Cairo University Data Analytics

Faculty of Computers and Artificial Intelligence Winter 2022

Operations Research and Decision Support Department Lab #6

**Lab Objectives**

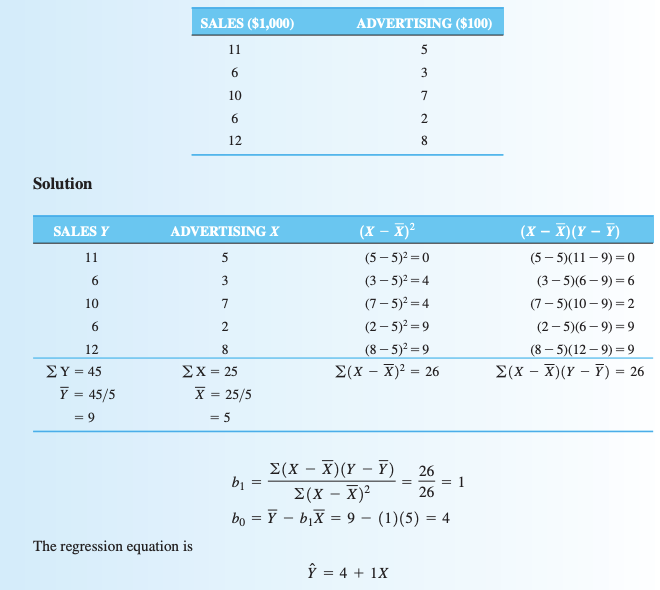
* **Regression Analysis**
* Regression analysis is a valuable tool that can be used to:
  + Understand the relationship between variables.
  + Predict the value of one variable based on another.
* Types of regression models:
  + Simple linear regression models contain only two variables.
  + Multiple regression models have more variables.
* Important terminologies:
  + The variable to be predicted is called the: dependent or response variable.
  + The other variable is called the: independent, explanatory or predictable variable.
* Scatter diagrams (plots) are a graphical way to investigate the relationship between variables:
  + The independent variable is normally plotted on the X-axis.
  + The dependent variable is normally plotted on the Y-axis.
* **Simple Linear Regression Model:**
* Regression models are used to test, if there is a relationship between variables.

𝒀=𝜷𝟎+𝜷𝟏𝑿

* + - 𝒀 = dependent variable (response) or Actual value “Sales of ice cream”
    - 𝑿 = independent variable (explanatory or predictable) “temp”
    - 𝜷𝟎 = intercept “value of Y when X is zero”
    - 𝜷𝟏 = slope of the regression line “unit increase in Y given unit increase in X”

**Question 1:**

Judith Thompson runs a florist shop on the Gulf Coast of Texas, specializing in floral arrangements for weddings and other special events. She advertises weekly in the local newspapers and is considering increasing her advertising budget. Before doing so, she decides to evaluate the past effectiveness of these ads. Five weeks are sampled, and the advertising dollars and sales volume for each of these is shown in the following table.



Build the Regression Model:

* Draw a scatter diagram
* Develop a regression equation that would help Judith evaluate her advertising
* Use the model to predict sales if the advertising budget is increased to 30

**Answer:**

Y = 4 + 1 X

The sales when the advertising budget is increased to 30 is: Y = 4 + (1) (30) = 34

* **Measuring the fit of the regression model:**
* Coefficient of Determination (𝒓𝟐):
  + - The proportion of the variability in Y explained by regression equation
* Correlation Coefficient (𝒓): >0.7 strong <0.5 weak
  + - An expression of the strength of the linear relationship between the variables.
    - It will always be between +1 and -1

𝒓= ± √𝒓𝟐



**Question 1 Cont.:**

* Measure how the model is fit by getting r2 and r.

**Answer:**

About 81% of the variability in sales can be explained by the regression model with advertising as the independent variable

𝑟=0.901 indicating that there is a very strong positive (directly proportional) linear relationship between the sales and the advertising

**Question 2:**

Given the following Scatter Diagram and Linear Regression Line:



Which of the following statements are true?

1. r = 0

2. r = 1

3. r = -1

4. 𝑟2 = 100%

5. 𝑟2 = 0

**Answer:**

3 & 4

* **Analysis of Variance (ANOVA) Table:**



* **3 Measures of Variability:** 
  + **SST: Sum of the squares total** [Total variability about the mean (Model Variability)]



* + **SSE: Sum of the squared error [**Variability about the regression line (Unexplained Variability)]

****

* + **SSR: Sum of squares due to regression [**Total variability that is explained by the regression model (Explained Variability)]



* **Important relationship: SST = SSR + SSE**
* **Therefore, we can conclude that: r2 var reg=**

**Question 1 Cont.:**

* Calculate r2 using the ANOVA table

**Answer:**

**r2 =**

* **Regression Model Significance**
* The F-test of overall significance is performed to indicate whether the resulting linear regression model is a true representation of the population or not (to avoid sampling error)

𝒀^=𝜷𝟎+𝜷𝟏𝑿

* + Hypothesis Testing:
    1. H0: population slope coefficient 𝜷𝟏 = 0
    2. H1: population slope coefficient 𝜷𝟏 ≠ 0
  + If (𝜷𝟏 = 0) the null hypothesis is that there is no relationship between Y and X
  + The alternate hypothesis is that there is a linear relationship (𝜷𝟏 ≠ 0)
  + If the null hypothesis is rejected, then the independent variable is responsible in the variation in Y
* An F-statistic is the ratio of two variances,

* MSR (mean squared regression)
* MSE (mean squared error)
  + Whenever the F value is large, the significance level (p-value) will be low, indicating that the model is useful!
* = value from F distribution table
  + df1= k (1)
  + df2= n-k-1 (5-1-1=3)
  + k is number of independent variables, n is the sample size
* = Probability ()
* Alpha or α, is the significance level
* A significance level of 0.05 signifies a 5% risk of deciding that an effect exists when it does not exist (95% confidence)
  + You can find F and p-Value in the ANOVA table

**Question 3:**

The dataset "Cereal" contains, among other variables, the consumer reports ratings of 77 cereals available in many grocery stores and the number of grams of sugar contained in each serving.

Considering "Sugars" as the explanatory variable and "Rating" as the response variable, generate the regression model and the ANOVA table. Comment on the results. (Let alpha = 0.05)

**Answer:**

* **Rating = 59.3 - 2.40 Sugars**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Degrees of  Freedom** | **Sum of  Squares** | **Mean of  Squares** | **F** | **p-Value** |
| ***ANOVA Table*** |
| **Explained** | 1 | 8654.737293 | 8654.737293 | 102.3492334 | < 0.0001 |
| **Unexplained** | 75 | 6342.063107 | 84.56084143 |  |  |

* **In the ANOVA table, the**
* **From the F table,**
* **Since** , **therefore our model is significant (there is strong evidence that** 𝜷𝟏 **is not equal to zero)**
* **Another Solution: In the ANOVA table, the**
* **Since** , **therefore our model is significant**
* **r = 0.759, therefore there is a strong positive relationship**
* **The r² term = (0.7592) = 0.577 indicating that 57.7% of the variability in the response variable “rating” is explained by the explanatory variable “sugar”.**