COMMERCE ACTIVATION

DEPARTMENT OF ANALYTICS

IT'S NOT MY SQUARED ANYMORE HONEY, IT'S R^2

INFLUENCER FORECASTING WITH REGRESSION

March 28, 2019

Pat Barry, Sr. Director
Ben Jones, Sr. Analyst

Contents

1 Introduction

"All models are wrong, but some models are useful." -George E.P. Box

Social data is hard. Like, really hard. The outcomes can vary wildly: CNN has 40+ million followers on their Twitter account and regularly gets a couple hundred engagements per post. On the other hand, a random post with 13K followers can go viral with 300K engagements.

While the world of social networking is beautiful, it is exceptionally difficult to predict. Unreliable data collection, sparse sample size, and the volatile nature of the industry means even the best forecasting models will struggle with large residuals. But, that does not mean we should make no attempt. When used properly, forecasting models constantly evolve and are an extremely powerful tool. It is a necessity to be as scientifically rigorous as possible.

This document is intended to serve as an introduction to *regression modeling*. While it does not demonstrate every intricate aspect, it is meant to educate the reader and provide a solid starting point. After all, experience is the best teacher.

Note: 'benchmark' and 'benchmarking' will be used interchangeably with 'forecast', 'predicted', and other similar terms throughout the document.

2 The Art of Forecasting

While the *modeling* aspect will be demonstrated in the following sections, it is just a piece of *forecasting*. The final ingredient is not shoving another variable into an equation, but an artistic component. Consider the following example, paraphrased from the Ben Reiter book *Astroball: The New Way to Win It All*.

You have two baseball players who on paper perform identically. Both players have the same batting average, on base percentage, and are equal in every baseball metric. However, one player practices hard and is a good teammate while the other is lazy and parties every night. If you could only keep one player, which would you choose?

Most would pick the hard working player. While it doesn't come across in the box score, selecting that player means a player's personality does hold some value.

Coming back to a social influencer example, there was a program promoting Lipton® via a buy-one-get-one deal. Using the exact models from this document, it projected the program to receive 5,000 engagements and 250 conversions. The actual results of the program were 17,000 engagements and 2,000 conversions. The art of forecasting comes into play here in the forecaster recognizing the program has the potential to perform better than the projection/typical program. Knowing that Lipton® is a premium, popular brand and the fact it included an enticing offer could cause the forecaster to

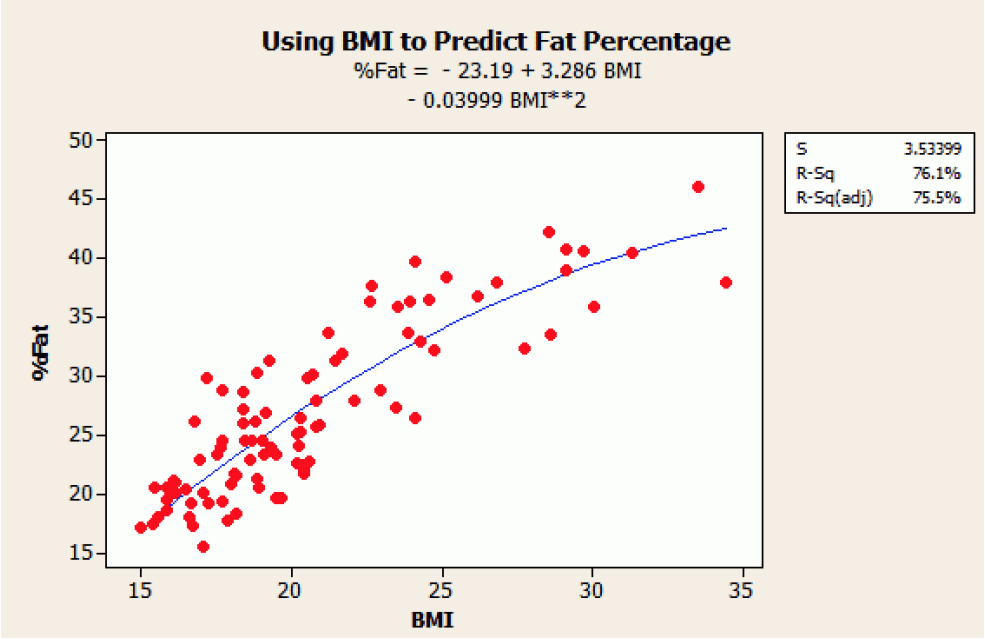
consider making an increase to the predictions the models spit out. If a program was for a small, unknown brand, it may be wise to consider making the final forecast a value lower than the initial output from the models. How does one know when to adjust a forecast based on gut-feeling? Experience.

