MTH5120 Statistical Modelling 1

LSR 2023 – Solutions

Question 1

(a) the y_i are usually assumed to be

- normally distributed
- with mean $\beta_0 + \beta_1 x_i$
- with constant variance σ^2

(b) β_1-hat = S_{xy} / S_{xx}

$Sxy = \sum xy - \sum x \sum y/10 = 5.876$

$Sxx = \sum x^2 - (\sum x)^2/10 = 6.264$

β_1-hat = 0.938059

β_0-hat = mean(y) - β_1-hat.mean(x) = 8.11 – 0.938059(2.84) = 5.445913

(c) the 95% confidence interval is [a,b] where

[a,b] = β_1-hat +/- t(0.025; 8) s.e.(β_1-hat) = 0.938059 +/- 1.8595(0.4074)

= [0.180498, 1.695619]

(d) the value 0 does not lie within the 95% confidence interval for β_1-hat therefore we can reject H0: $\beta_1 = 0$ at the 95% significance level

(e) we could also test the Variance Ratio in the ANOVA table against the critical value of the F distribution on 1 and 8 d.f.

Notes

(a) lecture notes
(b) similar exercise sheet
(c) similar exercise sheet
(d) unseen
(e) application of lecture notes

IFoA CS1 syllabus 4.1.1, 4.1.2, 4.1.3, 4.1.4

Question 2

(a) There are n=32 observations so

d.f. for regression = 1

d.f. for residuals = n – 2 = 30

total d.f. = n – 1 = 31

$R^2 = 0.6307 = 1 – (SS_E / SS_T)$ therefore $SS_T = SS_E /(1 – 0.6307) = 60.977 / 0.3693 = 165.1151$

$SS_R = SS_T – SS_E = 104.1381$

$MS_R = SS_R / 1 = 104.1381$

$MS_E = SS_E / 30 = 2.032567$

$VR = F = MS_R / MS_E = 51.23477$

(b) MS_E is an unbiased estimator of $\sigma^2 = 2.032567$

(c) we first need y-hat at x=5

y-hat = 9.2928 – 5(1.2269) = 3.1583

then the prediction interval is [a,b] where

$$[a,b] = \text{y-hat} +/- t(0.025,n-2) \sqrt{S^2(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}})}$$

now $(x_i - \bar{x})^2 = (5 – 3.7)^2 = 1.69$, Sxx = 69.18, n=32, $S^2 = 2.032567$

so $[a,b] = 3.1583 +/- 1.812\{ 2.032567(1 + 1/32 + 1.69/69.18)\}^{1/2}$

$= 3.1583 +/- 2.654279 = [0.504021, 5.812579]$

(d) The 1/n term in the prediction interval would change from 1/32 to 1/48 so all other things being equal this would lead to a very small reduction in the width of the prediction interval but because mean(x), Sxx and S^2 would also change we cannot be certain – it is possible that the interval could be wider or narrower.

Notes

(a) similar exercise sheet
(b) lecture notes
(c) similar exercise sheet
(d) unseen, higher order

IFoA CS1 syllabus 4.1.4

Question 3

(a) crude residuals = observed y – fitted y = $y_i – y_i$hat

(b) the crude residuals have a variance (lower) and covariances (higher, non-zero) than those assumed for the error terms epsilon_i in the original model specification. Therefore we standardise to bring the variance / covariances closer to the assumed values (σ^2 and zero respectively)

(c) These lines of R code

- construct a simple linear regression of y on x and store it as object slr
- calculate the standardised residuals of slr and store them in a vector di
- calculated the fitted values for each x in slr and store them as yh
- plot the standardised residuals against x
- plot the standardised residuals against the fitted y
- creates a QQ plot from the standardised residuals
- plots the QQ line

(d) the first plot is standardised residuals against x

this is to test the assumption of linearity in the model

we are looking for the plot to appear random with no discernible pattern

this seems to be the case with this plot

the second plot is standardised residuals against fitted y values

this is to test the assumption of constant variance

again we look for the plot to appear random

our plot has the funnel shape with more dispersion in residuals are fitted y increases

this is indication that the variance is not constant but increasing with y

the third plot is the QQ plot

this tests the assumption of normally distributed residuals

we look for the plot to sit along the straight line

there is variability away from the line throughout but particularly at the lowest and highest residual values

this suggests the distribution may not be normal but have fatter tails

all three plots can also be used to detect outliers

there is no evidence of outliers here

Notes

(a) lecture notes
(b) lecture notes
(c) similar IT lab practical
(d) similar IT lab practical

IFoA CS1 syllabus 4.1.4

Question 4

(a) in matrix form

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_6 \end{pmatrix} \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_6 \end{pmatrix} \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_6 \end{pmatrix}$$

(b) in matrix form the normal equations become

$$\boldsymbol{X^T y} = \boldsymbol{X^T X}\, \widehat{\boldsymbol{\beta}}$$

and for this to have a unique solution for $\widehat{\boldsymbol{\beta}}$ we need $\boldsymbol{X^T X}$ to be invertible

(c) $\boldsymbol{X^T X} = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_6 \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_6 \end{pmatrix} = \begin{pmatrix} 6 & \sum x \\ \sum x & \sum x^2 \end{pmatrix}$

for $\boldsymbol{X^T X}$ to be invertible we need its determinant to be non zero

here the determinant = $6\sum x^2 - (\sum x)^2 = 6S_{xx}$ which is not zero

(d) the hat matrix is $\boldsymbol{H} = \boldsymbol{X}\,(\boldsymbol{X^T X})^{-1}\,\boldsymbol{X^T}$

$\mathbf{H} = \mathbf{HH}$ and $\mathbf{H} = \mathbf{H}^T$

Notes

all parts – lecture notes, beyond IFoA CS1 syllabus

Question 5

(a) the R code required is

```
sales <- lm(y~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11)
summary(sales)
anova(sales)
```

(b) $SS_R = SS_T - SS_E = 15890 - 318 = 15572$

there are n = 16 x 9 = 144 observations

d.f. regression = 12 – 1 = 11

d.f. residuals = 144 – 12 = 132

$MS_R = 15572 / 11 = 1415.636$

$MS_E = 318 / 132 = 2.409091$

$VR = MS_R / MS_E = 587.6226$

(c) when the model assumptions (linear relationship, constant variance, normal residuals) are true

(d) this is the principle of parsimony that we seek the simplest model that explains the data well

(e) it is unchanged by the number of explanatory variables so 15890

(f) our full model is `y~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11`

our reduced model is `y~ x1+x3+x5+x8+x9+x11`

we need regression models and anova tables for each

```
sales1 <- lm(y~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11)
summary(sales1)
anova(sales1)
sales2 <- lm(y~ x1+x3+x5+x8+x9+x11)
summary(sales2)
anova(sales2)
```

from the first anova we extract $SS_E(1)$ and MS_E [which will be our S^2 variance estimate]

from the second anova we extract $SS_E(2)$

the extraSS = $SS_E(2) - SS_E(1)$

the number of additional parameters = 11 – 6 = 5

therefore our F statistic = (extraSS / 5) / S^2

H0: the 5 parameters not in the full model are all zero

H1: at least one of these parameters is not zero

under H0 our F statistic follows Fisher's F on 5 and 144 – 12 = 132 d.f.

we find the critical value at p=0.05 in R with

```
qf(0.05, 5, 132, lower.tail=FALSE)
```

and reject H0 if our F statistic is > this value

(g) adjusted $R^2 = 100\%(1 - (n-1)\frac{MS_E}{SS_T})$

this is better when comparing models with different numbers of explanatory variables because all additional variables will improve R^2, but only statistically significant ones (by an F test) will improve adjusted R^2

Notes

(a) similar IT lab
(b) similar exercise sheet
(c) lecture notes
(d) lecture notes, higher order
(e) unseen
(f) some IT lab, some unseen
(g) lecture notes

IFoA CS1 syllabus 4.1.5, 4.1.6