Introduction to Empirical Methods

Lecture 2, Part 2: Ordinary Least Squares

Where do we go from here?

- Theories $\rightarrow$ Hypotheses $\rightarrow$ Models
- How do we think the world works?
- We typically assume additive:
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$
    - Some independent variables (X_1 & X_2) cause changes in some dependent variable (Y)

Bivariate Regression

- Our Hypothesis: An independent variable causes changes in the dependent variable.
- $Y = \alpha + \beta X$
- Y: Dependent Variable
- X: Independent Variable
- α: Constant
- β: Coefficient
  - A one unit change in X results in a β unit change in (the expected value) of Y.

A Blast from the Past

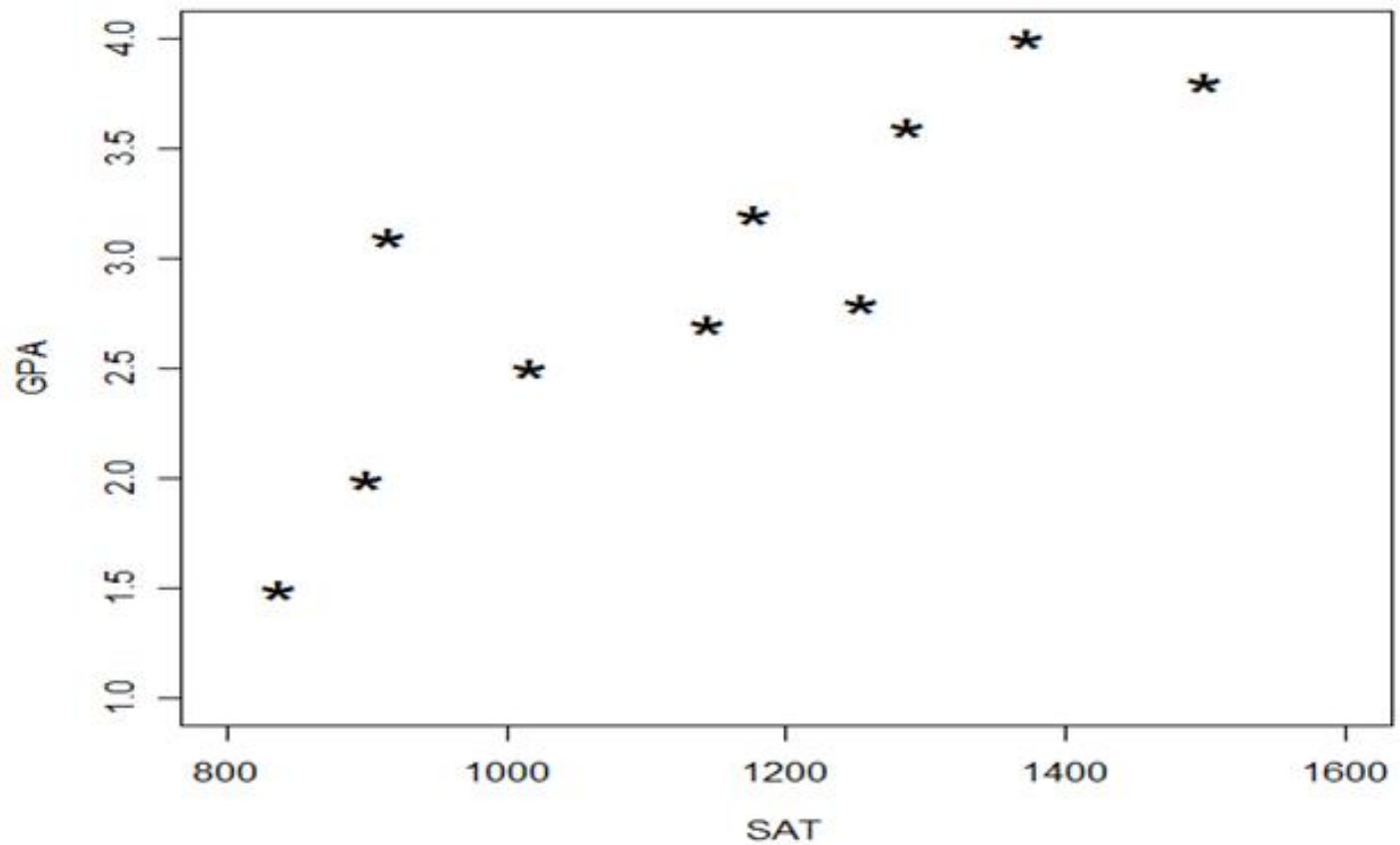
- Think back to high school algebra.
- This is just like lines and cartesian coordinates.
- $Y = \alpha + \beta X$
  - α is also known as the Y-Intercept.
  - β is also known as the slope.

Bivariate Regression Example with Fake Data

- SAT scores are supposed to estimate a student's first year GPA.
  - Independent Variable?
  - Dependent Variable?
- $GPA = \alpha + (\beta \times SAT)$

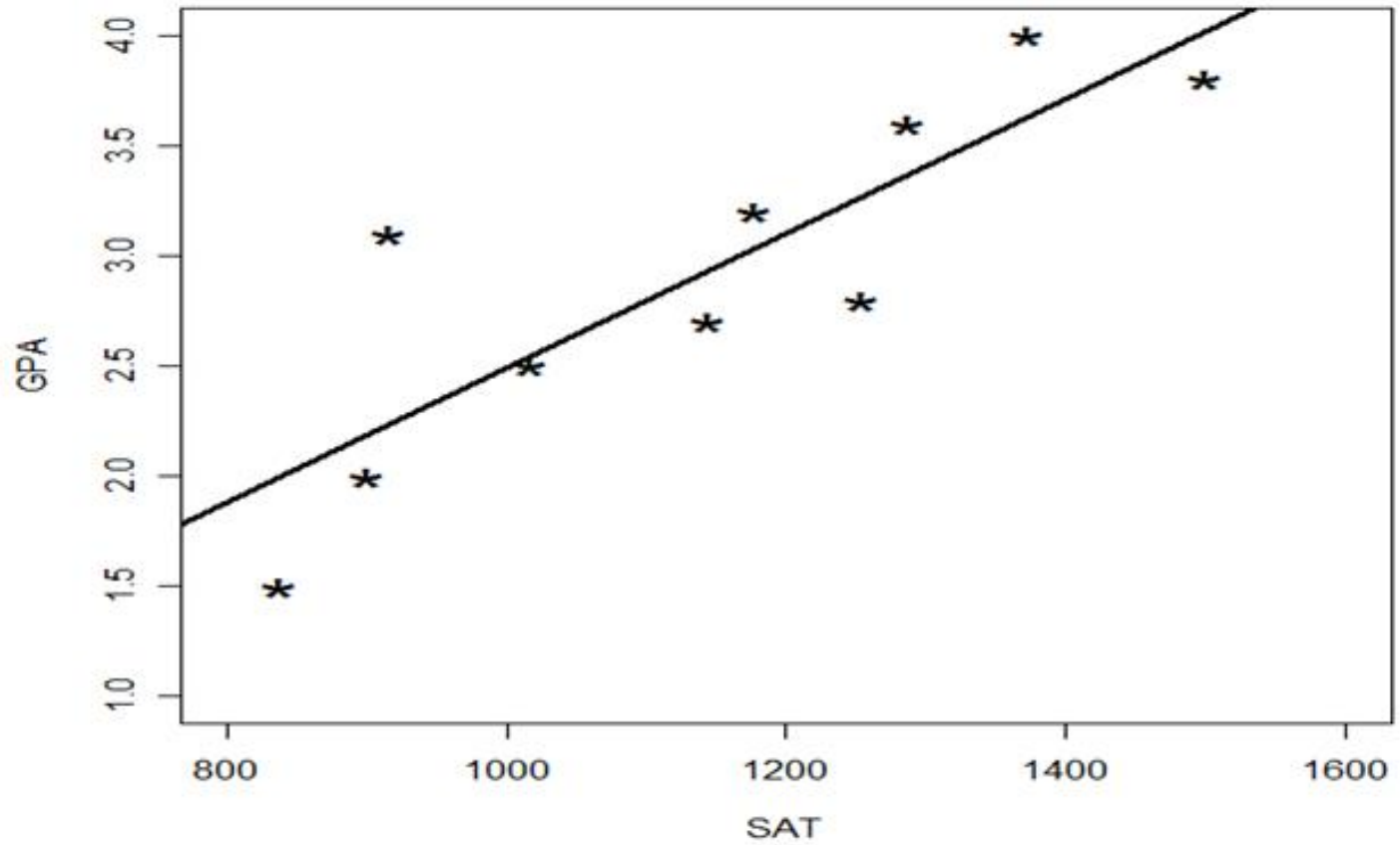
What are we doing?

- Start with a scatterplot.



