mathbb{X}^T\mathbb{X}]$ is invertible. each element of $f(A, B)$ is

Consistency of Multiple OLS

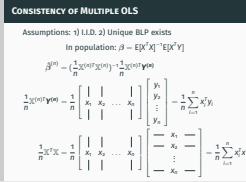
Assumptions: 1) I.I.D. 2) Unique BLP exists

In population: $\beta = E[X^T X]^{-1} E[X^T Y]$

$$\hat{\beta}^{(n)} = (\frac{1}{n}\mathbb{X}^{(n)T}\mathbb{X}^{(n)})^{-1}\frac{1}{n}\mathbb{X}^{(n)T}\boldsymbol{Y^{(n)}}$$

$$\frac{1}{n}\mathbb{X}^{(n)T}\boldsymbol{Y^{(n)}} = \frac{1}{n}\begin{bmatrix} | & | & & | \\ x_1 & x_2 & \dots & x_n \\ | & | & & | \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} x_i^T y_i$$

$$\frac{1}{n}\mathbb{X}^T\mathbb{X} = \frac{1}{n}\begin{bmatrix} | & | & & | \\ x_1 & x_2 & \dots & x_n \\ | & | & & | \end{bmatrix}\begin{bmatrix} — & x_1 & — \\ — & x_2 & — \\ & \vdots & \\ — & x_n & — \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} x_i^T x$$



1. Is this useful, or should we just do the bivariate case? it's a lot of content for one lightboard and I think it's too hard to actually prove continuity for the CMT
2. Maybe just call this a sketch, but show key part about turning the matrix products into sample means
3. Can write $\boldsymbol{Y} = \mathbb{X}\beta + \epsilon$ where $E(X^T\epsilon) = 0$
4. First, here's a useful way to write our vector of coefficients. as the true parameter plus a random error term.
5. Next, we have to open up these matrices and think about what's inside. Let the rows of $\mathbb{X}$ be $x_1, \dots, x_n$
6. To apply the continuous mapping theorem, we need to know that $f(A, B) = A^{-1}B$ is continuous. unique BLP tells us that $E[\mathbb{X}^T\mathbb{X}]$ is invertible. each element of $f(A, B)$ is

Consistency of OLS

$$\text{WLLN} \implies \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i \xrightarrow{p} \mathrm{E}[X^T X] \qquad \frac{1}{n}\sum x_i^T y_i \xrightarrow{p} \mathrm{E}[X^T Y]$$

$$\text{CMT} \implies \hat{\boldsymbol{\beta}} = \beta + (\frac{1}{n}\sum_{i=1}^{n} x_i^T x_i)^{-1}(\frac{1}{n}\mathbb{X}^T \epsilon) \xrightarrow{p} \beta + \mathrm{o} = \beta$$

Consistency of OLS
WLLN $\implies \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i \xrightarrow{p} \mathrm{E}[X^T X] \qquad \frac{1}{n}\sum x_i^T y_i \xrightarrow{p} \mathrm{E}[X^T Y]$
CMT $\implies \hat{\boldsymbol{\beta}} = \beta + (\frac{1}{n}\sum_{i=1}^{n} x_i^T x_i)^{-1}(\frac{1}{n}\mathbb{X}^T \epsilon) \xrightarrow{p} \beta + \mathrm{o} = \beta$

Unique Variation and Regression Anatomy

How Can We Understand a Specific $\hat{\beta}_i$?

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y}$$

How Can We Understand a Specific $\hat{\beta}_i$?
$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}$

1. We've written before that OLS is a *linear* regression. As a result of this, we showed that $\frac{\partial Y}{\partial X_i} = \beta_i \Delta \cdot X_i$.
2. Another, very interesting consequence of the underlying geometry of OLS regression is that each of the coefficients are fitted only on the *unique* variation in Y.
3. Agrist and Pichke (2009) term this the *Regression Anatomy* formula.
4. You've seen the matrix solution for ols regression, so you can now go and compute ols coefficients.
5. But what if you want to understand a specific coefficient? what makes it go up and down? You could invert the matrix, but that's really complicated
6. You might wonder, can you write down a formula for a single coefficient, that will help you understand what's going on.

