



Building a Linear Regression

This demonstration illustrates building, exploring, and refining a linear regression model in SAS Visual Statistics using the **VS_Bank** data.

Creating and Exploring the Linear Regression Model

1. Click **Explore and Visualize Data** in the application shortcut area. Alternatively, first select **SAS Home** from the bookmarks bar or from the link on the page.
2. Start a new report. In the upper right corner, click **⋮ (Menu)** and select **New**.
3. Click **Data**.
4. Select **VS_BANK** ⇒ **OK**.
5. Select the **Objects** pane.
6. Under SAS Visual Statistics, either double-click or drag **Linear Regression** onto the canvas.
7. Select the **Roles** pane. Add **tgt interval New Sales** as the response variable.
Note: You can also drag **tgt interval New Sales** onto the Linear Regression canvas or use **Assign Data**.
8. Add the twelve variables that begin with **logi_rfm** as continuous effects.
9. Add **category 1 Account Activity Level** and **category 2 Customer Value Level** as classification effects.

Target and input variable roles for the model are summarized below.

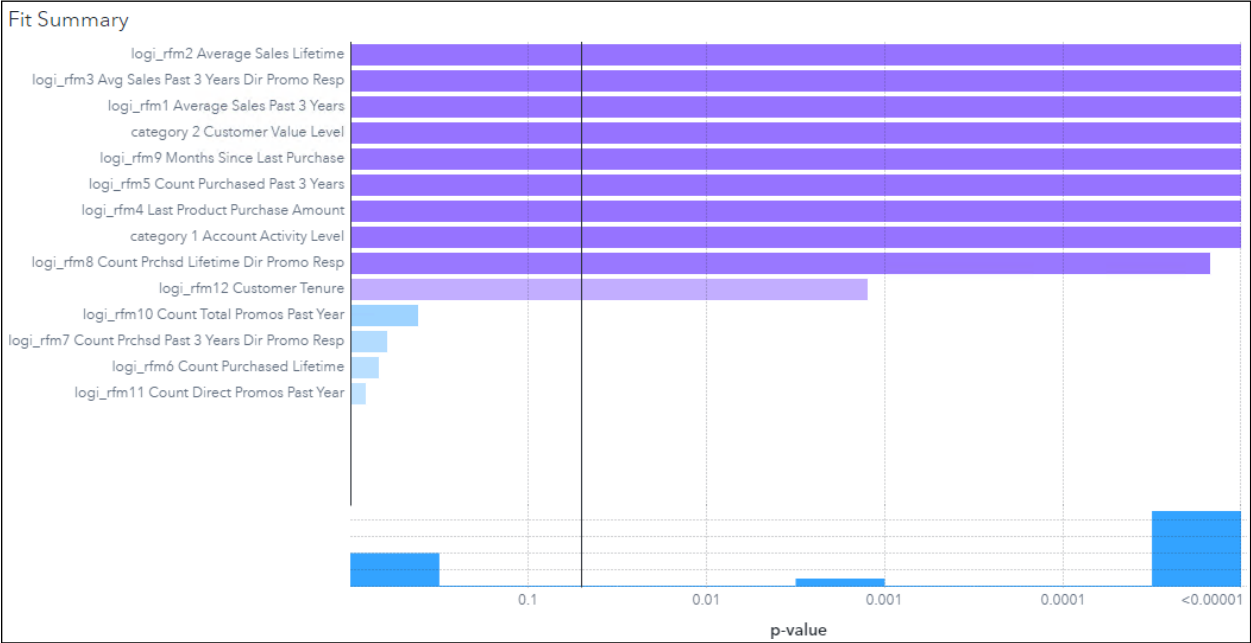
- Response
 - tgt Interval New Sales
- Continuous effects
 - logi_rfm1 Average Sales Past 3 Years
 - logi_rfm10 Count Total Promos Past Y
 - logi_rfm11 Count Direct Promos Past
 - logi_rfm12 Customer Tenure
 - logi_rfm2 Average Sales Lifetime
 - logi_rfm3 Avg Sales Past 3 Years Dir P
 - logi_rfm4 Last Product Purchase Amo
 - logi_rfm5 Count Purchased Past 3 Year
 - logi_rfm6 Count Purchased Lifetime
 - logi_rfm7 Count Prchsd Past 3 Years D
 - logi_rfm8 Count Prchsd Lifetime Dir P
 - logi_rfm9 Months Since Last Purchase
 - + Add
- Classification effects
 - category 1 Account Activity Level
 - category 2 Customer Value Level
 - + Add

10. On the summary bar, select **ASE** and change it to **R-Square**. The R-square is .0933. It appears the model explains less than 10% of the variability in the data. The R-square measure seems low, but further model refinement, illustrated below, might improve it. The model uses 211,509 observations.

