SDS 291 - Fat Bear 2020-2021

Laura Mora, Mariem Snoussi, Maya Crandall-Malcolm

2021-12-13

Contents


```
library(tidyverse)
library(openintro)
library(mosaic)
library(modeest)
library(GGally)
library(moderndive)
```

```
bears<-read.csv("FullBearData.csv")
survey20<-read.csv("Survey2020.csv")
survey21<-read.csv("Survey2021.csv")
```

```
lagged_bears <- bears %>%
  arrange(Year,BearNumber,Round) %>%
  group_by(Year,BearNumber)%>%
  mutate(votes_last_round=lag(Number.of.Votes)) %>%
  ungroup()

lagged_bears <- lagged_bears %>%
  mutate(BeforeAvg=BeforeSizeSum/BeforeSizeCount)

lagged_bears <- lagged_bears %>%
  mutate(AfterAvg=AfterSum/AfterCount)
```

Purpose

The purpose of this analysis is to evaluate what factors play a major role in the winner of Fat Bear Week. Fat Bear Week is an online single elimination tournament held in Katmai National Park, Alaska. This year

it was held from September 29th to October 5th. According to the website, where the voting takes place, who to vote for as the winner is subjective. Some voters may choose to vote for a particular bear based on dramatic change in weight over the year. Other voters might choose to vote based on additional challenges the bears may face in order to gain weight. But in general the idea is to vote for the fattest bear, in the voter's opinion. We hypothesize that the more angled away from the camera the bear is, i.e. the closer the butt of the bear is to the camera, the more votes that bear will receive. We also hypothesize that the probability that a bear will win the second round increases as the number of votes that bear obtains in the first round increases. That is, the likelihood of success, where success is that a particular bear wins in the second round, increases as the number of votes the bear won in the first round increases. We will also explore other variables that might play a role in the who named winner of Fat Bear Week; such as, fur color, change in size, and more.

Data

We plan to use data about the last two years of the Fat Bear Week competition that we compiled ourselves from various sources (Fat Bear Week Voting Page, Katmai Park publications, Katmai Bearcams Fan Wiki). We will at some point link to our dataset. There are many more bears at Katmai National Park than are featured in the Fat Bear Week competition. Former park ranger Mike Fitz elaborated on the selection of each year's competitors, saying, < the bears are typically chosen by the park rangers, based on a variety of factors such as their time of arrival at Brooks River. An early summer arrival at the river ensures that the bear is photographed before and after their weight gain. A compelling storyline is also a driving factor when choosing a contestant. The audience loves to see a veteran bear with skill and resilience, like Otis, competing year after year > [1]. Although the contest began in 2014, we restricted our investigation to the contests that took place in 2020 and 2021. Contests prior to those years took place on Facebook with the bear whose post received the most "Likes" winning each round, where we have no way to control for bias due to the Facebook algorithm for amplifying posts given previous likes. In 2020 and 2021, voting took place on a designated webpage [2] which guarantees that all respondents saw both bears for each match-up they voted in.

Respondent perception of a bear's size, color, and position in relation to the camera is subjective and vary between respondents. To determine the attributes size (before & after), color, and angle of each bear relative to the camera, we conducted two brief surveys (one for each contest year) with human respondents. The survey asked respondents to look at the official before and after photos of the bears from the contests, and report each bear's size, angle in relation to the camera, and fur color. We used convenience sampling of primarily our friends and family. We obtained the results of the survey for each bear by identifying the most frequent response for the questions related to fur color and size, and took the mean of the responses for angle to the camera.

Population

Each row of the dataset is one half of a match-up in the tournament. Each row contains the bear's designated number, name (if applicable), contest year, round, opponent's designated number, number of votes received in that match-up, outcome of the match-up, outcome of the whole tournament, number of previous Fat Bear Week wins, sex, age, fur color, summer size (before), fall size(after), and angle to the camera. It is our hope that our findings could be applied to predicting the winner of future Fat Bear Week tournaments.

Outcome Variable

The outcome variable for the first hypothesis is number of votes (Number.of.Votes) obtained in the first round. This is a quantitative variable in the positive integers, each vote will count as one unit. The outcome variable for the second hypothesis is RoundWinBinary (associated with Round 2). It's a binary response with 0 representing a loss and 1 representing a win.

Explanatory Variable

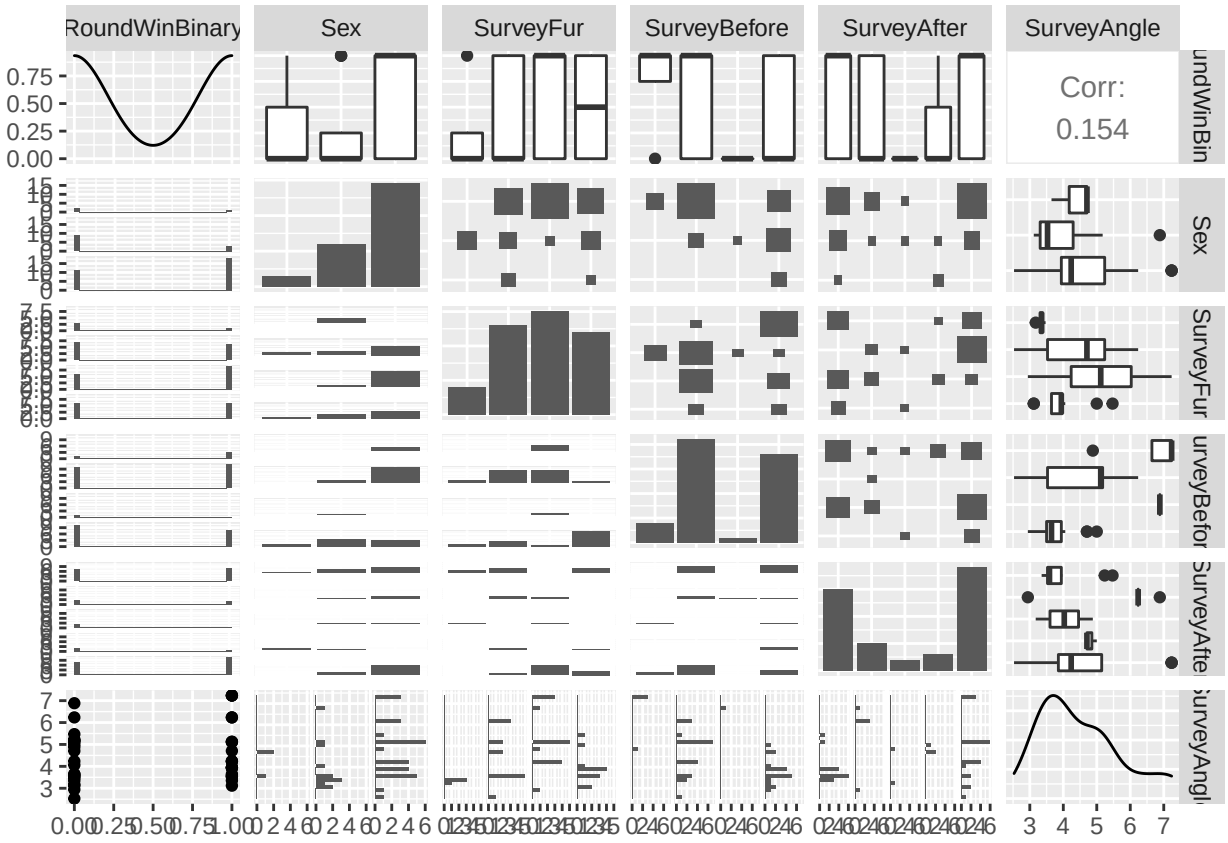
The explanatory variables are: Sex, a binary variable with values "M" for male and "F" for female. Color is a categorical variable. It's possible values are "Blond", "Light Brown", "Brown", "dark Brown" Age is a quantitative variable. Its unit is years. Presence of names is a binary variable. It will be 0 if the bear does not have a name and 1 when they have a name. Number of Votes is a quantitative variable. Each one point increase is equivalent to 1 additional vote towards the bear.

Exploratory Visualizations

A preliminary exploration of our data shows that there is some relationship between winning a round and some of our predictor variables using simple logistic regression. There appears to be positive associations between angle to camera and the likelihood of winning a round, bear age and likelihood of winning a round, and number of votes and winning a round.

