HW3 Team 04

Linlin Li (ll360), Bingruo Wu (bw199), and Jae Hyun Lee (jl914)

Sep 24th

We will explore logistic regression with the National Election Study data from Gelman & Hill (GH). (See Chapter 4.7 for descriptions of some of the variables and 5.1 of GH for initial model fitting). The link here may also be useful for background information http://gking.harvard.edu/files/preelection.pdf or http://www.icpsr.umich.edu/cgi-bin/file?comp=none&study=8475&ds=1&file_id=1196048&path=ICPSR

1. Summarize the data for 1992 noting which variables have missing data. Can you think of any reasons why they might be missing? Which variables are categorical but are coded as numerically?

```
na_count <- summary(nes1992)[7,]
na_count <- as.matrix(cbind(attributes(nes)$var.label,na_count))
na_count <- na_count[!is.na(na_count[,2]),]
na_count[,2] <- as.numeric(unlist(str_extract_all(na_count[,2],"[1-9]+[0-9]*")))
kable(na_count, col.names = c("variable description","number of NA"),
      caption = "Summary of variables have missing data")
```

Table 1: Summary of variables have missing data

	variable description	number of NA
occup1	respondent occupation (1)	47
union	household union membership	5
religion	religion of r (1)	2
martial_status	marital status of r	2
occup2	respondent occupation (2)	25
icpsr_cty	county of iw 1968-1982	1179
partyid7	7-pt scale party identification	2
partyid3	party id collapsed (1)	1
partyid3_b	party id collapsed (2)	1
str_partyid	strength of r partisanship	1
father_party	r father party id	173
mother_party	r mother party id	170
dem_therm	dem presidential candidate thermometer	12
rep_therm	rep presidential candidate thermometer	4
regis	is r registered to vote (pre)	1179
presvote_intent	r intent for presidential vote (pre)	14
ideo_feel	liberal-conservative thermometer index	53
ideo7	liberal-conservative 7pt scale	45
ideo	r position lib/cons 3-category summary	45
cd	congressional district of residence	12
rep_pres_intent		91
real_ideo		231
presapprov	approve presidential performance	15
perfin1	personal financial situation in past yr	3
perfin2	r fin situation last few yrs 1956-1964	1179
perfin		3
newfathe		173
newmoth		170
parent_party		230

Variables that have missing data are:

`black`, `female`, `educ1`, `age`, `state`, `income`, `presvote`, `occup1`, `union`, `religion`, `martial_status`, `occup2`, `icpsr_cty`, `partyid7`, `partyid3`, `partyid3_b`, `str_partyid`, `father_party`, `mother_party`, `dem_therm`, `rep_therm`, `regis`, `presvote_intent`, `ideo_feel`, `ideo7`, `ideo`, `cd`, `rep_pres_intent`, `real_ideo`, `presapprov`, `perfin1`, `perfin2`, `perfin`, `newfathe`, `newmouth`, `parent_party`.

One reason is that if there are no questions about these variables in the 1992 survey, then these variables should have missing in `nes1992`. For instance, `icpsr_cty variable` only includes data from 1968 to 1982, thus it automatically have `NA` in `nes1992`. This is also the case with `perfin2` and `regis`.

Secondly, some of the respondents may refuse to answer questions about personal privacy, such as occupation(`occup1` and `occup2`), religion, marital status(`marital_status`), and union membership(`union`).

The third reason is that some of the respondent may refuse to answer the ideo related questions, so variables like `ideo_feel` and `ideo7` have missing data.

Fourthly, maybe some of the respondents don't know others' political party preference, so `father_party`, `mother_party`, and `parent_party` have missing data.

Categorical variables but coded numerically are:

`gender`, `race`, `educ1`, `urban`, `region`, `income`, `occup1`, `union`, `religion`, `educ2`, `educ3`, `martial_status`, `occup2`, `partyid7`, `partyid3`, `partyid3_b`, `str_partyid`, `father_party`, `mother_party`, `dlikes`, `rlikes`, `regisvote`, `presvote`, `presvote_2party`, `presvote_intent`, `ideo7`, `ideo`, `cd`, `state`, `inter_pre`, `inter_post`, `female`, `rep_presvote`, `rep_pres_intent`, `south`, `real_ideo`, `presapprov`, `perfin1`, `perfin`, `presadm`, `newfathe`, `newmoth`, `parent_party`, `white`.

2. Fit the logistic regression to estimate the probability that an individual would vote Bush (Republican) as a function of `income` and provide a summary of the model.

We treated `income` as a categorical variable, even though it is coded as numerically in the data set, because the difference between each level of income may be different.

```
glm_fit.1 <- glm(vote ~ factor(income), data = nes1992, family = binomial(link=logit))
kable(summary(glm_fit.1)$coef, digits=4,
      caption = "Summary for glm(vote ~ factor(income))")
```

Table 2: Summary for glm(vote ~ factor(income))

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.1838	0.2087	-5.6733	0.0000
factor(income)2	0.4906	0.2555	1.9203	0.0548
factor(income)3	0.7509	0.2345	3.2023	0.0014
factor(income)4	1.1243	0.2312	4.8634	0.0000
factor(income)5	1.2378	0.3125	3.9616	0.0001

Based on the summary, at the significance level of 0.05, `factor(income)3`, `factor(income)4`, and `factor(income)5` have a statistically significant effect on the odds ratio for voting Republican, showing that there is a correlation between voting and income. Since the coefficients of `factor(income)2`, `factor(income)3`, `factor(income)4`, and `factor(income)5` are estimated to be positive and incremental, it indicates that a rich person is more likely to vote Republican than a poor person.

However, the residual deviance of this model is 1555.8 on 1174 degrees of freedom. Compared to the null model, the reduction of deviance is not very large, we may need to add new variables to the model.

3. Obtain a point estimate and create a 95% confidence interval for the odds ratio for voting Republican for a rich person (income category 5) compared to a poor person (income category 1). Provide a sentence interpreting the result.

Model:

$$log(\pi_{poor}/(1-\pi_{poor})) = \beta_0,$$
$$log(\pi_{rich}/(1-\pi_{rich})) = \beta_0 + \beta_4,$$
$$log(\frac{\pi_{rich}/(1-\pi_{rich})}{\pi_{poor}/(1-\pi_{poor})}) = \beta_4.$$

So, in order to compare the odds ratio for a rich person and a poor one, we need to estimate β_4.

Using the invariance property of the MLE allows us to exponentiate to get $e^{\beta_j \pm z^* SE(\beta_j)}$, where $j = 0, 1, 2, 3, 4$, which is the confidence interval on the odds ratio.

```
odds <- exp(glm_fit.1$coefficients[5]) # point estimate
names(odds)=c("Estimate")
table=c(odds,exp(confint(glm_fit.1, level =0.95)[5,])) # CI
kable(t(table),caption = "Estimate of factor(income)5",digits = 4)
```

Table 3: Estimate of factor(income)5

Estimate	2.5 %	97.5 %
3.4481	1.8797	6.4162

The point estimate for β_4 is 3.4481, indicating that the odds ratio for a rich person (income category 5) to vote Republican is 3.4481 times that for a poor person (income category 1). A 95% confidence interval for β_4 is [1.8797, 6.4162], which means that we are 95% confident that the odds ratio of voting Republican for a rich person is 1.8797 to 6.4162 times that of a poor person.

4. Obtain fitted probabilities and 95% confidence intervals for the income categories using the `predict` function. Use `ggplot` to recreate the plots in figure 5.1 of Gelman & Hill.

```
incomes <- sort(unique(nes1992$income),decreasing = F)
fit_CI_income <- data.frame(matrix(rep(0,3*length(incomes)),nrow = length(incomes)))
for(i in seq_along(incomes)){
  glm_pred <- predict(glm_fit.1, newdata = data.frame(income = i),
                      type = "response",se.fit = T)
  fit_CI_income[i,1] <- glm_pred$fit
  fit_CI_income[i,2] <- glm_pred$fit + glm_pred$se.fit * qnorm(0.025)
  fit_CI_income[i,3] <- glm_pred$fit + glm_pred$se.fit * qnorm(0.975)
}
colnames(fit_CI_income) <- c("fitted","2.5%","97.5%")
rownames(fit_CI_income) <- paste("factor(income)",1:5)
kable(fit_CI_income, digits = 4,
      caption = "Fitted probabilities and 95% confidence intervals for
      the odds ratio for voting Republican with different income categories in 1992")
```

Table 4: Fitted probabilities and 95% confidence intervals for the odds ratio for voting Republican with different income categories in 1992

	fitted	2.5%	97.5%
factor(income) 1	0.2344	0.1610	0.3078
factor(income) 2	0.3333	0.2691	0.3976
factor(income) 3	0.3934	0.3434	0.4435
factor(income) 4	0.4851	0.4364	0.5339
factor(income) 5	0.5135	0.3996	0.6274

