WORK SAMPLE

By

Suhasini Singh (1811068)

Department of Commerce

CHRIST (Deemed to be University)

Bangalore

2019-2020

Business Stats (COH233)

Multiple Linear Regression Analysis:

Gender Pay Gap in the Organized Working Sector

SUMMARY:

The document "Linear Regression Models Practical (STA551)" focuses on analyzing the gender pay gap in the organized working sector using multiple linear regression models. It involves the use of a dataset from Glassdoor, examining the relationship between gender and pay, incorporating factors like age, performance evaluation, and seniority. The analysis includes constructing and comparing different regression models to understand how these variables influence bonus and base pay, with a particular emphasis on identifying gender-based disparities in compensation. The results provide insights into the extent of the gender pay gap, highlighted through statistical measures and model coefficients.

OBJECTIVES:

Consider a dataset in the Gender Sensitization domain and perform the following:

1. Take your own data having two independent variables (one continuous and one categorical variable) and a dependent variable.
2. Apply the suitable multiple linear regression model and analyse the data.
3. Write a report on it.

DATA DESCRIPTION

Source: https://www.kaggle.com/datasets/nilimajauhari/glassdoor-analyze-gender-pay-gap

Glassdoor- Analyze Gender Pay Gap | Kaggle

Context

Want to know the base pay for different job roles, then this data set will be useful.

About the data set:

The data set has been taken from glassdoor and focuses on income for various job titles based on gender. As there have been many studies showcasing that women are paid less than men for the same job titles, this data set will be helpful in identifying the depth of the gender-based pay gap. The features of the data set are:

1. Job Title
2. Gender
3. Age
4. PerfEval
5. Education

6. Dept
7. Seniority
8. Base Pay
9. Bonus

Acknowledgements

The data set has been taken from the website of Glassdoor. The license was not mentioned on the source.

Inspiration

To find out the pay gap between the genders.

Dataset Head, Summary and Structure:

```
Data<-read.csv("Glassdoor Gender Pay Gap.csv")
Data2<-read.csv("numericpay.csv")
head(Data)
```

```
##             JobTitle Gender Age PerfEval Education          Dept Seniority
## 1    Graphic Designer Female  18        5   College    Operations         2
## 2   Software Engineer   Male  21        5   College    Management         5
## 3 Warehouse Associate Female  19        4       PhD Administration         5
## 4   Software Engineer   Male  20        5   Masters         Sales         4
## 5    Graphic Designer   Male  26        5   Masters   Engineering         5
## 6                  IT Female  20        5       PhD    Operations         4
##   BasePay Bonus
## 1   42363  9938
## 2  108476 11128
## 3   90208  9268
## 4  108080 10154
## 5   99464  9319
## 6   70890 10126
```

```
summary(Data)
```

```
##    JobTitle            Gender               Age           PerfEval
##  Length:1000        Length:1000        Min.   :18.00   Min.   :1.000
```

```
##  Class :character   Class :character   1st Qu.:29.00   1st Qu.:2.000
##  Mode  :character   Mode  :character   Median :41.00   Median :3.000
##                                        Mean   :41.39   Mean   :3.037
##                                        3rd Qu.:54.25   3rd Qu.:4.000
##                                        Max.   :65.00   Max.   :5.000
##   Education             Dept             Seniority        BasePay
##  Length:1000        Length:1000        Min.   :1.000   Min.   : 34208
##  Class :character   Class :character   1st Qu.:2.000   1st Qu.: 76850
##  Mode  :character   Mode  :character   Median :3.000   Median : 93328
##                                        Mean   :2.971   Mean   : 94473
##                                        3rd Qu.:4.000   3rd Qu.:111558
##                                        Max.   :5.000   Max.   :179726
##      Bonus
##  Min.   : 1703
##  1st Qu.: 4850
##  Median : 6507
##  Mean   : 6467
##  3rd Qu.: 8026
##  Max.   :11293
```

```
str(Data)
```

```
## 'data.frame':    1000 obs. of  9 variables:
##  $ JobTitle : chr  "Graphic Designer" "Software Engineer" "Warehouse Associate"
"Software Engineer" ...
##  $ Gender   : chr  "Female" "Male" "Female" "Male" ...
##  $ Age      : int  18 21 19 20 26 20 20 18 33 35 ...
##  $ PerfEval : int  5 5 4 5 5 5 5 4 5 5 ...
##  $ Education: chr  "College" "College" "PhD" "Masters" ...
##  $ Dept     : chr  "Operations" "Management" "Administration" "Sales" ...
##  $ Seniority: int  2 5 5 4 5 4 4 5 5 5 ...
##  $ BasePay  : int  42363 108476 90208 108080 99464 70890 67585 97523 112976 106524
...
##  $ Bonus    : int  9938 11128 9268 10154 9319 10126 10541 10240 9836 99
```

