MATH 423/533
EXAMPLE MIDTERM 2016

Name:

ID:

Q1 (a) Write down the form of the least squares criterion for fitting a simple linear regression model, and explain how the criterion can be deduced from assumptions about the residual errors $\boldsymbol{\epsilon}$ in the formulation

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}.$$

4 Marks

(b) Show that, in the usual notation from lectures, the least squares minimization is equivalent to finding β that solves the equation

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}_2.$$

3 Marks

(c) In simple linear regression, explain why the estimate of error variance σ^2 given by

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

is most commonly used, rather than the maximum likelihood estimate which takes the form

$$\widehat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Explain how to compute the quantity $\widehat{y}_i$ that appears in these expressions. 2 Marks

(d) In a study of the impact of poverty on the standardized test scores students in a US school, data on 22 schools were collected. For each school the data comprised an average standardized math. score `MATH`, an average standardized reading score `READING`, and a measure of poverty (`POVERTY`, percentage of households below the national poverty line).

The data were analyzed in `R` and produced the following output:

```
 1 > head(Fcat) #List first six elements
 2    MATH READING POVERTY
 3 1 166.4   165.0    91.7
 4 2 159.6   157.2    90.2
 5 3 159.1   164.4    86.0
 6 4 155.5   162.4    83.9
 7 5 164.3   162.5    80.4
 8 6 169.8   164.9    76.5
 9 .  .       .        .
10 .  .       .        .
11 > summary(lm(MATH~POVERTY,data=Fcat))
12 Coefficients:
13             Estimate Std. Error t value Pr(>|t|)
14 (Intercept) 189.81582    3.02148  62.822  < 2e-16 ***
15 POVERTY      -0.30544    0.04759  -6.418 2.93e-06 ***
16 ---
17 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
18
19 Residual standard error: 5.366 on 20 degrees of freedom
20 Multiple R-squared:  0.6731,    Adjusted R-squared:  0.6568
21 F-statistic: 41.19 on 1 and 20 DF,  p-value: 2.927e-06
```

(listing continued on the next page)

```
> summary(lm(READING~POVERTY,data=Fcat))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 187.01262    1.92762  97.017  < 2e-16 ***
POVERTY      -0.27081    0.03036  -8.919 2.09e-08 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 3.423 on 20 degrees of freedom
Multiple R-squared:  0.7991,    Adjusted R-squared:  0.789
F-statistic: 79.55 on 1 and 20 DF,  p-value: 2.088e-08

> summary(lm(MATH~READING,data=Fcat))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -21.152     18.663  -1.133     0.27
READING        1.128      0.109  10.352 1.76e-09 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 3.722 on 20 degrees of freedom
Multiple R-squared:  0.8427,    Adjusted R-squared:  0.8349
F-statistic: 107.2 on 1 and 20 DF,  p-value: 1.763e-09
```

Three separate analyses have been carried out. Summarize the results of the analyses in terms of the regression relationships between the variables in each analysis, assuming that all the simple linear regression model assumptions are correct.

6 Marks

(e) From the output, compute the sample correlation coefficient for the relationship between

(i) `POVERTY` and `MATH`

(ii) `READING` and `MATH`

Note: a positive correlation between x and y arises if y ***increases*** as x increases; a negative correlation arises when y ***decreases*** as x increases. 2 Marks

(f) On the basis of the information provided, compute the three terms in the sums-of-squares decomposition

$$SS_T = SS_{Res} + SS_R$$

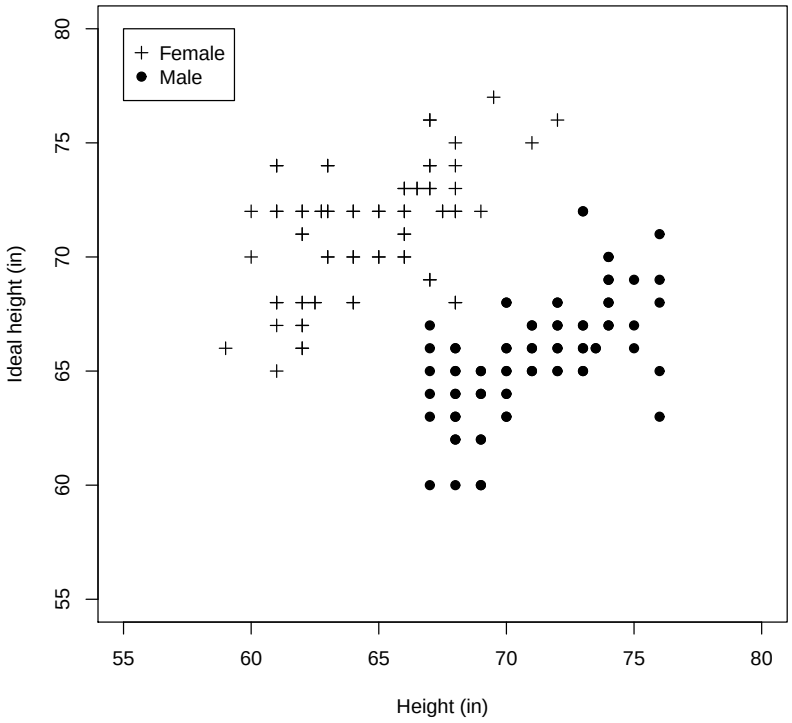
for the model regressing `MATH` on `READING`.

Explain exactly how each term is computed from the output. 3 Marks

MATH 423/533
EXAMPLE MIDTERM 2016

Name:
ID:

Q2 (a) The data in the plot below record the results of a survey of 147 US University students who were asked to select the height (in inches) of their ideal spouse or life partner. The specified value is recorded as `IdealHt`. Then the students' own height was recorded as `Height`, as well as their gender, recorded in the study using a binary indicator (`Gender`) taking the values `'F'` – recorded female – and `'M'` – recorded male. The data were stored in the `R` data frame `idh` and analyzed.




```
 1 > summary(lm(IdealHt~Height,data=idh))
 2 Coefficients:
 3             Estimate Std. Error t value Pr(>|t|)
 4 (Intercept) 86.01770    4.78629  XXXXXX  < 2e-16 ***
 5 Height      -0.26046    0.07035  -3.702 0.000303 ***
 6 ---
 7 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
 8
 9 Residual standard error: 3.643 on XXX degrees of freedom
10 Multiple R-squared:  0.08636,   Adjusted R-squared:  0.08006
11 F-statistic: 13.71 on 1 and 145 DF,  p-value: 0.0003031
12
13 > summary(lm(IdealHt~Height,subset=(Gender=='F'),data=idh)) #Females only
14 Coefficients:
15             Estimate Std. Error t value Pr(>|t|)
16 (Intercept) 39.30353    6.20615   6.333 2.00e-08 ***
17 Height       0.49261    0.09602   5.130 2.47e-06 ***
18 ---
19 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
20
21 Residual standard error: 2.322 on 70 degrees of freedom
22 Multiple R-squared:  0.2732,    Adjusted R-squared:  0.2629
23 F-statistic: 26.32 on 1 and 70 DF,  p-value: 2.474e-06
24
```

```
25 > summary(lm(IdealHt~Height,subset=(Gender=='M'),data=idh)) #Males only
26 Coefficients:
27             Estimate Std. Error t value Pr(>|t|)
28 (Intercept) 23.27364    6.33336   3.675 0.000451 ***
29 Height       0.59630    0.08902   6.698 XXXXXXXX ***
30 ---
31 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
32
33 Residual standard error: 2.057 on 73 degrees of freedom
34 Multiple R-squared:  0.3807,    Adjusted R-squared:  0.3722
35 F-statistic: 44.87 on 1 and 73 DF,  p-value: 3.758e-09
36
37 > confint(lm(IdealHt~Height,subset=(Gender=='F'),data=idh))  #Compute CIs
38                  2.5 %    97.5 %
39 (Intercept) 26.9257619 51.681305
40 Height       0.3010997  0.684121
41 > confint(lm(IdealHt~Height,subset=(Gender=='M'),data=idh))   #Compute CIs
42                  2.5 %     97.5 %
43 (Intercept) 10.6512597 35.8960136
44 Height       0.4188794  0.7737227
```

Note that the `R` command `subset=(Gender=='M')` means perform the analysis for those data for which `Gender` takes the value `M` (that is, analyze the males separately).

(i) How many males were in the study ?

(ii) There is an entry `XXXXXX` on line 4 - what is the correct numerical value to enter in that part of that coefficients table ?

(iii) What is the value `XXX` on line 10?

(iv) What is the value of the sum of squares for the residuals, SS_{Res}, in the analysis of the Females only subset ?

(v) What is the missing p-value marked `XXXXXXXX` on line 29?

5 Marks

(b) Explain how the results on lines 5, 17 and 29 should be interpreted.

Note that the `confint` results reported between lines 37 and 44 report the confidence intervals for the slope and intercept parameters for the relevant data subsets.

3 Marks

(c) On the basis of these results, what can an analyst conclude about the relationship between height of subject and selected ideal height of partner in females and males ? Justify your answer. 4 Marks

(d) On the basis of these results, predict (showing working) the selected ideal height of partner for

(i) a female of height 62 inches,

(ii) a male of height 70 inches.

4 Marks

(e) In prediction from the simple linear model, explain the difference between a *confidence interval* and a *prediction interval* for the prediction at a given $x = x^{\text{new}}$.

4 Marks