Simple Linear Regression III: Inference

Some chatter from the internets

2016 Election

Question at hand: How will Obama's 46% approval rating effect his party's candidate for the 2016 presidential election?

natesilver: I guess I look at it more like this. My prior is that elections with a term-limited incumbent are 50-50. I'm looking for evidence that persuasively overcomes that prior. An extremely popular or extremely unpopular incumbent would clearly matter. But Obama's popularity is about average.

micah: But this was one of my questions: Obama isn't running; how much of an effect will he have on the race? Positive or negative — is Obama's popularity really a big factor?

natesilver: He'll have a fairly neutral effect, given his current popularity level.

harry: We only have approval rating data at this point in a campaign (September/October the year before) for six instances when an incumbent president didn't run for re-election. Now, I took those and plugged them into a simple little regression. With Obama's approval at 46 percent, the GOP is expected to win by about 2 percentage points. Again, there's a huge margin of error, but signs point to a slight GOP edge.

How would you visualize this data?

natesilver: Dude. It's not even six examples really. It's four.

harry: Who are your four?

natesilver: Dwight Eisenhower, Ronald Reagan, Bill Clinton and George W. Bush are the only presidents in American history to be term-limited. Obama will be the fifth.

And I don't care if you get the same regression results with four.

harry: And did you know that Obama's approval rating is below the average approval rating for those four? And it's not particularly close either.

natesilver: The problem is that running a regression model based on an n of four is inherently kind of ridiculous.

NERD FIGHT!

Why is it ridiculous?

Inference for Regression

We can fit a line through any cloud of points that we please, but if we just have a *sample* of data, any trend we detect doesn't necessarily demonstrate that the trend exists in the *population* at large.

Plato's Allegory of the Cave

Statistical Inference

Goal: use *statistics* calculated from data to makes inferences about the nature of *parameters*.

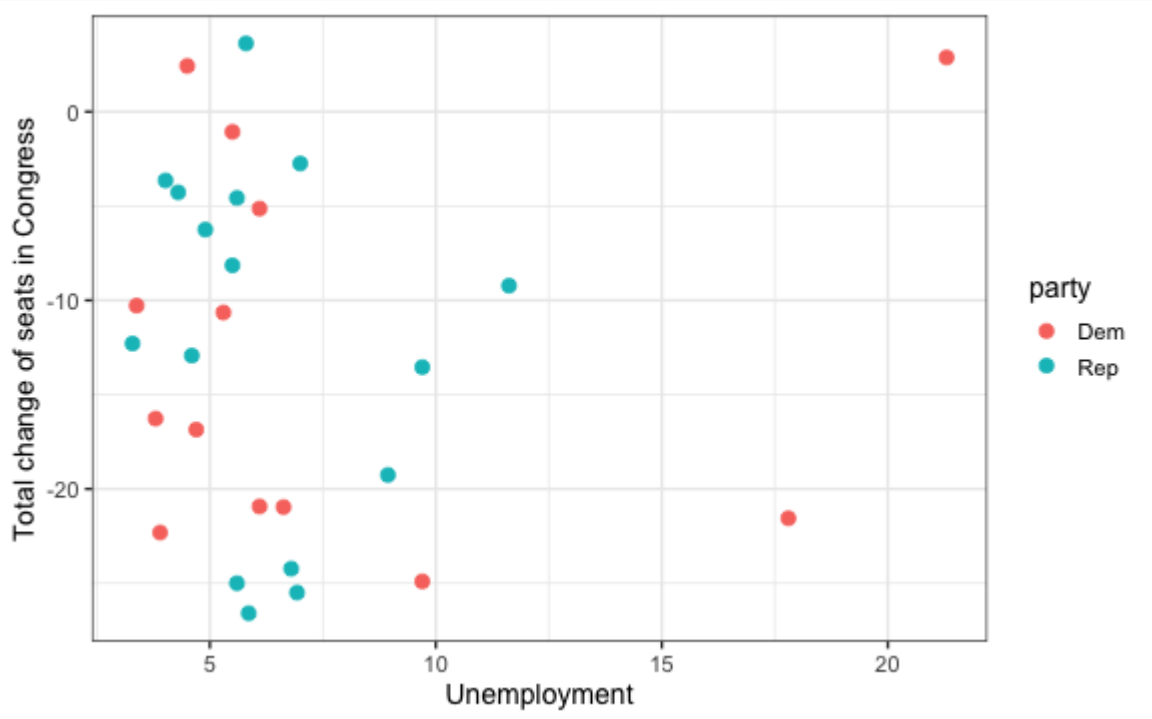
In regression,

- parameters: β_0, β_1
- statistics: b_0, b_1

Classical tools of inference:

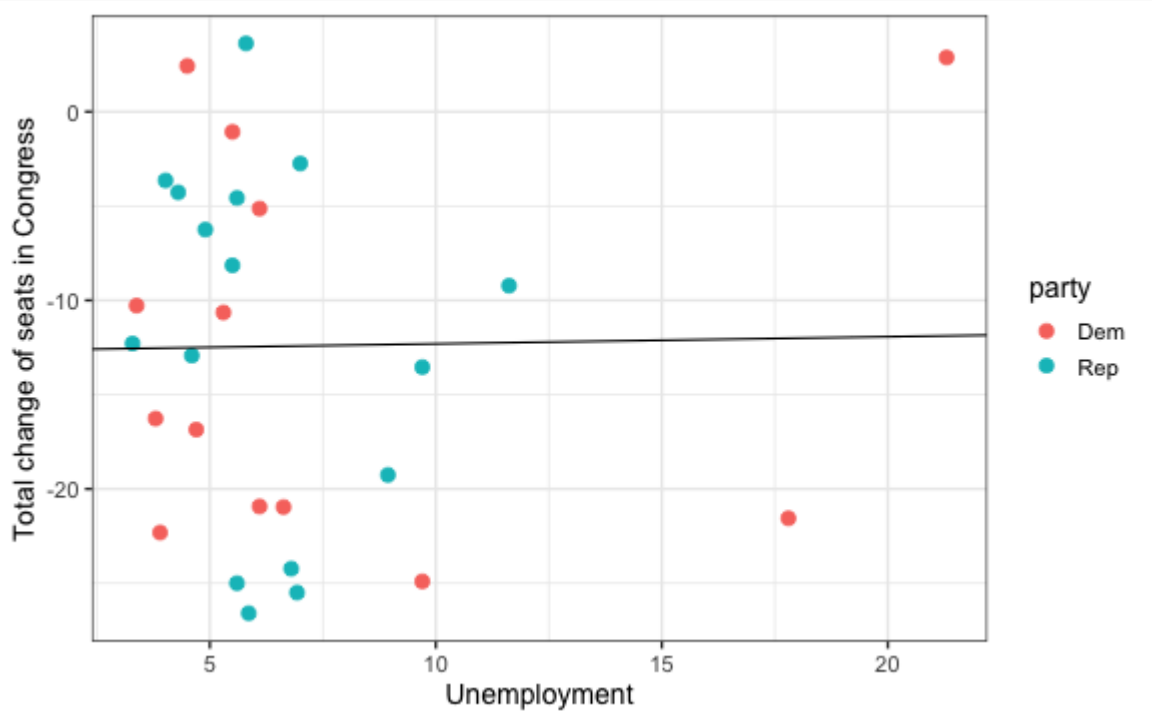
- Confidence Intervals
- Hypothesis Tests

