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

REGRESSION MODELLING

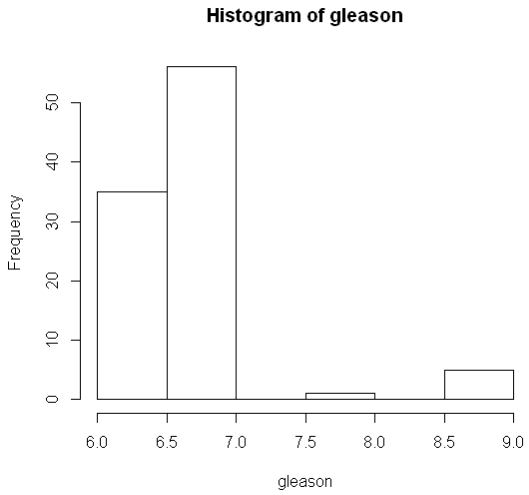
(STAT2008/STAT6038)

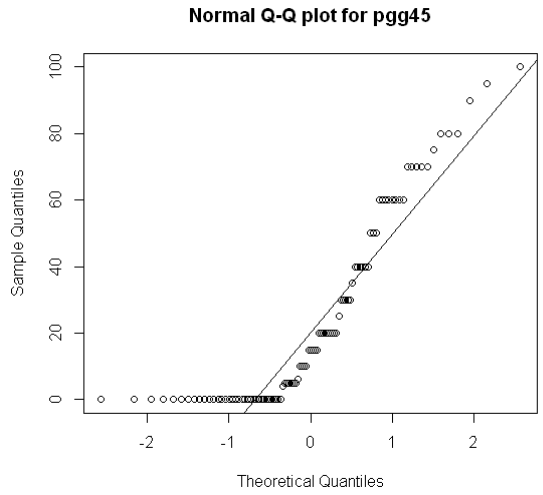
Solutions to Assignment 2 for 2014

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). In this assignment we are going to fit an appropriate multiple linear regression model to examine factors affecting `lcavol` (log of the cancer volume), which is a measure of the size of the cancer tumour (measured in ml).

(a) All of the other variables in the `prostate` dataset could potentially be included as predictors (explanatory variables) in a multiple regression model with `lcavol` as the response variable. Produce suitable plots and/or summary R output to investigate the contents of the variables `svi`, `gleason` and `pgg45`. How are these variables distributed? Discuss any potential problems with including these variables in a multiple regression model. **(3 marks)**






```
> table(svi)
svi
 0  1
76 21
> table(gleason)
gleason
 6  7  8  9
35 56  1  5
> table(pgg45)
pgg45
  0   4   5   6  10  15  20  25  30  35  40  50  60  70  75  80  90  95 100
 35   1   6   1   4   5   9   1   5   1   6   3   8   5   1   3   1   1   1
```

All three of these variables are definitely not normally distributed:

`svi` is an indicator variable which only takes on the values 0 and 1 (similar to a Bernoulli trial);

`gleason` is a categorical score, which a little internet research will reveal can only take on discrete values in the range 2 to 10, but in these data only takes on the values 6, 7, 8 or 9; and

`pgg45` is a percentage and as such cannot be less than 0 or more than 100 and in these data there are values close to both bounds (which is not an ideal situation for a normal approximation to a binomial distribution).

Question 1(a) continued

Fortunately we do not make strong normality assumptions about the explanatory variables in a regression model, rather we assume them "known", at least relative to the response variable. There are approaches for including all of these variables in regression models:

`svi` turns out to be not be a significant addition in this example to a model that already includes both `lpsa` and `lcp`, but we will consider a different example of an indicator variable in question 2;

`gleason` is measured on a scale, so could potentially be included as a continuous explanatory variable, but given it only takes on 4 unique values, might be better coded as a categorical factor variable (the indicator variable approach can be extended to variables with more than two categories, but that is outside the scope of this course); and

`pgg45` could also potentially be included as a continuous explanatory variable (which like `svi` and `gleason` turns out to be not be a significant addition to my preferred model), but might also possibly benefit from some scale transformation.

