Problem Set 5

Jingyuan Zhou

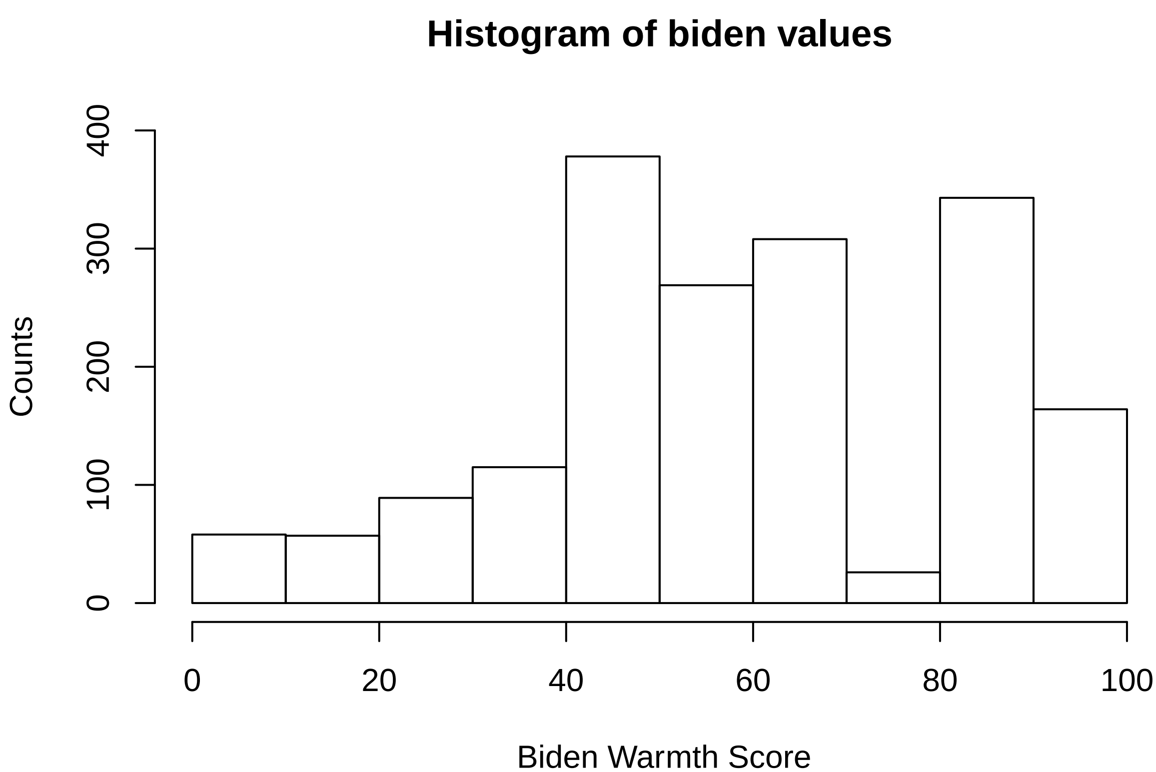
2/9/2017

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(modelr)
library(broom)

options(na.action = na.warn)
set.seed(1234)

theme_set(theme_minimal())
```

Describe the data

According to the histogram shown below, very few people have bad feelings towards Biden, most people are neutral and some have very positive feelings towards him. This is shown by the break for scores between neutral values and extremly high values.

```
data <- read.csv(file="biden.csv",head=TRUE)
hist(data$biden,main="Histogram of biden values",xlab=" Biden Warmth Score",
     ylab = 'Counts', ylim = c(0, 400))
```



Simple linear regression

Parameters and standard errors are shown below.

