# Predictive Modelling I – Regression

**Scenario:**

ABC Departmental Store is a well-known local departmental store. ABC runs advertisement on their products on local media (TV, radio, newspapers), social media, billboard and roadshows. In order to know the effectiveness of advertising campaigns, ABC monitors sales after each campaign as well as the amount spent on each campaign.

ABC wants to determine how advertising spending is related to sales performance, so that it can predict future sales performance, given advertising budget.

## Theory

**Learning Activity 1 – How does simple linear regression work**

For our scenario, identify the following:

|  |  |
| --- | --- |
| Dependent variable, *y* (aka the Target) |  |
| Independent variable, *x* (aka the Input) |  |

**Takeaway**:

|  |
| --- |
| In the regression equation, y = a + bx, y is the target, x is the input. |

**Learning Activity 2 – Least Square Method**

State whether the following statement is true or false.

|  |  |
| --- | --- |
| **Statement** | **True / False** |
| Error = Actual value – Predicted value | TRUE |
| Value of error is always positive | FALSE |
| Squaring error ensures they do not cancel each other out when we sum up all the error | TRUE |
| Best fit line aims to minimize the sum of square of all errors | TRUE |

**Takeaway**:

|  |
| --- |
|  |

**Learning Activity 3 – Evaluating Regression Model**

State whether the following statement is true or false.

|  |  |
| --- | --- |
| **Statement** | **True / False** |
| P-value < 0.05 indicates input is useful | TRUE |
| High R2value indicates better model | TRUE |
| High RMSE indicates better model | FALSE |
| The training partition provides a truer test of the model performance | FALSE |

**Takeaway**:

|  |
| --- |
|  |

## Hands-on

About the data:

We are using a fictitious dataset containing past advertising campaigns at ABC Department Store. The dataset consists of 500 rows (each row represents a past campaign) and 8 columns:

* CampaignID – unique identifier for each past advertising campaign
* TV – amount (in thousand) spent on TV ads
* Radio – amount (in thousand) spent on radio ads
* Newspaper – amount (in thousand) spent on newspaper ads
* SocialMedia – amount (in thousand) spent on social media ads
* RoadShow – amount (in thousand) spent on road shows
* BillBoard – amount (in thousand) spent on bill board ads
* Sales – sales performance (in thousand) observed after each campaign

1. Create a new SAS EM Project, you can name it after the topic, call it “Predictive Modeling”

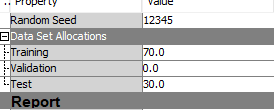
* Log in to the virtual machine, click on SAS Enterprise Miner Workstation 14.1
* Select New Project
* Enter “Predictive Modeling” as the project name
* Choose to save the project to Desktop in the virtual machine. We will copy the entire project folder back to our own machine at the end of this practical.

1. Create a new SAS EM Diagram and read the file *Advertising.csv* into SAS EM using the **File Import** Node

* File > New Diagram (name it *LinearRegression* or something meaningful)
* Add File Import node (found in Sample group) into your diagram
  + Delimiter = comma (,)
  + Name Row = Yes
  + Table Role = Train
* Click on the ellipsis on Train > Variables and set the Data role / level of measurement as shown below:

|  |  |  |
| --- | --- | --- |
| **Data Field** | **Role** | **Level of Measurement** |
| BillBoard | Input | Interval |
| CampaignID | ID | Nominal |
| Newspaper | Input | Interval |
| Radio | Input | Interval |
| RoadShow | Input | Interval |
| Sales | Target | Interval |
| SocialMedia | Input | Interval |
| TV | Input | Interval |

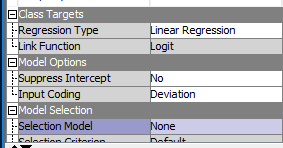
1. Connect a **Data Partition** node (found in Sample group) downstream from the File Import node. In the properties pane, under the Data Set Allocations group, set Training to be 70 (70%) and Test to be 30 (30%).

