**Introduction**

In this project, the dataset used is the “Business Sales” dataset available on <http://roycekimmons.com/tools/generated_data/sales>. It contains information about Salesperson performance reports for an electronic company. The various fields are Salary, Division, Level Of Education, Training Level, Work Experience and Sales.

Various questions have been addressed are as follows:

1) Is the mean / median salary higher of employees who have a bachelor’s degree as compared to employees who have a high school education?

2) What percentage of employees make sales between $ 400,000 and $ 500,000?

3) Are employees who make more sales given a higher salary?

4) What are the parameter values that maximize the likelihood function of the distribution of the difference of the salaries of employees who have a bachelor’s degree and who have a high school education?

5) What percentage of variance in Salary can be explained by Division, Level Of Education, Training Level, Work Experience and Sales

**To understand the distribution of the data**

The data was imported into R and a histogram was drawn for the field sales and salary. The plots are as follows:

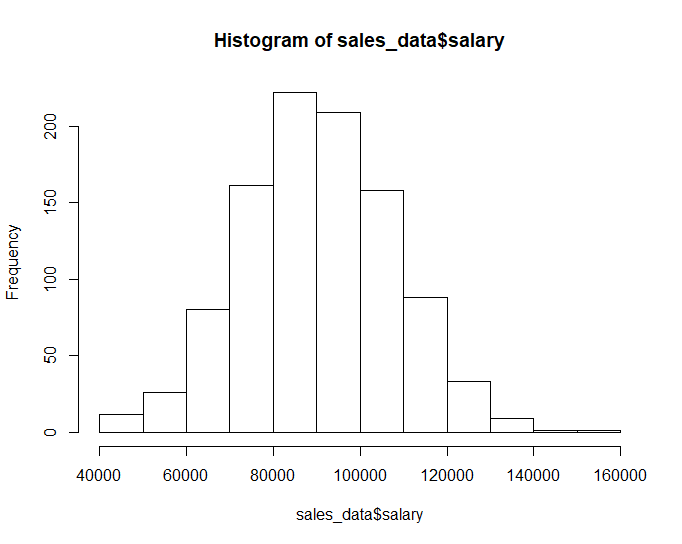
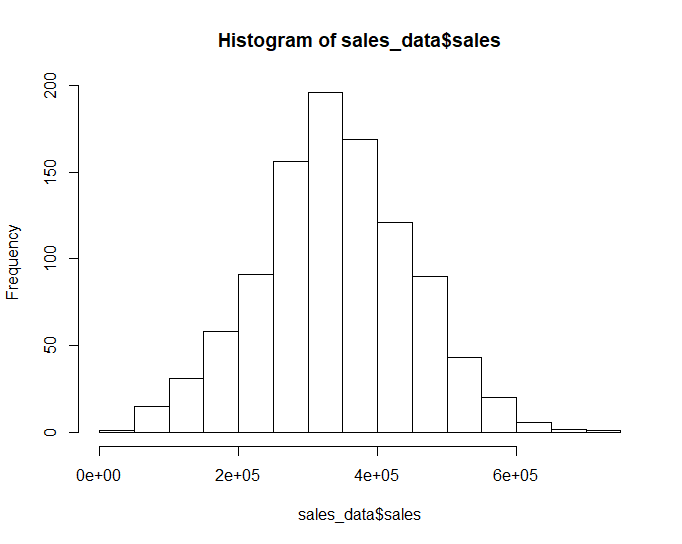


Figure 1 Figure 2

From the plot above, we can see that the data follows a normal distribution.

**Test for Normality**:

To check and confirm for normality, Sharpiro Test was conducted in R. The p value obtained was .5305 for Sales and .6768 for Salary indicating that the distribution is not significantly different from a normal distribution.

Conclusion: The sales and salary data from the dataset are normally distributed.

**To find out what percentage of employees make sales between 400000$ and 500000$**

**Empirical CDF**:

Next an empirical CDF was found and plotted. The plot is as follows:

A non parametric confidence band was plotted for the Sales Data. Also a confidence interval was plotted on the same graph. The plot is as follows:

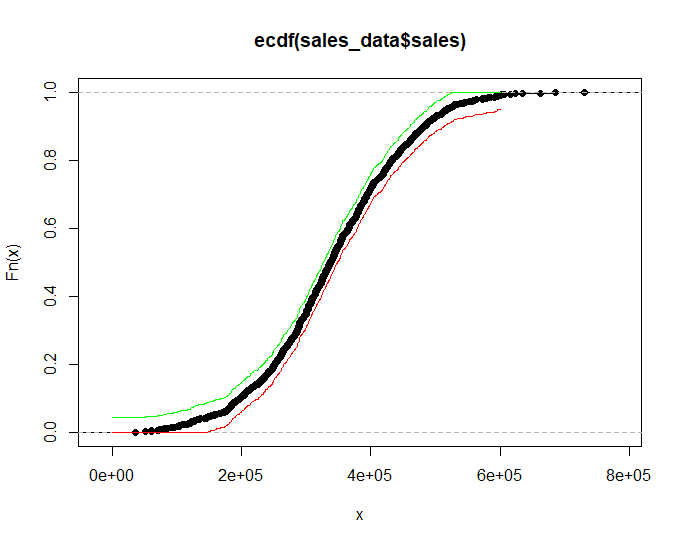


Figure 3

The green line represents the upper limit and the red line represents the lower limit of the non parametric confidence band.

The difference of the ecdf of the employees who make a sale of $500000 and that of employees who make a sale of $400000 was taken. This difference is 21.1 %.

Conclusion: 21.1% of the employees make sales between 400000$ and 500000$.

**Bootstrapping standard errors and Confidence Intervals:**

To check if the salary of an employee has a correlation with the they make, bootstrapping was done to find the correlation between salary and sales. The histogram of the correlations is as follows:

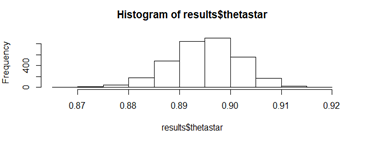


Figure 4

We can see that the correlation between salary and sales follows a normal distribution.The variance of the correlation is .0001846575 and the standard error is 0.01358887. The mean of the correlation is 0.8727259. The normal, pivotal and quantile confidence intervals were found. The confidence intervals are as follows:

|  |  |  |
| --- | --- | --- |
| **Type of CI** | **Lower Limit** | **Upper Limit** |
| Normal CI | 0.8822581 | 0.9081879 |

Table 1

|  |  |  |
| --- | --- | --- |
| **Type of CI** | **97.50%** | **2.50%** |
| Pivotal CI | 0.8831008 | 0.9085404 |

Table 2

|  |  |  |
| --- | --- | --- |
| **Type of CI** | **2.50%** | **97.50%** |
| Quantile CI | 0.8819057 | 0.9073452 |

Table 3

Conclusion: There is a high correlation between sales and salary with a mean of 0.8727259, variance of .0001846575 and standard error of 0.01358887

**To find out the MLE of the distribution consisting of the difference in salaries between employees who have a high school degree and those who have a bachelor’s degree**

**MLE and its asymptotic distributions:**

The distribution of the employees who have a bachelors degree and the distribution of the employees who have only a high school degree were separated from the main data. The MLE of the difference between these two distributions was found.

The values are as follows:

MLE of the mean is 319.7966

MLE of the standard deviation is 2256.814

Conclusion: The parameter values that maximize the likelihood function of the distribution of the difference of the salaries is given by a mean of 319.7966 and a standard deviation of 2256.814

Parametric bootstrapping

Parametric bootstrapping was used to find the standard error of the MLE. The value obtained was 2291.314. A confidence interval was formulated using the value of standard error obtained by parametric bootstrapping. It is given by (6975.676 , 16140.933)

Asymptotic distributions

The histogram of the distribution of the MLE of the mean found from bootstrapping is plotted below:

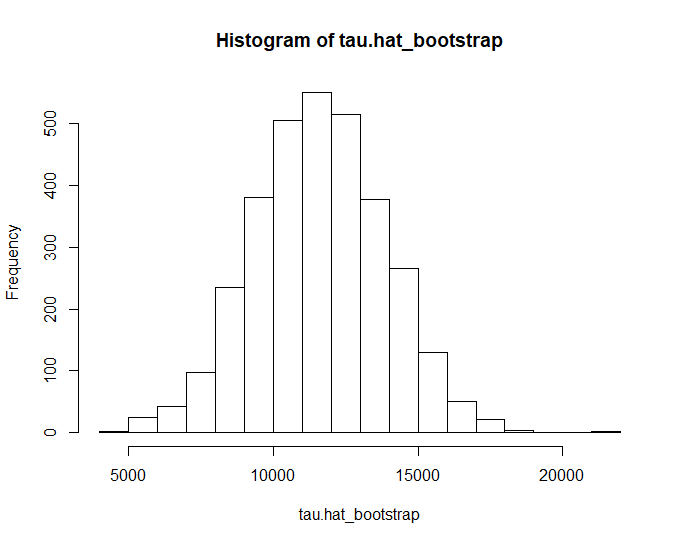


Figure 5

We can see that it appears to follow a normal distribution. The histogram of the distribution of the MLE of the standard deviation found from bootstrapping is plotted below:

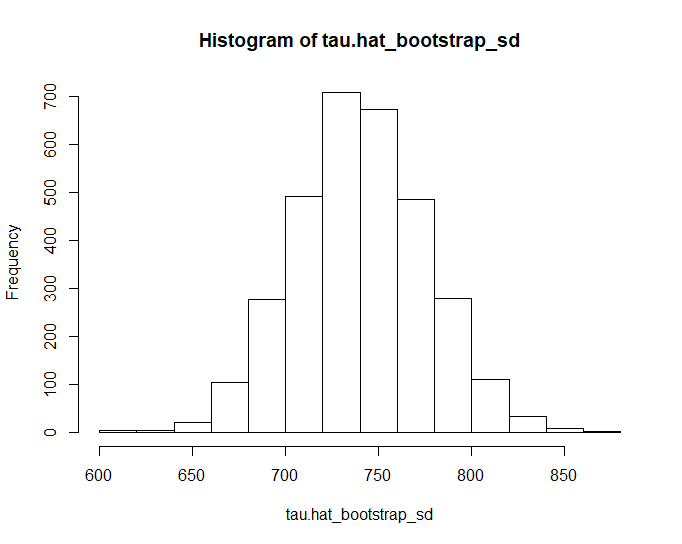


Figure 6

We can see that it appears to follow a normal distribution

Conclusion: The standard error of the MLE is 2291.314 and the confidence interval is given by (6975.676 , 16140.933). The distributions of the MLE of the mean and that of the MLE of the standard deviation obtained by bootstrapping follow a normal distribution.

**Hypothesis Tests:**

**To find out if there is a difference between the mean salary of employee who have a bachelor’s degree and those who have a high school degree**

**Wald Test :**A Wald Test was conducted. The null and alternate hypothesis were as below:

Null Hypothesis: The difference between the mean of salary of employees with a bachelor’s degree and the mean of salary of employees with a high school degree is zero

Alternate Hypothesis: The difference between the mean of salary of employees with a bachelor’s degree and the mean of salary of employees with a high school degree is not zero

The p value is 3.030915e-07. Since the p value is less than .05, we reject the null hypothesis and conclude that the difference between the means is significantly different from zero.

**To find out if median salary of employees with a bachelor’s degree is higher than the median salary of employees with a high school degree**

**Wilcoxon rank sum test**

The Wilcoxon tank sum test was conducted to compare between the medians.

Null Hypothesis: The median of salaries of employees with bachelor’s degree is greater than the median of salaries of employees with high school education.

Alternate Hypothesis: The median of salaries of employees with bachelor’s degree is not greater than the median of salaries of employees with high school education.

The p value is 4.428e-07. Since the p value is less than .05, we reject the null hypothesis and conclude that the median of salaries of employees with bachelor’s degree is not greater than the median of salaries of employees with high school education.

**Bayesian Analysis**

Assuming the priors , simulation was used to find the posterior mean and 95% posterior interval for . The mean of the posterior distribution is 11491.33. The 95 % confidence interval is given by

|  |  |  |
| --- | --- | --- |
| **Measure** | **2.50%** | **97.50%** |
| CI | 7006.031 | 15637.684 |

Table 4

**Scatter Plot**

A scatter plot was drawn between the sales made by employees and their salaries. The plot is as follows:

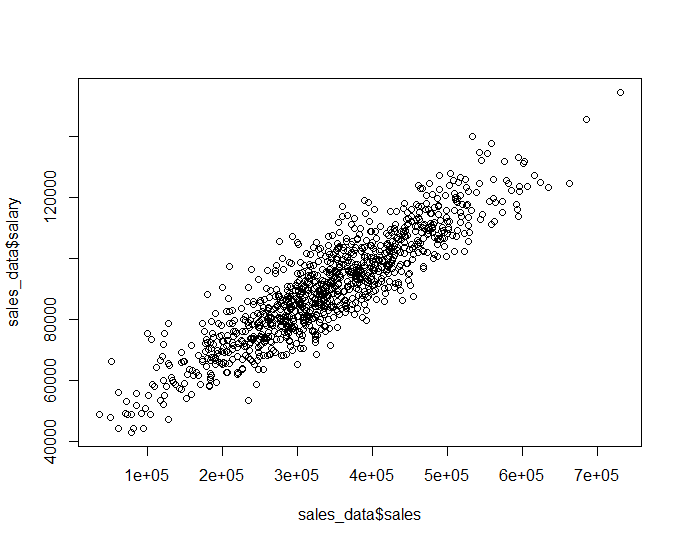


Figure 7

We can see that there is an upward linear trend between the sales made and the salaries of employees.

