BUDT 730

Data, Models and Decisions

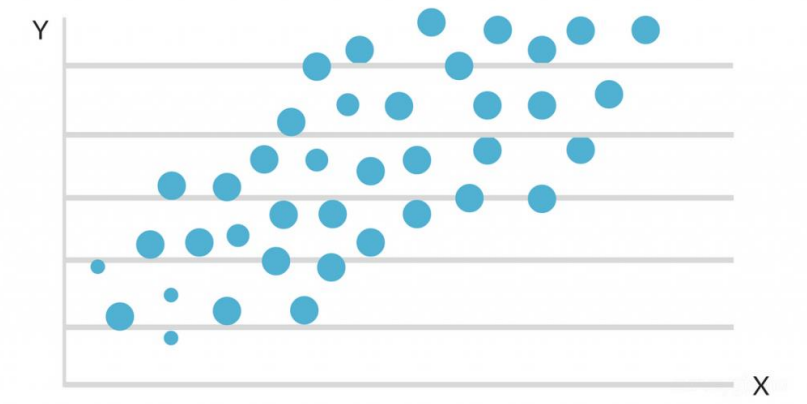
Lecture 9

Regression Analysis (I)

Prof. Sujin Kim

Learning Objective

- Introduce linear regression as a study of relationships between variables
- Example:
  - `Catalog Marketing_Reg.xlsx`



Overview of Linear Regression Analysis

Introduction to Linear Regression

Regression analysis is the study of relationships between variables

It is one of the most useful tools for a business analyst because it applies to many situations

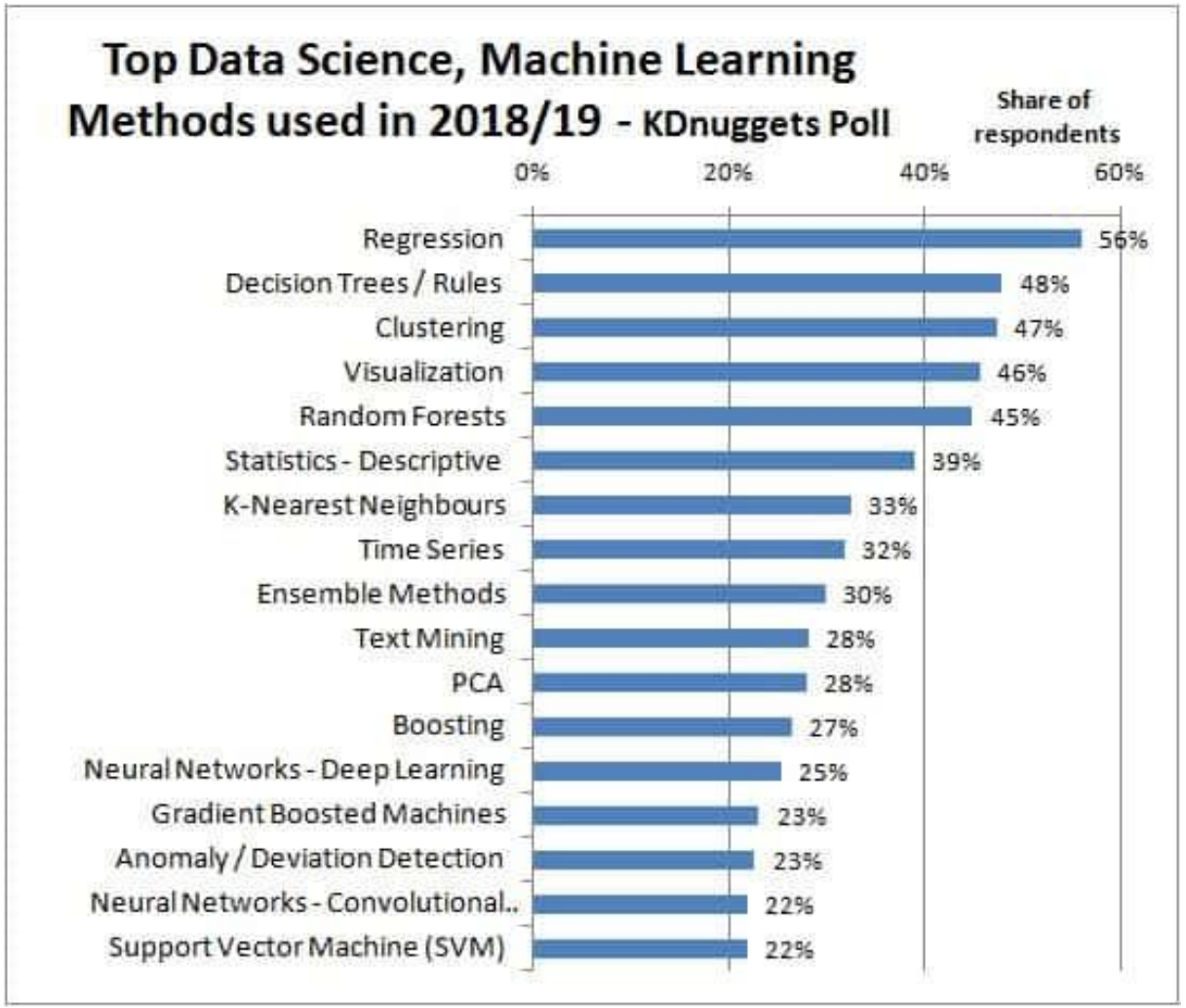
Regression is used for two primary purposes:

Explanatory model: Explain why and how a thing works

Predictive model: Estimate what will happen for future observations

Regression can be applied to cross-sectional or time series data

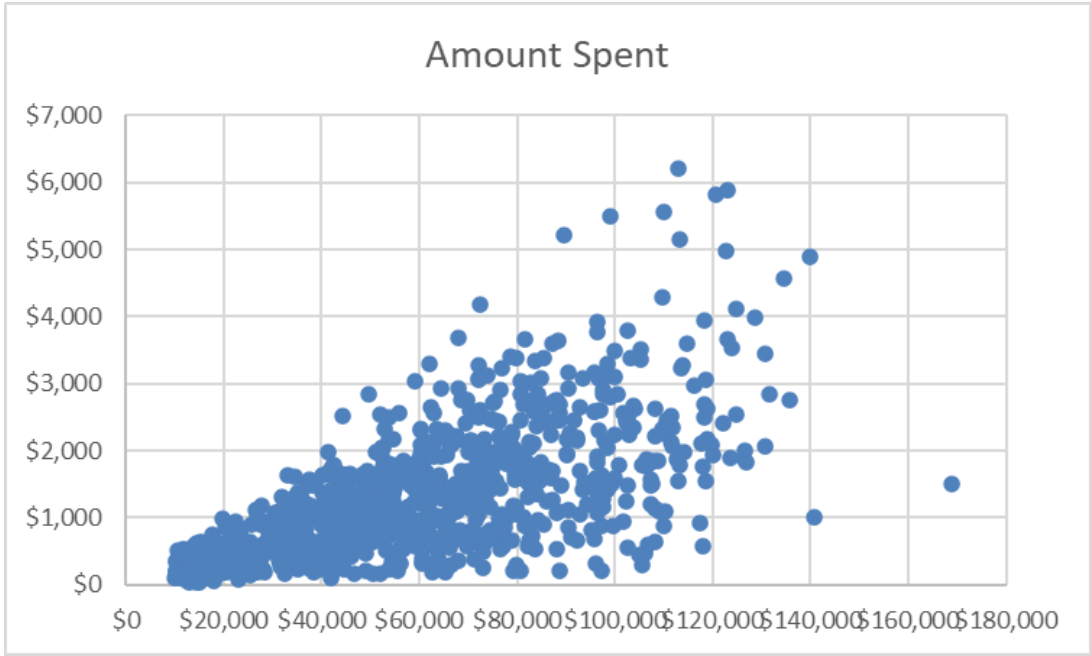
Statistical Methods for Business Analytics



