Linear Regression

November 16, 2022

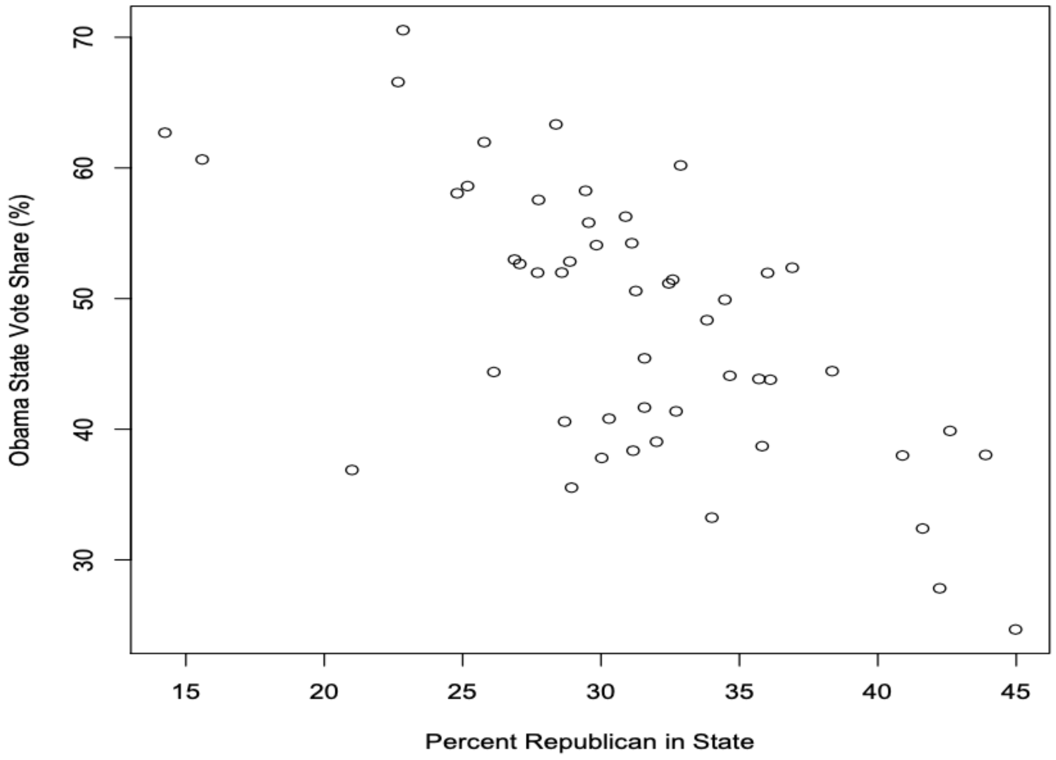
Today

- Discussion of Linear Regression
- Review of basic approach
  - Relate to difference-of-means testing
- Inference for regression
  - Statistical Significance
  - P-values
  - T-tests
  - Confidence Intervals
- Regression & Prediction
- Things to Keep in Mind with Linear Regression

