Josef Triman & Evan Hollier

3/17/21

Analyzing the Relationship between Website Engagement and Graduate Program Applications with Beta Regression

Data source and definitions

The data for our project was obtained from Google Analytics results for the Daniels College of Business. Google Analytics is a web analytics service offered by Google that helps website owners and marketers track and analyze their website traffic. Data is selected from a given time period. For our project we looked at weekly data from the five months leading up to the final submission deadline for the Fall 2022 quarter.

One of the features Google Analytics offers is conversion tracking, which is the number of "goals" met. Goals are defined by the website owner. The goal we looked at was "Grad Apply" which represents the clicks on the "Apply" button. Our data was based on "Grad Apply" given as a rate of the number of sessions in a week. A session is a single period of time a user is actively engaged with the website, as determined by Google Analytics. To give an example of what the data for a week looks like: the last week of April (right before Priority deadline 4) had 156 sessions (the highest in our data set) with 1.28% resulting in a "conversion" (2 out of the 156 sessions clicked the apply button). We focused on the data only from the Professional MBA program instead of the entire Daniels College of Business site. Appendix A is a screenshot of what Google Analytics looks like.

Main features of data set

Appendix B shows the data after being cleaned, including the two zero values we later removed. We performed a preliminary check on the relationship between Sessions and Grad Apply with a linear regression model (Appendix G). There did seem to be a positive relationship, however it was clear the linear regression would not be suitable for the data set.

Research question

The goal with this project was to explore the relationship between online sessions with a graduate program website and the rate at which the apply button is clicked. The question our research was focused on was, "Using a beta regression, can the number of applications be modeled as a function of sessions within a week, and is the relationship statistically significant?"

The method we chose to analyze our research question with was beta regression. Beta regression is frequently used when the data ranges between zero and one. That's because this model uses a beta distribution, which is a flexible probability distribution that uses two parameters (alpha and beta) to model continuous random variables that have values between zero and one (often probabilities, proportions, and percentages). When researching our method, we found that it is commonly used to model conversion rates, click-through rates, and other similar metrics in digital marketing and advertising. Beta regression was developed in 2004 by Silvia Ferrari and Francisco Cribari-Neto and implemented in R six years later as the package "betareg".

Data satisfaction of requirements of method

Beta regressions have multiple requirements. First, the data should be independent. We checked this by plotting a fitted and residuals graph for a beta regression, as shown in Appendix E. We visually assessed whether the model fits the data well or whether there is a pattern in the residuals that suggests the model is not appropriate. There was no clear pattern, indicating independence.

Second, the response variable should follow a beta distribution. The response variable, Grad Apply, should be continuous and bounded between zero and one so that a beta distribution can be made. Our data contained Grad Apply values of exactly zero, so we chose to remove them. However, we could have perturb these values so that they could be included. Afterwards, we were left with nineteen weeks of data that were ready to apply to a beta distribution model. The betareg package creates a beta distribution by estimating the alpha and beta parameters based on the input variables.

Third, the residuals of the model should be normally distributed, meaning that the errors in the model should follow a normal distribution after the data have been transformed by the logit function. Appendix C contains a quantile-quantile plot comparing the residuals to Normality. The data points lie in the gray envelope without being skewed, indicating this assumption is met.

Fourth, the variance of the response variable "Grad Apply" should be a function of the mean, which is known as the dispersion parameter. The variance of the response variable should increase or decrease with the mean of the response variable. Additionally, the phi coefficient for the precision model with the identity link was estimated to be 165.65, with a standard error of 54.65, and a z-value of 3.031. This was significant at the 0.05 level, indicating that there was some evidence of overdispersion in the data (Appendix F). Therefore, it may be necessary to fit an alternative model that accounts for overdispersion.

