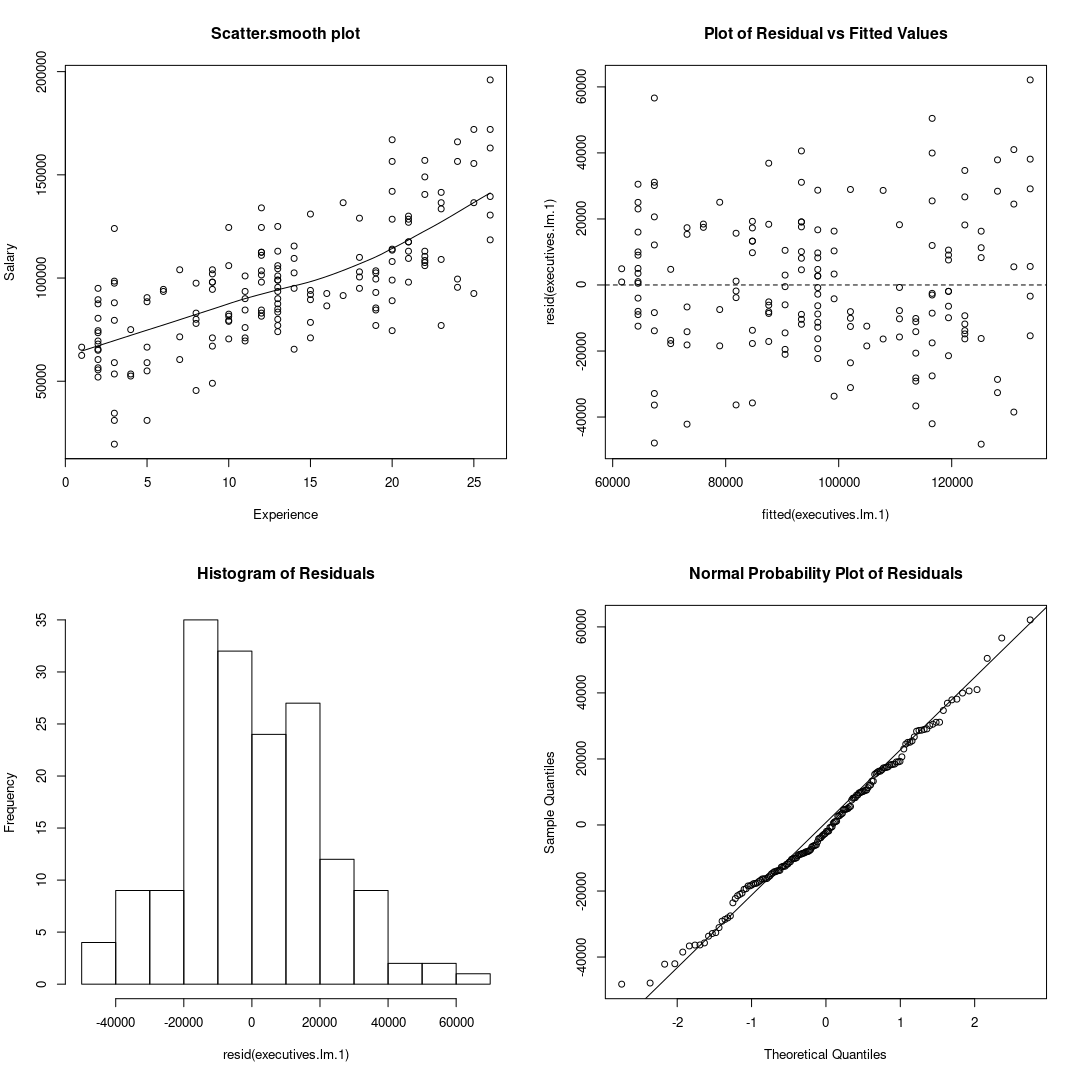
|  |  |
| --- | --- |
| **Module:** | ST2053 |
| **Name:** | Jack O’Connor |
| **Student Number:** | 119319446 |
| **Chapter:** | 4 |

