Problem Set #5: Linear Regression

William L. Guzman

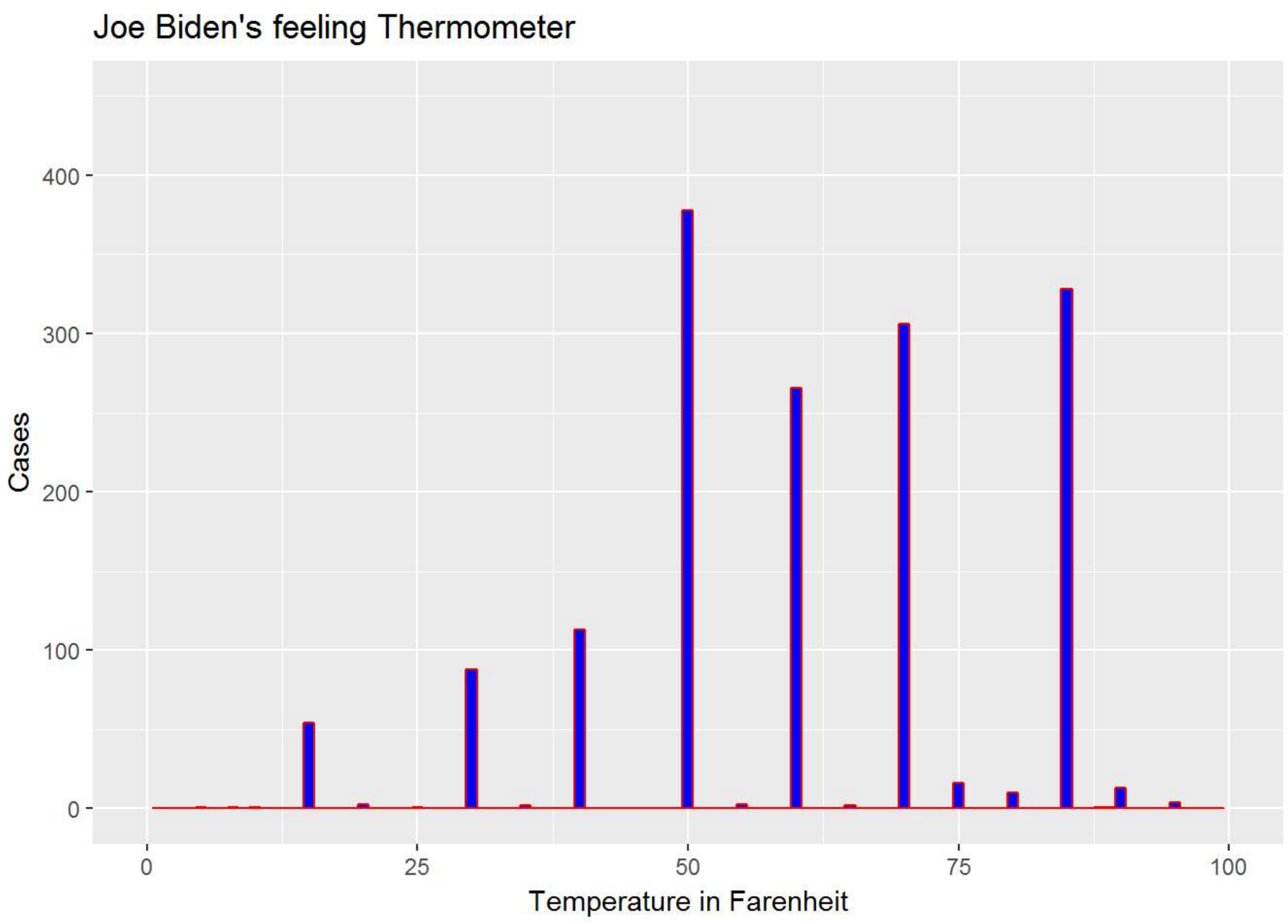
February 12, 2017

Part 1. Describe the data by plotting an histogram (1 point)

