Predicting Purchase Amount in Response to Catalogs

MSIA 401 Project Report

Jill Fan, Lauren Gardiner, Rishabh Joshi, Will Song

I. Executive Summary

To increase the return on investment (ROI) of mailed catalogs, we analyzed sales data from existing customers who received catalogs to identify high value targets for our marketing efforts in the future. Our analysis discovered the following key predictors for determining whether customers would be likely to respond to a catalog with a purchase: consistency of sales in consecutive years, recency of last purchases, the activeness of a customer, sales within the past year, and their average spend per order.

The combination of our final logistic regression and multiple linear regression models resulted in a payoff of $51,465.45. In theory, a perfect model would result in a payoff of $120,252.40 from the test set. While not perfect, our model still has predicting power as seen through its predictions capturing 42.80% of the maximum payoff.

II. Introduction

In order to increase traffic to online e-commerce websites, catalogs are often mailed to existing customers to remind them of a store's offerings. However, only 10% of our customers made a purchase in response to receiving a catalog in the mail. The low response rate prompted us to figure out how to not waste money on ineffective widespread marketing efforts through targeted marketing. Not only do we want to identify customers likely to respond, we also want to identify customers who are likely to respond with a high spend. The Pareto principle tells us that 80% of a company's profits often comes from only 20% of their customers. Therefore, our company should focus our investment on those high value

customers to improve our overall ROI. By identifying likely high value customers in response to receiving a catalog, the mailing list can be reduced. This will minimize the cost of marketing and mailing efforts while still capitalizing on the effect of mailed catalogs.

To complete this analysis, we looked at a dataset of customers who were sent a catalog on September 1, 2012. It included information about a customer's purchase history and whether they made a purchase by December 1, 2012. The data includes 101,532 customers split into a train set of 50,418 customers and a test set of 51,114 customers. The response variable for this analysis is `targdol`, which is the dollar purchase resulting from catalog mailing. Before feature engineering, there were fifteen potential predictors related to customer information existing in the original data set.

The approach for this analysis was split into three steps: data preprocessing, model fitting (Section III), and model validation (Section IV). Data preprocessing included data exploration to better understand the predictors and their relationships with each other, data cleaning to handle errors and inconsistencies, and data manipulation to transform variables and derive new predictors. Since only 10% of customers responded with a purchase, a multiple regression model alone was not sufficient for this analysis. We had to fit the following two models:

1. Binary logistic regression model to predict the probability that a customer will respond with a purchase to a mailed catalog. This gave us $P(targdol > 0)$.
2. Multiple linear regression model to predict the `targdol` for customers that did respond with a purchase. This gave us $E(targdol \mid targdol > 0)$.

The combination of the two model outputs provided the necessary information for the following formula, which calculates the expected `targdol` for each customer:

$$E(y) = E(y|y > 0)P(y > 0) \qquad \text{where y = } \texttt{targdol}$$

By predicting the expected `targdol` or purchase dollars, we were able to select the customers with the top 1,000 predict values and consider them as our high value customers for targeted marketing. We assessed the performance of our model by comparing the expected values to the actual values of `targdol` from the test set, corresponding to the top 1,000 customers according to our model. The performance was measured by both the statistical and financial criterion. While a statistical criterion is always necessary for assessing the validity of a model, the additional financial criterion demonstrates how our model achieves the goal of maximizing the dollar purchases or payoff from sending only 1,000 catalogs.

III. Model Fitting

Data Preprocessing

Before fitting the model, we explored the data to better understand the existing predictor variables and discover data integrity issues. Based on our exploration, we prepared the dataset by adding new predictor variables, managing inconsistencies, and imputing NA values.

First, we looked at the missing values. In order to figure out how to impute the missing values for the year of last purchase (`lpuryear`), we compared it to the full date of the last

purchase (datelp6) and noticed inconsistencies between their year values. This was due to only updating one variable when they both needed to be updated. However, the most recently updated variable was not consistent between data records. We decided to update NA values in lpuryear with the year parsed from datelp6. To handle the inconsistencies within the non NA values, we chose to set lpuryear to the most recent year found in either lpuryear or datelp6.

Another inconsistency issue we found was between the sum of fall orders (falord) and spring orders (sprord), and all orders (ordhist). Again, this issue stemmed from only one variable being updated with no consistency over which one was properly updated. If the total of falord and sprord was greater than ordhist, ordhist must not have been updated properly and the difference was added to ordhist. If the total orders were greater than the sum of the seasonal orders, we needed to update falord or sprord. In order to manage this discrepancy, we needed to determine which season to attribute the additional orders. We utilized datelp6 to determine which season their last purchase took place and updated either falord or sprord with respect to that information.

Since we were modelling only the customers who made a purchase using linear regression, we investigated and found some inconsistencies between this subset of the data set and the entire data set. We noticed that the range of datelp6 and datead6 differed between the subset and the full data set. There were last purchases dating back to 1980s in the full dataset, but the earliest last purchases in the subset were in 2003. There were customers added as early as the 1930s in the full dataset, but the oldest customers in the subset were added in 1985. This indicated that there might be errors in the dates for

customers who did not make a purchase. These customers were not included in our multiple linear regression model, so these potentially erroneous observations only impacted our logistic regression model. Since we could not eliminate such observations from the test set, we decided to not remove those observations from the test set either.

When looking at the histogram of the two date variables with `targdol` > 0 (Figures 1 and 2 in Appendix), we noticed that `datelp6` was not continuous as expected, but occurred every six months probably due to data entry practices. This led us to creating a variable for the season of the last purchase year and bin the data in six month increments called `lpurseason`. If the last purchase occurred later than June, `lpurseason` was fall, otherwise it was spring.

When looking at the correlation between variables (Table 1, R Output 1, Figures 14 and 15 in the Appendix) to see if there was multicollinearity in our candidate predictors, we saw a high correlation between the sales and order history totals and their totals for this year, last year, two years ago, and three years ago. Since the more granular totals provided highly valuable information, we could have simply removed the `ordhist` and `slshist` variable. However, we would have lost some information if we did that. To remove multicollinearity and retain information, we added two more variables called `sls4bfr` and `ord4bfr` (R Output 1), which represented the total sales dollars and number of orders before 4 years. These variables were obtained by subtracting the sum of `slstyr`, `slslyr`, `sls2ago`, and `sls3ago` from `slshist` and similarly for orders. With these new variables, we removed redundant `ordhist` and `slshist` variables at the time of model building.

We also derived some additional predictor variables that we felt useful and important in our analysis. To better capture the activeness of a customer over time, we created the following predictors: `recency`, `pur3yr`, `lifetime`, `active`, and `slscmp`. The variable `recency` represents how long it has been since a customer's last purchase derived from `datelp6`. We also created a boolean variable indicating whether the customer purchased within the last three years, called `pur3yr`. The variable `lifetime` represented the number of days a customer has been a customer derived from `datead6`. To represent the percentage of customers' lifetime that they were actively making purchases, we created a variable `active` by dividing `lifetime` by `recency`. The boolean variable `slscmp` indicates whether the customer spent more than last year. In addition to knowing the activeness of a customer, we also wanted more information about if he/she is a high value customer. The variable `avg_amount` represented the average spend per order for each customer (`slshist/ordhist`). A boolean variable `large_avg` indicated whether the average amount per order was greater than the mean of `avg_amount`. In addition, two more variables were added to aid the process of building our two models: an integer variable `id` to identify each customer uniquely and a boolean variable `responded` to indicate whether the customer has made a purchase or not. `responded` is 1 when `targdol>0` and 0 otherwise.

Logistic Regression

First, we fitted several logistic regression models on the dataset to predict whether a customer would make a purchase or not in response to receiving a catalog in mail. Specifically

our predictions were in the form of probabilities rather than binary classifications. We built four logistic regression models.

Our first logistic model served as a baseline for the logistic regression (R Output 2). For this regression model, we included all the initial predictor variables in our data except `ordhist`, `slshist`, `datelp6`, `datead6`, `pur3yr`, `slscmp`, as these variables are correlated with the sales and orders of the past years, as mentioned in the data preprocessing section. We computed the AIC, CCR (Correct Classification Rate or Accuracy), F1 Score and χ^2 for this model. These statistics are included in Table 3. The higher the CCR, F1 score, and χ^2, the better our model is. The lower the AIC, the better our model is.

After plotting the histograms of the predictor variables, we observed that many of the sales and order history predictor variables were skewed. Log transforming these predictor variables after adding a small constant of 0.0001 to avoid NAs helped with the skewness of the distribution apart from a large number of observations still having the same value (Figures 3 and 4). In addition, in the first logistic regression model, the coefficients of `ord4bfr` was NA, which indicated a perfect correlation with other predictors, which was consistent with our discussion in data preprocessing section. We fit a second logistic regression model with all the skewed variables log transformed except for the predictor variables `large_avg`, `active`, `lifetime`, `recency` and `lpurseason` since these variables did not exhibit skewed distributions (R Output 3). As displayed in the summary of this model and Table 3, the second model gave us a lower AIC, a higher CCR, a higher F-1 score and a higher χ^2. Therefore, we concluded that the second model performs better than our baseline model.

The consistency between sales in previous years might have a correlation with the customer response, therefore, we added interaction terms between sales of last 1, 2 and 3 years to capture consistency. We added these interactions one by one to see if they increased the CCR, the F1-Score and decreased the AIC. We added interactions between `slstyr` and `slslyr`, `slslyr` and `sls2ago`, `sls2ago` and `sls3ago`, and `sls3ago` and `sls4bfr`. We kept all the four interactions terms since they resulted in an increase in the CCR, F1-Score, and χ^2 , and a decrease in AIC. Adding similar interactions between the orders of last 1, 2 or 3 years did not result in such improvements, hence, we did not add them (R Output 4).

Since we achieved significant improvements from the second logistic regression model to the third one, we added more interaction terms into our model including interactions between sales and orders as well as sales between recency, because a customer having more recent and higher sales would be more likely to respond. This logistic model (R Output 5) increased the CCR and χ^2 but decreased the F1 Score. Since it also helped decrease our AIC, we decided to keep it as a candidate for our final logistic model. We noticed that there were some insignificant predictors in our model. So we performed lasso regression on this logistic model to see if there were any other predictor variables that should be removed. The lasso regression did not drop any predictors from the model. Therefore, we decided to move forward with this logistic regression model and continue model diagnostics by computing the AUC value. We plotted an ROC curve for this final regression model, with the AUC value of 0.815, which was the best among all the other models (Figure 5). This model also had the best AIC, CCR and χ^2 values, and the second largest F1 Score.

Multiple Linear Regression

Our initial step in building a multiple linear regression model was to build a baseline model with all nonredundant predictors and no transformations or interactions (R Output 6). It resulted in a baseline R^2 of 0.09509 and a R^2_{adj} of 0.09153. These values, along with the same statistics of all the linear models we fit, are given in Table 4. Next, we wanted to see if any interactions would improve the performance of our model. In order to explore the possible interactions, we screened all possible two-factor interactions and reduced the set through backwards stepwise regression. Even if an interaction was a significant predictor, we did not add it to our model unless it was supported by subject matter knowledge to ensure the explainability and interpretability of our model. The resulting interactions were added to the second iteration of our model and increased the R^2 and R^2_{adj} values (R Output 7). Through data exploration (Figures 3 and 4), we saw that many of the sales and order history data was skewed. To handle the skewness, we performed a log transformation on the predictors and added a small constant to handle the zeroes in the data to build our third model. The helped with the skewness of the predictors apart from many observations still having the same values. This resulted in a decrease in R^2 and R^2_{adj} (R Output 8). Even though our R^2 and R^2_{adj} both decreased, the skewness of the data still justified the transformation. By looking at the normal Q-Q plot (Figure 6) from the third model, we saw that there was curvature in the right side of the plot indicating the need for a log transformation on our response variable targdol. The normal Q-Q plot (Figure 8) for the log transformed model showed a more linear trend. We also checked for homoscedasticity by comparing the residual plot (Figures 7 and 9) and found that

while neither were perfectly randomly spread, the log model was more randomly spread throughout the plot. This fourth model had a higher R^2 and R^2_{adj} (R Output 9). The plots coupled with this increase confirmed our decision to move forward with the log transformed model.

In order to have a more parsimonious model, we iterated through a process of performing backward stepwise regression, removing the most insignificant predictors and performing stepwise regression again. This resulted in our fifth model which had a lower R^2 and R^2_{adj} (R Output 10). When performing the ANOVA (R Output 11), we can see the performance of the models are significantly different through the p-value of 0.02852. However, the drop in performance is outweighed by improving the explainability and parsimony of the model with dropping ten predictors.

