Topic 3: Simple Linear Regression

ECON 4753 — University of Arkansas

Prof. Kyle Butts

Fall 2024

Roadmap

- Bivariate Regression
  - Prediction vs Causation
- Regression Inference
- Goodness of Fit
- Influential Observations
- Discrete Variables
- $\log$ transformations

Covariance and Correlation

Recall the ways we discussed relationships between two random variables X and Y:

Covariance, σ_{XY} (sample analogue: s_{XY})

- Direction matters, but magnitude is hard to interpret

Correlation, ρ_{XY} (sample analogue: r_{XY})

- Direction and magnitude matter
- Correlation is always value between $[-1, 1]$

Covariance and Correlation

The correlation is calculated as

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}} \tag{1}$$

- Correlation is a function of covariance, just normalizes the magnitudes so we can interpret.

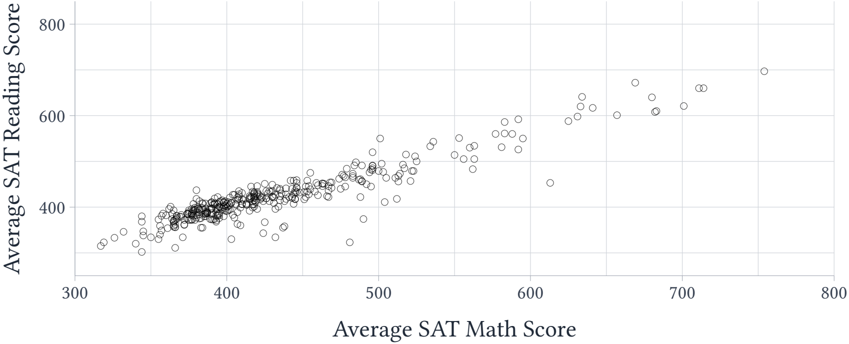
Practice question

Suppose you calculate the sample covariance, $s_{XY} = 1.2$, and the sample standard deviations $s_X = 2$ and $s_Y = 2.5$. What is the sample correlation, r_{XY}?

- 0.0576
- 0.24
- 0.048
- 4.17

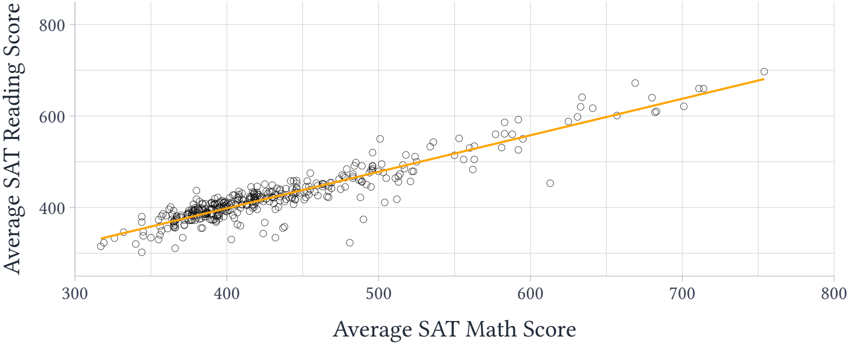
Relationship between X and Y

Consider this plot of NYC Math and Reading SAT Scores



Relationship between X and Y

Consider this plot of NYC Math and Reading SAT Scores. The easiest way to summarize the relationship between X and Y is using a regression line, aka the "line of best fit".



Regression line

We can write this linear model as

$$y = f(X) + \varepsilon = \beta_0 + \beta_1 * X + \varepsilon$$

The model says y is a linear function of X. β_0 is the ‘intercept’ and β_1 is the ‘slope’ of the line.

Regression line

We can write this linear model as

$$y = f(X) + \varepsilon = \beta_0 + \beta_1 * X + \varepsilon$$

The model says y is a linear function of X. β_0 is the 'intercept' and β_1 is the 'slope' of the line.

We use the following terminology:

- y is called 'the dependent variable', 'the response variable', or 'the predicted variable'
- X is called 'the independent variable', 'the explanatory variable', 'the control variable', or 'the predictor variable'

Motivation for regression line

$$y = \beta_0 + \beta_1 * X + \varepsilon$$

There are a few advantages to using a line:

1. Often time does a good job at prediction (like in our NYC example)
2. Easy to interpret
3. A simple model faces less risk of overfitting the data.

The cost is that the model might be *too simplistic* and fail to caputure many non-linear relationships between X and y. It might yield poor predictions.

Regression Line Example

In the previous example, the regression 'line of best fit' (we will talk about how to find this line later) is given by

$$\text{Average }\widehat{\text{SAT}}\text{ Reading} = 78.87 + 0.7983 * \text{Average SAT Math}$$

The $\widehat{}$ symbol means that we are *predicting* average SAT reading score with our model

Regression Line Example

Predictions

If a school has an average SAT math score of 600, we would predict their SAT reading score would be

$$\text{Average } \widehat{\text{SAT}} \text{ Reading} = 78.87 + 0.7983 * 600 = 557.85$$

Regression Line Example

Predictions

If a school has an average SAT math score of 600, we would predict their SAT reading score would be

$$\text{Average } \widehat{\text{SAT}} \text{ Reading} = 78.87 + 0.7983 * 600 = 557.85$$

That is, our linear model would predict an average SAT reading score of 558.

Slope of Line

How does y change with X? Take X and $X+1$, we have the following predicted values:

$$\hat{y} = \beta_0 + \beta_1 X \quad \text{and} \quad \hat{y}_{\text{new}} = \beta_0 + \beta_1 (X+1)$$

So y changes by

$$\begin{aligned} \Delta y &= [\beta_0 + \beta_1(X+1)] - [\beta_0 + \beta_1 X] \\ &= \beta_1 X + \beta_1 - \beta_1 X \\ &= \beta_1 \end{aligned}$$

$\Longrightarrow$ marginal effect of X on y is constant and equal to β_1

Slope of Line

Example of Constant Marginal Effects

$$\text{Wage} = \beta_0 + \beta_1 \text{ Education} + \varepsilon$$

Implies that each year of education leads to the same change in wages

- Do you think that is reasonable?

Slope of Line

Example of Constant Marginal Effects

$$\text{Wage} = \beta_0 + \beta_1 \text{ Education} + \varepsilon$$

Implies that each year of education leads to the same change in wages

- Do you think that is reasonable?
- Might there be a jump at high-school degree ("signaling")?
- Returns to schooling might get smaller as we get more educated?

Prediction Error

Given our line, we will want to be able to evaluate how good our model does at predicting observations y

Define the prediction error as

$$\hat{\varepsilon} = \underbrace{y}_{\text{true value}} - \underbrace{\hat{y}}_{\text{predicted value}}$$

Prediction Error

In the case of a linear prediction model

$$\hat{\varepsilon} = \underbrace{y}_{\text{true value}} - \underbrace{\hat{y}}_{\text{predicted value}} = y - b_0 - b_1 X,$$

where b_0 and b_1 are any numbers (for now).

Large $\hat{\varepsilon}$ mean you did a poor job of predicting that observation. That could be because

1. The linear model is bad at predicting y
2. Or, the true noise ε is making y far away from the systematic component $f(X)$.

Mean-square Error

Just like in Topic 2, we can form the mean-square prediction error of our linear model (in our training sample):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 X_i)^2$$

A line does a good job at predicting if MSE is (relatively) small.

Mean-square Error

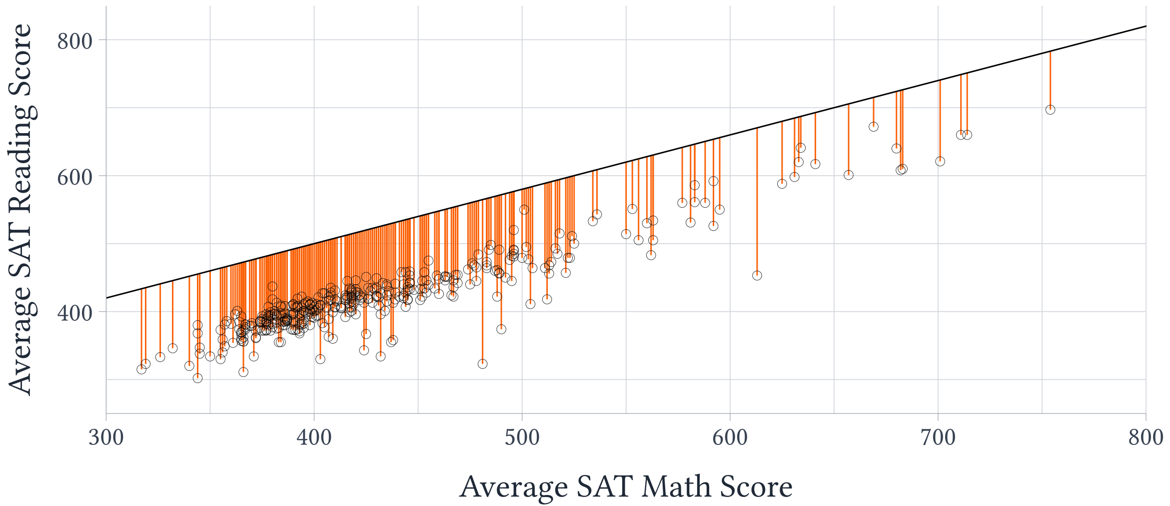
Just like in Topic 2, we can form the mean-square prediction error of our linear model (in our training sample):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 X_i)^2$$

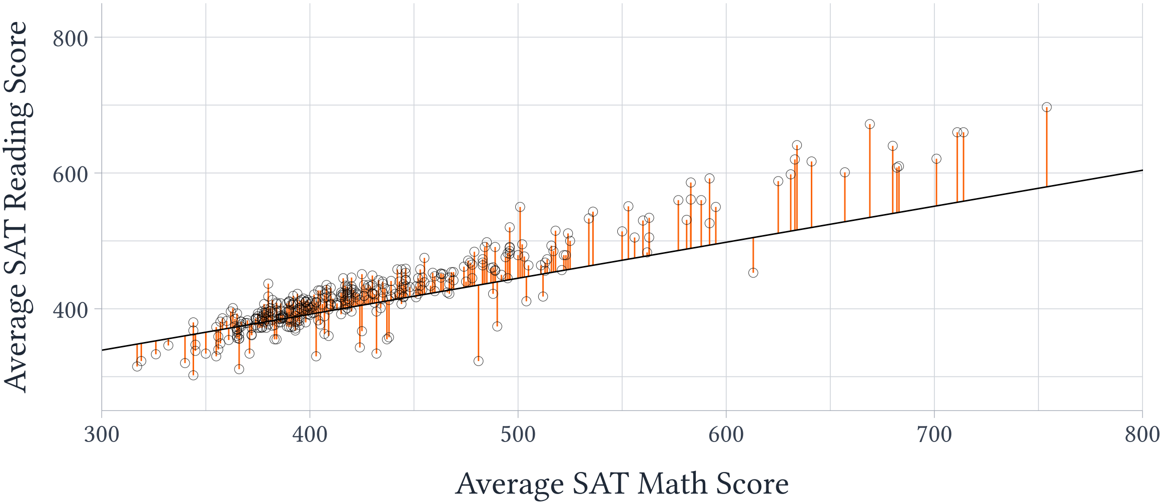
A line does a good job at predicting if MSE is (relatively) small.