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
datapath <- "C:/Users/Walle/Documents/RScript/Data"

dat <- read.csv(file=paste(datapath,"biden.csv",sep="/"))

qplot(dat$biden, geom = "histogram", main = "Joe Biden's feeling Thermometer", xlab = "Temperatu
re in Farenheit" , ylab = "Cases",
      binwidth = 1, col=I("red"), fill=I("blue"),
      xlim = c(0,100) , ylim = c(0,450))
```



Part 2. Simple linear regression (2 points)

1. Is there a relationship between the predictor and the response?

After applying the model, we can see that by looking at the p-value(0.0563) the alternative hypothesis is rejected by the null hypothesis, there is no statistical relationship between the predictor and the response.

2. How strong is the relationship between the predictor and the response?

There is no statistical relationship with the predictor and the response variable. We can say that age does not affect biden warmth, but if it could affect it in someway, it will be a positive relationship by only 0.006241.

3. Is the relationship between the predictor and the response positive or negative?

If there will be a relationship, it will be positive because the coeficient is positive.

4. Report the R2 of the model. What percentage of the variation in biden does age alone explain? Is this a good or bad model?

The R-Squared of the model is 0.001465 and the adjusted R-squared is 0.002018. This means that the model explain around 0.145% of the model. This is a bad model. The model does not explain at least 1% of the variation. We can clearly see that age does not affects bidens feeling thermometer.

5. What is the predicted biden associated with an age of 45? What are the associated 95% confidence intervals?

With a 95% prediction interval, we have that at the age of 45, biden thermometer wit a 61.50680 fit will be between 15.50680 and 107.5059 farenheit.

6. Plot the response and predictor. Draw the least squares regression line.

```
#Simple Linear regresion with just one variable.
linearModel1 <- lm(dat$biden ~ dat$age)

summary(linearModel1)
```

```
##
## Call:
## lm(formula = dat$biden ~ dat$age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.876 -12.318  -1.257  21.684  39.617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.19736    1.64792   35.92   <2e-16 ***
## dat$age      0.06241    0.03267    1.91   0.0563 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.44 on 1805 degrees of freedom
## Multiple R-squared:  0.002018,   Adjusted R-squared:  0.001465
## F-statistic: 3.649 on 1 and 1805 DF,  p-value: 0.05626
```

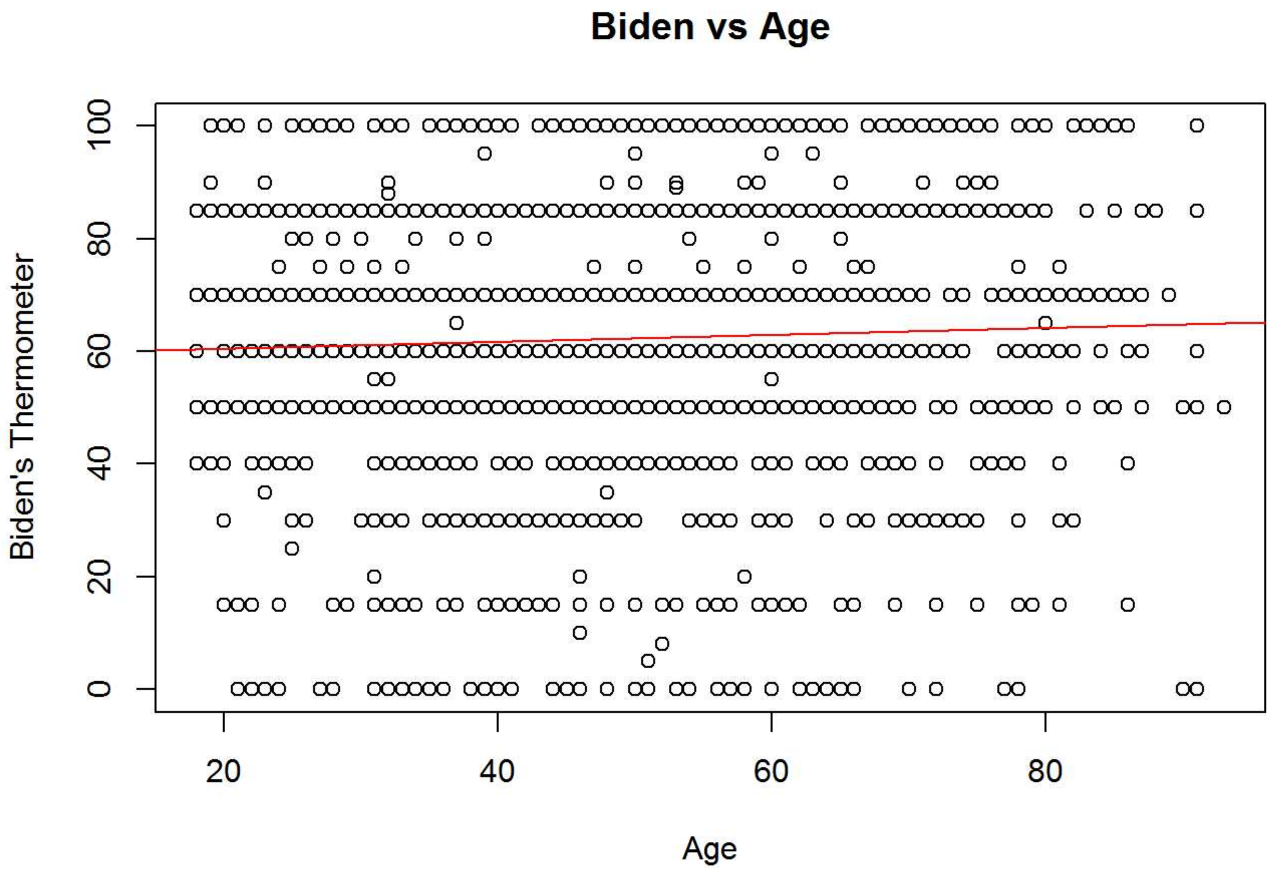
```
#Q5:
newdata = data.frame(age=45)

