|  |  |
| --- | --- |
| **Module:** | ST2053 |
| **Name:** | Jack O’Connor |
| **Student Number:** | 119319446 |
| **Chapter:** | 1 |

**Maximum 2 pages! Do not delete the page number in the footer.**

**Setup**

> executives.df <- read.table("Executives.txt", header=T)

> attach(executives.df)

> executives.lm <- lm(Salary ~ Experience, data=executives.df)

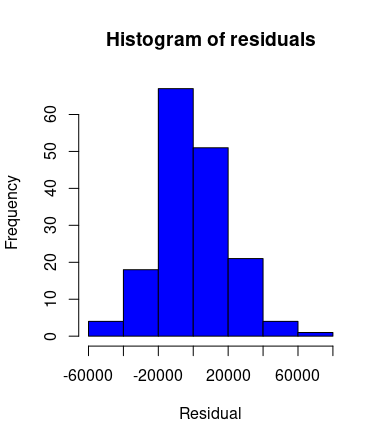
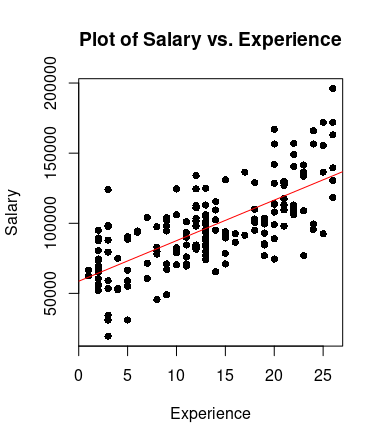
**a) Scatterplot**

> cor(Salary, Experience)

*[1] 0.7088156*

> plot(Experience, Salary, main="Plot of Salary vs. Experience", pch=16)

> abline(executives.lm, col='red')



Yes, this scatter plot does suggest that it is appropriate to fit a linear regression model to these data.

Salary very clearly trends upwards at a steady rate as experience grows.

In fact, these data have a strong correlation coefficient of 0.71.

It is also safe to assume that the common variance principle required for linear regression is upheld by the data since a histogram of the magnitudes of the residuals follows a normal distribution.

**b) Intercept and slope**

*\*Unnecessary text removed from output.*

> summary(executives.lm)

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 58699.0 3393.1 17.30 <2e-16 \*\*\**

*Experience 2892.0 224.7 12.87 <2e-16 \*\*\**

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 21130 on 164 degrees of freedom*

*Multiple R-squared: 0.5024, Adjusted R-squared: 0.4994*

**Interpretation**

The y-intercept beta-0 of the fitted regression line is equal to 58699 and the slope beta-1 is equal to 2892.

This means that we can expect the starting salary of an executive to be approximately €58,699 and that this salary will grow roughly at a rate of €2,892 per year of experience.

**c) 98% confidence interval for slope**

# alpha = 2%, two-tailed test -> critical point = qt(1 - alpha/2, df)

# standard error 224.7 from summary output

# Lower bound

> 2892 - qt(0.99, 164)\*224.7

*[1] 2364.11*

> # Upper bound

> 2892 + qt(0.99, 164)\*224.7

*[1] 3419.89*

**Interpretation**

After performing some statistical analysis on our sample data, we expect that there is a 98% chance that the true value of the slope lies in the interval 2364 <= beta-1 <= 3419.

**d) Test hypothesis H0: beta-1 = 0**

The null hypothesis that the value of beta-1 = 0 should be rejected because it lies well outside our previously computed 98% confidence interval of 2364 <= beta-1 <= 3419.

In practice this means that the likelihood that salary does not depend on experience (H0: beta-1 = 0) is very small and it is appropriate to model the relationship of these two variables using simple linear regression.

Using a 2% level of significance means that our confidence interval of the values we expect the true value of beta-1 to lie in is quite broad. It is extremely unlikely for beta-1 to actually equal 0 when it is so far outside our generous estimate of likely values.

**e) Interpret value of R^2**

The value of R^2 is 0.502 (from summary output) which means that about 50% of the variability in Salary can be explained by Experience, which is good but leaves room for improvement.

My recommendation for the current model would be to see if any of the other features in the executives data frame such as Education, Sales, etc. correlate with Salary and to see if a multiple linear regression model using these features as well as Experience would create better predictions.