What if we select a line based on making mean-squared prediction error as small as possible?

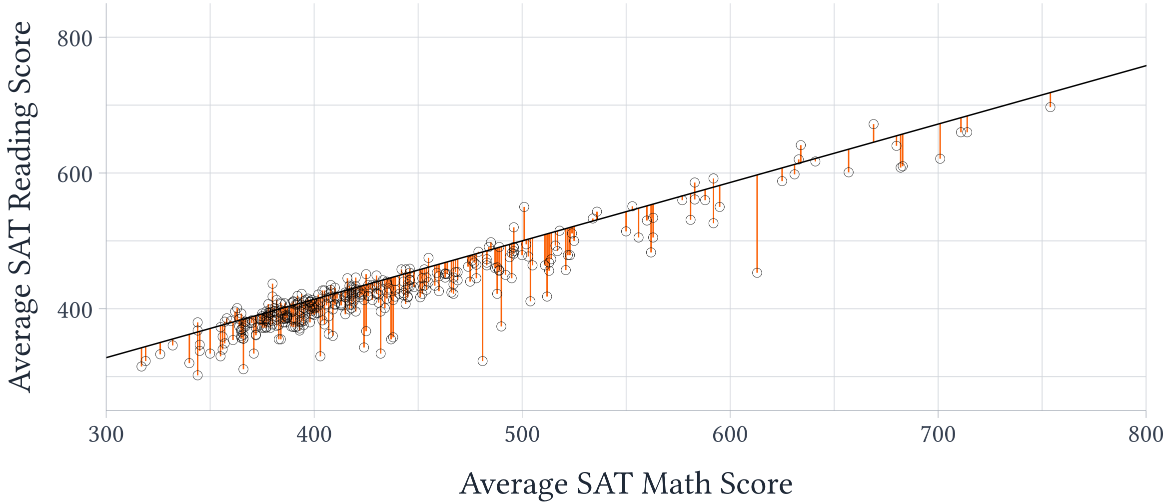
$\mathrm{MSE}(b_0, b_1) = 10902$



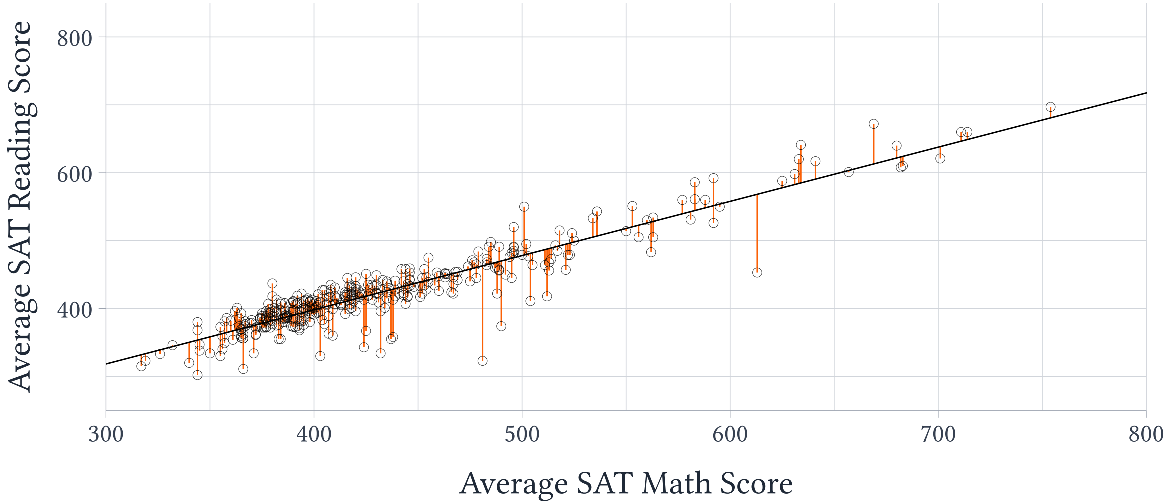
$\text{MSE}(b_0, b_1) = 1126$



$$\mathrm{MSE}(b_0, b_1) = 866$$



$\text{MSE}(b_0, b_1) = 528$



"Least Squares" Regression

This is the basis for the ordinary least squares regression estimator:

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

"Least Squares" Regression

This is the basis for the ordinary least squares regression estimator:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

$\hat{\beta}_0$, $\hat{\beta}_1$ are the values of the intercept and slope that minimize prediction error

- Do you see where the term "least squares" comes from?

Deriving Least Squares Formula

To minimize the function, we will take derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set equal to zero. First, $\hat{\beta}_0$:

$$\frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0$$

Deriving Least Squares Formula

To minimize the function, we will take derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set equal to zero. First, $\hat{\beta}_0$:

$$\frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0$$

$$\implies \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\implies \sum_{i=1}^{n} \hat{\varepsilon}_i = 0$$

Deriving Least Squares Formula

Continuing our first-order conditions for $\hat{\beta}_0$: $0 = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$

$$0 = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

$$\implies 0 = \left(\sum_{i=1}^{n} y_i\right) - n\hat{\beta}_0 - \left(\sum_{i=1}^{n} \hat{\beta}_1 X_i\right)$$

Deriving Least Squares Formula

Continuing our first-order conditions for $\hat{\beta}_0$: $0 = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$

$$0 = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

$$\implies 0 = \left(\sum_{i=1}^{n} y_i\right) - n\hat{\beta}_0 - \left(\sum_{i=1}^{n} \hat{\beta}_1 X_i\right)$$

$$\implies n\hat{\beta}_0 = \left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} \hat{\beta}_1 X_i\right)$$

Deriving Least Squares Formula

Continuing our first-order conditions for $\hat{\beta}_0$: $0 = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\
\implies 0 &= \left(\sum_{i=1}^{n} y_i\right) - n\hat{\beta}_0 - \left(\sum_{i=1}^{n} \hat{\beta}_1 X_i\right) \\
\implies n\hat{\beta}_0 &= \left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} \hat{\beta}_1 X_i\right) \\
\implies \hat{\beta}_0 &= \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right) - \hat{\beta}_1 \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)
\end{aligned}
$$

Deriving Least Squares Formula

All our work lead to

$$\hat{\beta}_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right) - \hat{\beta}_1 \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)$$

This we can write as our first least-squares formula

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

Deriving Least Squares Formula

To minimize the function, we will take derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set equal to zero. Second, $\hat{\beta}_1$:

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0$$

Deriving Least Squares Formula

To minimize the function, we will take derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set equal to zero. Second, $\hat{\beta}_1$:

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0$$

$$\implies \sum_{i=1}^{n} 2X_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

Deriving Least Squares Formula

To minimize the function, we will take derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set equal to zero. Second, $\hat{\beta}_1$:

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0$$

$$\implies \sum_{i=1}^{n} 2X_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\implies \sum_{i=1}^{n} X_i \hat{\varepsilon}_i = 0$$

Deriving Least Squares Formula

Taking $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$ and plugging into our first order condition for $\hat{\beta}_1$:

$$0 = \sum_{i=1}^{n} X_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

$$\implies 0 = \sum_{i=1}^{n} X_i(y_i - \bar{y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i)$$

Deriving Least Squares Formula

Taking $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$ and plugging into our first order condition for $\hat{\beta}_1$:

$$0 = \sum_{i=1}^{n} X_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

$$\implies 0 = \sum_{i=1}^{n} X_i (y_i - \bar{y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i)$$

$$\implies 0 = \sum_{i=1}^{n} X_i \left((y_i - \bar{y}) + \hat{\beta}_1 \left(\bar{X} - X_i \right) \right)$$

$$\implies 0 = \sum_{i=1}^{n} X_i \left(y_i - \bar{y} \right) - \hat{\beta}_1 \sum_{i=1}^{n} X_i \left(X_i - \bar{X} \right)$$

Deriving Least Squares Formula

$$0 = \sum_{i=1}^{n} X_i \left(y_i - \bar{y}\right) - \hat{\beta}_1 \sum_{i=1}^{n} X_i \left(X_i - \bar{X}\right)$$

$$\implies \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i \left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} X_i \left(X_i - \bar{X}\right)}$$

With a bit of algebra, you can find:

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right) \left(y_i - \bar{y}\right)}{\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} = \frac{\operatorname{cov}(X, y)}{\operatorname{var}(X)}$$

Least Squares Formula

With that, we have a formula for OLS coefficients:

$$\hat{\beta}_1 = \frac{\text{cov}(X, y)}{\text{var}(X)} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

Least Squares Formula

With that, we have a formula for OLS coefficients:

$$\hat{\beta}_1 = \frac{\text{cov}(X, y)}{\text{var}(X)} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

Note that the slope $\hat{\beta}_1$ describes how y changes with X

- That is what the covariance tells us!

