Homework Three

Statistics 151a (Linear Models)

Due on 4 October 2018 at 1:59pm

September 19, 2018

1. Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ for $i = 1, \dots, n$ where $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I_n$. Suppose that the explanatory variable values $x_1, \dots, x_n$ are not all constant.

   a) Show that both parameters β_0 and β_1 are estimable. (**0.2 points**)

   b) Show that the fitted regression line passes through the point $(\bar{x}, \bar{y})$ where $\bar{x} = \sum_i x_i/n$ and $\bar{y} = \sum_i y_i/n$. (**0.2 points**)

   c) Suppose $\bar{x} = 0$. Then show that $\hat{\beta}_0$ and $\hat{\beta}_1$ (these represent the least squares estimators) are uncorrelated. (**0.25 points**)

   d) Suppose $\bar{x} = 0$ and $e_1, \dots, e_n$ are jointly normal. Show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are independent. (**0.1 point**)

2. Consider the linear model $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i$ 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 $\beta_0 + \beta_1 \bar{x}_1 + \cdots + \beta_p \bar{x}_p$ is estimable. (**0.25 points**)

   b) What is the least squares estimate of $\beta_0 + \beta_1 \bar{x}_1 + \cdots + \beta_p \bar{x}_p$ and why? (**0.25 points**)

   c) What is the variance of the least squares estimate in (b) and how would you estimate it from the regression data? (**0.25 points**)

3. Do not use R for this problem. Consider the body fat dataset that we used in class. I want to fit the model for $BODYFAT$

$$\beta_0 + \beta_1 AGE + \beta_2 WEIGHT + \beta_3 HEIGHT + \beta_4 (WEIGHT + 3 * HEIGHT) + \beta_5 WRIST + e$$

which I accomplish by the following R code resulting in the output given below:

```
> model = lm(BODYFAT ~ AGE + WEIGHT + HEIGHT + I(WEIGHT + 3*HEIGHT) + WRIST, data = body)
> summary(model)
```

```
Call:
lm(formula = BODYFAT ~ AGE + WEIGHT + HEIGHT + I(WEIGHT + 3*HEIGHT) + WRIST, data = body)

Residuals:
     Min       1Q   Median       3Q      Max
-20.5918  -3.3673  -0.0016   3.4240  12.8823

Coefficients: (1 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           47.21461    8.89363   5.309 2.46e-07 ***
AGE                    0.20629    0.02807   7.349 2.91e-12 ***
WEIGHT                 0.24341    0.01672  14.562  < 2e-16 ***
HEIGHT                -0.44389    0.09706  -4.574 7.59e-06 ***
I(WEIGHT + 3 * HEIGHT)       NA         NA      NA       NA
WRIST                 -2.73998    0.55167  -4.967 1.27e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.142 on 247 degrees of freedom
Multiple R-squared: 0.5669,Adjusted R-squared: 0.5599
F-statistic: 80.82 on 4 and 247 DF,  p-value: < 2.2e-16
```

a) Why does R produce NAs in the output? (**0.25 points**)

b) The estimate for β_2 is apparently 0.24341. Does this make sense? Explain. (**0.25 points**)

c) I decide against including the variable $WEIGHT + 3 * HEIGHT$ in the model and just intend to fit

```
Model M: BODYFAT ~ AGE + WEIGHT + HEIGHT + WRIST
```

What is the RSS for this model? Why? (**0.25 points**)

d) The model M has too many parameters for my liking; so I decide to consider the following model:

```
Model m: BODYFAT ~ AGE + WEIGHT
```

which gave me the following R output:

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

Residuals:
         Min       1Q   Median       3Q      Max
    -15.3171  -4.3293   0.2917   3.9898  18.5237


    Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
    (Intercept) -18.37392    2.57545  -7.134 1.06e-11 ***
    AGE           0.18269    0.02853   6.403 7.54e-10 ***
    WEIGHT        0.16271    0.01224  13.298  < 2e-16 ***
    ---
    Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


    Residual standard error: 5.696 on 249 degrees of freedom
    Multiple R-squared: 0.4642,Adjusted R-squared: 0.4599
    F-statistic: 107.9 on 2 and 249 DF,  p-value: < 2.2e-16
```

Find the p-value for testing the model m against the model M. If you do not have a calculator that can calculate the p-value, write the answer in terms of the F-statistic. (**0.5 points**).

4. Do not use R for this problem. 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

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

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    10.30553    -0.104
```

```
AGE           0.18901    0.03033      6.233
WEIGHT        0.12373    XXXXXX      XXXXXX
HEIGHT       -0.46074    0.10478     -4.397
THIGH         XXXXXX     0.14952      2.444
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: XXXXXX on 247 degrees of freedom
Multiple R-squared: 0.5349
F-statistic: XXXXXX on 4 and 247 DF,  p-value: < 2.2e-16
```

a) Fill the six missing values (one in the $(X^TX)^{-1}$ matrix and five in the R summary; all indicated by XXXXX) above. (**0.75 points**)

b) Based on this dataset and the above linear model, I want to predict the bodyfat percentage for a new individual who is 30 years of age, weighs 180 lbs, is 72 inches tall and who thigh circumference is 60 cm. For this consider the following output:

```
> x0 = data.frame(AGE = 30, WEIGHT = 180, HEIGHT = 72, THIGH = 60)
> predict(M, x0, interval = "confidence")
    fit        lwr       upr
  XXXXXX    14.56726    XXXXXX

> predict(M, x0, interval = "prediction")
    fit       lwr         upr
  XXXXXX    XXXXXXX     XXXXXX
```

Fill in the four missing values above. (**0.75 points**)

c) Consider the following R output for testing Model 1 against Model 2 where

```
Analysis of Variance Table

Model 1: BODYFAT ~ I(AGE + THIGH) + WEIGHT + HEIGHT
Model 2: BODYFAT ~ AGE + WEIGHT + HEIGHT + THIGH
  Res.Df    RSS   Df   Sum of Sq     F     Pr(>F)
1    XXX  XXXXX
2    XXX  XXXXX   XX    XXXXXX    1.6203   0.2042
```

Fill in the six missing values. (**0.75 points**)

5. Consider the “GPA1.Rdata” dataset that we used in class. The dataset is about GPAs and other variables

for students at Michigan State University. The response variable is *colGPA* (college GPA). The explanatory variables are *hsGPA* (high-school GPA), *ACT* (ACT score) and *skipped* (average number of lectures skipped per week).

a) I fitted a linear model for *colGPA* based on all the explanatory variables and I got the following R output:

```
Call:
lm(formula = colGPA ~ hsGPA + ACT + skipped, data = gpa)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85698 -0.23200 -0.03935  0.24816  0.81657

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.38955    0.33155   4.191 4.95e-05 ***
hsGPA        0.41182    0.09367   4.396 2.19e-05 ***
ACT          0.01472    0.01056   1.393  0.16578
skipped     -0.08311    0.02600  -3.197  0.00173 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3295 on 137 degrees of freedom
Multiple R-squared:  0.2336,Adjusted R-squared:  0.2168
F-statistic: 13.92 on 3 and 137 DF,  p-value: 5.653e-08
```

Consider the following output for testing the model m against M where:

```
Analysis of Variance Table

Model 1 (m): colGPA ~ hsGPA + ACT
Model 2 (M): colGPA ~ hsGPA + ACT + skipped

   Res.Df    RSS    Df   Sum of Sq      F     Pr(>F)
1    XXX   XXXXXX
2    XXX   XXXXXX   X     XXXXXX    XXXXXX  XXXXXXX
```

Fill in the seven missing values in the above R output giving proper reasons. (**1 points**).

b) Based on this dataset and the linear model M above, I want to predict the college GPA for a new student whose high school GPA is 3.4, ACT score is 25 and who intends to skip 2 lectures per week on average. For this, consider the following R output:

```
> x0 = data.frame(hsGPA = 3.4, ACT = 25, skipped = 2)
> predict(M, x0, interval = "confidence")
       fit        lwr       upr
   XXXXXXXX    XXXXXXX   3.064408
> predict(M, x0, interval = "prediction")
       fit        lwr       upr
  XXXXXXXX     XXXXXXXX   XXXXXXXX
> qt(0.975, 137)
  1.977431
```

Fill in the five missing values giving proper reasons (**1 points**).

6. Determine whether each of following statements is true or false. Provide reasons in each case. (**3.75 points: 0.55 points for (h); 0.16 point for every other question. No point will be awarded if no reason is provided.**)

a) There are no solutions to the normal equations $X^T X\beta = X^T Y$ when X does not have full column rank.

b) There are infinitely many solutions to the normal equations $X^T X\beta = X^T Y$ when X does not have full column rank.

c) If the ith leverage is close to one, then the ith fitted value should be close to the ith response value.

high leverage

d) Under the assumption that $Y \sim N(X\beta, \sigma^2 I)$, the sum of squares of the residuals (normalized by σ^2) follows a chi-squared distribution. RSS/sigma^2 -- Chisquare df = n-p-1

e) Under the assumption that $Y \sim N(X\beta, \sigma^2 I)$, the sum of squares of the fitted values (normalized by σ^2) follows a chi-squared distribution. Multivariate normal

f) Suppose that $\lambda^T\beta$ is estimable. Then it is possible to obtain an estimator for $\lambda^T\beta$ that has strictly smaller variance than the least squares estimate. BLUE, best is in terms of variance

g) Under the assumption that $Y \sim N(X\beta, \sigma^2 I)$ and that $\beta_1 = \cdots = \beta_p = 0$, a suitably scaled version of R^2 has the F-distribution.

does scale mean scalar multiplicaiton?

h) Under the assumption that $Y \sim N(X\beta, \sigma^2 I)$, the variance of $\hat{\sigma}^2$ equals $2\sigma^4/(n-p-1)$ where $\hat{\sigma}$ is the residual standard error. variances of RSS?

i) The p-value given by a permutation test for testing $H_0 : \beta_1 = \cdots = \beta_p$ will change every time it is implemented.

j) The $(1, 2)$th entry of the hat matrix H should always lie between 0 and 1. Yes

k) In simple linear regression (i.e., when there is only one explanatory variable), the slope of the regression line can never be larger than one. why not?

l) Again consider simple linear regression. Suppose that the response and explanatory variable values are standardized to have mean zero and unit standard deviation. Then the slope of the regression line can never be larger than one. what does it mean by standardize response and explanatory variable?
Why would we even do this? we usually assume errors are normal

m) The residual standard error always increases when explanatory variables are removed from the linear model. i think it's correct

n) Any linear function of $\beta = (\beta_0, \ldots, \beta_p)$ is estimable when the matrix $X^T X$ is invertible. yes

o) Because of the assumptions underlying the linear model, the residuals $\hat{e}_1, \ldots, \hat{e}_n$ all have the same variance. no, even without homoskedasticity assumption, we can still assume linear model
Gauss-Markov theorem requires homoskedasticity

p) If the normality assumption is violated, then the vector of residuals and the vector of fitted values may not be orthogonal. under assumption of normality of error, we can conclude a lot distributions, like Y, beta-hat.
Residual distribution e hat is multivariate normal. We can show e hat and y hat are independent
and independent implies uncorrelated

q) If the normality assumption is violated, then the vector of residuals and the vector of fitted values may not be uncorrelated.

r) If the normality assumption is violated, then the vector of residuals and the vector of fitted values may not be independent.

s) A small p-value for the F-statistic in the regression summary validates the linear model.
it tests on all betas are zero, so .

t) An archaelogist fits a regression model rejecting the hypothesis that $\beta_2 = 0$ after getting a p-value less than 0.005. This must mean that β_2 must be large.
How large?, it doesn't need to be large, as long as it's not zero(null hypothesis)

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

References