Qualifying Exam: CAS MA575, Linear Models

Boston University, Spring 2017

1. Consider the data `Wool`, which contains the following variables:

   `logcycles`: logarithm of the number of cycles until the specimen fails;

   `len`: length of test specimen (250, 300, 350 mm);

   `amp`: amplitude of loading cycle (8, 9, 10 mm);

   `load`: load put on the specimen (40, 45, 50 g).

   Each of the three factors (`amp`, `len` and `load`) was set to one of three levels, and all $3^3 = 27$ possible combinations of the three factors were used exactly once in the experiment. The response variable is `logcycles`, and we will treat each of the three predictors as a factor with 3 levels. The associated `R` output is given below.

```
> summary(lm(logcycles ~ len + amp + load))

Call:
lm(formula = logcycles ~ len + amp + load)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36860 -0.13002  0.00902  0.10129  0.30469

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.48287    0.09644  67.225  < 2e-16 ***
len300       0.91833    0.08928  10.286 1.97e-09 ***
len350       1.66477    0.08928  18.646 4.10e-14 ***
amp9        -0.65521    0.08928  -7.339 4.31e-07 ***
amp10       -1.26173    0.08928 -14.132 7.19e-12 ***
load45      -0.32529    0.08928  -3.643  0.00162 **
load50      -0.78524    0.08928  -8.795 2.62e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1894 on 20 degrees of freedom
Multiple R-squared:  0.9691,    Adjusted R-squared:  0.9598
F-statistic: 104.5 on 6 and 20 DF,  p-value: 4.979e-14
```

   Let $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$ denote values of the response variable `logcycles`, and let $\bar{y}_n = n^{-1} \sum_{i=1}^{n} y_i$ be the average.

(a) Based on the above information, is it possible to compute $\texttt{SYY} = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2$? If so, find its value.

(b) Based on the above information, is it possible to compute $\sum_{i=1}^{n} y_i^2$? If so, find its value.

(c) Suppose we consider the fit without an intercept. Compute the new regression summary by filling the template below. Use XXX to fill entries that you think cannot be computed from the provided information.

```
> summary(lm(logcycles ~ len + amp + load - 1))

Call:
lm(formula = logcycles ~ len + amp + load - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36860 -0.13002  0.00902  0.10129  0.30469

Coefficients:
         Estimate Std. Error    t value   Pr(>|t|)

len250  _________  _________  _________  _________

len300  _________  _________  _________  _________

len350  _________  _________  _________  _________

amp9    _________  _________  _________  _________

amp10   _________  _________  _________  _________

load45  _________  _________  _________  _________

load50  _________  _________  _________  _________
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: _________ on ____ degrees of freedom
Multiple R-squared:  0.9994,    Adjusted R-squared:  0.9991
F-statistic:  4405 on 7 and 20 DF,  p-value: < 2.2e-16
```

(d) Based on the above information including those in part (c), is it possible to compute $\texttt{SYY} = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2$? If so, find its value. Note that you only need to do this problem if you answered "No" in part (a).

(e) Based on the above information including those in part (c), is it possible to compute $\sum_{i=1}^{n} y_i^2$? If so, find its value. Note that you only need to do this problem if you answered "No" in part (b).

2. Consider the linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, where

$$\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$$

for some $\boldsymbol{X}_1 \in \mathbb{R}^{n \times p_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{n \times p_2}$. To accommodate for this block matrix form, we write

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix},$$

where $\boldsymbol{\beta}_1 \in \mathbb{R}^{p_1}$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^{p_2}$. Throughout this problem, assume that the design matrix $\boldsymbol{X}$ is deterministic with full column rank, and that the error vector $\boldsymbol{e}$ has a multivariate normal distribution with zero mean and diagonal covariance matrix with common diagonal elements σ^2. Let

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p_1+p_2}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|, \quad \hat{\boldsymbol{\beta}}_1 = \operatorname*{argmin}_{\boldsymbol{\beta}_1 \in \mathbb{R}^{p_1}} \|\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}_1\|, \quad \hat{\boldsymbol{\beta}}_2 = \operatorname*{argmin}_{\boldsymbol{\beta}_1 \in \mathbb{R}^{p_2}} \|\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2\|,$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector.

(a) Construct an example of $(\boldsymbol{Y}, \boldsymbol{X})$ where $\hat{\boldsymbol{\beta}}^\top \neq (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top)$.

(b) Prove that $\hat{\boldsymbol{\beta}}^\top = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top)$ if $\boldsymbol{X}_1^\top \boldsymbol{X}_2$ is a zero matrix in $\mathbb{R}^{p_1 \times p_2}$.

(c) Prove that $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are independent if $\boldsymbol{X}_1^\top \boldsymbol{X}_2$ is a zero matrix in $\mathbb{R}^{p_1 \times p_2}$.

(d) Prove that $\|\boldsymbol{Y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\|^2 + \|\boldsymbol{Y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\|^2 \geq \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$ if $\boldsymbol{X}_1^\top \boldsymbol{X}_2$ is a zero matrix in $\mathbb{R}^{p_1 \times p_2}$.

(e) Construct an example of $(\boldsymbol{Y}, \boldsymbol{X})$ where $\|\boldsymbol{Y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\|^2 + \|\boldsymbol{Y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$.