Least Squares example

$$\hat{\beta}_1 = \frac{\text{cov}(X, y)}{\text{var}(X)} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

In our NYC example, calculate the regression coefficients of $y =$ Average SAT Reading Score and $X =$ Average SAT Math Score by hand. Here are the following statistics:

$\text{cov}\left(X, y\right) = 4132.97$

$\text{var}\left(X\right) = 5177.14$

$\bar{X} = 432.94$

$\bar{y} = 424.50$

Least Squares example

$$\hat{\beta}_1 = \frac{\text{cov}(X, y)}{\text{var}(X)} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

In our NYC example, calculate the regression coefficients of $y =$ Average SAT Reading Score and $X =$ Average SAT Math Score by hand. Here are the following statistics:

$\text{cov}\,(X, y) = 4132.97$

$\text{var}\,(X) = 5177.14$

$\bar{X} = 432.94$

$\bar{y} = 424.50$

Interpreting a Regression

$$y = \beta_0 + \beta_1 X$$

- β_0 is the value of y whenever $X = 0$.
- β_1 is the amount y changes when X increases by one.

Interpreting a Regression

Consider this hypothetical regression:

$$\hat{\text{Wins}}_i = 20.783 + 0.00913 * \text{ 3-point shots}$$

Our intercept, 20.783, is the predicted number of wings for an NBA team that shoots no 3-point shots

Interpreting a Regression

Consider this hypothetical regression:

$$\hat{\text{Wins}}_i = 20.783 + 0.00913 * \text{ 3-point shots}$$

Our intercept, 20.783, is the predicted number of wings for an NBA team that shoots no 3-point shots

Our slope, 0.00913, is the number of additional wins predicted for every 1 shot increase in the number of per-game 3-point shots

Interpreting a Regression

Say we calculate the following regression line from hours studied and final exam grades:

$$\widehat{\text{Final Exam}} = 38 + 5.7 * \text{Hours of Studying}$$

Interpret the two regression coefficients

Interpreting a Regression

Say we calculate the following regression line from hours studied and final exam grades:

$$\widehat{\text{Final Exam}} = 38 + 5.7 * \text{Hours of Studying}$$

Interpret the two regression coefficients

- 38 is the predicted score with no studying.

Interpreting a Regression

Say we calculate the following regression line from hours studied and final exam grades:

$$\widehat{\text{Final Exam}} = 38 + 5.7 * \text{Hours of Studying}$$

Interpret the two regression coefficients

- 38 is the predicted score with no studying.
- Each hour of studying increases the predicted final exam score by 5.7 points.

Practice Question

Given that same regression line, $\widehat{\text{Final Exam}} = 38 + 5.7 * \text{Hours of Studying}$, what is the predicted final exam score if you study 8 hours?

Practice Question

Given that same regression line, $\widehat{\text{Final Exam}} = 38 + 5.7 * \text{Hours of Studying}$, what is the predicted final exam score if you study 8 hours?

$38 + 5.7 * 8 = 83.6$

Practice Question

A convenience store calculates a least squares line that describes how price (in dollars) of juuls affects the quantity sold;

$$\widehat{\text{Juuls sold}} = 117 - 12.4 * \text{price}$$

If price *decreases* by 1 dollar, what happens to number of juuls sold?

Practice Question

A convenience store calculates a least squares line that describes how price (in dollars) of juuls affects the quantity sold;

$$\widehat{\text{Juuls sold}} = 117 - 12.4 * \text{price}$$

If price *decreases* by 1 dollar, what happens to number of juuls sold?

Quantity decreases by 12.4 units

Algebraic properties of OLS

There are three properties of OLS we will cover. The first two are our first-order conditions

1. $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$
   - $\rightarrow$ The residuals sum to 0

Algebraic properties of OLS

There are three properties of OLS we will cover. The first two are our first-order conditions

1. $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

   $\rightarrow$ The residuals sum to 0

2. $\sum_{i=1}^{n} X_i \hat{\varepsilon}_i = 0$

   $\rightarrow$ The residual is uncorrelated with the X variable

Algebraic properties of OLS

There are three properties of OLS we will cover. The first two are our first-order conditions

1. $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

   $\rightarrow$ The residuals sum to 0

2. $\sum_{i=1}^{n} X_i \hat{\varepsilon}_i = 0$

   $\rightarrow$ The residual is uncorrelated with the X variable

3. $(\bar{X}, \bar{y})$ is on the regression line

Algebraic properties of OLS

$(\bar{X}, \bar{y})$ is on the regression line comes from:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i)$$

$$= \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i$$

$$= \hat{\beta}_0 + \hat{\beta}_1 \bar{X} + 0$$

Cautions about Correlation and Regression

Our regression line is fit by comparing individuals with larger and smaller X values and seeing if units with larger X have larger or smaller values of y.

Cautions about Correlation and Regression

Our regression line is fit by comparing individuals with larger and smaller X values and seeing if units with larger X have larger or smaller values of y.

Units with larger values of X might have larger values of other variables and those other variables can affect y

- Which variable is driving the change in y? We do not know

Do not confuse *prediction* with *causation*!!!

Example of Prediction vs. Causation

Units with more years of schooling have higher wages

- Is this because of schooling?
- Or, is this because people with more schooling have higher intelligence? Differing home backgrounds? More responsible?

Example of Prediction vs. Causation

Units with more years of schooling have higher wages

- Is this because of schooling?
- Or, is this because people with more schooling have higher intelligence? Differing home backgrounds? More responsible?

Correlation and regression are powerful tools for describing the relationship between two variables, but you must be careful!

Correct regression interpretation

In general, you should use the following language:

> ✓ Our regression model predicts that a one unit increase in X is associated with a $\hat{\beta}_1$ units increase/decrease in Y

Do not say!!!!!!

> × Increase X by one unit increases/decreases Y by $\hat{\beta}_1$ units

Learning about Causation

If you are interested in learning how to estimate *causal effects*, you should take my Master's level class, ECON 5783 :-)

Roadmap

Bivariate Regression

Prediction vs Causation

Regression Inference

Goodness of Fit

Influential Observations

Discrete Variables

$\log$ transformations

Regression Inference

As we have seen, the regression coefficient $\hat{\beta}_1$ is often of interest

- Predicted change in y when you increase X by one unit

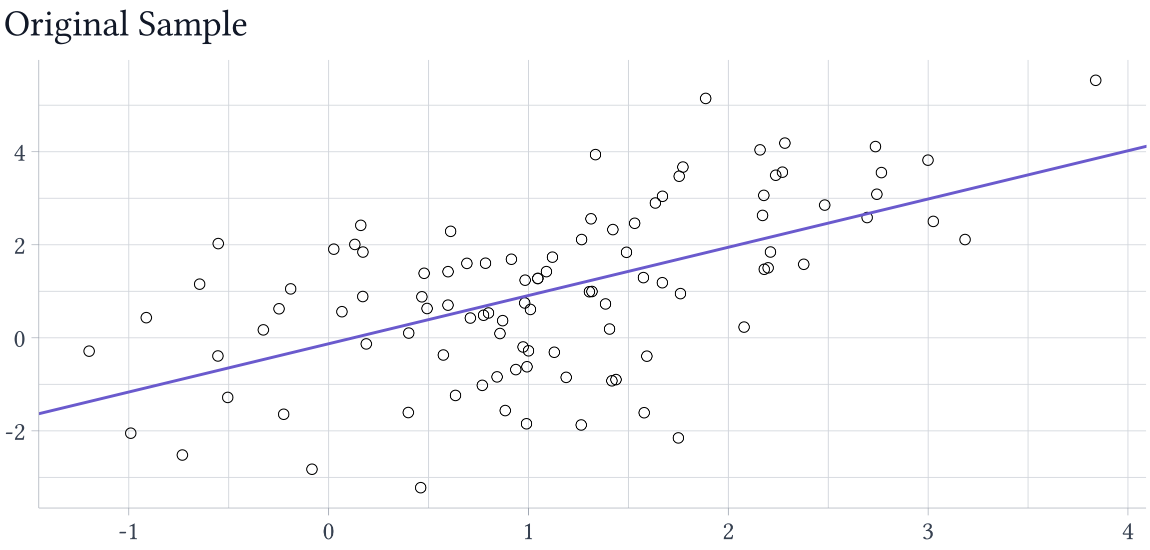
Regression Inference

As we have seen, the regression coefficient $\hat{\beta}_1$ is often of interest

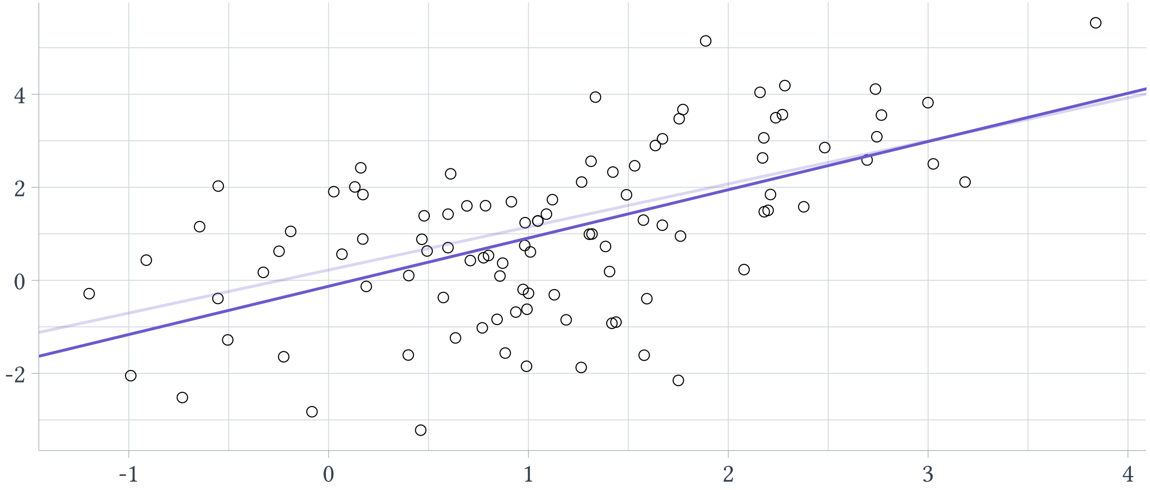
- Predicted change in y when you increase X by one unit

We want to be able to describe the uncertainty around this estimate. How does $\hat{\beta}_1$ change under repeated sampling?