(b) Find an appropriate multiple linear regression model with `lcavol` as the response variable and `lweight`, `age`, `lbph`, `lcp` and `lpsa` as possible predictors. To simplify this exercise, exclude the variables mentioned in part (a) from consideration, assume that all the other variables are already measured on an appropriate scale (i.e. no further transformations are necessary), that an additive model is appropriate (i.e. no interaction terms or quadratic/higher order terms are needed), but do NOT exclude any potential outliers. Do NOT present output for multiple models, choose just ONE model! Produce the ANOVA (Analysis of Variance) table for your chosen model and summary output showing the estimated coefficients and use these to justify your choice of model. Why have you included the explanatory variables that are included in your model and why have you chosen to exclude other possible predictors? **(4 marks)**

```
> prostate.lm <- lm(lcavol ~ lpsa + lcp)
> anova(prostate.lm)
Analysis of Variance Table

Response: lcavol
          Df Sum Sq Mean Sq F value    Pr(>F)
lpsa       1 71.938  71.938 143.030 < 2.2e-16 ***
lcp        1 14.143  14.143  28.119 7.538e-07 ***
Residuals 94 47.278   0.503
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> summary(prostate.lm)

Call:
lm(formula = lcavol ~ lpsa + lcp)

Residuals:
     Min       1Q   Median       3Q      Max
-1.65744 -0.54398 -0.05502  0.57163  2.07959

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09135    0.20527   0.445    0.657
lpsa         0.53162    0.07501   7.087 2.49e-10 ***
lcp          0.32838    0.06193   5.303 7.54e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7092 on 94 degrees of freedom
Multiple R-squared:  0.6455,      Adjusted R-squared:  0.6379
F-statistic: 85.57 on 2 and 94 DF,  p-value: < 2.2e-16
```

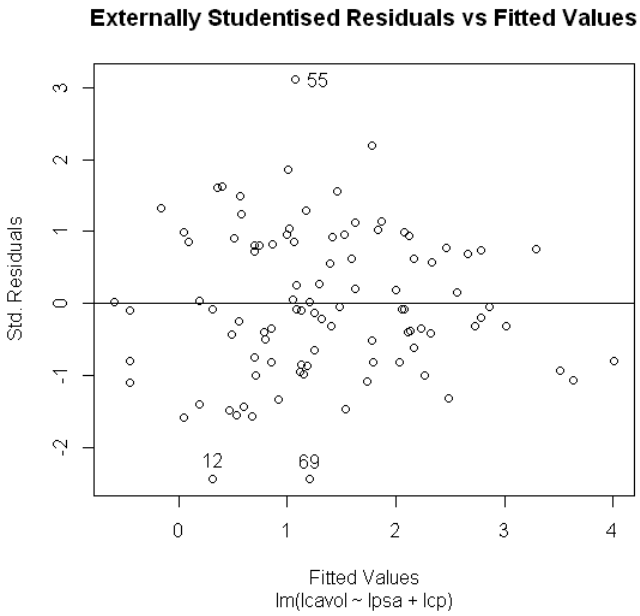
Question 1(b) continued

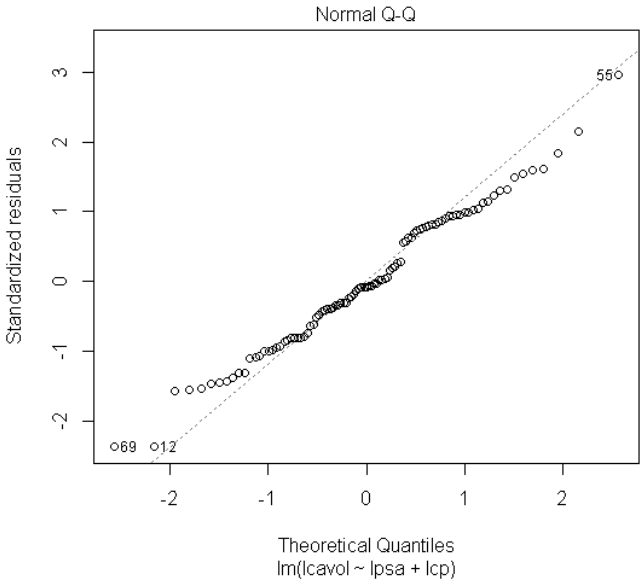
```
> vif(prostate.lm)
    lpsa      lcp
1.431016 1.431016
```

Trying all the potential explanatory variables in different orders, only `lpsa` and `lcp` turn out to be consistently significant in the ANOVA table and it does not matter which of these two variables we include first in the model. As can be seen from the above output, both of these variables explain significant proportions of the overall variance and have coefficients which are significantly different from 0 (at the 0.05 level).

As we will see in part (d) of this question, none of the other possible predictors are significant additions to this model.

(c) For your chosen multiple regression model, construct a plot of the externally studentised residuals against the fitted values and a normal Q-Q plot of the internally standardised residuals and use these plots to comment on the model assumptions.
Also produce selected statistics and/or a plot to investigate and discuss possible outliers and influential observations. Do NOT try to present a table of various statistics showing all 97 observations (though you could select just one statistic and present a relevant plot which shows all 97 observations). **(5 marks)**

