Homework 5

Denis Ostroushko

2022-10-20

```
library(tidyverse)
library(kableExtra)
library(readxl)
library(gridExtra)
library(ggeffects)
```

10.2

We enter the data below. We display the code in case we make a error and need to trace the mistake back to the origin.

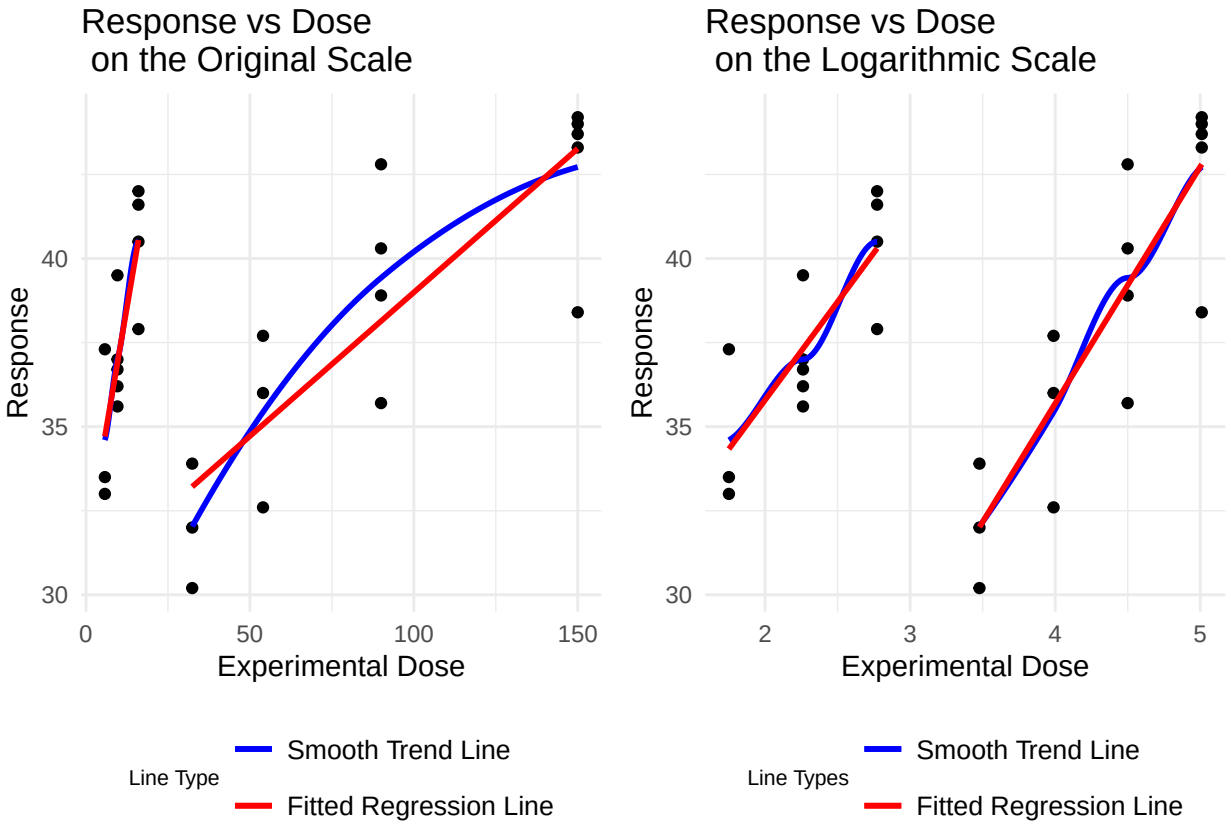
```
#put in the data

dose <- c(rep(5.76,3),
          rep(9.6, 5),
          rep(16, 4),
          rep(32.4, 3),
          rep(54, 3),
          rep(90, 4),
          rep(150, 5))

treat <- c(rep("Vitamin D3", 12),
           rep("Cod-liver Oil", 15))

response <- c(33.5, 37.3, 33,
              36.2, 35.6, 36.7, 37, 39.5,
              41.6, 37.9, 40.5, 42,
              32, 33.9, 30.2,
              32.6, 37.7, 36,
              35.7, 42.8, 38.9, 40.3,
              44, 43.3, 38.4, 44.2, 43.7)

vit_data <- data.frame(dose, response, treat)
```

