Regression Models for Ad Data

Joseph Parks

Introduction

This project looks at modeling the effect of Facebook advertising for a Self-Storage Business. The business and advertising are for a small local area around Albuquerque, NM. Approximately 1 million people live within a 50-mile radius. National average says 1 in 9 people rent a storage unit. I made and run the ads, and I built and manage the website. Therefore, I have full access to all of the data. The purpose of the ads is to drive interested traffic to the business website. The goal is to discover a model that best describes the data. To this end, the project will explore many different regression models and model selection methods.

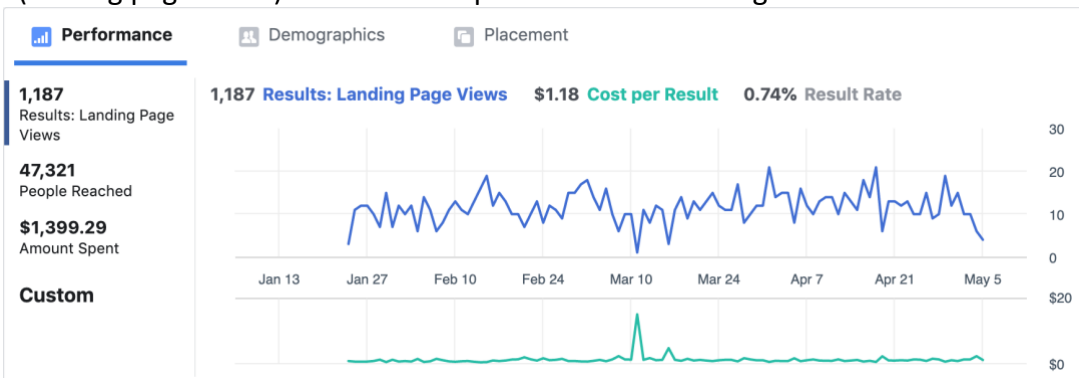
There are three things I want to gain more insight about:

1. How to increase legitimate traffic to my website (Landing page views).
2. How to increase awareness of my brand (ad impressions).
3. Optimizing the first two objectives while minimizing cost (spent).

I have data going back to January 24th, 2019. I will be looking at the following observational data:

- results : Number of landing page views per day
- impressions : Number of impressions per day
- reach : Number of unique impressions per day
- linkClicks : Number of link clicks per day
- spent : Amount spent per day
- CPC : Average cost per link click per day
- CPR : Average cost per result (landing page view) per day
- freq : Number of times Ad was shown to same person per day
- CTR : Number of people who click the link and wait for page to load

Results (landing page views) will be the response for all of the regression models.

