Thesis Update Tues Jun 28 2016

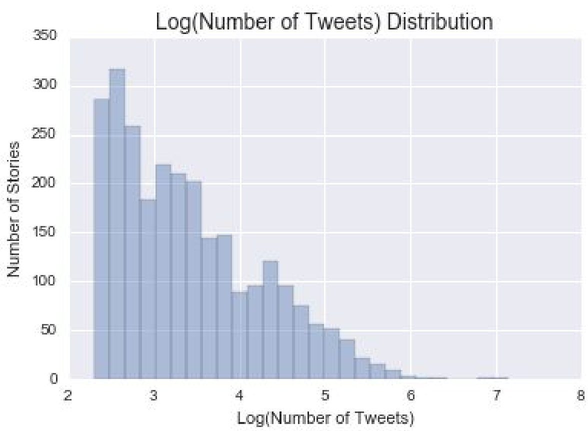
Part I

Negative Binomial Regressions

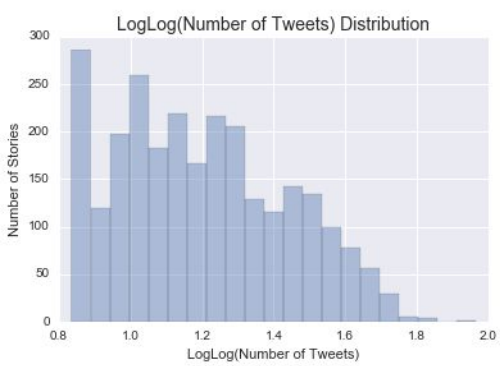
Part 0: Recap

As you may recall from last week, one of the difficulties I had with modelling the relationship between tweet volume (number of story shares) and the different factors was that the number of tweets did not follow a normal distribution and the R^2 values calculate for each regression were very small.

Even after a log transformation the data still did not appear normal and our results were still not very strong.



Although a log(log) transformation made the data appear more normal,



I was hesitant to do this for sake of interpretation.

Because of times a story is shared is a *discrete count*, and we see overdispersion and skew in our data, it makes more sense to analyze it as a negative binomial distribution, which is commonly used for counts data that's overdispersed.

(Poisson models are a subset of negative binomial models without the dispersion parameter. We know we have overdispersed data, as the dispersion parameter > 1 and also the negative binomial provides a better fit than Poisson.)

Below, recalculate each of the correlations as negative binomial:*

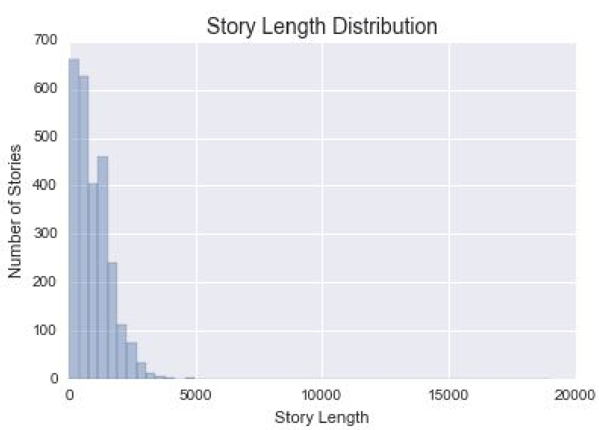
Part II: Summary of Results (Negative Binomial)

Note: When modelling using negative binomial regression a directly analogous R2 is not available and such comparisons are not possible.