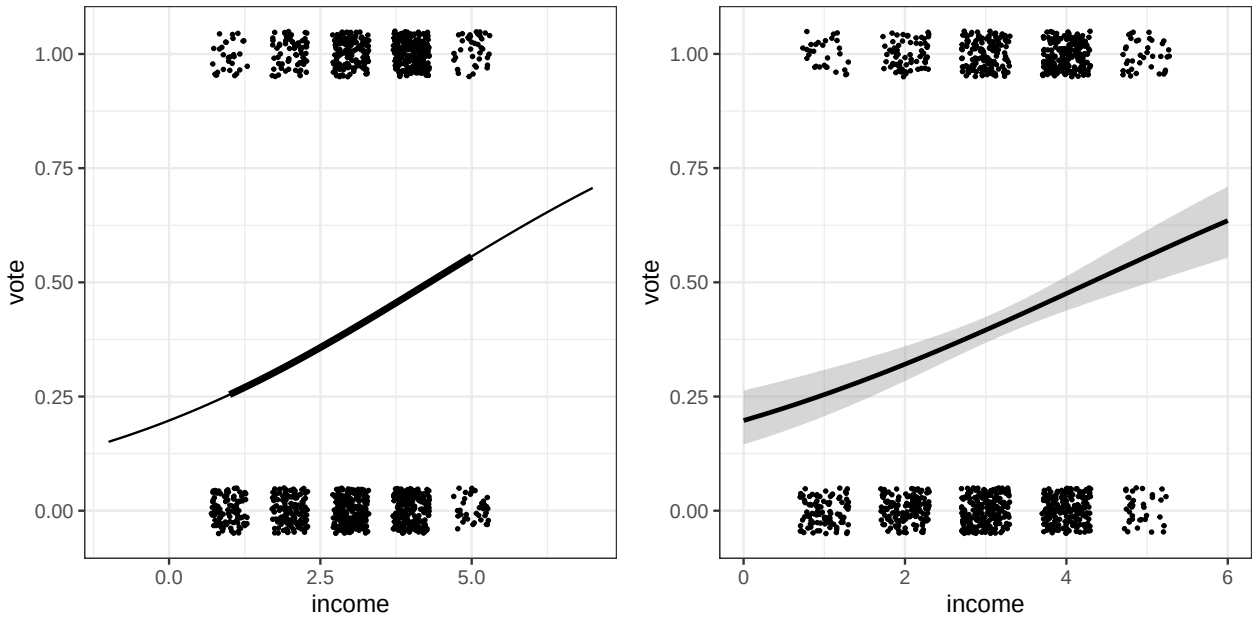

Figure 1: Logistic regression estimating the probability of supporting George Bush in the 1992 presidential election, as a function of discretized income level. (a) Fitted logistic regression: the thick line indicates the curve in the range of the data; the thinner lines at the end show how the logistic curve approaches 0 and 1 in the limits. (b) In the range of the data, the solid line shows the best-fit logistic regression, and the light lines show uncertainty in the fit.

```
plot1 <- ggplot(data = nes1992,
                mapping = aes(x = income, y = as.numeric(vote))) +
          geom_jitter(width = 0.3, height = 0.05, size = 0.5) +
          geom_smooth(method = "glm",
                      method.args = list(family = "binomial"),
                      size = 1.5, se = F, col = "black") +
          geom_smooth(method = "glm",
                      method.args = list(family = "binomial"),
                      size = 0.5 , fullrange = T, se = F, col = "black") +
          xlim(-1,7) +
          labs(x = "income" , y = "vote") +
          theme_bw()

plot2 <- ggplot(data = nes1992,
                mapping = aes(x = income, y = as.numeric(vote))) +
  geom_jitter(width = 0.3, height = 0.05, size = 0.5) +
  geom_smooth(method = "glm",
              method.args = list(family = "binomial"),
              size = 1, se = T, col = "black", fullrange = T) +
  labs(x = "income" , y = "vote") +
  theme_bw() +
  xlim(0,6)
plots <- plot1 + plot2
plots
```

According to Table 4 and Figure 1, the fitted probabilities for the income categories are incremental, showing the positive correlation between wealth and voting Republican.

5. What does the residual deviance or any diagnostic plots suggest about the model? (Do provide code for p-values and output and plots)

```
kable(anova(glm_fit.1, test = "Chisq"), caption = "Analysis of deviance for the model")
```

Table 5: Analysis of deviance for the model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	1178	1591.238	NA
factor(income)	4	35.4557	1174	1555.782	4e-07

```
p_val <-pchisq(glm_fit.1$deviance,glm_fit.1$df.residual,lower.tail=F)
p_val
```

```
## [1] 3.726962e-13
```

```
par(mfrow=c(2,2))
plot(glm_fit.1)
```

When we see the result of chisq test of residual deviance, it shows that it is very unlikely to have this large residual deviance if the model is good model. Thus we can conclude that the model have lack of fit problem.

Diagnostic plot of this model does not provide adequate assessment about model at all. For normal QQ plot, it is not needed in this model, because response variable is binary data. Scale-location plot is also not useful because variance of binary data varies according to $\hat{\pi}$. Residual vs fitted plot only shows 2 distinct lines which consist of cases when $y_i = 1$ and $y_i = 0$.

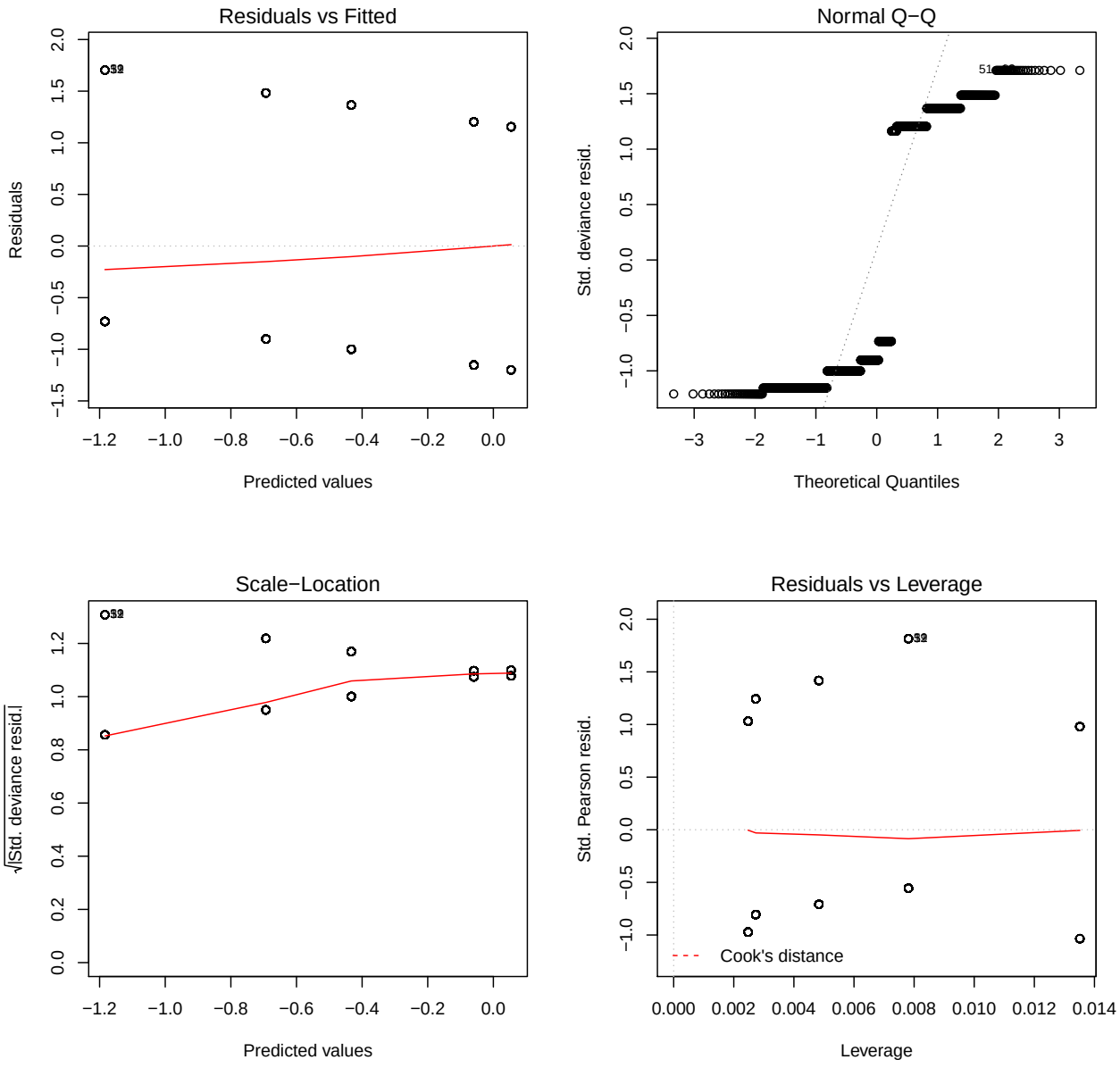

Figure 2: Diagnostic plots

6. Create a new data set by the filtering and mutate steps above, but now include years between 1952 and 2000.

```
newnes = nes %>% filter(!is.na(black)) %>%
              filter(!is.na(female)) %>%
              filter(!is.na(educ1)) %>%
              filter(!is.na(age)) %>%
              filter(!is.na(state)) %>%
              filter(!is.na(income)) %>%
              filter(presvote %in% 1:2) %>%
              filter(year >= 1952 & year <= 2000) %>%
              mutate(female = gender -1,
                     black=race ==2,
                     vote = presvote == 2)
```