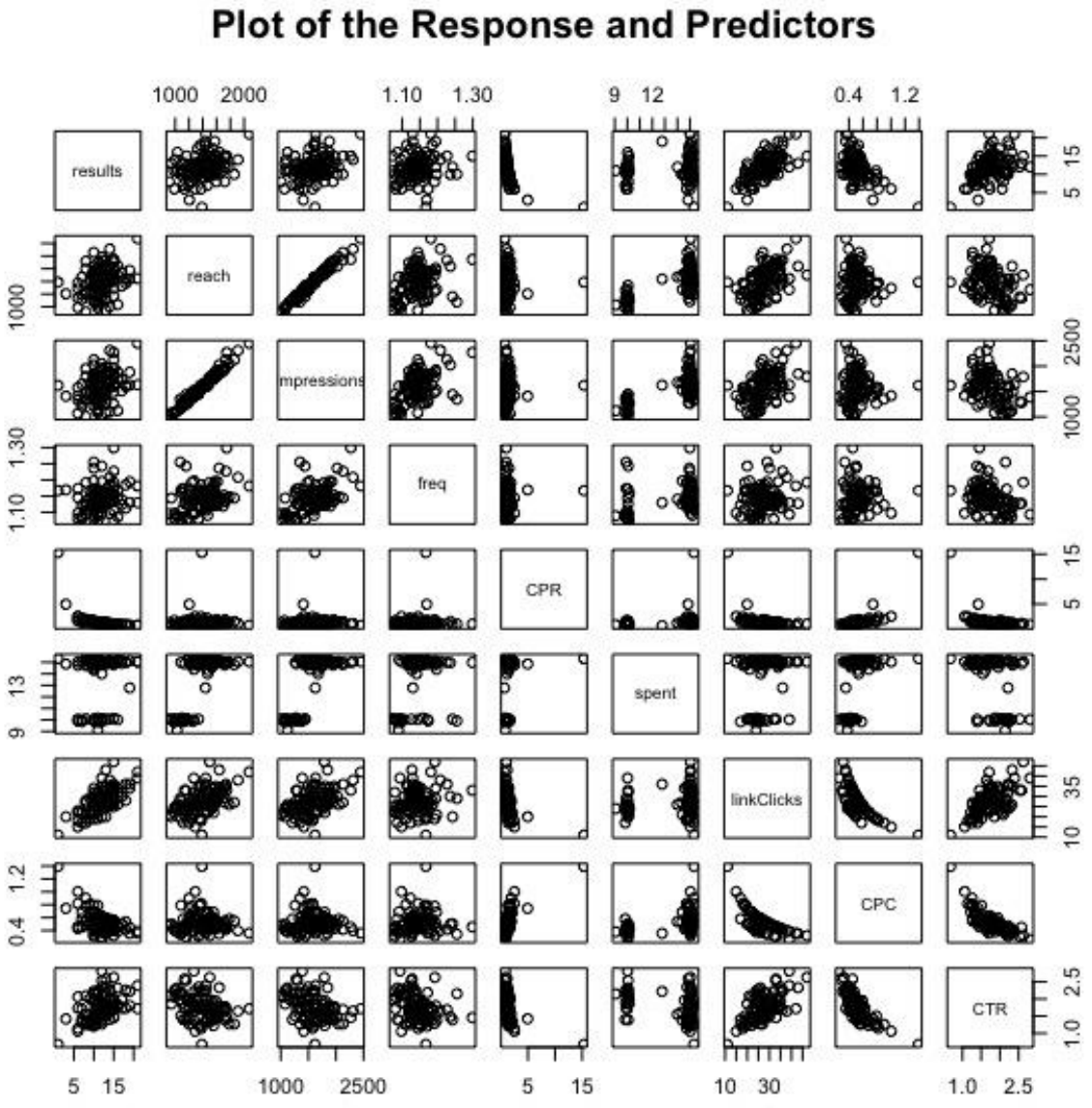


The above graph shows the results (Landing Page Views) per day in blue. And the cost, or amount spent per day per result, is in green.

Landing page views means that someone clicked the link in the ad and waited for the page to load and interacted with the page longer than a bot or crawler, meaning it was an actually interested human. A large difference in link clicks versus landing page views could also mean slow loading times for the website causing the users to close the browser window before the page loads.

Methods

The first step is to explore the dataset. I did this by plotting all the data.

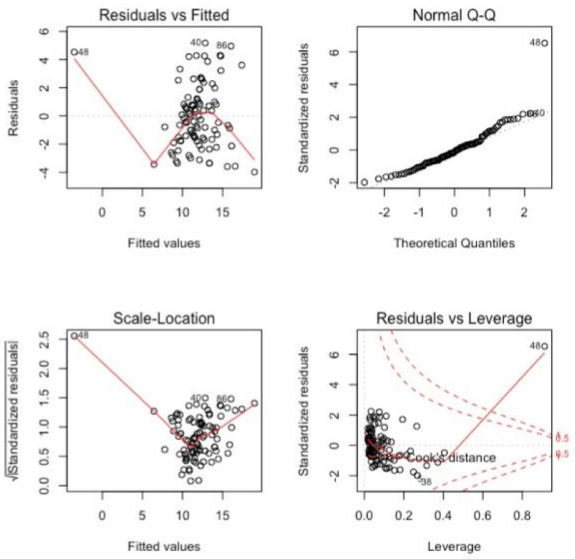


There is clear collinearity among several predictors, as is clearly seen between reach and impressions. This makes sense as reach is the number of unique people the ad was shown to

