|  |  |
| --- | --- |
| **Module:** | ST2053 |
| **Name:** | Cian Linehan |
| **Student Number:** | 119301381 |
| **Chapter:** | 4 |

(a)

> salary = executives.df$Salary

> experience = executives.df$Experience

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

> #model 1

> scatter.smooth(experience, salary, main='Smooth.scatter Plot')

> lm1 = lm(salary~experience)

> plot(fitted(lm1), resid(lm1), main='Plot of Residual V Fitted Values')

> abline(h=0,lty=2)

> hist(resid(lm1), main='Histogram of Residuals')

> qqnorm(resid(lm1),main='Normal Probability Plot of Residuals')

> qqline(resid(lm1))

> #log y

> scatter.smooth(experience, log(salary), main='Smooth.scatter Plot')

> lm2 = lm(log(salary)~experience)

> plot(fitted(lm2), resid(lm2), main='Plot of Residual V Fitted Values')

> abline(h=0,lty=2)

> hist(resid(lm2), main='Histogram of Residuals')

> qqnorm(resid(lm2),main='Normal Probability Plot of Residuals')

> qqline(resid(lm2))

> #sqrt y

> scatter.smooth(experience, sqrt(salary), main='Smooth.scatter Plot')

> lm3 = lm(sqrt(salary)~experience)

> plot(fitted(lm3), resid(lm3), main='Plot of Residual V Fitted Values')

> abline(h=0,lty=2)

> hist(resid(lm3), main='Histogram of Residuals')

> qqnorm(resid(lm3),main='Normal Probability Plot of Residuals')

> qqline(resid(lm3))

See graphs below.

Model 1 (Y~X):

Scatter plot: We see non-linearity due to a kink upwards in the slope of the line around 15 years' experience. There is roughly constant variance here.

Residual v fitted values: There is non-constant variance here because there is increasing variance as y approaches its min and max values. It does not appear to be sufficiently linear due to the smaller residuals towards the mean Y values (mid-graph), I have indicated this below in red in figure 1 which resembles a double U-shape.

Histogram: Non-normal (right skew)

Normal probability plot: Normal - some minor departures at both extremes.

Model 2 (log(Y)~X):

Scatter plot: Roughly linear despite a possible outlier around x=3 but clearly decreasing variance.

Residual v Fitted values: Non-linear due to the sideways v shape as I have drawn in the figure below. Also decreasing variance.

Histogram: Non-normal (extreme left skew and too few extremal values)

Normal probability plot: non normal - several substantial departures at both extremes and also systematic departures.

Model 3 (sqrt(Y)~X):

Scatter plot: Non-linear due to a kink in the plot. Constant variance if we ignore an outlier or two around x=3.

Residual v fitted values: Linear and constant variance ignoring the two-extreme y-values around x=260.

Histogram: Normal despite a slight left skew.

Normal probability plot: some minor departures at both ends - approximately normal.

![Chart, scatter chart

Description automatically generated]() ![Chart, scatter chart

Description automatically generated]()



![Chart, scatter chart

Description automatically generated]()

From top left to right;

Model 1: Y ~ X

Model 2: log(Y) ~ X

Model 3: sqrt(Y) ~ X

(b)

I would recommend model 3 (sqrt(Y)~X) as it mostly satisfies all 3 assumptions (linearity, constant variance and normality) which are in our aggregate analysis. Although some of the assumptions in this model may be borderline not satisfied (subject to opinion and best judgement), it is clear it is a better model than 1 or 2. In other words, it is the least worst. Model 1 is a mixed conclusion as it was found to be both normal and non-normal and also to have both constant and non-constant variance but was clearly non-linear. Model 2 was found to be clearly non-normal with decreasing variance and also probable non-linearity. Therefore model 3's 6 conclusions were by far the most satisfactory with only one of which was not (non-linearity in the smooth scatter plot). In comparison, model 1 had just 2/6 satisfactory conclusions whilst model 2 had 1/6.

(c)

Whilst model 3 is the best out of these 3 for this data, I would recommend exploring further options as 1 conclusion was not satisfactory for our aggregate analysis. I did make some conclusions here which seemed almost too close to call, because I had to ignore what I thought could be possible outliers or cases of high influence to come to these conclusions. Therefore, it is not perfectly clear if this is the most suitable model available for this data. It is also worth exploring whether these points were outliers and/or cases of high influence or not and whether they can be ignored.

It is clear from model 3 that the square root on the Y variable has a positive impact on the model, we should perhaps explore combining this with sqrt(X) or other combinations like 1/X or log(X) to see if these could possibly improve the model further and make it a clear winner. Other combinations of transformations may also work and need to be explored.