In short, forecasting is both an art and a science. Finding the right balance can only be achieved through experience and is a constant process of improvement.

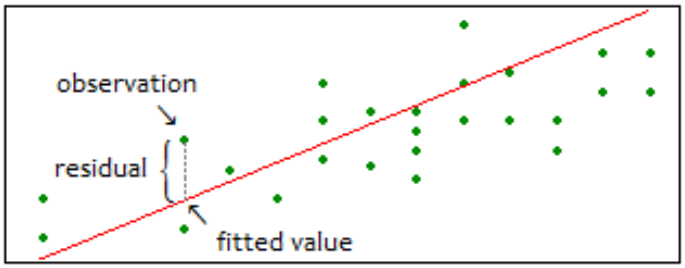
3 What is Regression?

Regression is a method of measuring the relationship between variables. Temperature and ice cream consumption. Height and weight. Dog breed and speed. You can even measure the affect of several variables at one time; this is known as **Multiple Regression** (ex. home price based on the # of rooms, sq. ft, neighborhood safety, etc.)

This technique is very flexible and thus is used across all types of industries and situations. In fact, you can use regression to model the relationship between any variables no matter how nonsensical. The number of letters in your name and income. Lunar cycle and internet speed. While you can build a model with whatever inputs you desire, the result will be painfully truthful if no legitimate relationship exists. Below is an example which forecasts the % Body Fat based on BMI. Note that regression does **not** need to be a straight line.



Regression aims to create a line of projected or fitted values (red line in graph below) which minimizes the difference between the projected value with the observed value. This difference between the observed and projected value is the **residual**. Regression will choose the line which minimizes the overall residual for all points as much as possible. This is the most common form of regression known as **Ordinary Least Squares**. See below for an illustration.



4 Forecasting Influencer Metrics

In this section we will demonstrate how to predict influencer metrics using regression. The metrics of interest are reach, actions, engagements, and conversions. We will go into detail for reach and engagements, but the same process follows for actions and conversions.

4.1 Gathering and Formatting Data

Your data source can come in a variety of forms. Automated data feeds (i.e. APIs) are generally preferred but is not always available. In this example, we are given a raw excel file. Here's a snapshot of the file:

	A	B	C	D	E	F	G	H	I	J
1	Job Number	Client Code	Client Descri	Division Cod	Division Des	Product Cod	Product Des	Job Descript	Brand	Start Date
2	6534	pgo	Perrigo	natl	National	otc	OTC	P: ODT Influe	Scale	1/31/19
3	6560	uni	Unilever HPC	wm	Walmart	skin	Skin	2019 WMT S	Scale	1/31/19
4	5227	nb	National Bra	NB	National Bra	pitch	Pitch	WMT Family	Tracfone	4/25/17
5	5484	tf	Tracfone Wi	wm	Walmart	mobile	Mobile	WMT Family	Tracfone	7/25/17
6	5949	Huaw	Huawei	Nat	National	Mob	Mobile	New Carrier	Scale	1/8/18
7	6533	pgo	Perrigo	natl	National	otc	OTC	P: ODT Influe	Scale	11/19/18
8	5469	pgo	Perrigo	natl	National	BYN	Baby	Infant Hypoa	Scale	7/21/17
9	5470	pgo	Perrigo	wm	Walmart	otc	OTC	2018 Omepr	Scale	7/21/17
10	6559	uni	Unilever HPC	wm	Walmart	hair	Hair	2019 Walma	Scale	1/8/19
11	6528	pgo	Perrigo	wm	Walmart	pet	Pet	P: PetArmor	Pet Armor	1/31/19
12	5468	pgo	Perrigo	natl	National	otc	OTC	NRT Nationa	Scale	7/21/17
13	5759	pgo	Perrigo	cos	Costco	otc	OTC	P: 2018 Ome	Kirkland Sigr	10/26/17
14	5410	mat	Mattel, Inc.	wm	Walmart	toys	Toys	Mattel Socia	Scale	6/26/17
15	6527	pgo	Perrigo	wm	Walmart	pet	Pet	P: PetArmor	Pet Armor	1/31/19
16	5463	ent	Entertainme	ENT	Entertainme	pitch	Pitch	Triad Social -	Pitch	7/19/17
17	5625	mirum	Mirum	natl	National	interc	Intercompar	P: Triad Soci	Pitch	10/12/17
18	6562	uni	Unilever HPC	wm	Walmart	scale	Scale	2019 WMT S	Scale	1/31/19

Since this is a small data set, we can do initial steps within excel. It's best to start by getting familiar with the raw data and metadata. How many records are there? How many variables (columns) are there? What data types are present? Are there many missing values? If the data is small enough,

as in this case, it is often easier to do as much reformatting (but not cleaning!) as possible in excel directly. Here we can already get a sense that many variables will not be used and we can focus on four variables that will *probably* be useful for modeling: planned impressions, program length, budget, and product category.

4.2 Exploratory Analysis and Cleaning

The data will now be loaded into the tool of choice. ***R***[1] will be used for this example. It was mentioned in the earlier section not to begin cleaning on the raw file. This is because countless modifications will be done and undone, so it is recommended to leave the original file untouched.

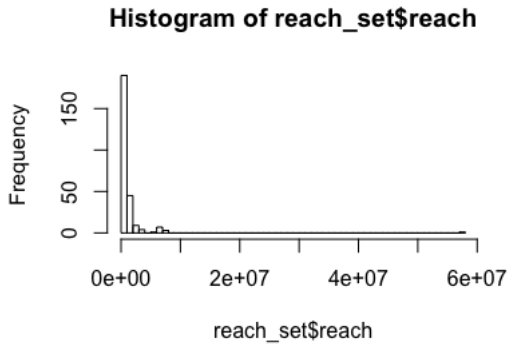
Right off the bat we notice some records have missing data (0's). It is best practice to try to correct these rows, but for now we'll simply remove them from the analysis.

	AJ	AK	AL
1	actions	conversions	engagements
94	7424	0	1810
95	15230	153	3039
96	37782	112	8172
97	26504	0	1627
98	20665	0	4575
99	12396	142	7046
100	1376	0	773

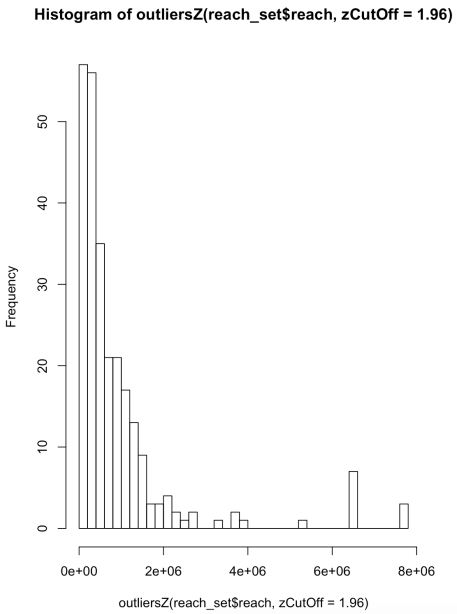
After removing incomplete records, what else should be cleaned? It is easy to spot missing data, but what about data that is present but potentially problematic? This is where exploratory analysis comes into play.

