STA 412 Group Project Step 3

Mark, Tanu, Pete

4/17/2020

# Facebook Data Project

Our group has decided to research the topic of “Fake News” on Facebook and how much it is reacted to and shared on the social media site. We stumbled upon this dataset on Kaggle, and with the 2020 election approaching, we thought a topic like this would be both relevant and informative. We hope we can shed light on how false information can be spread through social media.

Our dataset comes from Kaggle, and contains one csv file with over 2000 data points. The source of the data is from BuzzFeed News, who analyzed stories from hyperpartisan political Facebook pages. These pages contain right and left wing sources, and mainstream news sites, giving us a great mix of viewpoints. The dataset can be found at the following link: <https://www.kaggle.com/mrisdal/fact-checking-facebook-politics-pages>

### Import Data and Data Processing

# Read in Dataset - change to your PCs data location  
fb\_data = read.csv("Data/facebook-fact-check.csv")  
  
# Drop unnecessary columns  
fb\_data = as.data.frame(fb\_data[,c(3,4,7,8,10,11,12)])  
  
# Omit NaNs  
fb\_data = na.omit(fb\_data)  
head(fb\_data)

## Category Page Post.Type Rating share\_count  
## 2 mainstream ABC News Politics link mostly true 1  
## 3 mainstream ABC News Politics link mostly true 34  
## 4 mainstream ABC News Politics link mostly true 35  
## 5 mainstream ABC News Politics video mostly true 568  
## 6 mainstream ABC News Politics link mostly true 23  
## 7 mainstream ABC News Politics video mostly true 46  
## reaction\_count comment\_count  
## 2 33 34  
## 3 63 27  
## 4 170 86  
## 5 3188 2815  
## 6 28 21  
## 7 409 105

# Display feature types and classes  
sapply(fb\_data, mode)

## Category Page Post.Type Rating share\_count   
## "numeric" "numeric" "numeric" "numeric" "numeric"   
## reaction\_count comment\_count   
## "numeric" "numeric"

sapply(fb\_data, class)

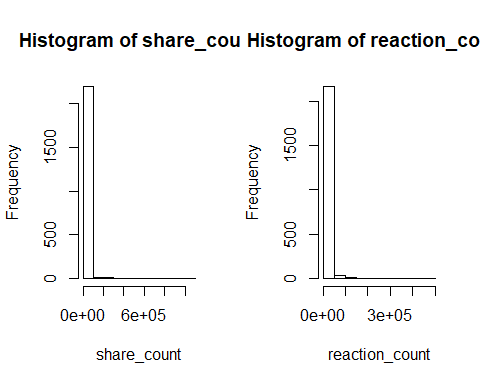
## Category Page Post.Type Rating share\_count   
## "factor" "factor" "factor" "factor" "integer"   
## reaction\_count comment\_count   
## "integer" "integer"

# Fix levels of Rating variable for plots  
  
fb\_data$Rating = factor(fb\_data$Rating, levels = c("mostly true",  
 "mixture of true and false",  
 "mostly false",  
 "no factual content"))

# Descriptive Statistics

Histogram for share\_count and reaction count

attach(fb\_data)  
par(mfrow=c(1,2))  
hist(share\_count)  
hist(reaction\_count)



Max, Min, Median values and Means for each rating for share\_count

# share\_count max  
max(share\_count)

## [1] 1088995

# share\_count min  
min(share\_count)

## [1] 1

# share\_count median  
median(share\_count)

## [1] 96

# share\_count means for each rating  
aggregate(share\_count~Rating,fb\_data,mean)

## Rating share\_count  
## 1 mostly true 1675.950  
## 2 mixture of true and false 5083.900  
## 3 mostly false 3570.301  
## 4 no factual content 19648.272

# share\_count variances for each rating  
aggregate(share\_count~Rating,fb\_data,var)

## Rating share\_count  
## 1 mostly true 123492375  
## 2 mixture of true and false 602339837  
## 3 mostly false 35339719  
## 4 no factual content 6629476927

# share\_count standard deviation for each rating  
aggregate(share\_count~Rating,fb\_data,sd)

## Rating share\_count  
## 1 mostly true 11112.712  
## 2 mixture of true and false 24542.613  
## 3 mostly false 5944.722  
## 4 no factual content 81421.600

