Sophie Thesis Update Tues Jun 21, 2016

Examining Dependent and Independent Variables, Correlations, and Data Transforms

[Note to self: iPython Notebook: DataTransforms]

Recap

Main Questions from Last Meeting:

1. What's up with small R squared value?
   a. I performed several data transformations on the data (see results below). Results are significant (p-values are very small), but overall the R^2 values are also small. Those data transformations helped make larger R^2.
   b. I think the small R^2 is not a reason to not report the results; what we are seeing is that I'm trying to make a direct linear correlation between one variable at a time and tweet volume; it makes sense that that one factor only explains a small portion of the correlation between X and Y. Small pseudo-R^2 values are also reported in Milkman's study (0.0, 0.04, 0.07...0.36).

Open Questions:

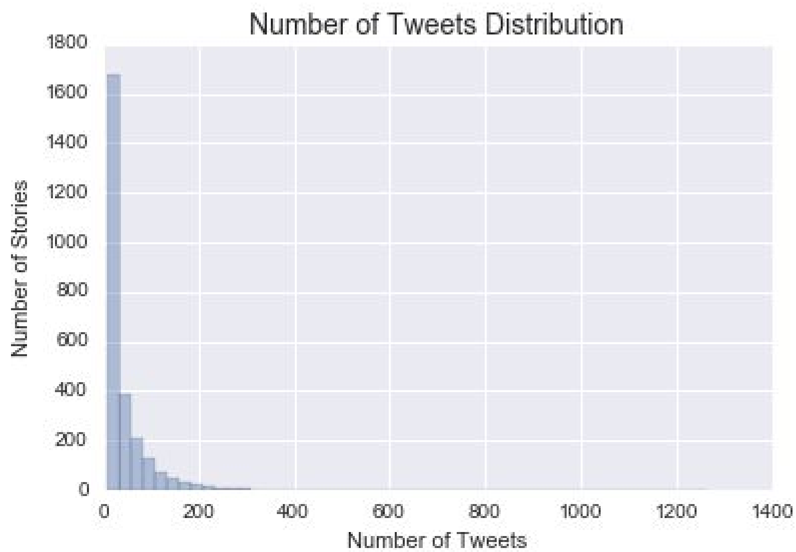
1. Are these independent variables normal?
   a. Emotionality, story length don't look normal
2. Can we assume Y is a normal distribution? -- No
3. Y (tweet volume) is a power-law distribution -- or lognormal? Testing MLE not significantly power law.
4. How can we model given that it's lognormal/power law?
   a. glm/glmer package-- use family log normal
   b. Take log(y) and plot-- if it looks close to normal, transform all X & Y and then use linear model (lm(log(y) ~ log(x))
5. How to interpret transformations correctly?

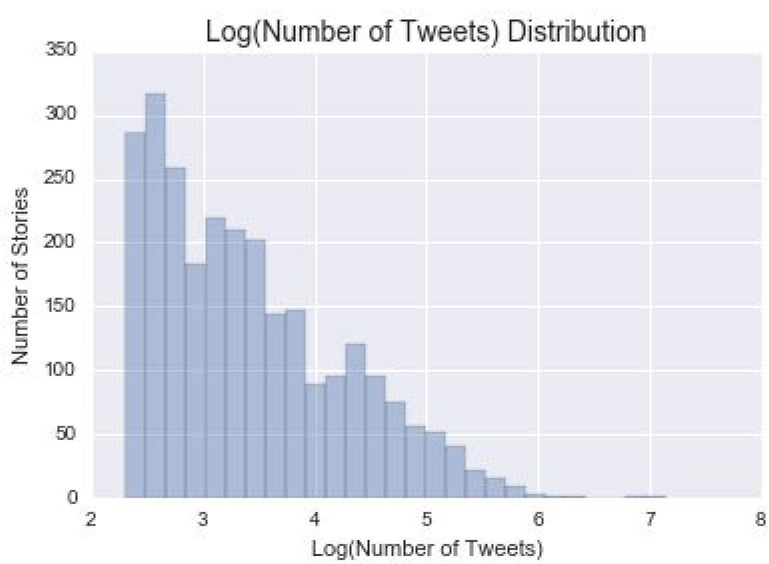
Next steps:

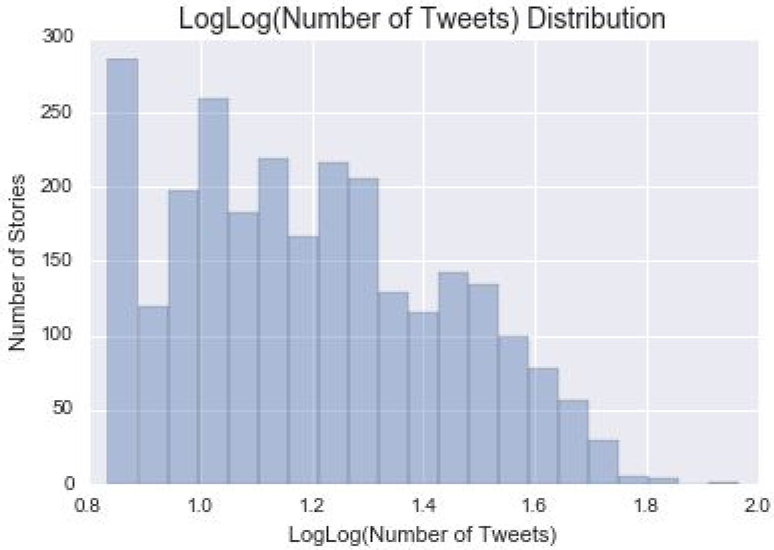
1. Log(number of tweets) looks more like lognormal family, try generalized linear model with family lognormal.
2. Write chapters about data pipeline and motivation

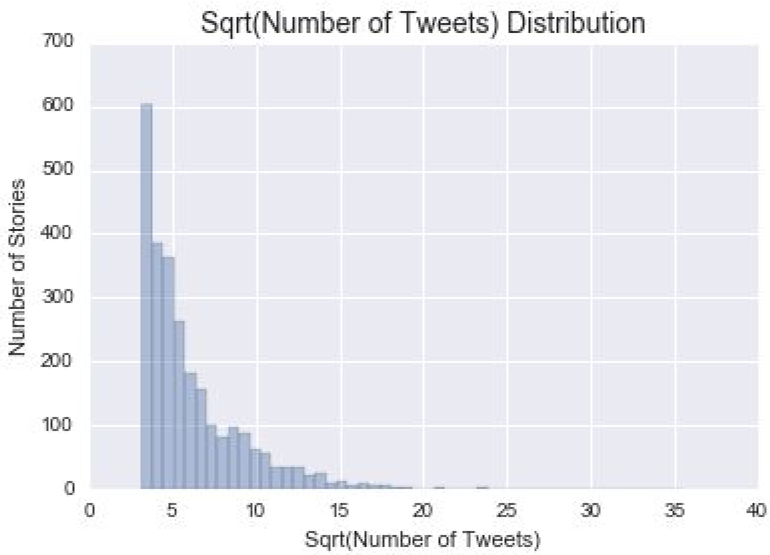
Dependent Value, Tweet Volume

Number of tweets in our dataset: 2.6 K (with 10-tweet cutoff)
Average number of times a story is tweeted: 46.3
Maximum number of times a story is tweeted: 1261
Standard deviation: 60.2









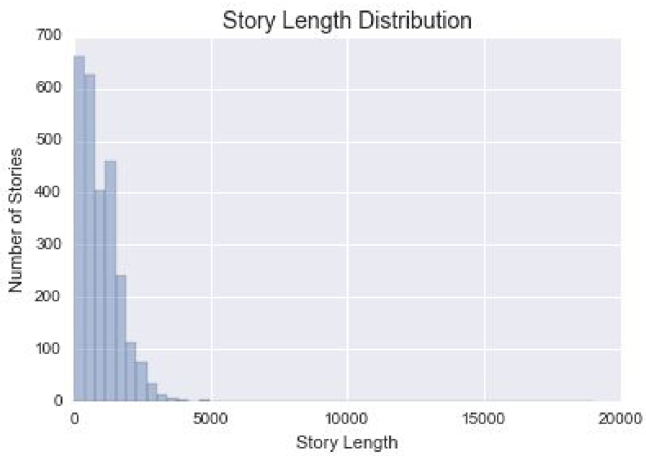
Power law dist? Lognormal?
MLE Power Law / MLE lognormal dist