7. Fit a separate logistic regression for each year from 1952 to 2000, using the `subset` option in `glm`, i.e. add `subset=year==1952`. For each find the 95% Confidence interval for the odds ratio of voting republican for rich compared to poor for each year in the data set from 1952 to 2000.

Similar to what we have done in Q5, the confidence interval on the odds ratio is $\exp\{\beta_j \pm z^* SE(\beta_j)\}$. Thus using same algorithms, we calculated the confidence interval for odds ratio for each year.

```
years <- unique(newnes$year)
OR_ci <- data.frame(matrix(rep(0,3*length(years)),ncol = 3))
for (i in 1:length(years)){
  glm_fit.2 <- glm(vote ~ factor(income), data = newnes,
                   family = binomial(link = logit),
                   subset = year==years[i])
  OR_ci[i,1] <- exp(glm_fit.2$coefficients[5])
  OR_ci[i,2:3] <- exp(confint(glm_fit.2,level = 0.95)[5,])
}
colnames(OR_ci) <- c("fitted","2.5%","97.5%")
rownames(OR_ci) <- paste(years,"yr")
kable(OR_ci, digits = c(4,4,4),
      caption = "Fitted probabilities and 95% confidence intervals for the
      odds ratio for voting Republican for each year from 1952 to 2000")
```

Table 6: Fitted probabilities and 95% confidence intervals for the odds ratio for voting Republican for each year from 1952 to 2000

	fitted	2.5%	97.5%
1952 yr	2.5473	1.2973	5.2809
1956 yr	2.3817	1.3770	4.1915
1960 yr	2.6975	1.2907	5.9457
1964 yr	2.5587	1.4735	4.4674
1968 yr	1.8342	0.9242	3.7375
1972 yr	3.5156	2.0188	6.3958
1976 yr	6.8205	3.6921	13.0457
1980 yr	9.1429	3.9485	24.0483
1984 yr	7.5408	4.0056	14.8758
1988 yr	4.9176	2.2977	11.3365
1992 yr	3.4481	1.8797	6.4162
1996 yr	4.3022	2.2428	8.4274
2000 yr	5.0682	2.5351	10.4668

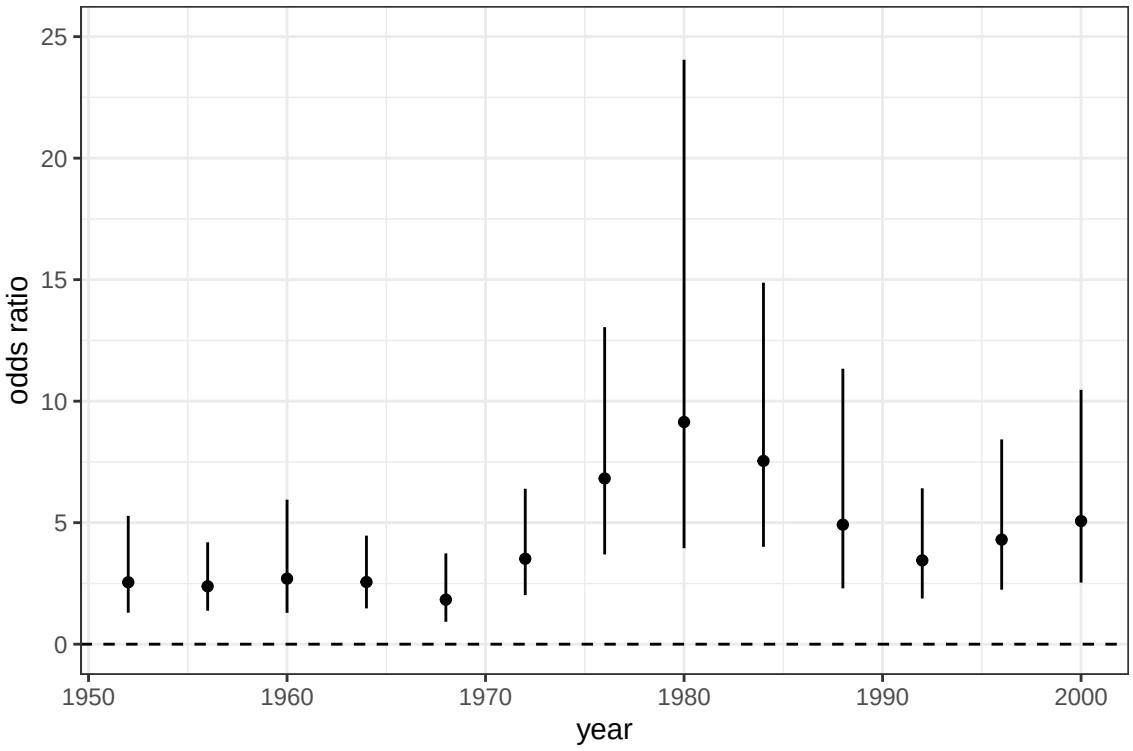

Figure 3: Coefficient of income (on a 1–5 scale) with 95% confidence interval in logistic regressions predicting Republican preference for president, as estimated separately from surveys from 1952 to 2000.

8. Using `ggplot` plot the confidence intervals over time similar to the display in Figure 5.4.

You can find our plot at Figure 3

```
ggplot(data = OR_ci, mapping = aes(x = seq(1952,2000,4), y = fitted)) +
  geom_point() +
  geom_linerange(ymin = OR_ci[,2],ymax = OR_ci[,3]) +
  ylim(0,25) +
  labs(x = "year", y = "odds ratio") +
  theme_bw() +
  geom_abline(slope = 0, intercept = 0, linetype = "dashed")
```

9. Fit a logistic regression using income and year as a factor with an interaction i.e. `income*factor(year)` to the data from 1952-2000. Find the log odds ratio for income for each year by combining parameter estimates and show that these are the same as in the respective individual logistic regression models fit separately to the data for each year.

```
glm_fit.3 <- glm(vote ~ factor(income)*factor(year),data = newnes,
                 family = binomial(link = logit))
OR <- rep(0,length(years))
for (i in 1:length(OR)){
  if (i == 1){
    OR[i] <- glm_fit.3$coefficients[5]
    next
  }
  OR[i] <- glm_fit.3$coefficients[5]+glm_fit.3$coefficients[13+4*i]
}
result <- cbind(OR,log(OR_ci[,1]))
```

```
result <- round(result,4)
result=cbind(result,check=ifelse(all.equal(result[,1],result[,2]),"TRUE","FALSE"))
colnames(result)=c("simultaneously","respectively","check simultaneously = respectively")
rownames(result) <- paste(years,"yr")
kable(result, digits = c(4,4),
      caption = "Comparison between the estimate for coefficient of income in the
      all model with interaction and that in the respective individual models")
```

Table 7: Comparison between the estimate for coefficient of income in the all model with interaction and that in the respective individual models

	simultaneously	respectively	check simultaneously = respectively
1952 yr	0.935	0.935	TRUE
1956 yr	0.8678	0.8678	TRUE
1960 yr	0.9923	0.9923	TRUE
1964 yr	0.9395	0.9395	TRUE
1968 yr	0.6066	0.6066	TRUE
1972 yr	1.2572	1.2572	TRUE
1976 yr	1.9199	1.9199	TRUE
1980 yr	2.213	2.213	TRUE
1984 yr	2.0203	2.0203	TRUE
1988 yr	1.5928	1.5928	TRUE
1992 yr	1.2378	1.2378	TRUE
1996 yr	1.4591	1.4591	TRUE
2000 yr	1.623	1.623	TRUE

In Table 7, we found that the estimated coefficients of log odds ratio for income for each year are the same as in the respective individual logistic regression models for each year, which is consistent with our intuition.