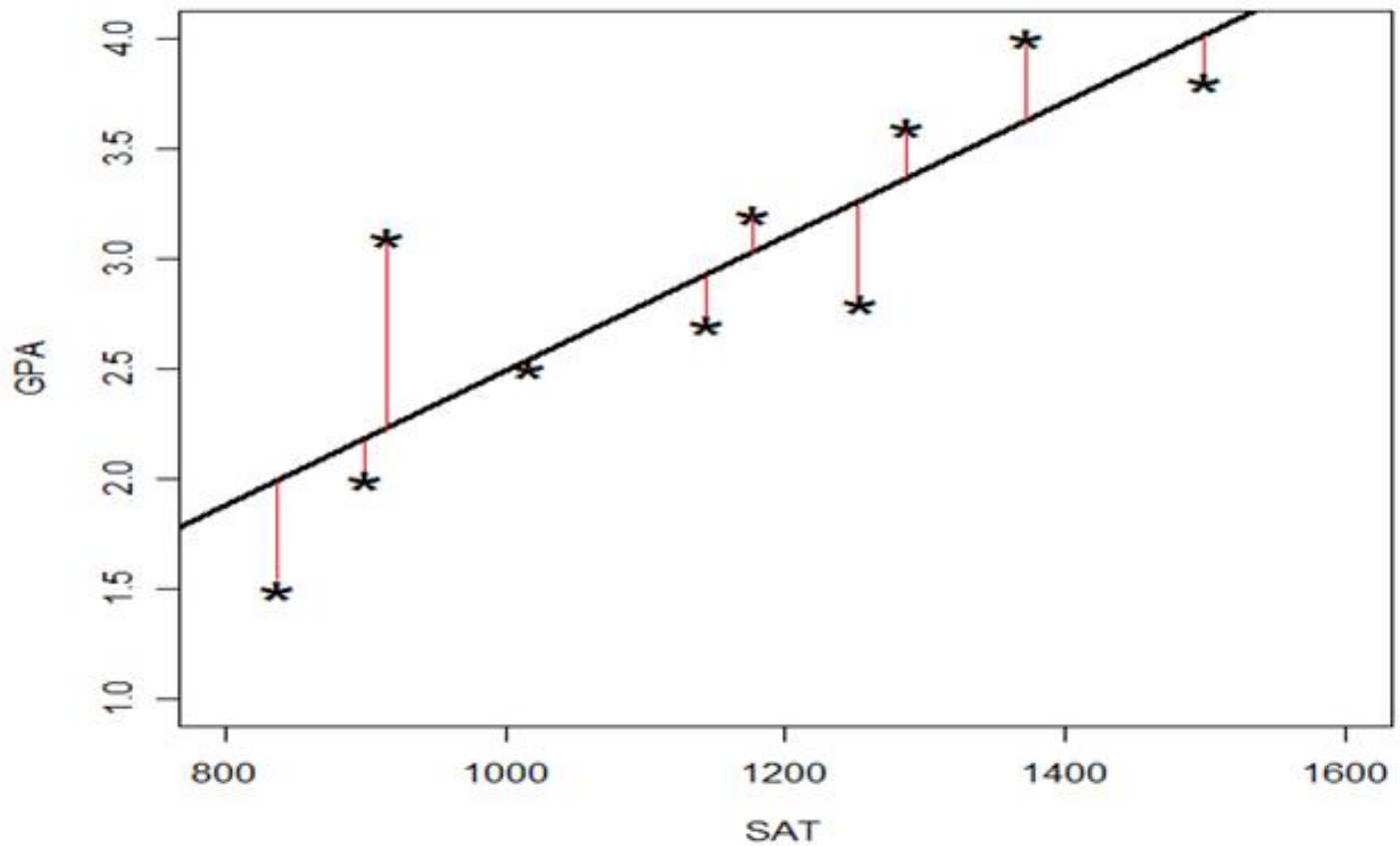
What are we doing?

- Draw the line that...



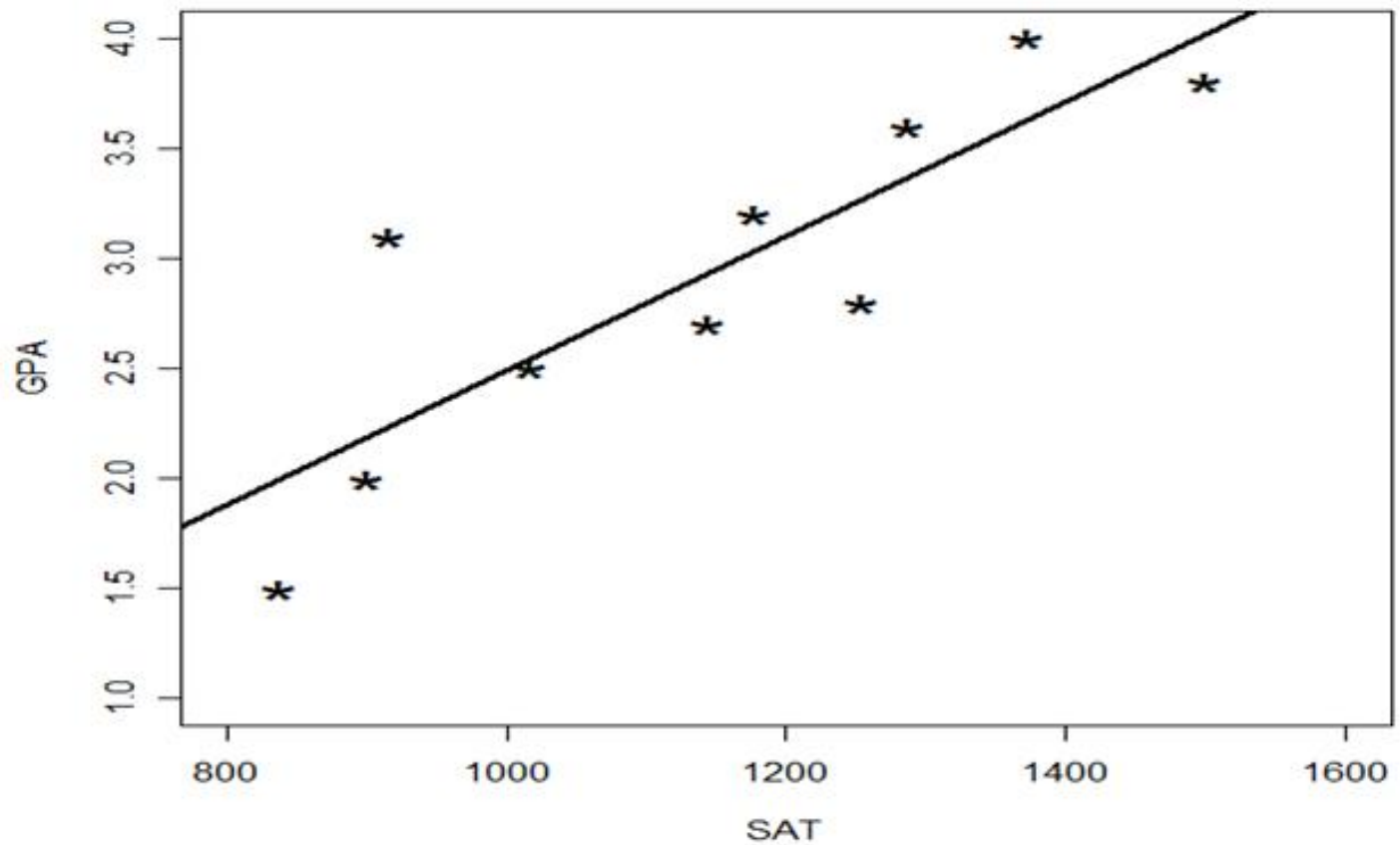
What are we doing?

- ...minimizes the sum of the squared errors.



What are we doing?

- $\hat{\alpha} = -0.57$ and $\hat{\beta} = 0.003$



Constants and Slopes

- a is our estimate of α
- b is our estimate of β
- $Y = a + bX$
- $a = -0.57$
  - a is one's expected GPA if that person had an SAT score of zero.
- $b = 0.003$
  - b is the increase in one's expected GPA for each additional point one does better on the SAT

O.L.S.

- This method of minimizing the sum of the squared errors is called "ordinary least squares" (O.L.S.)
- We may use O.L.S. if:
  - Our dependent variable is continuous and unbounded.
  - Our dependent variable is normally distributed.

Estimate in Bivariate Regression

- You do not need to memorize these.
- $\hat{\beta} = b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$
- $\hat{\alpha} = a = \bar{Y} - \hat{\beta}\bar{X}$
- Things to note:
  - The only information you need to calculate α and β are the values of the dependent and independent variables.
  - The formulas both rely on the means of the variables.
    - A single unusual value (outlier) can really affect the mean, so a single outlier can substantially affect these statistics as well.
  - These are estimates of the true, unknown α and β in the population.

Uncertainty

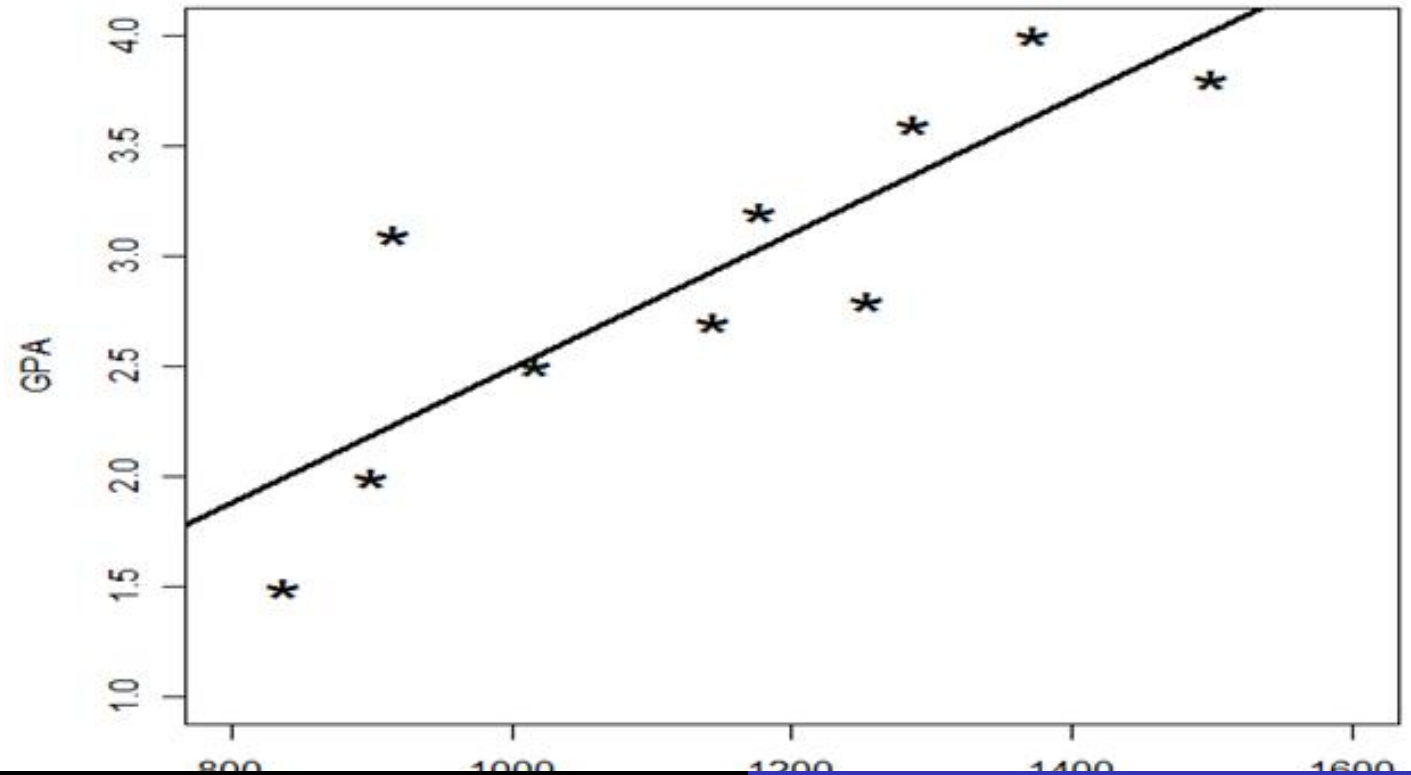
- Since these are estimated based on a sample, we are uncertain about the actual values of α and β.
  - That uncertainty is reduced as our sample size increases.
- We call our measure of uncertainty the standard error.
  - As usual.
- There is a standard error for α and a standard error for β.
  - The smaller the standard errors, the more confident we are that our estimates are equal to the true values.