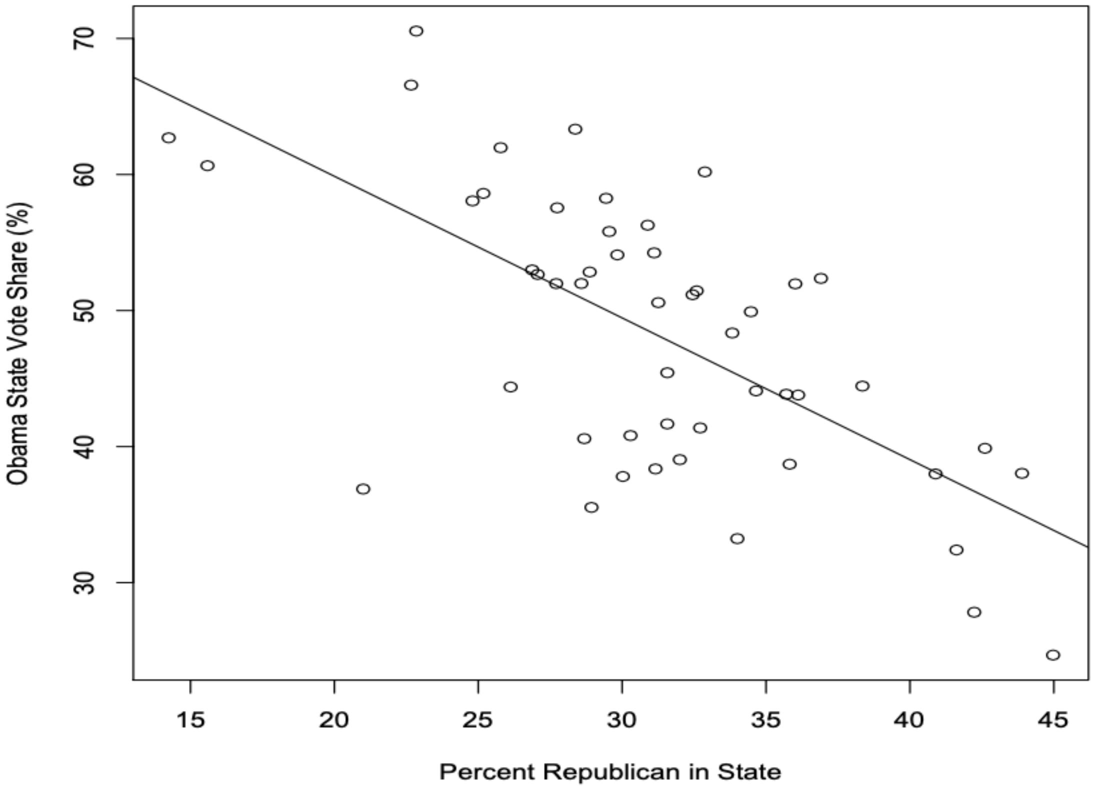
Linear Regression

- Linear regression assumes an interval DV
- First step: examine data graphically
- Scatterplot:
  - Values of IV on X-axis
  - Values of DV on Y-axis
- Can plot the "best fit" line through the scatterplot
- Gives regression equation: $y=\beta_0+\beta_1 x$
  - β_0="constant" or "intercept"
  - β_1="slope"
- Can use regression equation to generate predicted values

Scatterplot



Scatterplot with regression line



Regression Output

```
Call:
lm(formula = Obama2012 ~ reppct_m, data = states)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9484  -5.1864   0.8801   5.0077  13.7267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.7026     5.5661  14.499  < 2e-16 ***
reppct_m     -1.0415     0.1745  -5.968 2.81e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.875 on 48 degrees of freedom
Multiple R-squared:  0.426,	Adjusted R-squared:  0.414
F-statistic: 35.62 on 1 and 48 DF,  p-value: 2.805e-07
```

Regression and means

- Linear regression is focused on the mean value of *Y*
- Interested in difference in the mean of *Y* across different values of *X*
  - Similar to a difference-of-means test, but here with many (not just two) subpopulations
- Regression equation gives us prediction of mean value of *Y* for all units sharing the same value of *X*
  - Example—mean percentage of the vote received by President Obama in states with 40% Republican identifiers
- Linear regression assumes the means all fall on a line—i.e., one constant slope
  - That's the "linear" part

Linear regression

- Assumption of linear regression—a one-unit change in *X* has the same effect on the mean of *Y* across all values of *X*
  - Ex.—Difference in the (expected) mean vote for President Obama in states between 29% and 30% Republican identifiers is the same as the difference in the mean vote between 39% and 40% Republican identifiers
- Determining if this is appropriate requires first plotting your data (and some theory)
- If relationship does not appear to be linear, some transformation of a variable is necessary (we will discuss more later)

Residuals

- The regression equation gives us a prediction of the mean value of Y at different values of X
  - The individual values of Y will vary around this mean
- The residual is the difference between the actual Y and the predicted Y generated by the regression equation
  - Residual=$Y_i - \hat{Y}_i$
- The regression equation is the "best fit" line—it minimizes the sum of the squared residuals

Inference and Regression

- There is a regression equation in our population:
  - $Y=\beta_0+\beta_1 X$
- This regression equation means that the observed response for each Y is:
  - $Y_i=\beta_0+\beta_1 x_i+\varepsilon_i$
- But, we generally do not observe population, so we draw a sample from that population
- In that sample, we can generate a regression equation:
  - $\hat{Y} = \beta_0 + \beta_1 x$