that day, and impressions accounts the number of times the ad was shown to the same person. Another interesting predictor is the cost per result (CPR) which shows some strange behavior. This predictor is related to the results, or landing page views, and the amount spent per day.

Linear Regression

```
Call:
lm(formula = results ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9772 -1.6120 -0.2669  1.1634  5.1706

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.27989   44.60063   0.387 0.699348
reach        -0.02259    0.02802  -0.806 0.422153
impressions   0.02135    0.02273   0.939 0.350253
freq        -19.39792   34.21652  -0.567 0.572182
CPR          -1.10022    0.32121  -3.425 0.000927 ***
spent        -0.06450    0.31164  -0.207 0.836506
linkClicks    0.29547    0.31879   0.927 0.356483
CPC           8.36171    7.21463   1.159 0.249524
CTR           2.30013    5.43674   0.423 0.673253
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.35 on 90 degrees of freedom
Multiple R-squared:  0.5738,    Adjusted R-squared:  0.5359
F-statistic: 15.14 on 8 and 90 DF,  p-value: 7.397e-14
```



High Collinearity shown by the Variance Inflation Factor

```
> vif(lmod)
     reach impressions       freq        CPR      spent  linkClicks        CPC        CTR
748.708948 773.141194  32.979303   4.150698   7.431242  68.871099  22.942267  75.112591
```

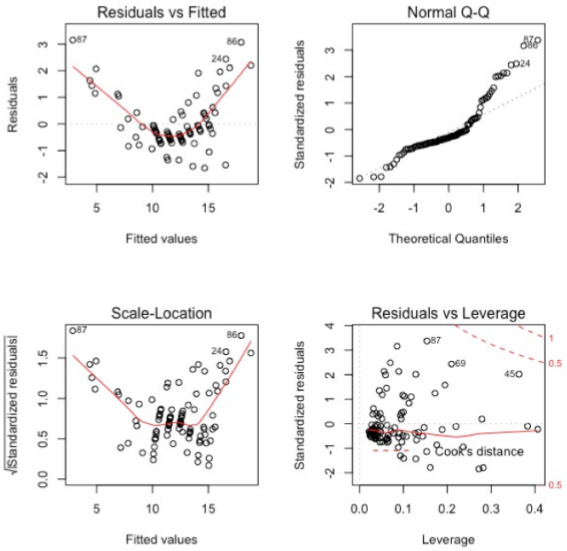
MSE = 5.020052

First, I performed a linear regression on the whole dataset with results (landing page views) as the response and the rest of the data as predictors. The p-values are all very large except for cost per result, CPR. The adjusted R-squared is only 0.5359, so much of the variability in the model is not accounted for by this linear model. Looking at the residuals versus fitted values plot, point 48 looks like a possible outlier, and there appears to be some curvature among the point cluster. Excluding point 48, the assumption of constant variance seems to hold. The normal Q-Q plot shows more deviance from the normality assumption. And the residuals versus leverage plot shows point 48 to be a high leverage point with a large Cook's distance. Note the large variance inflation factors that again point to collinearity between predictors. The mean squared error for this model is not too bad at 5.020052. However, this model does not fit the data very well

The data was fit again excluding point 48 and another outlier, point 53, was identified as a potential influential point. Point 53 was also a high leverage point with a large Cook's distance. Next, I ran the linear regression model while excluding points 48 and 53.

