Recommended Exercise 8 in Statistical Linear Models, Spring 2021

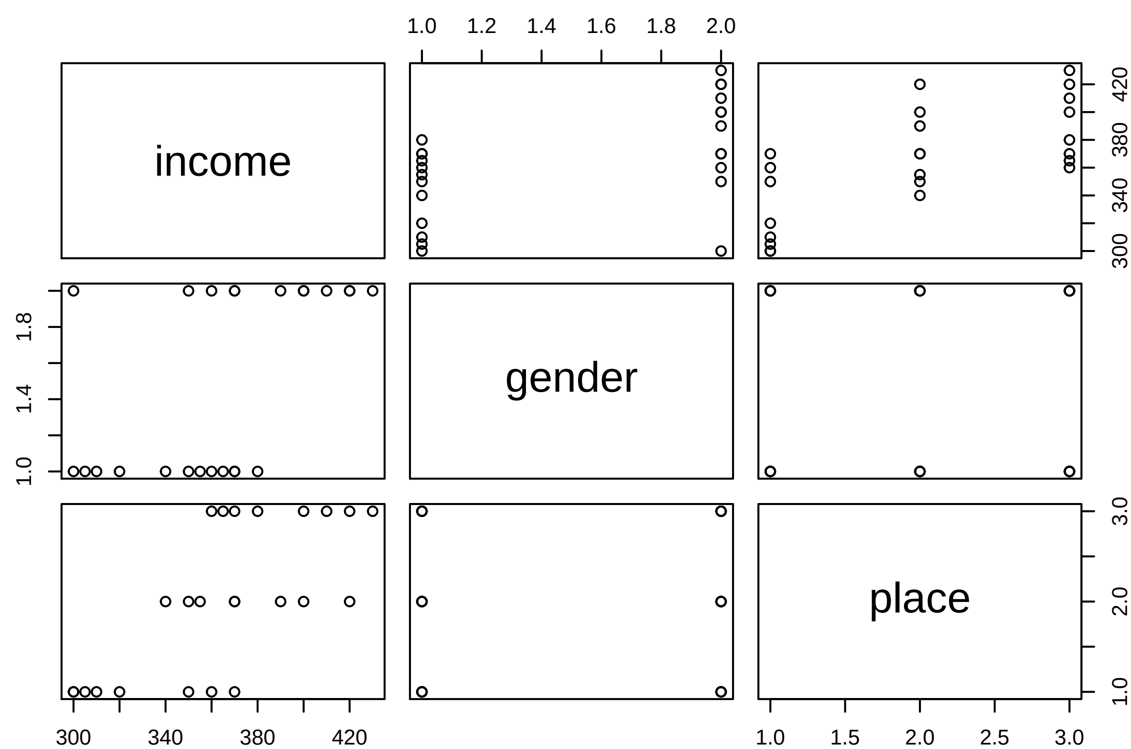
alexaoh

13 mai, 2021

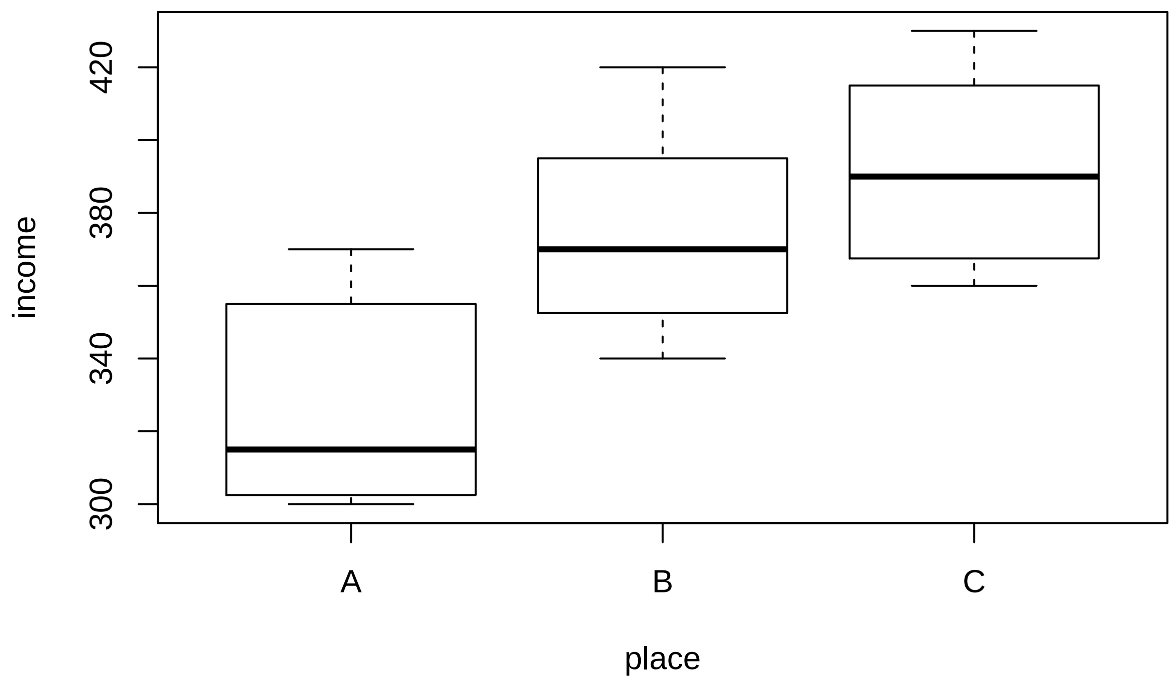
Problem 1 One- and two-way ANOVA - and the linear model

a)

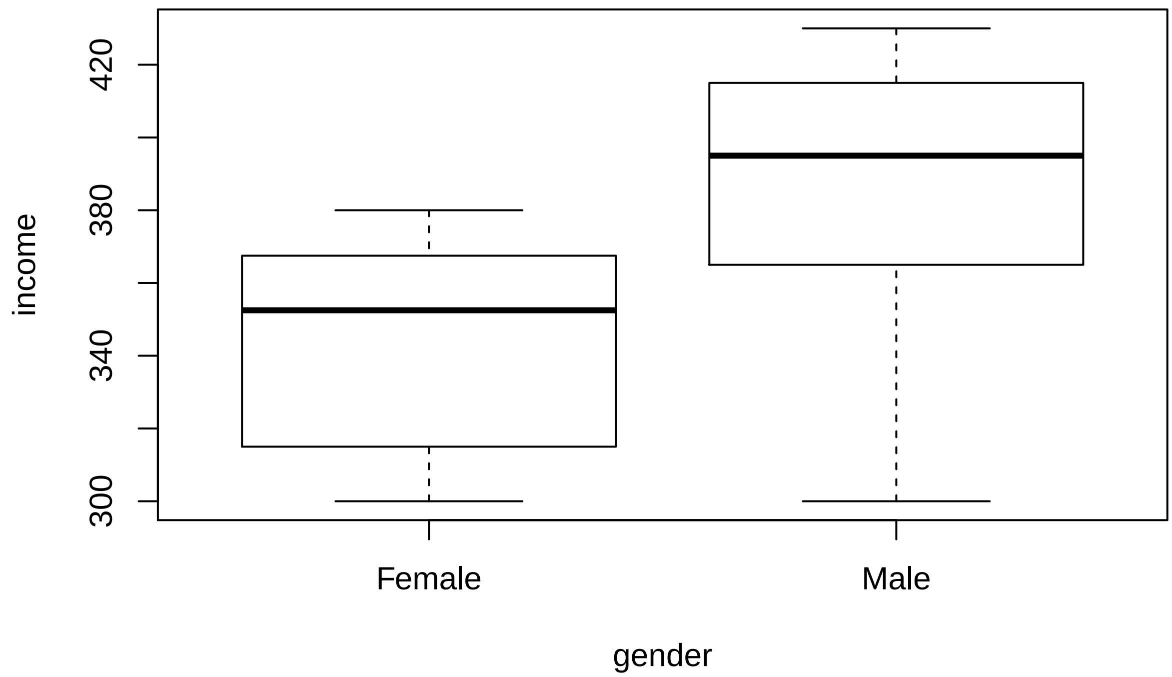
```
income <- c(300, 350, 370, 360, 400, 370, 420,
            390,400, 430, 420, 410, 300, 320, 310,
            305,350, 370, 340, 355, 370, 380, 360, 365)
gender <- c(rep("Male", 12), rep("Female",12))
place <- rep(c(rep("A",4), rep("B",4), rep("C",4)),2)
data <- data.frame(income, gender, place)

pairs(data)
```