**Maximum 4 pages! Do not delete the page number in the footer.**

**A) Diagnostic plots**

**Diagnostic plots for executives.lm.1**



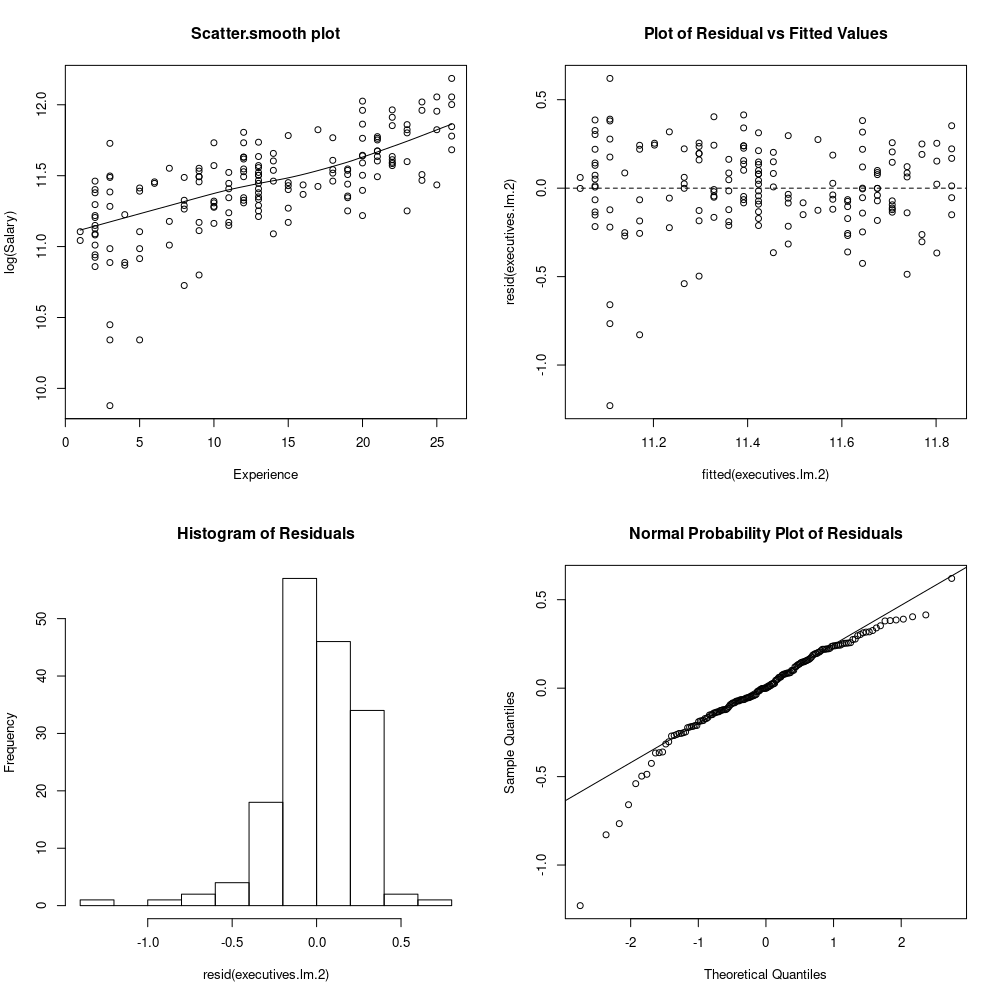
Scatter Plot: Linear & constant variance.

Residuals Vs Fitted Values: Linear & constant variance.

Histogram of residuals: Approximately normal but two peaks

Normal Probability Plot: Minor departures – approximately normal.