[1] R is a language and environment for statistical computing and graphics

Exploratory analysis is quickly gathering information on the data (ex. mean, median, max). Visualizations are tremendously useful in quickly getting a feel for the data and identifying potential problems. A good place to start is with a histogram[2]. Let's take a look at the distribution of the variable reach.



We quickly notice the vast majority of values are less than 1 million. It's also obvious there is at least one extremely large value. Identifying and cleaning **outliers** can greatly improve model performance. There are many methods on deciding how extreme a value is before it is considered an outlier; here we'll use **z-score**[3]. Identifying and removing these data points returns a much more digestible histogram and ultimately a better model. Side note: the distribution here is log-normal[4].

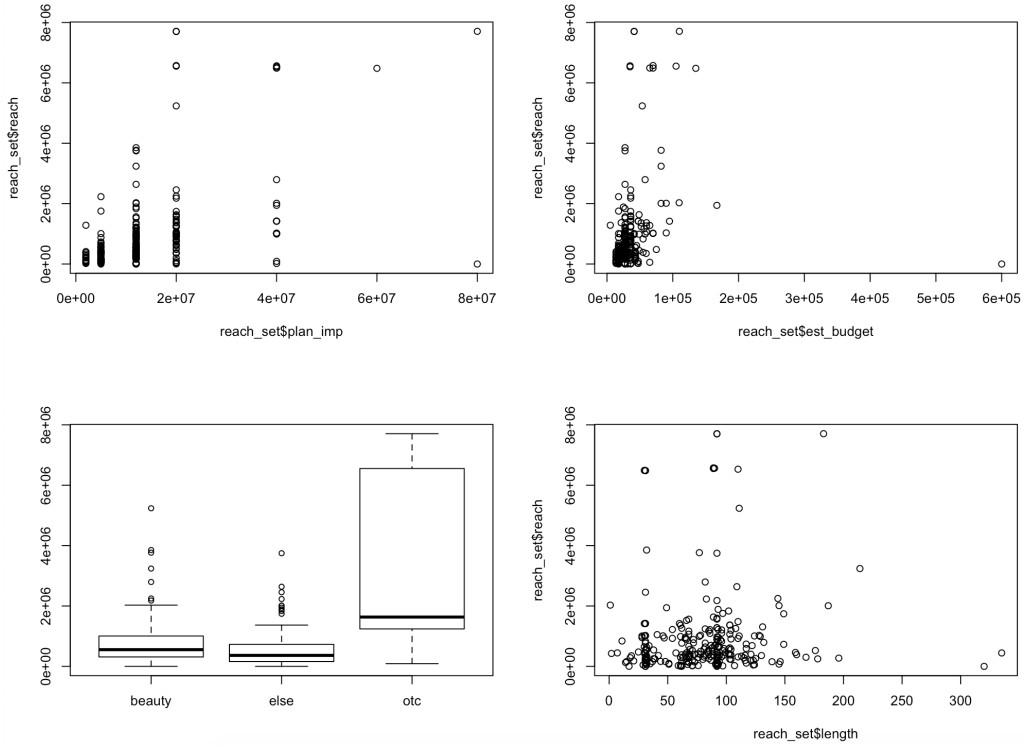


[2]Histograms visualize the distribution of data
[3]Z-score measures the distance (standard deviation) from the mean; 1.96 is a common cutoff
[4]log-normal distribution is when the log of the variable is normal

Outliers are removed for the other metrics in a similar fashion. It's also worth noting that R is an extremely powerful tool which can produce beautiful visualizations. The histograms previously shown made no effort to impress the eye but are useful enough for these first steps.

4.3 Modeling Reach - Iteration 1

Now that the data is more or less cleaned up, we can begin building a model to forecast reach. Similar to the exploratory analysis, plotting the data is a great first step. We'll start by plotting reach against the four variables we noted early as likely to be useful for building a model: planned impressions, program length, budget, and product category.