10.2 - A

In this section we need to use scatter plots and test to verify that the response measurement fits better against the dose on the logarithmic scale rather than the original one. We will also use this as a chance to use an F test for lack of fit.

Visual Examination of Dose Scale Against Respose

First, we begin with the visual examination of the relationship between the response and dose on different scales. Two plots are presented below, type of fit is stated in the title.



Right away we can see the two main problems with fitting the model to the original scale of dose:

1. We are conducting a Parallel-line assays analysis, and as we can see, the slopes for two groups are not parallel. Therefore, there is an effect modification present, an interaction between treatment type and a dose given. So, it will be inappropriate to fit the model with one common slope and two different intercepts to such data. We can see that this is not the case when we obtain look at the relationship between dose on the logarithmic scale and the response.
2. There is potentially an issue with the linearity of the fit. It is hard to tell if the Vitamin D3 Dose is indeed linearly related to the response without zooming in, but there is clearly an issue for the Cod-Liver treatment. The relationship is curved, which will cause issues with the residuals. This issue goes away when we consider dose on the logarithmic scale. The smooth line dips below and above the regression line, in a random fashion.

Lack of Fit Test for Original-Dose-Scale Based Model

Now we take a chance to conduct a statistical F-test and assess model's fit. Due to models being multivariate, I do not present calculation of the test statistic by hand and use built in `R` functions instead. Recall that the idea of the test is to see if the overall fit of the model is good for each replicate level, or quasi-level, of the predictor variable. If the group-wise deviations accumulate to a greeter total error than the pure error, i.e. SSE, than the model does not fit well for certain groups, or levels, of the predictor variables, and we need to consider a different model.

We first conduct a lack of fit test for a regression model against the original dose values.