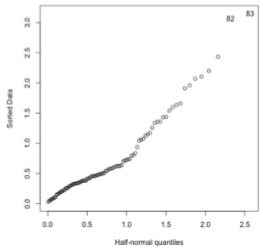
Linear Regression After Removing Points 48 and 53

```
Call:
lm(formula = results ~ ., data = df3)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6599 -0.5443 -0.3078  0.3299  3.1535

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.805453  19.375943   1.590 0.115448
reach        -0.014230   0.012140  -1.172 0.244319
impressions   0.011096   0.009834   1.128 0.262242
freq        -19.932857  14.791772  -1.348 0.181258
CPR          -8.944795   0.438700 -20.389  < 2e-16 ***
spent         0.287180   0.140800   2.040 0.044386 *
linkClicks    0.337655   0.138199   2.443 0.016553 *
CPC          12.831265   3.306184   3.881 0.000201 ***
CTR          -1.646529   2.395180  -0.687 0.493617
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.015 on 88 degrees of freedom
Multiple R-squared:  0.9061,    Adjusted R-squared:  0.8976
F-statistic: 106.2 on 8 and 88 DF,  p-value: < 2.2e-16
```



MSE = 0.9344043

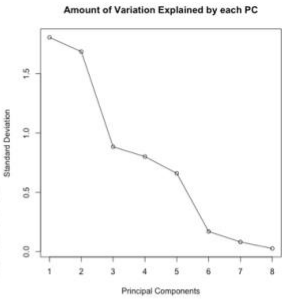
The half-normal plot shows some other potential outliers



Points 48 and 53 were discovered to be influential points due to the large change in the intercept and parameter values. Now, the same linear model, but with points 48 and 53 removed, shows four significant predictors for significance level of 5%: Cost per result (CPR), amount spent per day (spent), link clicks per day (linkClicks), and cost per click (CPC). The adjusted R-squared is improved a lot, but the model still does not look right.

The residuals versus fitted plot clearly shows a non-linear relationship between the response and the predictors. It also shows violation of the constant variance assumption. The normal Q-Q plot shows the normality assumption is also violated. The residuals versus leverage plot shows a few more points with high leverage and Cook's distance, but now they all have a Cook's distance less than 0.5. And the half normal plot above shows more potential outliers. The MSE is low at 0.9344043, but this model violated too many assumptions, and therefore, does not model the data very well.

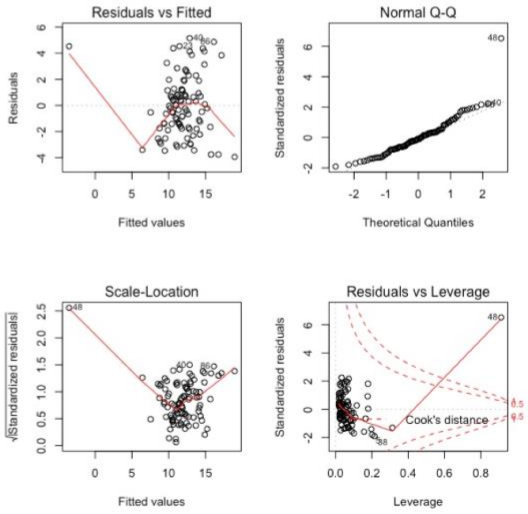
Principal Component Regression




```
> summary(s)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     1.8063 1.6864 0.88371 0.80088 0.65952 0.16919 0.08076 0.02549
Proportion of Variance 0.4079 0.3555 0.09762 0.08018 0.05437 0.00358 0.00082 0.00008
Cumulative Proportion  0.4079 0.7634 0.86098 0.94116 0.99553 0.99910 0.99992 1.00000
```

First 6 PCs Used

```
Call:
lm(formula = results ~ ., data = pc)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9498 -1.6281 -0.2162  1.1818  5.1535

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.85859    0.23460  50.547   <2e-16 ***
PC1         -0.05595    0.13054  -0.429   0.6692
PC2          1.47533    0.13982  10.552   <2e-16 ***
PC3          0.03940    0.26683   0.148   0.8829
PC4          0.22446    0.29442   0.762   0.4478
PC5          0.92247    0.35753   2.580   0.0115 *
PC6          2.49875    1.39371   1.793   0.0763 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.334 on 92 degrees of freedom
Multiple R-squared:  0.5701,    Adjusted R-squared:  0.542
F-statistic: 20.33 on 6 and 92 DF,  p-value: 5.175e-15
```

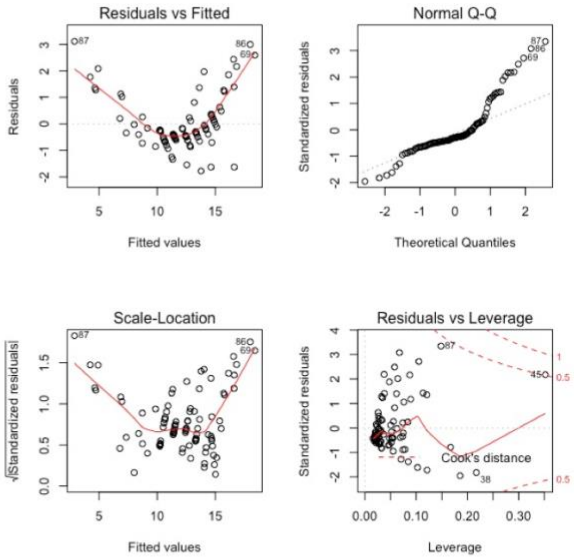


MSE = 5.06359

Next, I ran a principal component regression just for fun. I chose to use 6 principal components for a cumulative proportion of variance of 99.9%. The data for this model includes the influential points 48 and 53. The results are very similar to the linear model above that includes points 48 and 53, but here the MSE is a little higher. This is probably due to having only used 6 of the 8 principal components.

AIC Regression After Removing Points 48 and 53

```
Call:
lm(formula = results ~ ., data = dataAIC2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7779 -0.5108 -0.3025  0.2363  3.1144

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.86436    1.51541   3.210  0.00183 **
CPR         -8.89884    0.42348 -21.013  < 2e-16 ***
spent        0.28614    0.12223   2.341  0.02139 *
linkClicks   0.25672    0.05463   4.699 9.14e-06 ***
CPC         13.48679    2.96659   4.546 1.66e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.008 on 92 degrees of freedom
Multiple R-squared:  0.9032,    Adjusted R-squared:  0.899
F-statistic: 214.6 on 4 and 92 DF,  p-value: < 2.2e-16
```



High Collinearity shown by the Variance Inflation Factor

```
> vif(lmodAIC2)
      CPR     spent linkClicks       CPC
 2.265093  6.169702  10.003141 14.136274
```