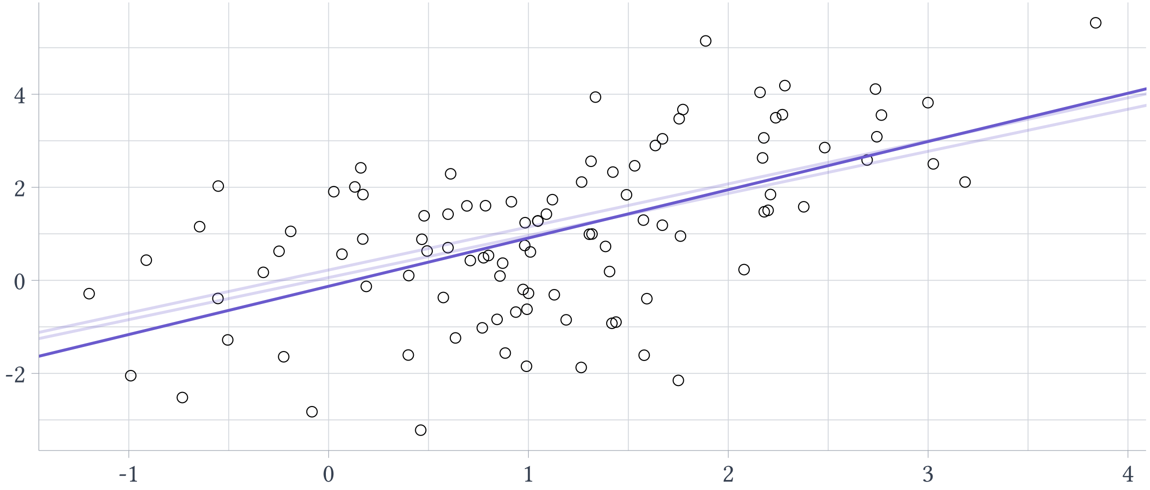
- That is, what is the *sampling distribution* of $\hat{\beta}_1$?


Original Sample
4
2
0
-2
-1
0
1
2
3
4

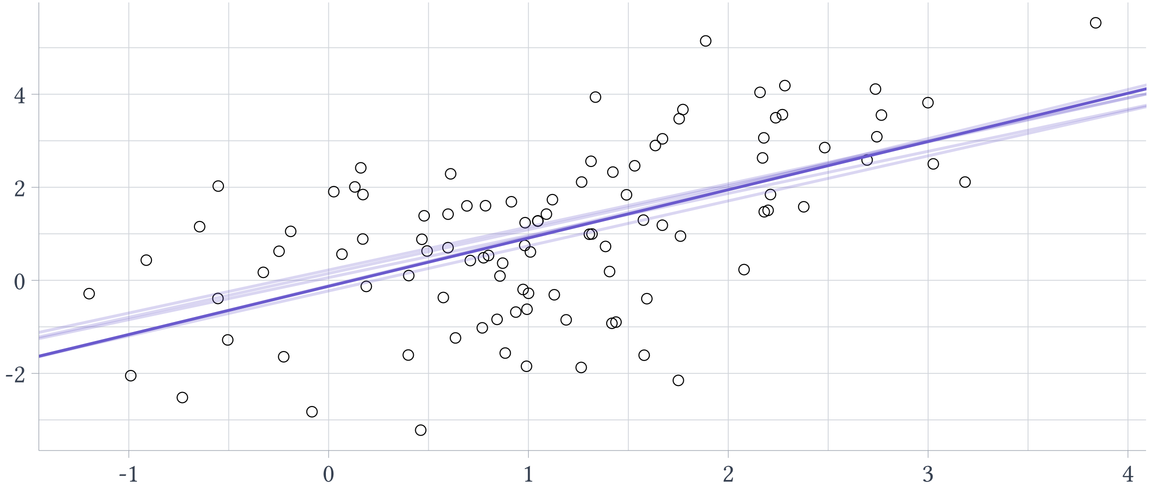
Original Sample + 1 Extra Sample



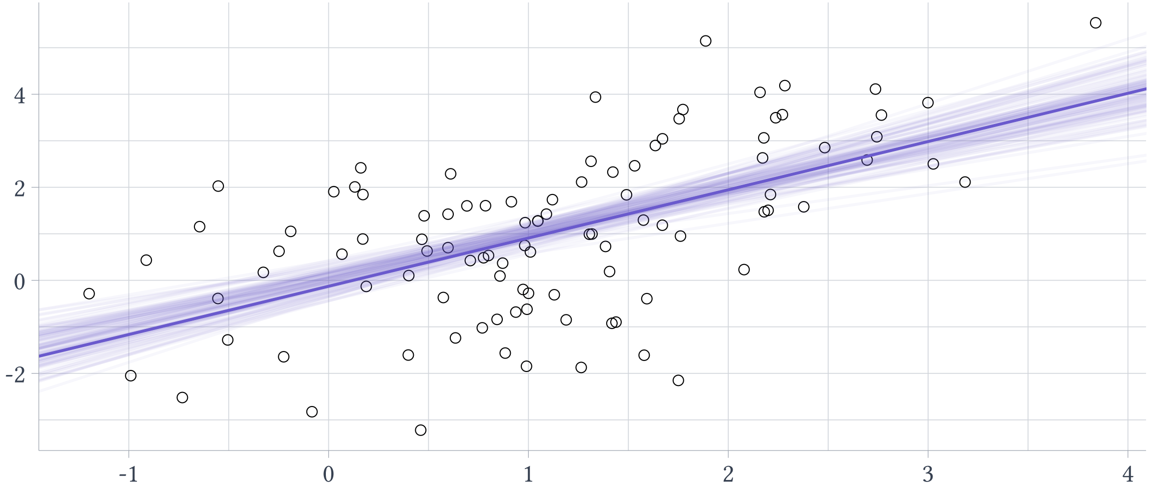
Original Sample + 2 Extra Samples

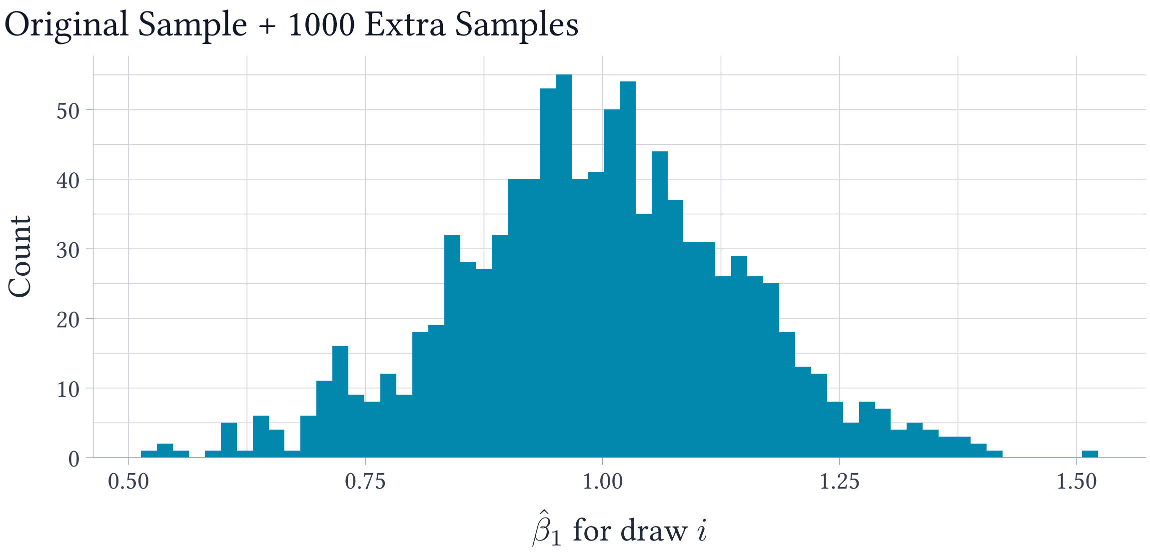


Original Sample + 5 Extra Samples



Original Sample + 100 Extra Samples




Original Sample + 1000 Extra Samples
Count
0
10
20
30
40
50
0.50
0.75
1.00
1.25
1.50
$\hat{\beta}_1$ for draw i

Regression Inference

For each sample of size n, the regression coefficient estimate $\hat{\beta}_1$ is different

- As n gets large, the noise of the estimate should get smaller

Sample Distribution of Sample Mean

Recall that we have the sample distribution of the sample mean (provided n is ‘big enough’):

$$\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

- What is the equivalent for regression estimates?

Sample Distribution of Regression Coefficients

Say the true regression line is

$$y_i = \beta_{0,0} + X_i\beta_{1,0} + \varepsilon_i$$

- $\beta_{0,0}$ and $\beta_{1,0}$ denotes the true regression coefficient for the population
- ε is the error term from the true regression line

Sample Distribution of Regression Coefficients

Say the true regression line is

$$y_i = \beta_{0,0} + X_i \beta_{1,0} + \varepsilon_i$$

- $\beta_{0,0}$ and $\beta_{1,0}$ denotes the true regression coefficient for the population
- ε is the error term from the true regression line

The sampling distribution of $\hat{\beta}_1$ (for n ‘big enough’) is:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_{1,0}, \frac{\operatorname{var}(\varepsilon)/n}{\operatorname{var}(X)}\right)$$

Sample Distribution of Regression Coefficients

The sampling distribution of $\hat{\beta}_1$ (for n ‘big enough’) is:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_{1,0}, \frac{\mathrm{var}\left(\varepsilon\right)/n}{\mathrm{var}\left(X\right)}\right)$$

Since we have a statistic $\hat{\beta}_1$ that has a sample distribution that is normally-distributed, we can do standard statistical techniques: confidence intervals, hypothesis testing, and form rejection region.