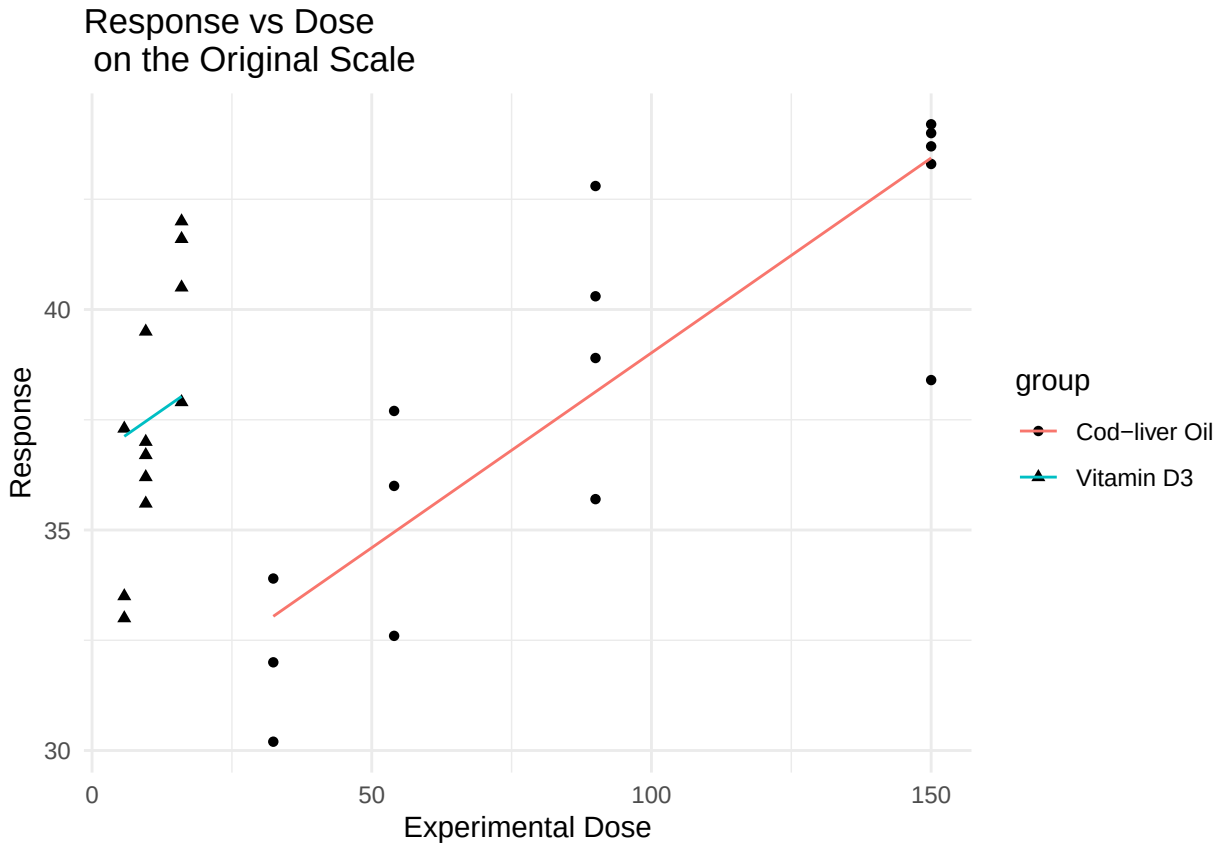
```r
reduced <- lm(response ~ dose + treat, data = vit_data)
full <- lm(response ~ 0 + as.factor(dose) + treat,data = vit_data)

res <- data.frame(anova(reduced, full) )

res$name <- c("Linear Fit","Within Group Fit")

res <- res %>% dplyr::select(name, everything())

colnames(res)[1] <- "Model Type"

res %>%
  kbl(booktabs = T, align = 'c') %>%
  kable_styling(latex_options = c("HOLD_position", "striped"))
```

Model Type	Res.Df	RSS	Df	Sum.of.Sq	F	Pr..F.
Linear Fit	24	158.2473	NA	NA	NA	NA
Within Group Fit	20	100.6488	4	57.59849	2.861359	0.0502598

Using code in the chunk above, we get all the data and statistics to conduct a test. We assume that the function of response is a linear model.

- Overall, Dose and Treatment explain 61.53% of variation in response measurements
- Null Hypothesis: $H_0 : E[Y] = \beta_0 + \beta_1 * Dose + \beta_2 * Treatment$
- Alternative Hypothesis: $H_a : E[Y] \neq \beta_0 + \beta_1 * Dose + \beta_2 * Treatment$
- Test Statistic: $F = 2.8614$
- $P(F^* > F) = 0.0502598$
- Conclusion: P-value is too close to 0.05. Therefore, we will reject the null hypothesis and conclude that response, Y, can't be reasonably expressed as a linear combination of dose and treatment levels. We need a different model or different scales of predictors.

We can also see that the fit is bad on the graph below. When fitting the model with the common slope, there is too much error and variation around the regression line for Vitamin D3 doses. This is something that we can tie back to the initial scatter plot, where we saw that we need two different slopes for two different treatments.



