RESEARCH SCHOOL OF
FINANCE, ACTUARIAL STUDIES AND APPLIED STATISTICS
College of Business & Economics, The Australian National University

REGRESSION MODELLING

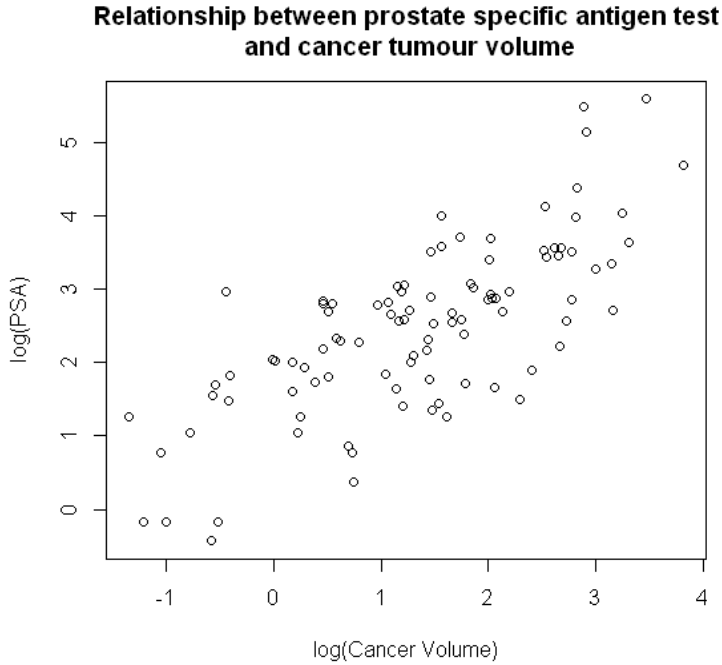
(STAT2008/STAT4038/STAT6038)

Solutions to Sample Assignment 1

Question 1 **(20 marks)**

The dataset `prostate` comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy (a surgical procedure). Of the variables included in this dataset, `lcavol` (log of the cancer volume) is a measure of the size of the cancer tumour and `lpsa` (log of the prostate specific antigen measure) is the result of a diagnostic blood test for prostate cancer.

(a) Plot `lpsa` against `lcavol`. Is there a significant correlation between `lpsa` and `lcavol`? Use *R* to conduct a suitable hypothesis test and present and interpret the results. **(4 marks)**



Plot shows a reasonably strong positive (and apparently linear) relationship.

```
> cor.test(lpsa, lcavol)

   Pearson's product-moment correlation

data:  lpsa and lcavol
t = 10.5483, df = 95, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6268370 0.8145819
sample estimates:
      cor
0.7344603
```

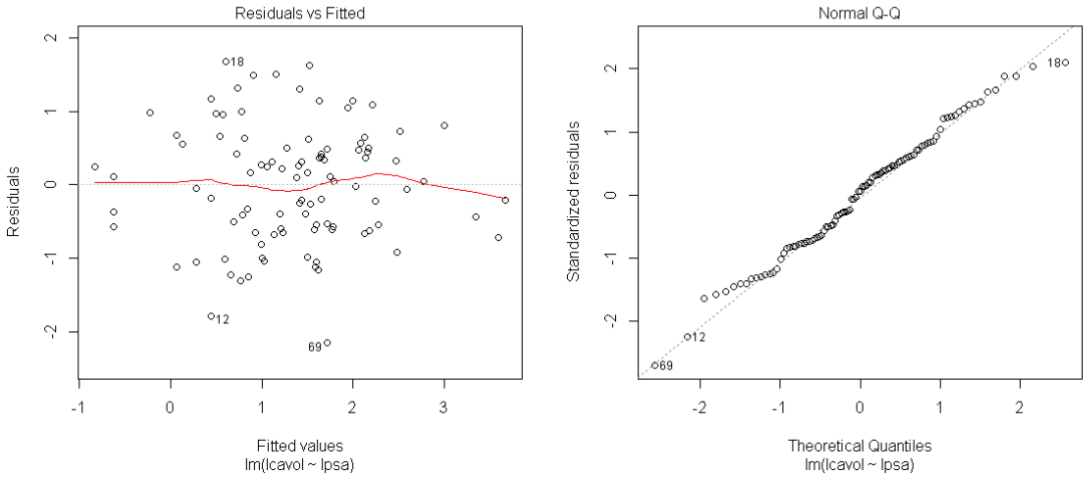
Test H_0: $\rho = 0$ vs H_A: $\rho \neq 0$, where ρ is correlation between log(PSA) and log(CaVol).

