Predictive Analytics

Trace Smith

MSDS 6371 Final Project

Date: December 10, 2016

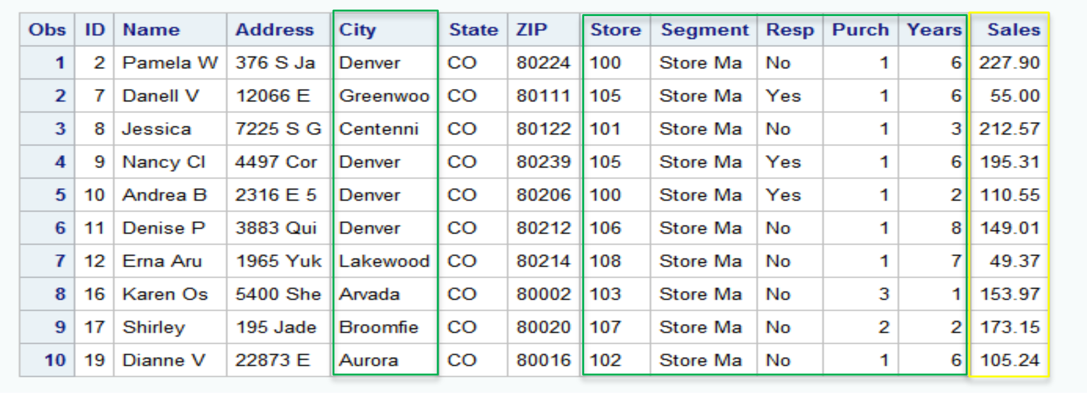
TABLE OF CONTENTS:

I. OVERVIEW

A recently acquired high-end furniture company in Colorado is looking to make changes on how the company is marketing merchandises to their new customers. As a data analytic consulting firm, our organization has been tasked with providing recommendations on whether or not to send a group of 250 new customers a monthly catalog advertising their products. The management team will approve the distribution of the marketing catalogs as long as the expected profit exceeds $30,000 a month generated from purchases through these catalogs. In addition to developing a predictive model as a method of justifying our recommendations related to the expected revenue generated, the company would like to analyze if a significant difference among the customer segments (i.e. reward programs) exists. If so, would it be necessary to develop a marketing scheme for each of these groups independently based on average transactions. The aim of this report is to provide an in-depth statistical analysis of the current customer's purchase history in order predict the total estimated profit for a new collection of customers along with investigating whether or not there is a substantial difference among the various reward programs as an attempt to enhance marketing strategies.

II. INTRODUCTION

The first phase in data analysis is to gather the data, which in most cases can be very time consuming and is not always readily available. Fortunately, the prior owners of the furniture company had maintained a very effective system for gathering and storing data related to their customer's purchase history. The data provided by our client includes the following: name, address, city, state, zip code, store number, the type of customers (i.e. Credit Card Only, Loyalty Card Only, Loyalty and Credit Card, or Store Mailing List), whether or not the customer has purchased an item in previous catalogs, the average number of years as a customer, the average amount of items the consumer buys from the company, the total dollar amount that the consumer spent ordering from the catalogues, and the store location each customer is registered with. Shown in the table below is a snapshot of the data utilized in the analysis consisting of 12 various attributes for 2,375 customers. Note, data points defining the customers such as customer ID, name, address, state, and zip code are not included in the model.



Obs	ID	Name	Address	City	State	ZIP	Store	Segment	Resp	Purch	Years	Sales
1	2	Pamela W	376 S Ja	Denver	CO	80224	100	Store Ma	No	1	6	227.90
2	7	Danell V	12066 E	Greenwoo	CO	80111	105	Store Ma	Yes	1	6	55.00
3	8	Jessica	7225 S G	Centenni	CO	80122	101	Store Ma	No	1	3	212.57
4	9	Nancy Cl	4497 Cor	Denver	CO	80239	105	Store Ma	Yes	1	6	195.31
5	10	Andrea B	2316 E 5	Denver	CO	80206	100	Store Ma	Yes	1	2	110.55
6	11	Denise P	3883 Qui	Denver	CO	80212	106	Store Ma	No	1	8	149.01
7	12	Erna Aru	1965 Yuk	Lakewood	CO	80214	108	Store Ma	No	1	7	49.37
8	16	Karen Os	5400 She	Arvada	CO	80002	103	Store Ma	No	3	1	153.97
9	17	Shirley	195 Jade	Broomfie	CO	80020	107	Store Ma	No	2	2	173.15
10	19	Dianne V	22873 E	Aurora	CO	80016	102	Store Ma	No	1	6	105.24

Customer Data: Attributes highlighted in green are inputs into the predictive model; yellow = response variable

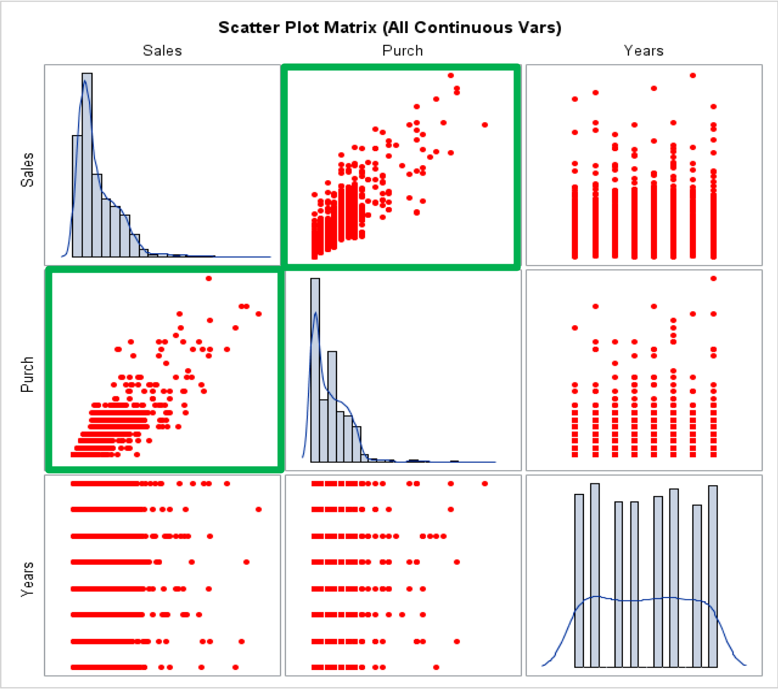
III. QUESTIONS OF INTEREST

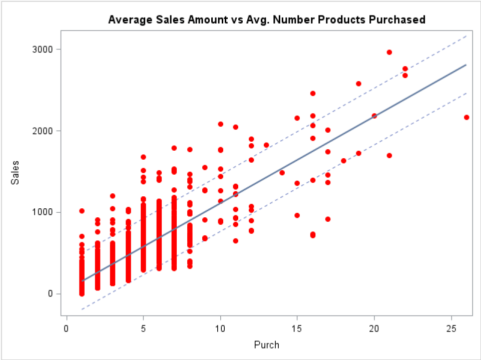
Given there are a total of 250 new customers who signed up to receive a catalog showcasing the company's furniture, our objective is to determine the expected profit generated from these new customers. Depending on the predicted profit we can expect in return as a result of sending a catalog to these new customers, does it exceed the $30,000 threshold set by the management team? To proceed with addressing this question, multiple linear regression is ideal for this scenario as the main objective is to predict a continuous value, average sales revenue, given a series of continuous and categorical predictor variables. Another aspect that is of interest is determining statistically if there is a difference among the four customer segments related to the average sales revenue. Depending on which assumptions are met, parametric test for paired samples with more than two samples like

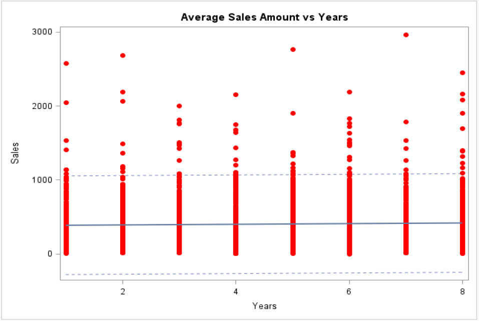
Welch's ANOVA, one-way-ANOVA, or a nonparametric test like Kruskal-Wallis would be appropriate to implement.

IV. EXPLORATORY DATA ANALYSIS

To get a better idea of the dataset at hand, a visual inspection of the two continuous predictor variables versus the response variable was assessed. The scatter matrix below illustrates that out of the two variables (average number of products purchased and number of years as a customer) only the products purchased relates linearly with the average sales. In other words, as the number of products a customer purchases increases, the more likely the customer is to spend more, which is a trend that would generally be expected (unless more customers purchase less products that are more expensive and vice versa -- possible confounder). One other observation made here is that the sales and number products purchased are not normally distributed, rather heavily right skewed. However, obtaining a normal distribution for each explanatory observation is not an assumption of linear regression.





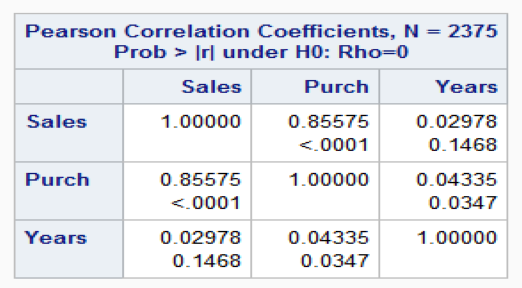


VI. BUILD PREDICTIVE MODEL

The underlying assumption of linear regression is that a continuous predictor variable is related linearly with the response variable. As illustrated in the previous section, a clear relationship between average sales and the average number of products purchased is depicted. In addition to visually reaching this conclusion, one way to quantify the relationship statistically would be to examine the linear correlation coefficient (r). As shown in the SAS output, Pearson correlation is a measurement of how far away the observed data points are to the best fit line. A strong relationship (r=0.855) is observed between sales and number products purchased, confirming our intuition a linear correlation exists.

Null Hypothesis: $\rho_1 = 0$
Alt. Hypothesis: $\rho_0 ! = 0$

$$r = \frac{N\sum xy - \sum x * \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2] * [N\sum y^2 - (\sum y)^2]}}$$



Pearson Correlation Coefficients, N = 2375
Prob > |r| under H0: Rho=0

	Sales	Purch	Years
Sales	1.00000	0.85575 <.0001	0.02978 0.1468
Purch	0.85575 <.0001	1.00000	0.04335 0.0347
Years	0.02978 0.1468	0.04335 0.0347	1.00000

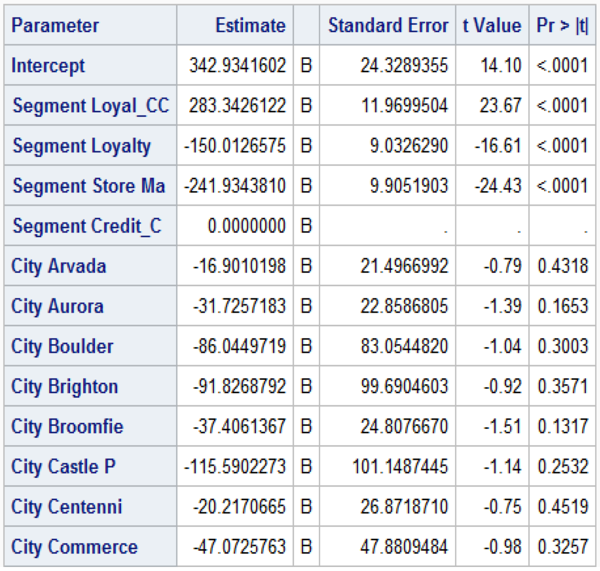
In simple linear regression, the model is used to describe the distribution of values of one continuous variable (i.e. the response) as a function of one explanatory predictor variable. The regression of the continuous response variable on the explanatory variables is a linear relationship between the mean response for a subpopulation of y's conditioned on each value of the explanatory variable (x). On the other hand, multiple regression, which is an extension of simple linear regression, however one of main difference is the difficulty of visualizing the data as a result of the feature space being in higher dimensions (i.e. introduces "hyperplanes" for a feature space greater than three or more dimensions). In this work, multiple linear regression will be the model implemented, which is ideal for handling categorical and continuous variables.