Uncertainty Around α

- The uncertainty around α is often important.
- However, for our purposes we are less concerned about α.
  - Why is that?
- α is not directed related to our hypothesis test.
  - Usually.
- α is important for calculating predictions of the dependent variable, but that's not usually our goal.
- Our goal is to determine if the independent variable affects the dependent variable.
  - And that's related to the uncertainty around β.

What are we doing?

- Remember the example
- $\hat{\alpha} = -0.57$ and $\hat{\beta} = 0.003$



Estimates in Bivariate Regression

Variable	Coefficient	Standard Error
Constant	-0.5668	0.7786
SAT	0.0031	0.0007

- We can now determine whether the relationship between SAT and GPA is "Statistically Significant."

Review

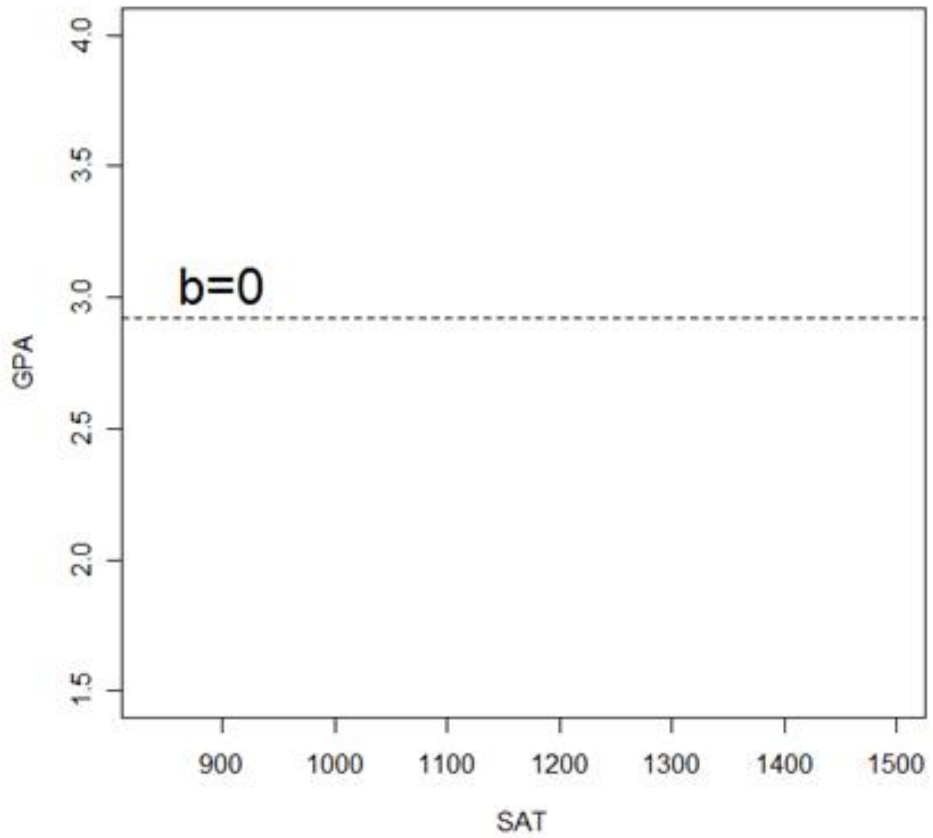
- The structure of this hypothesis test is like our previous hypothesis tests.
- The χ^2 test compared the observed table to the table we would have expected to have seen had there been no relationship.
  - This time we are going to compare our observed β to the value of β we would expect if there were no relationship.
- The difference of means test was a *t*-test in which we compared two values and divided the difference by the standard error.
  - This will be a *t*-test in which we compare two values and divide the difference by the standard error.
    - Though, in reality it will be even easier than that.

Statistical Significance

- Once again, we are looking to reject the null hypothesis.
- Non-Directional Hypothesis?
  - The null hypothesis is $\beta = 0$.
  - We use a **two-tailed** test to check if the coefficient is different from zero in either direction.
- Directional Hypothesis?
  - The null hypothesis is either that $\beta = 0$, or that the relationship is in the other direction.
  - We can use a **one-tailed** test to check if the coefficient is different from zero in the direction of our hypothesis.

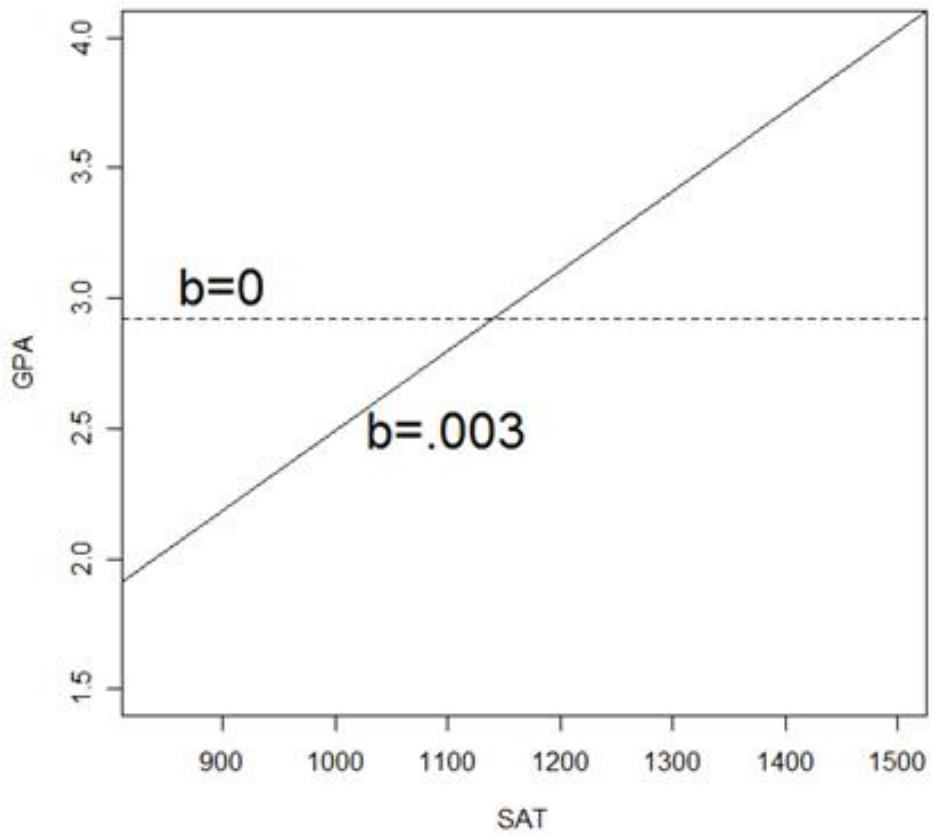
The Null Hypothesis

- The null hypothesis is $\beta = 0$.
- What does β tell us about the impact of X upon Y?



The Null Hypothesis

- The null hypothesis is $\beta = 0$.
- What does β tell us about the impact of X upon Y?



Testing for Statistical Significance

- T-Test
  - $t = \frac{\hat{\beta} - 0}{s_{\hat{\beta}}}$
  - $t = \frac{Coefficient}{Standard\ \ Error\ \ of\ \ Coeff}$
- Degrees of Freedom
  - $d.f. = n - \#$ of parameters
    1. Constant
    2. # of Coefficients
  - For bivariate regression
    - $d.f. = n - 2$

Statistical Significance

Variable	Coefficient	Standard Error
Constant	-0.5668	0.7786
SAT	0.00306	0.00067

- $t = \frac{Coefficient}{Standard\ \ Error\ \ of\ \ Coeff} = \frac{0.00306}{0.00067} = 4.552$

Statistical Significance

Variable	Coefficient	Standard Error
Constant	-0.5668	0.7786
SAT	0.00306	0.00067

- $t = \frac{Coefficient}{Standard\ \ Error\ \ of\ \ Coeff} = \frac{0.00306}{0.00067} = 4.552$

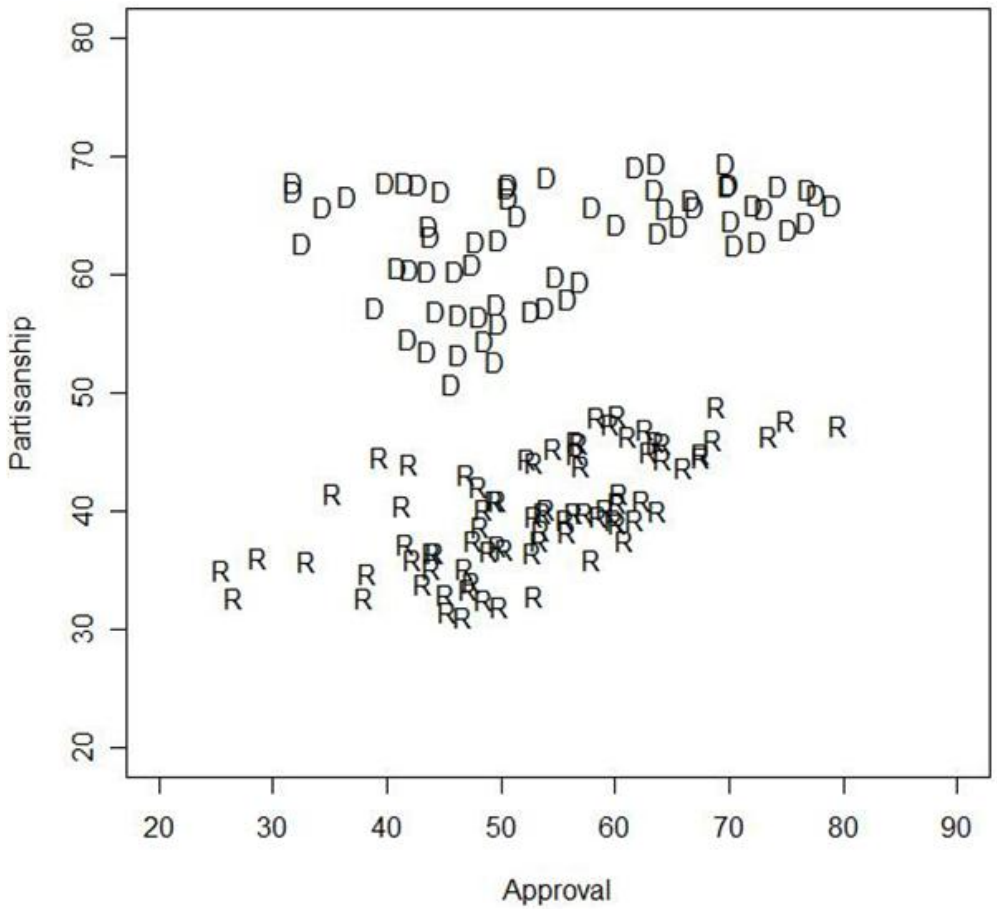
d.f.	.10	.05	.01
8	1.869	2.306	2.896

- We can reject the null hypothesis at the .10 level, the .05 level and the .01 level in a two-tailed test.
- So, we can say there is a statistically significant relationship between SAT score and GPA.