$t_{95} = 10.5$, $p << 0.05$, so reject H_0 in favour of H_A and conclude ρ is significantly different from 0. The observed sample correlation $r = 0.73$ suggests a strong positive correlation between log(Cancer Volume) and log(PSA).

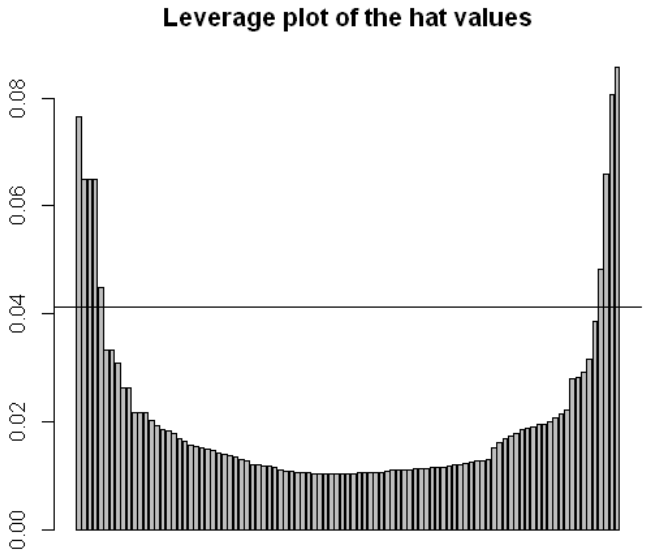
Question 1 continued

(b) Fit a simple linear regression model with `lcavol` as the response variable and `lpsa` as the predictor. Construct a plot of the residuals against the fitted values, a Q-Q plot of the residuals and a bar plot of the leverages for each observation. Comment on the model assumptions and on any unusual data points. **(4 marks)**

```
> prostate.lm

Call:
lm(formula = lcavol ~ lpsa)

Coefficients:
(Intercept)         lpsa
    -0.5086       0.7499
```



The residual plots do not show any real problems with the assumptions.



The leverage plot looks a little unusual, however, it is possible for points to have high leverage without being highly influential in the fit of the model and the residual plots didn't show any obvious problems.

The `lpsa` values were sorted in ascending order, so the observations with high leverage are the relatively few low and high `lpsa` values – we will return to the issue of discordant observations later in the course in context of multiple regression.

Question 1 continued

(c) Produce the ANOVA (Analysis of Variance) table for the SLR model in part (b) and interpret the results of the F test. Are these results consistent with the hypothesis test you conducted in part (a)? **(4 marks)**

```
> anova(prostate.lm)
Analysis of Variance Table

Response: lcavol
          Df Sum Sq Mean Sq F value    Pr(>F)
lpsa       1 71.938  71.938  111.27 < 2.2e-16 ***
Residuals 95 61.421   0.647
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$H_0 : \frac{\sigma^2_{Model}}{\sigma^2_{Error}} = 1 \quad H_A : \frac{\sigma^2_{Model}}{\sigma^2_{Error}} > 1$$

$F_{1,95} = 111.3$, $p << 0.05$, so reject H_0 in favour of H_A and conclude the variance explained by the model is large compared to the error variance, which means that the model involving `lpsa` is explaining a significant proportion of the variability in `lcavol`.

The results are definitely consistent with the test on the correlation in part (a) (identical p-values); as in the context of simple linear regression, the two tests are equivalent.

(d) What are the estimated coefficients of the SLR model in part (b) and the standard errors associated with these coefficients? Interpret the values of these estimated coefficients and perform t-tests to test whether or not these coefficients differ significantly from zero. What do you conclude as a result of these t-tests? **(4 marks)**

```
> summary(prostate.lm)

Call:
lm(formula = lcavol ~ lpsa)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15948 -0.59383  0.05034  0.50826  1.67751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.50858    0.19419  -2.619   0.0103 *
lpsa         0.74992    0.07109  10.548   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8041 on 95 degrees of freedom
Multiple R-squared:  0.5394,    Adjusted R-squared:  0.5346
F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

$$\text{Model}: \texttt{lcavol} = \beta_0 + \beta_1 \texttt{lpsa} + \varepsilon \quad \varepsilon \sim i.i.d.\ N\left(0, \sigma^2\right)$$