Note: The number of observations that are used seems low. The data set contains more than a million accounts or customers. Note that the data has a 20% response rate. For the interval-valued target that is used in this analysis, non-responders are coded with a missing value. The linear regression model is fitted using only responders in the data.

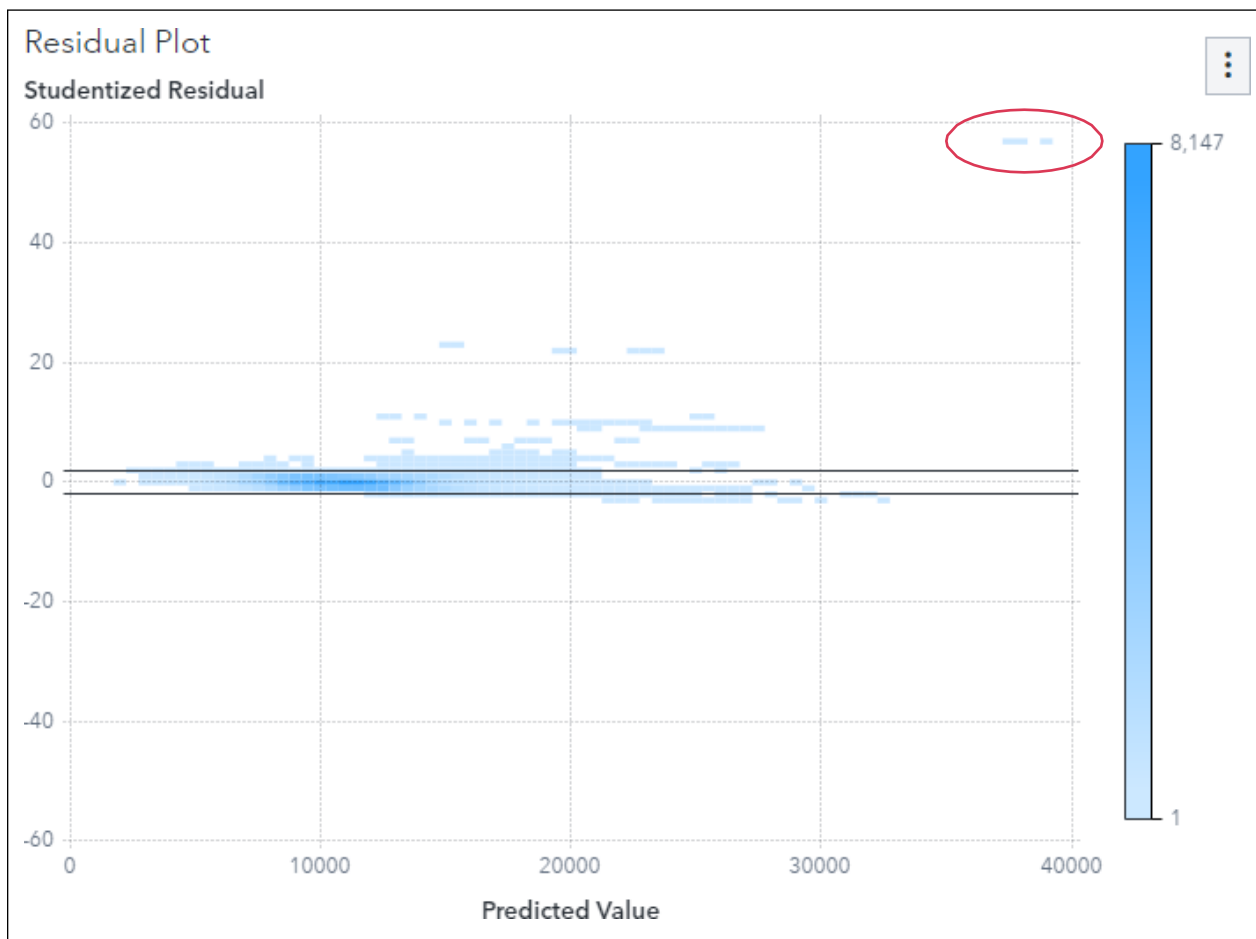
11. You can investigate other fit measures in addition to R-square. Click **R-Square** and select **Root MSE** (Mean Square Error.) The RMSE is *8086.21*. This statistic tells you that, on average and over the range of the data, the difference between the prediction and the actual value is approximately \$8,086. This seems imprecise and might be an artifact of how the data were collected. (Methods to improve this measure are explored in the model refinement part of this demonstration.)
12. In the Options pane, under Model Display, select **General** and change the plot layout to **Stack** to expand the Fit Summary window on the canvas.