The first step in constructing the regression model is to input the following variables: Customer Segment (4 levels), City (27 levels), Store (10 level), Responded to Catalog (2 levels), average years as customer, and average number of products purchased. The parameter estimate table from SAS shown below is only a snapshot of the entire table, which is then summarized in Table 1. The corresponding p-value for each slope is attributed to a hypothesis test: Null hypothesis $H_o : \beta_n = 0$ and Alternative Hypothesis: H_a: β_n != 0. Subsequently, parameters which are regarded as insignificant (i.e. p-value > 0.05), meaning there is no statistical evidence to suggest a linear relationship with the response variable exists, are removed from the model. Likewise a hypothesis test can be conducted for the intercept ($H_o : \beta_0 = 0$ and Alternative Hypothesis: H_a: β_0 != 0) which test whether or not the y-intercept differs significantly from zero. The parameter estimates are simply the least squared estimates of the intercept and slope that minimizes the sum of square residuals; residuals are defined as the squared distance between the observed y-value and predicted y-value for a given value of x. The parameter estimates for the slope and intercept are conditional on all of the variables identified below in the model; removing or the addition of other predictors can alter the coefficient estimates.

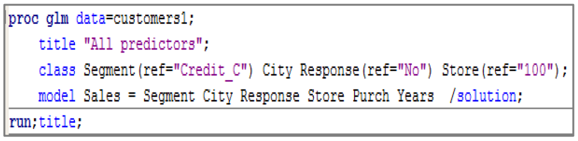
Notation: SegLoyCC = Customer Segment: Loyalty and Credit Card Member (categorical)
RespN = Responded to prior catalog: "No" (categorical)
City = Geographical Location of store (categorical)
Purch = Average number of products purchased (continuous)
Year = Average number of years as customer (continuous)

First Iteration (All variables)

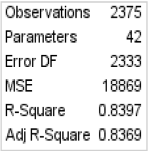
$$\hat{Y} = \beta_0 + \beta_1\text{SegLoyCC} + \beta_2\text{SegLoy} + \beta_3\text{SegStoreM} + \beta_4\text{SegCC} + \beta_5 * \text{CityAr} + \ \ldots..\beta_{31}\text{CityWheatRi} + \beta_{32} * \text{RespY} + \beta_{33}\text{RespN} + \beta_{34}Store100 + \cdots..\ \beta_{43}Store110 + \beta_{44}Purchase + \beta_{45}\text{Year}$$



Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	342.9341602	B	24.3289355	14.10	<.0001
Segment Loyal_CC	283.3426122	B	11.9699504	23.67	<.0001
Segment Loyalty	-150.0126575	B	9.0326290	-16.61	<.0001
Segment Store Ma	-241.9343810	B	9.9051903	-24.43	<.0001
Segment Credit_C	0.0000000	B	.	.	.
City Arvada	-16.9010198	B	21.4966992	-0.79	0.4318
City Aurora	-31.7257183	B	22.8586805	-1.39	0.1653
City Boulder	-86.0449719	B	83.0544820	-1.04	0.3003
City Brighton	-91.8268792	B	99.6904603	-0.92	0.3571
City Broomfie	-37.4061367	B	24.8076670	-1.51	0.1317
City Castle P	-115.5902273	B	101.1487445	-1.14	0.2532
City Centenni	-20.2170665	B	26.8718710	-0.75	0.4519
City Commerce	-47.0725763	B	47.8809484	-0.98	0.3257



```
proc glm data=customers1;
    title "All predictors";
    class Segment(ref="Credit_C") City Response(ref="No") Store(ref="100");
    model Sales = Segment City Response Store Purch Years  /solution;
run;title;
```