$$H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0$$

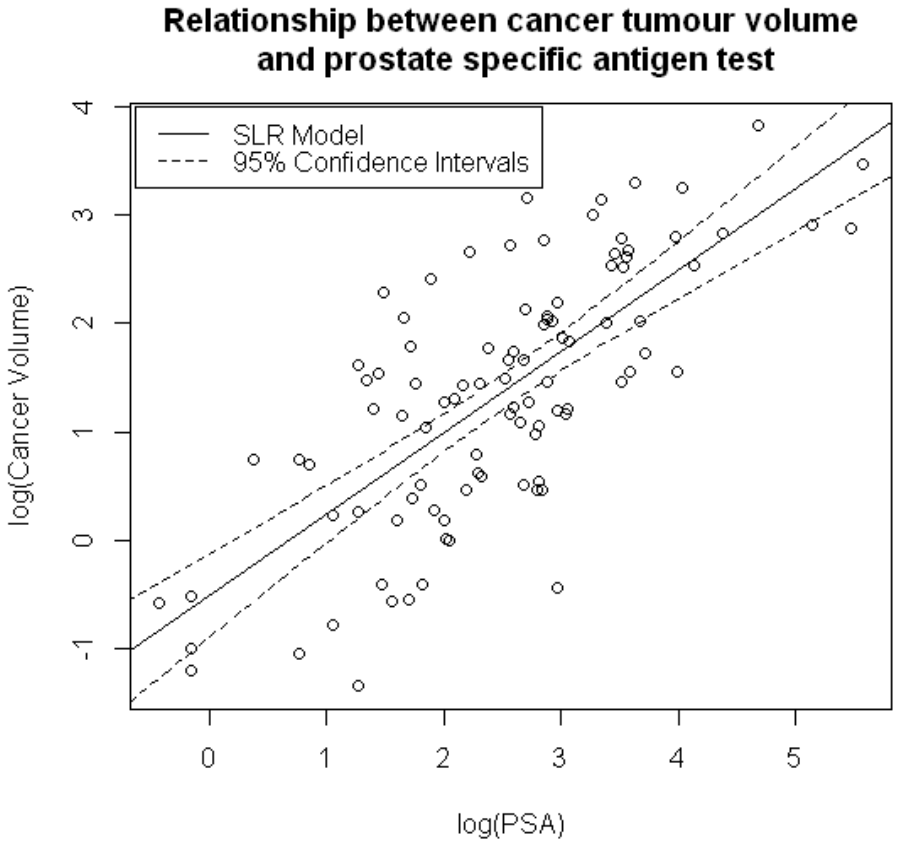
$t_{95} = 10.5$, $p << 0.05$, so reject H_0 in favour of H_A and conclude that the slope coefficient of `lpsa` is significantly different from 0, implying there is a significant linear relationship between `lcavol` and `lpsa`. Note this test is again directly equivalent to the tests in parts (a) and (c).

$$H_0 : \beta_0 = 0 \quad H_A : \beta_0 \neq 0$$

$t_{95} = -2.6$, $p < 0.05$, so reject H_0 in favour of H_A and conclude that the intercept is also significantly different from 0. On the log-log scale we do have values around the intercept, but it is a little difficult to interpret: `lpsa` = 0 implies PSA = 1 (a relatively low reading on the PSA scale) and `lcavol` = –0.50858 implies a relatively small cancer tumor, exp(–0.50858) = 0.6.

Question 1 continued

(e) Plot `lcavol` against `lpsa`. Include the fitted SLR model from part (b) as a line on the plot and also show 95% confidence intervals for the mean or expected value of `lcavol` (do NOT plot the 95% prediction intervals). Do the results of a PSA test appear to be a reliable predictor of the size of the prostate cancer tumour? **(4 marks)**