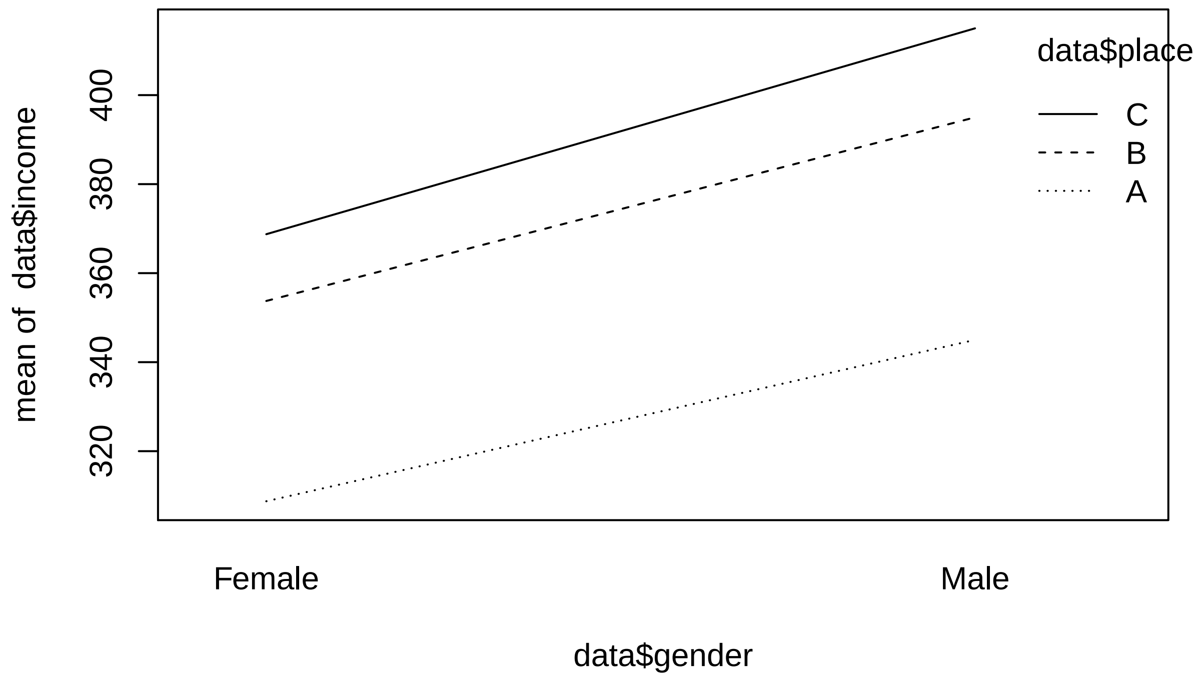
```r
plot(income~place, data=data)
```



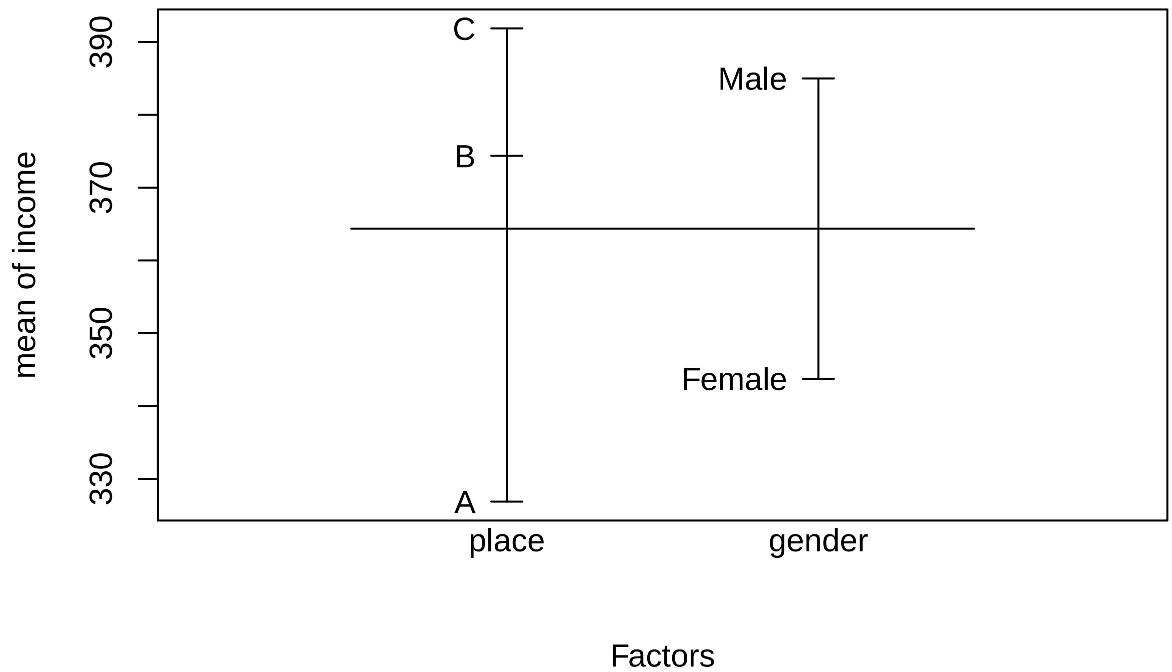

```r
plot(income~gender, data=data)
```




```
interaction.plot(data$gender, data$place, data$income)
```




```
plot.design(income~place+gender, data=data)
```



b)

```
X <- cbind(rep(1,length(data$income)), data$place=="A", data$place=="B",data$place=="C")
XTX <- t(X) %*% X
qr(XTX)$rank
```

```
#> [1] 3
```

The rank of $X^T X$ is 3. We need it to have full rank in order to be able to estimate the coefficients in the model. Problems with non-full rank can be solved by different encodings of the coefficients, e.g. dummy coding, which is standard in R.

c)

```
model <- lm(income~place-1, data=data, x = T)
summary(model)
```

```
#>
#> Call:
#> lm(formula = income ~ place - 1, data = data, x = T)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -34.375 -22.500  -5.625  23.750  45.625
#>
#> Coefficients:
```

```
#>         Estimate Std. Error t value Pr(>|t|)
#> placeA  326.875      9.733   33.58   <2e-16 ***
#> placeB  374.375      9.733   38.46   <2e-16 ***
#> placeC  391.875      9.733   40.26   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 27.53 on 21 degrees of freedom
#> Multiple R-squared:  0.9951, Adjusted R-squared:  0.9944
#> F-statistic:  1409 on 3 and 21 DF,  p-value: < 2.2e-16
```

```r
anova(model)
```

```
#> Analysis of Variance Table
#>
#> Response: income
#>           Df  Sum Sq Mean Sq F value    Pr(>F)
#> place      3 3204559 1068186  1409.4 < 2.2e-16 ***
#> Residuals 21   15916     758
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The intercept is removed, which gives the parametrization in this case. This means that each coefficient estimate includes the mean, i.e. the model has no mean (it is set to zero). The null hypothesis tested in `anova` is $\alpha_A = \alpha_B = \alpha_c = 0$. The result is that the null hypothesis is discarded, because the p-value is significant, which means that the model has some merit.

d)

```r
options(contrasts=c("contr.treatment", "contr.poly"))
model1 <- lm(income~place, data=data, x=TRUE)
summary(model1)
```

