**Quality Methods for Business-Multiple Regression Project**

**Team:**

**Sampreeth Maturi – U46217668  
Sai Ranadheer goud Mosanpally – U61091386  
Akhila Siddabatula – U73439960**

**Purpose**: The purpose of this project is to demonstrate how multiple regression model fits into Professors salary prediction and interpret their coefficients using R studio.

**Data set**: The dataset contains

* 397 observations with 6 variables.
* 5 independent variables and 1 dependent variable.

**Dataset source and reference**: <https://www.kaggle.com/code/lindleylawrence/regression-prof-salaries/data>

Below are the variable wise descriptions:

* Rank: observations are at “Prof”, “Asst Prof” and “Assoc Prof level.
* Discipline: Namely two departments A and B.
* Yrs.since.Phd: Years since PHD completion
* Yrs.service: Number of years in service
* Sex: Gender of the disciple
* Salary: Salary of the disciple.

**Load libraries:**

**library('rio','moments','tidyverse')**

**Setting up working directory:**

**setwd('C:/Users/sampreethmaturi/Downloads')**

**Import required dataset:**

**professor = import('salaries.xlsx')**

**Understanding structure of the Data frame:**

**#stucture of the dataframe**

**str(professor)**

**Data frame output:**

Text

Description automatically generated

**Simple Regression Models:**

**Simple linear regression model 1:**

**salary1 <- lm(salary~sex, data = professor)**

**summary(salary1)**

Text

Description automatically generated

**ANALYSIS:**

**From the above summary, we can say that the average salary for Female is 101002 dollars. A male gender earns 14008 more dollars compared to a Female person.**

**Simple linear regression model 2:**

**salary2 = lm(salary~discipline, data = professor)**

**summary(salary2)**

Text

Description automatically generated

**ANALYSIS:**

**From the above summary, we can say that the average salary for discipline A is 108548 dollars. It is noticed that Discipline B earns 9480 more dollars compared to Disciple B.**

**Simple linear regression model 3:**

**salary3 = lm(salary~rank, data = professor)**

**summary(salary3)**

Text

Description automatically generated

**ANALYSIS:**

**From the above summary, we can say that average salary for Associate Professor is 93876 dollars. Asst professor earns 13100 dollars less than Associate Professor. Professor earns 32986 dollars more than Associate Professor.**

**Multiple Regression models:**

**Multiple regression model 1:**

**msalary1 = lm(salary~sex+discipline, data = professor)**

**summary(msalary1)**

Text

Description automatically generated

**ANALYSIS:**

**From the above summary, we can state that predictor variable is explain only 5% variance in data. P value has significance in both the cases. Males are expected to earn 14029 dollars and discipline B are expected to increase by 9449 dollars.**

**Multiple regression model 2:**

**msalary2 = lm(salary~discipline+rank, data = professor)**

**summary(msalary2)**

Text

Description automatically generated

**ANALYSIS:**

**From the above summary, we can say that P value for all the variables are significant. Predictor variable explains 44.5% of variance in the data.**

**Multiple regression model 3:**

**msalary3 = lm(salary~rank+sex, data = professor)**

**summary(msalary3)**

Text

Description automatically generated

**ANALYSIS:**

**From above summary for rank and sex combination, we can state that prof has higher salary compared to others and Male is expected to earn by 4943 dollars more compared to female.**

**Full regression:**

full\_salary <- lm(salary~., data = professor)

summary(full\_salary)

Text

Description automatically generated

**ANALYSIS:**

**From the above summary, we can state that Faculty gets an average salary of 78862 dollars. Sex variable has least significant (0.21584) and other variables have significance. The coefficient of years of service is in negative (-489.5). As year of service increases, salary drops by 489 dollars. Discipline B salary increases by 14417 compared to discipline B.**

**Regression model with squared terms:**

**Regression model with squared terms 1:**

**professor$squared = professor$yrs.service\*professor$yrs.service**

**sssalary1 = lm(salary~yrs.service+squared, data = professor)**

**summary(sssalary1)**

Text

Description automatically generated

**ANALYSIS:**

**From the above summary, we can state that, intercept value is at 85115. Salary is expected to increase by 3011 dollars for every year but squared term salary expected to drop by 51 dollars.**

**Regression model with squared terms 2:**

**professor$squaredtwo = professor$yrs.since.phd\*professor$yrs.since.phd**

**sssalary2 = lm(salary~yrs.since.phd+squaredtwo, data = professor)**

**summary(sssalary2)**

Text

Description automatically generated

**ANALYSIS:**

**From the above summary, we can state that intercept is at 65051. P values for all variables has significant. Person who has PHd, salary expected to increase by 4075 dollars and squared metric has expected to drop by 63 dollars.**

**Best fit model:**

**full\_salary <- lm(salary~., data = professor)**

**summary(full\_salary)**

Text

Description automatically generated

**ANALYSIS:**

**Above model is the best fit model. Predictor variable is able to explain 45.47% of variance in the salary. Sex variable has least significant (0.21584) and other variables have significance. As year of service increases, salary drops by 489 dollars. Discipline B salary increases by 14417 compared to discipline B.**

**Linearity**:

Chart, scatter chart

Description automatically generated

**Normality:**

Chart, line chart

Description automatically generated

**Equality**:

Chart, scatter chart

Description automatically generated

**ANALYSIS:**

**From the above assumptions, we can say that Linearity condition is satisfied. Data is slightly deviated in normality condition but it can be improved by making changes in independent variables and having more observations. Equality condition of the dataset is satisfied but, we can see few outliers in the dataset.**

**Two types of predictions Predict and Confidence:**

**new\_salary\_pred <- data.frame(rank="Prof",discipline="A",sex="Male",yrs.since.phd=20,yrs.service=21)**

**predict(full\_salary,new\_salary\_pred,interval="predict")**

**predict(full\_salary,new\_salary\_pred,interval="confidence")**

Text

Description automatically generated

**ANALYSIS:**

**Predict: This can be used for only particular point. Fit point of the predict interval is 116226. We are 95% confident that lower bound is 71607 and upper bound is 160844.**

**Confidence: Fit line for Predict and confidence is same. We are 95% confident that lower bound is 111009 and upper bound is 121442.**

**Thank you**