STAT7203: Week 12 Practical Questions

1. A study examined the relationship between a student's grades and SAT scores from high school and the grades they achieve in university. The study surveyed 105 students majoring in computer science at an American university. The data is taken from `onlinestatbook.com/2/case_studies/sat.html`.

   The data contains the following variables:

   `high_GPA` High school grade point average (scale 0.0 - 4.0)

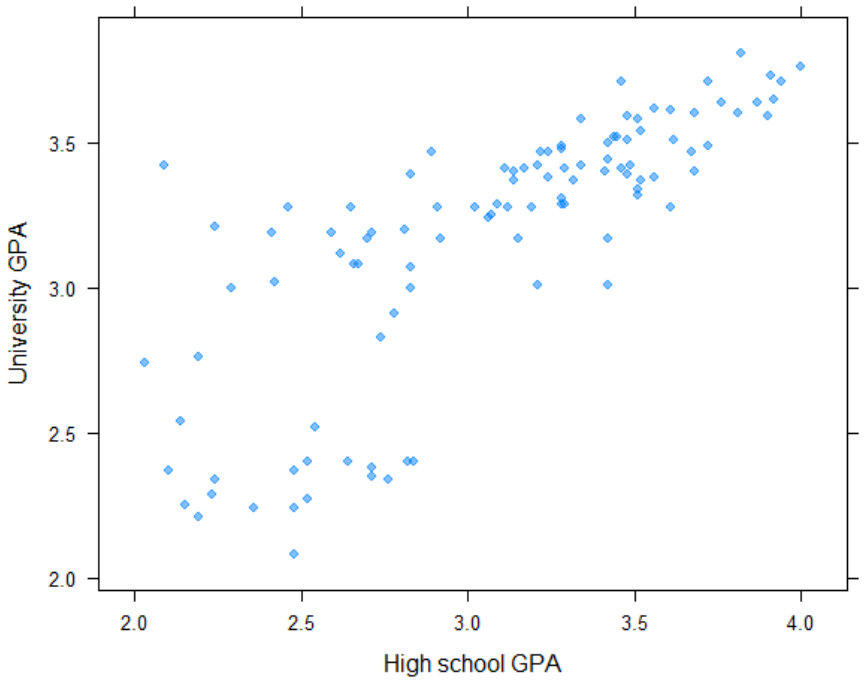
   `math_SAT` Math SAT score (scale 200 - 800)

   `verb_SAT` Verbal SAT score (scale 200 - 800)

   `comp_GPA` Computer science grade point average (scale 0.0 - 4.0)

   `univ_GPA` Overall university grade point average (scale 0.0 - 4.0)

   (a) Construct a plot of the students' university GPA against their high school GPA. Describe the plot. Do you think the assumptions of the linear regression model (with university GPA as the response variable and high school GPA as the explanatory variable) are satisfied?



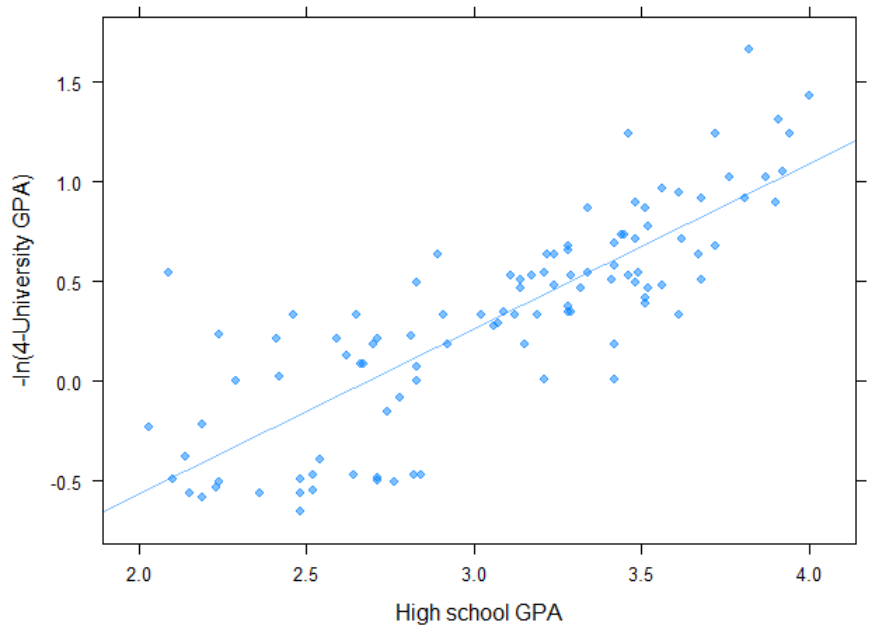
   (b) It appears that the variance of the university GPA decreases as the high school GPA increases. One way to deal with this is to apply a transformation to our response variable (university GPA) so that the variance of the transformed variable appears constant.

