DS 6371 Spring 2019 Final Analysis Questions

Social media plays such a big role in advertising these days. Some individual social media influencers can make a living by accruing followers and then advertising on their accounts. Twitter is one such social media platform. A social media consultant is interested in determining how to predict the number of Twitter followers based on a few characteristics. He painstakingly scoured the internet to find several hundred Twitter accounts and recorded their number of followers (how many people follow their account-this is the money maker), followees (the number of people the influencer follows), the average number of tweets per week (collected over the course of a year), and the gender of the account holder. We don’t know much about how this data was collected. This data set is found in the file “Twitterdata.csv”.

* You may use SAS or R in your analyses.
* Please save your work in a Word document (preferably this one) with your NAME in the file name.
* Include your relevant output at the end of each response (or within, if appropriate).
* Include code in the appendix of your work.
* Unless otherwise specified, assume α=0.05 and 95% confidence.
* Be careful that software treats the variable gender as you intend, categorical or numeric. (Its values are 0 for a man and 1 for a woman.)

## Part A (60 points)

The marketing consultant wishes to determine which variables (of those collected) are useful for predicting the number of Twitter followers for a given account.

* Consider the possibility that the relationship between followers and the explanatory variables (both tweets per week and number of followEES) can vary based on gender.
* No other interactions should be considered (besides those above).
* You will need to create a new variable that takes the log (base e) of the followers variable.
  1. Address All assumptions.  
     Examine the assumptions with no transformations and after logging (base e) the number of followers.  
     You should address any outliers, but do NOT delete any data points.  
     After assumptions are discussed, proceed (using logged followers variable) as if all assumptions are met, even if they are not.

Before Transformations:

proc glm data = twitter plots = all;

class gender;

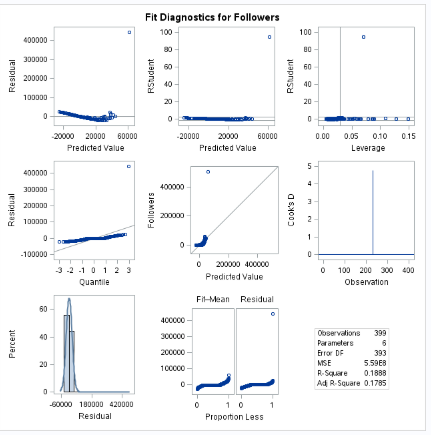
model followers = gender|FollowEES gender| Tweetsperwk/solution;

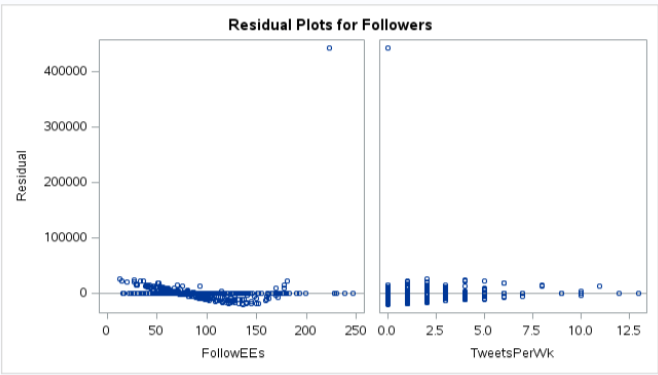
run;

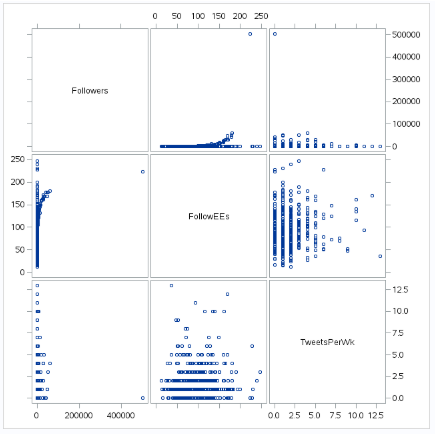
proc sgscatter data = twitter;