Estimating the Regression Equation

- Where does the regression equation come from?
- The regression slope (β_1) is the correlation coefficient between X and Y, multiplied by the standard deviation of Y divided by the standard deviation of X
  - $\beta_1 = (r_{yx})(\frac{s_y}{s_x})$
  - $\beta_1 = (\frac{\sum(\frac{x_i-\overline{x}}{s_x})(\frac{y_i-\overline{y}}{s_y})}{n-1})(\frac{s_y}{s_x}) = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}$
- The intercept (β_0) is the mean of Y minus β_1 times the mean of X:
  - $\beta_0 = \bar{Y} - \beta_1(\bar{X}_1)$
- The intercept is the mean value of Y when X=0—this may (or may not) have a meaningful interpretation

Goodness of fit

- $Y_i = \hat{Y}_i + e_i$
- Total Sum of Squares (TSS): Sum of difference between observed Y and mean of Y, squared
- Model Sum of Squares (MSS): Sum of difference between predicted Y and mean of Y, squared
- Residual sum of squares (RSS)—sum of the squared residuals
- $TSS = MSS + RSS$
- Can judge goodness-of-fit using R^2
  - $R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$
    - R^2 indicates the fraction of the variation in y that is explained by x
- Can also use sum of squares for an F-test (will discuss later)

Regression Output

```
Call:
lm(formula = Obama2012 ~ reppct_m, data = states)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9484  -5.1864   0.8801   5.0077  13.7267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.7026     5.5661  14.499  < 2e-16 ***
reppct_m     -1.0415     0.1745  -5.968 2.81e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.875 on 48 degrees of freedom
Multiple R-squared:  0.426,     Adjusted R-squared:  0.414
F-statistic: 35.62 on 1 and 48 DF,  p-value: 2.805e-07
```

Inference for Regression

- Basic problem of statistical inference: interested in population, but observe data from a sample
- Want to infer from results in sample to population
  - Given results of regression from sample, how confident can we be about the effect of X on Y in the population?
- Similar approach to that used for sample means:
  - In large samples, we can assume that sampling distributions for β_0 and β_1 are normally distributed
  - Can estimate standard deviations for β_0 and β_1 from the data

Determining the Standard Errors

- Residuals (e_i) correspond to model deviations ε_i
- $\sum e_i = 0$
- Variation of Y around the population regression line—measured by the standard deviation of the model deviations (σ):
  - Estimate is based on the residuals
  - $s^2 = \frac{\sum {e_i}^2}{n-k-1}$
  - *n-k-1* is the degrees of freedom
- So, the estimate of the model standard deviation is: $s = \sqrt{s^2}$
  - $s = \sqrt{\frac{\sum {e_i}^2}{n-k-1}}$
- Can use this to find standard errors for β_0 and β_1

Determining the Standard Errors

- We can use the estimate of the model standard deviation (*s*) to estimate a standard error for the intercept and the slope coefficients
- Standard Error for slope (β_1)
  - $SE_{\beta_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$
- Standard Error for intercept (β_0)
  - $SE_{\beta_0} = (s)(\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$
- We can use the standard error of the coefficient to generate a confidence interval around the coefficient

Confidence Interval

- Based on our sample, β_1 is the best estimate of the population parameter (coefficient)
- But, we are interested in the range within which we are confident β_1 actually lies
- A level-C confidence interval for β_1 is:
  - $\beta_1 +/-(t)(SE_{\beta_1})$
  - t is from the t distribution with ($n - k - 1$) degrees of freedom
- Why do we use a t-distribution?
  - We have estimated the standard error from the sample, it is not from the population
- We can also generate a confidence interval around the intercept
  - Sometimes useful, but generally not

Regression Output

```
Call:
lm(formula = Obama2012 ~ reppct_m, data = states)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9484  -5.1864   0.8801   5.0077  13.7267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.7026     5.5661  14.499  < 2e-16 ***
reppct_m     -1.0415     0.1745  -5.968 2.81e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.875 on 48 degrees of freedom
Multiple R-squared:  0.426,     Adjusted R-squared:  0.414
F-statistic: 35.62 on 1 and 48 DF,  p-value: 2.805e-07
```

Testing the null hypothesis

- Hypothesis tested through linear regression is that X has some linear effect on Y
    - As values of X increase, values of Y increase (directional, positive)
    - As values of X increase, values of Y decrease (directional, negative)
    - As values of X change, values of Y change as well (non-directional)
- Null hypothesis is that X has no effect on Y
    - Or, a change in X is not associated with a change in Y
- If null hypothesis true, then β_1 in the population = 0
- How likely is it that we observe β_1 in our sample if the null is true?
    - Generate a p-value