```
slr_mod <- lm(biden ~ age, data = data)
tidy(slr_mod)
```

```
##          term    estimate  std.error statistic       p.value
## 1 (Intercept) 59.19736008 1.64791889 35.922496 1.145056e-213
## 2         age  0.06240535 0.03266815  1.910281  5.625534e-02
```

```
glance(slr_mod)
```

```
##     r.squared adj.r.squared    sigma statistic    p.value df    logLik
## 1 0.002017624   0.001464725 23.44485  3.649174 0.05625534  2 -8263.475
##        AIC      BIC deviance df.residual
## 1 16532.95 16549.45 992137.7        1805
```

1.2.According to the summary, there is a relationship between the predictor and the response, but the relationship is very weak because age has a p-value of 5.625534e-02, which is larger than 0.025. Thus, it's not statistically significant at 95% significance level.

3.The relationship is positive since the coefficient of "age" is 0.06240535, which is positive.

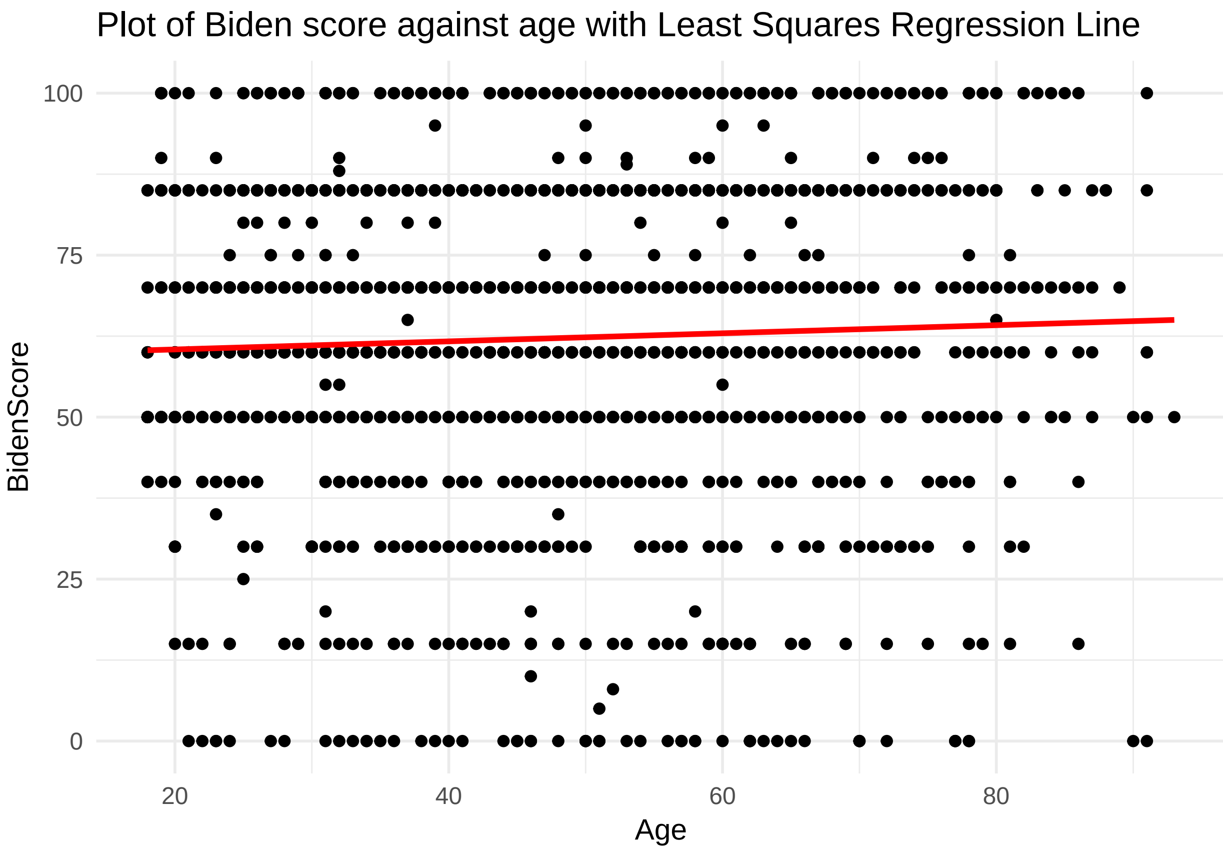
4.R^2 of this model is only 0.002017624. This means that only 0.2% of the variation of biden is explained by age. This shows that it is a really bad model.

```
pred_aug <- augment(slr_mod, newdata = data.frame(age = c(45)))
(pred_ci <- mutate(pred_aug,
                   ymin = .fitted - .se.fit * 1.96,
                   ymax = .fitted + .se.fit * 1.96))
```

```
##   age .fitted   .se.fit     ymin     ymax
## 1  45 62.0056 0.5577123 60.91248 63.09872
```

5.The predicted biden score with an age of 45 is 62.0056. The associated 95% confidence interval is (60.91248, 63.09872).

