exercises_week6

Lindsey Greenhill

3/3/2021

Exercises 5.1, 5.4

Question 1

Part a

I created 4 logistic models of presidential preference vs multiple independent variables.

- mod_1 is a logistic model of bush_vote vs. education level, party identification, ideology, and income.
- mod_2 is a logistic model of bush_vote vs. education level, party identification, ideology, income, race, and gender.
- mod_3 is a logistic model of bush_vote vs. education level, party identification, ideology, income, race, gender, and an interaction between income and gender.
- mod_4 is a logistic model of bush_vote vs. education level, party identification, ideology, race, gender, and an interaction between education and gender.

The table below shows the results of these models.

```
# model 1 is a regression of bush vote vs gender, education, party, ideology,
# race, and income

mod_1 <-
  glm(bush_vote ~ educ1 + partyid3 + ideo7 + income,
      data =  df_clean,
      family = "binomial")

mod_2 <-
  glm(
    bush_vote ~ educ1 + partyid3 + ideo7 + income + race + gender,
    data =  df_clean,
    family = "binomial"
  )

# model 3 is a regression of bush vote vs gender, education, party, ideology,
# race, and an interaction between income and gender

mod_3 <-
  glm(
```

```
    bush_vote ~ gender + educ1 + partyid3 + ideo7 + race + income + income:gender,
    data =  df_clean,
    family = "binomial"
  )

# model 4 is a regression of bush vote vs gender, education, party, ideology,
# race, and an interaction between education and gender. I took out income
# because it wasn't statistically significant in the past 2 models.

mod_4 <-
  glm(
    bush_vote ~ gender + educ1 + partyid3 + ideo7 + race + educ1:gender,
    data =  df_clean,
    family = "binomial"
  )

# regression table

stargazer(mod_1, mod_2, mod_3, mod_4, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Mar 10, 2021 - 00:58:00

Part b: model comparison

Coefficient estimates

- the education variable (educ1) is a 4 point scale of education (going from less education to more education). the education coefficient is positive and relatively similar in all 4 models (.124, .174, .174, .163), however, it is not statistically significant in any model.
- the party identification factor variable (partyid3) is statistically significant. The coefficient for the Independents are positive and similar across all models and is also statistically significant. The coefficient for the Republicans are positive and similar across all models and is also statistically significant. Without interpreting the specific coefficients, both of these coefficients mean that Independents and Republicans are more likely to vote for bush than Democrats. This effect is much larger than the education effect.
- the ideology variable (ideo7) is a 7 point scale of ideology (from very liberal to very conservative). It is positive and relatively similar across all 4 models and statistically significant. It has less of an effect compared to party identification but more of an effect than education.
- the income variable (income) is a 5 point scale of incomes. the income coefficient is positive but very small in models 1, 2, and 3 and is not statistically significant (which is why I didn't use it in model 4).
- the race factor variable (race) have 5 levels, with the model setting asian as the base level. The variable was only used in models 2,3, and 4. It is negative and statistically significant in these three models, meaning that black respondents were less likely to vote for bush. The hispanic, native_american, and white coefficients were not statistically significant.
- the gender variable (gender) coefficient is used in models 2, 3, and 4. It is positive in all 3 models but only statistically significant in model 2. It's effects are largest in model 3 (but again it is not statistically significant) and smallest in model 4.

Table 1:

	Dependent variable:			
	bush_vote			
	(1)	(2)	(3)	(4)
educ1	0.124	0.174	0.172	0.165
	(0.119)	(0.123)	(0.124)	(0.162)
partyid3I	1.724***	1.856***	1.860***	1.859***
	(0.299)	(0.310)	(0.311)	(0.311)
partyid3R	4.058***	4.060***	4.057***	4.065***
	(0.220)	(0.235)	(0.235)	(0.235)
ideo7	0.462***	0.481***	0.479***	0.482***
	(0.076)	(0.079)	(0.079)	(0.079)
income	0.023	0.013	0.059	
	(0.100)	(0.103)	(0.158)	
gender:income			−0.075	
			(0.194)	
raceblack		−2.098**	−2.127**	−2.104**
		(0.929)	(0.934)	(0.929)
racehispanic		0.562	0.559	0.554
		(0.914)	(0.915)	(0.913)
racenative_american		0.593	0.569	0.599
		(1.017)	(1.020)	(1.017)
racewhite		−0.193	−0.206	−0.193
		(0.820)	(0.823)	(0.819)
gender:educ1				0.028
				(0.223)
gender		0.408*	0.642	0.328
		(0.211)	(0.641)	(0.656)
Constant	−4.688***	−4.846***	−4.963***	−4.784***
	(0.534)	(1.019)	(1.066)	(1.056)
Observations	1,133	1,133	1,133	1,133
Log Likelihood	−354.198	−337.749	−337.674	−337.749
Akaike Inf. Crit.	720.397	697.497	699.348	697.497
Note:				*p<0.1; **p<0.05; ***p<0.01

- the interaction between income and gender (-.075) in model 3 is relatively small and negative, suggesting that an increase iin ncome has a small additional negative effect on the probability of voting for Bush when gender = 1. However, it is not statisticaly significant.
- the interaction between education and gender (.028) in model 4 is relatively small and positive, suggesting that an increase education has a small additional positive effect on the probability of voting for Bush when gender = 1. However, it is not statisticaly significant.

Deviances

The table below shows the null and residual deviances for the four models. If a variable does not add explanatory power to the model, and is random noise, the deviance will decrease by 1 on average. If the variable is a good predictor, we expect the deviance to decrease by more than 1. A lower deviance means that the model is a better fit. We see that the residual deviance for model 1 is significantly higher than the residual deviances of mdoels 2, 3, and 4. This means that adding race and gender to the model added significant explanatory power. However, the residual deviance for model_3 does not decrease by more than 1, meaning that the interaction term between income and gender does is not an informative predictive. Similarly, the residual deviance for model 4 does not change from that of model 2, meaning that the interaction term between gender and education is not an informative predictor and that income is not an informative predictor either (as I took it out in model 4).

```
## # A tibble: 4 x 3
##   Model `Null Dev` `Resid. Dev`
##   <chr>      <dbl>        <dbl>
## 1 mod_1      1534.         708.
## 2 mod_2      1534.         675.
## 3 mod_3      1534.         675.
## 4 mod_4      1534.         675.
```

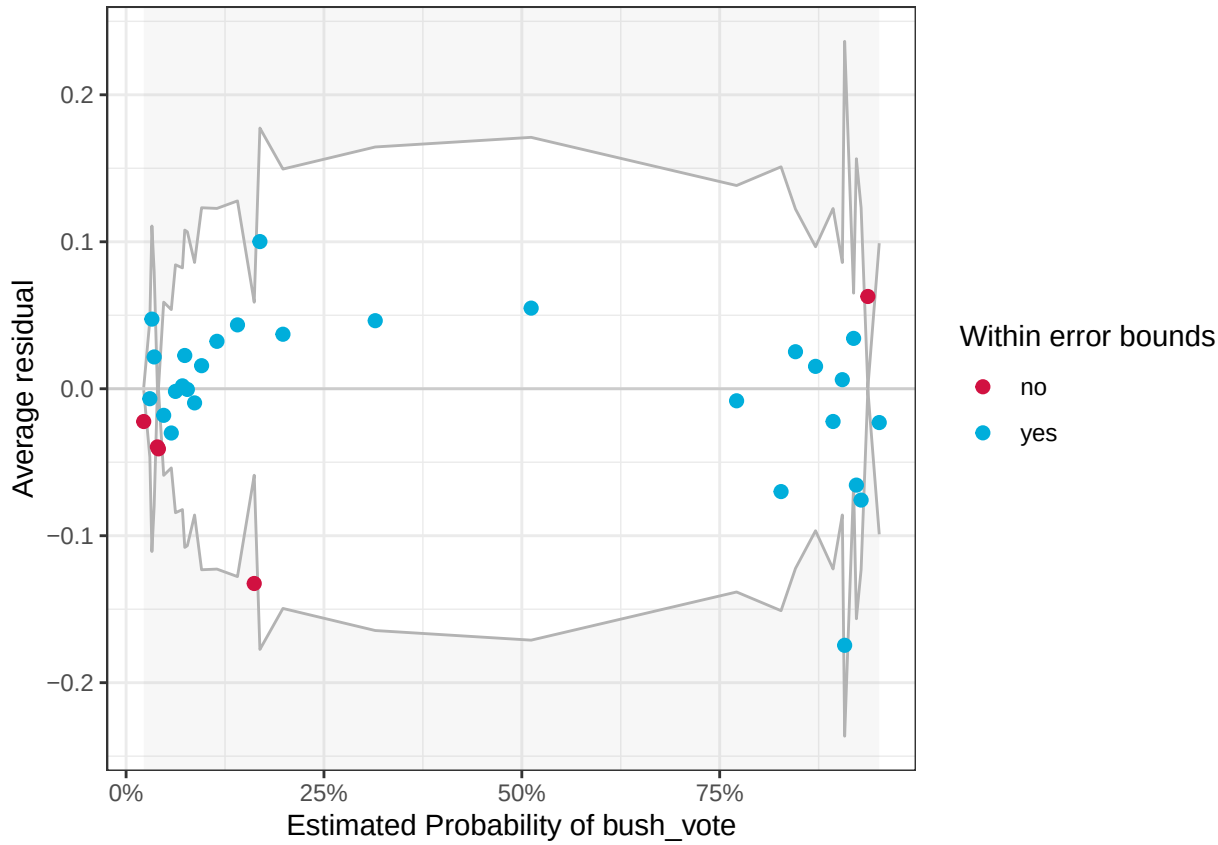
Binned Residual Plots

The binned residual plots divides the data into categories and plots the average residual vs the average fitted value for each category. The dotted lines shows a 2 standard-error bound. If the model is true, we would expect about 95% of the points to fall within the bounds. We would also expect no dramatic patterns to appear in the plot. (from the textbook).

Model 1:

```
binned_residuals(mod_1)
```

```
## Warning: About 85% of the residuals are inside the error bounds (~95% or higher would be good).
```

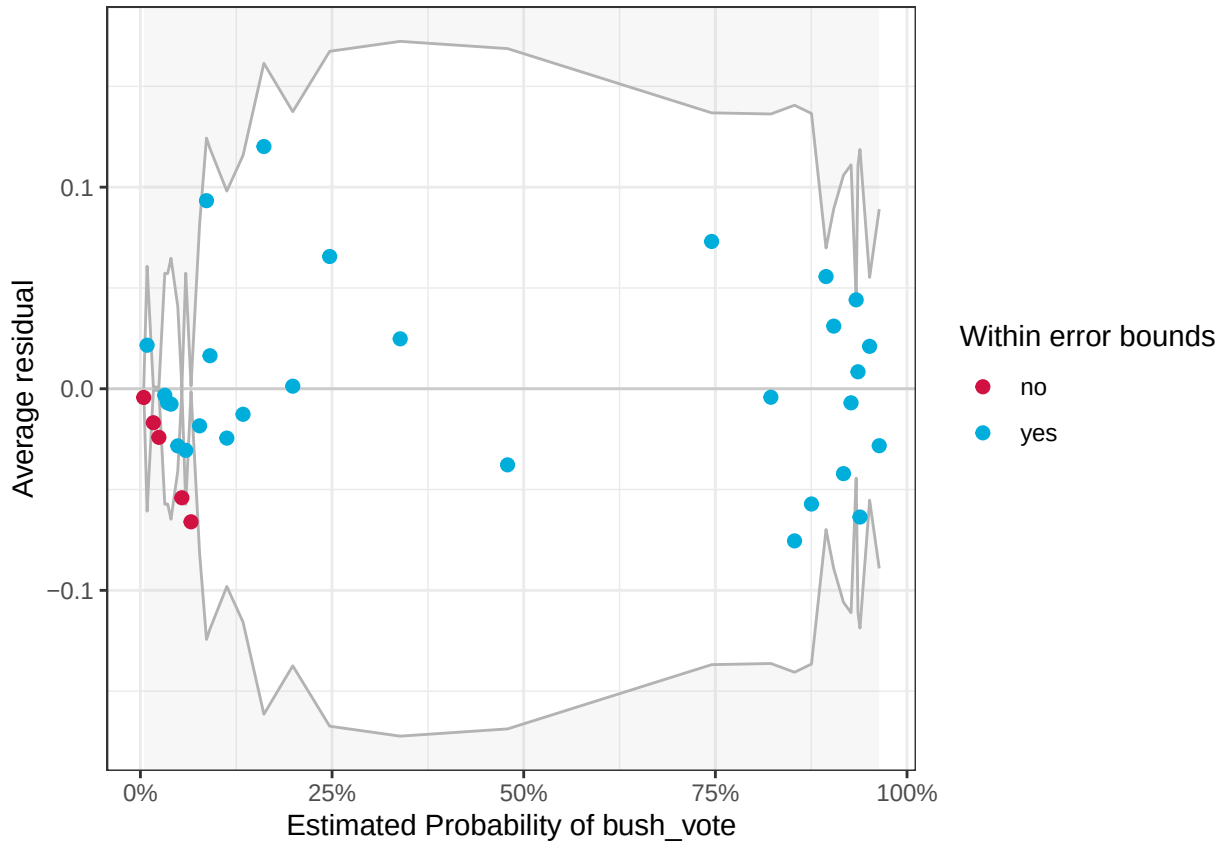


About 85% of the residuals are inside the error bounds for model 1

Model 2:

```
binned_residuals(mod_2)
```

```
## Warning: About 85% of the residuals are inside the error bounds (~95% or higher would be good).
```

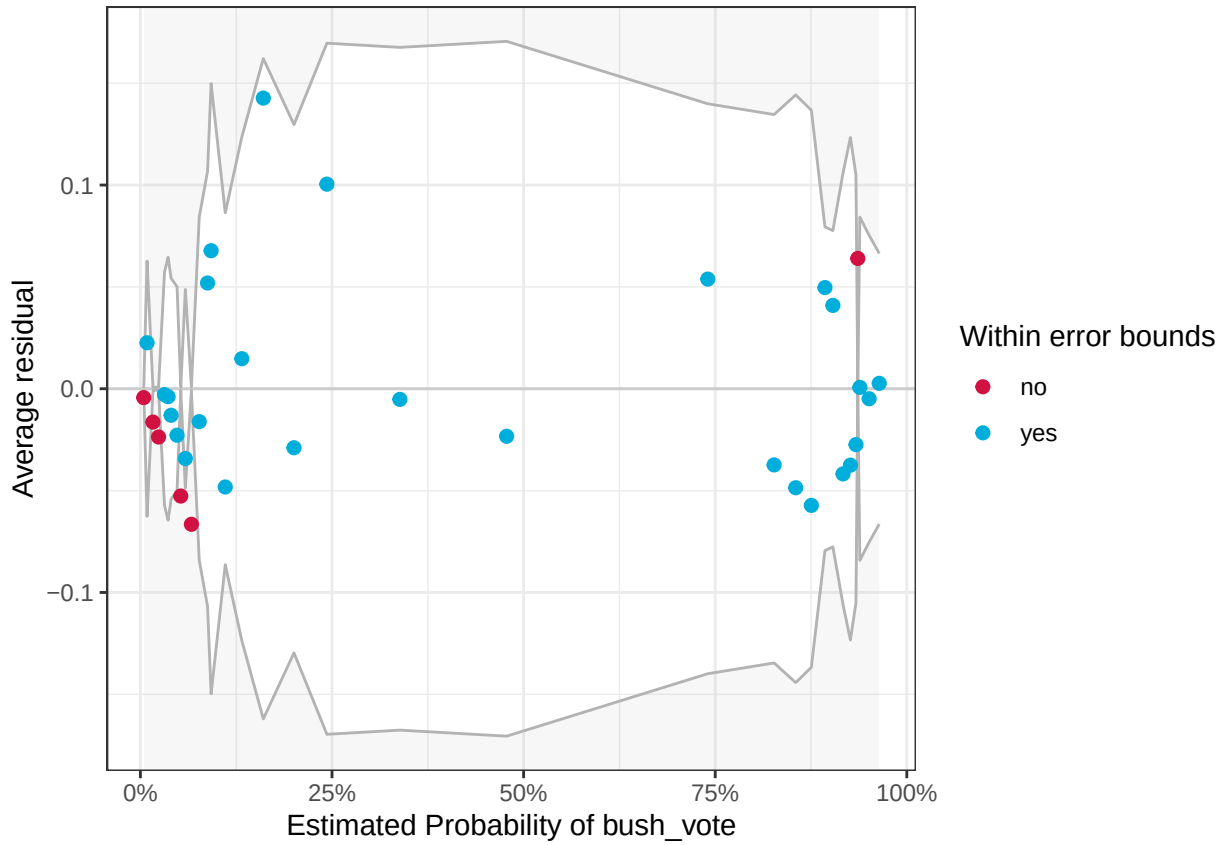


About 85% of the residuals are inside the error bounds for model 2.

Model 3:

```
binned_residuals(mod_3)
```

```
## Warning: About 82% of the residuals are inside the error bounds (~95% or higher would be good).
```

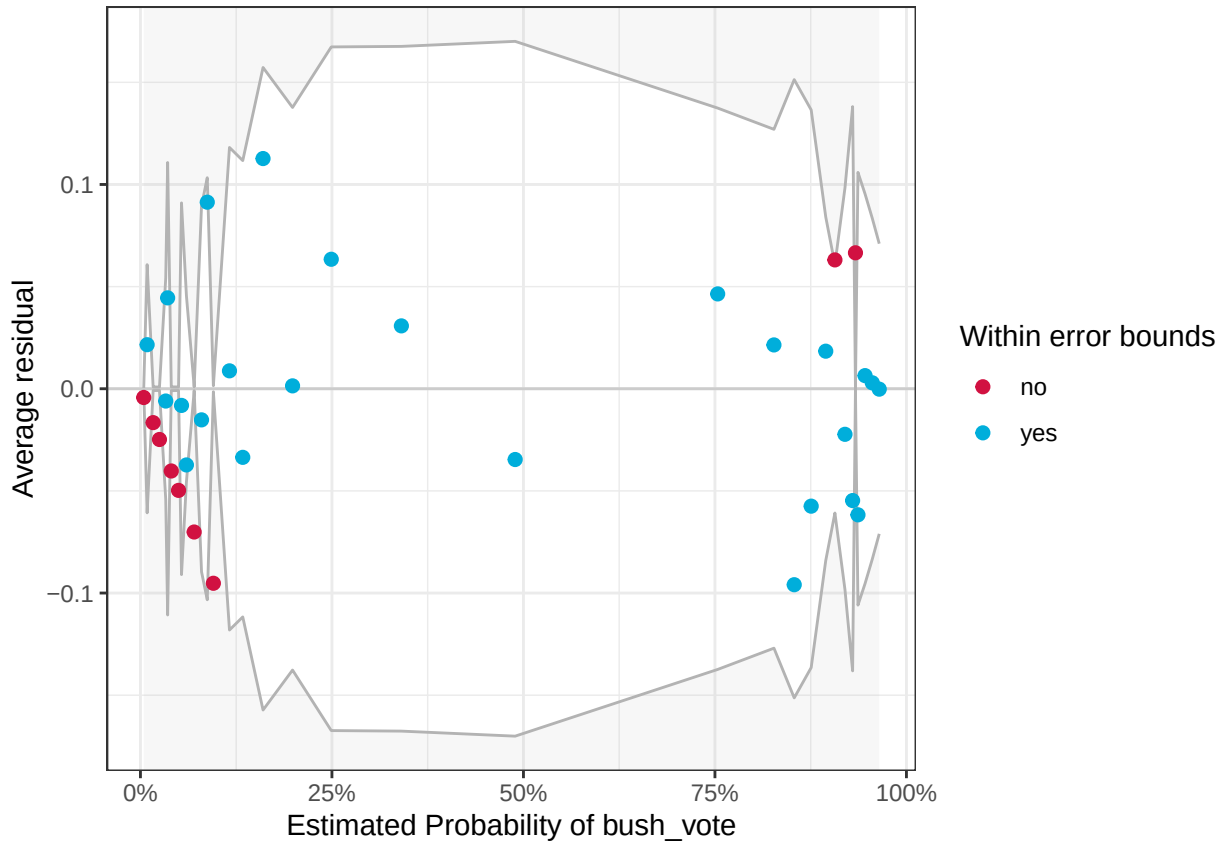


About 82% of the residuals are inside the error bounds for model 3.

Model 4:

```
binned_residuals(mod_4)
```

```
## Warning: Probably bad model fit. Only about 74% of the residuals are inside the error bounds.
```



About 74% of the residuals are inside the error bounds for model 4.

Part C: Model Choice

I think model 2 is the best model because it has the lowest residual deviance and does not include interaction terms that do not add explanatory power. It also has the highest percentage of residuals within the error bounds between models 2, 3, and 4.

```
stargazer(mod_2, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Mar 10, 2021 - 00:58:04

Coefficient interpretations

- Constant term: $logit^{-1}(-4.846) = $ The coefficient for educ1 is .174. In context, this means that a difference in 1 of education level corresponds with a .174 positive difference in the logit probability of voting for Bush. However, the intercept is an impossible condition, so I will not try and interpret the constant term.
- Coefficient for educ 1: to quickly interpret the coefficient on the probability scale, we use the divide by 4 rule. Using this rule (.174/4 = .043), an increase of 1 in education is associated with a 4% positive difference in the probability of voting for Bush holding all else constant, on average. This coefficeint is not statistically significant.

Table 2:

	Dependent variable:
	bush_vote
educ1	0.174
	(0.123)
partyid3I	1.856***
	(0.310)
partyid3R	4.060***
	(0.235)
ideo7	0.481***
	(0.079)
income	0.013
	(0.103)
raceblack	−2.098**
	(0.929)
racehispanic	0.562
	(0.914)
racenative_american	0.593
	(1.017)
racewhite	−0.193
	(0.820)
gender	0.408*
	(0.211)
Constant	−4.846***
	(1.019)
Observations	1,133
Log Likelihood	−337.749
Akaike Inf. Crit.	697.497
Note:	*p<0.1; **p<0.05; ***p<0.01

- Coefficient for partyid3I: to quickly interpret the coefficient on the probability scale, we use the divide by 4 rule. Using this rule (1.856/4 = .464), being an independent is associated with a 46% positive difference in the probability of voting for Bush holding all else constant, on average. This coefficient is statistically significant.
- Coefficient for partyid3R: to quickly interpret the coefficient on the probability scale, we use the divide by 4 rule. Using this rule (4.06/4 = 1.01), being a Republican is associated with a 100% positive difference in the probability of voting for Bush holding all else constant, on average. This coefficient is statistically significant.
- Coefficient for ideo7: to quickly interpret the coefficient on the probability scale, we use the divide by 4 rule. Using this rule (.481/4 = .12), an increase in 1 on the ideology scale (going from very liberal to very conservative) is associated with a 12% positive difference in the probability of voting for Bush, holding all else constant on average. This coefficient is statistically significant.
- Coefficient for income: to quickly interpret the coefficient on the probability scale, we use the divide by 4 rule. Using this rule (.013/4 = .003), an increase in 1 on the income scale is associated with a .03% positive difference in the probability of voting for Bush holding all else constant, on average. This coefficient is not statistically significant.
- Coefficient for raceblack: to quickly interpret the coefficient on the probability scale, we use the divide by 4 rule. Using this rule (-2.098/4 = .-.52), being black is associated with a 52% negative difference in the probability of voting for Bush holding all else constant, on average. This coefficient is statistically significant.
- Coefficient for racehispanic: to quickly interpret the coefficient on the probability scale, we use the divide by 4 rule. Using this rule (.562/4 = .14), being hispanic is associated with a 14% positive difference in the probability of voting for Bush holding all else constant, on average. This coefficient is not statistically significant.
- Coefficient for racenative_american: to quickly interpret the coefficient on the probability scale, we use the divide by 4 rule. Using this rule (.593/4 = .148), being native american is associated with a 14.8% positive difference in the probability of voting for Bush holding all else constant, on average. This coefficient is not statistically significant.
- Coefficient for racewhite: to quickly interpret the coefficient on the probability scale, we use the divide by 4 rule. Using this rule (-.193/4 = -.048), being white is associated with a 5% negative difference in the probability of voting for Bush holding all else constant, on average. This coefficient is not statistically significant.
- Coefficient for gender: to quickly interpret the coefficient on the probability scale, we use the divide by 4 rule. Using this rule (.408/4 = .1), being male is associated with a 10% positive difference in the probability of voting for Bush holding all else constant, on average. This coefficient is statistically significant.

Of all of these variables, the republican indicator variable seems to have the most predictive power. The independent indicator variable and ideology variables also have a relatively large effect on the probability of voting for Bush. The income variable seems to have the least predictive power. Except for the raceblack variable, the race variables do not seem to have a large impact relative to the party and ideology variables (as they are not statistically significant). The gender variable also seems to have a significant impact, although its effect is not as large as party membership. Considering all of this information, party membership seems to be the most important variable in predicting Bush vote.

Question 4

Background

My data for this questions from a 2017 study on how exposure to poverty or affluence affects support for redistributive economic poliicy.

Sands, Melissa L. 2017. "Exposure to inequality affectsredistribution.

I got this idea from one of my Gov51 problem sets. The paper analyzes an experiment designed to test whether exposure to poverty generated more support for higher taxes on the rich. In the experiment, an actor dressed as either an impoverished of affluent person positioned themselves near a person asking people to sign a petition to enact the millionaire's tax (a policy that would raise taxes for people with incomes over $1,000,000). The main outcome of interest is whether or not the person signed the petition.

```
# Data from Sands, Melissa L. 2017. "[Exposure to inequality affects
# redistribution.](http://dx.doi.org/10.1073/pnas.1615010113)" *Proceedings of
# the National Academy of Sciences*, 114(4): 663-668 from gov51 pset 4. The
# paper analyzed an experiment designed to test whether exposure to poverty
# generated more support for redistribution policy. In the experiment, an actor
# dressed as either a impoverished person or an affluent person was positioned
# around a person asking people to sign a petition to enact the millionaire's
# tax (a tax that would theoretically tax affluent Americans more).


ineq_df <- read.csv("inequality-exposure.csv")

#|:-------------|:-------------------------------------------------------------------------|
#| `signed`     | 1 if the respondent signed the petition, 0 otherwise                     |
#| `mill_tax`   | 1 if petitioned about the millionaire's tax, 0 for plastic bag petition. |
#| `blackactor` | 1 if actor was Black for this respondent, 0 for white                    |
#| `pooractor`  | 1 if actor was in poverty condition, 0 for affluent condition            |
#| `black`      | 1 if petitioner guessed respondent was Black                             |
#| `white`      | 1 if petitioner guessed respondent was non-Hispanic white                |
#| `asian`      | 1 if petitioner guessed respondent was Asian                             |
#| `hisp`       | 1 if petitioner guessed respondent was Hispanic                          |
#| `young`      | 1 if petitioner guessed respondent was 18-35 years old                   |
#| `middle`     | 1 if petitioner guessed respondent was 36-65 years old                   |
#| `old`        | 1 if petitioner guessed respondent was >65 years old                     |
#| `female`     | 1 if petitioner guessed respondent was female                            |
#| `clust`      | Cluster number of respondent (see question 6)                            |



# changing the data to make the age variable continuous

mill_df <- ineq_df %>%
  filter(mill_tax == 1) %>%
  mutate(age = case_when(young == 1 ~ 1,
                         middle == 1 ~ 2,
                         old == 1 ~ 3)) %>%
  drop_na()
```

Models

I created three different models of signed vs. a variety of independent variables. Independent variables:

- pooractor: a variable that indicates whether or not the actor was dressed as poor or affluent
- blackactor: a variable that indiciates whether or not the actor
- black/white/asian/hisp: the race of the respondennt
- female: the gender of the respondent
- age: 1 if the respondent young, 2 if the respondent middle aged, 3 if the respondent old

Models:

- Model 1 is a model of signed vs. pooractor (a variable that indicates whether or not the actor was dressed as poor or affluent)
- Model 2 is a model of signed vs. pooractor + blackactor
- Model 3 is a model of signed vs. pooractor + blackactor + black + white + asian + hisp + female + age + age:female

```
# model 1 looks at main treatment effect (the affluence or poverty of the actor).


mod_4_1 <- glm(signed ~ pooractor, data = mill_df,
               family = "binomial")

# model 2 looks at signed vs pooractor and demographic info (race and age)

mod_4_2 <-
  glm(
    signed ~ pooractor + blackactor,
    data = mill_df,
    family = "binomial"
  )

# model 3 adds an interaction between age and female

mod_4_3 <-
  glm(
    signed ~ pooractor + blackactor + black + white + asian + hisp + female + age + age:female,
    data = mill_df,
    family = "binomial"
  )


stargazer(mod_4_1, mod_4_2, mod_4_3, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Mar 10, 2021 - 00:58:04

Model Choice: deviances

The residual deviance for Model 1 is about 2 less than the Null Deviance, suggesting that pooractor is an informative predictor. The residual deviance for Model 2 is with 1 of the residual deviance for model 1,

Table 3:

	Dependent variable:		
	signed		
	(1)	(2)	(3)
pooractor	−0.285	−0.298	−0.298
	(0.207)	(0.208)	(0.209)
blackactor		−0.157	−0.116
		(0.214)	(0.216)
black			0.160
			(0.854)
white			−0.286
			(0.763)
asian			−1.136
			(0.912)
hisp			−0.181
			(0.967)
female			−0.455
			(0.581)
age			0.046
			(0.236)
female:age			0.226
			(0.309)
Constant	−2.356***	−2.249***	−2.006**
	(0.139)	(0.199)	(0.853)
Observations	1,334	1,334	1,334
Log Likelihood	−359.284	−359.020	−355.296
Akaike Inf. Crit.	722.568	724.041	730.591
Note:			*p<0.1; **p<0.05; ***p<0.01

suggesting that blackactor does not add predictive power to the model. The residual deviance for Model 3 is about 7.5 than the residual deviance for model 2. Because we added 7 new variables to the model, we would expect the deviance to drop by about 7 points, meaning that it is likely that the demographic variables for the respondent are random noise.

```
# showing the null and residual deviances. The Residual Dev in lowest for model
# 3. Well first should point out that they are all super close

an_4 <- anova(mod_4_1, mod_4_2, mod_4_3)

tibble(
  "Model" = c("mod_4_1", "mod_4_2", "mod_4_3"),
  "Null Dev" = c(720.5, 720.6, 720.5),
  "Resid. Dev" = an_4$"Resid. Dev"
)
```

```
## # A tibble: 3 x 3
##   Model   `Null Dev` `Resid. Dev`
##   <chr>        <dbl>        <dbl>
## 1 mod_4_1       720.         719.
## 2 mod_4_2       721.         718.
## 3 mod_4_3       720.         711.
```

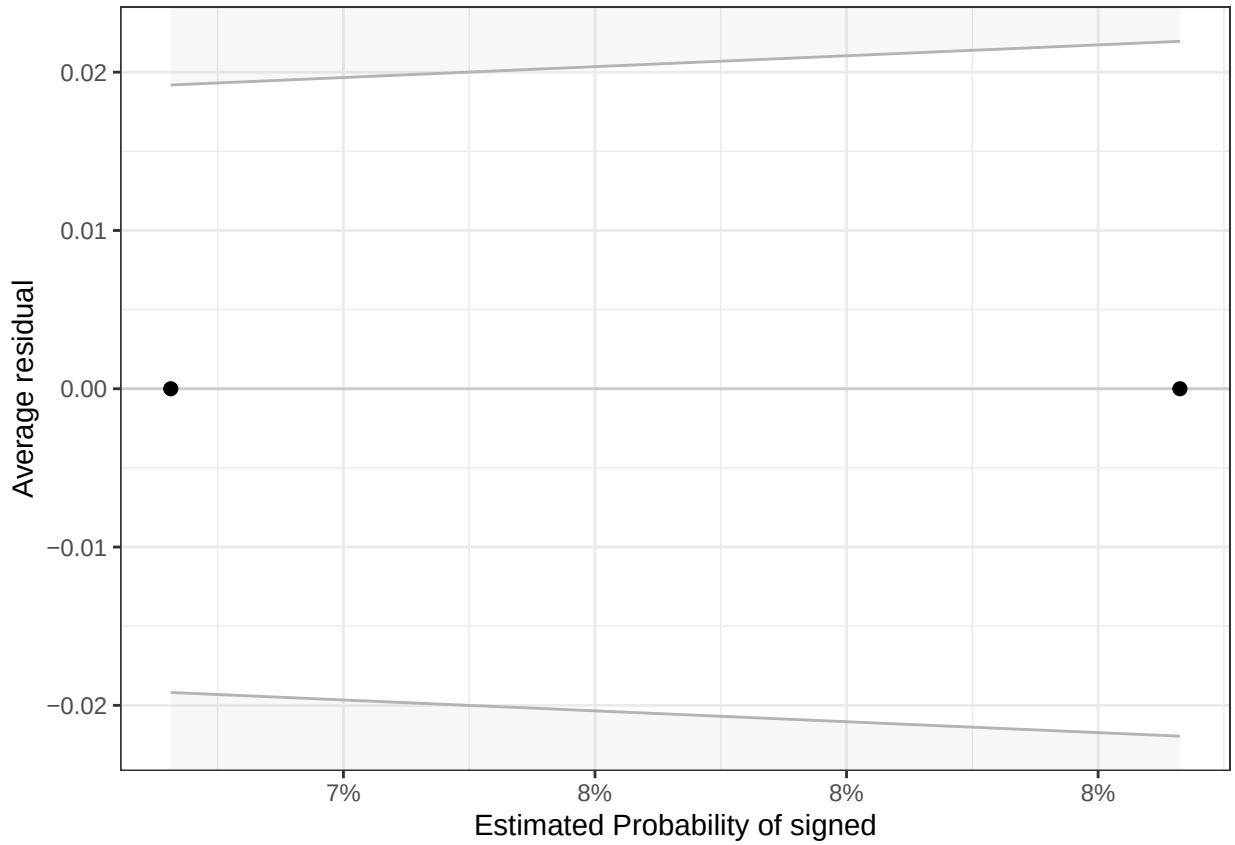
Model Choice: residuals

Model 1:

The plot looks a bit odd because the only explanatory variable is a binary variable. However, all of the residuals are within the error bounds.

```
binned_residuals(mod_4_1)
```

```
## Ok: About 100% of the residuals are inside the error bounds.
```

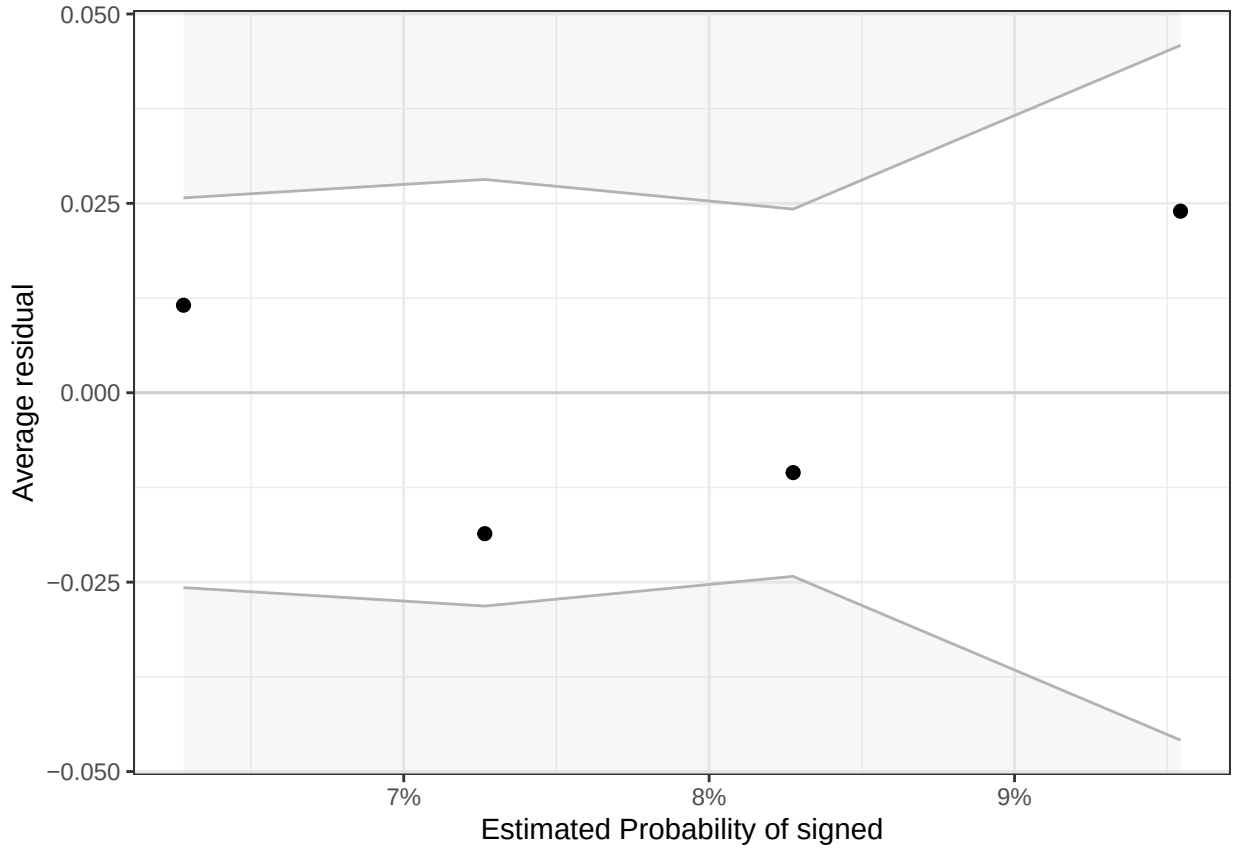


Model 2:

The plot still looks a bit odd because both the explanatory variables are binary variable. However, all of the residuals are within the error bounds.

```
binned_residuals(mod_4_2)
```

```
## Ok: About 100% of the residuals are inside the error bounds.
```

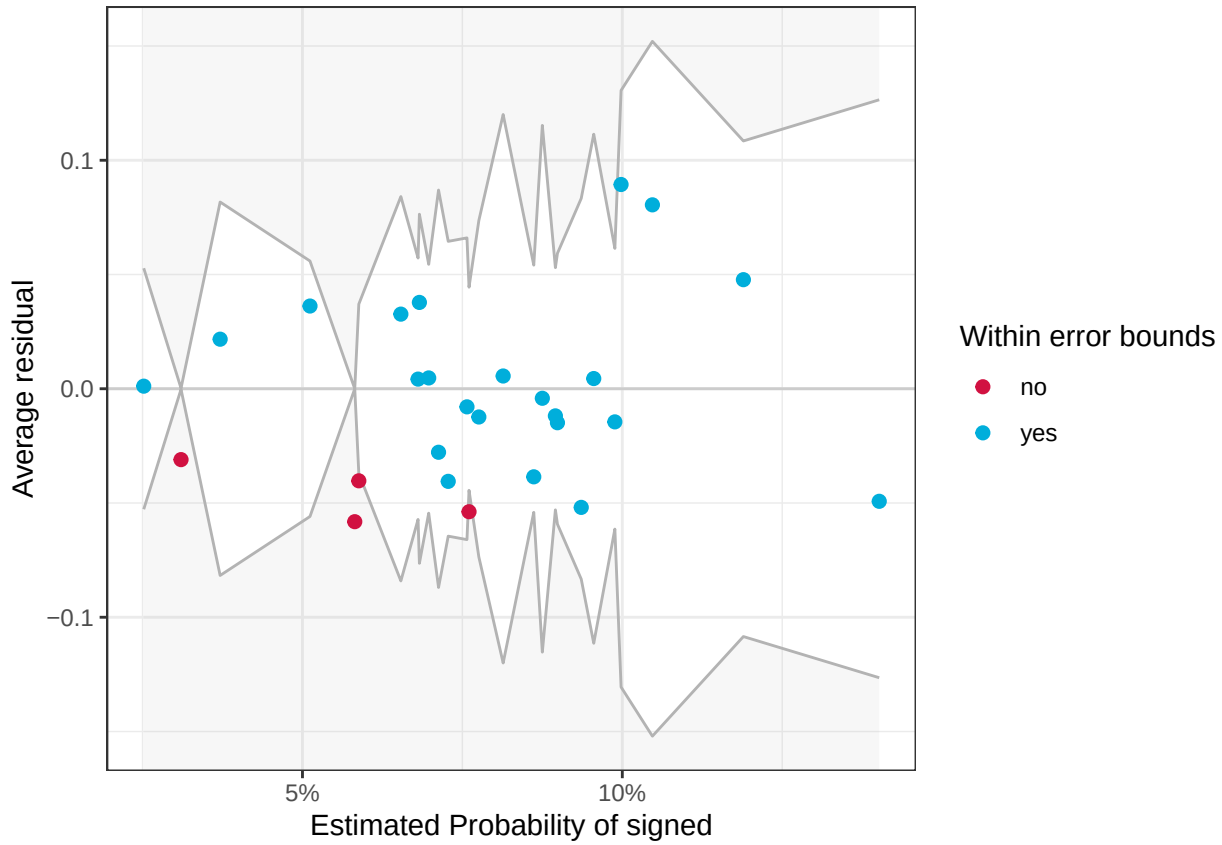


Model 3:

About 85% of the residuals fall within the error bounds.

```
binned_residuals(mod_4_3)
```

```
## Warning: About 85% of the residuals are inside the error bounds (~95% or higher would be good).
```



Model Choice:

While none of these models seem great, I think that it is best to choose model 1, because the residual deviances in models 2 and 3 suggested that all of the variables except for pooractor were not informative predictors of signing the petition.

Coeficient Interpretation

- Constant: the intercept in this context can be interpreted as a respondent's average probability of signing the petition given they saw an actor dressed as affluent. Using the divide by 4 rule, this average probability is -59%.
- pooractor: This coefficient tells us how seeing a pooractor effects a respondent's probability of singing the petition given they saw an actor dressed as poor. Using the divide by 4 rule, seeing an actor dressed as poor is associated with a 7% negative difference in the probability of signing the petition on average.

From these estimate, it seems like the respondents have a low probability of signing the petition whether or not they saw an affluent or poor actor.

```
stargazer(mod_4_1, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Mar 10, 2021 - 00:58:06

Table 4:

	Dependent variable:
	signed
pooractor	−0.285
	(0.207)
Constant	−2.356***
	(0.139)
Observations	1,334
Log Likelihood	−359.284
Akaike Inf. Crit.	722.568
Note:	*p<0.1; **p<0.05; ***p<0.01

Error Rate

The null error rate is the proportion of 1's in the data. In case, the null error rate is 7%, which means that very few respondents signed the survey in the study.

The model only predicted that respondents would not sign the survey and did not predict any successes. It correctly predicted 1232 out of 1334 failures sign. As such, the error rate is calculated by dividing the incorrectly classified observations by the sum of all of the observations. In this case, the error rate is equal to 7%, the same as the null error rate, suggesting that the model does not have any meaningful explanatory power.

```
null_error <- mean(mill_df$signed)

error_rate <- predict(mod_4_1, mill_df, type = "response")

# tab_mat <- table(mill_df$signed, predict > 0.5)

# tab_mat
```

Test case

In order to create a test case for the model, I creating a training data set out of the first half of the observations and a test data set out of the last half of the observations and fitted the model using only the training set.

The model correctly predicted 614 of 657 outcomes in the test data set. This equates to an error rate of about 7.9%.

```
train <- head(mill_df, nrow(mill_df)/2)
test <- tail(mill_df, nrow(mill_df)/2)

mod_fin <- glm(signed ~ blackactor, data = train)

predictions <- predict(mod_fin, newdata = test)

pred_df <- cbind(predictions, test)
```

```
pred_df %>%
  mutate(result = if_else(predictions < .5, 0, 1),
         correct = if_else(result == signed, 1, 0)) %>%
  count(correct)
```

```
##   correct   n
## 1       0  53
## 2       1 614
```

Concluding thoughts

As I went through this exercise, I realized that I had chosen a bad data set to perform a logistic model on. I wish I had chosen a different data set, but I was already too far down the path with this one.