10. Create a plot of fitted probabilities and confidence intervals as in question 4, with curves for all years in the same plot.

```
plot1 <- ggplot(data = newnes,
                mapping = aes(x = income, y = as.numeric(vote),
                              color = factor(year))) +
          geom_jitter(width = 0.3, height = 0.05, size = 0.1) +
          geom_smooth(method = "glm", method.args = list(family = "binomial"),
                      size = 1.5, se = F) +
          geom_smooth(method = "glm", method.args = list(family = "binomial"),
                      size = 0.5 ,
                      fullrange = T, se = F) +
          xlim(-1,7) +
          labs(x = "income" , y = "Probability of voting Republican",
               color = "year") + theme_bw()

plot2 <- ggplot(data = newnes,
                mapping = aes(x = income, y = as.numeric(vote),
                              color = factor(year))) +
         geom_jitter(width = 0.3, height = 0.05, size = 0.1) +
         geom_smooth(method = "glm", method.args = list(family = "binomial"),
                     size = 1, se = T, alpha=0.25) +
         labs(x = "income" , y = "Probability of voting Republican",
```

```
              color = "year") +
         theme_bw()

plot1 / plot2
```

Figure 4 (b) shows the fitted probabilities and 95% confidence intervals for different colors in different years. We can see that the lines for recent years(1976-2000) seem to have a larger slope than the lines for the previous years(from 1952 to 1972), indicating that the correlation between income and voting preferences is stronger than past.

11. Return to the 1992 year data. Filter out rows of `nes1992` with NA's in the variables below and recode as factors using the levels in parentheses: + gender (1 = "male", 2 = "female"), + race (1 = "white", 2 = "black", 3 = "asian", 4 = "native american", 5 = "hispanic", 7 = "other"), + education (use `educ1` with levels 1 = "no high school", 2 = "high school graduate", 3 = "some college", 4 = "college graduate"), + party identification (`partyid3` with levels 1= "democrats", 2 = "independents", 3 = "republicans", 9 = "apolitical" , and + political ideology (`ideo` 1 = "liberal", 3 ="moderate", 5 = "conservative")

```
newnes_1992 <- nes1992 %>%
  filter(!is.na(gender)) %>%
  filter(!is.na(race)) %>%
  filter(!is.na(educ1)) %>%
  filter(!is.na(partyid3)) %>%
  filter(!is.na(ideo))

newnes_1992 <- newnes_1992 %>%
  mutate(gender = recode(gender, '1' = "male", '2' = "female")) %>%
  mutate(race = recode(race, '1' = "white", '2' ="black", '3' = "asian",
                       '4' = "native american", '5' = "hispanic",
                       '7' = "other")) %>%
  mutate(educ1 = recode(educ1, '1' = "no high school",
                        '2' = "high school graduate",'3' = "some college",
                        '4' = "college graduate")) %>%
  mutate(partyid3 = recode(partyid3, '1' = "democrats", '2' = "independents",
                           '3' = "republicans", '9' = "apolitical")) %>%
  mutate(ideo = recode(ideo, '1' = "liberal", '3' = "moderate",
                       '5' = "conservative"))
```

12. Fit a logistic regression model predicting support for Bush given the variables above and income as predictors and also consider interactions among the predictors. You do not need to consider all possible interactions nor should you use automatic methods for model selection at this point, but suggest a couple from the predictors above that might make sense intuitively.

We considered 3 2-way interaction terms for predicting support for Bush.

The first one is `factor(income)*factor(educ1)`, because we believe that education has effects on income, so these two variables are likely to be correlated.

The second one is `factor(partyid3)*factor(ideo)`, since we think that party identification and political ideology are all about political preferences, so it is possible for these to be correlated.

The third one is `factor(gender)*factor(race)`, sometimes we hear that the political preferences between white women and black women are quite different, so there may be a correlation between these two variables.

```
glm_fit.4 <- glm(vote ~ factor(income)*factor(educ1) +
                   factor(partyid3)*factor(ideo) +
                   factor(gender)*factor(race),
```

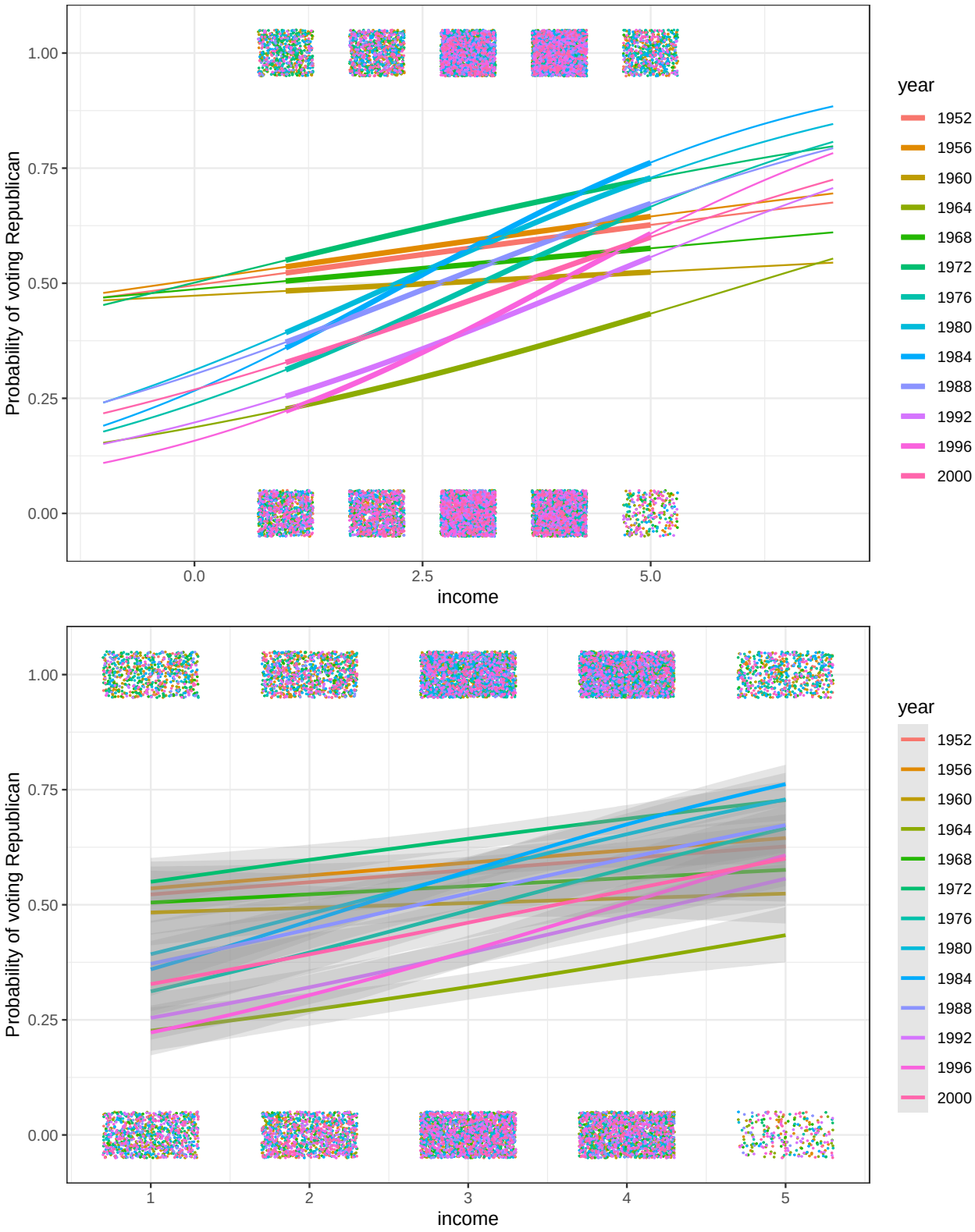

Figure 4: Logistic regression estimating the probability of supporting George Bush from 1952 to 2000, as a function of discretized income level. (a) Fitted logistic regression: the thick line indicates thecurve in the range of the data; the thinner lines at the end show how the logistic curve approaches 0 and 1 inthe limits. (b) In the range of the data, the solid line shows the best-fit logistic regression, and the light linesshow uncertainty in the fit.

```
                data = newnes_1992,
                family = binomial(link = "logit"))
kable(summary(glm_fit.4)$coef,digits=4,
      caption = "Summary for glm(vote ~ factor(income)*factor(educ1)
      +factor(partyid3)*factor(ideo) +factor(gender)*factor(race))")
```

Table 8: Summary for glm(vote ~ factor(income)*factor(educ1) +factor(partyid3)*factor(ideo) +factor(gender)*factor(race))

