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

$\log$ 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.

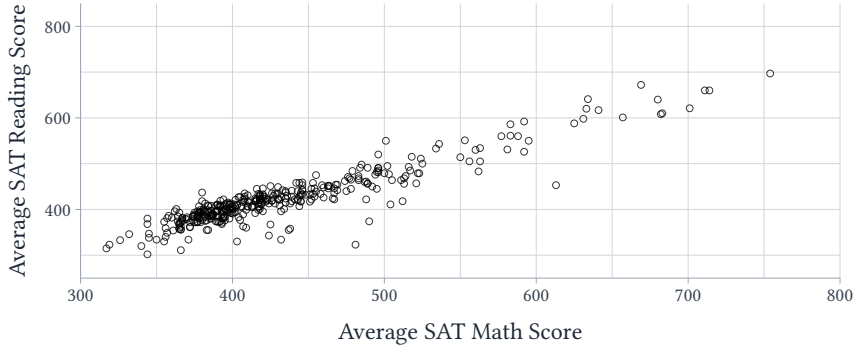
Practice question

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

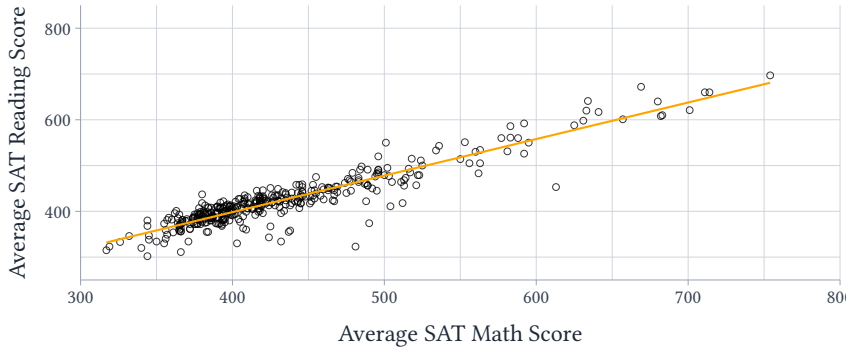
Relationship between X and Y

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

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

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

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

Regression Line Example

Predictions

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

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

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

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

$\implies$ 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

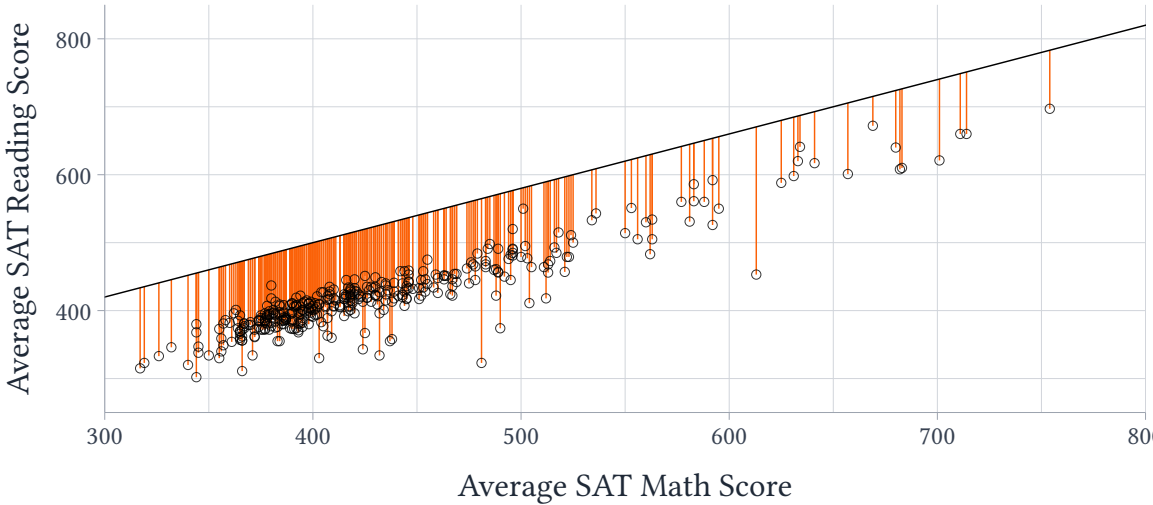
Just like in Topic 2, we can form the mean-square prediction error of our linear model (in our training sample):

