EDA Long Form

John Lee

December 9, 2018

Introduction

My data source is the General Social Survey (GSS), which is run by the National Opinion Research Center (NORC) at the University of Chicago. The GSS is one of the best known sources of nationally representative public opinion survey data in the U.S., with data that dates back to the 1970s. The GSS uses a stratified probability-based sampling method and asks a large number of questions related to social, cultural, political, and economic issues. The survey is conducted once every two years. Since 2006, the GSS has been running a series of three-wave panels. I am using the first set of panel data, which surveyed the same subjects in 2006, 2008, and 2010.

My research questions are the following: 1. What predicts changes in political ideology (political conservatism) over time? 2. What predicts changes in party ID (Republican party ID) over time? 3. How does gender shape the effects of the factors?

Here is a link to the dataset: https://gssdataexplorer.norc.org/

The citation for the data source is the following:

Smith, Tom W, Michael Davern, Jeremy Freese, and Michael Hout. General Social Surveys, 1972-2016 [machine-readable data file] /Principal Investigator, Tom W. Smith; Co-Principal Investigator, Michael Davern; Co-Principal Investigator, Jeremy Freese; Co-Principal Investigator, Michael Hout; Sponsored by National Science Foundation. –NORC ed.– Chicago: NORC at the University of Chiago [producer]; Storrs, CT: The Roper Center for Public Opinion Research, University of Connecticut [distributor], 2018.

Load Packages

```
# Loading package(s)
library(tidyverse)
library(forcats)
library(modelr)
library(haven)
library(plm)
```

Part 1: Exploring Variable Distributions

In this first section, I explore the distributions of the continuous variables in my dataset using faceted histograms: age, years of education, logged family income, conservative ideology, and Republican party ID. The data from each wave are shown separately for each variable: Wave 1 in red, Wave 2 in green, and Wave 3 in blue. I want to make sure that none of the variables have a serious skew; if so, I would consider logging the variables before proceeding to the modeling stage. Family income is already logged (natural log).

By looking at the distributions below, we can see that most variables either have a normal (or quasi-normal) distribution; unsurprisingly, the distributions look similar across the waves. The most interesting finding is that while the scale for conservative ideology is normally distributed (i.e., with the most common value being a 4, for "Moderate"), the most common values for the Republican party ID scale are 1 and 2, which stand for "Strong Democrat" and "Not Strong Democrat," respectively. This may suggest that ideology does not cleanly map onto party ID, as many people who self-identify as ideological moderates seem to prefer the Democratic Party (see Figures 4 and 5). I'll return to this topic again soon.

Figure 1: Distribution of Age

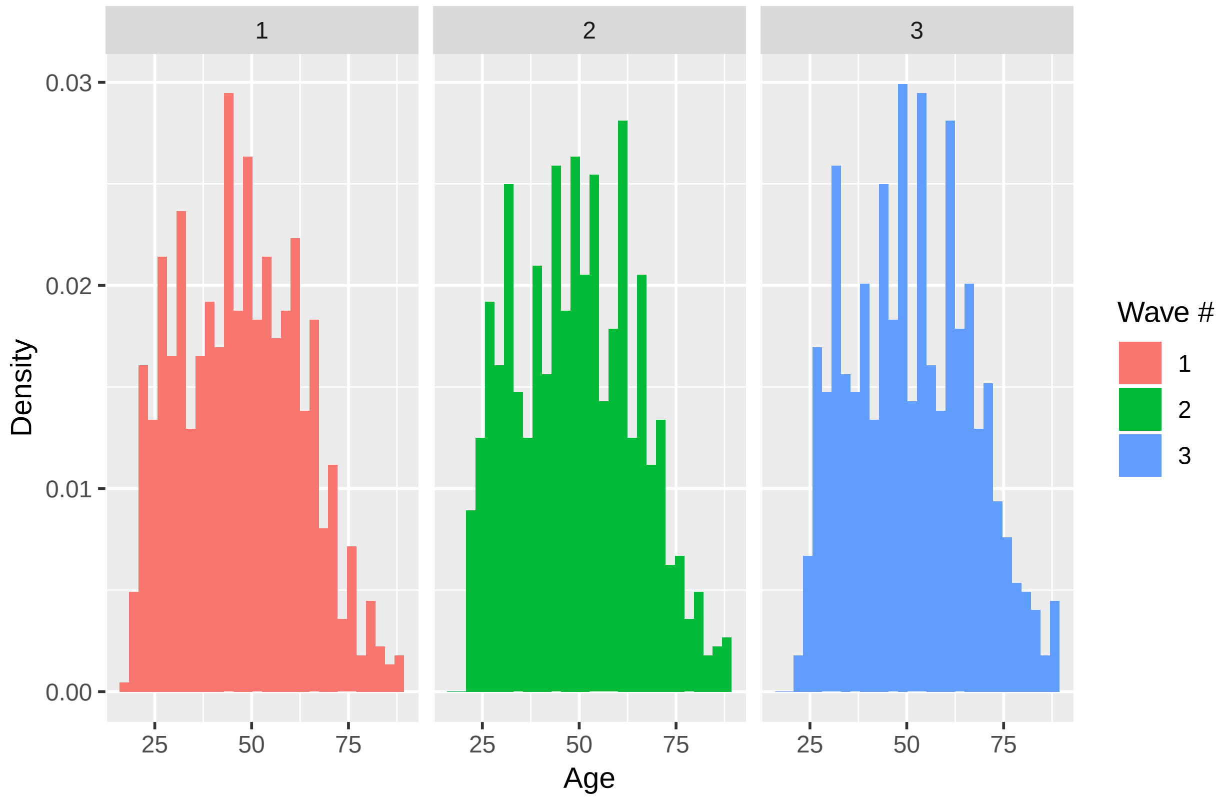


Figure 2: Distribution of Years of Education

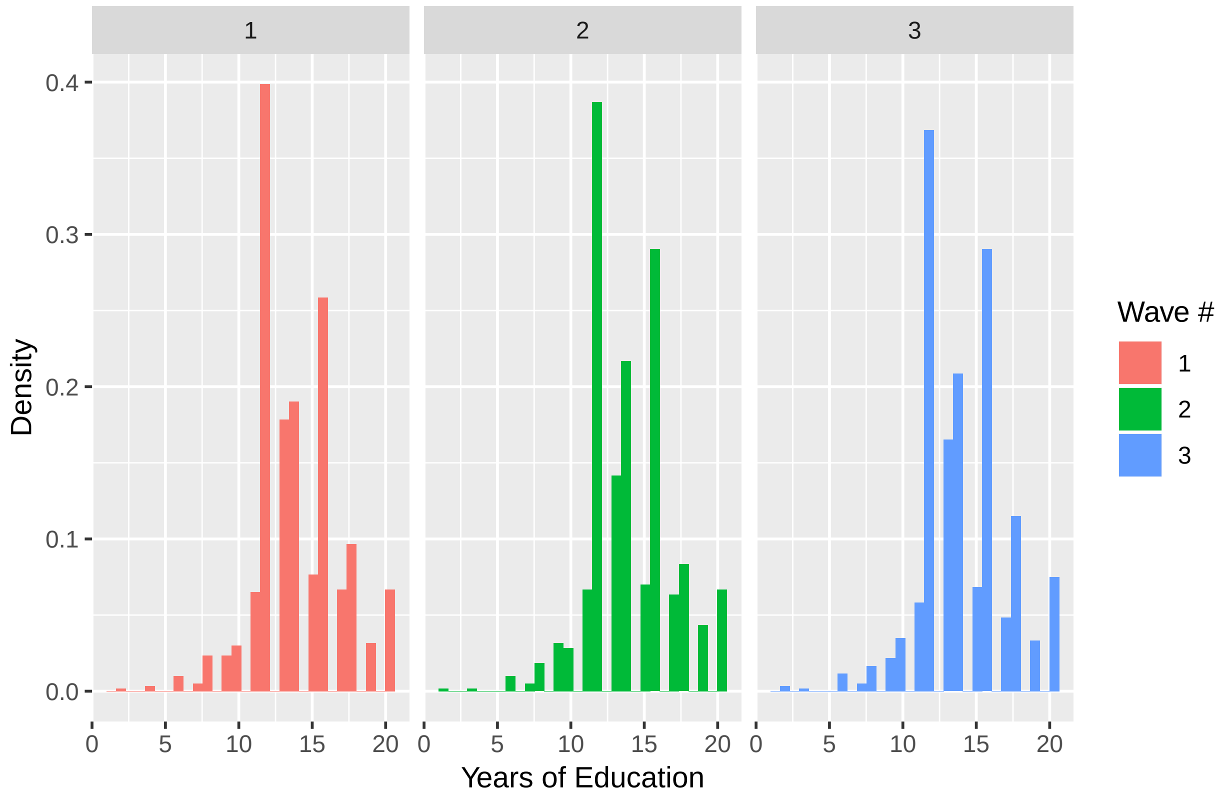
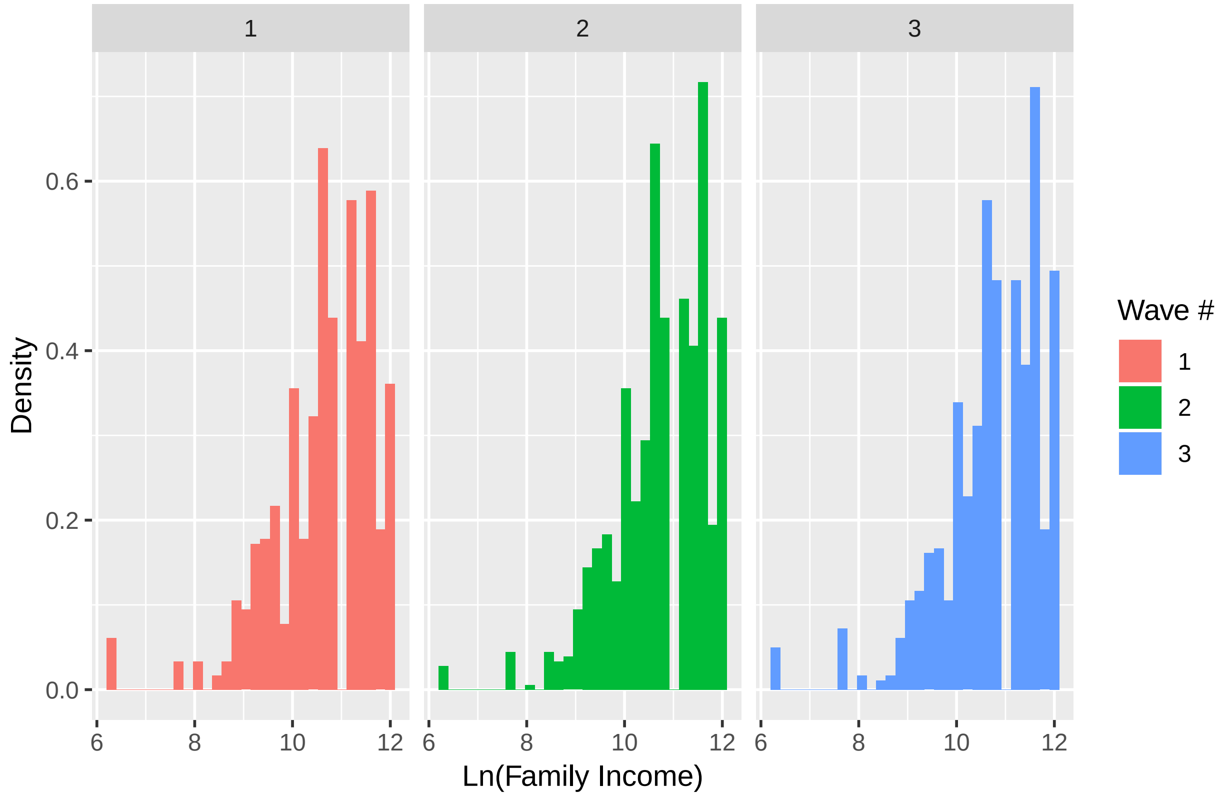


Figure 3: Distribution of Ln(Family Income)