Article

Recall: HyTex Catalog Marketing Data

- We can use scatterplots to visualize the relationship between “ Amount Spent” and “Salary”
- What can you say about the relationship between a customer’s salary and the amount he/she spends?
- We can use association measures to quantify the relationship between two variables



Measuring Associations

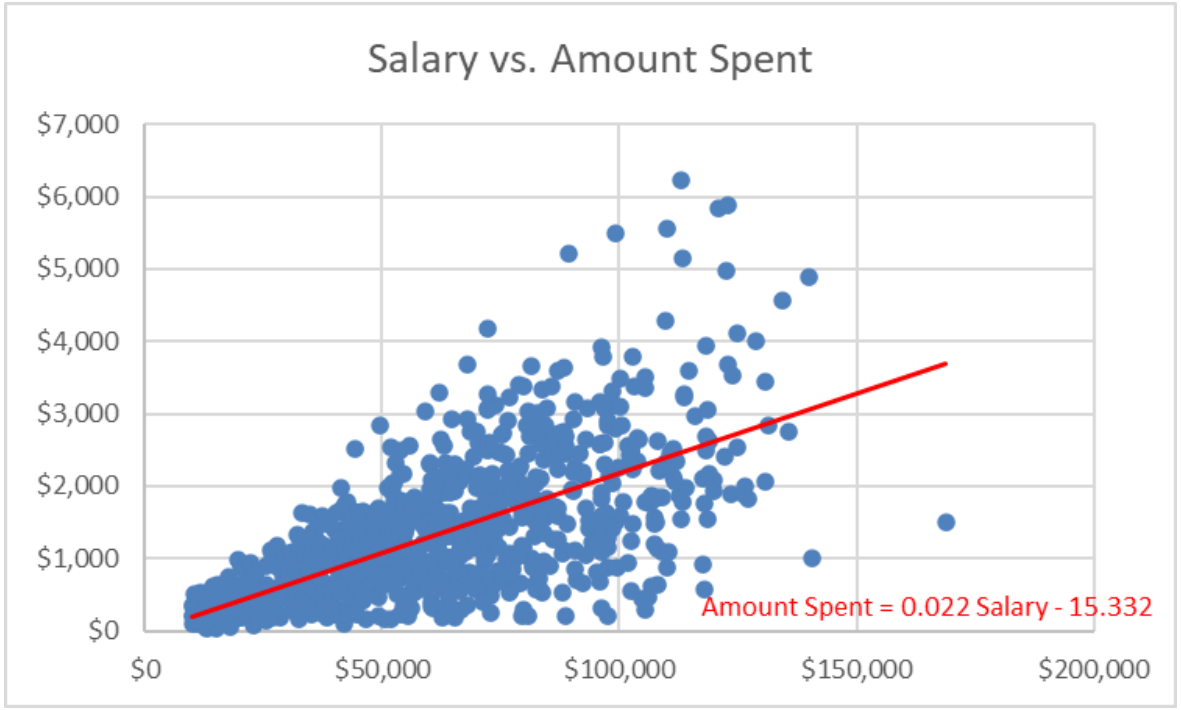
- A Scatterplot tells us:
  - that there is some relationship between Salary and Amount Spent
- The Correlation coefficients tells us:
  - how strong the linear relationship is between Salary and Amount Spent
- However, we still don't know the precise relationship
  - For example, we still don't know exactly by how much Amount Spent will increase for each additional $ amount of increase in Salary
- We would like to mathematically express the relationship between Salary and Amount Spent => Regression Analysis

Announcements

- Download:
  - Catalog marketing data and r script
  - Wine data

- Pick up the worksheet

Example: HyTex Catalog Marketing Data



- Linear regression quantifies the relationship between X (Salary) and Y (Amount Spent) variables.
- For a fixed change in X, how does Y change?

