MSiA401: Statistical Analysis for Data Mining Project

Fall 2014

Iris Ye Tu, Lily Qian Zhao, Sanjeevni Wanchoo, Shawn Xiang Li

Executive Summary

This project aims at selecting the set of significant predictors that best predict candidate customers and forecast overall sales from catalog mailing using regression models. The fitting process consists of two steps: the first one is using the logistic regression model to predict the probability of a particular customer making purchases; next step is using the linear regression model to predict the amount that the buyers would pay on their purchases. Candidate customers with high expected outcomes should be considered potential buyers and thus further marketing strategies should target them specifically.

From the results of the analysis, customer orders in recent years (ordtyr,ordlyr,ord2ago), recencyoflast purchase (recencymon), consistency of purchase (ordtyr:ordlyr) fall and spring orders (falord, sprord), and date the customer was added (datead6) were good predictors to identify buyers versus non-buyers. For forecasting spending amounts, similarly, recent and historical sales and orders (slstyr, slslyr, sls2ago, sls3ago, slshist, ordtyr, ordlyr, ord2ago,ord3ago), the consistency in sales (ordtyr:ordlyr, slslyr:sls2ago), and date the customer was added (datead6) were useful in predictions.

Our best model was evaluated by predicting the dollar purchase resulting from catalog mailing for the top 5000 customers in the test set, and comparing this to the amount resulting from top 5000 customers from actual sales. Our model predicted $107,587 of the total $224,005 in sales revenue.

Introduction

The aim of the project was to use sales data from an upscale clothing retail company's online sales to predict which customers were most likely to buy their products from catalog, and how much they were likely to spend. All customers were sent a catalog mailing on Sep 1, 2012 and on Dec 1, 2012 and it was recorded whether or not they responded by making a purchase. The purchase amounts by each customer in response to the mailed catalog, along with other potential predictor variables based on their sales history were provided.

In building a predictive model for the catalog sales data, a classification and a prediction model was fitted and combined to forecast the purchase behaviors of the customers.

Building a predictive model based solely on multiple regression was considered. However, it was observed that about 90% of the customers did not purchase from catalog mailing. Due to the uneven distribution of data skewed largely in favor of no-buys, it was better to first classify the customers into two groups - buyers and non-buyers - and then fit a multiple linear model on the buyers to predict the amount they would spend on purchase.

The rest of the report is organized in the following structure:

- Ÿ Data cleansing

A discussion of how inconsistencies were addressed, and the additional predictor variables created for analysis.

- Ÿ Logistic Regression Model

1) Explanatory Analysis
2) Fitting Logistic Regression Model based on the cleaned training dataset
3) Model Diagnostics
4) Refitting and Finalizing the Logistics Regression Model

- Ÿ Multiple Regression Model

1) Explanatory Analysis
2) Fitting Multiple Linear Regression Model based on the cleaned training dataset
3) Descriptive analysis, univariable analysis, testing of collinearity, multivariable model building, model diagnostics such as residual analysis and outlier elimination

- Ÿ Model Validation

Data Cleaning:

The data provided contained several inconsistencies. Some of these were addressed during our data cleaning process, and are outlined below:

- **datelp6 vs lpuryear**: datelp6 is the date of last purchase, while lpuryear is the last purchase year. It was expected that the year of the former would match the latter in all cases. However, in a significant number of cases this was not true. After analysis of the spread for each, and feedback from client, datelp6 was chosen as the more accurate option for our analysis.
- **falord, sprord, & ordhist**: In about 6% of the cases, fall orders (falord) and spring orders (sprord) did not add up to total historical orders (ordhist). Initially, an additional variable was introduced in the dataset, that accounted for this difference, and attributed it any additional seasons. This was called "othord". However, it was determined that the company only had two shopping seasons - fall and spring. So we decided to rely only on falord and sprord, and disregard ordhist values to avoid discrepancies.
- **Reference years for orders**: For fields such as ordtyr, ordlyr, ord2ago and ord3ago, 2011-2012 was expected to be the reference year. However, in several cases, it was seen that ordtyr amounted to non-zero values even though the year of last purchase was before 2011. To account for this discrepancy, the data relating to the following four cases was removed from our dataset for model building (please note that datelp6_year corresponds to the year of datelp6). A total of 302 train and 274 test inconsistencies were removed.
    - Case 1: (ordtyr>0) & (datelp6_year is not 2011 or 2012)
    - Case 2: datelp6_year<2010 & ordlyr>0
    - Case 3: datelp6_year<2009 &ord2ago>0
    - Case 4: datelp6_year<2008 &ord3ago>0

- **Recency variables**: Three recency-related columns were calculated: recencyyear (recency in years), recencymon (recency in months) and recencydays (recency in days). These were calculated going back from Sep 1, 2012 as the point of reference (which is when the catalogs were first mailed out).

Logistic Regression Model:

As discussed in the data cleaning section, some inconsistencies exist which suggest that the variables lpuryear and ordhist are not reliable. As a result, lpuryear and ordhist would not be treated as candidate predictor variables in the model building process.

For the purpose of selecting remaining predictor variables, we implemented a both-ways stepwise procedure, which resulted in the model listed below.

```
Call:  glm(formula = buy ~ falord + sprord + slshist + datead6 + recencymon +
    recencydays + sls3ago + slstyr + slslyr + ord3ago + sls2ago +
    ord2ago + ordlyr, family = binomial, data = train)

Coefficients:
(Intercept)       falord       sprord      slshist      datead6   recencymon
 -2.757e-01    3.602e-01    1.945e-01   -1.950e-03   -4.827e-05   -2.510e+00
recencydays      sls3ago       slstyr       slslyr      ord3ago      sls2ago
  8.119e-02    2.404e-03    1.766e-03    2.028e-03   -1.049e-01    1.973e-03
    ord2ago       ordlyr
 -7.823e-02   -5.519e-02

Degrees of Freedom: 50115 Total (i.e. Null);  50102 Residual
Null Deviance:     31480
Residual Deviance: 24300  AIC: 24330
```

The model's AIC statistic is 24330.

Multicollinearity issue arises when there exists approximate linear relationships among two or more predictor variables. In order to test the multicollinearity assumption, the VIF table of the above model has been examined as below.

```
   falord      sprord     slshist     datead6  recencymon recencydays     sls3ago
 3.645990    1.844748    6.127401    1.779992 4572.716217 4575.535746    2.399068
   slstyr      slslyr     ord3ago     sls2ago     ord2ago      ordlyr
 1.702466    2.360348    1.980376    2.574839    2.361564    2.031606
```

Based on the VIF results, recencymon and recencydays have a VIF larger than 10. We then checked the analysis of deviance table (Appendix 1.1). The deviance of rececymon is 2403.14, and the deviance of recencydays is 1587.98. Thus, the recencydays variable was removed from the model.

Interaction terms might also have significant predicting powers, especially the interaction between year over year orders and sales. We built a model using only the interaction terms between year over year orders and sales as explanatory variables.A backwards stepwise procedure has been implemented to select top three significant interaction terms, which are ordtyr:ordlyr, ordlyr:ord2go and ord2ago:ord3ago.

```
Call:  glm(formula = buy ~ ordtyr:ordlyr + ordlyr:ord2ago + ord2ago:ord3ago,
    family = binomial, data = train)

Coefficients:
    (Intercept)    ordtyr:ordlyr   ordlyr:ord2ago  ord2ago:ord3ago
        -2.4546           0.6657           0.2132           0.3754

Degrees of Freedom: 50115 Total (i.e. Null);  50112 Residual
Null Deviance:     31480
Residual Deviance: 29810  AIC: 29820
```

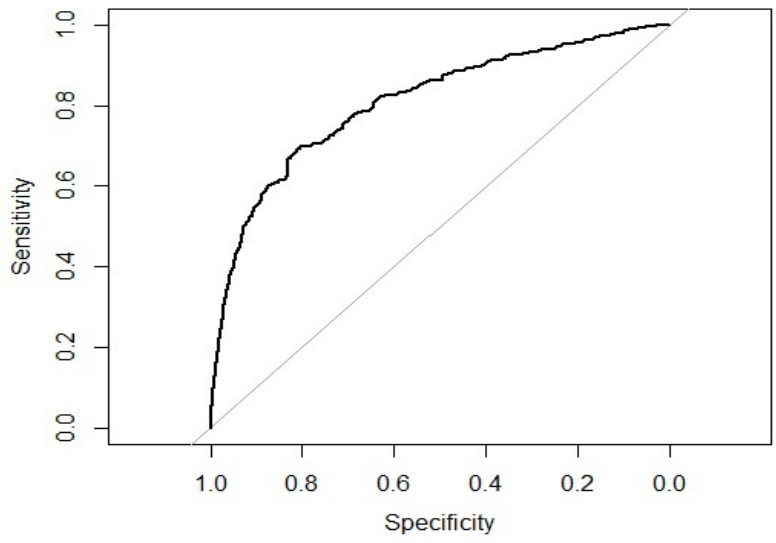
We added the top three interaction terms into the original model, and examined the updated model summary report in Appendix 1.2, and detected several variables with a p-value larger than 0.05 (5% significance level), which includes slstyr, slslyr, ord2ago, ord2ago:ordlyr and ord3ago:ord2ago. The non-significant explanatory variables have been removed from the model. Because the interaction term ordtyr:ordlyr is a significant predictor variable, we added back ordtyr variables and generated a summary report for the updated model in Appendix 1.3. Based on the result in Appendix 1.3, the p-values of sls3ago and ord3ago are higher than 0.05 (5% significance level) in the updated model. After removing the non-significant variables, a final model has been built. The summary report of the finalized logistics model is listed below.

```
Call:
glm(formula = buy ~ falord + sprord + datead6 + recencymon +
    ord2ago + ordtyr + ordlyr + ordtyr:ordlyr, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.9938  -0.4771  -0.3300  -0.1398   3.7614

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.038e-01  1.843e-01  -1.648  0.09938 .
falord         2.341e-01  1.179e-02  19.860  < 2e-16 ***
sprord         2.163e-01  1.462e-02  14.794  < 2e-16 ***
datead6       -5.172e-05  1.169e-05  -4.424 9.67e-06 ***
recencymon    -5.590e-02  1.566e-03 -35.699  < 2e-16 ***
ord2ago       -8.887e-02  2.844e-02  -3.125  0.00178 **
ordtyr        -4.360e-01  4.679e-02  -9.318  < 2e-16 ***
ordlyr        -3.037e-01  3.737e-02  -8.126 4.43e-16 ***
ordtyr:ordlyr  3.461e-01  3.828e-02   9.041  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31477  on 50115  degrees of freedom
Residual deviance: 25852  on 50107  degrees of freedom
AIC: 25870

Number of Fisher Scoring iterations: 7
```