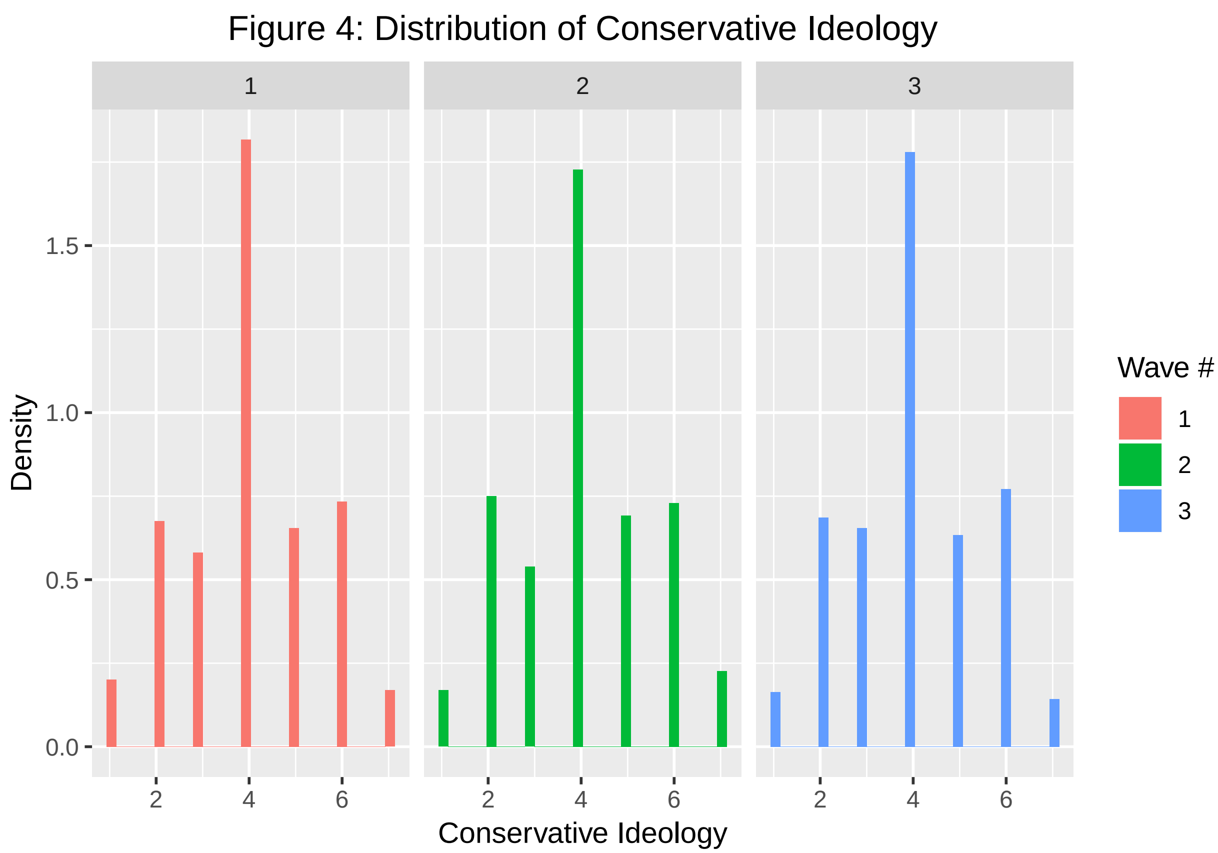
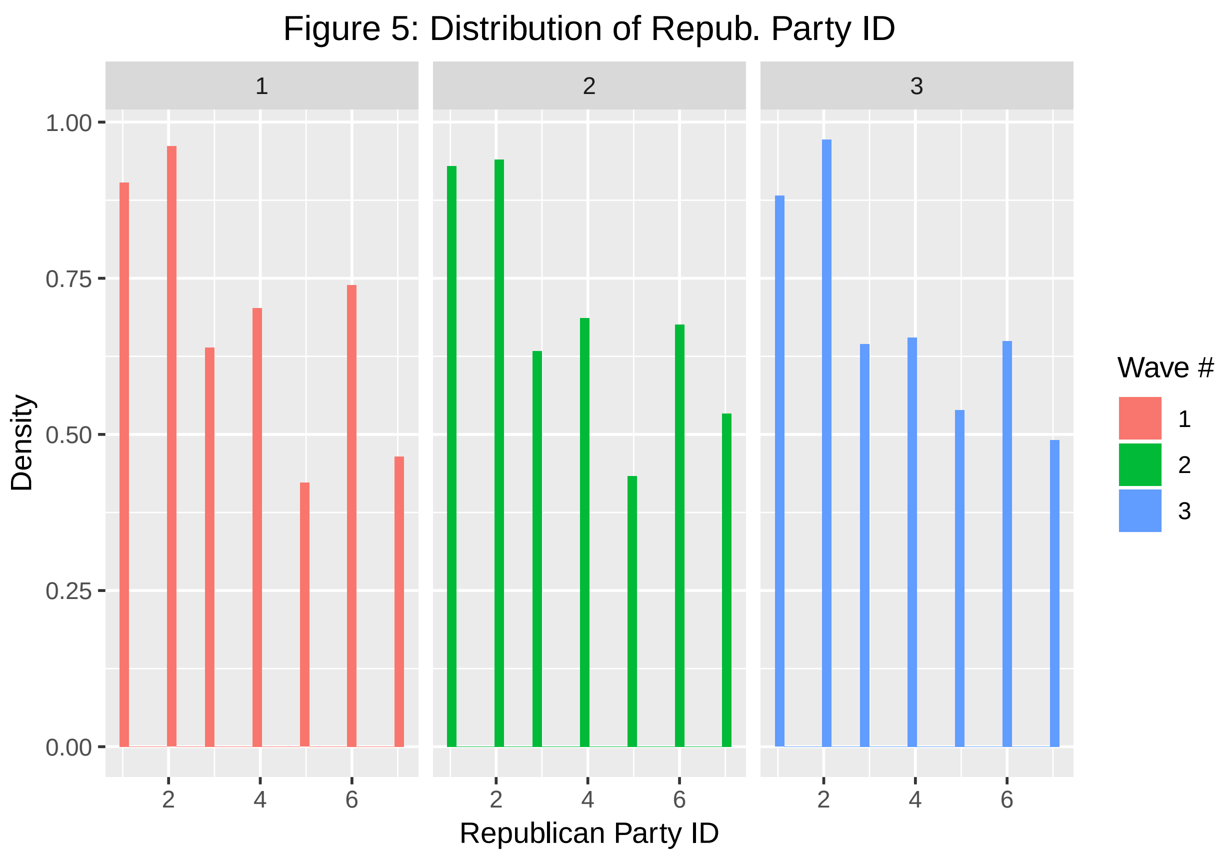


Figure 4: Distribution of Conservative Ideology
1
2
3
Density
1.5
1.0
0.5
0.0
2
4
6
Conservative Ideology
Wave #
1
2
3



Part 2: Exploring Bivariate Relationships

In Part 2, I explore bivariate relationships. At this point, I want to clarify that I'll only focus on bivariate relationships between time-variant predictors (i.e., variables that may change over time) and my two dependent variables. I do this because in the modeling stage of the EDA, I plan to run linear fixed effects (FE) models, not OLS models. One common threat to causal inference in the analysis of survey data is the presence of unobserved heterogeneity (e.g., omitted variables, reverse causation). However, given the panel structure of my data, I can use FE models that only use within-unit variance over time to compute model parameters – which provide a stronger basis for causal inference. One downside to this is that FE models automatically drop variables that don't change over time (e.g., race). This is why in this section, I'll only focus on examining the bivariate relationship between time-variant predictors (e.g., years of education, religiosity, marital status) and the two DVs of interest: conservative ideology and Republican party ID.

First, I look at the potential effects of two personal characteristics: religiosity and and marital status. The intuition here is that individuals who go through a divorce may adopt more liberal, progressive attitudes; also, it's possible that becoming less religious may also be associated with a move towards more liberal beliefs (e.g., on social/cultural issues). I use boxplots because I am displaying the link between a categorical variable (e.g., marital status) and a continuous variable (e.g., political conservatism scale). The plots are also faceted by gender, because that is one of my key research questions: i.e., whether the link between the predictors and the two DVs is shaped in some way by gender.

Please note that in general, we want to display the DV on the Y axis. However, I had to use `coord_flip()` for the four plots (Figures 6a - 7b) below because the text for religiosity and marital status did not fit on the x-axis. For religiosity, we can observe what we may have expected: indeed, religiosity appears to have a positive association with both political conservatism and a stronger tie to the Republican party (Figures 6a, 6b).

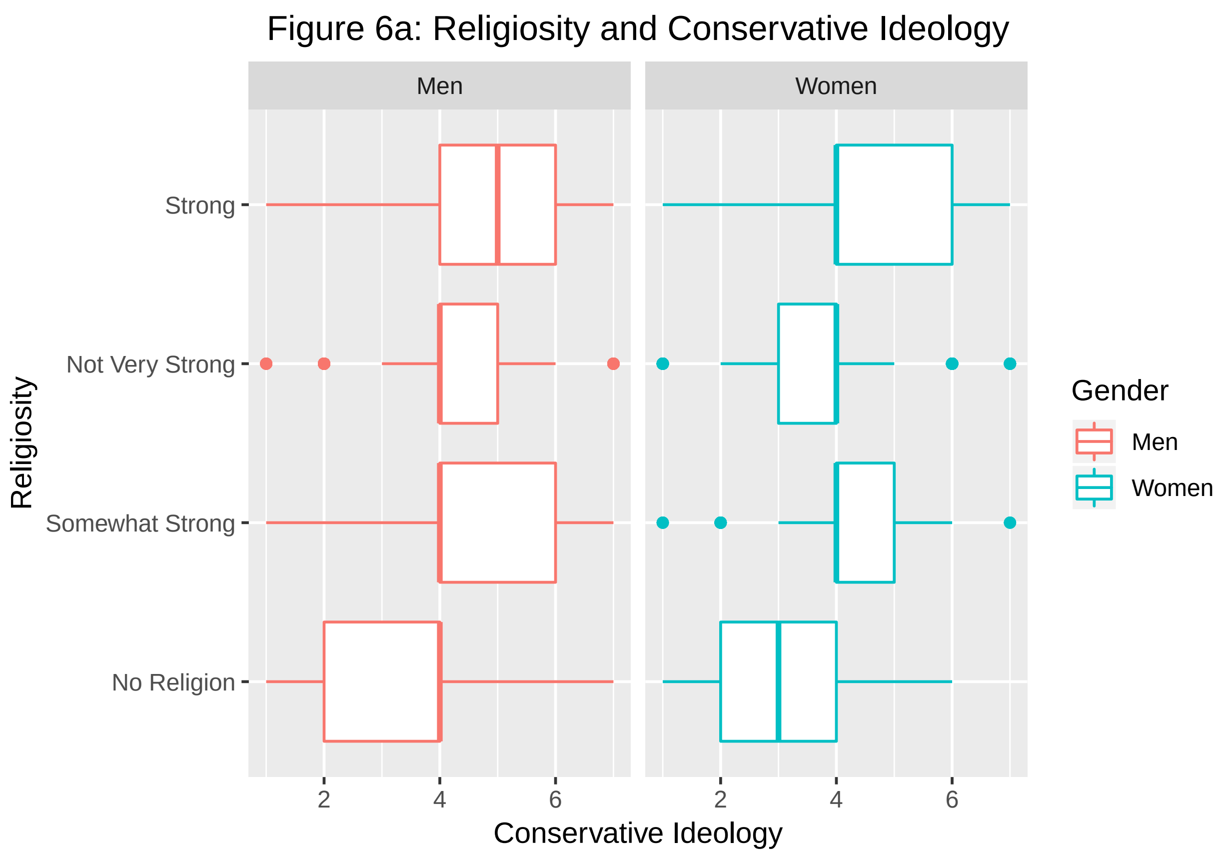
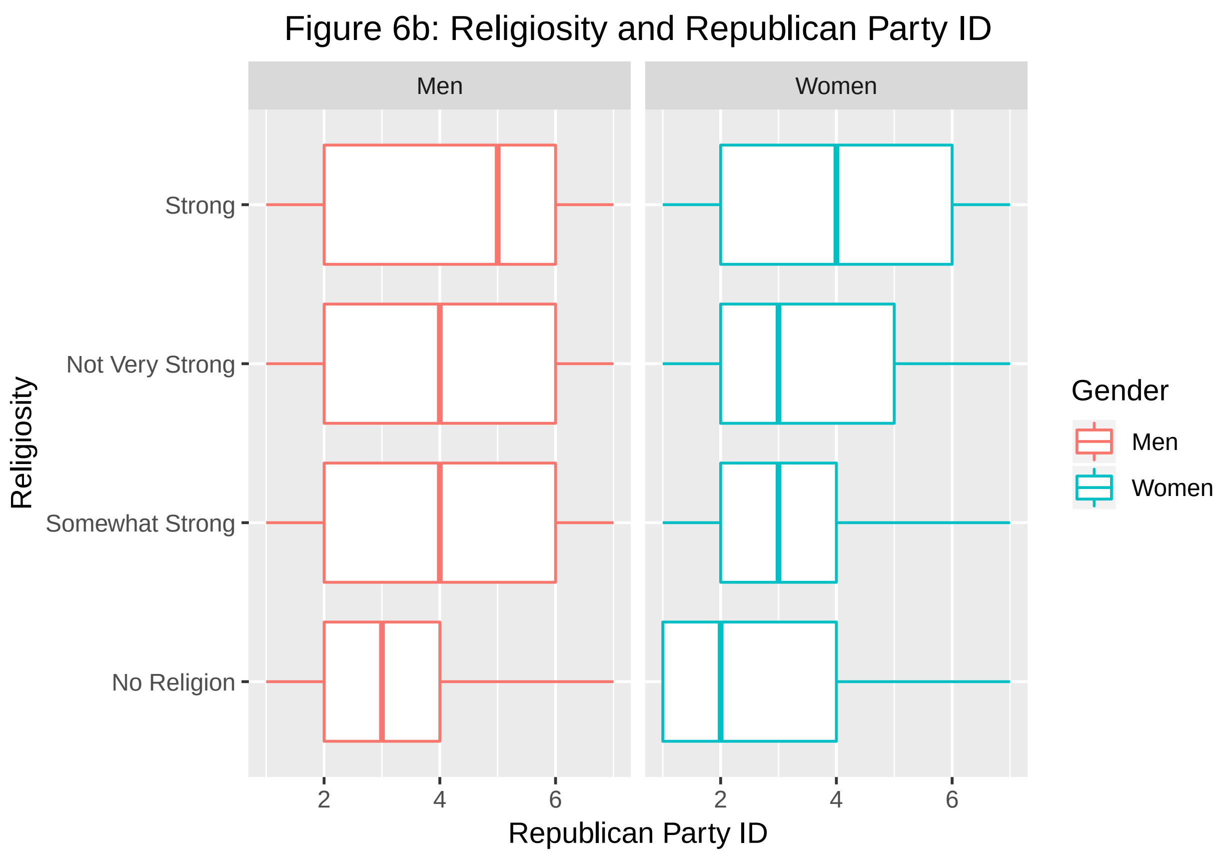


Figure 6a: Religiosity and Conservative Ideology
Men
Women
Strong
Not Very Strong
Somewhat Strong
No Religion
Religiosity
2
4
6
2
4
6
Conservative Ideology
Gender
Men
Women



However, the link between marital status and the DVs appears to be shaped by gender. For men, across all categories of marital status, the mean score is a 4/7 for the Republican party ID scale; however, men who have never been married are much less likely to be Republican identifiers (Figure 7b). This is likely because marital status is somewhat correlated with age, and men who've never been married are also more likely to be in their 20s and early 30s. Women are less likely to be Republican than men across the board, with one exception: for some reason, women who are married have a mean score of 4/7 on the Republican party ID scale (Figure 7b).