A default criterion, $p\text{-value} < .05$, is used as a threshold for variable importance. Effects with $p\text{-values}$ less than .05 have purple horizontal bars that extend to the right side of the plot. Effects with $p\text{-values}$ greater than .05 have blue bars. A bar chart below shows the distribution of effects in various ranges of $p\text{-values}$ on the negative \log_{10} scale.



Some of the logged, imputed RFM variables (1, 2, 3, 4, 5, 8, 9, and 12), and both categorical inputs, are important predictors in the model according to the criterion listed above. (Functionality for removing irrelevant input variables from the model is discussed below.)

13. Click the **Residual Plot** tab. Right-click in the residual plot and change the residual measure to **Studentized Residual**.



Note: A studentized residual is a raw residual that is divided by its estimated standard deviation.

Two features of this plot warrant further investigation.

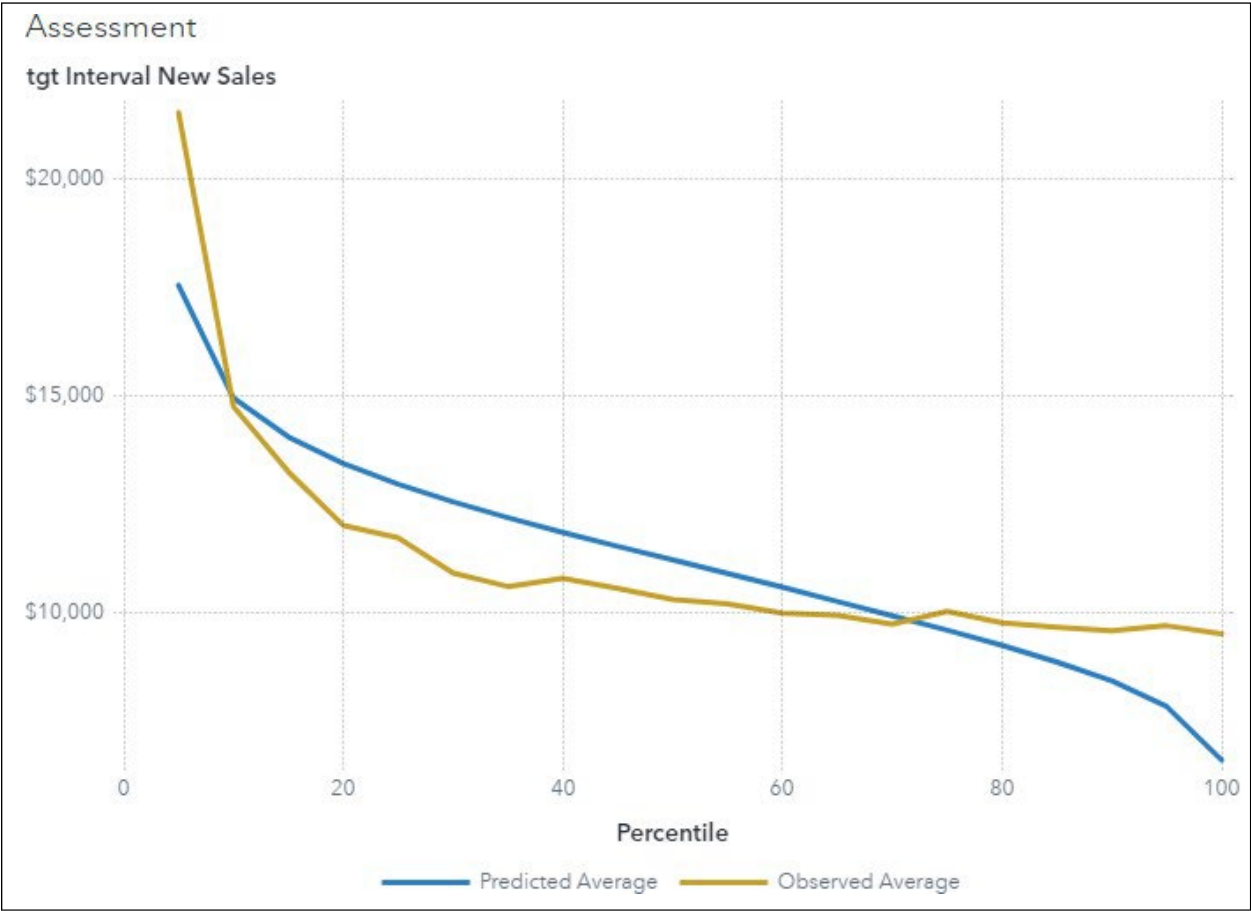
- The variance of the residuals seems to increase as a function of the predicted value. (Causes and possible remedies for heteroscedasticity are explored later in this demonstration.)
- There are some very large, positive outliers in the residuals. The influence plot can help you explore the effect of these outliers on the fitted model.