The next model diagnostic we checked was the VIFs to see if there was unwanted multicollinearity in our model. We noticed that some of our interactions were causing high VIFs. This is an example of structural multicollinearity. Structural multicollinearity is a multicollinearity that is a mathematical artifact caused by creating new predictors from other predictors, such as, creating the predictor x^2 from the predictor x or multiplying different predictors. The interactions in our model are nothing but products of the corresponding predictors. This is because these interactions terms will be strongly correlated with their original predictors, resulting in very high VIFs. Structural multicollinearity is not something to be concerned about, however, because the *p*-value for the interaction terms is not affected by this multicollinearity. A popular way of handling such multi collinearities is by centering/scaling the predictors before taking the interaction.

By centering those values in our model, we reduced the majority of our VIF scores below the desired threshold of 10. The remaining high VIF scores resulted from including all predictors from an interaction independently even if they are not significant. Removing the independent variables appearing in the interactions reduced all VIFs below 10. However, since they are a part of an interaction we still included them in the model. This is generally a good practice in linear regression. Centering the data also caused two predictors to become non-significant and we removed `falord` and `lifetime` from the model. Even though removing these predictors reduced our R^2 and R^2_{adj}, they had to be removed to ensure the generalizability of our model with future data (R Output 12).

The final model diagnostics we performed was looking at outliers and influential observations. There were twenty-seven observations that would be considered outliers since their standardized residuals from the `rstandard()` function were more than three standard deviations away from the mean . However, twenty-two of those observations are in the top 5% of `targdol`. This indicated that they were high value customers we wanted our model to predict and removing them would hurt the performance of our model. Instead, we removed four outliers that had low `targdol` since we considered them low value customers. To determine which observations were influential, we looked at Cook's distance (Figure 10) and Cook's distance versus leverage (Figure 11). Using the threshold of `4/(n-p-1)` for Cook's Distance, 190 points were considered influential. Since we did not want to weaken our model by removing all influential observations, we only removed three customers with the highest Cook's distance: 433, 2602, and 4672. We looked at the fitted values plot (Figure 13) and the normal Q-Q plot (Figure 12) for the new model and still saw about the same level of random

spread in the fitted values versus residual plot and a roughly linear trend in the normal Q-Q plot. Removing these outliers and influential observations improved our final model giving us a R^2 of 0.141 and a R^2_{adj} of 0.1376 (R Output 13).

IV. Model Validation

To assess the models we fitted, we predicted the `targdol` on our test set of 51,114 customers. The prediction for `targdol` came from the following formula:

$$E(y) = E(y|y > 0)P(y > 0) \qquad \text{where y = } \texttt{targdol}$$

The $E(y|y > 0)$ came from our multiple linear regression model, which predicted an expected value for `targdol` for customers who made a purchase in response to a catalog. Our logistic regression model predicted the probability $P(y > 0)$ of whether they responded with a purchase or not. By multiplying these two values, we calculated the expected `targdol` for every customer in the test set. We then sorted the test set by their expected values of purchase. Once those values were predicted, we selected the customers with the 1000 highest expected `targdol` as our high value customers for targeted marketing. For our top 1000 predicted customers, we compared their expected `targdol` with actual `targdol` and calculated the total payoff (sum of the corresponding actual `targdol` values) to use as the financial criterion. For the multiple regression model we also calculated the mean squared prediction error (MSPE) for those customers whose actual `targdol` was greater than 0 to use as the statistical criterion.

By looking at the statistics of the four logistic regression model we built (Table 3), we decided to go with the fourth logistic regression since it gave us the lowest AIC, highest CCR,

highest χ^2, and a second highest F1- Score. While our model diagnostics led us to choose the fourth logistic regression model and the seventh multiple linear regression model, we still evaluated our other multiple linear regression model iterations using the statistical and financial criterion (Table 5). Even though multiple linear regression model 1 had the lowest MSPE (2496.46) and model 2 had the highest payoff ($53,555,5), both of those models had many insignificant predictors. Not only did they include insignificant predictors, but they both had the largest amount of predictors with 33 explanatory variables. This large number of predictors reduced desired parsimony and explainability in a model and this tradeoff was not worth the 1.74% increase in payoff.

When looking at the MSPE and payoffs for each iteration of the multiple linear regression model, we noticed that there was an increase in MSPE and decrease in payoff once we built the log model. For one last final check, we rebuilt multiple linear regression model 7 without the log transformation on `targdol`. While the payoff increased to $52,729 and the MSPE decreased to 2,554, the residual and normal Q-Q plots (Figures 16 and 17) showed us that there was still a lack of normality if we did not log the response variable. The log transformation stabilized the variance, which will make it easier for the model to predict in the future. Therefore, we chose the seventh model as our final multiple linear regression model. For this model, the MSPE was 2,710.869 and the total payoff was $51,465.45.

V. Conclusion

Our final logistic regression (4th logistic model) and multiple linear regression (7th linear regression model) resulted in a payoff of $51,465.45. By looking at the top 1,000

targdol in the test set, we determined the highest possible payoff is $120,252.41 indicating that our prediction achieved 42.80% of the potential maximum payoff.

Our logistic regression model included a mixture of numeric, categorical and boolean variables. Referring to the summary for our logistic model, the following predictors had the smallest p-values below 2e-16: log(sls4bfr+K), log(ord4bfr+K), active, log(slslyr+K):log(sls2ago+K), log(sls2ago+K):log(sls3ago+K), and log(slslyr+K):recency. These predictors were considered to be statistically significant. Among them, log(sls4bfr+K), log(ord4bfr+K), and active captured the historical dollar purchase and orders 4 years ago, and demonstrated the historical activity of the customers. We found that the two interactions log(slslyr+K):log(sls2ago+K) and log(sls2ago+K):log(sls3ago+K) captured how consistent a customer was in terms of making purchases over the past three years. The predictor log(slslyr+K):recency captured the interactive relationship between a customer's purchase amount of last year and how recent he/she has made the latest purchase.

Our final multiple linear regression model includes nineteen predictors. Referring to the summary for our multiple linear regression model, the following predictors had the smallest p-values: scale(log(slstyr+K)), active:scale(log(avg_amount+K)), scale(log(slstyr+K)):scale(log(avg_amount+K)), scale(log(avg_amount+K)), and scale(log(avg_amount+K)):pur3yr1. The centered log transformation of last year's sales (scale(log(slstyr+K))) was easily explainable because if a customer purchased a lot last year, they are probably more likely to be a return

customer this year. Additionally, the effect of `avg_amount`, both independently and within interactions, seemed reasonable. One would expect the average amount a customer average spent in the past should indicate the purchase dollars of a customer in the future. This average spend combined with their activeness and recent activity spoke to a customer's value in the future.

While the models we built were the best given the data provided, additional predictors could provide information that would improve the model. Our dataset only provided us with information about the customers had to do with their purchase history. Additional demographic information, like gender, age, ethnicity, and income could tell us more about the customer and in turn, their behavior. While demographic information is sometimes hard to obtain unless customers volunteer it, we already have geographic data on the customer based on the catalog mailing address. Not only would a customer's location provide potentially useful information, but we could also cross reference this information with Census data to obtain income data for an area. Ths additional financial information should have predicting power for a consumer's `targdol`.

In conclusion, our logistic and linear regression models identified several significant predictor variables about the expected dollar purchase for customers in response to receiving catalogs. Key predictors included consistency of sales in consecutive years, recency of last purchases, the activeness of a customer, sales within the past year, and their average spend per order. We believe that our models can identify future target customers to send catalogs in order to maximize the payoff of mailing catalogs.

VI. Appendix

Tables

Table 1: Highest Correlation Between Predictors

Var1 <fctr>	Var2 <fctr>	value <dbl>
falord	ordhist	0.8848995
ord4bfr	ordhist	0.8802502
sls4bfr	slshist	0.8371075
active	lifetime	0.8252264
ord4bfr	falord	0.7711792
ord4bfr	sls4bfr	0.7344765
ord2ago	sls2ago	0.6971758
ordhist	slshist	0.6943522
sprord	ordhist	0.6910660
sls4bfr	ordhist	0.6655982
ordlyr	slslyr	0.6615865
falord	slshist	0.6518204
active	ordhist	0.6411321
sls4bfr	falord	0.6336376
ord4bfr	sprord	0.6203430
ord4bfr	active	0.6127549
ord4bfr	lifetime	0.6015377
ord4bfr	slshist	0.5931184
active	falord	0.5930456
ordtyr	slstyr	0.5717693
avg_amount	slstyr	0.5464207
slshist	sls3ago	0.5457004
lifetime	falord	0.5437675
slshist	sls2ago	0.5390541
ord3ago	sls3ago	0.5361029
ordhist	ord3ago	0.5353875
lifetime	ordhist	0.5296377
ordhist	ord2ago	0.5130801

Table 2: Additional Predictors

Variable Name	Rationale
recency	Difference between the date of last purchase and Dec 1, 2012 in days. A lower value of recency means a more recent purchase by the customer.
pur3yr	A boolean variable indicating whether the customer purchased within the last 3 years (between 2010 and 2012).
lifetime	Difference between the date of customer added to file and Dec 1, 2012 in days.
active	Proportion of the lifetime of a customer in which they were active, active = (lifetime - recency)/lifetime
slscmp	A boolean variable indicating whether the customer spent more money this year as compared to last year.
avg_amount	A numeric variable indicating the average amount spend per order, avg_amount = slshist / ordhist
large_avg	A boolean variable whether the avg_mount is greater than the mean of avg_amount of all customers in the database.
id	A unique id for each customer
responded	A boolean variable indicating whether the customer responded or not, i.e. whether their corresponding targdol>0 or not.
sls4bfr	A numeric variable indicating the total sales dollars from 4 years ago, sls4bfr = slshist-slstyr-slslyr-sls2ago-sls3ago
ord4bfr	A numeric variable indicating the total number of orders from 4 years ago ord4bfr = ordhist-ordtyr-ordlyr-ord2ago-ord3ago

Table 3: Logistic Regression Model Comparison

Model Iteration	Description	Number of Predictors	AIC	CCR	F1 Score	χ^2
1	Baseline	18	25996	0.914	0.303	5947
2	Log transformed predictors	18	24332	0.924	0.390	7613
3	Added interactions between sales	22	24080	0.928	0.448	7873
4	Added more interactions	29	23873	0.929	0.444	8094

Table 4: Multiple Linear Regression Model Comparison

Model Iteration	Description	Number of Predictors	R^2	R^2_{adj}
1	Baseline	19	0.09509	0.09153
2	Added interactions	31	0.1086	0.1029
3	Log transformed predictors	31	0.0904	0.08454
4	Log transformed response	31	0.1395	0.134
5	Iterative backward stepwise	21	0.1359	0.1322
6	Centered predictors	19	0.1349	0.1314
7	Removed outliers and influential observations	19	0.141	0.1376

Table 5: Multiple Linear Regression Model Validation

Model Iteration	MSPE	Payoff
1	2,496.46	$52,715.35
2	2,598.30	$53,555.55
3	2,565.06	$51,733.86
4	2,731.60	$52,037.34
5	2,719.25	$51,868.61
6	2,715.68	$51,421.39
7	2,710.87	$51,465.45

