Stepwise Linear Regression

Stepwise Linear Regression is a variant on classical Linear Regression in which variables are only included in the model if they have a significant effect on the variable we are trying to predict.

For instance, when trying to predict the price a house will sell for, we can try to include many variables such as the number of bathrooms, number of bedrooms, material of the house, size of the lot etc. But rather than using all these variables to try to predict the price of the house, we only want to include those that have a demonstrated effect on the price. Stepwise Regression automatically determines which variables from our dataset have an effect on price and returns predictions based on those variables.

The Stepwise Linear Regression script uses backward elimination based on the Akaike Information Criterion (AIC) [4]. By using "backward elimination", all variables are initially included in the model and are removed one by one until only the significant variables remain.

For more detailed information on Stepwise Regression, please consult the Stepwise Regression Wikipedia page [3].

This R Script has two functional modes:

- The user opts for stepwise linear regression.
- The user opts not to perform stepwise regression. In such a case, standard linear regression is performed and all independent variables are included in the model.

How to Deploy to MicroStrategy:

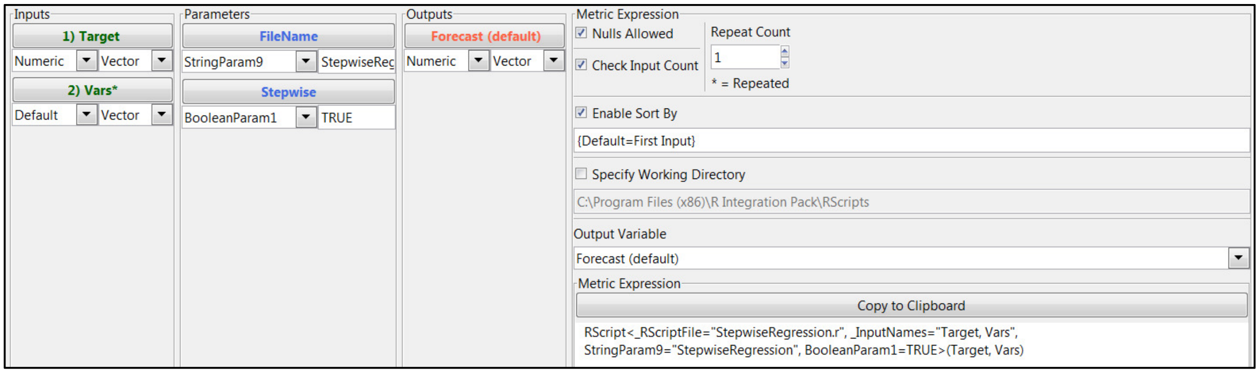
Prerequisite: Please follow the instructions in the R Integration Pack User Guide [1] for configuring your MicroStrategy environment with R and that the R Script functions have been installed in your MicroStrategy project(s).

1) Download the StepwiseRegression.R file from the Off-The-Shelf R Script Repository. [2].
2) From the R console, run the StepwiseRegression.R script to verify the script runs correctly. For details, see the "**Running from the R Console**" section below.
3) Cut-and-paste the appropriate metric expression below in any MicroStrategy metric editor. Map the arguments to the appropriate MicroStrategy metrics. A description of each output is shown below.
4) Use the new metric in reports, dashboards and documents.

Metric Expressions:

1) **Forecast:** Returns the forecasted values:
RScript<_RScriptFile="StepwiseRegression.r",_InputNames="Target,Vars",StringParam9="StepwiseRegression", BooleanParam1=TRUE>(Target, Vars)

Analytic Signature:



1) Target: Numeric, Vector
2) Vars*: Default, Vector
FileName: StringParam9, StepwiseReg
Stepwise: BooleanParam1, TRUE
Forecast (default): Numeric, Vector

☑ Nulls Allowed
☑ Check Input Count
Repeat Count: 1
* = Repeated
☑ Enable Sort By
{Default=First Input}
□ Specify Working Directory
C:\Program Files (x86)\R Integration Pack\RScripts
Output Variable: Forecast (default)

Metric Expression

```
RScript<_RScriptFile="StepwiseRegression.r", _InputNames="Target, Vars",
StringParam9="StepwiseRegression", BooleanParam1=TRUE>(Target, Vars)
```

Inputs:

Target: The variable that we are trying to predict.

