Results

Oliver

5/6/2019

Results

Typically, results sections start with descriptive statistics, e.g. what percent of the sample is male/female, what is the mean GPA overall, in the different groups, etc. Figures can be nice to illustrate these differences! However, information presented must be relevant in helping to answer the research question(s) of interest. Typically, inferential (i.e. hypothesis tests) statistics come next. Tables can often be helpful for results from multiple regression. Do not give computer output here! This should look like a peer-reviewed journal article results section. Tables and figures should be labeled, embedded in the text, and referenced appropriately. The results section typically makes for fairly dry reading. It does not explain the impact of findings, it merely highlights and reports statistical information.

Data Frame

The researchers' initial data frame was obtained from http://files.pushshift.io/reddit/comments/daily/ and contained all of the comments from the entirety Reddit for one day. They collected four days worth of data (two weekends and two weekdays) and took a random sample of 100,000 comments from each day to end up with a total of roughly 400,000 comments across four days of Reddit content. With this data frame the researchers then added columns that used the R package SentimentAnalysis to keep track of the sentiment of each comment according to different sentiment dictionaries. They also created metrics that kept track of the average sentiment of each comments Subreddit as well as the difference between a comment's sentiment and the average sentiment of their Subreddit. The final dataframe then contained 397,998 observations with 18 variables. The variable of interest that the researchers hoped to predict was controversiality, a binary variable where 1 indicates a comment is controversial, and 0 indicates a comment is not controversial. In the entire dataset, 2.28% of comments were controversial (9,064 controversial comments).

```
## Controversiality
##      0      1
## 388934   9064
```

The researchers then decided to split the dataframe into a training set and testing set so that they could determine a model based on half of the data and then test the accuracy of that model on the second half of the data.

Initial Modeling

For variable selection, the researchers decided to start with a model that used all of the newly created variables (as well as all of the original variables from the initial dataframe) and use likelihood ratio tests to determine which variables were the best at predicting controversiality. The initial model showed that only the variables *Gilded* and *Moderator* were not significant at the 0.05 level. A series of likelihood ratio tests were thus preformed to determine whether or not these variables should be included in our final model. The researchers determined that the variables *Gilded* and *Moderator* both were not significant predictors and did not add anything to the model. From the first series of models, the best model that was found thus included 7 predictor variables and can be seen below in *Table 1.1.*

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

```
## (Intercept)                 -3.980e+00  2.620e-02 -151.890  < 2e-16 ***
## WordCount                    2.552e-03  4.414e-04    5.781 7.41e-09 ***
## WKNDTRUE                     8.505e-02  3.036e-02    2.802 0.005085 **
## comments_in_subreddit       -6.117e-06  2.879e-06   -2.124 0.033631 *
## subreddit_scaled_total_mean -1.016e+00  6.524e-02  -15.575  < 2e-16 ***
## subreddit_diff_QDAP         -7.723e-01  1.230e-01   -6.280 3.38e-10 ***
## subreddit_diff_LM           -7.616e-01  2.017e-01   -3.776 0.000159 ***
## relative_sentiment_diff      1.528e-01  5.241e-02    2.916 0.003549 **
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 42616  on 196192  degrees of freedom
## Residual deviance: 42253  on 196185  degrees of freedom
##   (2806 observations deleted due to missingness)
## AIC: 42269
##
## Number of Fisher Scoring iterations: 6
```

Table 1.1

Randomization Tests

In the final model only three variables had p-values that were greater than 0.001: *comments_in_subreddit*, *WKND*, and *relative_sentiment_diff* (with p-values of 0.0336, 0.0051, and 0.0035 respectively). To confirm that these variables were significant, the researchers decided to perform a randomization test on each to determine if the observed coefficients were statistically significant or not. The results of these randomization tests can be seen in Figures *2.1*, *2.2*, and *2.3* below where the red line indicates the observed coefficient and the density plot shows the randomization distribution. If the red line falls far outside of the density plot, this means that the observed coefficient would not be expected to found by chance and is thus statistically significant. While the researchers were only able to run 100 randomization tests (as opposed to the 10,000 that they hoped to) for each variable, this was due to the limit on computing power that they had access to. This explains the imperfect normal distribution of the plots below, but as the observed coefficients for each predictor fell outside of each randomization distribution, the researchers were able to conclude that each of the three predictors were indeed statistically significant, as the initial p-values for each suggested.