Figures

Figure 1. Histogram of `datelp6`

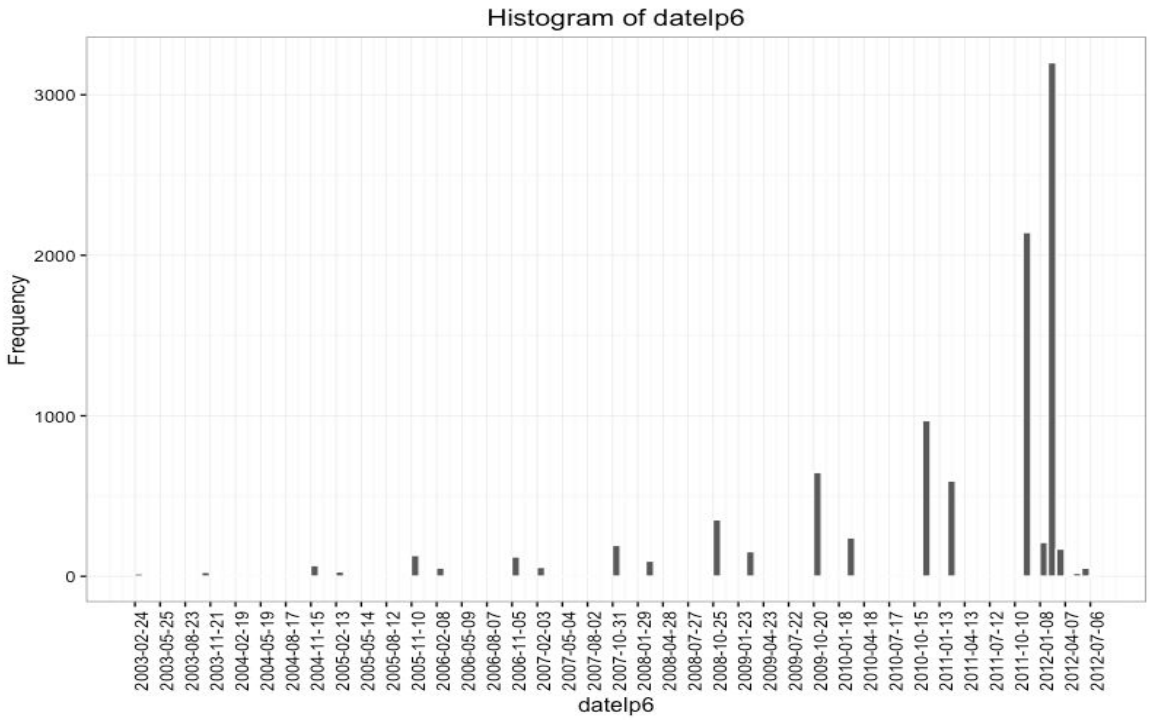


Figure 2. Histogram of `datead6`

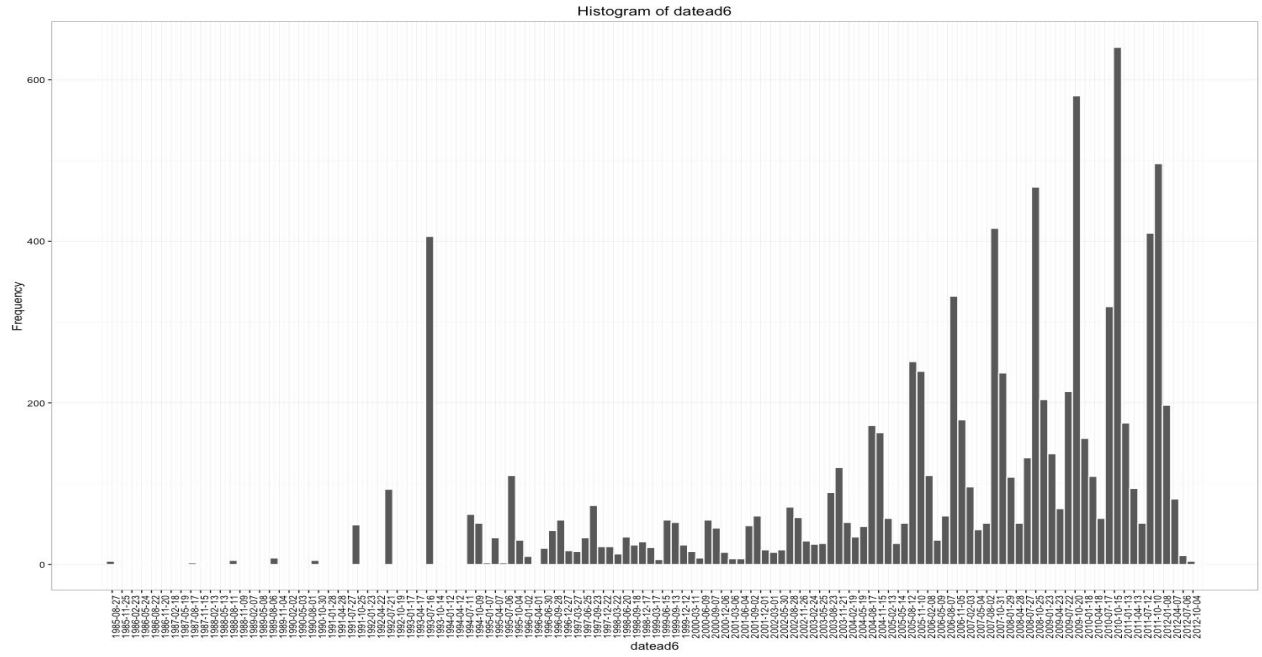


Figure 3. Histogram of Original Predictors

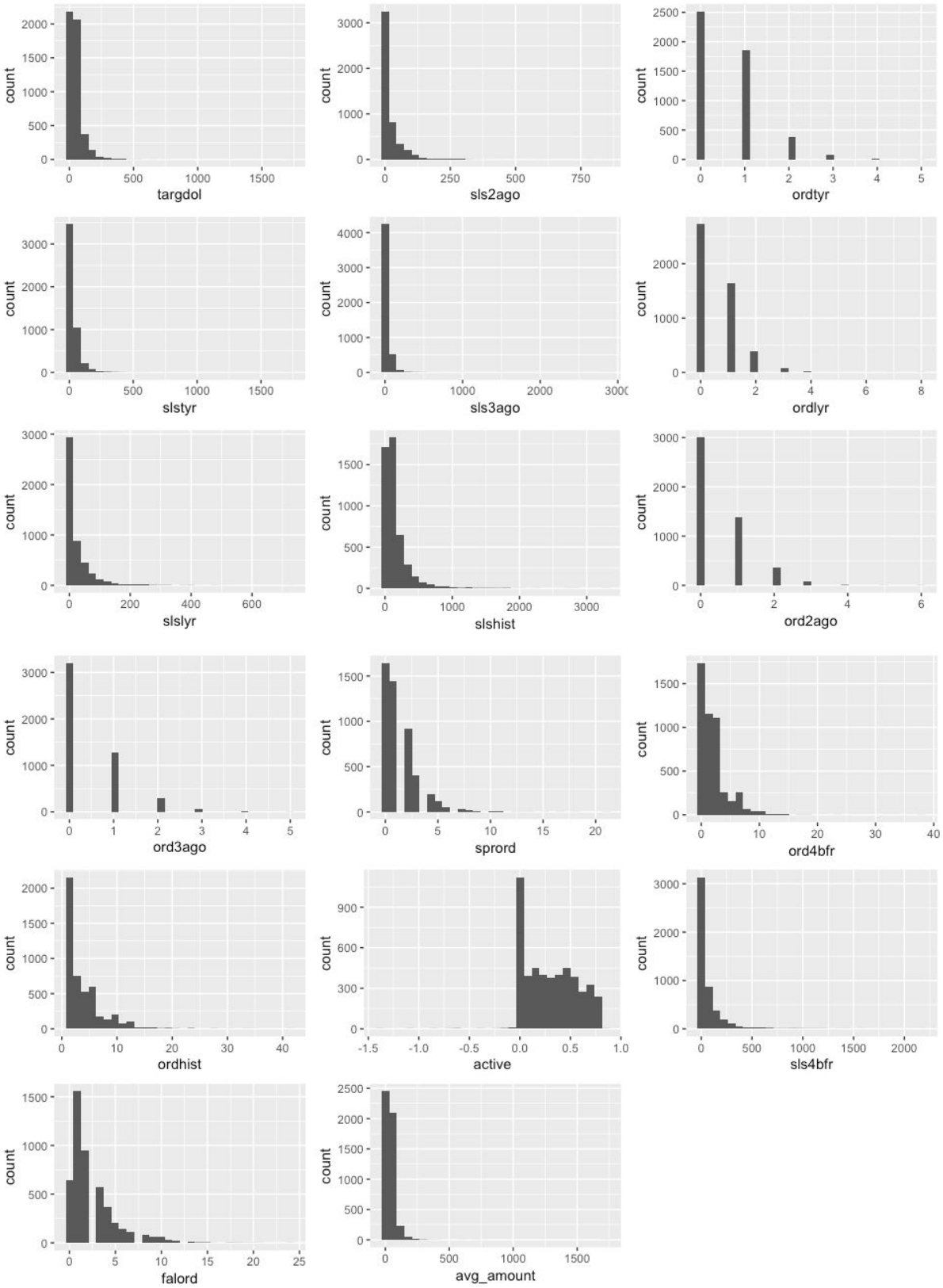


Figure 4. Histogram of log transformed Predictors

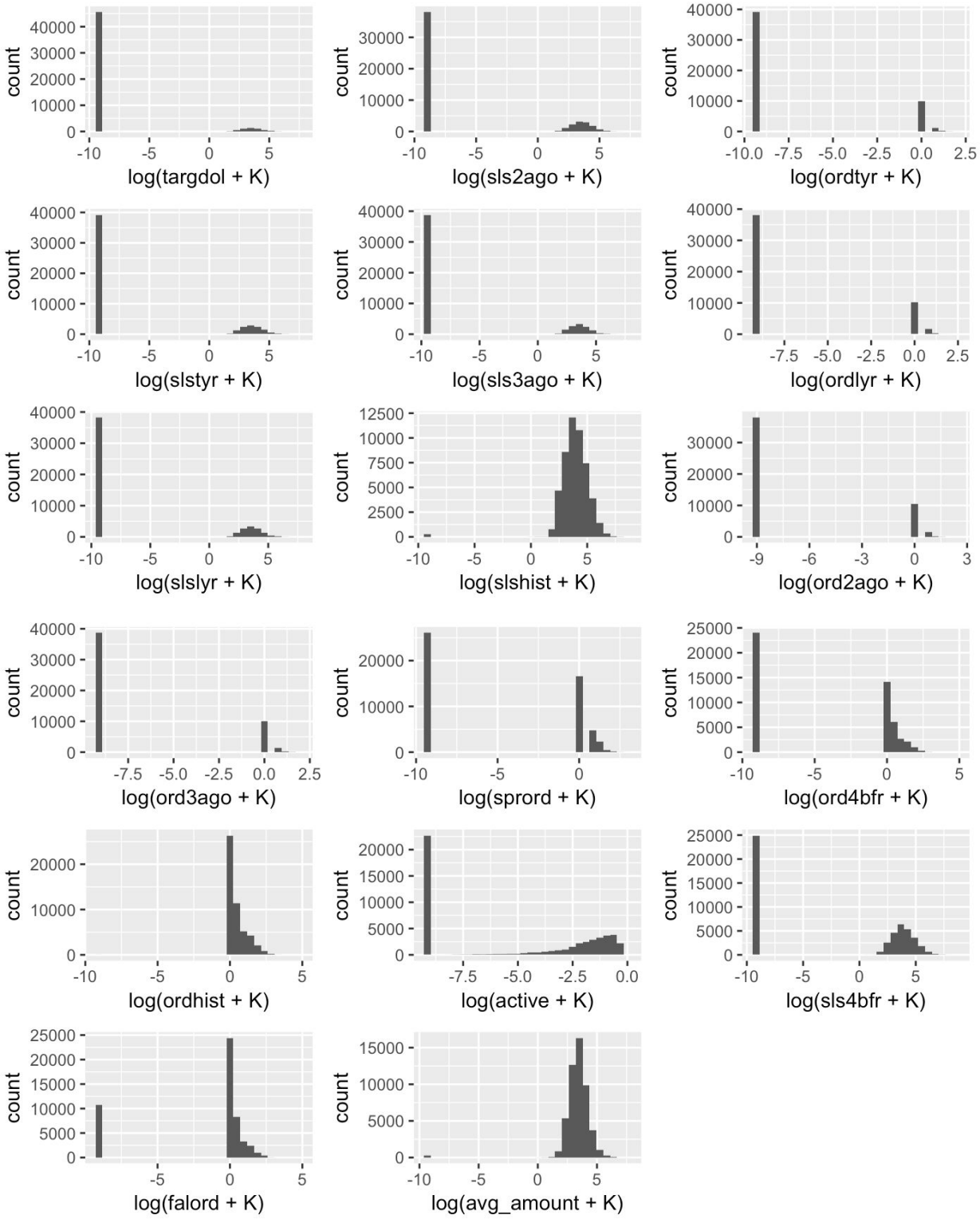


Figure 5. ROC Curve of Logistic Regression Model

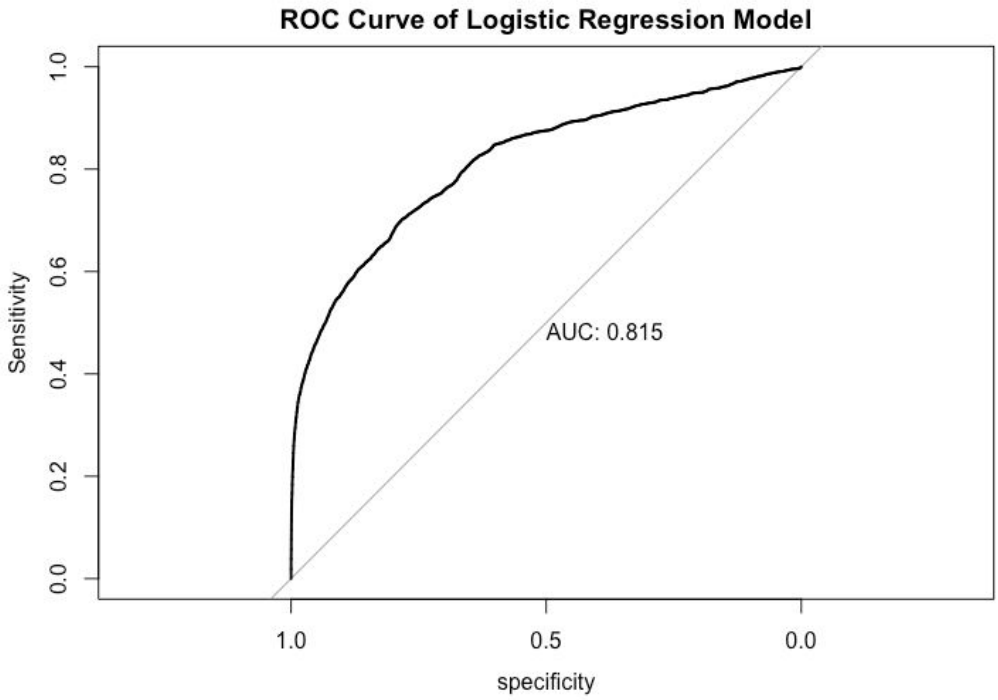


Figure 6. Normal Q-Q Plot for Multiple Linear Regression Model

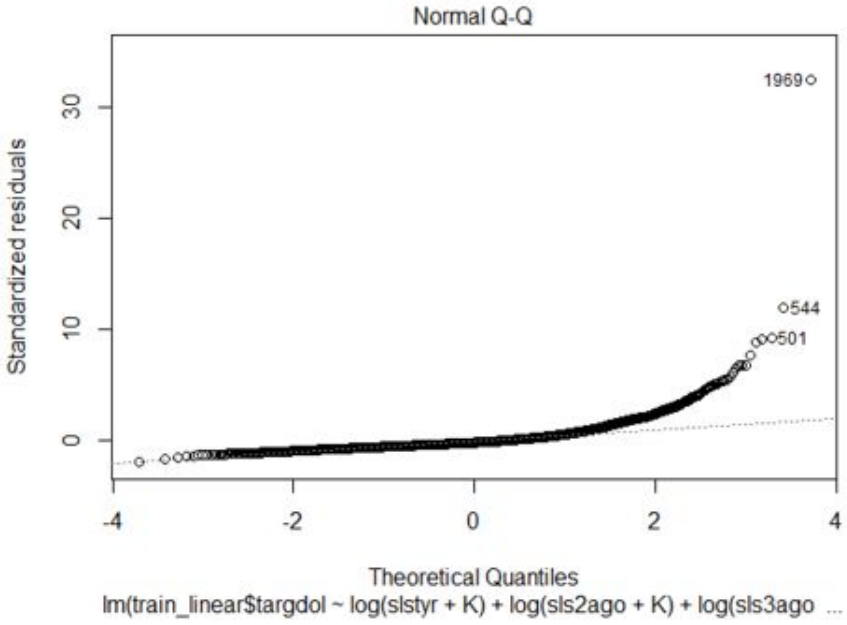


Figure 7. Fitted Residual Plot for Multiple Linear Regression Model

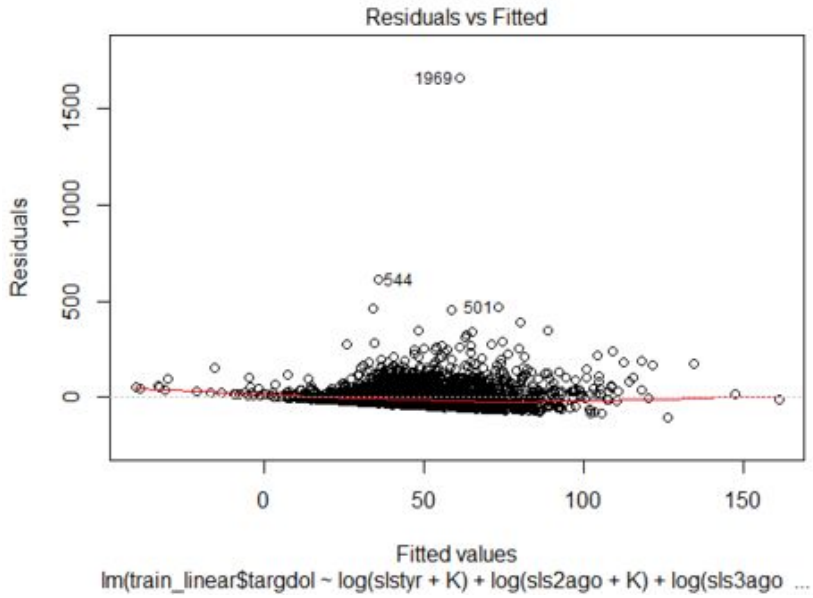


Figure 8. Normal Q-Q Plot for Log Multiple Linear Regression Model

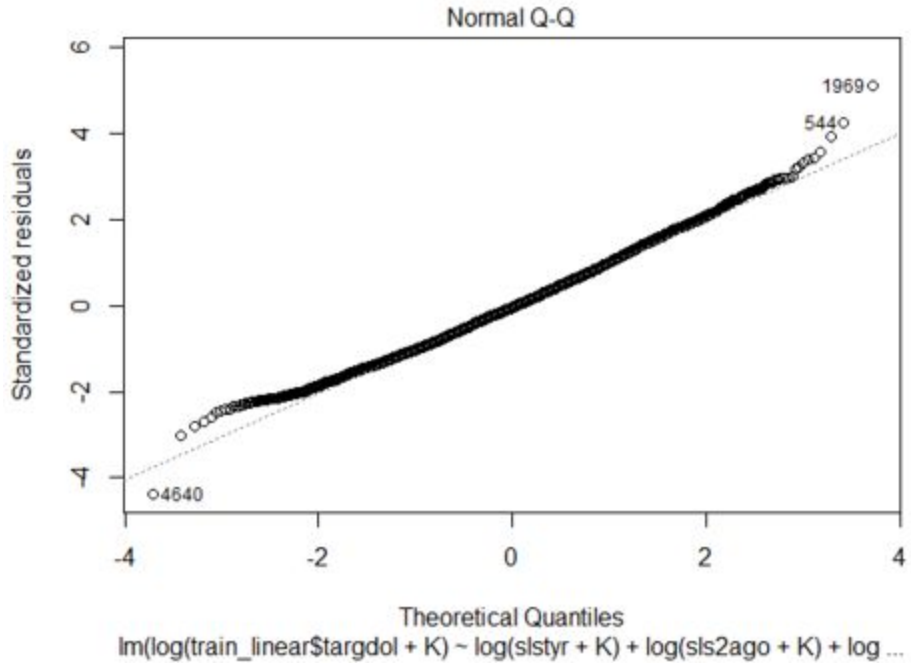


Figure 9. Fitted Residual Plot for Log Multiple Linear Regression Model

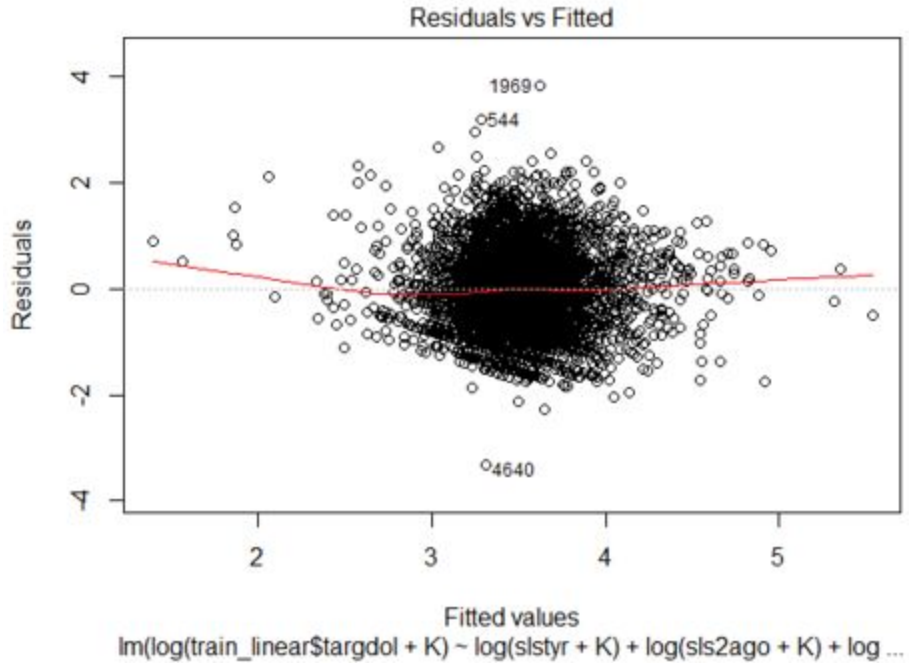


Figure 10. Cook's Distance Log Multiple Linear Regression Model

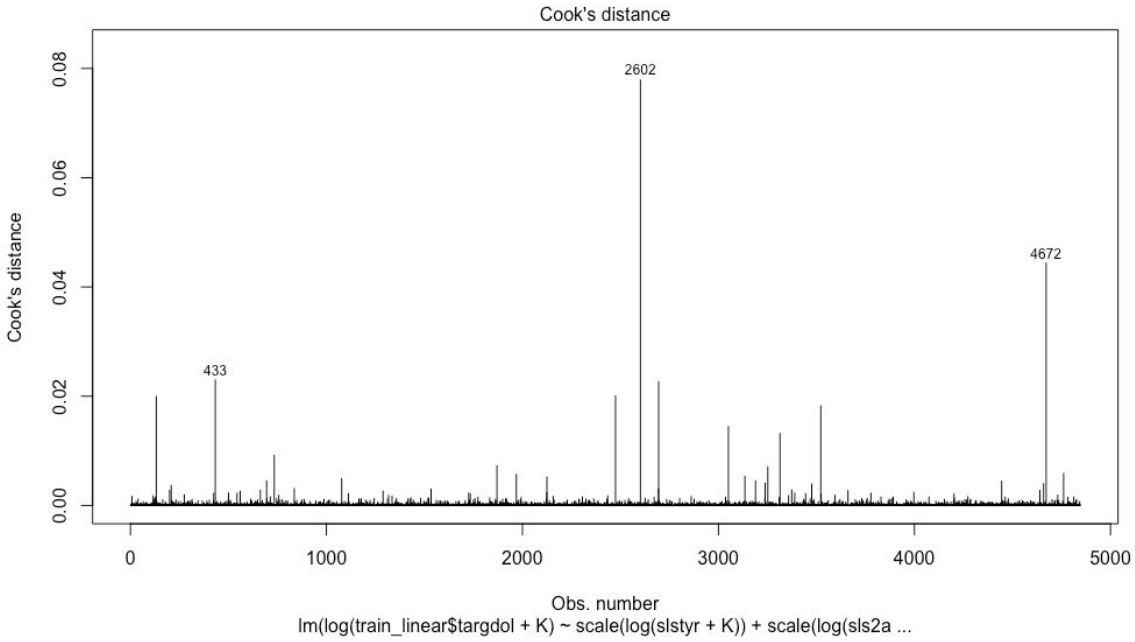


Figure 11. Cook's Distance versus Leverage Log Multiple Linear Regression Model

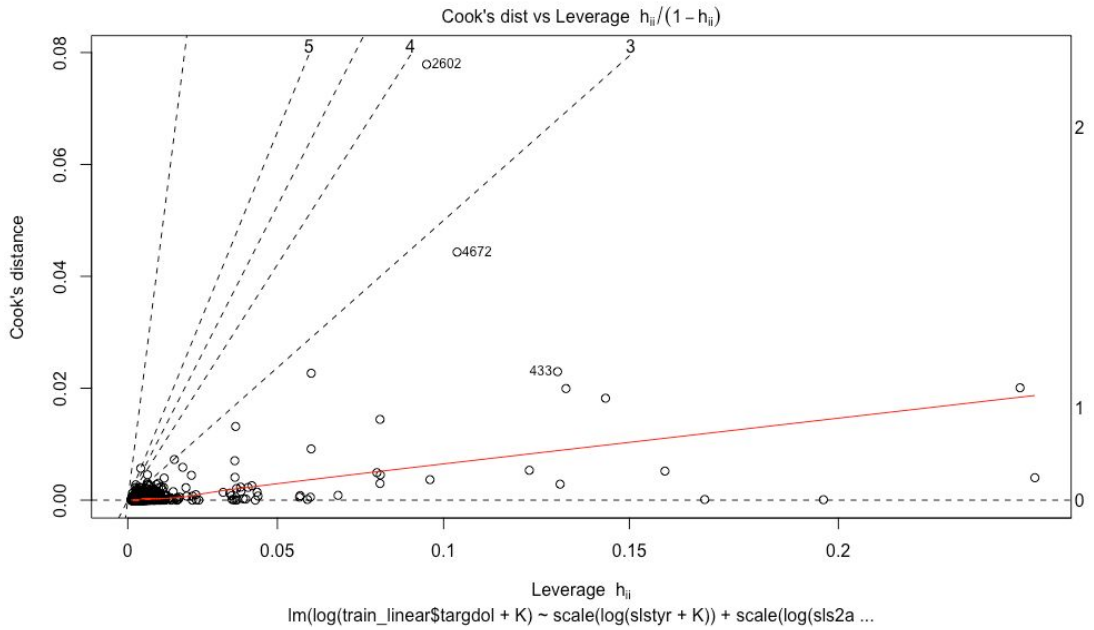


Figure 12. Normal Q-Q Plot for Final Multiple Linear Regression Model

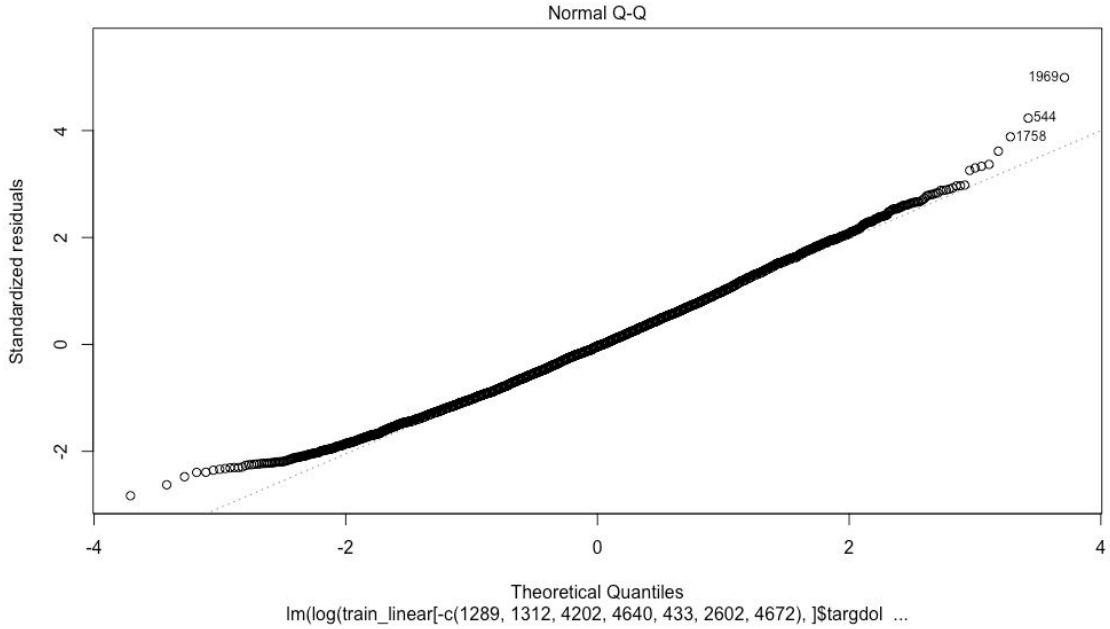


Figure 13. Fitted Residual Plot for Final Multiple Linear Regression Model

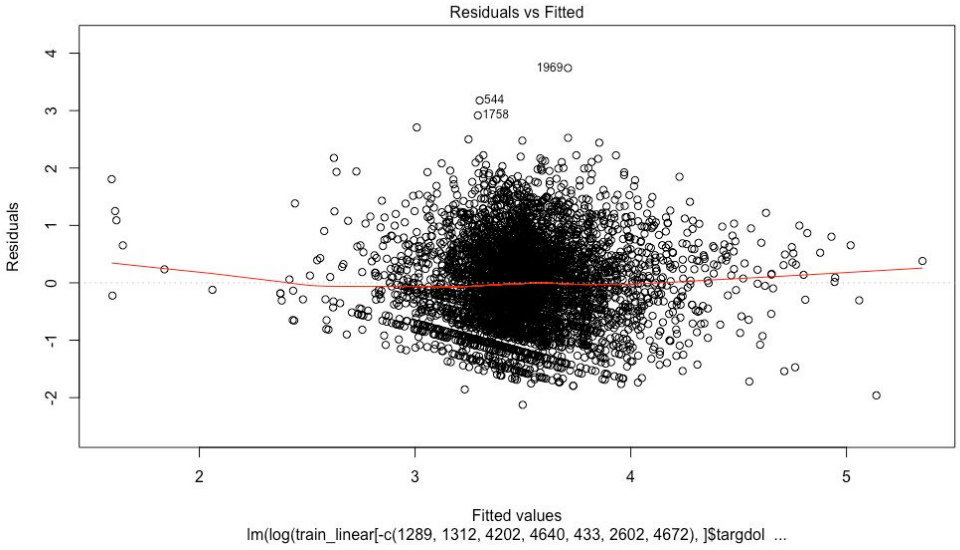


