MSDS 6371 Final Exam – Spring 2019



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 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.
2. **Build** the regression model with parameters to answer the question of interest (Don’t plug in estimates/values of βi’s yet.):
3. **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):
4. **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):
5. **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.

| Beta | Estimated Beta (Value) | Term/variable | P-value | Significant in this model at alpha = 0.05? | Interpretation |
| --- | --- | --- | --- | --- | --- |
| β0 | 1000 | y-intercept | 0.003 | Yes | When followees and tweets per week are negligible, the predicted mean of followers for men is 2. |
| β1 |  |  |  |  |  |
| β2 |  |  |  |  |  |
| β3 |  |  |  |  |  |
| β4 |  |  |  |  |  |
| β5 |  |  |  |  |  |
| 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)
2. **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.
3. **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?).

## 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.

| **Source** | **Df** | **SS** | **MS** | **F** | **P-value** |
| --- | --- | --- | --- | --- | --- |
| Model |  |  |  |  |  |
| Error |  |  |  |  |  |
| Total |  |  |  |  |  |