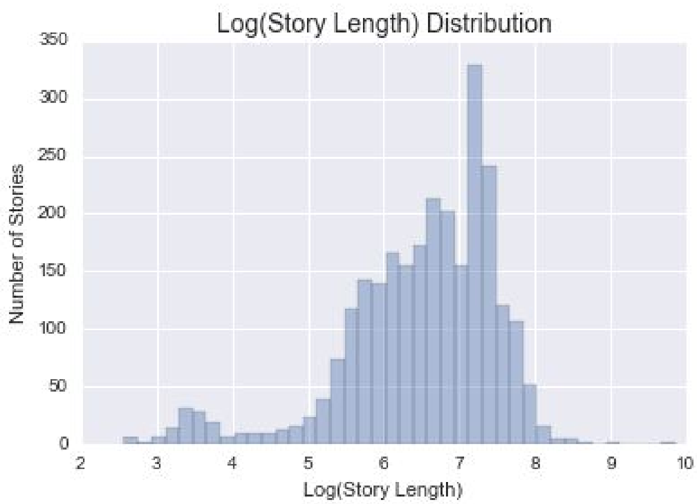
R 0.202683453995
p 0.839382456554

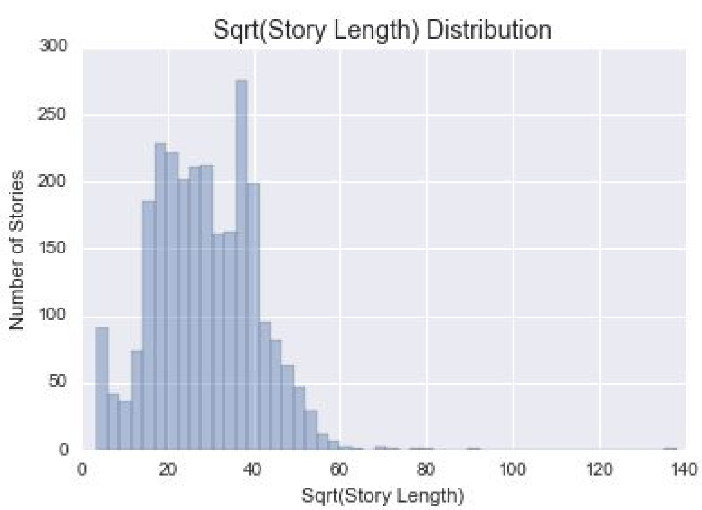
Not significantly more likely to be Power Law than Lognormal.
Analyse as lognormal?

Independent Variables & Data Transforms

Story Length (Word count)