14. Use the mouse to drag and select the large, positive residuals. Right-click in the residual plot and select **Show selected** to open a details table. Column values associated with these observations are listed for examination.

Show Selected			
Predicted Value	tgt Interval New Sales	category 1 Account Activity ...	category 2 Customer Value
38832.398277	\$500,000.00	X	E
37930.487508	\$500,000.00	X	E
37930.487508	\$500,000.00	X	E
37922.978595	\$500,000.00	X	E
37494.80967	\$500,000.00	X	E
37494.80967	\$500,000.00	X	E
37771.817823	\$500,000.00	Y	E
37494.80967	\$500,000.00	X	E
37494.80967	\$500,000.00	X	E

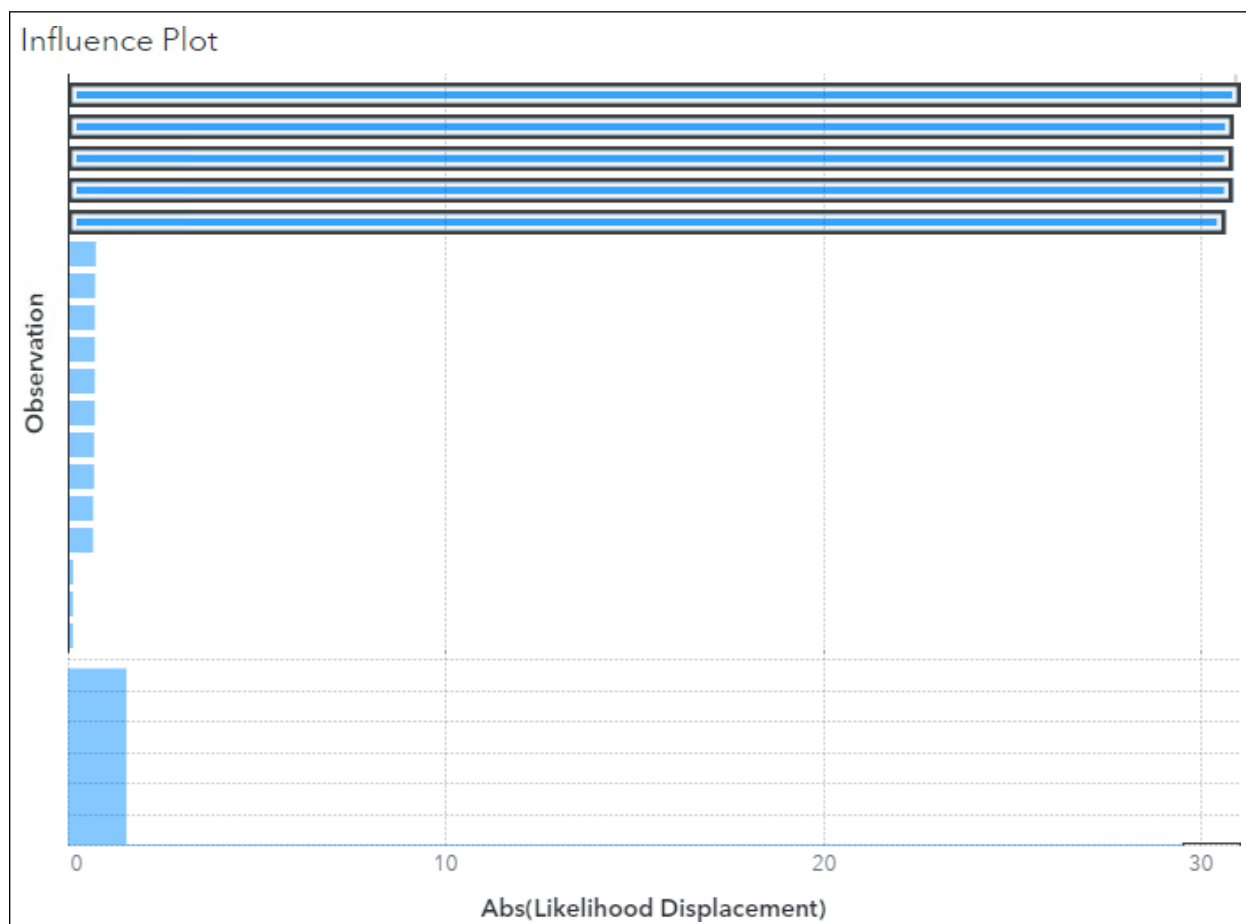
15. Close the Show selected window.