Rejecting the Null Hypothesis

- Want to determine how confidently we can reject null hypothesis
- First, find a t-statistic
  - $t = \frac{\beta_{11}}{SE_{\beta_1}}$
- Then, use the t-distribution chart to determine the p-value
  - Remember, degrees of freedom = $(n - k - 1)$
  - Decide whether you are looking at one or two tails
- Again, the p-value tells you the probability you would get a coefficient of the estimated size if the null hypothesis were true
- Compare to level of statistical significance chosen
  - Example—if $p \leq 0.05$, statistically significant at the .05 level; can reject null hypothesis with 95% confidence

Three Ways to Tell Statistical Significance (at .05 level)

1. Is $p \leq 0.05$
   - Can be 0.10 if using a one-tailed test (and looking at regression output with default two-tailed p-values)
2. Is t greater than 1.96 (with a large N)?
   - Affected by degrees of freedom, if N is smaller, need a larger t
   - If one-tailed test, is t greater than 1.645?
3. Does a 95% confidence interval have the same sign on the lower and upper bound?
   - If a one-tailed test, use a 90% confidence interval

- Important—all of these will show the same thing, they are all functions of one another

Regression Output

```
Call:
lm(formula = Obama2012 ~ reppct_m, data = states)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9484  -5.1864   0.8801   5.0077  13.7267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.7026     5.5661  14.499  < 2e-16 ***
reppct_m     -1.0415     0.1745  -5.968 2.81e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.875 on 48 degrees of freedom
Multiple R-squared:  0.426,     Adjusted R-squared:  0.414
F-statistic: 35.62 on 1 and 48 DF,  p-value: 2.805e-07
```

Statistical Significance

- What affects statistical significance?

1. Size of the coefficient
   - The bigger effect *X* has on *Y* in the sample, the less likely it is that the true effect in the population is zero
2. Size of the standard error
   - t-statistic is coefficient divided by standard error, smaller standard error means bigger *t*
3. The number of observations
   - Smaller *t* gives bigger p-value as degrees of freedom increases
   - This increase diminishes quickly and (essentially) stops at about 1,000 degrees of freedom

Type I and Type II Errors

- Remember:
  - Type I Error means we **mistakenly reject** the null hypothesis;
  - Type II Error means we **mistakenly fail to reject** the null hypothesis
- Accepting a 95% significance level means that there is 5% chance we will have a p-value of .05 or less if the null hypothesis is true
  - This means a greater than 5% chance that we will commit a Type II Error
    - Will frequently reject null when *X* does have an effect on *Y*