predict1 <- predict(linearModel1, newdata, interval="predict")
```

```
## Warning: 'newdata' had 1 row but variables found have 1807 rows
```

```
predict1[45, ]
```

```
##       fit       lwr       upr
##  61.50636  15.50680 107.50592
```

```
#Part 6
plot(dat$age, dat$biden, xlab = "Age", ylab="Biden's Thermometer", main = "Biden vs Age")
abline(linearModel1, col="red")
```



Part 3. Multiple linear regression (2 points)

1. Is there a statistically significant relationship between the predictors and response?

Between gender and education, there is a significant relationship with biden temperature.

2. What does the parameter for female suggest?

The parameter suggest that if a person gender is female (1), bidden thermometer will increase by 6.19607 and if its male(0), it will not be affected since is a categorical value.

3. Report the R2 of the model. What percentage of the variation in biden does age, gender, and education explain? Is this a better or worse model than the age-only model?

The R-Squared of the model is 0.02561 and the adjusted R-squared is 0.02723. This means that the model explain around 2.723% of biden thermometer. This is a bad model, still, we can see that it explain more than the last model.

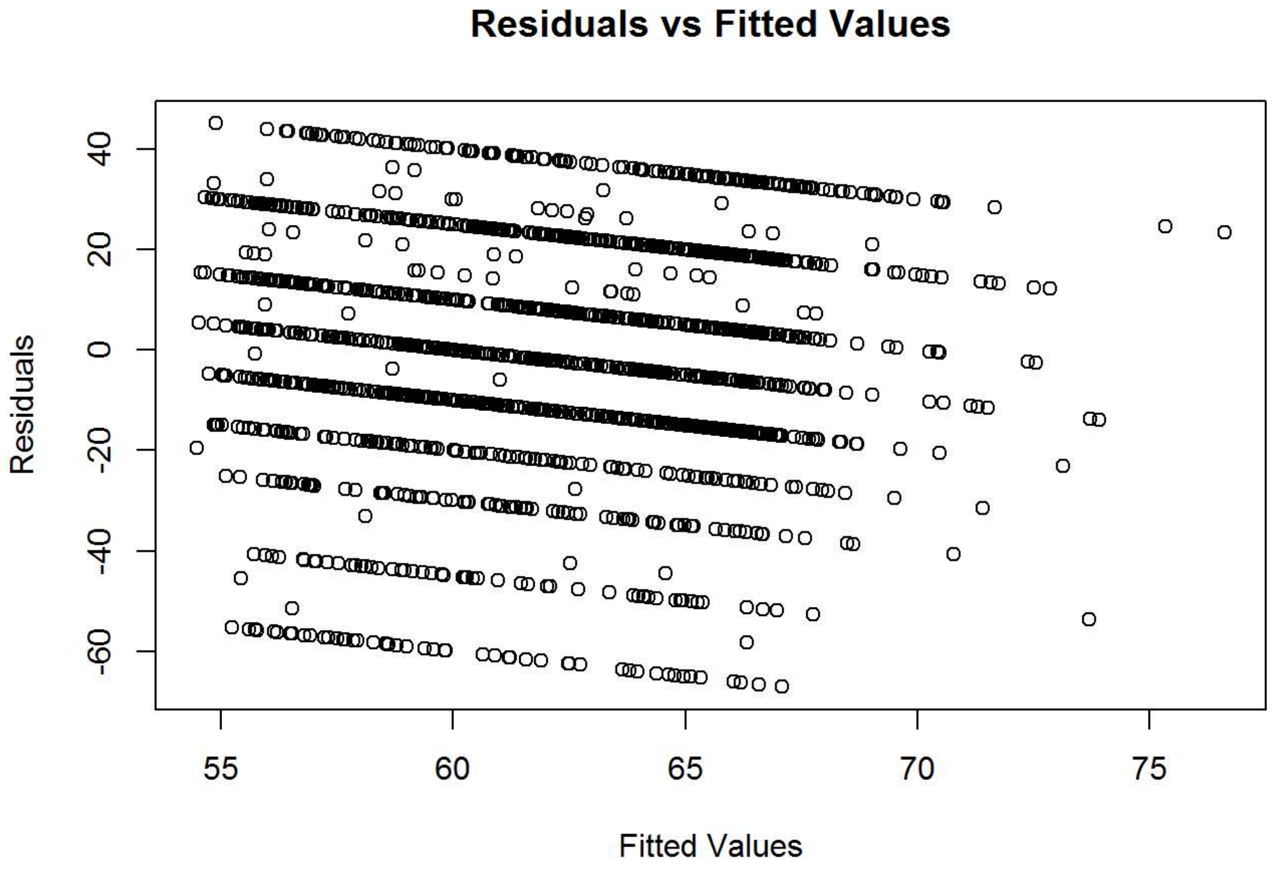
4. Generate a plot comparing the predicted values and residuals, drawing separate smooth fit lines for each party ID type. Is there a problem with this model? If so, what?

Yes, the model does not have a clear linear relationship. Still, we can see that there could be a negative relationship.

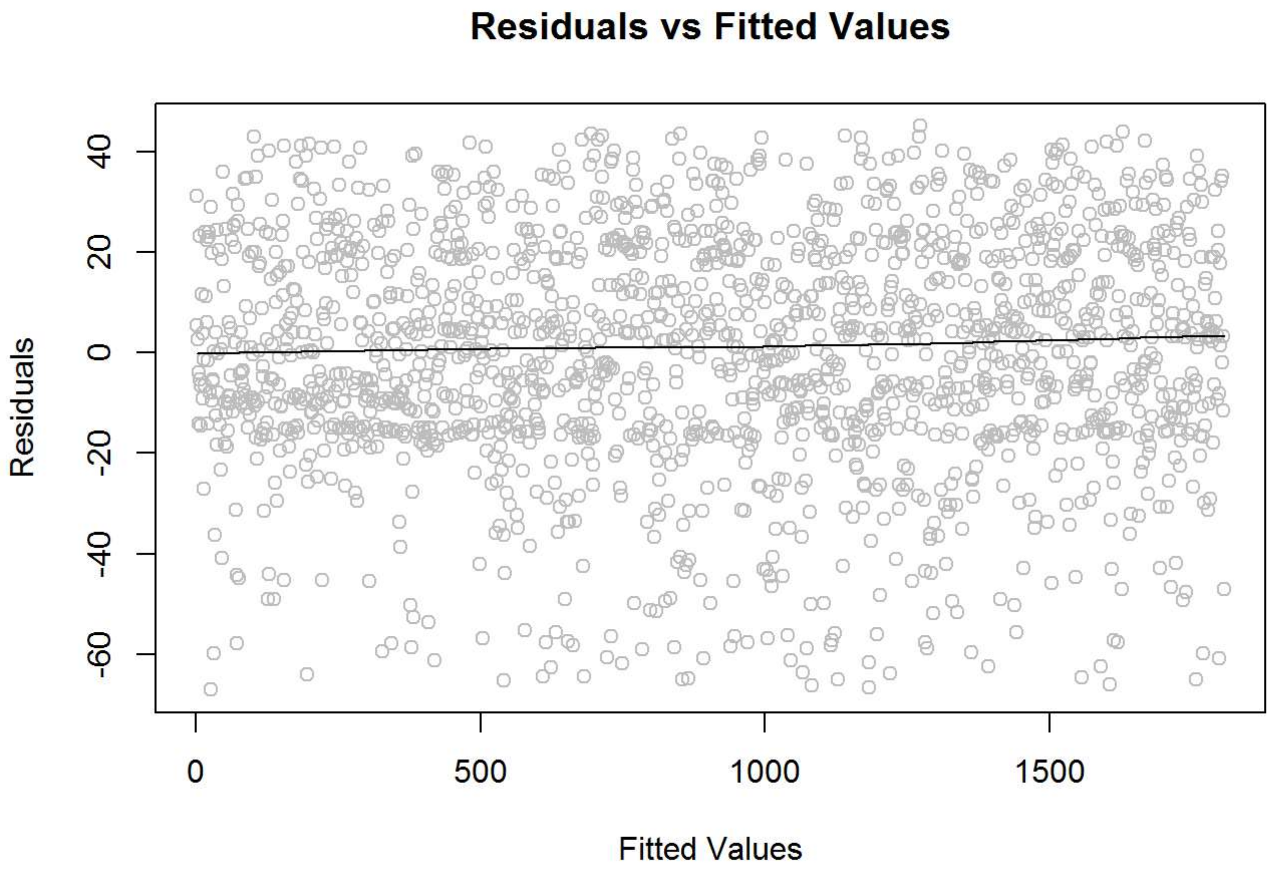
```
#Linear model 2
linearModel2 <- lm(biden~age+female+educ, data = dat)