16. Click the **Assessment** tab.



The line plots show model predictions versus the actual response amount in the data. To create the lines, both outcomes and predictions are binned into percentiles. Although the model's predictions seem consistent with responder outcomes over the middle range of the plot, the model seems to under-predict the response amount at the high and low end of the response range.

17. In the Options pane, select **Influence Plot/ Variable Selection Plot**. Select **Influence Plot** from the Plot to show pull-down. Then click the **Influence** tab.
18. Right-click the plot. Change the influence measure to **Likelihood Displacement**. Then select the top five bars in the influence plot.



The bars in the influence plot correspond to observations (accounts) in the data. The selected observations are **high leverage**. That is, when they are removed from the data, the maximized value of the function that is used to generate the parameter estimates changes substantially—hence, likelihood displacement.

Note: Each bar on the influence plot can represent more than one observation. This happens when two or more observations are influential and they have the same column (target and predictor) values.

Refining the Linear Regression Model and Exporting Results

Model Selection

- 1. In the Options pane, select **Informative missingness**.
- 2. Then select **Backward** in the **Variable selection method** field. Change the **Selection criterion** field value to **Significance level**. Enter **.01** as the significance level (if it is not the default.)

▼ General

✓ Informative missingness ⓘ

Variable selection method:

Backward ▼

Selection criterion:

Significance level ▼

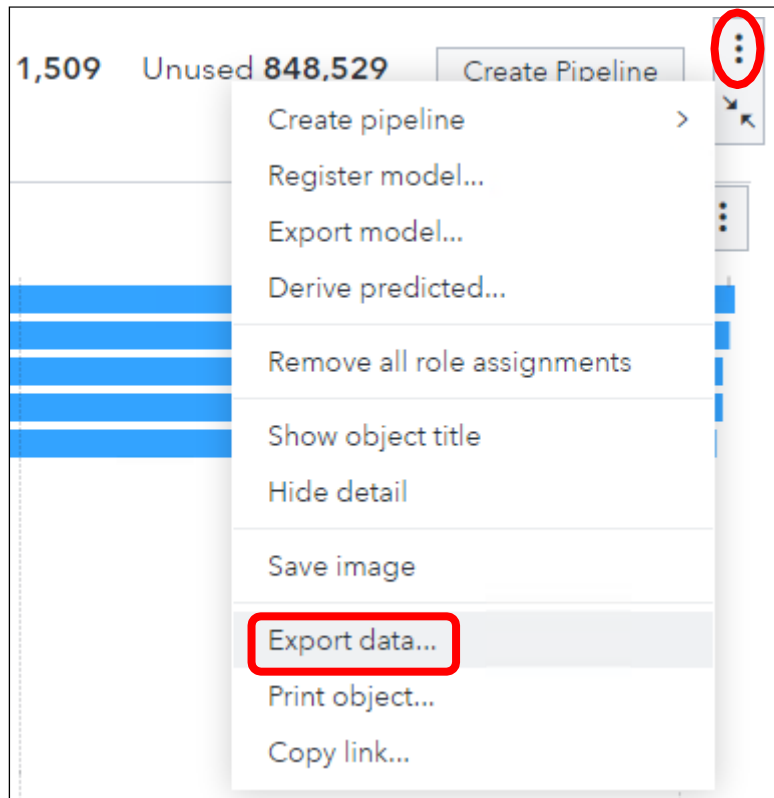
Significance level:

0.01

- 3. Select the **Selection Summary** tab to examine the variables that were removed from the model as well as the order in which they were removed during the backward elimination.