Standard Error

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_{1,0}, \frac{\operatorname{var}(\varepsilon)/n}{\operatorname{var}(X)}\right)$$

With this, we can calculate the standard error, i.e. the standard deviation of the sample distribution of $\hat{\beta}_1$:

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\operatorname{var}(\hat{\varepsilon})/n}{\operatorname{var}(X)}}$$

- We use the residual $\hat{\varepsilon}$ because we do not observe the true error term

Standard Error

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\text{var}\,(\hat{\varepsilon})/n}{\text{var}\,(X)}}$$

- As our sample size gets larger, $n \to \infty$, we have the distribution converges to the true value (*consistency*)

Confidence intervals for $\hat{\beta}_1$

Since we have an approximately normally distributed random variable, we can form confidence intervals just like before:

$$\left[\hat{\beta}_1 - 1.96 * \mathsf{SE}(\hat{\beta}_1), \hat{\beta}_1 + 1.96 * \mathsf{SE}(\hat{\beta}_1)\right]$$

Confidence intervals for $\hat{\beta}_1$

Since we have an approximately normally distributed random variable, we can form confidence intervals just like before:

$$\left[\hat{\beta}_1 - 1.96 * \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 1.96 * \text{SE}(\hat{\beta}_1)\right]$$

The interpretation is as before: across repeated samples, 95% of samples' confidence intervals will contain the true value $\beta_{1,0}$

Confidence Interval Example

In our NYC example, the regression coefficients were $\hat{\beta}_0 = 78.87$ and $\hat{\beta}_1 = 0.798$. The standard error on $\hat{\beta}_1$ is 0.020.

Form the 95% confidence interval for $\hat{\beta}_1$:

Confidence Interval Example

In our NYC example, the regression coefficients were $\hat{\beta}_0 = 78.87$ and $\hat{\beta}_1 = 0.798$. The standard error on $\hat{\beta}_1$ is 0.020.

Form the 95% confidence interval for $\hat{\beta}_1$:

$$[0.798 - 1.96 * 0.205, 0.798 + 1.96 * 0.205] \approx [0.756, 0.837]$$

- With 95%, we think the true value of β_1 falls within the confidence interval $[0.756, 0.837]$

Hypothesis Testing

Consider the test $H_0 : \beta_{1,0} = b_1$, that the slope coefficient equals b_0 with the (two-sided) alternative hypothesis $H_A : \beta_{1,0} \neq b_1$

Form the test statistic:

$$\hat{t} \equiv \frac{\hat{\beta}_1 - b_1}{\text{SE}(\hat{\beta}_1)}$$

With n approximately large, t is distributed standard-normal, and can look up the t-statistic in the Z-table to form the p-value.

Hypothesis Testing and p-value

The p-value tells you the probability of observing an estimate as or more extreme as the one you *did* observe under the null that $\beta_{1,0} = b_1$. So we want to look up

$$p\text{-value} = \mathbb{P}\big(Z \leq -\left|\hat{t}\right|\big) + \mathbb{P}\big(Z \geq \left|\hat{t}\right|\big) = 2 * \mathbb{P}\big(Z \leq -\left|\hat{t}\right|\big)$$

Hypothesis Testing and p-value

The p-value tells you the probability of observing an estimate as or more extreme as the one you *did* observe under the null that $\beta_{1,0} = b_1$. So we want to look up

$$p\text{-value} = \mathbb{P}\left(Z \leq -\left|\hat{t}\right|\right) + \mathbb{P}\left(Z \geq \left|\hat{t}\right|\right) = 2 * \mathbb{P}\left(Z \leq -\left|\hat{t}\right|\right)$$

If p-value is less than the critical value (typically 0.05), then we reject the null.

Rejection Region

Again, we could form the 95% confidence interval using the null value b_1 to find all values of $\hat{\beta}_1$ you would not reject at the 5% significance level:

$$\left[b_1 - 1.96 * \mathsf{SE}(\hat{\beta}_1), b_1 + 1.96 * \mathsf{SE}(\hat{\beta}_1)\right]$$

Regression in `R`

In `R`, we can do regression using the `lm` function that is built into base `R`. But we are going to use a package called `fixest` since it has a lot of extra useful features

- Install it (only need to do this once) using `install.packages("fixest")`
- At the top of your `.Rmd` files, load the pacakge using `library(fixest)`

Regression in `R`

You can call either `lm` or the fixest function `feols` with the exact same arguments:

```
feols(y ~ x, data = df)
```

- `y ~ x` is the formula where `y` and `x` are the name of the variables in `df`
- `df` is the name you called your dataframe.

Regression output with `feols`

When you run a regression, you get the following output (some lines are cut off):

```
                     Estimate Std. Error  t value  Pr(>|t|)
(Intercept)         78.879717   8.485841  9.29545 < 2.2e-16 ***
average_score_sat_math  0.798312   0.020525 38.89494 < 2.2e-16 ***
```

Regression output with `feols`

When you run a regression, you get the following output (some lines are cut off):

```
                        Estimate Std. Error  t value  Pr(>|t|)
(Intercept)            78.879717   8.485841  9.29545 < 2.2e-16 ***
average_score_sat_math  0.798312   0.020525 38.89494 < 2.2e-16 ***
```

Each row tells you the:

- The estimate $\hat{\beta}_i$
- The standard error $SE(\hat{\beta}_i)$
- The test-statistic $\hat{t}$ for $H_0 : b_i = 0$ and the corresponding p-value

“Stars” in regression

```
                       Estimate Std. Error  t value  Pr(>|t|)
average_score_sat_math  0.798312   0.020525 38.89494 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The stars, $\star$, correspond to different levels of significance (as shown at the bottom)

- More stars means you can reject the null that $b_i = 0$ with more significance

“Stars” in regression

```
                          Estimate Std. Error  t value  Pr(>|t|)
(Intercept)              78.879717   8.485841  9.29545 < 2.2e-16 ***
average_score_sat_math  0.798312   0.020525 38.89494 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The stars, $\star$, correspond to different levels of significance (as shown at the bottom).

You get stars when you reject the null at a critical value

- More stars means you can reject the null that $b_i = 0$ with more significance

In this example, can you reject the null that the slope coefficient on the Average SAT Math Score is 0? How do you know?

Roadmap

Bivariate Regression

Prediction vs Causation

Regression Inference

Goodness of Fit

Influential Observations

Discrete Variables

$\log$ transformations

R^2

We want Next we define a measure to evaluate how well the regression line fits:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\text{ESS}}{\text{TSS}}$$

- The ESS is the explained sum of squares, i.e. the variance of the prediced $\hat{Y}$
- The TSS is the total sum of squares, i.e. the variance of Y

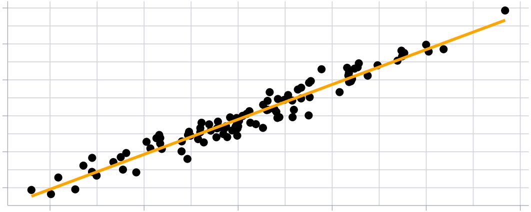
Intuition of R^2

Intuitively, R^2 measures the percent of variation in Y explained by the model

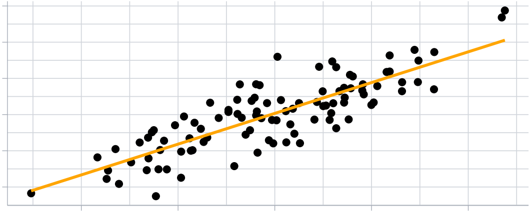
$$R^2 = \frac{\text{variation in } \hat{y} \text{ along the regression line as x varies}}{\text{total variation in observed values of y}}$$

Comparisons of R^2

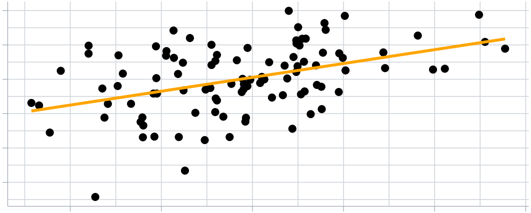
$R^2 = 0.943$



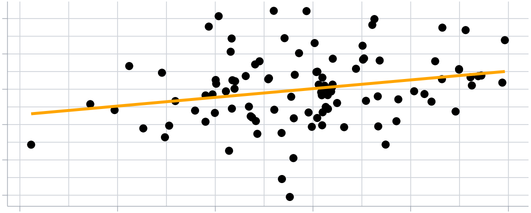
$R^2 = 0.698$



$R^2 = 0.189$



$R^2 = 0.058$



r and R^2

Correlation, r, describes the strength of a straight-line relationship between two variables

R^2, is the fraction of the variation in the values of y that is explained by the least-squares regression of y on X. In the case of a single-variable regression, we have

$$R^2 = r^2$$

r and R^2

Lets say we have $r = -0.7786$ and $R^2 = (-0.7786)^2 = 0.6062$ between exercise and weight loss.

- $r = -0.7786$, there is a strong negative linear relationship between time exercised and amount of weight gained
- $R^2 = 0.6062$, about 61% of the variation in weight losseis accounted for by the linear relationship between weight loss and exercise. This means about 39% of the change in weight lossed is not explained by this relationship

R^2 Sidebar

A small R^2 does not mean the result is uninteresting. All it means is that the x variable alone does not explain a large portion of the variation in y.

R^2 Sidebar

A small R^2 does not mean the result is uninteresting. All it means is that the x variable alone does not explain a large portion of the variation in y.

Example: You find a significant relationship between exercise and income, but it has a small R^2.

We know income is determined by a variety of variables – parent's income, education, innate ability, experience, etc.

- Your result isn't uninteresting; it just means there is a lot of variation in income *not due* to exercise, which is exactly what we'd expect

R^2 Practice Question

Say a researcher calculated a correlation coefficient 0.503 between SAT scores and college freshman GPA. This implies an R^2 of 0.253.

Practice interpreting what this R^2 mean?

- Does this make sense? What other things could explain the variation in freshman year GPA?

