HW 2

Maria Luisa Klobongona

```
library(knitr) # compiling .qmd files
library(pander) # format tables for markdown
library(ggplot2) # graphics
library(ggeasy) # help with graphics:
# - `ggeasy::easy_labs()` applies same logic as `ggplot2::labs()` but uses as default the "la
library(dplyr) # manipulate data
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(haven) # import Stata files
library(tidyr) # Tools to help to create tidy data
library(plotly) # interactive graphics
```

```
Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':

    last_plot
```

```
The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout
```

```
library(dobson) # datasets from Dobson and Barnett 2018
library(parameters) # format model output tables for markdown
library(conflicted) # check for conflicting function definitions
library(magrittr) # `%>%` and other additional piping tools
library(ggfortify)
```

1 Definitions

1.1

Memorize and write down from memory the definition of the standard error of an estimator , using natural-language prose (do not use any math symbols other than for this sub-section).

The standard error of an estimator is the standard deviation of the error of that estimator ( ).

1.2

Memorize and write down from memory the definition of the standard error of an estimator , using mathematical notation.

$$SE(\hat{\theta}) = SD(\hat{\theta})$$

2 Linear Regression

(Adapted from Dobson & Barnett, 2018, ex. 6.4)

It is well known that the concentration of cholesterol in blood serum increases with age, but it is less clear whether cholesterol level is also associated with body weight. The cholesterol dataset in the dobson package contains serum cholesterol (chol, millimoles per liter), age (age, years)

and body mass index (bmi, weight divided by height squared, where weight was measured in kilograms and height in meters), for thirty women.

```
library(pander)
 library(dobson)
 data(cholesterol, package = "dobson")
 head(cholesterol) |> pander()
```

chol	age	bmi
5.94	52	20.7
4.71	46	21.3
5.86	51	25.4
6.52	44	22.7
6.8	70	23.9
5.23	33	24.3

2.1

Create scatterplots of the bivariate relationships between these variables.

```
library(GGally)
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```
ggpairs(cholesterol)
```