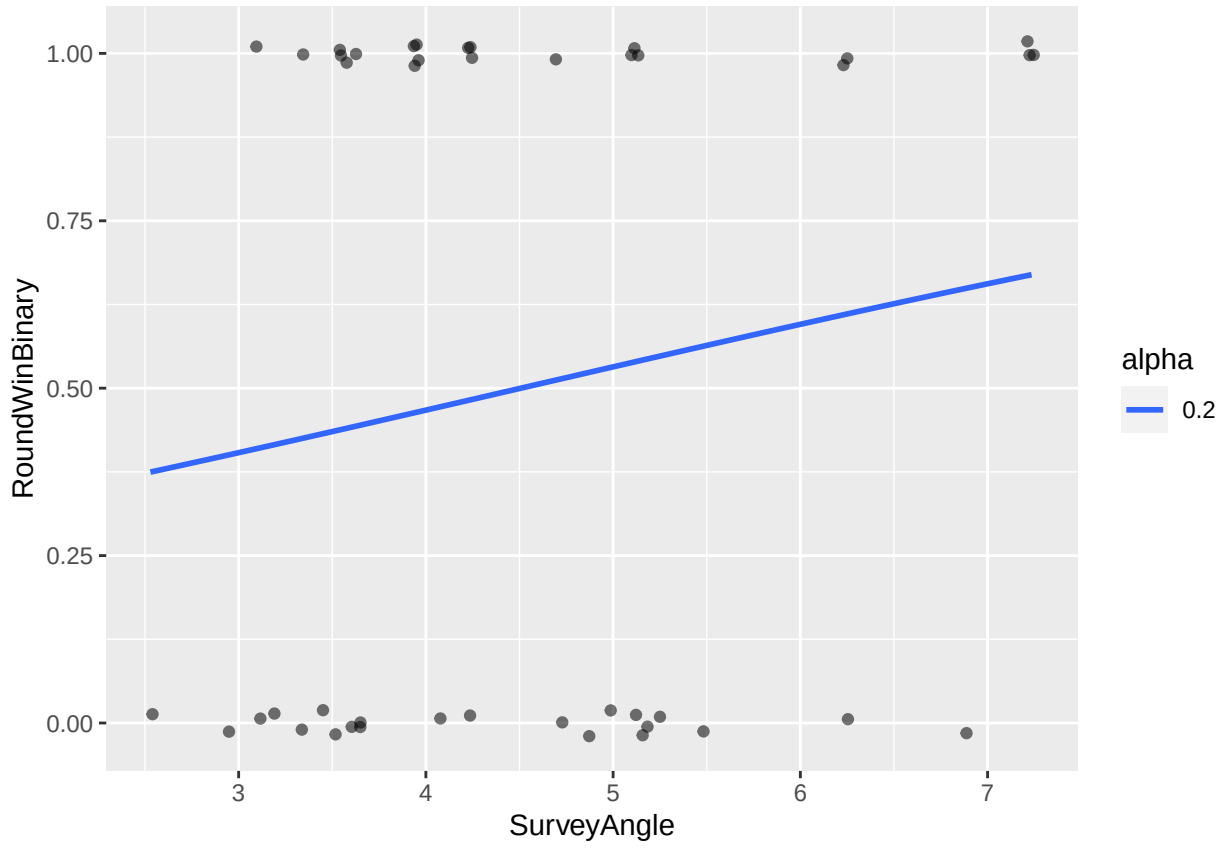
```
bears %>%
  select(RoundWinBinary, Sex, SurveyFur, SurveyBefore, SurveyAfter, SurveyAngle) %>%
  ggpairs()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



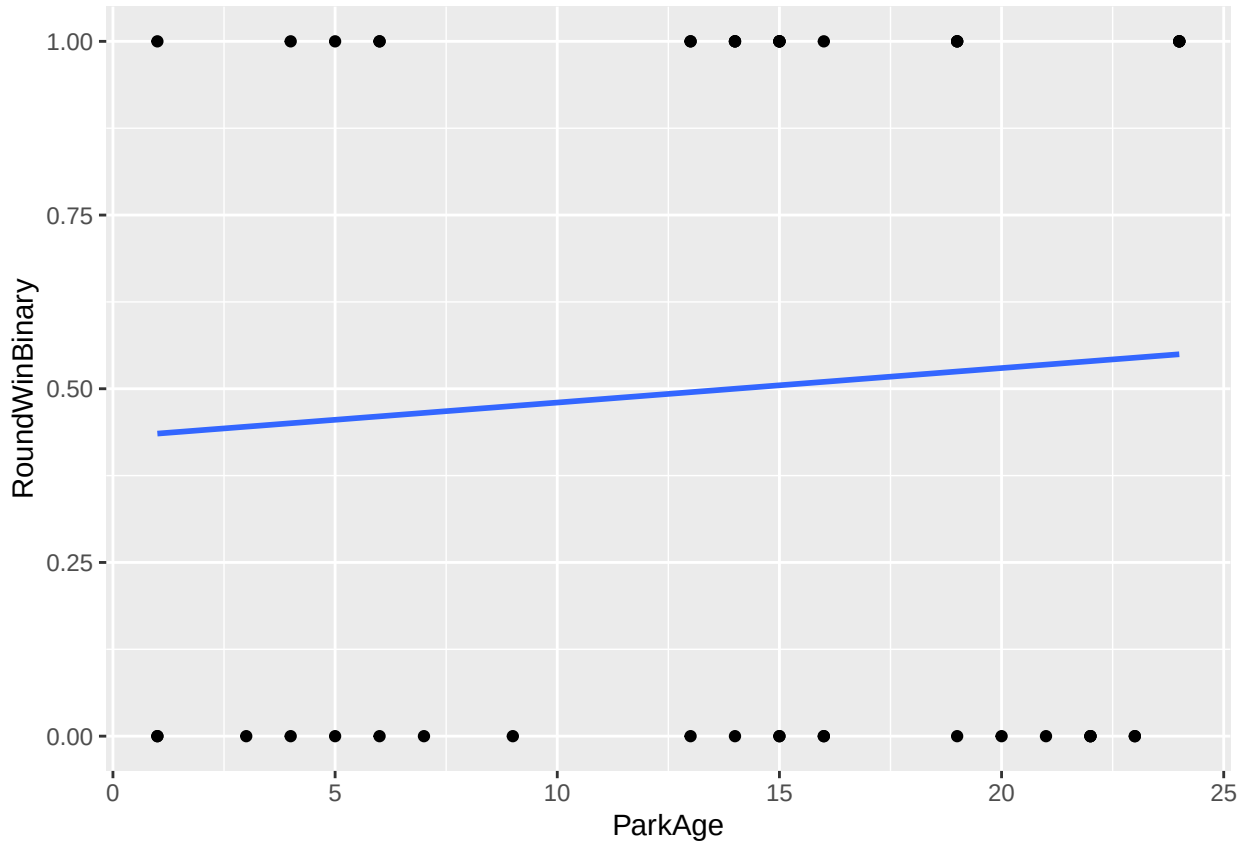

```
plot <- qplot(x = SurveyAngle, y = RoundWinBinary, data = bears,
              alpha = 0.2, show.legend = FALSE,
              geom = "jitter", height = 0.02, width = 0)
plot + geom_smooth(method = "glm",
method.args = list(family = "binomial"), se = 0)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



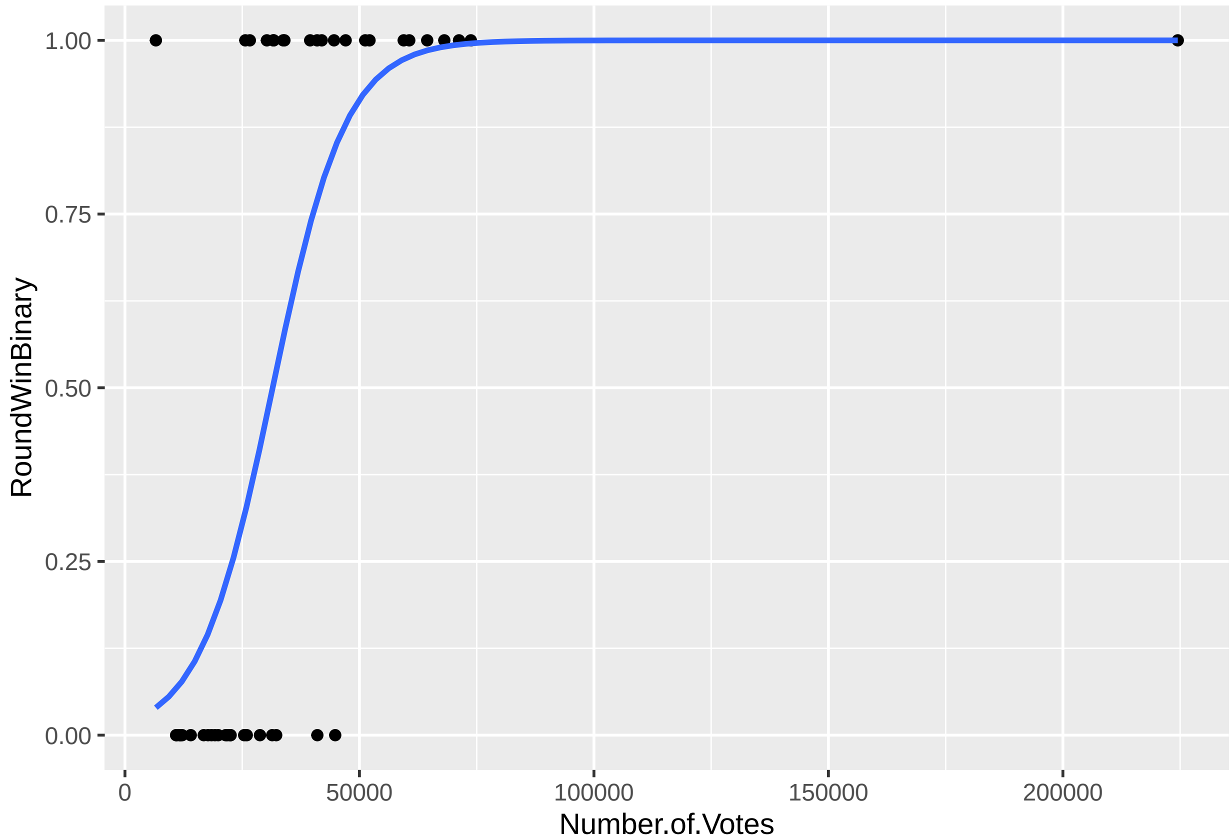

```
qplot(x=ParkAge, y=RoundWinBinary, data=bears) + geom_smooth(method="glm", se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



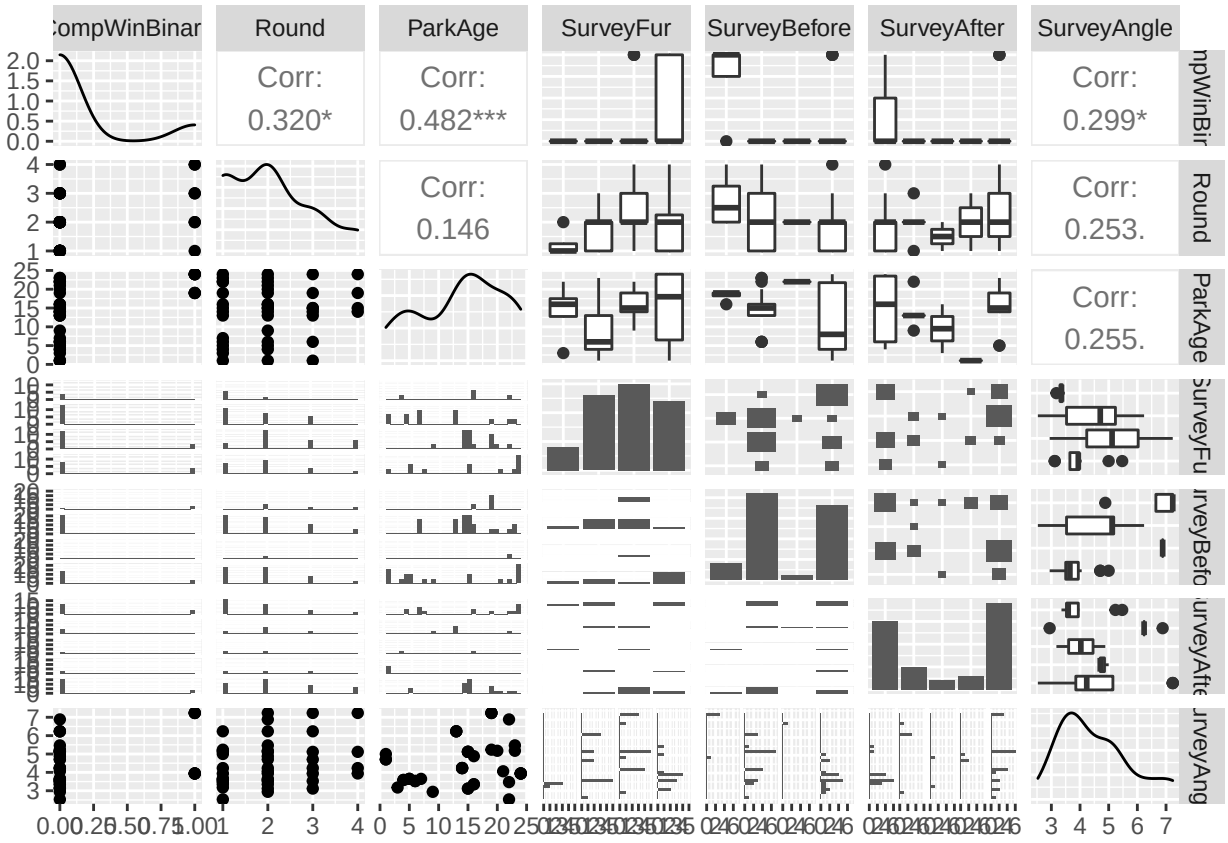