Another Example

- Our theory: When the president is more popular, more people say that they are members of the president's party.
- Data:
  - Quarterly Gallup survey data on average presidential approval and percent of the respondents who say they are members of the president's party.
  - Years: 1960-1996
- Independent Variable: %Approval
- Dependent Variable: %Party
  - $Y = \alpha + \beta X$
  - $\%Party = \alpha + (\beta \times \%Approval)$

Approval and Partisanship



O.L.S. Model Results

Estimate	Coefficient	Standard Error
Constant	36.58	4.688
Approval	0.255	0.085

- Does presidential approval have a statistically significant effect on partisanship?

O.L.S. Model Results

Estimate	Coefficient	Standard Error
Constant	36.58	4.688
Approval	0.255	0.085

- Does presidential approval have a statistically significant effect on partisanship?
- $t = \frac{0.255}{0.085} = 2.99$

d.f.	.10	.05
142	1.656	1.98

Predicted Values

- We can use predicted values to help understand the size of the effects.
- $Y = \alpha + \beta X$
- Pick two values of interest:
- George H.W. Bush
  - Highest Approval: 80%
  - Lowest Approval: 35%
- $Y = 36.58 + 0.255 * 80 = 56.98$
- $Y = 36.58 + 0.255 * 35 = 45.5$
- So, one of the largest popularity swings changes partisanship by about 11.5% according to our model.
  - Remember, there is uncertainty around these predictions.

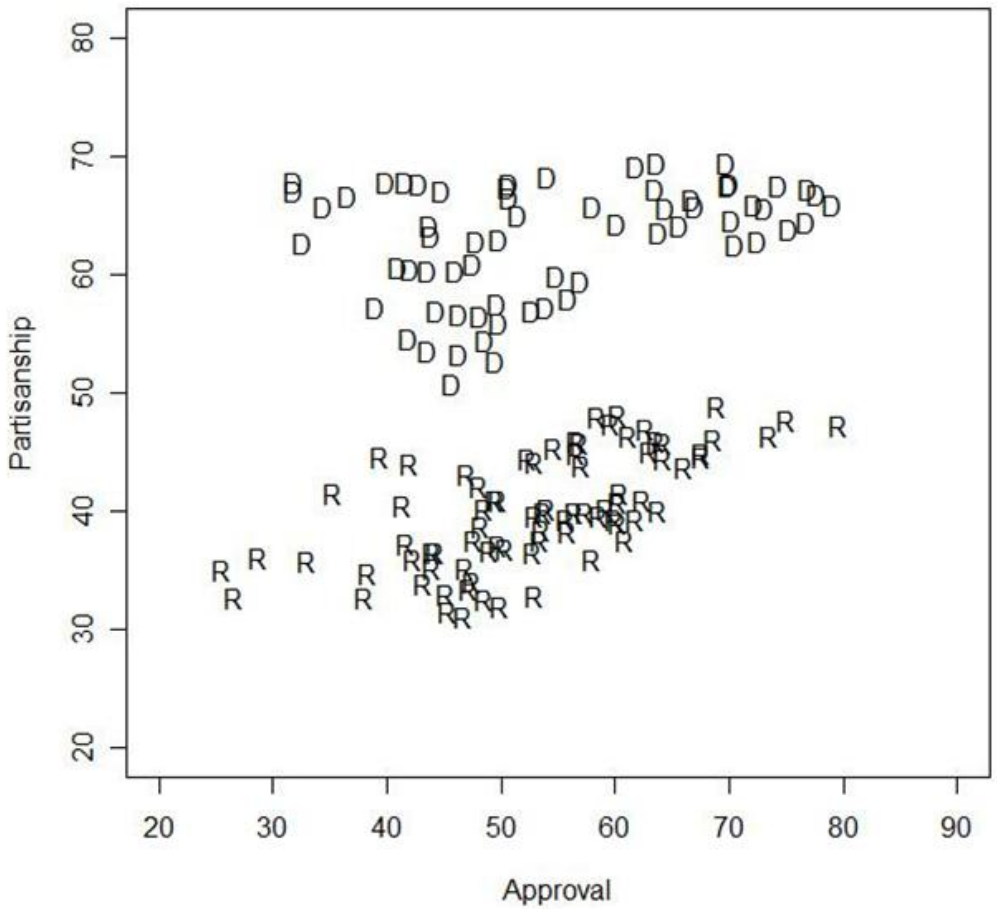
More Predicted Values

- Ronald Reagan
- Highest Approval: 64%
- Lowest Approval: 38%
- $Y = 36.58 + 0.255 * 64 = 52.97$
- $Y = 36.58 + 0.255 * 38 = 46.3$

Predicted Value

- The expected value of the dependent variable given the specified value of the independent variable.
- Remember, there is error.
  - e.g., Look at the Bush and Reagan partisanship estimates at their highest approval levels.
    - Bush=56.98; Reagan=52.97

Approval and Partisanship



Bivariate Regression Model

- What did the model do?
  - It drew a straight line through the points.
  - None of the actual observed points fall on that line.
- So, the model does not do a perfect job of predicting values of the dependent variable, though it might get close in some cases.
- But it can tell us if the independent variable is affecting the dependent variable.

Assumptions of the O.L.S. Model

- Linearity
  - A straight line adequately represents the relationship in the population
  - Fitting a linear model to a nonlinear relationship results in biased estimates

Assumptions of the O.L.S. Model

- Independent Observations
  - More specifically, the values of the dependent variable are independent of each other.
  - Time series data, panel data, and clustered data often do not satisfy this condition.
  - The estimates are unbiased, but the standard errors are typically biased downwards.
    - This means we're more likely to *mistakenly* reject the null hypothesis.

Things To Watch Out For

- Linearity
  - Is the relationship between the independent and dependent variable in a straight line?
- Outliers
  - Does a case have an unusual value of its dependent value given the value of its independent variable?
- Leverage
  - Does the case have an unusual value for its independent variable?
    - i.e., is it far from the mean of the independent variable?
- Influence
  - Does an outlier case have high leverage?

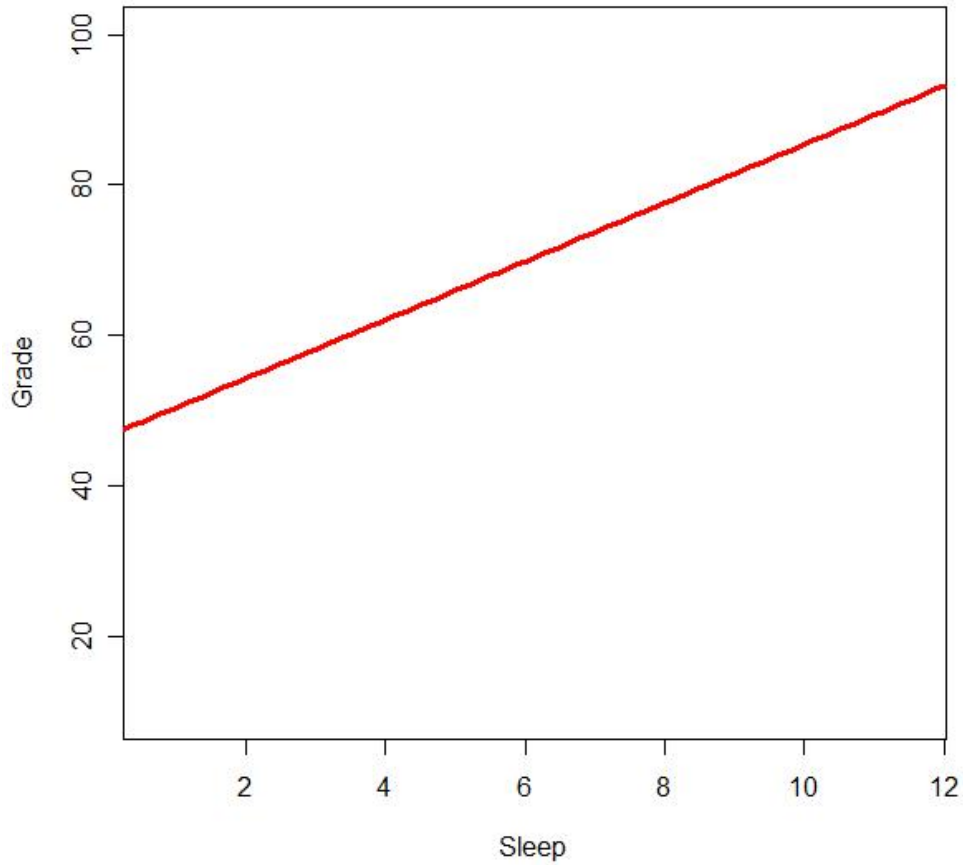
Linearity

- O.L.S. assumes the relationship between the independent variable and the dependent variable is linear.
  - Research Question: What is the relationship between sleep the night before an exam and grade on the exam?