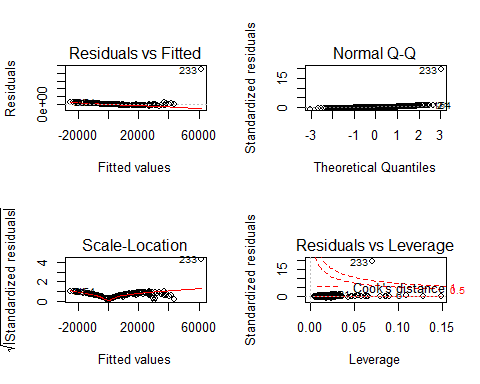
matrix followers followees tweetsperwk;

run;









**Assumptions of untransformed data:**

**Normality: From the q-q plot of residuals and histogram of residuals, we can see that the residuals are not normally distributed.**

**Linearity: Looking at the scatter plots, there does not seem to be a linear relationship between Followers and FollowEES.**

**Equal Standard Deviation: The residuals do not appear to be a random cloud, but that may be due to the outlying point.**

**Independence: We will assume the observations are independent, although we don’t have much information about how this data was collected.**

**Outliers/Influential Points: There appears to be one point (#233) with high Cooke’s D and studentized residual. This point is likely influencing the model substantially. It should not be left as is.**

**After taking the natural log of the Followers:**

**data twitter;**

**set twitter;**

**logfollowers = log(followers);**

**run;**

**proc sgscatter data = twitter;**

**matrix logfollowers followees tweetsperwk;**

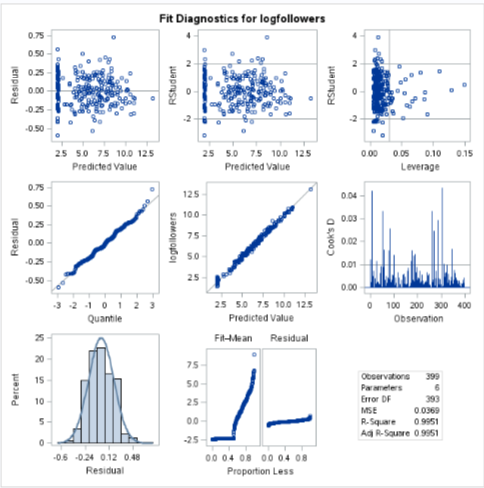
**run;**

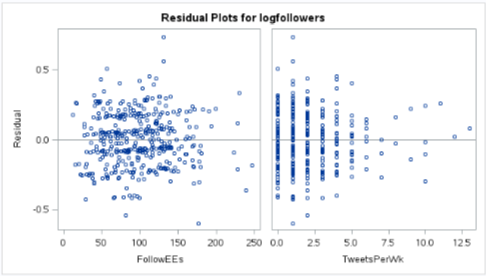
**proc glm data = twitter plots = all;**

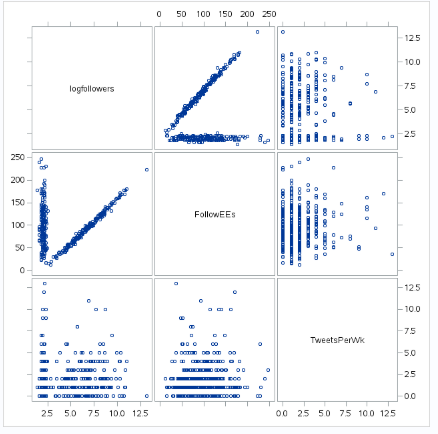
**class gender (ref="0");**

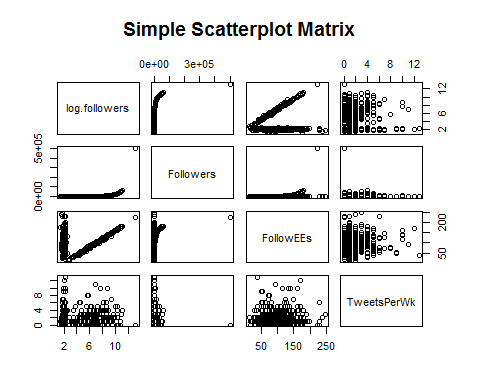
**model logfollowers = gender|FollowEES gender| Tweetsperwk/solution clparm;**