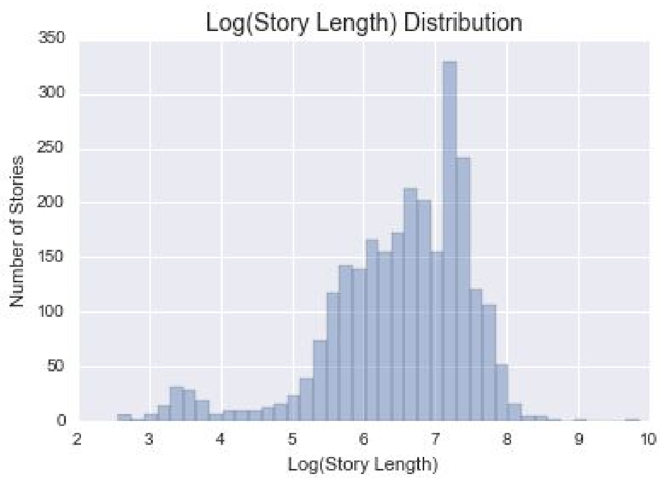
<u>Story Length</u>

Note on story length:

Word count of stories also follows a similar distribution (not normal), as it's also count data.



In this case applying a log transformation to the data yields a more normal distribution



So we apply that transformation to the independent variable, story length.

Negative Binomial Model, Number of Tweets vs Log(story length)

```
glm.nb(formula = num_tweets ~ log(wc), data = stories, control = glm.control(maxit = 100),
    init.theta = 1.351900484, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5919  -1.0711  -0.6085   0.1396   8.0219

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.40668    0.10991  40.092  < 2e-16 ***
log(wc)     -0.08826    0.01671  -5.283 1.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Dispersion parameter for Negative Binomial(**1.3519**) family taken to be 1)

```
 Null deviance: 2941.6  on 2649  degrees of freedom
Residual deviance: 2913.8  on 2648  degrees of freedom
AIC: 25548

Number of Fisher Scoring iterations: 1
```

Theta: 1.3519
Std. Err.: 0.0346

2 x log-likelihood: -25541.5120

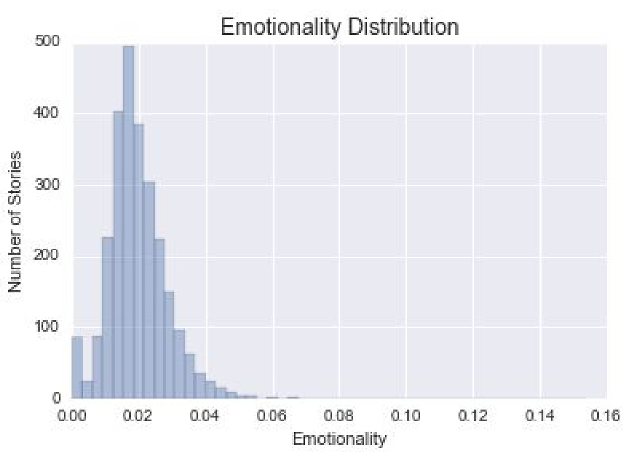
Goodness of Fit:
1 - pchisq(model deviance, residual degrees of freedom) = 0.0001959767
We can reject the null hypothesis that the model is no better than the null model.

Comparing to Linear Model:
Negative binomial performs significantly better, use chi-squared test:

2 * [log likelihood(negative binomal model(story length) - log likelihood(linear model(story length)] =
3686.066 (degrees of freedom=3)

We can reject null hypothesis with p ~ 0.

Emotionality



Negative Binomial Model, Number of Tweets vs Emotionality

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.4859	-1.0705	-0.6031	0.1306	7.7904

Coefficients:

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.716      0.039  95.276  < 2e-16 ***
emotionality    6.019      1.777   3.388 0.000705 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.3452) family taken to be 1)

    Null deviance: 2927.6  on 2649  degrees of freedom
Residual deviance: 2915.4  on 2648  degrees of freedom
AIC: 25563

Number of Fisher Scoring iterations: 1


              Theta:  1.3452
          Std. Err.:  0.0344

 2 x log-likelihood:  -25557.0260
```