Figure 2.1

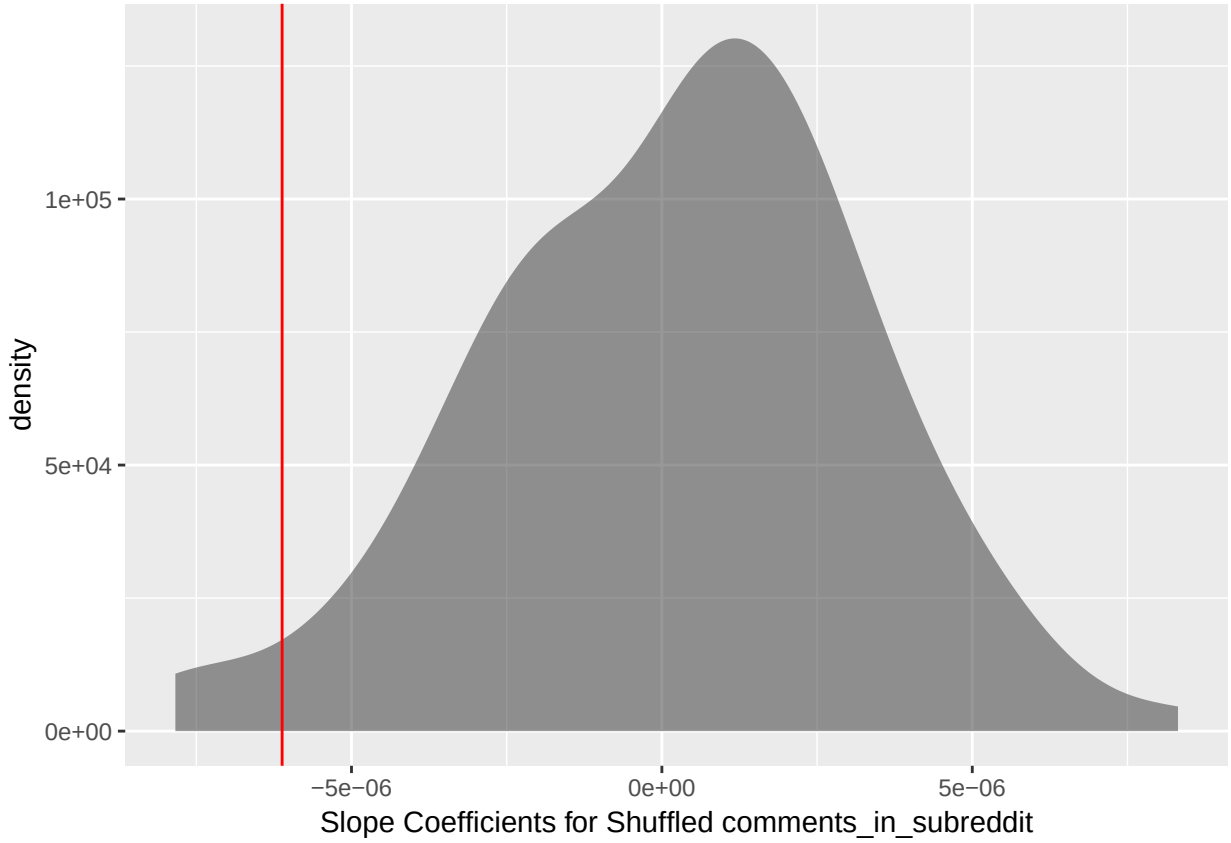

density
1e+05
5e+04
0e+00
−5e−06
0e+00
5e−06
Slope Coefficients for Shuffled comments_in_subreddit