```
qplot(x=Number.of.Votes, y=RoundWinBinary, data=bears) +
  geom_smooth(method="glm", method.args = list(family = "binomial"), se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
bears %>%
  select(CompWinBinary, Round, ParkAge, SurveyFur, SurveyBefore, SurveyAfter, SurveyAngle) %>%
  ggpairs()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#logit model with Win and Age
logitAge <- glm(CompWinBinary ~ ParkAge + Round, data = bears, family = binomial)
summary(logitAge)
```

```
##
## Call:
## glm(formula = CompWinBinary ~ ParkAge + Round, family = binomial,
##     data = bears)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.46644  -0.20842  -0.03179  -0.00085   2.19164
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -20.3528     8.4679  -2.404   0.0162 *
## ParkAge       0.7411     0.3224   2.299   0.0215 *
## Round         1.9822     0.9911   2.000   0.0455 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38.558  on 43  degrees of freedom
## Residual deviance: 15.995  on 41  degrees of freedom
## AIC: 21.995
```

```
##
## Number of Fisher Scoring iterations: 8
```

```r
exp(cbind(OR = coef(logitAge), confint(logitAge)))
```

```
## Waiting for profiling to be done...

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                       OR        2.5 %       97.5 %
## (Intercept) 1.448359e-09 1.540043e-19 2.163392e-04
## ParkAge     2.098335e+00 1.320573e+00 4.997619e+00
## Round       7.258639e+00 1.576049e+00 9.875248e+01
```

```r
qnorm(p=.05/2, lower.tail=FALSE)
```

```
## [1] 1.959964
```

fitted logistic regression equation in logit form:

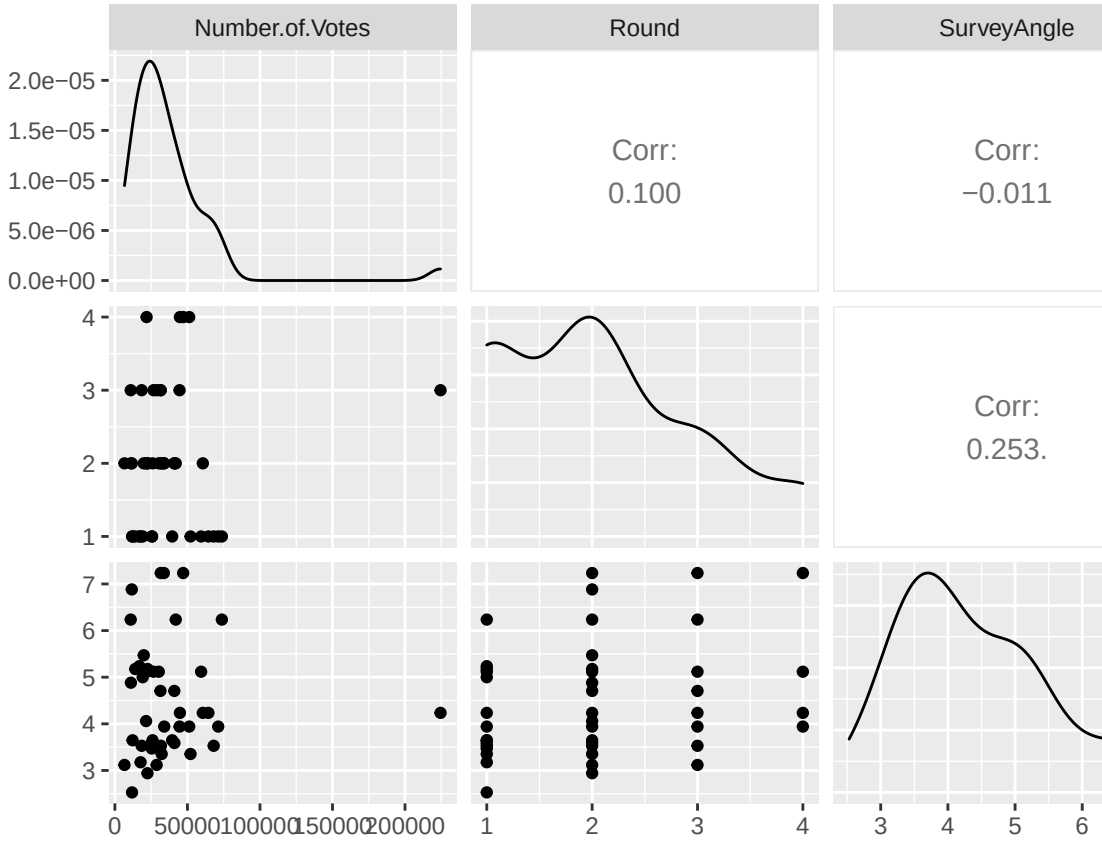
$$log(\frac{\widehat{\pi}_{Win}}{1-\widehat{\pi}_{Win}}) = -20.3528 + 0.7411(Age) + 1.9822(Round)$$

Adjusting for round, the association between winning the tournament and age is statistically significant. Specifically, the odds of a bear winning the tournament are 2.09 times higher for each year increase in age,

adjusting for round, on average in the population. The z-statistic is slightly larger (z=2.00) and the pvalue is small (p=0.0215) compared to their critical values (z=1.96, p<0.05), so we can reject the null hypothesis that the relationship between the odds of winning the tournament and age was statistically significant. We also see that the 95% CI for this estimate (OR:2.098, 95% CI: 1.32, 4.99), does not include the null of 1.

Adjusting for age, the odds of a bear winning the tournament are 1.98 times higher for each additional round passed, on average in the population.

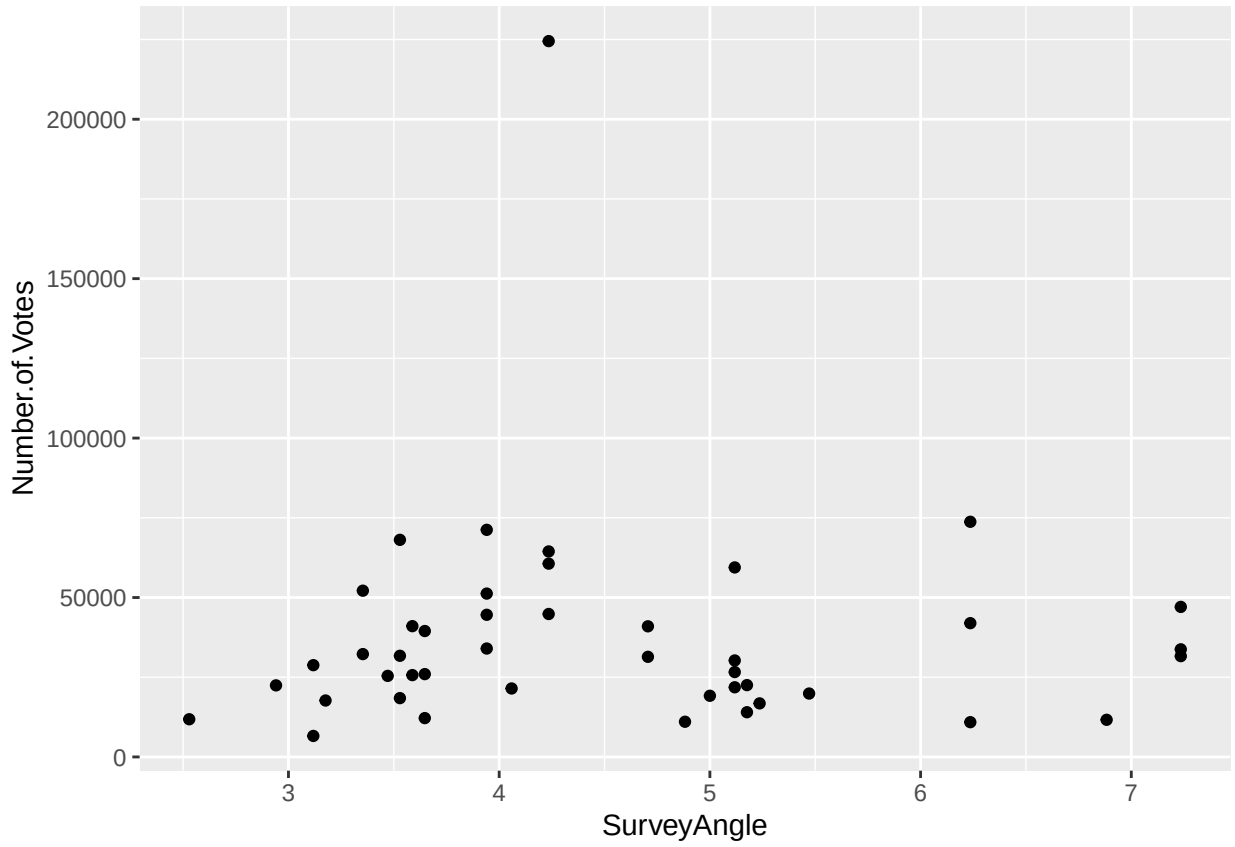
```
bears %>%
  select(Number.of.Votes, Round, SurveyAngle) %>%
  ggpairs()
```



HYPOTHESIS TESTS