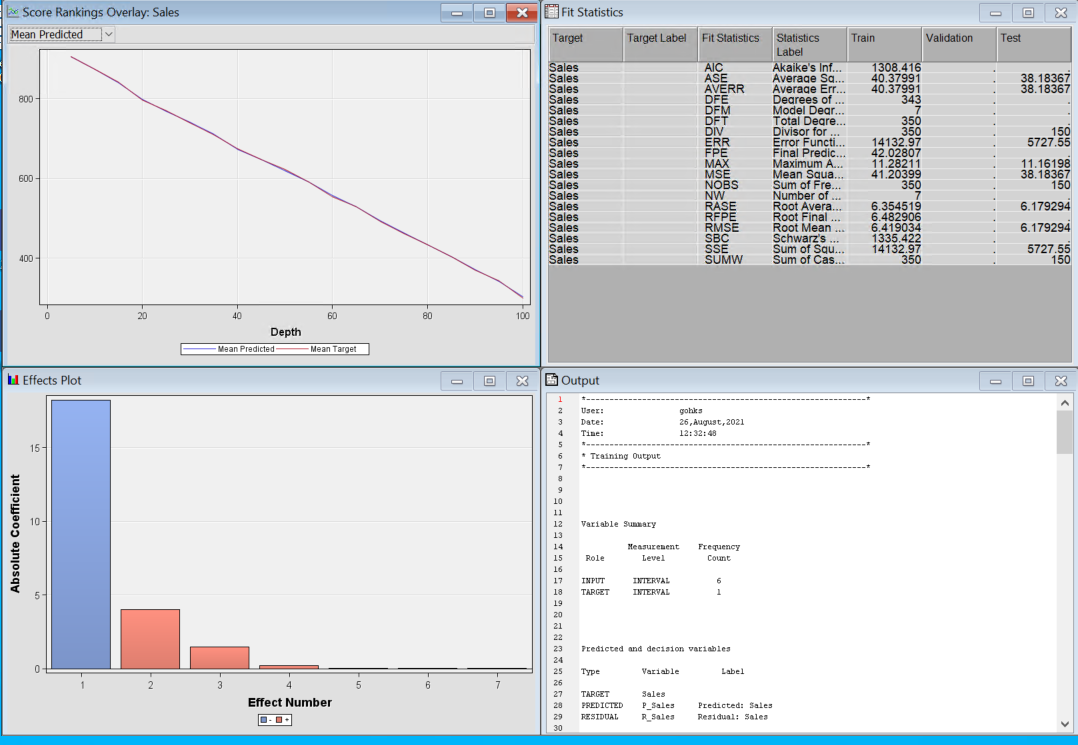


1. Connect a **Regression** node (found in Model group) downstream from the **Partition** node.

In the properties pane, select Linear Regression as the Regression Type.

Run the **Regression** node and view the results.





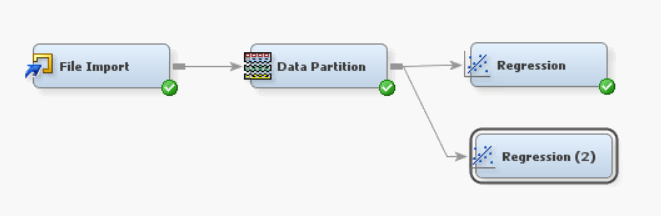
Expand the Effects Plot window. This shows you the relative importance of inputs in helping to predict the target variable.

Expand the Fit Statistics window. You can see the Root Mean Square Error (RMSE) of the model, take note of the RMSE for both the Train and Test partition.

Expand the Output window. Look for the P-value for each of the inputs (line 69). Look for the value of R2 (line 61)

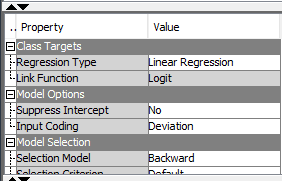
Close the results window.

1. Add another **Regression** node into your diagram. Connect the Partition node to it. Your diagram should resemble the one shown below.



As per before, select Linear Regression as the Regression Type.

Modify the Selection Model method to be “*Forward*”.