```
grid <- data %>%
  data_grid(age)
grid <- grid %>%
  add_predictions(slr_mod)

#plot
ggplot(data, aes(age)) +
  geom_point(aes(y = biden)) +
  geom_line(aes(y = pred), data = grid, color = "red", size = 1) +
  labs(title = 'Plot of Biden score against age with Least Squares Regression Line',
        x = 'Age',y = 'BidenScore')
```



Multiple linear regression

```
mlr_mod <- lm(biden ~ age + female + educ, data = data)
tidy(mlr_mod)
```

```
##          term    estimate  std.error statistic      p.value
## 1 (Intercept) 68.62101396 3.59600465 19.082571 4.337464e-74
## 2         age  0.04187919 0.03248579  1.289154 1.975099e-01
## 3      female  6.19606946 1.09669702  5.649755 1.863612e-08
## 4        educ -0.88871263 0.22469183 -3.955251 7.941295e-05
```

```
glance(mlr_mod)
```

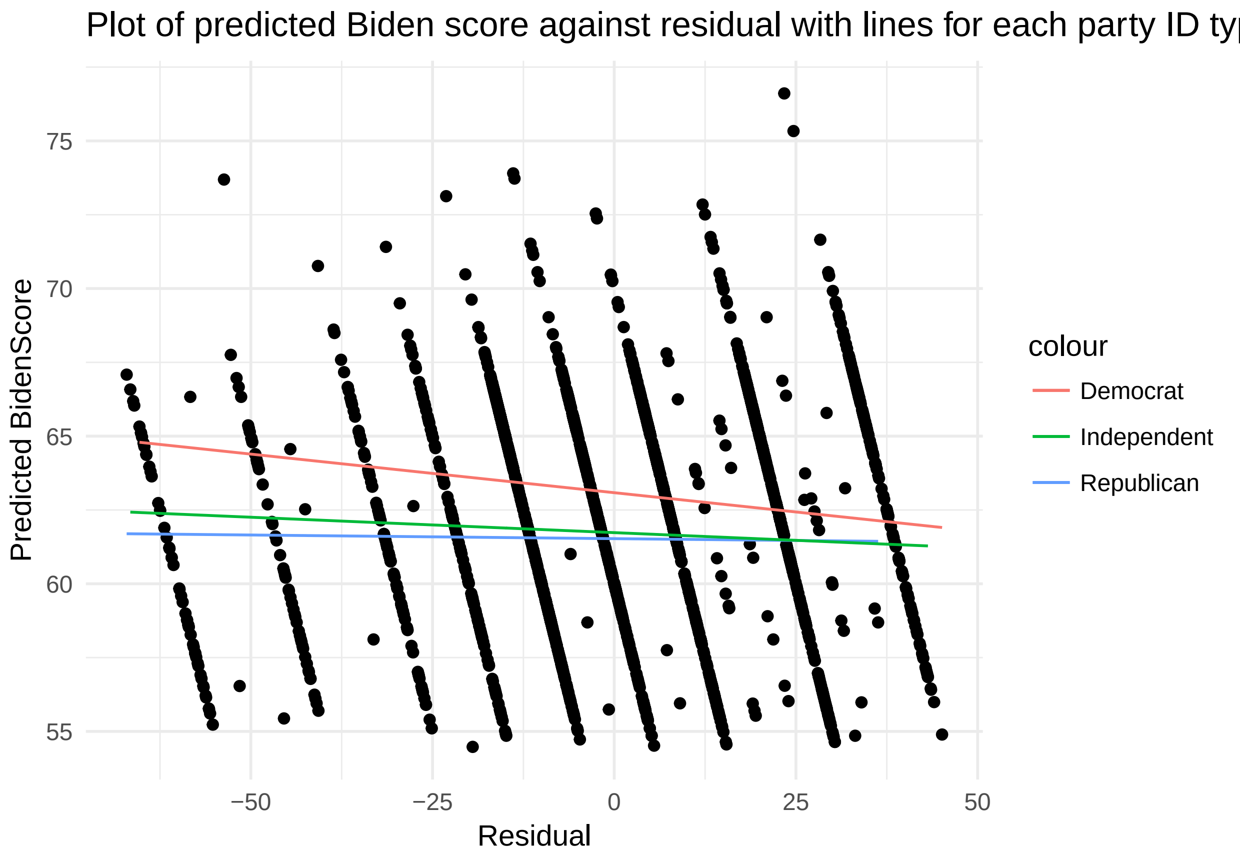
```
##    r.squared adj.r.squared    sigma statistic      p.value df    logLik
## 1 0.02722727    0.02560868 23.15967  16.82159 8.876492e-11  4 -8240.359
##        AIC      BIC deviance df.residual
## 1 16490.72 16518.22 967075.7        1803
```

1.'Age' is not a statistically significant predictor at 95% significance level because its p-value is 0.198, which is larger than 0.025; both 'female' and 'educ' are statistically significant predictors at 95% significance level because their p-values, 1.863612e-08 and 7.941295e-05, are both smaller than 0.025.

2.The estimated parameter of 'female' is 6.19606946. It shows that every unit increase of "female" will on avaerage increase 6.20 units of biden score. In other words, when two people have same age and same years of education, a female will on average give 6.196 scores higher than a male.

3.Adjusted R^2 of this model is 0.02560868. This shows that around 2.56% of variation in biden score is explained by age, gender, and education. It is a better model compared to the previouse one because its R^2 value is larger than that of the previous model.

