MTH5120 Statistical Modelling 1

Assessed Coursework 1 - Solution

This solution uses an example dataset to show an answer that would score full marks on this coursework. Your own solutions will use your own datasets and will therefore all be different to this one. In particular your analysis of how well the regression model explains the data in part (e) will be particular to your data and modelling results.

Question

(a) Load the data set that you submitted previously to QM Plus into R and assign the data to an explanatory variable and a response variable. [4]

```
> setwd("C:/Downloads")
> cwdata <- read.csv("AI.csv")
> x <- cwdata[,1]
> y <- cwdata[,2]
```

(b) Explain briefly (in less than 50 words) why you chose this data set. [3]

The dataset compares different AI large language models and contains their scores in a test of "knowledge" (the MMLU benchmark score) as the response variable and the size of training database (billions of data points) as the explanatory variable. I am interested to see the extent to which development in LLMs like ChatGPT can be attributed to increased data processing capabilities in development.

(c) Construct a simple linear regression model using your data and display a summary of the model results. Copy your R code and output into your Word document and then write down the values of the least squares estimates of the two regression parameters. [4]

```
> model <- lm(y~x)
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-20.3389  -9.7144  -0.1738  10.2485  20.2485

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.263e+01  3.297e+00  12.930 7.27e-11 ***
x           3.855e-03  7.599e-04   5.073 6.75e-05 ***
```

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

Residual standard error: 12.83 on 19 degrees of freedom
Multiple R-squared:  0.5753,  Adjusted R-squared:  0.5529
F-statistic: 25.74 on 1 and 19 DF,  p-value: 6.754e-05
```

The least squares estimates of the regression parameters are

$\hat{\beta}_0 = 42.63$ and $\hat{\beta}_1 = 0.003855$

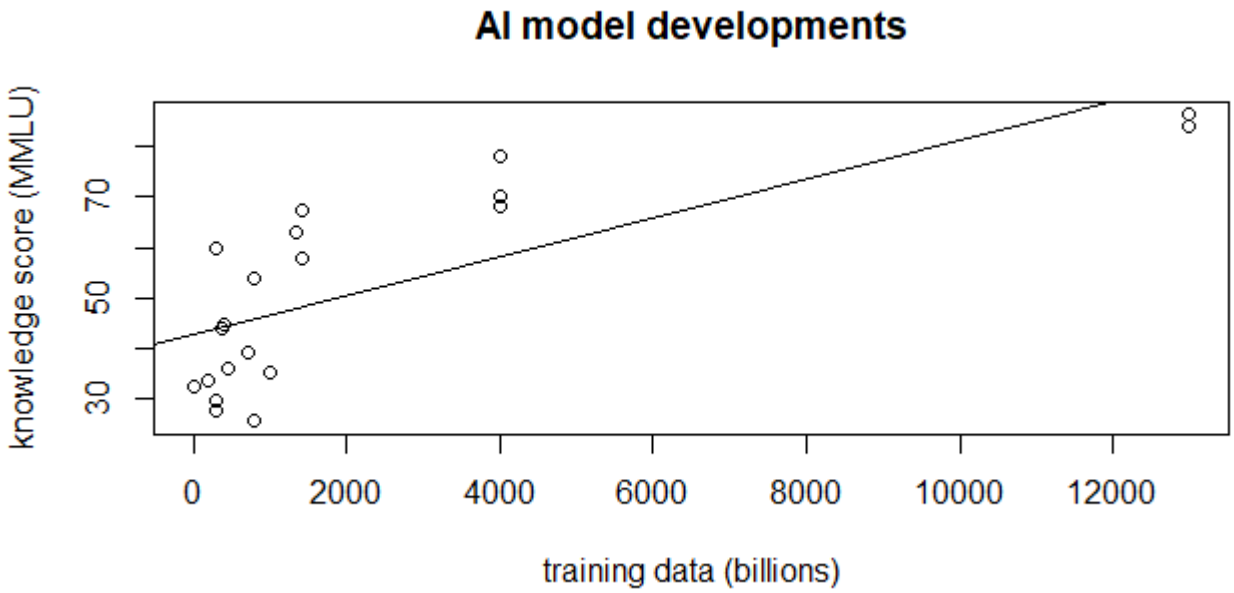
(d) Write in one sentence an interpretation of the model in (c) and its two parameters. [2]

A large language model with no training data would score 42.63% on the knowledge test and for every 1 billion data points in the training data that score would increase by 0.003855%.

(e) How well does the model in (c) explain your data? Your answer to this part should be typed into your Word document in no more than 500 words. You should use the methods covered in the module lectures and IT labs for assessing a simple linear regression model and then make your own conclusions. Note that the marks for this part will be awarded for the quality of your analysis and for conclusions made from evidence generated in R and not for how well the model explains the data. Where you use output and plots from R to support your conclusions these should be copied into your Word document along with the R code used to generate them. [12]

From the initial summary model output the coefficient of regression or R^2 is 57.5% suggesting that training data size explains some but by no means all of the advance in AI capabilities. A plot of the data and the regression line demonstrates this.