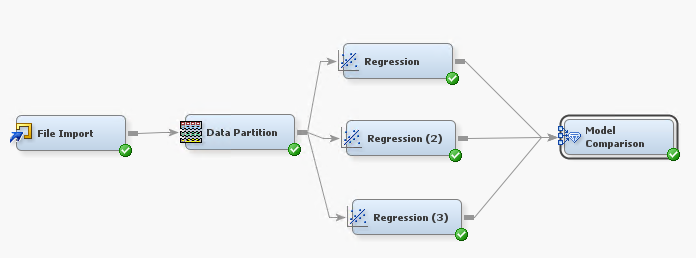


Run this second **Regression** node.

In the results window, notice how fewer inputs were included in the final model.

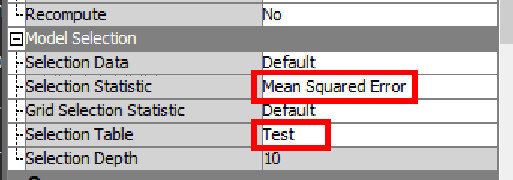
1. Repeat step 5, but this time, set Selection Model for the **Regression** node to be *‘Backward*. You should now have 3 regression nodes in your diagram.
2. Add a **Model Comparison** node (found in the Assess group) to your diagram.

Connect all three Regression nodes to it.



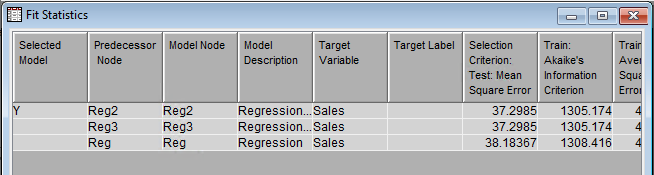
In the Properties pane, set the Selection Statistic to be *Mean Squared Error*.

Set the Selection Table to be *Test*.



The **Model Comparison** node will rank the three regression models based on the selected criteria, in this case, the Mean Squared Error.

In the results window under Fit Statistics, we can see that Regression Model 2 (the one with *Backward* selection) is ranked as the better performing mode.



Close the results window.

**Question**

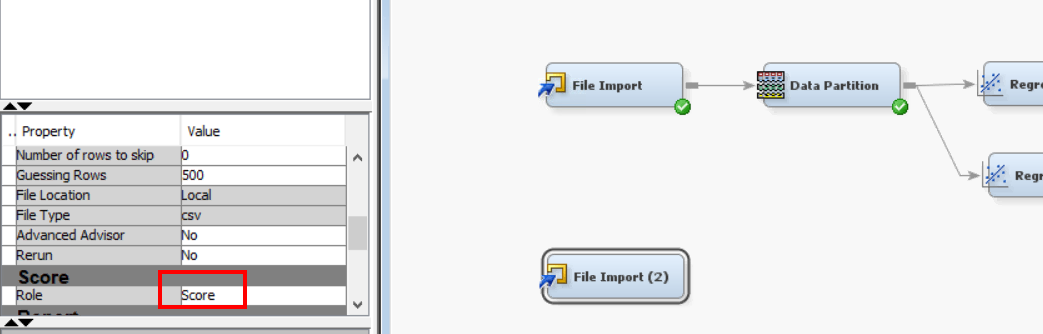
How can we confirm that Reg2 performed better than Reg?

|  |
| --- |
|  |

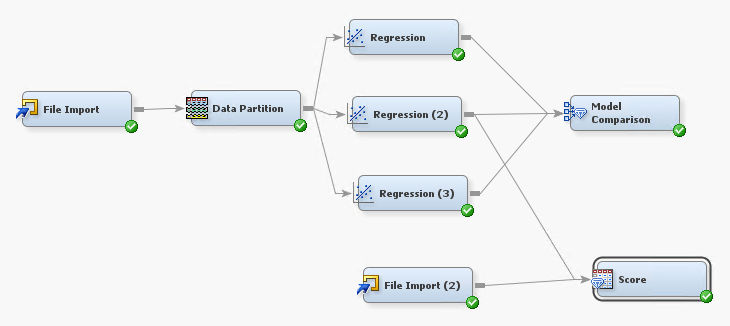
1. Let’s go ahead and get our regression model to predict a few cases for us.

Add a second **File Import** node to your diagram. Read in the file *Advertising Score.csv*

Set the Role in the property pane to be *Score*.



