PPOL561 Quasi-experimental Designs: Regression Discontinuity

Merykokeb Belay

Using Regression Discontinuity (RD) to estimate the effect of political party on ideology.

Congressional elections are decided by a clear rule: whoever gets the most votes in November wins. Because virtually every congressional race in the United States is between two parties, whoever gets more than 50 percent of the vote wins. We can use this fact to estimate the effect of political party on ideology. Some argue that Republicans and Democrats are very distinctive; others argue that members of Congress have strong incentives to respond to the median voter in their districts, regardless of party. We can assess how much party matters by looking at the ideology of members of Congress in the 112th Congress (which covered the years 2011 and 2012).

```
congress <- read_csv("Data/congress.csv")
head(congress)
```

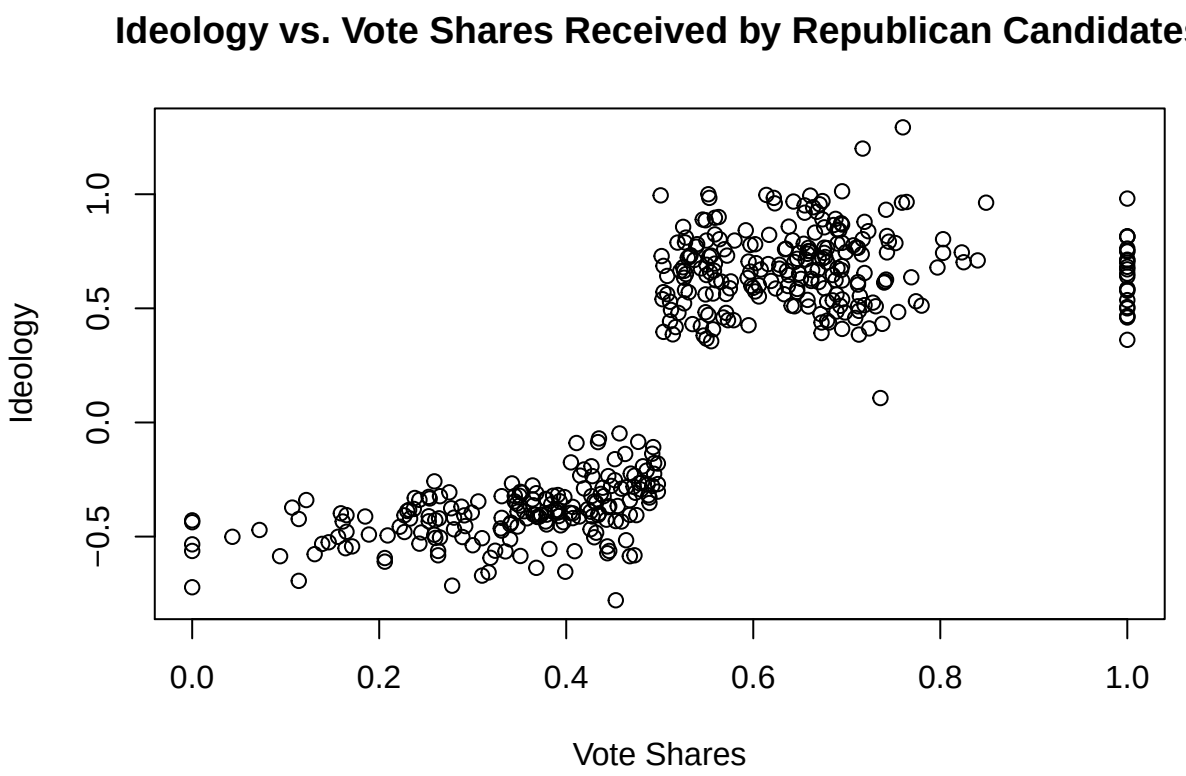
```
## # A tibble: 6 x 7
##   ChildPoverty MedianIncome Obama2008 GOP2party2010 GOPwin2010 WhitePct Ideology
##          <dbl>        <dbl>     <dbl>         <dbl>      <dbl>    <dbl>    <dbl>
## 1        0.270        39597     0.390         1              1    0.680    0.460
## 2        0.290        37289     0.360         0.511          1    0.670    0.443
## 3        0.25         38079     0.430         0.595          1    0.650    0.426
## 4        0.260        35719     0.23          1              1    0.900    0.467
## 5        0.19         43832     0.38          0.580          1    0.780    0.797
## 6        0.12         56552     0.23          1              1    0.890    0.586
```