Dimensions	Overall ANOVA	Fit Statistics	Parameter Estimates	Type III Test	Selection Info	Selection Summary	Assessment	Assessment Statistics
Control	Step	Effect Removed				Number Of Effects	Significance Level	Optimal P Value
	0	logi_rfm10 Count Total Promos Past Year_miss				25	.	0
	0	logi_rfm11 Count Direct Promos Past Year_miss				24	.	0
	0	logi_rfm12 Customer Tenure_miss				23	.	0
	0	logi_rfm2 Average Sales Lifetime_miss				22	.	0
	0	logi_rfm3 Avg Sales Past 3 Years Dir Promo Resp_miss				21	.	0
	0	logi_rfm4 Last Product Purchase Amount_miss				20	.	0
	0	logi_rfm5 Count Purchased Past 3 Years_miss				19	.	0
	0	logi_rfm6 Count Purchased Lifetime_miss				18	.	0
	0	logi_rfm7 Count Prchsd Past 3 Years Dir Promo Resp_miss				17	.	0
	0	logi_rfm8 Count Prchsd Lifetime Dir Promo Resp_miss				16	.	0
	0	logi_rfm9 Months Since Last Purchase_miss				15	.	0
	1	logi_rfm11 Count Direct Promos Past Year				14	0.815696	0
	2	logi_rfm6 Count Purchased Lifetime				13	0.717673	0
	3	logi_rfm7 Count Prchsd Past 3 Years Dir Promo Resp				12	0.498028	0
	4	logi_rfm10 Count Total Promos Past Year				11	0.137549	0

- 4. From the object toolbar, click  (**More**) and select **Export data**.



5. In the Export Data window, validate that **Formatted data** is selected and click **OK**.
6. In Chrome, wait until the exported file appears in the lower left corner. Select **Open** in the menu beside the exported Linear Regression 1.xlsx file.

7. Click the **Parameter Estimates** tab in Microsoft Excel.


	A	B	C	D	E
1	Parameter	Estimate	Standard Error	t Value	Pr > t
2	Intercept	-1814.89177	313.7246515	-5.78498298	0.00000
3	category 1 Account Activity Level X	-740.5268673	96.01228053	-7.71283489	0.00000
4	category 1 Account Activity Level Y	-880.5769616	112.9913292	-7.79331448	0.00000
5	category 1 Account Activity Level Z	0			
6	category 2 Customer Value Level A	-1276.638689	49.89432487	-25.5868517	0.00000
7	category 2 Customer Value Level B	-695.4534215	51.87636195	-13.405979	0.00000
8	category 2 Customer Value Level C	-209.7935784	55.59368562	-3.77369437	0.00016
9	category 2 Customer Value Level D	-429.2543825	64.57358743	-6.64752261	0.00000
10	category 2 Customer Value Level E	0			
11	logi_rfm1 Average Sales Past 3 Years	3154.553511	117.1169675	26.93506823	0.00000
12	logi_rfm12 Customer Tenure	207.9554504	61.45120184	3.384074585	0.00071
13	logi_rfm2 Average Sales Lifetime	3693.767283	108.0529157	34.18479973	0.00000
14	logi_rfm3 Avg Sales Past 3 Years Dir Promo Resp	-2553.609557	87.60486085	-29.1491766	0.00000
15	logi_rfm4 Last Product Purchase Amount	597.7835631	51.86909844	11.52484969	0.00000
16	logi_rfm5 Count Purchased Past 3 Years	-1680.12347	58.59540679	-28.6732965	0.00000
17	logi_rfm8 Count Prchsd Lifetime Dir Promo Resp	366.7816247	51.01692568	7.18941057	0.00000

◀ ▶ Dimensions Overall ANOVA Fit Statistics **Parameter Estimates** Model ANC ... +

8. Exit Excel and do not save the changes.

Interactions

Domain experts hypothesize interactions between input variables in the data. The interaction functionality in SAS Visual Statistics is accessed under the Roles portion of the interface.

- On the report, click  (**Restore**) so that the Data pane is available.
- Click the **Data** pane. Select **New data item** ⇨ **Interaction Effect**.
- Highlight **logi_rfm4 Last Product Purchase Amount**. Select the arrow to move it under the Effect elements side of the table. Do the same for **logi_rfm9 Months Since Last Purchase**.