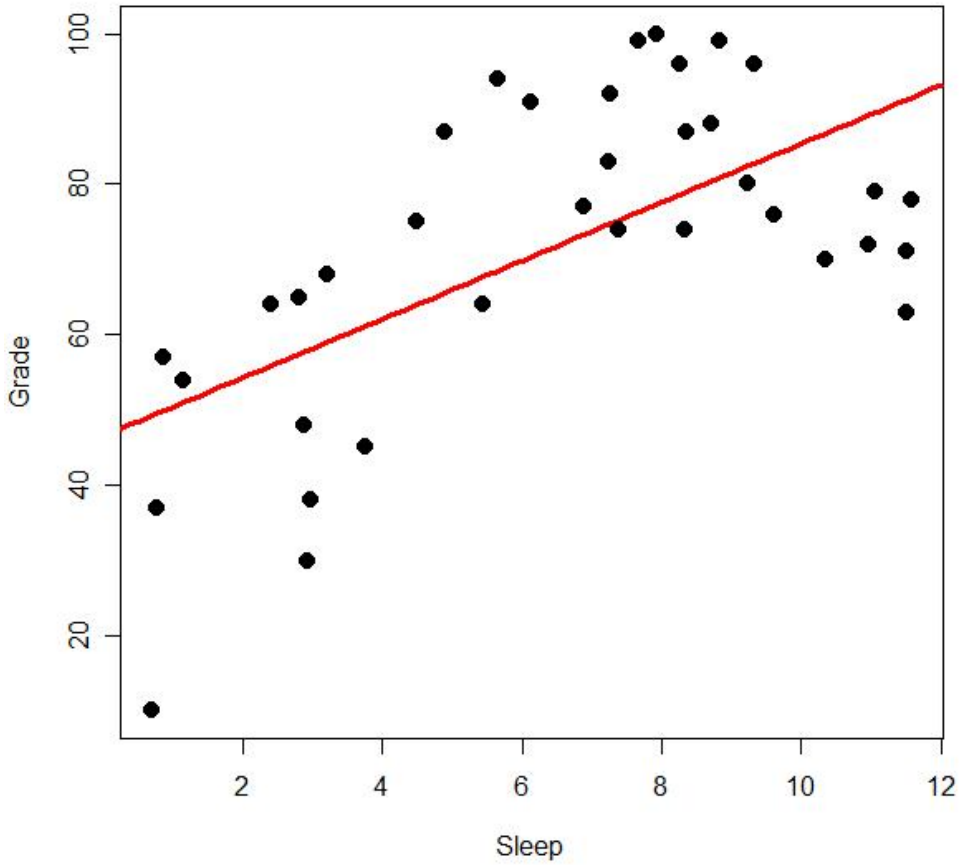
	Estimate	Standard Error	T-Ratio
Constant	46.53	6.11	7.62
Sleep	3.89	0.84	4.64

- For every hour of sleep....

Linearity

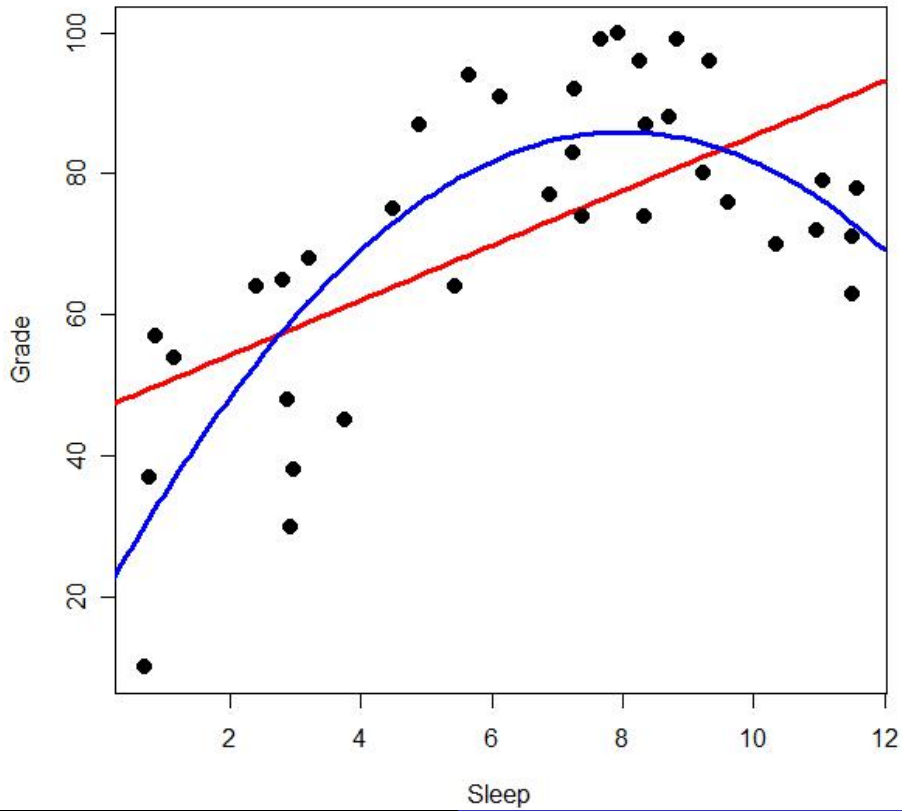


Linearity



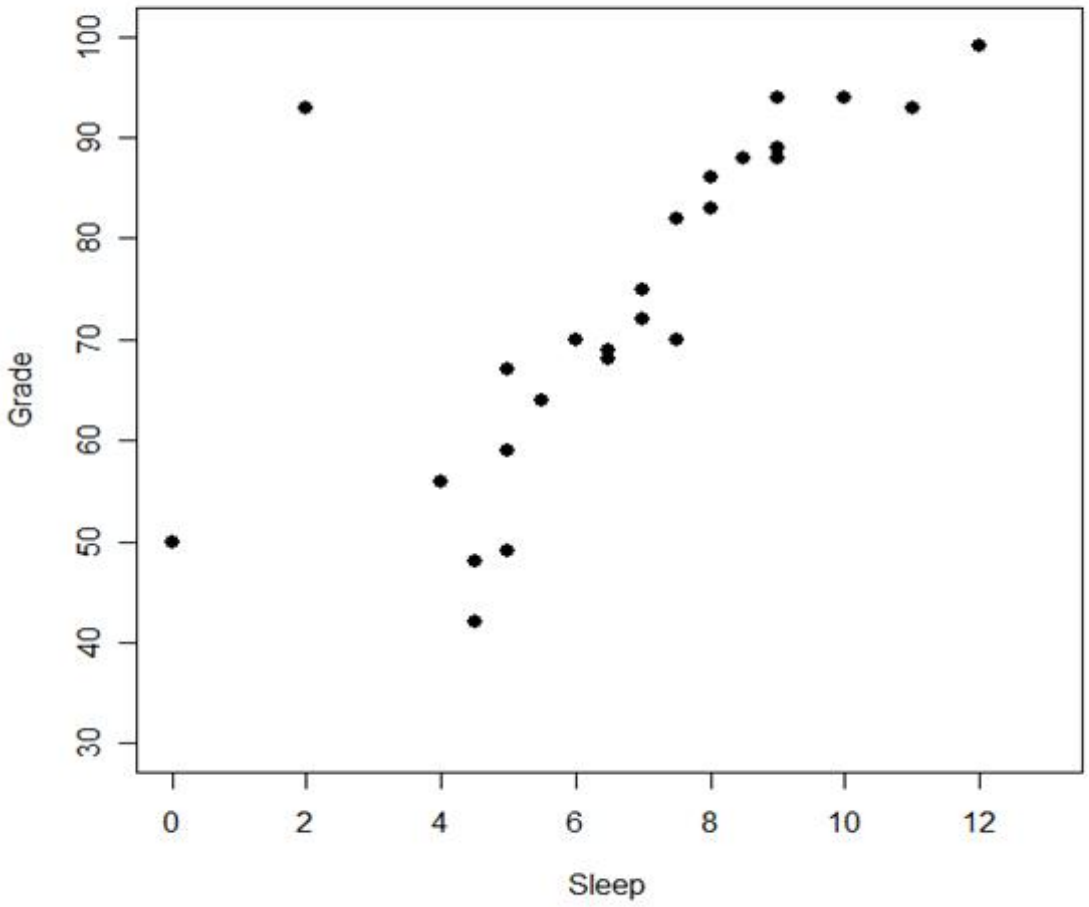
Coping with Linearity

- This can be dealt with...
  - $Grade = \alpha + \beta_1 Sleep + \beta_2 Sleep^2$
- But don't worry about the fix, understand the problem.



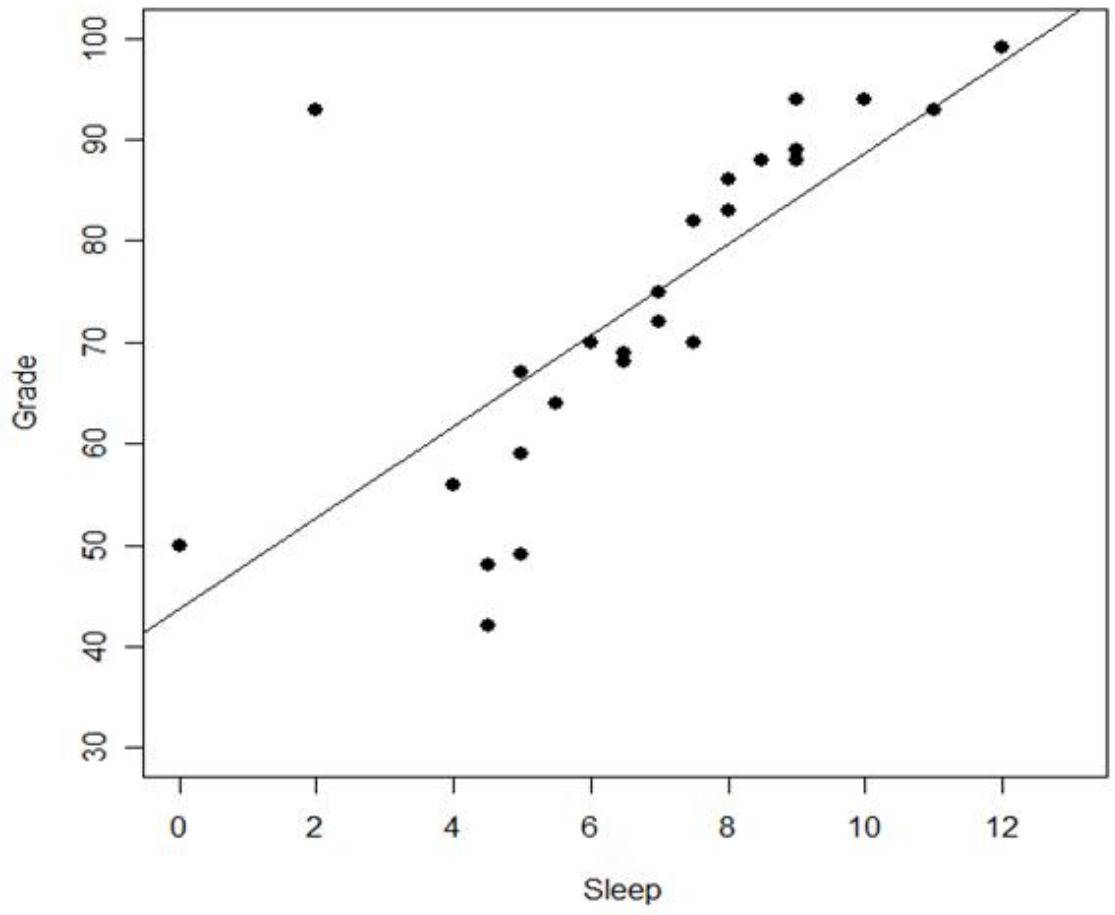
Outliers

- When a case has an unusual Y value given its X value.