summary(linearModel2)
```

```
##
## Call:
## lm(formula = biden ~ age + female + educ, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.084 -14.662   0.703  18.847  45.105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.62101    3.59600  19.083  < 2e-16 ***
## age          0.04188    0.03249   1.289    0.198
## female       6.19607    1.09670   5.650 1.86e-08 ***
## educ        -0.88871    0.22469  -3.955 7.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.16 on 1803 degrees of freedom
## Multiple R-squared:  0.02723,    Adjusted R-squared:  0.02561
## F-statistic: 16.82 on 3 and 1803 DF,  p-value: 8.876e-11
```

```
#Plot the model
plot(linearModel2$fitted.values,linearModel2$residuals, main = "Residuals vs Fitted Values", xla
b = "Fitted Values",
     ylab ="Residuals" )
```



```
#smooth line plot
scatter.smooth(x=1: length(linearModel2$fitted.values), y=linearModel2$residuals, col="grey",
               main = "Residuals vs Fitted Values", xlab = "Fitted Values",
               ylab ="Residuals" )
```



Part 4. Multiple linear regression model (with even more variables!) (3 points)

1. Did the relationship between gender and Biden warmth change?

Yes, the gender decrease from 6.19607 to 4.10323. Still, the gender does come in factor for changing biden warmth

2. Report the R2 of the model. What percentage of the variation in biden does age, gender, education, and party identification explain? Is this a better or worse model than the age + gender + education model?

The R-Squared of the model is 0.2795 and the adjusted R-squared is 0.2815 This means that the model explain around 28.15% of biden thermometer. This still is a bad model, but so far, is better than the last two model.

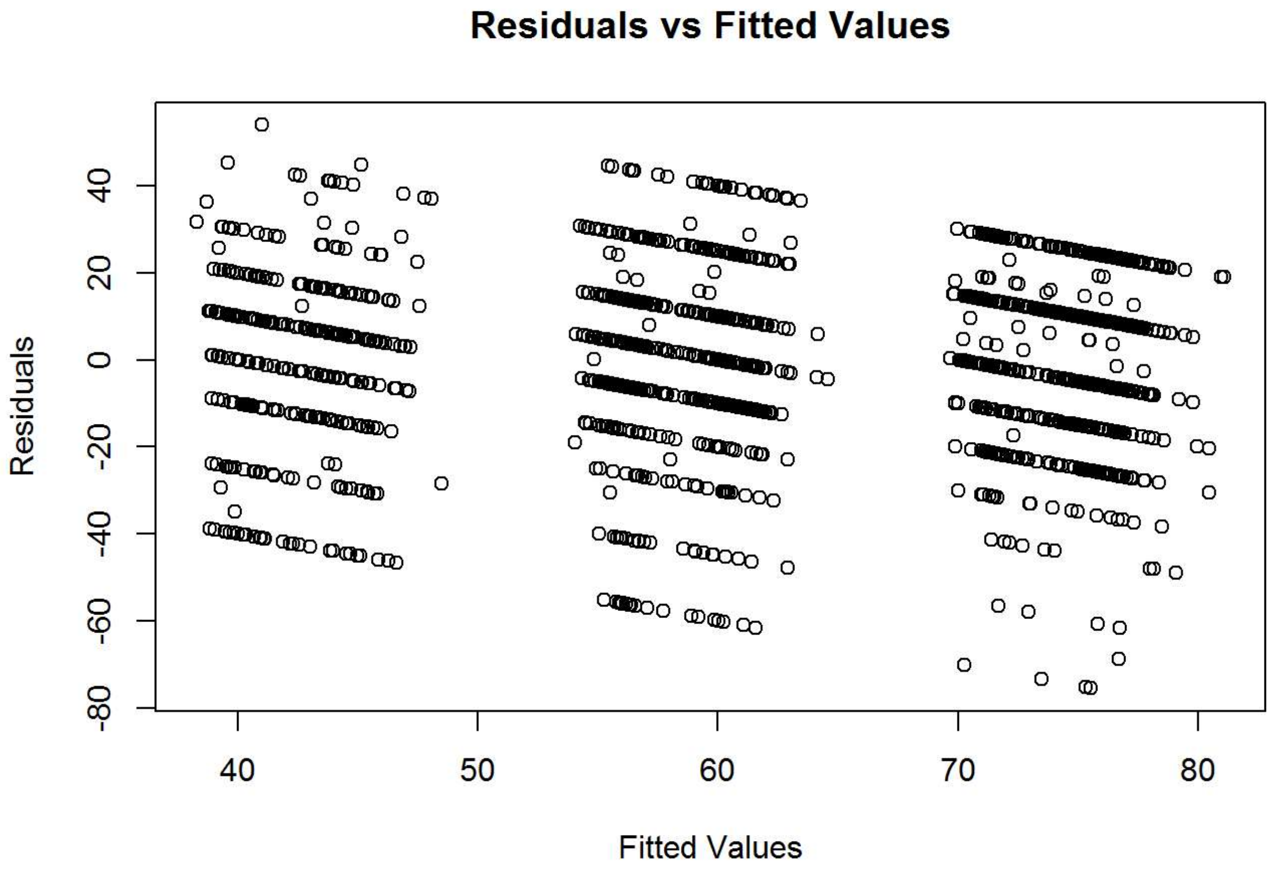
3. Generate a plot comparing the predicted values and residuals, drawing separate smooth fit lines for each party ID type. By adding variables for party ID to the regression model, did we fix the previous problem?

After seeing the two plots, we can see that adding more variables, we still don have a model that shows a linear relationship in. This did not fix our problem.

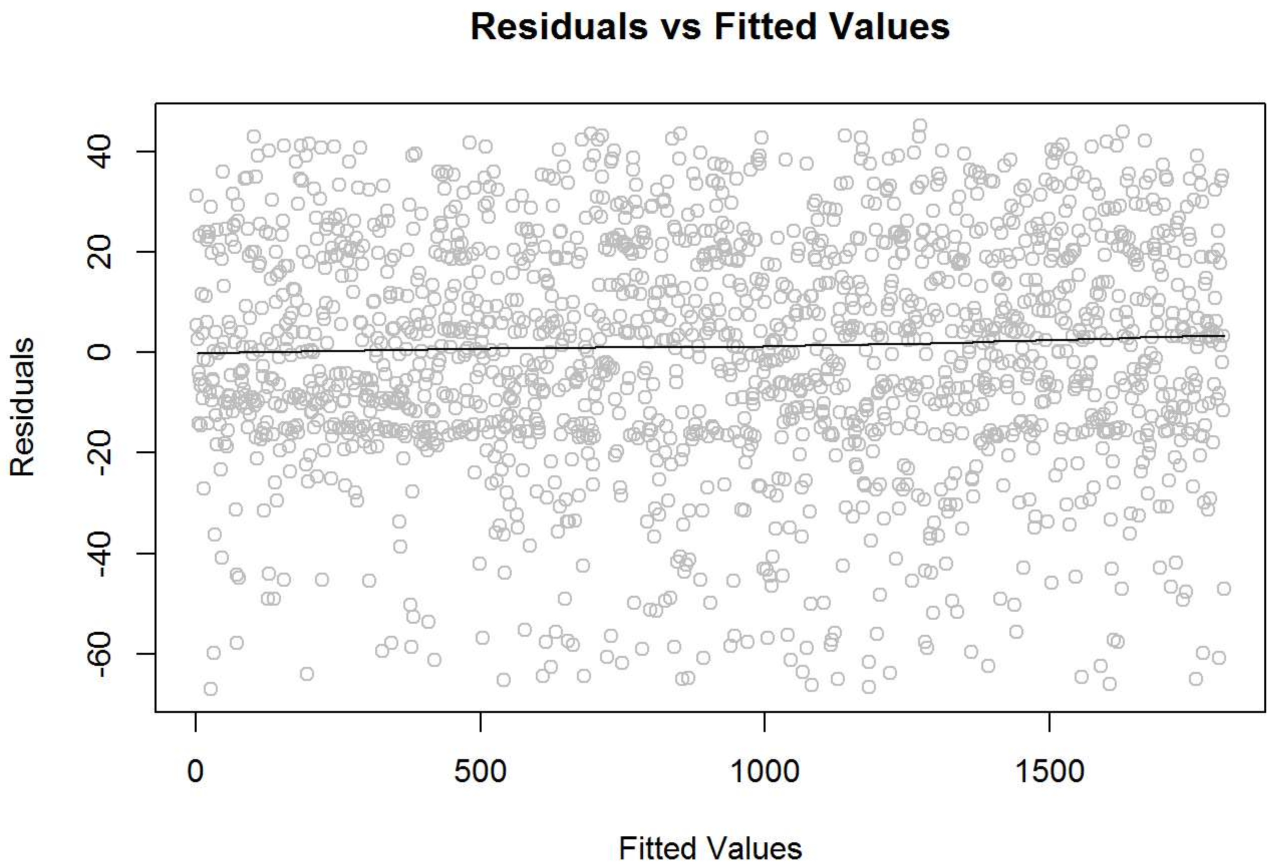
```
#Multiuple linear regression with more variables.
linearModel3 <- lm(biden~., data = dat)

