1.

h) I)

The residuals of the model:

5.7799025 2.7542508 7.5352349 0.2211861 4.1598581 -4.7693149 5.0913501

5.7798776 -5.2108344 -7.1044882 -0.1435089 -5.7125891 -3.2637759 9.5596010

-5.0920929 -2.4950478 4.8176434 -7.6871192 -2.4763595 -5.9227218 -1.9773457

2.0566616 -3.8529424 -2.6122556 3.4157971 -3.4296492 -3.1863599 3.6791308

-3.6947257 2.9871647 -4.4117913 2.9276166 -4.9451392 3.5326971 -2.9330061

4.4062412 -1.1967962 7.3897875 -1.9310299 7.9548926

The standardized residuals of the model:

1.20058166 0.57210360 1.56519332 0.04594403 0.86407156 -0.99066583

1.05755791 1.20057649 -1.08237677 -1.47572009 -0.02980918 -1.18659955 -

0.67794041 1.98568775

-1.05771218 -0.51826283 1.00070447 -1.59674222 -0.51438096 -1.23024759 -

0.41072751 0.42720273 -0.80032008 -0.54260883 0.70951774 -0.71239505 -

0.66185982 0.76421652

-0.76745584 0.62048367 -0.91640227 0.60811456 -1.02718748 0.73379981 -

0.60923403 0.91524941 -0.24859443 1.53498148 -0.40110695 1.65236320

The fitted values of the model:

47.33510 37.82475 43.13377 48.26581 46.49614 47.36131 49.15065 54.39412

49.35383 57.81549 28.93051 35.09159 23.29378 49.90440 51.55609 52.99805

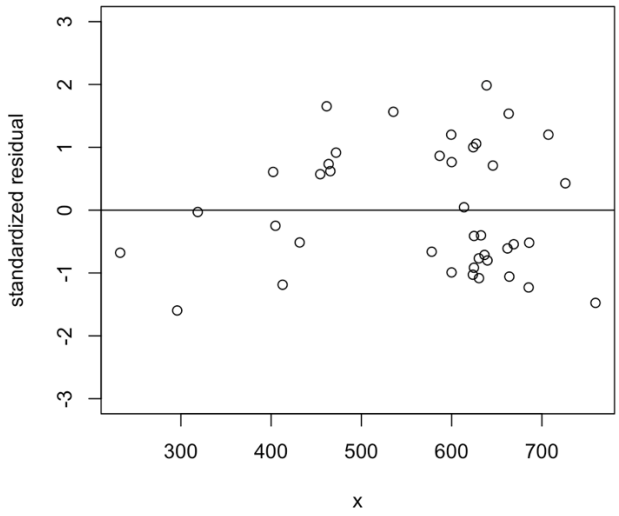
48.93436 27.43612 36.33036 52.95872 48.99335 55.62634 49.96994 51.87726

50.36320 49.75365 45.91936 47.36787 49.34073 38.55884 48.98679 34.40338

48.90814 38.43430 51.42501 38.97176

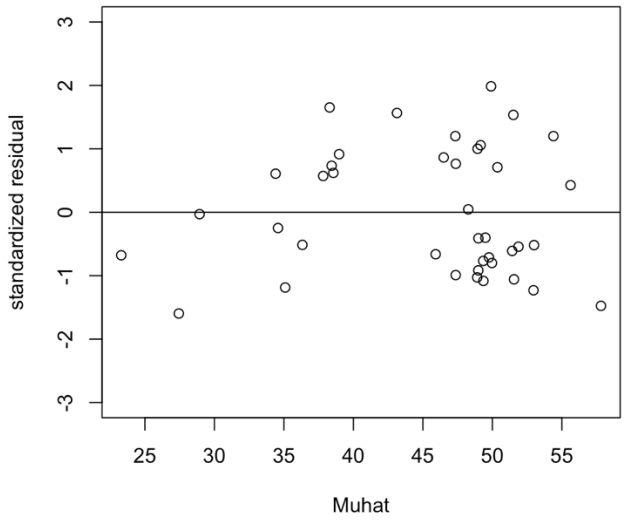
34.57380 51.51021 49.49803 38.29011

II)



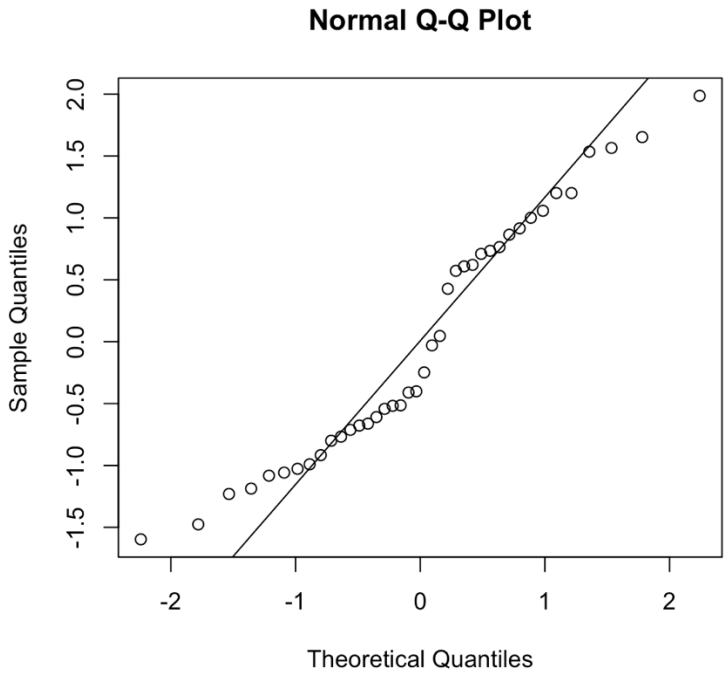
This plot is testing the assumption that whether the response variate, which is the fuel consumption, can be modeled by a Gaussian random variable whose mean is a linear function of the explanatory variate, which is the distance driven, and whose standard deviation is constant over the range of values of the explanatory variate. From this plot, I think they are satisfied because the points lie roughly within a horizontal band of constant width between -3 and 3 and approximately half the points lie on either side of the horizontal line at 0.