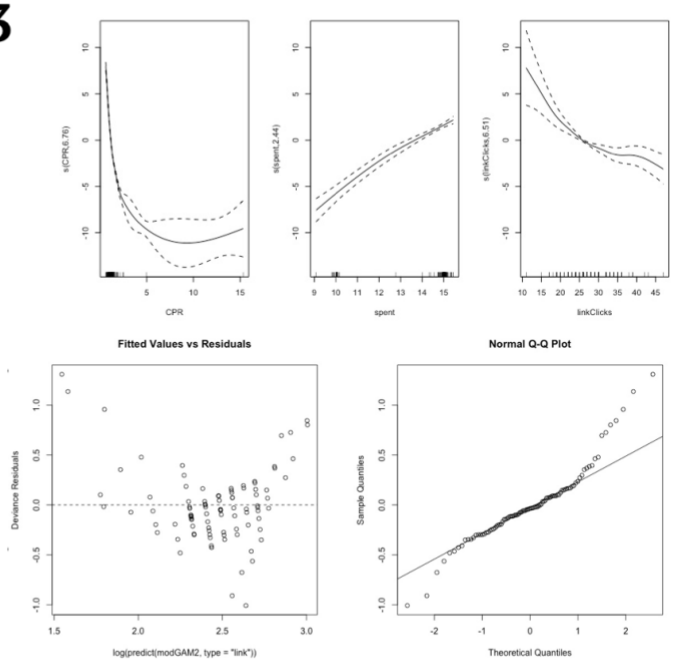
MSE = 0.963585

Due to the collinearity between predictors, I ran AIC model selection and performed regression on the predictors selected. This gave four significant predictors and a significant intercept. The adjusted R-squared is the highest yet at 0.899. This is slightly higher than the full model with all of the predictors. Note, the adjusted R-squared adjusts for the number of predictors allowing the comparison between models with different numbers of predictors.

The variance inflation factors are still pretty large and from the plots the model is breaking all the same assumptions from above, non-linear, non-Normal, non-constant variance, and more points with large Cook’s distances.

GAM Regression After Removing Points 48 and 53

```
> summary(modGAM2)

Family: gaussian
Link function: identity

Formula:
results ~ s(CPR) + s(spent) + CPC + s(linkClicks)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.570      1.433  12.963  < 2e-16 ***
CPC          -12.550      2.761  -4.545 1.91e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Approximate significance of smooth terms:
               edf Ref.df       F  p-value
s(CPR)       5.830  6.956 439.505  < 2e-16 ***
s(spent)     2.257  2.828  59.209  < 2e-16 ***
s(linkClicks) 6.342  7.513   4.213 0.000418 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

R-sq.(adj) =  0.984   Deviance explained = 98.7%
GCV = 0.1939  Scale est. = 0.16106   n = 97
```



MSE = 0.1337817

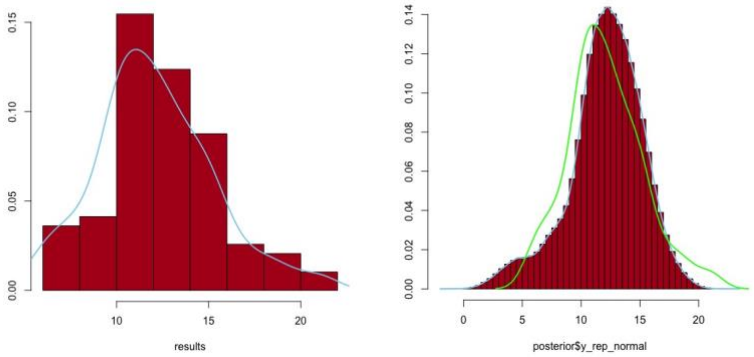
As the linearity assumption was violated in all of the models above, I decided to run a Generalized Additive Model (GAM) so I could treat some of the predictors as non-parametric. First, I used the full model with all 8 predictors. All of the predictors were wrapped in the s() function to be treated non-parametrically. After viewing the plots, many of the predictors behaved linearly. The predictor click through rate (CTR) could have a straight line drawn within the confidence region at zero, which indicated that this predictor should be removed from the model.

Next, as seen above, I removed the influential points 48 and 53 and used the four predictors found with the AIC method. From the plots of the previous GAM, I decided to keep cost per click (CPC) as parametric. I then treated cost per result (CPR), amount spent per day (spent), and link clicks per day (linkClicks) as non-parametric. I ran this model using both the gaussian family and the poisson family. In the end, the gaussian family gave a slightly better fit and lower MSE. As can be seen in the linear predictor versus residuals plot, the data still show quadratic like curvature. All of the predictors used have very small p-values, and the adjusted R-squared at 0.984 is higher than any of the other models. Looking at the plots of the non-parametric predictors, spent shows very small curvature and could probably be treated parametrically. But cost per result (CPR) and link clicks (linkClicks) show some very interesting non-linear behavior. And the normal Q-Q plot shows normal behavior between -1 and 1, but otherwise the data diverges quickly away from normal. With and MSE = 0.1337817, the GAM above seems to be the best model so far.