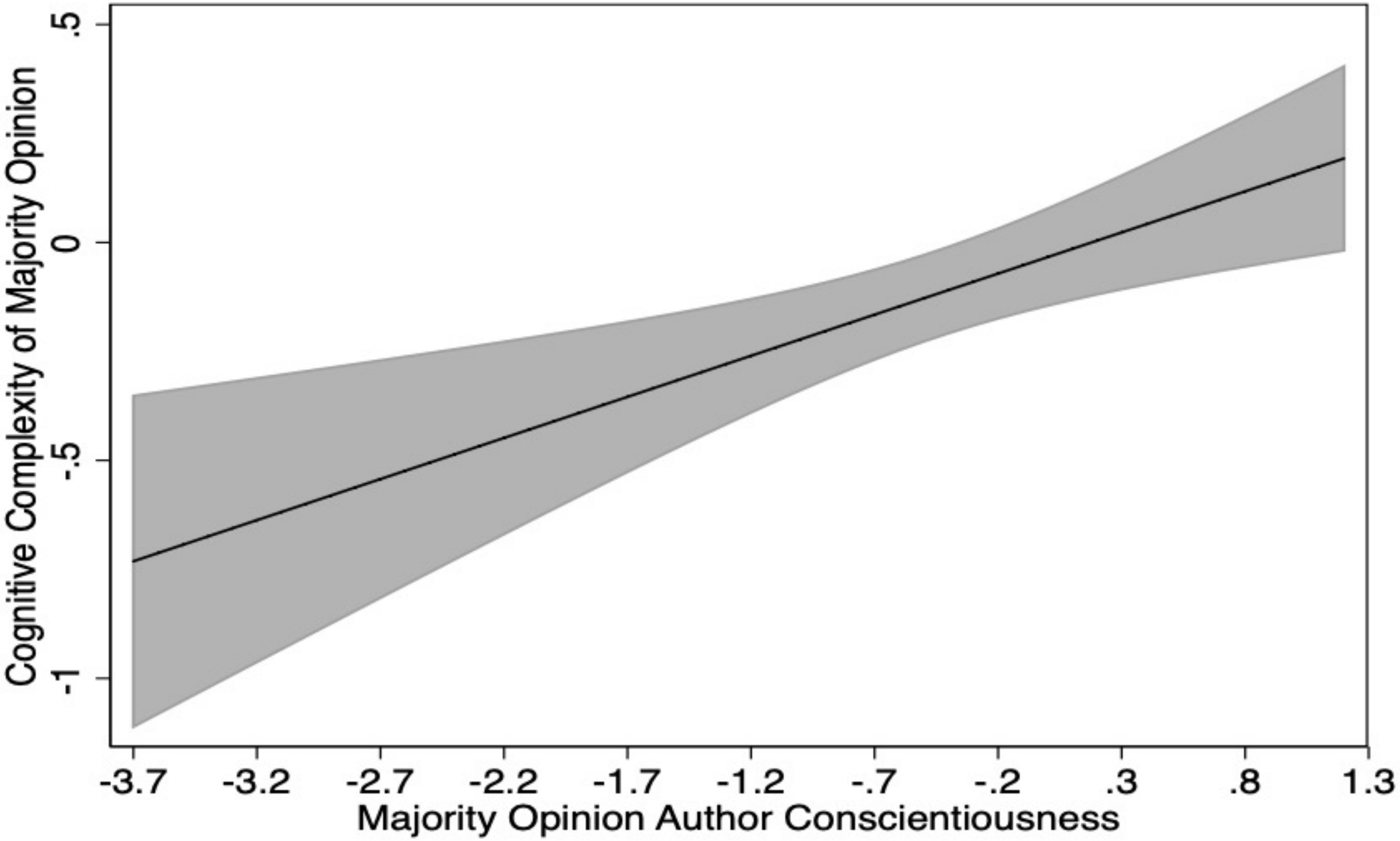
Substantive Significance

- Statistical significance does not tell us how **large, or meaningful** of an effect *X* has on *Y*
  - Example—each additional $1,000 spent on advertising increases vote share by 0.001%
  - Example—each additional $1,000 spent on GOTV increases vote share by 0.01%
- Regression allows for evaluating substantive significance:
  - Coefficient is the effect of a one-unit change in *X* on *Y*
  - Have to think critically about what a "one-unit change" means
  - Also want to evaluate confidence interval
  - **Best practice**—display predicted values graphically with confidence intervals and use a figure to help with substantive significance
- R^2 can also be a measure of explanatory power