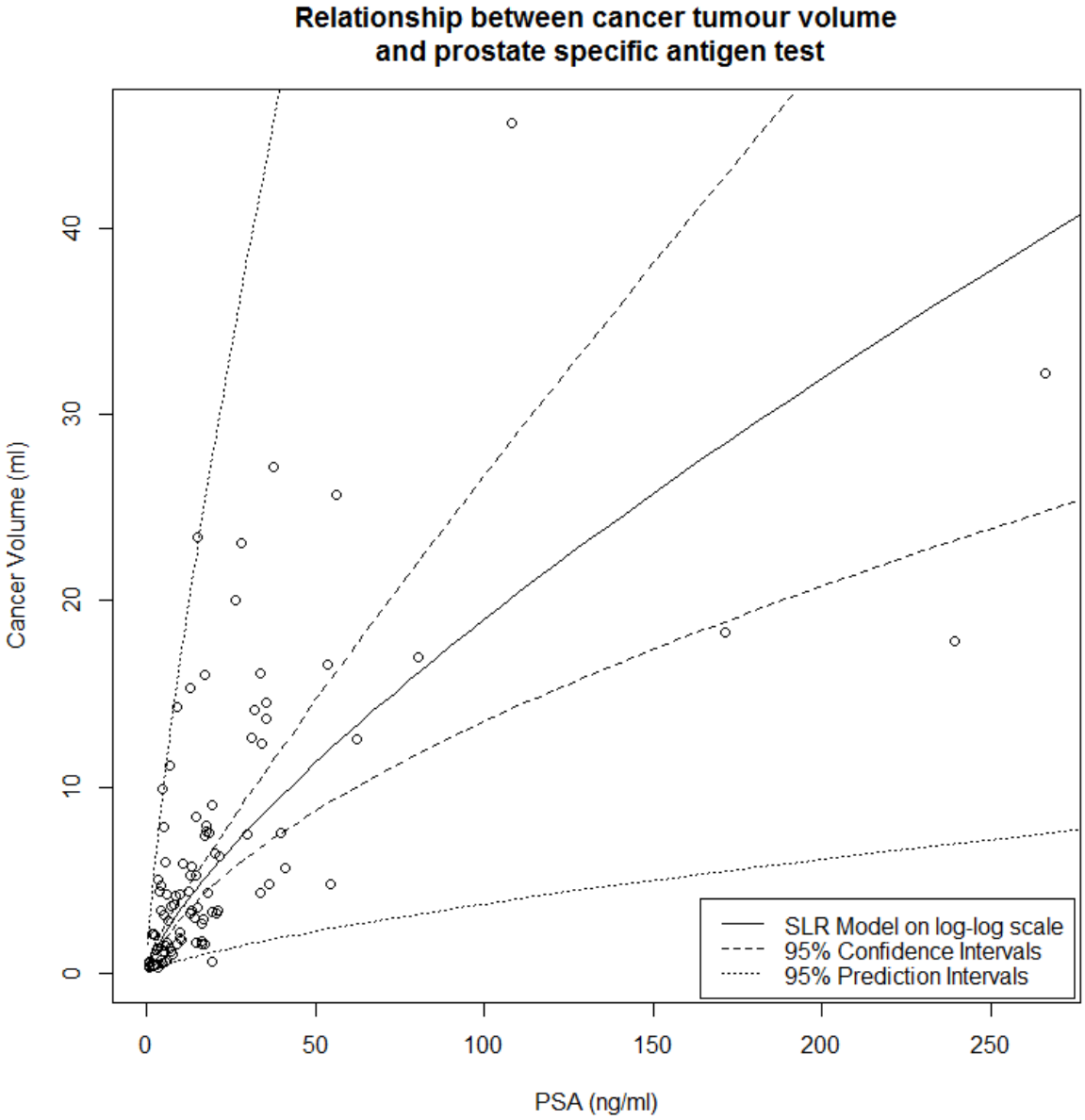


The highly significant hypothesis tests in parts (a), (c) and (d) and the fairly tight confidence intervals certainly indicate that there is a relationship between the PSA results and the volume of the cancer tumour, so tumour size does tend to increase as the PSA tests results increase.

A lot of observations lie outside the confidence intervals indicating that there is a lot of variability around this increasing relationship, so PSA is not necessarily a reliable indicator of tumour size. This would not be a good model to use to try and predict cancer size for an individual with a particular PSA result, rather than the mean or expected value over a group with the same PSA result. If you ignore the assignments instructions and plot the 95% prediction intervals, you find they span almost the full range of the observations.

The above results are consistent with the general advice available on the Internet that the PSA test is not considered to be reliable enough to use in isolation as a population screening tool. However, these data were all collected from men who actually have prostate cancer (i.e. they definitely had a tumour of some size), so this is not a model we could use to directly address the research question of how useful the PSA test would be in screening the general male population to diagnose prostate cancer. We might have a look at also including the other variables in a multiple regression model in another sample assignment.