Max, Min, Median values and Means for each rating for reaction\_count

# reaction\_count max  
max(reaction\_count)

## [1] 456458

# reaction\_count min  
min(reaction\_count)

## [1] 2

# reaction\_count median  
median(reaction\_count)

## [1] 554.5

# reaction\_count means for each rating  
aggregate(reaction\_count~Rating,fb\_data,mean)

## Rating reaction\_count  
## 1 mostly true 2934.925  
## 2 mixture of true and false 6277.158  
## 3 mostly false 5939.835  
## 4 no factual content 21491.689

# reaction\_count variances for each rating  
aggregate(reaction\_count~Rating,fb\_data,var)

## Rating reaction\_count  
## 1 mostly true 105406647  
## 2 mixture of true and false 442714116  
## 3 mostly false 106135933  
## 4 no factual content 2006644836

# reaction\_count standard deviation for each rating  
aggregate(reaction\_count~Rating,fb\_data,sd)

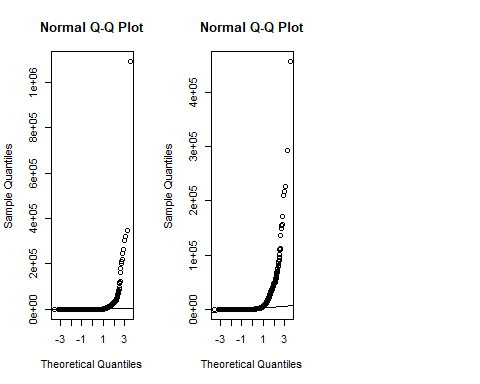
## Rating reaction\_count  
## 1 mostly true 10266.77  
## 2 mixture of true and false 21040.77  
## 3 mostly false 10302.23  
## 4 no factual content 44795.59

Q-Q Plots for Normality

par(mfrow=c(1,3))  
attach(fb\_data)

## The following objects are masked from fb\_data (pos = 3):  
##   
## Category, comment\_count, Page, Post.Type, Rating,  
## reaction\_count, share\_count

qqnorm(share\_count)  
qqline(share\_count)  
qqnorm(reaction\_count)  
qqline(reaction\_count)



Logarithmic Transformation

fb\_data$log\_share\_count= log(fb\_data$share\_count)   
fb\_data$log\_reaction\_count= log(fb\_data$reaction\_count)

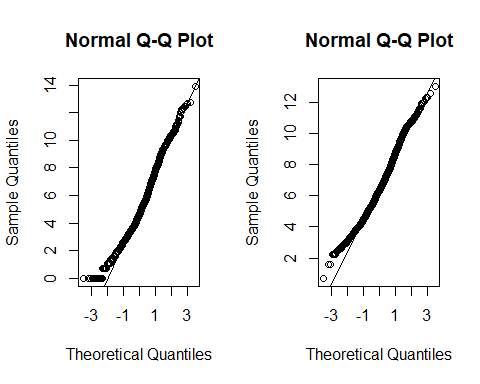
New Q-Q Plots

attach(fb\_data)

## The following objects are masked from fb\_data (pos = 3):  
##   
## Category, comment\_count, Page, Post.Type, Rating,  
## reaction\_count, share\_count

## The following objects are masked from fb\_data (pos = 4):  
##   
## Category, comment\_count, Page, Post.Type, Rating,  
## reaction\_count, share\_count

par(mfrow=c(1,2))  
qqnorm(log\_share\_count)  
qqline(log\_share\_count)  
qqnorm(log\_reaction\_count)  
qqline(log\_reaction\_count)



### Our First Hypothesis

Facebook posts, regardless of what political viewpoint, are shared and reated to differently for each level of factual content. In other words, posts with at least one level of factual content have different mean values of shares and reactions.

In order to test our first hypothesis, we will use a One-Way Analysis of Means. This will test whether the mean values of share\_count and reaction\_count are different for each individual rating.

H0: mu1 = mu2 = mu3 = mu4  
Ha: at least one muk is different

This test will show us if, on average, posts are shared, reacted to, and commented on more for each rating.

