ADVANCED DATA ANALYSIS

HW4

Fan Yang
UNI: fy2232
02/22/2018

Problem 1

(6pt) (data in file mileage.csv) This problem is designed to review regression with categorical variables. International Oil Inc. Is attempting to a develop a reasonably priced minimum unleaded gasoline that will deliver higher gasoline mileage than can be achieved by its current premium unleaded gasolines. As part of its development process, International Oil Inc. wishes to study the effect of one qualitative variable, x_1, premium gasoline unleaded type (A, B, C) and one quantitative variable x_2 amount of gasoline additive VST (0, 1, 2, 3 units) on the gasoline mileage y obtained by an automobile called Encore. For testing purposes a sample of 22 Encores is randomly selected and driven under normal driving conditions. The combination of x_1 and x_2 used in the experiment along with the corresponding values of y are in file mileage.csv. Define $\mu_{[A,x]}, \mu_{[A,x]}$ and $\mu_{[B,x]}$ to be the mean unleaded gasoline mileage by Encore when using AST amount x and premium unleaded gasoline types A, B and C, respectively. Consider the model

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 x_2 + \epsilon$$

where $D_{1i} = 1$ gas type is B and 0 otherwise and $D_{2i} = 1$ is gas type is C and 0 otherwise.

Load the data first:

```
In [1]: mileage <- read.csv("mileage.csv", header = T)
```

(a)

(2pt) Estimate the $\beta_i s$ and interpret your result (see note for how to fit this model)

`lm(y ~ factor(x1) + x2)`

In [2]:
```
lm.1a <- lm(y ~ factor(x1) + x2, data = mileage)
summary(lm.1a)
```

```
Call:
lm(formula = y ~ factor(x1) + x2, data = mileage)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6171 -1.6321  0.5508  1.3756  4.0021

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.0171     1.0005  32.002   <2e-16 ***
factor(x1)B   1.5218     1.2650   1.203    0.245
factor(x1)C   0.5252     1.6194   0.324    0.749
x2           -0.4192     0.6042  -0.694    0.497
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.532 on 18 degrees of freedom
Multiple R-squared:  0.09453,   Adjusted R-squared:  -0.05638
F-statistic: 0.6264 on 3 and 18 DF,  p-value: 0.6072
```

From the output above, $\beta_0 = 32.0171,\ \beta_1 = 1.5218,\ \beta_2 = 0.5252,\ \beta_3 = -0.4192$. Therefore the fitted model is

$$Y_i = 32.0171 + 1.5218D_{1i} + 0.5252D_{2i} - 0.4192x_2 + \epsilon$$

(b)

(2pt) Construct a 95% confidence interval for β_1 and interpret your result

A $100(1-\alpha)\%$ confidence interval for β_i is

$$b_i \pm t_{n-p-1}(\alpha/2)SE(b_i)$$

In [3]:
```
confint(lm.1a)
```

	2.5 %	**97.5 %**
(Intercept)	29.915164	34.1189970
factor(x1)B	-1.135886	4.1795680
factor(x1)C	-2.877095	3.9274823
x2	-1.688644	0.8502126

So the 95% confidence interval for β_1 is $[-1.135886, 4.1795680]$.
Notice that this interval covers 0. Which means under confidence level $\alpha = 0.05$, we fail to reject $\beta_1 = 0$ and conclude $\beta_1 = 0$.

(c)

(2pt)Test $H_0 : \beta_1 = \beta_2 = 0$ ***against*** H_a **: *Not*** H_0 ***using*** $\alpha = 0.05$***.***

In [4]:
```
Full.1c <- lm(y ~ factor(x1) + x2, data = mileage)
Reduced.1c <- lm(y ~ x2, data = mileage)
anova(Reduced.1c,Full.1c)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
20	125.1361	NA	NA	NA	NA
18	115.4223	2	9.713798	0.7574291	0.4832412

Or we can compute by defination:

$$F = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F}$$

In [5]:
```
Fullaov.1c <- anova(Full.1c)
SSEF <- Fullaov.1c$'Sum Sq'[3]
Reducedaov.1c <- anova(Reduced.1c)
SSER <- Reducedaov.1c$'Sum Sq'[2]
(SSER - SSEF)/2/(SSEF/Full.1c$df.residual)
```

0.757429118121028

In [6]:
```
qf(0.95,2,Full.1c$df.residual)
1-pf(0.757429,2,Full.1c$df.residual)
```

3.55455714566179

0.483241236142742

The F^* statistic is 0.757429 which is less than $F(1-\alpha, df_R - df_F, df_F) = 3.55455$ with p-value 0.483241. Therefore, we fail to reject H_0 and conclude $\beta_1 = \beta_2 = 0$.

Problem 2

(5pt) In this problem we study the grain yield of rice at six seeding rates (kg/ha): The seeding rates are 25, 50, 75, 100, 125 and 150 kilograms per acre. Assume that four fields were chosen and each field was divided into 6 plots and each plot was planted at a seeding rate assigned to it at random. Besides the seeding rate, all other agricultural practices are the same. The data is

Seeding rate (kg/ha)

Filed	25	50	75	100	125	150
1	5.1	5.3	5.3	5.2	4.8	5.3
2	5.4	6.0	5.7	4.8	4.8	4.5
3	5.3	4.7	5.5	5.0	4.4	4.9
4	4.7	4.3	4.7	4.4	4.7	4.1

Fit an appropriate model to this data and test H_0 : the average yields are the same for the 6 seeding rates against the alternative H_a : There are not the same. Use $\alpha = 0.05$.

The model for the randomized block design is

$$y_{ij} = \mu + \alpha_i + b_j + \epsilon_{ij},\ i = 1, 2, \ldots k;\ j = 1, 2, \ldots b.$$

where

y_{ij} = the observation in the jth block receiving the ith treatment

μ = the over all mean

α_i = the effect of the ith treatment b_j the effect of the jth block

ϵ_{ij} = random error

In [7]:
```
response2 <- c(5.1,5.3,5.3,5.2,4.8,5.3,
               5.4,6.0,5.7,4.8,4.8,4.5,
               5.3,4.7,5.5,5.0,4.4,4.9,
               4.7,4.3,4.7,4.4,4.7,4.1)
seed.level = c(25,50,75,100,125,150)   # treatment levels
k = 6                      # number of treatment levels
n = 4                      # number of control blocks
seeding = gl(k, 1, n*k, factor(seed.level))   # matching treatment
blk2 = gl(n, k, k*n)              # blocking factor
aov.2 = aov(response2 ~ seeding + blk2)
summary(aov.2)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
seeding      5  1.267  0.2534   2.126 0.11837
blk2         3  1.965  0.6549   5.494 0.00949 **
Residuals   15  1.788  0.1192
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value is 0.11837. Under $\alpha = 0.05$ we fail to reject H_0 and conclude that the average yields are the same for the 6 seeding rates.

Problem 3

(6pt) The cutting speeds of four types of tools are being compared in an experiment. Five cutting materials of varying degree of hardness are to be used as experimental blocks. The data giving the measurement of cutting time in seconds appear in the table below

Block

Treatment	**1**	**2**	**3**	**4**	**5**
1	12	2	1	8	7
2	20	14	17	12	17
3	13	7	13	8	14
4	11	5	10	3	6

Create needed data:

In [8]:
```
response <- c(12,2,1,8,7,
              20,14,17,12,17,
              13,7,13,8,14,
              11,5,10,3,6)
response3 = c(t(matrix(response,5)))
tm3 = c(1,2,3,4)   # treatment levels
k = 4                    # number of treatment levels
n = 5                    # number of control blocks
treatment3 = gl(k, 1, n*k, factor(tm3))   # matching treatment
blk3 = gl(n, k, k*n)             # blocking factor
```

(a) (2pt) Fit an appropriate model to this data and test H_0 :The mean cutting speeds are the same for the four tools H_a : There difference. Use $\alpha = 0.05$.

The model for the randomized block design is

$$y_{ij} = \mu + \alpha_i + b_j + \epsilon_{ij},\ i = 1, 2, \ldots k;\ j = 1, 2, \ldots b.$$

where

y_{ij} = the observation in the jth block receiving the ith treatment

μ = the over all mean

α_i = the effect of the ith treatment b_j the effect of the jth block

ϵ_{ij} = random error

In [9]:
```
aov.3 = aov(response3 ~ treatment3 + blk3)
summary(aov.3)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
treatment3   3  310.0  103.33  14.850 0.000242 ***
blk3         4  124.5   31.12   4.473 0.019217 *
Residuals   12   83.5    6.96
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value is 0.000242. Under $\alpha = 0.05$ we can reject H_0 and conclude that the mean cutting speeds are different for the four tools.

(b) (4pt) Use the Bonferroni method to determine where the differences are

In [10]:
```
pairwise.t.test(response3,treatment3,pool.sd=TRUE,p.adjust.method="bonf"
 )
```

```
        Pairwise comparisons using t tests with pooled SD

data:  response3 and treatment3

  1      2      3
2 0.0028 -      -
3 0.2608 0.2608 -
4 1.0000 0.0069 0.5912

P value adjustment method: bonferroni
```

From the Bonferroni adjustment result above, we conclude that treatment 1 against treatment 2 and treatment 2 against treatment 4 are significant. That's to say, treatment 2 is quite different from treatment 1 and 4. While the rest levels of treatment are not different in our test.

Problem 4

(3pt) An experiment to investigate the effects of various dietary starch levels on milk production was conducted on four cows. The four diets, D1, D2, D3, and D4, (in order of increasing starch equivalent), were fed for three weeks to each cow and the total yield of milk in the third week of each period was recorded (i.e. third week to minimize carry-over effects due to the use of treatments administered in a previous period). That is, the trial lasted 12 weeks since each cow received each treatment, and each treatment required three weeks. The investigator felt strongly that time period effects might be important (i.e earlier periods in the experiment might influence milk yields differently compared to later periods). Hence, the investigator wanted to block on both cow and period. However, each cow cannot possibly receive more than one treatment during the same time period; that is, all possible cow-period blocking combinations could not logically be considered. It is decided to use a 4x4 latin square design and the data is

cow

Treatment	**1**	**2**	**3**	**4**
Period 1	D4(192)	D1(195)	D3(292)	D2(249)
Period 2	D1(190)	D4(203)	D2(218)	D3(210)
Period 3	D3(214)	D2(139)	D1(245)	D4(163)
Period 4	D2(221)	D3(152)	D4(204)	D1(134)

(each cell provides the treatment applied and response between the parentheses). Fit an appropriate model to this data and test H_0 : there is no difference between the four diets against H_a there is a difference.

The model for the randomized block design is

$$y_{ij} = \mu + \rho_i + \gamma_j + \tau_k + \epsilon_{ijk},\ i = 1, 2, \ldots p;\ j = 1, 2, \ldots p;\ k = 1, 2, \ldots k.$$

where

ρ_i = row effect

γ_j = column effect

τ_k = treatment effect

ϵ_{ijk} = the error term

First create needed data:

In [11]:
```
row<-c(rep("P1",4), rep("P2",4), rep("P3",4), rep("P4",4))
row
```

'P1' 'P1' 'P1' 'P1' 'P2' 'P2' 'P2' 'P2' 'P3' 'P3' 'P3' 'P3' 'P4' 'P4' 'P4' 'P4'

In [12]:
```
col<-rep(c("C1","C2","C3","C4"),4)
col
```

'C1' 'C2' 'C3' 'C4' 'C1' 'C2' 'C3' 'C4' 'C1' 'C2' 'C3' 'C4' 'C1' 'C2' 'C3' 'C4'

In [13]:
```
yield<-c(192,195,292,249,
         190,203,218,210,
         214,139,245,163,
         221,152,204,134)
trt <- c('D4','D1','D3','D2',
         'D1','D4','D2','D3',
         'D3','D2','D1','D4',
         'D2','D3','D4','D1')
```

In [14]:
```
fit4 <- lm(yield~trt+row+col)
anova(fit4)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
trt	3	1995.687	665.2292	0.5376766	0.6735933
row	3	6539.187	2179.7292	1.7617829	0.2540218
col	3	9929.187	3309.7292	2.6751141	0.1408863
Residuals	6	7423.375	1237.2292	NA	NA

From the ANOVA table above, we get the p-value is 0.6735933. Under $\alpha = 0.05$, we fail to reject H_0 and conclude there is no difference between the four diets.