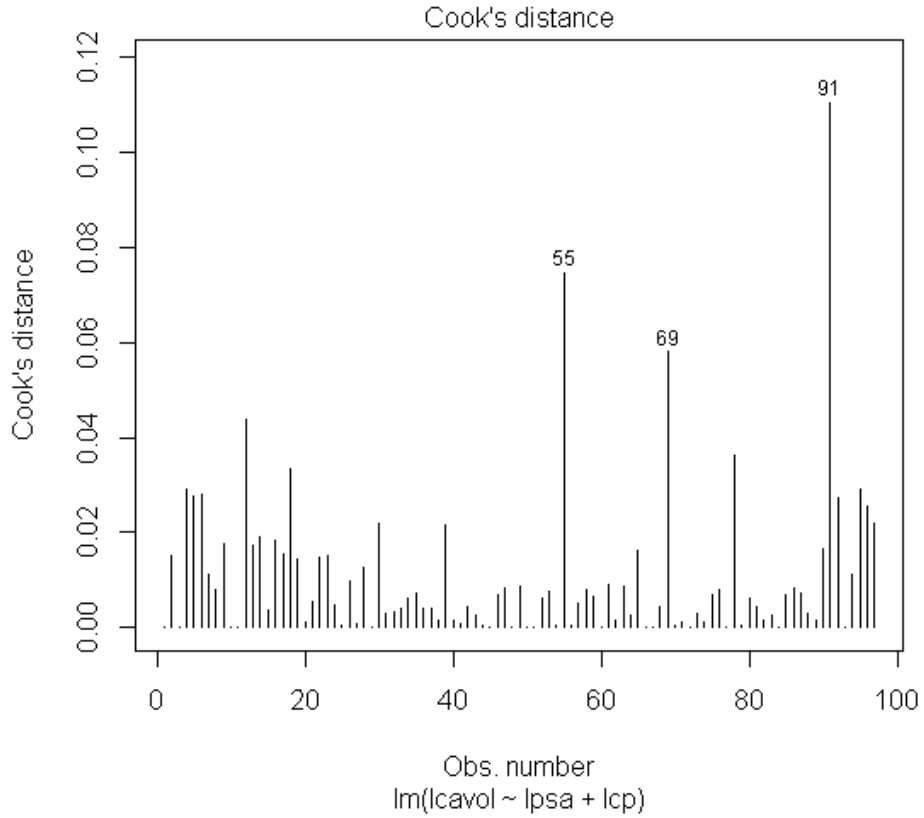




From the "Residuals vs Fitted" plot, there are no apparent problems with the assumptions of either independence or constant variance, but there are some potential problems with individual observations, which I have identified on the plot. Observations 12 and 69 have slightly outlying negative standardised residuals, but are only just less than –2; however, observation 55 appears more likely to be a real vertical outlier with a standardised residual of more than +3.

From the "Normal Quantile" plot, there are no apparent problems with the assumption of normally distributed residuals, but the same potential problem observations have been identified (for this plot, I used the default plot, rather than creating a customised one). A good way to further investigate problems with potential problem is to look at the default bar plot of Cook's distances:

Question 1(c) continued



The observation with the largest Cook's distance turns out not to be one of the earlier identified potential problem observations. Observation 91 turns out to be the one with the second highest positive externally studentised residual and therefore appears as the unidentified point second from the top on the earlier "Residuals vs Fitted" plot.

However, observation 91 has only fourth highest standardised residual in absolute value and has only the seventh highest leverage value:

```
> sort(cooks.distance(prostate.lm), decreasing=TRUE)[1:7]
        91         55         69         12         78         18         95
0.11054045 0.07460762 0.05797709 0.04385529 0.03630946 0.03346528 0.02911686
> sort(abs(rstudent(prostate.lm)), decreasing=TRUE)[1:7]
      55       69       12       91       18       16       14
3.102554 2.434265 2.423942 2.186070 1.857459 1.632679 1.616596
> sort(hatvalues(prostate.lm), decreasing=TRUE)[1:7]
        97         96          1         92         95         47         91
0.09358372 0.08112663 0.08050114 0.07211666 0.07161414 0.06912912 0.06732295
>
> dfbetas(prostate.lm)[c(91,55,69,12),]
   (Intercept)       lpsa         lcp
91  -0.3982803  0.5023575 -0.44240094
55  -0.1409238  0.2569120 -0.37236383
69   0.1739459 -0.2695535  0.32995189
12  -0.2388168  0.1723856  0.08599187
```

The standardised deletion coefficients (DFBETAS) shown above suggest that none of the identified potential problem observations have been very influential in the fit of the model, and the other influence measures (not shown, but included in the appendix of R code) tell a similar story. If we are only intending to use this model as an explanatory model to investigate the relationships between the variables (rather than a more precise predictive model), then it is not really important to do anything to treat these potential problems.