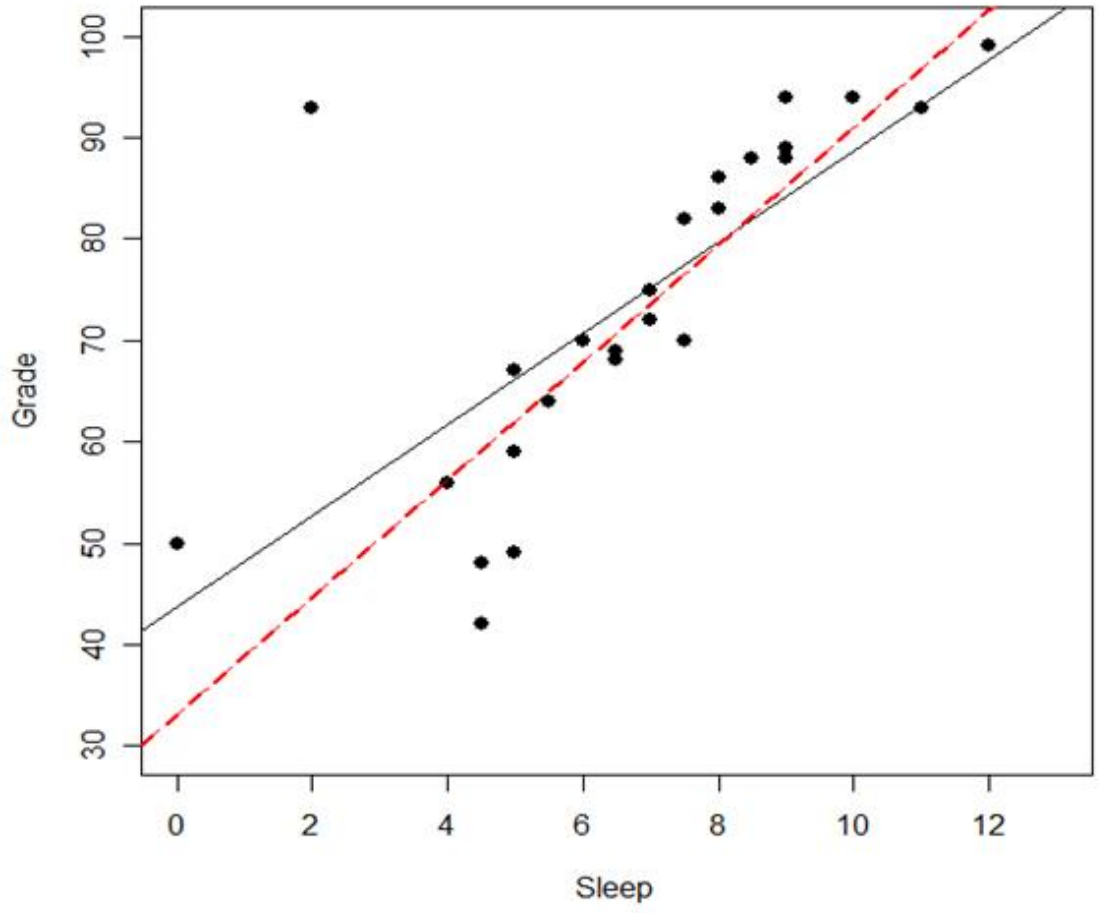
Outliers

- When a case has an unusual Y value given its X value.



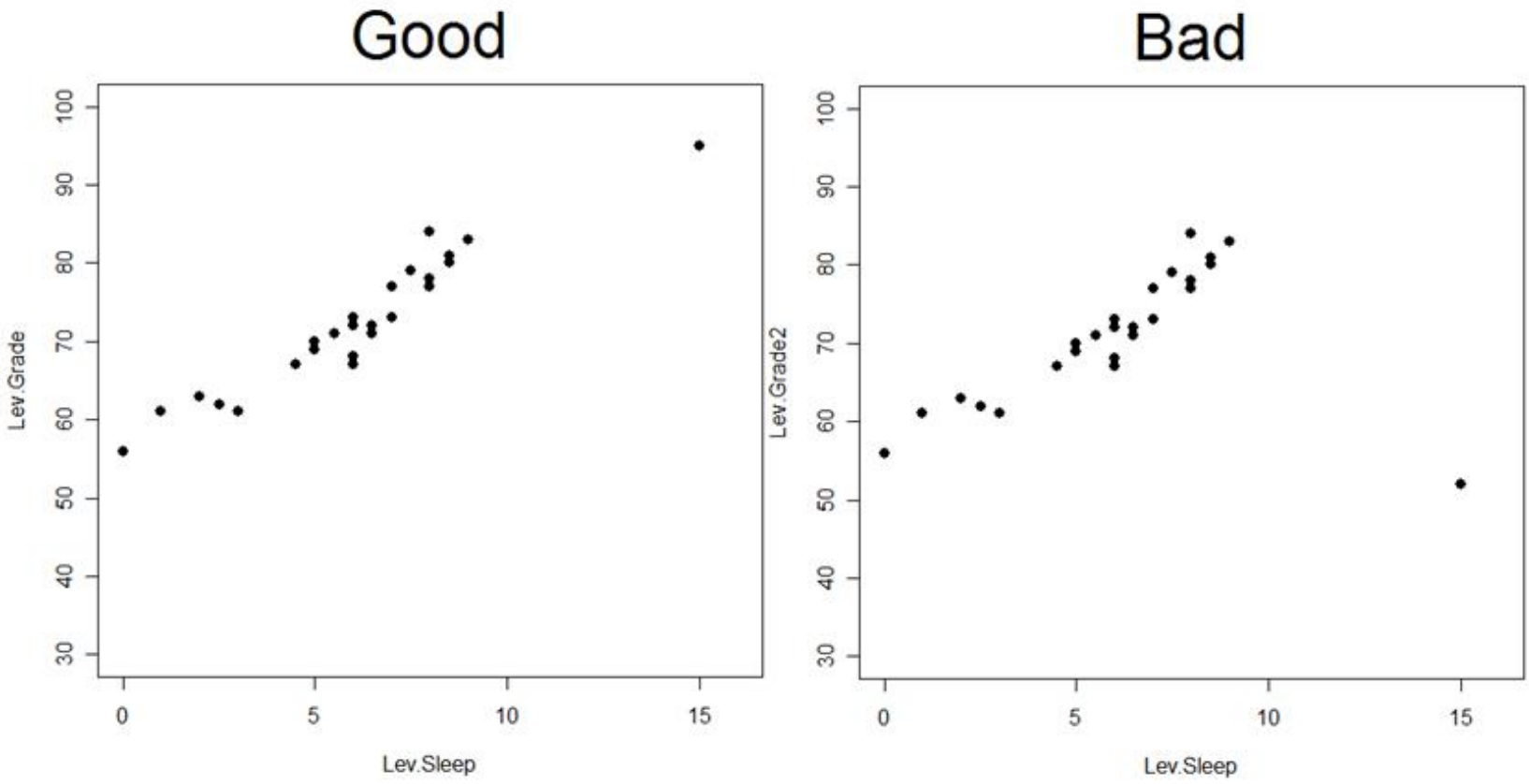
Outliers

- When a case has an unusual Y value given its X value.



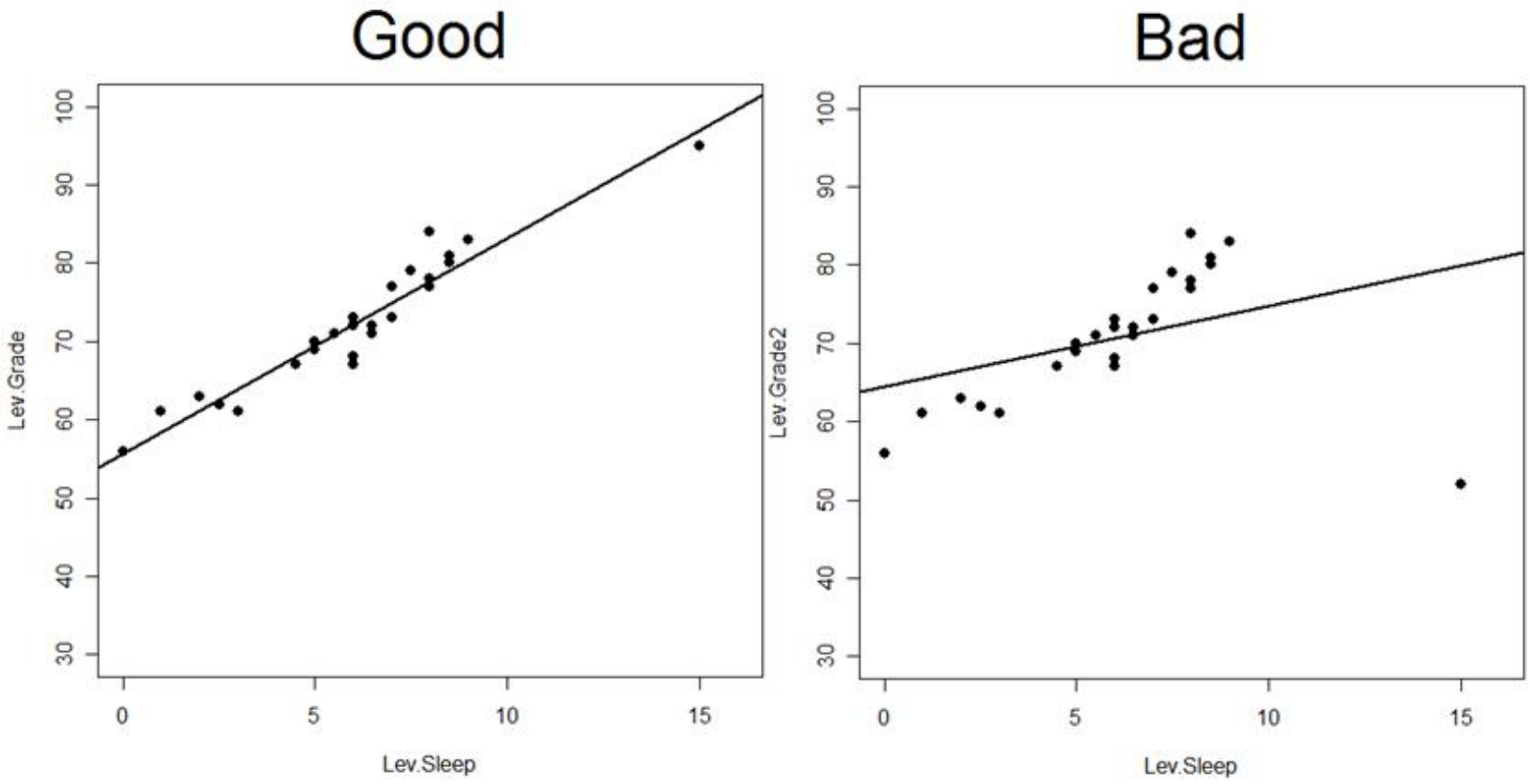
Leverage

- When a case has an unusual X value.
  - Leverage is not always bad.