```
bears_out <- lagged_bears %>%
     filter(Number.of.Votes<224496)
```

```
qplot(x = SurveyAngle, y = Number.of.Votes, data = bears)
```




```
# Linear regression model for SurveyAngle and Votes controlling for round
hyp1mod <- lm(Number.of.Votes ~ SurveyAngle + Round, data = bears)
summary(hyp1mod)
```

```
##
## Call:
## lm(formula = Number.of.Votes ~ SurveyAngle + Round, data = bears)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32121 -16345  -7613   5035 183087
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    34265      20944   1.636    0.109
## SurveyAngle    -1072       4399  -0.244    0.809
## Round           3894       5704   0.683    0.499
##
## Residual standard error: 34450 on 41 degrees of freedom
## Multiple R-squared:  0.01137,    Adjusted R-squared:  -0.03686
## F-statistic: 0.2358 on 2 and 41 DF,  p-value: 0.791
```

fitted:

$$\widehat{NumberOfVotes} = 34265 - 1072(Angle) + 3894(Round)$$

We predict that a bear would win 34,364 votes at Round 0 and at 0 degrees rotation from camera (??).

For every 1 point in angle rotation (fix this part), we'd predict a 1,072 point decrease in votes, holding round constant. However, we fail to reject the null hypothesis that $\beta_1 = 0$. Therefore, we cannot say there is a statistically significant effect of SurveryAngle on Number of Votes when controlling for round, t(41) = -0.244, p= 0.809.

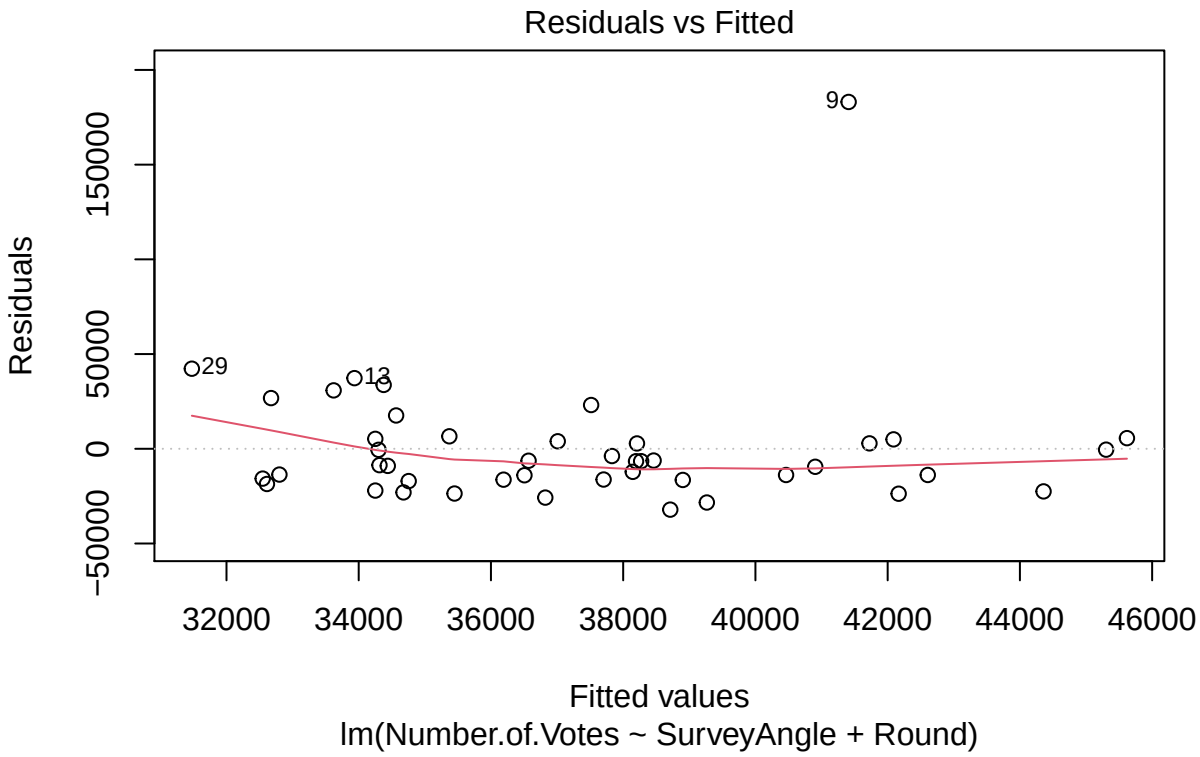
For every 1 round increase, we'd predict a 3,894 point increase in votes, holding SurveyAngle constant. However, we fail to reject the null hypothesis that $\beta_2 = 0$. Therefore, we cannot say there is a statistically significant effect of Round on Number of Votes, holding SurveyAngle Constant, t(41) = 0.683, p= 0.499.

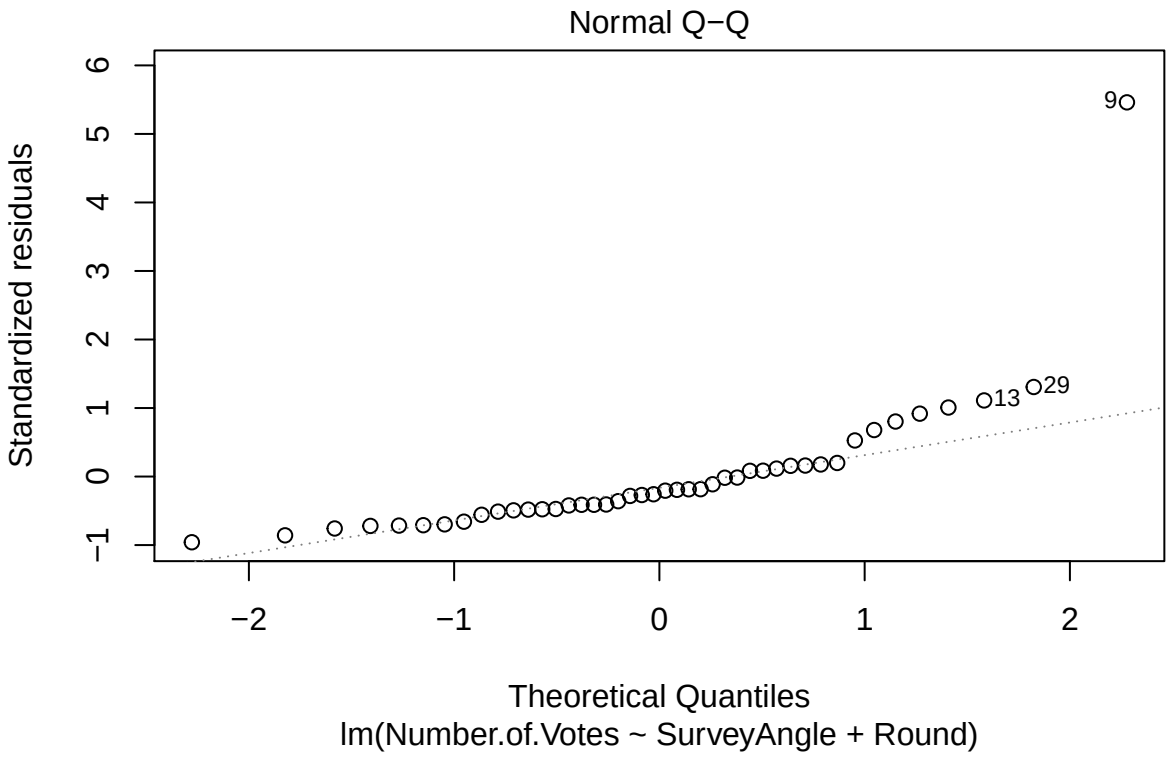
Our model does not explain a statistically significant amount of variation in number of votes, F(2, 41) = 7.71 , p =0.791.

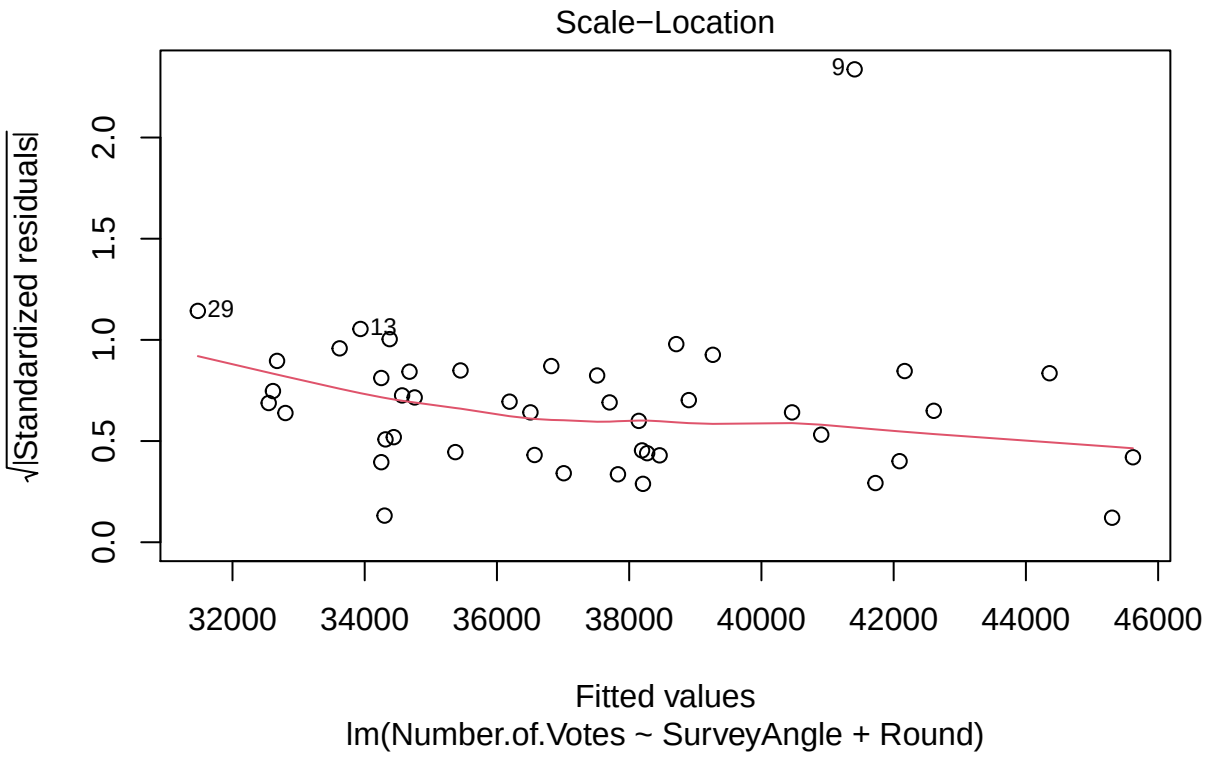
```
# Linear regression model for SurveyAngle and Votes controlling for round
hyp1modOut <- lm(Number.of.Votes ~ SurveyAngle + Round, data = bears_out)
summary(hyp1modOut)
```

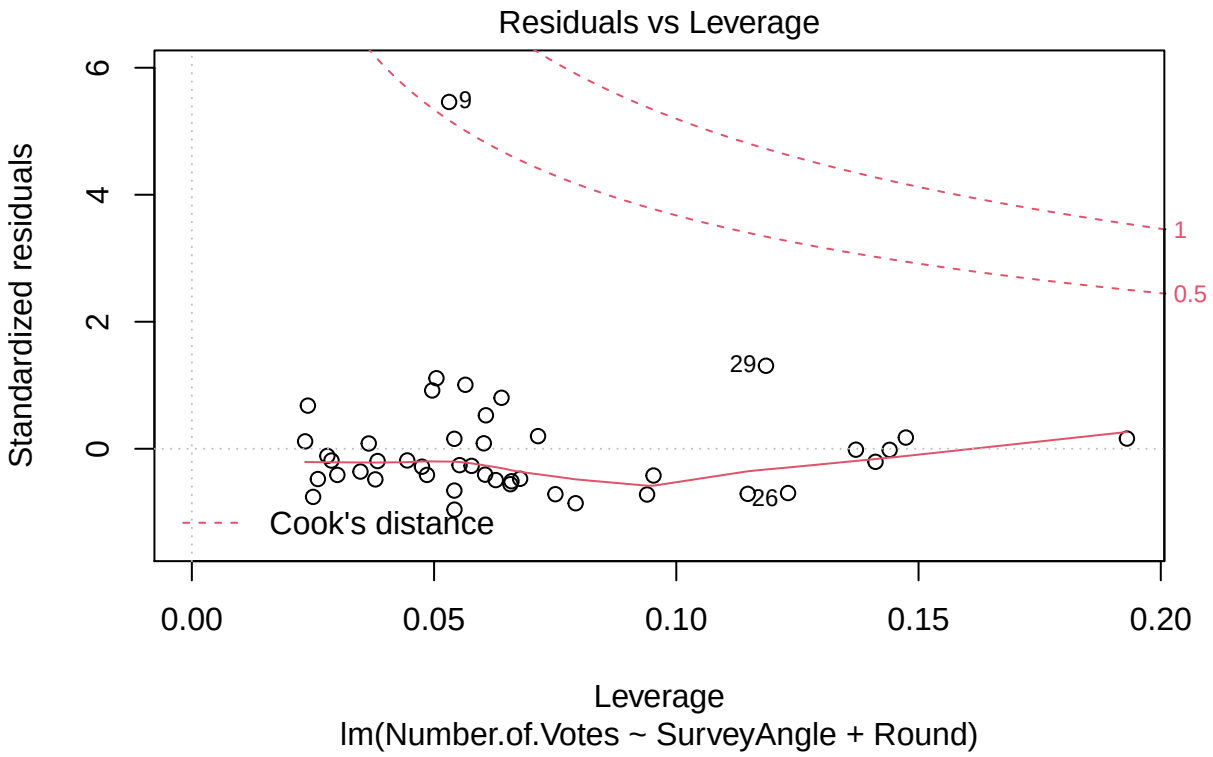
```
##
## Call:
## lm(formula = Number.of.Votes ~ SurveyAngle + Round, data = bears_out)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25209 -12980  -1984  11259  37819
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32391.3    11073.2   2.925  0.00565 **
## SurveyAngle    819.8     2332.7   0.351  0.72710
## Round        -1573.7     3061.5  -0.514  0.61007
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18210 on 40 degrees of freedom
## Multiple R-squared:  0.007806,   Adjusted R-squared:  -0.0418
## F-statistic: 0.1573 on 2 and 40 DF,  p-value: 0.8549
```