Amount Spent
= 0.022 * Salary-15.332

Linear Regression Models

- The **dependent variable (Y)** is the variable that we are trying to explain or predict.
  - Also called the **response** or target variable
- We use one or more **independent variables (X)** to help explain or predict the dependent variable
  - Also called **explanatory** or **predictor** variables

Linear Regression Models (10/20 (W))

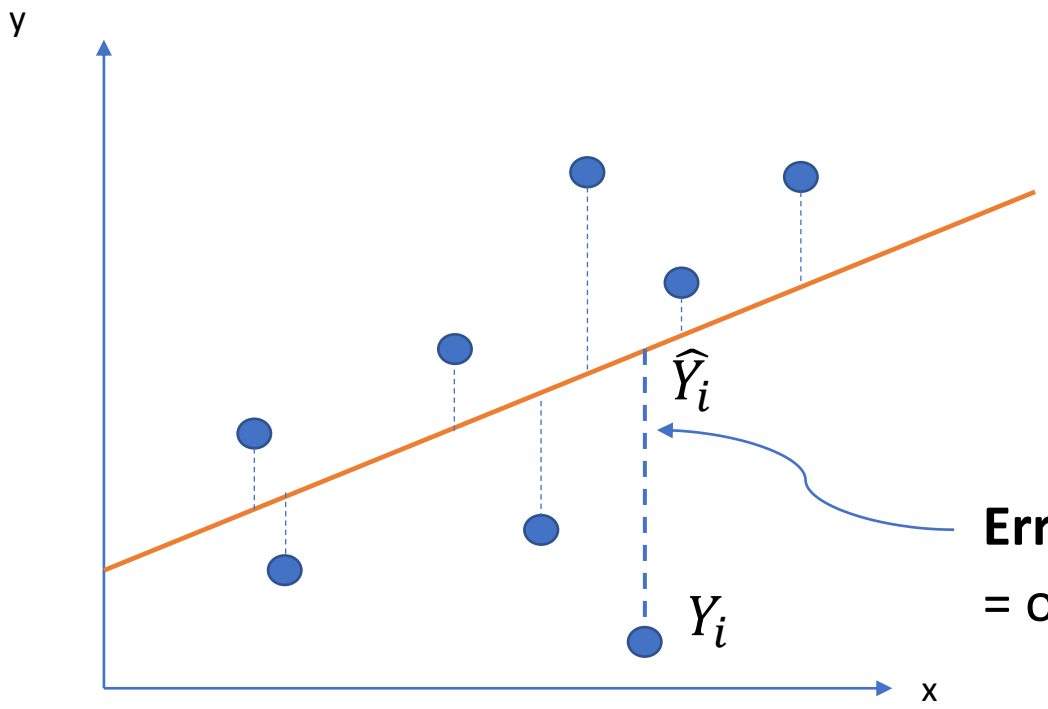
- We can include more than one independent variable to obtain a better fit:

$$Y = a + b_1 X_1 + \ \ldots + b_k X_k$$

- One independent variable -- **simple regression**
- More than one independent variable -- **multiple regression**

- How to identify the best line (or best linear regression model)?

“Best Line”



The regression line minimizes the sum of the squared errors/residuals:

$$\sum e_i^2$$

Error or residual, e_i
= observed value(Y_i) – predicted/fitted value $\left(\widehat{Y}_i\right)$