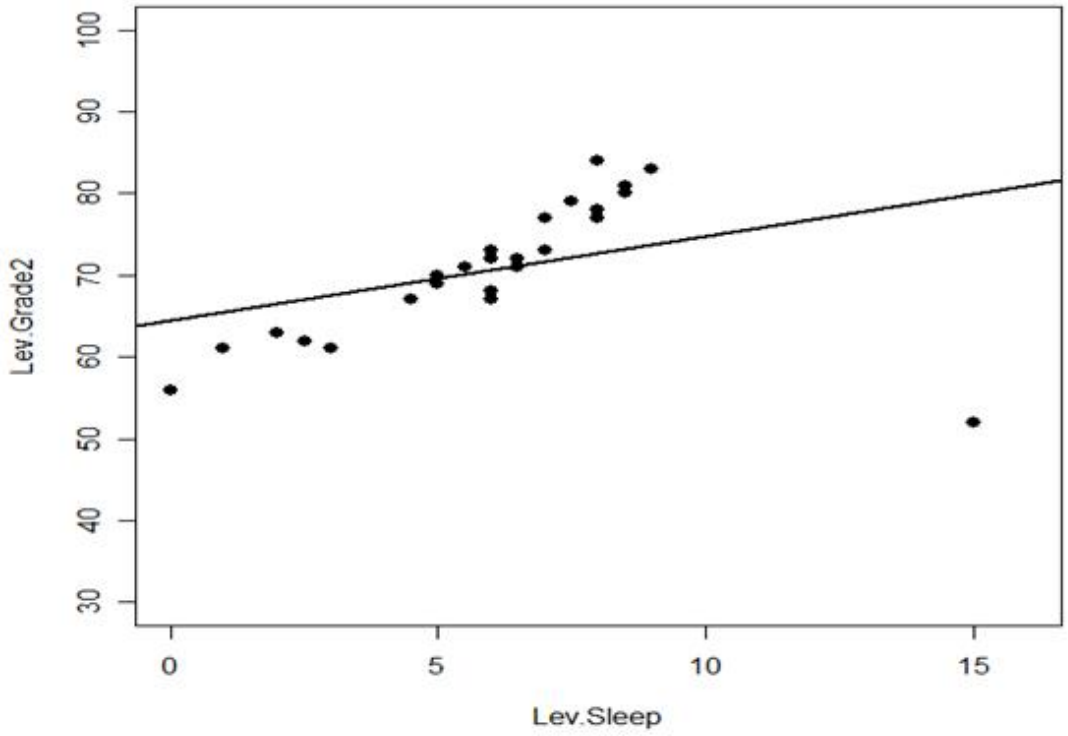
Leverage

- When a case has an unusual X value.
  - Leverage is not always bad.



Influence

- A case that is both an outlier and has leverage is said to "influence" the regression line.
  - It effects both the constant and the slope.



Predicting the Dependent Variable

- The regression model allows us to predict values of the dependent variable based on the value of an independent variable.
  - This prediction is going to better than just using the mean of the dependent variable in prediction.
- However, a good prediction is typically not our goal.
  - The model could be a good test of our hypothesis even if it does a poor job predicting the dependent variable.
  - This is because our hypothesis is typically only directional.

Residuals

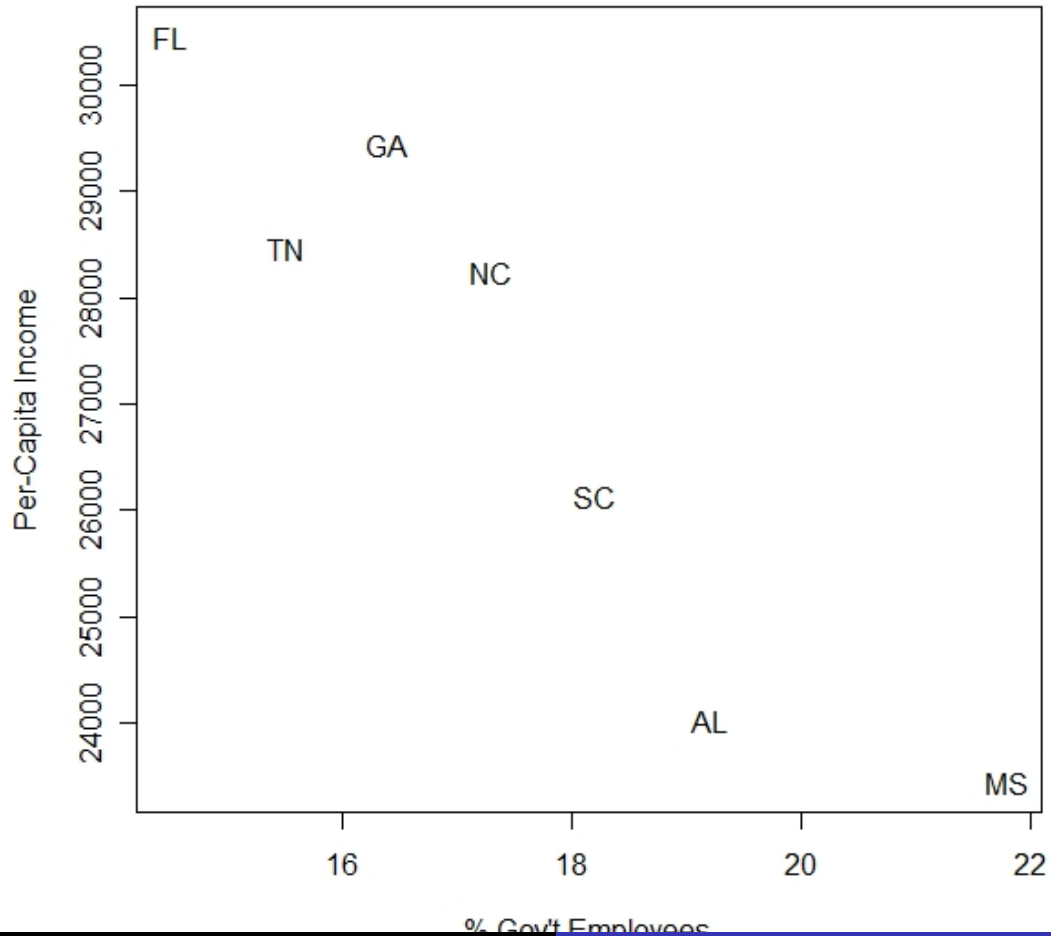
- We can use the model to predict the dependent variable for every case in our data set.
  - $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
- Those predictions are not going to be perfect.
- The difference between the actual value and the predicted value is called the *residual.*
  - $\hat{u}_i = Y_i - \hat{Y}_i$

Revisiting an Earlier Example

- (Flipped) Hypothesis: States with governments that employ more workers will have lower per-capita incomes.

State	Per-Capita Income (Y)	% Gov't Employees (X)
Alabama	$24,028	19.2%
Florida	$30,446	14.5%
Georgia	$29,442	16.4%
Mississippi	$23,448	21.8%
North Carolina	$28,235	17.3%
South Carolina	$26,132	18.2%
Tennessee	$28,455	15.5%

The Data Visually



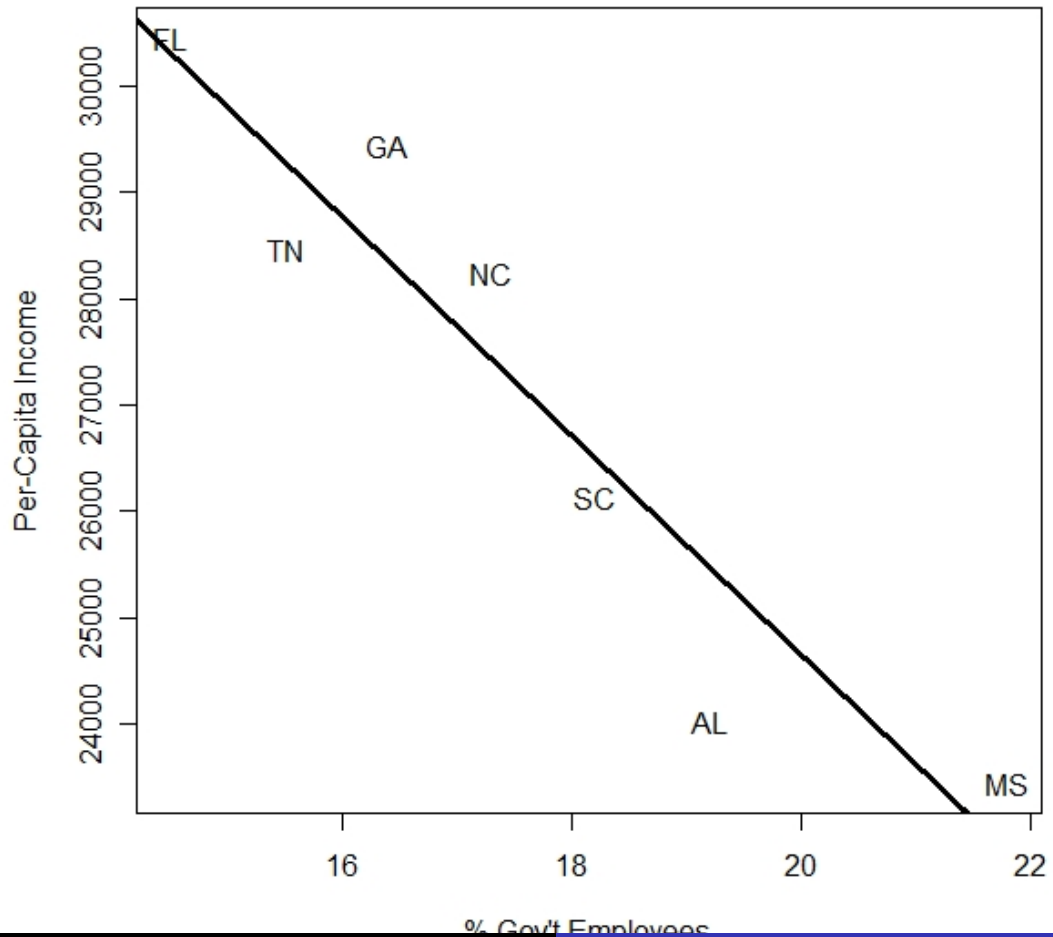
The Model

- *Per-Capita Income*$= \alpha + \beta$*% Gov't Employees*

	Coefficient	Standard Error	T-Value
% Gov't Employees	-1030.07	169.4421	-6.08
Constant	45254.51	2999.699	15.09