```
plot(hyp1mod)
```

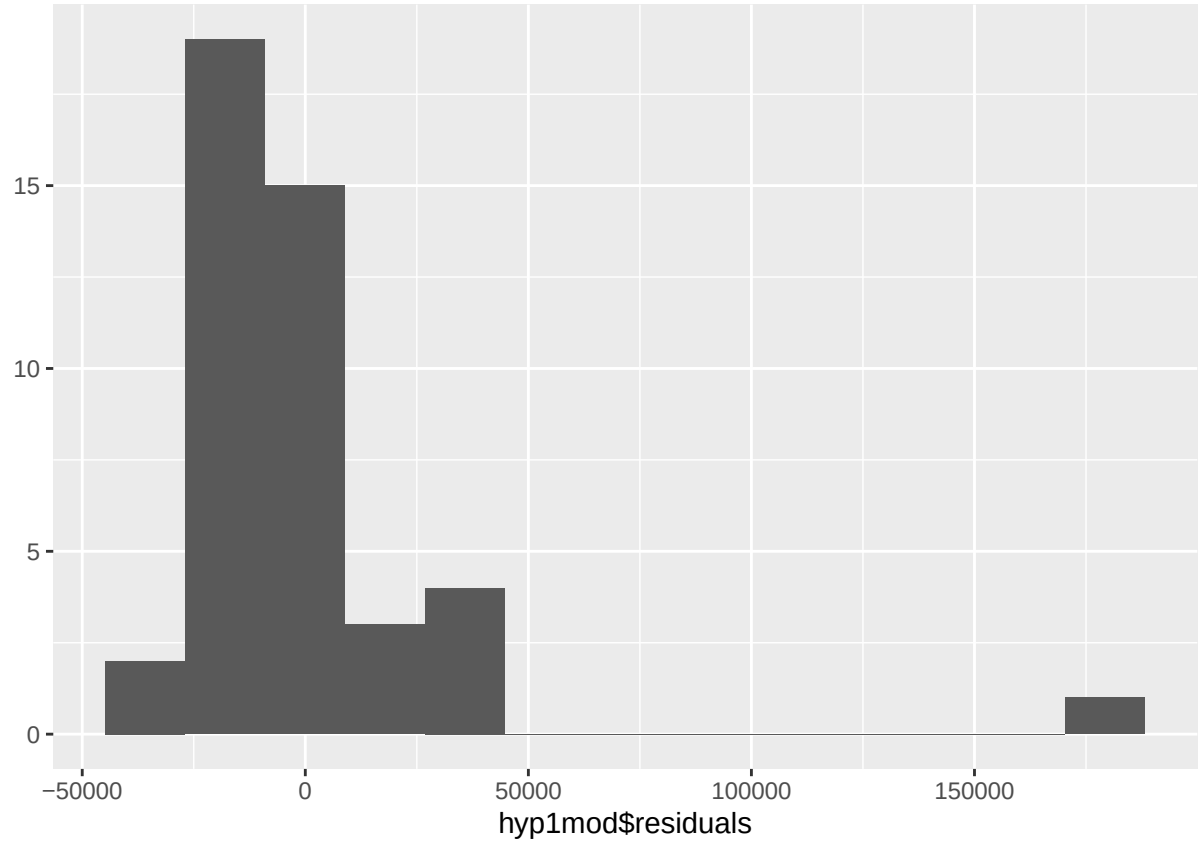

Residuals vs Fitted
Residuals
150000
50000
0
−50000
9
29
13
32000
34000
36000
38000
40000
42000
44000
46000
Fitted values
lm(Number.of.Votes ~ SurveyAngle + Round)


Normal Q−Q
Standardized residuals
−1
0
1
2
3
4
5
6
9
13
29
−2
−1
0
1
2
Theoretical Quantiles
lm(Number.of.Votes ~ SurveyAngle + Round)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
9
29
13
32000
34000
36000
38000
40000
42000
44000
46000
Fitted values
lm(Number.of.Votes ~ SurveyAngle + Round)

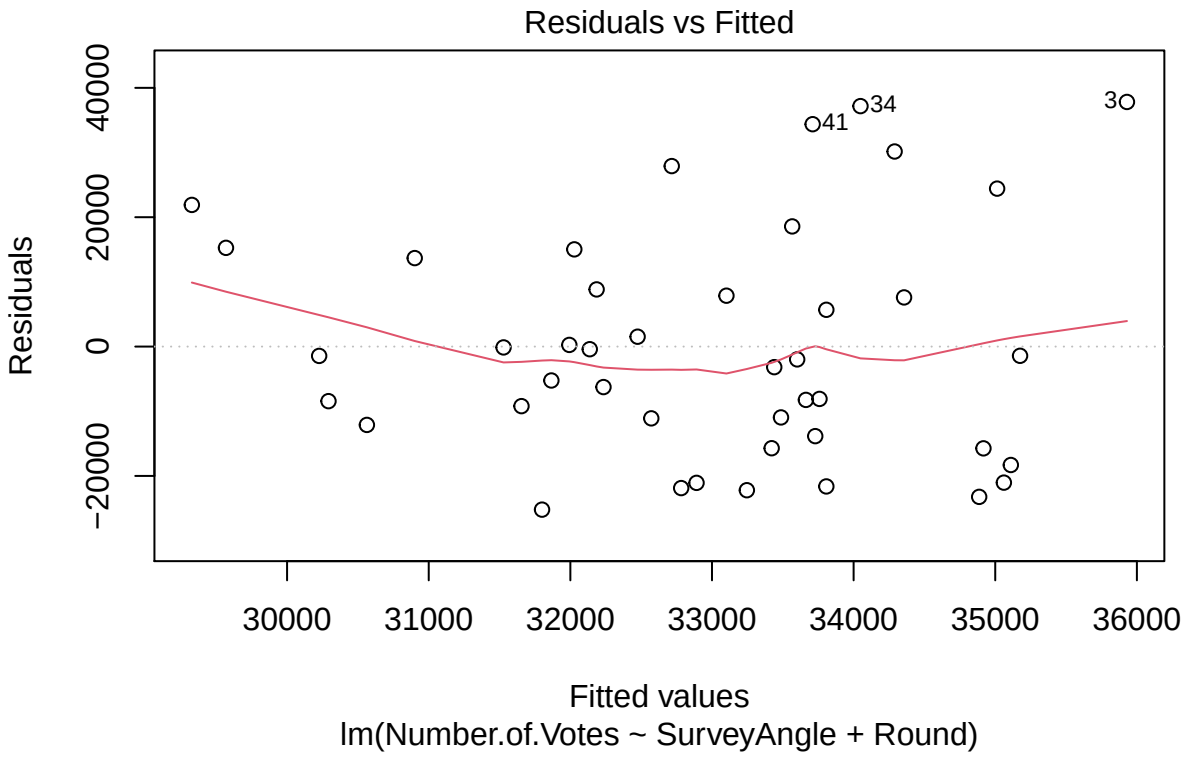


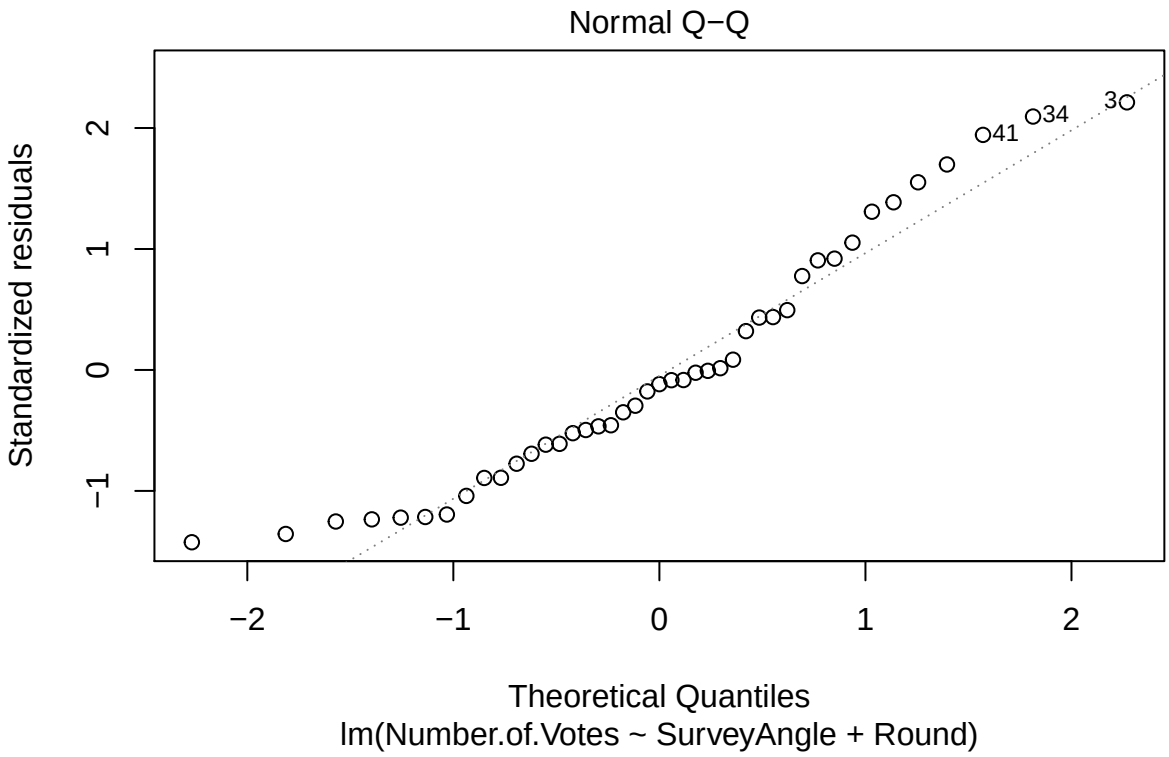

