Pam Sue

Frederick Pickering, CFA

6/18/2019

## Introduction

PamSue is a client department store operating in the south-eastern United States. The store primarily serves lower income residents. Recently PamSue’s forecasts have been missing the market when it comes to selecting the placement for new stores. They are looking to add more analytics to the process of selecting the location for new stores. In addition to regression analysis they are looking at a new metric created by the real estate team where they analyze the competitive locations.

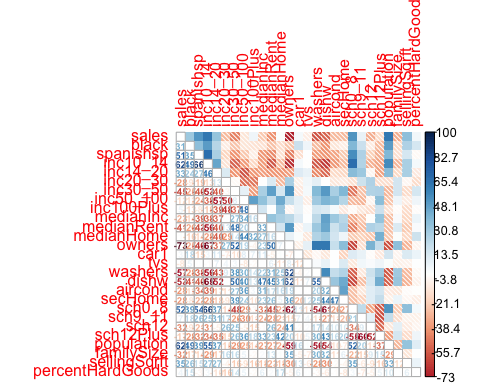
This analysis was conducted in R and the document was build using RMarkdwon. To see the web version of this report please go to XYZ.com.

## Data Prep

### Correlation Analysis

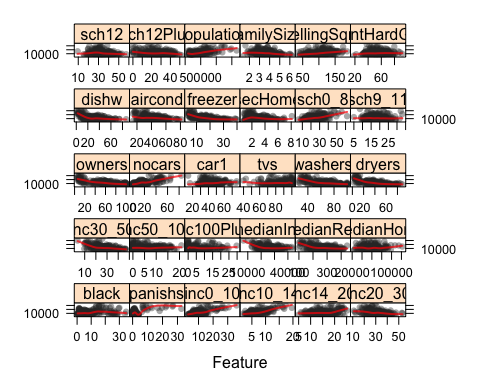
Before we can start running regressions we need to fimaliraze ourselves with the underlying data. Here we will look at scatter plots of the various regressors (features) and howt hey compare against the dependent variable (sales). We’ll also look at correlation matricies to check for possible multicolinearity in the variables.

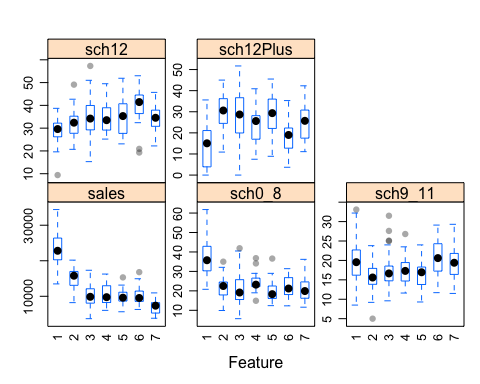
The correlation plot of the various regressors to sales shows us that there is a strong correlation between the sales of a store and the percentage of spanish speaking people in a neighborhood. The correlation also shows us that in neighborhoods where the percentage of residents earn between 10 and 14 thousand dollars has a strong relationship with sales in the sore. We also see strong negative correlations in the data set. Home ownership has a strong negative correlation with sales and the more in home utilizes. For example homes with dishwashers and dryers are more likely to see a strong negative relationship with sales.

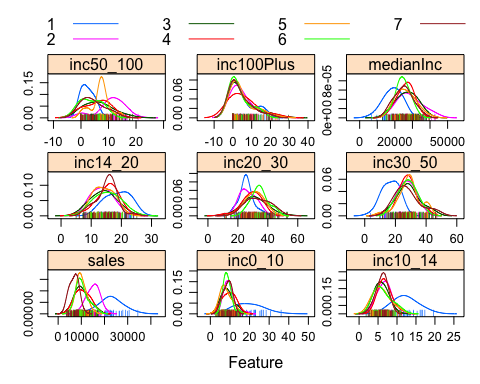


### Feature Engineering

# featurePlot(x = rawData %>% select\_at(),   
 # y = rawData$comType,   
 # plot = "ellipse",  
 # ## Add a key at the top  
 # auto.key = list(columns = 3))







## Queston 1

*As a possible alternative to the subjective “competitive type” classification, how well can you forecast sales using the demographic variables (along with the store size and the percentage of hard goods)? What does your model reveal about the nature of location sites that are likely to have higher sales?*

Before we can begin to assess the accuracy of a model with all of these variables we nee to find the significant variables in the dataset. Not all variables are going to be useful in predicting the sales of a store. In this model I start by making a linear regression model in R with all of the variables minus *Competitive Type*. The data has already been loaded into the R enviroment and the resulting line for a linear regression model is lm(sales ~ . -comType, data = regData) and assign the model to q1Reg. This provides a linear regression model with all of the varialbes minus compType. The next step is to assess the varialbes importance. Caret provides useful funciton for pulling variables importance from a model the function varImp(q1Reg) provides a ranking of the most important varialbes. This funciton looks ranks the varialbes by their contribution to lower the standard error of the model. I’ve taken the top 15 largest contributors to the model and passed them into a second linear regression model and I’ve come with with a linear regression model in the exhibits below.

This model is able to explain approximately 68% of the variation in the dataset and has an adjusted R^2 of 66%. This top down moethod of modeling sales provides a slightly higher R^2 and adjusted R^2 values than the defaul stepwise methodology provided by the RESESSIONS.XLSX file.