Lack of Fit Test for Logarithmic-Dose-Scale Based Model

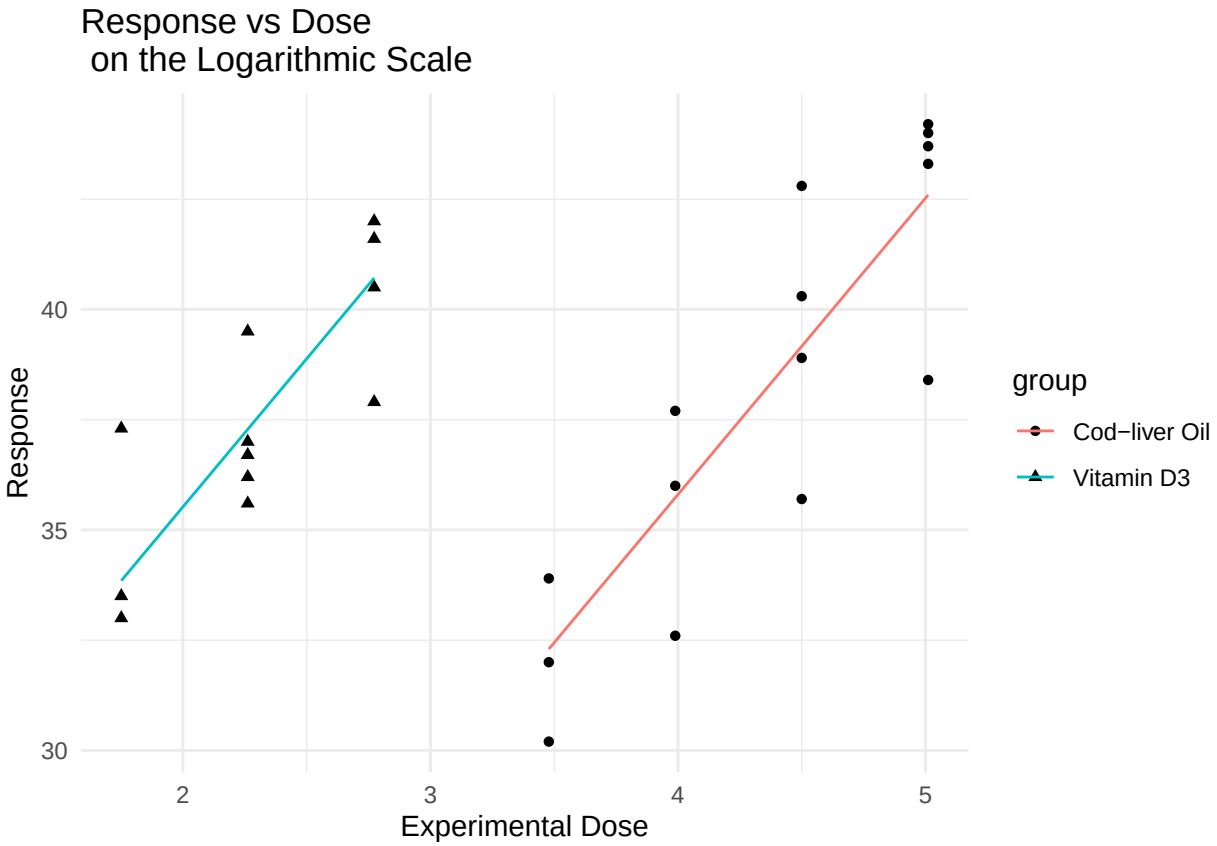
Now we will consider a fit of response against treatment level and log-transformed doses. Again, we use `R` code to conduct the test and evaluate hypotheses.

```
vit_data$log_dose <- log(vit_data$dose)

reduced <- lm(response ~ log_dose + treat, data = vit_data)
full <- lm(response ~ 0 + as.factor(log_dose) + treat,data = vit_data)

res <- data.frame(anova(reduced, full) )

res$name <- c("Linear Fit","Within Group Fit")

res <- res %>% dplyr::select(name, everything())

colnames(res)[1] <- "Model Type"

res %>%
  kbl(booktabs = T, align = 'c') %>%
  kable_styling(latex_options = c("HOLD_position", "striped"))
```

Model Type	Res.Df	RSS	Df	Sum.of.Sq	F	Pr..F.
Linear Fit	24	103.7474	NA	NA	NA	NA
Within Group Fit	20	100.6488	4	3.098517	0.1539271	0.958987