```
data_d <- subset(data, dem ==1)
res <- add_residuals(data_d, mlr_mod)['resid']
grid_d <- add_predictions(data_d, mlr_mod)
grid_d['resid']<- res
d_m <- lm(pred ~ resid, data = grid_d)
grid_d <- add_predictions(grid_d, d_m)

data_r <- subset(data, rep ==1)
res <- add_residuals(data_r, mlr_mod)['resid']
grid_r <- add_predictions(data_r, mlr_mod)
grid_r['resid']<- res
r_m <- lm(pred ~ resid, data = grid_r)
grid_r <- add_predictions(grid_r, r_m)

data_i <- subset(data, (!dem==1)&(!rep==1))
res <- add_residuals(data_i, mlr_mod)['resid']
grid_i <- add_predictions(data_i, mlr_mod)
grid_i['resid']<- res
i_m <- lm(pred ~ resid, data = grid_i)
grid_i <- add_predictions(grid_i, i_m)

res <- add_residuals(data, mlr_mod)['resid']
grid <- add_predictions(data, mlr_mod)
grid['resid']<- res

# plot
ggplot(grid, aes(resid)) +
  geom_point(aes(y = pred))+
  geom_line(aes(y= pred, color = 'Democrat'), data = grid_d) +
  geom_line(aes(y=pred, color = 'Republican'), data = grid_r) +
  geom_line(aes(y=pred, color = 'Independent'), data = grid_i) +
  labs(title = 'Plot of predicted Biden score against residual with lines for each party ID type',
       x = 'Residual',y = 'Predicted BidenScore')
```



4. These is a problem for this model because when we separate people according to their party IDs, we can observe from the plot that the three smooth fit lines have different slopes. This suggests that changes in residual could affect the predicted Biden score of people with different party IDs differently.

Multiple linear regression model (with even more variables!)

```
mlr2_mod <- lm(biden ~ age + female + educ + dem + rep, data = data)
tidy(mlr2_mod)
```

