Statistical Modelling Coursework 1

# Load the data set and assign the data to an explanatory variable and a response variable

library(readxl)  
studentscores <- read\_excel("C:/Users/shay/Downloads/Book1.xlsx")  
View(studentscores\_1)  
# checking the column names for each column  
dimnames(studentscores)  
explanatory <- studentscores$'Sleep Hours'  
response <- studentscores$'Performance Index'

# Explain briefly why you chose this data set

I chose this data set as I’m interested in seeing the correlation between sleep and performance, I find it relevant to my own lifestyle. I could use this to better understand the differences between certain hours of sleep, and if there is a significant difference between neighbouring sleep values.

# Construct a simple linear regression model and display a summary of the model results, showing the values of the least squares estimates of the two regression parameters

## Code

# Plot the graph  
plot(explanatory, response, main='Sleep vs Performance', xlab='Amount of Sleep', ylab='Performance Index')  
# Find the least squares estimates of regression parameters using linear model function  
linmod <- lm(response ~ explanatory)  
print(linmod)  
summary(linmod)  
abline(linmod) # Add line of best fit

## Summary (Output)

Call:

lm(formula = response ~ explanatory)

Residuals:

Min 1Q Median 3Q Max

-40.279 -12.426 1.916 13.574 31.916

Coefficients:

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

(Intercept) 50.9630 8.7882 5.799 3.77e-07 \*\*\*

explanatory 0.9023 1.3019 0.693 0.491

---

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

Residual standard error: 17.26 on 53 degrees of freedom

Multiple R-squared: 0.008982, Adjusted R-squared: -0.009717

F-statistic: 0.4804 on 1 and 53 DF, p-value: 0.4913

The two regression parameters are 50.9630 and 0.9023 which respectively define the y-intercept and the gradient.

# Write in one sentence an intepretation of the model and its two parameters

Students with the minimum amount of sleep have an expected performance index rating of about 51 with this model, and it claims students on average increase their performance by 0.9 for each extra hour of sleep.

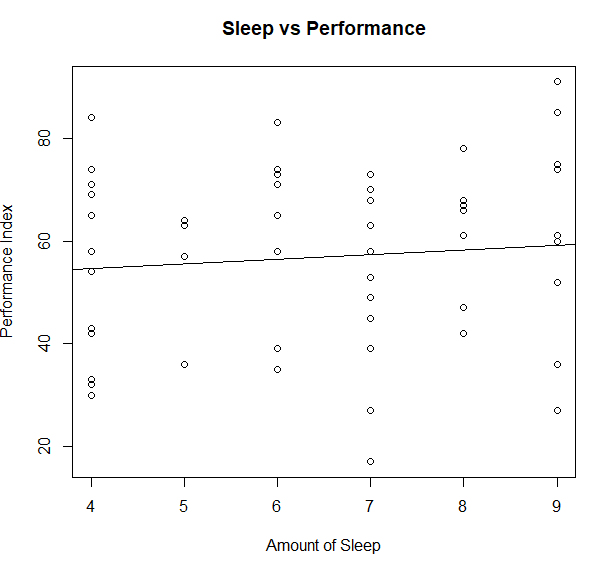
# How well does the model explain your data? (500 words)

## Anova Table & Summary

### Code

plot(explanatory, response, main='Sleep vs Performance', xlab='Amount of Sleep', ylab='Performance Index')  
abline(linmod) # Add line of best fit  
summary(linmod)  
anova(linmod) # Analysis of variance table

### Summary

An important thing to consider is the R^2 value, which is 0.8982%, meaning my model doesn’t account for much of the variation at all, this is also reflected in the fitted line plot with respect to the data (visible on the right).

We can also look at the p-values of the intercept (3.77e-07), which indicates our observed relationships are not due to chance, allowing us to be a little more confident in estimations/conclusions drawn from the model.