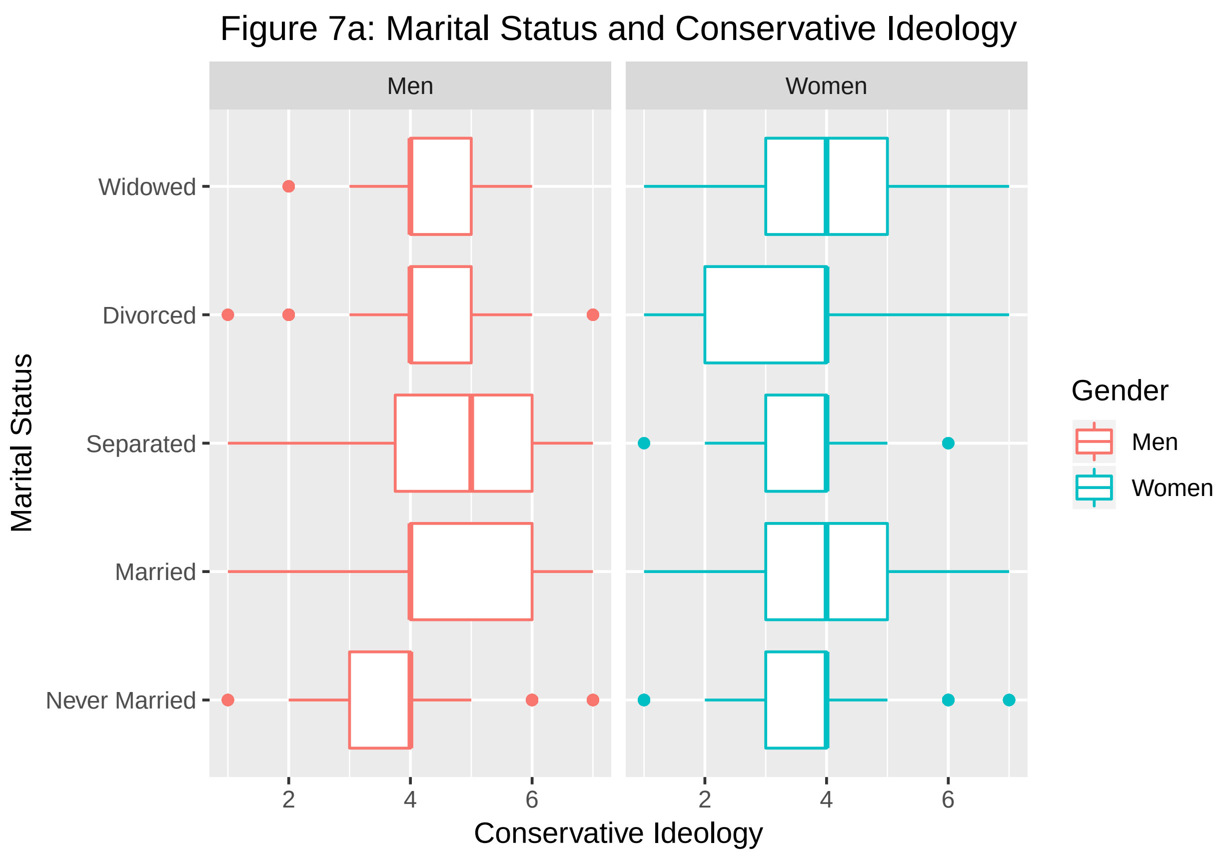
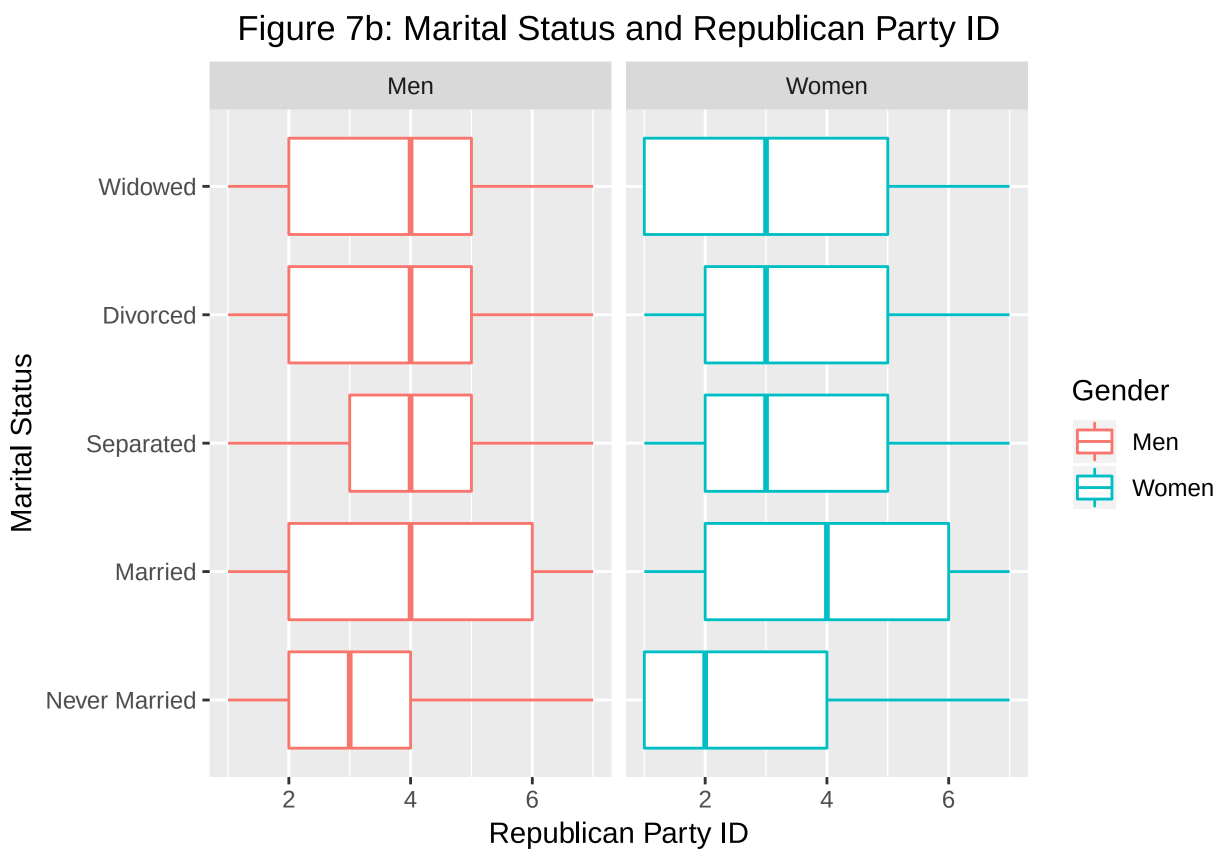
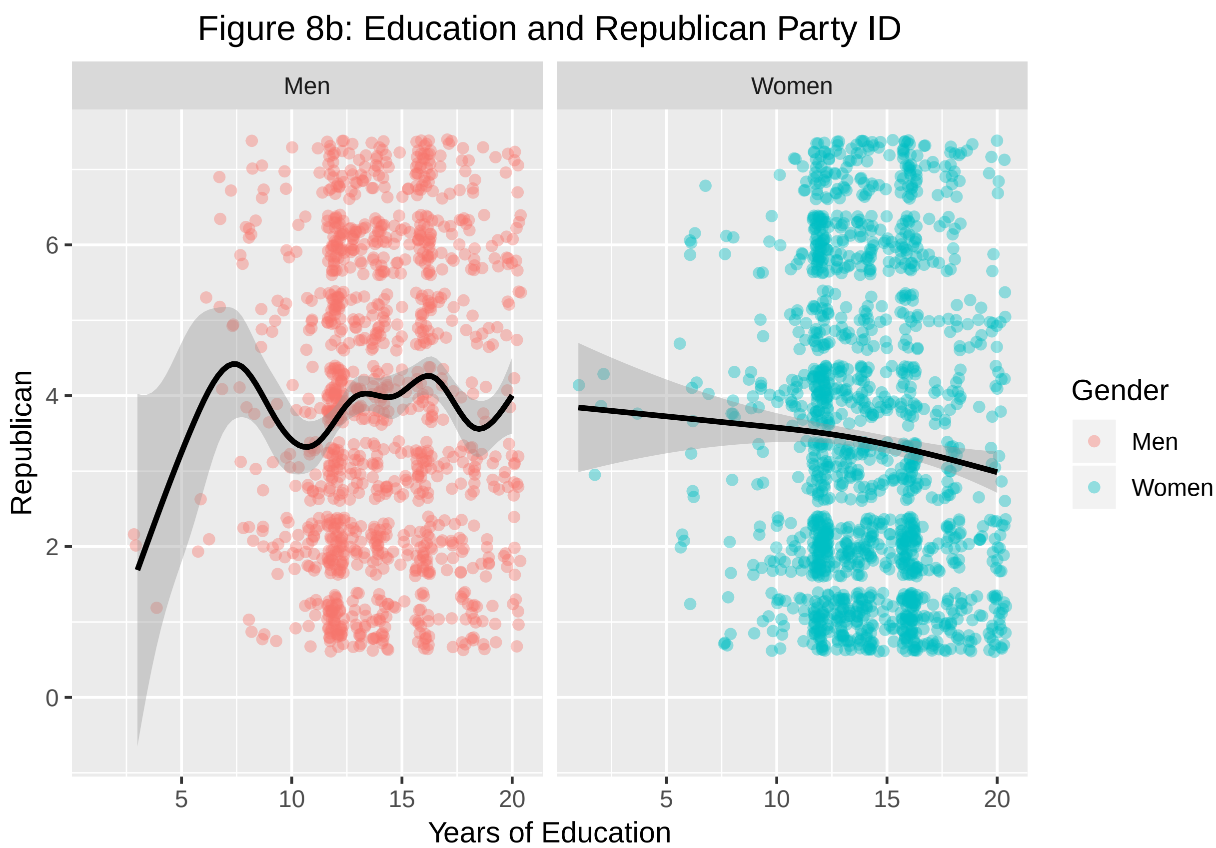


Figure 7a: Marital Status and Conservative Ideology
Men
Women
Widowed
Divorced
Separated
Married
Never Married
Marital Status
2
4
6
Conservative Ideology
Gender
Men
Women



Next, I explore the potential effects of various dimensions or indicators of socioeconomic status: education, family income, and subjective class identity. It seems as though educational attainment is negatively associated with political conservatism and Republican Party ID, although the effect of education appears to be more steep among women (Figure 8a). It's worth noting, of course, that there is a great deal of heterogeneity here: i.e., although those with more education are generally more liberal/Democratic Party-leaning, Figures 8a and 8b clearly show that there are many highly educated individuals who are ideologically conservative and/or strong Republican identifiers as well. Among women, eduational attainment seems to be negatively associated with support for Republicans; among men, the link seems less clear (Figure 8b).

Figure 8a: Education and Conservative Ideology



Among men, family wealthy is positively associated with both political conservatism and support for Republican Party membership; among women, family income has a positive linear association with Republican Party ID, but a nonlinear relationship with political conservatism (Figures 9a and 9b). This suggests that the lack of alignment between political ideology and party ID may be more common among women than among men. I investigate this issue further using FE linear models in a subsequent section.

The link between subjective class identity and the DVs is very interesting. Class identity appears to have a weak negative association with political conservatism and a nonlinear relationship with Republican Party ID (Figures 10a and 10b). That is, on average, respondents who self-identify as working class are more somewhat more conservative than those who self-identify as upper class. While initially surprising, it might be because Americans who are highly class conscious are also more likely to be ideologically liberal. That is, upper-middle class conservatives are more likely to self-identify as “middle class” than their similarly situated counterparts who are more liberal. This working hypothesis is also supported by Figures 9a and 9b: family income generally has a positive association with political conservatism and support for the Republican Party.