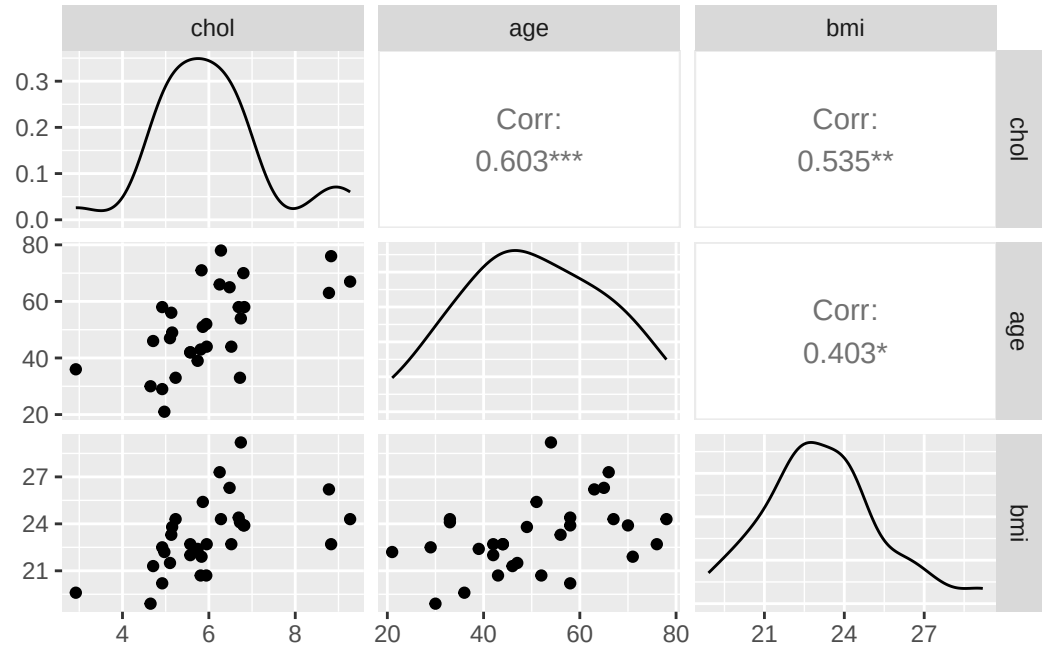
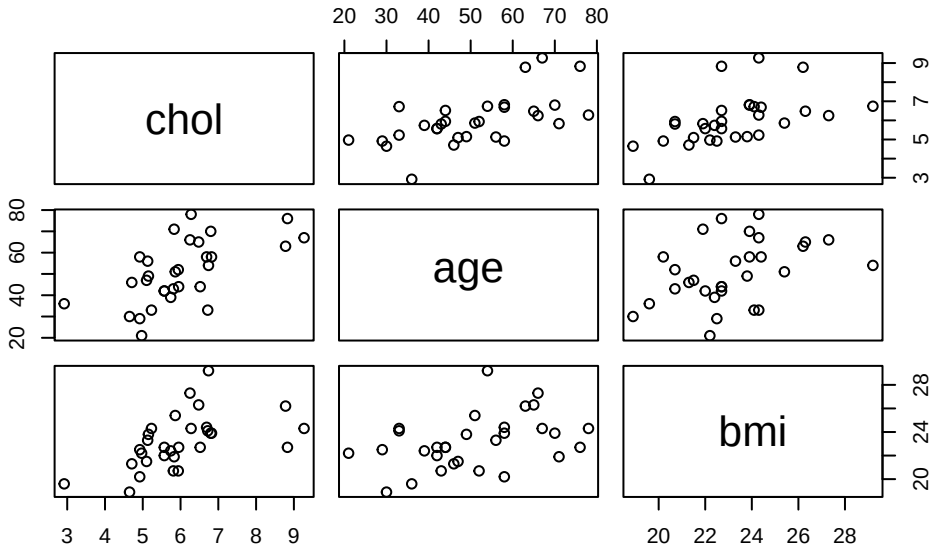


Figure 1: Scatterplots f the bivariate relationships

```
pairs(cholesterol)
```



2.2

Use multiple regression to assess whether serum cholesterol might be associated with body mass index, adjusting for age. Interpret the coefficient estimates, and state your scientific conclusions.

```
chol_lm <- lm(chol ~ bmi + age, data = cholesterol)
#summary (chol_lm)
chol_lm |> parameters()|> print_md()
```

Parameter	Coefficient	SE	95% CI	t(27)	p
(Intercept)	-0.74	1.90	(-4.63, 3.15)	-0.39	0.700
bmi	0.20	0.09	(0.02, 0.38)	2.27	0.031
age	0.04	0.01	(0.01, 0.07)	3.01	0.006

The Intercept (β_0): The mean serum cholesterol millimoles per liter is -0.74 when the body mass index and age is 0.The p-value for intercept is not significant, the interpration does not practically matter and applicable.

The Slope coefficient of bmi (β_1): The difference in the mean serum cholesterol per one kg/m difference in body mass index was 0.20 millimoles per liter, holding the age constant.

The slope coefficient of age (β_2): The average difference in mean serum cholesterol levels associated with a one-year difference in age was 0.04 millimoles per liter, while controlling for the effects of bmi.

Because the p-value for BMI was smaller than $\alpha = 0.5$ (p=0.03), we reject the null and conclude that there is enough evidence that there is significant and positive association between BMI and the serum cholesterol, after adjusting for age. The p-value for age was also smaller than $\alpha = 0.5$ (p=0.006), we reject the null and conclude there is enough evidence that there is significant and positive association between age and the serum cholesterol.