Unemployment and elections



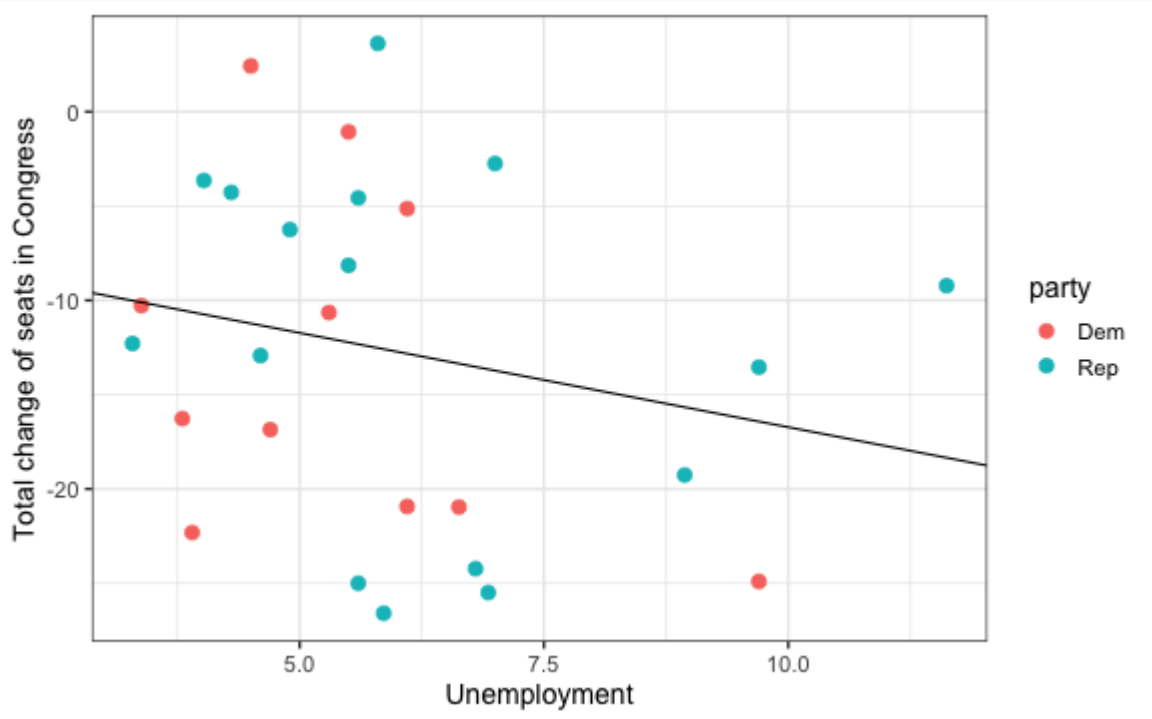
Reigning theory: voters will punish candidates from the Presidents party at the ballot box when unemployment is high.

Unemployment and elections



Reigning theory: voters will punish candidates from the Presidents party at the ballot box when unemployment is high.

Unemployment and elections, cont.



Some evidence of a negative linear relationship between unemployment level and change in party support - or is there?

H-test for Regression

H_0 : There is no relationship between unemployment level and change in party support (or: change in party support is independent of unemployment).

$H_0 : \beta_1 = 0$

Method

If there is no relationship, the pairing between X and Y is artificial and we can permute:

1. Create synthetic data sets under H_0 by shuffling X.
2. Compute a new regression line for each data set and store each b_1.

Your turn

Take a moment to sketch out the infer pipeline that will results in a collection of 500 slopes that would might see in a world where the null hypothesis was true.

First shuffle

```
library(infer)
ump %>%
  specify(change ~ unemp) %>%
  hypothesize(null = "independence") %>%
  generate(1, type = "permute")
```

```
## Response: change (numeric)
## Explanatory: unemp (numeric)
## Null Hypothesis: independence
## # A tibble: 27 x 3
## # Groups:   replicate [1]
##    change unemp replicate
##     <dbl> <dbl>     <int>
##  1 -22.3  11.6          1
##  2   3.62  4.3          1
##  3 -25     3.29         1
##  4  -4.57  5.86         1
##  5 -10.3   6.63         1
##  6  -4.28  3.38         1
##  7 -24.2   6.93         1
##  8 -12.9   4.02         1
##  9  -8.14  8.94         1
## 10 -12.3   4.7          1
## # … with 17 more rows
```

