Statistics 151 (Linear Modelling -Theory and Applications)
Homework Two

Due on 02 March, 2015

20 February, 2015

1. Consider simple linear regression where there is one response variable y and one explanatory variable x and there are n subjects with values $y_1, \dots, y_n$ and $x_1, \dots, x_n$. The model is $y_i = \beta_0 + \beta_1 x_i + e_i$ where $e_1, \dots, e_n$ are independent $N(0, \sigma^2)$. Show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are independent if $x_1 + \cdots + x_n = 0$. Here, of course, $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the least squares estimates of β_0 and β_1.

2. In the Bodyfat dataset, consider the linear model:

$$\text{BODYFAT} = \beta_0 + \beta_1 \text{KNEE} + \beta_2 \text{THIGH} + \beta_3 \text{HIP} + \beta_4 \text{ANKLE} + e$$

   Assume that the errors are i.i.d normal.

   (a) Construct an F-test for testing $H_0 : \beta_1 + \beta_2 = \beta_3 + \beta_4$. Describe your method and report the value of the F-statistic, its degrees of freedom and the p-value.

   (b) Construct a t-test for testing $H_0 : \beta_1 + \beta_2 = \beta_3 + \beta_4$. Describe your method and report the value of the t-statistic, its degrees of freedom and the p-value.

   (c) How is the value of your t-test statistic related to the value of the F-test statistic?

3. In the following regression output, the value of the F-statistic (last line) and its p-value are missing. Fill them in, providing proper reasoning, based on the available information.

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9307  -2.8923  -0.3829   3.1778   9.5804

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.63014    5.95379   0.274    0.784
```

```
x1           0.85682    0.05065  16.916  < 2e-16 ***
x2          -2.02587    0.39720  -5.100 6.77e-07 ***
x3           0.04083    0.14899   0.274    0.784
x4          -0.33431    0.08191  -4.082 6.05e-05 ***
x5           0.24481    0.18236   1.342    0.181
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.122 on 246 degrees of freedom
Multiple R-squared: 0.7228,Adjusted R-squared: 0.7172
F-statistic: XXXXX on X and 246 DF,  p-value: XXXXX
```

4. In the linear model, show that the square of the (sample) correlation between the response values $(y_1, \ldots, y_n)$ and the fitted values $(\hat{y}_1, \ldots, \hat{y}_n)$ equals the coefficient of determination, R^2.

5. 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
```

The instructor wants to know if these different average scores for the four groups are caused merely by randomness or if there is really a connection between the performance ability of students and their year. Let $y_1, \ldots, y_n$ (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. She wants to test the hypothesis $H_0 : \mu_1 =$

$\mu_2 = \mu_3 = \mu_4$ against its complement H_1. Following the steps outlined below, show that this test can be carried out via the F-test that we learned for the linear model.

(a) Define four explanatory variables x_1, x_2, x_3 and x_4 in the following way: x_j takes the value $x_{ij} = 1$ for the ith subject if the ith subject is in year j; otherwise x_j takes the value $x_{ij} = 0$. Show that $y_i \sim N(\mu_i, \sigma^2)$ is equivalent to the statement that $y_i = \mu_1 x_{i1} + \mu_2 x_{i2} + \mu_3 x_{i3} + \mu_4 x_{i4} + e_i$.

(b) Calculate the RSS in this linear model.

(c) Calculate the RSS in the reduced model under the constraint $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

(d) Calculate the p-value for the F-test.

(e) Is there enough evidence in this data to reject the instructor's null hypothesis?

6. Consider the following R output:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data)

Residuals:
     Min       1Q   Median       3Q      Max
-17.3214  -3.8831  -0.0002   3.6401  16.1967

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -0.74860   10.34432  -0.072   0.9424
x1                                       0.18853    0.03039   XXXXX    XXXX
x2                                     -15.17748   32.12529   XXXXX    XXXX
x3                                      15.30167   32.12624   XXXXX    XXXX
x4                                      -0.45922    0.10500  -4.374 1.81e-05 ***
x5                                       0.35741    0.15070   2.372   0.0185 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.337 on 246 degrees of freedom
Multiple R-squared: 0.5353,Adjusted R-squared: 0.5259
F-statistic: 56.68 on 5 and 246 DF,  p-value: < 2.2e-16
```

(a) What is the p-value for the F-test for testing $H_0 : \beta_2 = 0$?

(b) What is the p-value for the F-test for testing $H_0 : \beta_3 = 0$?

(c) What is the p-value for the F-test for testing $H_0 : \beta_2 = \beta_3 = 0$? You may use information from the following R output corresponding to the same dataset as above.

```
Call:
lm(formula = y ~ x1 + x4 + x5, data)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8578  -4.0721  -0.0354   3.6837  20.1068

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.38716    7.93171  -4.083 6.00e-05 ***
x1            0.24649    0.02860   8.619 8.09e-16 ***
x4           -0.24999    0.09748  -2.564   0.0109 *
x5            0.97294    0.06839  14.227  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.534 on 248 degrees of freedom
Multiple R-squared: 0.4963,Adjusted R-squared: 0.4902
F-statistic: 81.46 on 3 and 248 DF,  p-value: < 2.2e-16
```