Conclusion: As the sales made by employees increases, their salary also seems to increase.

**Linear Regression**:

A linear regression model was fit with Salary being the dependent variable and Sales, Work Experience, Division, Level Of Education and Training Level being the independent variables. The estimated coefficients, standard errors, t values and p values are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** |
| (Intercept) | 49200 | 641.5 | 76.698 | < 2e-16 |
| sales\_data$sales | 0.09211 | 0.004676 | 19.697 | < 2e-16 |
| sales\_data$work.experience | 2208 | 148 | 14.919 | < 2e-16 |
| sales\_data$divisioncomputer software | 9331 | 570 | 16.371 | < 2e-16 |
| sales\_data$divisionoffice supplies | -5528 | 426.4 | -12.964 | < 2e-16 |
| sales\_data$divisionperipherals | -2988 | 444.2 | -6.726 | 2.95E-11 |
| sales\_data$divisionprinters | -5090 | 441.5 | -11.528 | < 2e-16 |
| sales\_data$level.of.educationbachelor's degree | 4525 | 467.9 | 9.671 | < 2e-16 |
| sales\_data$level.of.educationhigh school | -8265 | 466.4 | -17.721 | < 2e-16 |
| sales\_data$level.of.educationmaster's degree | 6733 | 1450 | 4.642 | 3.91E-06 |
| sales\_data$level.of.educationsome college | -3579 | 310.9 | -11.512 | < 2e-16 |
| sales\_data$training.level | 591.5 | 322.6 | 1.833 | 0.0671 . |

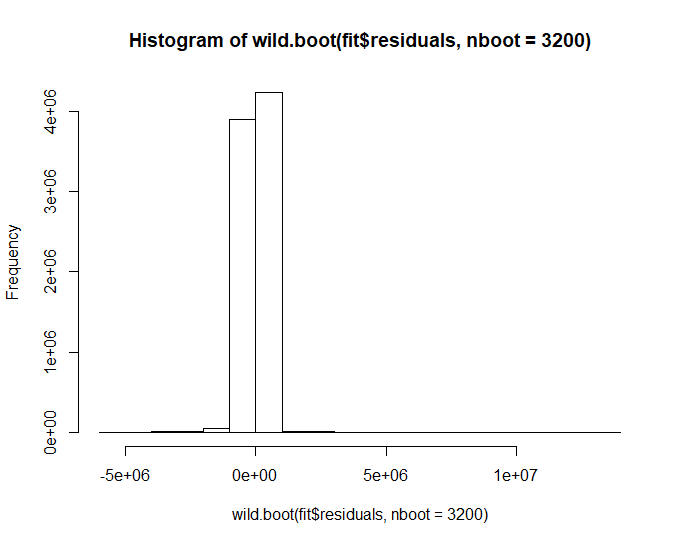
From the p values, we can see that all the independent variables except training level have statistically significant linear relationship with salary. The estimated equation is as follows: Salary = .09211 \* Sales + 2208 \*Work Experience + 9331 \* Division Computer Software - -5528 \* Division Office Supplies - 2988 \* Division Peripherals - 5090 \* Division Printers + 4525 \* Education Bachelor’s Degree - -8265 \* Education High SchoolThe R squared value is .9395

This means that 93.95 % of the variance in the Salary can be explained by the independent variables. The MSE is 18193189

Conclusion: 93.95 % of the variance in the Salary can be explained by the independent variables . The average of the squares of the errors ( difference between actual values of Salary and the predicted values of Salary is 18193189.

**Wild Bootstrap**

Wild bootstrap was done to generate bootstrap samples of the regression residuals. The histogram is as follows:



We can see that the bootstrapped residuals are not normal.

**Conclusions:**

The main conclusions are summarized below:

1. There is statistically significant difference in the mean salaries of sales persons who have a bachelor’s degree and those who have a high school degree
2. The median salaries of sales persons who have a bachelor’s degree is not greater than those of salespersons who have a high school education.
3. 93.95 % of the variance in the Salary can be explained by the variables Division, Level Of Education, Training Level, Work Experience and Sales.