Question 1 continued

(d) Perform a "nested model" F test to see whether or not any of the subset of possible predictors you have excluded from your chosen model would be a significant addition to your chosen model. If your chosen model includes 4 or 5 of the possible predictors (`lweight`, `age`, `lbph`, `lcp` and `lpsa`) then perform a test of the last two or three predictors as an addition to a model that already contains the other variables. Review the above test results and the output in parts (b), (c) and (d) and also compare your chosen model with the simple linear regression model shown in the model solutions to Question 1 of Assignment 1. Is your chosen a model an improvement in terms of reliably predicting the size of a prostate cancer tumour? **(5 marks)**

```
> prostate.lmplus <- lm(lcavol ~ lpsa + lcp + cbind(lweight, age, lbph))
> anova(prostate.lmplus)
Analysis of Variance Table

Response: lcavol
                           Df Sum Sq Mean Sq  F value    Pr(>F)
lpsa                        1 71.938  71.938 146.1643 < 2.2e-16 ***
lcp                         1 14.143  14.143  28.7355 6.232e-07 ***
cbind(lweight, age, lbph)   3  2.490   0.830   1.6866    0.1754
Residuals                  91 44.788   0.492
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\text{Model}: \texttt{lcavol} = \beta_0 + \beta_1 \texttt{lpsa} + \beta_2 \texttt{lcp} + \beta_j \texttt{X}_j + \varepsilon \quad \varepsilon \sim i.i.d.\ N\left(0, \sigma^2\right)$$

$$\texttt{X}_j = \left[\texttt{lweight} \quad \texttt{age} \quad \texttt{lbph}\right],\ j = 3,4,5$$

$$\text{H}_0 : \frac{\sigma^2_{\texttt{X}_j}}{\sigma^2_{Error}} = 1 \quad \text{H}_\text{A} : \frac{\sigma^2_{\texttt{X}_j}}{\sigma^2_{Error}} > 1 \quad OR \quad \text{H}_0 : all\ \beta_j = 0 \quad \text{H}_\text{A} : at\ least\ one\ \beta_j \neq 0$$

$F_{3, 91} = 1.6866$, $p > 0.05$, so do NOT reject H_0 in favour of H_A and conclude that the additional terms in the model do not significantly increase the proportion of the variance explained by the model and so are not significant additions to the model. Note you can do a similar test for all the excluded variables (`svi`, `gleason` and `pgg45`, as well as `lweight`, `age` and `lbph`) and get the same results.

The model in Assignment 1 was the simple linear regression (SLR) of `lcavol` on `lpsa`, and the results of part (b) clearly show that `lcp` is a significant addition to this model, which is why I chose the expanded multiple regression model fitted in part (b).

However, even with this expanded model there still remains a lot of unexplained variation: the R^2 for the SLR model in Assignment 1 was only 54%, whilst the R^2 for the expanded model has increased significantly, but only marginally to 65%. As argued in Assignment 1, there would be problems in using even this expanded model for making precise predictions.

Question 1 continued

(e) Add an interaction term between `lpsa` and `lcp` to your chosen model (if your chosen model does not already include linear terms in `lpsa` and `lcp`, also add those terms to the model). Is this term a significant addition to the model? Interpret the coefficients of the terms involving both `lpsa` and `lcp` (and their interaction) in this expanded model and in your chosen model. **(3 marks)**

```
> prostate.lmint <- lm(lcavol ~ lpsa * lcp)
> anova(prostate.lmint)
Analysis of Variance Table

Response: lcavol
          Df Sum Sq Mean Sq  F value    Pr(>F)
lpsa       1 71.938  71.938 149.7905 < 2.2e-16 ***
lcp        1 14.143  14.143  29.4484 4.542e-07 ***
lpsa:lcp   1  2.614   2.614   5.4431    0.0218 *
Residuals 93 44.664   0.480
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> summary(prostate.lmint)

Call:
lm(formula = lcavol ~ lpsa * lcp)

Residuals:
     Min       1Q   Median       3Q      Max
-1.73890 -0.47743 -0.04935  0.53685  2.02766

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.26030    0.21326   1.221   0.2253
lpsa         0.50364    0.07427   6.781 1.08e-09 ***
lcp          0.63612    0.14513   4.383 3.07e-05 ***
lpsa:lcp    -0.10276    0.04404  -2.333   0.0218 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.693 on 93 degrees of freedom
Multiple R-squared:  0.6651,      Adjusted R-squared:  0.6543
F-statistic: 61.56 on 3 and 93 DF,  p-value: < 2.2e-16

> vif(prostate.lmint)
    lpsa      lcp lpsa:lcp
1.469335 8.230951 7.277963
```

Adopting the approach to hypothesis testing based on the ANOVA table used in part (d), the interaction term is a significant addition to the model at the 0.05 level ($F_{1, 93} = 5.4431$, $p < 0.05$).