In trying to answer this problem, if we choose to explain congressional ideology as a function of political party only endogeneity may creep in because there are plenty of factors in the error term that are correlated with both ideology and political party, creating an environment where we have omitted variable bias. For instance, the demographic distribution of a candidates congressional district. Therefore, we would worry that any estimate we obtain would actually reflect the impact of that hidden factor rather than just the impact of political party on ideology.

Alternatively, a RD model can fight endogeneity by including an assignment variable that essentially breaks off the connection between the treatment variable (political party) and the error term. This assignment variable would partition our observations into two categories that we can compare by assessing their differences at the cutoff for assignment into either category. RD not only provides a way to break that connection between our key independent variable of interest and the error term, but it also provides us with a visual representation of the effect of political party on ideology.

EDA.

```r
## generate scatterplot
plot(congress$GOP2party2010, congress$Ideology,
     main="Ideology vs. Vote Shares Received by Republican Candidates",
     ylab = 'Ideology',
     xlab = 'Vote Shares')
```



Based on this plot, I suspect the RD will show that there is a statistically significant difference in ideology between Republicans who were just short of winning and Republicans who got 50% or more of the vote share. It looks like Republicans who have a vote share of 50% or higher tend to vote more conservative in Congress.

Model Specification.

$Ideology_i = \beta_0 + \beta_1 GOPwin2010 + \beta_2 (GOP2party2010_i - 0.50) + \nu_i$

- Ideology: the dependent variable with a range from –0.779 to 1.293. Higher values indicate more conservative voting in Congress.
- β_0: the average ideology of Republicans who have a vote share that is less than 50%.
- β_1: indicates the difference in ideology between Republicans who have a vote share of 50% or higher and those who have a vote share that is less than 50%.
- GOPwin2010: a dummy variable which takes a value of 1 if the Republican candidate receives 50% or more of the vote share
- β_2: represents the slope on either side of the graph; represents the relationship between vote shares and ideology.
- (GOP2party2010_i - 0.50): this is the assignment variable which indicates if observations will be place below the cutoff (those with values less than 0) or above the cutoff (those with values above 0)
- ν_i: the error term

Estimating a basic RD model.

```
## determine how much above or below an observation is from the cutoff
congress<-congress%>%
  mutate(X=GOP2party2010 - 0.50)

## view modified dataframe
congress%>%
  head()
```

```
## # A tibble: 6 x 8
##   ChildPoverty MedianIncome Obama2008 GOP2party2010 GOPwin2010 WhitePct Ideology
##          <dbl>        <dbl>     <dbl>         <dbl>      <dbl>    <dbl>    <dbl>
## 1        0.270        39597     0.390         1              1    0.680    0.460
## 2        0.290        37289     0.360         0.511          1    0.670    0.443
## 3        0.25         38079     0.430         0.595          1    0.650    0.426
## 4        0.260        35719     0.23          1              1    0.900    0.467
## 5        0.19         43832     0.38          0.580          1    0.780    0.797
## 6        0.12         56552     0.23          1              1    0.890    0.586
## # ... with 1 more variable: X <dbl>
```

```
## basic RD model
mod_1<-lm(Ideology~GOPwin2010+X, data=congress)
broom::tidy(mod_1)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   -0.360    0.0141    -25.5  4.11e- 88
## 2 GOPwin2010     0.995    0.0238     41.8  2.52e-153
## 3 X              0.230    0.0571      4.04 6.46e-  5
```