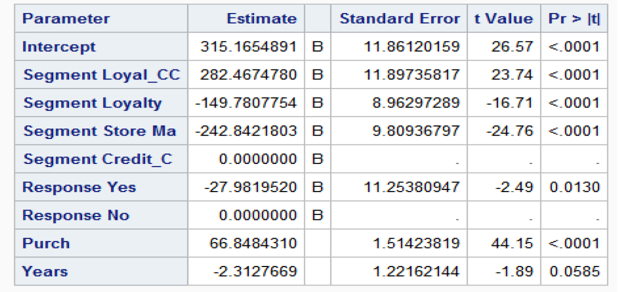
```
Observations    2375
Parameters        42
Error DF        2333
MSE            18869
R-Square      0.8397
Adj R-Square  0.8369
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|	Sign/Non
Intercept	342.93416	24.32894	14.1	<.0001	Significant
Segment Loyal_CC	283.34261	11.96995	23.67	<.0001	Significant
Segment Loyalty	-150.01266	9.03263	-16.61	<.0001	Significant
Segment Store Ma	-241.93438	9.90519	-24.43	<.0001	Significant
City Henderso	-290.66386	138.98567	-2.09	0.0366	Significant
Response Yes	-31.16374	11.37304	-2.74	0.0062	Significant
Purch	66.89780	1.52900	43.75	<.0001	Significant
Years	-2.51717	1.23585	-2.04	0.0418	Significant

Table 1: Summary of Parameters

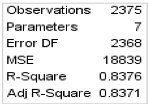
As previously discussed, if evidence suggest that the mean response is unrelated to the explanatory variable (i.e. reject the null hypothesis; p-value < 0.05), then the variable itself is removed. In table 1, zero stores and one out of the 27 levels for cities are significant, thus these categorical variables do not add any value to the model and should be omitted. The question then arises as to why remove any data from the model. According to Occam's Razor and the principle of parsimony, the simplest explanation is usually the right one. Moreover, if we have a model with a large number of attributes, including every explanatory variable will likely lead to overfitting as the model essentially just memorizes the data; the model will not perform well on out-of-sample data. Ideally, those parameters which are significant predictors of the mean response and are not redundant should be considered, leading to a model that better fits the dataset.

On the right displays the final variables included in the model without exploring any interaction terms; the inclusion of interactions and interpretation of the coefficients will be discussed in the following section. Using 'Proc Reg' in SAS requires the variable names to be hard coded and so it should be noted that "Segment Credit_C" refers to the "Credit Card Members" in the customer segments class and is labeled as the reference point. Likewise, "Response_Yes" is the reference for "Responded_to_Last_Catalog" categorical variable.



Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	315.1654891	B	11.86120159	26.57	<.0001
Segment Loyal_CC	282.4674780	B	11.89735817	23.74	<.0001
Segment Loyalty	-149.7807754	B	8.96297289	-16.71	<.0001
Segment Store Ma	-242.8421803	B	9.80936797	-24.76	<.0001
Segment Credit_C	0.0000000	B	.	.	.
Response Yes	-27.9819520	B	11.25380947	-2.49	0.0130
Response No	0.0000000	B	.	.	.
Purch	66.8484310		1.51423819	44.15	<.0001
Years	-2.3127669		1.22162144	-1.89	0.0585

$$R^2_{adj} = 1 - [\frac{(1-R^2)(n-1)}{n-k-1}]$$

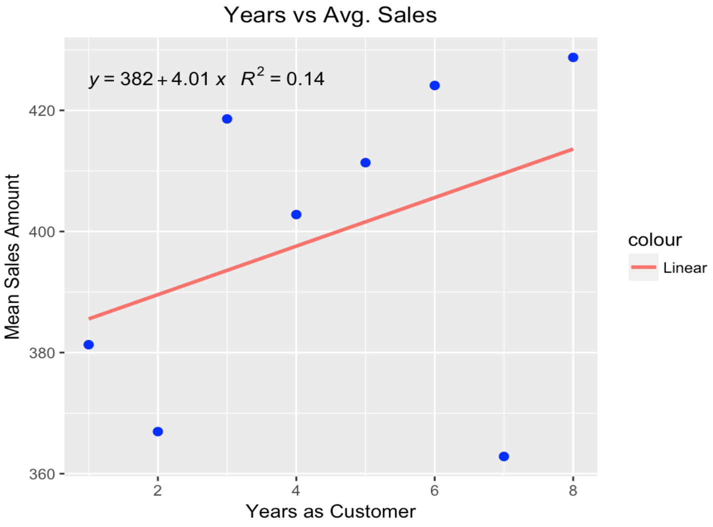


Observations	2375
Parameters	7
Error DF	2368
MSE	18839
R-Square	0.8376
Adj R-Square	0.8371

The corresponding adjusted R^2 increases slightly from 0.8369 (model with all variables) to 0.8371 when the cities and store classes are both removed from the model. The adjusted R^2 is a modification of R^2 which explains the variation in 'Y' that is explained by the linear model. A higher R^2 corresponds to more variance being accounted for as the data points will be nearer to the regression line. R squared is computed by taking one minus the ratio of sum of squares residual to the total sum of squares (i.e. $\frac{\sum((y_i-\hat{y})^2}{\sum(y_i-\bar{y})^2}$). By minimizing the residual sum of squares through the best fit, the ratio decreases and in return, the R^2 increases. As an added comment, the value of R^2 (coefficient of determination) can be misleading and indicates the model fits the data well as more data is added to the model (does not decrease). A solution is the adjusted R^2 as it accounts for the addition of the number of predictors added into the model and if a new term added improves the model, the adjusted R^2 will increase.

From this mode, the mean squared error is 18,839, which is determined by summing the sum of square residuals and dividing by the degrees of freedom. This value is the total variance for the fitted regression line.

Interestingly, in the parameter estimate table above, the model indicates "Number of Years as a Customer" is a significant predictor. Recall in the exploratory data analysis section, visual interpretation of the scatter plot, years vs average sales amount, was suggestive of no linear relationship. The distribution of years as customers was re-examined more closely by removing noise affiliated with the respected variable and only interpreting the mean average sales amount conditional on the year. The scatter plot to the right appears so show a minimal relationship (r = 0.38). To check statistically if the model with years is better than the model without years, we will conduct an extra sum of squares test by comparing the full and reduced models. The results from this test suggest that there is strong evidence (p-value = 0.029) that the model with "Years as Customer" is better.

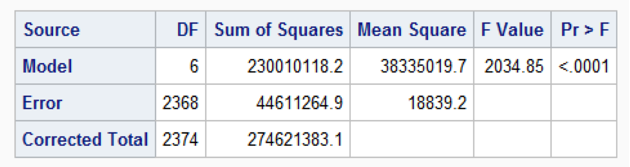


Extra Sum of Squares Test:

Ho: $\beta_{years} = 0$ (if null is rejected, the model with “year as customer” is better)
Ha: $\beta_{years} \; ! = 0$

Full Model:

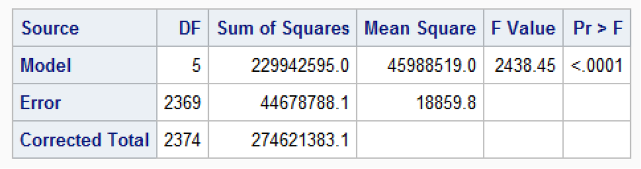
$$\hat{Y} = \beta_0 + \beta_1\text{SegLoyCC} + \beta_2\text{SegLoy} + \beta_3\text{SegStoreM} + \beta_4\text{SegCC} + \\ + \beta_5 * \text{RespY} + \beta_6\text{RespN} + \beta_7 Purchase + \beta_8 Years$$



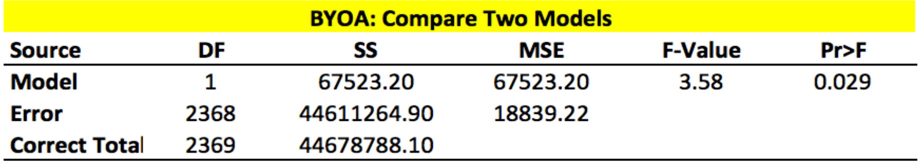
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	230010118.2	38335019.7	2034.85	<.0001
Error	2368	44611264.9	18839.2		
Corrected Total	2374	274621383.1			

Reduced Model:

$$\hat{Y} = \beta_0 + \beta_1\text{SegLoyCC} + \beta_2\text{SegLoy} + \beta_3\text{SegStoreM} + \beta_4\text{SegCC} + \\ + \beta_5 * \text{RespY} + \beta_6\text{RespN} + \beta_7 Purchase$$



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	229942595.0	45988519.0	2438.45	<.0001
Error	2369	44678788.1	18859.8		
Corrected Total	2374	274621383.1			

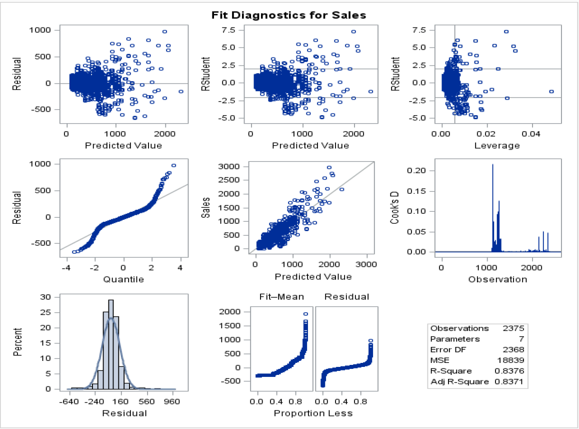


BYOA: Compare Two Models

Source	DF	SS	MSE	F-Value	Pr>F
Model	1	67523.20	67523.20	3.58	0.029
Error	2368	44611264.90	18839.22		
Correct Total	2369	44678788.10			

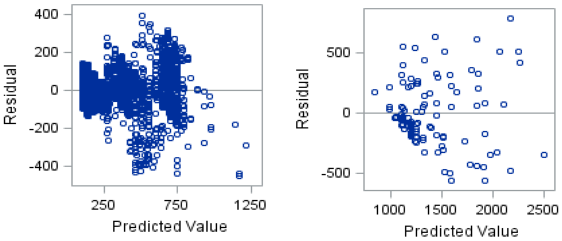
Test the Full model (with years) to the Reduced model (without years)

Model Assessment: To model the data with a linear fit, several key assumptions must be met. First, the variance for a subpopulation of ‘Y’ conditioned on ‘X’ is constant. In the “Predicted Values vs. Residuals” plot, there appears to be heteroskedasticity (non-constant) standard deviation for predicted values. Values greater than 1000 appear to be the most suspect, which could be the result of fewer predicted values larger than this threshold. In this case, one option would be to look at transforming the axes or implementing weighted regression as the variability increases with an increase in the response variable. Another option that could be considered here is that there appears to be two separate clusters (0-1000, and 1000-2000+) and each group could be analyzed separately (bottom right).



Observations	2375
Parameters	7
Error DF	2368
MSE	18839
R-Square	0.8376
Adj R-Square	0.8371

Fit Diagnostic (Linear Model)

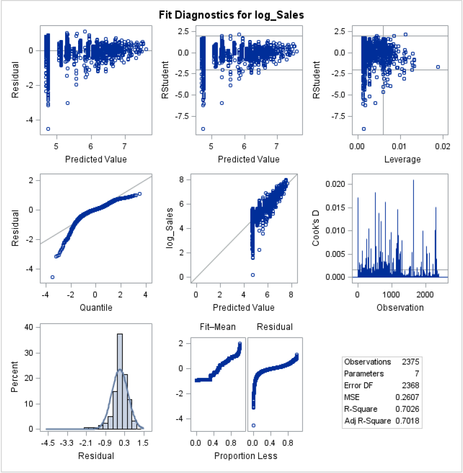


Left: Only response observation > 1000; Right: only Observations > 1000

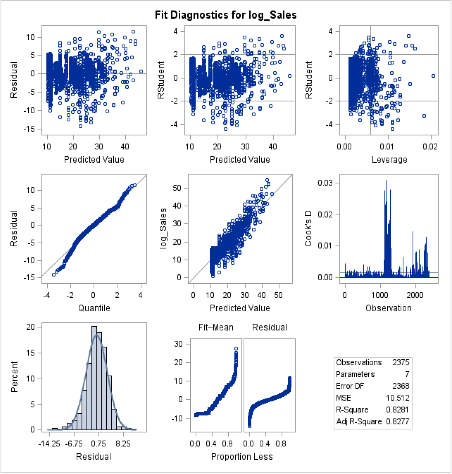
Moreover, the residuals plot also seems to indicate that a linear fit would be adequate, however under and over fitting for predicted value >1500 is anticipated. In reference to the subpopulation of response observations for each value of the explanatory variable being normally distributed, the histogram and QQplot indicate this assumption is met. Note, the cubic shaped QQplot is indicative of long-tailed skewness which could yield misleading results when inference is being made to the normal approximation. Under the assumption that for each subpopulation of ‘y’ conditional on x, the mean responses are normally distributed and hence, for long-tailed distributions the p-values would be overestimated. The data is suggesting that observing a mean response as extreme or more extreme than the observed value is not as rare; this ultimately has an impact on the confidence interval given that each subpopulation is assumed to have the same standard deviation (pooled variance). Finally, independence is the observations is assumed to be valid as well.

It's also import to note that when performing regression that the model is not resistant to outliers or "influential" observations. Even just one or two observations can strongly influence the parameter estimates of the slope and intercept and thus lead to errors in predictions. If possible, any influential point identified through diagnostics such as studentized residuals, leverage, and Cook's D will require further investigation into the observation before deleting the data point. Based on the fit diagnostic plots, there seems to be no influential data points for this model.

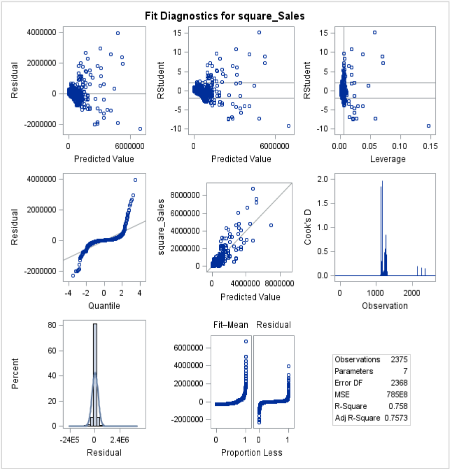
Transformation: In some instances, transformation of the scales is perhaps necessary in order to make using linear regression models more appropriate, nevertheless this can enable some difficulties in interpreting the scales of the transformed axes. As seen in the predicted values vs sales scatter plot, distribution of the response observations is skewed around the regression line. In this scenario, transformation could create additional problems from an interpretation standpoint; thus when making inference to this data set, inference of the mean response is for a skewed distribution. Several transformations to the response and explanatory axes were considered in this work: logarithmic, square root, and squared (X^2). The diagnostic plots shown below provide evidence against a log and X^2 transformation, respectively as a linear fit is not valid (residuals plot). One could make the argument for the square root transformation as the heteroskedasticity does not appear to be as severe, but the distribution of the residuals now become slightly left skewed. The interpretation of the parameter estimates and confidence intervals in this case would also be more difficult, thus we will proceed with the linear-linear multiple regression model but with caution.



Log-Log



$\sqrt{sqrt} - \sqrt{sqrt}$



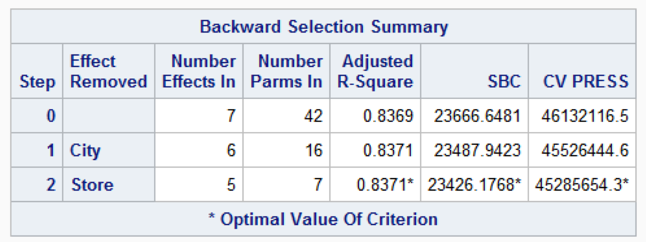
$X^2 – X^2$

VII. VARIABLE SELECTION

To summarize, in the prior section the steps taken to finalize which predictor variables were included in the multiple regression model was discussed in detail. One particular variable that is of concern is the number of years as customer. The linearity between this variable and the response is very poor (r=0.03) except when looking specifically at the mean sales conditional on the year (r increases to 0.38). When customer segments, average purchases and responded to previous catalogs are in the model, the number of years as a customer is borderline significant (p-value = 0.058). The extra sum of squares test confirms keeping years in the model with the adjusted R^2 increasing by nearly one, one thousandths (0.001).

As another layer of validation on the final model before investigating any interactions, variable selection was implemented in SAS. Running 'Proc GLMSelect' with backward selection, which starts with all variables in the model, then at each step removes those variables with the highest p-values and then re-runs the model. The removal of variables will stop once a local minimum is reach (CV Press score). Moreover, to test the validity of the model, a 5-fold cross validation was performed as well. The result of the backward (removal of city and store) and forward selection process is depicted below; the results of the variable selection agrees with the model described in the previous section.

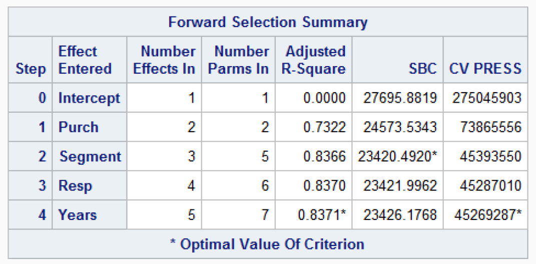
$$\hat{Y} = \beta_0 + \beta_1 SegLoyCC + \beta_2 SegLoy + \beta_3 SegStoreM + \beta_4 SegCC + \beta_5 * RespY + \beta_6 RespN + \beta_7 Purchase + \beta_8 Years$$



Backward Selection Summary

Step	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0		7	42	0.8369	23666.6481	46132116.5
1	City	6	16	0.8371	23487.9423	45526444.6
2	Store	5	7	0.8371*	23426.1768*	45285654.3*

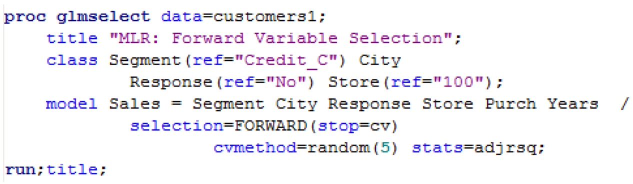
* Optimal Value Of Criterion



Forward Selection Summary

Step	Effect Entered	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0	Intercept	1	1	0.0000	27695.8819	275045903
1	Purch	2	2	0.7322	24573.5343	73865556
2	Segment	3	5	0.8366	23420.4920*	45393550
3	Resp	4	6	0.8370	23421.9962	45287010
4	Years	5	7	0.8371*	23426.1768	45269287*

* Optimal Value Of Criterion