Partialling Out

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k$$

Step 1: Regress X_1 on other Xs

$$\hat{X}_1 = \hat{\delta}_0 + \hat{\delta}_2 X_2 + ... + \hat{\delta}_k X_k + r_1$$

Step 2: Regress Y on the residuals from Step 1

$$\hat{Y} = \hat{\gamma}_0 + \hat{\beta}_1 r_1$$

Regression anatomy: $\hat{\beta}_1 = \frac{\widehat{\text{cov}}(Y, r_1)}{\widehat{\text{V}}(r_1)}$

Partialling Out
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k$
Step 1: Regress X_1 on other Xs
$\hat{X}_1 = \hat{\delta}_0 + \hat{\delta}_2 X_2 + ... + \hat{\delta}_k X_k + r_1$
Step 2: Regress Y on the residuals from Step 1
$\hat{Y} = \hat{\gamma}_0 + \hat{\beta}_1 r_1$
Regression anatomy: $\hat{\beta}_1 = \frac{\widehat{\text{cov}}(Y, r_1)}{\widehat{\text{V}}(r_1)}$

1. Let's say we're interested in computing β_1
2. It turns out that we can compute it in stages:
3. First, regress X_1 on the other X's (we can regress any var on any other vars)
4. We have some residuals from that regression, r_1. those represent the unique variation in X_1 - the part not explainable by other variables.
5. Next, we take those residuals and regress Y on them.
6. It turns out that we get the same β_1. We'll prove that soon
7. This is a powerful idea. ols works on unique variation in each X, variation that is collinear with the other variables doesn't contribute, you may as well subtract it out. As we'll see later, the more unique variation we have in X, the more precision we'll have.

Deriving the Regression Anatomy Formula

Deriving the Regression Anatomy Formula

Deriving the Regression Anatomy Formula

Use content from old course:

10.5 Regression Anatomy

Deriving the Regression Anatomy Formula
Use content from old course:
10.5 Regression Anatomy

Segment for Consideration: Applying the Regression Anatomy Formula

Segment for Consideration: Applying the Regression Anatomy Formula

Applying the Regression Anatomy Formula

Consider using the old concept check: 10.6 Applying the Regression Anatomy Formula

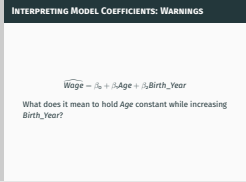
Applying the Regression Anatomy Formula

Consider using the old concept check: 10.6 Applying the Regression Anatomy Formula

Interpreting Model Coefficients: Warnings

$$\widehat{Wage} = \beta_0 + \beta_1 Age + \beta_2 Birth\_Year$$

What does it mean to hold *Age* constant while increasing *Birth_Year*?



1. Ceteris paribus hints at some of the care we need when putting variables into a linear model.
2. Here's a model for wage, with two variables, Age and Birth Year, how do you interpret β_1?
3. You can fit a model like this (many do), but is it telling you something about the joint distribution?
4. Alex, this slide is a bit of a struggle - my current suggestion is to postpone it till late in the unit. We haven't mentioned fitting yet, so can't really say things like "picking up on noise" or multicollinearity. What I think we can address here feels like a small point.

Evaluating the Large-Sample Linear Model

Evaluating the Large-Sample Linear Model

The Large-Sample Linear Model

- I.I.D. data
- A unique BLP exists

The Large-Sample Linear Model
- I.I.D. data
- A unique BLP exists

1. The great thing about regression with large samples, is that you don't need a lot of assumptions.
2. A lot of books heavily stress the CLM, and people get the impression that regression takes a lot of assumptions
3. Actually, if you're seeing this for the first time, you might be really surprised that we are doing everything with just two assumptions, which we are calling the large sample linear model.
4. But we do need these two assumptions, so it's important that we stop to assess them.
5. We can't just blindly assume they're true. we have to look at the situation we're modeling and ask, are they plausible? or are they realistic.

What Does I.I.D. Mean?

Imagine selecting each new datapoint...

- from the same distribution
- with no memory of any past datapoints

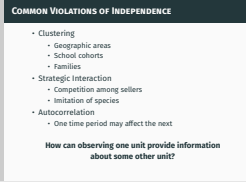
What Does I.I.D. Mean?
Imagine selecting each new datapoint...
• from the same distribution
• with no memory of any past datapoints

1. This is a model, and we have to look at the real world and see how much it resembles this model

Common Violations of Independence

- Clustering
  - Geographic areas
  - School cohorts
  - Families
- Strategic Interaction
  - Competition among sellers
  - Imitation of species
- Autocorrelation
  - One time period may affect the next

How can observing one unit provide information about some other unit?