```
> plot(x,y, main = "AI model developments", xlab = "training
data (billions)", ylab = "knowledge score (MMLU)")
> abline(model)
```



We complete the Analysis of Variance for the model.

```
> anova(model)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 4238.0  4238.0  25.736 6.754e-05 ***
Residuals 19 3128.7   164.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 25.736 > qf(0.95, 1, 19)
[1] TRUE
```

The Variance Ratio (25.736) is greater than the 5% upper critical value of Fisher's F distribution on 1 and n – 2 = 19 degrees of freedom (4.38) allowing us to reject the null hypothesis H0: $\beta_1 = 0$ at 95% significance. This is important here because numerically β_1 is close to zero but the F test assures us that the regression parameter as statistical significance.

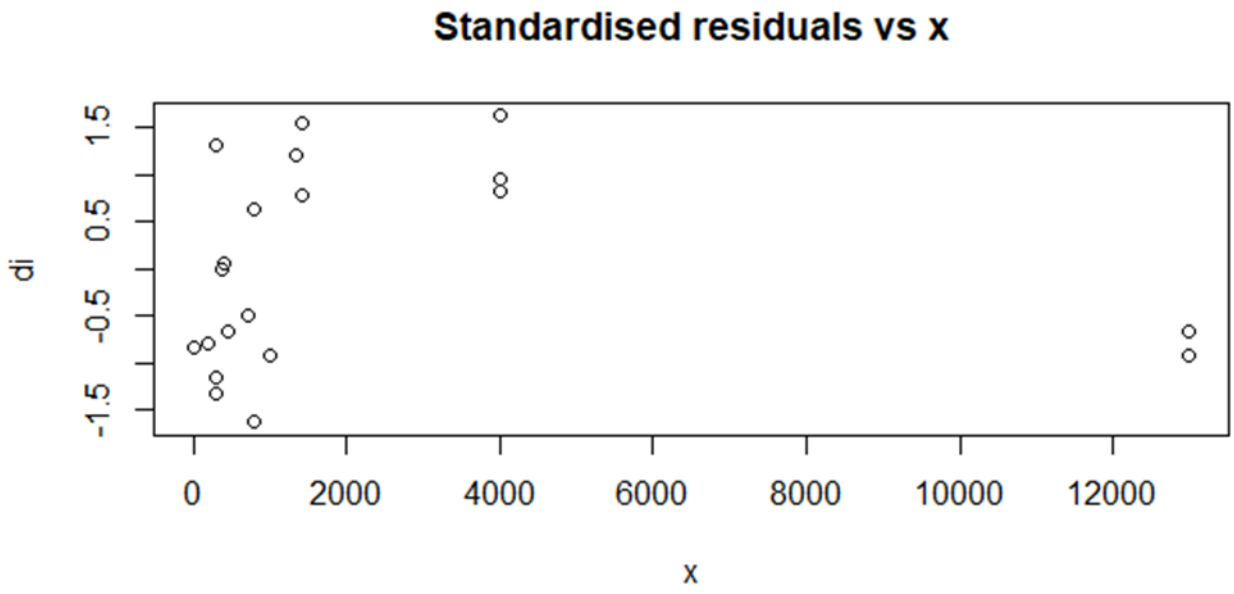
From ANOVA our Mean Square for Residuals MS_E = 164.7 which is our unbiased estimator for the model variance σ^2.

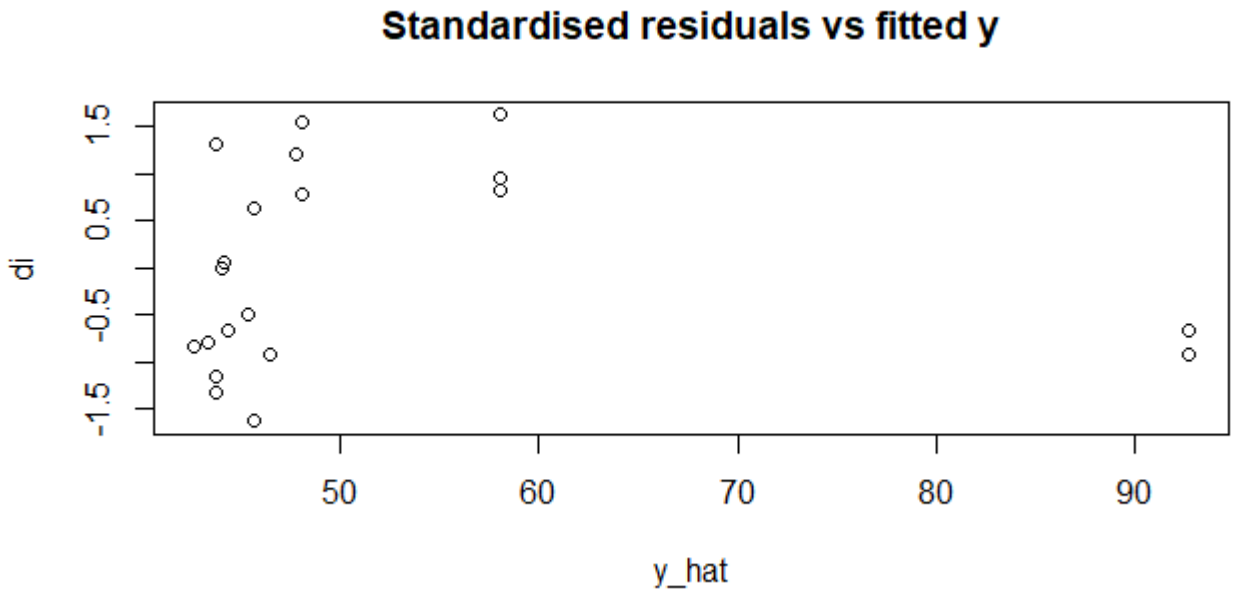
Next, we look at the standardised residuals and three residual plots.

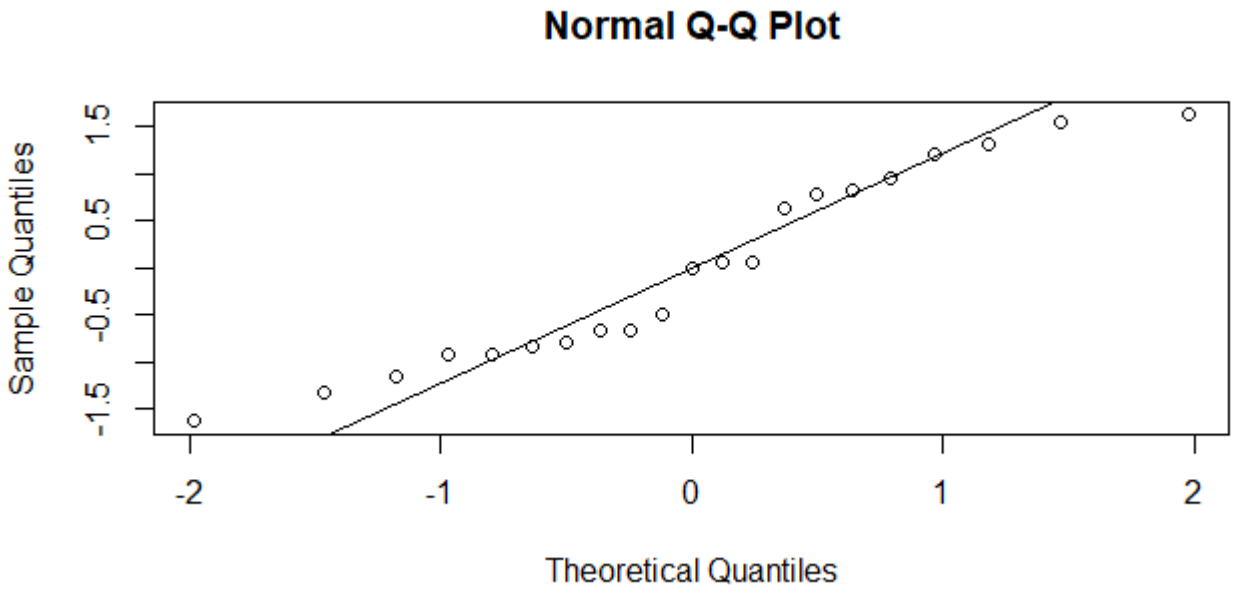
```
> di <- rstandard(model)
> y_hat <- fitted(model)
```