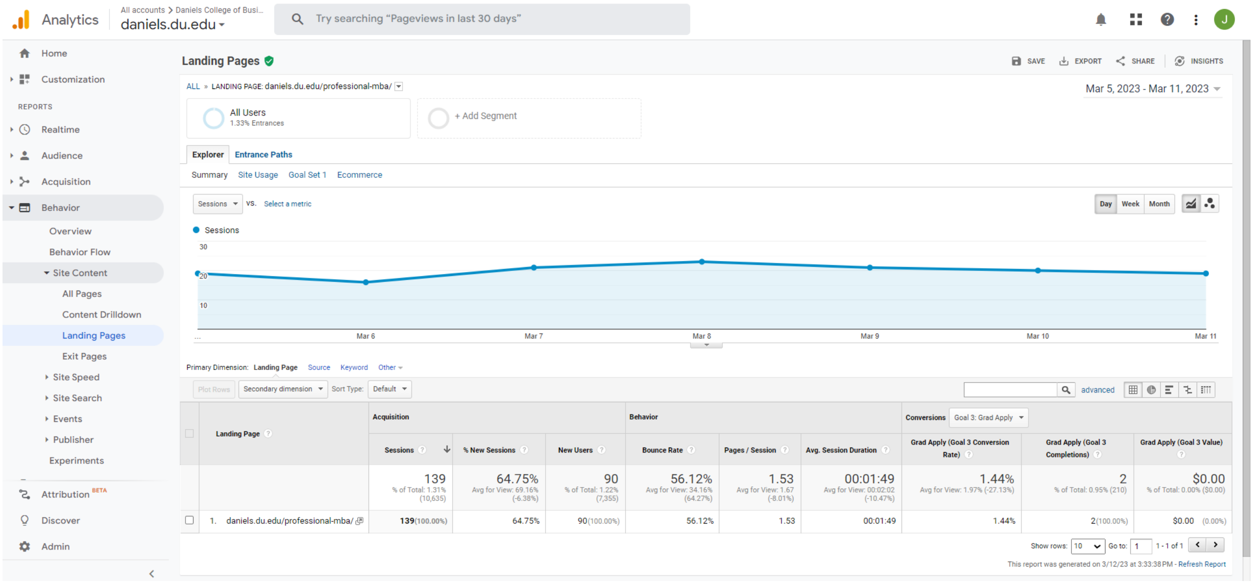
Lastly, the data should not be overly skewed or have extreme outliers, as this can lead to biased estimates and incorrect inferences. We used a Cook's distance plot to identify influential points in the dataset. Cook's distance measures the influence of each observation on the fitted values of the regression model. We can see in Appendix D that the third data point exceeds the 4/n threshold (roughly 0.2). Thus, we would identify this observation as an influential data point that gives the regression model a potential bias. We choose not to remove the third data point.

Method applied and interpreted

Based on the summary shown in Appendix F, the standardized weighted residuals are in the range of -1.9659 to 1.5133, suggesting that the model is a good fit for the data. The coefficients indicate that the intercept has an estimated value of -3.787243 with a standard error of 0.410627, and Sessions has an estimated value of 0.001983 with a standard error of 0.003858. However, the p-value for the Sessions coefficient is not significant (p = 0.607), which indicates that the Sessions variable does not have a significant effect on the proportion of people clicking apply. The Phi coefficient estimates the precision of the model and has a significant value of 165.65 (p = 0.00244), indicating that there is significant overdispersion in the model. The log-likelihood value is 57.62 with 3 degrees of freedom, and the pseudo R-squared value is 0.01531. The number of iterations to fit the model was 95 (BFGS) plus 3 (Fisher scoring). In summary, based on the results, we can conclude that the Sessions variable does not have a significant effect on the proportion of people clicking the apply button and that there is significant overdispersion in the model. The model's pseudo R-squared indicates that only a small amount of the variation in the dependent variable (Grad Apply) can be explained by the independent variable (Sessions).

To conclude, we have insufficient conditions to prove that the number of sessions within a week does not predict the number of people clicking on the apply button. There could be other possible variables such as deadlines to submit applications or outside marketing that affect the number of people clicking the apply button on a graduate website. Further analysis would need to be conducted or included with our test to truly see if the number of sessions has an effect on the percentage of people clicking the apply button. Namely, a time-series analysis could be conducted since proximity to deadlines could affect the number of applications to a program.

Appendix A.