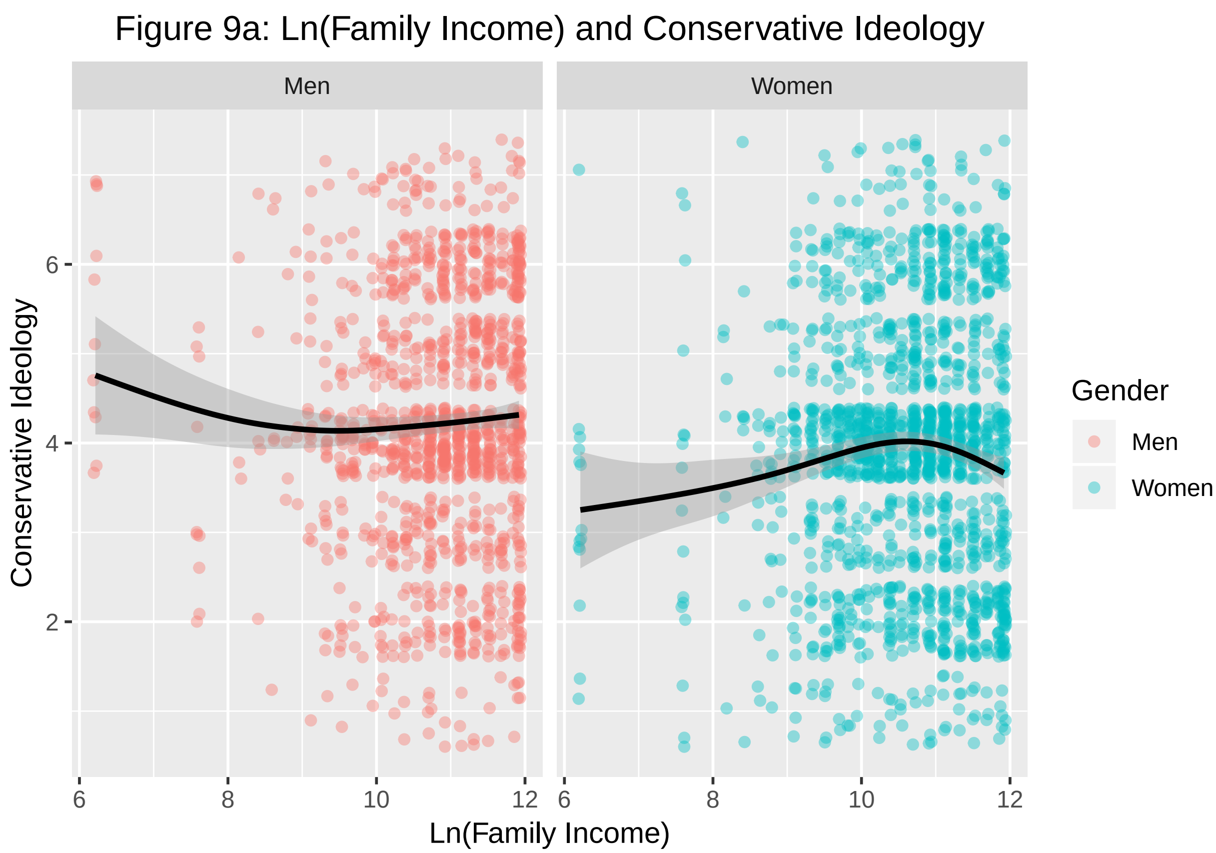


Figure 9a: Ln(Family Income) and Conservative Ideology
Men
Women
Conservative Ideology
6
4
2
6
8
10
12
6
8
10
12
Ln(Family Income)
Gender
Men
Women

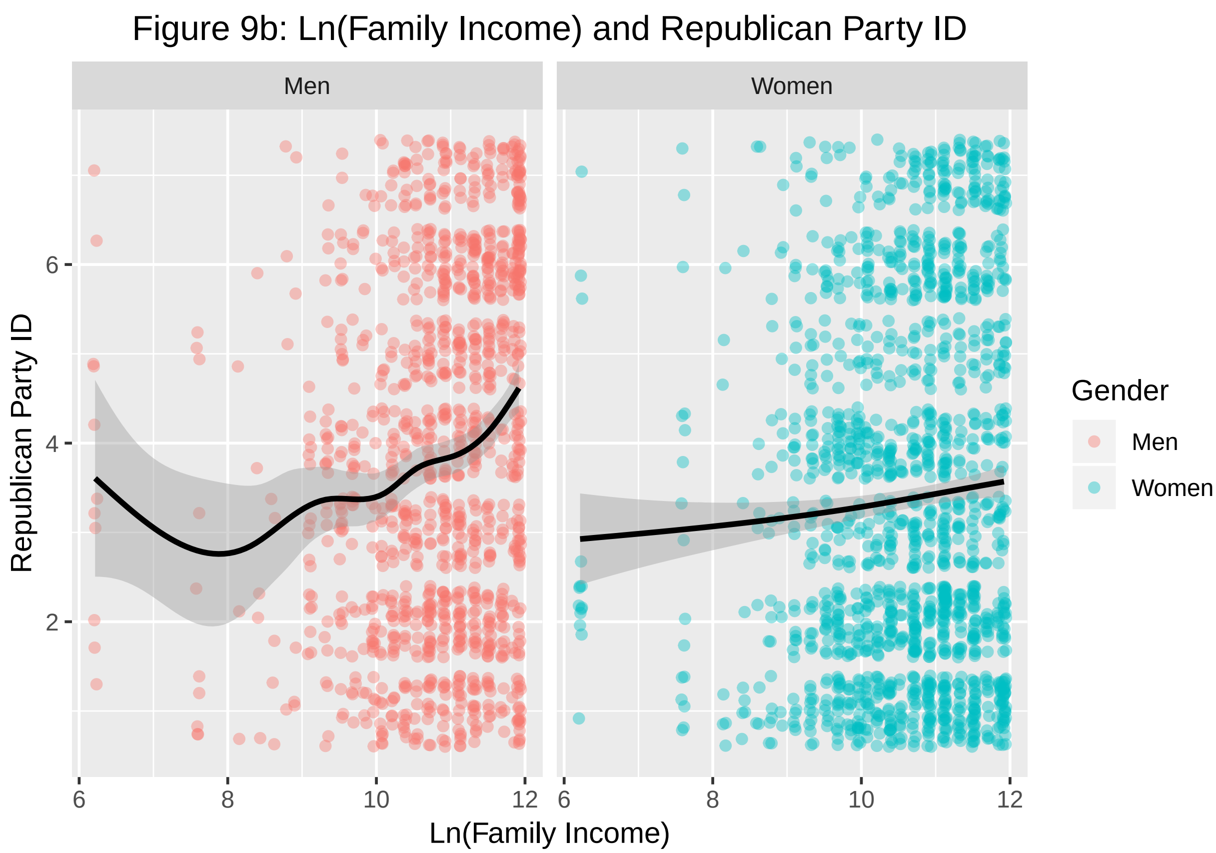


Figure 9b: Ln(Family Income) and Republican Party ID
Men
Women
Republican Party ID
6
4
2
6
8
10
12
Ln(Family Income)
Gender
Men
Women

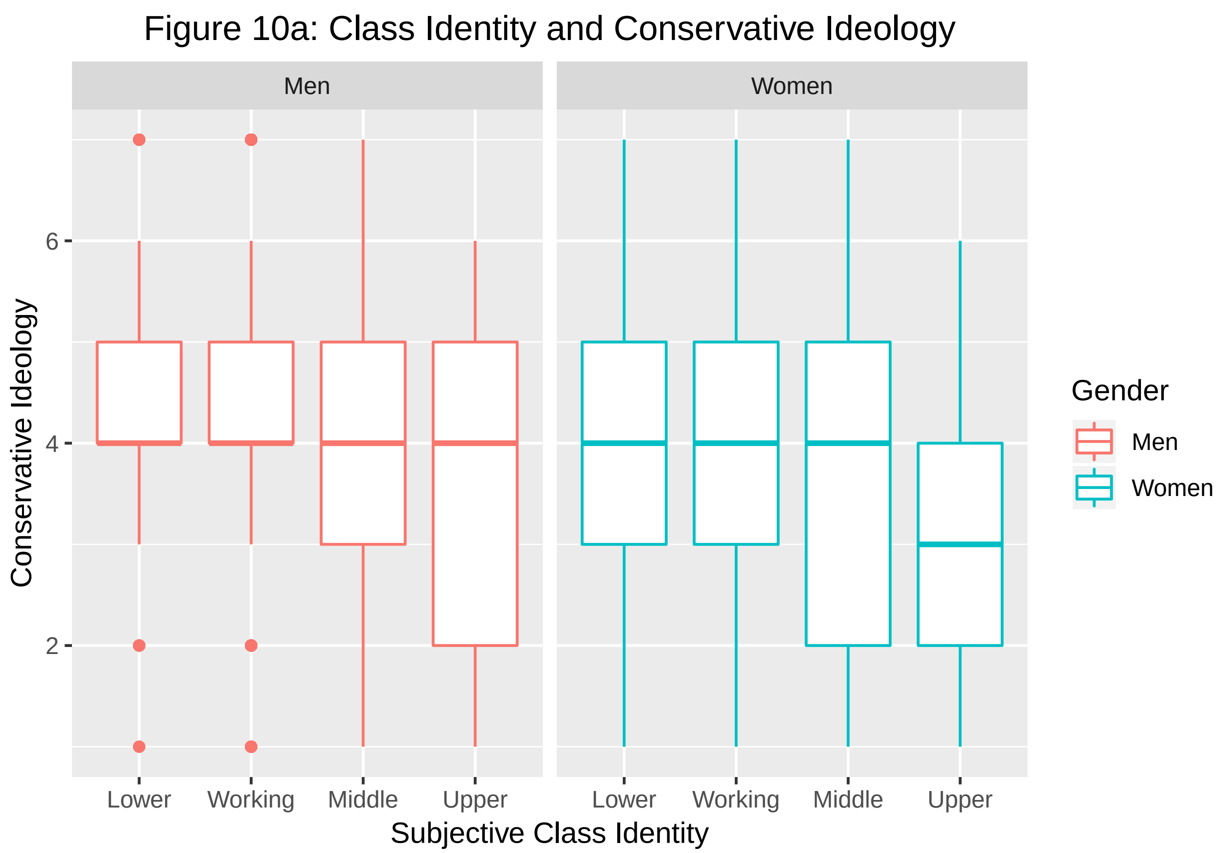
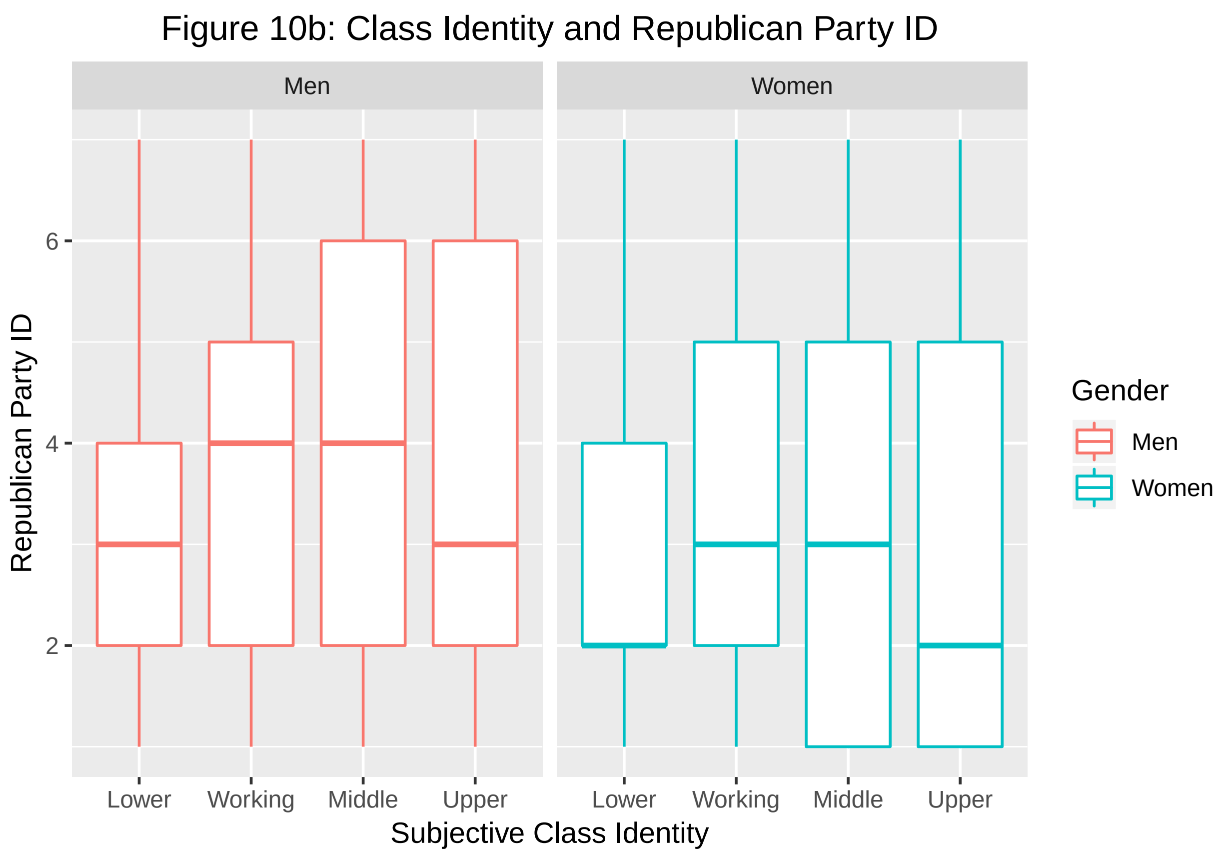