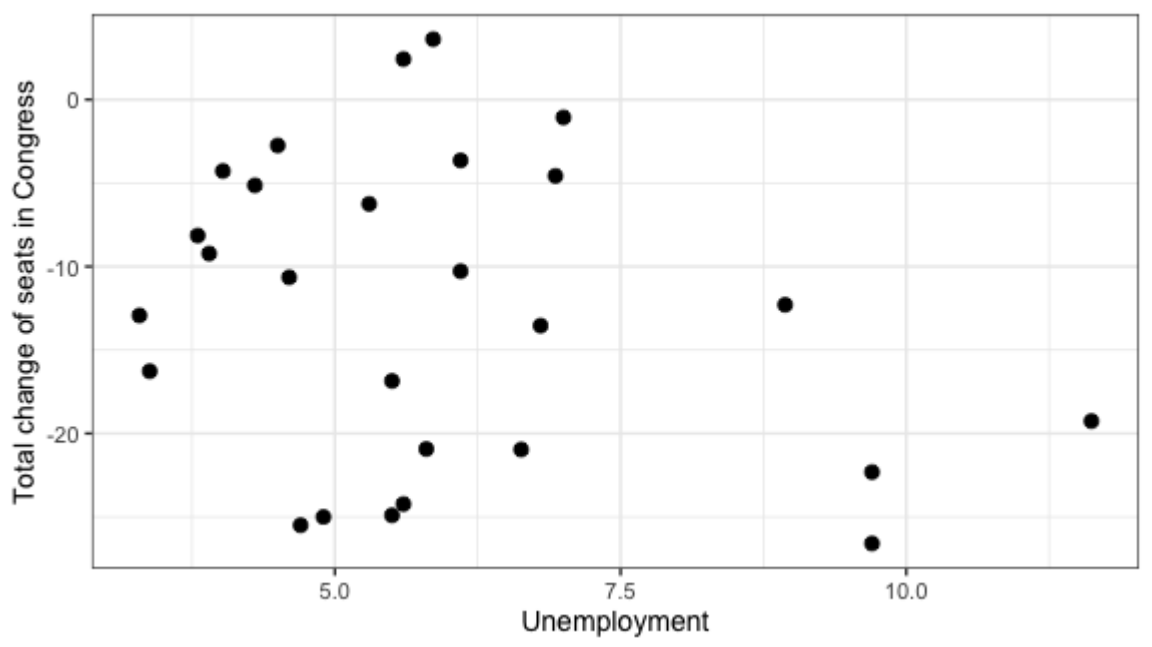
Second shuffle

```
shuffle2 <- ump %>%
  specify(change ~ unemp) %>%
  hypothesize(null = "independence") %>%
  generate(1, type = "permute")
shuffle2
```

```
## Response: change (numeric)
## Explanatory: unemp (numeric)
## Null Hypothesis: independence
## # A tibble: 27 x 3
## # Groups:   replicate [1]
##    change unemp replicate
##     <dbl> <dbl>     <int>
##  1 -19.3  11.6          1
##  2  -5.14  4.3          1
##  3 -12.9   3.29         1
##  4   3.62  5.86         1
##  5 -21.0   6.63         1
##  6 -16.3   3.38         1
##  7  -4.57  6.93         1
##  8  -4.28  4.02         1
##  9 -12.3   8.94         1
## 10 -25.5   4.7          1
## # … with 17 more rows
```

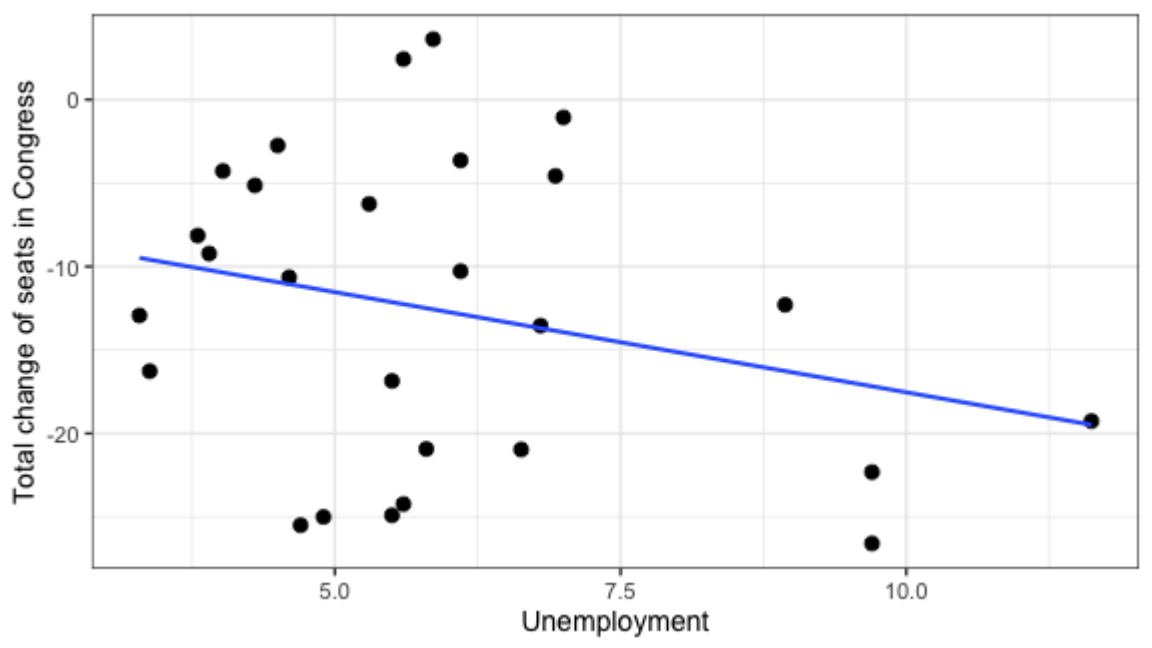
Second shuffle, visualized

```
shuffle2 %>%
  ggplot(aes(x = unemp, y = change)) +
  geom_point(size = 3) +
  theme_bw(base_size = 14) +
  xlab("Unemployment") +
  ylab("Total change of seats in Congress")
```