Landing Pages

ALL » LANDING PAGE: daniels.du.edu/professional-mba/

Mar 5, 2023 - Mar 11, 2023

All Users
1.33% Entrances

+ Add Segment

Primary Dimension: Landing Page Source Keyword Other

Sort Type: Default

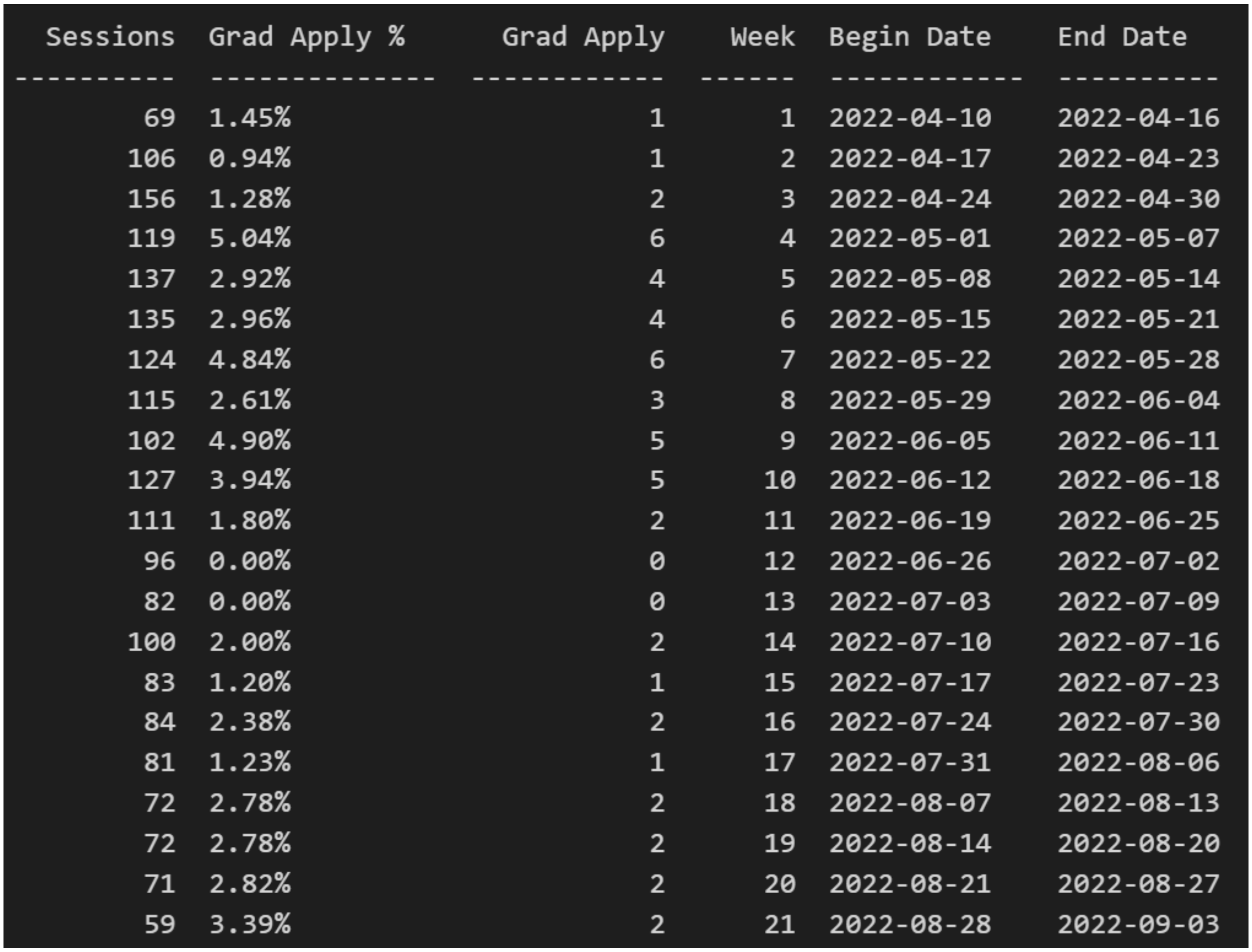
Landing Page	Acquisition			Behavior			Conversions Goal 3: Grad Apply		
	Sessions	% New Sessions	New Users	Bounce Rate	Pages / Session	Avg. Session Duration	Grad Apply (Goal 3 Conversion Rate)	Grad Apply (Goal 3 Completions)	Grad Apply (Goal 3 Value)
	139 % of Total: 1.31% (10,635)	64.75% Avg for View: 69.16% (-6.38%)	90 % of Total: 1.22% (7,355)	56.12% Avg for View: 34.16% (64.27%)	1.53 Avg for View: 1.67 (-8.01%)	00:01:49 Avg for View: 00:02:02 (-10.47%)	1.44% Avg for View: 1.97% (-27.13%)	2 % of Total: 0.95% (210)	$0.00 % of Total: 0.00% ($0.00)
1. daniels.du.edu/professional-mba/	139(100.00%)	64.75%	90(100.00%)	56.12%	1.53	00:01:49	1.44%	2(100.00%)	$0.00 (0.00%)