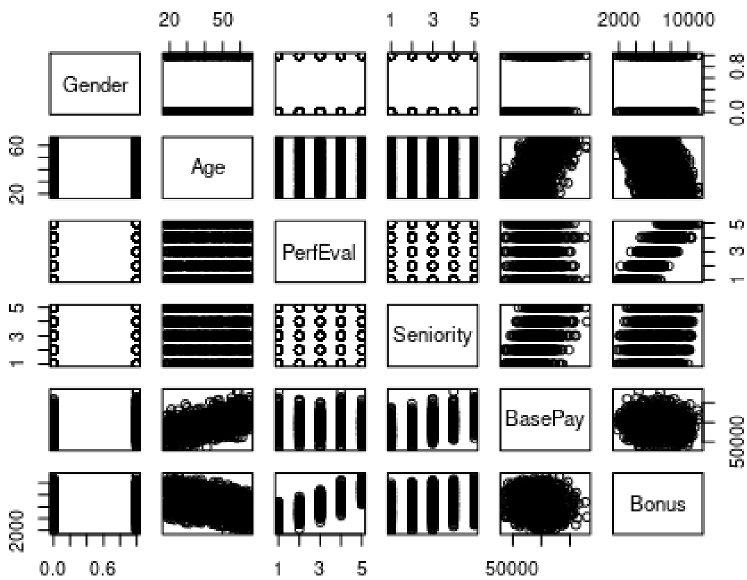
One can use the summary and structure statistics of the data to note the various summary statistics of the data which include Minimum Value, Quartiles, Mean, Median and Maximum Value.

Data$Gender~as.factor(Data$Gender)

table(Data$Gender)

```
##
## Female   Male
##    468    532
```

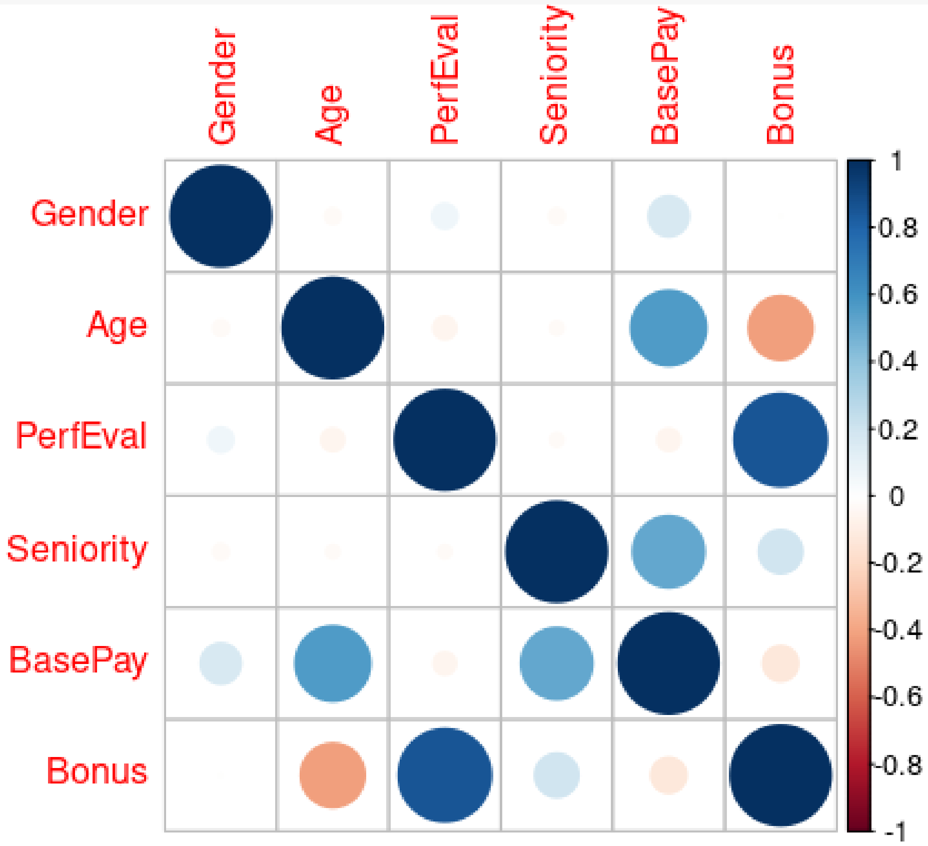
pairs(Data2)



(Figure 1.1)

The pair plot showcases the linearity between different data columns. We can see that Age and BasePay have a strong linear relationship, so do Bonus and Age and other than that any strong linear trend is not seen.

```
mtrix<-cor(Data2)
corrplot(mtrix, method="circle")
```