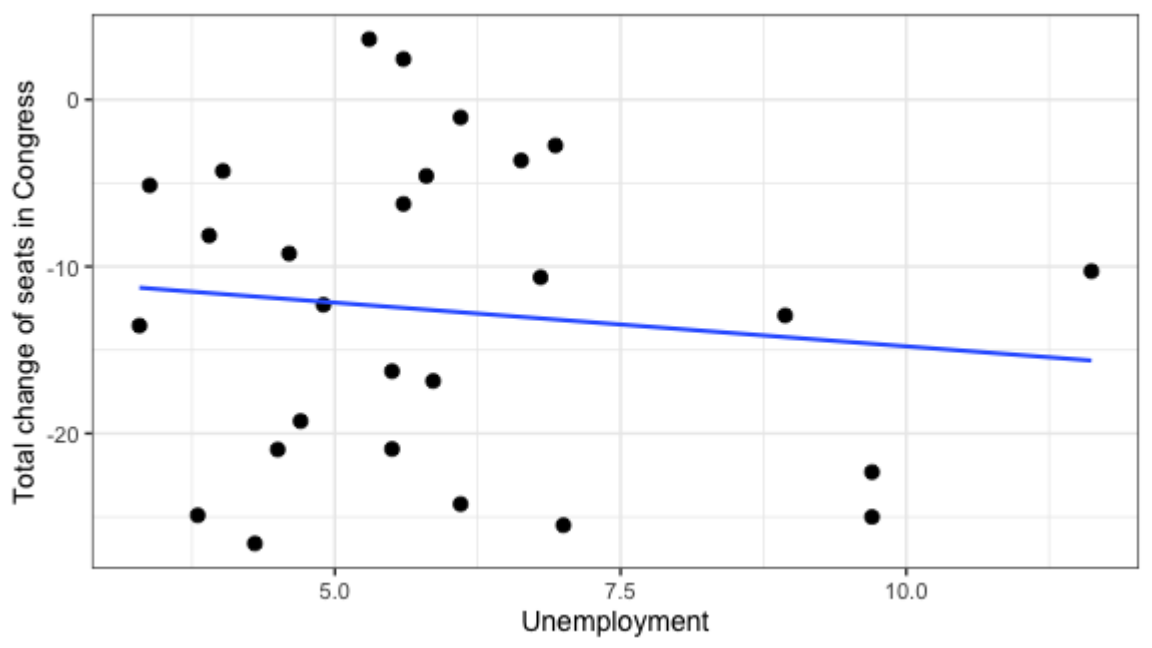
Second shuffle, visualized

```
shuffle2 %>%
  ggplot(aes(x = unemp, y = change)) +
  geom_point(size = 3) +
  theme_bw(base_size = 14) +
  xlab("Unemployment") +
  ylab("Total change of seats in Congress") +
  stat_smooth(method = "lm", se = FALSE)
```



