STAT495 HW#8

YOUR NAME HERE

December 3, 2015

This homework is due at midnight on Tuesday, December 8th (one report per pair). Please add, commit, and push the Rmd and pdf files to your public github repository as well as create and then close an issue for "HW#8". Your task is to improve upon your model from HW#6, in conjunction with your partner.

The files for this assignment can be found in the **STAT495-Horton** github repository (called **hw08.Rda**). This is a slightly modified dataset: fewer stores, no days with zero sales, and no missing values.

It will be important to control for store in your model.

```
load("hw08.Rda")
glimpse(merged)
```

```
## Observations: 65,631
## Variables: 8
## $ store     (int) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1...
## $ dow       (int) 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ sales     (int) 6206, 10107, 11299, 14451, 7692, 6803, 11956, 8752, ...
## $ customers (int) 589, 850, 862, 1724, 659, 661, 969, 802, 697, 546, 1...
## $ promotion (int) 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ noschool  (int) 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1...
## $ type      (chr) "c", "a", "a", "c", "a", "a", "a", "a", "a", "a", "a...
## $ distance  (int) 1270, 570, 14130, 620, 29910, 310, 24000, 7520, 2030...
```

```
mod1 <- lm(sales ~ customers + promotion  + noschool + type + distance, data=merged)
msummary(mod1)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.302e+01  1.353e+01   3.920 8.87e-05 ***
## customers    7.938e+00  1.702e-02 466.453  < 2e-16 ***
## promotion    1.119e+03  9.082e+00 123.215  < 2e-16 ***
## noschool     4.469e+01  1.105e+01   4.044 5.27e-05 ***
## typeb       -1.380e+03  3.953e+01 -34.923  < 2e-16 ***
## typec       -9.119e+01  1.360e+01  -6.706 2.02e-11 ***
## typed        1.240e+03  1.015e+01 122.220  < 2e-16 ***
## distance     3.464e-02  6.410e-04  54.040  < 2e-16 ***
##
## Residual standard error: 1108 on 65623 degrees of freedom
## Multiple R-squared:  0.819,  Adjusted R-squared:  0.819
## F-statistic: 4.243e+04 on 7 and 65623 DF,  p-value: < 2.2e-16
```

```
sqrt(mean((predict(mod1) - merged$sales)^2))
```

```
## [1] 1107.689
```

```
# another (better) model
mod2 <- lm(sales ~ as.factor(store) + customers + promotion  + noschool + type + distance, data=merged)
msummary(mod2)
```

```
##                     Estimate Std. Error  t value Pr(>|t|)
## (Intercept)       -1.485e+03  2.573e+01  -57.694  < 2e-16 ***
## as.factor(store)2   1.656e+01  3.378e+01    0.490  0.62395
## as.factor(store)3   2.224e+02  3.400e+01    6.541 6.15e-11 ***
## as.factor(store)4  -3.056e+03  3.651e+01  -83.698  < 2e-16 ***
## as.factor(store)5   2.172e+02  3.385e+01    6.417 1.40e-10 ***
## as.factor(store)6   1.893e+00  3.386e+01    0.056  0.95541
## as.factor(store)7   3.532e+01  3.445e+01    1.025  0.30518
## as.factor(store)8  -1.998e+02  3.381e+01   -5.908 3.48e-09 ***
## as.factor(store)9   1.612e+03  3.383e+01   47.651  < 2e-16 ***
## as.factor(store)10  4.886e+02  3.378e+01   14.464  < 2e-16 ***
## as.factor(store)11 -2.642e+03  3.538e+01  -74.688  < 2e-16 ***
## as.factor(store)12 -6.371e+02  3.431e+01  -18.569  < 2e-16 ***
## as.factor(store)13  2.132e+03  3.646e+01   58.464  < 2e-16 ***
## as.factor(store)14  5.178e+01  3.385e+01    1.530  0.12612
## as.factor(store)15  8.761e+02  3.382e+01   25.903  < 2e-16 ***
## as.factor(store)16 -9.031e+01  3.424e+01   -2.637  0.00836 **
## as.factor(store)17 -7.928e+02  3.406e+01  -23.278  < 2e-16 ***
## as.factor(store)18  2.232e+02  3.390e+01    6.583 4.64e-11 ***
## as.factor(store)19  1.050e+03  3.384e+01   31.034  < 2e-16 ***
## as.factor(store)20 -5.412e+00  