```
qplot(hyp1mod$residuals, bins = 13)
```

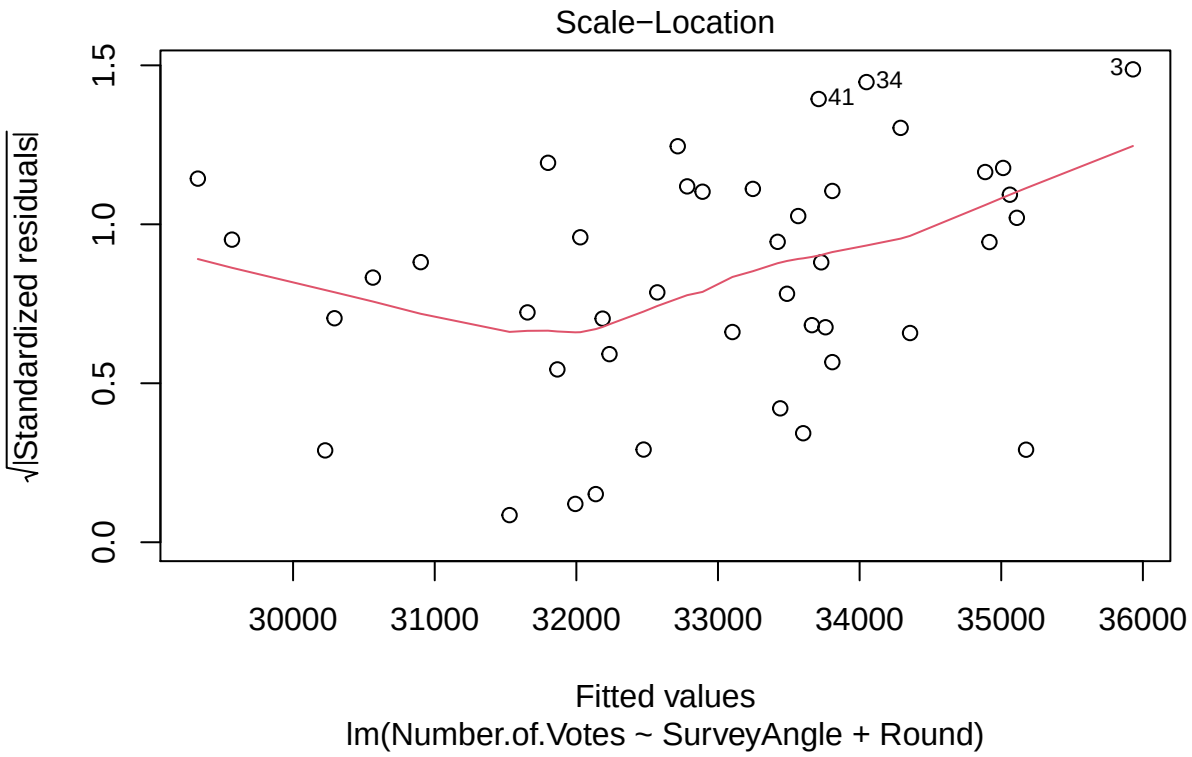


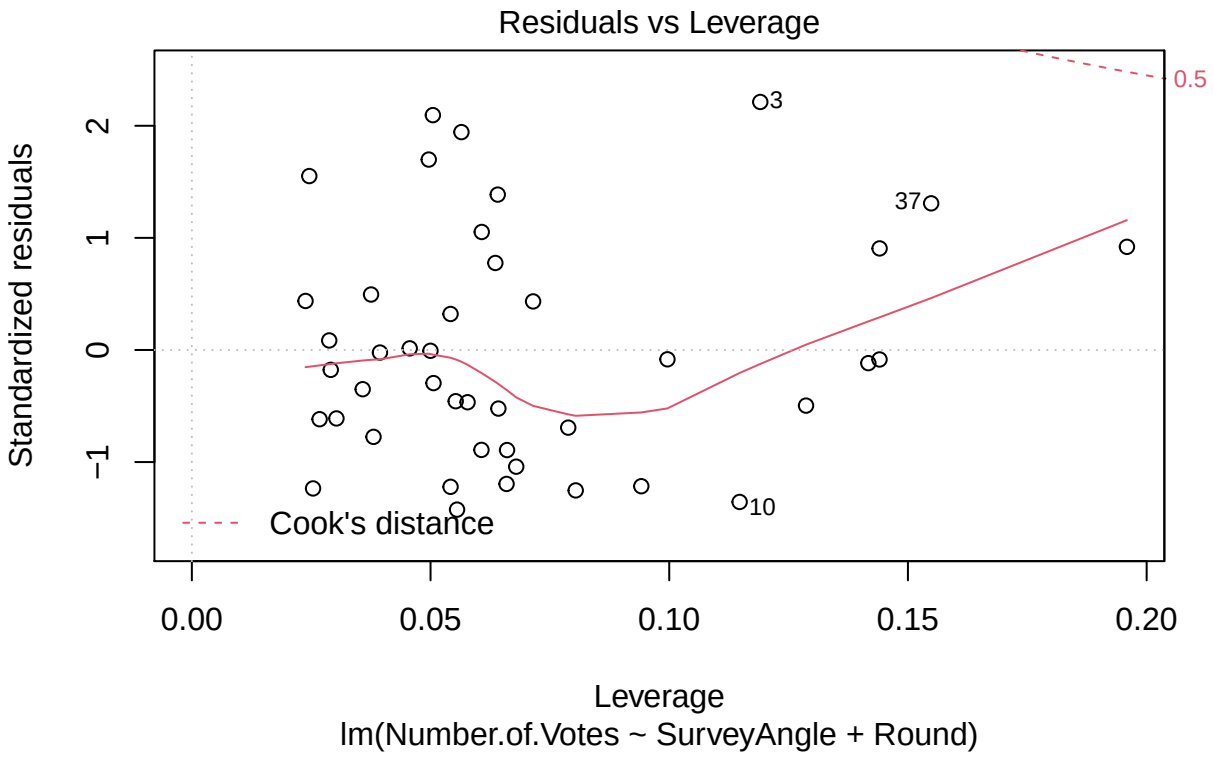
Homogeneity of errors fails. Linearity fails. Normal distribution fails kinda. Obvious outlier. Independence fails.

