STA 5207 Assignment 4

Due Friday October 8

The data set gamble (in the faraway package) has 47 observations and 5 variables. The columns do not have headings. A survey was conducted to study teenage gambling in Britain. The data set is saved as gamble.txt in SAS Studio.

The variables, in order, are:

Sex: 0=male, 1=female

Status: Socioeconomic status score based on parents' occupation

Income: in pounds per week

Verbal: verbal score in words out of 12 correctly defined

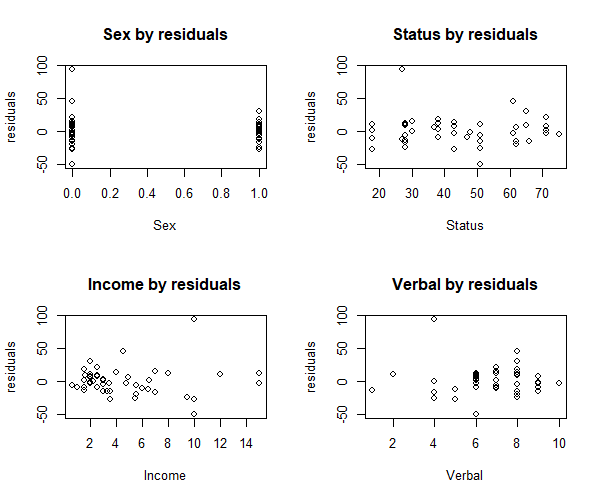
Gamble: expenditure on gambling in pounds per year

Gamble is the response variable, and the others are the predictors.

Some of these questions are subjective, so there may not be a “right” answer. Just make sure that your answer and explanation agree.

Turn in the R output specifically requested in each question. You may turn in more, but the tables and plots asked for are sufficient.

Each question is 10 points.

1. Create a plot of residuals by predicted values. Based on this plot, does it appear that the assumption of constant error variance reasonably satisfied? Why or why not? Does a linear regression model seem to be appropriate? Why or why not? Submit the plot from R output.
   1. It does not appear that the assumption is satisfied because we do not see a random variance around zero. Therefore, a linear regression model does not seem to be appropriate since our usual assumptions are violated.
2. Perform regression using the following model:

Give the hypotheses, test statistic, and p-value. What do the results of this test tell you about the error variance? Does this agree with your answer to Question 1?

* 1. Hypotheses: H0: = 0 vs H1: ≠ 0
  2. F-statistic: 7.466 on 1 and 45 DF, p-value: 0.008952
  3. Reject H0. Variance is not constant in agreeance with my answer from Q1

1. Give the p-values and conclusions from the normality tests of the residuals. Create a histogram, boxplot, and normal probability plot of the residuals. Based on the tests and plots, does it appear that the assumption of normally distributed errors is reasonably satisfied? Why or why not? Submit the three plots from R. You do not have to turn in the output from the tests.

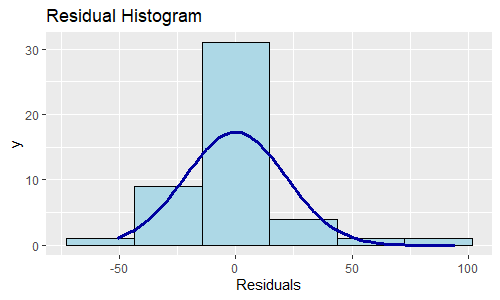
|  |  |  |
| --- | --- | --- |
| **Test** | **Statistic** | **pvalue** |
| **Shapiro-Wilk** | **0.8684** | **1e-04** |
| **Kolmogorov-Smirnov** | **0.1451** | **0.2507** |
| **Cramer-von Mises** | **3.6713** | **0.0000** |
| **Anderson-Darling** | **1.2829** | **0.0022** |

Reject H0

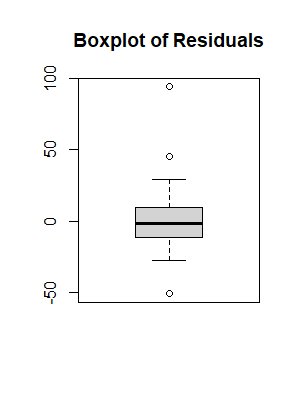
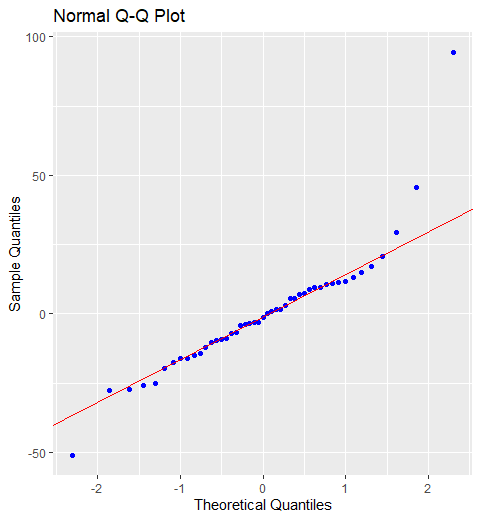
Do not reject H0

Reject H0

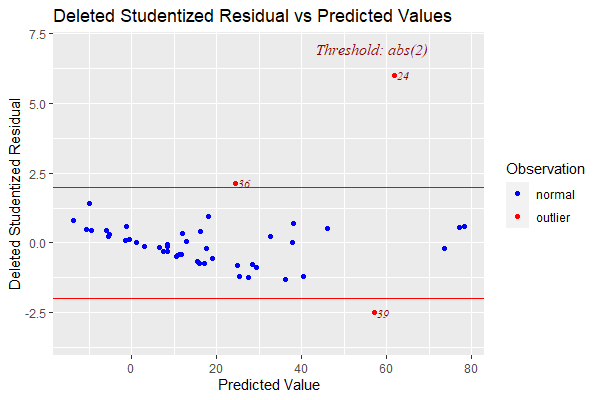
Reject H0

****

* 1. The assumption of normally distributed errors is not reasonably satisfied. The normal probability plot has long tails with a large departure from the fitted line, so a different distribution may provide a better fit. Additionally, only the p-value for the KS normality test is large, indicating we reject the hypothesis of normality.

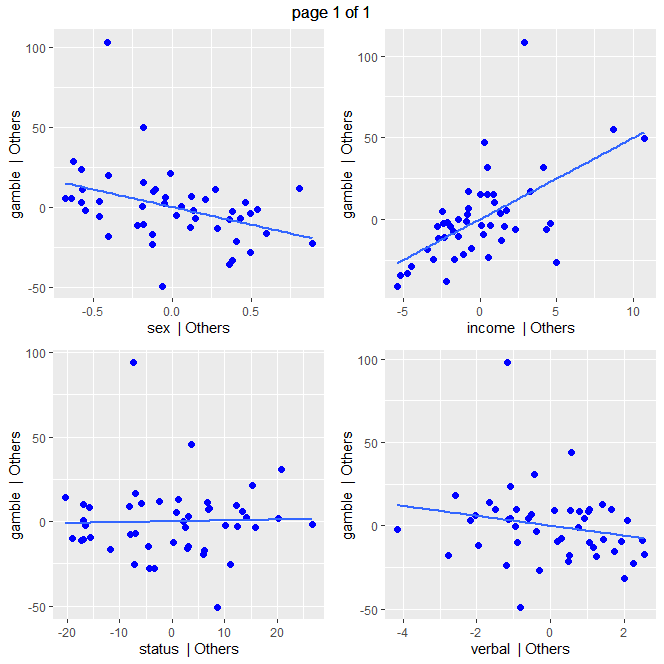
 ****

1. Create a plot of studentized residuals by fitted values to determine which, if any, observations are outliers in terms of the response. Submit the plot.



* 1. We can see that observations 24, 36, and 39 are outliers in terms of the response

1. Which, if any, observations have undue influences on one fitted value? No output needed.
   1. There are 2 observations that have undue influence on one fitted value: 24 and 39
2. Which, if any, observations are outliers in terms of the predictors? No output needed.
   1. We can see that observations 31, 33, 35, 42 are outliers in terms of the predictors.
3. Which, if any, observations have undue influence on all fitted values? No output needed.
   1. There are 2 observations that have undue influence on one fitted value: 24 and 39
4. Do any observations have undue influence on coefficient estimates? If so, which observations and which estimates? No output needed.
   1. Observations 5 and 24 have undue influence on the estimate of sex
   2. Observations 5 and 24 have undue influence on the estimate of status
   3. Observations 24 and 39 have undue influence on the estimate of income
   4. Observations 24, 27, and 35 have undue influence on the estimate of verbal
5. Obtain the partial regression plots. For each of the predictors, determine if it appears to have a linear relationship with the response after removing the effects of the other predictors. Turn in the partial regression plots.



* 1. The relationship between sex and gamble appears to be linear after adjusting for the other predictors.
  2. There does not appear to be a relationship between status and gamble after adjusting for the other predictors.
  3. The relationship between income and gamble appears to be linear after adjusting for the other predictors.
  4. There does not appear to be a relationship between verbal and gamble after adjusting for the other predictors.

1. Obtain the t-tests and determine which predictors are significant in the model. Does this match your answer to question 9? No output needed. Just write down the p-values of the tests.
   1. Coefficients:

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

(Intercept) 22.55565 17.19680 1.312 0.1968

sex -22.11833 8.21111 -2.694 0.0101

status 0.05223 0.28111 0.186 0.8535

income 4.96198 1.02539 4.839 1.79e-05

verbal -2.95949 2.17215 -1.362 0.1803

For an assumed 5% significance level, only sex and income are significant in the model. This does match my answer to Q9.