Show rows: 10 Go to: 1 1 - 1 of 1

This report was generated on 3/12/23 at 3:33:38 PM - Refresh Report

Google Analytics page displaying our data. Each data table associated with a week was exported as a .CSV, resulting in 21 .CSV's

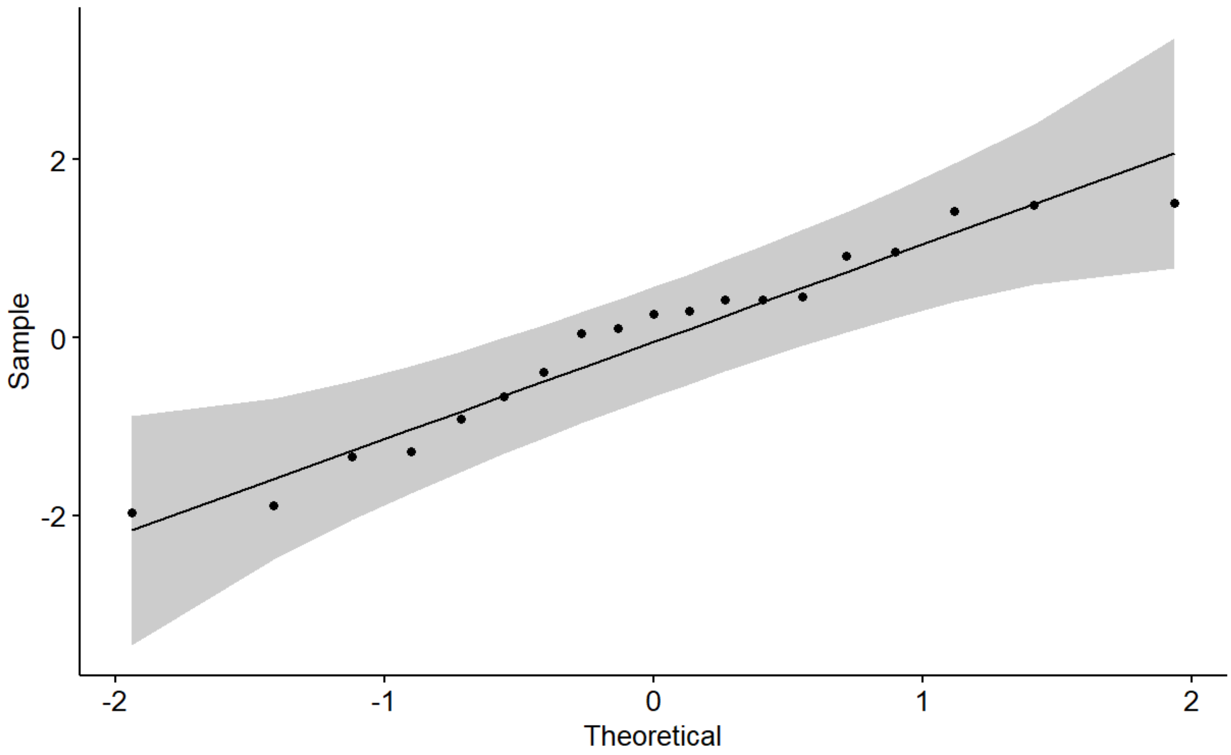
Appendix B.



```
  Sessions  Grad Apply %     Grad Apply    Week  Begin Date   End Date
----------  -------------  ------------  ------  -----------  ----------
        69  1.45%                     1       1  2022-04-10   2022-04-16
       106  0.94%                     1       2  2022-04-17   2022-04-23
       156  1.28%                     2       3  2022-04-24   2022-04-30
       119  5.04%                     6       4  2022-05-01   2022-05-07
       137  2.92%                     4       5  2022-05-08   2022-05-14
       135  2.96%                     4       6  2022-05-15   2022-05-21
       124  4.84%                     6       7  2022-05-22   2022-05-28
       115  2.61%                     3       8  2022-05-29   2022-06-04
       102  4.90%                     5       9  2022-06-05   2022-06-11
       127  3.94%                     5      10  2022-06-12   2022-06-18
       111  1.80%                     2      11  2022-06-19   2022-06-25
        96  0.00%                     0      12  2022-06-26   2022-07-02
        82  0.00%                     0      13  2022-07-03   2022-07-09
       100  2.00%                     2      14  2022-07-10   2022-07-16
        83  1.20%                     1      15  2022-07-17   2022-07-23
        84  2.38%                     2      16  2022-07-24   2022-07-30
        81  1.23%                     1      17  2022-07-31   2022-08-06
        72  2.78%                     2      18  2022-08-07   2022-08-13
        72  2.78%                     2      19  2022-08-14   2022-08-20
        71  2.82%                     2      20  2022-08-21   2022-08-27
        59  3.39%                     2      21  2022-08-28   2022-09-03
```