3.669e+01   -0.148  0.88272
## as.factor(store)21  1.187e+03  3.387e+01   35.059  < 2e-16 ***
## as.factor(store)22  3.858e+02  3.636e+01   10.610  < 2e-16 ***
## as.factor(store)23  1.624e+03  3.387e+01   47.937  < 2e-16 ***
## as.factor(store)24  1.754e+03  3.417e+01   51.337  < 2e-16 ***
## as.factor(store)25 -4.149e+03  3.895e+01 -106.525  < 2e-16 ***
## as.factor(store)26  1.992e+03  3.381e+01   58.917  < 2e-16 ***
## as.factor(store)27 -6.570e+02  3.509e+01  -18.723  < 2e-16 ***
## as.factor(store)28  3.621e+02  3.398e+01   10.654  < 2e-16 ***
## as.factor(store)29  1.631e+03  3.386e+01   48.176  < 2e-16 ***
## as.factor(store)30 -1.741e+03  3.412e+01  -51.034  < 2e-16 ***
## as.factor(store)31  8.720e+02  3.378e+01   25.811  < 2e-16 ***
## as.factor(store)32  2.330e+02  3.634e+01    6.411 1.45e-10 ***
## as.factor(store)33  8.269e+02  3.419e+01   24.183  < 2e-16 ***
## as.factor(store)34 -1.214e+03  3.466e+01  -35.030  < 2e-16 ***
## as.factor(store)35  3.094e+03  3.394e+01   91.144  < 2e-16 ***
## as.factor(store)36 -9.540e+02  3.754e+01  -25.416  < 2e-16 ***
## as.factor(store)37  8.317e+01  3.407e+01    2.441  0.01464 *
## as.factor(store)38  1.538e+03  3.384e+01   45.459  < 2e-16 ***
## as.factor(store)39 -5.715e+02  3.384e+01  -16.886  < 2e-16 ***
## as.factor(store)40 -9.155e+02  3.392e+01  -26.987  < 2e-16 ***
## as.factor(store)41  2.293e+03  3.643e+01   62.930  < 2e-16 ***
## as.factor(store)42 -1.956e+02  3.521e+01   -5.557 2.76e-08 ***
## as.factor(store)43  1.613e+03  3.384e+01   47.652  < 2e-16 ***
## as.factor(store)44 -6.764e+01  3.381e+01   -2.001  0.04541 *
## as.factor(store)45  2.150e+03  3.395e+01   63.321  < 2e-16 ***
## as.factor(store)46 -4.212e+02  3.632e+01  -11.595  < 2e-16 ***
## as.factor(store)47 -1.354e+03  3.438e+01  -39.386  < 2e-16 ***
## as.factor(store)48  5.888e+02  3.387e+01   17.383  < 2e-16 ***
```

```
## as.factor(store)49    2.866e+03  3.386e+01   84.638  < 2e-16 ***
## as.factor(store)50    1.649e+03  3.406e+01   48.412  < 2e-16 ***
## as.factor(store)51    2.567e+03  3.631e+01   70.682  < 2e-16 ***
## as.factor(store)52    3.329e+03  3.630e+01   91.724  < 2e-16 ***
## as.factor(store)53    4.857e+02  3.378e+01   14.377  < 2e-16 ***
## as.factor(store)54    1.373e+03  3.391e+01   40.487  < 2e-16 ***
## as.factor(store)55   -9.512e+01  3.383e+01   -2.812  0.00493 **
## as.factor(store)56    3.122e+03  3.383e+01   92.285  < 2e-16 ***
## as.factor(store)57    1.452e+03  3.492e+01   41.590  < 2e-16 ***
## as.factor(store)58    2.036e+03  3.628e+01   56.114  < 2e-16 ***
## as.factor(store)59    7.376e+02  3.381e+01   21.819  < 2e-16 ***
## as.factor(store)60    1.672e+03  3.392e+01   49.287  < 2e-16 ***
## as.factor(store)61   -1.129e+02  3.386e+01   -3.336  0.00085 ***
## as.factor(store)62    1.362e+02  3.394e+01    4.013 6.00e-05 ***
## as.factor(store)63    1.393e+03  3.380e+01   41.200  < 2e-16 ***
## as.factor(store)64    3.147e+03  3.406e+01   92.377  < 2e-16 ***
## as.factor(store)65    1.486e+03  3.390e+01   43.847  < 2e-16 ***
## as.factor(store)66    1.776e+03  3.387e+01   52.439  < 2e-16 ***
## as.factor(store)67    6.021e+02  3.411e+01   17.651  < 2e-16 ***
## as.factor(store)68   -3.334e+03  3.542e+01  -94.120  < 2e-16 ***
## as.factor(store)69   -1.737e+03  3.565e+01  -48.719  < 2e-16 ***
## as.factor(store)70    6.298e+02  3.384e+01   18.612  < 2e-16 ***
## as.factor(store)71    2.405e+03  3.399e+01   70.750  < 2e-16 ***
## as.factor(store)72    5.975e+02  3.631e+01   16.459  < 2e-16 ***
## as.factor(store)73    5.354e+02  3.381e+01   15.835  < 2e-16 ***
## as.factor(store)74   -1.533e+03  3.428e+01  -44.712  < 2e-16 ***
## as.factor(store)75    4.622e+02  