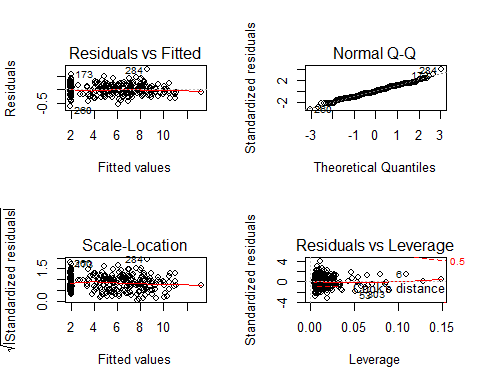
**run;**











**Assumptions of transformed data (log of Followers):**

**Normality: From the q-q plot of residuals and histogram of residuals, we can see that the residuals appear normally distributed.**

**Linearity: Looking at the scatter plots, there does seem to be a linear relationship between Followers and FollowEES, although there appears to be two distinct groups of data.**

**Equal Standard Deviation: The residuals appear to be more of a random cloud.**

**Independence: We will assume the observations are independent, although we don’t have much information about how this data was collected. (Didn’t change when data was transformed.)**

**Outliers/Influential Points: Point (#233) appears to have a much lower Cooke’s D and studentized residual, both of which are now reasonable. No other point appears to unduly influence the model.**

* 1. Build the regression model with parameters to answer the question of interest (Don’t plug in estimates/values of ’s yet.):

**i.e. a total of 6 parameters (including the intercept), with 1 of the parameters dedicated to gender, 1 dedicated to the interaction between gender and number of followees, 1 dedicated to the interaction between gender and number of tweets per week, and main effects for number of followees and number of tweets per week.**

* 1. Copy and paste the typical software output that includes the beta estimates, p-values for each beta, and confidence interval, etc. for each beta (screenshot of a table should suffice):

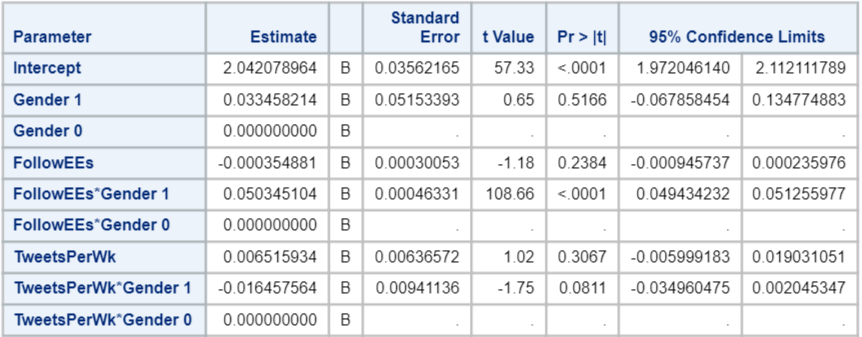
Two tables for two different references.

proc glm data = twitter plots = all;

class gender (ref="0");

model logfollowers = gender|FollowEES gender| Tweetsperwk/solution clparm;

run;

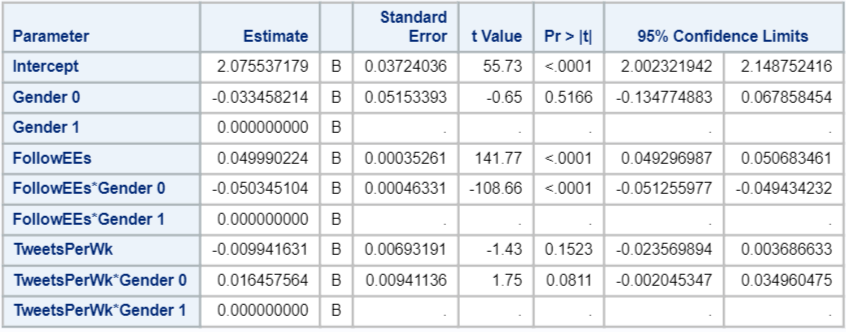


proc glm data = twitter plots = all;

class gender (ref="1");

model logfollowers = gender|FollowEES gender| Tweetsperwk/solution clparm;

run;