The VIF table is listed below:

```
   falord        sprord       datead6     recencymon       ord2ago        ordtyr
2.023865      1.324012      1.785359      2.188708      1.337186      3.397774
   ordlyr ordtyr:ordlyr
2.358175      2.735336
```

The ROC statistic of the final logistic model is 0.8073.



Logistic Regression Diagnostics

There are mainly two steps in assessing the fit of the model: first is to determine if the model fits using summary measures of goodness of fit or by assessing the predictive ability of the model; second is to determine if there's any observations that do not fit the model or that have an influence on the model. Printouts for such model diagnostics are included in Appendix 1.4~.

- **Goodness-of-fit test**

All goodness-of-fit tests are based on the premise that the data will be divided into subsets and within each subset the predicted number of outcomes will be computed and compared to the observed number of outcomes. The Pearson \chi^2 and the deviance \chi^2 are based on dividing the data up into the natural covariate patterns. In our model, the p-value is less than 2.2e-16. The p-value is small indicating no evidence of lack of fit. (Appendix 1.4)

- **Outlier and Influential Points**

Outliers as well as influential points are taken into account and addressed in the model. Looking at the graph for outliers, it is obvious that a certain point (69850) is away from the fitted line, which we will delete. (Appendix 1.5)

Then it comes to the influential points. Like for linear regression, large positive or negative standardized residuals allow to identify points, which are not well fit by the model. A plot of Pearson residuals as a function of the logit for fit.newtar is drawn here, with bubbles relative to size of the covariate pattern. The plot should be a horizontal band with observations between a parallel bands. Covariate patterns 69850, 29944 and 49039 are problematic. (Appendix 1.6)

Because there are multiple observations that satisfy multiple of the above criterion, 3 observations were removed, for a new training set total of 50113 customers.

- **Refitting and Finalizing the Logistics Regression Model**

The model was refit, and the results are displayed in Appendix 1.7.

From the plots, it is evident that Deviance and Pearson residuals are more homogenous, and while there are new leverage points, Deviance and Pearson Chi-Square deletion plots show more continuity. AIC has dropped from 25870 to 25810 after deleting the outliers and high leverage points.

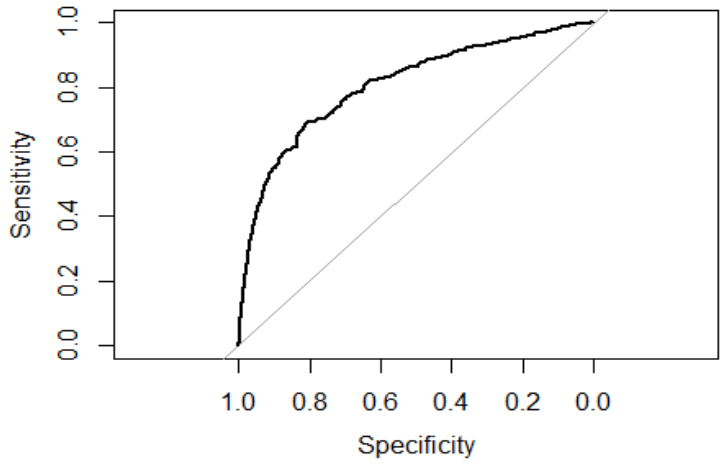
The final model, which will be seeked to validate on the test set with, consists of 8 predictors: falord + sprord + datead6 + recencymon + ord2ago + ordtyr + ordlyr + ordtyr:ordlyr. All predictors are significant at the $\alpha = 0.05$ level. The AIC for our final model is 25810 (the lowest yet), and area under ROC curve is 0.8072. The summary report of the final logistics model is listed below.

```
Call:
glm(formula = buy ~ falord + sprord + datead6 + recencymon +
    ord2ago + ordtyr + ordlyr + ordtyr:ordlyr, family = binomial,
    data = newtrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8252  -0.4739  -0.3315  -0.1396   3.7721

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.052e-01  1.860e-01  -2.179 0.029357 *
falord         2.470e-01  1.186e-02  20.831  < 2e-16 ***
sprord         2.195e-01  1.467e-02  14.963  < 2e-16 ***
datead6       -4.355e-05  1.180e-05  -3.691 0.000223 ***
recencymon    -5.624e-02  1.572e-03 -35.777  < 2e-16 ***
ord2ago       -1.068e-01  2.856e-02  -3.740 0.000184 ***
ordtyr        -4.710e-01  4.727e-02  -9.963  < 2e-16 ***
ordlyr        -3.318e-01  3.774e-02  -8.792  < 2e-16 ***
ordtyr:ordlyr  3.884e-01  3.901e-02   9.955  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31477  on 50112  degrees of freedom
Residual deviance: 25792  on 50104  degrees of freedom
AIC: 25810

Number of Fisher Scoring iterations: 7
```

The VIF table is listed below:

```
     falord       sprord      datead6    recencymon      ord2ago       ordtyr       ordlyr
   2.028029     1.321956     1.797163     2.204653     1.337759     3.428597     2.379858
ordtyr:ordlyr
   2.763328
```

The ROC statistic of the final logistic model is 0.8072.



We then tested the model in the test data set again, and created a classification table using a cut off of 0.2.

test_y	FALSE	TRUE	Sum
0	43224	2957	46181
1	2483	2176	4659
Sum	45707	5133	50840

The overall classification rate is 0.9004721 and the sensitivity rate is 0.4491228.

> total.test.rate

[1] 0.9004721

> test.buy.rate

[1] 0.4491228

Multiple Regression Model

In building the multiple regression, the interaction terms between year over year orders and sales are included as candidate explanatory variables.A both-ways stepwise procedure has been implemented to select variables. Below is the model resulted from the stepwise procedure.

```
Call:
lm(formula = log(targdol + 1) ~ datead6 + datelp6 + slstyr +
    slslyr + sls2ago + sls3ago + slshist + ordtyr + ordlyr +
    ord2ago + ord3ago + falord + recencymon + ordtyr:ordlyr +
    ordlyr:ord2ago + slslyr:sls2ago + sls2ago:sls3ago, data = multiple_project)

Coefficients:
    (Intercept)          datead6          datelp6          slstyr          slslyr
      6.782e+01        3.276e-05       -4.157e-03        1.959e-03        2.468e-03
        sls2ago          sls3ago          slshist          ordtyr          ordlyr
      3.692e-03        1.859e-03        5.916e-04       -1.015e-01       -1.553e-01
        ord2ago          ord3ago           falord       recencymon    ordtyr:ordlyr
     -1.632e-01       -1.116e-01        1.577e-02       -1.257e-01        5.686e-02
 ordlyr:ord2ago   slslyr:sls2ago  sls2ago:sls3ago
      2.413e-02       -1.446e-05       -7.266e-06
```

We examined the VIF table of this model.

datead6	datelp6	slstyr	slslyr	sls2ago
1.701901	6846.852160	2.035198	2.587550	3.430613
sls3ago	slshist	ordtyr	ordlyr	ord2ago
5.666405	5.200016	2.953411	3.582332	3.438059
ord3ago	falord	recencymon	ordtyr:ordlyr	ordlyr:ord2ago
2.231814	4.096927	6821.493948	3.036595	3.074210
slslyr:sls2ago	sls2ago:sls3ago			
1.996983	3.753089			

Based on the VIF results, datelp6 and recencymon have a VIF larger than 10. We decided to delete the variable datelp6.

After deleting variables datelp6, the model was refitted. The model summary is attached in Appendix 2.1. Based on the summary, falord ,recencymon, and ordlyr:ord2ago, which have p-values larger than 0.05 (5% significance level), are not significant predictor variables. Thus, they have been removed from the model.

