MA2261 Linear Statistical Models - DLI, Year 2022-2023

Coursework 2

INSTRUCTIONS AND DEADLINE:

Please submit *electronically* one piece of written/typed work per person in a single PDF file by **Friday 14 April 2023 at 4pm UK time/23:00 China time.**

Please use this page as the cover page for your submission. Write below your student ID and sign it.

Student ID:

Signature:

MARKING CRITERIA:

- This problem sheet is worth 100 points. Scores for each main question are indicated at the beginning of each.
- Clearly justify and explain your answers. If you are using the R software for calculations, a printout of your answers without a full explanation of the formulas you are using and your reasoning will not score full marks.
- A true/false question answered without justification will get zero marks.
- Computational mistakes will be penalized more in coursework than in exam marking, since you have plenty of time and tools to check your calculations when doing the coursework.

Question 1 [30 marks]

Decide if the following statements are true or false. Justify your answers, absence of justification counts as incorrect justification.

i) **[5 marks]** In a simple linear regression model, the point $(\bar{x}+1, \bar{y}+\hat{b})$ is on the fitted regression line.

ii) **[5 marks]** The sampling distributions of $\hat{a}$ and $\hat{b}$ are the distributions of the observations y_i's.

iii) **[5 marks]** In a simple linear regression model, the hypothesis $b = 0$ is accepted if and only if the confidence interval for b does not contain 0.

iv) **[5 marks]** For simple linear regression model, there are 4 parameters to be estimated.

v) **[5 marks]** If in a simple linear regression model $\hat{\rho} = 0.35$, then 35% of the variation of Y is explained by x.

vi) **[5 marks]** A simple linear regression model is fitted to data consisting of ten observations. Then it is possible to have 7 negative residuals less than -1 and 3 positive residuals less than 2.

Question 2 [10 marks]

Consider a linear regression model with intercept only,

$$Y_i = a + \varepsilon_i, \qquad i = 1, \ldots, n$$

where ε_i are independent random variables with normal distribution $N(0, \sigma^2)$. Using the method of least squares estimation, derive the maximum likelihood estimator of a for this model.

Question 3 [10 marks]

Consider a linear regression model with a fixed zero intercept,

$$Y_i = b \sin(X_i^2) + \varepsilon_i, \qquad i = 1, \ldots, n$$

where ε_i are independent random variables with normal distribution $N(0, \sigma^2)$. Using the method of least squares estimation, derive the maximum likelihood estimator of b for this model.

Question 4 [50 marks]

In order to investigate the feasibility of starting a Sunday edition for a Daily regional newspaper, information was obtained from a sample of 25 regional newspapers concerning their Daily (x) and Sunday (y) circulation (in millions of copies). Data are shown in Table 1 below.

Newspapers	Daily (x)	Sunday (y)
Baltimore Sun	0.392	0.489
Boston Globe	0.517	0.798
Charlotte Observer	0.239	0.299
Chicago Sun Times	0.538	0.559
Cincinnati Enquirer	0.199	0.349
Denver Post	0.253	0.418
Des Moines Register	0.206	0.345
Hartford Courant	0.231	0.323
Houston Chronicle	0.450	0.621
Kansas City Star	0.289	0.423
Los Angeles Daily News	0.186	0.203
Miami Herald	0.445	0.553
Minneapolis Star Tribune	0.413	0.686
New Orleans Times-Picayune	0.272	0.324
Omaha World Herald	0.224	0.285
Orange County Register	0.355	0.408
Portland Oregonian	0.338	0.441
Providence Journal-Bulletin	0.197	0.268
Rochester Democrat & Chronicle	0.133	0.262
Rocky Mountain News	0.374	0.433
Sacramento Bee	0.274	0.338
San Francisco Chronicle	0.570	0.704
St. Louis Post-Dispatch	0.391	0.586
St. Paul Pioneer Press	0.202	0.268
Tampa Tribune	0.322	0.408

Table 1: Daily and Sunday circulation of newspapers (Source: Gale Directory of Publications, 1994).

Assume the significance level is $\alpha = 0.05$.

a) i) **[5 marks]** Obtain the estimated simple linear regression line.

ii) **[5 marks]** Is there a significant relationship between Sunday circulation and Daily circulation? Justify your answer by testing whether the slope of the simple linear regression is zero and indicate your conclusion.

iii) **[5 marks]** Obtain the 95% confidence intervals for $\hat{a}$ and $\hat{b}$. Justify whether the regression line goes through the origin?

b) i) **[5 marks]** What proportion of the variation in the Sunday circulation is explained by the Daily circulation?

ii) [**5 marks**] Calculate a 95% confidence interval for the correlation coefficient between the two variables. And test the hypothesis that the correlation coefficient between the two variables is zero.