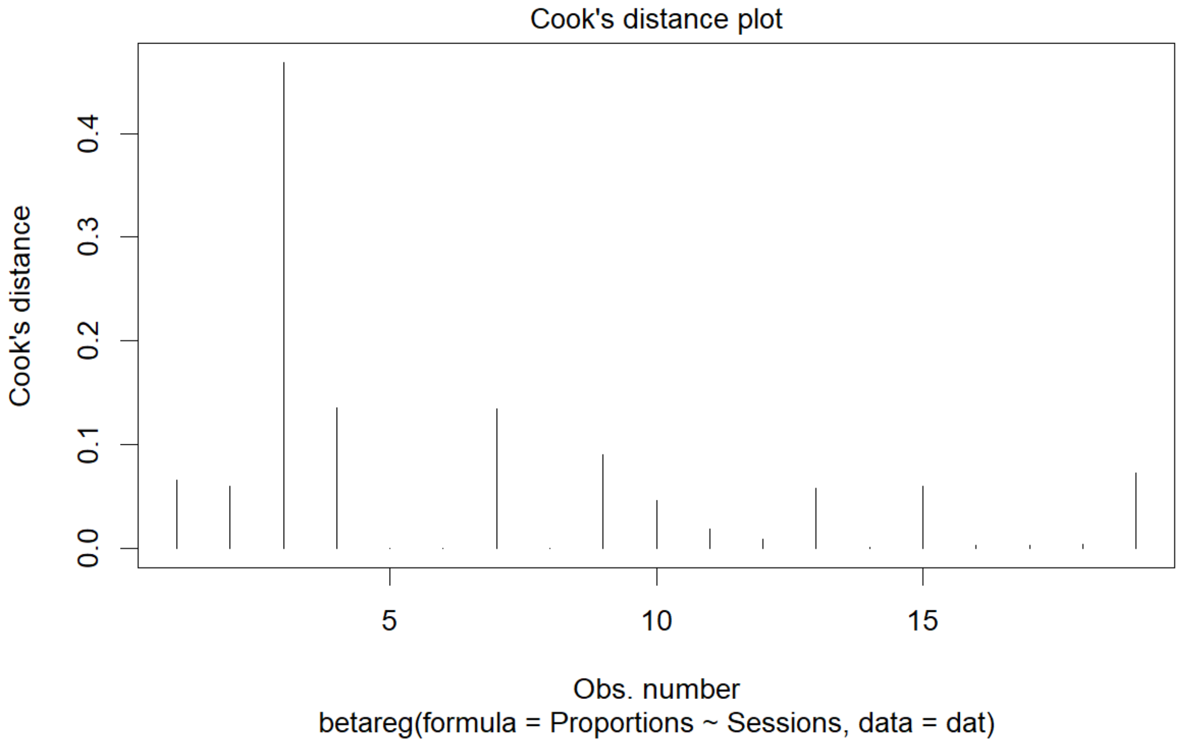
Professional MBA program data after being cleaned