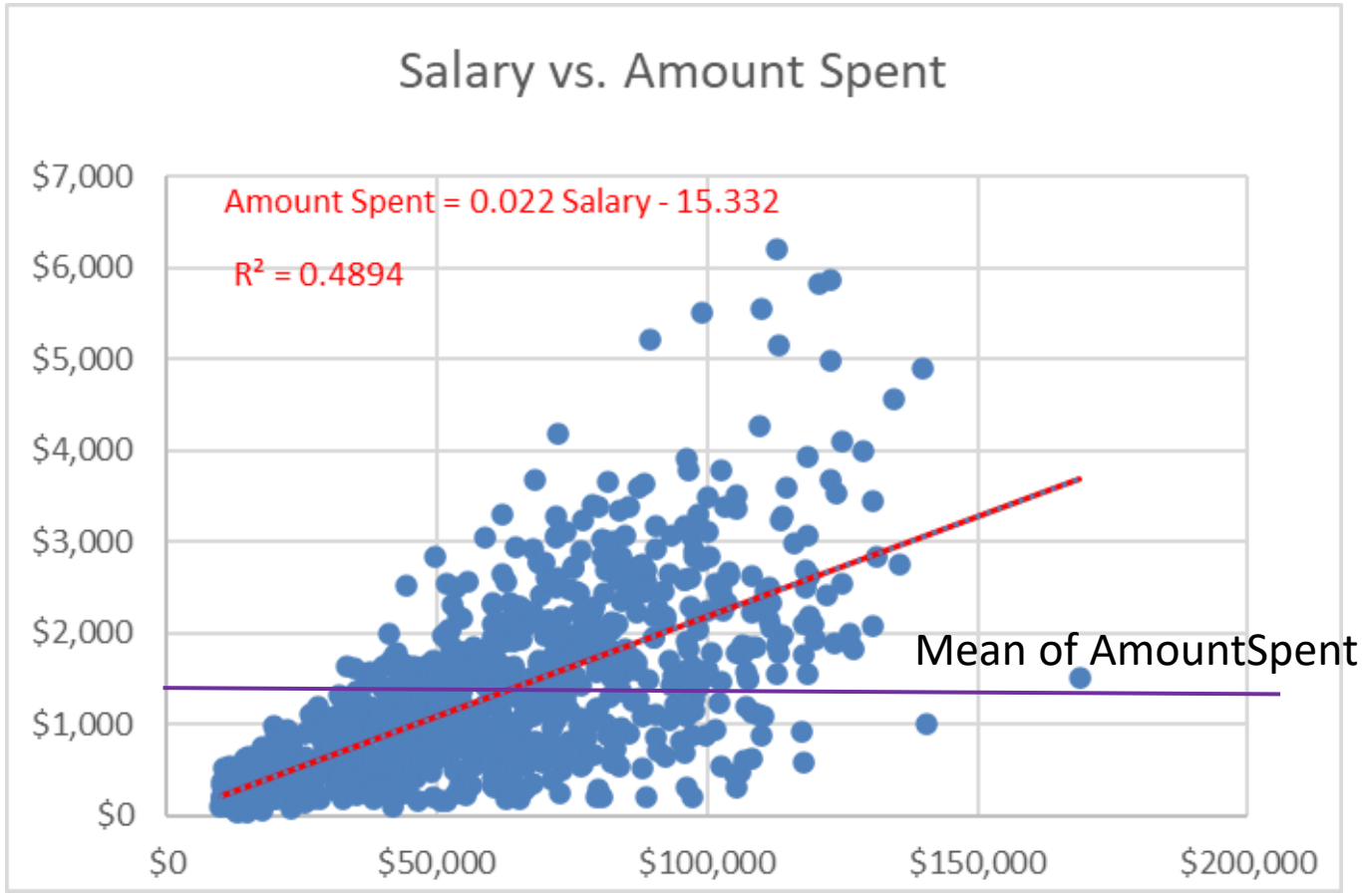
Linear Regression Models

- How to measure the linear relationship between dependent and independent variables for multiple regression?

  => R^2: measure of fit,

  it is the square of correlation in simple regression

Example: HyTex Catalog Marketing Data

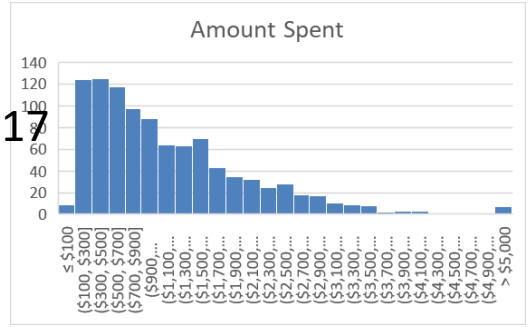


$$Correl\ (AmountSpent, Salary) = 0.66598$$

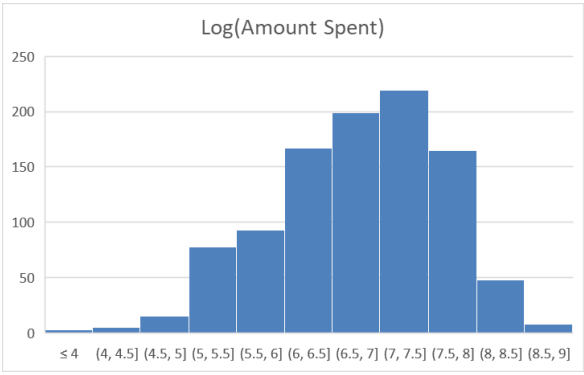
$$R^2 = \left(Correl\ (AmountSpent, Salary)\right)^2 = 0.4894$$