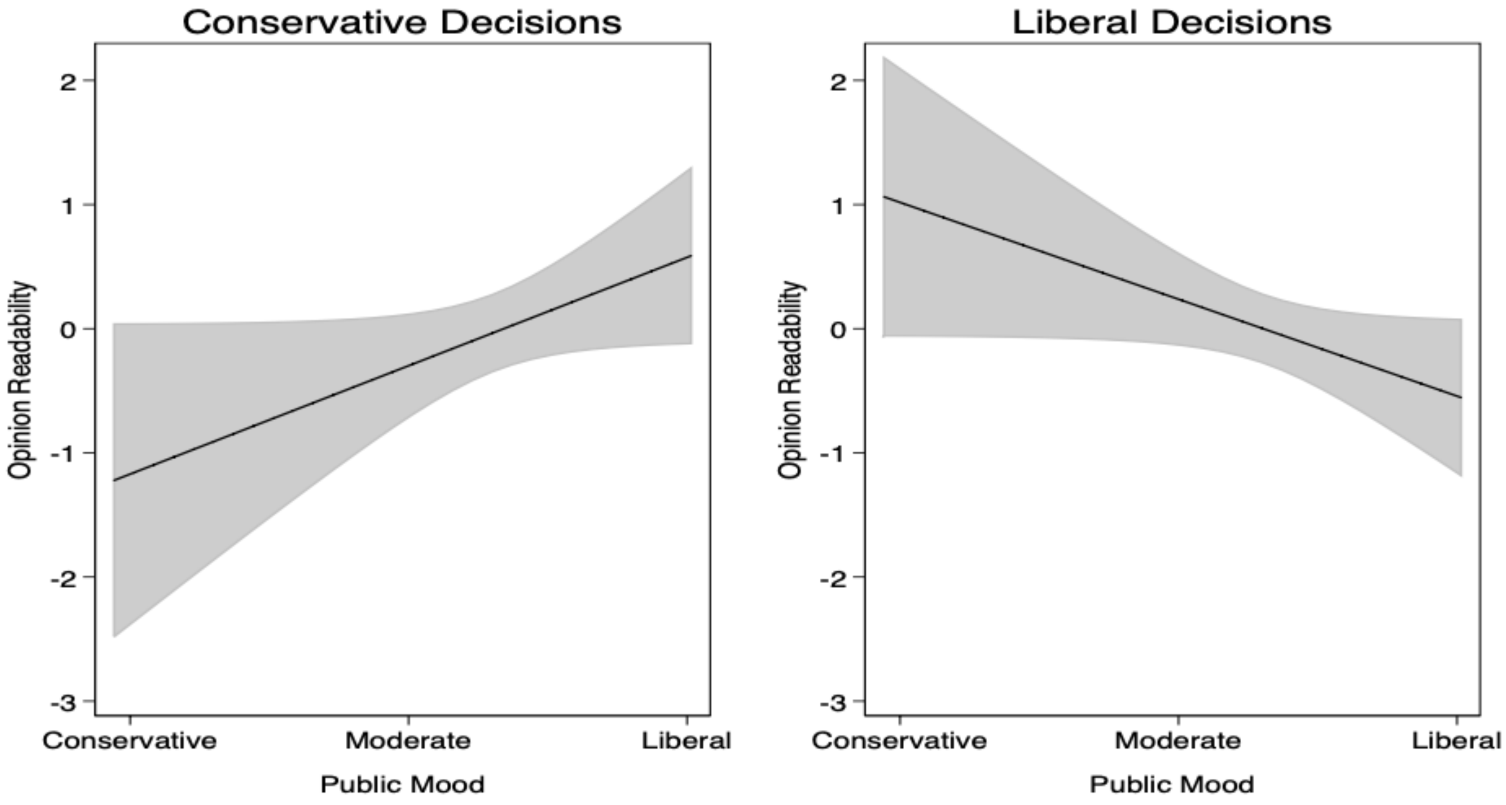
Confidence Intervals for Mean Response

- The regression line gives us a prediction for any value of X (X*)
  - This is a prediction of the mean value of Y at a given value of X
  - $\hat{\mu}_y = \beta_0 + \beta_1 x^*$
- Can also generate a confidence interval around these predictions
- To do that, we need the standard error for the predicted mean of Y ($\hat{\mu}_y$)
  - $SE_{\hat{\mu}} = (s)(\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \overline{x})^2}})$
- A level-C confidence interval for the mean response ($\hat{\mu}_y$) when X takes the value X* is:
  - $\hat{\mu}_y$ +/- (t)($SE_{\hat{\mu}}$)
- Can use this technique to generate confidence intervals around the regression line