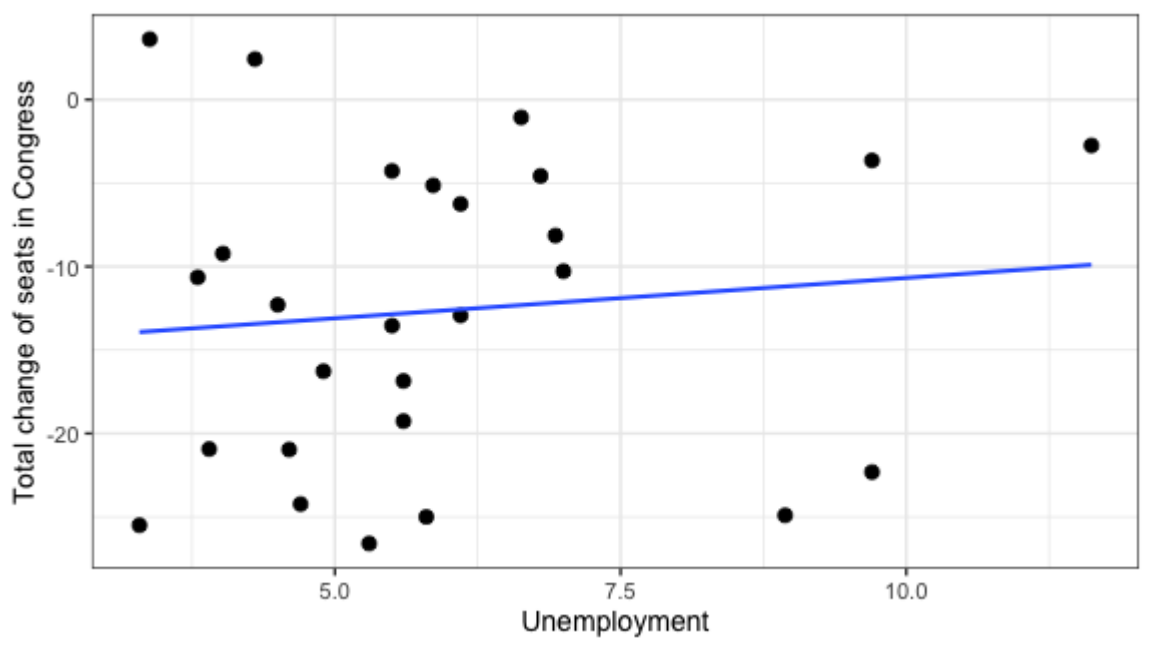
Third shuffle, visualized

```
shuffle3 %>%
  ggplot(aes(x = unemp, y = change)) +
  geom_point(size = 3) +
  theme_bw(base_size = 14) +
  xlab("Unemployment") +
  ylab("Total change of seats in Congress") +
  stat_smooth(method = "lm", se = FALSE)
```



Fourth shuffle, visualized

```
shuffle4 %>%
  ggplot(aes(x = unemp, y = change)) +
  geom_point(size = 3) +
  theme_bw(base_size = 14) +
  xlab("Unemployment") +
  ylab("Total change of seats in Congress") +
  stat_smooth(method = "lm", se = FALSE)
```