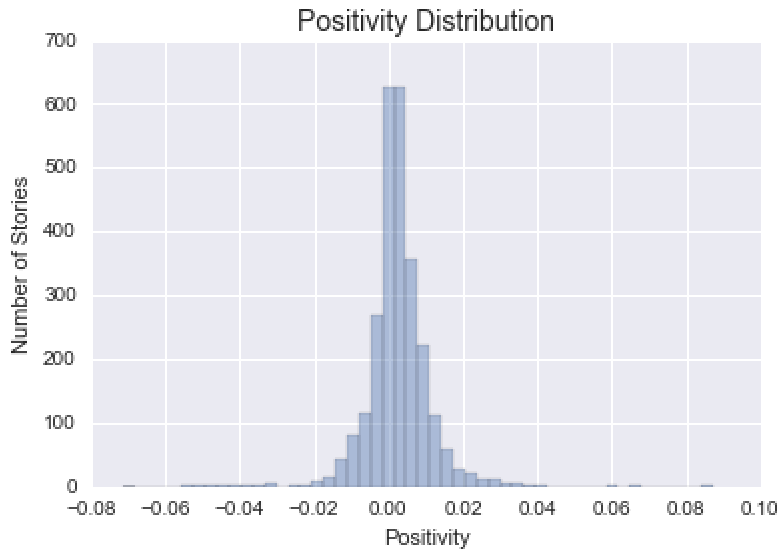
Goodness of Fit:
1 - pchisq(model deviance, residual degrees of freedom) = 0.0001810102
We can reject the null hypothesis that the model is no better than the null model.

Comparing to Linear Model:
Negative binomial performs significantly better, use chi-squared test:

2 * [log likelihood(negative binomal model(emotionality) - log likelihood(linear model(emotionality)] =
3677.356 (df=3)

We can reject null hypothesis with p ~ 0.

Positivity



Negative Binomial Model, Number of Tweets vs Positivity

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.4567	-1.0696	-0.6086	0.1272	7.8548

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.84910	0.01774	217.017	< 2e-16	***
positivity	**-5.39140**	2.02900	-2.657	0.00788	**

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

(Dispersion parameter for Negative Binomial(**1.3434**) family taken to be 1)

Null deviance: 2923.8 on 2649 degrees of freedom
Residual deviance: 2915.8 on 2648 degrees of freedom
AIC: 25567

Number of Fisher Scoring iterations: 1

Theta: 1.3434
Std. Err.: 0.0344

2 x log-likelihood: -25561.2700

Goodness of Fit:
1 - pchisq(model deviance, residual degrees of freedom) = 0.0001770943
We can reject the null hypothesis that the model is no better than the null model.

Comparing to Linear Model:
Negative binomial performs significantly better, use chi-squared test:

2 * [log likelihood(negative binomal model(positivity) - log likelihood(linear model(positivity)] = 3675.27 (df=3)

We can reject null hypothesis with p ~ 0.

***Full Results for Negative Binomial Analysis**

<u>Story Length</u>

Linear Model:
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
```

Modeling as Poisson:

```
Call:
glm(formula = num_tweets ~ wc, family = poisson, data = stories)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-7.039  -5.308  -3.255   0.869  77.100

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.936e+00  4.653e-03   846.0   <2e-16 ***
wc          -1.059e-04  4.041e-06   -26.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 116893  on 2649  degrees of freedom
Residual deviance: 116154  on 2648  degrees of freedom
AIC: 130,089

Number of Fisher Scoring iterations: 5
```

Modeling as Negative Binomial:

```
Call:
glm.nb(formula = num_tweets ~ wc, data = stories, control = glm.control(maxit = 100),
    init.theta = 1.345997917, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4248  -1.0712  -0.6054   0.1330   7.8343

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.895e+00  2.651e-02 146.932  < 2e-16 ***
wc          -6.167e-05  2.103e-05  -2.933  0.00335 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.346) family taken to be 1)

    Null deviance: 2929.2  on 2649  degrees of freedom
Residual deviance: 2915.4  on 2648  degrees of freedom
AIC: 25,561

Number of Fisher Scoring iterations: 1


              Theta:  1.3460
          Std. Err.:  0.0345

 2 x log-likelihood:  -25555.4870
```

Which Model Fits Better?