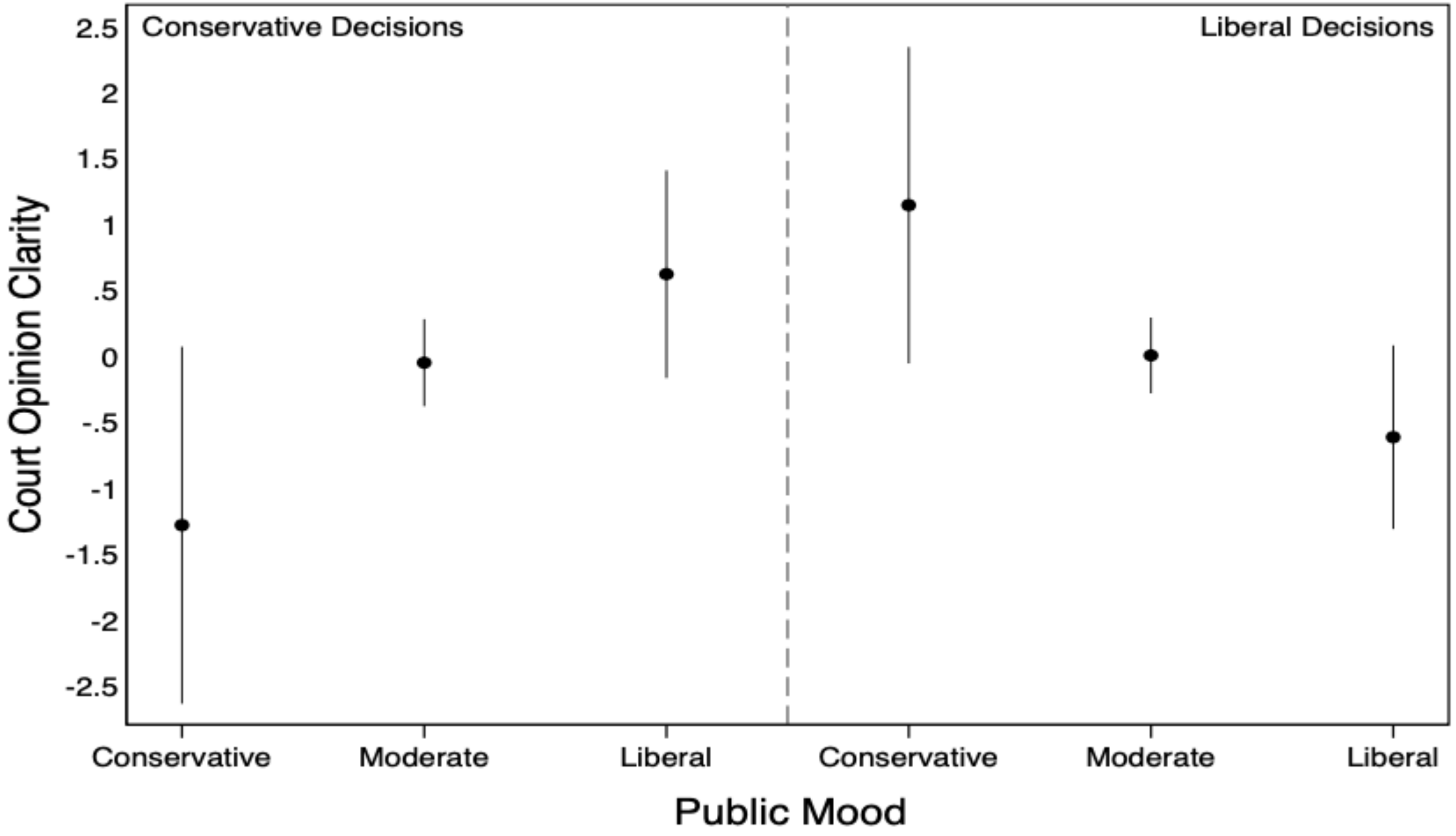
Mean Response Example



Mean Response Example



Mean Response Example



Prediction Intervals

- One advantage of regression is that we can use it to generate specific predictions about future values
  - Example—income of an individual based on his/her level of education
- Can also generate a prediction interval for a future observation
  - Includes a margin of error
- To calculate this, need the standard error of $\hat{y}$
  - $SE_{\hat{y}} = (s)(\sqrt{1 + \frac{1}{n} + \frac{(x^{*} - \bar{x})^2}{\sum(x_i - \overline{x})^2}})$
- A level-C prediction interval for a future observation $\hat{y}$ is:
  - $\hat{y}$ +/- (t)($SE_{\hat{y}}$)

