Homework 7

Sarah Adilijiang

Problem 1

```r
library(faraway)
data(teengamb)
str(teengamb)
```

```
## 'data.frame':    47 obs. of  5 variables:
##  $ sex   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ status: int  51 28 37 28 65 61 28 27 43 18 ...
##  $ income: num  2 2.5 2 7 2 3.47 5.5 6.42 2 6 ...
##  $ verbal: int  8 8 6 4 8 6 7 5 6 7 ...
##  $ gamble: num  0 0 0 7.3 19.6 0.1 1.45 6.6 1.7 0.1 ...
```

```r
# change the quantitative variable "sex" into a factor variable
teengamb$sex = as.factor(teengamb$sex)
str(teengamb)
```

```
## 'data.frame':    47 obs. of  5 variables:
##  $ sex   : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ status: int  51 28 37 28 65 61 28 27 43 18 ...
##  $ income: num  2 2.5 2 7 2 3.47 5.5 6.42 2 6 ...
##  $ verbal: int  8 8 6 4 8 6 7 5 6 7 ...
##  $ gamble: num  0 0 0 7.3 19.6 0.1 1.45 6.6 1.7 0.1 ...
```

```r
# remove the two-way interaction terms between "sex" and other varialbes with different sequences
model1 = lm(gamble~sex+status+income+verbal+sex:status+sex:income+sex:verbal, teengamb)
model2 = lm(gamble~sex+status+income+verbal+sex:status+sex:verbal+sex:income, teengamb)
model3 = lm(gamble~sex+status+income+verbal+sex:income+sex:status+sex:verbal, teengamb)
model4 = lm(gamble~sex+status+income+verbal+sex:income+sex:verbal+sex:status, teengamb)
model5 = lm(gamble~sex+status+income+verbal+sex:verbal+sex:income+sex:status, teengamb)
model6 = lm(gamble~sex+status+income+verbal+sex:verbal+sex:status+sex:income, teengamb)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: gamble
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## sex         1  7598.4  7598.4 17.2655 0.0001717 ***
## status      1  3613.0  3613.0  8.2096 0.0066802 **
## income      1 11898.6 11898.6 27.0367 6.657e-06 ***
## verbal      1   955.7   955.7  2.1717 0.1485994
## sex:status  1  2103.3  2103.3  4.7793 0.0348704 *
## sex:income  1  2189.5  2189.5  4.9751 0.0315396 *
## sex:verbal  1   167.4   167.4  0.3804 0.5409650
## Residuals  39 17163.5   440.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model2)
```

```
## Analysis of Variance Table
```

```
##
## Response: gamble
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## sex          1  7598.4  7598.4 17.2655 0.0001717 ***
## status       1  3613.0  3613.0  8.2096 0.0066802 **
## income       1 11898.6 11898.6 27.0367 6.657e-06 ***
## verbal       1   955.7   955.7  2.1717 0.1485994
## sex:status   1  2103.3  2103.3  4.7793 0.0348704 *
## sex:verbal   1   215.5   215.5  0.4897 0.4882132
## sex:income   1  2141.4  2141.4  4.8658 0.0333540 *
## Residuals   39 17163.5   440.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: gamble
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## sex          1  7598.4  7598.4 17.2655 0.0001717 ***
## status       1  3613.0  3613.0  8.2096 0.0066802 **
## income       1 11898.6 11898.6 27.0367 6.657e-06 ***
## verbal       1   955.7   955.7  2.1717 0.1485994
## sex:income   1  3898.9  3898.9  8.8594 0.0049886 **
## sex:status   1   393.9   393.9  0.8950 0.3499569
## sex:verbal   1   167.4   167.4  0.3804 0.5409650
## Residuals   39 17163.5   440.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model4)
```

```
## Analysis of Variance Table
##
## Response: gamble
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## sex          1  7598.4  7598.4 17.2655 0.0001717 ***
## status       1  3613.0  3613.0  8.2096 0.0066802 **
## income       1 11898.6 11898.6 27.0367 6.657e-06 ***
## verbal       1   955.7   955.7  2.1717 0.1485994
## sex:income   1  3898.9  3898.9  8.8594 0.0049886 **
## sex:verbal   1   379.6   379.6  0.8626 0.3587294
## sex:status   1   181.7   181.7  0.4128 0.5243068
## Residuals   39 17163.5   440.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model5)
```

```
## Analysis of Variance Table
##
## Response: gamble
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## sex          1  7598.4  7598.4 17.2655 0.0001717 ***
## status       1  3613.0  3613.0  8.2096 0.0066802 **
```

```
## income      1 11898.6 11898.6 27.0367 6.657e-06 ***
## verbal      1   955.7   955.7  2.1717 0.1485994
## sex:verbal  1  1087.3  1087.3  2.4705 0.1240773
## sex:income  1  3191.3  3191.3  7.2514 0.0103875 *
## sex:status  1   181.7   181.7  0.4128 0.5243068
## Residuals  39 17163.5   440.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model6)
```

```
## Analysis of Variance Table
##
## Response: gamble
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## sex         1  7598.4  7598.4 17.2655 0.0001717 ***
## status      1  3613.0  3613.0  8.2096 0.0066802 **
## income      1 11898.6 11898.6 27.0367 6.657e-06 ***
## verbal      1   955.7   955.7  2.1717 0.1485994
## sex:verbal  1  1087.3  1087.3  2.4705 0.1240773
## sex:status  1  1231.6  1231.6  2.7985 0.1023588
## sex:income  1  2141.4  2141.4  4.8658 0.0333540 *
## Residuals  39 17163.5   440.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

Here all the ANOVA tests show that "sex:income" is a significant interaction term and should be added into the model. And "sex:status" is sometimes significant and sometimes not. So we can compare between the model adding only "sex:income" and the model adding both of them. The result is shown in the results of anova(model3), where the "sex:income" is added first and the "sex:status" is added next. The p-value of F-test is 0.3499569, so we do not reject the reduced model and pick the final model:

$$gamble = \beta_0 + \beta_{sex1} sex1 + \beta_{status} status + \beta_{income} income + \beta_{verbal} verbal + \beta_{sex1:income} sex1 : income + noise$$

```r
model = lm(gamble~sex+status+income+verbal+sex:income, teengamb)
summary(model)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal + sex:income,
##     data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.109  -6.162  -0.938   2.267  86.503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.25943   15.79635   1.219  0.22972
## sex1         4.06362   11.51612   0.353  0.72600
## status      -0.04876    0.25978  -0.188  0.85203
## income       6.19885    1.02591   6.042 3.77e-07 ***
## verbal      -2.60864    1.99386  -1.308  0.19805
## sex1:income -6.43683    2.14337  -3.003  0.00454 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.79 on 41 degrees of freedom
## Multiple R-squared:  0.6121, Adjusted R-squared:  0.5647
## F-statistic: 12.94 on 5 and 41 DF,  p-value: 1.417e-07
```

Answer:

Here "male" (sex=0) is the reference level, β_{sex1} = 4.06362, β_{income} = 6.19885, $\beta_{sex1:income}$ = -6.43683.

So when "status" and "verbal" are the same:

(1) for male, the average change of "gamble" (expenditure on gambling in pounds per year) is 6.19885 pounds when there is an additional increase of "income" (in pounds per week);

(2) for female, the average change of "gamble" (expenditure on gambling in pounds per year) is 6.19885-6.43683 = -0.23798 pounds when there is an additional increase of "income" (in pounds per week).

Problem 2

```r
library(lattice)
```

```
##
## Attaching package: 'lattice'

## The following object is masked from 'package:faraway':
##
##     melanoma
```

```r
data(barley)
str(barley)
```

```
## 'data.frame':    120 obs. of  4 variables:
##  $ yield  : num  27 48.9 27.4 39.9 33 ...
##  $ variety: Factor w/ 10 levels "Svansota","No. 462",..: 3 3 3 3 3 3 7 7 7 7 ...
##  $ year   : Factor w/ 2 levels "1932","1931": 2 2 2 2 2 2 2 2 2 2 ...
##  $ site   : Factor w/ 6 levels "Grand Rapids",..: 3 6 4 5 1 2 3 6 4 5 ...
```

(a)

Answer:

There are $10 \times 2 \times 6 = 120$ possible combinations of "variety", "year", and "site" (including reference levels), thus 120 degrees of freedom would be used by the model with all interactions.

Since number of observations n=120, which is equal to the degrees of freedom used by the model, thus we will not be able to do significance testing on this full model (n-p=0).

(b)

Answer:

There are $10 \times 2 \times 6 - 9 \times 1 \times 5 = 120 - 45 = 75$ degrees of freedom would be used by the model with all factors and two-way interactions, but not three-way interactions.

Since number of observations n=120 > df=75, thus we now will be able to do significance testing on this reduced model.

(c)

First, we try to remove different two-way interaction terms first.