- Overall, Log - Dose and Treatment explain 74.78% of variation in response measurements
- Null Hypothesis: $H_0 : E[Y] = \beta_0 + \beta_1 * Log - Dose + \beta_2 * Treatment$
- Alternative Hypothesis: $H_a : E[Y] \neq \beta_0 + \beta_1 * Log - Dose + \beta_2 * Treatment$
- Test Statistic: $F = 0.1539$
- $P(F^* > F) = 0.958987$
- Conclusion: The p-value is very large, meaning that we do not have enough evidence to reject the null hypothesis. Therefore, we can use this model for the Parallel-line assays analysis.

We can see that the fit is so much better here. Parallel slopes fit their respective groups very well. Variance of data points around the regression lines is reduced, and visually appears constant. So, we will use this model to complete to the rest of this problem.



10.2 - B

From the lecture notes we know that parallel-line assays are those in which the response is linearly related to the log dose. If the response is linearly related to log dose, then the coefficient should be statistically different from 0. Using a model from the previous section, we can validate this. Looking at the summary table for the model, we can conduct a statistical test for the coefficient of log-dose.

Model Term	Estimate	Std. Error	T-value	P-value
Intercept	22.084	1.940	11.384	0
Log - Dose	6.719	0.801	8.393	0
Cod - Liver Oil Treatment	-13.156	1.835	-7.171	0

- Null Hypothesis: $\hat{\beta}_{log-dose} = 0$
- Alternative Hypothesis: $\hat{\beta}_{log-dose} \neq 0$
- Test Statistic: 8.393
- $P(t^* > t) = 0$
- Conclusion: P-value is less than 0.05, so we reject the null hypothesis in favor of the alternative hypothesis. We have enough statistical evidence to conclude that the response is linearly related to the log-dose variables.

We can also test if the lines are indeed parallel. The two lines are parallel if they have the same slope, which we verified with the previous test, and if they have two different intercepts, i.e. they are indeed two different lines.

We conduct this test using an estimate for the indicator variable, also obtained from the same model we used from the previous test.

- Null Hypothesis: $\hat{\beta}_{treatment} = 0$
- Alternative Hypothesis: $\hat{\beta}_{treatment} \neq 0$
- Test Statistic: -7.171
- $P(t^* > t) = 0$
- Conclusion: P-value is less than 0.05, so we reject the null hypothesis in favor of the alternative hypothesis. We have enough statistical evidence to conclude that the two lines are indeed parallel becuase they have coefficients that are statistically significantly different.

10.2 - C

Using model summary table, we know that the two estimates are:

- $\hat{\beta}_1 = 6.72$
- $\hat{\beta}_2 = -13.16$

Code below calculates $m = log[p]$

```
rel_pot <- coefficients(full_lm)[3] / coefficients(full_lm)[2]
```

- From slide 11, we obtain Relative potency $= m = log[p] = \frac{\hat{\beta}_2}{\hat{\beta}_1} = -1.958$

10.2 - D

Formula for variance and standard error of the estimate is much bigger, and requires that we obtain more estimates from the model. We save down $\hat{\beta}_1$ and $\hat{\beta}_2$ using the model we fit and used in the previous sections. Also, using this model we obtain a variance-covariance matrix for the estimates. Code chunk below obtains and saves all estimates that we need.

```
beta_1 <- coefficients(full_lm)[2]
beta_2 <- coefficients(full_lm)[3]

var_beta_1 <- vcov(full_lm)[2,2] # beta_1 variance
var_beta_2 <- vcov(full_lm)[3,3] # beta_2 variance

cov_beta_12 <- vcov(full_lm)[2,3] #covariance of beta_1 and beta_2
```