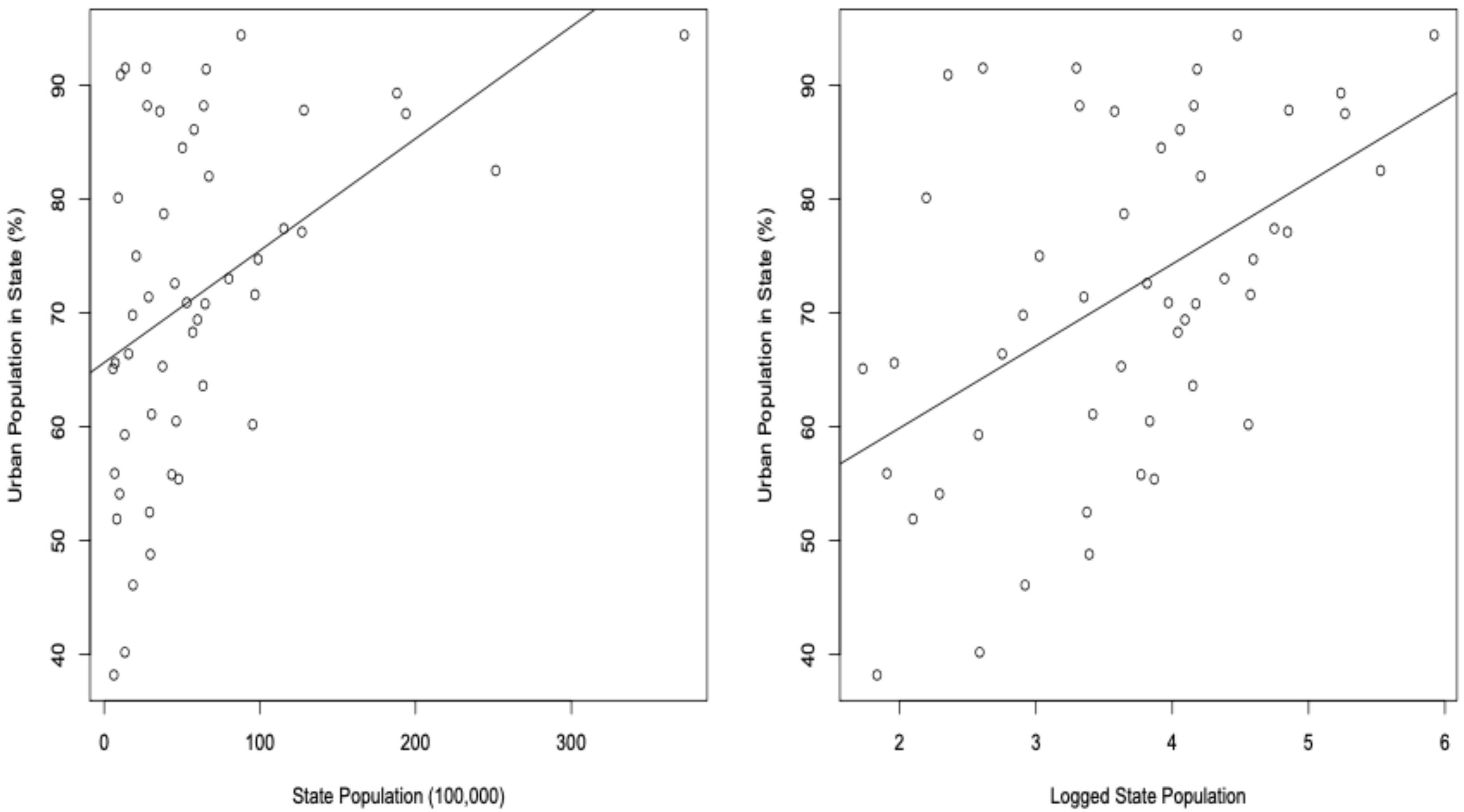
Three Types of Interval Estimates

- We have essentially three interval estimates—important to understand what they are:
  - Confidence interval around regression coefficient (β_1)
    - Tells us the range within which we are confidence the population regression coefficient lies
  - Confidence interval for mean response
    - Tells us the range within which we are confident the population mean of y falls for a specific value of x
  - Prediction interval
    - Tells us the range within which we are confident a *future value* of y would fall for a specific value of x

Things to Keep in Mind

- 1. Ordinary Least Squares is *linear* regression
  - Remember: means that effect of one-unit change in X on Y is constant across values of X
  - May not be true in theory or in practice
- Can transform variables to deal with non-linearity
- A common transformation is the natural log-transformation
  - $\ln(X_{orig}) = X_{log}$ is equivalent to: $e^{X_{log}} = X_{orig}$
  - Be careful with interpretation—the units change
- Why would we use a log-transformation?
  - If we think that effect of X on Y diminishes as X gets larger
  - If our data is skewed, and large observations on X and Y affect the overall pattern

Scatterplot with Log-Transformed Variable



Things to Keep In Mind

- 2. Still have to think about control variables
- Regression tells us how mean value of Y changes as values of X change
- But, both changes could be driven by lurking variable(s)
- One thing we like about regression is that it is easy to incorporate additional variables

Next Steps

- Next class we will discuss how to incorporate control variables into regression (i.e., multiple regression)
- Homework due Wednesday, November 30
- Section this week, no class or section Thanksgiving week