# 2018 Model Extension

*Sean Davern*

*June 18, 2019*

## Overview

This notebook extends the modeling done in 2013 to include data through much of 2018.

## Load Data

Loading cleaned and transformed giving data:

```
df <- readRDS("../data/Cleaned and Transformed Giving Data.rds")
head(df)
```

```
##   week.ending    month year    total monthly.giving.families
## 1  2010-01-31  January 2010 14241.15                      45
## 2  2010-02-28 February 2010 22437.50                      41
## 3  2010-03-28    March 2010 13317.00                      32
## 4  2010-04-25    April 2010 13232.50                      36
## 5  2010-05-30      May 2010 13478.15                      40
## 6  2010-06-27     June 2010 14930.50                      43
##   SundaysInMonth MonthsGivingPerWeek
## 1              5            2848.230
## 2              4            5609.375
## 3              4            3329.250
## 4              4            3308.125
## 5              5            2695.630
## 6              4            3732.625
```

## Columns to Accomodate Analysis

I'll add a few columns to enable asking analysis "what if's" and adjust for anomolies:

```
library("dplyr", quietly=TRUE, warn.conflicts = FALSE)
df <- mutate(df,
             excluded = FALSE,
             removed.amount = 0,
             original.total = total)
```

## Accommodating Unusual Gift

12/26/2016 had an unusual single gift skewing analyses. Other large gifts have been recieved over this period but this gift was at least twice all others. I'm removing the amount of the unusual gift from the month's total. Previously I excluded that month all-together, but I've commented that line out so no values are excluded.

```
df$removed.amount[as.Date(df$week.ending) == "2016-12-25"] <- 20000
df$total <- df$original.total - df$removed.amount
#df$excluded[as.Date(df$week.ending) == "2016-12-25"] <- TRUE
```

Note: we're not excluding the 3 months consider "outliers" in the 2013 analysis. Now with 9 years of history, those months don't appear as much to be outliers.
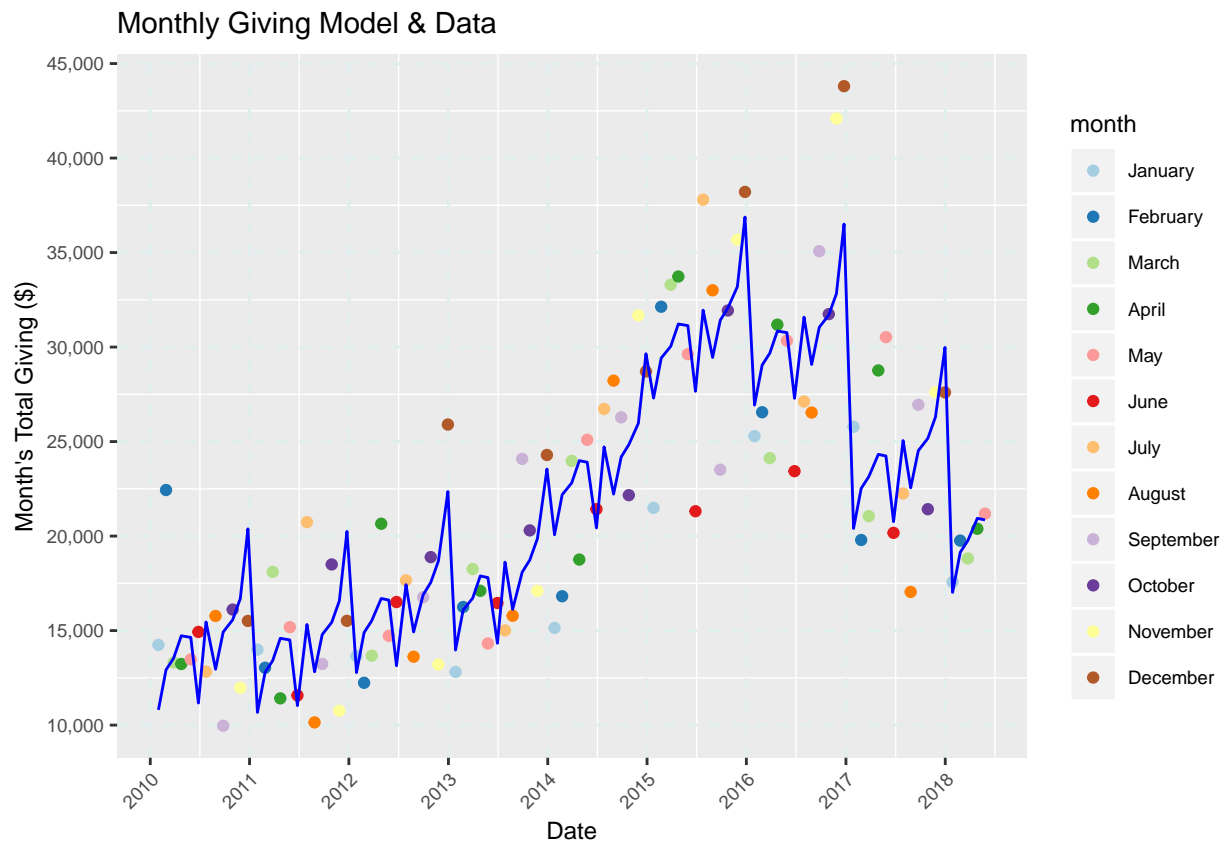
## Modeling