According to the regression output, Republicans who won the congressional race vote more conservatively in Congress than Republicans who did not win the congressional race. Specifically, Republican who won the race rank 0.99 units higher on the conservative range. This estimate is statistically significant at the 99.9% confidence level since the pvalue is less than α=0.001. The average ideology for Republicans who did not win the congressional race is approximately -0.4. This estimate is also statistically significant at the 99.9% confidence level since the pvalue is less than α=0.001. The estimate on the assignment variable X is positive and represents the slope of the line on either side of the cutoff. Because it is positive, it means the model estimates that an increase in vote share on either sides of the cutoff is accompanied by an increase in conservative voting in Congress.

Estimate a varying slopes model.

```
## varyng slopes model
mod_2<-lm(Ideology~GOPwin2010*X, data=congress)
broom::tidy(mod_2)
```

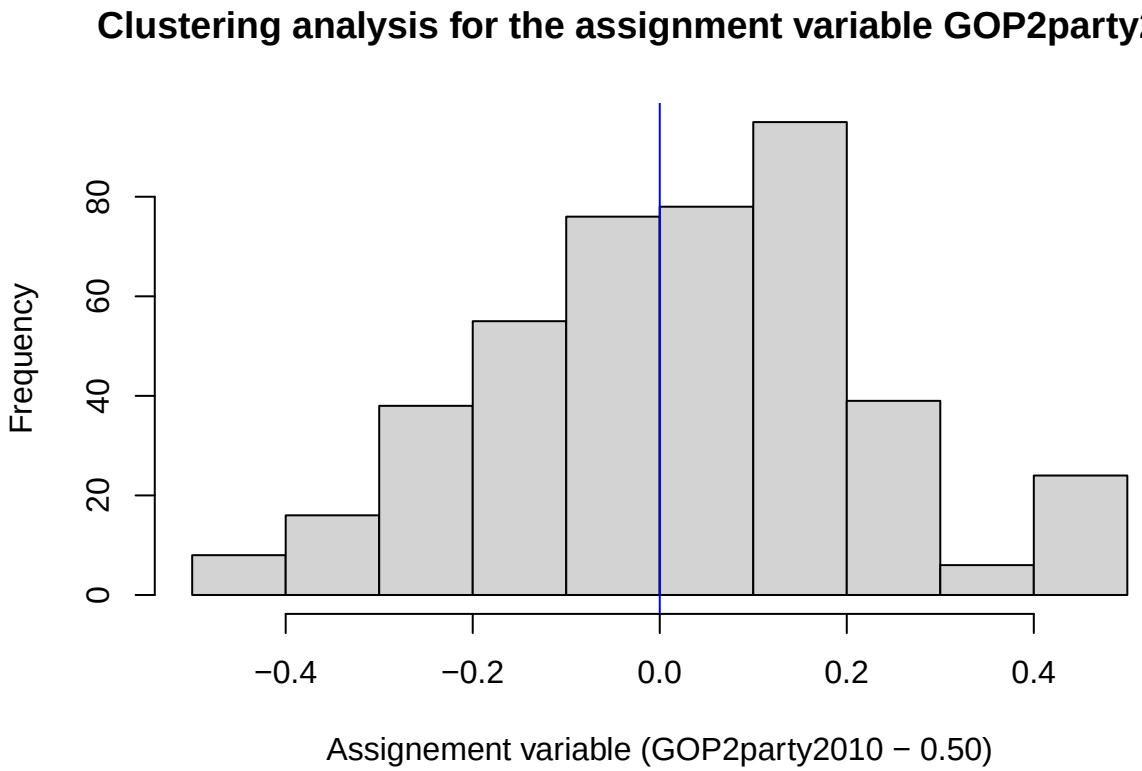
```
## # A tibble: 4 x 5
```

```
##   term          estimate std.error statistic   p.value
##   <chr>            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    -0.313    0.0176    -17.8  2.42e- 53
## 2 GOPwin2010      0.982    0.0236     41.6  1.22e-152
## 3 X               0.529    0.0896      5.90 7.39e-  9
## 4 GOPwin2010:X   -0.489    0.115      -4.26 2.48e-  5
```