Vars: The independent variables that are considered when creating the model and generating the prediction. Since the Vars argument is a repeated input, it can be mapped to any number of MicroStrategy metrics. In this way, we enable our algorithm to consider any number of variables when generating predictions.

Note: **Repeated inputs must all be the same type (i.e. all variables must be numeric or all variables must be strings)**

Note: **The _InputNames parameter must be the same length as the number of inputs. For instance, if there are 5 independent variables, then the _InputNames argument must reflect all six inputs (Target + 5 independent variables). Spaces will be automatically removed from the _InputNames.**

Parameters:

FileName: Uses StringParam9 with a default of "StepwiseRegression". This parameter specifies the name of both the .Rdata file used to persist relevant objects from the R environment such as the model and the data used to create it, allowing for additional inspection and analysis, and the PMML file that gets generated as part of script execution. Please note the R Script automatically appends the ".Rdata" file extension to this file name. If a value of "" is passed in, then no Rdata file is created.

Stepwise: Uses BooleanParam1 with a default of TRUE. If this value is set to TRUE, then stepwise regression will be performed and only the variables with a significant effect will be considered when generating predictions. If this value is set to FALSE, then traditional linear regression will be performed and all variables will be included in the model.

Outputs:

Forecast: A numeric value representing the predicted value.

Additional Results Generated by the R Script:

Two files are stored in the working directory:

Rdata File: This file persists the state of several objects from the R environment for later inspection, analysis, and reuse, including df (a data frame containing the data read in from MicroStrategy), model (the lm model object), and Forecast (the predictions).

Pmml File: This file contains PMML, an XML representation of our linear regression model. This file can be imported into MicroStrategy, which will create predictive metrics in MicroStrategy that can be used to score records contained in MicroStrategy.

Running from the R Console:

In addition to processing data from MicroStrategy during execution of a report or dashboard, the R script is also configured to run from the R console. Running the script for the R Console verifies that the script is functioning as expected, a good practice when initially deploying this analytic to a new system (for more details, see "Configuring dual execution modes" in [1]).

When run from the R Console, if the script is executing properly, a "Success!" message will appear in the console. If a "Success!" message does not appear, then please note the error in order to take appropriate action. For common pitfalls, please consult the **Troubleshooting** section below.

Troubleshooting:

This section covers certain situations you might encounter but it's not intended as a comprehensive list of possible errors:

If an error occurs, the report will either fail with an error message, or return nulls as the output. In these cases, please refer to the RScriptErrors.log file generated for further guidance and the DSSErrors.log. Please consult the User Guide [1] and the R documentation for additional guidance.

If Stepwise is set to TRUE, then the script will attempt to install the required R package. If the package is not successfully installed, you can install using the R console using the command:

```
    install.pacakges("MASS", repos="http://cran.rstudio.com/")
```

- **MASS**: This R package contains functions that support the execution of Stepwise Regression.

If a mix of string and numeric variables is passed in to the Vars argument, the report will fail with an error message indicating that a variable with unexpected type was passed in. This can be remedied by using all strings or all numeric variables corresponding to the Vars argument.

If an error message similar to "Duplicated name X in data frame" appears, that means that the _InputNames parameter does not have length corresponding to the number of inputs.