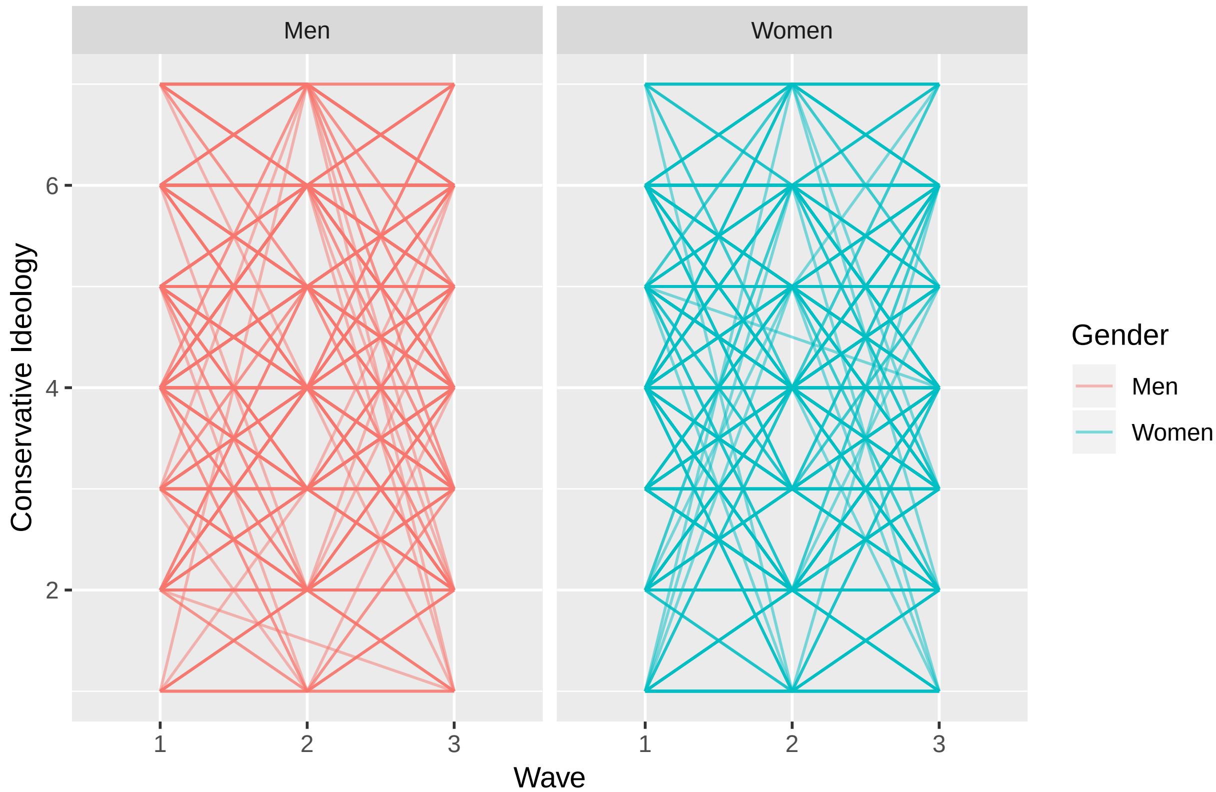
Figure 10a: Class Identity and Conservative Ideology
Men
Women
Conservative Ideology
6
4
2
Lower
Working
Middle
Upper
Lower
Working
Middle
Upper
Subjective Class Identity
Gender
Men
Women



Part 3a: Exploring Changes Over Time in Political Conservatism (DV 1)

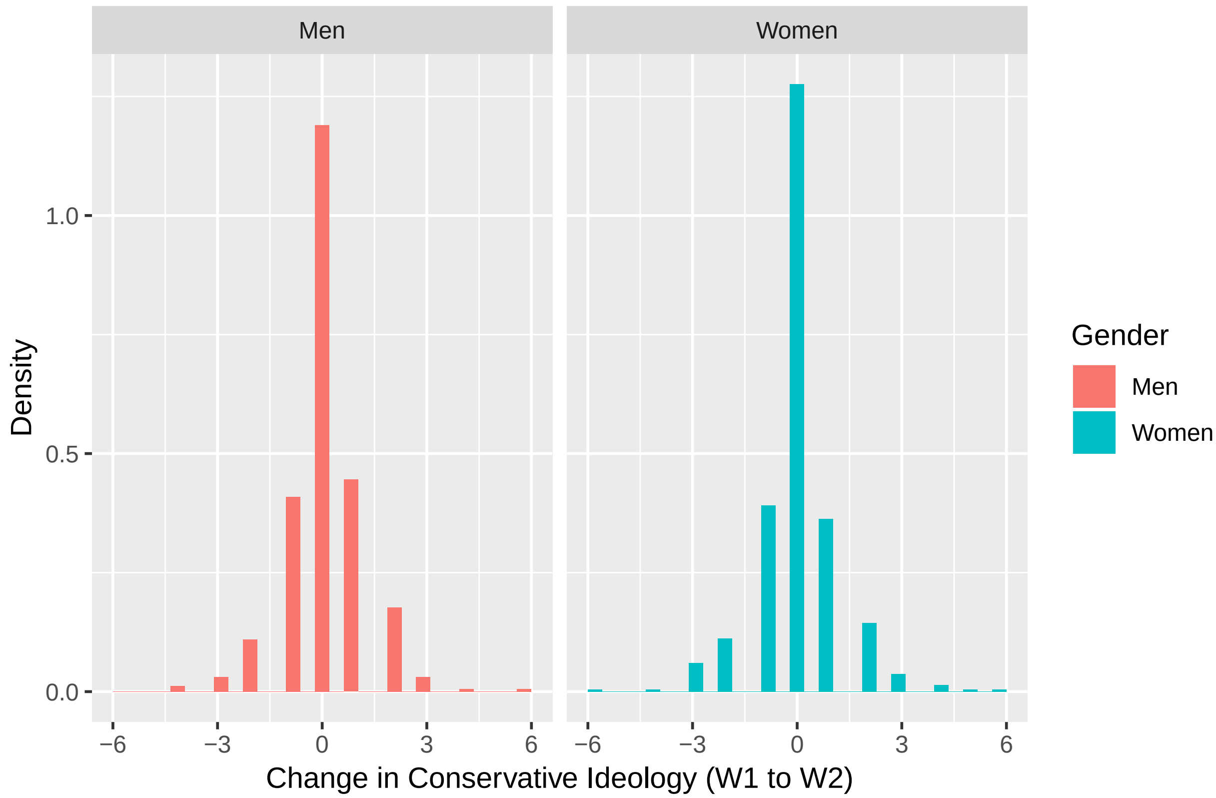
In this section, I look at how the values of the two dependent variables change over time. This is important because if there is not enough within-unit variation in the values of the DVs across the waves, then the FE linear models won't be able to detect the effects of the predictors – even if they really do exist. In Figure 11 below, I display changes in the values of conservative ideology over time. Please note: each line represents the trajectory or trend for each subject. The figure clearly shows that movement is taking place for this variable, and that it's happening in many different directions among both men and women. However, the weakness of this plot is that it doesn't do a good job of showing more than that. For example, it doesn't tell us the relative proportion of subjects who became more/less conservative – and by how much.

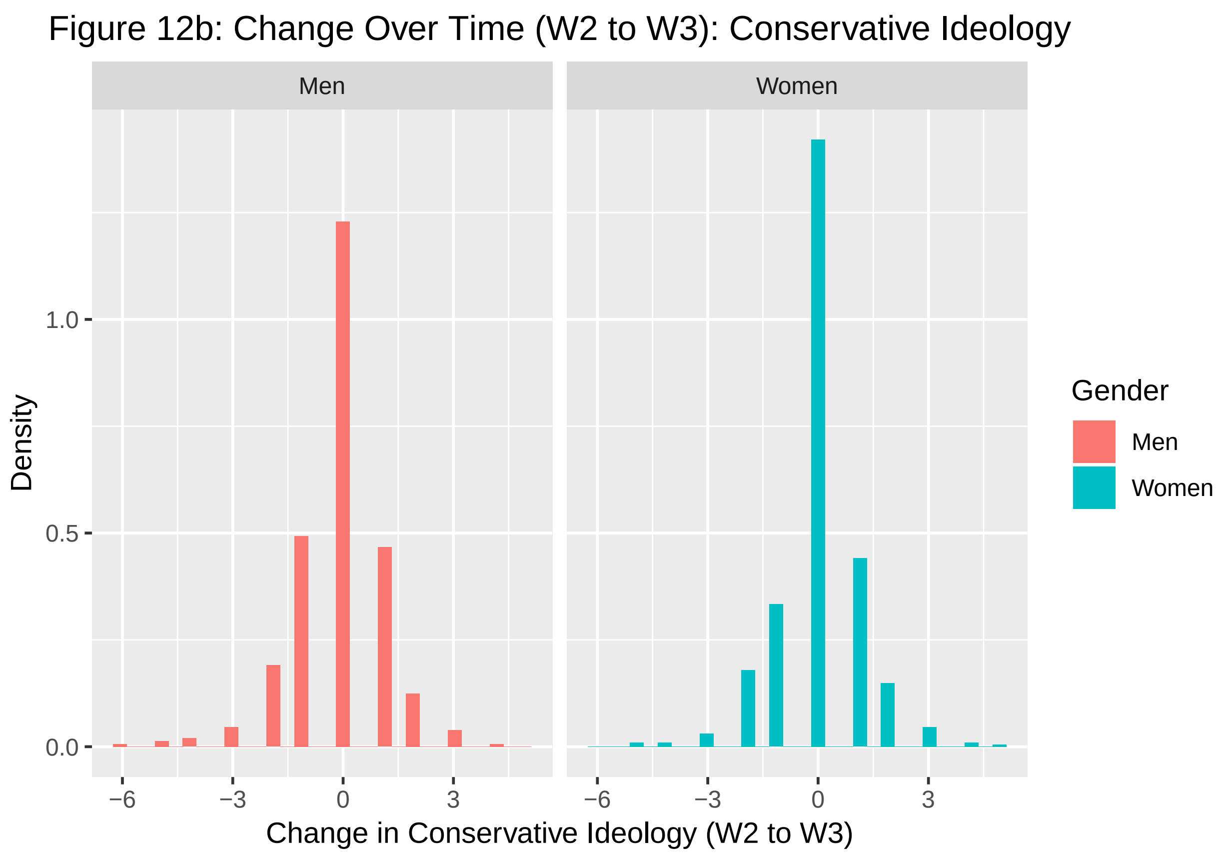
Figure 11: Change Over Time: Conservative Ideology



To address this problem, I wrote some code that computes the changes (or lack the lack thereof) in the values of conservative ideology for each subject across the three waves. Then, I present the data in two sets of histograms. The first set (Figure 12a) provides a distribution of the changes in conservative ideology between waves 1 and 2 for men and women; and the second set (Figure 12b) provides a similar distribution for changes between waves 2 and 3. The figures are insightful: among men and women, about half of the subjects did not change their views between the waves (there are two years in between each wave). Moreover, the changes are normally distributed, which suggests that people were just as likely to become more conservative as they were to become more liberal.

Figure 12a: Change Over Time (W1 to W2): Conservative Ideology





Part 3b: Exploring Changes Over Time in Republican Party ID (DV 2)

Next, I also examine changes in Republican Party ID within the subjects across the three waves. The results as shown in Figures 13a and 13b largely mirror those of the first DV: i.e., between waves 1 and 2 and waves 2 and 3, about half of the subjects changed their level of support for the Republican Party. The changes appear to be roughly normally distributed and there does not appear to be any strong gender differences.