New Interaction Effect

Available columns (12 of 55):

log

- logi_rfm3 Avg Sales Past 3 Year...
- logi_rfm4 Last Product Purchas...
- logi_rfm5 Count Purchased Pas...
- logi_rfm6 Count Purchased Life...
- logi_rfm7 Count Prchsd Past 3 Y...
- logi_rfm8 Count Prchsd Lifetime

Effect elements (2):

- logi_rfm4 Last Product Purchase ...
- logi_rfm9 Months Since Last Purc...

☒ Add complete interaction effect ⓘ

logi_rfm4 Last Product Purchase Amount*logi_rfm9 Months Since Last Purchase

☐ Add two-way interaction effects ⓘ

☐ Add all squared interaction effects (measures only) ⓘ

logi_rfm4 Last Product Purchase Amount*logi_rfm4 Last Product Purchase Amount, logi_rfm9 Months Since Last Purchase*logi_rfm9 Months Since Last Purchase

OK

Cancel

- Click **OK**. The new interaction term appears at the bottom of the list of variables.

Note: Domain experts suggest that there is a negative correlation between purchase amount and purchase recency. That is, customers who purchased recently tend to buy smaller amounts (obtain smaller loans) than customer who purchased less recently.

- In the Roles pane, add **RFMInteraction** as an interaction effect.

The Fit Summary plot provides evidence in favor of the hypothesis regarding purchase recency and amount. The interaction term is selected and included in the model. Also notice that although the individual effect that reflects purchase recency (logi_rfm9) is no longer significant, it is retained in the regression due to model hierarchy.

- Click the **Influence** tab.

- Right-click the **Influence Plot** and select **Variable Selection Plot**.

The variable selection plot displays the four effects that were removed from the model as well as the order in which they were removed. The last variable removed from the model is logi_rfm10 with a significance level of .08. After this term is removed, 11 effects remain in the model.

Variable Selection Plot

Significance Level

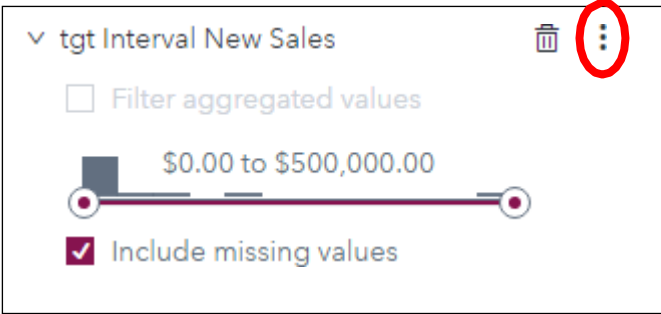


Filtering

Most of the accounts in the data are consumers of short- to medium-term loans. These include home equity lines of credit and automobile loans. Historically, these are in the \$2,000 to \$50,000 range for the bank. However, some loans in the data are less than \$2,000, and some are similar to “payday” loans. Also, some loan amounts greatly exceed \$50,000. If accounts at the high and low ends of the target variable range are very different in terms of systematic relationships between predictors and loan quantity demanded, then bias and heteroscedasticity (unequal variance across the regression line) could result. The analysis assesses the impact of observations in the tails of the target variable on linear regression model results.

- 1. Select the **Filters** pane. Select **New filter** ⇒ **tgt Interval New Sales**.

The existing range for new sales amounts is \$0 to \$500,000. To be consistent with the loan profile for this portfolio, modify the filter range.



- 2. The default filter does not lend the granularity that is necessary for the desired range, but it provides a good starting definition. On the filter, click **Options** and select **Advanced edit**.
- 3. Enter **2000** and **50000** for the filter values. Select **OK**.

Advanced Filter

Name: *

tgt Interval New Sales Filter

Operators

Functions

Data

New parameter

1

BetweenInclusive('tgt Interval New Sales'n, 2000, 50000) OR Missing('tgt Interval New Sales'n)

Results

Preview selection only

tgt Interval New Sales	tgt Interval New Sales Filter
\$7,000.00	True
\$7,000.00	True

lower:

2000

upper:

50000

Returned observations: 1047388

Total observations: 1060038

OK

Cancel

The diagnostics indicate these results:

- The filter greatly reduces the range of the target variable. However, only approximately 5% of the responders are excluded by the filter.
- Root mean square error is reduced to 6031.4372.
- Problems with bias and violations of error assumptions, noted earlier, are partially mitigated.