```
> coef(prostate.lm)
(Intercept)        lpsa         lcp
 0.09134534  0.53162111  0.32837535
> coef(prostate.lmint)
(Intercept)        lpsa         lcp    lpsa:lcp
  0.2603019   0.5036378   0.6361206  -0.1027554
```

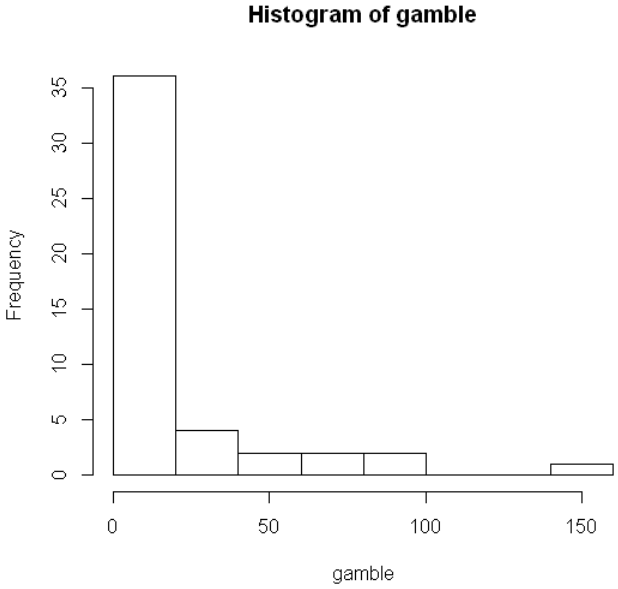
The coefficients of the main effects (linear) terms for both variables are positive in both models, suggesting that `lcavol` tends to increase as both `lpsa` and `lcp` increase, however, these two variables are also positively correlated ($r = 0.54$,with a correlation test included in the R appendix).

The total expected increase in the response variable is not as large as just adding together the increases that would result if we were to increase the two variables the two explanatory variables separately (which we can’t really do, as they are inter-related) and so the interaction term is negative.

Question 2 **(20 marks)**

The dataset `teengamb` concerns a study of teenage gambling in Britain. In this assignment we are going to fit an appropriate multiple linear regression model to examine factors affecting the amount that teenagers will `gamble` (gambling expenditure measured in UK £ per year), including both teenagers who do and who do not regularly gamble.

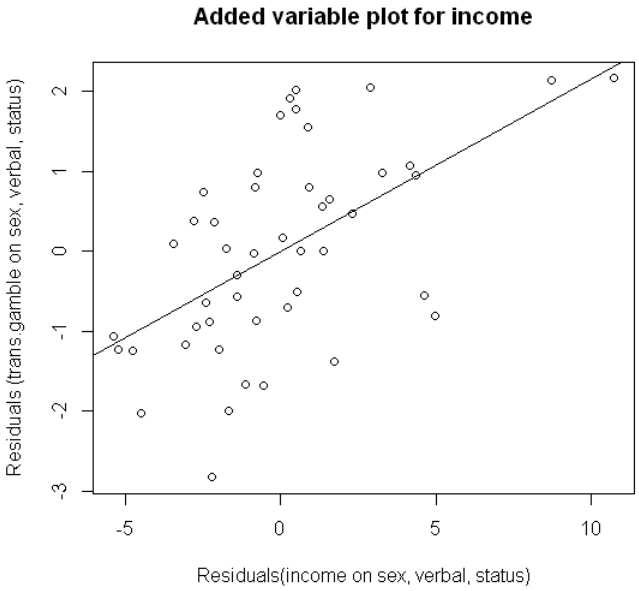
(a) Transform `gamble` by creating a new variable `trans.gamble <- log(gamble + 1)`. Compare histograms of `gamble` and `trans.gamble` and comment on which is more likely to be suitable for inclusion in a multiple regression model.
Assume that the researchers who collected the data believe that gambling expenditure differs by `sex` and is also strongly affected by factors such as education and socio-economic status. This is why they collected the variables `verbal` and `status` (as measures of education and socio-economic status respectively) and any multiple regression model will include `status`, `verbal` and `sex` as predictors so we can test these assertions (and control for the effects of these factors).
This leaves `income` as the only remaining observed variable (covariate). Construct an added variable plot to assess `income` as a possible addition to a multiple regression model for `trans.gamble` that already includes `sex`, `verbal` and `status` as predictors. Does this added variable plot suggest a transformation is required for income? The transformation we used in Question 2 of Assignment 1 was log(`income`). Construct a different added variable plot for log(`income`). Is this an improvement? **(5 marks)**

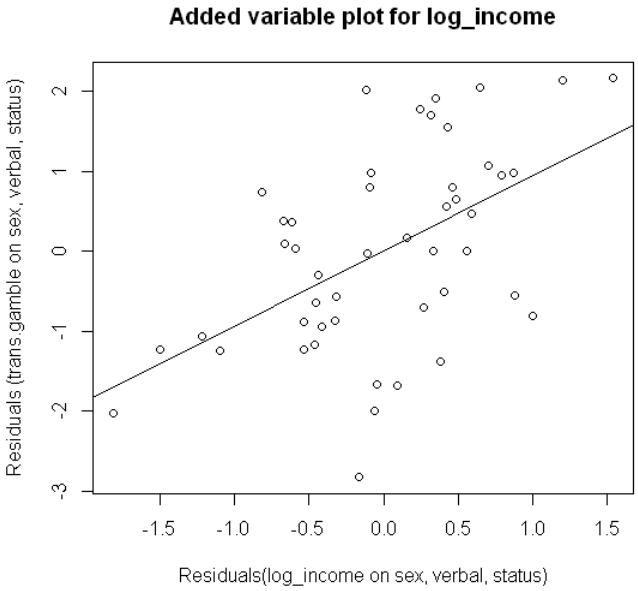