##### Hypothesis 1: Test 1

# One-Way Analysis of Means for share\_count  
  
# Test Using Logarithmic Transformation  
oneway.test(log\_share\_count~Rating, fb\_data, var.equal = F)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: log\_share\_count and Rating  
## F = 115.31, num df = 3.00, denom df = 317.64, p-value < 2.2e-16

This one-way analysis of means gives significant evidence that there is at least one different mean value of share\_count of a rating. This is significant to us because it concludes that at least one of the ratings are receiving more or less shares than the others.

##### Hypothesis 1: Test 2

# One-Way Analysis of Means for reaction\_count  
  
# Test Using Logarithmic Transformation  
oneway.test(log\_reaction\_count~Rating, fb\_data, var.equal = F)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: log\_reaction\_count and Rating  
## F = 64.653, num df = 3.00, denom df = 316.91, p-value < 2.2e-16

This one-way analysis of means gives significant evidence that there is at least one different mean value of reaction\_count of a rating. This is significant to us because it concludes that at least one of the ratings are receiving more or less reactions than the others.

#### First Hypotheses Conclusion

We have significant evidence to reject the null hypothesis in each test. This is great for our research, in that it concludes that based on the rating of the post, it may be shared or reacted to more or less than posts with other ratings. We still have yet to find out how a rating of false or true may influence these response variables, however we will address this in the next hypotheses.

### Our Second Hypothesis

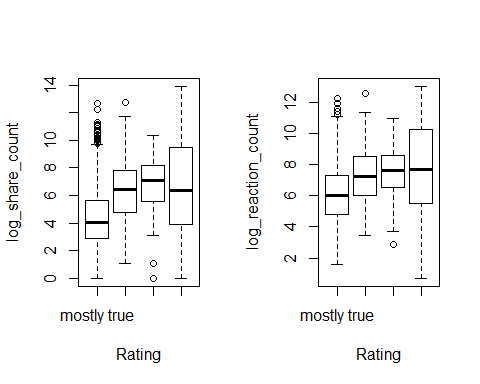
For our second question, we want to see how the factual content rating levels affect the number of shares and reactions Facebook posts receive. So our main question is, can the factual content rating level of a Facebook post determine how many shares and reactions it receives? First, we will use the F Test for Nested Models to determine if a linear regression model that includes the Rating predictor is more significant than a linear regression model with no predictors. We will conduct this for both of our response variables. If the conclusion of that test is that the model with the Rating predictor is significant, then we will see how the individual ratings affect the count of shares and reactions.

H0: Beta\_1 = 0 Ha: Beta\_1 does not equal zero

Our significance level (alpha) will be 0.05

### Plots for Linear Relationships

par(mfrow=c(1,2))  
plot(log\_share\_count~Rating, data=fb\_data)  
plot(log\_reaction\_count~Rating, data=fb\_data)



### F Test for Nested Models

F Test for Nested Models -> share\_count

shares.lm = lm(log\_share\_count ~ Rating, data = fb\_data)  
shares\_full = shares.lm  
shares\_reduced = lm(log\_share\_count ~ 1, data = fb\_data)  
anova(shares\_reduced, shares\_full)

## Analysis of Variance Table  
##   
## Model 1: log\_share\_count ~ 1  
## Model 2: log\_share\_count ~ Rating  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 2211 13690   
## 2 2208 11865 3 1825.4 113.23 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

F Test for Nested Models -> reaction\_count

reactions.lm = lm(log\_reaction\_count ~ Rating, data = fb\_data)  
reactions\_full = reactions.lm  
reactions\_reduced = lm(log\_reaction\_count ~ 1, data = fb\_data)  
anova(reactions\_reduced, reactions\_full)

## Analysis of Variance Table  
##   
## Model 1: log\_reaction\_count ~ 1  
## Model 2: log\_reaction\_count ~ Rating  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 2211 8980.5   
## 2 2208 8160.7 3 819.76 73.933 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Linear Regression

