Homework four

Yi Chen(yc3356)

February 19, 2018

Homework Four

problem one

```
setwd("C:/Users/cheny/Desktop/study/second term/Advanced Data Analysis/homework/homework four")
data <- read.csv("mileage.csv",header = TRUE)
```

(a)

```
fit_1 <- lm(y~factor(x1) + x2,data = data)
summary(fit_1)
```

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

analysis

1. the estimated value of $\beta_0 = 32.0171$. This means that, for premium unleaded gasoline types A, if gasoline additive VST is 0, the average of the gasoline mileage is estimated to be 32.0171.
2. the estimated value of $\beta_1 = 1.5218$. This means that, if fix gasoline additive VST to be the same, for premium unleaded gasoline types B on the average will have 1.5218 higher gasoline mileage than premium unleaded gasoline types A.

3. the estimated value of $\beta_2 = 0.5252$This means that, if fix gasoline additive VST to be the same, for premium unleaded gasoline types C on the average will have 0.5252 higher gasoline mileage than premium unleaded gasoline types A.

4. the estimated value of $\beta_3 = -0.4192$. This means that, if fix the premium unleaded gasoline types to be the same, if we increase 1 unit of gasoline additive VST, on the average the gasoline mileage will decrease 0.4192.

(b)

```
confint(fit_1,level = 0.95)
```

```
##                  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
```

analysis

Based on the data, we are 95% confident that the “true” β_1(marginal effect of premium unleaded gasoline types B compared with type A) is between -1.135886 and 4.1795680.

This also show that:for current significance level, we cannot reject the hyphothesis that $\beta_1 = 0$. There may have no difference for the marginal effect of premium unleaded gasoline types B compared and type A. Since, this confidence interval include the point 0.

(c)

```
# since here I use the factor method, thus I can use the anova directly
anova(fit_1)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq F value Pr(>F)
## factor(x1)  2   8.964  4.4819  0.6989 0.5101
## x2          1   3.087  3.0867  0.4814 0.4967
## Residuals  18 115.422  6.4123
```

analysis

As we can see from the anova table, the F-value for factor(x1) (which is related to β_1 and β_2) is 0.6989. And the p-value of this F test is 0.5101, which is bigger than $\alpha = 0.05$.

Thus, we can conclude that we fail to reject the hyphothesis that $\beta_1 = \beta_2 = 0$ for current data and significance level. And there may have no significant differenct between different premium unleaded gasoline types’ effect on the gasoline mileage.

```
# method two
fit_1_more <- lm(y~x2,data = data)
anova(fit_1_more,fit_1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x2
## Model 2: y ~ factor(x1) + x2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     20 125.14
## 2     18 115.42  2    9.7138 0.7574 0.4832
```

analysis

Since the F value is 0.7574 and the p value is 0.4832. We can say that we fail to reject that the Ho hyphothesis that the $\beta_1 = \beta_2 = 0$

problem two

```
# first input the data
average_yield <- 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)
seeding_rate <- c(rep(c(25,50,75,100,125,150),4))
field <- c(rep(1,6),rep(2,6),rep(3,6),rep(4,6))
data_2 <- as.data.frame(cbind(average_yield,seeding_rate,field))
fit_2 <- lm(average_yield~factor(seeding_rate)+factor(field),data = data_2)
anova(fit_2)
```

```
## Analysis of Variance Table
##
## Response: average_yield
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## factor(seeding_rate)  5 1.2671 0.25342  2.1261 0.118366
## factor(field)         3 1.9646 0.65486  5.4941 0.009488 **
## Residuals            15 1.7879 0.11919
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

analysis Here we can see that the factor of different fields is so called nuisance factors, which may affect the measured result but are not the primary interest.

H0: all average yields are the same for the 6 seeding rates.

As we can see from the anova table, the F-value for seed_rate is 2.1261 and the p-value for this F test is 0.118366, which is bigger than $\alpha = 0.05$.

Thus, we can conclude that we fail to reject the hyphothesis all average yields are the same for the 6 seeding rates that for current data and significance level. The average yield may be the same for the 6 different seeding rates.

problem three

(a)

```r
# first input the data
cutting_speed <- c(12,2,1,8,7,20,14,17,12,17,13,7,13,8,14,11,5,10,3,6)
block <- c(rep(c(1,2,3,4,5),4))
treatment <- c(rep(1,5),rep(2,5),rep(3,5),rep(4,5))
data_3 <- as.data.frame(cbind(cutting_speed,block,treatment))
fit_3 <- lm(cutting_speed~factor(block)+factor(treatment),data = data_3)
anova(fit_3)
```

```
## Analysis of Variance Table
##
## Response: cutting_speed
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## factor(block)      4  124.5  31.125  4.4731 0.0192167 *
## factor(treatment)  3  310.0 103.333 14.8503 0.0002421 ***
## Residuals         12   83.5   6.958
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

analysis

H0: all average cutting speed are the same for the 4 treatment.

As we can see from the anova table, the F-value for seed_rate is 4.4731 and the p-value for this F test is 0.0192167, which is smaller than $\alpha = 0.05$.

Thus, we can conclude that we reject the hyphothesis all average cutting speed are the same for the 4 treatments, for current data and significance level. The average cutting speed may be different for the 4 treatments.

(b)

```r
pairwise.t.test(data_3$cutting_speed,data_3$treatment,pool.sd=TRUE,p.adjust.method="bonf" )
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data_3$cutting_speed and data_3$treatment
##
##   1      2      3
## 2 0.0028 -      -
## 3 0.2608 0.2608 -
## 4 1.0000 0.0069 0.5912
##
## P value adjustment method: bonferroni
```

analysis

As we can see from the result, here we used the paired t test with the bonf method. The H0 hyphothesis is that for each give two treatment, the mean of them are same. Thus, for each pair of test, if the corresponding p-value if less than $\alpha = 0.05$. We can conclude that the difference between these two treatment is significantly different.

So, we can see that, we conclude that the treatment 1 differne fro the treatment 2. The treatment 2 is different from treatment 4.

problem four

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.2
```

```
# input the data
yield <- c(192,195,292,249,190,203,218,210,214,139,245,163,221,152,204,134)
treatment <- c('D4','D1','D3','D2','D1','D4','D2','D3','D3','D2','D1','D4','D2','D3','D4','D1')
cow <- rep(c('C1','C2','C3','C4'),4)
period <- c(rep('P1',4),rep('P2',4),rep('P3',4),rep('P4',4))
fit_4 <- lm(yield~treatment+cow+period)
Anova(fit_4,type = "II")
```

```
## Anova Table (Type II tests)
##
## Response: yield
##           Sum Sq Df F value Pr(>F)
## treatment 1995.7  3  0.5377 0.6736
## cow       9929.2  3  2.6751 0.1409
## period    6539.2  3  1.7618 0.2540
## Residuals 7423.4  6
```

analysis

H0: all average yield are the same for the 4 treatments.

As we can see from the anova table, the F-value for treatment is 0.0094 and the p-value for this F test is 0.92419, which is much bigger than $\alpha = 0.05$.

Thus, we can conclude that we fail to reject the hyphothesis all average cutting speed are the same for the 4 treatments, for current data and significance level. The average yield may be the same for the 4 treatments.