Note that it is often best to present the data to the "client" on the original scale rather than on some transformed scale such as the log scale. This is probably not necessary in this instance (as the data were provided already transformed to the log scale), but here is the plot from part (e) "back-transformed" and this time I have included the prediction intervals:

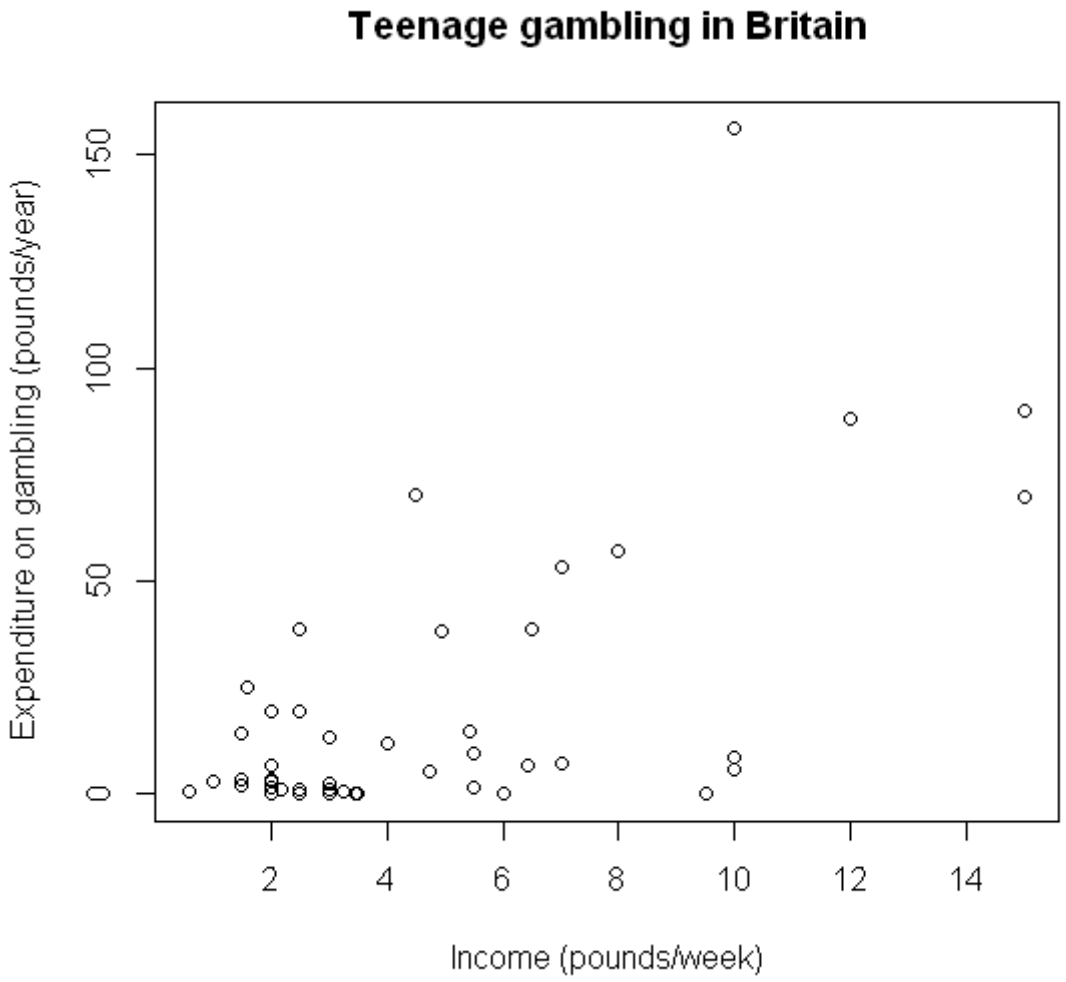


It is often useful to do some additional work in interpreting the model and the results of the analysis. This last plot was not a required part of the assignment and as such is not worth any marks, but done well, good and relevant additional work may redeem marks lost elsewhere in the same question.

Question 2 **(20 marks)**

The dataset `teengamb` concerns a study of teenage gambling in Britain. For this assignment, we are interested in whether a teenager's `income` (measured in UK £ per week) can be used to predict the amount they will `gamble` (gambling expenditure measured in UK £ per year), at least for the teenagers who do regularly gamble.