Regressing the R version of the 2013 model:

$$Monthly\ Giving = a + b_{year} + c_{month}$$

```r
mod <- lm(
  formula = total ~
    year + month,
  data = df[df$excluded!=TRUE,]
)
```

. . . gives the following

## Model Predictions



In this case it is far more clear that giving was increasing through 2016 which would be expected to improve the statistical significance of the *year* model term over what was seen in the 2013 analysis.

2

# Model Fit Assessment

```
## Analysis of Variance Table
##
## Response: total
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## year         8 3920266253  490033282  30.198 < 2.2e-16 ***
## month       11  588525930   53502357   3.297  0.000896 ***
## Residuals   81 1314433124   16227569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
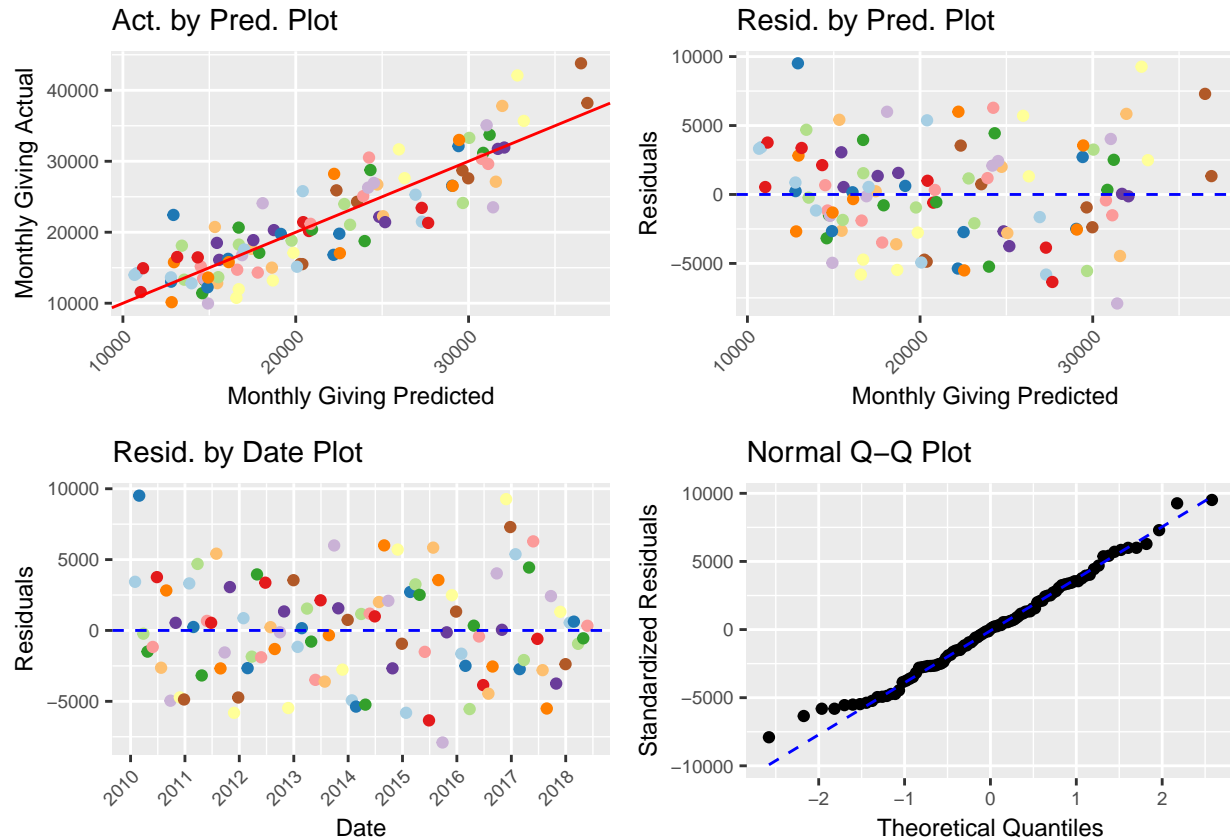


Both model terms are highly statistically significant. The residuals seem nicely randomly scattered. I don't see any particularly concerning patterns in the plots including model biases by month (dot color). The quantile plot shows very normally distributed residuals.

# Regressed Model Details