Figure 2.2

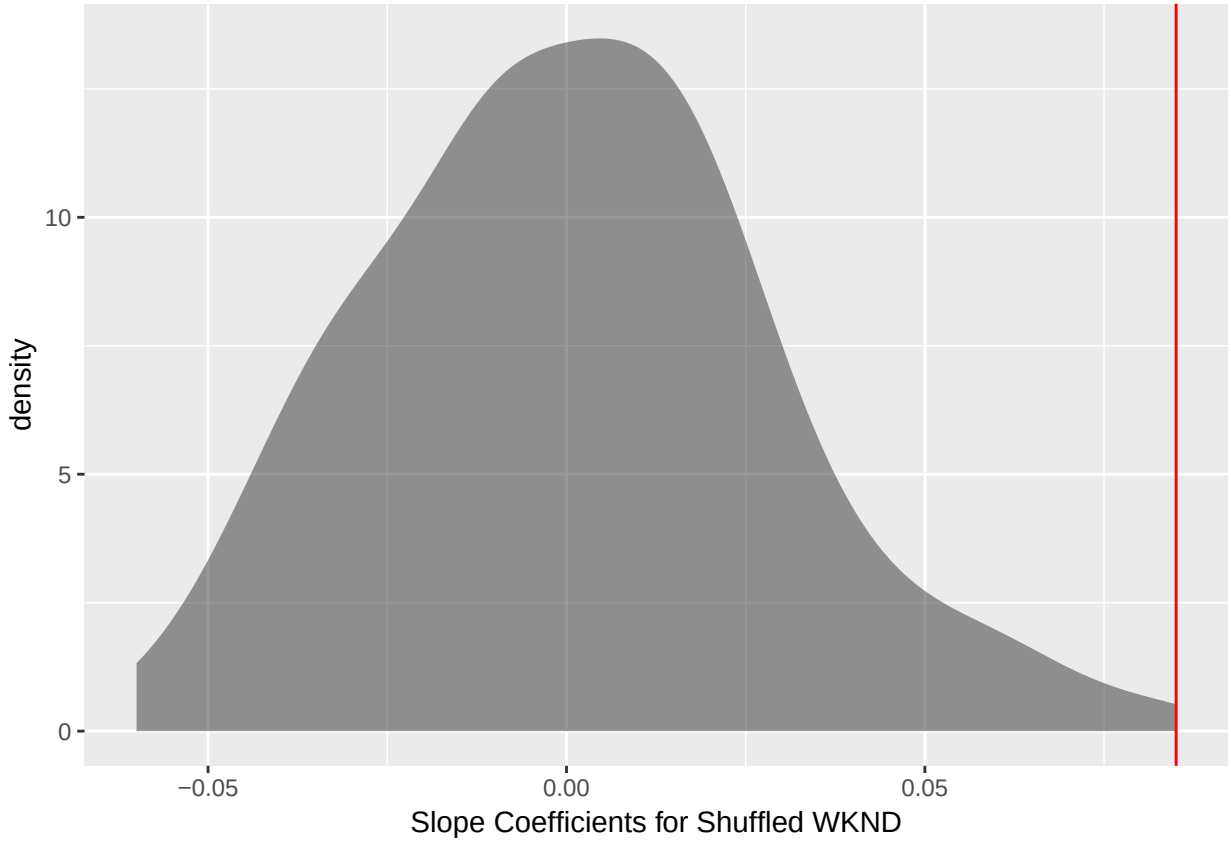
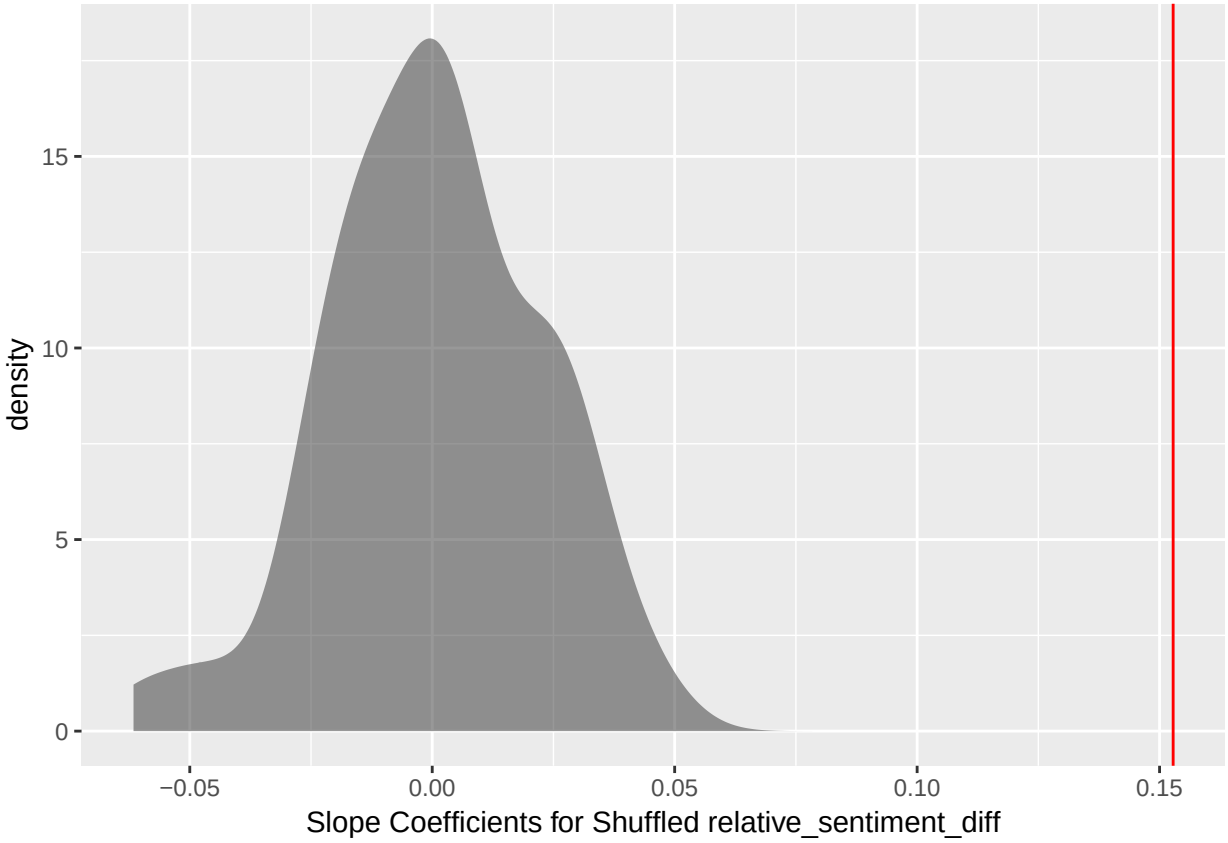