1. It helps to know some common violations so that you can watch out for them
2. In general, the question is how can observing one unit give information about some other unit.
3. For some of these violations, you can devise a statistical test. But there's no test can can detect an arbitrary network of dependencies. To assess this assumption you really need to use your background knowledge.

A Unique BLP Exists

A BLP exists:

- $\text{cov}[X_i, X_j]$ and $\text{cov}[X_i, Y]$ are finite (no heavy tails)

The BLP is unique:

- No perfect collinearity
- $E[X^T X]$ is invertible

$\implies$ No X_i can be written as a linear combination of the other X's.

A Unique BLP Exists

A BLP exists:

- $\text{cov}[X_i, X_j]$ and $\text{cov}[X_i, Y]$ are finite (no heavy tails)

The BLP is unique:

- No perfect collinearity
- $E[X^T X]$ is invertible

$\implies$ No X_i can be written as a linear combination of the other X's.

1. What does it mean for the BLP to exist? we need all the covariances in the solution to be finite. that basically means that the random variables can't have tails that are too heavy.
2. This is usually true. actually, some textbooks don't even mention this assumption.
3. I think it's good to mention it, because researchers do believe that some variables have heavy tails: wealth is one example. air emissions. financial returns...
4. What does this mean? Each X has to have some unique variation.
5. That makes sense because we know ols works on unique variation. If we combine this assumption with the previous one, we can prove that a unique BLP exists.

Perfect Collinearity Example 1

$$\widehat{Price} = .5\ Donuts + 0.0\ Dozens$$

or

$$\widehat{Price} = 0.0\ Donuts + 6.0\ Dozens$$

Perfect Collinearity Example 1
$\widehat{Price} = .5\ Donuts + 0.0\ Dozens$
or
$\widehat{Price} = 0.0\ Donuts + 6.0\ Dozens$

1. Here's an example to help understand why we need this assumption
2. You regress the price on both number of donuts and number of dozens of donuts.
3. it's 50 cents per donut, so you can write the price as 50 cents times number of donuts
4. But you can also write it as 6 dollars times number of dozens.
5. Both are equivalent for predicting the price - this problem doesn't affect prediction.
6. But you can't estimate coefficients, because they aren't uniquely defined.

Perfect Collinearity Example 2

$$\widehat{Voters} = 200\ Positive\_Ads + 100\ Negative\_Ads + 0\ Total\_Ads$$

or

$$\widehat{Voters} = 100\ Positive\_Ads + 0\ Negative\_Ads + 100\ Total\_Ads$$

Perfect Collinearity Example 2
$\widehat{Voters} = 200\ Positive\_Ads + 100\ Negative\_Ads + 0\ Total\_Ads$
or
$\widehat{Voters} = 100\ Positive\_Ads + 0\ Negative\_Ads + 100\ Total\_Ads$

1. Here's another example, to show you that multicollinearity isn't always about pairs of variables, it might be more variables
2. Here you have a regression with number of positive ads, number of negative ads, and the total number of ads.
3. Once again, you can find multiple ways to write the same model.

Goodness of Fit

Goodness of Fit

Goodness of fit: How well does a model fit the data?

- R^2
- Adjusted R^2
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)

Goodness of fit: How well does a model fit the data?

- R^2
- Adjusted R^2
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)

1. We've seen how to compute regression coefficients, and you know a little about interpreting them.
2. But we sometimes you also want to step back and assess the model as a whole.
3. One question you might ask is this: How well does a model fit the data?
4. I know that's a fuzzy question, but it seems important. Is there a really close match between model and data or are they far apart?
5. Can we come up with a single number to capture that idea?
6. There are a number of statistics for that purpose, and we can call them measures of fit.
7. The most famous one is R^2. you've probably heard of it. but exactly what does it measure? and how do you use it properly?

