Homework 5

Denis Ostroushko

2022-10-18

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

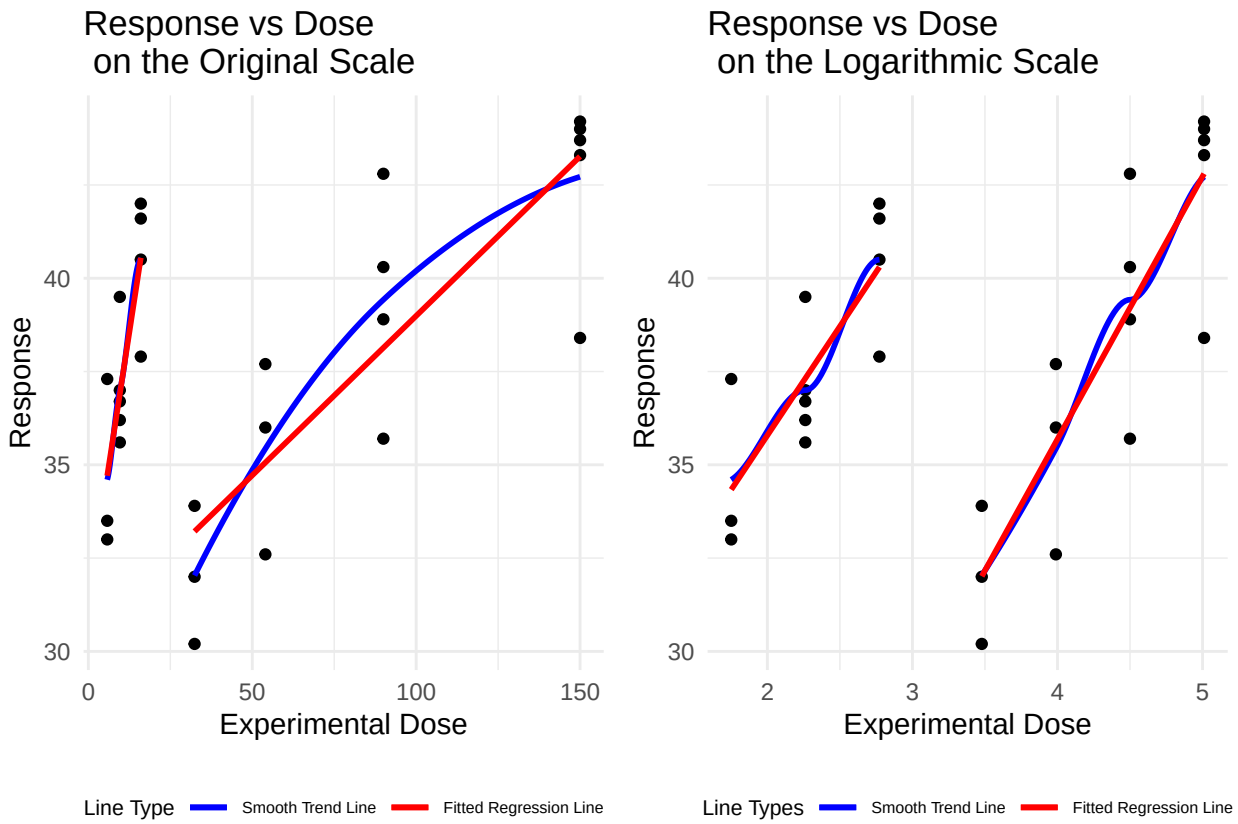
10.2

We enter the data below.

```
#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

Visual Examination of Dose Scale Against Respose asd f gasg sfd



asfvseg

Lack of Fit Test for Original-Dose-Scale Based Model sdfg sdf g

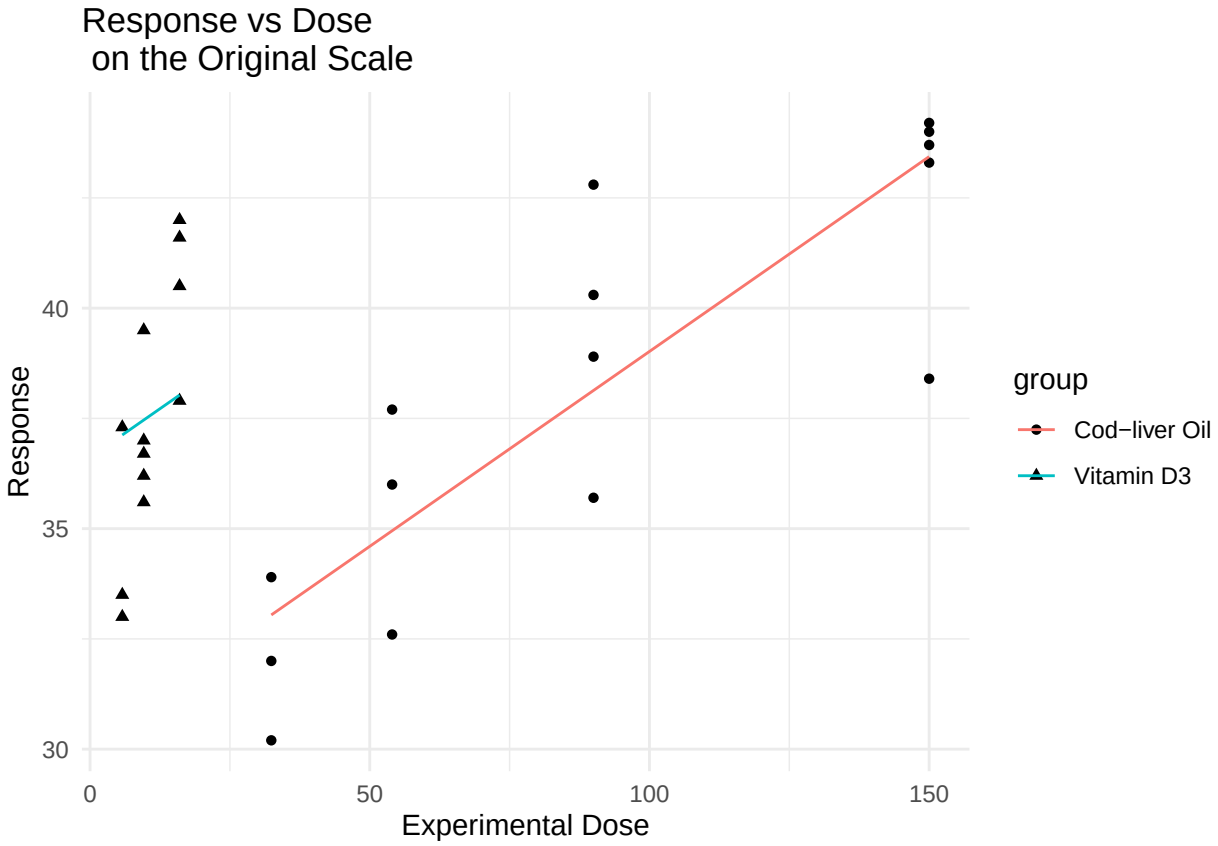
```
reduced <- lm(response ~ dose + treat, data = vit_data)
full <- lm(response ~ 0 + as.factor(dose) + treat,data = vit_data)

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

- 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:

Visualization of bad fit



- Observations

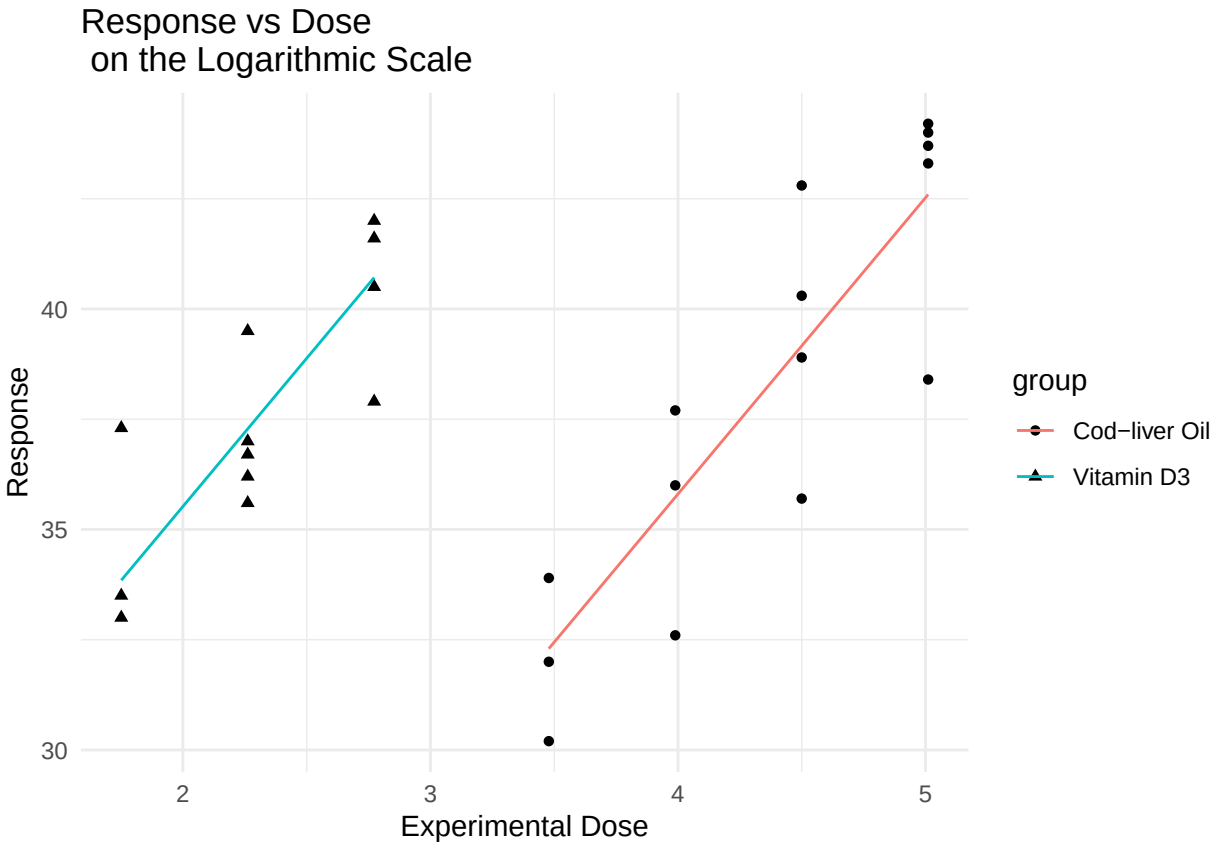
```
vit_data$log_dose <- log(vit_data$dose)

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

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

- 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:

Fitted Lines from the log based model



*Observations

10.2 - B

fit the multiple linear regression

FROM SLIDES Parallel-line assays are those in which the response is linearly related to the log dose

So, the lines are parallel if the response is related to the log dose

response is related if coefficient for log dose is not 0

