5-1 Exercise 5: JMP for Supervised Learning

Joel Meza

Professor Torre Roenne

April 2, 2021

DAT 220 -Fundamentals of Data Mining

Southern New Hampshire University

**Simple Linear Regression**

***Figure 1:***

**Bivariate Fit of Sales ($M) By # Employ**

Chart, line chart

Description automatically generated

A picture containing icon

Description automatically generated

**Linear Fit**

Sales ($M) = -603.6937 + 0.1579073\*# Employ

**Summary of Fit**

|  |  |
| --- | --- |
| RSquare | 0.947496 |
| RSquare Adj | 0.945746 |
| Root Mean Square Error | 2557.529 |
| Mean of Response | 5433.856 |
| Observations (or Sum Wgts) | 32 |

**Lack Of Fit**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Ratio** |
| --- | --- | --- | --- | --- |
| Lack Of Fit | 29 | 195409235 | 6738249 | 8.2241 |
| Pure Error | 1 | 819328 | 819328 | **Prob > F** |
| Total Error | 30 | 196228563 |  | 0.2702 |
|  |  |  |  | **Max RSq** |
|  |  |  |  | 0.9998 |

**Analysis of Variance**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Ratio** |
| --- | --- | --- | --- | --- |
| Model | 1 | 3541202728 | 3.5412e+9 | 541.3895 |
| Error | 30 | 196228563 | 6540952.1 | **Prob > F** |
| C. Total | 31 | 3737431292 |  | <.0001\* |

**Parameter Estimates**

| **Term** | **Estimate** | **Std Error** | **t Ratio** | **Prob>|t|** |
| --- | --- | --- | --- | --- |
| Intercept | -603.6937 | 521.2823 | -1.16 | 0.2560 |
| # Employ | 0.1579073 | 0.006787 | 23.27 | <.0001\* |

***Figure 2:***

**Bivariate Fit of Profits ($M) By # Employ**

Chart, scatter chart

Description automatically generated

A picture containing icon

Description automatically generated

**Linear Fit**

Profits ($M) = 48.238652 + 0.0094438\*# Employ

**Summary of Fit**

|  |  |
| --- | --- |
| RSquare | 0.671067 |
| RSquare Adj | 0.660103 |
| Root Mean Square Error | 454.913 |
| Mean of Response | 409.3188 |
| Observations (or Sum Wgts) | 32 |

**Lack Of Fit**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Ratio** |
| --- | --- | --- | --- | --- |
| Lack Of Fit | 29 | 6205673.9 | 213989 | 79.2221 |
| Pure Error | 1 | 2701.1 | 2701 | **Prob > F** |
| Total Error | 30 | 6208375.0 |  | 0.0887 |
|  |  |  |  | **Max RSq** |
|  |  |  |  | 0.9999 |

**Analysis of Variance**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Ratio** |
| --- | --- | --- | --- | --- |
| Model | 1 | 12665910 | 12665910 | 61.2040 |
| Error | 30 | 6208375 | 206945.83 | **Prob > F** |
| C. Total | 31 | 18874286 |  | <.0001\* |

**Parameter Estimates**

| **Term** | **Estimate** | **Std Error** | **t Ratio** | **Prob>|t|** |
| --- | --- | --- | --- | --- |
| Intercept | 48.238652 | 92.72158 | 0.52 | 0.6067 |
| # Employ | 0.0094438 | 0.001207 | 7.82 | <.0001\* |

From both figures illustrated above, we are given a model which in connection from each other are the two continuous variables. While using the JMP platform, this is created in Analyze > Fit Y by X, with the X for numbers of Employees and Y for numbers in Sales. Both queries are used to demonstrate valuable information in detail to the red line from the Bivariate Fit graph shown on figure two. Also, this graph demonstrates how Sales has a more linear connection to the number of employees rather from profit numbers.

**Simple Logistics Regression**

***Figure 3***

**Logistic Fit of Type By Profits ($M)**

Chart, scatter chart

Description automatically generated

**Parameter Estimates**

| **Term** | **Estimate** | **Std Error** | **ChiSquare** | **Prob>ChiSq** |
| --- | --- | --- | --- | --- |
| Intercept | 0.88299514 | 0.4574847 | 3.73 | 0.0536 |
| Profits ($M) | -0.0009002 | 0.0006549 | 1.89 | 0.1693 |

For log odds of Computer/Pharmaceutical

**Covariance of Estimates**

Cov

|  | **Intercept** | **Profits ($M)** |
| --- | --- | --- |
| Intercept | 0.2093 | -0.000 |
| Profits ($M) | -0.000 | 0.0000 |

***Figure 4***

**Logistic Fit of Size Co By profit/emp**

Chart, scatter chart

Description automatically generated

**Parameter Estimates**

| **Term** | **Estimate** | **Std Error** | **ChiSquare** | **Prob>ChiSq** |
| --- | --- | --- | --- | --- |
| Intercept[big] | -0.5659769 | 0.4977443 | 1.29 | 0.2555 |
| profit/emp[big] | -8.0483e-7 | 2.3415e-5 | 0.00 | 0.9726 |
| Intercept[medium] | -1.0122556 | 0.5865665 | 2.98 | 0.0844 |
| profit/emp[medium] | 1.33035e-5 | 2.5058e-5 | 0.28 | 0.5955 |

For log odds of big/small, medium/small

**Covariance of Estimates**

Cov

|  | **Intercept[big]** | **profit/emp[big]** | **Intercept[medium]** | **profit/emp[medium]** |
| --- | --- | --- | --- | --- |
| Intercept[big] | 0.2477 | -0.000 | 0.0897 | -0.000 |
| profit/emp[big] | -0.000 | 0.0000 | -0.000 | 0.0000 |
| Intercept[medium] | 0.0897 | -0.000 | 0.3441 | -0.000 |
| profit/emp[medium] | -0.000 | 0.0000 | -0.000 | 0.0000 |

From Figures 3 and 4 shown above, theses graphs and readings are indicators with calculations in possibility of series from event. While using the JMP platform, is used by Analyze through Fit Y by X was selected with X = with its category and Y = the profit from figure three given. Next, was a query which was created on X = the size of the company and Y = the profit as demonstrated in figure four given. Prior to the graph shown on figure three the greater possibility that is more profitable will be produced by a pharmaceutical company than a computer one. While on figure four the graph shown is greater in possibility than the lower sized companies which has a more profitable over employee proportion.

**Clustering**

***Figure 5***

**Hierarchical Clustering**

Method = Ward

**Dendrogram**

Diagram

Description automatically generated

For the above illustration it demonstrates what clustering is used for which are groups of rows with relevant characteristics collectively. While using the JMP platform, the mode is used as Analyze through Clustering through Hierarchical Cluster within application. Prior to continuous variables, they were chosen in comparison and labeled the type along with the profit which was inserted for Y axis. From the above illustration, it shows the clusters of four which matches the data in colors where the bigger cluster is red. This data shown differentiates between the bigger cluster between the computer companies having furthermore than the pharmaceutical companies have in their cluster.