Breaking Down Variance

Total variance = explained variance + residual variance

Breaking Down Variance
Total variance = explained variance + residual variance

1. R^2 is about breaking down the variance in the outcome variable.
2. This may remind you of the law of total variance, but it's a different decomposition, this only works for OLS
3. $\hat{V}[Y] = \hat{V}[\hat{Y} + \hat{\epsilon}] = \hat{V}[\hat{Y}] + \hat{V}[\hat{\epsilon}] + 2\widehat{cov}[\hat{Y}, \hat{\epsilon}]$
4. $\widehat{cov}[\hat{Y}, \hat{\epsilon}] = \widehat{cov}[\beta_0 + \beta_1 X_1 + ... + \beta_k X_k, \hat{\epsilon}] = 0$
5. One technical note: these are regular sample variances - so the denominator is n-1 everywhere. I'm saying that because when people say "residual variance" it usually includes a correction for the number of coefficients. If you do that, you get something called adjusted R squared. I want to talk about regular R squared so no corrections.

Defining R^2

$$R^2 = 1 - \frac{\hat{\mathbb{V}}(\hat{\epsilon})}{\hat{\mathbb{V}}(\boldsymbol{Y})} = 1 - \frac{\text{residual variance}}{\text{total variance}}$$

$$\text{For OLS: } R^2 = \frac{\text{explained variance}}{\text{total variance}}$$

How much of the variation in the outcome does the model explain?

Defining R^2

$R^2 = 1 - \frac{\hat{\mathbb{V}}(\hat{\epsilon})}{\hat{\mathbb{V}}(\boldsymbol{Y})} = 1 - \frac{\text{residual variance}}{\text{total variance}}$

For OLS: $R^2 = \frac{\text{explained variance}}{\text{total variance}}$

How much of the variation in the outcome does the model explain?

1. Now that you understand these components of variance, here's the definition of R^2.
2. R^2 is a number between 0 and 1. it's the fraction of the variation in the outcome that's explained by the model.
3. You can think about variance as information. The independent variables are giving you some information about the outcome. If you could predict the outcome perfectly, R^2 would be 1. but of course, the predictions usually don't match the outcomes, so R^2 is lower.

R is Correlation

$$R^2 = \frac{\hat{\mathbb{V}}(\hat{\boldsymbol{Y}})}{\hat{\mathbb{V}}(\boldsymbol{Y})}$$

R is Correlation
$R^2 = \frac{\hat{\mathbb{V}}(\hat{\boldsymbol{Y}})}{\hat{\mathbb{V}}(\boldsymbol{Y})}$

1. Here's another way to look at R^2. The R is a correlation r. For ols regression, you could really make it a lowercase r.
2. I can rewrite the numerator:

$$\begin{aligned}\hat{V}(\hat{Y}) &= \widehat{cov}(\hat{Y}, \hat{Y}) = \widehat{cov}(\hat{Y}, Y - \hat{\epsilon}) \\ &= \widehat{cov}(\hat{Y}, Y) - \widehat{cov}(\hat{Y}, \hat{\epsilon}) = \widehat{cov}(\hat{Y}, Y)\end{aligned}$$

3. $$R^2 = \frac{\hat{V}(\hat{Y})\hat{V}(\hat{Y})}{\hat{V}(Y)\hat{V}(\hat{Y})} = \frac{\widehat{cov}(\hat{Y}, Y)^2}{\hat{V}(Y)\hat{V}(\hat{Y})} = \widehat{corr}(\hat{Y}, Y)^2$$

4. So we learned that R^2 is also the squared correlation between our predicted and the actual outcomes. It's a measure of agreement between the model predictions and the data.

Understanding Sums of Squares

Total sum of squares: $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

Explained sum of squares: $\text{ESS} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Residual sum of squares: $\text{RSS} = \sum_{i=1}^{n}(\hat{\epsilon}_i)^2$

$$\text{For OLS: TSS} = \text{ESS} + \text{RSS}$$

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Understanding Sums of Squares
Total sum of squares: TSS = $\sum_{i=1}^{n}(y_i - \bar{y})^2$
Explained sum of squares: ESS = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
Residual sum of squares: RSS = $\sum_{i=1}^{n}(\hat{\epsilon}_i)^2$
For OLS: TSS = ESS + RSS
$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$

1. I want to warn you that when you read about R-squared, the formula is usually different.
2. Instead of variances, most people learn about these things called sums of squares

Things to Remember About R^2

- Adding variables always makes R^2 go up.
- With many variables, consider alternatives.
  - Adjusted R^2
- R^2 is not a measure of practical significance.
  - For example, regress hospital admissions on being shot
- A low R^2 is a negative, but assess it in context.