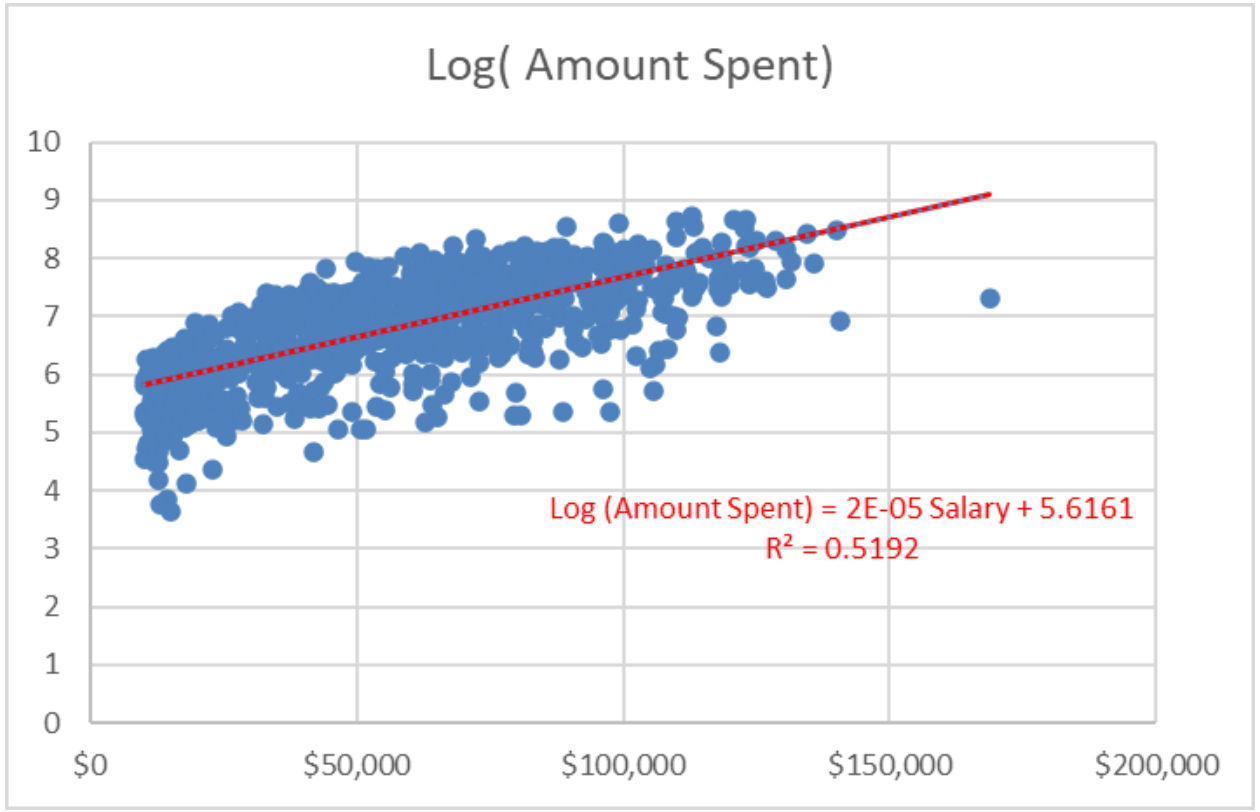
=> Measure of fit

The variance of error increases as salary increases



Logarithmic Transformation





To obtain a valid linear regression model we should check if our model satisfy all the regression model assumptions. We will discuss this later.