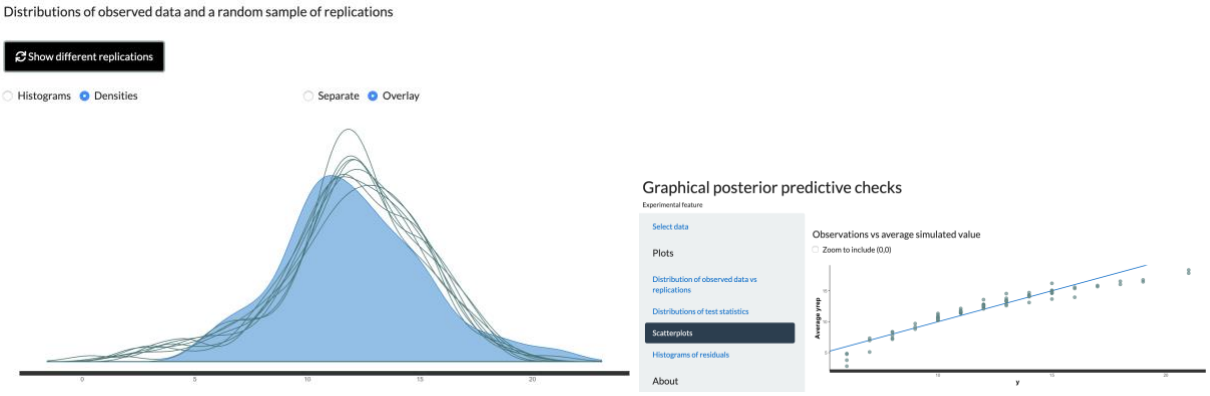
Conclusions

Linear models are inadequate as the data modeled consistently violated the assumptions needed for a good fit with linear regression. The predictors are not linearly independent, and in fact, many have high variance inflation factors. Using ANOVA, I compared some of the full models against the reduced model found with AIC, and, to no surprise, the reduced model was sufficient. Many of the predictors are a linear combination of another predictor. Then there is the case of outliers, leverage, and influential points. There are other factors influencing the data that are not accounted for. Holidays and big events like the Super Bowl can either drive more people to Facebook that day, or fewer people, causing either a spike in landing page views or none at all. I am not sure of how to account for this variability and just removing those days from the dataset, while giving better fits, does not seem accurate. What I really want is a model that can account for that variability. I think if I had a few years' worth of data, the influential points like 48 and 53 would no longer be such high leverage points as there would be a cluster of similar low result value points. I also split the data into a training and test set and the model fit to the training set did very poorly at predicting the results of the test set. For example, the linear regression after removing points 48 and 53 had an MSE = 0.9344043 while the predicted MSE on the test data gave MSE_prediction = 33.75004.

I made a quick attempt at a Bayesian linear regression and got some interesting results.



The plot on the left is a histogram of the results (landing page views) and the blue line is the estimated density. The plot on the right is after a Bayesian linear regression using Normal priors on the betas for the predictors and a Cauchy prior on the variance. The red histogram is from the posterior predictive function and the blue line is the estimated density of the posterior predictive, and the green line is the density of the true observations, i.e. the results.



The plot on the left shows the distribution of the observed data (results) shaded in blue and the lines overlaid are a small random sample of the posterior predictive densities used in the MCMC. The plot on the right is a scatter plot of the observed data y, where y is the results (landing page views), and the Average yrep is from the posterior predictive distribution from the Bayesian linear regression.

With more time I would like to compare the Bayesian results to that of the GAM. In the end, the more you spend, the more results you get. But the relation between these is non-linear. I would like to continue trying new models and collecting more data. In particular, I would explore more with the GAM and the Bayesian regressions.