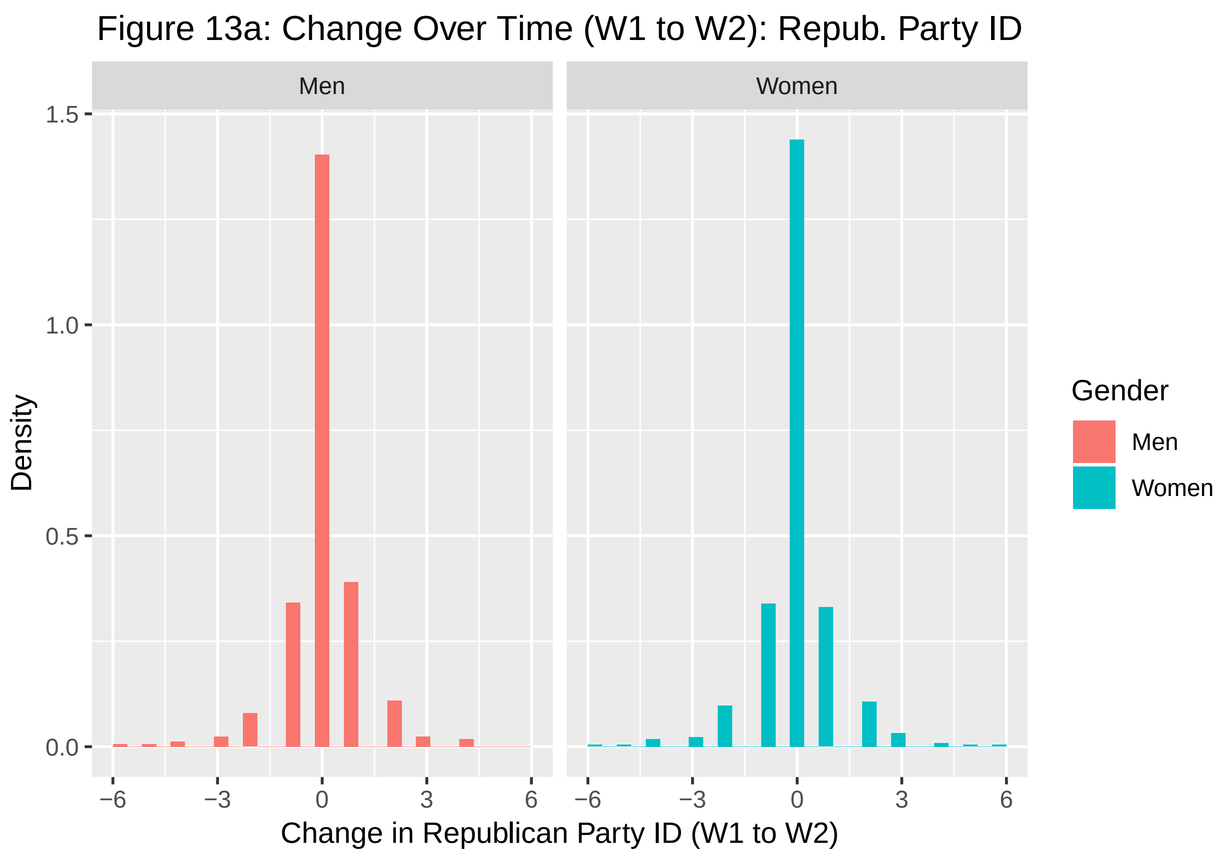
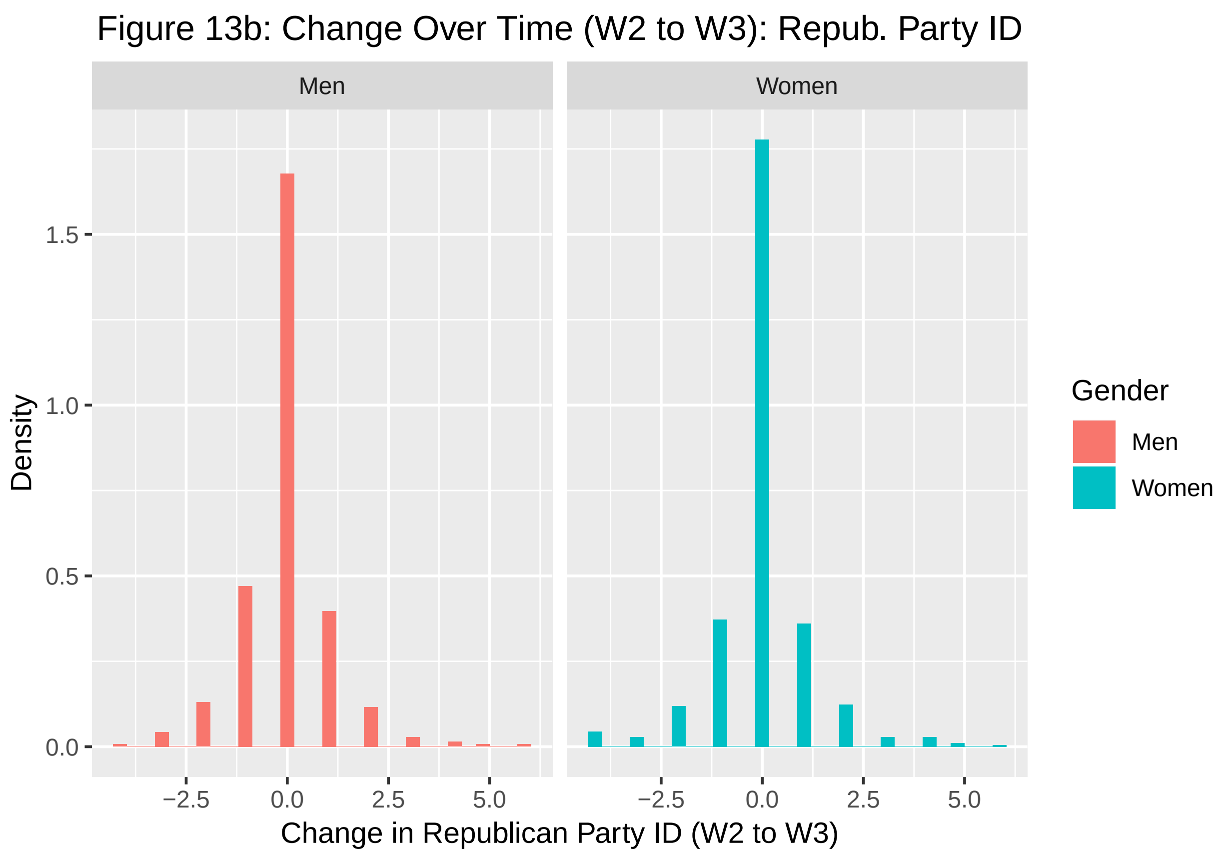


Figure 13a: Change Over Time (W1 to W2): Repub. Party ID
Men
Women
1.5
1.0
0.5
0.0
Density
−6
−3
0
3
6
Change in Republican Party ID (W1 to W2)
Gender
Men
Women



Part 4a: Modeling Changes in Conservative Ideology Over Time (DV 1)

In Part 4, I model changes in the two DV over time by using linear FE models. I use the `plm()` function implemented in the package `plm`. I run three sets of models for each DV. The model specifications are the same across all three models. The main difference is that model 1 uses the full dataset, model 2 only uses the sample of men, and model 3 only uses the sample of women. This is to see whether the link between the predictors and the DVs are shaped by gender. In theory, of course I could have included gender directly in the models and allowed it to interact with the other variables. I do not do this for two reasons: first, as mentioned earlier, the FE models automatically drop time-invariant variables such as gender (in the GSS dataset, gender is treated as time-invariant); second, it is computationally inefficient to allow gender to interact with many different predictors. The results of such a model (i.e., its coefficients) would be very difficult to interpret.

For each model, I use the following predictors: divorced (1 for yes, 0 otherwise); no religion (1 for yes, 0 otherwise); years of education (continuous variable); logged family income (continous variable); lower class ID (1 if either "Working" or "Lower", 0 otherwise); and unemployed (1 for yes, 0 otherwise).

According to the results of the three FE models below, none of the predictors cleared the conventional threshold for statistical significance ($p<.05$). This may indicate that none of the regressors are reliable predictors of political conservatism; it could also indicate that there wasn't enough within-unit variance (or statistical power) to detect real relationships that exist. The second point is plausible, given that we're looking at changes within individuals over only three waves or periods of data collection.

```
# conserv_ideol mod1: full dataset
conserv_mod1 <-plm(conserv_ideol ~ divorced + no_religion + educ + ln_famincome + lower_classid + unempl
conserv_mod1 %>%
  summary()
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = conserv_ideol ~ divorced + no_religion + educ +
##     ln_famincome + lower_classid + unempl + wave, data = complete_gss_df,
##     model = "within", index = c("subject.ID", "wave"))
##
## Balanced Panel: n = 915, T = 3, N = 2745
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -3.3522837 -0.3319821  0.0026084  0.3415866  4.0331956
##
## Coefficients:
##                 Estimate Std. Error t-value Pr(>|t|)
## divorced      -0.0899119  0.1305673 -0.6886  0.49115
## no_religion   -0.0688280  0.0911777 -0.7549  0.45042
## educ          -0.0179559  0.0219321 -0.8187  0.41306
## ln_famincome   0.0122280  0.0377840  0.3236  0.74626
## lower_classid -0.1041776  0.0632004 -1.6484  0.09945 .
## unempl        -0.2311679  0.1656882 -1.3952  0.16313
## wave2          0.0395677  0.0396437  0.9981  0.31837
## wave3         -0.0025152  0.0398366 -0.0631  0.94966
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1302.7
## Residual Sum of Squares: 1297
## R-Squared:      0.0043844
## Adj. R-Squared: -0.49943
## F-statistic: 1.00294 on 8 and 1822 DF, p-value: 0.43161
```

```r
# conserv_ideol mod2: only men
conserv_mod2 <-plm(conserv_ideol ~ divorced + no_religion + educ + ln_famincome + lower_classid + unempl
conserv_mod2 %>%
  summary()
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = conserv_ideol ~ divorced + no_religion + educ +
##     ln_famincome + lower_classid + unempl + wave, data = only_men,
##     model = "within", index = c("subject.ID", "wave"))
##
## Unbalanced Panel: n = 406, T = 1-3, N = 1192
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -3.29758 -0.33570  0.00000  0.34972  3.92756
##
## Coefficients:
##                 Estimate Std. Error t-value Pr(>|t|)
## divorced      -0.2102552  0.1892018 -1.1113   0.2668
## no_religion    0.1074434  0.1400004  0.7675   0.4430
## educ          -0.0096338  0.0333403 -0.2890   0.7727
```

```
## ln_famincome  -0.0157575  0.0602618 -0.2615   0.7938
## lower_classid -0.0848596  0.1011887 -0.8386   0.4019
## unempl        -0.0814153  0.2801202 -0.2906   0.7714
## wave2          0.0818222  0.0613290  1.3342   0.1825
## wave3         -0.0569090  0.0615599 -0.9244   0.3555
##
## Total Sum of Squares:    571.67
## Residual Sum of Squares: 566.14
## R-Squared:      0.0096737
## Adj. R-Squared: -0.51604
## F-statistic: 0.949956 on 8 and 778 DF, p-value: 0.47432
```

```
# conserv_ideol mod3: only women
conserv_mod3 <-plm(conserv_ideol ~ divorced + no_religion + educ + ln_famincome + lower_classid + unempl
conserv_mod3 %>%
  summary()
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = conserv_ideol ~ divorced + no_religion + educ +
##     ln_famincome + lower_classid + unempl + wave, data = only_women,
##     model = "within", index = c("subject.ID", "wave"))
##
## Unbalanced Panel: n = 526, T = 1-3, N = 1553
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -3.1058408 -0.3331765 -0.0025549  0.3123981  4.0454141
##
## Coefficients:
##                Estimate Std. Error t-value Pr(>|t|)
## divorced       0.024185   0.182466  0.1325   0.8946
## no_religion   -0.190336   0.121625 -1.5649   0.1179
## educ          -0.026716   0.029828 -0.8957   0.3706
## ln_famincome   0.033926   0.049050  0.6917   0.4893
## lower_classid -0.126847   0.082392 -1.5396   0.1240
## unempl        -0.310477   0.211678 -1.4667   0.1428
## wave2          0.015390   0.052570  0.2927   0.7698
## wave3          0.039763   0.053130  0.7484   0.4544
##
## Total Sum of Squares:    724.67
## Residual Sum of Squares: 718.5
## R-Squared:      0.0085079
## Adj. R-Squared: -0.5101
## F-statistic: 1.093 on 8 and 1019 DF, p-value: 0.36538
```

Part 4b: Modeling Changes in Republican Party ID Over Time (DV 2)

Next, I repeat the FE analysis for the second DV: Republican Party ID. Again, the model 1 uses the full sample; model 2 only uses the sample of men; and model 3 only uses the sample of women. According to model 1, conservative ideology and no religion are statistically significant predictors of stronger Republican Party ID. While the effect of conservative ideology is very significant in a statistical sense (i.e., it's p-value), the effect in a substative sense is actually arguably quite small. That is, a 1-point increase in political conservatism (out of a 7-point scale) is expected to increase an individual's support for the Republican Party by 0.1 points out of a 7-point scale.

Are there gender differences? Models 2 and 3 suggest that there are: for the sample of men, unemployment also has a statistically and substantively significant effect on support for the Republican Party. Moving from non-unemployed (e.g., having a full-time or part-time job) to unemployed is expected to yield a decrease of 0.63 points in the Republican Party ID scale. This is substantively significant when we consider that the average level of support for Republicans among men is 3.9. Among women, conservative ideology is also a statistically significant predictor of support for the Republican Party, although the effect is smaller than among men (coefficients of 0.085 and 0.122, respectively).