After implementing the varied slopes model, there is a slight adjustment to the coefficient estimate on GOPwin2010. This model estimates that Republicans who won the congressional race vote about 0.98 units more conservatively in Congress than Republicans who did not win. As such, the average conservativism of Republicans who won the congressional race is about 0.67. This estimate is statistically significant at the 99.9% confidence level since the pvalue is less than α=0.001. The average conservativism of Republicans who did not win is about -0.31. This estimate is also statistically significant at the 99.9% confidence level since the pvalue is less than α=0.001. The slope for the observations on the right side of the cutoff (i.e., Republicans with 50% or more votes) is about 0.49 units less than the slope on the left side of the cutoff. Note that, like the previous regression output, both slopes are positive, indicating that increase in vote share for Republicans is generally associated with an increase in conservativism.

Assess clustering of the dependent variable just above the cutoff.

```
## check for clustering
hist(congress$X, main = "Clustering analysis for the assignment variable GOP2party2010", xlab = "Assigne
abline(v=0, col="blue")
```



We can see that, right around the cutoff, the bars tend to be uniform. There isn't a pronounced bump up or down in the two bars on either sides of the blue line. However, we see some clustering on the far right side of the cutoff (the second bar in particular). Ideally, we would want the bars to be relatively similar across

the graph, but since we are mostly concerned about what is going on right around the cutoff, there doesn't seem to be a pronounced clustering.

To assess whether there are discontinuities in the other variables, I will run models with each control variable as the dependent variable.

```
## check for discontinuity for ChildPoverty
mod_3<-lm(ChildPoverty~GOPwin2010+X, data=congress)
broom::tidy(mod_3)
```

Child Poverty

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   0.208    0.00746    27.8   1.84e-98
## 2 GOPwin2010   -0.0122   0.0126     -0.963 3.36e- 1
## 3 X            -0.0977   0.0303     -3.22  1.37e- 3
```

The output above shows that the coefficient estimate on GOPwin2010 is not statistically significant. It has a pvalue value that is well above α=0.1, thus preventing the coefficient estimate from being statistically significant at any of the conventional confidence levels. This a good thing for our analysis. A key assumption in RD is that the only thing jumping at the point of discontinuity (right at the cutoff) is the dependent variable. Given that the estimate is not statistically significant, we can conclude that it is not jumping at discontinuity.

```
#### check for discontinuity for MedianIncome
mod_4<-lm(MedianIncome~GOPwin2010+X, data=congress)
broom::tidy(mod_4)
```

Median Income

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   52319.     1273.    41.1   2.61e-151
## 2 GOPwin2010     -388.     2154.    -0.180 8.57e-  1
## 3 X              2920.     5172.     0.565 5.73e-  1
```

Similar to the diagnostic test above, the coefficient estimate on GOPwin2010 is not statistically significant at any of the conventional confidence levels because the pvalue is greater than α=0.1. Again, this means that MedanIncome is not jumping at discontinuity; we can hesitantly continue to believe our assumption that the only thing jumping at discontinuity is the dependent variable Ideology.

```
### check for discontinuity for Obama2008

mod_5<-lm(Obama2008~GOPwin2010+X, data=congress)
broom::tidy(mod_5)
```

Obama2008

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   0.577    0.00624     92.4  1.42e-286
## 2 GOPwin2010   -0.0434   0.0106      -4.11 4.75e-  5
## 3 X            -0.552    0.0254     -21.8  1.57e- 71
```

Unlike the above tests, here we see that the coefficient estimate on GOPwin2010 is statistically significant–in fact, it is statistically significant at the 99.9% confidence level as it has a pvalue that is less than α=0.001. This means that Obama2008 jumps at the discontinuity.This is worrying because it violates a key assumption in RD. Although the partial fix for this violation is to control for Obama2008 in our modeling process going forward, we should still worry that if there is a jump at cutoff for this variable, that there could possible be a jump at cutoff for a variable we haven't measure/aren't accounting for. If that is the case, then we cant be sure that the jump at discontinuity for the dependent variable is only because of the treatment.

```
### check for discontinuity for WhitePct

mod_6<-lm(WhitePct~GOPwin2010+X, data=congress)
broom::tidy(mod_6)
```

WhitePct

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   0.645     0.0176     36.6  1.53e-134
## 2 GOPwin2010    0.0595    0.0298      2.00 4.64e-  2
## 3 X             0.494     0.0716      6.90 1.87e- 11
```

Like the test above, here we see that the coefficient estimate on GOPwin2010 is statistically significant at the 90% confidence level as it has a pvalue that is less than α=0.1. This means that WhitePct jumps at the discontinuity. This is worrying because it violates a key assumption in RD. Moving forward, we will need to control for this variable. The concerns mentioned above apply here as well: we should still worry that if there is a jump at cutoff for this variable, that there could possible be a jump at cutoff for a variable we haven't measure/aren't accounting for. If that is the case, then we cant be sure that the jump at discontinuity for the dependent variable is only because of the treatment.

Estimating a varying-slopes model controlling for ChildPoverty, MedianIncome, Obama2008, and WhitePct.

```
## varying slopes model; control for covariates
mod_7<-lm(Ideology~GOPwin2010*X+ChildPoverty+MedianIncome+Obama2008+WhitePct, data=congress)
broom::tidy(mod_7)
```

```
## # A tibble: 8 x 5
##   term            estimate  std.error statistic   p.value
##   <chr>              <dbl>      <dbl>     <dbl>     <dbl>
## 1 (Intercept)  -0.0134     0.127         -0.105 9.16e-  1
## 2 GOPwin2010    0.963      0.0229        42.1   1.57e-153
## 3 X            -0.0476     0.142         -0.336 7.37e-  1
## 4 ChildPoverty  0.265      0.189          1.41  1.60e-  1
## 5 MedianIncome  0.00000212 0.00000101     2.10  3.61e-  2
## 6 Obama2008    -0.791      0.124         -6.36  5.16e- 10
## 7 WhitePct     -0.0609     0.0498        -1.22  2.22e-  1
## 8 GOPwin2010:X -0.177      0.137         -1.30  1.96e-  1
```

After including the controls and using a varying slopes model, the coefficient estimate on the independent variable of interest GOPwin2010 remains stable at 0.96 and statistically significant at the 99.9% confidence level as it has a pvalue less than α=0.001. However, there is a change in the intercept–it was about -0.13 in the previous sections, but now it is about -0.013. This means that the average Ideology/convervativism of Republicans who get less than 50% of the vote share is -0.013. Note that unlike before, this coefficient estimate is not statistically significant at any of the conventional confidence levels since the pvalue is greater than α=0.1. Another notable difference is the difference in slope between the left and right side of the cutoff. Recall that previously, the slope of both sides was positive with the one the right side being lower than the one on the left. Now we see that the slope on both sides of the cutoff is negative, with the right having a lower slope.

The coefficient estimate on ChildPoverty is about 0.27, indicating that a one percent increase in child poverty rates in the district is associated with a 0.27 unit increase in the conservativism of the member in Congress. The coefficient estimate on WhitePct is about -0.061, indicating that a one percent increase in the non-Hispanic white population in the district is associated with a 0.061 unit decrease in the conservativism of the the member in Congress. Both of these estimates are not statistically significant at any of the conventional confidence levels as they have pvalues that are greater than α=0.1. However, recall that WhitePct was actually statistically significant in the diagnositc test–it had a somewhat meaningful relationship with the independent variable of interest. Here, see that its relationship with the dependent variable is not statistically significant.