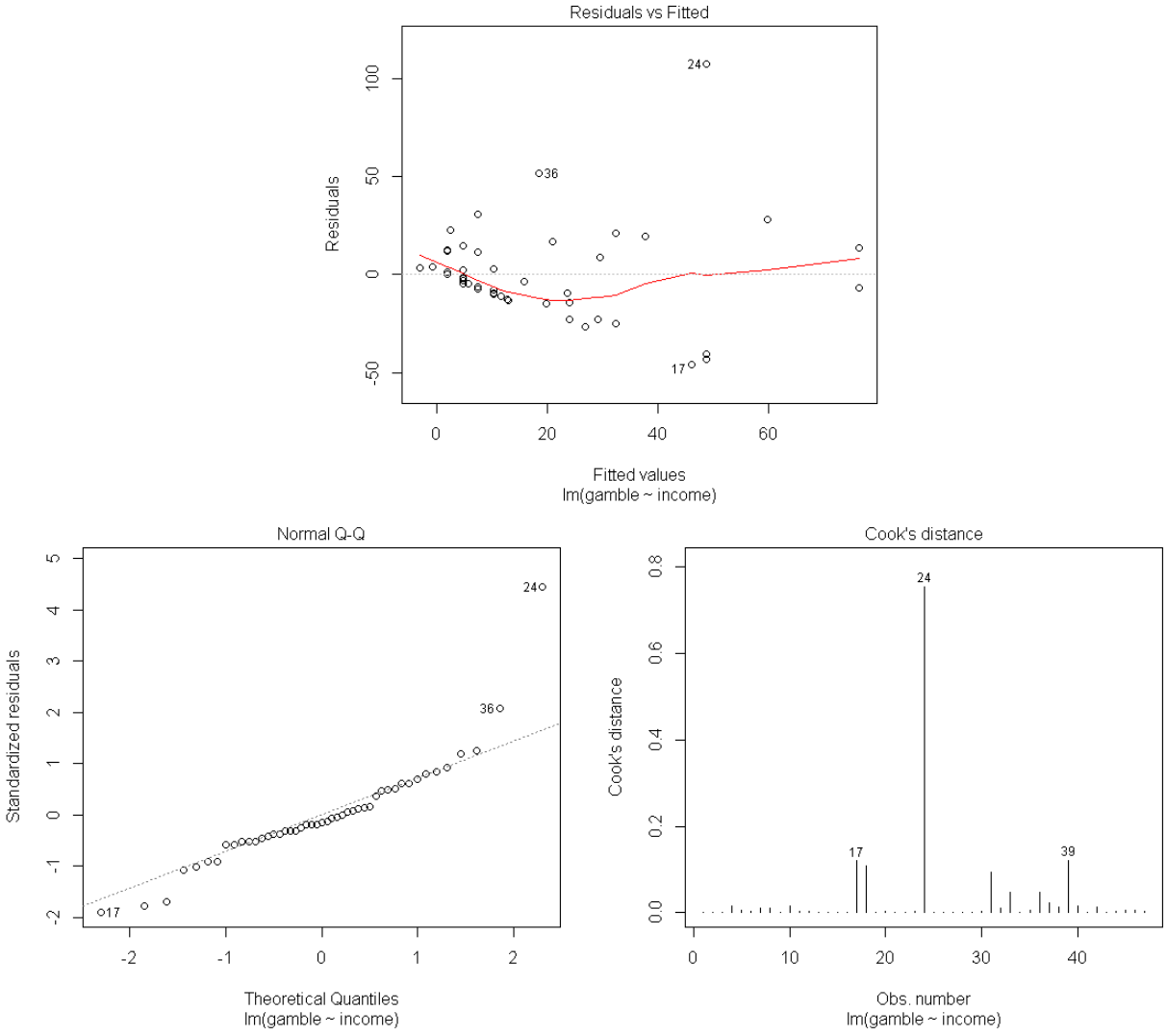
(a) Plot `gamble` against `income`. Describe the correlation shown in the plot. Would you expect a simple linear regression model to be a reasonable model for the relationship shown in the plot? **(4 marks)**



There is a moderate positive correlation, with expenditure on gambling increasing as income increases, however, the variability in this relationship is definitely greater for the higher incomes than it is for lower incomes, so I would not expect a SLR model to be a good fit.

(b) Fit a simple linear regression model with `gamble` as the response variable and `income` as the predictor. Construct a plot of the residuals against the fitted values, a Q-Q plot of the residuals and a bar plot of Cook's Distances for each observation. Comment on the model assumptions and on any unusual data points. **(4 marks)**