Things to Remember About R^2

- Adding variables always makes R^2 go up.
- With many variables, consider alternatives.
  - Adjusted R^2
- R^2 is not a measure of practical significance.
  - For example, regress hospital admissions on being shot
- A low R^2 is a negative, but assess it in context.

1. First, adding variables makes R^2 go up. this is because OLS is essentially an R^2 maximizing machine. you give it more variables and it can't do any worse.
2. If you have a lot of variables, you might want to consider alternatives. These are measures of fit that penalize extra variables.
3. For example, if you are trying to model the purchasing decision that someone makes, there are *all* kinds of things that affect that choice. Having a low R2 on a behavioral outcome is pretty common.
4. This is a general test – if you're including a lot of new model features, you would expect R2 to increase – but this could just happen as a result of chance!
5. The real concern is one of *overfitting* where your model is

Part 3: Measuring Uncertainty

Review: Review: Statistics, Distributions, and Uncertainty

Reading: Robust Standard Errors

Reading: Robust Standard Errors

Read section 4.2 Inference, stopping at the bottom of page 153.

Reading: Robust Standard Errors

Read section 4.2 Inference, stopping at the bottom of page 153.

Robust Standard Errors

Robust Standard Errors

You might want to split this segment into 2: one show the equation for $V[\beta_j]$, pointing out unique variation, error variance, etc, then another segment to talk about estimation with robust standard errors.

You might want to split this segment into 2: one show the equation for $V[\beta_j]$, pointing out unique variation, error variance, etc, then another segment to talk about estimation with robust standard errors.

Robust Standard Errors

Note: Lightboard of how changes in data shape increase and decrease the errors of regression

- More data
- More variance in X
- More unique variance in X
- Possible demo of bootstrapping

Robust Standard Errors

Note: Lightboard of how changes in data shape increase and decrease the errors of regression

- More data
- More variance in X
- More unique variance in X
- Possible demo of bootstrapping

Hypothesis Tests

Hypothesis Tests

Testing Improvement in a Model

Testing Improvements in a Model: The F-Test

Testing Improvements in a Model: The F-Test

Testing Individual Coefficients

Testing Individual Coefficients

Hypothesis Tests for OLS

Are the relationships we see in our regression true of the population, or just consequences of sampling variation?

Most common tests:

- Testing single coefficients
  - Usually $H_0 : \beta_i = 0$
- Testing multiple coefficients
  - Usually $H_0 : \beta_i = \beta_j = ... = 0$

Hypothesis Tests for OLS
Are the relationships we see in our regression true of the population, or just consequences of sampling variation?
Most common tests:
- Testing single coefficients
  - Usually $H_0 : \beta_i = 0$
- Testing multiple coefficients
  - Usually $H_0 : \beta_i = \beta_j = ... = 0$

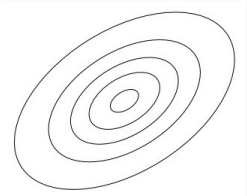
1. First, we often test single coefficients. We are usually interested in whether there is really a relationship between a var and the outcome. So we set $H_0 : \beta = 0$.
2. We hope to reject the null, which gives us evidence that there is a relationship. Lets us tell an interesting story
3. But remember, if you only report positive results, that's called publication bias. It can contribute to false discoveries and makes it harder to correct false discoveries. so for this class, negative results are just as interesting as positive results
4. The other common type of test is of multiple coefficients. For now, let's look at single coefficients.

Asymptotic Normality of OLS

Theorem

When the data points are I.I.D. from a distribution with unique BLP. $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a multivariate normal distribution.

- Each β_i is asymptotically normal.
- Linear combinations (e.g., $\beta_i - \beta_j$) are asymptotically normal.



Asymptotic Normality of OLS
Theorem
When the data points are I.I.D. from a distribution with unique BLP. $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a multivariate normal distribution.
- Each β_i is asymptotically normal.
- Linear combinations (e.g., $\beta_i - \beta_j$) are asymptotically normal.

1. Think of this as the CLT for ols coefficients. You can derive it from the CLT.
2. Remember that we know how to consistently estimate the variance of each coefficient (and covariances), so asymptotically we know the exact distribution better and better.