- **Square root transformation**

We conducted square root transformation on predictor variables to improve the fit .Square root transformations have been made for all the non-interaction terms in the model, which significantly improved the fit. The model summary is in Appendix 2.2. The R-Square of the transformed model increased to 13.93. The model summary suggests that the p-value of the interaction term sls2ago:sls3ago is above 0.05 (5% significance level) in the transformed model.After deleting the non-significant variable, the model was finalized. The summary of finalized multiple regression model is listed below.

```
Call:
lm(formula = log(targdol + 1) ~ datead6 + sqrt(slstyr) + sqrt(slslyr) +
    sqrt(sls2ago) + sqrt(sls3ago) + sqrt(slshist) + sqrt(ordtyr) +
    sqrt(ordlyr) + sqrt(ord2ago) + sqrt(ord3ago) + ordtyr:ordlyr +
    slslyr:sls2ago, data = multiple_project)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6639 -0.5236 -0.0339  0.4610  3.6374

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.974e+00  1.077e-01  27.616  < 2e-16 ***
datead6         3.009e-05  6.829e-06   4.406 1.08e-05 ***
sqrt(slstyr)    5.293e-02  6.176e-03   8.569  < 2e-16 ***
sqrt(slslyr)    4.308e-02  6.713e-03   6.418 1.52e-10 ***
sqrt(sls2ago)   6.182e-02  7.599e-03   8.136 5.19e-16 ***
sqrt(sls3ago)   2.415e-02  6.859e-03   3.521 0.000433 ***
sqrt(slshist)   2.401e-02  3.516e-03   6.830 9.58e-12 ***
sqrt(ordtyr)   -3.756e-01  4.073e-02  -9.223  < 2e-16 ***
sqrt(ordlyr)   -3.328e-01  4.358e-02  -7.636 2.69e-14 ***
sqrt(ord2ago)  -4.089e-01  4.705e-02  -8.690  < 2e-16 ***
sqrt(ord3ago)  -1.965e-01  4.418e-02  -4.448 8.86e-06 ***
ordtyr:ordlyr   6.521e-02  1.331e-02   4.897 1.00e-06 ***
slslyr:sls2ago -9.711e-06  2.120e-06  -4.580 4.76e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7299 on 4750 degrees of freedom
Multiple R-squared:  0.1386, Adjusted R-squared:  0.1365
F-statistic: 63.71 on 12 and 4750 DF,  p-value: < 2.2e-16
```

- **Comparison between Non-square-root-Transformation Model and Transformed Model**

The non-square-root-Transformation Model is:

fit.multi2 = lm(formula = log(targdol + 1) ~ datead6 + (slstyr) + (slslyr) + (sls2ago) + (sls3ago) + (slshist) + (ordtyr) + (ordlyr) + (ord2ago) + (ord3ago) + ordtyr:ordlyr + slslyr:sls2ago, data = multiple_project)

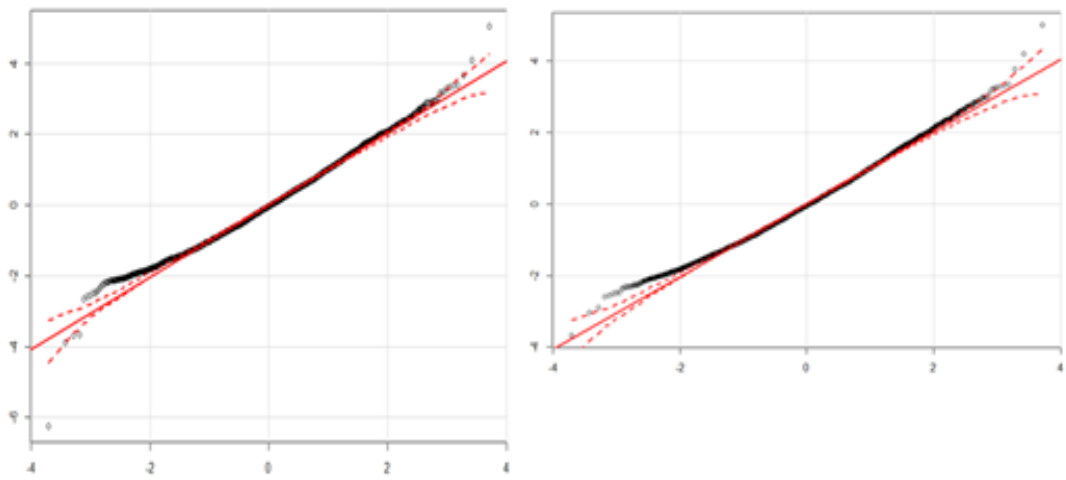
And the Transformed Model is:

fit.multi = lm(formula = log(targdol + 1) ~ datead6 + sqrt(slstyr) + sqrt(slslyr) + sqrt(sls2ago) + sqrt(sls3ago) + sqrt(slshist) + sqrt(ordtyr) + sqrt(ordlyr) + sqrt(ord2ago) + sqrt(ord3ago) + ordtyr:ordlyr + slslyr:sls2ago, data = multiple_project)

- **R-squared**

The non-square-root-Transformation Model is: 0.1063. The Transformed Model is: 0.1386, which means the transformed model's predictors have a better prediction of the sales.

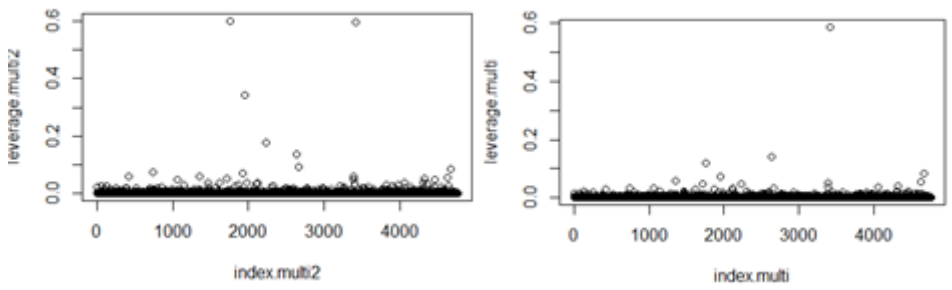
- **Normality: Q-Q Plot**



The non-square-root-Transformation Model The Transformed Model

A difference in the left corner could be seen. So the transformed model is better in normality.

- **Leverage Analysis**



The non-square-root-Transformation Model The Transformed Model

The transformed model is with less high leverage points. Based on the three major comparisons above, the transformed model is used as the final model.

Multiple Linear Diagnostics

After fitting a regression model, it is important to determine whether all the necessary model assumptions valid before performing inference. If there are any violations, subsequent inferential procedures may be invalid resulting in faulty conclusions. Therefore, it is crucial to perform appropriate model diagnostics.

There are four basic assumptions that underlie the multiple linear regression model:

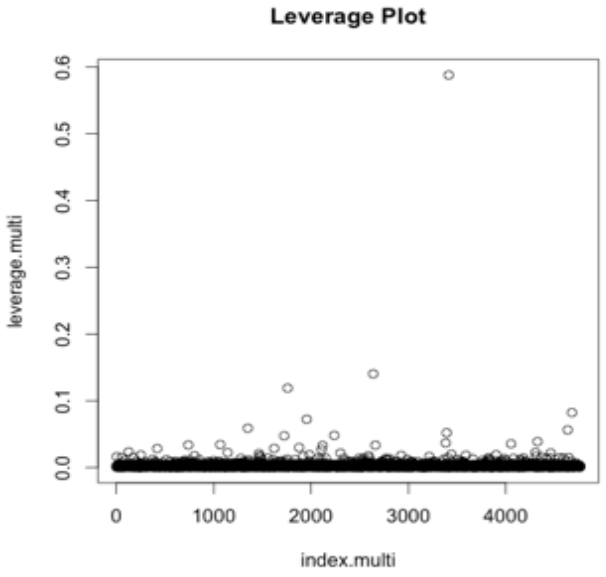
1. The response variable y to the explanatory variables are linear in the β parameters.
2. The errors are normally distributed with zero mean.
3. The errors have a constant variance σ^2.
4. The errors are independent.

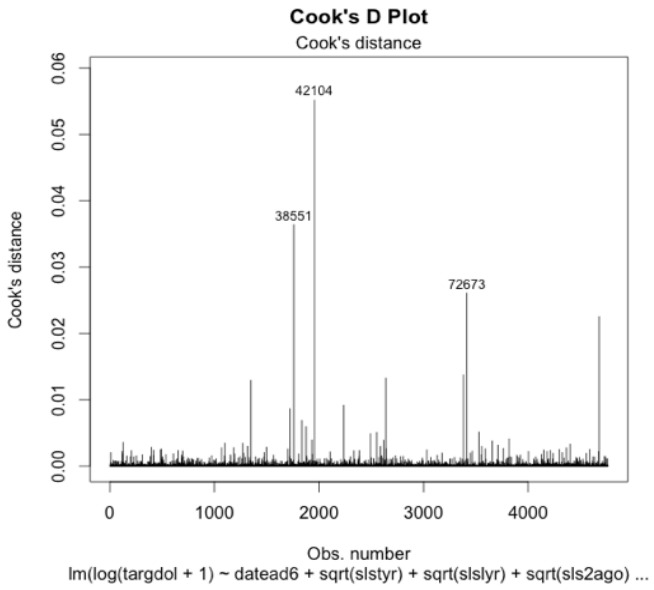
We checked these above four assumptions. Moreover, we checked if there were outliers and influential leverage and deleted such observations to ensure the model gets improved.