Further exploration using filtering can help analysts understand whether homogeneous groups of responder accounts exist within the distribution of the target variable. Subsequent prediction quality can be improved if separate modeling exercises are performed for different groups of responders.



Save your report as **VS_BankLinear**.



Performing Model Validation

This demonstration illustrates using the **VS_Bank** data set to perform model validation on a linear regression model in SAS Visual Statistics.

Using Partitioned Data in a Linear Regression Model

1. Open the previous report, **VS_BankLinear**. In the upper right corner, click  (**Menu**) and select **Open** ⇒ **My Folder** ⇒ **VS_BankLinear**.
2. Click **Open**.
3. On the Page 1 tab of the Linear Regression, click  (**Options**) and select **Duplicate Page**.

This is the same linear regression model with a filter that was created earlier. It yielded the following results:


- R-square: .0941
- RMSE: 6031.44
- ASE: 36,375,307.4
- Observations Used: 198,859


4. In the Data pane, select **New data item** ⇒ **Partition**.

- 5. In the New Partition window, use the default values and click **OK**.
- 6. In the Roles pane, add **Partition** as the partition ID.
- 7. Click the **Assessment** tab.



The Validation ASE (Average Square Error) of the new model is 36,358,640.9, and 198,859 observations were used (for the entire model.) Although it is still showing issues with bias, the validation assessment might also reveal an indication of model complexity.

- 8. Click  **Maximize**) on the object toolbar to enter the Maximize mode and see the details table.

< Dimensions Overall ANOVA Fit Statistics Parameter Estimates Type III Test Selection In > 	
Description	Value
Number of Model Effects	29
Number of Classification Effects	2
Number of Columns in X	35
Rank of Cross-product Matrix	17
Observations Read	1,047,388
Observations Used	198,859
Observations Used for Training	118,771
Observations Used for Validation	80,088

60 percent of the data, or 118,771 observations, were used in the validation. Results can vary due to the partition.

9. Click the **Assessment** tab in the details table.


Dimensions	Overall ANOVA	Fit Statistics	Parameter Estimates	Type III Test	Assessment	Assessment Statistics	
Percentile	Training Observations	Training Predicted Average	Training Observed Average		Validation Observations	Validation Predicted Average	Validation Observed Average
5	4958	\$16,334.51	\$19,435.97		4986	\$16,332.82	\$19,325.79
10	4958	\$14,639.34	\$15,312.53		4986	\$14,618.68	\$15,214.90
15	4958	\$13,919.98	\$13,661.15		4986	\$13,905.65	\$13,654.93
20	4958	\$13,428.64	\$12,628.58		4986	\$13,413.19	\$12,705.38
25	4958	\$13,048.60	\$12,257.77		4986	\$13,028.02	\$12,106.00
30	4958	\$12,718.97	\$11,290.24		4986	\$12,697.38	\$11,479.14
35	4958	\$12,422.13	\$11,528.64		4986	\$12,403.51	\$11,461.39
40	4958	\$12,154.65	\$11,154.30		4986	\$12,134.97	\$11,270.80
45	4958	\$11,899.69	\$11,146.53		4986	\$11,883.40	\$11,173.66
50	4958	\$11,660.59	\$10,908.03		4986	\$11,650.80	\$11,062.07
55	4958	\$11,419.69	\$10,999.90		4986	\$11,412.47	\$11,014.34

Predicted and observed averages for both the training data and the validation data by percentile are available for comparison on the Assessment tab of the details table. The number of observations in each percentile is also listed. For a graphical comparison of the training and validation assessment data, examine the assessment plot.

10. Click the **Assessment Statistics** tab in the details table.

Dimensions	Overall ANOVA	Fit Statistics	Parameter Estimates	Type III Test	Assessment	Assessment Statistics	
Partition		ASE	Observed Average		SSE	Observations Used	Unused
Training		36,395,639.5747	19,435.9722		3,608,882,433,305.2417	99,157	430,862
Validation		36,358,640.9218	19,325.7922		3,625,029,217,187.8457	99,702	430,317

The Assessment Statistics tab contains both the training and validation ASE (average square error) as well as the number of observations that are used to generate each statistic. Once again, results can vary due to the partition.

11. In the upper right corner, click  (**Menu**) and select **Save As**. Navigate to **Folders** ⇨ **My Folder** and name the report **VS_BankPartition**. Click **Save**.

End of Demonstration