MATH 423/533 MIDTERM 2016

Name:

ID:

Q1 (a) State the conditional mean and variance assumptions about response Y_i (or response vector $\mathbf{Y}$) that characterize simple linear regression. 4 Marks

(b) Derive the equations used for estimating $\beta = (\beta_0, \beta_1)^\top$ under the *least squares* criterion, and explain how to estimate the residual error variance σ^2. 6 Marks

(c) Define the *residuals*, e_i, that arise from the fit of a simple linear regression using least squares, and show that

$$\sum_{i=1}^{n} e_i = 0.$$

4 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)   2.2776     2.8665   0.795  0.44530
5 x               XXX      0.2197   3.648  0.00448 **
6 ---
7 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
8
9 Residual standard error: 4.374 on 10 degrees of freedom
10 Multiple R-squared:  XXX,     Adjusted R-squared:  0.5281
11 F-statistic: XXX on 1 and 10 DF,  p-value: 0.004476
```

From the output identify

(i) the sample size n; 1 Mark

(ii) the `Estimate` omitted from line 5; 1 Mark

(iii) the value of the F statistic omitted from line 11; 1 Mark

(iv) whether `x` is a useful predictor of `y`; 1 Mark

(v) a 95% confidence interval for parameter β_1.

Note that the 0.975 (right tail) quantile of the relevant Student-t distribution, denoted in lectures $t_{\alpha/2,n-2}$, is 2.228. 1 Mark

(vi) the value of the R^2 statistic omitted from line 10; 1 Mark

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

Use this page for extra answer or working space

Answer to Q1:

(a) Assumptions are

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

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

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

4 Marks

(b) Least squares computes the estimates $\widehat{\beta}$ as

$$\widehat{\beta} = \arg\min_{\beta} S(\beta) = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i\beta)^2 = \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta).$$

where $\mathbf{x}_i = [1\ x_{i1}]$ and $\beta = (\beta_0, \beta_1^\top$. We have

$$\frac{\partial \S(\beta)}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \mathbf{x}_i\beta) \qquad \frac{\partial \S(\beta)}{\partial \beta_1} = -2\sum_{i=1}^{n} x_{i1}(y_i - \mathbf{x}_i\beta)$$

and equating to zero simultaneously we have

$$\sum_{i=1}^{n} \mathbf{x}_i\beta = \sum_{i=1}^{n} y_i \qquad \sum_{i=1}^{n} x_{i1}\mathbf{x}_i\beta = \sum_{i=1}^{n} x_{i1}y_i$$

which we may write concisely as

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 \end{bmatrix} \beta = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \end{bmatrix}$$

or even more concisely as

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

All of these equations are different representations of the Normal Equations.

6 Marks

(c) We have for $i = 1, \ldots, n$, the fitted values and residuals respectively as

$$\widehat{y}_i = \mathbf{x}_i\widehat{\beta} \qquad e_i = y_i - \widehat{y}_i.$$

Then

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \mathbf{x}_i\widehat{\beta}) = 0$$

as $\widehat{\beta}$ is a solution the normal equations, which, for the simple linear regression including an intercept, has this equation as the first in the system.

4 Marks

(d) The complete output is this:

```
 1 > summary(lm(y~x))
 2 Coefficients:
 3              Estimate Std. Error t value Pr(>|t|)
 4 (Intercept)    2.2776     2.8665   0.795  0.44530
 5 x              0.8016     0.2197   3.648  0.00448 **
 6 ---
 7 Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1    1
 8
 9 Residual standard error: 4.374 on 10 degrees of freedom
10 Multiple R-squared:  0.571,     Adjusted R-squared:  0.5281
11 F-statistic: 13.31 on 1 and 10 DF,  p-value: 0.004476
```

(i) the sample size n: 12 (from line 9, as $n-2=10$); 1 Mark

(ii) the `Estimate` omitted from line 5: $0.8016 = 0.2197 \times 3.648$; 1 Mark

(iii) the value of the F statistic omitted from line 11: $13.31 = 3.648^2$; 1 Mark

(iv) whether `x` is a useful predictor of `y`; yes, it is useful, the p-value on 11 reveals this. 1 Mark

(v) a 95% confidence interval for parameter β_1; the interval, as always, is

$$\widehat{\beta}_1 \pm \text{e.s.e}(\widehat{\beta}_1) \times t_{\alpha/2,n-p} \equiv 0.8016 \pm 0.2197 \times 2.228 = (0.312, 1.291)$$

as the 0.975 (right tail) quantile of the relevant Student-t distribution is 2.228. 1 Mark

(vi) the value of the R^2 statistic omitted from line 10; we have

$$R^2_{\text{Adj}} = 1 - \frac{\text{SS}_{\text{Res}}/(n-p)}{\text{SS}_{\text{T}}/(n-1)} = 1 - \frac{\text{SS}_{\text{Res}}}{\text{SS}_{\text{T}}}\frac{n-1}{n-2} = 1 - \frac{\text{SS}_{\text{Res}}}{\text{SS}_{\text{T}}}\frac{11}{10} = 0.5281$$

from line 10. Therefore

$$\frac{\text{SS}_{\text{Res}}}{\text{SS}_{\text{T}}} = \frac{n-p}{n-1}(1 - R^2_{\text{Adj}})$$

and hence

$$R^2 = 1 - \frac{\text{SS}_{\text{Res}}}{\text{SS}_{\text{T}}} = 1 - \frac{n-p}{n-1}(1 - R^2_{\text{Adj}}) = 1 - \frac{10}{11}(1 - 0.5281) = 0.571.$$

1 Mark