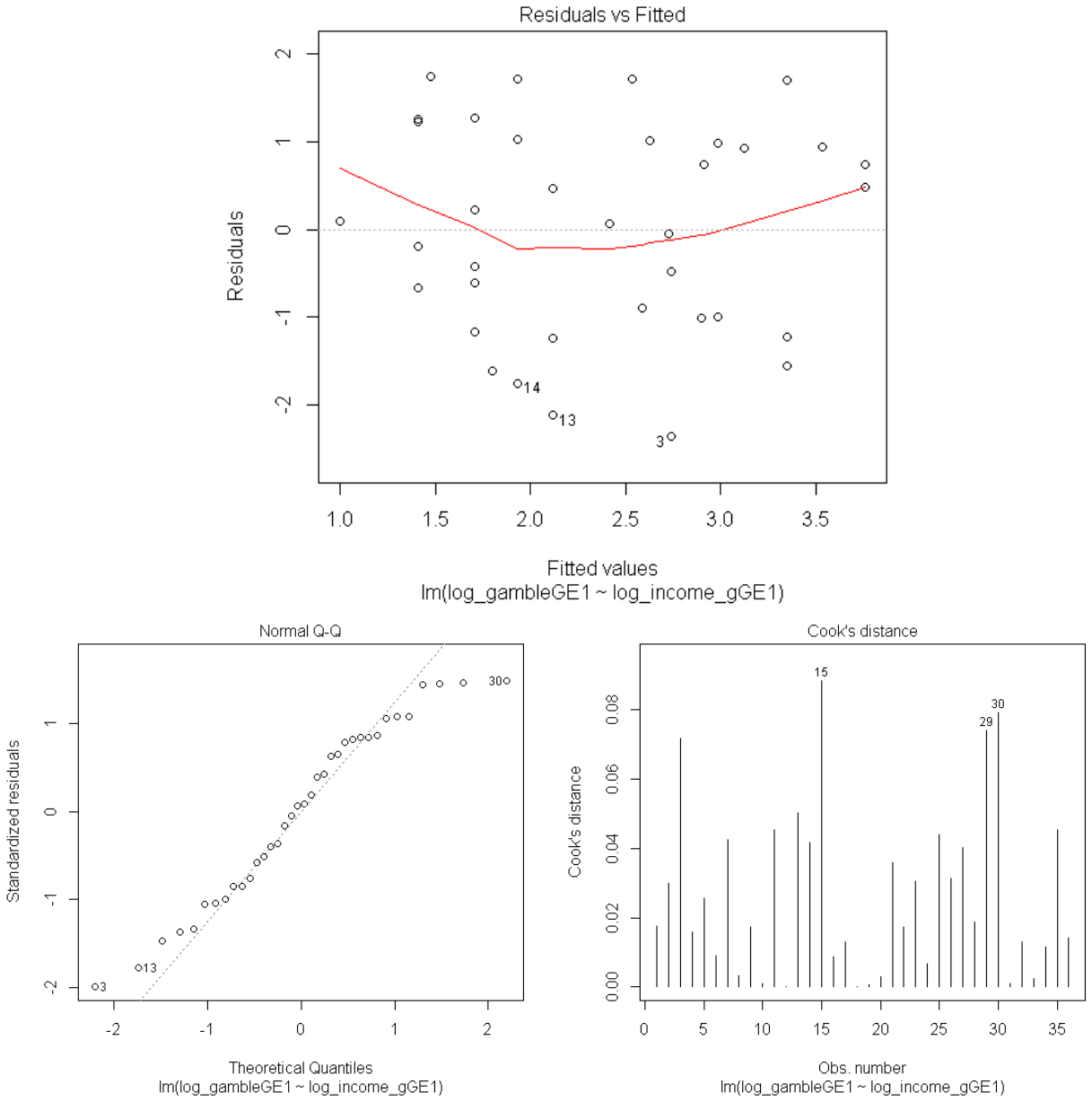
```
> gamble.lm

Call:
lm(formula = gamble ~ income)

Coefficients:
(Intercept)       income
     -6.325        5.520
```



The "Residuals vs Fitted" plots show definite problems with increasing variance (heteroscedasticity). The non-linearity is not that apparent so we can accept the linearity assumption (or you can say there is a possible non-linearity problem but not severe.) Also there is possibly a problem with an outlier (observation 24). These plots suggest we need to do something to fix the model, especially for the non-constant variance problem (heteroskedasticity). The "Normal Q-Q" plot looks ok except for a possible outlier (observation 24) and the "Cook's distance" plot shows that observation 24 would be a possible influential point and need further investigation, if we were happy with the other plots. Observation 24 is a male who spends £156 per year on gambling on a modest income of only £10 per week (no-one else spends more than £100).

