Qualifying Exam: CAS MA575, Linear Models

Boston University, Spring 2016

1. Consider a data set provided by the Wisconsin Department of Health and Family Services (DHFS), which involves the following variables:

   `TPY`: total patient years;

   `NUMBED`: number of beds.

   The sample size is $n = 717$. An analyst proposed to conduct a linear regression of `LOGTPY` = log(`TPY`) on `LOGNUMBED` = log(`NUMBED`), where $\log(\cdot)$ denotes the natural logarithm. The associated `R` output is given below.

```
> summary(lm(LOGTPY ~ LOGNUMBED))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.163315   0.036045  -4.531 6.88e-06 ***
LOGNUMBED    1.015739   0.008038 126.372  < 2e-16 ***
---
Residual standard error: 0.1058 on 715 degrees of freedom
Multiple R-squared:  0.9571,
F-statistic: 1.597e+04 on 1 and 715 DF
```

   (a) Provide a possible reasoning on why the logarithmic transform should be used in the linear regression.

   (b) Based on the provided information, is it possible to provide a predicted value for `TPY` when `NUMBED` = 200? If yes, find the predicted value. If no, explain. How about the associated predication interval?

   (c) Suppose you replace `LOGNUMBED` by `LOG2NUMBED` = $\log_2$(`NUMBED`), where $\log_2(\cdot)$ denotes the logarithm with base 2. 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(LOGTPY ~ LOG2NUMBED))
Coefficients:
               Estimate Std. Error    t value   Pr(>|t|)

(Intercept)  _________  _________  _________  _________

LOG2NUMBED   _________  _________  _________  _________
---
```

```
Residual standard error: _________ on ____ degrees of freedom

Multiple R-squared: _________,

F-statistic: _________ on _________ and _________ DF
```

(d) Suppose you further replace `LOGTPY` by `LOG2TPY` $= \log_2($`TPY`$)$. 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(LOG2TPY ~ LOG2NUMBED))
Coefficients:
            Estimate Std. Error    t value   Pr(>|t|)

(Intercept)  _________ _________ _________ _________

LOG2NUMBED   _________ _________ _________ _________
---

Residual standard error: _________ on ____ degrees of freedom

Multiple R-squared: _________,

F-statistic: _________ on _________ and _________ DF
```

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

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 2 \\ -1 & 1 & -2 \\ 1 & -1 & 2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \quad \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}.$$

Assume that e_i, $i = 1, \ldots, 4$, are independent normal random variables with mean zero and variance $\sigma^2 > 0$. Suppose you only observe $\boldsymbol{Y}$ and $\boldsymbol{X}$. Let $\boldsymbol{a} = (a_1, a_2, a_3)^\top$ be a column vector ($^\top$ denotes the transpose), we are interested in making statistical inference about the linear combination $\boldsymbol{a}^\top\boldsymbol{\beta} = \sum_{j=1}^{3} a_j\beta_j$.

(a) Is there any problem that you may have when computing the least squares estimate? If yes, then what causes it?

(b) Is it possible to obtain an unbiased estimate of β_2? If yes, provide the estimate. If not, explain. How about the quantity $\zeta = \beta_1 + \beta_2 + 2\beta_3$? *[Hint: The unbiased estimate here does not need to be the BLUE.]*

(c) The above system is not identifiable as there exists a nonzero vector $\boldsymbol{\gamma}$ such that $\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) + \boldsymbol{e}$ also holds. Find one such $\boldsymbol{\gamma}$.

(d) Is it possible to form a statistical test for the null hypothesis $H_0 : \beta_1 = 0$? If so, provide the test statistic and its distribution under the null. If not, explain.

(e) Is it possible to form a statistical test for the null hypothesis $H_0 : \beta_2 = 0$? If so, provide the test statistics and its distribution under the null. If not, explain.