In linear regression, we are assuming that the residuals from the regression model are normally distributed. As we also assume that the explanatory variables are known then the residuals are equivalent to an adjusted (mean-corrected) version of the original response variable, as the residuals are equal to the response variable minus the fitted values (which are a linear combination of the known explanatory variables).

So the assumption of normally distributed residuals is more likely, if the response variable is also normally distributed. Judging by the first histogram, `gamble` is definitely not normally distributed. The second histogram is also not normally distributed, but is closer than the first one and a similar transformation worked well in the first assignment.

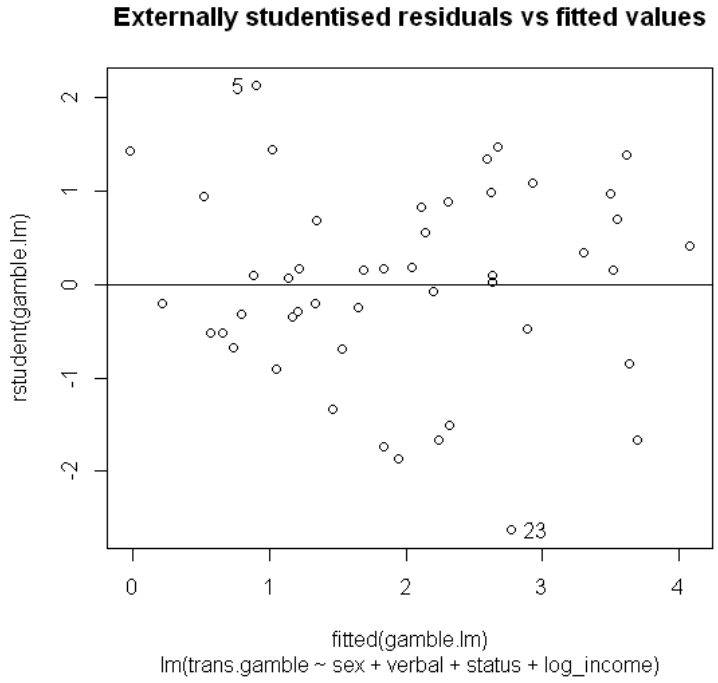
Question 2(a) continued



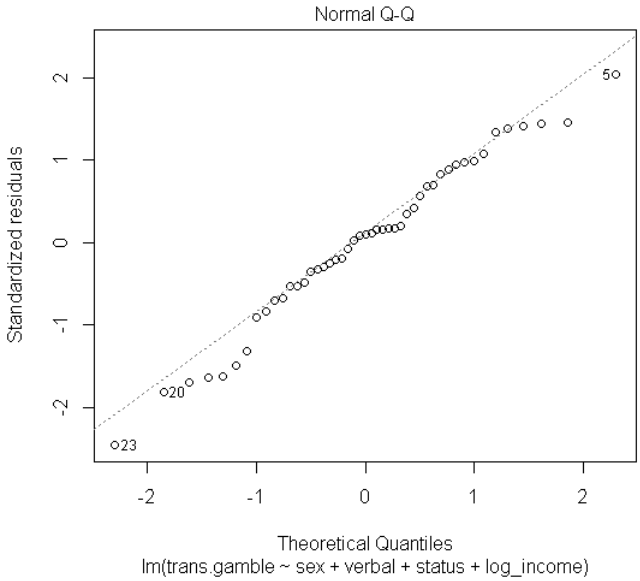


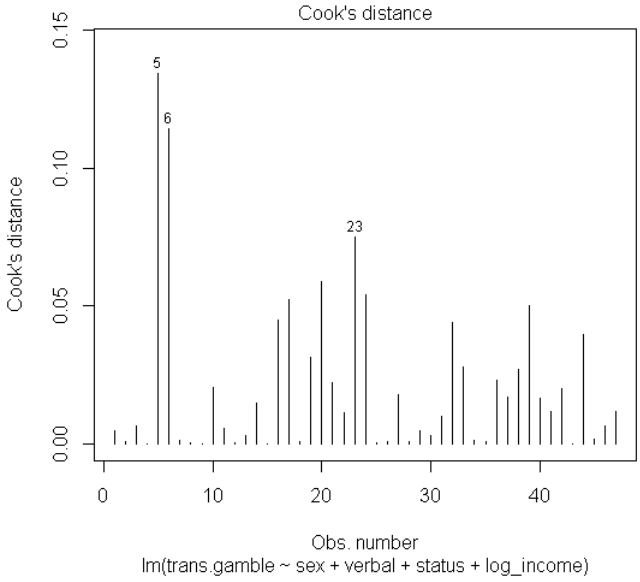
There is some indication of curvature in the first plot, though it may just be a suggestion caused by the two points in the upper right of the plot. However, the second plot doesn't have either of these apparent problems and is definitely an improvement, so it does seem a good idea to apply the suggested log transformation.

(b) Fit the multiple linear regression model with `trans.gamble` as the response variable and `sex`, `verbal`, `status` and log(`income`) as predictors. Construct a plot of the externally studentised residuals against the fitted values, a normal Q-Q plot of the internally studentised residuals and a bar plot of Cook's Distances for each observation. Comment on the model assumptions and on any unusual data points. Calculate appropriate influence statistics for the most unusual data point and comment on these statistics, but do NOT refine the model by removing this observation as a possible outlier. **(5 marks)**



Question 2(b) continued





The "Residuals vs Fitted" plot has an interesting absence of points in the bottom left hand corner, but there are no real signs of problems with the assumptions of independence or non-constant variance. The two observations (5 and 23) with the largest residuals values (both positive and negative) have been identified on the plot, but are only just outside ±2, so are probably not an issue. Similarly, the "Normal Q-Q" plot doesn't show any real problem with the assumption of normality.

The "Cook's distance" plot highlights both of the earlier identified observations (and also observation 6), but they do not seem extreme relative to the other observations. The identified observations are not the observations with the highest leverage and the standardised deletion coefficients (DFBETAS) suggest that none of the identified observations have been overly influential in the fit of the model:

```
> sort(hatvalues(gamble.lm), decreasing=TRUE)[1:7]
       35        31        29        42         6        28         5
0.3192015 0.2263684 0.1837170 0.1727774 0.1644503 0.1547624 0.1387434
>
> dfbetas(gamble.lm)[c(5,6,23),]
   (Intercept)        sex      verbal      status  log_income
5   -0.3709880  0.6594871 -0.09117119  0.56779080 -0.05565517
6    0.2419247 -0.6119417  0.39412153 -0.63439243 -0.15522729
23  -0.4701669  0.3398218  0.31450812  0.04866075  0.09421694
```

Question 2 continued

(c) Produce the ANOVA table and the summary table of estimated coefficients for the multiple linear regression model in part (b). Interpret the overall and sequential F tests and the t-tests and the values of the estimated coefficients of the model. Are the earlier assertions in part (a) about sex, education and socio-economic status supported in the context of this model? **(5 marks)**

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

Response: trans.gamble
           Df Sum Sq Mean Sq F value    Pr(>F)
sex         1 23.600 23.6005 18.8683 8.675e-05 ***
verbal      1  6.609  6.6086  5.2835 0.0265733 *
status      1  0.735  0.7349  0.5876 0.4476496
log_income  1 19.692 19.6921 15.7436 0.0002775 ***
Residuals  42 52.534  1.2508
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(gamble.lm)

Call:
lm(formula = trans.gamble ~ sex + verbal + status + log_income)

Residuals:
     Min       1Q   Median       3Q      Max
-2.67231 -0.56225  0.08561  0.80866  2.11944

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.82343    0.85258   2.139 0.038318 *
sex         -1.10598    0.39296  -2.814 0.007404 **
verbal      -0.25520    0.10704  -2.384 0.021713 *
status       0.02430    0.01354   1.795 0.079887 .
log_income   0.93798    0.23640   3.968 0.000278 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.118 on 42 degrees of freedom
Multiple R-squared:  0.4908,      Adjusted R-squared:  0.4423
F-statistic: 10.12 on 4 and 42 DF,  p-value: 7.906e-06

> vif(gamble.lm)
       sex     verbal     status log_income
  1.397388   1.452338   2.009824   1.105423
```