Testing a Single OLS Coefficient

Large-sample testing procedure

Let s_i be the robust standard error estimate for β_i.
Then, under $H_0 : \beta_i = \mu_0$,

$$z = \frac{\hat{\beta}_i - \mu_0}{s_i}$$

is asymptotically distributed $N(0, 1)$.

- Computed automatically in statistical software

Testing a Single OLS Coefficient
Large-sample testing procedure
Let s_i be the robust standard error estimate for β_i.
Then, under $H_0 : \beta_i = \mu_0$,
$z = \frac{\hat{\beta}_i - \mu_0}{s_i}$
is asymptotically distributed $N(0, 1)$.
- Computed automatically in statistical software

The t-Test for OLS Coefficients

In practice, we call the statistic t and test against a T-distribution.

- Asymptotically, the Z and T distributions are equal.
- Under the classical linear model assumptions, the T distribution is *exact* for small samples.

The t-Test for OLS Coefficients

In practice, we call the statistic t and test against a T-distribution.

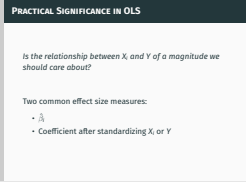
- Asymptotically, the Z and T distributions are equal.
- Under the classical linear model assumptions, the T distribution is exact for small samples.

Practical Significance in OLS

Is the relationship between X_i and Y of a magnitude we should care about?

Two common effect size measures:

- $\hat{\beta}_i$
- Coefficient after standardizing X_i or Y



1. As we've told you, when you conduct a test, it's not enough to report on statistical significance. it's very important to assess the practical significance of your result. How do you do that in ols?
2. Really there are two common effect size measures that you can focus on.
3. First, the coefficient itself! betas are effect size measures.
4. That's great when the units are understandable
5. If the units are confusing, you can still use the coefficients, but you will want to standardize your X or Y variable first.

Practical Significance Example 1

$$\widehat{Price} = 82,213 + 25,134\ Bedrooms + 8\ Gnomes$$

"Garden gnomes have a statistically significant relationship with house price ($t = 2.1$, $p = .03$)."

Practical Significance Example 1
$\widehat{Price} = 82,213 + 25,134\ Bedrooms + 8\ Gnomes$
"Garden gnomes have a statistically significant relationship with house price ($t = 2.1$, $p = .03$)."

Practical Significance Example 1

$$\widehat{Price} = 82,213 + 25,134\ Bedrooms + 8\ Gnomes$$

"Garden gnomes have a statistically significant relationship with house price ($t = 2.1$, $p = .03$)."

"The predicted price only changes by $8 per garden gnome, a negligible amount compared to the total house price. For comparison, the model predicts that it would take over 3,000 gnomes to match the effect of a single bedroom."

Practical Significance Example 1

$\widehat{Price} = 82,213 + 25,134\ Bedrooms + 8\ Gnomes$

"Garden gnomes have a statistically significant relationship with house price ($t = 2.1$, $p = .03$)."

"The predicted price only changes by $8 per garden gnome, a negligible amount compared to the total house price. For comparison, the model predicts that it would take over 3,000 gnomes to match the effect of a single bedroom."

Practical Significance Example 2

$$\widehat{Agility} = 132 + 3.4\ Sleep\_Hours$$

"We found evidence that more sleep is related to a higher agility score ($t = 3.8$, $p < .001$)."

"Each extra hour of sleep is associated with an extra 3.4 points."

Practical Significance Example 2

$\widehat{Agility} = 132 + 3.4\ Sleep\_Hours$

"We found evidence that more sleep is related to a higher agility score ($t = 3.8$, $p < .001$)."

"Each extra hour of sleep is associated with an extra 3.4 points."

1. What's wrong with this discussion?
2. The problem is we have no idea what 3.4 agility points means.
3. In this case, a good idea is to first standardize the agility score.

Practical Significance Example 2 (cont.)

Let $S\_Agility = \frac{Agility - \overline{Agility}}{\sqrt{\widehat{var}(Agility)}}$

$$\widehat{S\_Agility} = -1.21 + 0.92\ Sleep\_Hours$$

"Each extra hour of sleep is predicted to increase agility by 0.92 standard deviations."