(Figure 1.2)

The corrplot showcases the correlation between different attributes of data. We can see positive correlation between the following:

1. ***Age and Base Pay***
2. ***Bonus and Performance Evaluation***
3. ***Seniority and Base Pay***
4. ***Seniority and Bonus (Weak)***
5. ***Base Pay and Gender (Weak)***

Negative Correlation between:

1. ***Bonus and Age***

Constructing 4 different models, model 1 is the main one and the others are meant for comparison of different methods and variables.

```
model1=lm(Bonus~Gender+Age+Seniority+PerfEval,data=Data)
model1

##
## Call:
## lm(formula = Bonus ~ Gender + Age + Seniority + PerfEval, data = Data)
##
## Coefficients:
## (Intercept)  GenderMale        Age    Seniority     PerfEval
##     4243.12     -257.35     -51.02       291.97      1187.13
```

The Fitted Model is:

Bonus= (4243.12) + (-257.35) Gender Male+ (-51.02) Age + (291.97) Seniority+(1187.13) PerfEval

Average Bonus for Females = *4243.12* if other factors are 0.

Average Bonus for Males = *4243.12+(-257.35)* if other factors are 0.

For example Expected Test Value= *10126*

```
newdata=data.frame(Gender="Female",Age=20,Seniority=4,PerfEval=5)
predict(model1, newdata)

##        1
## 10326.29


newdata2=data.frame(Gender="Male",Age=20,Seniority=4,PerfEval=5)
predict(model1, newdata2)

##        1
## 10068.94
```

The gender changes the entire model by -257.35.

Hence, we can construct our regression model by the above model which provides the best approximation to associate between Bonus and Gender + Age + Seniority + PerfEval.

Now here we can see that:

Bo=2.194e+04 which shows that if we put Gender + Age + Seniority + PerfEval as 0, then the Bonus will be taken for females and the income will be approximately equal to 4243.12 and we can say that for each additional employee the Bonus increase by B1+B2+B3+B4.

summary(model1)

```
##
## Call:
## lm(formula = Bonus ~ Gender + Age + Seniority + PerfEval, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1850.58  -397.65   -14.75   416.73  1874.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4243.122     86.022  49.326  < 2e-16 ***
## GenderMale  -257.352     37.850  -6.799 1.81e-11 ***
## Age          -51.018      1.321 -38.624  < 2e-16 ***
## Seniority    291.965     13.517  21.600  < 2e-16 ***
## PerfEval    1187.134     13.284  89.369  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 595.5 on 995 degrees of freedom
## Multiple R-squared:  0.9121, Adjusted R-squared:  0.9117
## F-statistic:  2581 on 4 and 995 DF,  p-value: < 2.2e-16
```

The min and max residuals of the model are –1850.58 and -397.65

The residuals show that there is a difference in the observed value and estimated values.

Here Bo,B1,B2, B3 and B4 is divided by the standard error to get the following t statistic:

1. ***49.326***
2. ***-6.799***
3. ***-38.624***
4. ***21.600***
5. ***89.369***

The residual standard error is 595.5 which says the deviation of residuals from their mean on 995 df.

Since the mean of residuals will be 0 so we can say that the average deviation is 995. R^2 is 91.21% and 91.21% of the model fits the data.

Multiple R-Squared or The coefficent of deterimation is found to be 91.17%, and it shows that 91.17% of the bonus depends on the age, gender, seniority and Performance Index and the rest is error or other factors.

The p-value is not significant for any variable.

```
confint(model1,level=0.95)

##                  2.5 %     97.5 %
## (Intercept) 4074.31714 4411.92713
## GenderMale  -331.62591 -183.07748
## Age          -53.61052  -48.42637
## Seniority    265.44022  318.49057
## PerfEval    1161.06737 1213.20140
```

According to this we can see that the average value of B1, B2, B3 and B4 will be between:

2.5 % 97.5 %

(Intercept) 4074.31714 4411.92713

GenderMale -331.62591 -183.07748

Age -53.61052 -48.42637

Seniority   265.44022 318.49057

PerfEval   1161.06737 1213.20140

```
anova(model1)

## Analysis of Variance Table
##
## Response: Bonus
##            Df     Sum Sq    Mean Sq   F value Pr(>F)
## Gender      1      41300      41300    0.1165  0.733
## Age         1  689891349  689891349 1945.4862 <2e-16 ***
## Seniority   1  138545714  138545714  390.6974 <2e-16 ***
## PerfEval    1 2832194513 2832194513 7986.7583 <2e-16 ***
## Residuals 995  352838216     354611
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this we can see that the p value of Gender is greater than 0.05,then we can accept the null hypothesis which says the Bonus and Gender are correlated.

All other Null Hypothesis are rejected.

Now we look into the other 3 models:

```
model2=lm(Bonus~Gender,data=Data)
model2

##
## Call:
## lm(formula = Bonus ~ Gender, data = Data)
##
## Coefficients:
## (Intercept)   GenderMale
##     6474.01       -12.88
```

```
summary(model2)

##
## Call:
## lm(formula = Bonus ~ Gender, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4758.1 -1611.9    41.4  1564.9  4831.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6474.01      92.70  69.840   <2e-16 ***
## GenderMale    -12.88     127.09  -0.101    0.919
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2005 on 998 degrees of freedom
## Multiple R-squared:  1.029e-05,  Adjusted R-squared:  -0.0009917
## F-statistic: 0.01027 on 1 and 998 DF,  p-value: 0.9193

model3=lm(BasePay~Gender+Age+PerfEval+Seniority,data=Data)
model3

##
## Call:
## lm(formula = BasePay ~ Gender + Age + PerfEval + Seniority, data = Data)
##
## Coefficients:
## (Intercept)   GenderMale          Age     PerfEval    Seniority
##       19326        10187         1025         -408         9602

summary(model3)

##
## Call:
```

```
## lm(formula = BasePay ~ Gender + Age + PerfEval + Seniority, data = Data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42425 -10366  -1404   9580  52748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19325.91    2229.11   8.670   <2e-16 ***
## GenderMale  10187.43     980.81  10.387   <2e-16 ***
## Age          1025.29      34.23  29.954   <2e-16 ***
## PerfEval     -407.98     344.22  -1.185    0.236
## Seniority    9601.63     350.27  27.412   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15430 on 995 degrees of freedom
## Multiple R-squared:  0.6306, Adjusted R-squared:  0.6291
## F-statistic: 424.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

```
model4=lm(BasePay~Gender,data=Data)
model4
```

```
##
## Call:
## lm(formula = BasePay ~ Gender, data = Data)
##
## Coefficients:
## (Intercept)   GenderMale
##       89943         8515
```

```
summary(model4)
```

```
##
## Call:
## lm(formula = BasePay ~ Gender, data = Data)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -61816 -16995  -149  17036  81268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    89943       1155  77.859  < 2e-16 ***
## GenderMale      8515       1584   5.376 9.48e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24990 on 998 degrees of freedom
## Multiple R-squared:  0.02815,    Adjusted R-squared:  0.02717
## F-statistic:  28.9 on 1 and 998 DF,  p-value: 9.479e-08
```

Conclusion:

One can use the summary and structure statistics of the data to note the various summary statistics of the data which include Minimum Value, Quartiles, Mean, Median and Maximum Value.

The pair plot showcases the linearity between different data columns. We can see that Age and BasePay have a strong linear relationship, so do Bonus and Age and other than that any strong linear trend is not seen.

The corrplot showcases the correlation between different attributes of data. We can see positive correlation between the following:

1. Age and Base Pay
2. Bonus and Performance Evaluation
3. Seniority and Base Pay
4. Seniority and Bonus (Weak)
5. Base Pay and Gender (Weak)

Negative Correlation between:

1. Bonus and Age

Bonus= (4243.12) + (-257.35) Gender Male+ (-51.02) Age + (291.97) Seniority+(1187.13) PerfEval

Average Bonus for Females = 4243.12 if other factors are 0.

Average Bonus for Males = 4243.12+(-257.35) if other factors are 0.

The gender changes the entire model by -257.35.

Here Bo,B1,B2, B3 and B4 is divided by the standard error to get the following t statistic:

1. 49.326
2. -6.799
3. -38.624
4. 21.600
5. 89.369

The residual standard error is 595.5 which says the deviation of residuals from their mean on 995 df.

Since the mean of residuals will be 0 so we can say that the average deviation is 995. R^2 is 91.21% and 91.21% of the model fits the data.

Multiple R-Squared or The coefficent of deterimation is found to be 91.17%, and it shows that 91.17% of the bonus depends on the age, gender, seniority and Performance Index and the rest is error or other factors.

The p-value is not significant for any variable.

According to this we can see that the average value of B1, B2, B3 and B4 will be between:

##	2.5 %	97.5 %
## (Intercept)	4074.31714	4411.92713
## GenderMale	-331.62591	-183.07748
## Age	-53.61052	-48.42637

```
## Seniority    265.44022  318.49057
## PerfEval    1161.06737 1213.20140
```

From this we can see that the p value of Gender is greater than 0.05,then we can accept the null hypothesis which says the Bonus and Gender are correlated.

All other Null Hypothesis are rejected.