```
#>
#> Call:
#> lm(formula = income ~ place, data = data, x = TRUE)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -34.375 -22.500  -5.625  23.750  45.625
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  326.875      9.733  33.583  < 2e-16 ***
#> placeB        47.500     13.765   3.451 0.002394 **
#> placeC        65.000     13.765   4.722 0.000116 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 27.53 on 21 degrees of freedom
#> Multiple R-squared:  0.5321, Adjusted R-squared:  0.4875
#> F-statistic: 11.94 on 2 and 21 DF,  p-value: 0.000344
```

```r
anova(model1)
```

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

```
#>
#> Response: income
#>           Df Sum Sq Mean Sq F value   Pr(>F)
#> place      2  18100  9050.0  11.941 0.000344 ***
#> Residuals 21  15916   757.9
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
options(contrasts=c("contr.sum", "contr.poly"))
model2 <- lm(income~place, data=data, x=TRUE)
summary(model2)
```

```
#>
#> Call:
#> lm(formula = income ~ place, data = data, x = TRUE)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -34.375 -22.500  -5.625  23.750  45.625
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  364.375      5.619  64.841  < 2e-16 ***
#> place1       -37.500      7.947  -4.719 0.000117 ***
#> place2        10.000      7.947   1.258 0.222090
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 27.53 on 21 degrees of freedom
#> Multiple R-squared:  0.5321, Adjusted R-squared:  0.4875
#> F-statistic: 11.94 on 2 and 21 DF,  p-value: 0.000344
```

```
anova(model2)
```

```
#> Analysis of Variance Table
#>
#> Response: income
#>           Df Sum Sq Mean Sq F value   Pr(>F)
#> place      2  18100  9050.0  11.941 0.000344 ***
#> Residuals 21  15916   757.9
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When using `contr.treatment` the regular "dummy coding" is used, i.e. placeA is dropped/merged with the intercept. Thus the coefficient estimate for placeA is found in the intercept, while the estimates for placeB and placeC are found by adding the estimates from the model to the intercept, respectively. In essence, placeA is used as a baseline.

When using `contr.sum` the "zero-sum" or "effect coding" is used. This means that, in order to retrieve the estimate for placeA, the coefficient called place1 is added to the intercept, while, similarly, the estimate for placeB is retrieved by adding the coefficient called place2 to the intercept. The estimate for placeC can be retrieved by computing the intercept minus the other two coefficients (place1 and place2).

e)

```r
# Model 1
r <- 2
C <- cbind(rep(0,r), diag(r))
C
```

```
#>      [,1] [,2] [,3]
#> [1,]    0    1    0
#> [2,]    0    0    1
```

```r
d <- matrix(rep(0,r), ncol=1)
n <- length (data$income)

betahat <- matrix(model1$coefficients, ncol=1)
sigma2hat <- summary(model1)$sigma^2
X <- model.matrix(model1)
F1 <- (t(C%*%betahat-d)%*%solve(C%*%solve(t(X)%*%X)%*%t(C))%*%(C%*%betahat-d))/(r*sigma2hat)
F1
```

```
#>         [,1]
#> [1,] 11.9411
```

```r
1-pf(F1,r,n-length(betahat))
```

```
#>              [,1]
#> [1,] 0.0003439736
```

```r
# Model 2
betahat2 <- matrix(model2$coefficients, ncol=1)
sigma2hat2 <- summary(model2)$sigma^2
X2 <- model.matrix(model2)
F2 <- (t(C%*%betahat2-d)%*%solve(C%*%solve(t(X2)%*%X2)%*%t(C))%*%(C%*%betahat2-d))/(r*sigma2hat2)
F2
```

```
#>         [,1]
#> [1,] 11.9411
```

```r
1-pf(F2,r,n-length(betahat))
```

```
#>              [,1]
#> [1,] 0.0003439736
```

We can see that the test for both models gives the same result!