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

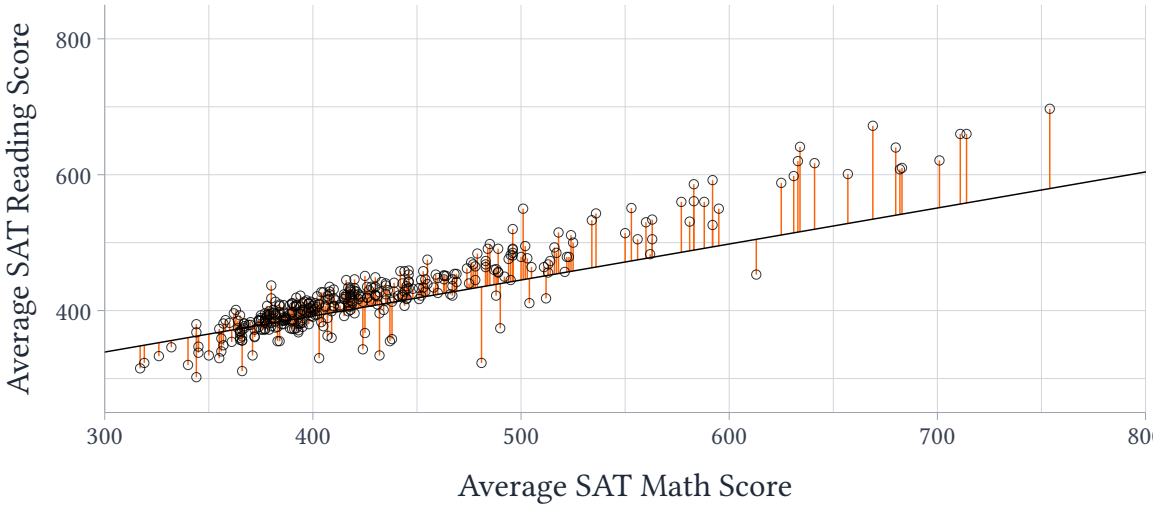
A line does a good job at predicting if MSE is (relatively) small.

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

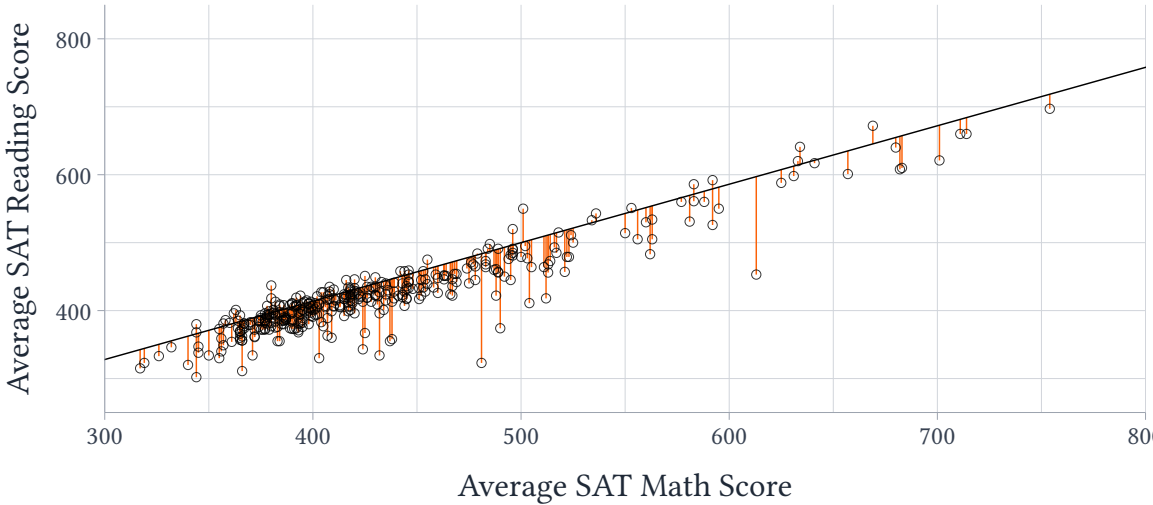
$\text{MSE}(b_0, b_1) = 10902$



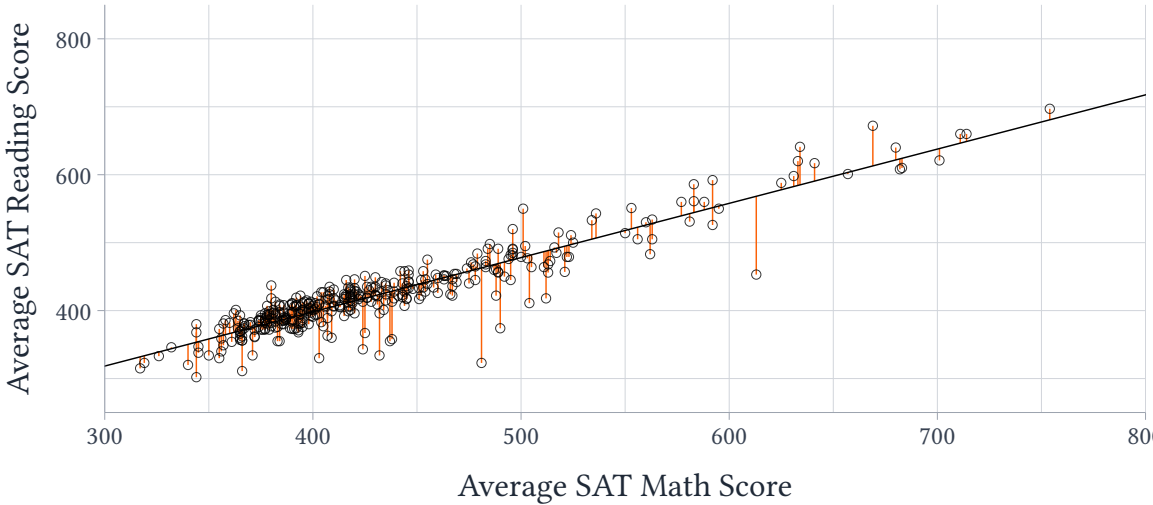
$\text{MSE}(b_0, b_1) = 1126$



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



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



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

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

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

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

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

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:

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

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

$\bar{X} = 432.94$

$\bar{y} = 424.50$

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:

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i) \\ &= \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i \\ &= \hat{\beta}_0 + \hat{\beta}_1 \bar{X} + 0\end{aligned}$$

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

$\log$ 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

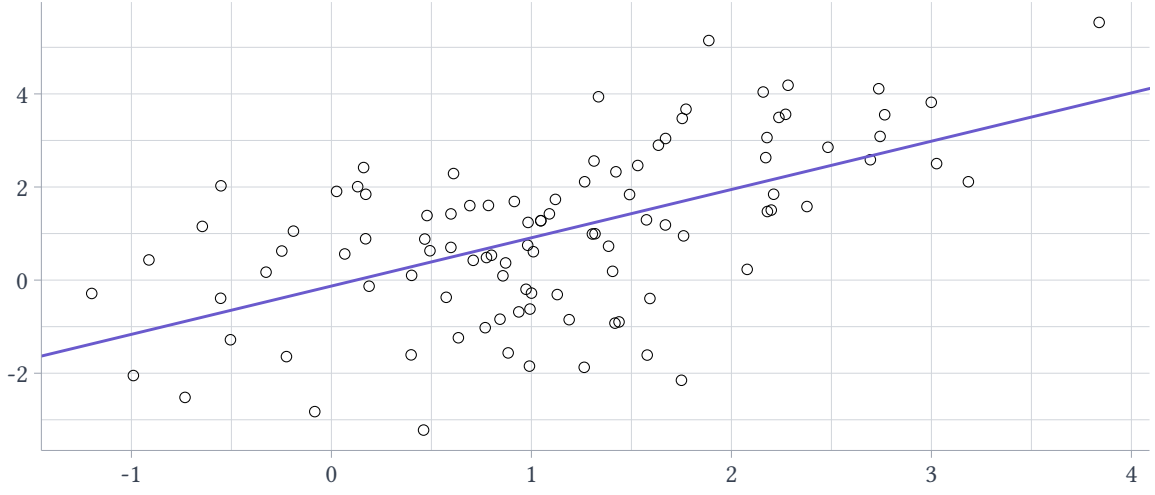
As we have seen, the regression coefficient $\hat{\beta}_1$ is often of interest