Appendix C.



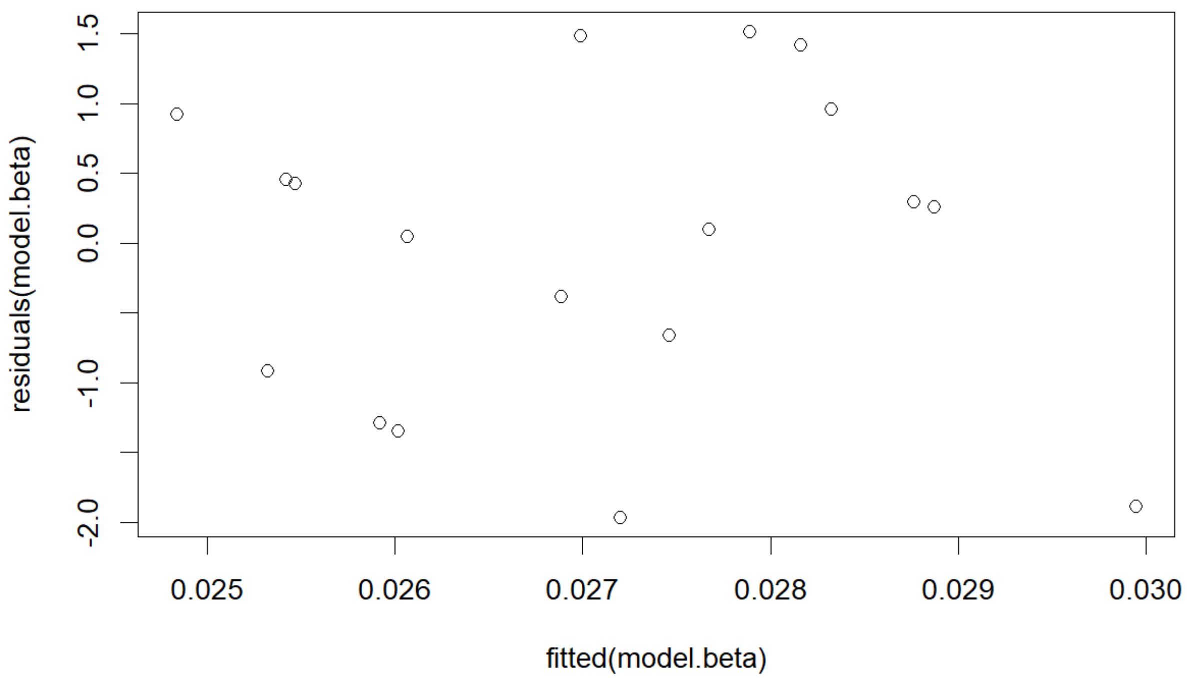
Q-Q plot of residuals to test Normality

Appendix D.