Figure 14. Scatter Plot of `ordhist` vs. `ordtyr` + `ordlyr` + `ord2go` + `ord3ago`

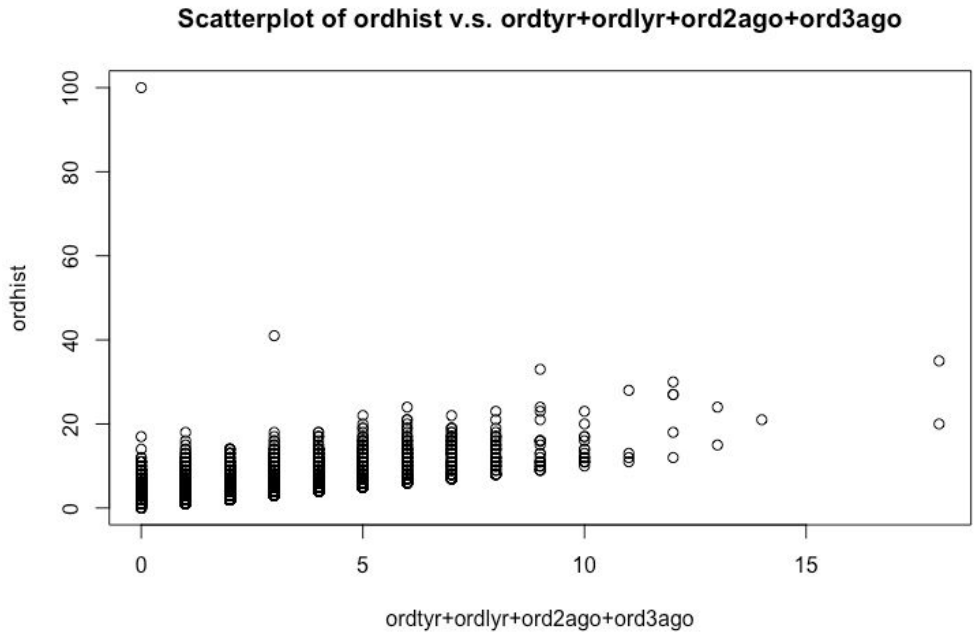


Figure 15. Scatter Plot of `slshist` vs. `slstyr` + `slslyr` + `sls2ago` + `sls3ago`

Figure 16. Normal Q-Q Plot for Final Multiple Linear Regression Model without Log Transformation

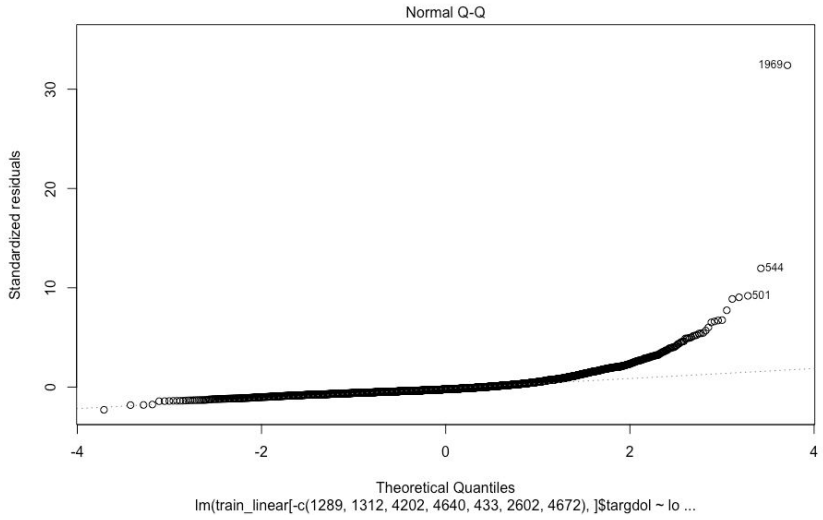
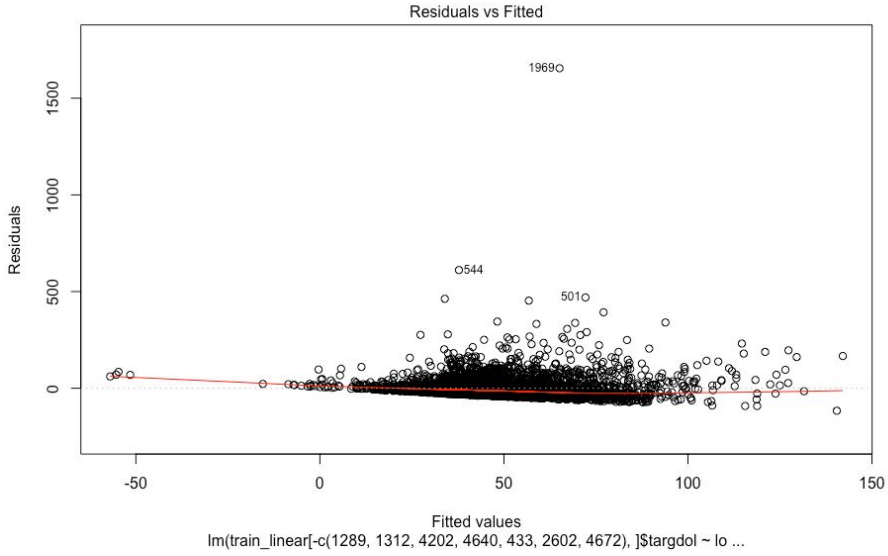


Figure 17. Fitted Residual Plot for Final Multiple Linear Regression Model without Log Transformation



R Outputs

R Output 1:

Correlation between the sum of sales for the past years, sales history, and sales before 4 years. The same correlations for orders.

```
>
> with(train, cor(slstyr + slslyr + sls2ago + sls3ago, slshist))
[1] 0.7075605
> with(train, cor(ordtyr + ordlyr + ord2ago + ord3ago, ordhist))
[1] 0.6135407
> with(train, cor(slstyr + slslyr + sls2ago + sls3ago, sls4bfr))
[1] 0.1286372
> with(train, cor(ordtyr + ordlyr + ord2ago + ord3ago, ord4bfr))
[1] 0.06160577
```

R Output 2:

Logistic Regression Model 1 Output

```
Call:
glm(formula = responded ~ . - ordhist - slshist - recency_bin -
    pur3yr - slscmp, family = binomial, data = select(train_logistic,
    -id, -datelp6, -datead6))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.5605  -0.4414  -0.2965  -0.1709   3.9878

Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.510e-01  1.364e-01   2.573  0.01007 *
slstyr            2.302e-03  5.517e-04   4.171 3.03e-05 ***
slslyr            1.547e-03  5.478e-04   2.824  0.00475 **
sls2ago           1.168e-03  6.175e-04   1.891  0.05862 .
sls3ago           1.293e-03  4.289e-04   3.015  0.00257 **
ordtyr           -2.007e-01  4.083e-02  -4.916 8.83e-07 ***
ordlyr           -1.622e-01  3.600e-02  -4.504 6.66e-06 ***
ord2ago          -1.834e-01  3.883e-02  -4.725 2.30e-06 ***
ord3ago          -2.011e-01  3.746e-02  -5.370 7.88e-08 ***
falord            3.325e-01  1.740e-02  19.113  < 2e-16 ***
sprord            1.175e-01  1.964e-02   5.985 2.17e-09 ***
lpurseasonspring  6.446e-01  4.130e-02  15.608  < 2e-16 ***
recency          -9.578e-04  5.281e-05 -18.136  < 2e-16 ***
lifetime         -2.704e-04  2.558e-05 -10.572  < 2e-16 ***
active            2.803e+00  1.994e-01  14.057  < 2e-16 ***
avg_amount       -3.917e-03  8.806e-04  -4.448 8.66e-06 ***
large_avg1        3.273e-03  4.877e-02   0.067  0.94650
sls4bfr          -1.887e-03  2.758e-04  -6.844 7.72e-12 ***
ord4bfr                  NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31907  on 50417  degrees of freedom
Residual deviance: 25960  on 50400  degrees of freedom
AIC: 25996

Number of Fisher Scoring iterations: 6
```