3.380e+01   13.671  < 2e-16 ***
## as.factor(store)76    4.274e+02  3.685e+01   11.598  < 2e-16 ***
## as.factor(store)77    1.793e+03  3.385e+01   52.982  < 2e-16 ***
## as.factor(store)78   -3.242e+02  3.386e+01   -9.576  < 2e-16 ***
## as.factor(store)79    7.675e+02  3.386e+01   22.670  < 2e-16 ***
## as.factor(store)80    2.711e+03  3.383e+01   80.146  < 2e-16 ***
## as.factor(store)81    2.212e+03  3.628e+01   60.973  < 2e-16 ***
## as.factor(store)82    1.395e+03  3.419e+01   40.800  < 2e-16 ***
## as.factor(store)83    1.416e+03  3.408e+01   41.553  < 2e-16 ***
## as.factor(store)84   -5.148e+02  3.707e+01  -13.886  < 2e-16 ***
## as.factor(store)85   -2.147e+03  3.342e+01  -64.251  < 2e-16 ***
## as.factor(store)86   -4.999e+02  3.385e+01  -14.769  < 2e-16 ***
## as.factor(store)87   -6.648e+02  3.391e+01  -19.601  < 2e-16 ***
## as.factor(store)88    7.569e+02  3.384e+01   22.367  < 2e-16 ***
## as.factor(store)89    5.679e+02  3.630e+01   15.645  < 2e-16 ***
## as.factor(store)90   -1.507e+03  3.501e+01  -43.052  < 2e-16 ***
## as.factor(store)91    2.232e+02  3.386e+01    6.593 4.35e-11 ***
## as.factor(store)92    1.391e+03  3.383e+01   41.127  < 2e-16 ***
## as.factor(store)93    7.900e+02  3.384e+01   23.347  < 2e-16 ***
## as.factor(store)94    1.551e+03  3.385e+01   45.802  < 2e-16 ***
## as.factor(store)95   -7.433e+00  3.422e+01   -0.217  0.82807
## as.factor(store)96    6.382e+02  3.383e+01   18.862  < 2e-16 ***
## as.factor(store)97    6.186e+02  3.390e+01   18.248  < 2e-16 ***
## as.factor(store)98    1.509e+03  3.391e+01   44.490  < 2e-16 ***
## as.factor(store)99    1.376e+03  3.634e+01   37.863  < 2e-16 ***
## as.factor(store)100   1.902e+03  3.667e+01   51.867  < 2e-16 ***
## customers             1.042e+01  1.843e-02  565.635  < 2e-16 ***
## promotion             7.418e+02  5.652e+00  131.246  < 2e-16 ***
```

```
## noschool                  -3.120e-01  6.238e+00   -0.050  0.96010
##
## Residual standard error: 624.2 on 65528 degrees of freedom
## Multiple R-squared:  0.9426, Adjusted R-squared:  0.9425
## F-statistic: 1.055e+04 on 102 and 65528 DF,  p-value: < 2.2e-16
```

```r
sqrt(mean((predict(mod2) - merged$sales)^2))
```

```
## [1] 623.757
```

Use this code as a starting point along with your two draft answers for HW#6 to improve anwers to the following questions.

Predicting sales

You've been promoted to a new position that gives you responsibility for monitoring sales at a chain of stores. Your goal is to predict sales at the store level. Data fields that are available include the `store`, `sales` (in dollars), `customers` (number of customers), `storeopen` (whether the store was open that day), `promotion` (whether there was a sale on), `noschool` (whether there was school that day), `type` (type of store: 4 levels), and `distance` (distance to nearest competitor).

Part one Part one of your challenge is to generate a model for sales that will be used to predict sales for future months. Your R Markdown file should generate a model called `finalmod()` which I will use to generate predicted values and calculate test error.

SOLUTION:

Part two Part two of your challenge is to interpret your model: what do you conclude about predictors of sales?

SOLUTION:

Part three Part three of your challenge is to assess your model: how well does it work for the training data?

SOLUTION:

Visualizing patterns of sales In addition to your predictive model, please generate one visualization of the data that tells a story related to sales. Please pay particular attention to the quality of your visualization. You should include a single paragraph interpretation of the figure.

SOLUTION: