MATH 556 HOMEWORK 2

Due: Wednesday, 09/12/2018

Friday, September 13th, 2019

1. p.426-436. 12 (do lack-of-fit, as well as t-test and F-test for model checking) Mr. A and Ms. B, both newly employed, pack crates of material in the basement. Let x_1 and x_2 be indicator variables showing which packer is on duty and y be the number of crates packed. On 9 successive working days these data were collected:

Day, t	Mr. A, x_{1t}	Ms. B, x_{2t}	y_t
1	1	0	48
2	1	1	51
3	1	0	39
4	0	1	24
5	0	1	24
6	1	1	27
7	0	1	12
8	1	0	27
9	1	1	24

Denote by β_1 and β_2 the number of crates packed per day by A and B, respectively. First entertain the models

$$\eta = \beta_1 x_1 + \beta_2 x_2$$

Carefully stating what assumptions you make, do the following:

(a) Obtain least squares estimates of β_1 and β_2,

Answer:

```
data<-c(1,2,3,4,5,6,7,8,9,1,1,1,0,0,1,0,1,1,0,1,0,1,1,1,1,0,1,48,51,39,24,24,27,12,27,24)
dim(data)=c(9,4)
colnames(data)<-c("Day","A","B","y")
lm1<-lm(y~A+B-1,data=as.data.frame(data))
(a)
coefficients(lm1)
> coefficients(lm1)
 A  B
30 12

```

By R output, β_1 and β_2 shoule be 30 and 12. **Is the answer right ?**

(b) Compute the residuals $y_t - \hat{y}_t$.

(c) Calculate $\sum_{t=1}^{9} x_{1t}(y_t - \hat{y}_t)$ and $\sum_{t=1}^{9} x_{2t}(y_t - \hat{y}_t)$.

(d) Show an analysis of variance appropriate for contemplation of the hypothesis that A and B pack at a rate of 30 crates per day, the rate laid down by the packers' union.

(e) Split the residual sum of squares into a lack-of-fit term $S_L = \sum_{u=1}^{3} 3(\hat{y}_u - \bar{y}_u)^2$ and an "error" term $\sum_{u=1}^{3}\sum_{i=1}^{3}(y_{ui} - \bar{y}_u)^2$. *Note:* The above design consists of three distinct sets of conditions, (x_1, x_2) = (0,1), (1,0), (1,1), which are designated by the subscripts $u = 1, 2, 3$, respectively. The subscripts $i = 1, 2, 3$ are used to designate the three individual observations included in each set.

(f) Make a table of the following quantities, each having nine values: $x_1, x_2, y_{ui}, \hat{y}_u, \bar{y}_u, \hat{y}_u - \bar{y}_u$, and $y_{ui} - \bar{y}_u$. Make relevant plots. Look for a time trend in the residuals.

(g) At some point it was suggested that the presence of B might stimulate A to do more (or less) work by an amount β_{12} while the presence of A might stimulate B to do more (or less) work by an amount β_{21}. Furthermore, there might be a time trend. A new model was therefore tentatively entertained:

$$\eta = (\beta_1 + \beta_{12}x_2)x_1 + (\beta_2 + \beta_{21}x_1)x_2 + \beta_4 t$$

that is.

$$\eta = \beta_1 x_1 + \beta_2 x_2 + (\beta_{12} + \beta_{21})x_1 x_2 + \beta_4 t$$

My comments.

Part (d) In view of part (a) it seems that $E(Y|X) = 30A + 12B = 30x1 + 12x2$

```
 A  B
30 12
> y=c(48, 51, 39, 24, 24, 27, 12, 27, 24)
> x1=c(1, 1, 1, 0, 0, 1, 0, 1, 1)
> x2=c(0, 1, 0, 1, 1, 1, 1, 0, 1)
> summary(lm(y~x1+x2-1))
   Estimate Std. Error t value Pr(>|t|)
x1   30.000      6.503   4.613  0.00245 **
x2   12.000      6.503   1.845  0.10749
> summary(lm(y-30~x1+x2-1))
   Estimate Std. Error t value Pr(>|t|)
x1   10.000      5.014   1.994   0.0863 .
x2   -8.000      5.014  -1.595   0.1546
> z=lm(y-30~x1+x2)
> Z=lm(y-30~1)
> summary(z)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.000     11.225  -0.535    0.612
x1            14.000      9.165   1.528    0.177
x2            -4.000      9.165  -0.436    0.678
> summary(Z)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6667     4.2361   0.157    0.879
> anova(z,Z)
Analysis of Variance Table

Model 1: y - 30 ~ x1 + x2
Model 2: y - 30 ~ 1
  Res.Df  RSS Df Sum of Sq     F Pr(>F)
1      6  756
2      8 1292 -2      -536 2.127 0.2003
> #model y-30~1

Another way:
> m2=lm(y~x1+x2)
> m5=lm(y~1)
> anova(m5,m2)
Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1 + x2
  Res.Df  RSS Df Sum of Sq     F Pr(>F)
```

```
1       8 1292
2       6  756  2       536 2.127 0.2003
> summary(m5)
lm(formula = y ~ 1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   30.667      4.236   7.239  8.9e-05 ***


> m1=lm(y~x1+x2+offset(30*x1*x2)-1) # make no sense
> summary(m1)
   Estimate Std. Error t value Pr(>|t|)
x1     20.0       10.8   1.852    0.106
x2      2.0       10.8   0.185    0.858
> summary(lm(y-30*x1*x2~x1+x2-1)) # make no sense
   Estimate Std. Error t value Pr(>|t|)
x1     20.0       10.8   1.852    0.106
x2      2.0       10.8   0.185    0.858
> summary(lm(y-30*x1*x2~x1+x2+x1*x2-1)) # make no sense
> summary(lm(y+30*x1*x2~x1+x2+x1*x2-1)) # make no sense
```

In (d) Model Z is better than model m5. **Why ?**

Part (g), (h)

```
> t=1:9
> r= c(1, 4, 1, 2, 2, 4, 2, 1, 4)
> r=factor(r)
> summary(lm(y~r+t-1))
   Estimate Std. Error t value Pr(>|t|)
r1  51.9518     3.2112  16.178 1.64e-05 ***
r2  38.6024     3.7466  10.303 0.000148 ***
r4  53.7651     3.8903  13.820 3.56e-05 ***
t   -3.4880     0.5471  -6.375 0.001405 **
> m8=lm(y~x1+x2+t)
> anova(m8,m2) # keep m8
Model 1: y ~ x1 + x2 + t
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1      5  82.83
2      6 756.00 -1   -673.17 40.638 0.001405 **
> m3=lm(y~t)
> anova(m8,m3) # keep m8
Model 1: y ~ x1 + x2 + t
Model 2: y ~ t
  Res.Df    RSS Df Sum of Sq     F  Pr(>F)
1      5  82.83
2      7 492.65 -2   -409.82 12.37 0.01159 *
> summary(m8)
lm(formula = y ~ x1 + x2 + t)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.7892     4.5376   8.108 0.000463 ***
x1           15.1627     3.3282   4.556 0.006080 **
x2            1.8133     3.4460   0.526 0.621252
t            -3.4880     0.5471  -6.375 0.001405 **
> m7=lm(y~x1+t)
> anova(m8,m7) # keep m8
Analysis of Variance Table
Model 1: y ~ x1 + x2 + t
Model 2: y ~ x1 + t
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      5 82.825
2      6 87.412 -1   -4.5865 0.2769 0.6213
> summary(m7)
```

```
lm(formula = y ~ x1 + t)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.1961     3.4381  11.109 3.17e-05 ***
x1           14.2941     2.7103   5.274  0.00188 **
t            -3.4118     0.4948  -6.895  0.00046 ***
> qqnorm(m7$resid)
> qqline(m7$resid)
> #The model is $Y=a+bx1+ct+e$
```

The model for the data is $E(Y|X) = 38 + 14x_1 - 3.4t$.

Mr. A and Ms. B, both newly employed, pack crates of material in the basement. Let x_1 and x_2 be indicator variables showing which packer is on duty and y be the number of crates packed. On 9 successive working days these data were collected:

y=c(48, 51, 39, 24, 24, 27, 12, 27, 24)

x1=c(1, 1, 1, 0, 0, 1, 0, 1, 1)

x2=c(0, 1, 0, 1, 1, 1, 1, 0, 1)

t=1:9

r= c(1, 4, 1, 2, 2, 4, 2, 1, 4)

How to interpret the following models:

(1) lm(y~r+t-1) LSE=(52,39,53,-3)

(2) lm(y~x1+x2+t) LSE=(36,15,2,-3)

(3) lm(y-30*x1*x2~x1+x2-1) LSE=(20,2)