f)

```r
options(contrasts=c("contr.treatment", "contr.poly"))
model3 <- lm(income~place+gender, data=data, x=TRUE)
anova(model3)
```

```
#> Analysis of Variance Table
#>
#> Response: income
#>           Df  Sum Sq Mean Sq F value    Pr(>F)
#> place      2 18100.0  9050.0  31.720 6.260e-07 ***
#> gender     1 10209.4 10209.4  35.783 7.537e-06 ***
#> Residuals 20  5706.2   285.3
```

```
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(model3)
```

```
#>
#> Call:
#> lm(formula = income ~ place + gender, data = data, x = TRUE)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -47.500  -6.250   0.000   9.687  25.000
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  306.250      6.896  44.411  < 2e-16 ***
#> placeB        47.500      8.446   5.624 1.67e-05 ***
#> placeC        65.000      8.446   7.696 2.11e-07 ***
#> genderMale    41.250      6.896   5.982 7.54e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 16.89 on 20 degrees of freedom
#> Multiple R-squared:  0.8322, Adjusted R-squared:  0.8071
#> F-statistic: 33.07 on 3 and 20 DF,  p-value: 6.012e-08
```

```r
options(contrasts=c("contr.sum", "contr.poly"))
model4 <- lm(income~place+gender, data=data, x=TRUE)
summary(model4)
```

```
#>
#> Call:
#> lm(formula = income ~ place + gender, data = data, x = TRUE)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -47.500  -6.250   0.000   9.687  25.000
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  364.375      3.448 105.680  < 2e-16 ***
#> place1       -37.500      4.876  -7.691 2.13e-07 ***
#> place2        10.000      4.876   2.051   0.0536 .
#> gender1      -20.625      3.448  -5.982 7.54e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 16.89 on 20 degrees of freedom
#> Multiple R-squared:  0.8322, Adjusted R-squared:  0.8071
#> F-statistic: 33.07 on 3 and 20 DF,  p-value: 6.012e-08
```

```r
anova(model4)
```

```
#> Analysis of Variance Table
#>
#> Response: income
```

```
#>           Df  Sum Sq Mean Sq F value    Pr(>F)
#> place      2 18100.0  9050.0  31.720 6.260e-07 ***
#> gender     1 10209.4 10209.4  35.783 7.537e-06 ***
#> Residuals 20  5706.2   285.3
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As before, the first model uses the "dummy coding", while the second model uses the "effect coding". Also, the ANOVA-tables look the same for both models, as expected based on the One-way ANOVA from earlier.

```r
interaction.model <- lm(income~place*gender, data = data, x = TRUE)
summary(interaction.model)
```

```
#>
#> Call:
#> lm(formula = income ~ place * gender, data = data, x = TRUE)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -45.000  -5.938   1.250  11.250  25.000
#>
#> Coefficients:
#>                    Estimate Std. Error t value Pr(>|t|)
#> (Intercept)         308.750      8.824  34.989  < 2e-16 ***
#> placeB               45.000     12.479   3.606 0.002020 **
#> placeC               60.000     12.479   4.808 0.000141 ***
#> genderMale           36.250     12.479   2.905 0.009446 **
#> placeB:genderMale     5.000     17.648   0.283 0.780168
#> placeC:genderMale    10.000     17.648   0.567 0.577963
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 17.65 on 18 degrees of freedom
#> Multiple R-squared:  0.8352, Adjusted R-squared:  0.7894
#> F-statistic: 18.24 on 5 and 18 DF,  p-value: 1.74e-06
```

```r
anova(interaction.model)
```

```
#> Analysis of Variance Table
#>
#> Response: income
#>              Df  Sum Sq Mean Sq F value    Pr(>F)
#> place         2 18100.0  9050.0 29.0569 2.314e-06 ***
#> gender        1 10209.4 10209.4 32.7793 1.988e-05 ***
#> place:gender  2   100.0    50.0  0.1605    0.8529
#> Residuals    18  5606.2   311.5
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Manual F-test on the interaction model
r <- 2
C2 <- matrix(c(0,0,0,0,1,0,0,0,0,0,0,1), byrow = T, nrow = 2)
C2
```

```
#>      [,1] [,2] [,3] [,4] [,5] [,6]
#> [1,]    0    0    0    0    1    0
#> [2,]    0    0    0    0    0    1
```