We copy down a formula from lecture Slide 17:

$$Var(m) = \frac{\hat{\beta}_2^2}{\hat{\beta}_1^4} \times Var(\hat{\beta}_1) + 2(-\frac{\hat{\beta}_2}{\hat{\beta}_1^2}) \times (\frac{1}{\hat{\beta}_1}) \times Cov(\hat{\beta}_1, \hat{\beta}_2) + \frac{1}{\hat{\beta}_1^2} \times Var(\hat{\beta}_2)$$

For this calculation we have the following estimates:

- $\hat{\beta}_1 = b_1 = 6.71875$
- $\hat{\beta}_2 = b_2 =$ -13.15563
- $Var(\hat{\beta}_1) = Var(b_1) = 0.64086$
- $Var(\hat{\beta}_2) = Var(b_2) = 3.36599$
- $Cov(\hat{\beta}_1, \hat{\beta}_2) = Cov(b_1, b_2) =$ -1.31969

Calculation of the estimate is given below:

```
Var_m <-
  (beta_2 ^ 2)/(beta_1 ^ 4) * var_beta_1 +
    2 * (-1) * (beta_2 / beta_1^2 ) * (1/beta_1) * cov_beta_12 + (1/(beta_1^2)) * var_beta_2
```

So, $Var(m) = 0.01451$, and the standard error is $se(m) = \sqrt{Var(m)} = 0.12046$

11.1

11.1 - A

In order to decompose SSR into Extra Sum of Squares we will do in sequence, fitting models one by one and calculating SSR along the way. We need the following components:

- Full Regression SSR, that involves all 3 predictors: $SSR(CPD, Age, Gender)$ and $df = p - 1 = 4 - 1 = 3$
- $SSR(CPD)$ and $df = p_{cpd} - 1 = 2 - 1 = 1$, because we have two model estimates that helped us obtain this SSR
- $SSR(Age|CPD)$ and $df = 1$, for one extra predictor that we include
  - $SSR(Age|CPD) = SSR(Age, CPD) - SSR(CPD)$
- $SSR(Gender|CPD, Age)$ and $df = 1$, again, for the extra predictor
  - $SSR(Gender|CPD, Age) = SSR(Age, CPD, Gender) - SSR(Age, CPD)$

Code below shows how we fit models one and by and save down SSR and Extra SSR along the way

```
full_model <- lm(nnal ~ cpd + age + gender, data = cig)
SSR_full <- sum((mean(cig$nnal) - full_model$fitted.values)^2) # full regression SSR

cpd_model <- lm(nnal ~ cpd, data = cig)
SSR_cpd <- sum((mean(cig$nnal) - cpd_model$fitted.values)^2) # SSR(X_1)

age_model <- lm(nnal ~ age, data = cig)
SSR_age <- sum((mean(cig$nnal) - age_model$fitted.values)^2) # SSR(X_2)

cpd_age_model <- lm(nnal ~ cpd + age, data = cig)
SSR_cpd_age <- sum((mean(cig$nnal) - cpd_age_model$fitted.values)^2) # (SSR X_1, X_2)

SSR_age_given_cpd <- SSR_cpd_age - SSR_cpd # SSR(X_2 | X_1)
SSR_cpd_given_age <- SSR_cpd_age - SSR_age # SSR(X_1 | X_2)

SSR_gender_given_cpd_age <- SSR_full - SSR_cpd_age # SSR(X_3 | X_1, X_2) = SSR(X_3 | X_2, X_1)
```

We also save down residual sum of squares and total sum of errors

```
SSE <- sum(full_model$residuals^2)

SSTO <- sum((mean(cig$nnal) - cig$nnal)^2)
```

And we finally can put together a table that we will use later in the analysis

Source	SS	DF	MS
CPD + Age + Gender	60.08376	3	20.02792
Extra SS			
CPD	33.02253	1	33.02253
Age|CPD	13.94347	1	13.94347
Gender|Age, CPD	13.11776	1	13.11776
Error			
Residual Error	823.32427	82	10.04054
Total Error	883.40803	85	NA