R Output 3:

Logistic Regression Model 2 Output

```
Call:
glm(formula = responded ~ I(log(slstyr + K)) + I(log(slslyr +
    K)) + I(log(sls2ago + K)) + I(log(sls3ago + K)) + I(log(sls4bfr +
    K)) + I(log(ordtyr + K)) + I(log(ordlyr + K)) + I(log(ord2ago +
    K)) + I(log(ord3ago + K)) + I(log(ord4bfr + K)) + I(log(sprord +
    K)) + I(log(falord + K)) + I(log(avg_amount + K)) + large_avg +
    active + lifetime + recency + lpurseason, family = binomial,
    data = select(train_logistic, -id))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7424  -0.4138  -0.2833  -0.1890   3.3989

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.745e+00  2.345e-01   7.438 1.02e-13 ***
I(log(slstyr + K))     1.185e-02  1.961e-02   0.604  0.54566
I(log(slslyr + K))    -1.090e-02  2.500e-02  -0.436  0.66275
I(log(sls2ago + K))   -5.756e-02  2.549e-02  -2.259  0.02390 *
I(log(sls3ago + K))    3.883e-03  2.437e-02   0.159  0.87339
I(log(sls4bfr + K))   -2.990e-01  1.125e-02 -26.578  < 2e-16 ***
I(log(ordtyr + K))     1.718e-02  2.716e-02   0.633  0.52706
I(log(ordlyr + K))     5.316e-02  3.431e-02   1.550  0.12124
I(log(ord2ago + K))    1.051e-01  3.497e-02   3.004  0.00266 **
I(log(ord3ago + K))    2.359e-02  3.339e-02   0.707  0.47977
I(log(ord4bfr + K))    4.637e-01  1.466e-02  31.632  < 2e-16 ***
I(log(sprord + K))    -3.532e-02  5.825e-03  -6.064 1.33e-09 ***
I(log(falord + K))     5.798e-02  6.758e-03   8.578  < 2e-16 ***
I(log(avg_amount + K)) 1.411e-01  2.258e-02   6.250 4.10e-10 ***
large_avg1             4.908e-03  4.547e-02   0.108  0.91405
active                 1.507e+00  2.666e-01   5.655 1.56e-08 ***
lifetime              -7.558e-05  2.652e-05  -2.850  0.00437 **
recency               -9.790e-04  6.742e-05 -14.522  < 2e-16 ***
lpurseasonspring       5.747e-01  5.168e-02  11.120  < 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: 31907  on 50417  degrees of freedom
Residual deviance: 24294  on 50399  degrees of freedom
AIC: 24332

Number of Fisher Scoring iterations: 6
```

R Output 4:

Logistic Regression Model 3 Output

```
Call:
glm(formula = responded ~ I(log(slstyr + K)) + I(log(slslyr +
    K)) + I(log(sls2ago + K)) + I(log(sls3ago + K)) + I(log(sls4bfr +
    K)) + I(log(ordtyr + K)) + I(log(ordlyr + K)) + I(log(ord2ago +
    K)) + I(log(ord3ago + K)) + I(log(ord4bfr + K)) + I(log(sprord +
    K)) + I(log(falord + K)) + I(log(avg_amount + K)) + large_avg +
    active + lifetime + recency + lpurseason + I(log(slstyr +
    K)):I(log(slslyr + K)) + I(log(slslyr + K)):I(log(sls2ago +
    K)) + I(log(sls2ago + K)):I(log(sls3ago + K)) + I(log(sls3ago +
    K)):I(log(sls4bfr + K)), family = binomial, data = select(train_logistic,
    -id))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7984  -0.4161  -0.2878  -0.1899   4.0007


Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            2.318e+00  2.611e-01   8.878  < 2e-16 ***
I(log(slstyr + K))                     1.874e-02  1.975e-02   0.949  0.34275
I(log(slslyr + K))                     7.756e-03  2.478e-02   0.313  0.75431
I(log(sls2ago + K))                   -3.939e-02  2.502e-02  -1.575  0.11531
I(log(sls3ago + K))                    6.465e-03  2.373e-02   0.272  0.78525
I(log(sls4bfr + K))                   -3.055e-01  1.148e-02 -26.618  < 2e-16 ***
I(log(ordtyr + K))                    -1.502e-02  2.731e-02  -0.550  0.58244
I(log(ordlyr + K))                     6.207e-02  3.388e-02   1.832  0.06694 .
I(log(ord2ago + K))                    1.102e-01  3.422e-02   3.219  0.00129 **
I(log(ord3ago + K))                    2.819e-02  3.244e-02   0.869  0.38482
I(log(ord4bfr + K))                    4.556e-01  1.480e-02  30.775  < 2e-16 ***
I(log(sprord + K))                    -3.198e-02  5.891e-03  -5.430 5.65e-08 ***
I(log(falord + K))                     5.944e-02  6.799e-03   8.742  < 2e-16 ***
I(log(avg_amount + K))                 1.829e-01  2.377e-02   7.696 1.41e-14 ***
large_avg1                             1.623e-02  4.597e-02   0.353  0.72403
active                                 2.410e+00  3.083e-01   7.817 5.42e-15 ***
lifetime                              -1.439e-04  2.975e-05  -4.836 1.33e-06 ***
recency                               -1.320e-03  8.284e-05 -15.932  < 2e-16 ***
lpurseasonspring                       4.613e-01  5.299e-02   8.705  < 2e-16 ***
I(log(slstyr + K)):I(log(slslyr + K))  6.925e-03  5.546e-04  12.487  < 2e-16 ***
I(log(slslyr + K)):I(log(sls2ago + K)) 4.460e-03  4.954e-04   9.002  < 2e-16 ***
I(log(sls2ago + K)):I(log(sls3ago + K)) 4.651e-03  5.012e-04   9.281  < 2e-16 ***
I(log(sls3ago + K)):I(log(sls4bfr + K)) 1.234e-03  5.235e-04   2.357  0.01845 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31907  on 50417  degrees of freedom
Residual deviance: 24034  on 50395  degrees of freedom
AIC: 24080

Number of Fisher Scoring iterations: 6
```

R Output 5:
Logistic Regression Model 4 Output

```
Call:
glm(formula = responded ~ I(log(slstyr + K)) + I(log(slslyr +
    K)) + I(log(sls2ago + K)) + I(log(sls3ago + K)) + I(log(sls4bfr +
    K)) + I(log(ordtyr + K)) + I(log(ordlyr + K)) + I(log(ord2ago +
    K)) + I(log(ord3ago + K)) + I(log(ord4bfr + K)) + I(log(sprord +
    K)) + I(log(falord + K)) + I(log(avg_amount + K)) + large_avg +
    active + lifetime + recency + lpurseason + I(log(slstyr +
    K)):I(log(slslyr + K)) + I(log(slslyr + K)):I(log(sls2ago +
    K)) + I(log(sls2ago + K)):I(log(sls3ago + K)) + I(log(sls3ago +
    K)):I(log(sls4bfr + K)) + recency:active + lifetime:slscmp +
    I(log(slstyr + K)):I(log(ordtyr + K)) + I(log(slstyr + K)):recency +
    I(log(slslyr + K)):recency + I(log(sls2ago + K)):I(log(ord2ago +
    K)) + I(log(avg_amount + K)):recency, family = binomial,
    data = select(train_logistic, -id))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3882  -0.3978  -0.2934  -0.2096   4.8638


Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          3.811e+00  6.171e-01   6.176 6.58e-10 ***
I(log(slstyr + K))                  -2.307e-01  4.737e-02  -4.870 1.12e-06 ***
I(log(slslyr + K))                   4.495e-01  5.619e-02   8.000 1.24e-15 ***
I(log(sls2ago + K))                 -2.354e-02  2.632e-02  -0.894 0.371061
I(log(sls3ago + K))                  1.030e-02  2.476e-02   0.416 0.677560
I(log(sls4bfr + K))                 -3.091e-01  1.180e-02 -26.201  < 2e-16 ***
I(log(ordtyr + K))                   7.193e-02  7.218e-02   0.996 0.319022
I(log(ordlyr + K))                   5.666e-03  3.623e-02   0.156 0.875751
I(log(ord2ago + K))                  3.197e-01  6.952e-02   4.599 4.25e-06 ***
I(log(ord3ago + K))                  2.298e-02  3.386e-02   0.679 0.497309
I(log(ord4bfr + K))                  4.487e-01  1.510e-02  29.727  < 2e-16 ***
I(log(sprord + K))                  -2.941e-02  6.045e-03  -4.866 1.14e-06 ***
I(log(falord + K))                   5.338e-02  6.992e-03   7.634 2.28e-14 ***
I(log(avg_amount + K))              -4.382e-01  1.019e-01  -4.301 1.70e-05 ***
large_avg1                           7.833e-02  4.764e-02   1.644 0.100117
active                               6.414e+00  5.766e-01  11.125  < 2e-16 ***
lifetime                            -1.693e-04  3.082e-05  -5.491 3.99e-08 ***
recency                             -2.021e-03  2.522e-04  -8.014 1.11e-15 ***
lpurseasonspring                     3.563e-01  5.575e-02   6.391 1.64e-10 ***
I(log(slstyr + K)):I(log(slslyr + K))  1.056e-03  8.192e-04   1.289 0.197225
I(log(slslyr + K)):I(log(sls2ago + K))  4.184e-03  5.073e-04   8.248  < 2e-16 ***
I(log(sls2ago + K)):I(log(sls3ago + K))  4.781e-03  5.123e-04   9.332  < 2e-16 ***
I(log(sls3ago + K)):I(log(sls4bfr + K))  2.598e-03  5.534e-04   4.695 2.67e-06 ***
active:recency                      -1.310e-03  1.656e-04  -7.910 2.58e-15 ***
lifetime:slscmp1                    -4.249e-05  1.563e-05  -2.718 0.006567 **
I(log(slstyr + K)):I(log(ordtyr + K))  2.236e-02  7.336e-03   3.048 0.002302 **
I(log(slstyr + K)):recency           1.339e-04  1.816e-05   7.375 1.64e-13 ***
I(log(slslyr + K)):recency          -1.805e-04  2.110e-05  -8.557  < 2e-16 ***
I(log(sls2ago + K)):I(log(ord2ago + K))  2.599e-02  6.835e-03   3.803 0.000143 ***
I(log(avg_amount + K)):recency       1.960e-04  3.233e-05   6.064 1.32e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31907  on 50417  degrees of freedom
Residual deviance: 23813  on 50388  degrees of freedom
AIC: 23873

Number of Fisher Scoring iterations: 7
```

R Output 6:

Linear Regression Model 1 Output

```
Call:
lm(formula = targdol ~ ., data = select(train_linear, -id, -slshist,
    -ordhist, -ord4bfr))

Residuals:
    Min      1Q  Median      3Q     Max
-190.46  -24.68  -11.84    9.01 1657.55

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.286e+01  9.124e+00   4.698 2.70e-06 ***
slstyr            1.692e-01  2.420e-02   6.988 3.16e-12 ***
slslyr            1.154e-01  2.319e-02   4.976 6.72e-07 ***
sls2ago           1.705e-01  2.473e-02   6.892 6.21e-12 ***
sls3ago           4.734e-02  1.569e-02   3.018  0.00256 **
ordtyr           -4.145e+00  1.838e+00  -2.255  0.02419 *
ordlyr           -4.821e+00  1.570e+00  -3.070  0.00215 **
ord2ago          -5.321e+00  1.634e+00  -3.257  0.00113 **
ord3ago          -4.431e+00  1.550e+00  -2.859  0.00427 **
falord            1.389e+00  6.331e-01   2.194  0.02828 *
sprord            6.224e-01  7.182e-01   0.867  0.38623
lpurseasonspring  6.436e+00  1.974e+00   3.260  0.00112 **
recency           2.990e-04  2.841e-03   0.105  0.91619
lifetime         -8.254e-04  1.046e-03  -0.789  0.42999
active           -1.095e+01  8.611e+00  -1.272  0.20359
avg_amount       -5.055e-02  2.958e-02  -1.709  0.08753 .
large_avg1        1.131e+01  2.083e+00   5.429 5.94e-08 ***
pur3yr1          -1.499e+00  3.681e+00  -0.407  0.68383
slscmp1          -2.922e+00  2.697e+00  -1.084  0.27857
sls4bfr           3.036e-02  9.312e-03   3.260  0.00112 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.18 on 4825 degrees of freedom
Multiple R-squared:  0.09509,   Adjusted R-squared:  0.09153
F-statistic: 26.69 on 19 and 4825 DF,  p-value: < 2.2e-16
```