All of the terms in the model appear significant, with the exception of `status`, which the sequential F test in the ANOVA table suggests is not a significant addition to a model already containing `sex` and `verbal`. So we have evidence for two of the assertions in part (a), but not for `status`.

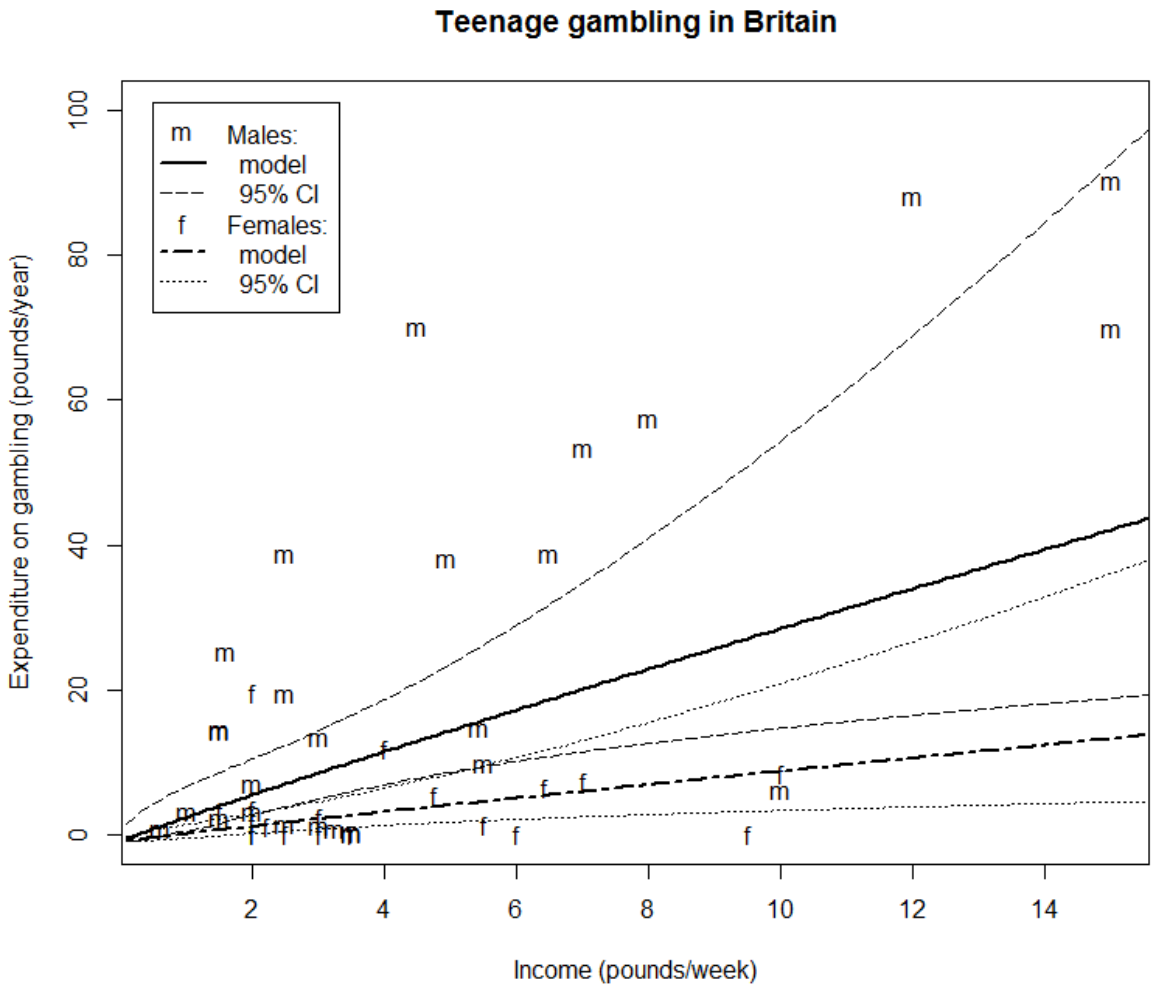
`log_income` is definitely a significant addition to the model and as in Assignment 1, the positive coefficients suggests that as income increases, so does expenditure on gambling (this is true in the model and also on the back-transformed scale, as the transformations we have used preserve the data in the same order, i.e. they are monotonically increasing).

The coefficients of `sex` and `verbal` are both negative, suggesting that gambling expenditure is significantly lower amongst females (`sex` = 1) than amongst males (`sex` = 0) and that higher levels of education (as measured by `verbal`) lead to significantly lower expenditure on gambling.

The coefficient of `status` is small and positive, suggesting a slight, but non-significant increase in gambling expenditure as socio-economic status increases.

Question 2 continued

(d) To help the researchers interpret the model, plot `gamble` against `income`, with different plotting symbols for the two values of `sex`. Include your model on this plot by calculating predicted values for `trans.gamble`, for the full range of `income` values and for both values of `sex`, holding `verbal` and `status` at their mean values. Suitably back-transform the predictions and include them on the plot separately for both males and females. Also include point-wise 95% confidence intervals (but not 95% prediction intervals) on the plot. **(5 marks)**



Plot indicates that the model does appear to have captured the general trend of the data, but also makes it obvious that whilst this model might be acceptable for the sort of exploratory analysis discussed in part (c); we should have concerns about using the model as a predictive one.