Recommended Exercise 8 in Statistical Linear Models, Spring 2021

alexaoh

04 mai, 2021

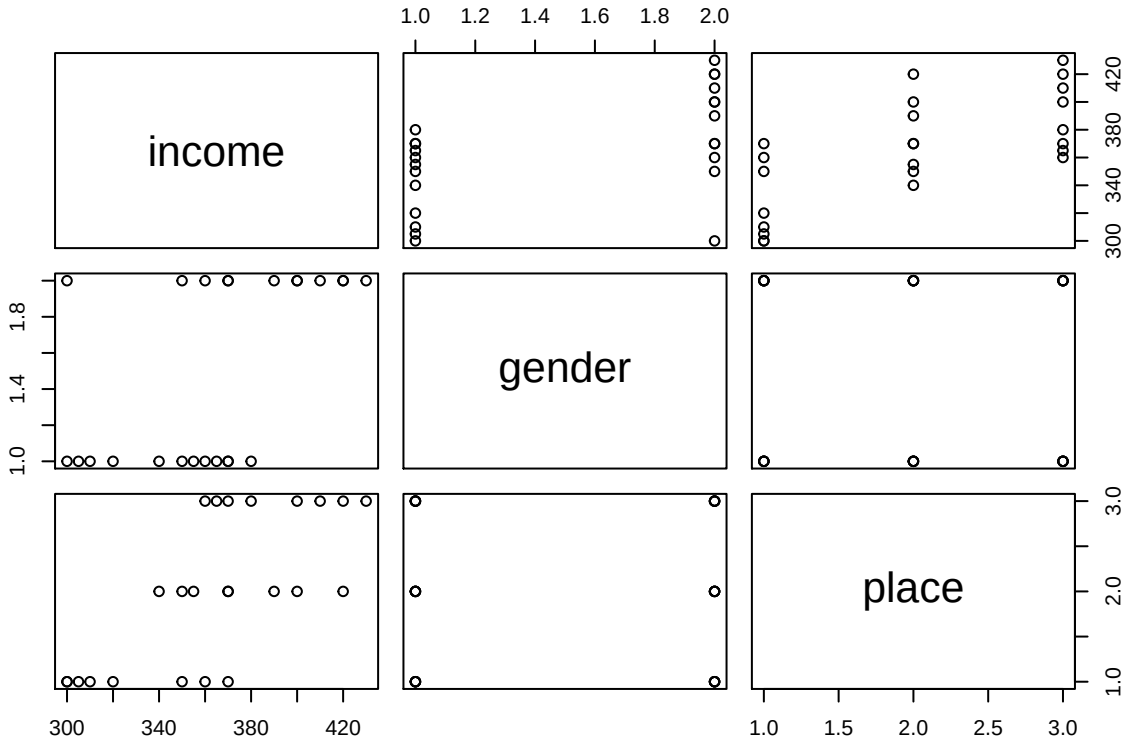
Problem 1 One- and two-way ANOVA - and the linear model

a)