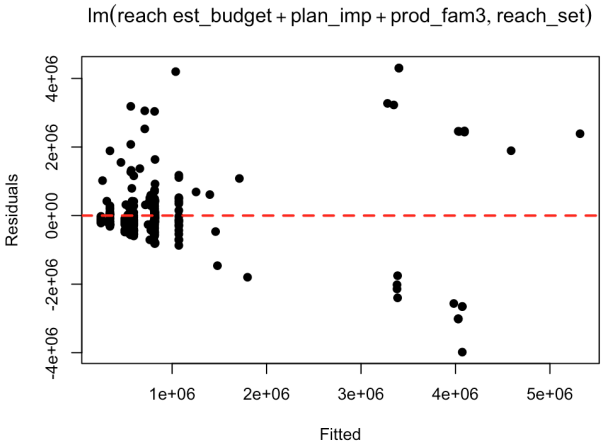
What we want to look for are patterns. The more obvious the pattern, the stronger the correlation and the more likely that variable is a significant predictor. There's a fairly noticeable pattern between reach and planned impressions. There might be a subtle trend with budget, but a very weak one. Products, when bucketed into these categories, do show signs that certain family of products perform differently than others. Lastly, there is no discernible pattern between program length and reach. As a result, we'll first attempt to model reach with the inputs of planned impressions, budget, and product category. Note that while not demonstrated in this document, these product buckets (beauty, else, otc) were chosen through trial and error, keeping the categories that improved the model the most.

```
Call:
lm(formula = reach ~ est_budget + plan_imp + prod_fam3, data = reach_set)

Residuals:
     Min       1Q   Median       3Q      Max
-3982935  -376267  -148720   191624  4306111

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.567e+05  1.273e+05   3.588 0.000399 ***
est_budget    -1.894e+00  2.280e+00  -0.831 0.406915
plan_imp       3.413e-02  1.014e-02   3.366 0.000880 ***
prod_fam3else -2.522e+05  1.382e+05  -1.825 0.069174 .
prod_fam3otc   2.340e+06  3.189e+05   7.337 2.94e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1058000 on 254 degrees of freedom
Multiple R-squared:  0.4344,    Adjusted R-squared:  0.4255
F-statistic: 48.77 on 4 and 254 DF,  p-value: < 2.2e-16
```



On the left we have the model summary. On the right a plot of the residuals. There's a lot of information provided in the model summary; the following is a brief explanation of some key components:

coefficients are the weights applied to the independent variables
p-value determines the level of significance of both the model overall and each independent variable (industry standard accepts values $<.05$)
adjusted R^2 explains how much variance the model can explain

Straight away there are some positive signs. The p-value is significant and the adjusted R^2 is very strong. We also see that budget variable has a negative coefficient and is not significant. This makes sense as we noted in the earlier plots there was not an obvious pattern between reach and budget. Even if budget was a significant variable, the negative coefficient means the projected reach volume decreases as the program budget increases (try explaining that one to the stakeholders!)

We also generated a scatter plot of the residuals. What we're looking for is a normal distribution with $\mu = 0$ (mean) and a constant σ^2 (variance). Another histogram (not shown here) can help with the mean, but due to the OLS method of regression we're using, this is guaranteed. Additionally, we can visually see that there's roughly the same amount of points above and below the red zero line (i.e. equal magnitudes of over and under predictions).

As for a constant variance σ^2, we can see this model does not satisfy that condition. Evidence of this is the "fan-shape": the residuals (distance from the zero line) increase as the values increase (move to the right). This is something to keep an eye on in the next iteration of the model.