- **Outliers and influential observations**

We used partial-regression plots between the response variable and each predictive variable in order to check if any outliers exist. Based on these partial-regression plots, we found there was one outlier. (Appendix 3.1) The outlier was indicated by *outlierTest* function from car package in R.

The leverage for each observation Y_i equals to the weight put by the observation on its own fitted value $\hat{y}_i$, which is the (i,j) entry of the hat matrix. We had a leverage plot by plotting observation number vs. its corresponding leverage. An influential observation is an observation has too much influence on its own fitted value then its leverage will be much larger in comparison to other observations' leverage. From the leverage plot, we concluded multiple influential leverages exited.





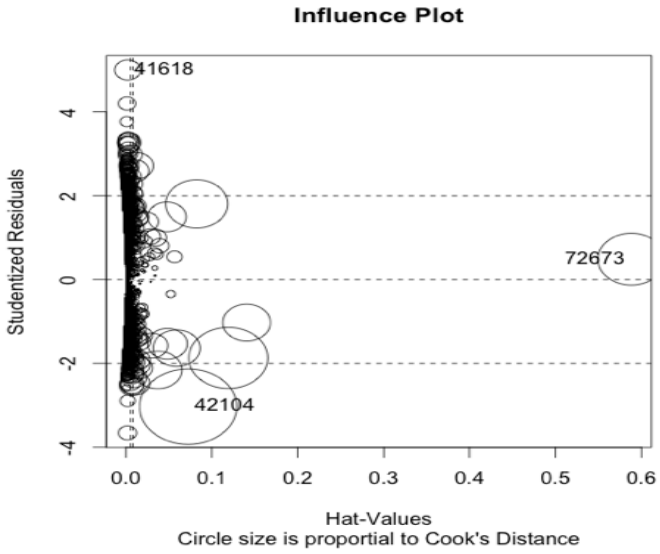
One method to indicate these influential points is using Cook's Distance (D). We can declare the ith observation as influential at significance level α if $D_i > 4/(n-p-1)$. We plot Cook's distance with observation number under the model. There were 3 influential observations were indicated out.

The other method we used to indicate influential observations is function influencePlot from car package in R. It gave an influence plot and table of influential observations. (Table 1)

	StudRes	Hat	CookD
41618	5.0013164	0.002050164	0.06271301
42104	-3.0315833	0.072483517	0.23484566
72673	0.4870216	0.588224560	0.16145546

Table 1

After deleting all duplicated result, we found out 1 outlier and 4 influential observations with high leverages. We deleted these 5 observations before checking the four assumptions. We fitted the model again and fount out the adjust R^2 increased from 13.65% to 13.71%. It shows the model fitted data better after deleting outliers and influential observations.



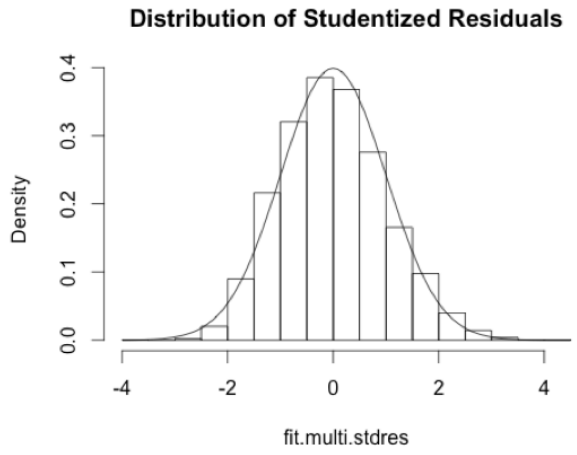
- **Checking the Linearity**

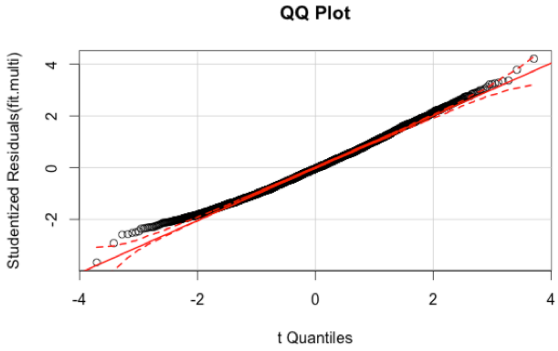
A partial residual plot shows the relationship between the response variable and each predictive variable, which is represented by the dash line in each plot. We used crPlots from car package in R to get partial residual plots by each predictive variable. (Appendix 3.2) From these plots, we can clearly see each predictive variable is linearly associated to the response variable. Therefore, all data satisfy the linear model we got.

- **Normality**

Method 1: we checked the distribution of studentized residuals. [Figure: Distribution of Studentized Residuals] Based on the plot, the studentized residuals are normally distributed with a mean zero.

Method 2: Normal quantile-quantile (Q-Q) plot was checked: plot normal distribution quantiles vs. quantiles of the studentized residuals. The plot is roughly linear and the line goes across origin. It shows the residuals are normally distributed with a mean zero.

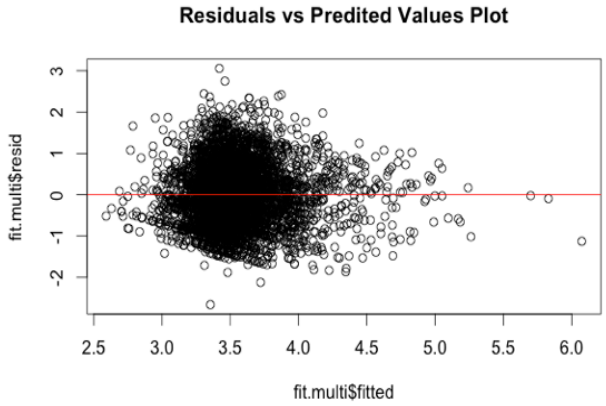




Overall, both results of above two methods help to conclude that the errors are normally distributed with zero mean.

- **Homoscedasticity**

If the errors have a constant variance σ^2, we call it homoscedasticity. If $Var(y_i) = Var(\varepsilon_i)$ is not constant then usually it is a function of $E(y_i)$. To see how $Var(y_i)$ depends on $E(y_i)$ we plot the residuals vs. fitted values. [Figure: Residuals vs. Predicted values Plot] The plot is almost random forming a parallel band, which shows the residuals do not increase with $E(y_i)$. We can conclude that the errors have a constant variance σ^2.



- **Time Independence**

The independence assumption is typically violated for time series data when the data are auto or serially correlated. Durbin-Watson (DW) test is a formal test for testing for independence. Errors are independent if D-W statistic is close to cutoff level 2. Following is the result from DW test. D-W Statistic equals 1.94544, which is very close to 2. (Table2) Therefore, assumption that errors are independent satisfies.

Lag	Autocorrelation	D-W Statistic	p-value
1	0.02711357	1.94544	0.056

Alternative hypothesis: rho != 0

Table 2

- **Multicollinearity**

Multicollinearity happens when two or more predictive variables in a multiple regression model are highly correlated. Multicollinearity usually exits when a model with all highly significant predictive variables has a low R^2. Variance inflation factor (VIF) is a good measure to declare any multicollinearity when VIF is larger than 10. We checked VIF for the model and get the following result. (Table 3)

datead6	sqrt(slstyr)	sqrt(slslyr)	sqrt(sls2ago)	sqrt(sls3ago)	sqrt(slshist)	sqrt(ordlyr)
1.586810	5.542567	6.466950	7.539817	6.147408	4.114532	5.710549

sqrt(ordtyr)	sqrt(ord2ago)	sqrt(ord3ago)	ordtyr:ordlyr	slslyr:sls2ago
5.077386	6.271373	5.428849	2.017471	1.762536

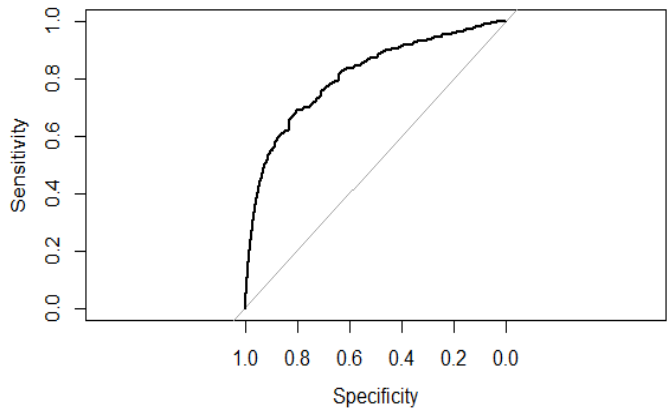
Table 3

We indicated and deleted all outliers and influential observation. Each of these four assumptions is satisfied under the multiple linear model. Moreover, there is no multicollinearity. Overall, these results from the model diagnostics ensure that the multiple linear model is good to use.

Model Validation

- **ROC Curve and Cut-Off Probabilities**
  - ROC Curve on test set

In order to validate the logistic regression model, the ROC curve for the test data was studied. The ROC area under the curve of the final logistic model is 0.8096. The graph is shown below:



Cutoff Probability

For the purpose of choosing the best cut off probability, a cost function C (C = c0n01+c1n10) has been introduced, where c0 is the cost of misclassifying 0 to 1 (a false positive), and c1 is the cost of misclassifying 1 to 0 (a false negative). The best cut off probability can be determined by minimizing the cost function C.