We can also double check our calculation using a build in `anova()` function.

```
anova_built_in <- data.frame(anova(full_model))
anova_built_in$Variable <- c("CPD", "Age", "Gender", "Residuals")

rownames(anova_built_in) <- NULL

anova_built_in <- anova_built_in %>% select(Variable, everything())

anova_built_in %>%
  kbl(booktabs = T, align = 'c') %>%
  kable_styling(latex_options = c("HOLD_position", "striped"))
```

Variable	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
CPD	1	33.02253	33.02253	3.288919	0.0734074
Age	1	13.94347	13.94347	1.388717	0.2420303
Gender	1	13.11776	13.11776	1.306480	0.2563592
Residuals	82	823.32427	10.04054	NA	NA

As we can see, Sum of Squares and Mean Square, match with the table that we have produced

11.1 - B

To test we need to get a few values for the F statistic

- We already have extra sum of squares $SSR(Gender|CPD, Age)$
- We also have $SSE(Gender, Age, CPD)$
- F - statistic is then:

$$\frac{\frac{SSR(Gender|CPD,Age)}{1}}{\frac{SSE(Gender,Age,CPD)}{n-4}}$$

Hypothesis and test results are given below:

- Null Hypothesis: $H_0 : \hat{\beta}_{gender} = 0$
- Alternative Hypothesis: $H_0 : \hat{\beta}_{gender} \neq 0$
- $F-$ statistic: 1.3065

- $P(F^* > F) = 0.2563592$
- For comparison, here is a model summary that provides a t-test for Gender covariate:

Model Term	Estimate	Std. Error	T-value	P-value
Intercept	2.054	1.906	1.077	0.285
CPD	0.052	0.029	1.788	0.077
Age	0.016	0.024	0.644	0.521
Gender	-0.877	0.767	-1.143	0.256

- Note how test statistic F equals squared t statistic from the model summary table above. P-value for the t-test also matches a p-value from the F-test that we conducted, which is obviously to be expected.
- Conclusion: Since p-value is above 0.05, we can not reject the Null hypothesis and conclude that the coefficient for gender is statistically different from 0. Therefore, we also can not conclude that the inclusion of gender into the model after cpd and age meaningfully contributes to the proportion of the variation in NNAL values that the model is explains.

11.1 - C

To test we need to get a few values for the F statistic

- We already have extra sum of squares $SSR(Gender, Age|CPD) = SSR(Age, Gender, CPD) - SSR(CPD)$
- We also have $SSE(Gender, Age, CPD)$
- F - statistic is then:

$$\frac{\frac{SSR(Gender,Age|CPD)}{2}}{\frac{SSE(Gender,Age,CPD)}{n-4}}$$

Hypothesis and test results are given below:

- Null Hypothesis: $H_0 : \hat{\beta}_{gender} = \hat{\beta}_{age} = 0$
- Alternative Hypothesis: $H_a : \hat{\beta}_{gender}$ and $\hat{\beta}_{age}$ are not all 0
- $F-$ statistic: 1.3476
- $P(F^* > F) = 0.1044811$
- Conclusion: the p-value is close to 0.05, but still twice as big as our accepted confidence level. Therefore, we do not have enough evidence to conclude that the inclusion of both age and gender after cpd meaningfully contributes to the proportion of variation of the response variable that are able to explain.

11.1 - D

Yes, it is always the case, because the order of the variables is arbitrary.

For example, in this problem we have

- $SSR(X_1) = SSR(CPD) = 33.022528$
- $SSR(X_2) = SSR(Age) = 12.1301697$
- $SSR(X_2|X_1) = SSR(Age|CPD) = 13.9434691$
- $SSR(X_1|X_2) = SSR(CPD|Age) = 34.8358274$

- Now we can show that

  $SSR(X_2|X_1) + SSR(X_1) =$

  $13.9434691 + 33.022528 =$

  $12.1301697 + 34.8358274 =$

  $SSR(X_1|X_2) + SSR(X_2)$