# Summary of shares.lm  
summary(shares.lm)

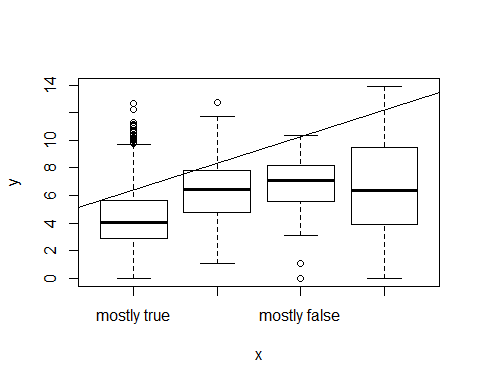
##   
## Call:  
## lm(formula = log\_share\_count ~ Rating, data = fb\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.8712 -1.6076 -0.2898 1.3857 8.2434   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.44082 0.05736 77.42 <2e-16 \*\*\*  
## Ratingmixture of true and false 1.93857 0.15996 12.12 <2e-16 \*\*\*  
## Ratingmostly false 2.43038 0.23550 10.32 <2e-16 \*\*\*  
## Ratingno factual content 2.01276 0.16173 12.45 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.318 on 2208 degrees of freedom  
## Multiple R-squared: 0.1333, Adjusted R-squared: 0.1322   
## F-statistic: 113.2 on 3 and 2208 DF, p-value: < 2.2e-16

contrasts(fb\_data$Rating)

## mixture of true and false mostly false  
## mostly true 0 0  
## mixture of true and false 1 0  
## mostly false 0 1  
## no factual content 0 0  
## no factual content  
## mostly true 0  
## mixture of true and false 0  
## mostly false 0  
## no factual content 1

# Plot shares.lm  
plot(fb\_data$Rating, fb\_data$log\_share\_count)  
abline(shares.lm)

## Warning in abline(shares.lm): only using the first two of 4 regression  
## coefficients



# Summary of reactions.lm  
summary(reactions.lm)

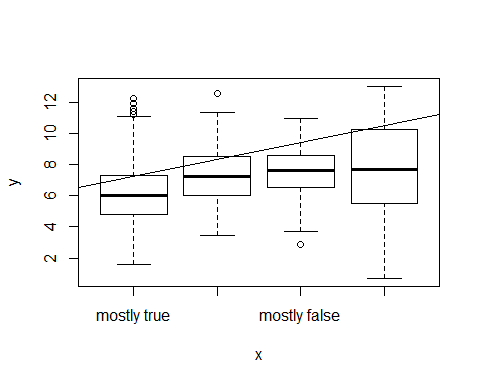
##   
## Call:  
## lm(formula = log\_reaction\_count ~ Rating, data = fb\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.0569 -1.3885 -0.1184 1.3116 6.1083   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.14663 0.04757 129.201 < 2e-16 \*\*\*  
## Ratingmixture of true and false 1.09514 0.13266 8.255 2.58e-16 \*\*\*  
## Ratingmostly false 1.41482 0.19531 7.244 5.99e-13 \*\*\*  
## Ratingno factual content 1.60340 0.13413 11.954 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.922 on 2208 degrees of freedom  
## Multiple R-squared: 0.09128, Adjusted R-squared: 0.09005   
## F-statistic: 73.93 on 3 and 2208 DF, p-value: < 2.2e-16

contrasts(fb\_data$Rating)

## mixture of true and false mostly false  
## mostly true 0 0  
## mixture of true and false 1 0  
## mostly false 0 1  
## no factual content 0 0  
## no factual content  
## mostly true 0  
## mixture of true and false 0  
## mostly false 0  
## no factual content 1

# Plot reactions.lm  
plot(fb\_data$Rating, fb\_data$log\_reaction\_count)  
abline(reactions.lm)

## Warning in abline(reactions.lm): only using the first two of 4 regression  
## coefficients



### Confidence Intervals

Confidence Intervals for share\_count model

# 95% CI for beta0 and beta1   
confint(shares.lm)

## 2.5 % 97.5 %  
## (Intercept) 4.328326 4.553311  
## Ratingmixture of true and false 1.624885 2.252264  
## Ratingmostly false 1.968551 2.892206  
## Ratingno factual content 1.695602 2.329922

Confidence Interval for reaction\_count model

# 95% CI for beta0 and beta1   
confint(reactions.lm)

## 2.5 % 97.5 %  
## (Intercept) 6.0533386 6.239928  
## Ratingmixture of true and false 0.8349841 1.355296  
## Ratingmostly false 1.0318114 1.797837  
## Ratingno factual content 1.3403635 1.866431