Moreover, while having no religion reduces support for the Republican Party among men, it does not have that effect among women. Empirically, we see another reason why it's often important to disaggregate the results by gender: in the full sample, no religion had a positive statistically significant effect on Republican Party ID, but the coefficient or effect was -0.19. However, we see that this marginal effect was driven almost entirely by changes in the sample of men: the coefficient for men is -0.45 (and statistically significant) among men but there is basically a null effect among women.

```
# repub_partyid mod1: full dataset
repub_mod1 <-plm(repub_partyid ~ conserv_ideol + divorced + no_religion + educ + ln_famincome + lower_cl
repub_mod1 %>%
  summary()
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = repub_partyid ~ conserv_ideol + divorced + no_religion +
##     educ + ln_famincome + lower_classid + unempl + wave, data = complete_gss_df,
##     model = "within", index = c("subject.ID", "wave"))
##
## Balanced Panel: n = 915, T = 3, N = 2745
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -3.7744e+00 -2.7621e-01  8.2054e-05  2.7527e-01  4.0205e+00
##
## Coefficients:
##                 Estimate Std. Error t-value Pr(>|t|)
## conserv_ideol  0.1003538  0.0224618  4.4678 8.39e-06 ***
## divorced       0.0024809  0.1252015  0.0198  0.98419
## no_religion   -0.1941466  0.0874329 -2.2205  0.02651 *
## educ          -0.0026391  0.0210319 -0.1255  0.90016
## ln_famincome   0.0430457  0.0362276  1.1882  0.23491
## lower_classid  0.0067570  0.0606404  0.1114  0.91129
## unempl        -0.0184052  0.1589433 -0.1158  0.90783
## wave2          0.0056774  0.0380199  0.1493  0.88131
## wave3          0.0117784  0.0381945  0.3084  0.75783
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1209.3
## Residual Sum of Squares: 1191.6
## R-Squared:      0.014679
## Adj. R-Squared: -0.48475
## F-statistic: 3.0143 on 9 and 1821 DF, p-value: 0.0014093
```

```r
# repub_partyid mod2: only men
repub_mod2 <-plm(repub_partyid ~ conserv_ideol + divorced + no_religion + educ + ln_famincome + lower_cl
repub_mod2 %>%
  summary()
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = repub_partyid ~ conserv_ideol + divorced + no_religion +
##     educ + ln_famincome + lower_classid + unempl + wave, data = only_men,
##     model = "within", index = c("subject.ID", "wave"))
##
## Unbalanced Panel: n = 406, T = 1-3, N = 1192
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -2.76843 -0.26043  0.00000  0.29477  3.93358
##
## Coefficients:
##                Estimate Std. Error t-value  Pr(>|t|)
## conserv_ideol  0.122691   0.032560  3.7681 0.0001769 ***
## divorced      -0.015712   0.171969 -0.0914 0.9272237
## no_religion   -0.445241   0.127196 -3.5004 0.0004910 ***
## educ           0.040722   0.030281  1.3448 0.1790853
## ln_famincome   0.021978   0.054732  0.4016 0.6881239
## lower_classid -0.107829   0.091941 -1.1728 0.2412331
## unempl        -0.630777   0.254419 -2.4793 0.0133755 *
## wave2          0.012476   0.055763  0.2237 0.8230283
## wave3          0.024072   0.055939  0.4303 0.6670748
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    487.67
## Residual Sum of Squares: 466.36
## R-Squared:      0.043684
## Adj. R-Squared: -0.46586
## F-statistic: 3.9437 on 9 and 777 DF, p-value: 6.3014e-05
```

```r
# repub_partyid mod3: only women
repub_mod3 <-plm(repub_partyid ~ conserv_ideol + divorced + no_religion + educ + ln_famincome + lower_cl
repub_mod3 %>%
  summary()
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = repub_partyid ~ conserv_ideol + divorced + no_religion +
##     educ + ln_famincome + lower_classid + unempl + wave, data = only_women,
##     model = "within", index = c("subject.ID", "wave"))
##
## Unbalanced Panel: n = 526, T = 1-3, N = 1553
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -3.7297757 -0.2565799  0.0024784  0.2378935  3.9834741
```

```
##
## Coefficients:
##                 Estimate Std. Error t-value Pr(>|t|)
## conserv_ideol  0.0850234  0.0310719  2.7363 0.006321 **
## divorced       0.0137219  0.1809841  0.0758 0.939579
## no_religion   -0.0150728  0.1207809 -0.1248 0.900711
## educ          -0.0374397  0.0295967 -1.2650 0.206163
## ln_famincome   0.0447545  0.0486624  0.9197 0.357951
## lower_classid  0.1005673  0.0818170  1.2292 0.219291
## unempl         0.3887391  0.2101793  1.8496 0.064667 .
## wave2         -0.0022456  0.0521450 -0.0431 0.965659
## wave3          0.0111038  0.0527122  0.2106 0.833204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    716
## Residual Sum of Squares: 706.17
## R-Squared:      0.013722
## Adj. R-Squared: -0.50364
## F-statistic: 1.57376 on 9 and 1018 DF, p-value: 0.11831
```

Part 5: Assessing Model Fit

In this final section, I want to assess model fit. Since only the results of the models for the 2nd DV were statistically significant, I'll just focus on those models (i.e., which predict Republican Party ID). First, I'll focus on model 1, which uses the full sample. I first save the residuals and plot the distribution with respect to (wrt) to conservative ideology (Figure 14a); I also repeat this analysis but segment by gender (Figure 14b). I plot the residuals in relation to conservative ideology because the latter appears to be the most consistent predictor of Republican Party ID.

It appears as though the residuals are distributed in a roughly random manner. The one exception is that the residuals are greater when conservative ideology is 4/7 points (or "Moderate"). Intuitively, this makes sense. Among ideological moderates, almost by definition we would expect the link between political ideology and partisanship to be weaker. Thus, the residuals will be greater. Overall, there is relatively little evidence of heteroskedasticity, which is good. This seems to be true across for model 1 when it comes to men and women.

Figure 14a: Model 1 Residuals

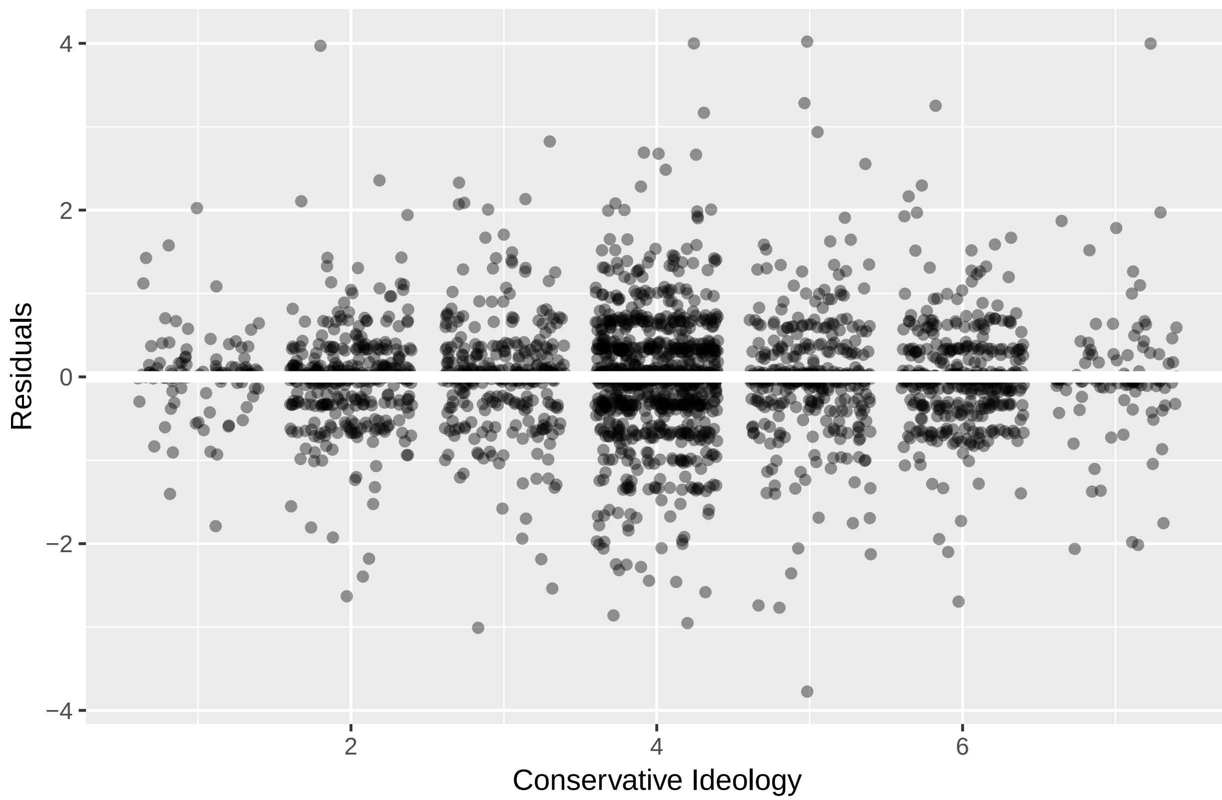
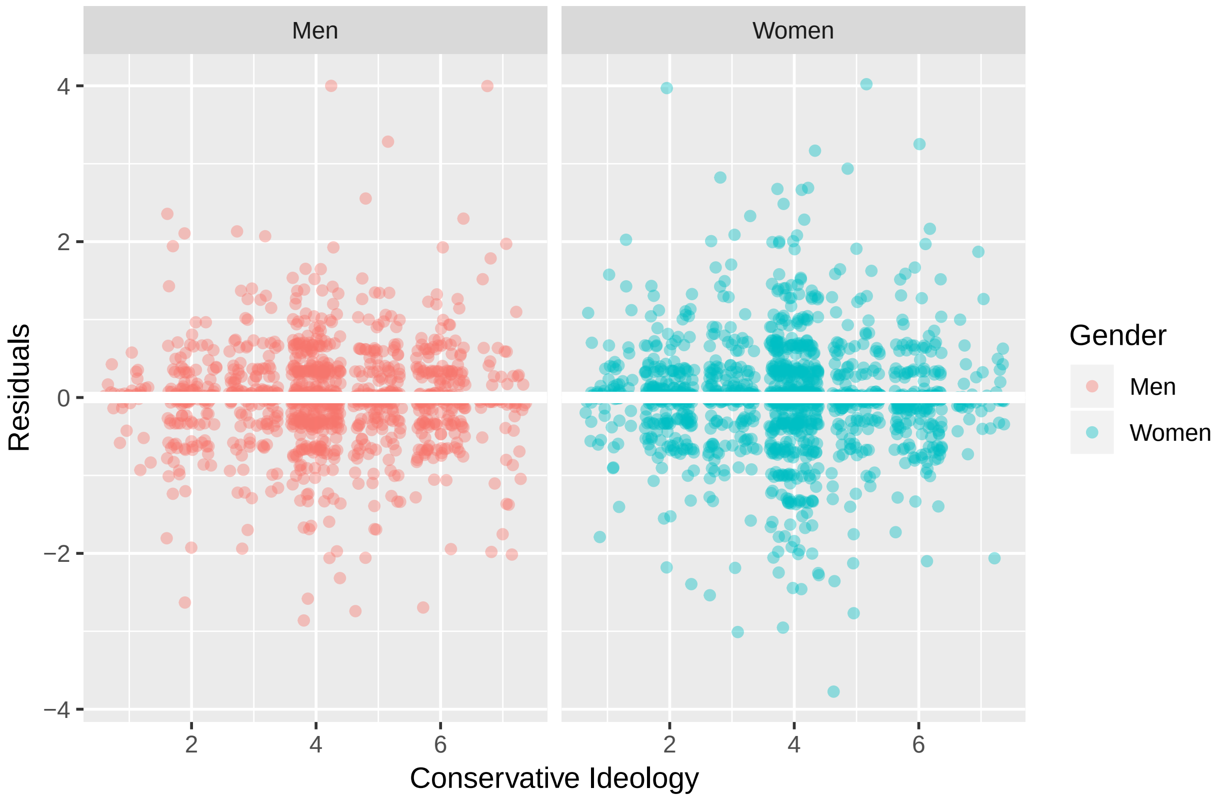