```
plot(hyp1modOut)
```

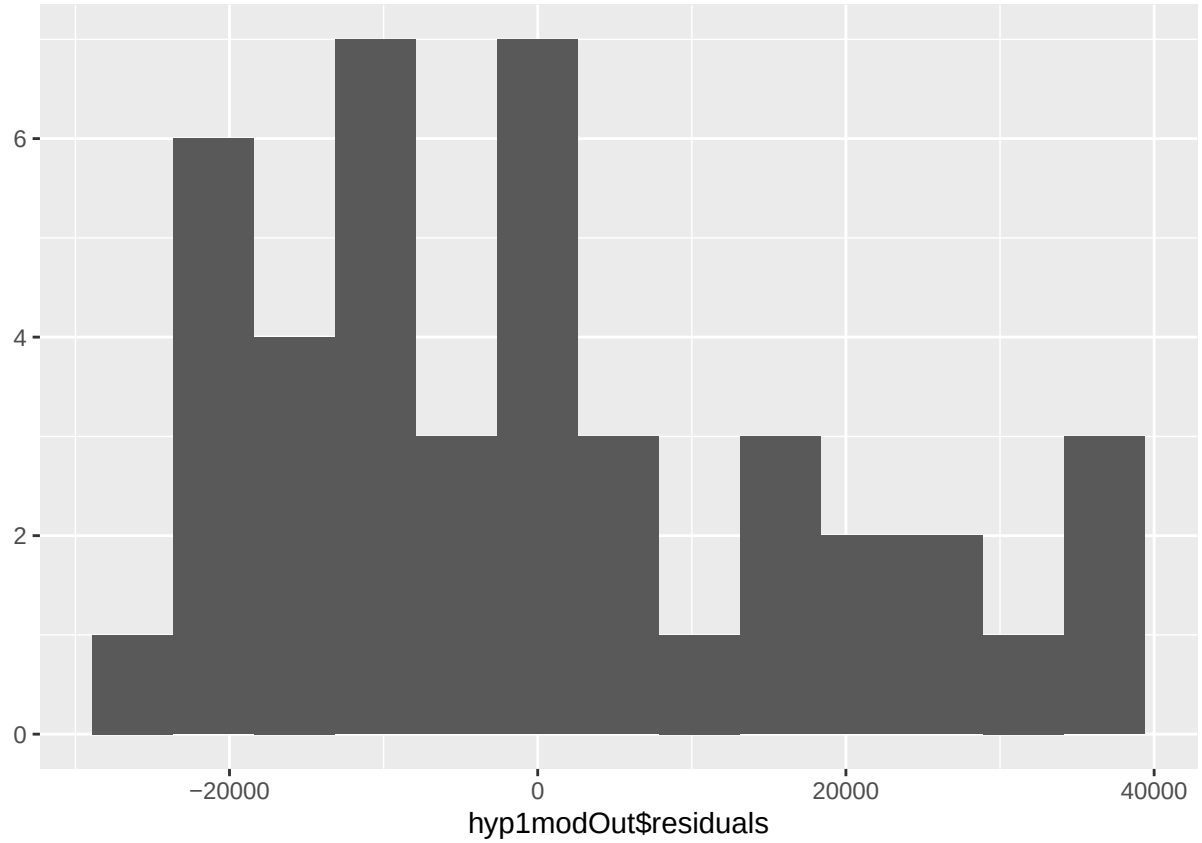

Residuals vs Fitted
Residuals
40000
20000
0
−20000
41
34
3
30000
31000
32000
33000
34000
35000
36000
Fitted values
lm(Number.of.Votes ~ SurveyAngle + Round)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Number.of.Votes ~ SurveyAngle + Round)
41
34
3
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
30000
31000
32000
33000
34000
35000
36000
3
34
41
Fitted values
lm(Number.of.Votes ~ SurveyAngle + Round)

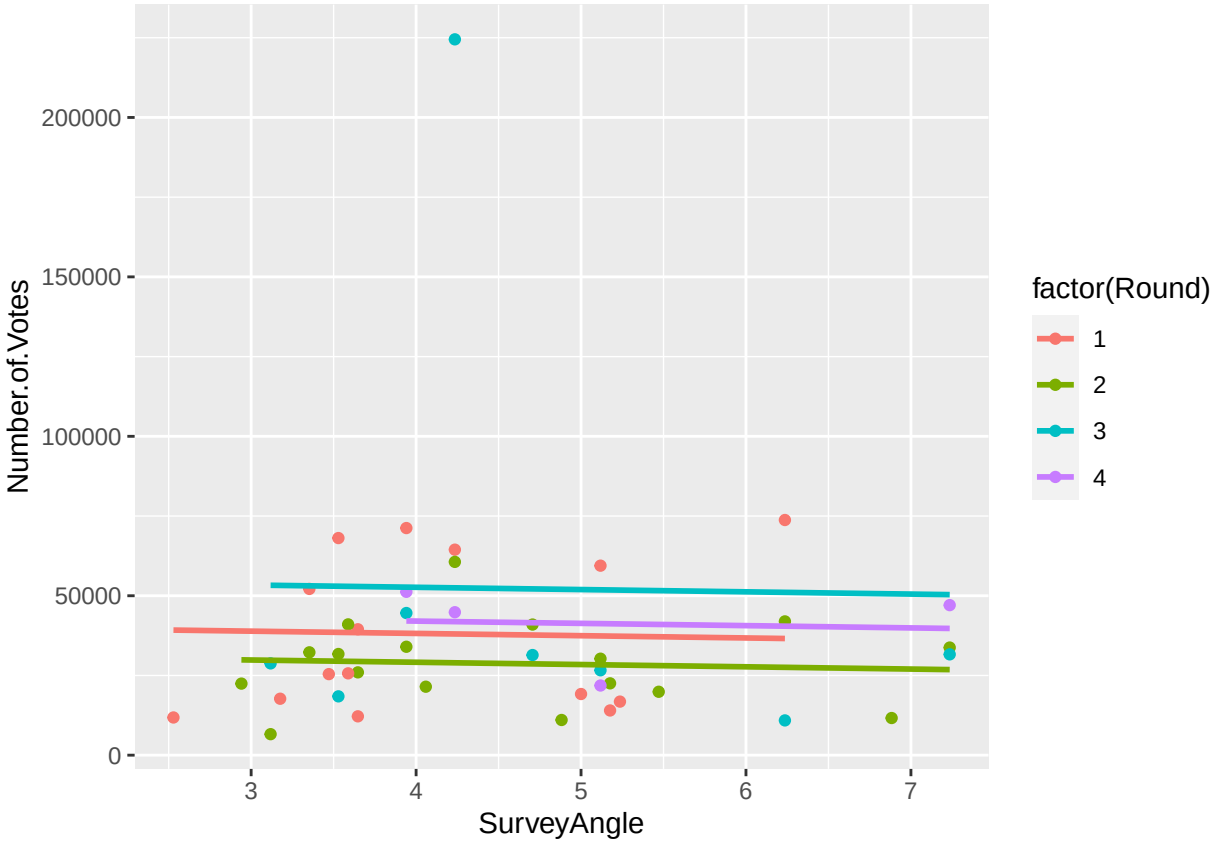