4.4 Modeling Reach - Iteration 2

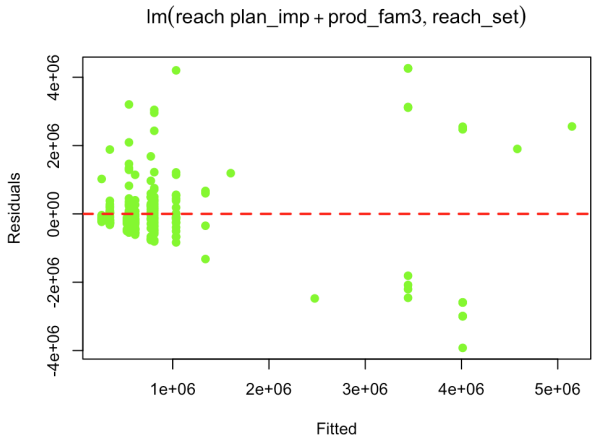
Since we saw that budget was not a significant predictor of reach, we will remove it from our first iteration.

```
Call:
lm(formula = reach ~ plan_imp + prod_fam3, data = reach_set)

Residuals:
     Min       1Q   Median       3Q      Max
-3923165  -362521  -159861   201544  4262700

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.675e+05  1.265e+05   3.695 0.000269 ***
plan_imp       2.838e-02  7.404e-03   3.833 0.000159 ***
prod_fam3else -2.626e+05  1.375e+05  -1.910 0.057279 .
prod_fam3otc   2.410e+06  3.076e+05   7.834 1.28e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1057000 on 255 degrees of freedom
Multiple R-squared:  0.4329,    Adjusted R-squared:  0.4262
F-statistic: 64.88 on 3 and 255 DF,  p-value: < 2.2e-16
```



Having removed budget from the equation, we can see from the summary the model is improved. It is still significant and has a strong R^2. Turning to the residual scatter plot, it looks unchanged from the first iteration. This should make sense, since the only change is the removal of the budget variable, which was not a significant predictor anyways.

However, we are still dealing with a non-constant variance σ^2 for our residuals. We'll attempt to correct this by employing a transformation. **Transformations** are operations which change the scale of measurement. Common examples are taking the log, square root, or adding polynomial terms. While blindly experimenting with various operations and seeing if the model improves can work, there does exist a method to determine an appropriate transformation.

4.5 Modeling Reach - Iteration 3

The **Box-Cox Transformation** is a good starting point for when transformations are suspected to be needed. The technique is named after George P.E. Box and Sir David Roxbee Cox, who developed the method in 1964. To be brief, the Box-Cox method considers a family of transformations:

$$g_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(y) & \lambda = 0 \end{cases}$$

λ is determined by maximizing the log-likelihood:

$$L(\lambda) = -\frac{n}{2}\log(RSS_\lambda/n) + (\lambda - 1)\sum \log(y_i)$$

The finer details of the method will not be discussed, but the original paper can be found here.

Applying the Box-Cox technique results in $\lambda = 0.3$, which means instead of modeling for reach, we will model for

$$\frac{reach^{0.3} - 1}{0.3}$$

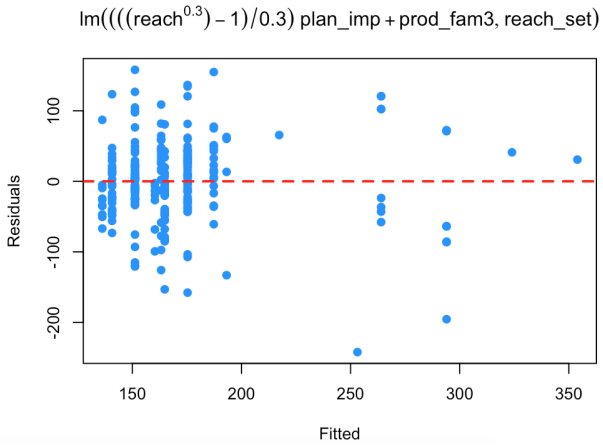
Applying this transformation results in

```
Call:
lm(formula = (((reach^0.3) - 1)/0.3) ~ plan_imp + prod_fam3,
    data = reach_set)

Residuals:
     Min       1Q   Median       3Q      Max
-242.150  -30.513    0.199   32.981  158.089

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.573e+02  6.923e+00  22.726  < 2e-16 ***
plan_imp       1.499e-06  4.051e-07   3.701 0.000263 ***
prod_fam3else -2.411e+01  7.524e+00  -3.205 0.001525 **
prod_fam3otc   7.670e+01  1.683e+01   4.557 8.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57.85 on 255 degrees of freedom
Multiple R-squared:  0.3025,    Adjusted R-squared:  0.2943
F-statistic: 36.87 on 3 and 255 DF,  p-value: < 2.2e-16
```