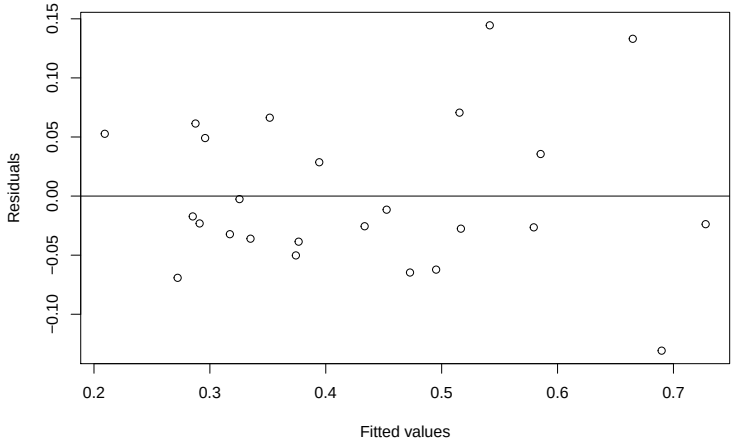
iii) [**5 marks**] Compare your conclusions in a) ii) and b) ii).

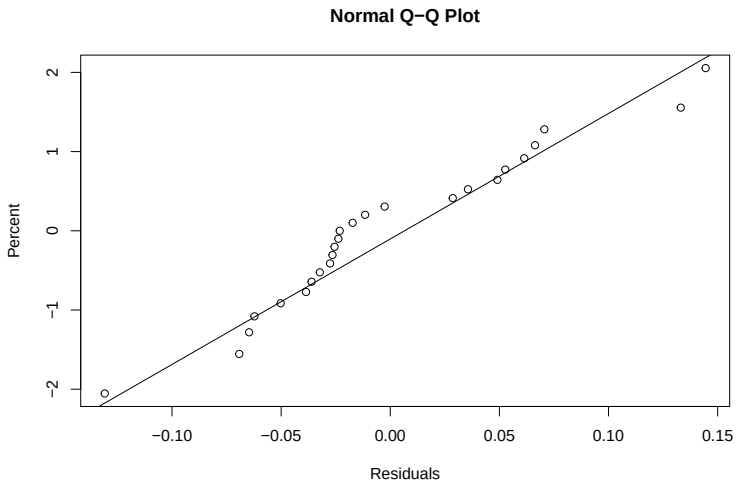
c) i) [**5 marks**] Calculate the 95% confidence interval for the expected number of Sunday circulation of a newspaper with a Daily circulation of 0.5 million copies.

ii) [**5 marks**] A particular newspaper that is considering a Sunday edition has a Daily circulation of 0.5 million. Provide a 95% prediction interval for the predicted Sunday circulation of this newspaper.

iii) [**5 marks**] Explain the difference in meaning between the intervals in part c) i) and c) ii) and why one interval is wider than the other.

d) [**5 marks**] On the basis of residual plots below, comment on the validity of simple linear regression model for these data.





Solution to Question 1

i) **[5 marks]** True. In fact, since $\bar{y} = \hat{b}\bar{x} + \hat{a}$, we have $\bar{y} + \hat{b} = \hat{b}\bar{x} + \hat{a} + \hat{b} = \hat{b}(\bar{x} + 1) + \hat{a}$.

ii) **[5 marks]** False. The sampling distributions are the ones obtained by repeating the experiment many times and each time calculating $\hat{a}$, $\hat{b}$ and then looking at the distributions of these values.

iii) **[5 marks]** False. The hypothesis $b = 0$ is accepted if and only if the confidence interval for b contains 0.

iv) **[5 marks]** False. There are 3 parameters, $a, b,$ and σ^2.

v) **[5 marks]** False. It is $\hat{\rho}^2 = 0.35^2 = 0.1225$, hence 12.25% of the variation of Y is explained by x.

vi) **[5 marks]** False. We have $\sum_{i=1}^{10} r_i \leq -1 \times 7 + 3 \times 2 = -1$. Hence in particular $\sum_{i=1}^{10} r_i \neq 0$. This contradicts the fact that in the simple linear regression model the sum of the residuals is zero.

Solution to Question 2

[10 marks] The given model is equivalent to $Y_i \sim N(a, \sigma^2)$. The the likelihood function is given by

$$\begin{aligned}\mathcal{L}(a, \sigma^2|\mathbf{Y}) &= \prod_{i=1}^{n} f(y_i|a, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_i - a)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a)^2\right\}\end{aligned}$$

To maximize the likelihood we need to minimize

$$SS = \sum_{i=1}^{n}(y_i - a)^2$$

which is a function of the parameter a. To minimize it, we need to equate to zero the partial derivative

$$\frac{dSS}{da} = \sum_{i=1}^{n} 2(-1)(y_i - a) = 0 \Rightarrow \sum_{i=1}^{n} y_i - na = 0$$

Therefore,

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

Solution to Question 3

[10 marks] The given model is equivalent to $Y_i \sim N(b\sin(X_i^2), \sigma^2)$. The the likelihood function is given by

$$\mathcal{L}(b, \sigma^2|\mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{n} f(y_i|b, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{[y_i - b\sin(x_i^2)]^2}{2\sigma^2}\right\}$$
$$= \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - b\sin(x_i^2)\right]^2\right\}$$