Example (using MicroStrategy Tutorial):

A retailer wants to understand the relationship between the profitability of a customer and the characteristics of that customer. Specifically, the retailer wants to determine how the following seven metrics affect profit:

a) Customer Age
b) Customer Gender

c) Household Count of Customer
d) Income Bracket of Customer
e) Marital Status of Customer
f) Whether the customer is new
g) Whether the customer is recently active

To do so, the retailer employs stepwise linear regression to determine which variables have a significant effect on profit. This can be done by creating and saving a report with the following design:

Rows (Attributes):

- **Customer**

Columns (All Metrics which are found in the folder **Public Objects -> Reports -> MicroStrategy Platform Capabilities -> MicroStrategy Data Mining Services -> Support Objects**):

- **Age Range ID**
- **Gender ID**
- **Household Count ID**
- **Income Bracket ID**
- **Household Count ID**
- **IsNewCustomer**
- **IsRecentCustomer**

Filter (Filter found in **Public Objects -> Reports -> MicroStrategy Platform Capabilities -> MicroStrategy Data Mining Services -> Sampling)**:

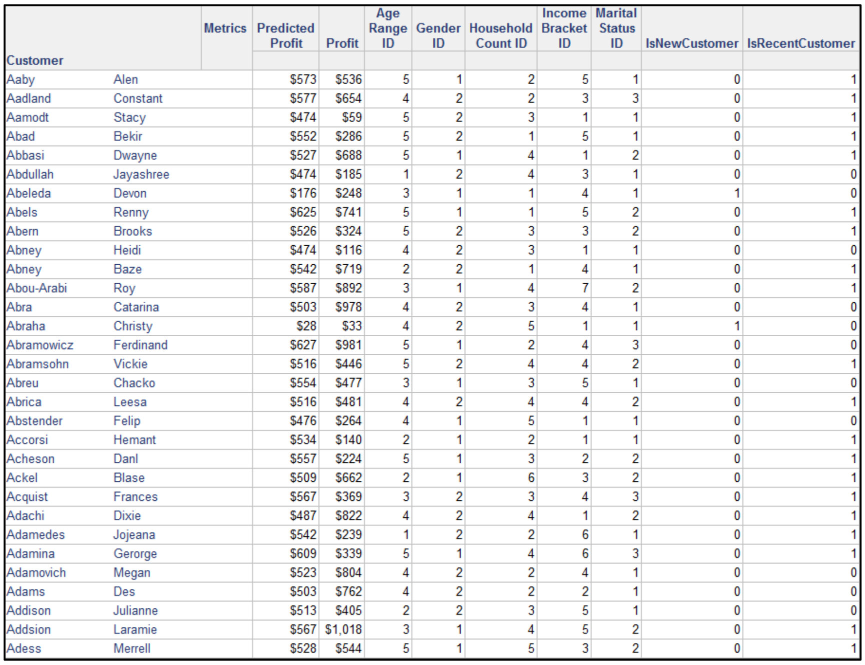
- **Add n-th record sample filter**

Run this report with a prompt answer of **4.**

Now, insert a derived metric named “Predicted Profit” with the following definition:

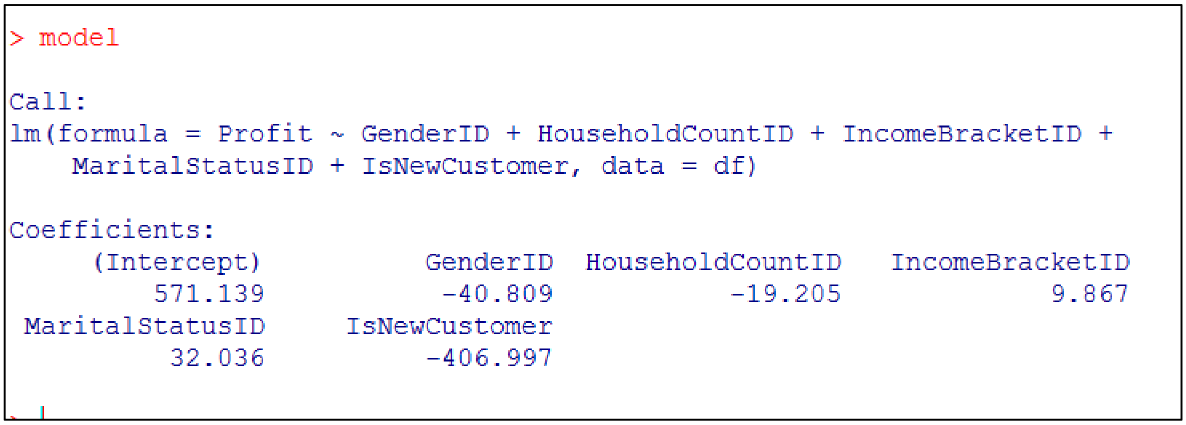
RScript<[BooleanParam1]=True,[StringParam9]="StepwiseRegression",[_RScriptFile]="StepwiseRegression.r", [_InputNames]="Profit, [Age Range ID], [Gender ID], [Household Count ID], [Income Bracket ID], [Marital Status ID], IsNewCustomer, IsRecentCustomer">(Profit, [Age Range ID], [Gender ID], [Household Count ID], [Income Bracket ID], [Marital Status ID], IsNewCustomer, IsRecentCustomer)

You should obtain the following results:



Customer		Metrics	Predicted Profit	Profit	Age Range ID	Gender ID	Household Count ID	Income Bracket ID	Marital Status ID	IsNewCustomer	IsRecentCustomer
Aaby	Alen		$573	$536	5	1	2	5	1	0	1
Aadland	Constant		$577	$654	4	2	2	3	3	0	1
Aamodt	Stacy		$474	$59	5	2	3	1	1	0	1
Abad	Bekir		$552	$286	5	2	1	5	1	0	1
Abbasi	Dwayne		$527	$688	5	1	4	1	2	0	1
Abdullah	Jayashree		$474	$185	1	2	4	3	1	0	0
Abeleda	Devon		$176	$248	3	1	1	4	1	1	0
Abels	Renny		$625	$741	5	1	1	5	2	0	1
Abern	Brooks		$526	$324	5	2	3	3	2	0	1
Abney	Heidi		$474	$116	4	2	3	1	1	0	0
Abney	Baze		$542	$719	2	2	1	4	1	0	1
Abou-Arabi	Roy		$587	$892	3	1	4	7	2	0	1
Abra	Catarina		$503	$978	4	2	3	4	1	0	0
Abraha	Christy		$28	$33	4	2	5	1	1	1	0
Abramowicz	Ferdinand		$627	$981	5	1	2	4	3	0	0
Abramsohn	Vickie		$516	$446	5	2	4	4	2	0	1
Abreu	Chacko		$554	$477	3	1	3	5	1	0	0
Abrica	Leesa		$516	$481	4	2	4	4	2	0	1
Abstender	Felip		$476	$264	4	1	5	1	1	0	0
Accorsi	Hemant		$534	$140	2	1	2	1	1	0	1
Acheson	Danl		$557	$224	5	1	3	2	2	0	1
Ackel	Blase		$509	$662	2	1	6	3	2	0	1
Acquist	Frances		$567	$369	3	2	3	4	3	0	1
Adachi	Dixie		$487	$822	4	2	4	1	2	0	1
Adamedes	Jojeana		$542	$239	1	2	2	6	1	0	1
Adamina	Gerorge		$609	$339	5	1	4	6	3	0	1
Adamovich	Megan		$523	$804	4	2	2	4	1	0	0
Adams	Des		$503	$762	4	2	2	2	1	0	0
Addison	Julianne		$513	$405	2	2	3	5	1	0	0
Addsion	Laramie		$567	$1,018	3	1	4	5	2	0	1
Adess	Merrell		$528	$544	5	1	5	3	2	0	1