   Let Y be a students university GPA and define the new random variable $\tilde{Y} = -\ln(4-Y)$. Suppose we have the linear regression model for $\tilde{Y}$:

$$\tilde{Y} = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

   Express the median of Y in terms of β_0, β_1, and x.

(c) Construct a plot of $-\ln(4-$ `univ_GPA`$)$ against high school GPA. Does the variance of the transformed variable $-\ln(4-$ `univ_GPA`$)$ appear relatively constant? Does there appear to be a linear relationship between the mean of $-\ln(4-$ `univ_GPA`$)$ and high school GPA?



(d) Fit the linear regression model with $-\ln(4-$ `univ_GPA`$)$ as the response variable and high school GPA as the explanatory variable. Construct the diagnostic plots and comment on whether the plots are consistent with the model assumptions.

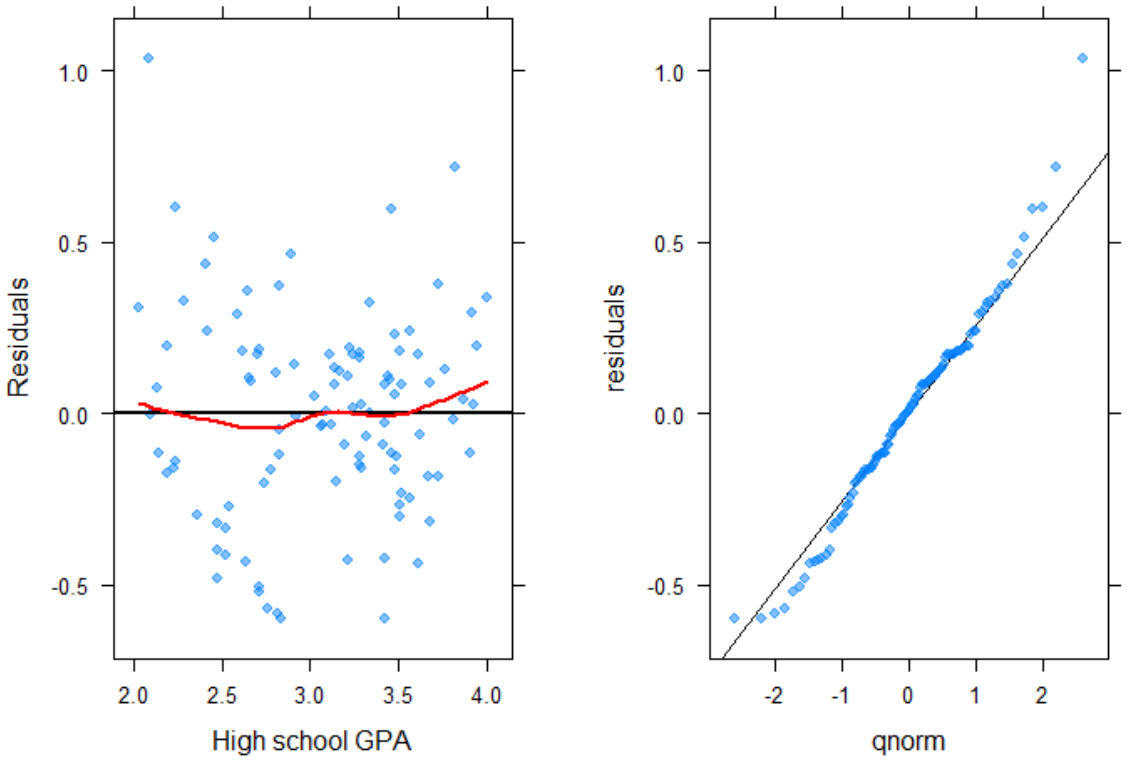
Henceforth, you may assume the assumptions of the linear regression model hold.