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	8.5188	2130.6096	0.0040	0.9968
factor(income)2	1.6811	2.0446	0.8222	0.4110
factor(income)3	1.5045	1.9549	0.7696	0.4415
factor(income)4	1.3244	1.9377	0.6835	0.4943
factor(income)5	0.4793	1.9694	0.2434	0.8077
factor(educ1)high school graduate	0.4584	1.9713	0.2325	0.8161
factor(educ1)no high school	1.2243	2.0395	0.6003	0.5483
factor(educ1)some college	-0.2210	2.1230	-0.1041	0.9171
factor(partyid3)democrats	4.0673	2058.2431	0.0020	0.9984
factor(partyid3)independents	5.6329	2058.2431	0.0027	0.9978
factor(partyid3)republicans	7.9035	2058.2431	0.0038	0.9969
factor(ideo)liberal	-1.2115	0.4212	-2.8764	0.0040
factor(ideo)moderate	-1.1739	0.5169	-2.2709	0.0232
factor(gender)male	-16.0400	550.5721	-0.0291	0.9768
factor(race)black	-16.8769	550.5717	-0.0307	0.9755
factor(race)hispanic	-13.0938	550.5716	-0.0238	0.9810
factor(race)native american	-13.9944	550.5717	-0.0254	0.9797
factor(race)white	-14.7335	550.5714	-0.0268	0.9787
factor(income)2:factor(educ1)high school graduate	-1.3228	2.1179	-0.6246	0.5323
factor(income)3:factor(educ1)high school graduate	-1.1432	2.0259	-0.5643	0.5726
factor(income)4:factor(educ1)high school graduate	-1.0783	2.0155	-0.5350	0.5927
factor(income)5:factor(educ1)high school graduate	-0.9527	2.1913	-0.4347	0.6638
factor(income)2:factor(educ1)no high school	-2.7511	2.2965	-1.1979	0.2309
factor(income)3:factor(educ1)no high school	-1.2732	2.2764	-0.5593	0.5760
factor(income)4:factor(educ1)no high school	-11.6387	1455.3991	-0.0080	0.9936
factor(income)2:factor(educ1)some college	-0.3074	2.2969	-0.1338	0.8935
factor(income)3:factor(educ1)some college	-0.2999	2.1865	-0.1372	0.8909
factor(income)4:factor(educ1)some college	-0.5002	2.1739	-0.2301	0.8180
factor(income)5:factor(educ1)some college	1.5820	2.4563	0.6440	0.5195
factor(partyid3)democrats:factor(ideo)liberal	-1.2202	0.5832	-2.0925	0.0364
factor(partyid3)independents:factor(ideo)liberal	-0.5574	0.8436	-0.6607	0.5088
factor(partyid3)democrats:factor(ideo)moderate	0.3541	0.7821	0.4528	0.6507
factor(partyid3)independents:factor(ideo)moderate	0.5957	0.8978	0.6635	0.5070
factor(gender)male:factor(race)black	16.2865	550.5728	0.0296	0.9764
factor(gender)male:factor(race)hispanic	12.7879	550.5730	0.0232	0.9815
factor(gender)male:factor(race)native american	15.0940	550.5738	0.0274	0.9781
factor(gender)male:factor(race)white	15.7786	550.5721	0.0287	0.9771

```
kable(anova(glm_fit.4, test = "Chisq"), caption = "Analysis of deviance for the model", digits = 4)
```

Table 9: Analysis of deviance for the model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	1132	1534.1028	NA
factor(income)	4	33.8670	1128	1500.2358	0.0000
factor(educ1)	3	2.3967	1125	1497.8391	0.4942
factor(partyid3)	3	756.4597	1122	741.3794	0.0000
factor(ideo)	2	54.4580	1120	686.9215	0.0000
factor(gender)	1	4.8366	1119	682.0848	0.0279
factor(race)	4	27.7674	1115	654.3175	0.0000
factor(income):factor(educ1)	11	6.5897	1104	647.7277	0.8313
factor(partyid3):factor(ideo)	4	5.8748	1100	641.8529	0.2087
factor(gender):factor(race)	4	13.8270	1096	628.0259	0.0079

Since Anova table indicates that 2-way interaction terms of income with educ1, partyid3 with ideo are insignificant, we decide to refit the model excluding those terms.

```
glm_fit.5 <- glm(vote ~ factor(income) + factor(educ1) + factor(partyid3) +
                   factor(ideo) + factor(gender)*factor(race),
                 data = newnes_1992, family = binomial(link = "logit"))

kable(anova(glm_fit.5, test = "Chisq"), caption = "Analysis of deviance for the model",digits=4)
```

Table 10: Analysis of deviance for the model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	1132	1534.1028	NA
factor(income)	4	33.8670	1128	1500.2358	0.0000
factor(educ1)	3	2.3967	1125	1497.8391	0.4942
factor(partyid3)	3	756.4597	1122	741.3794	0.0000
factor(ideo)	2	54.4580	1120	686.9215	0.0000
factor(gender)	1	4.8366	1119	682.0848	0.0279
factor(race)	4	27.7674	1115	654.3175	0.0000
factor(gender):factor(race)	4	14.2455	1111	640.0720	0.0066

Then we use ANOVA again and found that we need to remove "educ1" as well, because its p-value is too large and `glm_fit.6` is our final model, which are shown in the chunk below.

```
glm_fit.6 <- glm(vote ~ factor(income) + factor(partyid3) + factor(ideo) +
                   factor(gender)*factor(race), data = newnes_1992,
                 family = binomial(link = "logit"))

kable(anova(glm_fit.6, test = "Chisq"), caption = "Analysis of deviance for the model",digits=4)
```

Table 11: Analysis of deviance for the model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	1132	1534.1028	NA
factor(income)	4	33.8670	1128	1500.2358	0.0000
factor(partyid3)	3	757.2539	1125	742.9819	0.0000
factor(ideo)	2	53.4382	1123	689.5437	0.0000
factor(gender)	1	3.8615	1122	685.6823	0.0494

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
factor(race)	4	27.0289	1118	658.6534	0.0000
factor(gender):factor(race)	4	14.3495	1114	644.3039	0.0063

Now we have model which include only significant variables.

13. Plot binned residuals using the function `binnedplot` from package `arm` versus some of the additional predictors in the 1992 dataframe. Are there any suggestions that the mean or distribution of residuals is different across the levels of the other predictors and that these predictors should be added to the model? (Provide plots and any other summaries to explain).

Among candidate variables, we test 9 variables, `age`, `urban`, `occup1`, `religion`, `martial-status`, `dlikes`, `presvote`, `presapprove`, and `perfin`

To minimize distortion from removing NA's in each variables, we make datasets which only remove NAs in each variables.

```
newnes_1992_occup <- newnes_1992  %>%
  filter(!is.na(occup1))
newnes_1992_religion <- newnes_1992  %>%
  filter(!is.na(religion))
newnes_1992_martial <- newnes_1992  %>%
  filter(!is.na(martial_status))
newnes_1992_presvote <- newnes_1992  %>%
  filter(!is.na(presvote_intent))
newnes_1992_presapprov <- newnes_1992  %>%
  filter(!is.na(presapprov))
newnes_1992_perfin <- newnes_1992  %>%
  filter(!is.na(perfin))
```

```
par(mfrow = c(3,3))
binnedplot(newnes_1992$age,glm_fit.6$residuals, main = "age")
binnedplot(newnes_1992$urban, glm_fit.6$residuals, main = "urban")
binnedplot(newnes_1992_occup$occup1, glm_fit.6$residuals, main = "occup1")
binnedplot(newnes_1992_religion$religion, glm_fit.6$residuals, main = "religion")
binnedplot(newnes_1992_martial$martial_status, glm_fit.6$residuals, main = "martial")
binnedplot(newnes_1992$dlikes, glm_fit.6$residuals, main = "dlikes")
binnedplot(newnes_1992_presvote$presvote_intent, glm_fit.6$residuals, main = "presvote_intent")
binnedplot(newnes_1992_presapprov$presapprov, glm_fit.6$residuals, main = "presapprove")
binnedplot(newnes_1992_perfin$perfin, glm_fit.6$residuals, main = "perfin")
```

From these 9 binnedplots above, we can add `religion`,`perfin` and `presapprov` as our additional predictors because the binnedplots of these three variables show linear relationships between residuals and expected values of these predictors. It means that average residual of our previous model increase or decrease at different level of predictor variable. By adding new variables, we expect improvement in our model.

```
newnes_1992_temp <- newnes_1992 %>%
  filter(!is.na(religion)) %>%
  filter(!is.na(perfin)) %>%
  filter(!is.na(presapprov))

glm_fit.8 <- glm(vote ~ factor(income) + factor(partyid3) + factor(ideo)
                 +factor(gender)*factor(race) + factor(religion) + factor(perfin)
                 +factor(presapprov), data = newnes_1992_temp, family = binomial(link = "logit"))
kable(anova(glm_fit.8, test = "Chisq"),digits=4,caption="Analysis of deviance for the model")
```

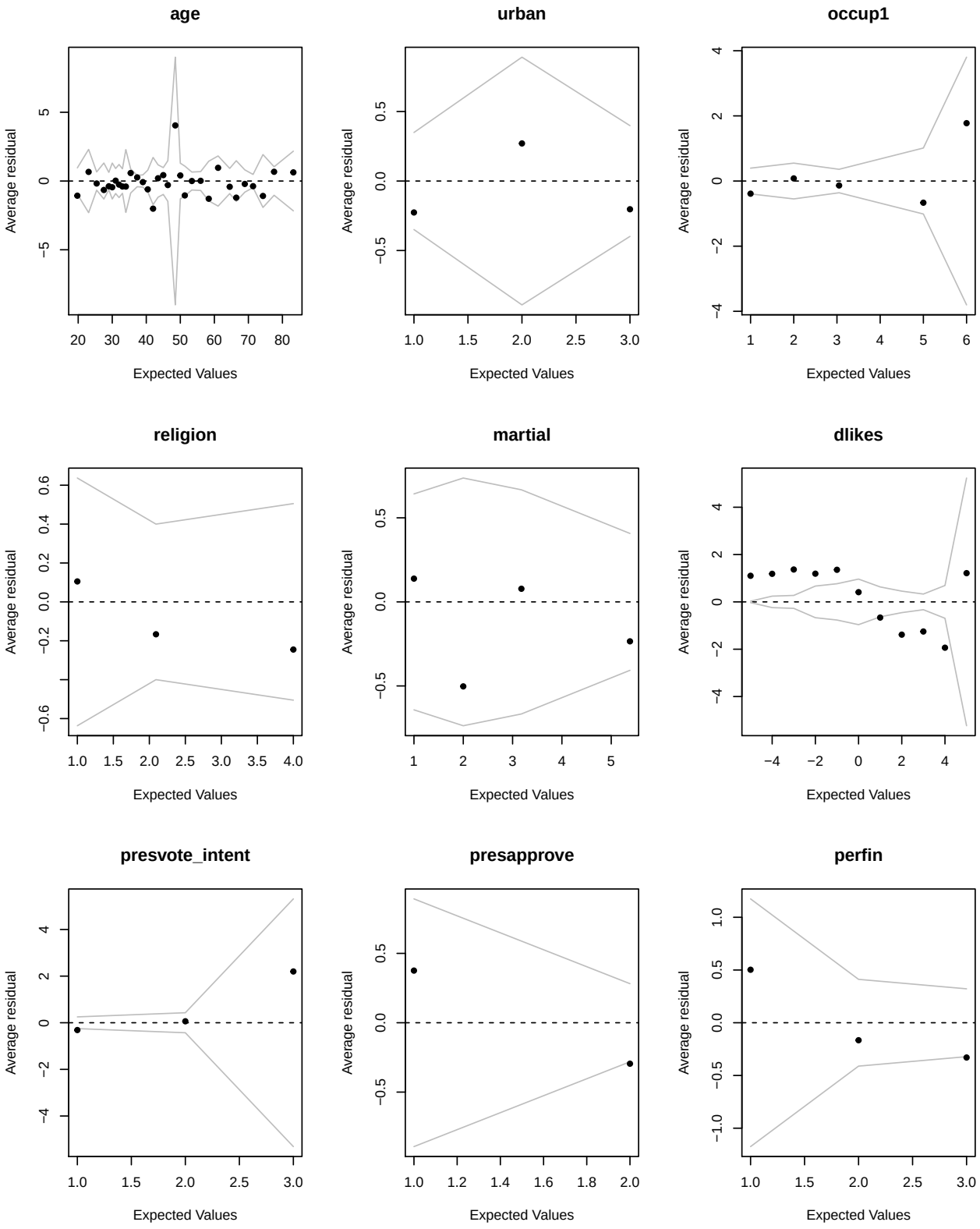

Figure 5: binned plots for candidate variables. Some variables show trends of increasing or decreasing in average residuals corresponding to levels of predictor variables

Table 12: Analysis of deviance for the model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	1115	1511.7888	NA
factor(income)	4	33.1652	1111	1478.6236	0.0000
factor(partyid3)	3	755.3895	1108	723.2341	0.0000
factor(ideo)	2	52.1233	1106	671.1107	0.0000
factor(gender)	1	4.7235	1105	666.3873	0.0298
factor(race)	4	24.8369	1101	641.5504	0.0001
factor(religion)	3	8.6139	1098	632.9365	0.0349
factor(perfin)	2	15.9171	1096	617.0195	0.0003
factor(presapprov)	1	132.7004	1095	484.3190	0.0000
factor(gender):factor(race)	4	15.4224	1091	468.8967	0.0039

Then we used ANOVA to check our new model and found these predictors are all signifiant. It means that by adding new variables, we significantly improve our model.

14. Evaluate and compare the different models you fit. Consider coefficient estimates (are they stable across models) and standard errors (any indications of identifiability problems), residual plots and deviances.

```
newnes_1992_fin <- newnes_1992 %>%
  filter(!is.na(religion)) %>%
  filter(!is.na(perfin)) %>%
  filter(!is.na(presapprov)) %>%
  mutate(income = as.factor(income)) %>%
  mutate(partyid3 = as.factor(partyid3)) %>%
  mutate(ideo = as.factor(ideo)) %>%
  mutate(gender = as.factor(gender)) %>%
  mutate(race = as.factor(race)) %>%
  mutate(religion = as.factor(religion)) %>%
  mutate(perfin = as.factor(perfin)) %>%
  mutate(presapprov = as.factor(presapprov))
```

```
glm_fit.6 <- glm(vote ~ income + partyid3 + ideo + gender*race,
                 data = newnes_1992_fin, family = binomial(link = "logit"))
glm_fit.8 <- glm(vote ~ income + partyid3 + ideo + gender*race + religion + perfin +
                   presapprov, data = newnes_1992_fin, family = binomial(link = "logit"))
```

```
old <- glm_fit.6$coefficients
new <- glm_fit.8$coefficients
old_ci <- data.frame(cbind(old,confint(glm_fit.6)),model = "old", names = names(old))
new_ci <- data.frame(cbind(new,confint(glm_fit.8)),model = "new", names = names(new))
colnames(old_ci) <- c("coefficient","2.5%","97.5%","model","name")
colnames(new_ci) <- c("coefficient","2.5%","97.5%","model","name")
kable(old_ci,digits = 4,
      caption = "coefficient of variables and their confidence interval table for old model")
```

Table 13: coefficient of variables and their confidence interval table for old model

	coefficient	2.5%	97.5%	model	name
(Intercept)	-0.3931	-24.9243	505.0672	old	(Intercept)
income2	0.4741	-0.3478	1.3169	old	income2
income3	0.5739	-0.1908	1.3673	old	income3

	coefficient	2.5%	97.5%	model	name
income4	0.3489	-0.4236	1.1424	old	income4
income5	-0.0253	-1.0568	1.0336	old	income5
partyid3democrats	13.7906	-282.7439	NA	old	partyid3democrats
partyid3independents	15.5990	-280.8801	NA	old	partyid3independents
partyid3republicans	17.9448	-278.6158	NA	old	partyid3republicans
ideoliberal	-1.7560	-2.2478	-1.2780	old	ideoliberal
ideomoderate	-0.8840	-1.6077	-0.1679	old	ideomoderate
gendermale	-16.3442	NA	49.4096	old	gendermale
raceblack	-17.2187	NA	49.6032	old	raceblack
racehispanic	-13.5004	NA	53.3124	old	racehispanic
racenative american	-14.5475	NA	52.1034	old	racenative american
racewhite	-15.2435	NA	51.9801	old	racewhite
gendermale:raceblack	16.5415	-23.3273	NA	old	gendermale:raceblack
gendermale:racehispanic	13.0491	-5.6012	241.7699	old	gendermale:racehispanic
gendermale:racenative american	15.3544	-39.8709	NA	old	gendermale:racenative american
gendermale:racewhite	16.0659	-49.5574	NA	old	gendermale:racewhite

```
kable(new_ci,digits = 4,
      caption = "coefficient of variables and their confidence interval table for new model")
```