Because the objective of the model is to capture potential buyers, the sensitivity rate is much more important than specificity rate for this model. Thus, we set the cost ratio c0/c1 as 3. The best cut-off turns out to be around 0.2.

We then tested the model in the test data set, and created a classification table using the cut off of 0.2. The overall classification rate is 0.9000555, and the sensitivity rate is 0.4474716.

test_y	FALSE	TRUE	Sum
0	43211	2970	46181
1	2491	2168	4659
Sum	45702	5138	50840

```
> total.test.rate
[1] 0.9000555
> test.buy.rate
[1] 0.4474716
```

- Overall Combined Model

In order to validate our model, two criteria were used. First, the statistical criteria was to look at our mean square error of prediction (MSEP). The second criteria was the financial criteria, which aimed to quantify how much of the actual sales amounts our model predicted. The values for both are given below.

- **Statistical Criteria**

The calculated Mean Square Error of Prediction (MSEP) for the given model (using a cutoff probability of 0.203 for our logistic model) is given below. Here, n is the total data points in test set and p is the number of predictors (p+1 is the number of β coefficients).

$[\sum_{i=1}^{n}(\hat{y}_i - y_i)^2]/[n-(p+1)] = 5212.832$, where n = 4963, and p = 12.

- **Financial Criteria:**

For this criteria, we calculated the actual total sales that will result if the models were applied to the test set and catalogs were mailed to the top 5000 prospects. Logistic model was used to first calculate the buyers in the test set. The earlier criteria gave us an optimal cut-off of about 0.20. Further analysis provided an optimal financial value at 0.203.This was used as the cut-off probability for this analysis. Next, their predicted purchase values were calculated using the multiple regression model. Of these, the top 5000 predictions were taken, and the actual targdol (purchase value) was summed. The values are shown below:

```
> sum(test.predicted.targdol$targdol)
[1] 107587
```

This was compared with the actual total purchase amount of the actual top customers in test data. These values are given below:

Thus, our model predicted $107,587 of the total $224,005 of customer spending on catalogue-based purchases. Please note that about 4700 of the values in the test data had purchases.

Conclusion

From the results of the analysis, customer orders in recent years (ordtyr,ordlyr,ord2ago), recency of last purchase (recencymon), consistency of purchase (ordtyr:ordlyr) fall and spring orders (falord, sprord), and date the customer was added (datead6) were good predictors to identify buyers versus non-buyers. For forecasting spending amounts, similarly, recent and historical sales and orders (slstyr, slslyr, sls2ago, sls3ago, slshist, ordtyr, ordlyr, ord2ago,ord3ago), the consistency in sales (ordtyr:ordlyr, slslyr:sls2ago), and date the customer was added (datead6) were useful in predictions.

In geneneral, the predictor variables were often interrelated, and let to high VIF values. A large part of the analysis was using the best predictors without introducing multicollinearity in our model.
Our best model was evaluated by predicting the dollar purchase resulting from catalog mailing for the top 5000 customers in the test set, and comparing this to the amount resulting from top 5000 customers from actual sales. Our model predicted $107,587 of the total $224,005 in sales revenue with a mean squared error of prediction (MSEP) of 5212.832.

The customer data provided was based only on sales. Having some additional customer demographic and financial data (such as age, address zipcode, salary, etc) would provide additional information, and could be useful in building any future models.

Appendix

Appendix 1.1

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: buy

Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev
NULL			50115	31477
falord	1	1926.01	50114	29551
sprord	1	840.63	50113	28711
slshist	1	44.24	50112	28667
datead6	1	338.33	50111	28328
recencymon	1	2403.14	50110	25925
recencydays	1	1587.98	50109	24337
sls3ago	1	5.18	50108	24332
slstyr	1	7.43	50107	24324
slslyr	1	7.83	50106	24317
ord3ago	1	5.83	50105	24311
sls2ago	1	4.11	50104	24307
ord2ago	1	3.47	50103	24303
ordlyr	1	2.54	50102	24301

Appendix 1.2

```
Call:
glm(formula = buy ~ falord + sprord + slshist + datead6 + recencymon +
    sls3ago + slstyr + slslyr + ord3ago + sls2ago + ord2ago +
    ordtyr:ordlyr + ordlyr:ord2ago + ord2ago:ord3ago, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.6194  -0.4804  -0.3148  -0.1563   3.5673

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -6.008e-01  1.843e-01  -3.259  0.00112 **
falord           2.723e-01  1.580e-02  17.233  < 2e-16 ***
sprord           2.558e-01  1.741e-02  14.691  < 2e-16 ***
slshist         -1.477e-03  2.447e-04  -6.037 1.57e-09 ***
datead6         -6.842e-05  1.158e-05  -5.908 3.47e-09 ***
recencymon      -4.513e-02  1.202e-03 -37.535  < 2e-16 ***
sls3ago          1.912e-03  4.753e-04   4.022 5.76e-05 ***
slstyr           3.461e-04  4.961e-04   0.698  0.48537
slslyr           4.191e-04  4.992e-04   0.839  0.40124
ord3ago         -9.030e-02  4.150e-02  -2.176  0.02956 *
sls2ago          1.673e-03  6.383e-04   2.621  0.00878 **
ord2ago         -7.954e-02  4.222e-02  -1.884  0.05954 .
ordtyr:ordlyr    7.115e-02  2.705e-02   2.630  0.00853 **
ord2ago:ordlyr  -2.294e-03  2.612e-02  -0.088  0.93001
ord3ago:ord2ago  9.317e-03  3.267e-02   0.285  0.77550
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31477  on 50115  degrees of freedom
Residual deviance: 25894  on 50101  degrees of freedom
AIC: 25924

Number of Fisher Scoring iterations: 6
```

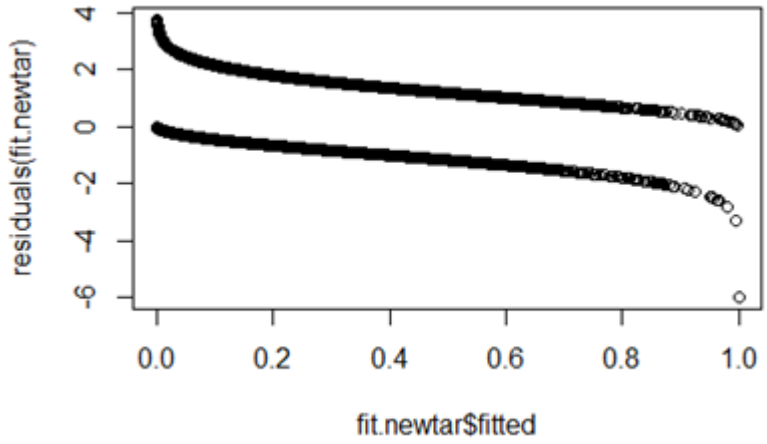
Appendix 1.3

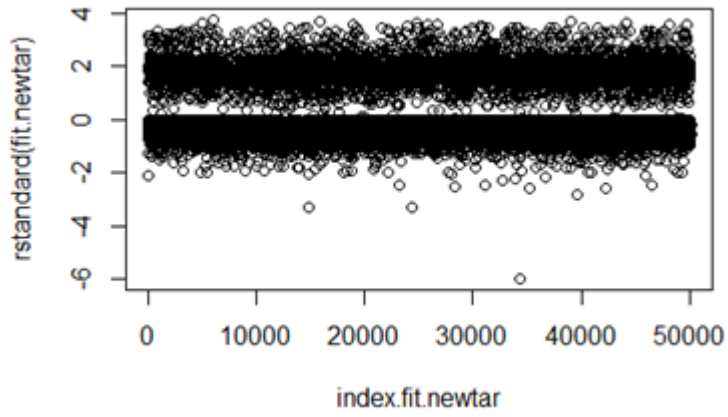
```
Call:
glm(formula = buy ~ falord + sprord + datead6 + recencymon +
    sls3ago + ord3ago + ord2ago + ordtyr + ordlyr + ordtyr:ordlyr,
    family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.1878  -0.4773  -0.3263  -0.1413   3.7554

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.386e-01  1.855e-01  -1.825 0.067992 .
falord         2.459e-01  1.298e-02  18.944  < 2e-16 ***
sprord         2.275e-01  1.552e-02  14.664  < 2e-16 ***
datead6       -4.896e-05  1.179e-05  -4.151  3.3e-05 ***
recencymon    -5.578e-02  1.559e-03 -35.775  < 2e-16 ***
sls3ago       -7.895e-05  3.757e-04  -0.210 0.833561
ord3ago       -6.408e-02  3.566e-02  -1.797 0.072346 .
ord2ago       -9.579e-02  2.861e-02  -3.349 0.000812 ***
ordtyr        -4.502e-01  4.724e-02  -9.530  < 2e-16 ***
ordlyr        -3.161e-01  3.783e-02  -8.356  < 2e-16 ***
ordtyr:ordlyr  3.498e-01  3.837e-02   9.117  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31477  on 50115  degrees of freedom
Residual deviance: 25847  on 50105  degrees of freedom
AIC: 25869

Number of Fisher Scoring iterations: 7
```