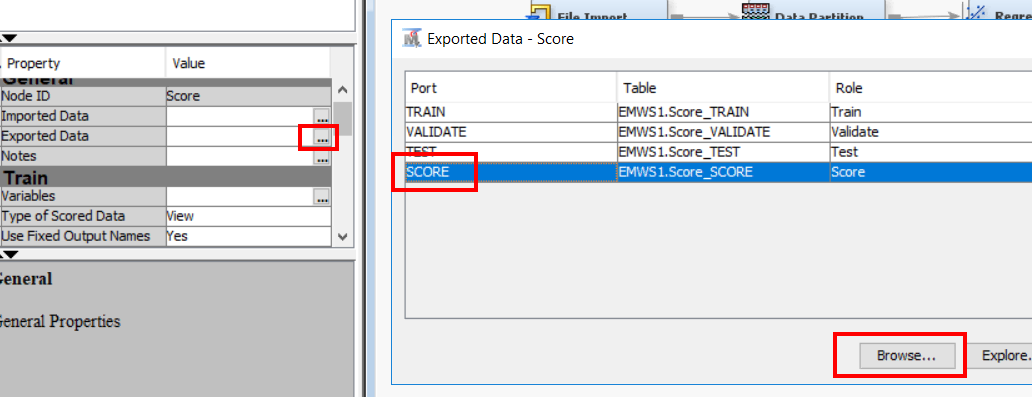
1. Add a **Score** node (found in Assess group) to your diagram. Connect the **Regression (2)** and **File Import (2)** nodes to the **Score** node.



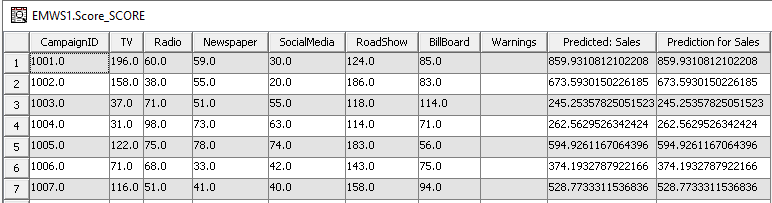
1. Run the **Score** node.

In the property pane of the **Score** node, locate the Exported Data item, click on the ellipsis.

In the resulting window, select SCORE port, then click *Browse* button.



This will display the predicted Sales values, carried out by Regression (2) model (the one with backward selection)



## Interpretation and Application

What is the predicted sales if we plan to run an advertising campaign with the following budget?

TV: $145 (thousands)

Radio: $46 (thousands)

Newspaper: $52 (thousands)

Social Media $51 (thousands)

Road Show $114 (thousands)

Bill Board $62 (thousands)

[Hint, look at the output from previous step]

|  |
| --- |
| Answer: |

Based on the effects plot of the champion regression model, what are the important factors influencing Sales performance for ABC Department Store?

|  |
| --- |
| Answer: |

## Summary

1. Linear regression allows us to build a predictive model based on historical data.
2. Data Partitioning allows us to hold-out some past data so that we can have a truer test of model’s learning.
3. Linear regression uses the least square criterion to find the best fit line.
4. Using the linear regression equation, we can plug in the various values to X to predict the value of Y.
5. Regression uses numeric inputs to predict a numeric target.

## Resources for further reading

1. Statistics 101: Simple Linear Regression, The Very Basics

[A YouTube video explaining the basics of Simple Linear Regression – 22 minutes]

<https://www.youtube.com/watch?v=ZkjP5RJLQF4>

You can continue on the other videos in the playlist if you are keen

1. Assumptions of Linear Regression

[Article explaining the underlying assumptions of linear regression]

<https://towardsdatascience.com/assumptions-of-linear-regression-algorithm-ed9ea32224e1>

1. Variable Selection in Multiple Regression

[Article explaining the difference between forward, backward, stepwise selection]

<https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-multiple-regression/variable-selection.html>

1. Partition Your Data for Predictive Modelling

[Short article explaining the purpose of data partitioning]

<https://amadeus.co.uk/tips/partition-your-data-for-predictive-modelling/>