R Output 7:

Linear Regression Model 2 Output

```
Call:
lm(formula = train_linear$targdol ~ slstyr + sls2ago + sls3ago +
    ordlyr + ord2ago + ord3ago + falord + active + lifetime +
    lpurseason + avg_amount + pur3yr + slstyr:ordtyr + slstyr:avg_amount +
    slstyr:slscmp + slslyr:ordlyr + slslyr:avg_amount + slslyr:lifetime +
    sls2ago:ord2ago + sls2ago:recency + ordtyr:ordlyr + ordtyr:avg_amount +
    ordlyr:lifetime + ord2ago:recency + avg_amount:large_avg +
    avg_amount:active + active:recency + active:pur3yr + lifetime:recency +
    lifetime:pur3yr + avg_amount:pur3yr, data = train_linear)

Residuals:
    Min      1Q  Median      3Q     Max
-174.36  -24.44  -11.68    9.83 1659.61


Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.486e+01  1.089e+01   4.120 3.86e-05 ***
slstyr                  3.505e-01  7.469e-02   4.693 2.77e-06 ***
sls2ago                 1.068e-01  1.864e-01   0.573 0.566666
sls3ago                -3.638e-03  1.668e-02  -0.218 0.827320
ordlyr                 -3.137e+00  2.654e+00  -1.182 0.237265
ord2ago                -2.331e+01  1.073e+01  -2.172 0.029920 *
ord3ago                -2.079e+00  1.471e+00  -1.414 0.157553
falord                  1.871e+00  4.940e-01   3.787 0.000154 ***
active                  9.744e+01  6.108e+01   1.595 0.110693
lifetime               -1.085e-02  6.546e-03  -1.658 0.097420 .
lpurseasonspring        6.720e+00  1.759e+00   3.821 0.000135 ***
avg_amount              1.756e-01  8.103e-02   2.167 0.030255 *
pur3yr1                -6.232e+00  1.121e+01  -0.556 0.578210
slstyr:ordtyr           4.728e-02  2.113e-02   2.237 0.025311 *
slstyr:avg_amount      -9.625e-05  3.243e-05  -2.968 0.003013 **
slstyr:slscmp1         -1.602e-01  6.476e-02  -2.474 0.013414 *
ordlyr:slslyr           1.514e-02  1.382e-02   1.096 0.273286
avg_amount:slslyr      -2.961e-04  2.661e-04  -1.113 0.265792
lifetime:slslyr         1.407e-05  8.165e-06   1.724 0.084833 .
sls2ago:ord2ago         6.018e-02  1.507e-02   3.993 6.63e-05 ***
sls2ago:recency        -3.214e-05  7.779e-05  -0.413 0.679519
ordlyr:ordtyr           4.757e-01  1.173e+00   0.405 0.685148
avg_amount:ordtyr      -2.471e-01  4.688e-02  -5.271 1.41e-07 ***
ordlyr:lifetime        -4.389e-04  6.040e-04  -0.727 0.467540
ord2ago:recency         7.269e-03  4.484e-03   1.621 0.105067
avg_amount:large_avg1   6.713e-02  6.091e-02   1.102 0.270485
active:avg_amount       4.654e-01  1.053e-01   4.421 1.01e-05 ***
active:recency         -2.949e-02  1.219e-02  -2.419 0.015612 *
active:pur3yr1         -4.282e+01  3.668e+01  -1.167 0.243148
lifetime:recency        2.090e-06  1.071e-06   1.952 0.051043 .
lifetime:pur3yr1        4.548e-03  4.398e-03   1.034 0.301183
avg_amount:pur3yr1     -5.565e-02  5.950e-02  -0.935 0.349621
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.85 on 4813 degrees of freedom
Multiple R-squared:  0.1086,    Adjusted R-squared:  0.1029
F-statistic: 18.92 on 31 and 4813 DF,  p-value: < 2.2e-16
```

R Output 8:

Linear Regression Model 3 Output

```
Call:
lm(formula = train_linear$targdol ~ log(slstyr + K) + log(sls2ago +
    K) + log(sls3ago + K) + log(ordlyr + K) + log(ord2ago + K) +
    log(ord3ago + K) + log(falord + K) + active + lifetime +
    lpurseason + log(avg_amount + K) + pur3yr + log(slstyr +
    K):log(ordtyr + K) + log(slstyr + K):log(avg_amount + K) +
    log(slstyr + K):slscmp + log(slslyr + K):log(ordlyr + K) +
    log(slslyr + K):log(avg_amount + K) + log(slslyr + K):lifetime +
    log(sls2ago + K):log(ord2ago + K) + log(sls2ago + K):recency +
    log(ordtyr + K):log(ordlyr + K) + log(ordtyr + K):log(avg_amount +
    K) + log(ordlyr + K):lifetime + log(ord2ago + K):recency +
    log(avg_amount + K):large_avg + log(avg_amount + K):active +
    active:recency + active:pur3yr + lifetime:recency + lifetime:pur3yr +
    log(avg_amount + K):pur3yr, data = train_linear)


Residuals:
    Min      1Q  Median      3Q     Max
-102.30  -24.91  -11.72    9.71 1658.88

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -4.347e+01  1.882e+01  -2.309 0.020983 *
log(slstyr + K)                    8.456e-01  1.107e+00   0.764 0.445180
log(sls2ago + K)                   8.641e+00  5.637e+00   1.533 0.125354
log(sls3ago + K)                   7.527e-01  1.075e+00   0.700 0.483934
log(ordlyr + K)                    2.782e+00  2.522e+00   1.103 0.270048
log(ord2ago + K)                  -7.037e+00  7.524e+00  -0.935 0.349717
log(ord3ago + K)                  -1.068e+00  1.469e+00  -0.727 0.467325
log(falord + K)                    1.981e-01  2.718e-01   0.729 0.466192
active                             8.936e+01  6.587e+01   1.357 0.174988
lifetime                          -1.691e-02  8.151e-03  -2.075 0.038085 *
lpurseasonspring                   5.601e+00  1.968e+00   2.845 0.004459 **
log(avg_amount + K)                1.588e+01  3.577e+00   4.439 9.22e-06 ***
pur3yr1                            2.019e+01  1.614e+01   1.251 0.210916
log(slstyr + K):log(ordtyr + K)    3.769e-01  1.354e-01   2.785 0.005377 **
log(slstyr + K):log(avg_amount + K)  6.212e-01  2.208e-01   2.814 0.004917 **
log(slstyr + K):slscmp1            1.509e+00  8.695e-01   1.736 0.082687 .
log(ordlyr + K):log(slslyr + K)    5.846e-01  2.542e-01   2.300 0.021491 *
log(avg_amount + K):log(slslyr + K)  6.335e-01  1.678e-01   3.774 0.000162 ***
lifetime:log(slslyr + K)           6.308e-05  2.895e-04   0.218 0.827545
log(sls2ago + K):log(ord2ago + K)  7.918e-01  2.585e-01   3.063 0.002205 **
log(sls2ago + K):recency          -1.450e-03  2.269e-03  -0.639 0.522798
log(ordlyr + K):log(ordtyr + K)    1.412e-01  4.529e-02   3.117 0.001837 **
log(avg_amount + K):log(ordtyr + K) -1.712e-01  3.037e-01  -0.564 0.573020
log(ordlyr + K):lifetime          -4.452e-05  4.027e-04  -0.111 0.911978
log(ord2ago + K):recency           2.946e-03  2.915e-03   1.011 0.312203
log(avg_amount + K):large_avg1     2.013e+00  5.834e-01   3.452 0.000562 ***
active:log(avg_amount + K)         1.898e+01  4.190e+00   4.530 6.03e-06 ***
active:recency                    -4.414e-02  1.348e-02  -3.275 0.001065 **
active:pur3yr1                    -5.624e+01  3.776e+01  -1.489 0.136435
lifetime:recency                   4.548e-06  1.762e-06   2.582 0.009851 **
lifetime:pur3yr1                   6.595e-03  4.775e-03   1.381 0.167278
log(avg_amount + K):pur3yr1       -9.883e+00  2.470e+00  -4.002 6.39e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.37 on 4813 degrees of freedom
Multiple R-squared:  0.0904,    Adjusted R-squared:  0.08454
F-statistic: 15.43 on 31 and 4813 DF,  p-value: < 2.2e-16
```

R Output 9:

Linear Regression Model 4 Output

```
Call:
lm(formula = log(train_linear$targdol + K) ~ log(slstyr + K) +
    log(sls2ago + K) + log(sls3ago + K) + log(ordlyr + K) + log(ord2ago +
    K) + log(ord3ago + K) + log(falord + K) + active + lifetime +
    lpurseason + log(avg_amount + K) + pur3yr + log(slstyr +
    K):log(ordtyr + K) + log(slstyr + K):log(avg_amount + K) +
    log(slstyr + K):slscmp + log(slslyr + K):log(ordlyr + K) +
    log(slslyr + K):log(avg_amount + K) + log(slslyr + K):lifetime +
    log(sls2ago + K):log(ord2ago + K) + log(sls2ago + K):recency +
    log(ordtyr + K):log(ordlyr + K) + log(ordtyr + K):log(avg_amount +
    K) + log(ordlyr + K):lifetime + log(ord2ago + K):recency +
    log(avg_amount + K):large_avg + log(avg_amount + K):active +
    active:recency + active:pur3yr + lifetime:recency + lifetime:pur3yr +
    log(avg_amount + K):pur3yr, data = train_linear)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3070 -0.5340 -0.0314  0.4873  3.8284


Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.592e+00  2.768e-01   5.751 9.44e-09 ***
log(slstyr + K)                    1.443e-03  1.628e-02   0.089 0.929392
log(sls2ago + K)                   2.165e-01  8.287e-02   2.613 0.009010 **
log(sls3ago + K)                   1.940e-02  1.581e-02   1.227 0.219712
log(ordlyr + K)                    3.527e-02  3.708e-02   0.951 0.341546
log(ord2ago + K)                  -2.650e-01  1.106e-01  -2.395 0.016649 *
log(ord3ago + K)                  -2.459e-02  2.160e-02  -1.139 0.254960
log(falord + K)                    6.951e-03  3.997e-03   1.739 0.082088 .
active                             9.354e-01  9.685e-01   0.966 0.334152
lifetime                          -2.902e-04  1.198e-04  -2.422 0.015485 *
lpurseasonspring                   8.768e-02  2.894e-02   3.029 0.002463 **
log(avg_amount + K)                3.682e-01  5.259e-02   7.002 2.87e-12 ***
pur3yr1                            4.374e-01  2.373e-01   1.843 0.065363 .
log(slstyr + K):log(ordtyr + K)    6.703e-03  1.990e-03   3.368 0.000763 ***
log(slstyr + K):log(avg_amount + K) 1.570e-02  3.246e-03   4.836 1.36e-06 ***
log(slstyr + K):slscmp1            3.673e-02  1.278e-02   2.873 0.004081 **
log(ordlyr + K):log(slslyr + K)    9.240e-03  3.737e-03   2.473 0.013448 *
log(avg_amount + K):log(slslyr + K) 1.281e-02  2.468e-03   5.190 2.19e-07 ***
lifetime:log(slslyr + K)           3.387e-06  4.257e-06   0.796 0.426222
log(sls2ago + K):log(ord2ago + K)  8.279e-03  3.801e-03   2.178 0.029456 *
log(sls2ago + K):recency          -5.059e-05  3.337e-05  -1.516 0.129577
log(ordlyr + K):log(ordtyr + K)    2.462e-03  6.659e-04   3.698 0.000220 ***
log(avg_amount + K):log(ordtyr + K) -4.704e-03  4.465e-03  -1.054 0.292054
log(ordlyr + K):lifetime          -4.367e-06  5.921e-06  -0.738 0.460818
log(ord2ago + K):recency           8.543e-05  4.286e-05   1.993 0.046280 *
log(avg_amount + K):large_avg1     1.625e-02  8.577e-03   1.894 0.058262 .
active:log(avg_amount + K)         3.785e-01  6.160e-02   6.145 8.65e-10 ***
active:recency                    -7.097e-04  1.982e-04  -3.581 0.000345 ***
active:pur3yr1                    -9.358e-01  5.552e-01  -1.686 0.091949 .
lifetime:recency                   8.298e-08  2.590e-08   3.204 0.001363 **
lifetime:pur3yr1                   1.181e-04  7.021e-05   1.682 0.092590 .
log(avg_amount + K):pur3yr1       -2.110e-01  3.631e-02  -5.811 6.60e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7553 on 4813 degrees of freedom
Multiple R-squared:  0.1395,    Adjusted R-squared:  0.134
F-statistic: 25.17 on 31 and 4813 DF,  p-value: < 2.2e-16
```