summary(linearModel3)
```

```
##
## Call:
## lm(formula = biden ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.546 -11.295   1.018  12.776  53.977
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  58.81126    3.12444  18.823  < 2e-16 ***
## female        4.10323    0.94823   4.327 1.59e-05 ***
## age           0.04826    0.02825   1.708   0.0877 .
## educ         -0.34533    0.19478  -1.773   0.0764 .
## dem          15.42426    1.06803  14.442  < 2e-16 ***
## rep         -15.84951    1.31136 -12.086  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.91 on 1801 degrees of freedom
## Multiple R-squared:  0.2815, Adjusted R-squared:  0.2795
## F-statistic: 141.1 on 5 and 1801 DF,  p-value: < 2.2e-16
```

```
#plot the model
plot(linearModel3$fitted.values,linearModel3$residuals, main = "Residuals vs Fitted Values", xla
b = "Fitted Values",
     ylab ="Residuals" )
```



```
#smooth line plot
scatter.smooth(x=1: length(linearModel3$fitted.values), y=linearModel2$residuals, col="grey",
               main = "Residuals vs Fitted Values", xlab = "Fitted Values",
               ylab ="Residuals" )
```



Part 5. Interactive linear regression model (2 points)

1. Report the values of the standard errors and the parameter.

```
#Subsetting data
filterData <- subset(dat, dem == 1 | rep == 1, select=c(biden,female, age,dem, rep,educ))

linearModel4 <- lm(biden~female+dem, data=filterData)

summary(linearModel4)
```

```
## 
## Call:
## lm(formula = biden ~ female + dem, data = filterData)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -76.028 -12.263   5.485  12.737  54.250 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   40.750      1.177  34.619  < 2e-16 ***
## female         3.765      1.166   3.229  0.00128 ** 
## dem           31.513      1.230  25.617  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 19.43 on 1148 degrees of freedom
## Multiple R-squared:  0.3742, Adjusted R-squared:  0.3731 
## F-statistic: 343.2 on 2 and 1148 DF,  p-value: < 2.2e-16
```

2. Estimate predicted Biden warmth feeling thermometer ratings and 95% confidence intervals for female Democrats, female Republicans, male Democrats, and male Republicans. Does the relationship between party ID and Biden warmth differ for males/females? Does the relationship between gender and Biden warmth differ for Democrats/Republicans?

After analyzing and comparing the different cases, we can clearly see that Biden warmth will differ more by the party ID than the gender of the person. For example, the fitted value for a Female democrat and a male democrat is 76.02831 vs 72.26313 with a 95% interval of (37.87119, 114.1854) vs (34.09054, 110.4357). If we compare these results with female/male republican, we can see that is the same case with a fitted value of 44.51537 vs 40.75019 with a 95% interval of (6.325838, 82.7049) vs (2.557887, 78.94249). We also can see that there is not much difference between the gender of a particular party, but there is between different party, between democrats vs republican.

```
#When female democrats
predictBiden1 <- data.frame(female=1, dem=1)

#When female republican
predictBiden2 <- data.frame(female=1, dem=0)

#When male Democrats
predictBiden3 <- data.frame(female=0, dem=1)

#When male republican
predictBiden4 <- data.frame(female=0, dem=0)

#predict
#Female democrats
predict(linearModel4, predictBiden1, interval = "predict")
```

```
##        fit      lwr      upr
## 1 76.02831 37.87119 114.1854
```

```
#Female republican
predict(linearModel4, predictBiden2, interval = "predict")
```

```
##        fit      lwr     upr
## 1 44.51537 6.325838 82.7049
```

```
#Male Democrats
predict(linearModel4, predictBiden3, interval = "predict")
```

```
##        fit      lwr      upr
## 1 72.26313 34.09054 110.4357
```

```
#Male Republican
predict(linearModel4, predictBiden4, interval = "predict")
```

```
##        fit      lwr      upr
## 1 40.75019 2.557887 78.94249
```