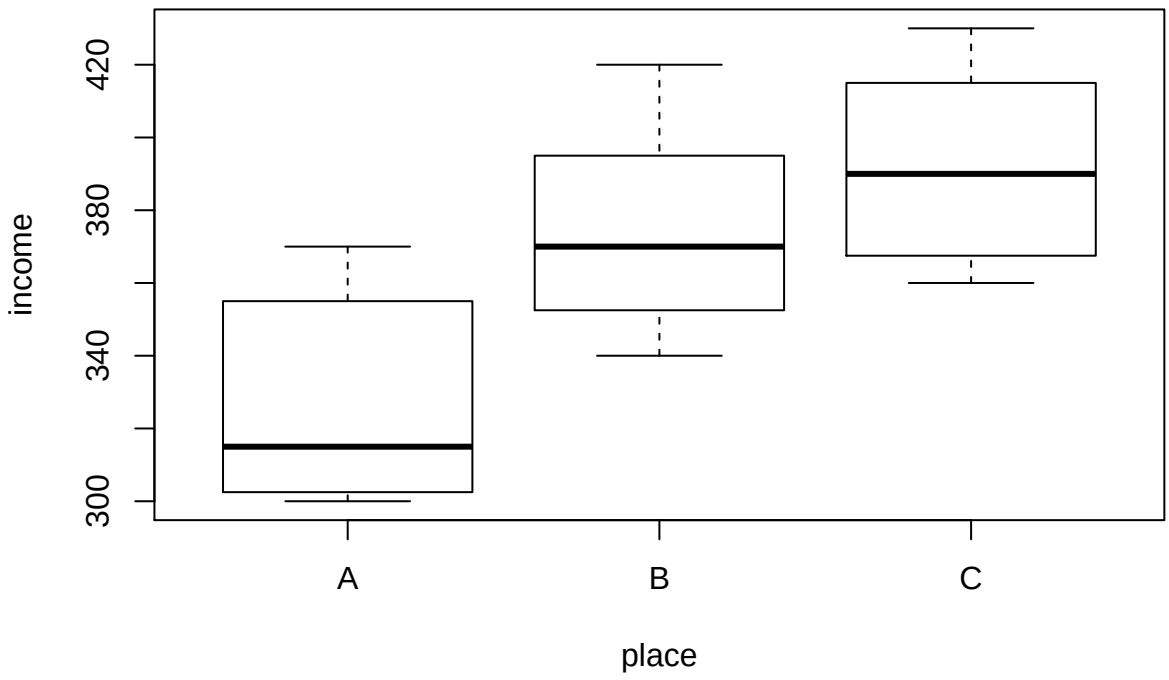
```
income <- c(300, 350, 370, 360, 400, 370, 420, 390,400, 430, 420, 410, 300, 320, 310, 305,350, 370, 340
gender <- c(rep("Male", 12), rep("Female",12))
place <- rep(c(rep("A",4), rep("B",4), rep("C",4)),2)
data <- data.frame(income, gender, place)
data
```

```
#>    income gender place
#> 1     300   Male     A
#> 2     350   Male     A
#> 3     370   Male     A
#> 4     360   Male     A
#> 5     400   Male     B
#> 6     370   Male     B
#> 7     420   Male     B
#> 8     390   Male     B
#> 9     400   Male     C
#> 10    430   Male     C
#> 11    420   Male     C
#> 12    410   Male     C
#> 13    300 Female     A
#> 14    320 Female     A
#> 15    310 Female     A
#> 16    305 Female     A
#> 17    350 Female     B
#> 18    370 Female     B
#> 19    340 Female     B
#> 20    355 Female     B
#> 21    370 Female     C
#> 22    380 Female     C
#> 23    360 Female     C
#> 24    365 Female     C
```