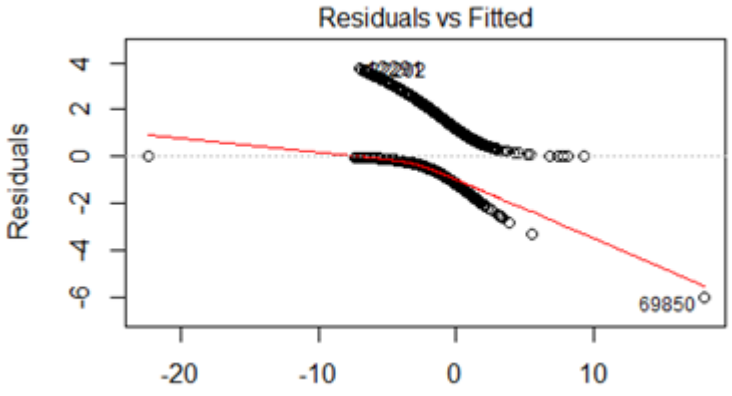
Appendix 1.4

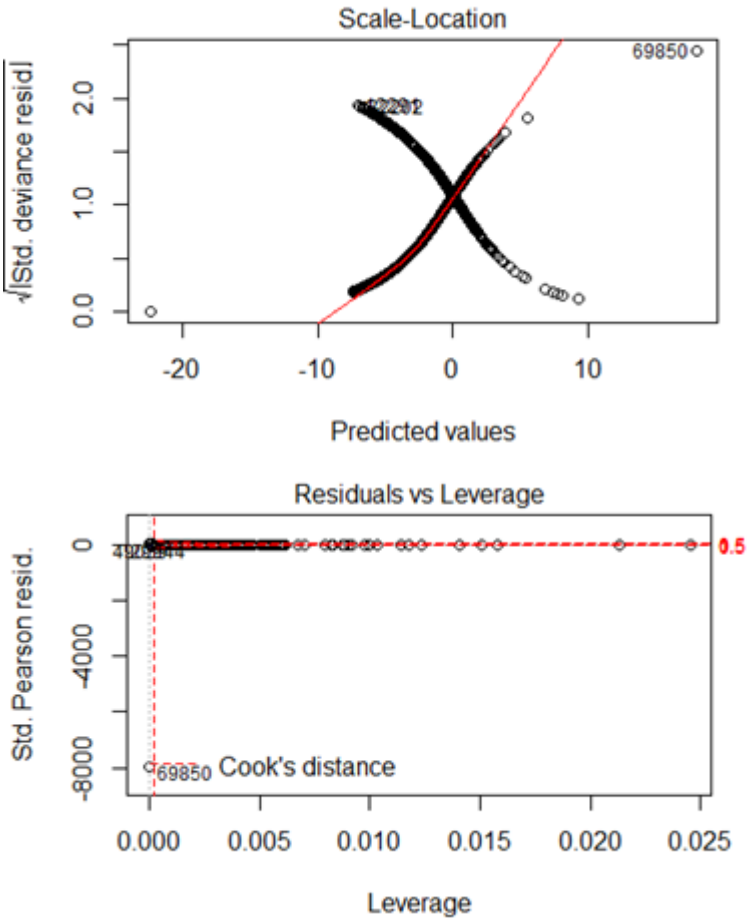
```
> chisq.test(dataframe.diag)

 Pearson's Chi-squared test

data:  dataframe.diag
X-squared = 1069878, df = 350805, p-value < 2.2e-16
```

Appendix 1.5

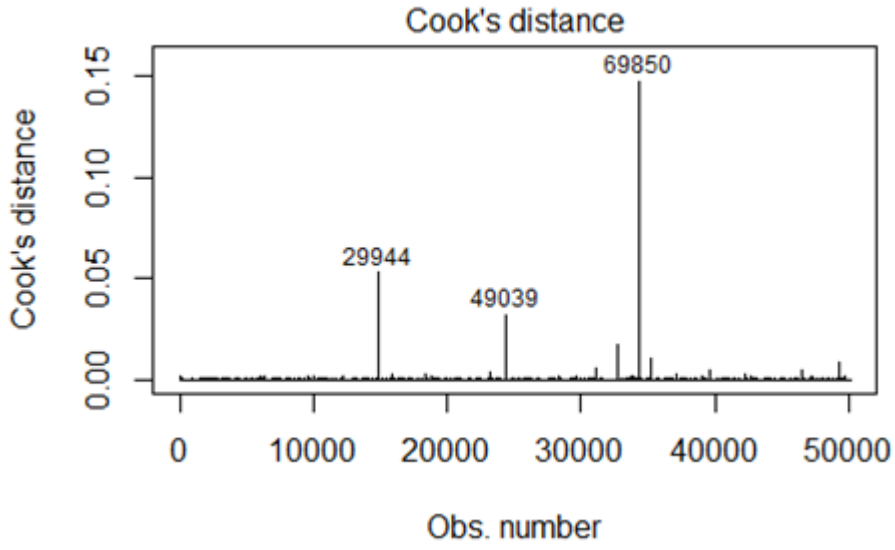

residuals(fit.newtar)
4
2
0
-2
-4
-6
0.0
0.2
0.4
0.6
0.8
1.0
fit.newtar$fitted

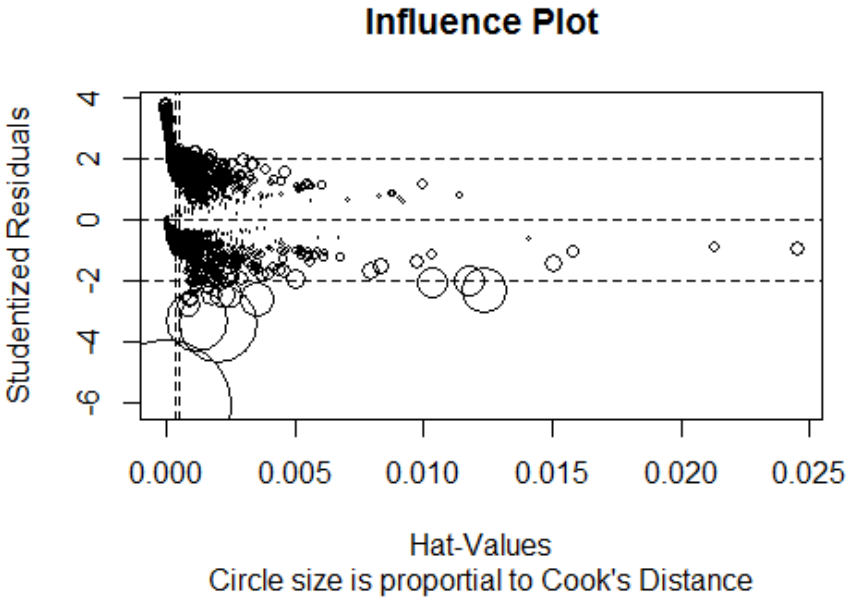

rstandard(fit.newtar)
4
2
0
-2
-4
-6
0
10000
20000
30000
40000
50000
index.fit.newtar


Residuals vs Fitted
Residuals
4
2
0
-2
-4
-6
69850
-20
-10
0
10

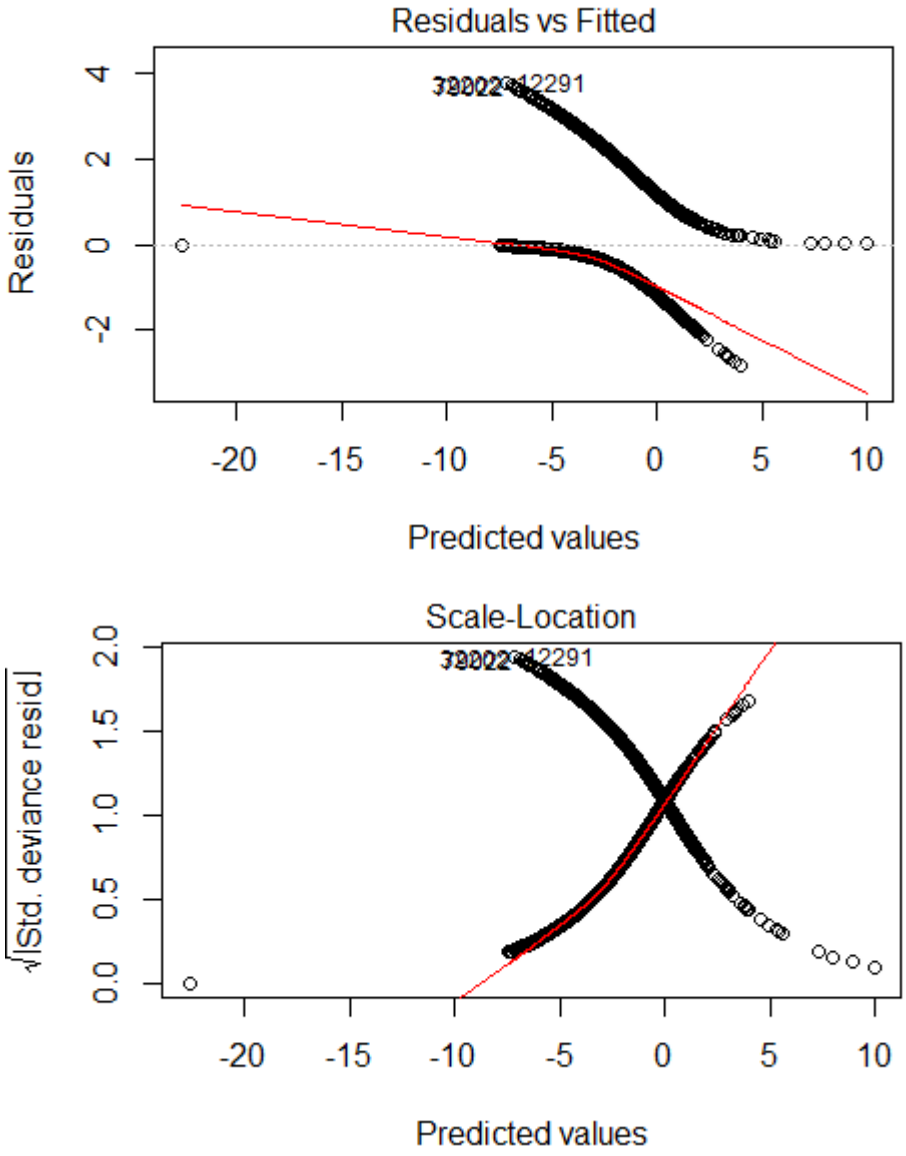

Scale-Location
69850
√|Std. deviance resid|
0.0
1.0
2.0
-20
-10
0
10
Predicted values
Residuals vs Leverage
0.5
Std. Pearson resid.
0
-4000
-8000
69850
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
Leverage

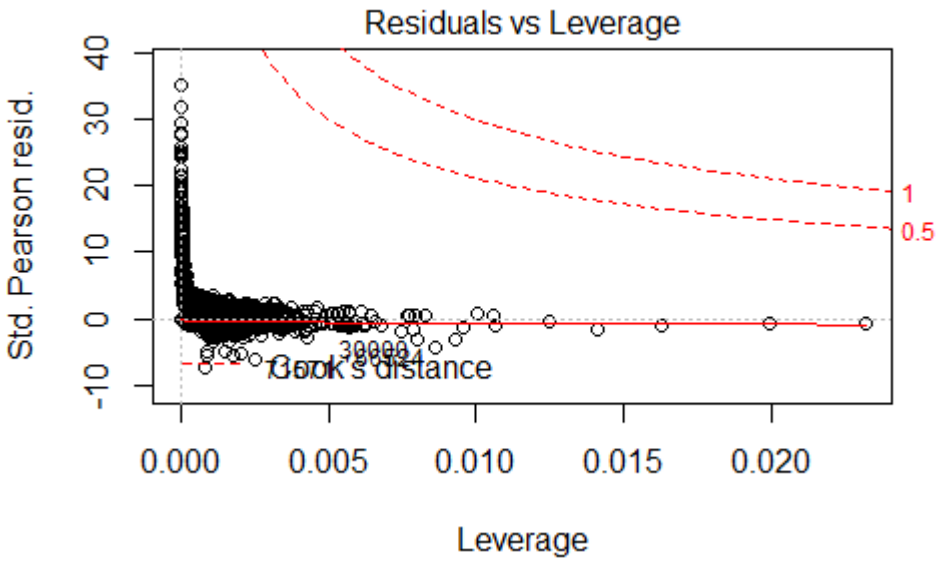
Appendix 1.6


Cook's distance
69850
29944
49039
Cook's distance
0.00
0.05
0.10
0.15
0
10000
20000
30000
40000
50000
Obs. number


Influence Plot
Studentized Residuals
-6
-4
-2
0
2
4
0.000
0.005
0.010
0.015
0.020
0.025
Hat-Values
Circle size is proportial to Cook's Distance

Appendix 1.7


Residuals vs Fitted
Residuals
Predicted values
Scale-Location
√|Std. deviance resid|
Predicted values


Residuals vs Leverage
Std. Pearson resid.
-10
0
10
20
30
40
1
0.5
30000
Cook's distance
0.000
0.005
0.010
0.015
0.020
Leverage