## Loading required package: carData

##   
## Call:  
## lm(formula = log.followers ~ Gender + FollowEEs + TweetsPerWk +   
## FollowEEs \* Gender + TweetsPerWk \* Gender, data = twitter)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.59949 -0.11717 -0.01021 0.14894 0.73396   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.0420790 0.0356216 57.327 <2e-16 \*\*\*  
## Gender 0.0334582 0.0515339 0.649 0.5166   
## FollowEEs -0.0003549 0.0003005 -1.181 0.2384   
## TweetsPerWk 0.0065159 0.0063657 1.024 0.3067   
## Gender:FollowEEs 0.0503451 0.0004633 108.664 <2e-16 \*\*\*  
## Gender:TweetsPerWk -0.0164576 0.0094114 -1.749 0.0811 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1922 on 393 degrees of freedom  
## Multiple R-squared: 0.9951, Adjusted R-squared: 0.9951   
## F-statistic: 1.61e+04 on 5 and 393 DF, p-value: < 2.2e-16

* 1. Fit the model using software output found above (regression equation with actual estimates for the betas-do not delete any variables even if some are not significant to this model):

For gender = 0 as the reference:

For gender = 1 as the reference:

* 1. Fill out the following table. For the final column, interpret the coefficients as if all coefficients in the model are significant.
* The first row is started for you (Replace the results with your own—information in the first row is fictional).
* There are MORE rows than you need.

This table is filled out for gender = 0 as the reference.

| Beta | Estimated Beta (Value) | Term/variable | P-value | Significant in this model at alpha = 0.05? | Interpretation |
| --- | --- | --- | --- | --- | --- |
| β0 | 2.04 | y-intercept | <.0001 | Yes | When followees and tweets per week are negligible, the predicted median of followers for men is e2.04=7.7. |
| β1 | 0.033 | gender | 0.5166 | No | When followees and tweets per week are negligible, the predicted median of followers for women men is e0.033=1.03 **TIMES MORE** than the median for men (when followees and tweets per week are negligible). |
| β2 | -0.000355 | FollowEES | 0.2384 | No | When Tweets per week are held constant, one more followEE results in a multiplicative change in predicted median followers of e-0.000355=0.999645 (or decrease of 0.0355%) for men. |
| β3 | 0.0065 | Tweets Per Week | 0.3067 | No | When followEEs are held constant, one more tweet per week results in a multiplicative change in predicted median followers of e0.0065=1.0065 (or increase of 0.65%) for men. |
| β4 | 0.05 | Gender\*FollowEEs | <0.0001 | Yes | When Tweets per week are held constant for WOMEN, one more followEE results in a multiplicative increase in multiplicative change (it’s a mouthful) in predicted median followers of e0.05=1.05 (or increase of 5%) MORE than the multiplicative change in predicted median followers for men when followEEs is increased by 1. |
| β5 | -0.016 | Gender\*TweetsPerWk | 0.0811 | No | When FollowEEs are held constant for WOMEN, one more Tweet per week results in a multiplicative decrease in multiplicative change in predicted median followers of e-0.016=0.98 (or decrease of 2%) MORE than the multiplicative change in predicted median followers for men when Tweets per week is increased by 1. |
| β6 |  |  |  |  |  |
| β7 |  |  |  |  |  |
| β8 |  |  |  |  |  |
| β9 |  |  |  |  |  |
| Β10 |  |  |  |  |  |
| Β11 |  |  |  |  |  |
| Β12 |  |  |  |  |  |
| Β13 |  |  |  |  |  |
| Β14 |  |  |  |  |  |
| Β15 |  |  |  |  |  |
| Β16 |  |  |  |  |  |

* 1. Rebuild the regression model, taking into account the p-values found when fitting the full model above (regression equation with parameters -GENERAL betas-this may be different or the same as question A-2)

**The model is re-specified, removing the terms that include the number of tweets per week and its interaction with gender but leaving gender, number of followees as main effects to maintain the hiearachical structure of the model as well as their interaction, due to the low p-value.**

* 1. Fit (model with estimates/values for betas instead of general betas) a simplified regression model for gender = 1. Begin with the solution for A-4 and find the equation for gender = 1; do not delete any nonsignificant terms for THIS equation.
  2. Find an appropriate prediction interval for the number of followers of a female Twitter account holder with 20 followees who averages 1 tweet per week. Be SPECIFIC about what the interval is for (prediction interval for what?).