```
proc glmselect data=customers1;
    title "MLR: Forward Variable Selection";
    class Segment(ref="Credit_C") City
            Response(ref="No") Store(ref="100");
    model Sales = Segment City Response Store Purch Years  /
            selection=FORWARD(stop=cv)
                    cvmethod=random(5) stats=adjrsq;
run;title;
```

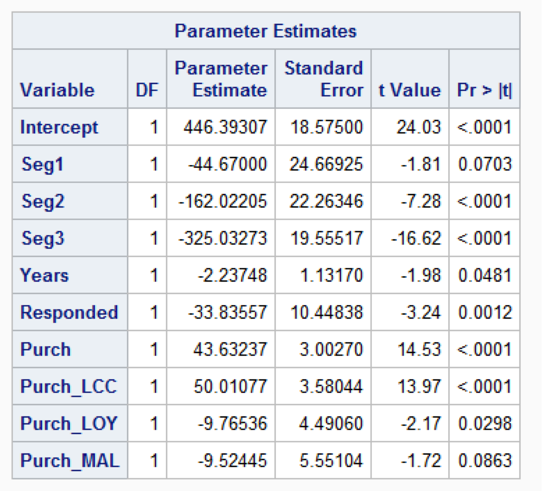
VII. INTERACTIONS

Including an interaction term accounts for the effect of one predictor variable that depends on the value of another variable. The question of interest here is to determine if there is any reason the customer segments should have different slopes when interacting with the average number of products purchased? When an interaction term in included in the model, an adjustment is being made to the slope only. To illustrate this, an interaction terms was added in the model for each level of the customer category with the average number of products purchased. Using 'proc Reg' to look at each interaction separately, the variables were hard-coded with the following notation:

Seg1 = Loyal and Credit Card Members
Seg2 = Loyalty Members
Seg3 = Store Mailing List
LCC = Loyal and Credit Card Members
LOY = Loyalty Members
MAL= Store Mailing List

Purch = Average number of products purchased
Resp_Y = Responded to catalog in past (yes)
Year = Average number of years as customer

Interaction example:
Purch_LCC = Purchase and Loyalty/Cred Card Members

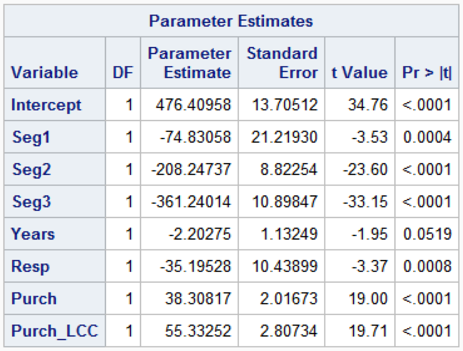


Parameter Estimates

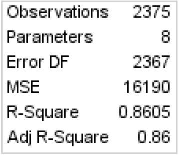
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	446.39307	18.57500	24.03	<.0001
Seg1	1	-44.67000	24.66925	-1.81	0.0703
Seg2	1	-162.02205	22.26346	-7.28	<.0001
Seg3	1	-325.03273	19.55517	-16.62	<.0001
Years	1	-2.23748	1.13170	-1.98	0.0481
Responded	1	-33.83557	10.44838	-3.24	0.0012
Purch	1	43.63237	3.00270	14.53	<.0001
Purch_LCC	1	50.01077	3.58044	13.97	<.0001
Purch_LOY	1	-9.76536	4.49060	-2.17	0.0298
Purch_MAL	1	-9.52445	5.55104	-1.72	0.0863

Model with the initial interaction terms considered in the model.

First Iteration: Shown on the right is the parameter estimate table for model with three interaction terms (each level of the customer segment with the avg. purchases) in addition to the other explanatory variables. It is important to note that "Credit Card Members" is the reference. To answer the question of interest, Store Mailing list is insignificant meaning there is no substantial difference between the two groups. Furthermore, if those categories are grouped together as one, and the model is then re-run, Purch_Loy becomes insignificant (p-value = 0.0958). Hence, when running the final iteration and comparing three levels of the customer group as equal with the the "Loyalty and Credit Card Members*Avg. Purchases", this interaction term remains significant. It can be concluded that there is no significant difference in the amount of products purchased between Loyalty, Mailing List, and Credit Card Members. A significant difference is observed however for the Loyalty and Credit Card Members and as a result, there is evidence to suggest this type of consumer produces $55.33 more on average for every one-unit increase in products purchased than the other three groups.

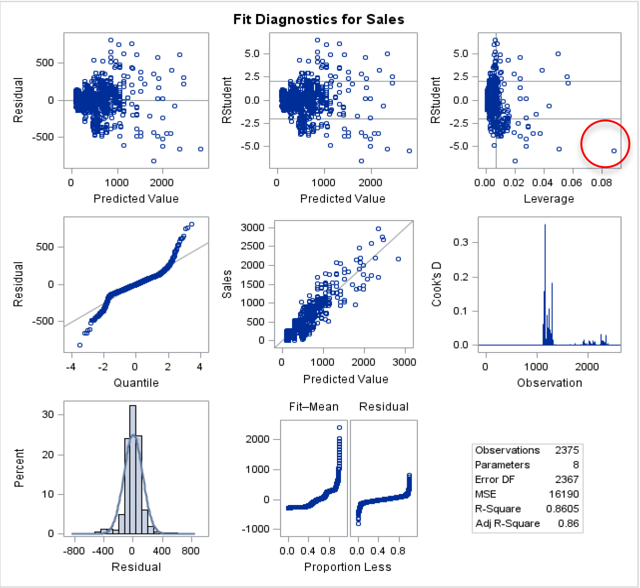


Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	476.40958	13.70512	34.76	<.0001
Seg1	1	-74.83058	21.21930	-3.53	0.0004
Seg2	1	-208.24737	8.82254	-23.60	<.0001
Seg3	1	-361.24014	10.89847	-33.15	<.0001
Years	1	-2.20275	1.13249	-1.95	0.0519
Resp	1	-35.19528	10.43899	-3.37	0.0008
Purch	1	38.30817	2.01673	19.00	<.0001
Purch_LCC	1	55.33252	2.80734	19.71	<.0001



Observations	2375
Parameters	8
Error DF	2367
MSE	16190
R-Square	0.8605
Adj R-Square	0.86

Right: Insignificant interaction terms removed from model (i.e final model)



Model Assessment: Very similar interpretation as discussed earlier when no interaction terms were included in the model. As before, there appears to be some minor heteroskedasticity occurring with the standard deviations, especially for values greater than 1000. The residuals plot is also conclusive that a linear fit would be adequate, however under and over fitting for predicted value >1500 is anticipated. In reference to the subpopulation of response observations for each value of the explanatory variable being normally distributed, the histogram and QQplot indicate this assumption is met; long tailed distribution of residuals. It should be pointed out that an important assumption of equal spread appears to be somewhat violated and when proceeding with making predictions and interpreting confidence intervals, it should be taken with caution.

Investigation into potential outlier through the Fit Diagnostics from the SAS output (shown above) identifies one observation that looks somewhat suspect. The observation highlighted in red has high residuals and relatively high residual. A rule of thumb to be considered "high leverage" is 4/n (0.0016). Leverage essentially measures the distance between the mean of the explanatory variable and each value of the explanatory variables in the model (x value). In this case, the Cook's distance (0.35) does not seem to be large enough to be deemed influential.

VII. INTERPRETING PARAMETER ESTIMATE

Final Mode:

$$\hat{Y} = \beta_0 + \beta_1 \mathrm{SegLoyCC} + \beta_2 \mathrm{SegLoy} + \beta_3 \mathrm{SegStoreM} + \beta_4 Year + \beta_5 * \mathrm{RespY} + \beta_6 Purchase + \beta_7 Purch\_SegLCC$$

$$\hat{Y} = 476.41 - 74.83 * \mathrm{SegLoyCC} - 208.24 * \mathrm{SegLoy} - 361.12 * \mathrm{SegStoreM} - 2.20 Year - 35.19 * \mathrm{RespY} + 38.32 * Purchase + 55.33 * Purch\_SegLCC$$

$\boldsymbol{\beta_o}$= The intercept is 476.41 for the model. For every zero products purchased and zero years as customer, the expected revenue is 476 .41. It does not make sense to interpret the slope at x=0 rather at x=1 since the minimum number of products purchased and years as customer is greater than 1. Simply interpolating the response variable for x=1 will result in a slightly higher value for the intercept and more interpretable.

$\boldsymbol{\beta_1}$= Referred to as “Loyalty and Credit Card Members” (categorical variable) and is an adjustment for intercept. For customers signed up for this respected reward program, the mean response decreases by $74.83 with respect to the reference, Credit Card Members.

$\boldsymbol{\beta_2}$= Referred to as “Loyalty Only Members” (categorical variable) and is an adjustment for the intercept. This category of customers results in an average revenue decreasing by $208.24 with respect to Credit Card members.

$\boldsymbol{\beta_3}$= Referred to as “Store Mailing List” (categorical variable) and is an adjustment for the intercept. The reward program for this type of customers decreases by $361.12 with respect to Credit Card members. It is strongly recommended that this reward program, which generates the least revenue, should be thoroughly investigated why and potentially restructuring how the company markets to this group.

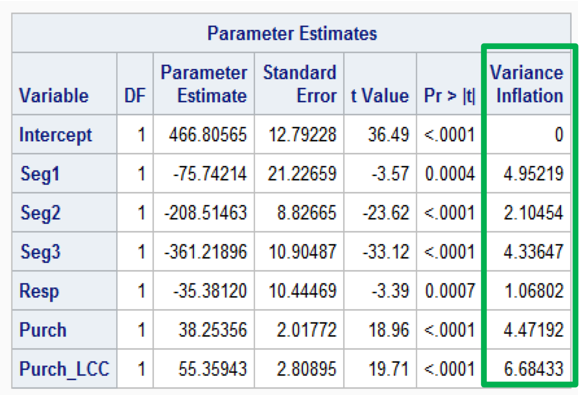
$\boldsymbol{\beta_4}$= The average number of years as a customer is a continuous variable and is an adjustment for the slope. For every one-unit increase in the number of year, the average revenue from each customer will decrease by a factor or -$2.20 (assuming every explanatory variable held constant). The management team should explore more why this is the case, longer customers are spending less money (could be that customers are buying their furniture sets upfront when signing up and thus had a less demand for purchasing additional furniture).

$\boldsymbol{\beta_5}$= Referred to as “Responded to Catalog - Yes” (categorical variable) and is an adjustment for the intercept. The mean response of sales decreases by $35.19 with respect to those who did not respond to the catalog. This is interesting, customers who do not make purchases in the catalogs spends less ($avg. 35.19 less)

$\boldsymbol{\beta_6}$= The average number of products purchased is a continuous variable and is an adjustment for the slope. With everything else held constant, for every one-unit increase in the number of products purchased, the revenue from each customer increases on average by $38.32.

$\boldsymbol{\beta_7}$= The interaction between “Loyalty and Credit Card Members” (i.e. Customer Segment group) and average number of products purchased is an adjustment to the slope. With the other variables held constant, for every one unit increase in the number of products purchased for this customer segment, the response increases by (β_6+β_7) $93.60.

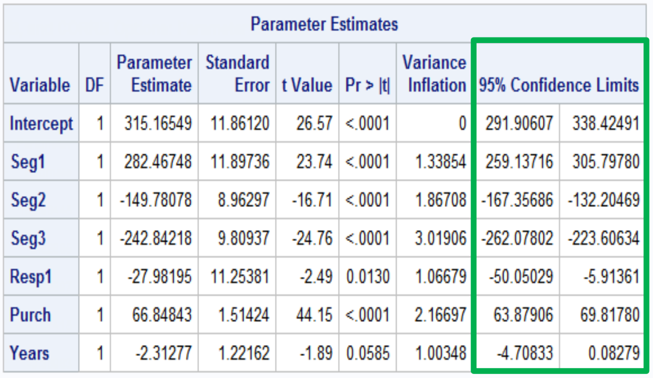
<u>Note</u>: The variance inflation factor is shown below in the parameter estimate table to check if any multicollinearity between the variables exist. A VIF larger than 10 is a rule of thumb for continuous variables being highly correlated and would inflate the parameter estimates leading to misleading coefficients. Also shown below is the 95% confidence intervals for each of the parameter estimates.



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	466.80565	12.79228	36.49	<.0001	0
Seg1	1	-75.74214	21.22659	-3.57	0.0004	4.95219
Seg2	1	-208.51463	8.82665	-23.62	<.0001	2.10454
Seg3	1	-361.21896	10.90487	-33.12	<.0001	4.33647
Resp	1	-35.38120	10.44469	-3.39	0.0007	1.06802
Purch	1	38.25356	2.01772	18.96	<.0001	4.47192
Purch_LCC	1	55.35943	2.80895	19.71	<.0001	6.68433

Variance Inflation Factor



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	315.16549	11.86120	26.57	<.0001	0	291.90607	338.42491
Seg1	1	282.46748	11.89736	23.74	<.0001	1.33854	259.13716	305.79780
Seg2	1	-149.78078	8.96297	-16.71	<.0001	1.86708	-167.35686	-132.20469
Seg3	1	-242.84218	9.80937	-24.76	<.0001	3.01906	-262.07802	-223.60634
Resp1	1	-27.98195	11.25381	-2.49	0.0130	1.06679	-50.05029	-5.91361
Purch	1	66.84843	1.51424	44.15	<.0001	2.16697	63.87906	69.81780
Years	1	-2.31277	1.22162	-1.89	0.0585	1.00348	-4.70833	0.08279

95% Confidence Interval

IX.

VALIDATE MODEL

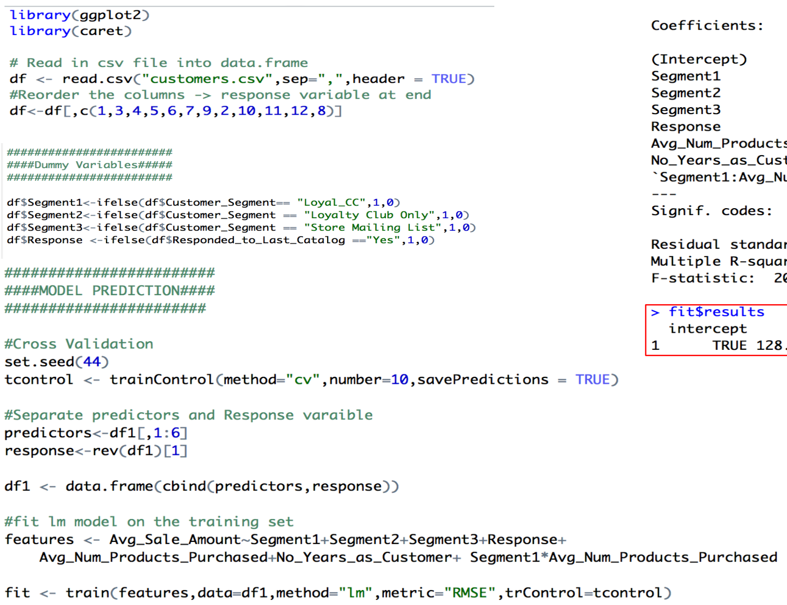
In any machine learning application, the model should be tested on an out of sample data set to monitor the performance of the model. There are various methods of randomly assigning the data samples into training and testing subsets such as stratified shuffle split. An alternative approach is to implement cross-validation, and for analyzing the performance of this model, 10-fold cross validation was considered. For example, the customer dataset consists of 2375 observations and thus each fold will consist of 238 observations. The initial iteration entails the first fold containing the testing data (238 observations) and the remaining folds (2-10) containing 2142 observations. In second iteration, the testing set is now fold #2 while fold #1 and fold #3-10 are the training set. We train K different models, with each time leaving out a single subset for measuring the cross-validation error. The final cross-validation error is calculated by taking the mean or median of the K models. The advantage of cross-validation provides a more accurate estimate of the out-of-sample accuracy and is more efficient as every observation is used for both training and testing as oppose to the train/test/split methods.

The input predictor variables for the model are as follows: customer segments, responded to previous catalog, average number of products purchased, average number of years as customer, and interaction term (average number of products purchased paired with loyalty and credit card members (i.e. continuous variable paired with categorical). The performance metric utilized to gauge the how well or poor the model performs on out of sample data is the root mean squared error. The RMSE is 128.62 and the amount of variation in the average sales explained by the explanatory variables is 85.1%.

Notation: Segment1 = Loyalty and Credit Card
Segment2= Loyalty Club Only
Segment3= Store Mailing List

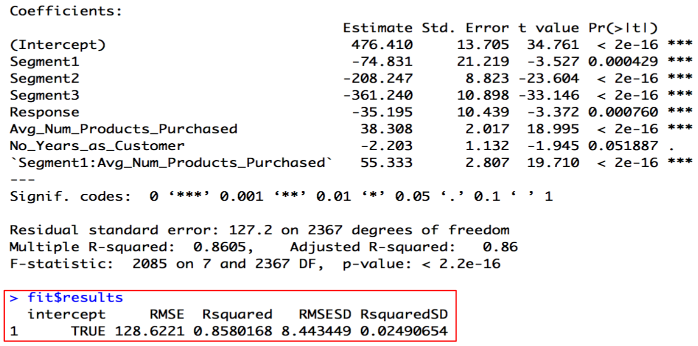
Responded= Responded to catalog in past ("Yes")
Variable of prediction= Average Sales (continuous)

R Code:



```
library(ggplot2)
library(caret)

# Read in csv file into data.frame
df <- read.csv("customers.csv",sep=",",header = TRUE)
#Reorder the columns -> response variable at end
df<-df[,c(1,3,4,5,6,7,9,2,10,11,12,8)]


########################
####Dummy Variables#####
########################

df$Segment1<-ifelse(df$Customer_Segment== "Loyal_CC",1,0)
df$Segment2<-ifelse(df$Customer_Segment == "Loyalty Club Only",1,0)
df$Segment3<-ifelse(df$Customer_Segment == "Store Mailing List",1,0)
df$Response <-ifelse(df$Responded_to_Last_Catalog =="Yes",1,0)

########################
####MODEL PREDICTION####
#######################

#Cross Validation
set.seed(44)
tcontrol <- trainControl(method="cv",number=10,savePredictions = TRUE)

#Separate predictors and Response varaible
predictors<-df1[,1:6]
response<-rev(df1)[1]

df1 <- data.frame(cbind(predictors,response))

#fit lm model on the training set
features <- Avg_Sale_Amount~Segment1+Segment2+Segment3+Response+
    Avg_Num_Products_Purchased+No_Years_as_Customer+ Segment1*Avg_Num_Products_Purchased

fit <- train(features,data=df1,method="lm",metric="RMSE",trControl=tcontrol)
```

Output:



```
Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                             476.410     13.705  34.761  < 2e-16 ***
Segment1                                -74.831     21.219  -3.527 0.000429 ***
Segment2                               -208.247      8.823 -23.604  < 2e-16 ***
Segment3                               -361.240     10.898 -33.146  < 2e-16 ***
Response                                -35.195     10.439  -3.372 0.000760 ***
Avg_Num_Products_Purchased               38.308      2.017  18.995  < 2e-16 ***
No_Years_as_Customer                     -2.203      1.132  -1.945 0.051887 .  
`Segment1:Avg_Num_Products_Purchased`    55.333      2.807  19.710  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 127.2 on 2367 degrees of freedom
Multiple R-squared:  0.8605,    Adjusted R-squared:   0.86 
F-statistic:  2085 on 7 and 2367 DF,  p-value: < 2.2e-16

> fit$results
  intercept     RMSE  Rsquared   RMSESD RsquaredSD
1      TRUE 128.6221 0.8580168 8.443449 0.02490654
```

X. PREDICT TOTAL PROFIT

Once the final model has been selected and examined which variables are significant when predicting the mean response (including interaction terms), predictions can then be made. The regression equation for each customer is shown below with the references being "Credit Card Members" and "Responded Yes" to catalog (i.e. Yes=1). Note: The new customers all responded to the catalog hence the input for the explanatory variable "RespY" is 1 given the reference is set equal to "No". In this case, the total number regression equations are reduced to 4 as opposed to 8.

Store Mailing List

$\hat{Y} = \beta_0 + \beta_1\text{SegLoyCC} + \beta_2\text{SegLoy} + \beta_3\text{SegStoreM} + \beta_4\text{Year} + \beta_5\text{RespY} + \beta_6 Purchase + \beta_7\text{Purch\_ SegLoyCC}$

$\hat{Y} = 476.40 + \beta_1(0) + \beta_2(0) - 361.24(1) - 2.20(\text{Year}) - 35.19(1) + 38.38(Purchase) + \beta_6\text{Purch}(0)$

Response (Yes): $\hat{Y} = 79.97 - 2.20(Year) + 38.25(Purch)$

Loyalty Members

$\hat{Y} = \beta_0 + \beta_1\text{SegLoyCC} + \beta_2\text{SegLoy} + \beta_3\text{SegStoreM} + \beta_4\text{Year} + \beta_5\text{RespY} + \beta_5 Purchase + \beta_6\text{Purch\_ SegLoyCC}$

$\hat{Y} = 476.40 + \beta_1(0) - 208.24(1) - \beta_3(0) - 2.20(\text{Year}) - 35.19(1) + 38.38(Purchase) + \beta_6\text{Purch}(0)$

Response (Yes): $\hat{Y} = 268.16 - 2.20(Year) + 38.25(Purch)$

Loyalty and Credit Card: (interaction term in red)

$\hat{Y} = \beta_0 + \beta_1\text{SegLoyCC} + \beta_2\text{SegLoy} + \beta_3\text{SegStoreM} + \beta_4\text{Year} + \beta_5\text{RespY} + \beta_6 Purchase + \beta_7\text{Purch\_ SegLoyCC}$

$\hat{Y} = 476.40 - 74.83(1) + \beta_2(0) - \beta_3(0) - 2.20(\text{Year}) - 35.19(1) + 38.38(Purchase) + 55.33(\text{Purch})$

Response (Yes): $\hat{Y} = 366.38 - 2.20(Year) + 93.58(Purchase)$

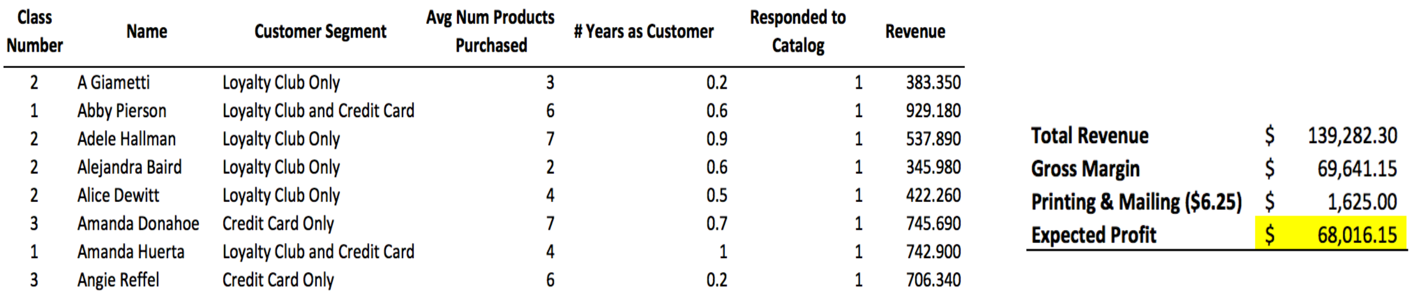
Credit Card Only

$\hat{Y} = \beta_0 + \beta_1\text{SegLoyCC} + \beta_2\text{SegLoy} + \beta_3\text{SegStoreM} + \beta_4\text{Year} + \beta_5\text{RespY} + \beta_5 Purchase + \beta_6\text{Purch\_ SegLoyCC}$

$\hat{Y} = 476.40 + \beta_1(0) - \beta_2(0) - \beta_3(0) - 2.20(\text{Year}) - \beta_5(0) + 38.38(Purchase) + \beta_6\text{Purch}(0)$

Response (Yes): $\hat{Y} = 476.40 - 2.20(Year) + 38.25(Purch)$;

Compute Total Expected Profit: To compute the expected profit of sending a catalog to the new 250 customers is to utilize the linear regression equation defined above and calculate the predicted revenue for each customer. Next, multiplying the predicted revenue by "Score_Yes" (probability of signing up for the catalog) yields the expected revenue for each customer. Assuming the gross margin is 50%, multiply the revenue for each customer by 0.50 and then deduct mailing and shipping expenses (assuming $6.50 per customer) to get a final expected profit for each customer. As shown below, the profit expected to be generated by these new customers is approximately $68,016.25 per month. Therefore, we would strongly advice our client to send the new customers a catalog.

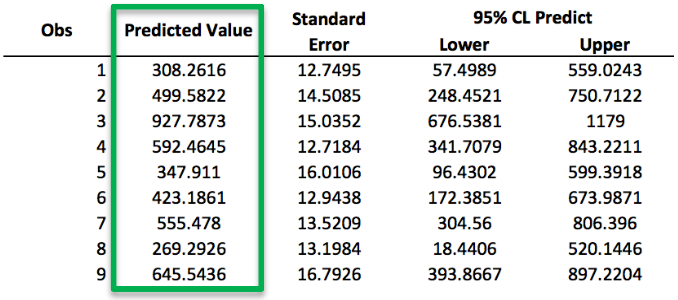


Class Number	Name	Customer Segment	Avg Num Products Purchased	# Years as Customer	Responded to Catalog	Revenue
2	A Giametti	Loyalty Club Only	3	0.2	1	383.350
1	Abby Pierson	Loyalty Club and Credit Card	6	0.6	1	929.180
2	Adele Hallman	Loyalty Club Only	7	0.9	1	537.890
2	Alejandra Baird	Loyalty Club Only	2	0.6	1	345.980
2	Alice Dewitt	Loyalty Club Only	4	0.5	1	422.260
3	Amanda Donahoe	Credit Card Only	7	0.7	1	745.690
1	Amanda Huerta	Loyalty Club and Credit Card	4	1	1	742.900
3	Angie Reffel	Credit Card Only	6	0.2	1	706.340

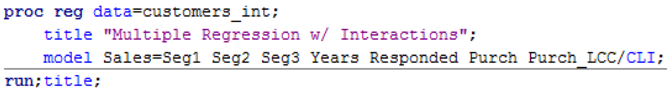
Total Revenue	$	139,282.30
Gross Margin	$	69,641.15
Printing & Mailing ($6.25)	$	1,625.00
Expected Profit	$	68,016.15

Excel calculations for Total Predicted Revenue and Final Expected profit; 7 observations out of 250 showing.

a.) Prediction Intervals: The predicted response from each consumer individually can be quantified in prediction intervals which accounts for any estimation errors and/or random sampling errors. With 95% certainty, the predicted revenue produced from each customer can be computed; these intervals are shown below for the first several customers as an illustration. To compare calculations in SAS, the predicted revenue computed utilizing the regression equations provided in section IX (Predict Total Profit) in Excel yielded marginal differences with SAS. It should also be noted that zero out of 250 predictions for average revenue per customer were outside of the 95% confidence interval. A screenshot of the first 9 predictions is shown below.



Obs	Predicted Value	Standard Error	95% CL Predict Lower	Upper
1	308.2616	12.7495	57.4989	559.0243
2	499.5822	14.5085	248.4521	750.7122
3	927.7873	15.0352	676.5381	1179
4	592.4645	12.7184	341.7079	843.2211
5	347.911	16.0106	96.4302	599.3918
6	423.1861	12.9438	172.3851	673.9871
7	555.478	13.5209	304.56	806.396
8	269.2926	13.1984	18.4406	520.1446
9	645.5436	16.7926	393.8667	897.2204