To maximize the likelihood we need to minimize

$$SS = \sum_{i=1}^{n}\left[y_i - b\sin(x_i^2)\right]^2$$

which is a function of the parameter b. To minimize it, we need to equate to zero the partial derivative

$$\frac{dSS}{db} = \sum_{i=1}^{n} -2\sin(x_i^2)\left[y_i - b\sin(x_i^2)\right] = 0 \Rightarrow \sum_{i=1}^{n}\left[y_i\sin(x_i^2) - b\sin^2(x_i^2)\right] = 0$$

Therefore,

$$\hat{b} = \frac{\sum_{i=1}^{n} y_i\sin(x_i^2)}{\sum_{i=1}^{n}\sin^2(x_i^2)}$$

Solution to Question 4

a) i) **[5 marks]** We firstly calculate $\bar{x} = 0.3204, \quad \bar{y} = 0.43164,$

$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 0.34788,$

$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 0.5886898,$

$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} = 0.4127466.$

We then calculate $\hat{a}$ and $\hat{b}$,

$\hat{b} = \frac{S_{xy}}{S_{xx}} = 1.1865, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} = 0.0515.$ The regression line is $\hat{y} = 0.0515 + 1.1865x.$

ii) **[5 marks]** We need to test the null hypothesis $H_0 : b = 0$ against $H_1 : b \neq 0$. For this we have to calculate $\hat{\sigma}$:

$\hat{\sigma}^2 = \frac{RSS}{(n-2)} = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = 0.0043035,$ therefore $\hat{\sigma} = 0.065601.$

Hence we can calculate

$$\frac{\hat{b}\sqrt{S_{xx}}}{\hat{\sigma}} = \frac{1.1865\sqrt{0.34788}}{0.065601} = 10.667 \sim t_{23}$$

we can use either the following ways to test

$$\text{p-value} < 0.001 < 0.05$$
$$\text{Critical region is } (-\infty, -2.069) \cup (2.069, +\infty)$$

Hence, we reject the null hypothesis and conclude that the Daily circulation is statistically significant in determining the mean Sunday circulation.

iii) **[5 marks]** We calculate

95% C.I. for a:

$$\hat{a} \pm t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}} = 0.0515 \pm 2.069 \times 0.065601\sqrt{\frac{1}{25}+\frac{0.3204^2}{0.34788}} = (-0.027; 0.13).$$

95% C.I. for b:

$$\hat{b} \pm t_{0.025,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}} = 1.1865 \pm 2.069 \times \frac{0.065601}{\sqrt{0.34788}} = (0.9564; 1.4165).$$

The C.I. for a contains 0, therefore we accept the hypothesis that the regression line goes through the origin.

b) i) **[5 marks]** The sample correlation coefficient is

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{0.4127466}{\sqrt{0.34788 \times 0.5886898}} = 0.91206$$

Therefore we obtain $\hat{\rho}^2 = 0.83186$, which means that 83.186% of variation in Sunday circulation is explained by Daily circulation.

ii) **[5 marks]** The 95% confidence interval for correlation ρ is (l_1,l_2) where

$$l_1 = \frac{e^{-\left(\frac{2\times 1.96}{\sqrt{n-3}}\right)}\frac{1+\hat{\rho}}{1-\hat{\rho}} - 1}{e^{-\left(\frac{2\times 1.96}{\sqrt{n-3}}\right)}\frac{1+\hat{\rho}}{1-\hat{\rho}} + 1} = 0.8082$$

$$l_2 = \frac{e^{\left(\frac{2\times 1.96}{\sqrt{n-3}}\right)}\frac{1+\hat{\rho}}{1-\hat{\rho}} - 1}{e^{\left(\frac{2\times 1.96}{\sqrt{n-3}}\right)}\frac{1+\hat{\rho}}{1-\hat{\rho}} + 1} = 0.9609$$

The C.I. for ρ does not contain 0, therefore we reject the null hypothesis $\rho = 0$.

iii) **[5 marks]** The hypothesis $\rho = 0$ is rejected, which is expected from the fact that the hypothesis $b = 0$ is equivalent to $\rho = 0$. A test of both hypotheses can not lead to contradictory conclusions. Both tests are alternatives to testing whether the explanatory is statistically significant to the response.

c) i) **[5 marks]** This question requires the calculation of the confidence interval of mean response for $x_0 = 0.5$, that is

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = (0.595289; 0.694168).$$

ii) **[5 marks]** Because the question deals with a particular newspaper that is considering a Sunday edition, we have to calculate the prediction interval for $x_0 = 0.5$, which is

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = (0.500297; 0.789161).$$

iii) **[5 marks]** The first is the confidence interval for the mean response corresponding to $x_0 = 0.5$ and takes into account only the variance of $\hat{a}$ and $\hat{b}$. The second is the prediction interval for the new observation corresponding to $x_0 = 0.5$ and also takes into account the variance of the error component of the model. This why the second interval is wider.

d) **[5 marks]** From eye inspection, the normal probability plot looks reasonably linear, indicating the assumptions of the linear regression model are satisfied. The plot of residuals versus fit shows no pattern. Hence we can reasonably conclude that the simple linear regression assumptions are satisfied.