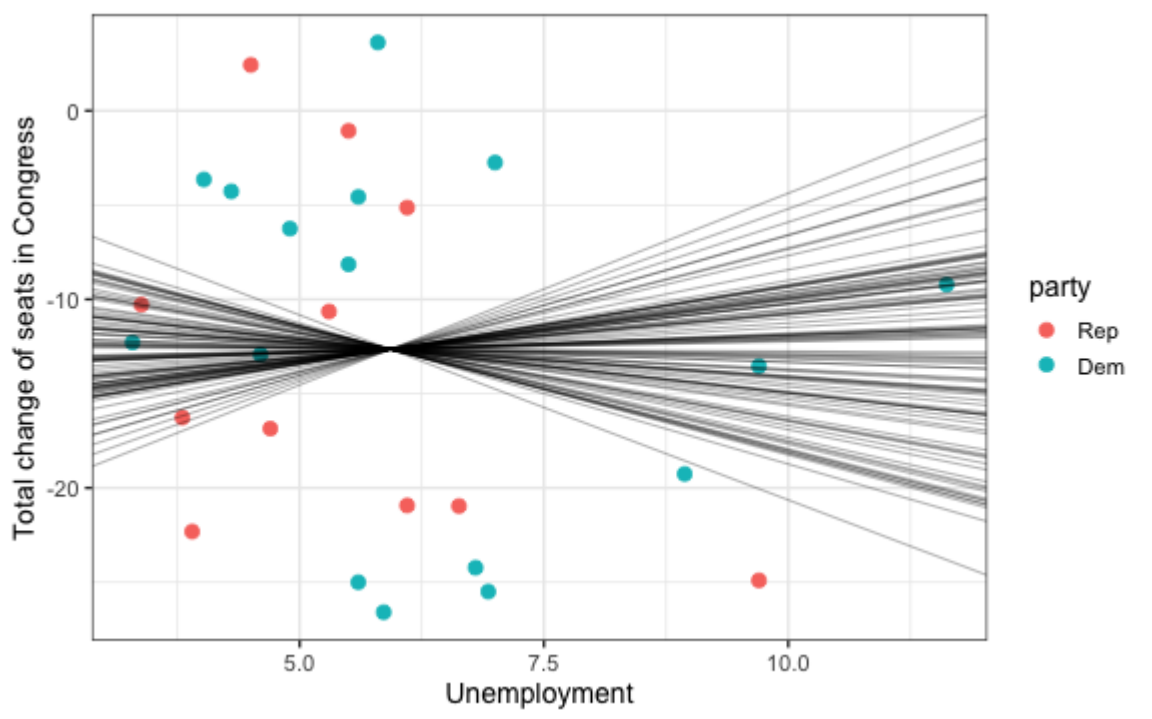


Generate 500 permuted b_1's

```
null <- ump %>%
  specify(change ~ unemp) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 500, type = "permute") %>%
  calculate(stat = "slope")
null
```

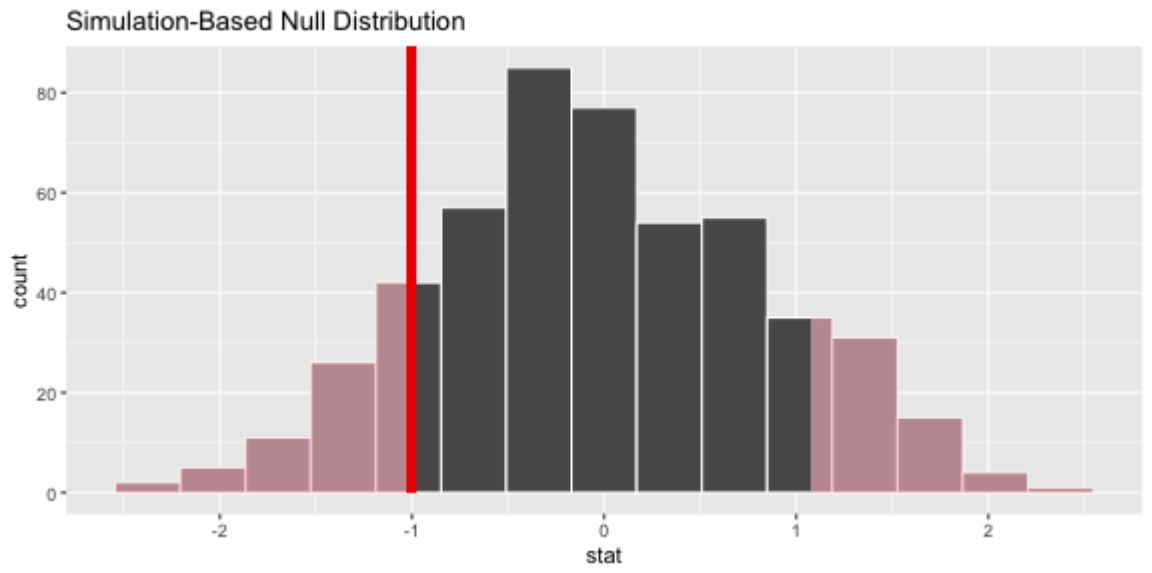
```
## # A tibble: 500 x 2
##    replicate     stat
##        <int>    <dbl>
##  1         1 -0.391
##  2         2  1.25
##  3         3  0.622
##  4         4  1.99
##  5         5 -0.876
##  6         6 -0.239
##  7         7 -0.00730
##  8         8 -0.479
##  9         9  1.04
```