(e) The edited outplut of the linear regression is below

```
Call:
lm(formula = -log(4 - univ_GPA) ~ high_GPA, data = SAT)

Coefficients:
            Estimate Std. Error
(Intercept) -2.22452    0.17631
high_GPA     0.82861    0.05653
---

Residual standard error: 0.2978 on 103 degrees of freedom
Multiple R-squared:  0.676,Adjusted R-squared:  0.6728
F-statistic: 214.9 on 1 and 103 DF,  p-value: < 2.2e-16
```

Write out the fitted regression line and briefly interpret each of the estimated coefficients.

(f) Does the data provide evidence of a linear association between $-\ln(4-$ `univ_GPA`$)$ and high school GPA?

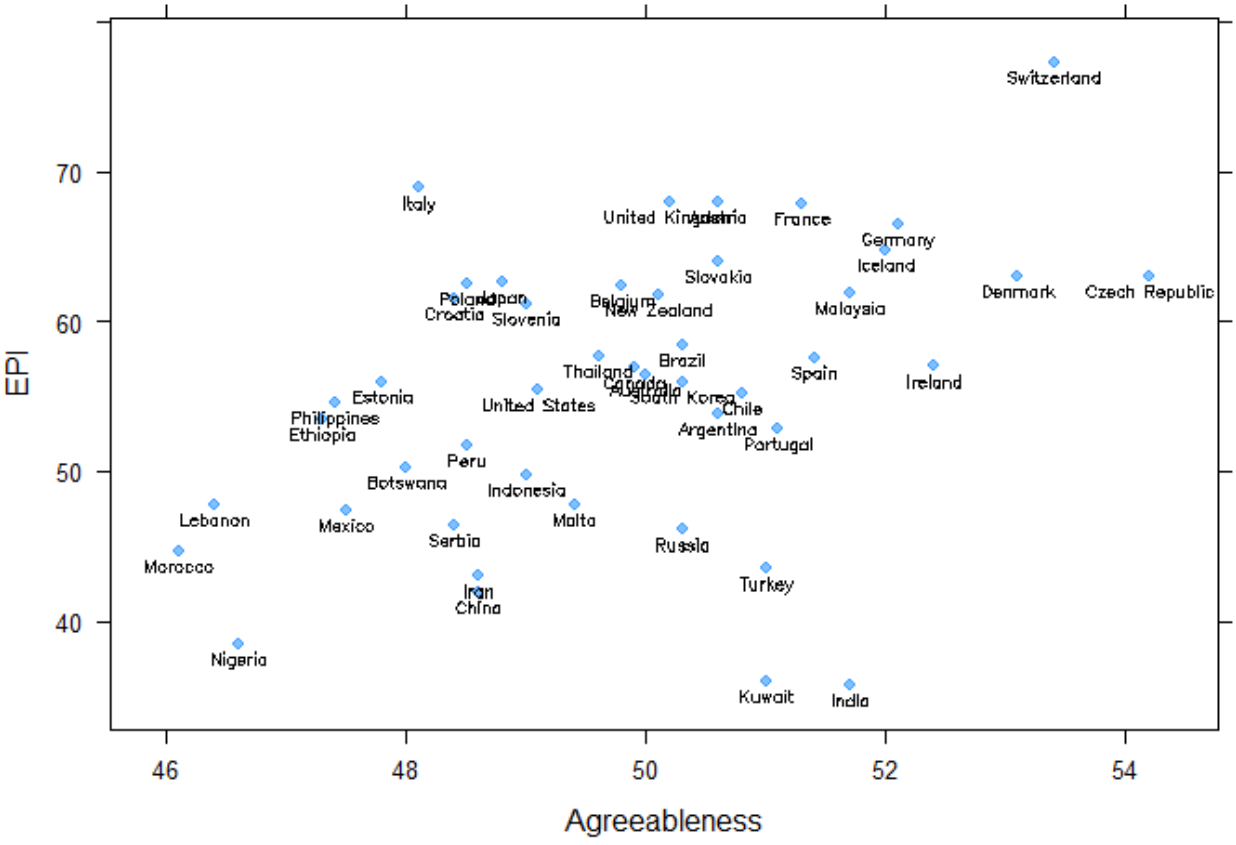


(g) Provide a 90% confidence interval for β_0.