Table 14: coefficient of variables and their confidence interval table for new model

	coefficient	2.5%	97.5%	model	name
(Intercept)	-0.3120	-186.0756	726.5884	new	(Intercept)
income2	0.7928	-0.1916	1.7973	new	income2
income3	0.6410	-0.2874	1.5907	new	income3
income4	0.5568	-0.3669	1.4946	new	income4
income5	0.1411	-1.0997	1.4038	new	income5
partyid3democrats	15.6889	-280.7628	NA	new	partyid3democrats
partyid3independents	18.0036	-278.3342	NA	new	partyid3independents
partyid3republicans	19.1543	-277.3200	NA	new	partyid3republicans
ideoliberal	-1.5506	-2.1408	-0.9803	new	ideoliberal
ideomoderate	-0.9850	-1.8719	-0.1119	new	ideomoderate
gendermale	-15.8716	NA	48.4169	new	gendermale
raceblack	-17.5876	NA	47.9404	new	raceblack
racehispanic	-13.5132	NA	52.0417	new	racehispanic
racenative american	-14.8666	NA	50.3559	new	racenative american
racewhite	-15.5651	NA	50.5046	new	racewhite
religion2	-0.0630	-0.6382	0.5156	new	religion2
religion3	-2.2155	-4.3259	0.0893	new	religion3
religion4	-0.9368	-1.8803	-0.0175	new	religion4
perfin2	-0.0203	-0.6213	0.5830	new	perfin2
perfin3	-0.4144	-1.0505	0.2214	new	perfin3
presapprov2	-2.8324	-3.3601	-2.3308	new	presapprov2
gendermale:raceblack	16.1274	-43.4028	NA	new	gendermale:raceblack
gendermale:racehispanic	11.6366	-53.2405	NA	new	gendermale:racehispanic
gendermale:racenative american	14.4587	-50.6205	NA	new	gendermale:racenative american
gendermale:racewhite	15.6460	-48.4799	NA	new	gendermale:racewhite

```
coeff_table <- rbind(old_ci,new_ci)
coeff_table <- na.omit(coeff_table)
ggplot(data = coeff_table, mapping = aes(x=model, y=as.numeric(coefficient)),
       labs(x="old & new models",y="coefficients")) +
  geom_point() +
  geom_errorbar(aes(ymin = coeff_table[,2],ymax = coeff_table[,3])) +
  facet_wrap(~name,scales = "free")
```

As we can see in tables and plots above, some variables' coefficients are showing instability because they have too large standard error compared to its coefficient. For instance, CI for partyid3, race, gender and their interaction terms does not converge and show too large error compared to their estimated values. Some variables even change their sign after new variables are added in model.

```
index_old <- apply(old_ci,1,function(x)sum(is.na(x)))>0
indicator_old <- rownames(old_ci)[index_old]
index_new <- apply(new_ci,1,function(x)sum(is.na(x)))>0
indicator_new <- rownames(new_ci)[index_new]
indicator <- list(old = indicator_old, new = indicator_new)
indicator
```

```
## $old
##  [1] "partyid3democrats"              "partyid3independents"
##  [3] "partyid3republicans"            "gendermale"
##  [5] "raceblack"                      "racehispanic"
##  [7] "racenative american"            "racewhite"
##  [9] "gendermale:raceblack"           "gendermale:racenative american"
## [11] "gendermale:racewhite"
##
## $new
##  [1] "partyid3democrats"              "partyid3independents"
##  [3] "partyid3republicans"            "gendermale"
##  [5] "raceblack"                      "racehispanic"
##  [7] "racenative american"            "racewhite"
##  [9] "gendermale:raceblack"           "gendermale:racehispanic"
## [11] "gendermale:racenative american" "gendermale:racewhite"
```

Above variables are showing the problem of identifiability, because their confidence intervaldoes not converge and are showing NA. We cannot estimate it.

```
par(mfrow=c(2,1))
binnedplot(glm_fit.6$fitted.values,glm_fit.6$residuals, main = "Old Model Binned plot")
binnedplot(glm_fit.8$fitted.values,glm_fit.8$residuals, main = "New Model Binned plot")
```

From our above binned plots, we couldn't see any severe violation of our model. So it is a good fit.

```
kable(anova(glm_fit.6,glm_fit.8,test = "Chisq"), digit = 4,
      caption = "table shows we need to choose new model")
```

Table 15: table shows we need to choose new model

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1097	627.2994	NA	NA	NA
1091	468.8967	6	158.4027	0

Through Anova table, our new model which added new variables is significantly better than old model.

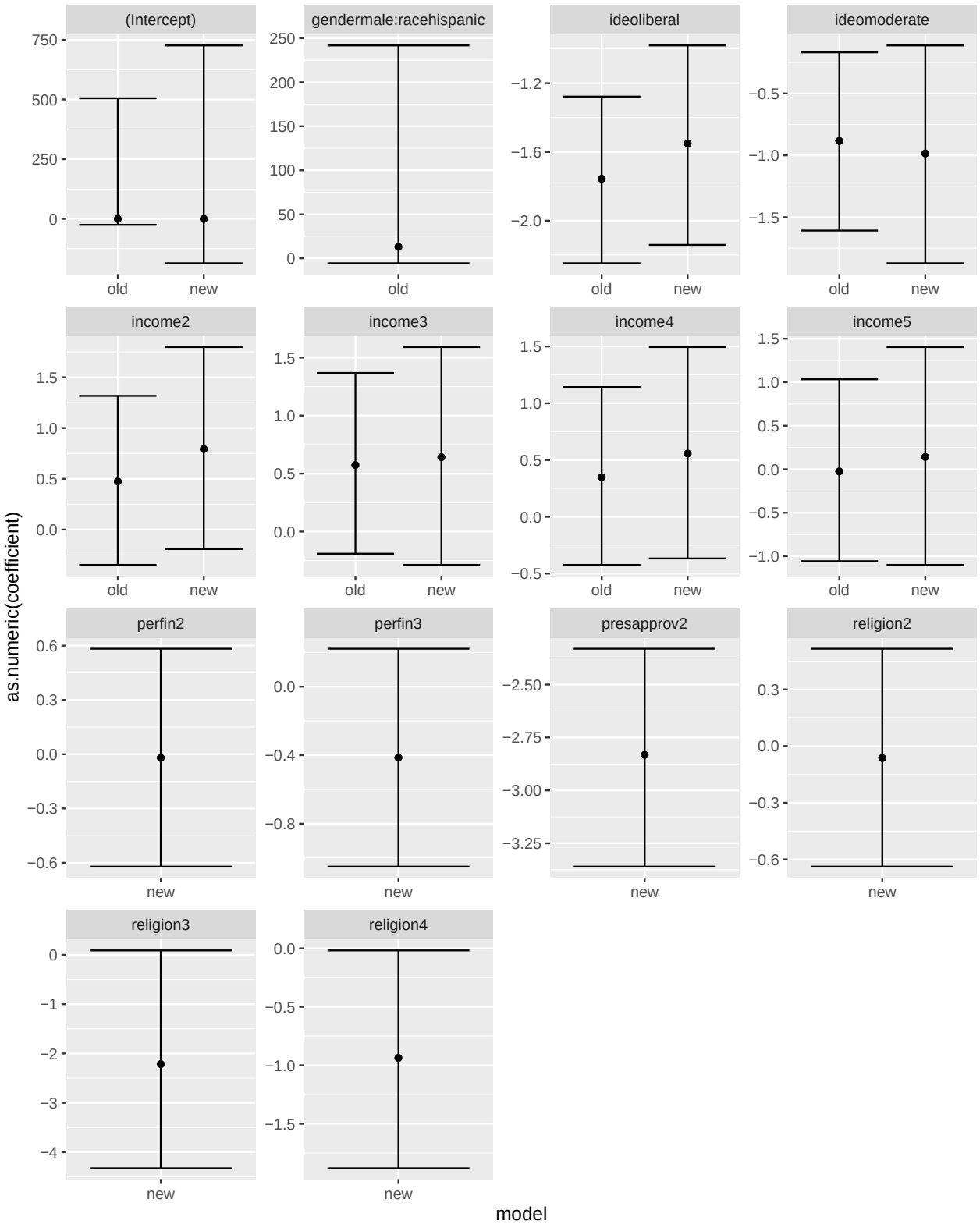

Figure 6: Coefficient comparison plot for each variables. It shows both point and interval estimates for variables.