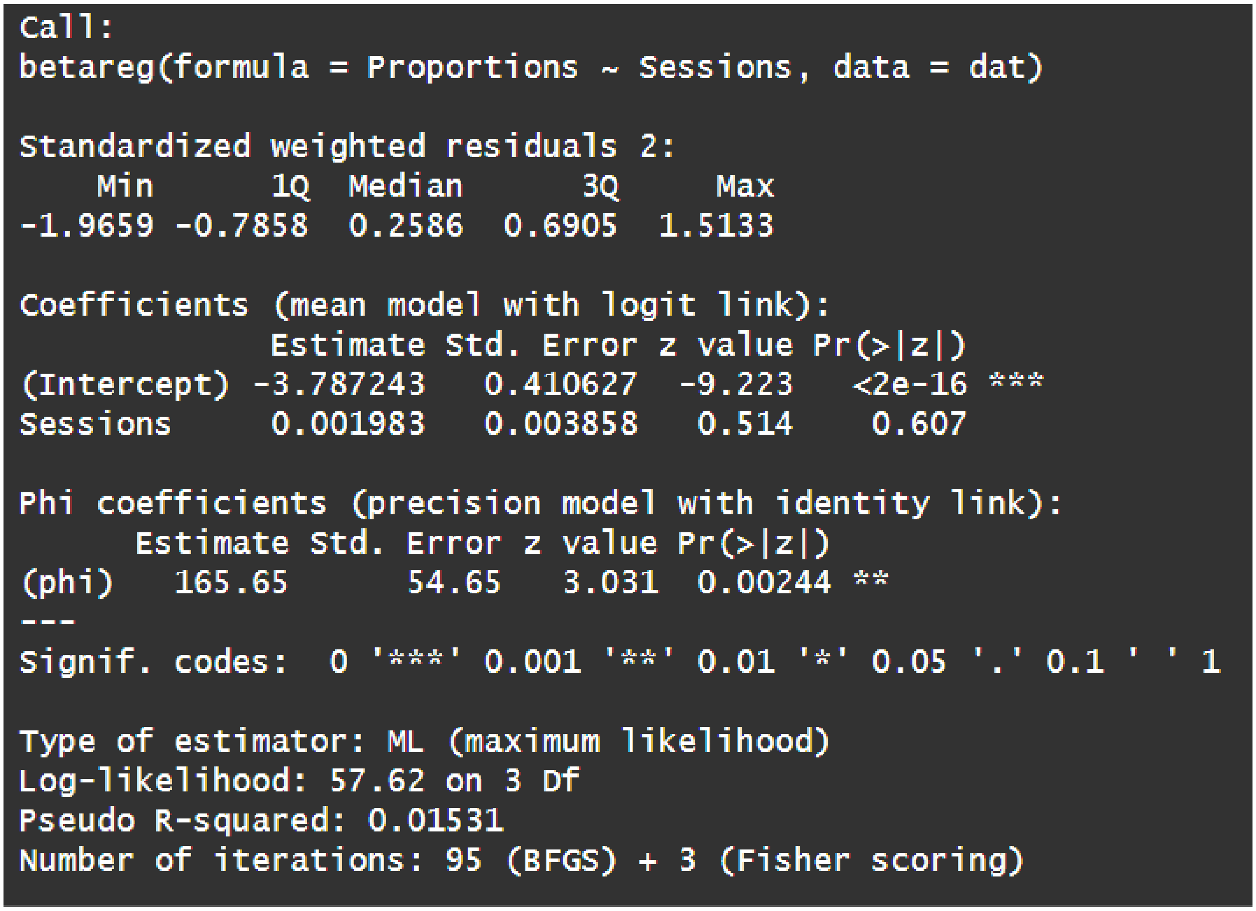
Cook’s distance plot of beta regression model

Appendix E.



Plot of residuals of beta regression model

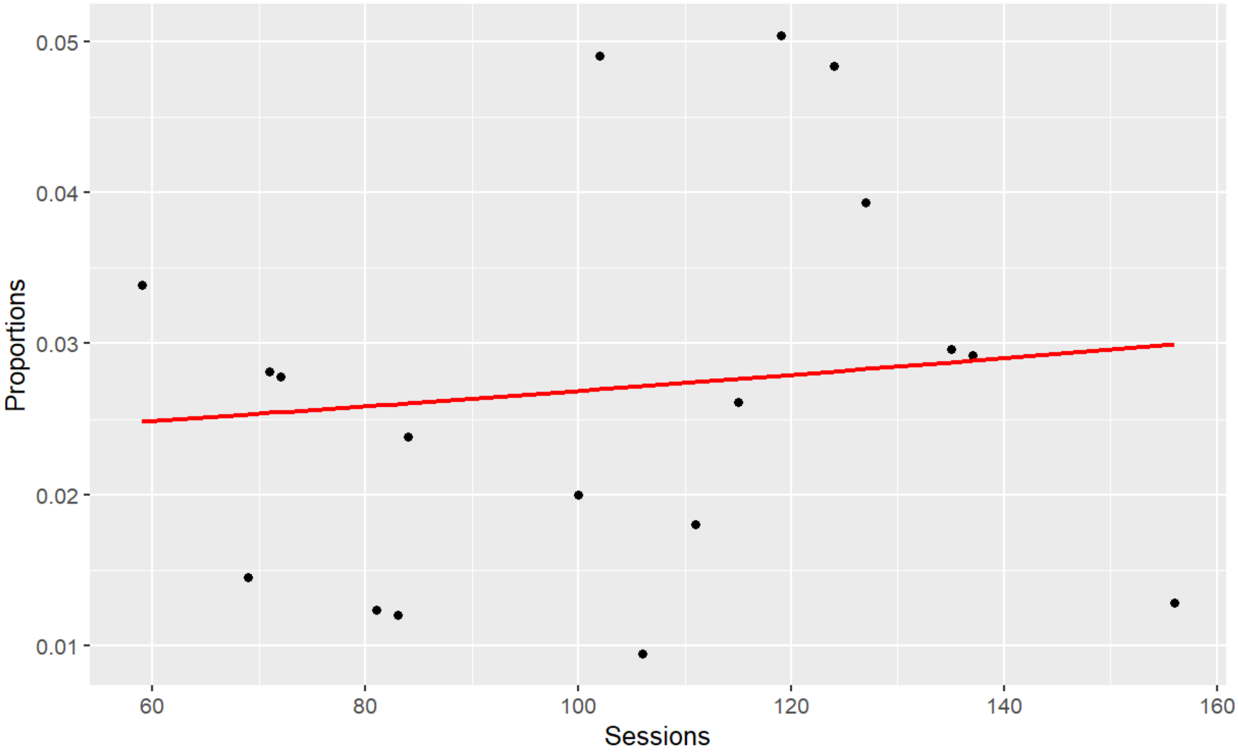
Appendix F.



```
Call:
betareg(formula = Proportions ~ Sessions, data = dat)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-1.9659 -0.7858  0.2586  0.6905  1.5133

Coefficients (mean model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.787243   0.410627  -9.223   <2e-16 ***
Sessions     0.001983   0.003858   0.514    0.607

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   165.65      54.65   3.031  0.00244 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 57.62 on 3 Df
Pseudo R-squared: 0.01531
Number of iterations: 95 (BFGS) + 3 (Fisher scoring)
```

Summary of the beta regression results

Appendix G.



Linear regression model with percentage of people clicking apply renamed to proportions

Sources:

- https://rcompanion.org/handbook/J_02.html - R Handbook: Beta Regression for Percent and Proportion Data
- https://towardsdatascience.com/beta-distribution-intuition-examples-and-derivation-cf00f4db57af - Beta Distribution—Intuition, Examples, and Derivation
- https://cran.r-project.org/web/packages/betareg/vignettes/betareg.pdf - Beta Regression in R
- https://www.frontiersin.org/articles/10.3389/fams.2021.780322/full - A New Two-Parameter Estimator for Beta Regression Model: Method, Simulation, and Application
- https://esajournals.onlinelibrary.wiley.com/doi/full/10.1002/ecs2.3940 - A case for beta regression in the natural sciences
- https://www.youtube.com/watch?v=juF3r12nM5A - The Beta distribution in 12 minutes!
- https://www.youtube.com/watch?v=1k8lF3BriXM - The Beta Distribution : Data Science Basics
- https://www.youtube.com/watch?v=V0zg_6DJ3gk - 6840-10-19-5: 3.6 Beta Regression
- https://support.google.com/analytics/answer/2731565?hl=en#zippy=%2Cin-this-article - How a web session is defined in Universal Analytics