```
proc reg data=customers_int;
    title "Multiple Regression w/ Interactions";
    model Sales=Seg1 Seg2 Seg3 Years Responded Purch Purch_LCC/CLI;
run;title;
```

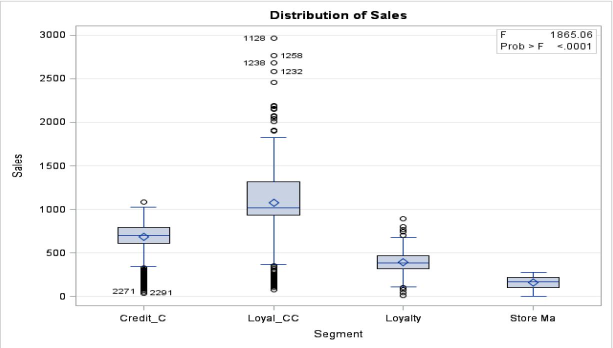
Prediction Interval: SAS output in excel: (Showing First 9 observations; 250 total)

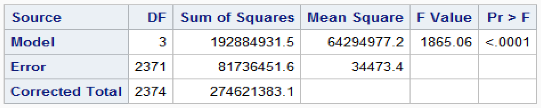
XI. ANALYSIS OF COVARIANCE (ANOVA) TEST

As a way of further optimizing how our client brands it's merchandize to certain types of customers, a test of equality of mean sales from multiple classes utilizing analysis of covariance. If the average sales generated from one group is different than the others, then perhaps the company should target these customers differently as opposed to treating the revenue generated from each class of customers equally. Analysis of variance test would be applicable in this scenario with the class variable being "Customer Segment" (categorical) and the response variable being the average sales amount (continuous). First, a lack of fit test was conducted in order to determine whether or not each group should have a separate mean or could it be represented by a linear best fit line. Performing an extra sum of squares test will provide further insight into whether linear regression or the separate means model yields the better fit.

Ho: Linear regression model is a good fit: Ha: The separate means model fits better

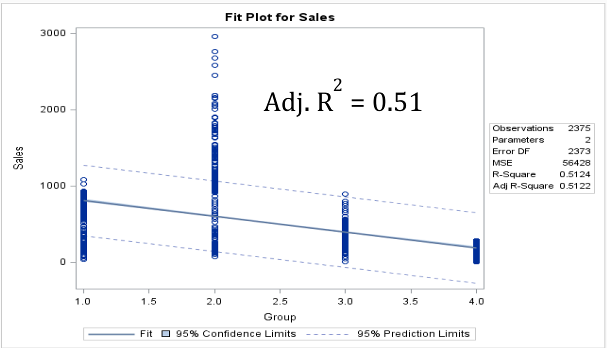
Separate Means Model

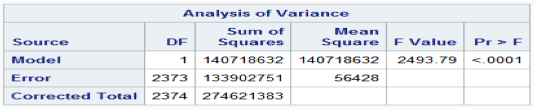




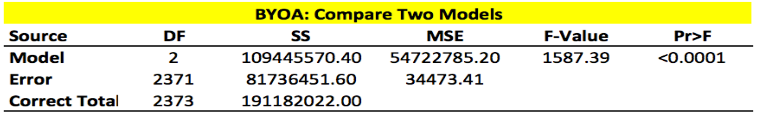
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	192884931.5	64294977.2	1865.06	<.0001
Error	2371	81736451.6	34473.4		
Corrected Total	2374	274621383.1			

Linear Regression





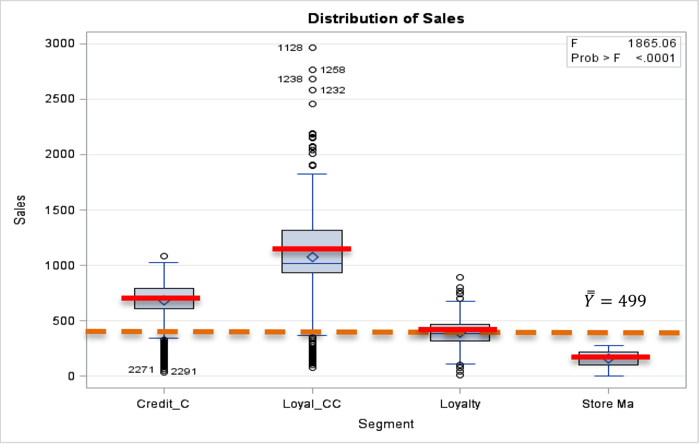
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	140718632	140718632	2493.79	<.0001
Error	2373	133902751	56428		
Corrected Total	2374	274621383			



BYOA: Compare Two Models					
Source	DF	SS	MSE	F-Value	Pr>F
Model	2	109445570.40	54722785.20	1587.39	<0.0001
Error	2371	81736451.60	34473.41		
Correct Total	2373	191182022.00			

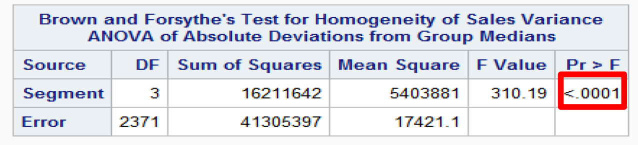
Based on the results from BYOA ("Build Your Own ANOVA table) to compare the two models shown above, there is strong evidence that the linear regression model has a "lack of fit" with respect to the separate means model (p-value < 0.001). With only two degrees of freedom (intercept and slope) 109,445,570 units of variance is explained by the linear regression model, which is essentially says the within variation is larger than the between variation for each group (F-statistic 1587.39). For instance, given the same distribution and the null is true, if the observations are resampled 1,000 times, then less than 0.01% of the time the between variation would be such that linear regression would be a better fit as opposed to the separate means model. Consequently, the extra sum of squares test confirms that fitting the mean response with a linear trend has a "lack of fit".

Along with the extra sum of squares test, visual inspection of the boxplots (right) is indicative of the separate means model being a better fit as opposed to the regression model. Now, let's confirm the intuition that at least one of the mean sales is significantly different than at least one other group. To statistically conclude there is a difference, a one-way analysis of variance test is conducted to compare the full model (separate means) with the reduced model (equal means). If the additional three degrees of freedom used to explain the between group variation is large enough, the resulting F-statistic will also be relatively large; if the p-value is below the level of significance (α=0.05) and the null is true $(\mu_1, \mu_2 \ldots . \mu_n)$, then the chance of observing a distribution such that the means of each customer segment is not statistically different (represented by one mean; i.e. equal means model) is very rare.



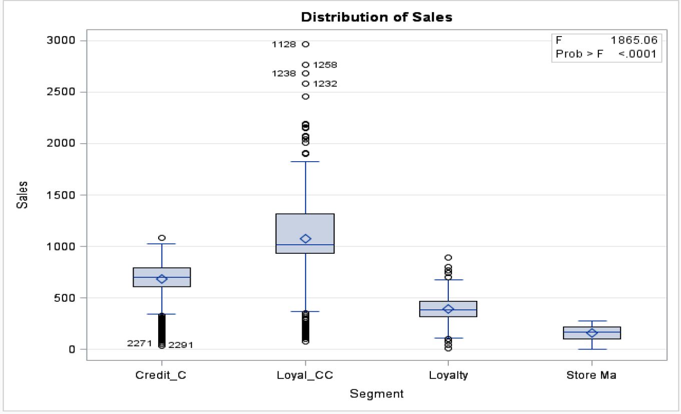
Boxplot: Comparing within variation with between variation

The assumptions of one-way ANOVA must be upheld in order for the parametric model to be valid, thus a brief examination is required. First, visual inspection of the box plots indicates that the data is not coming from the same distribution. As secondary evidence, Brown and Forsythe's Test of homogeneity of variances also confirms unequal spread (p-value = 0.001). Therefore, proceed with Welch's ANOVA, which does not pool the standard deviations. Furthermore, there is evidence against a normal distribution for Mailing List and Loyalty CC members shown in the histograms below. However, note for skewed distributions, normality does not appear to be a major concern least as long as the sample size are roughly equal. In the case of this dataset, the sample size is large enough for the Central Limit Theorem to hold. The CLT theorem states that if 'n' is large then the mean of the sampling distribution will be normal (or nearly) The ratio of the smallest to largest value less than 10, which could indicate a log transformation wouldn't be appropriate.

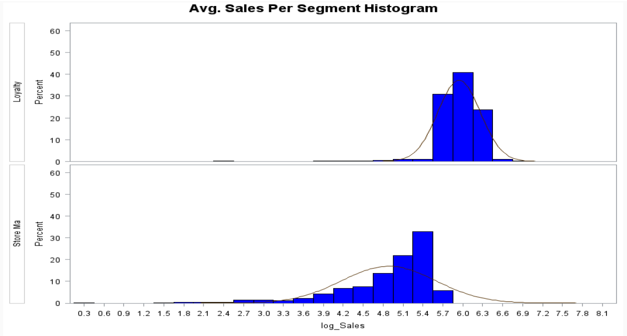


Brown and Forsythe's Test for Homogeneity of Sales Variance
ANOVA of Absolute Deviations from Group Medians

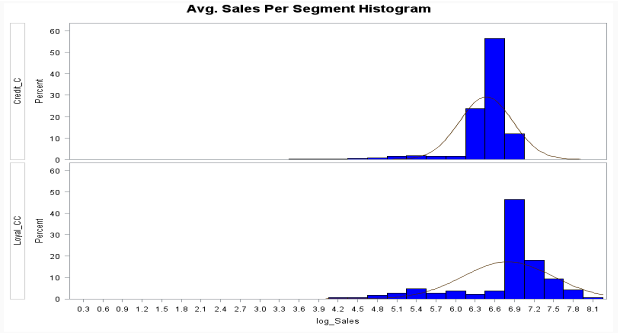
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Segment	3	16211642	5403881	310.19	<.0001
Error	2371	41305397	17421.1		



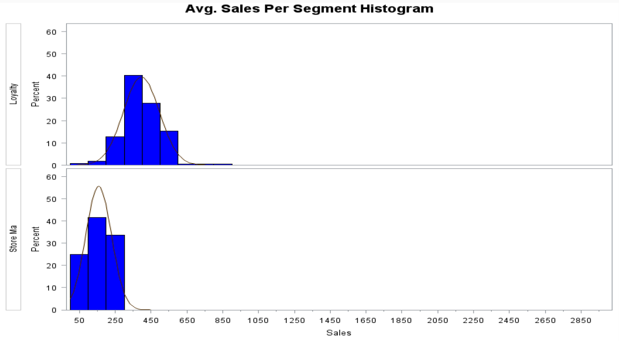
Test of homogeneity of variances



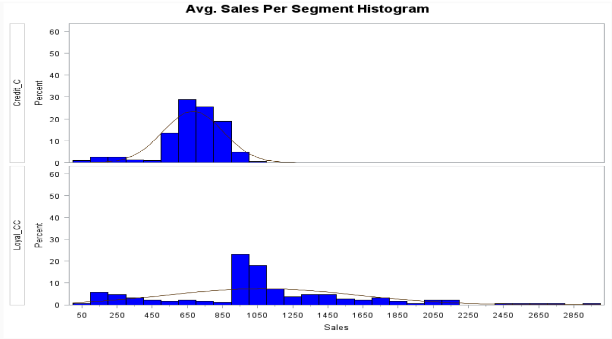
Linear-Linear



Linear-Linear



Log-Log Transformation



Log-Log Transformation

SIX STEP HYPOTHESIS TEST:

1.) Problem Statement: Test if there is a significant difference in the distribution of the response (average sale revenue) between the different levels Customer Segments

2.) Null and Alternative Hypothesis:

H_o: $\mu_{\text{credit card}}$ μ_{Loyalty} $\mu_{\text{LoyaltyCreditCard}}$ $\mu_{\text{Store Mailing List}}$

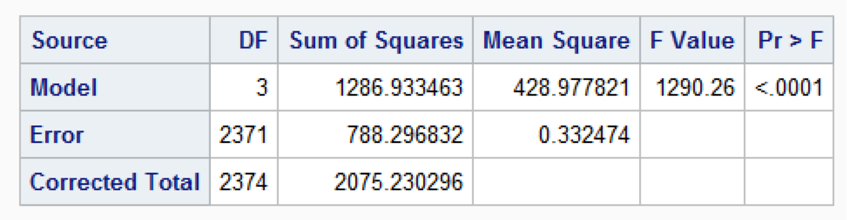
H_a: At least one of customer segments has a different mean

3.) F-statistic: 1290.26

4.) P-value: <0.001

5.) Decision: Reject Null

6.) Conclusion: There is significant evidence to suggest that at least one customer segment group has a mean response that is statistically different (p-value < 0.001). This experiment is observational, therefore inference to the mean response (avg. sales) is only an association, not causation.



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1286.933463	428.977821	1290.26	<.0001
Error	2371	788.296832	0.332474		
Corrected Total	2374	2075.230296			

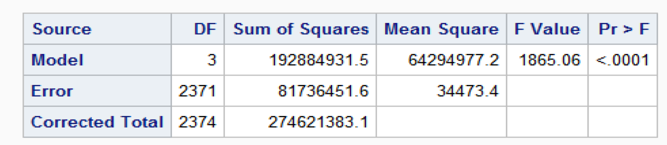
Extra Sum of Squares Test**:** In this last section, an extra sum of squares test will be performed to test whether the mean response of average sales amount for Credit Card, Loyalty, and Store Mailing List Members differs from Loyalty Credit Card Members. If so, it could be concluded that Credit Card, Loyalty, and Store Mailing List Members would all share the same mean and there would be no reason to suggest they are different.

$H_o: \mu_{\text{credit card}} = \mu_{\text{Loyalty}} = \mu_{\text{Store Mailing List}}$

H_a: At least two are different (credit card, loyalty, mailing list)

Separate Means: (Full)

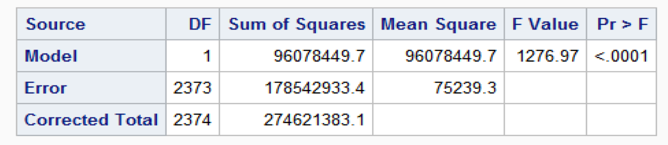
$\mu, \mu_{\text{Loyalty}}, \mu_{\text{LoyaltyCreditCard}}$



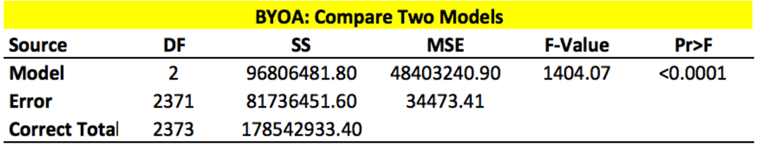
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	192884931.5	64294977.2	1865.06	<.0001
Error	2371	81736451.6	34473.4		
Corrected Total	2374	274621383.1			

Equal Means: (Reduced)

$\mu, \mu, \mu, \mu_{\text{LoyaltyCreditCard}}$



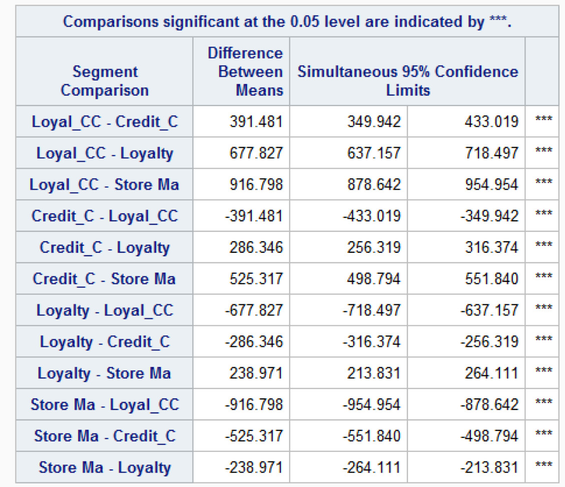
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	96078449.7	96078449.7	1276.97	<.0001
Error	2373	178542933.4	75239.3		
Corrected Total	2374	274621383.1			



BYOA: Compare Two Models

Source	DF	SS	MSE	F-Value	Pr>F
Model	2	96806481.80	48403240.90	1404.07	<0.0001
Error	2371	81736451.60	34473.41		
Correct Total	2373	178542933.40			

Conclusion: There is sufficient evidence at the significance level of 0.05 to suggest that the mean average sales amount for Credit Card Users, Loyalty Members, and Store Mailing List customers is different. Thus, an additional pairwise comparisons would be necessary to determine which Customer Segment differs significantly. The confidence intervals computed (right) were compared to the CL calculated in SAS with a Bonferroni adjustment. The table on the right shows all 12 possible comparisons and each comparison indicates there is a significant difference among the Customer Segments (adjusted alpha = 0.005). Based on this data and with no additional explanatory variables, the final recommendation to our client would be to market the various Customer Segments differently as opposed to treating each group as the same.



Comparisons significant at the 0.05 level are indicated by ***.

Segment Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Loyal_CC - Credit_C	391.481	349.942	433.019	***
Loyal_CC - Loyalty	677.827	637.157	718.497	***
Loyal_CC - Store Ma	916.798	878.642	954.954	***
Credit_C - Loyal_CC	-391.481	-433.019	-349.942	***
Credit_C - Loyalty	286.346	256.319	316.374	***
Credit_C - Store Ma	525.317	498.794	551.840	***
Loyalty - Loyal_CC	-677.827	-718.497	-637.157	***
Loyalty - Credit_C	-286.346	-316.374	-256.319	***
Loyalty - Store Ma	238.971	213.831	264.111	***
Store Ma - Loyal_CC	-916.798	-954.954	-878.642	***
Store Ma - Credit_C	-525.317	-551.840	-498.794	***
Store Ma - Loyalty	-238.971	-264.111	-213.831	***

XII. DATA SOURCE:

https://www.udacity.com/course/predictive-analytics-for-business--nd008

XIII. APPENDIX

A. SAS CODE:

```
*********************
*****Trace Smith*****
*****  Project  *****
*********************;
data customers;
infile
"\\Client\C$\Users\tracesmith\Desktop\Trace\SMU\Stats\Unit13\project\customers.csv"
dlm="," firstobs=2;
input Name $ Segment $ ID Address $ City $ State $ ZIP $ Sales Store $ Resp $ Purch
Years;
run;
proc print data=customers(obs=10);
      title "Raw Customer Data";
run;title;
data customers1;
      title "Reorder Variables";
      retain ID Name Address City State ZIP Store Segment Resp Purch Years Sales;
      set customers;
run;title;
proc print label data=customers1(obs=10);
      title "Raw Customer Data (re-ordered)";
run;title;
**************************
*****  Explore Data  *****
**************************;
proc print data=customers1(obs=10);
      title "Customer Data (reordered)";
run;title;
proc means data=customers1 mean median range var;
      var Sales;
run;
proc sgscatter data=customers1;
      title "Scatter Plot Matrix (All Continuous Vars)";
      matrix Sales Purch Years
      / diagonal=(histogram kernel)
                  markerattrs=(symbol=circlefilled size=8 color="red");
run;
proc sgscatter data=customers1;
      title "Average Sales Amount vs Avg. Number Products Purchased";
      plot (Sales)*(Purch)
      / reg=(degree=1 cli) markerattrs=(symbol=circlefilled color="Red");
run;title;
proc sgscatter data=customers1;
      title "Average Sales Amount vs Years";
      plot (Sales)*(Years)
      / reg=(degree=1 cli) markerattrs=(symbol=circlefilled color="Red");
run;title;
proc corr data=customers1 out=P;
      title "Pearson Correlation ";
      var Sales Purch Years;
run;
**************************
```

```
*****   Lack of Fit   ****
**************************;
data customers_LOF; set customers1;
      if Segment="Credit_C" then Group=1;
      if Segment="Loyal_CC" then Group=2;
      if Segment="Loyalty" then Group=3;
      if Segment="Store Ma" then Group=4;
run;
proc glm data=customers_LOF;
      title "Separate Means Model";
      class Purch;
      model Sales=Purch;
      means Purch/HOVTEST = Welch;
run;
proc reg data=customers_LOF;
      title "Linear Regression Model";
      model Sales=Purch;
run;title;
**************************
*****   Fit Model   ******
**************************;
/*ODS TAGSETS.EXCELXP*/
/*file='"\\Client\C$\Users\tracesmith\Desktop\Trace\SMU\Stats\Unit13\project\regression.x
ls';*/
proc glm data=customers1 plots=all;
      title "MLR: Full Model";
      class Segment(ref="Credit_C") City Response(ref="No") Store(ref="100");
      model Sales = Segment City Resp Store Purch Years  /solution;
run;title;
/*ods tagsets.excelxp close;*/
proc glm data=customers1 plots=all;
      title "MLR: Remove City & Store";
      class Segment(ref="Credit_C") Response(ref="No");
      model Sales = Segment Resp Purch Years  /solution;
run;title;
proc glm data=customers1 plots=all;
      title "MLR: Remove Years";
      class Segment(ref="Credit_C") Response(ref="No");
      model Sales = Segment Response Purch Years /solution;
run;title;
proc glm data=customers1 plots=all;
      title "MLR: Remove Years";
      class Segment(ref="Credit_C") Resp(ref="No");
      model Sales = Segment Resp Purch/solution;
run;title;
**************************
****   Dummy Vars    *****
**************************;
data customers2;set customers1;
      /*Credit Card = Reference*/
      if Segment="Loyal_CC" then Seg1=1;
      else Seg1=0;
      if Segment="Loyalty" then Seg2=1;
      else Seg2=0;
      if Segment="Store Ma" then Seg3=1;
      else Seg3=0;
      if Resp = "Yes" then Responded=1;
      else Responded=0;
run;
proc print data=customers2(obs=10);
      title "Dataset w/ Dummy Variables";
run;title;
```

```
proc reg data=customers2;
      title "Multiple Regression w/ LOF";
      model Sales=Seg1 Seg2 Seg3 Responded Purch Years /CLI;
run;title;
**************************
*****  Interaction   *****
**************************;
data customers_int;set customers2;
      title "Interaction Terms";
      Purch_LCC = Purch*Seg1;
      Purch_LOY = Purch*Seg2;
      Purch_MAL = Purch*Seg3;
      Year_LCC = Years*Seg1;
      Year_LOY = Years*Seg2;
      Year_MAL = Years*Seg3;
      RespY_Purch = Purch*Responded;
      RespY_Year = Responded*Years;
run;title;
proc reg data=customers_int;
      title "Multiple Regression w/ Interactions";
      model Sales=Seg1 Seg2 Seg3 Years Responded Purch Purch_LCC/CLI;
run;title;
**************************
*****   Transform   ******
**************************;
data log_customers;set customers1;
      log_Sales = log(Sales);
      log_Purch = log(Purch);
      log_Years = log(Years);
run;
proc glm data=log_customers plots=all;
      title "MLR: Final Model";
      title2 "Log Transformation:";
      class Segment(ref="Credit_C") Resp;
      model square_Sales = Segment Resp log_Purch log_Years /solution;
run;title2;
proc glm data=log_customers plots=all;
      title "MLR: Final Model";
      title2 "Log Transformation: Log-Linear";
      class Segment(ref="Credit_C") Resp;
      model log_Sales = Segment Resp Purch Years /solution;
run;title2;
proc glm data=log_customers plots=all;
      title "MLR: Final Model";
      title2 "Log Transformation: Linear-Log";
      class Segment(ref="Credit_C") Resp;
      model Sales = Segment Resp log_Purch log_Years /solution;
run;title2;
**************************
***** Variable Sel. ******
**************************;
proc glmselect data=customers1;
      title "MLR: Backward Variable Selection";
      class Segment(ref="Credit_C") City
                  Resp(ref="No") Store(ref="100");
      model Sales = Segment City Resp Store Purch Years  /
                  selection=backward(stop=cv)
                              cvmethod=random(5) stats=adjrsq;
run;title;
**************************
*****  6Step Test   ******
**************************;
```

```
/*Check Normality of the Data*/
proc glm data=customers1;
      title "ANOVA";
      class Segment;
      model Sales=Segment;
      means Segment/HOVTEST=WELCH bon cldiff;
      lsmeans Segment/pdiff;
run;
data customers3_log;set customers1;
      log_Sales = log(Sales);
run;
proc univariate data=customers;
      class Segment;
      var log_Sales;
      histogram log_Sales/ normal cfill='blue';
      title 'Avg. Sales Per Segment Histogram';
run;title;
proc glm data=customers3_log;
      title "ANOVA - Log/Log";
      class Segment;
      model log_Sales=Segment;
      means Segment/HOVTEST=Welch;
run;
**************************
*****   Extra SS    ******
**************************;
data customers_SS;set customers1;
      if Segment="Loyal_CC" then Group="Loyal_CC";
      if Segment="Loyalty" then Group="Other";
      if Segment="Store Ma" then Group="Other";
      if Segment="Credit_C" then Group="Other";
proc print data=customers_SS;
run;
proc glm data=customers_SS;
      title "ANOVA - Extra SS Test";
      class Group;
      model Sales=Group;
      means Group/HOVTEST=Welch;
run;
```

B. R CODE:

```
library(ggplot2)
library(caret)

# Read in csv file into data.frame
df <- read.csv("customers.csv",sep=",",header = TRUE)
#Reorder the columns -> response variable at end
df<-df[,c(1,3,4,5,6,7,9,2,10,11,12,8)]

dim (df)

#########################
####Dummy Variables#####
#########################

df$Segment1<-ifelse(df$Customer_Segment== "Loyal_CC",1,0)
df$Segment2<-ifelse(df$Customer_Segment == "Loyalty Club Only",1,0)
```

```
df$Segment3<-ifelse(df$Customer_Segment == "Store Mailing List",1,0)
df$Response <-ifelse(df$Responded_to_Last_Catalog =="Yes",1,0)


summary(df$Avg_Sale_Amount)

#Reorder; moving response to end
df1<-df[,c(1,2,3,4,5,6,7,8,9,10,11,13,14,15,16,12)]

########################
####Regression Model####
#######################

#Multiple Regression Model
model<-lm(Avg_Sale_Amount ~
Customer_Segment+Responded_to_Last_Catalog+Avg_Num_Products_Purchased+No_Years_as_Customer,
data=df)
#Summary of model (i.e. Parameter Estimates, Std. Error, P-value)
summary(model)

#Diagnostic Plots
plot(model,col="red")

############################
####Exploratory Analysis####
###########################

#Scatter Plots (Years as customer vs Sales)
plot(df$No_Years_as_Customer,df$Avg_Sale_Amount,
    main="Years as Customer vs Sales",
    xlab="No. Years as Customer",ylab="Avg. Sales Amount", col="red")

#Scatter Plot (Avg. Number Products Purchased vs Sales)
plot(df$Avg_Num_Products_Purchased,df$Avg_Sale_Amount,
    xlab="Avg. Number Products Purchased",ylab="Avg. Sales Amount",
    main="Products Purchased vs Sales",col="blue")

cust <- aggregate(df$Avg_Sale_Amount,by=list(df$No_Years_as_Customer),FUN=mean)
colnames(cust)<-c("Years","AvgSales")
cust

p <- ggplot(cust, aes(x = Years, y = AvgSales)) + geom_point(size=2,color="blue") +
  stat_smooth(method = 'lm', aes(colour = 'Linear'), se = FALSE)
p + stat_poly_eq(formula = cust$AvgSales~cust$Years,
               aes(label = paste(..eq.label.., ..rr.label.., sep = "~~~")),parse = TRUE) +
  labs(title="Years vs Avg. Sales", x="Years as Customer",y="Mean Sales Amount") +
  theme(plot.title=element_text(hjust=0.5))

#Check correlation between customers and average Sales
cor(cust$Years,cust$AvgSales)
```

```
########################
####MODEL PREDICTION####
#######################

#Cross Validation
set.seed(44)
tcontrol <- trainControl(method="cv",number=10,savePredictions = TRUE)

#Separate predictors and Response varaible
predictors<-df1[,1:6]
response<-rev(df1)[1]

df1 <- data.frame(cbind(predictors,response))

#fit lm model on the training set
features <- Avg_Sale_Amount~Segment1+Segment2+Segment3+Response+
   Avg_Num_Products_Purchased+No_Years_as_Customer+ Segment1*Avg_Num_Products_Purchased

fit <- train(features,data=df1,method="lm",metric="RMSE",trControl=tcontrol)

summary(fit)
fit$results
```