Practical Significance Example 2 (cont.)
Let $S\_Agility = \frac{Agility - \overline{Agility}}{\sqrt{\widehat{var}(Agility)}}$
$\widehat{S\_Agility} = -1.21 + 0.92\ Sleep\_Hours$
"Each extra hour of sleep is predicted to increase agility by 0.92 standard deviations."

1. We standardize by subtracting the mean and then dividing by the standard deviation.
2. Now the coefficient can be interpreted in terms of standard deviations. Let's read..
3. Of course you should try to add more context: is this a study for firefighters, for a wellness magazine? Remember that the larger goal is to communicate whether the effect is something that we should care about.

Testing Multiple Coefficients

Testing Multiple Coefficients

Testing Multiple Coefficients

	Dependent Variable: Crimes per 1000
Density	8.414***
	(1.140)
Federal Wage	0.027
	(0.030)
Service Wage	−0.008
	(0.007)
Manufacturing wage	0.003
	(0.019)
Intercept	10.768
	(11.964)

Testing Multiple Coefficients

	Dependent Variable: Crimes per 1000
Density	8.414*** (1.140)
Federal Wage	0.027 (0.030)
Service Wage	−0.008 (0.007)
Manufacturing wage	0.003 (0.019)
Intercept	10.768 (11.964)

1. Here's an example regression table, We're predicting the crime rate.
2. This is somewhat old data, but it's for a set of counties in North Carolina, in 1987
3. For our predictors, we have the density, and then a set of wage variables.
4. Only the density is statistically significant. is the relationship with wages really that small?
5. Well there's a potential problem here. These variables are closely related. there is a lot of multicollinearity, so the standard errors are high. That might be the reason they are non significant.

Testing Joint Significance

Full model: $Crime = \widehat{\beta_0} + \hat{\beta}_1 Density + \hat{\beta}_2 Fed\_Wage + \hat{\beta}_3 Ser\_Wage + \hat{\beta}_4 Man\_Wage + \hat{\epsilon}_f$

Restricted model: $Crime = \hat{\beta}_0 + \hat{\beta}_1 Density + \hat{\epsilon}_r$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$

Idea: if H_0 is true, the full model should not be much better at explaining the outcome

Testing Joint Significance
Full model: $Crime = \widehat{\beta_0} + \hat{\beta}_1 Density + \hat{\beta}_2 Fed\_Wage + \hat{\beta}_3 Ser\_Wage + \hat{\beta}_4 Man\_Wage + \hat{\epsilon}_f$
Restricted model: $Crime = \hat{\beta}_0 + \hat{\beta}_1 Density + \hat{\epsilon}_r$
$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$
Idea: if H_0 is true, the full model should not be much better at explaining the outcome

Total variance = explained variance + error variance

$$f = \frac{df_f}{df_r - df_f} \frac{\mathrm{V}[\epsilon_r] - \mathrm{V}[\epsilon_f]}{\mathrm{V}[\epsilon_f]}$$

Total variance = explained variance + error variance

$f = \frac{df_f}{df_r - df_f} \frac{\mathrm{V}[\epsilon_r] - \mathrm{V}[\epsilon_f]}{\mathrm{V}[\epsilon_f]}$

1. Let me remind you that we can break down the variance in the outcome like this. the smaller the error variance, the more the model explains
2. If the null hypothesis is true, we expect the two models to have similar error variance.
3. We make a test statistic by taking a fraction... again if the null is true, the numerator should be small
4. There is a degrees of freedom adjustment, it's just a constant, so not as important now
5. It turns out that this statistics has exactly an F-distribution under the null, so we can use it to test.
6. A big F statistic means that the full model is predicting the outcome much better using the extra variables. so we're more likely to reject the null.

F-Test Results

	RSS	Df	Sum of Sq	F	Pr(>F)
Full	14526				
Wages	14924	-3	-397.57	0.7846	0.5057

F-Test Results

	RSS	Df	Sum of Sq	F	Pr(>F)
Full	14526				
Wages	14924	-3	-397.57	0.7846	0.5057

1. Here you can see the results
2. This is a style called an Anova table.
3. A lot of people assume Anova is a different technique - it's not. an Anova is always based on a linear model. But anova means that you'll report F-tests to explain what different factors are doing.
4. In this case, you can see that our p-value is .51. so not at all significant. the problem was not multicollinearity, or not just multicollinearity. We can see that these variables actually do very little to explain the crime rate.

Making Predictions

Making Predictions