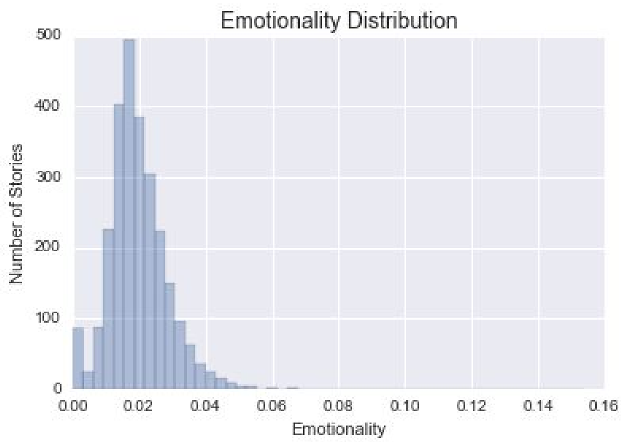


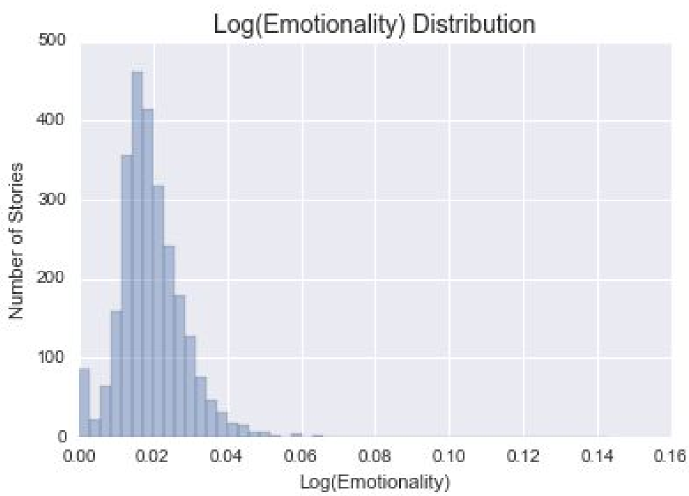


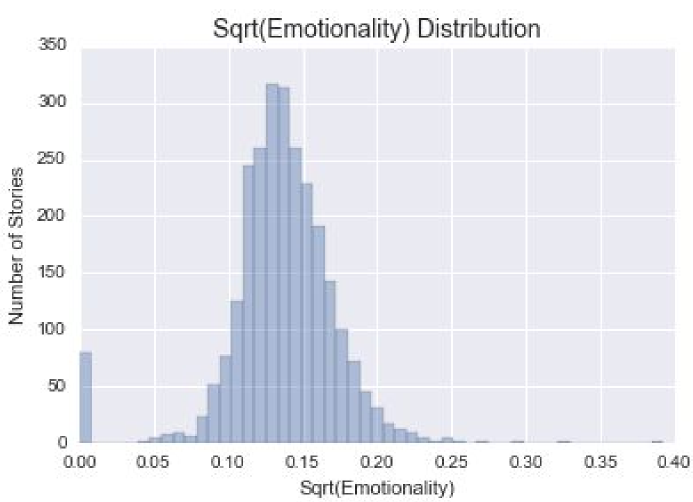
Does this look more normal? Statistical tests?

Emotionality

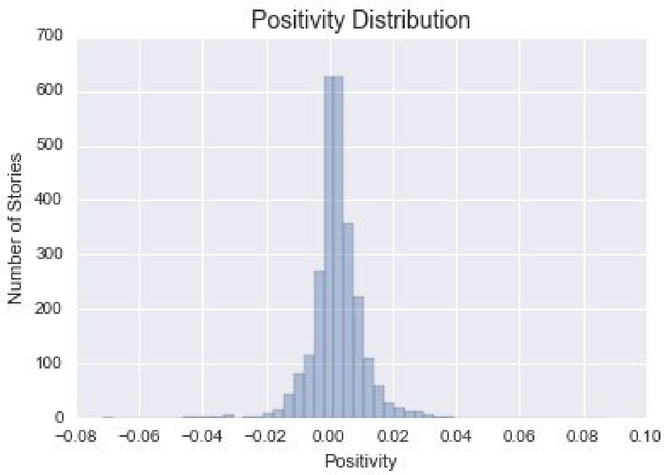
(Percent of words that are either positive or negative in an article)

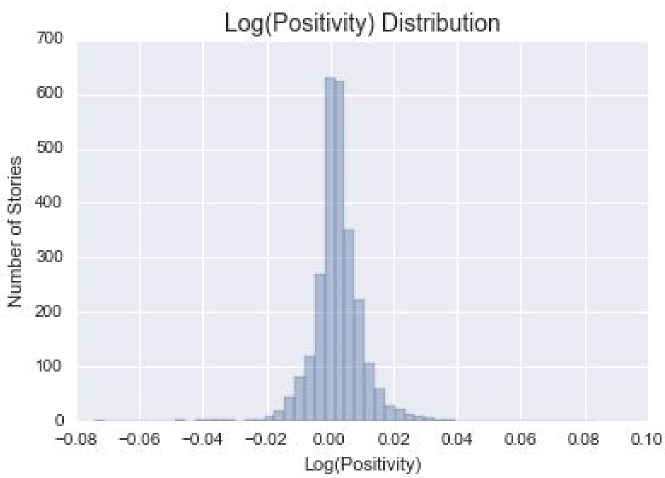


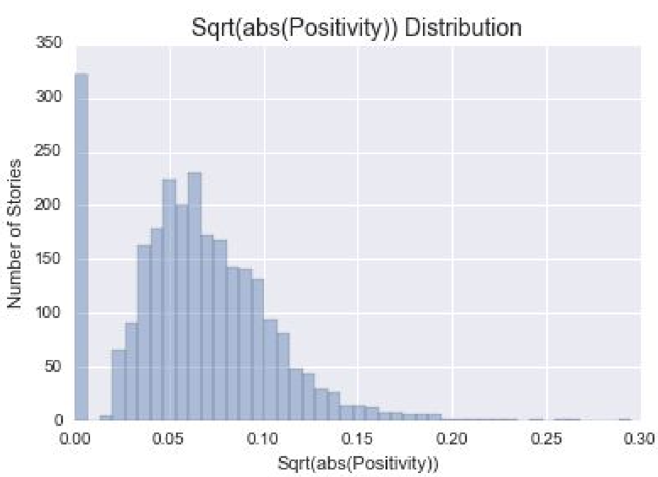




Positivity
(Percent difference between positive and negative words in an article)







Summary of Findings: (with simple univariate linear model; data transformations)

Q1: Does the length of a story have a correlation with how many times it's shared on Twitter?