Now, navigate to your default R scripts folder (usually C:/Program Files (x86)/R Integration Pack/R Scripts) and locate the StepwiseRegression.Rdata file that was created. Double click on the file and open it with an R program. Once it is open enter the command "**model**". The result should be:



```
> model

Call:
lm(formula = Profit ~ GenderID + HouseholdCountID + IncomeBracketID + 
    MaritalStatusID + IsNewCustomer, data = df)

Coefficients:
     (Intercept)           GenderID  HouseholdCountID   IncomeBracketID  
         571.139            -40.809           -19.205             9.867  
 MaritalStatusID      IsNewCustomer  
          32.036           -406.997  
```

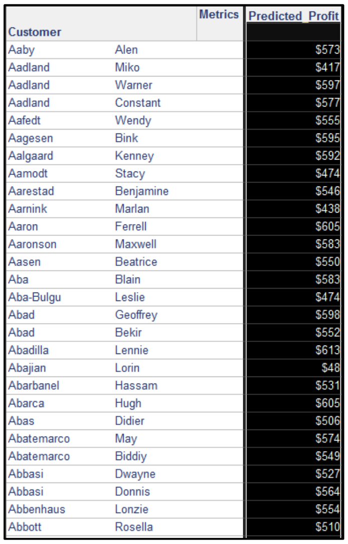
Observe that the stepwise regression model found that just five of the seven independent variables were significant. The stepwise procedure determined that **IsRecentCustomer** and **Age Range ID** did not have a significant effect on the profitability of a customer.

Since our script also saves the model as PMML, we can import the PMML to MicroStrategy and create predictive metrics that leverage the model to score existing customers within MicroStrategy. If you have access to MicroStrategy Developer, open Developer as an administrator and login to Tutorial. Then, click on **Tools->Import Data Mining Model.** Browse to your default R scripts folder and click on the StepwiseRegression.PMML file that was created as part of the script. Choose the default predicted value output and choose a folder to store the predictive metrics that will be created. Map the four input metrics to the four metrics in the folder **Public Objects -> Reports -> MicroStrategy Platform Capabilities -> MicroStrategy Data Mining Services -> Support Objects** as follows:

a) GenderID = Gender ID

b) HouseholdCountID = Household Count ID
c) IncomeBracketID = Income Bracket ID
d) MaritalStatusID = Marital Status ID

A new predictive metric named "Predicted Profit" should be created in the folder you selected. You can now create a report with Customer Rows and Predicted Profit on Columns to see the predicted profit for each customer. The report should look as follows:



	Metrics	Predicted Profit
Customer		
Aaby	Alen	$573
Aadland	Miko	$417
Aadland	Warner	$597
Aadland	Constant	$577
Aafedt	Wendy	$555
Aagesen	Bink	$595
Aalgaard	Kenney	$592
Aamodt	Stacy	$474
Aarestad	Benjamine	$546
Aarnink	Marlan	$438
Aaron	Ferrell	$605
Aaronson	Maxwell	$583
Aasen	Beatrice	$550
Aba	Blain	$583
Aba-Bulgu	Leslie	$474
Abad	Geoffrey	$598
Abad	Bekir	$552
Abadilla	Lennie	$613
Abajian	Lorin	$48
Abarbanel	Hassam	$531
Abarca	Hugh	$605
Abas	Didier	$506
Abatemarco	May	$574
Abatemarco	Biddiy	$549
Abbasi	Dwayne	$527
Abbasi	Donnis	$564
Abbenhaus	Lonzie	$554
Abbott	Rosella	$510

References:

1) MicroStrategy R Integration Pack User Guide: https://rintegrationpack.codeplex.com/documentation
2) R Integration Pack Analytic Shelf: Off-The-Shelf R Script Repository: https://rintegrationpack.codeplex.com/wikipage?title=R%20Script%20%22Shelf%22&referringTitle=Home#
3) Stepwise Regression Wikipedia Page: http://en.wikipedia.org/wiki/Stepwise_regression
4) Akaike Information Criterion Wikipedia Page: http://en.wikipedia.org/wiki/Akaike_information_criterion