Further, we have a pretty high RSE (Residual Standard Error), from which we can draw that the models predictions are deviating from the given values in a non-negligible manner. This highlights the variability that isn’t being accounted for in the model

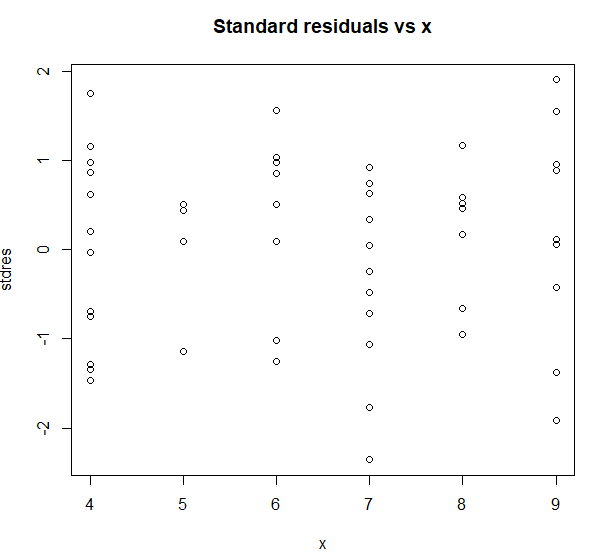
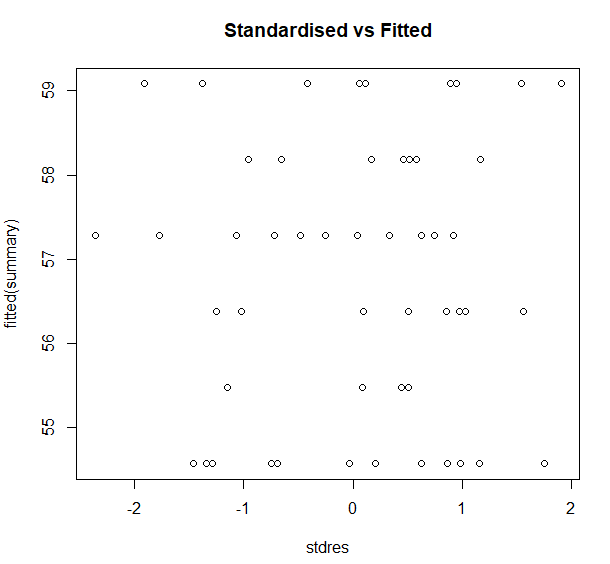
From this we can already understand that our linear model isn’t that reliable with this particular dataset, and that we should be very cautious when drawing inferences from the model.

## Residual Plot & Fitted vs Residuals

### Code

# Plot standard residuals vs x  
stdres <- rstandard(linmod)  
plot(explanatory, stdres, main="Standard residuals vs x")  
# Plot standard residuals vs fitted values  
plot(fitted(linmod), stdres, main="Standardised vs Fitted")

### Summary

In these graphs we’re looking for a completely random variance, and the absence of any correlation. The residual plot & fitted vs stdres graphs seem to be homoscedastic, there is no obvious pattern in the graph and all points seem somewhat random (or atleast random enough for it to not be a cause of concern). It also seems as if the variance is constant, or in particular, that it doesn’t resemble a funnel or trumpet of any sorts. This means the models predictive power could be satisfactory, and we may be appropriately capturing the data, however we know from the RSE and R^2 values that its predictive power is not as viable as these graphs make it seem.

## Quartile-Quartile Plot

### Code

# Quantile-Quantile plot  
qqnorm(stdres) # plot graph  
qqline(stdres) # add fitted line

### Summary

Next, we have our Quartile-Quartile plot, the closer our points resemble the line the better, as we can see our points are quite close to the line, but deviates a bit at the head and tail, this is a possible cause of concern and shows there might be a problem with our linear model; The presence of outliers or heavy tailed distributions, this places a small doubt on the assumption of normality. However our points are generally all close enough to the line to disregard this and say that the assumption of normality for the residuals is likely satisfied and supports the notion that prediction intervals may be satisfactorily reliable.

## Shapiro-Wilk normality test

### Code

# Shapiro-Wilk normality test  
shapiro.test(stdres)

#### Results

Shapiro-Wilk normality test

data: stdres

W = 0.97168, p-value = 0.2188

### Summary

Our Shapiro-wilk test should align well with our findings from the Quantile-Quantile plot, and this seems to be the case with our p value of 0.2188 which is above the significance level, providing re-assurance on the assumption of normality.

## Conclusion

All in all, some of our tests seem to contradict the others, my personal opinion on this is that more samples may be needed a cross a wider range of x values (hours of sleep) in order to more accurately determine if a linear model is suitable for this data and if it can be confidently used to make reliable inference and prediction intervals.

# Files