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

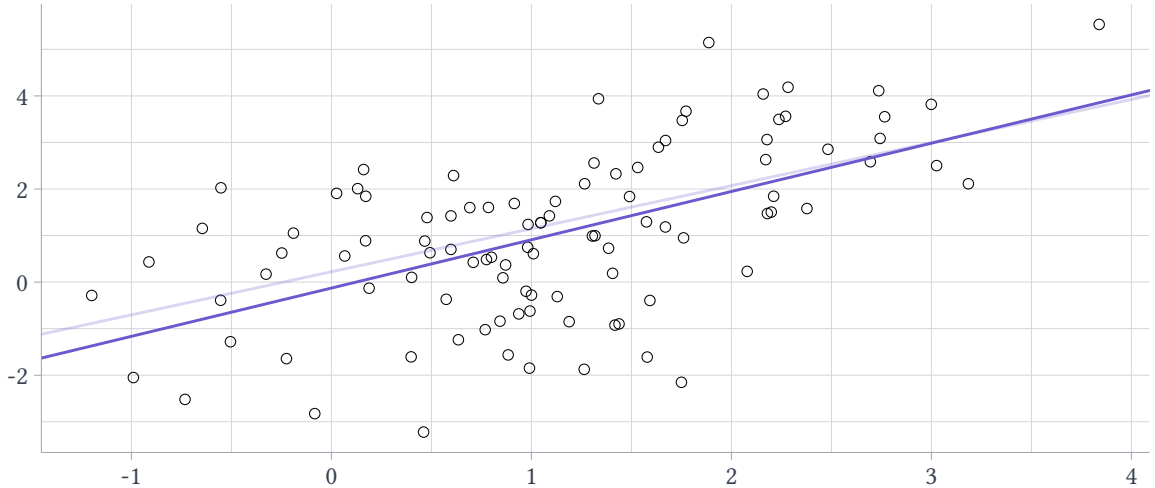
We want to be able to describe the uncertainty around this estimate. How does $\hat{\beta}_1$ change under repeated sampling?