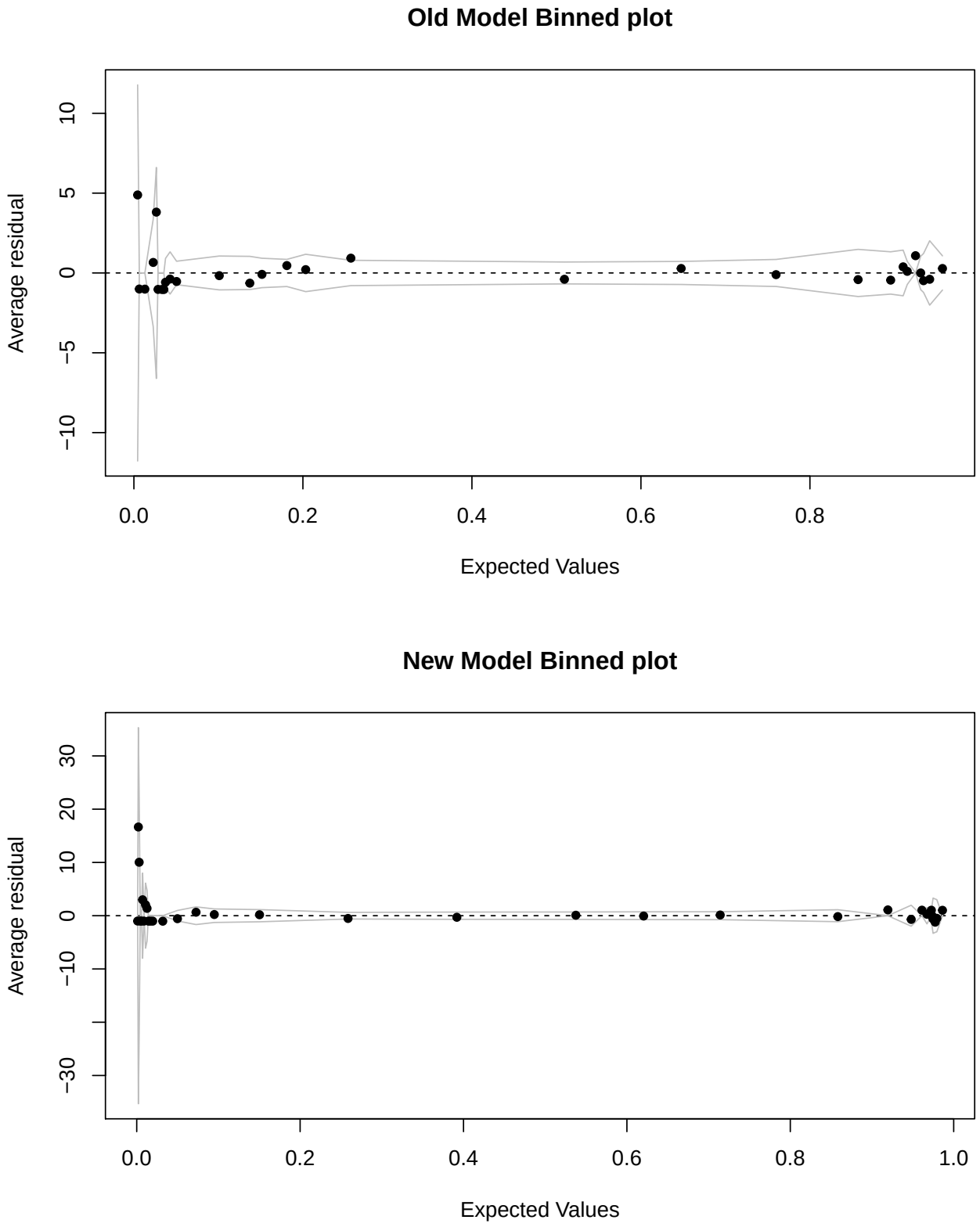

Figure 7: avergae residual of model versus their fitted value which shows how well model is fitted.

Because we check that it has much smaller deviance by doing chisq test.

15. Compute the error rate of your model (see GH page 99) and compare it to the error rate of the null model. We can define a function for the error rate as:

```
error.rate = function(pred, true) {
  mean((pred > .5 & true == 0) | (pred < .5 & true == 1))
}
glm_null <- glm(vote ~ 1, data = newnes_1992_fin, family = binomial(link = "logit"))
glm_fin <- glm(vote ~ income + partyid3 + ideo + gender*race + religion + perfin + presapprov, data = ne
null.pred <- predict(glm_null)
fin.pred <- predict(glm_fin)
null.error <- error.rate(null.pred, newnes_1992_fin$vote)
fin.error <- error.rate(fin.pred, newnes_1992_fin$vote)
error = data.frame(null = null.error,fin = fin.error)
kable(error,digits = 3, caption = "error rate reduce dramatically when we use fitted value for our model
```

Table 16: error rate reduce dramatically when we use fitted value for our model as prediction

null	fin
0.411	0.086

Our model's prediction is much more precise than null model.

16. Provide one to two paragraphs summarizing your findings. Provide a neatly formatted table of odds ratios and 95% confidence intervals for each predictor, and in the text interprete key coefficients (providing ranges of supporting values from above) in terms of the odds of voting for Bush. Comment on which variables had missing data, suggesting possible reasons why they may be missing and discuss whether you think that this may impact your analysis. Discuss any limitations of the models and its generalizability. Attempt to write this at a level that readers of the New York Times Upshot column could understand.

Our model evaluated the factors that influnce people voting for Republican in 1992 presidential election. We examined income, political ideology, gender, race, party choice, religion, perfin and presapprov as well as some potential interaction between these variables. We found that political party was significantly associated with a Republican vote:republicans are more likely to vote for Bush than other parties. i.e. odds of voting for Bush is larger than other party. In addition, political ideology was also assiciated with voting. Odds of voting for Bush in Moderated was 1 to 2.4 times larger than odds of voting for Bush in Liberals. When political party affiliation and ideology are held constant, we found religion is another factor to influence people's voting. Odds of voting for Bush in people from Religion2 is 9 times larger than people from Religion3 and 3 three times larger than people from Religion4. Apart from that, we also found black and white male like Bush better than native american and hispanic male. Also, people with less income but not the least income have the highest possibility to vote for Republican, which is two times more than the richest people. Another thing needs to be mentioned is that there are some variables in our model which cannot be estimated because of the problem called identifiability which occurred from large changing rate and that is one of the limitation of our model.

Most missing data occured in variables related to political parties and ideology, for example, parents' parties and ideology feel. We also have two columns of NA under the variable regis and icpsr_cty, because people didn't register and didn't agree with enthics were not eligible to vote. As for other variables with missing data, it might because data were collected based on a survey about people's political reference and other personal information. Some people were not sure which ideology or political parties were applied to them and some people didn't know their parents' parties, so they skipped these questions. If these were the cases, these missing data might cause a lower level of political knowledge and biased our analysis. Because when a

voter is not sure about his/her political preference, he/she has higher leverege on choosing Republican than other voters. This might influence our model's generalizability as well. Therfore, we may want to add a another variable indicating voter's political knowledge or sensitivity. Because this report are prepared for all US citizens and people with less political knowledge just a subset of it, which can't represent all Americans.

```
library(knitr)
options(digits = 4)
odds.ratio <- round(exp(glm_fit.8$coefficients),4)
confit<- round(exp(confint(glm_fit.8)),4)
df<-data.frame(odds.ratio,confit)
kable(df,col.names = c("odds ratio", "2.5%", "97.5%"),
      caption = "Odds ratio and confidence interval for each predictor")
```

Table 17: Odds ratio and confidence interval for each predictor

	odds ratio	2.5%	97.5%
(Intercept)	7.320e-01	0.0000	Inf
income2	2.209e+00	0.8256	6.033e+00
income3	1.899e+00	0.7502	4.907e+00
income4	1.745e+00	0.6929	4.458e+00
income5	1.151e+00	0.3330	4.071e+00
partyid3democrats	6.510e+06	0.0000	NA
partyid3independents	6.590e+07	0.0000	NA
partyid3republicans	2.083e+08	0.0000	NA
ideoliberal	2.121e-01	0.1176	3.752e-01
ideomoderate	3.734e-01	0.1538	8.941e-01
gendermale	0.000e+00	NA	1.065e+21
raceblack	0.000e+00	NA	6.611e+20
racehispanic	0.000e+00	NA	3.994e+22
racenative american	0.000e+00	NA	7.401e+21
racewhite	0.000e+00	NA	8.588e+21
religion2	9.389e-01	0.5282	1.675e+00
religion3	1.091e-01	0.0132	1.093e+00
religion4	3.919e-01	0.1525	9.826e-01
perfin2	9.799e-01	0.5372	1.791e+00
perfin3	6.607e-01	0.3497	1.248e+00
presapprov2	5.890e-02	0.0347	9.720e-02
gendermale:raceblack	1.009e+07	0.0000	NA
gendermale:racehispanic	1.132e+05	0.0000	NA
gendermale:racenative american	1.903e+06	0.0000	NA
gendermale:racewhite	6.237e+06	0.0000	NA