This regression model tells us several important things aabout the locations of PamSue stores. From the model I created population is the most important variable. This stands to reason as more populus locations are more likely to drive a higher level of foortraffic into the store and therefore more sales. The second and third most important vaiables are median home and selling square feet. These variables suggest that wealtheir neighborhoods and larger stores are going to see a higher level of sales. Interestingly, demographic variables such as a home having an air conditioner, freezer, dryers, or a car are more likely to see lower level of sales. This suggests that PamSue does cater to a lower wealth demographic compared to the overall population.

| **Metric** | **Value** |
| --- | --- |
| r.squared | 0.683 |
| adj.r.squared | 0.662 |
| sigma | 3167.674 |
| statistic | 33.580 |
| p.value | 0.000 |
| df | 16.000 |
| logLik | -2361.655 |
| AIC | 4757.311 |
| BIC | 4817.176 |
| deviance | 2347993814.365 |
| df.residual | 234.000 |

| **Term** | **Beta** | **Standard Error** | **T-Stat** | **P-Value** |
| --- | --- | --- | --- | --- |
| (Intercept) | 16916.622 | 2793.569 | 6.056 | 0.000 |
| population | 0.005 | 0.001 | 5.059 | 0.000 |
| medianHome | 0.057 | 0.015 | 3.883 | 0.000 |
| sellingSqrft | 21.864 | 7.609 | 2.873 | 0.004 |
| sch9\_11 | -142.762 | 49.355 | -2.893 | 0.004 |
| spanishsp | 266.625 | 70.510 | 3.781 | 0.000 |
| aircond | -52.976 | 20.193 | -2.623 | 0.009 |
| freezer | -118.013 | 48.807 | -2.418 | 0.016 |
| sch0\_8 | -68.139 | 34.728 | -1.962 | 0.051 |
| owners | -31.579 | 13.991 | -2.257 | 0.025 |
| car1 | -42.689 | 22.791 | -1.873 | 0.062 |
| secHome | -287.017 | 185.715 | -1.545 | 0.124 |
| dryers | -41.898 | 21.058 | -1.990 | 0.048 |
| black | -50.874 | 34.915 | -1.457 | 0.146 |
| inc10\_14 | 190.643 | 135.164 | 1.410 | 0.160 |
| inc50\_100 | 89.475 | 52.155 | 1.716 | 0.088 |

## Question 2

*How good is the “competitive type” classification method (along with using the store size and percentage of hard goods) at predicting sales.*

Competitive type is a rather good predictor at estimating sales. Just using competitive type, percent hard goods, and selling square feet provdes a model with an adjusted R^2 of 70%. This means the dummy variables, and selling square feet are better at predicting sales than the demographic data provided in the dataset. This means that the real estate deparment has come up with a very useful methodology for categorizing store locations. Interestingly percent hard goods does not produce a statistically significant result when included in the model. This either means it’s explanitory analysis is dwarfed by the relationship between sales and the other variables or sales is not strongly related to percent hard goods.

This is a strong model with all of the regressors having t-stats that are significnatlly above 2. The p-values also suggest that there is a s very small chance we are accepting a false Beta as true. Note these values show the estimated increase in sales over competitive type seven. Competitive type seven includes ***stores located along the sides of major roads*** meaning any other cometetitive type location will see higher sales than stores simply placed alongside major roads.

| **Metric** | **Value** |
| --- | --- |
| r.squared | 0.714 |
| adj.r.squared | 0.705 |
| sigma | 2959.262 |
| statistic | 86.180 |
| p.value | 0.000 |
| df | 8.000 |
| logLik | -2348.843 |
| AIC | 4715.686 |
| BIC | 4747.379 |
| deviance | 2119250366.690 |
| df.residual | 242.000 |

| **Term** | **Beta** | **Standard Error** | **T-Stat** | **P-Value** |
| --- | --- | --- | --- | --- |
| (Intercept) | 5219.535 | 725.403 | 7.195 | 0.000 |
| comType1 | 14895.319 | 837.641 | 17.782 | 0.000 |
| comType2 | 7192.957 | 774.730 | 9.284 | 0.000 |
| comType3 | 2544.779 | 685.767 | 3.711 | 0.000 |
| comType6 | 3030.626 | 868.942 | 3.488 | 0.001 |
| sellingSqrft | 24.234 | 7.192 | 3.370 | 0.001 |
| comType4 | 2677.296 | 907.221 | 2.951 | 0.003 |
| comType5 | 2076.237 | 822.162 | 2.525 | 0.012 |

## Q3

**Two sites, A and B, are currently under consideration for the next new store opening. Characteristics of the two sites are provided below in Table B. Which site would you reccommend? What sales forecasting approact would you reccommend.**

q3All <-  
 train(  
 sales ~ .,  
 regData,  
 method = "lm",  
 )

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading

impVars <-   
 varImp(q3All) %$%  
 importance %>%  
 rownames\_to\_column() %>%   
 arrange(Overall %>% desc()) %>%  
 slice(1:15) %>%   
 pull(rowname)  
  
q3Results <-   
 train(  
 glue("sales ~ {paste0(impVars, collapse = \" + \")}") %>% as.formula(),  
 data = regData,  
 method = "lm"  
 )  
  
 coefs <-   
 q3Results %$%  
 finalModel %>%   
 broom::tidy() %>%   
 set\_names(c("Term", "Beta", "Standard Error", "T-Stat", "P-Value"))  
  
myft <- coefs %>% flextable() %>% theme\_vanilla() %>% autofit()  
   
q3Results %$%   
 finalModel %>%   
 broom::glance() %>%   
 t() %>% data.frame() %>%   
 rownames\_to\_column() %>%   
 set\_names(c("Metric", "Value")) %>%   
 flextable() %>% theme\_vanilla() %>% autofit()

| **Metric** | **Value** |
| --- | --- |
| r.squared | 0.776 |
| adj.r.squared | 0.762 |
| sigma | 2661.145 |
| statistic | 54.084 |
| p.value | 0.000 |
| df | 16.000 |
| logLik | -2318.095 |
| AIC | 4670.190 |
| BIC | 4730.055 |
| deviance | 1657116716.947 |
| df.residual | 234.000 |