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

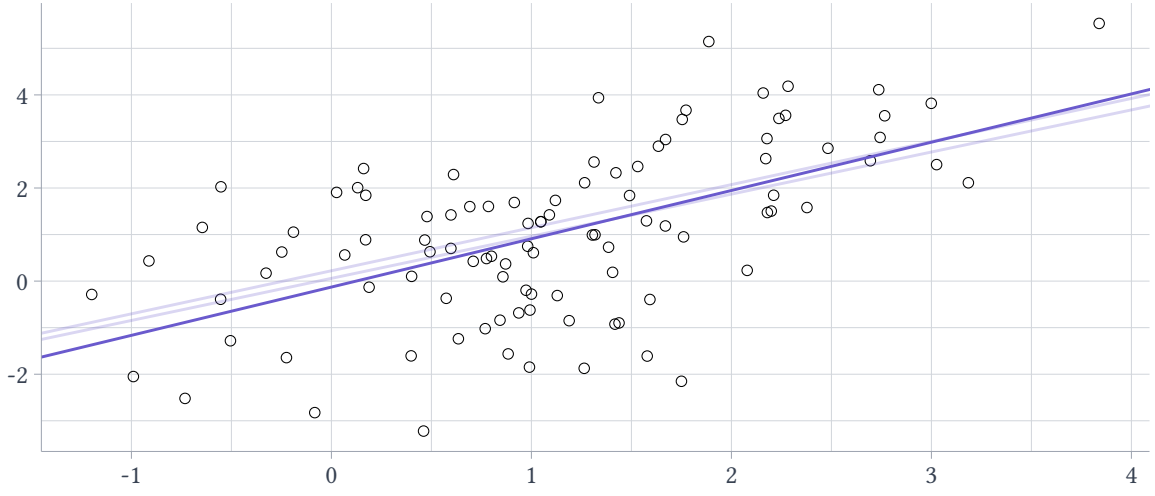
Original Sample



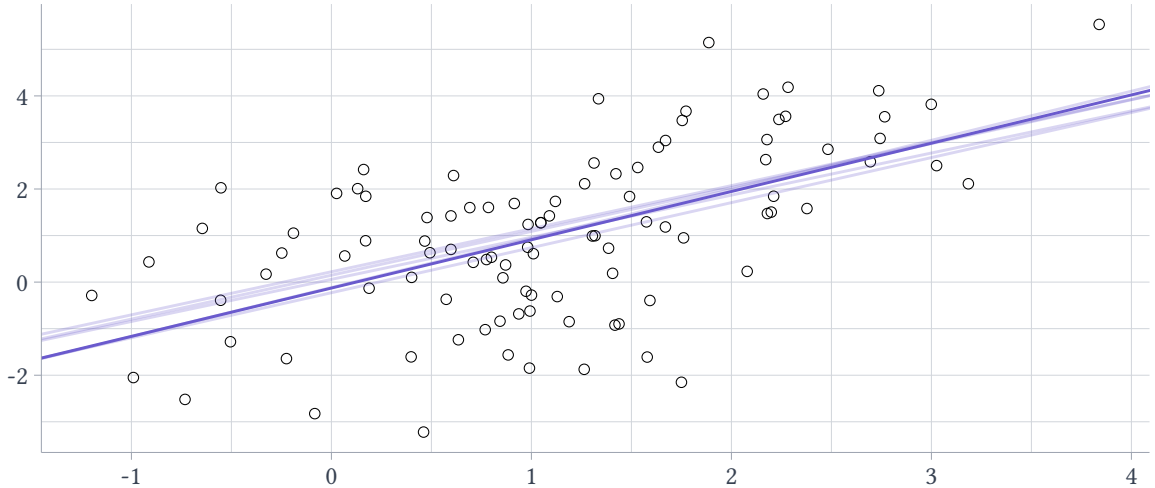
Original Sample + 1 Extra Sample



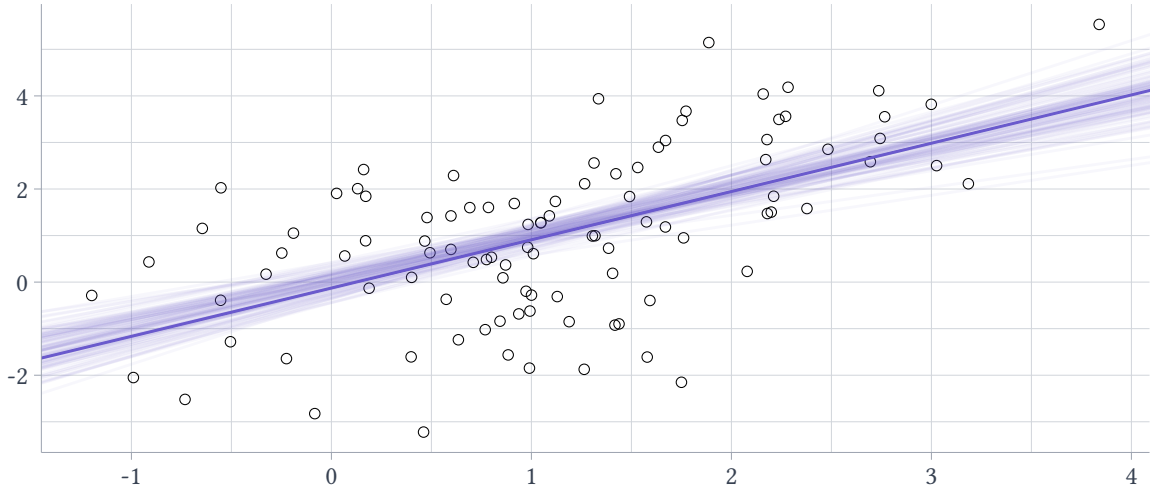
Original Sample + 2 Extra Samples



Original Sample + 5 Extra Samples



Original Sample + 100 Extra Samples



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:

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

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 $\hat{\beta}_1$:

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

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

Standard Error

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

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

Roadmap

Bivariate Regression

- Prediction vs Causation

Regression Inference

Goodness of Fit

Influential Observations

$\log$ transformations

R^2

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

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

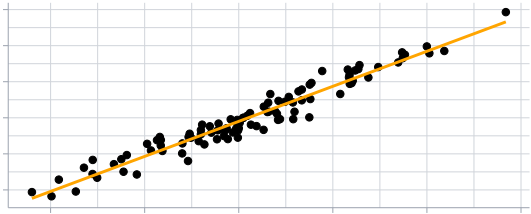
Intuition of R^2

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

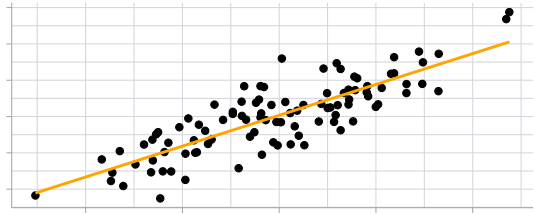
$$R^2 = \frac{\text{variation in } \hat{y} \text{ along the regression line as x varies}}{\text{total variation in observed values of y}}$$

Comparisons of R^2

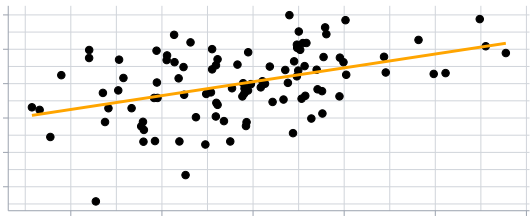
$R^2 = 0.943$



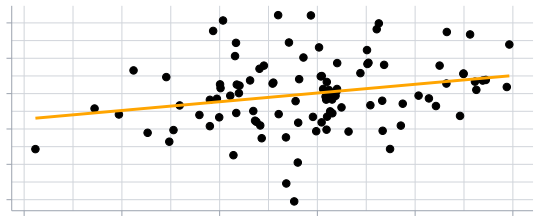
$R^2 = 0.698$



$R^2 = 0.189$



$R^2 = 0.058$



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

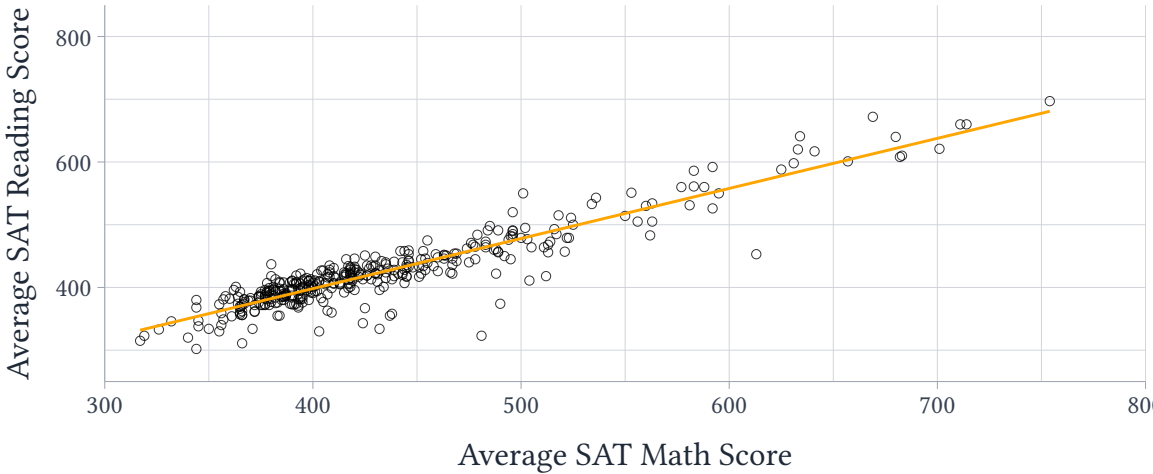
Influential Observations

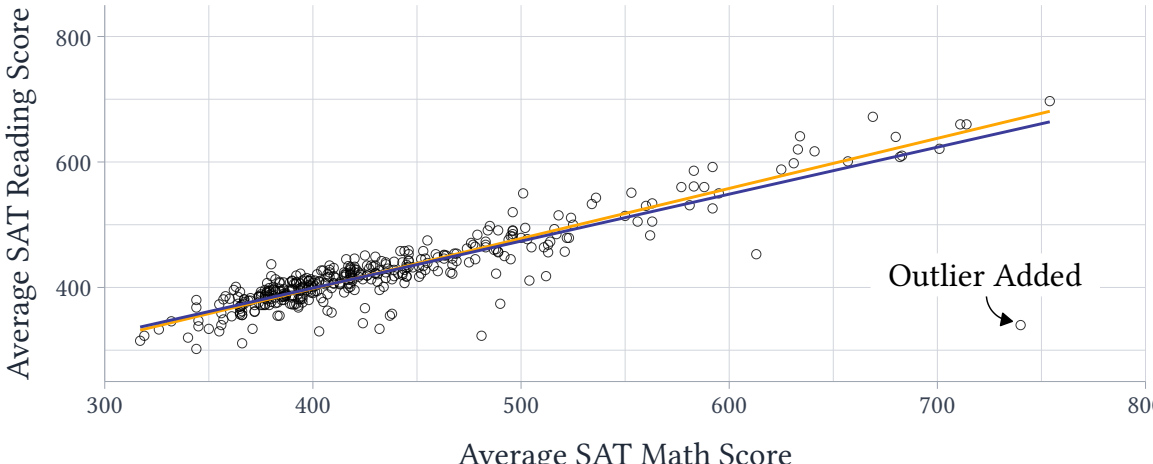
$\log$ transformations

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

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

Roadmap

Bivariate Regression

- Prediction vs Causation

Regression Inference

Goodness of Fit

Influential Observations

$\log$ transformations