**This should be taken from the model without any TweetsPerWeek terms.**

**This prediction interval is the range of plausible values for any *individual* person possessing these characteristics.**

**proc glm data = combined plots = all;**

**class gender (ref="0");**

**model logfollowers = gender|FollowEES/solution cli;**

**run;**

**The “best guess” is 21.4 followers with lower and upper bounds of 14.6 and 31.4, respectively. On the log scale, these values are 3.07, 2.68, and 3.45.**

## Length Class Mode   
## 0 NULL NULL

## Length Class Mode   
## 0 NULL NULL

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000000 0.0001428 0.0009530 0.0025698 0.0028457 0.0435218

##   
## Call:  
## lm(formula = log.followers ~ Gender + FollowEEs + FollowEEs \*   
## Gender, data = twitter)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.60539 -0.10879 -0.01342 0.15285 0.74629   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.0550374 0.0333427 61.634 <2e-16 \*\*\*  
## Gender 0.0050150 0.0488422 0.103 0.918   
## FollowEEs -0.0003579 0.0003009 -1.189 0.235   
## Gender:FollowEEs 0.0502963 0.0004625 108.741 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1924 on 395 degrees of freedom  
## Multiple R-squared: 0.9951, Adjusted R-squared: 0.9951   
## F-statistic: 2.675e+04 on 3 and 395 DF, p-value: < 2.2e-16

## fit lwr upr  
## 1 3.0654 2.683145 3.447655

## fit lwr upr  
## 1 21.44304 14.63104 31.42661

**proc glm data = combined plots = all;**

**class gender (ref="0");**

**model logfollowers = gender|FollowEES gender| Tweetsperwk/solution cli;**

**run;**

**If using the full model, the “best guess” is 21.3 followers with lower and upper bounds of 14.5 and 31.2, respectively. On the log scale, these values are 3.06, 2.68, and 3.44.**

## Part B (Bonus 10 points, up to a maximum of 100 exam points)

Build Your Own ANOVA (lack of fit test) that compares a model that predicts followers based on the three independent variables in the dataset (no interaction terms) to the model found in A-4. If these are exactly the same, state that. If not, build your own ANOVA.

H0:

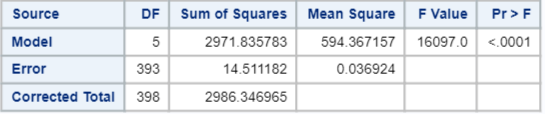
Ha: at least one of , does not equal 0

proc glm data = twitter;

class gender (ref="0");

model logfollowers = gender|FollowEES gender| Tweetsperwk/solution;

run;

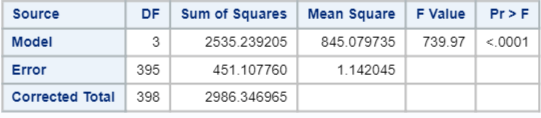


proc glm data = twitter plots = all;

class gender (ref="0");

model logfollowers = gender FollowEES Tweetsperwk/solution;

run;



data pval;

pvalue = 1-probF(5912.1, 2, 393);

run;

proc print data = pval;

run;

| **Source** | **Df** | **SS** | **MS** | **F** | **P-value** |
| --- | --- | --- | --- | --- | --- |
| Model | 2 | 436.6 | 218.3 | 5912.1 | <0.0001 |
| Error | 393 | 14.511182 | 0.036924 |  |  |
| Total | 395 | 451.107760 |  |  |  |

partb.1 <- lm(log.followers ~ Gender + FollowEEs + TweetsPerWk, data=twitter)  
anova(partb.1, parta.2)

## Analysis of Variance Table  
##   
## Model 1: log.followers ~ Gender + FollowEEs + TweetsPerWk  
## Model 2: log.followers ~ Gender + FollowEEs + TweetsPerWk + FollowEEs \*   
## Gender + TweetsPerWk \* Gender  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 395 451.11   
## 2 393 14.51 2 436.6 5912.1 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**From the BYOA above, we reject the null (p < 0.01) and conclude at least one of the interaction terms is significant.**