```
##
## Call:
## lm(formula = total ~ year + month, data = df[df$excluded != TRUE,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7904.4 -2659.0    46.1  2486.6  9513.2
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10807.3     1741.1   6.207 2.18e-08 ***
## year2011          -133.7     1644.6  -0.081  0.93542
## year2012          1974.0     1644.6   1.200  0.23352
## year2013          3165.1     1644.6   1.925  0.05779 .
## year2014          9265.5     1644.6   5.634 2.47e-07 ***
## year2015         16494.5     1644.6  10.030 7.42e-16 ***
## year2016         16124.2     1644.6   9.805 2.05e-15 ***
## year2017          9597.9     1644.6   5.836 1.06e-07 ***
## year2018          6215.8     2198.7   2.827  0.00592 **
## monthFebruary     2117.0     1899.0   1.115  0.26824
## monthMarch        2739.3     1899.0   1.443  0.15301
## monthApril        3918.4     1899.0   2.063  0.04227 *
## monthMay          3833.6     1899.0   2.019  0.04682 *
## monthJune          359.5     1968.9   0.183  0.85559
## monthJuly         4648.5     1968.9   2.361  0.02063 *
## monthAugust       2149.1     1968.9   1.092  0.27828
## monthSeptember    4115.6     1968.9   2.090  0.03973 *
## monthOctober      4765.5     1968.9   2.420  0.01774 *
## monthNovember     5895.7     1968.9   2.994  0.00365 **
## monthDecember     9574.9     1968.9   4.863 5.