MAST90104: A First Course in Statistical Learning

Assignment 2, 2023 Solution

1. (8 marks) Consider the full rank linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \ldots, n.$$

Assume that the errors are independent, normally distributed with mean 0 and variance σ^2. Derive an expression for a joint $100(1-\alpha)\%$ confidence region for parameter β_1 and β_2.

(*Hint: You can use the result that the marginal distribution over a subset of multivariate normal random variables is multivariate normal, and the mean and covariance matrix are obtained by dropping the irrelevant variables (the variables that one wants to marginalize out) from the mean vector and the covariance matrix.*)

Solution Note that the least square estimator $\mathbf{b} = \begin{bmatrix} b_0 & b_1 & b_2 \end{bmatrix}^T$ follows a multivariate normal distribution

$$\mathbf{b} \sim MVN(\beta, (X^T X)^{-1}\sigma^2).$$

Let $V = (X^T X)^{-1}$, so var $\mathbf{b} = \sigma^2 V$.

Let $\mathbf{d} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ and $\boldsymbol{\delta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$, then $\mathbf{d} \sim MVN\left(\boldsymbol{\delta}, \sigma^2 A\right)$, where

$$A = \begin{bmatrix} V_{22} & V_{23} \\ V_{32} & V_{33} \end{bmatrix}.$$

Alternatively note that $\mathbf{d} = C\mathbf{b}$ where $C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Then using the property that linear combination of multivariate normals results in another multivariate normal, then $\mathbf{d} \sim MVN(C\boldsymbol{\beta}, \sigma^2 CVC^T)$, which gives the same result.

We have :

$$\frac{1}{\sigma^2}(\mathbf{d}-\boldsymbol{\delta})^T A^{-1}(\mathbf{d}-\boldsymbol{\delta}) \sim \chi^2_2.$$

Because $SS_{Res}/\sigma^2 \sim \chi^2_{n-3}$, we have

$$\frac{\frac{1}{\sigma^2}(\mathbf{d}-\boldsymbol{\delta})^T A^{-1}(\mathbf{d}-\boldsymbol{\delta})/2}{\frac{SS_{Res}}{\sigma^2}/(n-3)} = \frac{(\mathbf{d}-\boldsymbol{\delta})^T A^{-1}(\mathbf{d}-\boldsymbol{\delta})}{2s^2} \sim F_{2,n-3}$$

Let f_α be the $1-\alpha$ quantile (or $100(1-\alpha)\%$ percentile) of the F distribution with degrees of freedom 2 and $n-3$, then the confidence region is

$$(\mathbf{d}-\boldsymbol{\delta})^T A^{-1}(\mathbf{d}-\boldsymbol{\delta}) \leq 2s^2 f_\alpha.$$

Or

$$\begin{bmatrix} b_1-\beta_1 & b_2-\beta_2 \end{bmatrix} A^{-1} \begin{bmatrix} b_1-\beta_1 \\ b_2-\beta_2 \end{bmatrix} \leq 2s^2 f_\alpha$$

2. (20 marks) A study was conducted to predict success in the early university years. Success was measured using the cumulative grade point average (GPA). The file *gpa.csv* contains the records of 100 students. The variables that we are considering are the students' GPA after three semesters, their average high school grades in Mathematics (HSM), Science (HSS) and English (HSE).

(a) Write down the formula of a linear model to predict a student's GPA at university based on their high school grades in Mathematics, Science and English.

$$GPA = \beta_0 + \beta_1 HSM + \beta_2 HSS + \beta_3 HSE + \epsilon$$

or

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i,$$

where $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are HSM, HSS and HSE. $\mathbf{y}$ is GPA.

(b) Fit the model in part (a) to the data and estimate the parameters and variance σ^2, using the formulas in the lecture notes.

Solution Note that $\mathbf{b} = (X^T X)^{-1} X^T y$. Where $X = \left[\begin{array}{c|c|c|c} 1 & \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{array} \right]$

```
> y = gpa$GPA
> n = nrow(gpa)
> X = cbind(rep(1,n),gpa$HSM,gpa$HSS,gpa$HSE)
> (b= solve(t(X)%*%X,t(X)%*%y))
             [,1]
[1,] -0.30174744
[2,]  0.18826042
[3,]  0.08457661
[4,]  0.09377269
```

The sample variance $s^2 = SS_{Res}/(100-4) = \mathbf{e}^T\mathbf{e}/96 = 0.5390297$

(c) Fit the model using the function `lm()` in R.

Solution

```
> m1 = lm(GPA~ HSM+ HSS + HSE, data = gpa)
> m1$coefficients
(Intercept)         HSM         HSS         HSE
-0.30174744  0.18826042  0.08457661  0.09377269
```

(d) Using the formulas in the lecture notes, calculate the standardised residual, leverage, and Cook's distance for the 15^{th} observation.

Solution

The standardised residual is 0.8445203 ; The leverage is 0.01965834; The Cook's distance is 0.003575442;

```
> p = 4
> e = m1$residuals
> (s2 = sum(e^2)/(n-p))
[1] 0.5390297
> H <- X %*% solve(t(X) %*% X) %*% t(X)
> id = 15
> # standardised residual
> z <- e/sqrt(s2*(1-diag(H)))
> z[id]
       15
0.8445203
> # leverage
> ( H[id,id])
[1] 0.01965834
> # Cook's distance
> (1/p * z[id]^2 * H[id,id]/(1-H[id,id]))
         15
0.003575442
```

(e) Estimate the GPA after three semesters of a student whose high school grades in Mathematics, Science and English are 8, 9 and 7 respectively.

Solution The point estimate is $(\mathbf{x}^*)^T\mathbf{b}$ where $\mathbf{t} = \begin{bmatrix} 1 & 8 & 9 & 7 \end{bmatrix}^T$

```
> xstar = c(1,8,9,7)
> (ystar = t(xstar)%*%b)
         [,1]
[1,] 2.621934
```

Alternatively, you can use R's function

```
predict(m1, newdata = data.frame(HSM = 8, HSS = 9, HSE = 7))
```

For part (f) and (g), please use both matrix calculations and `R` functions.

(f) Calculate a 99% confidence interval for the expected GPA after three semesters of a student whose high school grades in Mathematics, Science and English are 8, 9 and 7 respectively.

Solution

The confidence interval is $(\mathbf{x}^*)^T\mathbf{b} \pm t_{\alpha/2}s\sqrt{(\mathbf{x}^*)^T(X^TX)^{-1}(\mathbf{x}^*)} = (2.175524, 3.068345)$

In R:

```
halfwidth = qt(0.995,n-p)*sqrt(s2*(t(xstar)%*%solve(t(X)%*%X)%*%xstar))

c(ystar-halfwidth, ystar+halfwidth)
# [1] 2.175524 3.068345
```

Using `predict()` function gives

```
predict(m1, newdata = data.frame(HSM = 8, HSS = 9, HSE = 7),
         interval = "confidence",level = 0.99)
#        fit      lwr      upr
# 1 2.621934 2.175524 3.068345
```

(g) Calculate a 90% prediction interval for the GPA after three semesters of a student whose high school grades in Mathematics, Science and English are 8, 9 and 7 respectively.

Solution

The prediction interval is

$$(\mathbf{x}^*)^T\mathbf{b} \pm t_{\alpha/2}s\sqrt{1 + (\mathbf{x}^*)^T(X^TX)^{-1}(\mathbf{x}^*)} = (1.370326, 3.873543)$$

In R:

```
halfwidth = qt(0.95,n-p)*sqrt(s2*(1+t(xstar)%*%solve(t(X)%*%X)%*%xstar))
c(ystar-halfwidth, ystar+halfwidth)

# [1] 1.370326 3.873543
```

Using `predict()` function gives

```
predict(m1, newdata = data.frame(HSM = 8, HSS = 9, HSE = 7),
interval = "prediction",level = 0.9)
#       fit      lwr      upr
#1 2.621934 1.370326 3.873543
```

(h) Test the hypothesis that the parameter corresponding to HSS is 0, using both t test and F test.

The coefficient of HSS is β_2 in the model that I fitted, so the t-test statistic is

$$t^* = \frac{b_2}{s\sqrt{c_{22}}} = 1.03185$$

For the F test, one can compute the test statistic $H_0 : C\boldsymbol{\beta} = \delta^*$ where $C = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}$ and $\delta^* = 0$.

$$F^* = \frac{(C\mathbf{b})^T[C(X^TX)^{-1}C^T]^{-1}(C\mathbf{b})/1}{SS_{Res}/(n-4)}$$

Using the formula above, or the `linearHypothesis` function in R, gives the test statistics $F^* = 1.0647$

```
# t-test
XTXinv = solve(t(X)%*%X)
(tstar = (b[3])/sqrt(s2*XTXinv[3,3])) # less than 2.78
# [1] 1.03185
2*(1-(pt(abs(tstar),n-p)))
# [1] 0.3047348
# F-test
C = c(0,0,1,0)
dst = 0
library(car)
linearHypothesis(m1,C,dst)

# Linear hypothesis test
#
# Hypothesis:
# HSS = 0
#
# Model 1: restricted model
# Model 2: GPA ~ HSM + HSS + HSE
#
#   Res.Df    RSS Df Sum of Sq      F Pr(>F)
# 1     97 52.321
# 2     96 51.747  1   0.57391 1.0647 0.3047
```

Using a t distribution with $n - p = 96$ degrees of freedom or F distribution with degrees freedom 1 and 96, the p-value is 0.3047. We do not reject the null hypothesis that $\beta_{HSS} = 0$

3. (12 marks) Consider the data set in question 2. The data also contains the students' scores from three SAT tests: SAT Mathematics (SATM), SAT Critical Reading (SATCR) and SAT Writing (SATW).

   (a) Fit a linear model to predict GPA based on high school grades and SAT scores. You may use the function `lm()` in **R**.

   Solution

```
> m2 <- lm(GPA~HSM+ HSS + HSE + SATM + SATCR + SATW,data = gpa)
> summary(m2)

Call:
lm(formula = GPA ~ HSM + HSS + HSE + SATM + SATCR + SATW, data = gpa)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9796 -0.2542  0.1512  0.4695  1.6147

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.1600525  0.7601631  -1.526   0.1304
```

```
HSM          0.1521029  0.0713699   2.131   0.0357 *
HSS          0.0866930  0.0819746   1.058   0.2930
HSE          0.0970508  0.0822070   1.181   0.2408
SATM         0.0015559  0.0012657   1.229   0.2221
SATCR        0.0004277  0.0013344   0.320   0.7493
SATW        -0.0001732  0.0015274  -0.113   0.9100
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.734 on 93 degrees of freedom
Multiple R-squared:  0.3153,Adjusted R-squared:  0.2711
F-statistic: 7.138 on 6 and 93 DF,  p-value: 2.813e-06
```

(b) Test for model relevance using a corrected sum of squares.

Solution: We test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_6 = 0$

The test statistics is

$$F^* = \frac{SSReg - n\bar{y}^2}{SSRes/(n-p)} = 7.1378,$$

it follows a F distribution with degrees of freedom 6 and 93. The p value is almost 0, so we reject the null hypothesis. At least one of the β's is non-zero or the model is relevant.

```
SSReg_corrected = t(y)%*%m2$fitted.values - n*mean(y)^2
SSRes = sum((y- m2$fitted.values)^2)
p = 7
(Fstar = (SSReg_corrected/(p-1))/(SSRes/(n-p)))
#          [,1]
# [1,] 7.137796
(pf(Fstar,p-1,n-p,lower.tail = F))
#              [,1]
# [1,] 2.812708e-06
```

Alternatively

```
nullmodel <- lm(GPA~1,data =gpa)
anova(nullmodel,m2)
# Analysis of Variance Table
#
# Model 1: GPA ~ 1
# Model 2: GPA ~ HSM + HSS + HSE + SATM + SATCR + SATW
#   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
# 1     99 73.182
# 2     93 50.107  6    23.075 7.1378 2.813e-06 ***
#   ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(c) Starting from the full model in part (a), use stepwise selection with AIC to select variables for your model. Use this as your final model.

Solution

The model selected by stepwise is

$$GPA = \beta_0 + \beta_1 HSM + \beta_2 HSE + \beta_3 SATM + \epsilon$$

```
step(m2, scope = ~ . + HSM+ HSS + HSE + SATM + SATCR + SATW,steps = 1)
m3 <- lm(GPA~HSM+ HSS + HSE + SATM + SATCR ,data = gpa)
step(m3,scope = ~ . + HSM+ HSS + HSE + SATM + SATCR + SATW,steps = 1)
m4 <- lm(GPA~HSM+ HSS + HSE + SATM ,data = gpa)
step(m4,scope = ~ . + HSM+ HSS + HSE + SATM + SATCR + SATW,steps = 1)
```

```
m5 <- lm(GPA~HSM + HSE + SATM ,data = gpa)
step(m5,scope = ~ . + HSM+ HSS + HSE + SATM + SATCR + SATW,steps = 1)

summary(m5)

# Call:
# lm(formula = GPA ~ HSM + HSE + SATM, data = gpa)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -2.0479 -0.3279  0.1491  0.4530  1.7136
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) -1.142484   0.725220  -1.575  0.11846
# HSM          0.187554   0.059687   3.142  0.00223 **
# HSE          0.151646   0.062097   2.442  0.01643 *
# SATM         0.001739   0.001014   1.715  0.08952 .
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 0.7272 on 96 degrees of freedom
# Multiple R-squared:  0.3063,Adjusted R-squared:  0.2846
# F-statistic: 14.13 on 3 and 96 DF,  p-value: 1.059e-07
```

Alternatively, we can run

```
mfinal <- step(m2,scope = ~.)
```

which produces the same final model.

(d) Test whether the parameters corresponding to HSM and SATM are equal.
Solution. We test $H_0 : C\beta = \mathbf{0}$, where

$$C = \begin{bmatrix} 0 & 1 & 0 & -1 \end{bmatrix}$$

The test statistic is 9.5748, and follows an F distribution with degree of freedom 1 and 96. P-value is $0.002584 < 0.05$ so we reject this hypothesis.

```
C = c(0,1,0,-1)
dst = 0

# using R
linearHypothesis(m5,C,dst)

# Linear hypothesis test
#
# Hypothesis:
# HSM - SATM = 0
#
# Model 1: restricted model
# Model 2: GPA ~ HSM + HSE + SATM
#
#   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
# 1     97 55.828
# 2     96 50.765  1    5.0632 9.5748 0.002584 **
#   ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Alternatively

```
b_m5 = m5$coefficients
X_m5 = as.matrix(cbind(rep(1,n),gpa[,c('HSM','HSE','SATM')]))
num = t(C%*%b_m5-dst)%*%solve(C%*%solve(t(X_m5)%*%X_m5)%*%C)%*%(C%*%b_m5-dst)

(Fstar = num/(sum(m5$residuals^2)/(n-4)))
#          [,1]
# [1,] 9.574783
pf(Fstar,1,n-4,lower.tail = F)
#              [,1]
# [1,] 0.002584462
```