Recall that we experimented with a transformation because the residuals did not have a constant variance σ^2. The third iteration now displays a scatter plot with a constant variance (no longer fan-shaped). However, the model's predictive power dramatically decreased as the adjusted R^2 decreased from 0.4262 to 0.2943.

4.6 Modeling Reach - Final Model

Ultimately Iteration 2 is the best model. While forecasting model 3 did address the non-constant variance issue, solving that was achieved by sacrificing the model's overall predictive ability. Forecasting model 2 overall is superior enough over version 3, that we're able to accept a non-constant variance. Additionally, the vast majority of the data points, and thus programs, are on the far left (before the fan drastically opens). This means most of the programs the model will be predicting will not be greatly affected by this.

Having decided on Iteration 2, the following is the model equation:

$$reach = 467500 + .02838 * plannedimpressions + z$$

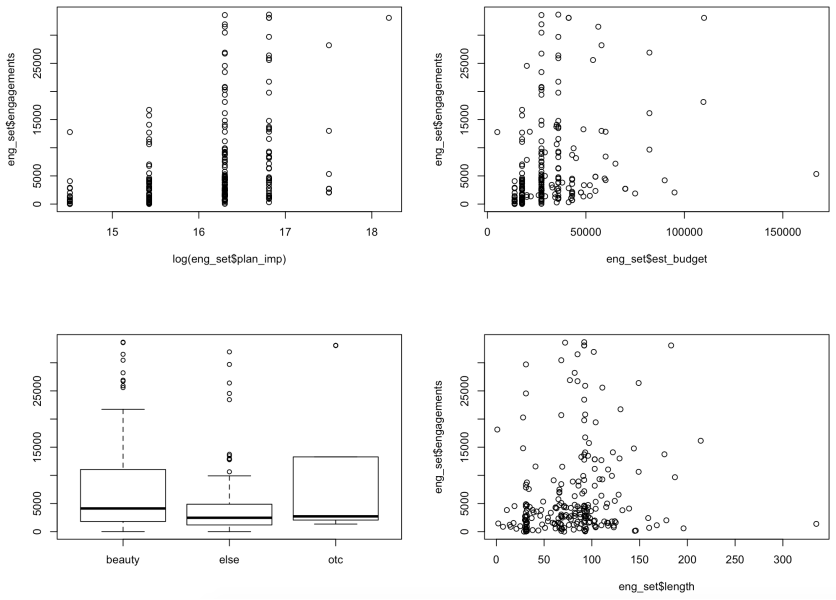
$$z = \begin{cases} 0 & \text{beauty} \\ 2410000 & \text{otc} \\ -262600 & \text{else} \end{cases}$$

where z is product category.