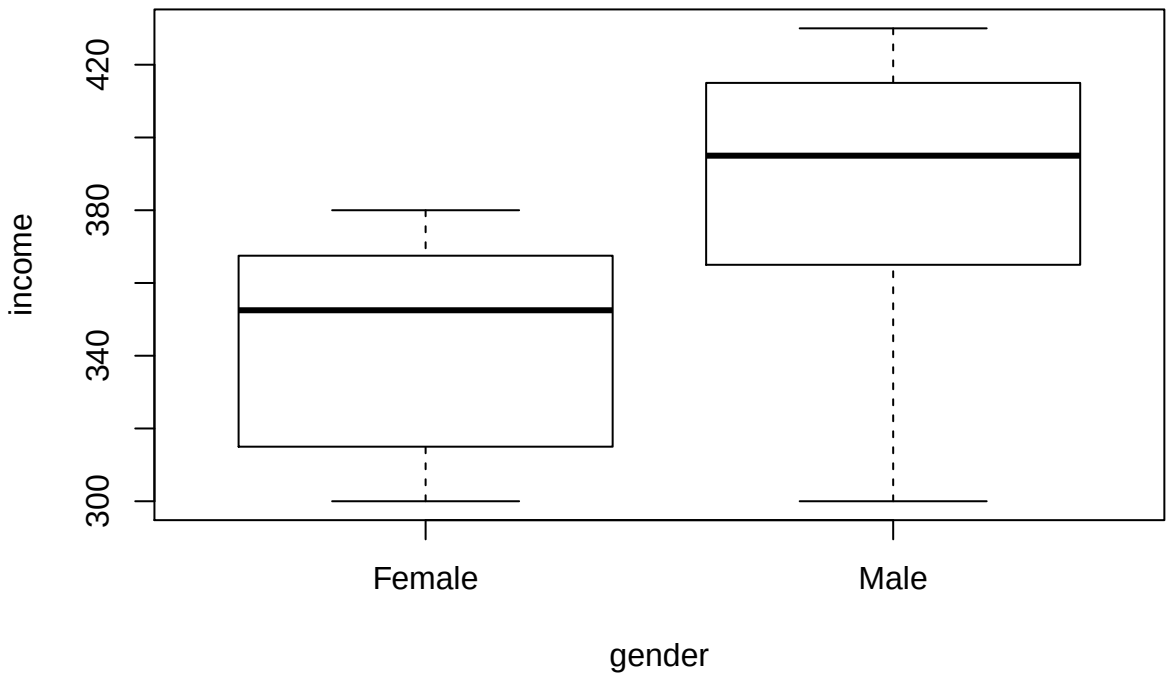
```
pairs(data)
```



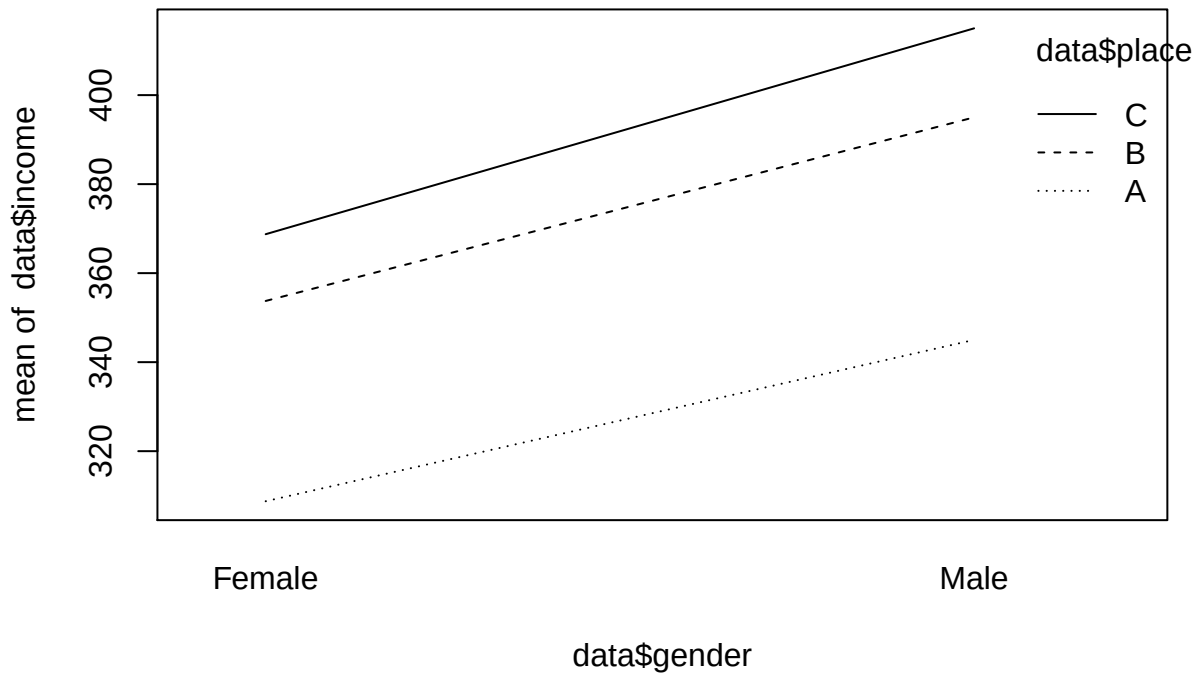

```
plot(income~place, data=data)
```



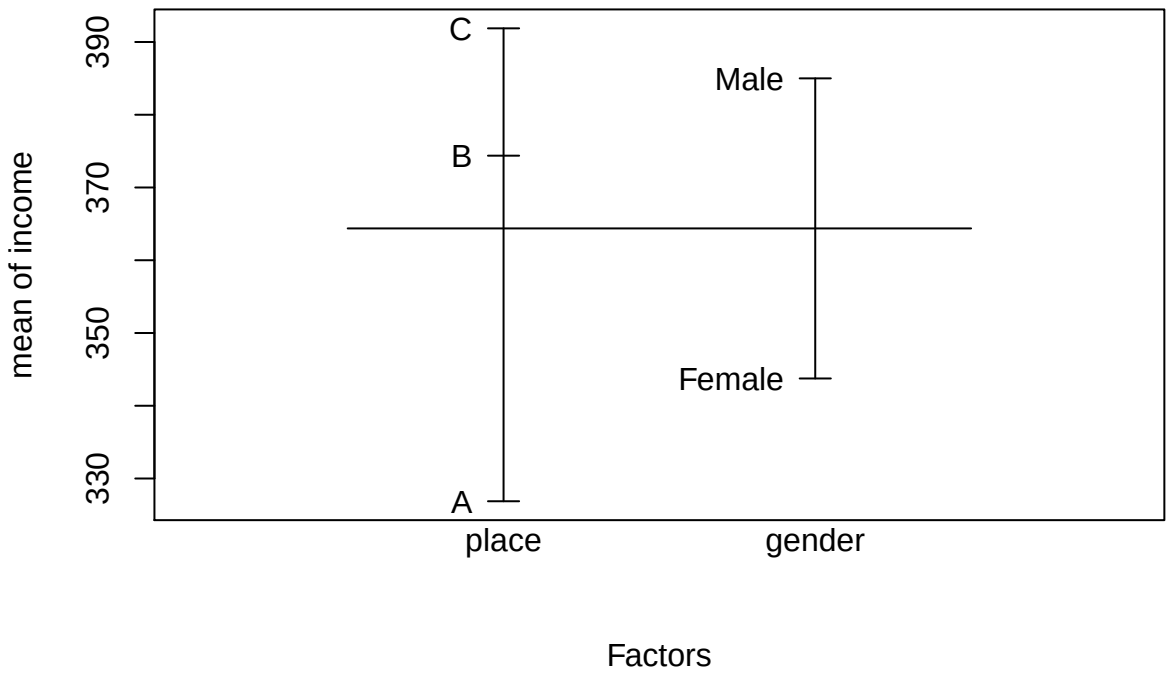

```
plot(income~gender, data=data)
```




```
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
```

```
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)

```
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
```

```
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)

```
# Må finne ut hva C og d skal være!
```

f)

```
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
```

```
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
```

```
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
```

```
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
```

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

# Gjør F-test + tolkning av modellene her også!
```

Problem 2 Teaching reading

a)

The hypothesis test is

$$H_0 : \alpha_A = \alpha_B = \alpha_C = 0 \quad \text{vs.} \quad H_1 : \text{ At least one } \alpha \neq 0.$$

Assumptions needed to make to perform the test are ...

Performing the test gives ...

The conclusion from the test is ...

b)

The suggested estimator, $\hat{\gamma}$, for γ is