```r
d <- matrix(rep(0,r), ncol=1)
n <- length (data$income)

betahat3 <- matrix(interaction.model$coefficients, ncol=1)
sigma2hat3 <- summary(interaction.model)$sigma^2
X3 <- model.matrix(interaction.model)
F3 <- (t(C2%*%betahat3-d)%*%solve(C2%*%solve(t(X3)%*%X3)%*%t(C2))%*%(C2%*%betahat3-d))/(r*sigma2hat3)
F3
```

```
#>           [,1]
#> [1,] 0.1605351
```

```r
1-pf(F3,r,n-length(betahat3))
```

```
#>           [,1]
#> [1,] 0.8528939
```

As is apparent, the interaction effect is not significant according to the F-test.

Problem 2 Teaching reading

a)

Define μ_A, μ_B and μ_C as the expected score when using methods A, B and C respectively.

Then, the hypothesis test is

$$H_0 : \mu_A = \mu_B = \mu_C \quad \text{vs.} \quad H_1 : \text{ At least one is different from the others.}$$

The *treatment sum of squares*, i.e. the SSR is given by $22 \cdot ((41.05-44.02)^2 + (44.27-44.02)^2 + (46.73-44.02)^2) = 357.005$ and the SSE is 2511.712.

The model under the null hypothesis only contains an intercept, where SSE_0 is the same as SST, since SSR vanishes. Thus the F-test reads

$$F = \frac{(\text{SSE}_0 - \text{SSE})/r}{\text{SSE}/(n-p)} = \frac{\text{SSR}/r}{\text{SSE}/(n-p)} = \frac{357.005/2}{2511.712/(66-3)} = 4.477.$$

The p-value is 0.015, so the null hypothesis is rejected (at least) at level 0.05, i.e. the teaching method matters.

Assumptions needed to perform the test are that the model has the form $X_{ij} = \mu + \alpha_i + \epsilon_{ij}$, for $i = 1, 2, 3$ and $j = 1, 2, \ldots, 22$, where the response is the reading score for subject j receiving teaching method i.

b)

Define $\gamma = \mu_B/\mu_C$. We suggest an estimator for this quantity: $\hat{\gamma} = \bar{X}_B/\bar{X}_C$. A first-order Taylor expansion yields

$$\frac{x}{y} = h(x, y) \approx h(\mu_B, \mu_C) + h_x(\mu_B, \mu_C)(x - \mu_B) + h_y(\mu_B, \mu_C)(y - \mu_C) = \frac{\mu_B}{\mu_C} + \frac{1}{\mu_C}(x - \mu_C) - \frac{\mu_B}{\mu_C^2}(y - \mu_C),$$

which implies the following approximations to the expected value and standard deviation of this estimator

$$
\begin{aligned}
\widehat{\mathrm{E}(\hat{\gamma})} &= \frac{\hat{\mu}_B}{\hat{\mu}_C} \\
\widehat{\mathrm{SD}(\hat{\gamma})} &= \left(\frac{1}{\hat{\mu}_C^2} \mathrm{Var}(\bar{X}_B) + \frac{\hat{\mu}_B^2}{\hat{\mu}_C^4} \mathrm{Var}(\bar{X}_C) \right)^{1/2} = \frac{1}{\hat{\mu}_c} \sqrt{\frac{\hat{\sigma}_B^2}{n_B} + \frac{\hat{\mu}_B^2}{\hat{\mu}_C^2} \frac{\hat{\sigma}_C^2}{n_C}},
\end{aligned}
$$

where $\mathrm{Var}(\bar{X}_B) = \frac{\sigma_B^2}{n_B}$. With the given numerical values, these estimates are $\widehat{\mathrm{E}(\hat{\gamma})} = \frac{46.73}{44.27} \approx 1.06$ and $\widehat{\mathrm{SD}(\hat{\gamma})} = \frac{1}{44.27} \sqrt{\frac{7.388^2}{22} + \frac{46.73^2}{44.27^2} \frac{5.767^2}{22}} \approx 0.046$.