```
X1 <- 2 * (logLik(model.nb.wc) - logLik(model.pois.wc))
'log Lik.' 104529.3 (df=3)
pchisq(X1, df = 0, lower.tail=FALSE)
'log Lik.' 0 (df=3)
```

This very large chi-square strongly suggests the negative binomial model,
which estimates the dispersion parameter, is more appropriate than the Poisson model.

Applying a log transformation to the independent variable

```
glm.nb(formula = num_tweets ~ log(wc), data = stories, control = glm.control(maxit =
100),
   init.theta = 1.351900484, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5919  -1.0711  -0.6085   0.1396   8.0219

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.40668    0.10991  40.092  < 2e-16 ***
log(wc)     -0.08826    0.01671  -5.283 1.27e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.3519) family taken to be 1)

    Null deviance: 2941.6  on 2649  degrees of freedom
Residual deviance: 2913.8  on 2648  degrees of freedom
AIC: 25548

Number of Fisher Scoring iterations: 1


              Theta:  1.3519
          Std. Err.:  0.0346

 2 x log-likelihood:  -25541.5120
```

Emotionality

Model as Poisson:

```
Call:
glm(formula = num_tweets ~ emotionality, family = poisson, data = stories)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-7.776  -5.339  -3.261   0.799  76.833
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.714313   0.006301  589.48   <2e-16 ***
emotionality 6.111343   0.276162   22.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 116893  on 2649  degrees of freedom
Residual deviance: 116430  on 2648  degrees of freedom
AIC: 130,365

Number of Fisher Scoring iterations: 5
```

Model as Negative Binomial:

```
glm.nb(formula = num_tweets ~ emotionality, data = stories, init.theta = 1.34524726,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4859  -1.0705  -0.6031   0.1306   7.7904

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.716      0.039  95.276  < 2e-16 ***
emotionality    6.019      1.777   3.388 0.000705 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.3452) family taken to be 1)

    Null deviance: 2927.6  on 2649  degrees of freedom
Residual deviance: 2915.4  on 2648  degrees of freedom
AIC: 25,563

Number of Fisher Scoring iterations: 1
```

```
              Theta:  1.3452
          Std. Err.:  0.0344

 2 x log-likelihood:  -25557.0260
```

What model fits better?

```
X2 <- 2 * (logLik(model.nb.emot) - logLik(model.pois.emot))
'log Lik.' 104803.9 (df=3)
pchisq(X2, df = 0, lower.tail=FALSE)
'log Lik.' 0 (df=3)
```

This very large chi-square strongly suggests the negative binomial model,
which estimates the dispersion parameter, is more appropriate than the Poisson model.

__**Positivity**__

Model as Poisson:

```
glm(formula = num_tweets ~ positivity, family = poisson, data = stories)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-7.306  -5.342  -3.292   0.786  77.113

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.850563   0.002937 1310.87   <2e-16 ***
positivity  -6.029223   0.337735  -17.85   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 116893  on 2649  degrees of freedom
Residual deviance: 116576  on 2648  degrees of freedom
```

AIC: 130,511

Number of Fisher Scoring iterations: 5

Model as Negative Binomial:

```
glm.nb(formula = num_tweets ~ positivity, data = stories, init.theta = 1.34343482,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4567  -1.0696  -0.6086   0.1272   7.8548

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.84910    0.01774 217.017  < 2e-16 ***
positivity  -5.39140    2.02900  -2.657  0.00788 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.3434) family taken to be 1)

    Null deviance: 2923.8  on 2649  degrees of freedom
Residual deviance: 2915.8  on 2648  degrees of freedom
AIC: 25,567

Number of Fisher Scoring iterations: 1


              Theta:  1.3434
          Std. Err.:  0.0344

 2 x log-likelihood:  -25561.2700
```

Which Model Fits Better?

```
X3 <- 2 * (logLik(model.nb.pos) - logLik(model.pois.pos))
'log Lik.' 104946 (df=3)

pchisq(X3, df = 0, lower.tail=FALSE)
```

'log Lik.' 0 (df=3)

This very large chi-square strongly suggests the negative binomial model,
which estimates the dispersion parameter, is more appropriate than the Poisson model.