```
> plot(x,di, main = "Standardised residuals vs x")
> plot(y_hat,di, main = "Standardised residuals vs fitted y")
> qqnorm(di)
> qqline(di)
```

Which generates the following plots







The first plot (standardised residuals versus x) checks for the linear relationship assumption and we seek a random scatter plot with no obvious pattern. Two much larger x values are noticeable on the right of the plot but the remaining 19 do not show an obvious pattern. We would want to investigate the two larger x values as potential influential observations [beyond the scope of this coursework] but do not have reason to doubt the assumption of linearity.

The second plot (standardised residuals versus fitted y) checks for a constant variance and we seek a random scatter plot an in particular no “funnel” type shape. Out plot gives no reason to doubt the constant variance assumption.

The QQ plot checks the Normal distribution assumption. We seek a plot close to the QQ line. Our QQ plot does exhibit something of an “S” shape with the largest and smallest standardised deviations beyond what we might expect from the tails of a normal distribution. Visually there is some concern that the distribution of residuals might not be Normal. Although the Shapiro-Wilk test suggests that the normal assumption is valid given a p-value greater than 0.05.

```
> shapiro.test(di)

     Shapiro-Wilk normality test

data:  di

W = 0.93303, p-value = 0.1583
```

In summary, our simple linear regression model somewhat explains the data, produces a statistically significant regression parameter and results that justify our three main modelling assumptions. However the explanatory power of the model is not that large with less than 60% of the variability in knowledge test results explained by training data size. There are also two observations with much larger training data which need further investigation. It would seem that our simple linear regression

model provides a starting point for understanding developments in AI models but that more modelling work is needed.

[Total 25]