```r
barley2 = barley[-c(23,83), ]

# first, try to remove different two-way interaction terms first
model1 = lm(yield~(variety+site+year)**2, barley2)
model2 = lm(yield~(variety+year+site)**2, barley2)
model3 = lm(yield~(site+year+variety)**2, barley2)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: yield
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## variety       9 1029.6  114.40   9.8935 4.271e-08 ***
## site          5 6607.1 1321.43 114.2814 < 2.2e-16 ***
## year          1  912.1  912.10  78.8815 2.271e-11 ***
## variety:site 44 1161.8   26.40   2.2835  0.003615 **
## variety:year  9  189.9   21.10   1.8244  0.090593 .
## site:year     5 2164.7  432.94  37.4421 8.767e-15 ***
## Residuals    44  508.8   11.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: yield
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## variety       9 1029.6  114.40   9.8935 4.271e-08 ***
## year          1  912.1  912.10  78.8815 2.271e-11 ***
## site          5 6607.1 1321.43 114.2814 < 2.2e-16 ***
## variety:year  9  189.9   21.10   1.8244  0.090593 .
## variety:site 44 1161.8   26.40   2.2835  0.003615 **
## year:site     5 2164.7  432.94  37.4421 8.767e-15 ***
## Residuals    44  508.8   11.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: yield
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## site          5 6556.4 1311.28 113.4036 < 2.2e-16 ***
## year          1  912.1  912.10  78.8815 2.271e-11 ***
## variety       9 1080.3  120.04  10.3812 2.201e-08 ***
## site:year     5 2164.1  432.83  37.4323 8.807e-15 ***
## site:variety 44 1161.8   26.40   2.2835  0.003615 **
## year:variety  9  190.4   21.16   1.8298  0.089547 .
## Residuals    44  508.8   11.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

Here the ANOVA tests show that only "year:variety" is not a significant interaction term at 0.05 significance level (F test p-value = 0.089547 > 0.05), thus it can be removed from the model.

Then we try to remove the other two interaction terms with different sequence.

```r
# Then, try to remove other two interaction terms with different sequence
model1 = lm(yield~variety+site+year+site:variety+site:year, barley2)
model2 = lm(yield~variety+site+year+site:year+site:variety, barley2)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: yield
##               Df Sum Sq Mean Sq  F value    Pr(>F)
## variety        9 1029.6  114.40   8.6716 7.427e-08 ***
## site           5 6607.1 1321.43 100.1663 < 2.2e-16 ***
## year           1  912.1  912.10  69.1387 3.525e-11 ***
## variety:site 44 1161.8   26.40   2.0015  0.008104 **
## site:year      5 2164.1  432.83  32.8090 4.377e-15 ***
## Residuals     53  699.2   13.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: yield
##               Df Sum Sq Mean Sq  F value    Pr(>F)
## variety        9 1029.6  114.40   8.6716 7.427e-08 ***
## site           5 6607.1 1321.43 100.1663 < 2.2e-16 ***
## year           1  912.1  912.10  69.1387 3.525e-11 ***
## site:year      5 2164.1  432.83  32.8090 4.377e-15 ***
## variety:site 44 1161.8   26.40   2.0015  0.008104 **
## Residuals     53  699.2   13.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

Here the ANOVA tests show that there is no evidence that "site:year" and "site:variety" should be removed from the model at 0.05 significance level (F test p-value = 4.377e-15 and 0.008104, both < 0.05), thus we keep them in the model. Therefore, the final reduced model is:

$$lm(yield \sim variety + site + year + site : variety + site : year)$$

Problem 3

(a)

```r
library(faraway)
data(pulp)
str(pulp)
```

```
## 'data.frame':    20 obs. of  2 variables:
##  $ bright  : num  59.8 60 60.8 60.8 59.8 59.8 60.2 60.4 59.9 60 ...
##  $ operator: Factor w/ 4 levels "a","b","c","d": 1 1 1 1 1 2 2 2 2 2 ...
```

```r
# calculate sample means of "bright" in each group of "operator"
means = tapply(X = pulp$bright, INDEX = pulp$operator, FUN = mean)
means
```

```
##     a     b     c     d
## 60.24 60.06 60.62 60.68
```

Answer:

$\hat{\alpha_A} = 60.24, \hat{\alpha_B} = 60.06, \hat{\alpha_C} = 60.62, \hat{\alpha_D} = 60.68$

(b)

```r
Levels = levels(pulp$operator)
RSS_bygroup = NULL
for (i in 1:length(Levels)) {
    RSS_bygroup[i] = sum((pulp$bright[pulp$operator==Levels[i]] - means[Levels[i]])^2)
}
RSS = sum(RSS_bygroup)
df = nrow(pulp) - length(Levels)
sigma_hat = sqrt(RSS/df);    sigma_hat
```

```
## [1] 0.3259601
```

Answer:

$\hat{\sigma} = 0.3259601$

(c)

If σ were known, we have $\sqrt{Var(\hat{\alpha_A} - \hat{\alpha_B})} = \sigma * \sqrt{1/5 + 1/5} = \sigma * \sqrt{2/5}$, so $SE(\hat{\alpha_A} - \hat{\alpha_B}) = \hat{\sigma} * \sqrt{2/5}$, and its value is calculated as shown below.

```r
SE_pair = sigma_hat * sqrt(2/5); SE_pair
```

```
## [1] 0.2061553
```

Answer:

Therefore, same as above, we have $SE(\hat{\alpha_A} - \hat{\alpha_B}) = SE(\hat{\alpha_A} - \hat{\alpha_C}) = SE(\hat{\alpha_A} - \hat{\alpha_D}) = SE(\hat{\alpha_B} - \hat{\alpha_C}) = SE(\hat{\alpha_B} - \hat{\alpha_D}) = SE(\hat{\alpha_C} - \hat{\alpha_D}) = \hat{\sigma} * \sqrt{2/5} = 0.2061553$

(d)

```r
L = length(Levels)
q = qtukey(0.95, L, nrow(pulp)-L)

# 95% CIs for each pair comparison
CIs = data.frame("diff"=rep(0,12), "lwr"=rep(0,12), "upr"=rep(0,12))
for (i in 1:(L-1)) {
    for (j in 1:L) {
        CIs$diff[(i-1)*4+j] = means[Levels[j]] - means[Levels[i]]
        CIs$lwr[(i-1)*4+j] = CIs$diff[(i-1)*4+j] - q/sqrt(2) * SE_pair
        CIs$upr[(i-1)*4+j] = CIs$diff[(i-1)*4+j] + q/sqrt(2) * SE_pair
        rownames(CIs)[(i-1)*4+j] =paste0(Levels[j],"-",Levels[i])
    }
}
CIs[-c(1,5,6,9,10,11), ]
```

```
##      diff         lwr       upr
## b-a -0.18 -0.76981435 0.4098143
## c-a  0.38 -0.20981435 0.9698143
## d-a  0.44 -0.14981435 1.0298143
## c-b  0.56 -0.02981435 1.1498143
## d-b  0.62  0.03018565 1.2098143
## d-c  0.06 -0.52981435 0.6498143
```

```r
# or using function TukeyHSD()
model = lm(bright~operator, pulp)
TukeyHSD(aov(model))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = model)
##
## $operator
##      diff         lwr       upr     p adj
## b-a -0.18 -0.76981435 0.4098143 0.8185430
## c-a  0.38 -0.20981435 0.9698143 0.2903038
## d-a  0.44 -0.14981435 1.0298143 0.1844794
## c-b  0.56 -0.02981435 1.1498143 0.0657945
## d-b  0.62  0.03018565 1.2098143 0.0376691
## d-c  0.06 -0.52981435 0.6498143 0.9910783
```

Answer:

Therefore, the 95% Tukey HSD confidence intervals of $\alpha_B - \alpha_A$, $\alpha_C - \alpha_A$, $\alpha_D - \alpha_A$, $\alpha_C - \alpha_B$, $\alpha_D - \alpha_C$ all cover zero, which means that there are no significant differences between these pairs of production methods at 0.05 significance level.

However, the 95% Tukey HSD confidence interval of $\alpha_D - \alpha_B$ does not cover zero and is greater than zero, which means that the brightness is significantly higher for production method D than for B at 0.05 significance level.

Problem 4

see next page

STAT34300 HW7 Sarah Adilijiang

Problem 4:

Full model:

$$Y_i = \beta_0 + \beta_{A1}\cdot 1_{A_i=1} + \beta_{B1}\cdot 1_{B_i=1} + \beta_{C1}\cdot 1_{C_i=1}$$
$$+ \beta_{A1:B1}\cdot 1_{A_i=1\&B_i=1} + \beta_{A1:C1}\cdot 1_{A_i=1\&C_i=1} + \beta_{B1:C1}\cdot 1_{B_i=1\&C_i=1}$$
$$+ \beta_{A1:B1:C1}\cdot 1_{A_i=1\&B_i=1\&C_i=1} + \text{noise}$$

① without any medication, $E(Y_i)=150$

i.e. when $A_i=B_i=C_i=0$, $E(Y_i)=\beta_0=150 \Rightarrow \beta_0=150$

② any one drug on its own has no effect

i.e. $\begin{cases} \text{when } A_i=1, B_i=C_i=0, & E(Y_i)=\beta_0+\beta_{A1}=\beta_0 \Rightarrow \beta_{A1}=0 \\ \text{when } B_i=1, A_i=C_i=0, & E(Y_i)=\beta_0+\beta_{B1}=\beta_0 \Rightarrow \beta_{B1}=0 \\ \text{when } C_i=1, A_i=B_i=0, & E(Y_i)=\beta_0+\beta_{C1}=\beta_0 \Rightarrow \beta_{C1}=0 \end{cases}$

③ drug A in combination with B or C will reduce blood pressure to 140, and it doesn't matter which one is used in combination with drug A

i.e. $\begin{cases} \text{when } A_i=B_i=1, C_i=0, & E(Y_i)=\beta_0+\beta_{A1:B1}=140 \Rightarrow \beta_{A1:B1}=140-150=-10 \\ \text{when } A_i=C_i=1, B_i=0, & E(Y_i)=\beta_0+\beta_{A1:C1}=140 \Rightarrow \beta_{A1:C1}=140-150=-10 \end{cases}$

④ There's no effect of using both B and C – it's equivalent to just using one.

i.e. when $A_i=1$: $\begin{cases} B_i=1, C_i=0 \text{ or } B_i=0, C_i=1, \ E(Y_i)=140 & ① \\ B_i=C_i=1, \ E(Y_i)=\beta_0+\beta_{A1:B1}+\beta_{A1:C1}+\beta_{B1:C1}+\beta_{A1:B1:C1}=130+\beta_{B1:C1}+\beta_{A1:B1:C1} & ② \end{cases}$

$① = ② \Rightarrow \beta_{B1:C1}+\beta_{A1:B1:C1}=10$

when $A_i=0$: $\begin{cases} B_i=1, C_i=0 \text{ or } B_i=0, C_i=1, \ E(Y_i)=\beta_0=150 & ③ \\ B_i=C_i=1, \ E(Y_i)=\beta_0+\beta_{B1:C1}=150+\beta_{B1:C1} & ④ \end{cases}$

$③ = ④ \Rightarrow \beta_{B1:C1}=0$

$\Rightarrow \beta_{A1:B1:C1}=10-\beta_{B1:C1}=10$

$\Rightarrow$ Results: $\begin{cases} \beta_0=150 \\ \beta_{A1}=\beta_{B1}=\beta_{C1}=0 \\ \beta_{A1:B1}=\beta_{A1:C1}=-10, \ \beta_{B1:C1}=0 \\ \beta_{A1:B1:C1}=10 \end{cases}$

Scanned with CamScanner