Midterm One

Statistics 151, Fall 2013

08 October, 2013

1. Last year, 80 students took this particular course at Berkeley of whom 20 were freshmen, 20 were sophomores, 20 juniors and 20 seniors. In R, I have saved the scores for the 20 freshmen in the vector `g1`, for the 20 sophomores in `g2`, juniors in `g3` and seniors in `g4`. Consider the following output:

```
> mean(g1)
[1] 58.53768
> sd(g1)
[1] 5.024681
> mean(g2)
[1] 64.72989
> sd(g2)
[1] 4.43851
> mean(g3)
[1] 64.06235
> sd(g3)
[1] 5.264511
> mean(g4)
[1] 66.27922
> sd(g4)
[1] 4.192543
```

Let $y_1, \dots, y_n$ (for $n = 80$) denote the scores of the students. The instructor makes the assumption that these are independent and that y_i is distributed according to $N(\mu_j, \sigma^2)$ if the ith student is in the jth year.

(a) What is a good estimate of σ and why?

(b) What is the least squares estimate of $\mu_1 - \mu_2$ and what is its standard error?

(c) Give a 95% confidence interval for $\mu_1 - \mu_2$.

(d) Describe a test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ and give its p-value.

(e) Mike is currently a sophomore who is taking this year's version of this course. Give a 95% prediction interval for Mike's score in the class.

2. Consider the linear model $y_i = \beta_0 + \beta_1 x_{i1} + \dots \beta_p x_{ip} + e_i$ for $i = 1, \dots, n$ where $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I_n$. Let $\bar{x}_j := \sum_{i=1}^n x_{ij}/n$ denote the sample mean of the jth explanatory variable for $j = 1, \dots, p$.

   (a) Show that $\beta_0 + \beta_1 \bar{x}_1 + \dots \beta_p \bar{x}_p$ is estimable.

   (b) What is the least squares estimate of $\beta_0 + \beta_1 \bar{x}_1 + \dots + \beta_p \bar{x}_p$ and why?

   (c) What is the variance of the least squares estimate in (b) and how would you estimate it from the regression data?

3. For the Bodyfat dataset used in class, consider the linear model

$$\text{BODYFAT} = \beta_0 + \beta_1\text{AGE} + \beta_2\text{WEIGHT} + \beta_3\text{HEIGHT} + \beta_4\text{THIGH} + e.$$

If X denotes the X-matrix for this regression, then R tells me that $(X^TX)^{-1}$ equals

$$\begin{pmatrix} 3.740212022 & -5.908839e-03 & 6.662131e-03 & -3.218478e-02 & -4.048954e-02 \\ -0.005908839 & 3.238651e-05 & -1.222844e-05 & 3.416435e-05 & 7.148358e-05 \\ 0.006662131 & -1.222844e-05 & 2.632523e-05 & -4.483900e-05 & -1.292477e-04 \\ -0.032184784 & 3.416435e-05 & -4.483900e-05 & 3.866749e-04 & 1.944136e-04 \\ -0.040489539 & 7.148358e-05 & -1.292477e-04 & 1.944136e-04 & 7.872727e-04 \end{pmatrix}$$

The regression summary given by R is as follows:

```
Call:
lm(formula = BODYFAT ~ AGE + WEIGHT + HEIGHT + THIGH, data)

Residuals:
     Min       1Q   Median       3Q      Max
-17.3699  -3.9361  -0.0351   3.6796  16.0833

Coefficients:
            Estimate   Std. Error   t value
(Intercept) -1.07425     XXXXX       XXXXX
AGE          0.18901    0.03033      6.233
WEIGHT       0.12373     XXXXX       XXXXX
HEIGHT      -0.46074    0.10478     -4.397
THIGH        0.36546    0.14952      2.444
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.329 on XXX degrees of freedom
Multiple R-squared: XXXXX
F-statistic: 71.01 on 4 and 247 DF,  p-value: < 2.2e-16
```

(a) Fill the six missing values (indicated by XXXXX) in the above R output.

(b) Find the value of the Residual Sum of Squares (RSS) from the above output. Also, find the value of $\sum_{i=1}^{n}(y_i - \bar{y})^2/n$ where $y_1, \dots, y_n$ are the values of the response variable (BODYFAT).

(c) Suppose I add the variable KNEE to the regression. Would the RSS increase or decrease? Explain with reason. Would the Residual Standard Error increase or decrease? Explain with reason.

(d) Find the p-value for testing model 1 against model 2 where

```
Model 1: BODYFAT ~ I(AGE + WEIGHT) + I(WEIGHT + HEIGHT) + I(HEIGHT + THIGH)
Model 2: BODYFAT ~ AGE + WEIGHT + HEIGHT + THIGH
```

4. Determine whether each of following statements is true or false. Provide reasons in each case.

   (a) The errors e_i in the linear model are uncorrelated.

   (b) The residuals $\hat{e}_i$ are uncorrelated.

   (c) The sum of the residuals is zero only when there is an intercept term in the linear model.

   (d) The residual standard error is an unbiased estimator of σ.

   (e) The vector of fitted values $\hat{Y}$ belongs to the column space of the X matrix.

   (f) The vector of response values Y belongs to the column space of the X matrix.

   (g) If the errors e_i are not i.i.d normal, the least squares estimator does not necessarily have the smallest variance among all unbiased estimators.

   (h) The hat matrix is idempotent.

   (i) An archaeologist fits a regression model rejecting the null hypothesis that $\beta_2 = 0$ after getting a p-value less than 0.005. This must mean that β_2 must be large.

   (j) An archaeologist fits a regression model rejecting the null hypothesis that $\beta_2 = 0$ after getting a p-value less than 0.005. This must mean that $\hat{\beta}_2$ must be large.