MTH5120 Statistical Modelling 1

May 2023 – Solutions

Question 1

(a) normal distribution, zero mean [hence the relationship between y and x is linear], constant variance some σ^2

(b) $\beta 1\text{-hat} = S_{xy} / S_{xx}$

$S_{xy} = \sum xy - \sum x \sum y/n = 760359 – (4854)(1875)/12$

$S_{xx} = \sum x^2 – (\sum x)^2/n = 1964356 – 4854^2/12$

$\beta 1\text{-hat} = 2.1046$

$\beta 0\text{-hat} = \text{mean}(y) – \beta 1\text{-hat}.\text{mean}(x) = -695.0608$

(c) Coefficient of determination = $R^2 = 1 – SS_E / SS_T = 1 – 3532.3 / 7576.3 = 0.53377 = 53.38\%$

(d) d.f. total = n – 1 =11

d.f. regression = 1

d.f. residuals = n – 2 = 10

$SS_T = 7576.3$

$SS_R = SS_T – SS_E = 7576.3 – 3532.3 = 4044$

$SS_E = 3532.3$

$MS_R = SS_R / 1 = 4044$

$MS_E = SS_E / 10 = 353.23$

$VR = F = MS_R / MS_E = 4044 / 353.23 = 11.4486$

(e) MS_E is an unbiased estimator for σ^2 so our estimate from ANOVA is 353.23

(f) H0: $\beta 1 = 0$ versus H1: $\beta 1 \neq 0$

(g) under H0 the variance ratio follows a Fisher F distribution on 1 and 10 d.f.

the critical value F(0.05) = 4.965

our F statistic 11.4486 from ANOVA > 4.965

therefore we reject H0

there is evidence that $\beta 1$ is significantly different from 0 at the 95% level

(h) we want y-hat at x=419 = -695.0608 + (2.1046)(419) = 186.77

(i) the 95% confidence interval is [a,b] = y-hat +/- t(0.025,10) $\sqrt{S^2(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}})}$

= 186.77 +/- 2.228139$\sqrt{353.23(\frac{1}{12}+\frac{(419-404.5)^2}{913})}$

= 186.77 +/- 2.228139(10.52518)

= [163.3184, 210.2216]

(j) the 95% prediction interval is [a,b] = y-hat +/- t(0.025,10) $\sqrt{S^2(1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}})}$

= 186.77 +/- 2.228139$\sqrt{1+353.23(\frac{1}{12}+\frac{(419-404.5)^2}{913})}$

= 186.77 +/- 2.228139(21.54088)

= [138.7739, 234.7661]

(k) the point estimate simply takes its value from the regression parameter estimates. The confidence interval allows for the number of observations in the model, the model variance estimate and the amount of dispersion amongst x values. The prediction interval is considerably wider than the confidence interval because it allows for the fact that we do not know how the random error component of the model will behave at the new observation.

(l) There are 3 residual plots one for each model assumption

first we need to calculate standardised residuals and fitted values at each observed x

the first plot is standardised residuals versus x

this is to check the assumption of a linear relationship

we are looking for the standardised residuals to appear random with no discernible pattern with x

the next plot is standardised residuals versus fitted y values

this is to check the assumption of a constant variance

again we look for the standardised residuals to appear random with no discernible pattern with y-hat – in particular we are concerned to look for a “funnel” shape indicating increasing variance with y

the final plot is the QQ Plot

this tests the assumption of normally distributed residuals

we look for the QQ plot to be close to the straight line (QQ line) throughout

Notes

(a) lecture notes
(b) similar exercise sheet
(c) similar exercise sheet
(d) similar exercise sheet
(e) method in lecture
(f) method in lecture
(g) similar coursework
(h) similar exercise sheet
(i) similar exercise sheet
(j) similar exercise sheet
(k) unseen, higher order
(l) applies lecture material