```
qplot(hyp1modOut$residuals, bins = 13)
```

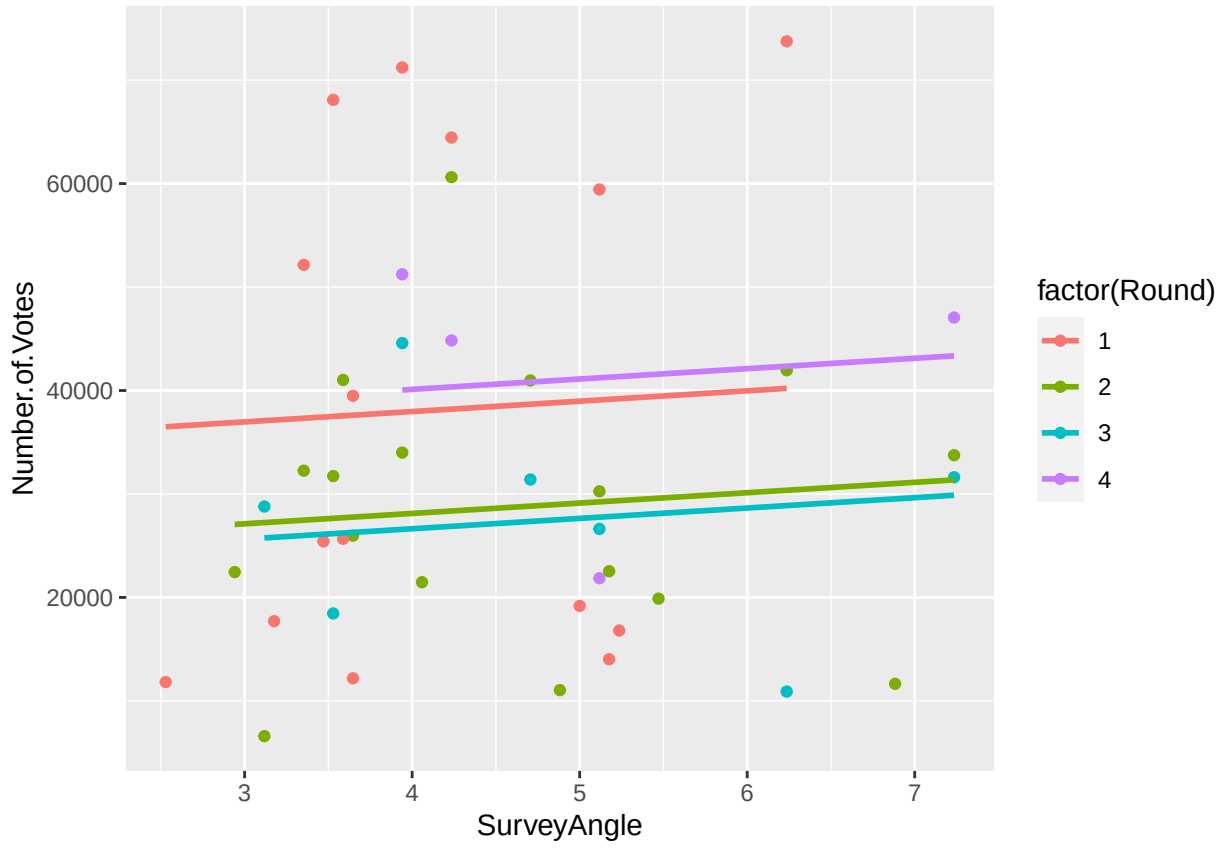


Better normality but still off. Residuals distributed above and below zero line more evenly but fans out. Linearity fails.

```
ggplot(bears, aes(x = SurveyAngle, y = Number.of.Votes, color = factor(Round))) + geom_point() +
geom_parallel_slopes(se = FALSE)
```




```
ggplot(bears_out, aes(x = SurveyAngle, y = Number.of.Votes, color = factor(Round))) + geom_point() +
geom_parallel_slopes(se = FALSE)
```




```
logitHyp2 <- glm(RoundWinBinary ~ votes_last_round, data = lagged_bears, family = binomial)
summary(logitHyp2)
```

```
##
## Call:
## glm(formula = RoundWinBinary ~ votes_last_round, family = binomial,
##     data = lagged_bears)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.290  -1.271   1.082   1.097   1.105
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.702e-01  7.141e-01   0.378    0.705
## votes_last_round -1.308e-06  1.042e-05  -0.125    0.900
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.526  on 19  degrees of freedom
## Residual deviance: 27.510  on 18  degrees of freedom
##   (24 observations deleted due to missingness)
## AIC: 31.51
##
## Number of Fisher Scoring iterations: 3
```

```
exp(cbind(OR = coef(logitHyp2), confint(logitHyp2)))
```

```
## Waiting for profiling to be done...
```

```
##                         OR     2.5 %   97.5 %
## (Intercept)      1.3102486 0.3078355 5.792590
## votes_last_round 0.9999987 0.9999742 1.000023
```

```
qnorm(p=.05/2, lower.tail=FALSE)
```

```
## [1] 1.959964
```

fitted logistic regression equation in logit form:

$$log(\frac{\widehat{\pi}_{RoundWin}}{1-\widehat{\pi}_{RoundWin}}) = 0.2702 - 1.308 * 10^{-6}(VotesLastRound)$$

The association between winning a round and number of votes obtained in previous round is not statistically significant. Specifically, the odds of a bear winning a round are 0.0000013 times lower or approximately equal for every 1 increase in vote, on average in this population. However, the z-statistic was small (-0.125) and pvalue was large (p=0.900) relative to their critical values (z=1.96, p<0.05), so we fail to reject the null hypothesis that the relationship between the odds of a bear winning a round in the tournament and the number of votes obtained by the bear in the previous round is statistically significantly. We also see that the 95% CI for this estimate (OR:0.9999987, 95% CI: 0.9999742, 1.000023) includes the null of 1.

```
bears_2 <- lagged_bears %>%
     filter(Round==2)
logitHyp2test <- glm(RoundWinBinary ~ votes_last_round, data = bears_2, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logitHyp2)
```

```
##
## Call:
## glm(formula = RoundWinBinary ~ votes_last_round, family = binomial,
##     data = lagged_bears)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.290  -1.271   1.082   1.097   1.105
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.702e-01  7.141e-01   0.378    0.705
## votes_last_round -1.308e-06  1.042e-05  -0.125    0.900
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.526  on 19  degrees of freedom
```

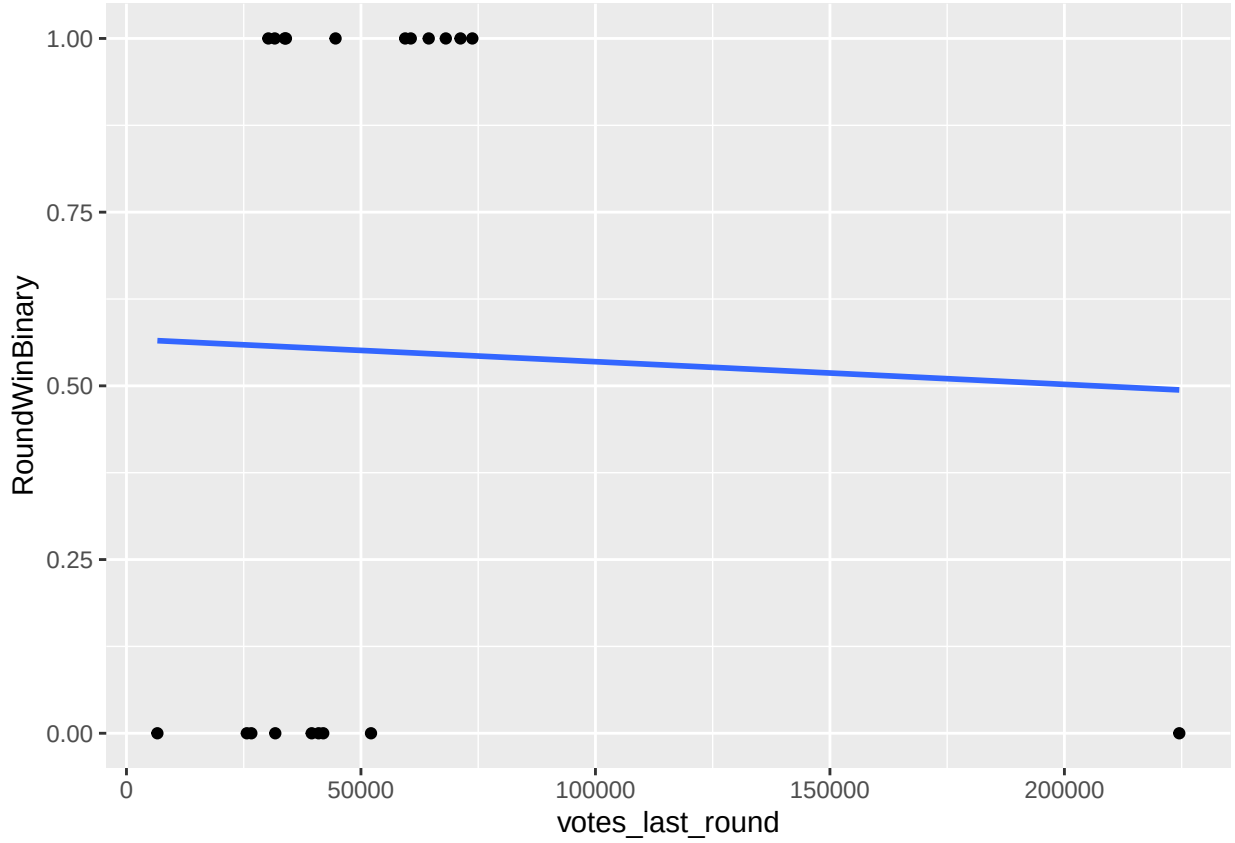
```
## Residual deviance: 27.510  on 18  degrees of freedom
##   (24 observations deleted due to missingness)
## AIC: 31.51
##
## Number of Fisher Scoring iterations: 3
```

```
qplot(x=votes_last_round, y=RoundWinBinary, data=lagged_bears) +
  geom_smooth(method="glm", method.args = list(family = "binomial"), se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 24 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 24 rows containing missing values (geom_point).
```



References

[1] Sarkar, D. (2021, October 08). A Glimpse Inside Fat Bear Week. Retrieved November 22, 2021, from https://www.discovermagazine.com/planet-earth/a-glimpse-inside-fat-bear-week

[2] Fat Bear Week 2021. (n.d.). Retrieved November 22, 2021, from https://explore.org/fat-bear-week

[3] Spencer, C. (2021, October 6). 480 Otis [Photograph found in National Park Service, Katmai National Park]. Retrieved December 8, 2021, from https://www.nps.gov/katm/learn/fat-bear-week-2021.htm (Originally photographed 2021, September 16)