(h) Build a model for University GPA using high school GPA and math SAT scores. Is there an association between University GPA and math SAT scores, after accounting for high school GPA?

2. An international study published in 2005 measured personality profiles for a range of countries. For example, Australia received an 'Agreeableness' score of 50.0 compared to 51.7 for Malaysia, suggesting Australians are less 'agreeable' than Malaysians, while the 'Openness' score for Australia was 50.7, higher than the corresponding score of 47.5 for Malaysia.

   A researcher in 2014 hypothesised that populations with higher levels of Agreeableness and Openness would be characterized by more sustainable environmental policies. He combined the data from the earlier study with scores on the Environmental Performance Index (EPI), a measure of national environmental sustainability. There were 46 countries for which both sets of scores were available. The following figure shows the relationship between EPI and Agreeableness from this combined data:



   The researcher also modelled EPI by Agreeableness and Openness together using multiple linear regression. R gave the following output for this analysis:

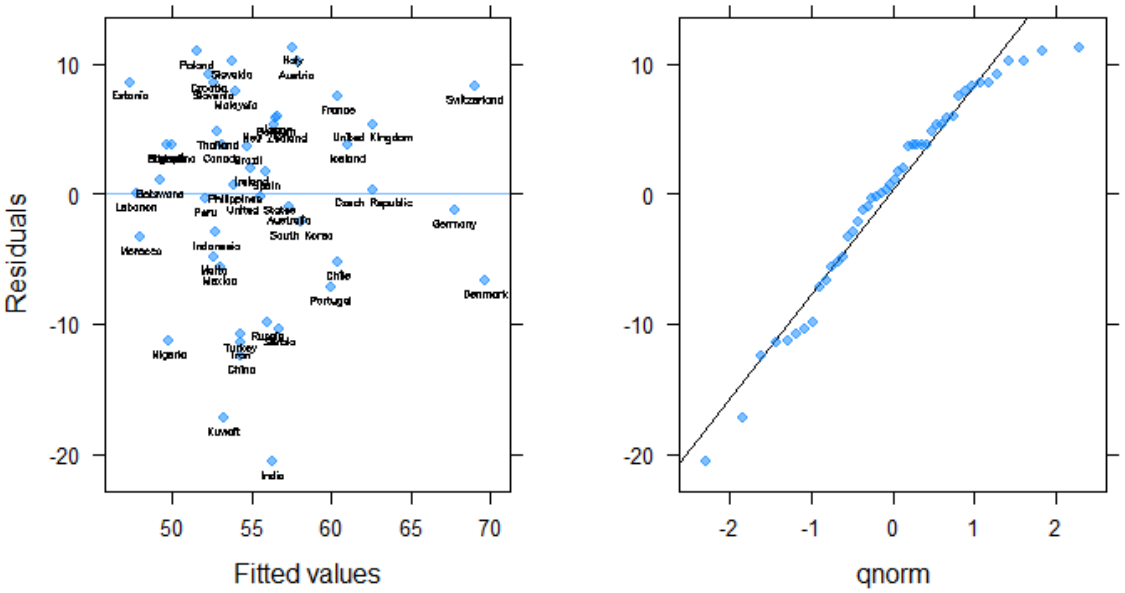
```
lm(formula = EPI ~ Agreeableness + Openness, data = EPI)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -101.9496    37.5253  -2.717  0.00946 **
Agreeableness    1.3826     0.6810   2.030  0.04854 *
Openness         1.7787     0.6252   2.845  0.00678 **
```

(a) The output from the multiple regression analysis includes three p-values. Which of these are relevant to the researcher? Give a brief interpretation of the evidence they provide.

(b) The linear regression model for EPI with Agreeableness as the only explanatory variable had an R^2 value of 0.165. For the multiple regression model, with Openness added, will the R^2 value be higher or lower than 0.165? Briefly justify your answer.

(c) Australia has an Agreeableness score of 50.0, an Openness score of 50.7 and an Environmental Performance Index of 56.4. Calculate the residual associated with Australia in the multiple regression model.

(d) The following figures were generated by `R` to help check the assumptions underlying the linear regression: Comment on the validity of the assumptions underlying linear



regression for this data with reference to these figures.