III)



Similarly, the plot is testing the assumption that whether the response variate, which is the fuel consumption, can be modeled by a Gaussian random variable whose mean is a linear function of the explanatory variate, which is the distance driven, and whose standard deviation is constant over the range of values of the explanatory variate. According to this plot, they are satisfied because the points lie roughly within a horizontal band of constant width between -3 and 3 and approximately half the points lie on either side of the horizontal line at 0.

IV)



Similarly, this plot is testing the assumption that whether the response variate, which is the fuel consumption, can be modeled by a Gaussian random variable whose mean is a linear function of the explanatory variate, which is the distance driven, and whose standard deviation is constant over the range of values of the explanatory variate. Based on this plot, they are satisfied. This is because the points in this plot lie roughly along a straight line. Also, there is more variability in the points at both ends of the line.

V)

Interval	Observed	Expected
(-∞,19)	0	0.09
[19,42)	10	14.39
[42,45)	6	5.03
[45,50)	11	8.24
[50,55)	9	6.32
[55,65)	4	5.25
[65,+∞)	0	0.68
Total	40	40.00

We wish to test whether a Gaussian model for Y = the fuel consumption per fill-up is consistent with these data. In math, we have

$$H_0: P(Y \in [a_{j-1}, a_j)) = \int_{a_{j-1}}^{a_j} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

$$H_A: P(Y \in [a_{j-1}, a_j)) \neq \int_{a_{j-1}}^{a_j} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

The null hypothesis is Y has a Gaussian distribution with mean is 45.28343 and standard deviation is 9.426723. The alternative hypothesis is Y does not follow a Gaussian distribution with mean is 45.28343 and standard deviation is 9.426723. The test statistic is

$$\lambda = 2\sum_{j=1}^{k} f_j \log\frac{f_j}{e_j}$$

$$= 2\left[10\log\frac{10}{14.39} + 6\log\frac{6}{5.03} + 11\log\frac{11}{8.24} + 9\log\frac{9}{6.32} + 4\log\frac{4}{5.25}\right]$$

$$= 5.3804$$

The distribution under the null hypothesis for the test statistic is approximately chi-squared distribution with 5-1-3=1 degree of freedom. In math, we have

$$df = k-1-p = 5-1-3 = 1 \qquad \Lambda \sim \chi^2(1)$$

For p-value, we have

$$\begin{aligned} \text{p-value} &= P(\Lambda \geq 5.3804 ; H_0) \\ &\approx P(W \geq 5.3804) \quad \text{where } W \sim \chi^2(1) \\ &= 0.02036421 \end{aligned}$$

Because 0.02036421>0.1, there is no evidence against the Gaussian model based on the observed data. Thus, Y has a Gaussian distribution with mean is 45.28343 and standard deviation is 9.426723.

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

i) I) Based on , The estimate $\hat{\beta}$ is 0.07906653.

II) The confidence interval for β based on the model is from 0.07628627 to 0.08184679.

III) This interval is narrower than the interval for the same parameter obtained in c. The first reason is that the number that $\hat{\beta}$ plus or minus of this interval is smaller than the number of the interval obtained in part c. To be specific, because

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2$$

,

S_{xx} should be larger than $\sum_{i=1}^{n} x_i^2$. Moreover, this interval is obtained by

$$\left[\hat{\beta} - a\frac{Se}{\sqrt{\sum_{i=1}^{n} x_i^2}},\ \hat{\beta} + a\frac{Se}{\sqrt{\sum_{i=1}^{n} x_i^2}}\right]$$

, and the interval in part c is obtained by

$$\left[\hat{\beta} - a\frac{Se}{\sqrt{Sxx}},\ \hat{\beta} + a\frac{Se}{\sqrt{Sxx}}\right]$$

. Thus, the number that $\hat{\beta}$ plus or minus in this interval is smaller than the number that $\hat{\beta}$ plus or minus in the interval obtained in part c because the denominator of the number in this interval is larger than the denominator of the number in the interval obtained in part c. This causes that this interval is narrower. Secondly, in part c, because of the presence of α and it is positive, β should be adjusted to make up the difference between the actual value and the expected value since α is positive means that there is fuel consumption when the distance driven is 0 kilometer, which does not match the common fact. However, in part i, we assume α = 0 and so there is no deviation caused by α. Thus, the interval in part c would be wider to adjust the difference. The model in part i is preferable because α = 0 match the common fact that there is no fuel consumption if we do not drive car. Also, the interval of this model is narrower would increase the accuracy.