2.3

Does the relationship between BMI and cholesterol depend on age? To answer this question, add an interaction term and refit the model. Interpret the coefficient estimates, and state your scientific conclusions.

```
chol_lm2 <- lm(chol ~ bmi + age + bmi:age, data = cholesterol)
chol_lm2|> parameters() |> print_md()
```

Parameter	Coefficient	SE	95% CI	t(26)	p
(Intercept)	-6.55	8.95	(-24.94, 11.85)	-0.73	0.471
bmi	0.46	0.40	(-0.36, 1.27)	1.16	0.258
age	0.15	0.17	(-0.20, 0.51)	0.90	0.375
bmi × age	-4.93e-03	7.43e-03	(-0.02, 0.01)	-0.66	0.512

The Intercept (β_0): The mean serum cholesterol was -6.55 millimoles per liter when the body mass index and age are 0. This p-value is not significant and intercept is also not interpretable.

The Slope coefficient of bmi (β_1): The difference in the mean serum cholesterol per one kg/m difference in body mass index was 0.46 millimoles per liter, holding the age constant.

The slope coefficient of age (β_2): The difference in mean serum cholesterol levels associated with a one-year difference in age was 0.15 millimoles per liter, while controlling for the effects of bmi.

The slope coefficient of age and bmi (β_3): The difference in mean serum cholesterol levels for 1 kg/m difference in bmi at different ages was -4.93e-03. Since the p-value is 0.512, this differential effect is not statistically significant, suggesting no enough evidence to conclude that there is influence of bmi on serum cholesterol across different age in the population studied.

Because the p-value for all slopes is greater than $\alpha = 0.5$, we rejected the null and stated we did not have enough evidence to conclude that bmi, age, and the interaction of bmi and age have a significant association with the difference in mean serum cholesterol.

2.4

If you haven't already done so, improve the precision of your coefficient estimates by recentering the covariates as needed. Reinterpret the coefficient estimates and state your revised scientific conclusions.

```
chol= cholesterol %>%
  mutate(
age.c = age - mean(age), # mean age = 50.7
bmi.c = bmi - mean(bmi) # mean bmi = 23.18
)

chol_lm3 <- lm(chol ~ bmi.c + age.c + bmi.c:age.c, data = chol)
chol_lm3|> parameters() |> print_md()
```

Parameter	Coefficient	SE	95% CI	t(26)	p
(Intercept)	6.07	0.21	(5.64, 6.50)	29.30	$< .001$
bmi c	0.21	0.09	(0.02, 0.39)	2.30	0.030
age c	0.04	0.01	(0.01, 0.07)	2.87	0.008
bmi c × age c	-4.93e-03	7.43e-03	(-0.02, 0.01)	-0.66	0.512

Here, we recentered the age and bmi by its mean (age-50.7, bmi-23.18) to the model with interaction of age and bmi. Then, the coefficient estimate as below;

The Intercept (β_0): The mean serum cholesterol was 6.07 millimoles per liter when the body mass index is 23.18 kg/m and age are 50.7. The intercept is significantly different from zero (p $< .001$), suggesting a high mean serum cholesterol level at the mean values of BMI and age.

The Slope coefficient of bmi (β_1): The slope of BMI is statistically significant with p-value of 0.03 less than α of 0.05 suggesting that the difference in the mean serum cholesterol per one kg/m difference in body mass index from its mean was 0.21 millimoles per liter, holding the age constant at its mean (50.7)

The slope coefficient of age (β_2): The slope of age is statistically significant with p-value of 0.008 less than α of 0.05 suggesting the difference in mean serum cholesterol levels is significantly associated with a one-year difference in age from its mean was 0.04 millimoles per liter, while controlling for the effects of bmi at its mean.

The slope coefficient of age and bmi (β_3): The difference in mean serum cholesterol levels for 1 kg/m difference in bmi from its mean at different ages was -4.93e-03. Since the p-value is 0.512, this differential effect is not statistically significant, suggesting no enough evidence to conclude that there is influence of bmi on serum cholesterol across different age in the population studied.