- Critical Value with 5 degrees of freedom is 2.57
- $\hat{Y}_i = 42,254.51 + (-1,030.07) * X_i$

The Regression Line



Predictions from the Model

- $\hat{Y}_i = 42,254.51 + (-1,030.07) * X_i$

State	Model	Prediction
Alabama	42,254.51+(-1,030.07)*19.2%	25,477.17
Florida	42,254.51+(-1,030.07)*14.5%	30,318.50
Georgia	42,254.51+(-1,030.07)*16.4%	28,361.37
Mississippi	42,254.51+(-1,030.07)*21.8%	22,798.99
North Carolina	42,254.51+(-1,030.07)*17.3%	27,434.30
South Carolina	42,254.51+(-1,030.07)*18.2%	26,507.24
Tennessee	42,254.51+(-1,030.07)*15.5%	29,288.43

Residuals

State	$Y_i - \hat{Y}_i$	Residual
Alabama	$24,028 - 25,477.17$	-1449.17
Florida	$30,446 - 30,318.50$	127.50
Georgia	$29,442 - 28,361.37$	1080.63
Mississippi	$23,448 - 22,798.99$	649.01
North Carolina	$28,235 - 27,434.30$	800.70
South Carolina	$26,132 - 26,507.24$	-375.24
Tennessee	$28,455 - 29,288.43$	-833.43

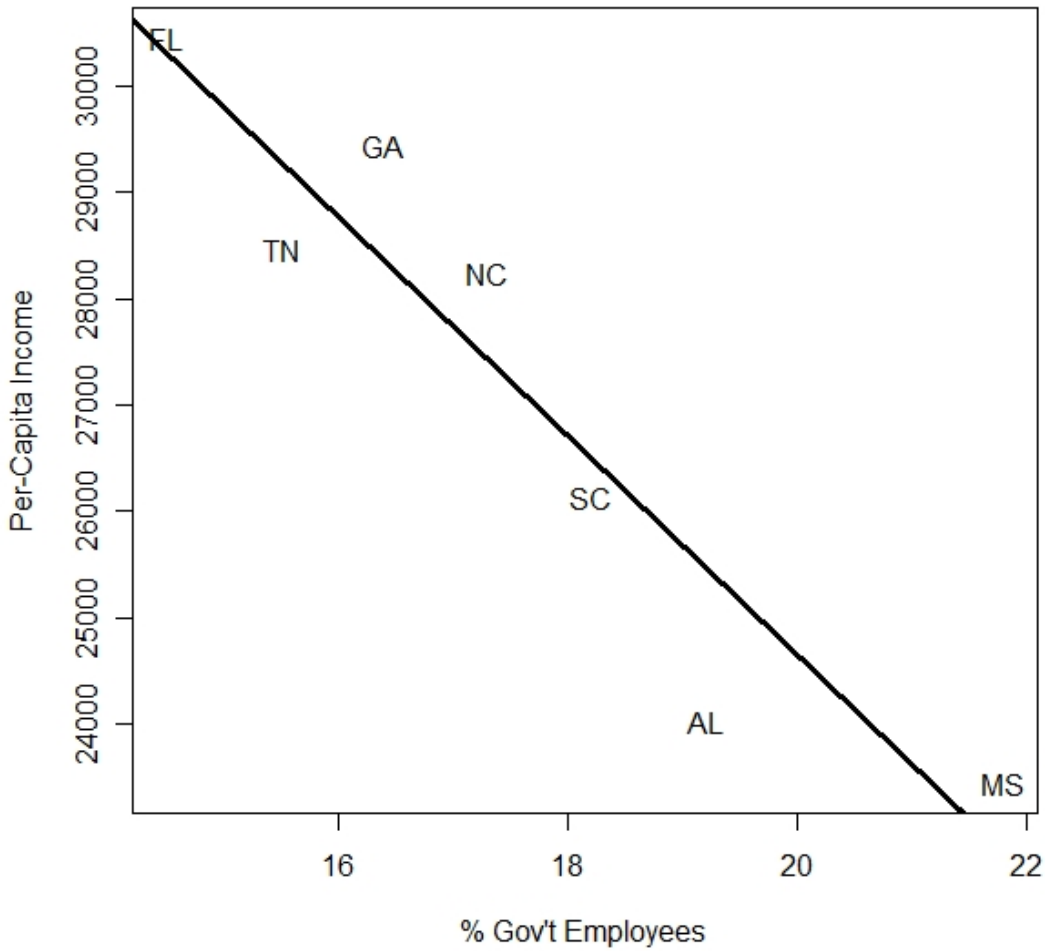
Goodness of Fit

- When we talk about the "goodness of fit" of a model, we're talking about how well the model predicts the dependent variable.
- We can use the residuals to figure out the "goodness of fit."
  - The smaller the residuals, the better the "goodness of fit".
- We will look at two measures of model fit.
  - Root Mean Squared Error (MSE)
  - R^2

What is the root Mean Squared Error?

- The root Mean Squared Error is a measure of the typical deviations from the regression line.
- root MSE$=\sqrt{\frac{\Sigma \hat{u}_i^2}{n-k}}$
  - k is equal to the number of parameters. For bivariate regression $k = 2$.
    - This is different from the book
  - Remember, $\hat{u}_i^2$ are the residuals
- So, alternatively, root MSE$=\sqrt{\frac{\Sigma(Y_i-\hat{Y}_i)^2}{n-k}}$

Our Example


FL
GA
TN
NC
SC
AL
MS
Per-Capita Income
24000
25000
26000
27000
28000
29000
30000
16
18
20
22
% Gov't Employees

Residuals

State	$Y_i - \hat{Y}_i$	$Residual^2$
Alabama	-1449.17	2100093.69
Florida	127.5	16256.25
Georgia	1080.63	1167761.20
Mississippi	649.01	421213.98
North Carolina	800.7	641120.49
South Carolina	-375.24	140805.06
Tennessee	-833.43	694605.56
		$\Sigma = 5181856.23$

- root MSE$=\sqrt{\frac{5181856.23}{7-2}} = 1018.02$

What Does This Tell Us?

- root MSE$=\sqrt{\frac{5181856.23}{7-2}} = 1018.02$
- On average, the model we estimated is off by \$1,018.02 in predicting the per-capita income in a state.
  - Is that good?
- root MSE is in the metric of the dependent variable.
  - The dependent variable ranges from about \$24,000-\$30,000.
  - Is a difference of \$1,018.02 small given that range?
    - The standard deviation of the dependent variable is \$2,692.04.

Another Goodness-of-Fit Measure

- R^2 is the proportion of the variance in the dependent variable that our model explains.
  - Therefore, it ranges from 0-1.
- The closer our R^2 is to 1, the more of the variation our model explains.
- The closer our R^2 is to 1, the better our model is at predicting the dependent variable.

Calculating R^2, Part 1

- R^2=Regression Sum of Squares/Total Sum of Squares
  - Deviations from the mean predicted by our model over the total deviations from the mean.
- Regression Sum of Squares=Total Sum of Squares-Residual Sum of Squares
  - Total deviations from the mean minus deviations that our model does not explain.
- $$R^2 = \frac{\Sigma(Y_i - \bar{Y})^2 - \Sigma(Y_i - \hat{Y}_i)^2}{\Sigma(Y_i - \bar{Y})^2}$$

Calculating R^2 Part 2

- We calculate the total sum of squares like we did before.
  - Subtract the mean of the dependent variable from all the observed values of the dependent variable.
  - Square that and add them all up.
- Calculate our residual sum of squares.
  - We use our model to predict values of the dependent variable for every case.
  - Calculate the residuals like we did before.
  - Square them and add them all up.

Total Sum of Squares

State	$Y_i - \bar{Y}_i$	$(Y_i - \bar{Y}_i)^2$
Alabama	24,028-27,169.43=-3,141.43	9,868,573.47
Florida	30,446-27,169.43=3,276.57	10,735,920.33
Georgia	29,442-27,169.43=2,272.57	5,164,580.90
Mississippi	23,448-27,169.43=-3,721.43	13,849,030.61
North Carolina	28,235-27,169.43=1,065.57	1,135,442.47
South Carolina	26,132-27,169.43=-1,037.43	1,076,258.04
Tennessee	28,455-27,169.43=1,285.57	1,652,693.90
		$\Sigma = 43,482,499.71$

Putting It Together.

- Total Sum of Squares=43,482,499.71.
- Residual Sum of Squares=5,181,856.23.
  - We figured that out earlier with the root MSE.
- $R^2 = \frac{43{,}482{,}499.71 - 5{,}181{,}856.23}{43{,}482{,}499.71} = .88$
- Our model explains 88% of the variation in the dependent variable.

Is R^2 everything we want to know?

- R^2 will get larger when you add more variables.
  - But this does not mean we should just add more variables to the model to increase the value of R^2
- "The 'right model' depends entirely on the use to which it is put."
  - King 1991, 1050
- The size of R^2 is most important when we are trying to build the model that is the most predictive.
  - If we are simply hypothesis testing, the value of R^2 is less important.