```
##          term     estimate std.error  statistic      p.value
## 1 (Intercept)  58.81125899 3.1244366  18.822996 2.694143e-72
## 2         age   0.04825892 0.0282474   1.708438 8.772744e-02
## 3      female   4.10323009 0.9482286   4.327258 1.592601e-05
## 4        educ  -0.34533479 0.1947796  -1.772952 7.640571e-02
## 5         dem  15.42425563 1.0680327  14.441745 8.144928e-45
## 6         rep -15.84950614 1.3113624 -12.086290 2.157309e-32
```

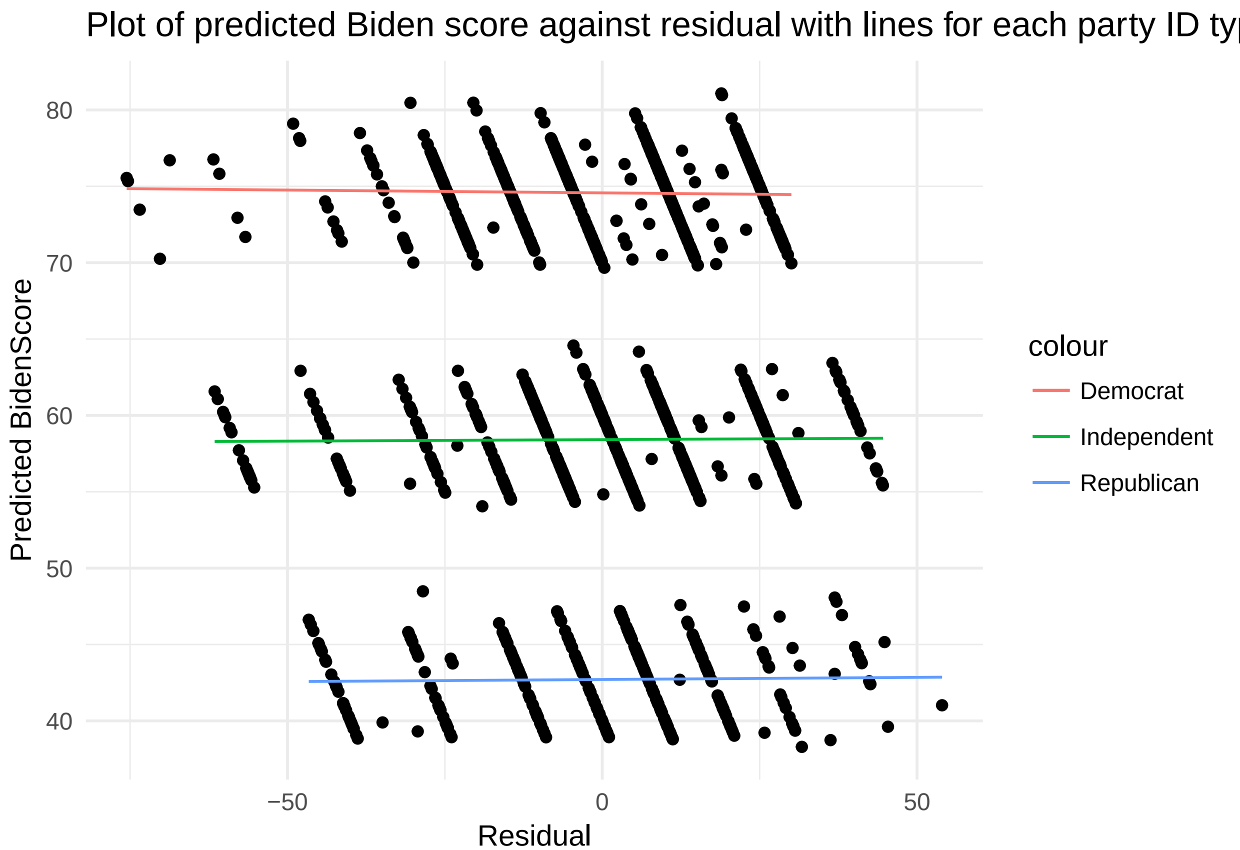
```
glance(mlr2_mod)
```

```
##   r.squared adj.r.squared    sigma statistic       p.value df    logLik
## 1 0.2815391     0.2795445 19.91449  141.1495 1.500182e-126  6 -7966.563
##        AIC      BIC deviance df.residual
## 1 15947.13 15985.62 714253.2        1801
```

1.Yes, the relationship between gender and Biden score changed. With this model, per unit increase in 'female' will on average increase Biden score with 4.10323009, which is smaller than the amount, 6.19606946, from last model.

2.Adjusted R^2 of this model is 0.2795445. This suggests that age, gender, education, and party identification explains 28.0% variability of the data. Since this number is larger than that of the previous model, it is a better model than age + gender + education model.