Lastly, the coefficient estimates on MedianIncome is about 0.00, indicating that a one unit increase in MedianIncome in the district is associated with a 0 unit increase in conservativism (i.e, median income of the district doesn't seem to have an impact on conservativism). This estimate is statistically significant at the 90% confidence level as it has a pvalue less than α=0.1. Similarly, Obama2008, which jumped at the discontinuity in the diagnostic tests, has a coefficient estimate of about -0.79. This indicates that a one unit increase in the percent of votes for Barack Obama in the district in 2008 presidential election is associated with a 0.79 unit decline in conservativism of the member of Congress. This coefficient estimate is statistically significant at the 99.9% confidence level as it has a pvalue less than 0.001.

Estimating a quadratic RD model.

```
## quadratic RD model
mod_8<-lm(Ideology~GOPwin2010*X+GOPwin2010*I(X^2), data=congress)
broom::tidy(mod_8)
```

```
## # A tibble: 6 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)         -0.282    0.0243    -11.6  3.59e-27
## 2 GOPwin2010           0.906    0.0336     27.0  4.03e-94
## 3 X                    1.01     0.275       3.66 2.85e- 4
## 4 I(X^2)               1.13     0.614       1.84 6.71e- 2
## 5 GOPwin2010:X        -0.403    0.358      -1.12 2.61e- 1
## 6 GOPwin2010:I(X^2)   -2.24     0.749      -2.99 2.96e- 3
```

After implementing a more flexible model, there is a slight difference in the coefficient estimate on GOPwin2010–the coefficient estimate was about 0.96, while now it is 0.90. Accordingly, Republicans who win the congressional race vote 0.90 units more conservatively in Congress. This estimate remains statistically significant at the 99.9% confidence level as it has a pvalue less than α=0.001. Our estimate of the treatment effect is fairly stable, with a very slight change when allowing the model fit on either side of the cutoff to be more flexible. In terms of the slopes, we see that slope before the cutoff is approximately 1.1 while the slope after the cutoff is -2.6. This means that, as we move across the graph starting from the left side (i.e.increasing in vote shares), conservativism tends to rise, however, after the cutoff, we see that the relationship between vote shares and conservativism seems to be negative.

Estimating a varying-slopes model with a window of GOP vote share from 0.4 to 0.6.

```
## varying slopes with restrictions
mod_9<-lm(Ideology~GOPwin2010*X, data=congress%>%filter(GOP2party2010>0.4 & GOP2party2010<0.6))
broom::tidy(mod_9)
```

```
## # A tibble: 4 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    -0.270    0.0349    -7.74  1.39e-12
## 2 GOPwin2010      0.893    0.0489    18.2   8.34e-40
## 3 X               1.10     0.632      1.75  8.27e- 2
## 4 GOPwin2010:X   -0.649    0.906     -0.716 4.75e- 1
```

After imposing a restriction and running a varying slopes model, the coefficient estimate on GOPwin2010 has slightly changed–it is now about 0.89, while it was about 0.90 in the previous section. Despite a reduction in sample size, it remains statistically significant at the 99.9% confidence level as it has a pvalue less than α=0.001. Furthermore, the slope on the left side of the cutoff is positive and has a value of 1.1, while the slope on the right side of the cutoff, which is also positive, has a value of about 0.5.

In conclusion, the most credible estimate would be 0.89 from the varying-slopes model with a window. This is because it allows us to assess the difference in outcome between those who were just short of getting the treatment and those who barely passed the cutoff and qualified for treatment without having to worry about the functional form for the observations far away from the cutoff. We only care about those closest to the cutoff on either side– since they are very similar to one another, any differences between their outcome would be attributed to the treatment. Additionally, since we still retain statistical significance after imposing a window, 0.89 is the most credible estimate.