density
10
5
0
−0.05
0.00
0.05
Slope Coefficients for Shuffled WKND

Figure 2.3



Testing Accuracy of Model

With a final model in hand with statistically significant predictors, the researchers then created a function to predict the accuracy of that model using the testing set that was created at the beginning. The researcher's found that the accuracy of the final model was 96.3%, but that the accuracy of the intercept only model was 97.7%.

```
## [1] 0.9633465
```

```
## [1] 0.977226
```

Next Steps: Balanced Dataset

Researchers were concerned with the efficacy of a model that predicted a variable with so few observations. The researchers then decided to resample the initial dataframe in order to have more observations where a comment was marked as controversial (as compared with the 2.28% of comments). This new, balanced dataset, contained exactly 50% of comments that were controversial and 50% of comments that were not controversial. The researchers again split the data into training and testing subsets.

```
## Controversiality
##    0    1
## 9064 9064
```

Modeling with Balanced Dataset

The researchers used the same process with the balanced dataset as with the original dataset to determine the best model to predict controversiality. The final model that was determined by the researchers contained 5 predictor variables and the can be seen below in *Table 3.1.* It's important to note that the *WKND* and *comments_in_subreddit* variables were found to be non-statistically significant using this balanced dataset (while they were found to be statistically significant in the researcher's final model using the original dataset).

```
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.2888174  0.0324828  -8.891  < 2e-16 ***
## WordCount                    0.0022532  0.0008622   2.613  0.00897 **
## subreddit_scaled_total_mean -1.8597738  0.1453709 -12.793  < 2e-16 ***
## subreddit_diff_QDAP         -0.8391867  0.1828052  -4.591 4.42e-06 ***
## subreddit_diff_LM           -0.8567301  0.3130343  -2.737  0.00620 **
## relative_sentiment_diff      0.1700559  0.0789073   2.155  0.03115 *
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12444  on 8976  degrees of freedom
## Residual deviance: 12189  on 8971  degrees of freedom
##   (86 observations deleted due to missingness)
## AIC: 12201
##
## Number of Fisher Scoring iterations: 4
```

The researchers then found the accuracy of this model to be 56.4%, as compared with the 50% accuracy of the intercept only model.

```
## [1] 0.5640375
```

```
## [1] 0.5
```