R Output 10:

Linear Regression Model 5 Output

```
Call:
lm(formula = log(train_linear$targdol + K) ~ log(slstyr + K) +
    log(sls2ago + K) + log(ordlyr + K) + log(ord2ago + K) + log(falord +
    K) + active + lifetime + lpurseason + log(avg_amount + K) +
    pur3yr + log(slstyr + K):log(ordtyr + K) + log(slstyr + K):log(avg_amount +
    K) + log(slstyr + K):slscmp + log(slslyr + K):log(ordlyr +
    K) + log(slslyr + K):log(avg_amount + K) + log(sls2ago +
    K):log(ord2ago + K) + log(ordtyr + K):log(ordlyr + K) + log(avg_amount +
    K):large_avg + log(avg_amount + K):active + active:recency +
    log(avg_amount + K):pur3yr, data = train_linear)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3587 -0.5399 -0.0294  0.4846  3.8019

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.199e+00  1.802e-01   6.656 3.14e-11 ***
log(slstyr + K)                    1.719e-03  1.624e-02   0.106 0.915697
log(sls2ago + K)                   9.440e-02  1.723e-02   5.478 4.51e-08 ***
log(ordlyr + K)                    2.188e-02  3.471e-02   0.630 0.528445
log(ord2ago + K)                  -6.808e-02  4.016e-02  -1.695 0.090082 .
log(falord + K)                    7.001e-03  3.951e-03   1.772 0.076440 .
active                            -1.259e+00  2.759e-01  -4.564 5.16e-06 ***
lifetime                           1.864e-05  1.317e-05   1.415 0.156996
lpurseasonspring                   9.833e-02  2.771e-02   3.549 0.000390 ***
log(avg_amount + K)                4.176e-01  4.436e-02   9.414  < 2e-16 ***
pur3yr1                            8.715e-01  1.253e-01   6.956 3.96e-12 ***
log(slstyr + K):log(ordtyr + K)    7.501e-03  1.865e-03   4.023 5.85e-05 ***
log(slstyr + K):log(avg_amount + K) 1.406e-02 2.424e-03   5.801 7.00e-09 ***
log(slstyr + K):slscmp1            3.545e-02  1.259e-02   2.815 0.004899 **
log(ordlyr + K):log(slslyr + K)    8.774e-03  3.711e-03   2.364 0.018111 *
log(avg_amount + K):log(slslyr + K) 1.460e-02 1.931e-03   7.561 4.75e-14 ***
log(sls2ago + K):log(ord2ago + K)  7.179e-03  3.784e-03   1.897 0.057889 .
log(ordlyr + K):log(ordtyr + K)    2.386e-03  6.590e-04   3.620 0.000297 ***
log(avg_amount + K):large_avg1     1.624e-02  8.452e-03   1.921 0.054776 .
active:log(avg_amount + K)         4.080e-01  5.915e-02   6.899 5.92e-12 ***
active:recency                    -1.699e-04  6.839e-05  -2.484 0.013023 *
log(avg_amount + K):pur3yr1       -2.402e-01  3.348e-02  -7.174 8.36e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7561 on 4823 degrees of freedom
Multiple R-squared:  0.1359,    Adjusted R-squared:  0.1322
F-statistic: 36.13 on 21 and 4823 DF,  p-value: < 2.2e-16
```

R Output 11:

Anova Test for Linear Models 4 & 5

```
> anova(linear4, linear5)
Analysis of Variance Table

Model 1: log(train_linear$targdol + K) ~ log(slstyr + K) + log(sls2ago +
    K) + log(sls3ago + K) + log(ordlyr + K) + log(ord2ago + K) +
    log(ord3ago + K) + log(falord + K) + active + lifetime +
    lpurseason + log(avg_amount + K) + pur3yr + log(slstyr +
    K):log(ordtyr + K) + log(slstyr + K):log(avg_amount + K) +
    log(slstyr + K):slscmp + log(slslyr + K):log(ordlyr + K) +
    log(slslyr + K):log(avg_amount + K) + log(slslyr + K):lifetime +
    log(sls2ago + K):log(ord2ago + K) + log(sls2ago + K):recency +
    log(ordtyr + K):log(ordlyr + K) + log(ordtyr + K):log(avg_amount +
    K) + log(ordlyr + K):lifetime + log(ord2ago + K):recency +
    log(avg_amount + K):large_avg + log(avg_amount + K):active +
    active:recency + active:pur3yr + lifetime:recency + lifetime:pur3yr +
    log(avg_amount + K):pur3yr
Model 2: log(train_linear$targdol + K) ~ log(slstyr + K) + log(sls2ago +
    K) + log(ordlyr + K) + log(ord2ago + K) + log(falord + K) +
    active + lifetime + lpurseason + log(avg_amount + K) + pur3yr +
    log(slstyr + K):log(ordtyr + K) + log(slstyr + K):log(avg_amount +
    K) + log(slstyr + K):slscmp + log(slslyr + K):log(ordlyr +
    K) + log(slslyr + K):log(avg_amount + K) + log(sls2ago +
    K):log(ord2ago + K) + log(ordtyr + K):log(ordlyr + K) + log(avg_amount +
    K):large_avg + log(avg_amount + K):active + active:recency +
    log(avg_amount + K):pur3yr
  Res.Df    RSS  Df Sum of Sq      F  Pr(>F)
1   4813 2745.7
2   4823 2757.2 -10   -11.469 2.0103 0.02852 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R Output 12:

Linear Regression Model 6 Output

```
Call:
lm(formula = log(train_linear$targdol + K) ~ scale(log(slstyr +
    K)) + scale(log(sls2ago + K)) + scale(log(ordlyr + K)) +
    scale(log(ord2ago + K)) + active + lpurseason + scale(log(avg_amount +
    K)) + pur3yr + scale(log(slstyr + K)):scale(log(ordtyr +
    K)) + scale(log(slstyr + K)):scale(log(avg_amount + K)) +
    scale(log(slstyr + K)):slscmp + scale(log(slslyr + K)):scale(log(ordlyr +
    K)) + scale(log(slslyr + K)):scale(log(avg_amount + K)) +
    scale(log(sls2ago + K)):scale(log(ord2ago + K)) + scale(log(ordtyr +
    K)):scale(log(ordlyr + K)) + scale(log(avg_amount + K)):large_avg +
    scale(log(avg_amount + K)):active + active:recency + scale(log(avg_amount +
    K)):pur3yr, data = train_linear)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3521 -0.5367 -0.0241  0.4920  3.7683

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      2.483e+00  1.543e-01  16.094  < 2e-16 ***
scale(log(slstyr + K))                          -8.464e-02  2.394e-02  -3.535 0.000412 ***
scale(log(sls2ago + K))                          3.248e-01  1.668e-01   1.947 0.051594 .
scale(log(ordlyr + K))                          -8.674e-02  2.362e-02  -3.672 0.000243 ***
scale(log(ord2ago + K))                         -4.458e-01  1.300e-01  -3.430 0.000610 ***
active                                           3.167e-01  1.774e-01   1.785 0.074350 .
lpurseasonspring                                 6.559e-02  2.511e-02   2.612 0.009018 **
scale(log(avg_amount + K))                       2.726e-01  3.684e-02   7.401 1.59e-13 ***
pur3yr1                                          2.824e-02  4.611e-02   0.613 0.540227
scale(log(slstyr + K)):scale(log(ordtyr + K))    3.828e-01  7.429e-02   5.152 2.67e-07 ***
scale(log(slstyr + K)):scale(log(avg_amount + K)) 8.191e-02  1.413e-02   5.799 7.10e-09 ***
scale(log(slstyr + K)):slscmp1                   1.086e-01  4.757e-02   2.282 0.022503 *
scale(log(ordlyr + K)):scale(log(slslyr + K))    3.271e-01  8.244e-02   3.968 7.36e-05 ***
scale(log(avg_amount + K)):scale(log(slslyr + K)) 7.284e-02  1.548e-02   4.707 2.59e-06 ***
scale(log(sls2ago + K)):scale(log(ord2ago + K))  2.123e-01  1.074e-01   1.976 0.048164 *
scale(log(ordlyr + K)):scale(log(ordtyr + K))    4.791e-02  1.449e-02   3.308 0.000947 ***
scale(log(avg_amount + K)):large_avg1            8.055e-02  3.504e-02   2.299 0.021551 *
active:scale(log(avg_amount + K))                3.751e-01  5.321e-02   7.049 2.05e-12 ***
active:recency                                  -1.816e-04  6.811e-05  -2.666 0.007709 **
scale(log(avg_amount + K)):pur3yr1              -2.057e-01  3.048e-02  -6.750 1.65e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7564 on 4825 degrees of freedom
Multiple R-squared:  0.1349,    Adjusted R-squared:  0.1314
F-statistic: 39.58 on 19 and 4825 DF,  p-value: < 2.2e-16
```

R Output 13:

Linear Regression Model 7 Output

```
Call:
lm(formula = log(train_linear[-c(1289, 1312, 4202, 4640, 433,
    2602, 4672), ]$targdol + K) ~ log(slstyr + K) + scale(log(sls2ago +
    K)) + scale(log(ordlyr + K)) + scale(log(ord2ago + K)) +
    active + lpurseason + scale(log(avg_amount + K)) + pur3yr +
    scale(log(slstyr + K)):scale(log(ordtyr + K)) + scale(log(slstyr +
    K)):scale(log(avg_amount + K)) + scale(log(slstyr + K)):slscmp +
    scale(log(slslyr + K)):scale(log(ordlyr + K)) + scale(log(slslyr +
    K)):scale(log(avg_amount + K)) + scale(log(sls2ago + K)):scale(log(ord2ago +
    K)) + scale(log(ordtyr + K)):scale(log(ordlyr + K)) + scale(log(avg_amount +
    K)):large_avg + scale(log(avg_amount + K)):active + active:recency +
    scale(log(avg_amount + K)):pur3yr, data = train_linear[-c(1289,
    1312, 4202, 4640, 433, 2602, 4672), ])

Residuals:
    Min      1Q  Median      3Q     Max
-2.1253 -0.5344 -0.0272  0.4858  3.7415

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     2.402e+00  1.579e-01  15.217  < 2e-16 ***
log(slstyr + K)                                -1.345e-02  3.697e-03  -3.639 0.000277 ***
scale(log(sls2ago + K))                         5.491e-01  1.723e-01   3.187 0.001445 **
scale(log(ordlyr + K))                         -9.153e-02  2.350e-02  -3.896 9.92e-05 ***
scale(log(ord2ago + K))                        -6.804e-01  1.381e-01  -4.926 8.67e-07 ***
active                                          3.305e-01  1.763e-01   1.874 0.060937 .
lpurseasonspring                                6.084e-02  2.497e-02   2.437 0.014862 *
scale(log(avg_amount + K))                      2.621e-01  3.591e-02   7.300 3.34e-13 ***
pur3yr1                                         3.283e-02  4.582e-02   0.717 0.473696
scale(log(slstyr + K)):scale(log(ordtyr + K))   3.786e-01  7.418e-02   5.103 3.47e-07 ***
scale(log(avg_amount + K)):scale(log(slstyr + K))  8.116e-02  1.389e-02   5.844 5.45e-09 ***
scale(log(slstyr + K)):slscmp1                  1.109e-01  4.732e-02   2.345 0.019085 *
scale(log(ordlyr + K)):scale(log(slslyr + K))   3.427e-01  8.197e-02   4.180 2.96e-05 ***
scale(log(avg_amount + K)):scale(log(slslyr + K))  7.303e-02  1.510e-02   4.837 1.36e-06 ***
scale(log(sls2ago + K)):scale(log(ord2ago + K))  2.350e-01  1.068e-01   2.199 0.027895 *
scale(log(ordlyr + K)):scale(log(ordtyr + K))   4.612e-02  1.440e-02   3.203 0.001370 **
scale(log(avg_amount + K)):large_avg1           8.363e-02  3.421e-02   2.445 0.014540 *
active:scale(log(avg_amount + K))               3.698e-01  5.249e-02   7.046 2.10e-12 ***
active:recency                                 -1.817e-04  6.768e-05  -2.684 0.007289 **
scale(log(avg_amount + K)):pur3yr1             -2.104e-01  3.011e-02  -6.988 3.17e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7514 on 4818 degrees of freedom
Multiple R-squared:  0.141,     Adjusted R-squared:  0.1376
F-statistic: 41.61 on 19 and 4818 DF,  p-value: < 2.2e-16
```