```r
vit_data$treat <- factor(vit_data$treat , levels = c("Vitamin D3", "Cod-liver Oil"))

full_lm <- lm(response ~ log(dose) + treat, data = vit_data)

sum_data <- data.frame(summary(full_lm)$coefficients)

sum_data$names <- c("Intercept", "Log - Dose", "Cod - Liver Oil Treatment")

rownames(sum_data) <- NULL

sum_data <- sum_data %>%  dplyr::select(names, everything())

round_3 <- function(x){round(x,3)}
sum_data[,2:5] <- lapply(sum_data[,2:5], round_3)

colnames(sum_data) <-c("Model Term", "Estimate", "Std. Error", "T-value", "P-value")



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

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:

10.2 - C

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

- $\hat{\beta}_1 = 6.72$
- $\hat{\beta}_2 = -13.16$
- Relative potency $= m = log[p] = \frac{\hat{\beta}_2}{\hat{\beta}_1} = -1.958$

10.2 - D

save all needed estimates from the model

```r
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 varinace

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

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

```
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

```
cig <- read_xls("/Users/denisostroushko/Desktop/UofM MS/MS Fall 2022/Puhb 7405/Data Sets/Cigarettes.xls
colnames(cig) <- c("age", "gender", "cpd", "carbon_mono", "cotinine", "nnal")

cig <- cig %>% dplyr::select(nnal, cpd, age, gender )
```

11.1 - A

We need:

- Full Regression $SSR(CPD, Age, Gender)$ and $df = 3$
- $SSR(CPD)$ and $df = 1$
- $SSR(Age|CPD)$ and $df = 1$
- $SSR(Gender|CPD, Age)$ and $df = 1$

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

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

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

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

```r
SSR_age_given_cpd <- SSR_cpd_age - SSR_cpd
SSR_cpd_given_age <- SSR_cpd_age - SSR_age

SSR_gender_given_cpd_age <- SSR_full - SSR_cpd_age
```

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

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

```r
anova_tab <-
  data.frame(
    Source = c("CPD + Age + Gender", "CPD", "Age|CPD", "Gender|Age, CPD", "Residual Error", "Total Erro
    SS = c(SSR_full, SSR_cpd, SSR_age_given_cpd, SSR_gender_given_cpd_age, SSE, SSTO),
    DF = c(3,1,1,1,nrow(cig)-4,nrow(cig)-1)
  )

anova_tab$MS <- anova_tab$SS / anova_tab$DF

anova_tab$MS[6] <- NA

anova_tab %>%
  kbl(booktabs = T) %>%
  kable_styling(latex_options = c("striped", "HOLD_position")) %>%
  pack_rows("Extra SS", 2, 4) %>%
  pack_rows("Error", 5, 6)
```

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

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

- Conclusion:

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
- $P(F^* > F) = 0.1044811$

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

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

  $13.9434691 + 33.022528 =$

  $12.1301697 + 34.8358274 =$

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