(c) In question 1, a natural log (to the base *e*) transformation had already been applied to both the response and predictor variables and appeared to produce reasonable results. In this example, there are a number of teenagers who are not regular gamblers (their annual expenditure on gambling is very small or even zero). What is the problem with applying a log transformation in this situation? Exclude any teenager who spends less than £1 per year on gambling and fit another simple linear regression model with log(`gamble`) as the response variable and log(`income`) as the predictor. Check the same plots you produced for the earlier model in part (b). Are the same problems still apparent? **(4 marks)**

```
log_gambleGE1 <- log(gamble[gamble>=1])
log_income_gGE1 <- log(income[gamble>=1])
```



Now there are no obvious problems apart from the normal distribution having a slightly shorter than expected tail. We could experiment with a slightly weaker transformation, but this appears to be a far more appropriate model.

Question 2 continued

(d) Produce the ANOVA table and the table of the estimated coefficients for the revised SLR model in part (c). Interpret the values of the estimated coefficients for this SLR model and the results of the overall F test and the t-tests on the estimated coefficients.

(4 marks)

$\text{Model}: \texttt{log\_gambleGE1} = \beta_0 + \beta_1 \texttt{log\_income\_gGE1} + \varepsilon \quad \varepsilon \sim i.i.d.\ N\left(0, \sigma^2\right)$

```
> anova(gambleGE1.lm)
Analysis of Variance Table

Response: log_gambleGE1
                Df Sum Sq Mean Sq F value    Pr(>F)
log_income_gGE1  1 20.199  20.198  13.834 0.0007179 ***
Residuals       34 49.642   1.460
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$H_0: \frac{\sigma^2_{Model}}{\sigma^2_{Error}} = 1 \quad H_A: \frac{\sigma^2_{Model}}{\sigma^2_{Error}} > 1$$

$F_{1,34} = 13.834$, $p << 0.05$, so reject H_0 in favour of H_A and conclude the variance explained by the model is large compared to the error variance, which means that the model involving `log_income_gGE1` is explaining a significant proportion of the variability in `log_gambleGE1`.

```
> summary(gambleGE1.lm)

Call:
lm(formula = log_gambleGE1 ~ log_income_gGE1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.36611 -0.99926  0.08747  0.99493  1.74254

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.9962     0.4183   2.382 0.022968 *
log_income_gGE1   1.0215     0.2746   3.719 0.000718 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.208 on 34 degrees of freedom
Multiple R-squared:  0.2892,     Adjusted R-squared:  0.2683
F-statistic: 13.83 on 1 and 34 DF,  p-value: 0.0007179
```

$$H_0: \beta_1 = 0 \quad H_A: \beta_1 \neq 0$$

$t_{34} = 3.7$, $p << 0.05$, so reject H_0 in favour of H_A and conclude that the slope coefficient of `log_income_gGE1` is significantly different from 0, implying there is a significant linear relationship between `log_income_gGE1` and `log_gambleGE1`. Note this test is directly equivalent to the above F test.

$$H_0: \beta_0 = 0 \quad H_A: \beta_0 \neq 0$$

$t_{95} = 2.4$, $p < 0.05$, so reject H_0 in favour of H_A and conclude that the intercept is also significantly different from 0. On the log-log scale, 0 is around the minimum of our data: `log_income_gGE1` = 0 implies `income` = 1 and `log_gambleGE1` = 0.9962 implies `gamble` = 2.708 or yearly expenditure on gambling of £2.70 (rounding to the same accuracy as the data – which was to the nearest £0.10).

Question 2 continued

(e) Use the revised SLR model from part (c) to predict the annual expenditure on gambling for three British teenagers, who were not included in the original study, but who have weekly incomes of £1, £5 and £20, respectively. Find 95% prediction intervals for these predictions. Do you think this revised SLR model is a good model for making all three of these predictions? **(4 marks)**

```
> incomes <- c(1,5,20)
> log_incomes <- log(incomes)

> predictions <- predict(gambleGE1.lm,
newdata=data.frame(log_income_gGE1=log_incomes), interval="prediction")

> exp(predictions)
        fit       lwr       upr
1  2.708015 0.2014226  36.40776
2 14.017508 1.1573202 169.78060
3 57.768871 4.0549933 822.99581
```

As with the model in question 1, we again have a statistically significant linear relationship between `log_gambleGE1` and `log_income_gGE1`, but the model is not really a good predictive model (note the very wide prediction intervals). The prediction for a weekly income of £20 is extrapolating outside the range of the original data, in which the largest income was only £15.

This is NOT a required part of the assignment, but the above discussion can be reinforced by producing a plot of the data, model and selected predictions on the original scale: