MATH 423/533
MIDTERM 2016

Name:

ID:

Q1 (a) List the modelling assumptions that are required for fitting a simple linear regression model to observed data using least squares, and describe how the least squares fit estimates are computed.

5 Marks

(b) Define the *hat matrix*, $\mathbf{H}$, used in aspects of simple linear regression, and explain its relevance to the construction of fitted values from the model.

4 Marks

(c) Show that, in the usual notation,

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

that is, $\mathrm{SS_T} = \mathrm{SS_{Res}} + \mathrm{SS_R}$, say.

5 Marks

(d) The following `R` output records the analysis of a small sample of data using simple linear regression. Some output has been removed and replaced by `XXX`.

```
 1 > summary(lm(y~x))
 2 Coefficients:
 3             Estimate Std. Error t value Pr(>|t|)
 4 (Intercept)   8.3375     1.8735   4.450 0.000467 ***
 5 x                XXX     0.1645   6.736      XXX ***
 6 ---
 7 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
 8
 9 Residual standard error: XXX on 15 degrees of freedom
10 Multiple R-squared:  0.7515,    Adjusted R-squared:  0.735
11 F-statistic: 45.37 on 1 and 15 DF,  p-value: 6.687e-06
```

It was also computed that

$$\mathrm{SS_{Res}} = 83.853 \qquad \mathrm{SS_R} = 253.612$$

From the output and the sums of squares,

(i) identify the three omitted entries on lines 5 and 9;

3 Marks

(ii) interpret line 10.

2 Marks

State clearly which pieces of output you use when giving answers.

(e) Predict the response y^{new} at a future value $x^{\text{new}} = 9.2$.

1 Mark

Use this page for extra answer or working space

Answer to Q1:

(a) We assume that

$$\mathbb{E}_{Y_i|X_i}[Y_i|\mathbf{x}_i] = \beta_0 + \beta_1 x_{i1} \qquad \mathrm{Var}_{Y_i|X_i}[Y_i|\mathbf{x}_i] = \sigma^2$$

with $Y_1, \ldots, Y_n$ presumed either uncorrelated or independent. In vector form

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \mathbf{X} \qquad \mathrm{Var}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \sigma^2 \mathbf{I}_n$$

The least squares fit is justified by considering the aggregate squared differences between the observed value y_i and fitted value under the model, $\widehat{y}_i = \beta_0 + \beta_1 x_{i1}$, to yield the function $S(\beta)$ 5 Marks

(b) The hat matrix, $\mathbf{H}$, is the matrix that linearly transforms the observed data into the fitted values. We have $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{H}\mathbf{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

4 Marks

(c) We have

$$\begin{aligned}
\sum_{i=1}^n (y_i - \overline{y})^2 &= \sum_{i=1}^n (y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2 \\
&= \sum_{i=1}^n (y_i - \widehat{y}_i)^2 + \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2 + 2\sum_{i=1}^n (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) \\
&= \sum_{i=1}^n (y_i - \widehat{y}_i)^2 + \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2
\end{aligned}$$

as the third term is zero. 5 Marks

(d) (i) On line 5: first entry is $0.1645 \times 6.736 = 1.1080$, second entry is the p-value from line 11, 6.687e-06, as the t-test and F-test are equivalent. On line 9, we have that $\widehat{\sigma} = \sqrt{\mathrm{SS}_{\mathrm{Res}}/(n-p)} = \sqrt{83.832/15} = 2.364$.

3 Marks

(ii) Line 11 is the result of the global F-test, which confirms that the continuous predictor is an influential variable in the fit, and that the hypothesis that $\beta_1 = 0$ is rejected.

2 Marks

(e) Prediction is $\widehat{y} = 8.3375 + 1.1080 \times 9.2 = 18.5311$.

1 Mark

Q1

Q2 In a physiological study of the immune system, the capability of aerobic fitness level (measured by maximal oxygen uptake, `MAXOXY`) to predict immunoglobulin levels in the blood, `IGG`, was investigated. 30 human subjects underwent a physical challenge and measurements of `MAXOXY` and `IGG` were made. An analysis in `R` is presented below:

```
> head(igg) #Print the first six entries in the data frame.
  SUBJECT  IGG MAXOXY
1       1  881   34.6
2       2 1290   45.0
3       3 2147   62.3
4       4 1909   58.9
5       5 1282   42.5
6       6 1530   44.3
> x1<-igg$MAXOXY
> y<-igg$IGG
> fit.igg<-lm(y~x1);summary(fit.igg)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -100.345    100.450  -0.999    0.326
x1            32.743      1.932  16.947 2.97e-16 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 124.8 on 28 degrees of freedom
Multiple R-squared:  0.9112,    Adjusted R-squared:  0.908
F-statistic: 287.2 on 1 and 28 DF,  p-value: 2.973e-16
```

The residuals from the straight line fit are depicted below, plotted against the predictor.