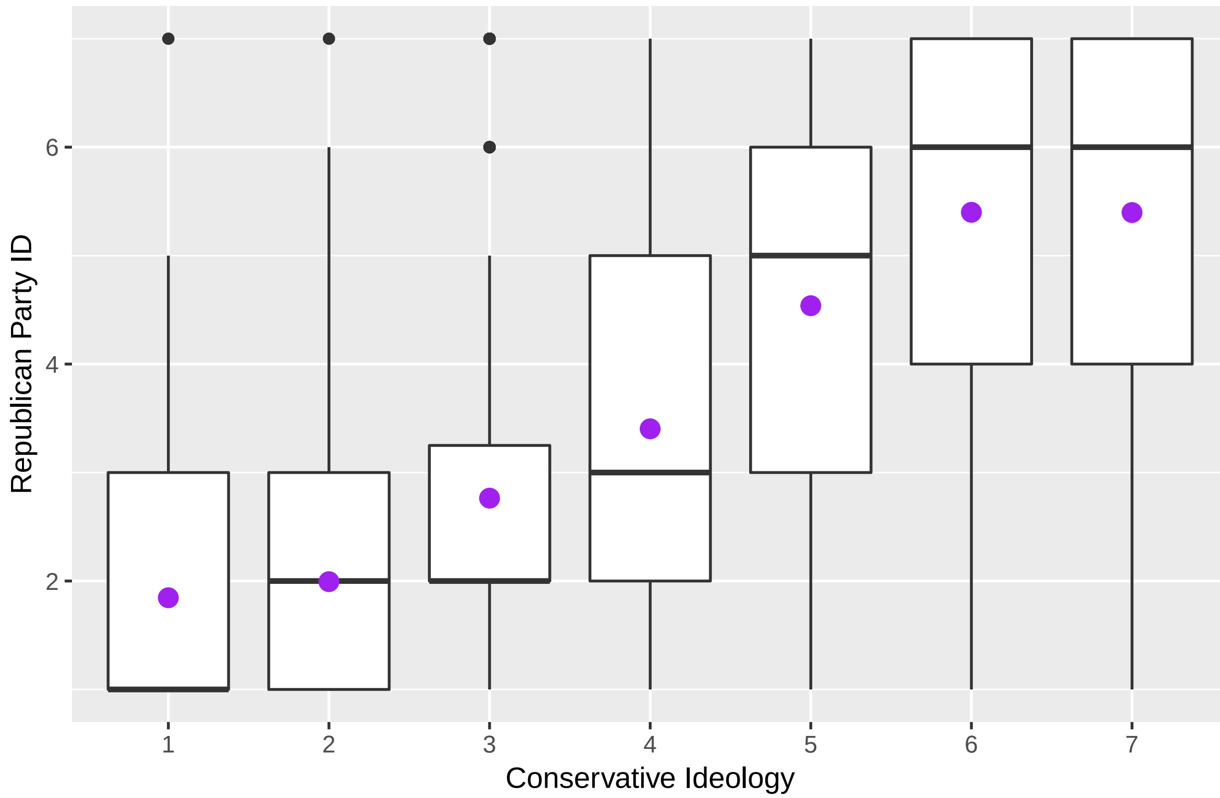
Figure 14b: Model 1 Residuals

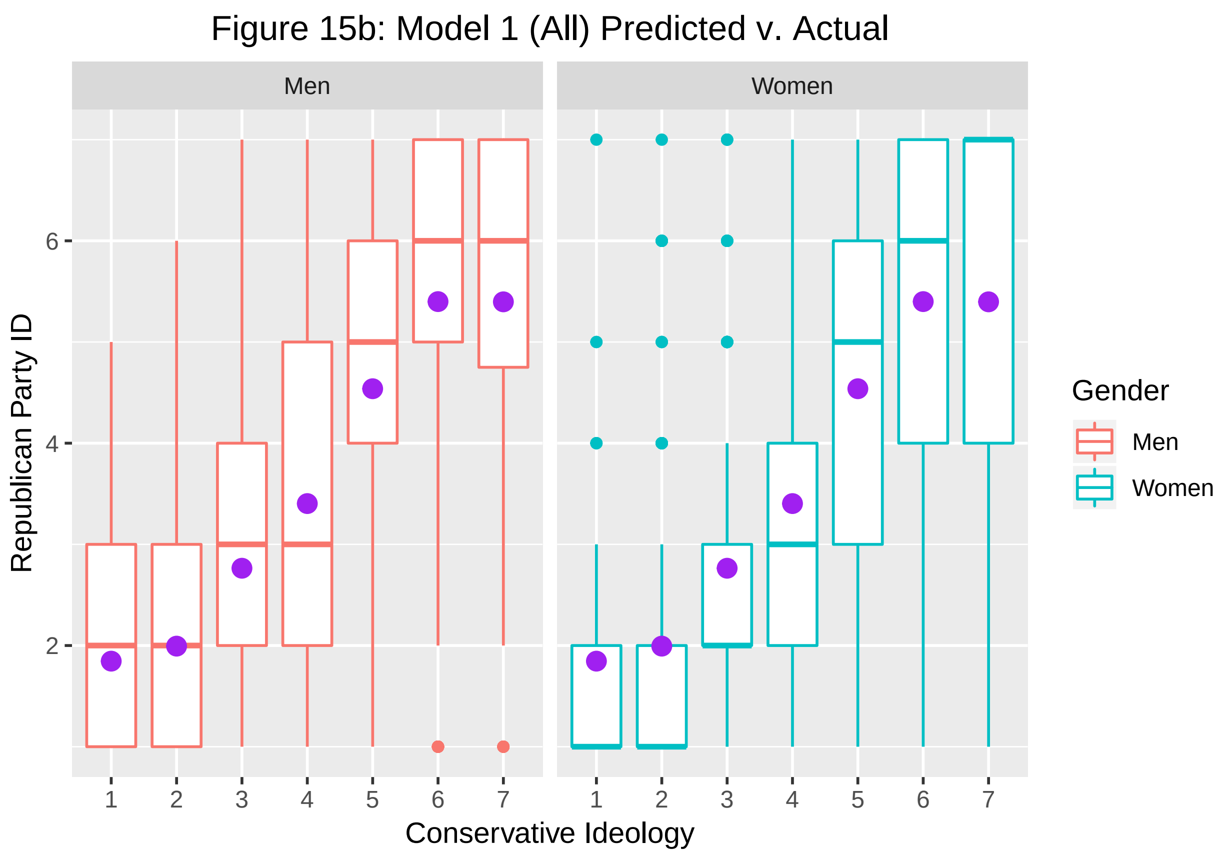


Next, I want to compare the actual values of the DV with the predicted values generated by model 1. I do this in two ways. First, I compare the distribution of actual values for the DV for each value of conservative ideology (i.e., I treat conservative ideology as a factor) – with the values of the DV predicted by model 1. In Figure 15a, the actual values are from the full sample, which were used to generate the predictions.

In Figure 15b, I plot the predicted values of the DV from model 1 but also compare them to actual values from men and women. What we can see is that while model 1 (full sample) did a reasonably good job of using conservative ideology to predict Republican Party ID at the aggregate level, it did not do a good job among women. This suggests that the link between conservative ideology and Republican Party ID is shaped by gender, and researchers should either explicitly account for this in the models (e.g., by including an interaction term) or by running separate regressions on samples of men and women (which is what I've done).

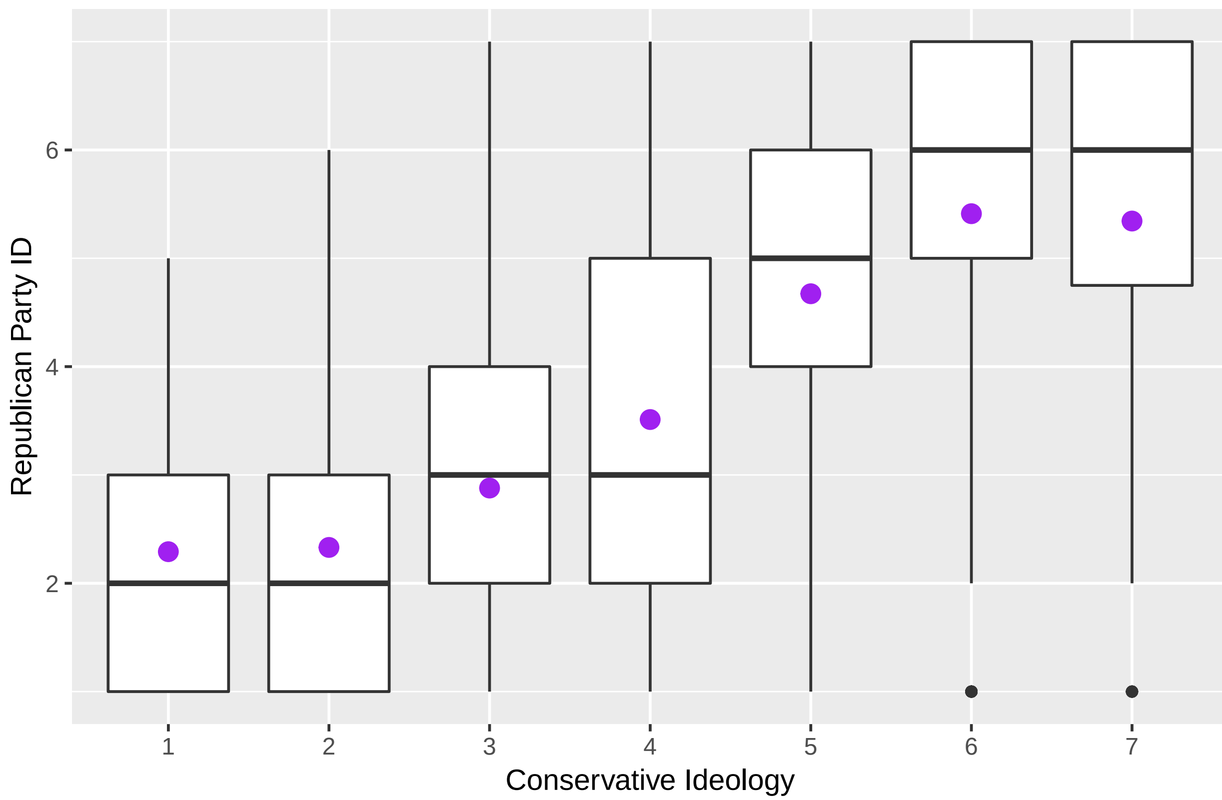
Figure 15a: Model 1 (All) Predicted v. Actual





To demonstrate that my preferred modeling strategy produces better predictions, I will use Figures 16a and 16b below. Figure 16a compares the predicted values of Republican Party ID generated from model 2 (which was only run on the sample of men) and the actual values of Republican Party ID among men. Both the predicted and actual values of the DV are in relation to conservative ideology; again, this is because this is the most consistent predictor of the DV as identified during the previous FE modeling stage. Figure 16b repeats this analysis but using the predicted values of Republican Party ID generated from model 3 (which only used data from the sample of women) and the actual values of of Republican Party ID among women. Please note that the models run on single-gender samples provide more reliable predictions. The improvements are particularly noticeable among women (e.g., compare Figures 15b and Figures 16b).

Figure 16a: Model 2 (Men Only) Predicted v. Actual



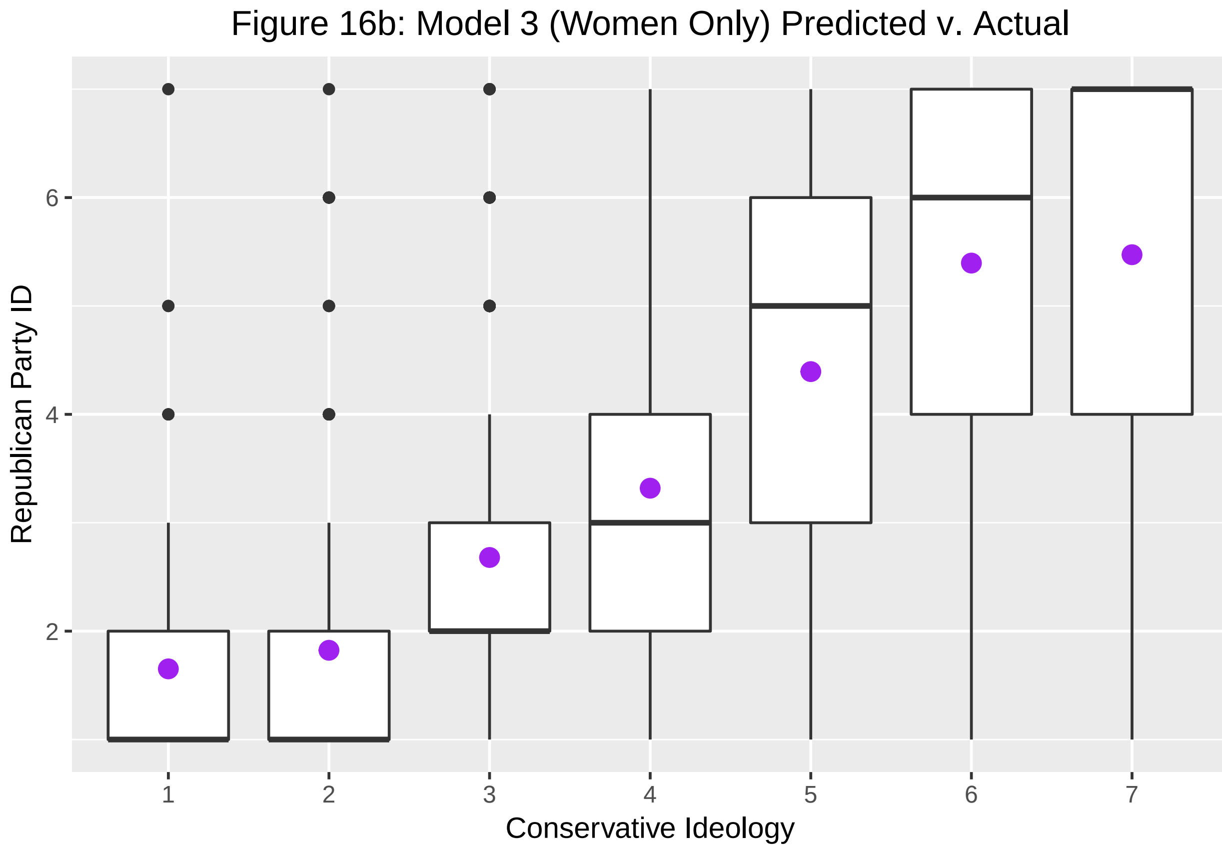


Figure 16b: Model 3 (Women Only) Predicted v. Actual
Republican Party ID
6
4
2
1
2
3
4
5
6
7
Conservative Ideology