Regression Analysis in R

- Dataset: Catalog Marketing_Reg.xlsx
- Environment: R and RStudio
- R functions:

Function Name	Description
lm()	Fitting Linear Models
abline()	Add Straight Lines to a Plot
plot()	Draw a scatter plot
summary()	Produce result summaries of the results of model fitting functions.
resid()	Extract Model Residuals

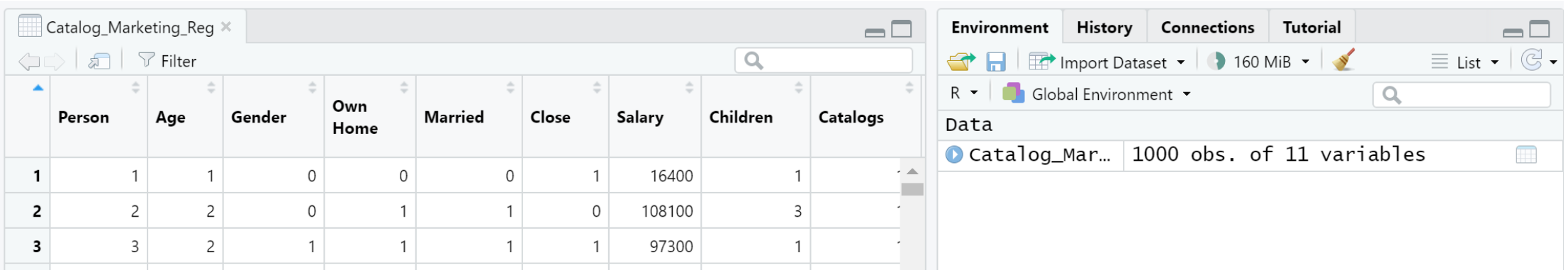
Regression Analysis in R

- Steps:
  - Upload dataset
  - Run a simple linear regression model
  - Create a scatterplot and plot the regression line

Step 1: Upload Dataset

- Use Dataset From Excel
- Import Catalog Marketing_Reg.xlsx

  There are 11 columns in the dataset,



Step 2: Run a simple linear regression model

- The command to fit a simple linear regression model is lm()
- You can get information about the lm() command by typing

```
>?lm()
```