IFoA CS1 syllabus 4.1.1, 4.1.2, 4.1.3, 4.1.4

Question 2

(a) There does appear to be some evidence that inflation increases with interest rates

however there is quite a lot of variation at similar interest rates

there is one observation with a much higher x value

we should check whether this is an 'influential' or high leverage observation

we note that the regression line is close to this observation

without this observation the fitted model may be quite different

(b) average leverage = 2/n so here is 2/14 = 0.142857
(c) the six lines of R code:
- fits a linear regression model to x, y and calls is econ
- calculates standardised residuals for econ and assigns them to di
- calculates leverage values for econ and assigns them to vi
- calculates Cook's Statistic for each econ observation and assigns them to Di
- creates a sequence 1, 2, .., n and assigns it to i
- plots the leverage values vi in observation order

(d) At 0.89 the leverage of this observation is > 6 times average so this is a very high leverage observation

so we should consider repeating the regression without this observation

a plot of the Cook's Statistic values will give a visual clue as to how much higher 7.98 is than the other observations' values

we can formally test the significance of Cook = 7.98

we compare it to the 50^{th} percentile of the F distribution on 2 and n – 2 = 12 d.f.

If 7.98 > this median F then the Cook's Statistic is significantly large which would provide more formal evidence to repeat the regression without this observation.

Notes

(a) unseen
(b) lecture notes
(c) IT lab practical
(d) application of IT lab, higher order

IFoA CS1 4.1.4

Question 3

(a) the likelihood function is the joint probability function of the observations

$L(\theta) = \theta^k(1 - \theta)^{n-k}$

(b) we first take log of the likelihood

$\log L = k \log(\theta) + (n - k)\log(1 - \theta)$

we then differentiate with respect to θ

$$\frac{d\log L}{d\theta} = \frac{k}{\theta} - \frac{n-k}{1-\theta}$$

we set this to zero and solve to find the maximum at θ-hat

$$\hat{\theta} = k/n$$

(c) MLE's have strong asymptotic properties

unbiased, normal distribution, lowest possible variance for an estimator

but those asymptotic properties may not exhibit at small n

with n=10, a single observation will move θ-hat by 0.1

with n=500, a single observation will move θ-hat by 0.002

a biased estimator in a clinical trial could lead to incorrect biomedical decisions

Notes

(a) lecture notes
(b) lecture notes
(c) application of lecture material unseen

beyond IFoA CS1 syllabus

Question 4

(a) the model becomes

$$\begin{pmatrix} y_1 \\ \vdots \\ y_6 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_3 \end{pmatrix} \begin{pmatrix} 1 & \cdots & x_{31} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{36} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_6 \end{pmatrix}$$

(b) MS_E is an unbiased estimator of the variance

degrees of freedom for residuals = n – p = n – 4 = 11

$MS_E = SS_E / (n-4) = SS_E / 11$

$SS_E = (1 – R^2) SS_T = (1 – 0.9879)(7675.5) = 92.87355$

so $MS_E = 8.44305$

(c) we use the extra sum of squares principle

for each reduced model the extra SS = SS_E(reduced) - SS_E(full)

for x1+x2, extra SS = 863.8 – 92.87355 = 770.9265

for x1+x3, extra SS = 6451.3 – 92.87355 = 6358.43

for x2+x3, extra SS = 700.7 – 92.87355 = 607.8265

then for each reduced model we can test

H0: the removed parameters are zero

H1: the removed parameters are not zero

we calculate the F statistic, F = (extra SS / parameters removed) / (estimated variance)

under H0 that F statistic follows a Fisher F distribution on 4 – 3 = 1 and 15 – 4 = 11 d.f.

So the critical value we need from the table is F(0.05) = 4.84

for x1+x2, F = 770.9265/8.44305 = 91.309 > 4.84 therefore we reject H0

for x1+x3, F = 6358.43/8.44305 = 753.096 > 4.84 therefore we reject H0

for x2+x3, F = 607.8265/8.44305 = 71.991 > 4.84 therefore we reject H0

for each of the three possible reduced models we reject H0

therefore we are unable to remove any of the explanatory variables at the 95% significance level and we need to stay with the 3 variable, 4 parameter model.

Notes

(a) lecture notes
(b) unseen application of lecture
(c) part similar exercise sheet, part unseen application

IFoA CS1 syllabus 4.1.5, 4.1.6 and part beyond CS1 syllabus