```
data_d <- subset(data, dem ==1)
res <- add_residuals(data_d, mlr2_mod)['resid']
grid_d <- add_predictions(data_d, mlr2_mod)
grid_d['resid']<- res
d_m <- lm(pred ~ resid, data = grid_d)
grid_d <- add_predictions(grid_d, d_m)

data_r <- subset(data, rep ==1)
res <- add_residuals(data_r, mlr2_mod)['resid']
grid_r <- add_predictions(data_r, mlr2_mod)
grid_r['resid']<- res
r_m <- lm(pred ~ resid, data = grid_r)
grid_r <- add_predictions(grid_r, r_m)

data_i <- subset(data, (!dem==1)&(!rep==1))
res <- add_residuals(data_i, mlr2_mod)['resid']
grid_i <- add_predictions(data_i, mlr2_mod)
grid_i['resid']<- res
i_m <- lm(pred ~ resid, data = grid_i)
grid_i <- add_predictions(grid_i, i_m)

res <- add_residuals(data, mlr2_mod)['resid']
grid <- add_predictions(data, mlr2_mod)
grid['resid']<- res

# plot
ggplot(grid, aes(resid)) +
  geom_point(aes(y = pred))+
  geom_line(aes(y= pred, color = 'Democrat'), data = grid_d) +
  geom_line(aes(y=pred, color = 'Republican'), data = grid_r) +
  geom_line(aes(y=pred, color = 'Independent'), data = grid_i) +
  labs(title = 'Plot of predicted Biden score against residual with lines for each party ID type',
       x = 'Residual',y = 'Predicted BidenScore')
```



3.We have fixed the problem by using party ID as a factor. Observing from the new plot, we can see that three smooth fit lines referring to three party IDs have almost the same slope but only different y-intercept. This suggests that residual does not change based on party ID.

Interactive linear regression model

```
data_noind <- subset(data, !(dem==0 & rep==0))
ilr_mod <- lm(biden ~ female*dem, data = data_noind)
tidy(ilr_mod)
```

```
##          term  estimate std.error statistic       p.value
## 1 (Intercept) 39.382022  1.455363 27.059928 4.045546e-125
## 2      female  6.395180  2.017807  3.169371  1.568102e-03
## 3         dem 33.687514  1.834799 18.360328  3.295008e-66
## 4  female:dem -3.945888  2.471577 -1.596506  1.106513e-01
```

```
pred_data_ <- data_frame(female = c(1,1,0,0), dem = c(0,1,0,1))
# use augment to generate predictions
pred_aug_ <- augment(ilr_mod, newdata = pred_data_)
# Calculate 95% confidence intervals
pred_ci <- mutate(pred_aug_,
                  ymin = .fitted - .se.fit * 1.96,
                  ymax = .fitted + .se.fit * 1.96)
pred_ci
```

```
##   female dem  .fitted   .se.fit     ymin     ymax
```

```
## 1       1    0 45.77720 1.3976638 43.03778 48.51662
## 2       1    1 75.51883 0.8881114 73.77813 77.25953
## 3       0    0 39.38202 1.4553632 36.52951 42.23453
## 4       0    1 73.06954 1.1173209 70.87959 75.25949
```

Estimate predicted Biden warmth feeling thermometer ratings and 95% confidence interval for female Democrats is (73.77813, 77.25953); that for female Republicans is (43.03778, 48.51662); that for male Democrats is (70.87959, 75.25949); that for male Republicans is (36.52951, 42.23453).

The relationship between party ID and Biden warmth differ for males/females. For females, Democrats give around 30 points higher than Republicans; for males, Democrats give around 34 points higher than Republicans. We could also see that "female", the variable that indicates gender, has a p-value of 1.568102e-03, which shows that it makes statistically significant difference on the Biden score at the confidence interval of 95%. And since the interactive term is not statistically significant with a p-value of 1.106513e-01, we can conclude that gender makes statistically significant difference.

The relationship between gender and Biden warmth also differ for Democrats/Republicans. For Democrats, females give around 3 points higher than males; for Replublicans, females give around 6.3 points higher than males. We could also see that "dem", the variable that indicates party ID, has a p-value of 3.295008e-66, which shows that it makes statistically significant difference on the Biden score at the confidence interval of 95%. Due to the insignificant interactive term, we could conclude that party affiliation makes significant difference on the relationship between gender and Biden warmth.