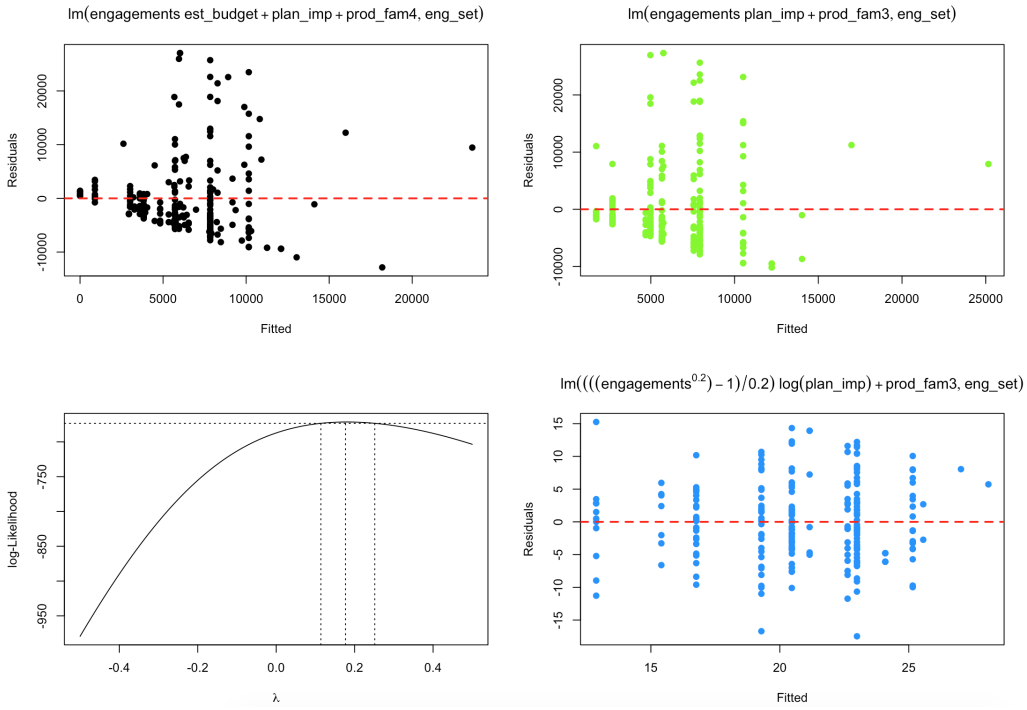
Additionally, to avoid **overfitting**, it is best to separate the data set into training and validation sets. Building the model would occur on training and the resulting model would be evaluated on validation (e.g. Monte Carlo, K-Fold). Admittedly the data set here is small enough where a split would not be useful, but it is recommended to follow this best practice.

4.7 Modeling Engagements

This second example will demonstrate building a forecast model to predict engagements. Similar to the first example, we begin by cleaning the data and removing outliers. Next, we plot the suspected variable of interest, each in its own plot against the dependent variable, engagements.



Again we note there appears to be no pattern between program length and engagements, so that variable will not be included in any iterations. We then run several iterations as we did before.



Starting at the top left plot, the same iterative process is used. First attempt finds that budget is not useful. Second iteration has non-constant variance. Using the Box-Cox technique, we find the transformation where $\lambda = 0.2$ is appropriate to fix the variance issue. And while not shown here, in this example the third iteration not only fixes the variance problem, but also is the best performing model overall. Unlike the model of reach, the model for engagements will keep the transformation. The resulting equation of the model is as follows:

$$\frac{reach^{0.2} - 1}{0.2} = -46.0215 + 4.2338 * \ln(plannedimpressions) + z$$

$$z = \begin{cases} 0 & \text{beauty} \\ -3.9985 & \text{otc} \\ -2.5262 & \text{else} \end{cases}$$

Note that we also transformed the independent variable of planned impressions by taking the natural log. This was found through experimentation and finding that this also improves the model's performance.

5 Closing Remarks

Forecasting is an extremely difficult task. It is also one of the coolest projects to work on: you get to literally predict the future.

This document demonstrated two examples of regression applied to social influencers. A helpful thing to keep in mind is remind yourself of the industry: An $R^2 = 0.32$ would be difficult to justify for an experimental drug test but is not terrible for social. There will even be instances where no matter how hard you try, no useable model is possible with the provided data.

To reiterate, this is not a comprehensive record of all the work that goes into building a model via regression. The purpose is provide an introduction and starting point for those new to statistics and modeling/forecasting. The key takeaway is not the result but the process.

"The irony is that by being less focused on your results, you may achieve better ones." -Nate Silver