Because the p-value for all slopes is greater than $\alpha = 0.5$, we failed to reject the null and stated we did not have enough evidence to conclude the interaction of bmi and age have a significant association with the difference in mean serum cholesterol. This means the relationship between BMI and serum cholesterol does not significantly change at different levels of age, vice versa. the BMI and

2.5

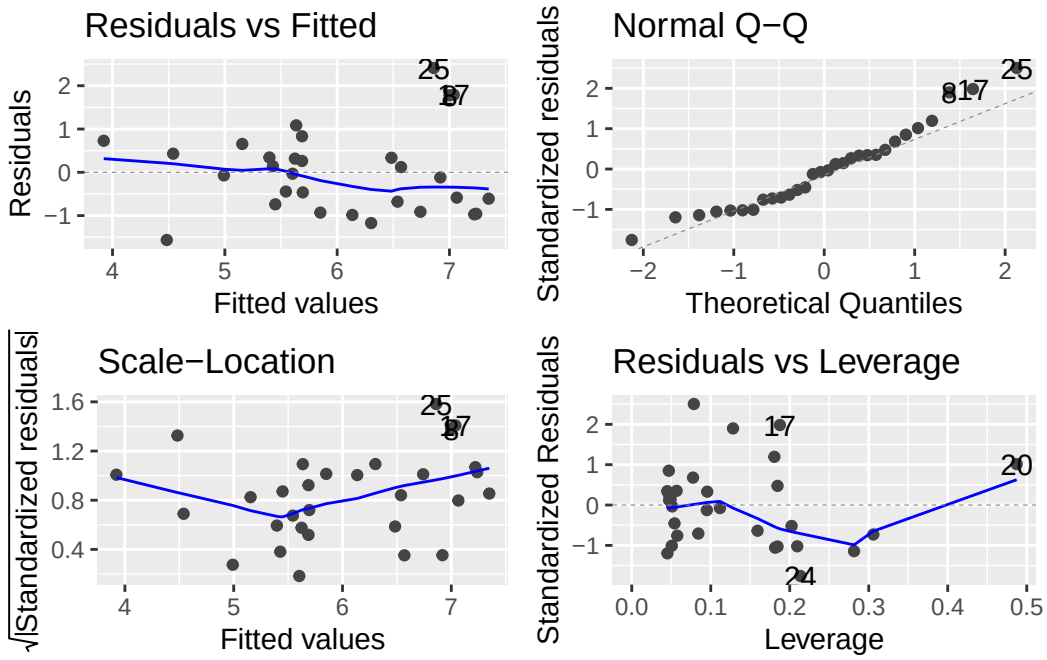
How did centering change your results?

We found that recentering the covariates still did not improved the interaction of age and bmi. However, each covariate individually improved, here, the p-values of age and bmi are smaller than significantly different from 0, indicating a positive association between both BMI and age with serum cholesterol levels when considering the mean values of these predictors.

2.6

Create graphs of regression diagnostics for your final model, and assess whether it seems to be a good model.

```
autoplot(chol_lm3)
```



The Residuals vs Fitted plot shows no clear pattern, which is good. However, a couple of points stand out, which might be outliers or influential points i.e 25

The Q-Q plot still shows some deviation from normality at the tails, especially for the points labeled 8, 17 and 25.