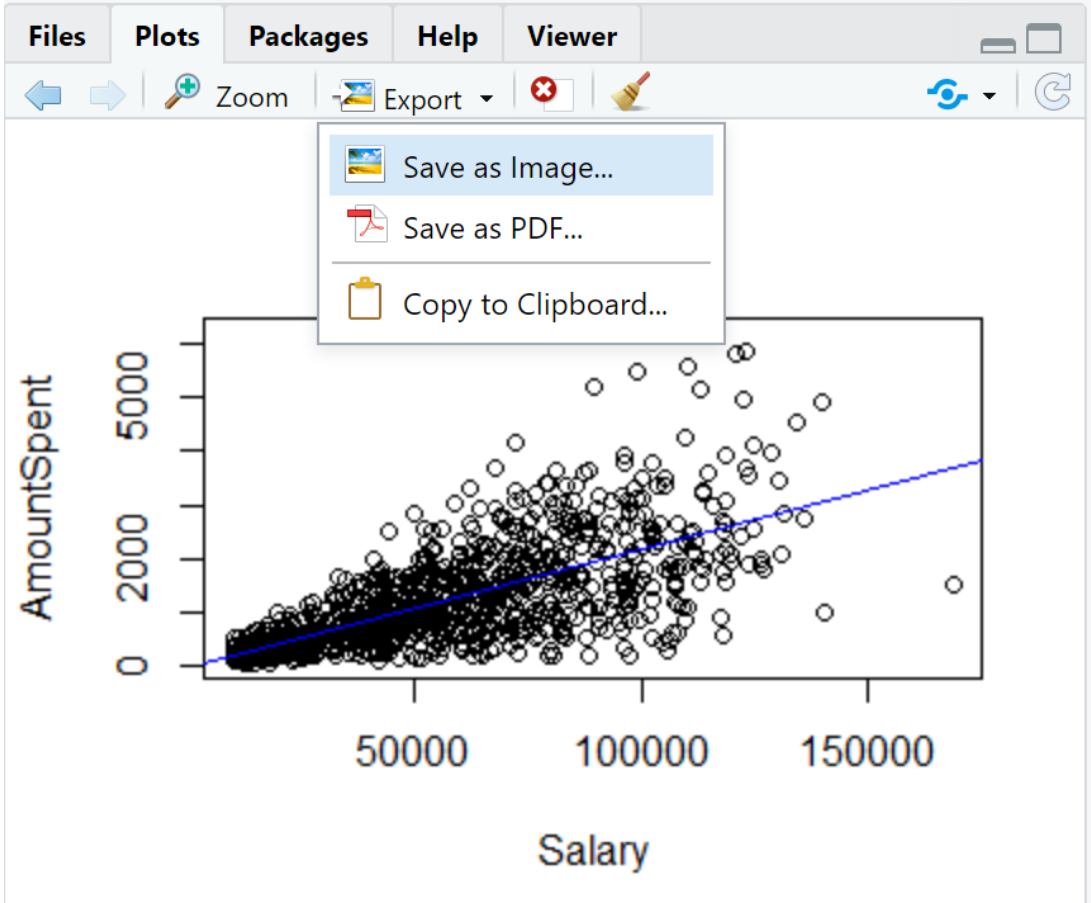
- We will fit a simple linear regression model to predict AmountSpent using Salary as the predictor.

```
>slr <- lm(AmountSpent~Salary, data = Catalog Marketing_Reg)
>summary(slr)
…
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.33243   45.37435  -0.338    0.736
Salary        0.02196    0.00071  30.931   <2e-16 ***
Multiple R-squared:  0.4894,          Adjusted R-squared:  0.4889
```

Step 3: Plot the regression line

- We can create a scatterplot and plot the regression line with the following commands

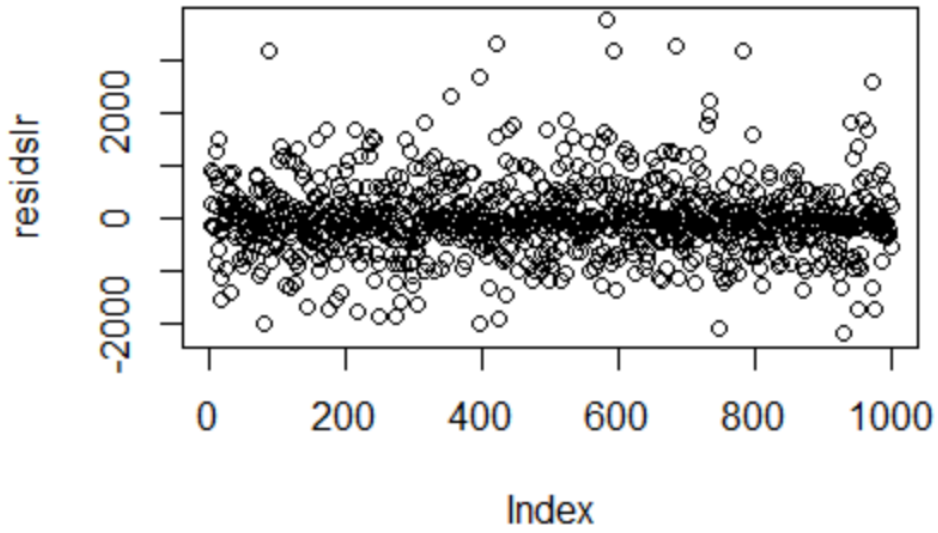
```
> plot(AmountSpent ~ Salary, main ="“AmountSpent ~ Salary “, data = Catalog_Marketing_Reg)

> abline(slr, col="blue")
```



Step 3: Residual Plots

- We can plot the residuals in the order given in the data with the following commands:

```
> residslr <-resid(slr)
> plot(residslr)
```



Next …

- More on Explanatory Regression Models