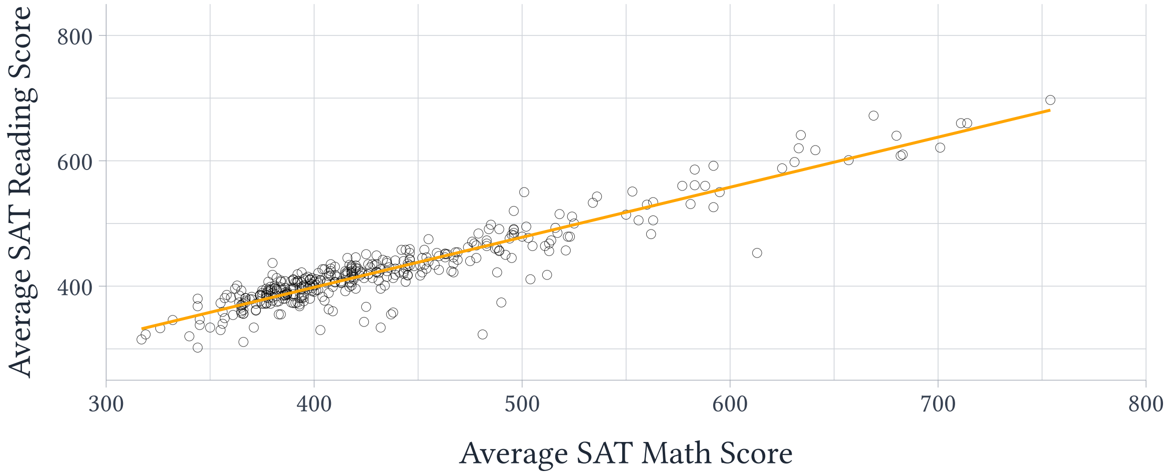
Roadmap

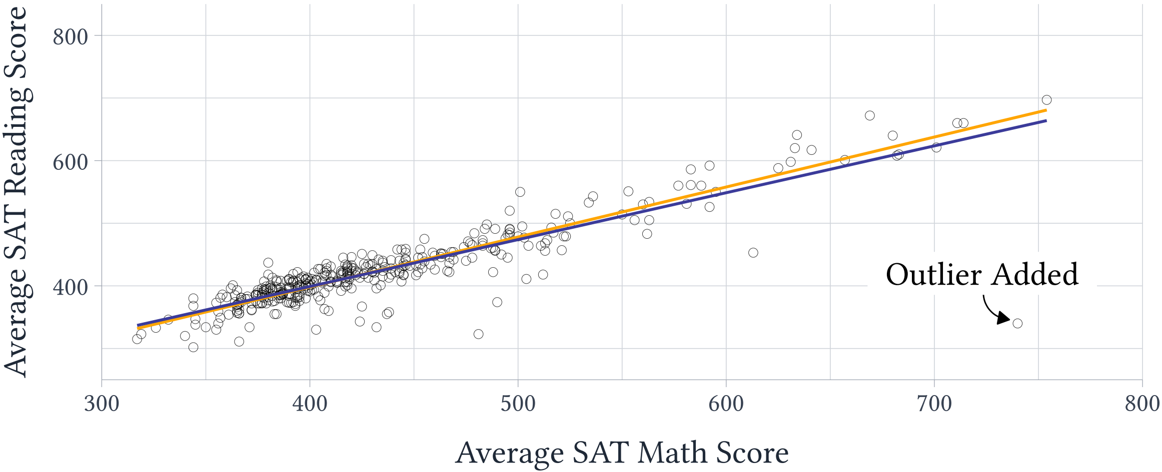
- Bivariate Regression
  - Prediction vs Causation
- Regression Inference
- Goodness of Fit
- **Influential Observations**
- Discrete Variables
- $\log$ transformations

Influential Observations

Our regression line is sensitive to outliers, either in the X or y dimension

- We say an outlier is influential if deleting it changes our regression line substantially
- The amount by which the line changes is called the leverage an influential observation has


Average SAT Reading Score
800
600
400
300
400
500
600
700
800
Average SAT Math Score



Average SAT Reading Score
800
600
400
300
400
500
600
700
800
Average SAT Math Score
Outlier Added

Outliers and large samples

In this example, since we have a relatively large number of observations, this single outlier did not move our regression line by much

- This was one of the benefits of the regression model for $f(X)$; it does not overfit any individual observation

Outliers tend to matter more for small samples!

Outliers

It is always good practice to *plot* the raw data. In a world full of dirty data, you will be amazed at how quickly you can spot oddities in the data

For example, `NA`s might be stored as 99 in a dataset

- While one single outlier might not move the regression line by much, a large number of them will!!

Roadmap

- Bivariate Regression
  - Prediction vs Causation
- Regression Inference
- Goodness of Fit
- Influential Observations
- Discrete Variables
- $\log$ transformations

Discrete Variables

So far, we have thought about X variables that are continuously distributed. Now, we turn to the other extreme where X is a discrete variable

- Remember, discrete means the variable takes on finitely many values

Regression and Sample Means

First, we will have an aside on a sort-of peculiar regression: `lm(y ~ 1, data = df)`

- Regress `y` on a variable that is equal to 1 for all observations

Regression and Sample Means

First, we will have an aside on a sort-of peculiar regression: `lm(y ~ 1, data = df)`

- Regress `y` on a variable that is equal to 1 for all observations

This model is written as $y_i = \alpha + u_i$. What is the best estimate of $\hat{\alpha}$?

- Well, we observe *no* information about the individual, so our best guess at y_i is the sample mean of $y \implies \hat{\alpha} = \bar{y}$

```r
library(fixest)
feols(mpg ~ 1, data = mtcars)
#> OLS estimation, Dep. Var.: mpg
#> Observations: 32
#> Standard-errors: IID
#>             Estimate Std. Error t value  Pr(>|t|)
#> (Intercept)  20.0906    1.06542 18.8569 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 5.93203

mean(mtcars[["mpg"]])
#> [1] 20.09062
```

Regression and Sample Means

```
lm(y ~ 1, data = df)
```

Running this regression is useful because it will estimate the mean of `y` and give us the standard error estimate $\frac{\sigma}{\sqrt{n}}$

- This makes inference easier: hypothesis testing and confidence intervals

Indicator variable

We now understnad the simplest regression on just an intercept. What about an indicator variable?

An *indicator variable* is a variable that can only equal 0 and 1

- X "indicates" when a unit is of type 0 or type 1

E.g. include being born male (=1) or female (=0); being White (=1) or not (=0); having a high-school degree (=1); being over 6 foot tall (=1) or under (=0); etc.

Indicator variable

Let’s work through some properties of an indicator variable. First, The sample mean of an indicator variable is the proportion of units with a 1:

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

Indicator variable

Let’s work through some properties of an indicator variable. First, The sample mean of an indicator variable is the proportion of units with a 1:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\#\text{ of }1s}{n} = \text{\% of sample with 1}$$

Define π as the fraction of units with $X_i = 1$

Indicator variable

Let's work through some properties of an indicator variable. First, The sample mean of an indicator variable is the proportion of units with a 1:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\#\text{ of } 1s}{n} = \text{% of sample with 1}$$

Define π as the fraction of units with $X_i = 1$

Second, I'll assert without proof (write down the formula and you can figure it out):

$$\mathrm{var}(X_i) = \pi(1-\pi)$$

Covariance with an indicator variable

What is $\operatorname{cov}(X_i, Y_i)$? Rembmer

$$\operatorname{cov}(X_i, Y_i) = \mathbb{E}(X_i Y_i) - \mathbb{E}(X_i)\,\mathbb{E}(Y_i)$$

Covariance with an indicator variable

What is $\text{cov}(X_i, Y_i)$? Rembmer

$$\text{cov}(X_i, Y_i) = \mathbb{E}(X_i Y_i) - \mathbb{E}(X_i)\,\mathbb{E}(Y_i)$$

Again, skipping the math:

$$\text{cov}(X_i, Y_i) = \pi(1-\pi)\left(\mathbb{E}(Y_i \mid X_i = 1) - \mathbb{E}(Y_i \mid X_i = 0)\right)$$

Covariance with an indicator variable

Math details (if you're curious):

$$
\begin{aligned}
\mathrm{cov}(X_i, Y_i) &= \mathbb{E}(X_i Y_i) - \mathbb{E}(X_i)\,\mathbb{E}(Y_i) \\
&= \mathbb{E}(X_i Y_i) - \pi\left(\pi\,\mathbb{E}(Y_i \mid X_i = 1) + (1-\pi)\,\mathbb{E}(Y_i \mid X_i = 0)\right) \\
&= \pi\,\mathbb{E}(Y_i \mid X_i = 1) - \pi\pi\,\mathbb{E}(Y_i \mid X_i = 1) - \pi(1-\pi)\,\mathbb{E}(Y_i \mid X_i = 0) \\
&= \pi(1-\pi)\left[\mathbb{E}(Y_i \mid X_i = 1) - \mathbb{E}(Y_i \mid X_i = 0)\right]
\end{aligned}
$$

Regression with an indicator variable

Say you have a regression of $Y_i = \beta_0 + \beta_1 * X_i + u_i$. What does $\hat{\beta}_0$ and $\hat{\beta}_1$ equal?

$$\hat{\beta}_1 = \frac{\mathrm{cov}(X_i, Y_i)}{\mathrm{var}\left(X_i\right)}$$

Regression with an indicator variable

Say you have a regression of $Y_i = \beta_0 + \beta_1 * X_i + u_i$. What does $\hat{\beta}_0$ and $\hat{\beta}_1$ equal?

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\operatorname{cov}(X_i, Y_i)}{\operatorname{var}(X_i)} \\
&= \frac{\pi(1-\pi)\left(\mathbb{E}(Y_i \mid X_i = 1) - \mathbb{E}(Y_i \mid X_i = 0)\right)}{\pi(1-\pi)} \\
&= \mathbb{E}(Y_i \mid X_i = 1) - \mathbb{E}(Y_i \mid X_i = 0)
\end{aligned}$$

The coefficient $\hat{\beta}_1$ tells me the difference in sample means between the group with $X_i = 1$ and the group with $X_i = 0$

Regression with an indicator variable

Say you have a regression of $Y_i = \beta_0 + \beta_1 * X_i + u_i$. From the last slide, we have:

$$\hat{\beta}_1 = \mathbb{E}(Y_i \mid X_i = 1) - \mathbb{E}(Y_i \mid X_i = 0)$$

Solving our other first-order condition for $\hat{\beta}_0$, we have:

$$\hat{\beta}_0 = \mathbb{E}(Y_i \mid X_i = 0)$$

Math Details

$$
\begin{aligned}
\beta_0 &= \mathbb{E}(Y) - \hat{\beta}_1 \, \mathbb{E}(X) \\
&= \mathbb{E}(Y) - \hat{\beta}_1 \, \mathbb{E}(X) \\
&= \mathbb{E}(Y) - \hat{\beta}_1 \pi \\
&= \pi \, \mathbb{E}(Y_i \mid X_i = 1) + (1 - \pi) \, \mathbb{E}(Y_i \mid X_i = 0) - \pi \, \mathbb{E}(Y_i \mid X_i = 1) - \pi \, \mathbb{E}(Y_i \mid X_i = 0) \\
&= \mathbb{E}(Y_i \mid X_i = 0)
\end{aligned}
$$

Interpreting the coefficients

Our model (without the error term) is $\hat{Y}_i = \beta_0 + \beta_1 X_i$.

Since X_i contains only two values, we can just compare them directly:

- When $X_i = 0$, $\hat{Y}_i = \beta_0 + \beta_1 * 0 = \beta_0$

Interpreting the coefficients

Our model (without the error term) is $\hat{Y}_i = \beta_0 + \beta_1 X_i$.

Since X_i contains only two values, we can just compare them directly:

- When $X_i = 0$, $\hat{Y}_i = \beta_0 + \beta_1 * 0 = \beta_0$
- When $X_i = 1$, $\hat{Y}_i = \beta_0 + \beta_1 * 1 = \beta_0 + \beta_1$

β_0 is our predicted value for Y_i for the group with $X_i = 0$ and $\beta_0 + \beta_1$ is our predicted value for Y_i for the group with $X_i = 1$

Intuition

β_0 is our predicted value for Y_i for the group with $X_i = 0$ and $\beta_0 + \beta_1$ is our predicted value for Y_i for the group with $X_i = 1$

Given this, then our regression coefficients make sense:

- $\hat{\beta}_0$ is the average value of Y_i for the group with $X_i = 0$
- $\hat{\beta}_1$ is the difference in the means between the two groups
- This makes $\hat{\beta}_0 + \hat{\beta}_1$ is the average value of Y_i for the group with $X_i = 1$

Example

Let's revisit our example with the `mtcars` dataset. There is an indicator variable, `am` for being an automatic (=1) or manual (=0). Regress the miles per gallon a car gets, `mpg`, on `am`.

- In `fixest`, we can use `i(am)` to make it print out more nicely

```r
feols(mpg ~ i(am), data = mtcars)
#> OLS estimation, Dep. Var.: mpg
#> Observations: 32
#> Standard-errors: IID
#>             Estimate Std. Error  t value   Pr(>|t|)
#> (Intercept) 17.14737    1.12460 15.24749 1.1340e-15 ***
#> am::1        7.24494    1.76442  4.10613 2.8502e-04 ***

mean(mtcars[mtcars$am == 1, ]$mpg)
#> [1] 24.39231
mean(mtcars[mtcars$am == 0, ]$mpg)
#> [1] 17.14737
```

Multi-valued discrete variables

This intuition will extend directly to settings where we have a discrete variable that obtains K distinct values:

- E.g. race, 10-year bins of age, number of cylinders in engine

Multi-valued discrete variables

This intuition will extend directly to settings where we have a discrete variable that obtains K distinct values:

- E.g. race, 10-year bins of age, number of cylinders in engine

We can construct a *set of* indicator variables for each value that X can obtain. For $k = 1, \dots, K$

$$X_{ik} \equiv \mathbb{1}[X_i = x_k]$$

- K such variables $X_{i1}, \dots, X_{iK}$

Multi-valued variable regression

Now we are in a setting where we have multiple regressors on the right-hand side (K indicators).

$$y_i = \sum_{k=1}^{K} X_{ik}\beta_k + u_i$$

Multi-valued variable regression

Now we are in a setting where we have multiple regressors on the right-hand side (K indicators).

$$y_i = \sum_{k=1}^{K} X_{ik}\beta_k + u_i$$

We don't know what these regressions estimate yet

- Note we are in a very special case since these variables are mutually exclusive (only one of them is non-zero per unit)

Multi-valued variable regression

$$y_i = \sum_{k=1}^{K} X_{ik}\beta_k + u_i$$

From the same intuition as before, we have $\hat{\beta}_k$ is the sample average of y_i for individuals with $X_i = x_k$

Example

Let's revisit our example with the `mtcars` dataset. Let's see if `mpg` differs based on the number of cylinders a car has, `cyl`.

- In `fixest`, we can use `i(am)` to make indicators for each value of a variable
- Otherwise, we could for $4, 6$, and 8 create the indicator variables with `mtcars$cyl4 = (mtcars$cyl == 4)`

Interpret these coefficients:

```
library(fixest)
feols(mpg ~ 0 + i(cyl), data = mtcars)
#> OLS estimation, Dep. Var.: mpg
#> Observations: 32
#> Standard-errors: IID
#>        Estimate Std. Error t value   Pr(>|t|)
#> cyl::4  26.6636   0.971801 27.4373  < 2.2e-16 ***
#> cyl::6  19.7429   1.218217 16.2064 4.4933e-16 ***
#> cyl::8  15.1000   0.861409 17.5294  < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 3.0683   Adj. R2: 0.704784
```

Intercept and Multicollinearity

In the future, we will want to include many right-hand side variables (beyond our multi-valued discrete variable). In this case, we want to include an *intercept*

$$y_i = \alpha + \sum_{k=1}^{K} X_{ik}\beta_k + u_i$$

Multicollinearity

Our X variables look like this. Note the 3 `cyl` indicator variables sum to the intercept

```
#> (Intercept)   cyl::4   cyl::6   cyl::8
#>           1        0        1        0
#>           1        0        1        0
#>           1        1        0        0
#>           1        0        1        0
#>           1        0        0        1
#>           1        0        1        0
#>           1        0        0        1
#>           1        1        0        0
```

Multicollinearity

$$\hat{y}_i = \hat{\alpha} + \sum_{k=1}^{K} X_{ik}\hat{\beta}_k$$

It turns out that we face a non-uniqueness problem because of the multicollinearity we identified

- We can add 10 to $\hat{\alpha}$ and subtract 10 from $\hat{\beta}_4$, $\hat{\beta}_6$, and $\hat{\beta}_8$ and get the same $\hat{y}$

Therefore, we will typically need to drop one of the X_{ik} variables (or `R` will do it for you)

```r
library(fixest)
feols(mpg ~ 1 + i(cyl), data = mtcars)
#> OLS estimation, Dep. Var.: mpg
#> Observations: 32
#> Standard-errors: IID
#>              Estimate Std. Error  t value   Pr(>|t|)
#> (Intercept)  26.66364   0.971801 27.43735  < 2.2e-16 ***
#> cyl::6       -6.92078   1.558348 -4.44110 1.1947e-04 ***
#> cyl::8      -11.56364   1.298623 -8.90453 8.5682e-10 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 3.0683   Adj. R2: 0.714009
```

Interpreting Multicollinearity

In the previous example, we dropped $\mathbb{1}[X_i = 4]$. This is the omitted group. What happened to our coefficient estimates?

- Just like in the indicator variable case, $\hat{\alpha}$ estimated the mean of `mpg` for cars with $X_i = 4$

Interpreting Multicollinearity

In the previous example, we dropped $\mathbb{1}[X_i = 4]$. This is the omitted group. What happened to our coefficient estimates?

- Just like in the indicator variable case, $\hat{\alpha}$ estimated the mean of `mpg` for cars with $X_i = 4$
- The coefficients on the other $\hat{\beta}_k$ now represent the *difference* in means between the group for $X_i = 6$ and the 'omitted group' $X_i = 4$.
  - $\rightarrow$ The mean for $X_i = 6$ is $19.742 = 26.663 - 6.921$

Specifying `ref` option

```r
library(fixest)
feols(mpg ~ i(cyl, ref = 6), data = mtcars)
#> OLS estimation, Dep. Var.: mpg
#> Observations: 32
#> Standard-errors: IID
#>             Estimate Std. Error  t value   Pr(>|t|)
#> (Intercept) 19.74286    1.21822 16.20636 4.4933e-16 ***
#> cyl::4       6.92078    1.55835  4.44110 1.1947e-04 ***
#> cyl::8      -4.64286    1.49200 -3.11182 4.1522e-03 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 3.0683   Adj. R2: 0.714009
```

Significance with indicator variables

```
#>               Estimate Std. Error  t value   Pr(>|t|)
#> (Intercept)  26.66364   0.971801 27.43735  < 2.2e-16 ***
#> cyl::6       -6.92078   1.558348 -4.44110 1.1947e-04 ***
#> cyl::8      -11.56364   1.298623 -8.90453 8.5682e-10 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including an intercept also helps with certaint statistical inference. The estimates test that the average of y for the omitted group is *the same* for the other groups

- Rejecting this (p-value $< \alpha$) rejects the null that the two means are the same

Roadmap

Bivariate Regression

Prediction vs Causation

Regression Inference

Goodness of Fit

Influential Observations

Discrete Variables

$\log$ transformations

$\log$-transformation

In economics, it is common to see $\log$ transformed outcomes:

$$\log(w_i) = \beta_0 + \beta_1 \text{College Degree}_i + u_i$$

$\log$-transformation

In economics, it is common to see $\log$ transformed outcomes:

$$\log(w_i) = \beta_0 + \beta_1 \text{College Degree}_i + u_i$$

This specification changes our interpretation of the slope coefficients:

Having a college degree is associated with an increase in wages of $\beta_1 * 100$ percent

- E.g. if $\beta_1 = 0.02$, then a college degree is associated with a 2% increase in wages.

Derivation of $\log$-transformation interpretation

Compare two individuals: unit 1 with and unit 0 without a college degree. Then, we have

$$\log(w_1) - \log(w_0) = \beta_0 + \beta_1 - \beta_0$$

$$\implies \log(w_1/w_0) = \beta_1$$

$$\implies \log(1 + \frac{w_1 - w_0}{w_0}) = \beta_1$$

Derivation of $\log$-transformation interpretation

Compare two individuals: unit 1 with and unit 0 without a college degree. Then, we have

$$\log(w_1) - \log(w_0) = \beta_0 + \beta_1 - \beta_0$$

$$\implies \log(w_1/w_0) = \beta_1$$

$$\implies \log(1 + \frac{w_1 - w_0}{w_0}) = \beta_1$$

If you recall, exponentiating gets rid of the the $\log$

$$\frac{w_1 - w_0}{w_0} = \exp(\beta_1) - 1$$

Derivation of $\log$-transformation interpretation

$$\frac{w_1 - w_0}{w_0} = \exp(\beta_1) - 1$$

The left-hand side is our percent-change formula from high-school science class

Derivation of $\log$-transformation interpretation

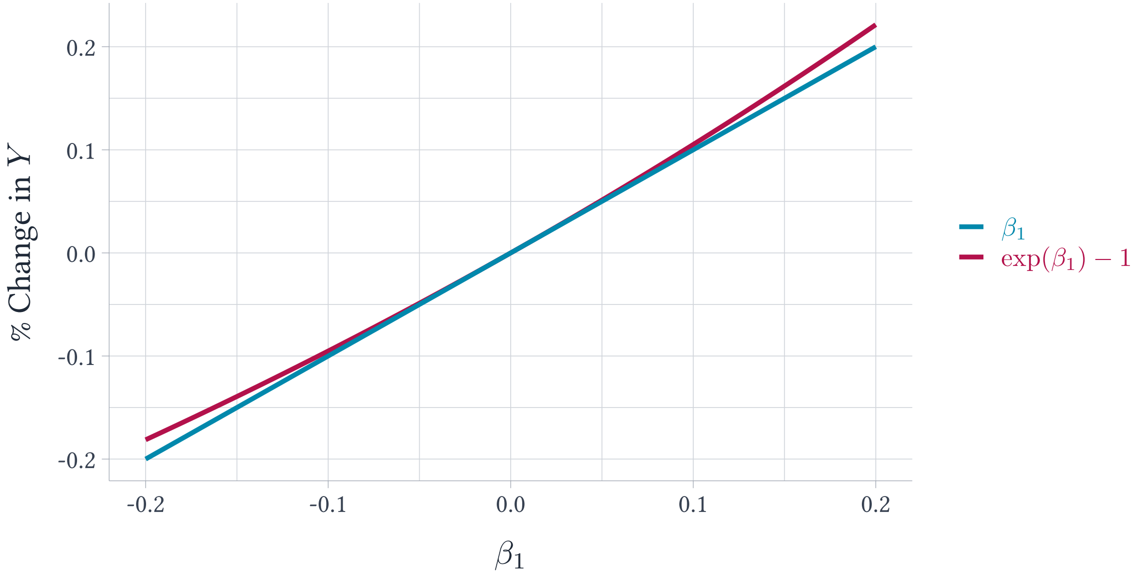
$$\frac{w_1 - w_0}{w_0} = \exp(\beta_1) - 1$$

The left-hand side is our percent-change formula from high-school science class

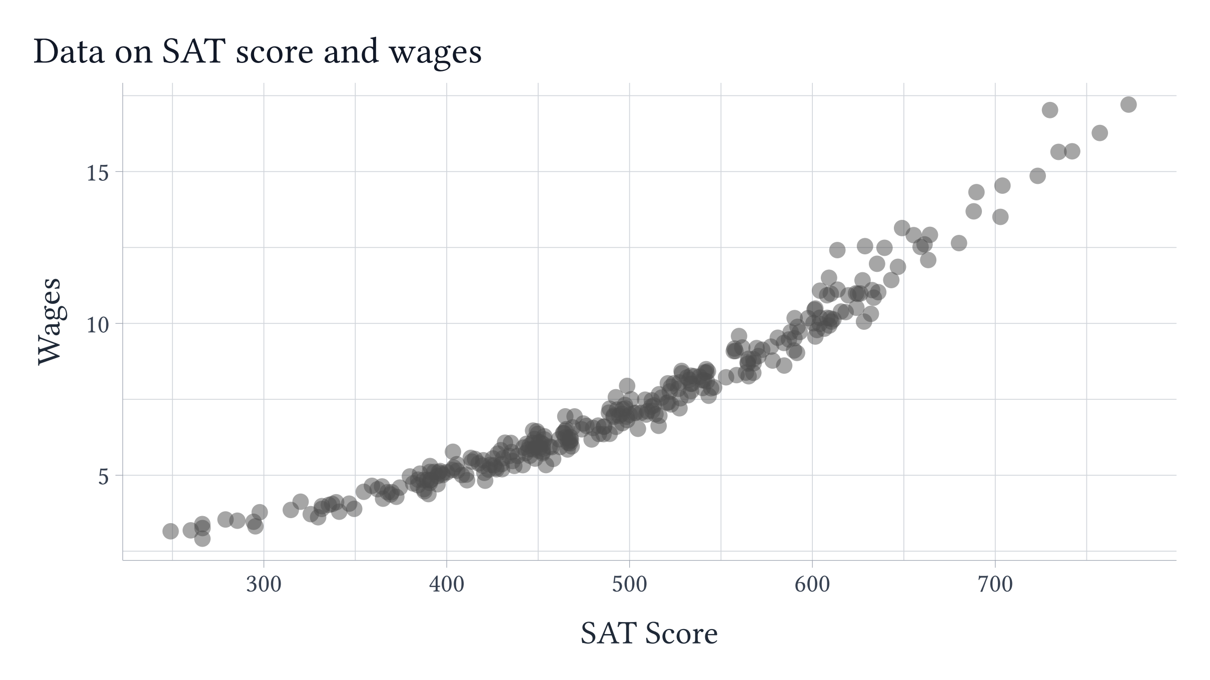
In this case, the more *precise* answer is that having a college degree is associated with an $\exp(\beta_1) - 1$ percent change in w

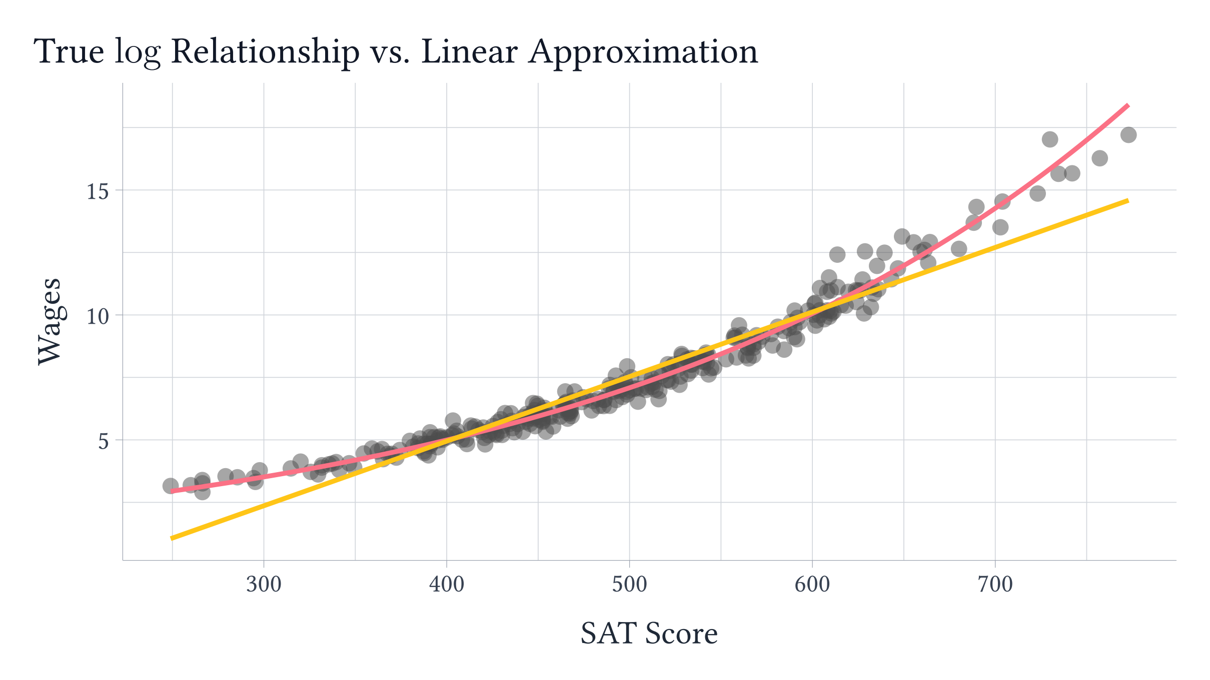
- But for $-0.10 < \beta_1 < 0.10$, $\exp(\beta_1) - 1$ is approximately equal to β_1 so it's simpler to use the latter

Comparison of $\exp(\beta_1) - 1$ and β_1



Example


Data on SAT score and wages
Wages
SAT Score
15
10
5
300
400
500
600
700


True log Relationship vs. Linear Approximation
Wages
15
10
5
300
400
500
600
700
SAT Score

When to use $\log$ transformations

You should take the $\log$ of an outcome varible when you think a 1 unit change in X is related to a % change in Y.

When to use $\log$ transformations

You should take the $\log$ of an outcome varible when you think a 1 unit change in X is related to a % change in Y.

$\log(Y)$ ~ X is called fitting an ‘exponential’ relationship. These are common in:

1. Financial markets where compounding returns imply $Y_t = Y_0 e^{rt}$
2. Epidemiology where disease growth rate is exponential (it is not actually, but early growth rate is approximately)

When to use $\log$ transformations

You should take the $\log$ of an outcome varible when you think a 1 unit change in X is related to a % change in Y.

$\log(Y)$ ~ X is called fitting an ‘exponential’ relationship. These are common in:

1. Financial markets where compounding returns imply $Y_t = Y_0 e^{rt}$
2. Epidemiology where disease growth rate is exponential (it is not actually, but early growth rate is approximately)
3. Settings with skewed distributions (e.g. home prices, GDP, population)
   - $\rightarrow$ Skewness makes a ‘unit’ change in X difficult to think about

$\log$-$\log$ transformations

Alternatively, You may see $\log$ transformations of both variables:

$$\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + u_i$$

The interpretation is now simpler: a 1% change in X_1 is associated with a β_1 % change in Y