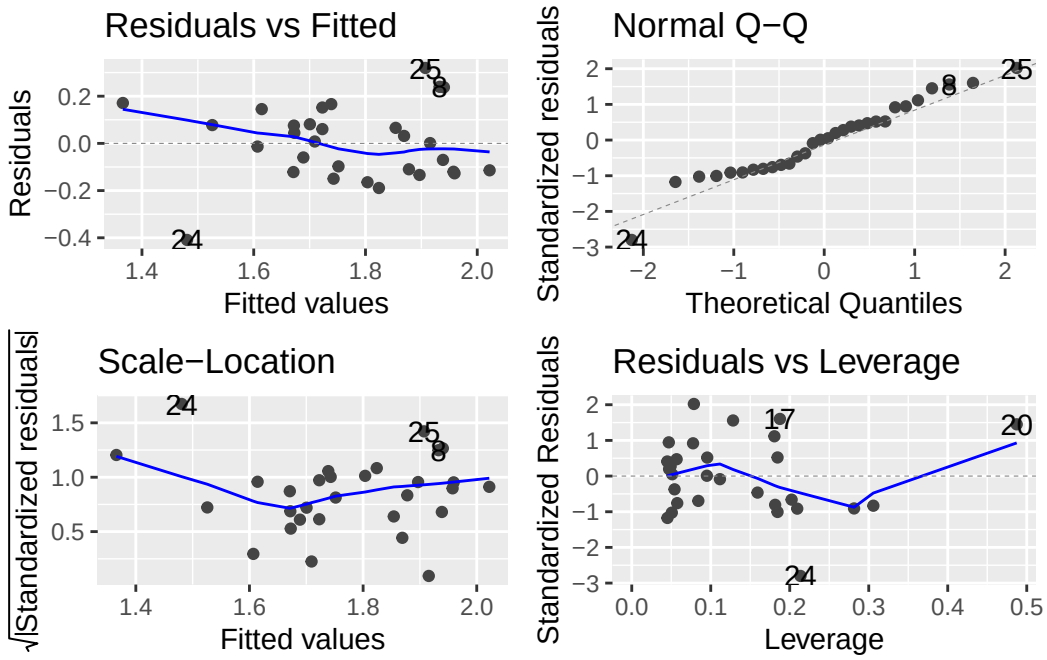
Scale-Location: This plot is showing a pattern that suggests possible heteroscedasticity, as the spread of the residuals increases with the fitted values.

Residuals vs Leverage: at least one point with high leverage, but it doesn't appear to have a large residual. The points labeled as 17, 25, and 26 might be a potential outliers

2.7

Try at least one change to the model that might improve the fit

```
chol_lm4 <- lm(log(chol) ~ bmi.c + age.c + bmi.c:age.c, data = chol)
autoplot(chol_lm4)
```



I modified the previous model by taking the log of cholesterol (y) in the model with centering covariates and interaction term of bmi and age. Here, the modified model did not have much improvement.

Residuals vs Fitted: The residuals appear randomly scattered around the horizontal line,but there are same outliers or influential points are still visible.

Normal Q-Q: The Q-Q plot still shows some deviation from normality at the tails, especially for the points labeled 24, 8, and 25, but the overall alignment along the line appears consistent.

Scale-Location: This plot is showing a pattern that suggests possible heteroscedasticity, as the spread of the residuals still increases with the fitted values.

Residuals vs Leverage: Point 24 has high leverage and a large residual, making it potentially influential. Points 17 and 20 also stand out, although to a lesser extent.

3 Stratification

```
data("birthweight", package= "dobson")

bw=
 birthweight|>
 pivot_longer(
 cols= everything(),
 names_to= c("sex",".value"),
 names_sep= "s "
 ) |>
 rename(age = 'gestational age') |>
  mutate(
 sex = ifelse( sex== "boy", "male","female"))

lm.bw= lm(weight ~ sex + sex:age - 1, data= bw)
 lm.bw|> parameters() |>print_md()
```

Parameter	Coefficient	SE	95% CI	t(20)	p
sex (female)	-2141.67	1163.60	(-4568.90, 285.56)	-1.84	0.081
sex (male)	-1268.67	1114.64	(-3593.77, 1056.42)	-1.14	0.268
sex (female) × age	130.40	30.00	(67.82, 192.98)	4.35	< .001
sex (male) × age	111.98	29.05	(51.39, 172.57)	3.86	< .001

Fittting two separate model, one for each sex:

```
 lm.bw.male=lm(
 formula= weight ~age,
 data = bw |> dplyr::filter(sex =="male"))
 lm.bw.male|>parameters() |> print_md()
```

Parameter	Coefficient	SE	95% CI	t(10)	p
(Intercept)	-1268.67	1239.97	(-4031.51, 1494.16)	-1.02	0.330
age	111.98	32.31	(39.99, 183.98)	3.47	0.006

```
lm.bw.female= lm(
 formula= weight ~ age,
 data = bw |> dplyr::filter(sex =="female"))
 lm.bw.female|> parameters()|> print_md()
```