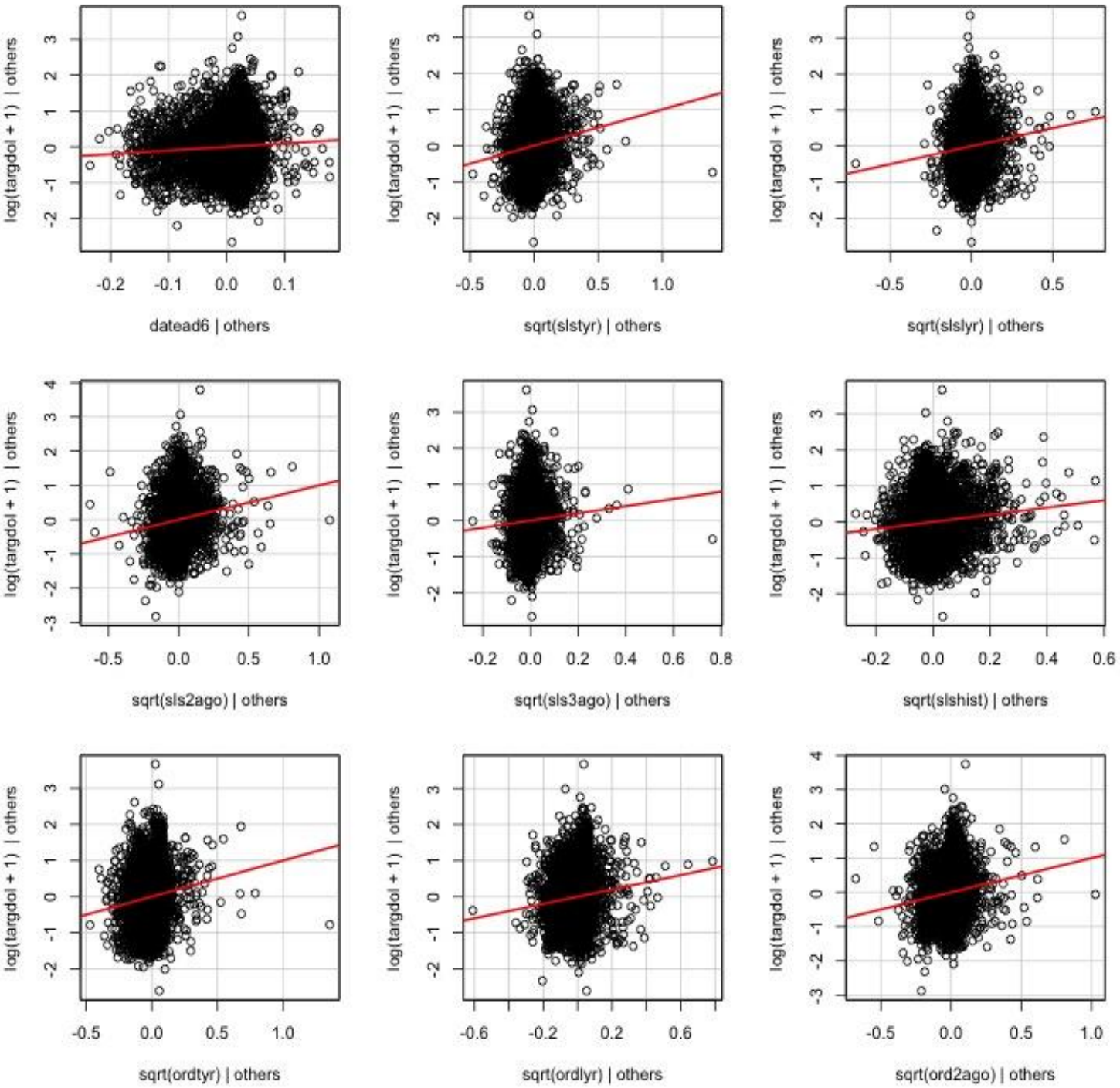
Appendix 2.1

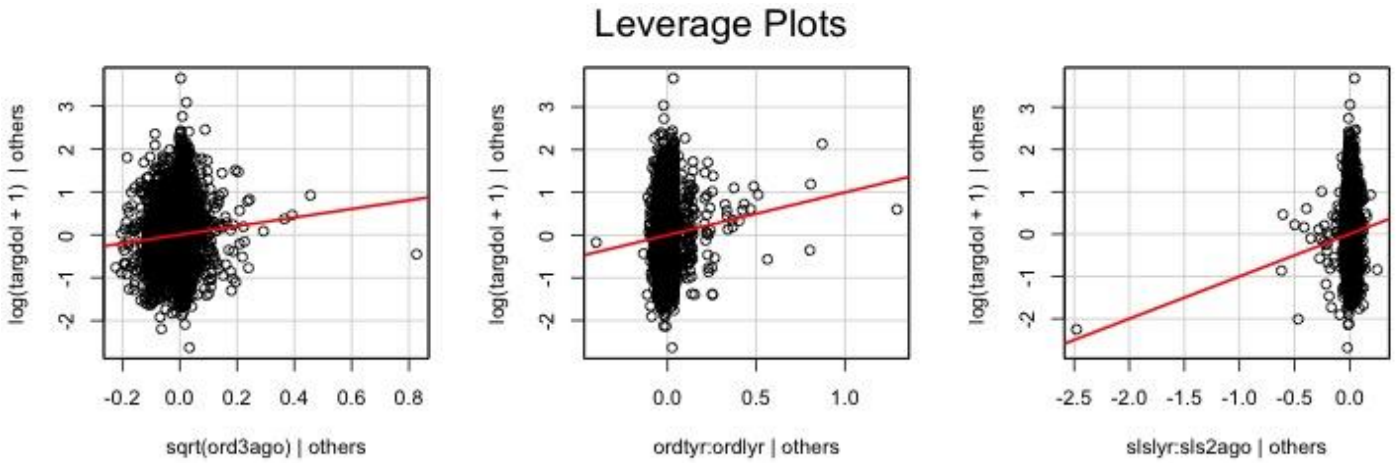
```
Call:
lm(formula = log(targdol + 1) ~ datead6 + slstyr + slslyr + sls2ago +
    sls3ago + slshist + ordtyr + ordlyr + ord2ago + ord3ago +
    falord + recencymon + ordtyr:ordlyr + ordlyr:ord2ago + slslyr:sls2ago +
    sls2ago:sls3ago, data = multiple_project)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6760 -0.5298 -0.0293  0.4733  3.7380

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.081e+00  1.068e-01  28.847  < 2e-16 ***
datead6          2.958e-05  7.106e-06   4.162 3.21e-05 ***
slstyr           1.878e-03  2.928e-04   6.413 1.57e-10 ***
slslyr           2.409e-03  3.641e-04   6.617 4.08e-11 ***
sls2ago          3.625e-03  4.424e-04   8.196 3.18e-16 ***
sls3ago          1.764e-03  4.285e-04   4.117 3.90e-05 ***
slshist          6.508e-04  1.137e-04   5.721 1.12e-08 ***
ordtyr          -1.069e-01  2.517e-02  -4.245 2.23e-05 ***
ordlyr          -1.543e-01  2.676e-02  -5.766 8.61e-09 ***
ord2ago         -1.611e-01  2.698e-02  -5.970 2.55e-09 ***
ord3ago         -1.053e-01  2.336e-02  -4.507 6.72e-06 ***
falord           7.696e-03  8.198e-03   0.939 0.347882
recencymon       2.032e-04  6.905e-04   0.294 0.768601
ordtyr:ordlyr    5.921e-02  1.685e-02   3.515 0.000444 ***
ordlyr:ord2ago   2.604e-02  1.604e-02   1.624 0.104454
slslyr:sls2ago  -1.441e-05  2.496e-06  -5.771 8.40e-09 ***
sls2ago:sls3ago -7.144e-06  1.565e-06  -4.565 5.13e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7418 on 4746 degrees of freedom
Multiple R-squared:  0.1111, Adjusted R-squared:  0.1081
F-statistic: 37.07 on 16 and 4746 DF,  p-value: < 2.2e-16
```

Appendix 2.2

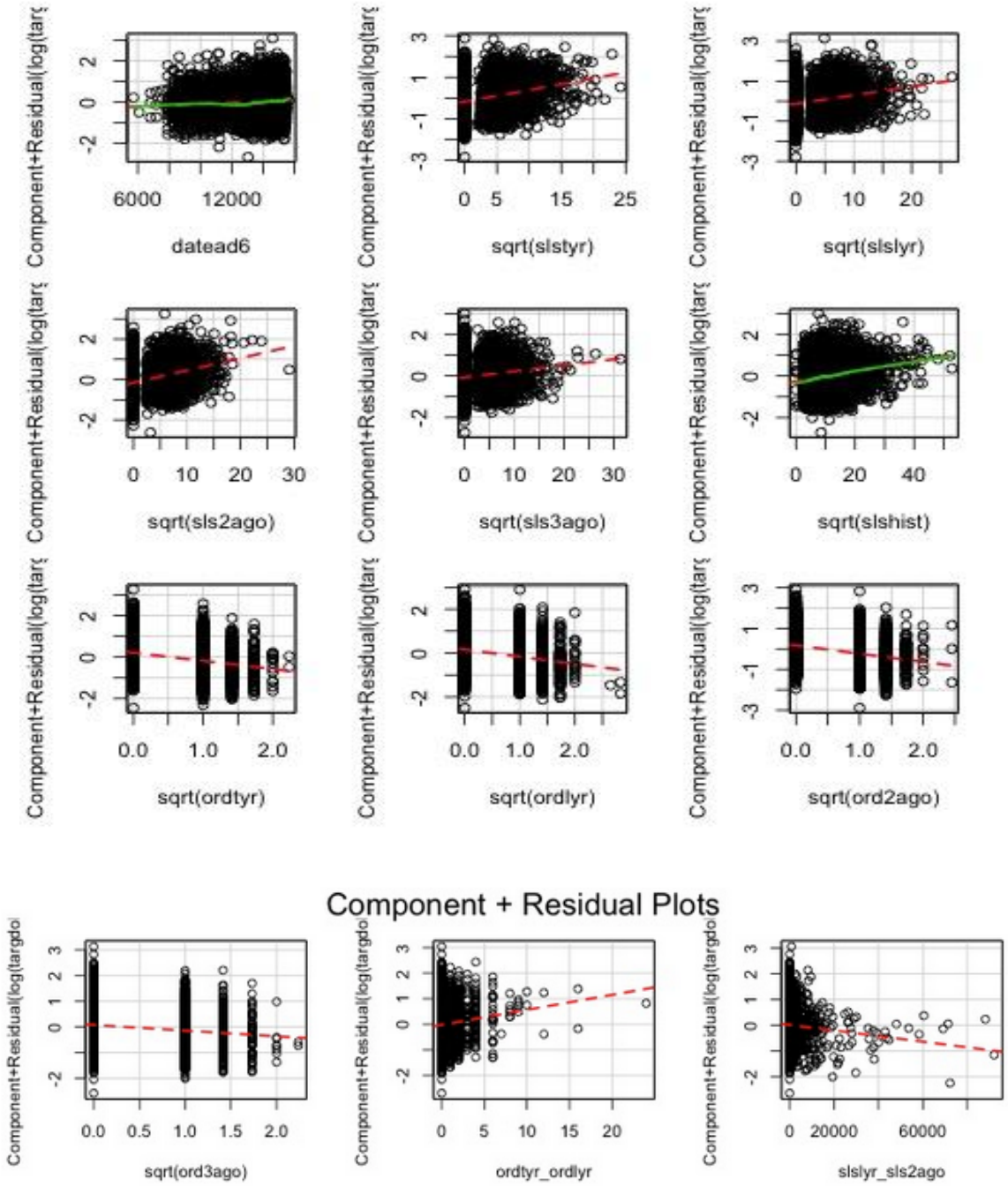
```
Call:
lm(formula = log(targdol + 1) ~ datead6 + sqrt(slstyr) + sqrt(slslyr) +
    sqrt(sls2ago) + sqrt(sls3ago) + sqrt(slshist) + sqrt(ordtyr) +
    sqrt(ordlyr) + sqrt(ord2ago) + sqrt(ord3ago) + ordtyr:ordlyr +
    slslyr:sls2ago + sls2ago:sls3ago, data = multiple_project)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6609 -0.5231 -0.0314  0.4564  3.6338

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.971e+00  1.077e-01  27.590  < 2e-16 ***
datead6          3.020e-05  6.828e-06   4.424 9.92e-06 ***
sqrt(slstyr)     5.272e-02  6.176e-03   8.537  < 2e-16 ***
sqrt(slslyr)     4.273e-02  6.714e-03   6.365 2.14e-10 ***
sqrt(sls2ago)    6.298e-02  7.620e-03   8.265  < 2e-16 ***
sqrt(sls3ago)    3.021e-02  7.535e-03   4.009 6.18e-05 ***
sqrt(slshist)    2.399e-02  3.515e-03   6.826 9.84e-12 ***
sqrt(ordtyr)    -3.733e-01  4.074e-02  -9.165  < 2e-16 ***
sqrt(ordlyr)    -3.303e-01  4.359e-02  -7.578 4.19e-14 ***
sqrt(ord2ago)   -4.135e-01  4.710e-02  -8.780  < 2e-16 ***
sqrt(ord3ago)   -2.273e-01  4.693e-02  -4.843 1.32e-06 ***
ordtyr:ordlyr    6.414e-02  1.332e-02   4.815 1.52e-06 ***
slslyr:sls2ago  -9.155e-06  2.139e-06  -4.280 1.91e-05 ***
sls2ago:sls3ago -1.818e-06  9.374e-07  -1.939   0.0526 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7297 on 4749 degrees of freedom
Multiple R-squared:  0.1393, Adjusted R-squared:  0.137
F-statistic: 59.13 on 13 and 4749 DF,  p-value: < 2.2e-16
```

Appendix 3.1


log(targdol + 1) | others
datead6 | others
sqrt(slstyr) | others
sqrt(slslyr) | others
sqrt(sls2ago) | others
sqrt(sls3ago) | others
sqrt(slshist) | others
sqrt(ordtyr) | others
sqrt(ordlyr) | others
sqrt(ord2ago) | others


Leverage Plots
log(targdol + 1) | others
sqrt(ord3ago) | others
ordtyr:ordlyr | others
slslyr:sls2ago | others

Appendix 3.2


Component + Residual Plots
datead6
sqrt(slstyr)
sqrt(slslyr)
sqrt(sls2ago)
sqrt(sls3ago)
sqrt(slshist)
sqrt(ordtyr)
sqrt(ordlyr)
sqrt(ord2ago)
sqrt(ord3ago)
ordtyr_ordlyr
slslyr_sls2ago