Visualize 100 permuted b_1's



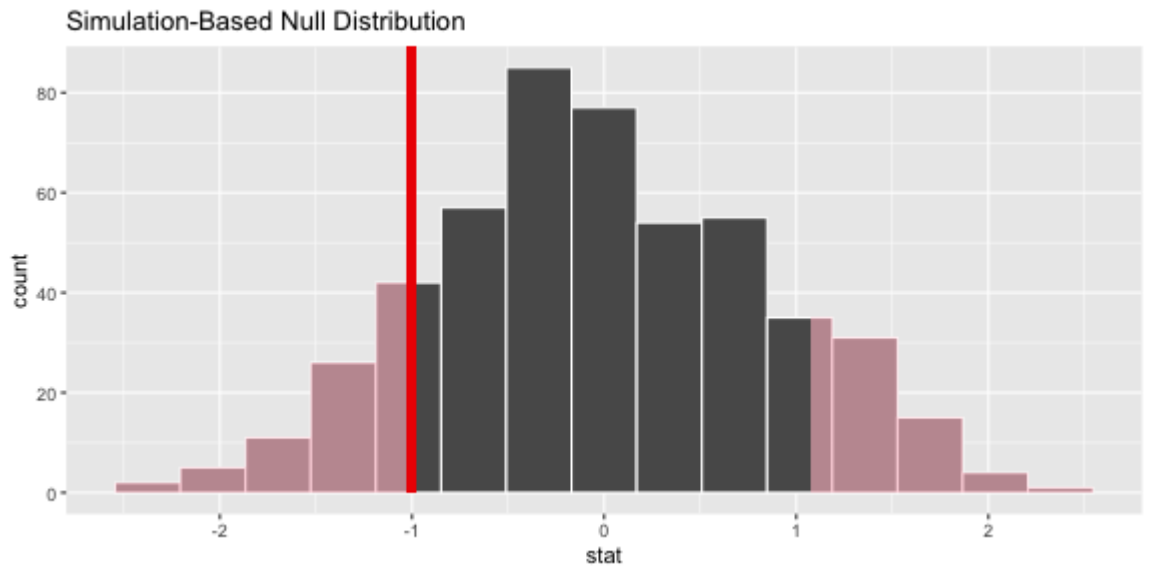
Sampling dist. of b_1

```
null %>%
  visualize() +
  shade_pvalue(obs_stat = obs_slope,
               direction = "both")
```



Sampling dist. of b_1

```
null %>%
  visualize() +
  shade_p_value(obs_stat = obs_slope,
                direction = "both")
```



Reigning theory: voters will punish candidates from the Presidents party at the ballot box when unemployment is high.

H-tests for regression

```
m0 <- lm(change ~ unemp, data = ump)
summary(m0)
```

```
##
## Call:
## lm(formula = change ~ unemp, data = ump)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.011  -7.861  -0.183   7.389  16.140
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.714      5.457   -1.23     0.23
## unemp         -1.001      0.872   -1.15     0.26
##
## Residual standard error: 9.11 on 25 degrees of freedom
## Multiple R-squared:  0.0501,    Adjusted R-squared:  0.0121
## F-statistic: 1.32 on 1 and 25 DF,  p-value: 0.262
```

H-tests for regression

- Each line in the summary table is a hypothesis test that the parameter is zero.
- Under certain conditions, the test statistic associated with b's is distributed like t random variables with $n - p$ degrees of freedom.

$$\frac{b - \beta}{SE} \sim t_{df=n-p}$$

```
t_stat <- (-1.0010 - 0)/0.8717
pt(t_stat, df = 27 - 2) * 2
```

```
## [1] 0.262
```

Conditions for inference

1. **Linearity**: linear trend between X and Y, check with residual plot.
2. **Independent errors**: check with residual plot for serial correlation.
3. **Normally distributed errors**: check for linearity in qq-plot.
4. **Errors with constant variance**: look for constant spread in residual plot.

natesilver: Dude. It's not even six examples really. It's four.

harry: Who are your four?

natesilver: Dwight Eisenhower, Ronald Reagan, Bill Clinton and George W. Bush are the only presidents in American history to be term-limited. Obama will be the fifth.

And I don't care if you get the same regression results with four.

harry: And did you know that Obama's approval rating is below the average approval rating for those four? And it's not particularly close either.

natesilver: The problem is that running a regression model based on an n of four is inherently kind of ridiculous.

NERD FIGHT!

Postscript: Sieve Metaphor