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4028 on 81 degrees of freedom
## Multiple R-squared:  0.7743, Adjusted R-squared:  0.7213
## F-statistic: 14.62 on 19 and 81 DF,  p-value: < 2.2e-16
```

# Model Generalization

We can generalize the model again, as described in the "Transfer of 2013 Analysis" Notebook, by shifting the coefficients to the average month and year like this:

$$Monthly\ Giving = \left[a + \overline{b_{year}} + \overline{c_{month}}\right] + \left(b_{year} - \overline{b_{year}}\right) + \left(c_{month} - \overline{c_{month}}\right)$$

where $\overline{b_{year}}$ is the mean of $b_{year}$ coefficients, including $b_{2010} = 0$, and $\overline{c_{month}}$ is the mean of the $c_{year}$ coefficients, including $c_{January} = 0$.

## Monthly Variation

This plot shows the resulting *month* regression coeffiecients shifted by $\overline{c_{month}}$ capturing month-to-month variation. $((c_{month} - \overline{c_{month}}))$ The shape of the yearly repeated pattern in the prediction are the result of these terms.

### Giving Variation By Month
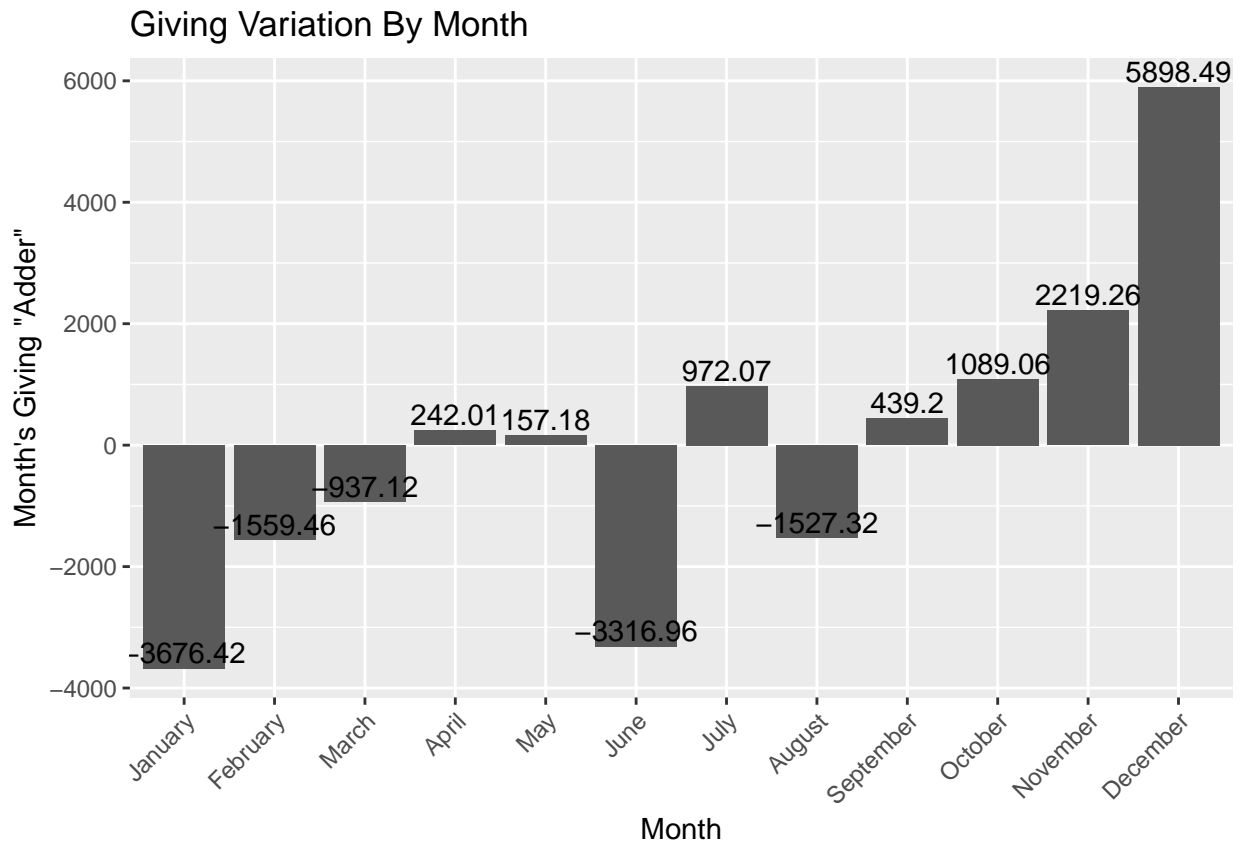


This plot is generally similar to Figure 13 in the 2013 Analysis with the exception of November. Since it is based on 9 years of Seed history, I'll judge that it is more representative of an average year.