**Diagnostic plots for executives.lm.2**

****

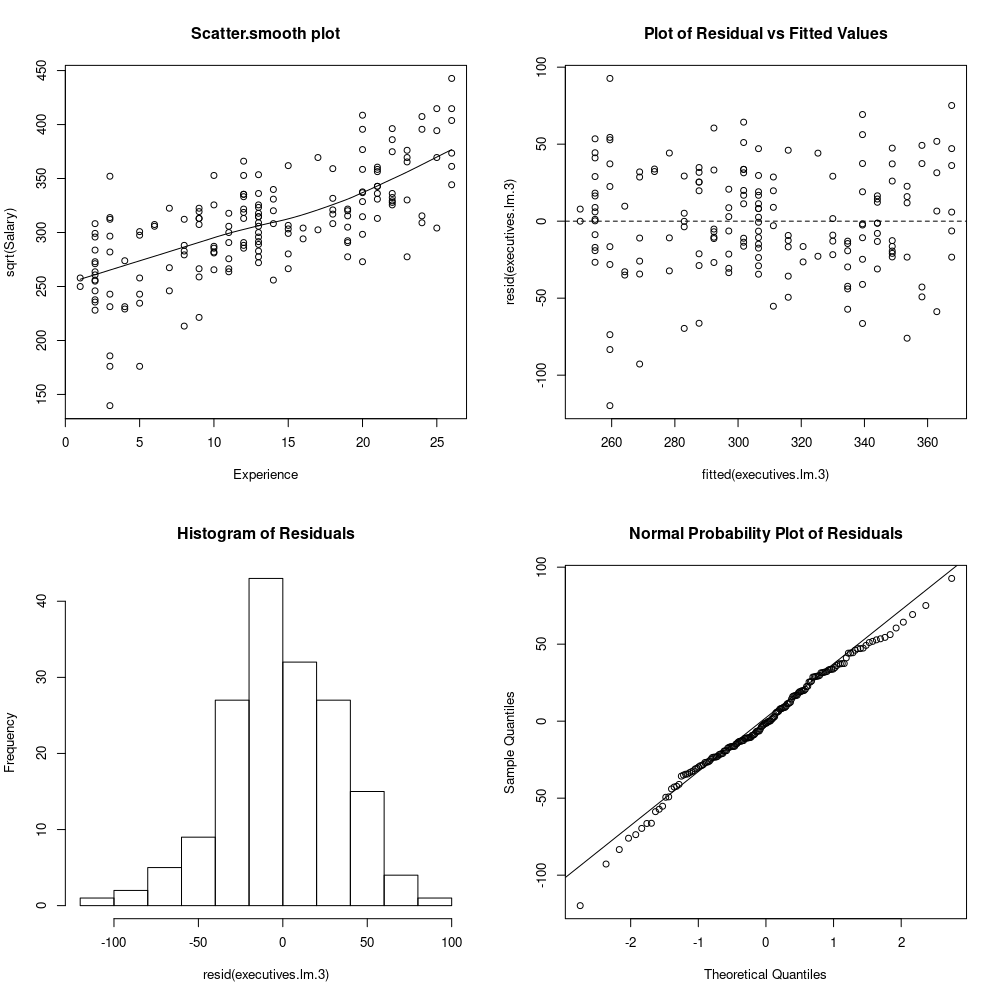
Scatter Plot: Linear & decreasing variance (including outliers)

Residuals Vs Fitted Values: Linear & decreasing variance (including outliers)

Histogram of residuals: Heavily skewed left - non-normal

Normal Probability Plot: Major departures at ends - non-normal

**Diagnostic Plors for exectuives.lm.3**



Scatter Plot: Linear & constant variance.

Residuals Vs Fitted Values: Linear & constant variance.

Histogram of residuals: Normal (slight skew left)

Normal Probability Plot: Moderate departures, mostly left tail – approximately normal.

**B) Which model to choose?**

If I had to pick one of the three currently I would go with model 3 (sqrt).

From visually inspecting the three sets of diagnostics plots I think model 2 (log) is obviously an inferior choice to 1 and 3 with it’s variance and normality issues.

It is a much harder choice to make between model 1 and 3. Both models appear to be linear with constant variance, so the deal-breaker between them will be how normally distributed the residuals are.

Even though model 3 has a bit more departure at the tails in the normal probability plot than model 1, it seems to stick closer to the normal distribution at the centre of the distribution.

The reason ultimately that I am not choosing model 1 is due to the histogram having two peaks, which could indicate a bimodal distribution, and a bimodal distribution will never be normal.

**C) Further action required?**

There are two things I would recommend to do currently.

The first thing I would do is see if there is some sort of test statistic that can be computed for determining the normality of a dataset. I would perform this test on both models 1 and 3 to see if the tests agree with my conclusion from just visually inspecting the graphs.

It is possible model 1’s residuals are actually normally distributed and not bimodal, in which case I would choose that model over 3 since model 1 would satisfy all the necessary assumptions.

The next thing I would do is try out some other combinations of the data transformations such as using the inverse transformation, or applying transformations to both Salary and Experience to see if any of those perform better on the diagnostic plots than the three current models.

The purpose of this is to see if the departures in the normal probability plots can be minimized further.

I am not sure how likely it is that better overall performance will be found since model 1 and 3 are quite good really but it can’t hurt to try one or two more transformations.

**Code to make plots**

**Model 1**

*> par(mfrow=c(2, 2))*

*> scatter.smooth(Experience, Salary, main="Scatter.smooth plot")*

*> plot(fitted(executives.lm.1), resid(executives.lm.1), main="Plot of Residual vs Fitted Values")*

*> abline(h=0, lty=2)*

*> hist(resid(executives.lm.1), main="Histogram of Residuals")*

*> qqnorm(resid(executives.lm.1), main="Normal Probability Plot of Residuals")*

*> qqline(resid(executives.lm.1))*

**Model 2 and Model 3**

Replace Salary in scatter.smooth with the appropriate transformations and use executives.lm.2 and executives.lm.3 in place of executives.lm.1.s