Answer: Yes. There is a significant negative correlation between story length and the number of times it's shared.

Possible explanation: people have short attention spans.
Shorter stories have different content than longer ones.

With Log-Log Transform:
(more significance plus increase in R^2 but R^2 still small)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.99898    0.10634  37.604  < 2e-16 ***
log_wc      -0.09031    0.01616  -5.589 2.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.8428 on 2648 degrees of freedom
Multiple R-squared: 0.01166, Adjusted R-squared: 0.01128
F-statistic: 31.23 on 1 and 2648 DF, p-value: 2.521e-08

Q2: Does the emotionality of a story have a correlation with how many times it's shared on Twitter?

Answer: *Significant positive correlation* between emotionality and number of shares.
Possible explanation: high emotional content gets more shares!

With Log-Log Transform

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0228177  0.0011690  19.518   <2e-16 ***
log_wc      -0.0005089  0.0001776  -2.865   0.0042 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.009265 on 2648 degrees of freedom
Multiple R-squared: 0.00309, Adjusted R-squared: 0.002713
F-statistic: 8.208 on 1 and 2648 DF, p-value: 0.004204

Q3: Does the positivity of a story have a correlation with how many times it's shared on Twitter?

Answer: *Significant negative correlation* between positivity and number of shares.
Possible explanation: negative content gets more shares! This supports the hate-linking/negativity of the internet research Q/hypothesis.

With Log-Log Transform

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0228177  0.0011690  19.518   <2e-16 ***
log_wc      -0.0005089  0.0001776  -2.865   0.0042 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009265 on 2648 degrees of freedom
Multiple R-squared:  0.00309,	Adjusted R-squared:  0.002713
F-statistic: 8.208 on 1 and 2648 DF,  p-value: 0.004204
```

Testing for Correlations: full report

Story Length x Tweet Volume

Significant negative correlation between length of story and number of shares.
Possible explanation: people have short attention spans

```
Call:
lm(formula = num_tweets ~ wc, data = stories)

Residuals:
    Min      1Q  Median      3Q     Max 
 -40.32  -30.90  -20.22    5.99 1215.21 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 50.425959   1.818541  27.729  < 2e-16 ***
wc          -0.004180   0.001438  -2.907  0.00368 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.16 on 2648 degrees of freedom
Multiple R-squared:  0.003182,	Adjusted R-squared:  0.002805 
F-statistic: 8.452 on 1 and 2648 DF,  p-value: 0.003677
```

Log Transform:

```
lm(formula = num_tweets ~ log_wc, data = stories)

Residuals:
    Min      1Q  Median      3Q     Max 
 -50.04  -30.40  -20.24    5.85 1216.74 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   73.432      7.583   9.683  < 2e-16 ***
log_wc        -4.161      1.152  -3.611 0.000311 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.1 on 2648 degrees of freedom
Multiple R-squared:  0.0049,	Adjusted R-squared:  0.004524 
F-statistic: 13.04 on 1 and 2648 DF,  p-value: 0.0003108
```

Sqrt Transform:

```
Call:
lm(formula = num_tweets ~ sqrt_wc, data = stories)

Residuals:
    Min      1Q  Median      3Q     Max 
 -44.78  -30.37  -20.19    5.58 1216.23 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 56.55345    3.04254  18.588  < 2e-16 ***
sqrt_wc     -0.35407    0.09774  -3.623 0.000297 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.1 on 2648 degrees of freedom
Multiple R-squared:  0.004932,	Adjusted R-squared:  0.004556 
F-statistic: 13.12 on 1 and 2648 DF,  p-value: 0.000297
```

Log-Log Transform:
(more significance plus increase in R^2 but R^2 still small)

```
lm(formula = log_num_tweets ~ log_wc, data = stories)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4057 -0.6953 -0.1635  0.5392  3.7737 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.99898    0.10634  37.604  < 2e-16 ***
log_wc      -0.09031    0.01616  -5.589 2.52e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8428 on 2648 degrees of freedom
```

**Multiple R-squared: 0.01166,	Adjusted R-squared: 0.01128**

```
F-statistic: 31.23 on 1 and 2648 DF,  p-value: 2.521e-08
```

<u>Emotionality x Tweet Volume</u>

Significant positive correlation between emotionality and number of shares.
Possible explanation: high emotional content gets more shares!

```
Call:
lm(formula = num_tweets ~ emotionality, data = stories)

Residuals:
    Min      1Q  Median      3Q     Max 
 -50.89  -30.98  -20.32    5.47 1214.59 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    40.349      2.685  15.027   <2e-16 ***
emotionality  305.229    122.427   2.493   0.0127 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.18 on 2648 degrees of freedom
Multiple R-squared:  0.002342,	Adjusted R-squared:  0.001965 
F-statistic: 6.216 on 1 and 2648 DF,  p-value: 0.01272
```

Log Transform:

```
lm(formula = num_tweets ~ log_emot, data = stories)

Residuals:
    Min      1Q  Median      3Q     Max 
 -50.36  -31.00  -20.32    5.46 1214.58 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   40.270      2.722  14.792   <2e-16 ***
log_emot     312.932    126.028   2.483   0.0131 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.18 on 2648 degrees of freedom
Multiple R-squared:  0.002323,	Adjusted R-squared:  0.001946 
F-statistic: 6.165 on 1 and 2648 DF,  p-value: 0.01309
```

LogLog Transform:

```
Call:
lm(formula = log_emot ~ log_wc, data = stories)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.021475 -0.005193 -0.001039  0.004456  0.121588 
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0228177  0.0011690  19.518   <2e-16 ***
log_wc      -0.0005089  0.0001776  -2.865   0.0042 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009265 on 2648 degrees of freedom
```

Multiple R-squared: 0.00309, Adjusted R-squared: 0.002713
F-statistic: 8.208 on 1 and 2648 DF, p-value: 0.004204

Square Root Transform:

```
Call:
lm(formula = num_tweets ~ sqrt_emot, data = stories)

Residuals:
    Min      1Q  Median      3Q     Max 
 -40.47  -31.25  -20.39    5.15 1214.32 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.882      4.397   8.843   <2e-16 ***
sqrt_emot     55.320     31.293   1.768   0.0772 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.22 on 2648 degrees of freedom
Multiple R-squared:  0.001179,	Adjusted R-squared:  0.0008016 
F-statistic: 3.125 on 1 and 2648 DF,  p-value: 0.0772
```

<u>Positivity X Tweet Volume</u>

Significant negative correlation between positivity and number of shares.
Possible explanation: negative content gets more shares! This supports the hate-linking/negativity of the internet research Q/hypothesis.

```
lm(formula = num_tweets ~ positivity, data = stories)

Residuals:
    Min      1Q  Median      3Q     Max 
 -46.27  -31.08  -20.38    5.41 1215.44 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   47.078      1.221  38.572   <2e-16 ***
positivity  -281.010    139.546  -2.014   0.0441 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.21 on 2648 degrees of freedom
Multiple R-squared:  0.001529,	Adjusted R-squared:  0.001152
F-statistic: 4.055 on 1 and 2648 DF,  p-value: 0.04414
```

Log Transform

```
Call:
lm(formula = num_tweets ~ log_pos, data = stories)

Residuals:
    Min      1Q  Median      3Q     Max
 -46.88  -31.08  -20.37    5.39 1215.46

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   47.076      1.219  38.612   <2e-16 ***
log_pos     -284.450    139.834  -2.034    0.042 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.2 on 2648 degrees of freedom
Multiple R-squared:  0.00156,	Adjusted R-squared:  0.001183
F-statistic: 4.138 on 1 and 2648 DF,  p-value: 0.04203
```

Log-Log Transform

```
Call:
lm(formula = log_emot ~ log_wc, data = stories)

Residuals:
      Min        1Q    Median        3Q       Max
-0.021475 -0.005193 -0.001039  0.004456  0.121588

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0228177  0.0011690  19.518   <2e-16 ***
log_wc      -0.0005089  0.0001776  -2.865   0.0042 **
---
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009265 on 2648 degrees of freedom
Multiple R-squared: 0.00309,   Adjusted R-squared: 0.002713
F-statistic: 8.208 on 1 and 2648 DF, p-value: 0.004204