Parameter	Coefficient	SE	95% CI	t(10)	p
(Intercept)	-2141.67	1016.05	(-4405.56, 122.23)	-2.11	0.061
age	130.40	26.19	(72.04, 188.76)	4.98	< .001

3.1 What is the key difference between this stratified approach and the interaction model above?

```
sigma(lm.bw)
```

```
[1] 180.6135
```

```
sigma(lm.bw.male)
```

```
[1] 200.9225
```

```
sigma(lm.bw.female)
```

```
[1] 157.7105
```

The standard error of each model is different. the model with only male has the largest standard error (200.92) compared to the full model with both male and female (180.62) and the model with only female (157.71). standard error change once we startified the model by sex.

```
library(survival)
lm.bw.strat=survreg(
Surv(time= weight) ~sex+ strata(sex) +sex:age + 0,
data = bw,
dist = "gaussian")
lm.bw.strat|> parameters() |> print_md()
```

Table 8: Fixed Effects

Parameter	Coefficient	SE	95% CI	z	p
sex (female)	-2141.67	927.52	(-3959.58, -323.76)	-2.31	0.021
sex (male)	-1268.67	1131.94	(-3487.23, 949.88)	-1.12	0.262
sex (female) × age	130.40	23.91	(83.53, 177.27)	5.45	< .001

Parameter	Coefficient	SE	95% CI	z	p
sex (male) × age	111.98	29.50	(54.17, 169.79)	3.80	< .001
female	4.97	0.20	(4.57, 5.37)	24.35	< .001
male	5.21	0.20	(4.81, 5.61)	25.53	< .001

The last two coefficients are the logs of $\hat{\theta}$ parameters for females and males, respectively. So we can get out the exponential version like so:

```
lm.bw.strat$scale
```

```
  female     male
143.9693 183.4163
```

If we multiply by sqrt(12/10), we will get unbiased estimates instead of MLEs:

```
lm.bw.strat$scale * sqrt(12/10)
```

```
  female     male
157.7105 200.9225
```

3.2

Compare these estimates to the ones we got from lm.bw.female and lm.bw.male above. Are they the same?

these estimates gave the same result to the one we got from m.bw.female (157.7105) and lm.bw.male(200.9225).

3.3

This survreg() approach has given us some extra information- namely, SEs and confidence intervals for the logarithms of the $\hat{\theta}$ estimates. If you exponentiate the CIs, you'll have get 95% confidence intervals for $\hat{\theta}$. Do that, and state your scientific conclusions about $\hat{\theta}_{male}$ and $\hat{\theta}_{female}$.

```
exp(4.57) #lower CI female
```

```
[1] 96.54411
```

```r
exp(5.37) #upper CI female
```

```
[1] 214.8629
```

```r
exp(4.81) #lower CI male
```

```
[1] 122.7316
```

```r
exp(5.61) #upper CI male
```

```
[1] 273.1442
```

```r
female.SE= sqrt(lm.bw.strat$var[5,5])
male.SE= sqrt(lm.bw.strat$var[6,6])

Female.CI=exp(log(lm.bw.strat$scale[1])+c(-1,1)*qnorm(0.975)*(female.SE))

Male.CI=exp(log(lm.bw.strat$scale[2])+c(-1,1)*qnorm(0.975)*(male.SE))

Female.CI
```

```
[1]  96.49817 214.79326
```

```r
Male.CI
```

```
[1] 122.9383 273.6458
```

the CI calculated in the previous question (lm.bw.strat) is in log scale, so we exponentiated the CI to get the CI in its original scale/normal scale. Based on the result above, the interpretation would be:

for female, we are 95% confidence that the true value of the birthweight falls between the range of 96.49 and 214.79.

for male, we are 95% confidence that the true value of the baby's birthweight falls between the range of 122.93 and 273.64.