HW2

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

Standard error is a statistic that approximates the standard deviation of a statistical sample population.

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

$$SE(\hat{\theta}) = \sqrt{VAR(\hat{\theta})} = \sqrt{E(\hat{\theta}^2) - [E(\hat{\theta})]^2}$$

2 Linear regression

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 wasmeasured in kilograms and height in meters), for thirty women.

```r
library(knitr) # compiling .qmd files
library(ggplot2) # graphics
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(dobson) # datasets from Dobson and Barnett 2018
library(conflicted) # check for conflicting function definitions
library(magrittr) # `%>%` and other additional piping tools
library(pander)
library(ggeasy)
library(parameters)
library(GGally)
```

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

```
data(cholesterol, package = "dobson")
print(cholesterol)
```

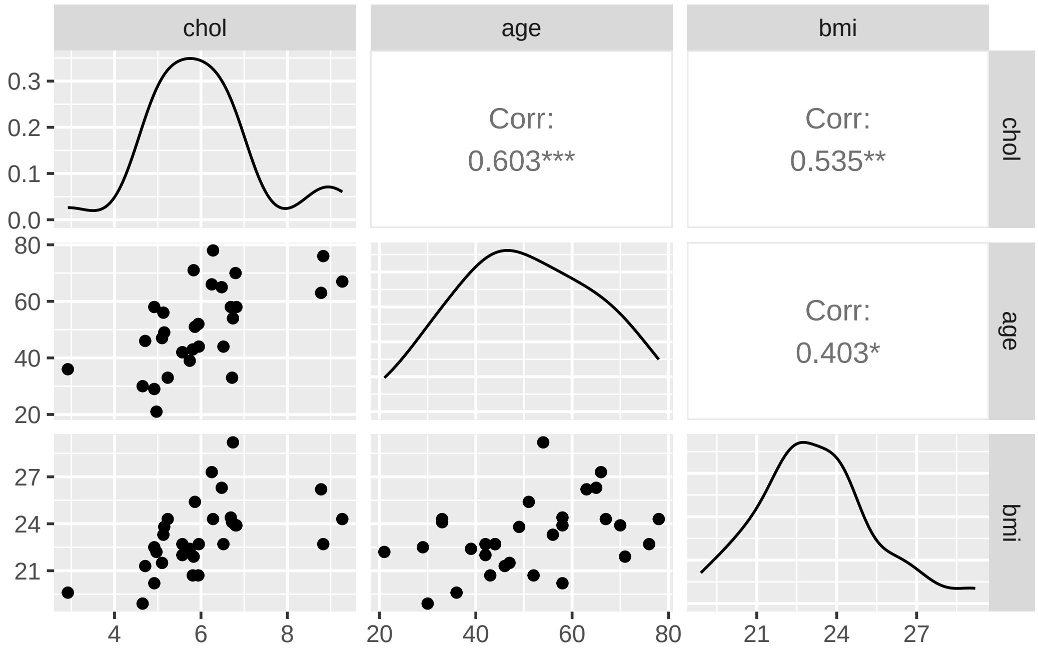
```
# A tibble: 30 x 3
    chol   age   bmi
   <dbl> <dbl> <dbl>
 1  5.94    52  20.7
 2  4.71    46  21.3
 3  5.86    51  25.4
 4  6.52    44  22.7
 5  6.8     70  23.9
 6  5.23    33  24.3
 7  4.97    21  22.2
 8  8.78    63  26.2
 9  5.13    56  23.3
10  6.74    54  29.2
# i 20 more rows
```

```
#> # A tibble: 30 x 3
#> chol age bmi
#> <dbl> <dbl> <dbl>
#> 1 5.94 52 20.7
#> 2 4.71 46 21.3
#> 3 5.86 51 25.4
#> 4 6.52 44 22.7
#> 5 6.8 70 23.9
#> 6 5.23 33 24.3
#> 7 4.97 21 22.2
#> 8 8.78 63 26.2
#> 9 5.13 56 23.3
#> 10 6.74 54 29.2
#> # i 20 more rows
```

2.1

Create scatterplots of the bivariate relationships between these variables.

```r
ggpairs(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.

```r
bmi_model <- lm(chol ~ age + bmi, data=cholesterol)
summary(bmi_model)
```

```

Call:
lm(formula = chol ~ age + bmi, data = cholesterol)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7619 -0.7353 -0.0205  0.3772  2.3717 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.73983    1.89641  -0.390  0.69951   
age          0.04097    0.01363   3.006  0.00567 **
bmi          0.20137    0.08876   2.269  0.03149 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.992 on 27 degrees of freedom
Multiple R-squared:  0.4654,    Adjusted R-squared:  0.4258 
F-statistic: 11.75 on 2 and 27 DF,  p-value: 0.000213
```

Intercept: -0.73983. Interpretation: the mean cholesterol level for female at age 0 and bmi 0 is -0.73983.

β_{age} = 0.04097. Interpretation: the mean difference in cholesterol level per 1 year in age difference is 0.04097 for female who have the same bmi.

β_{bmi} = 0.20137. Interpretation: the mean difference in cholesterol level per 1 unit difference in bmi is 0.20137 for female who have the same age.

Conclusion: The p-value of the coefficients age and bmi are smaller than the 0.05 significance value. Therefore, there is evidence to reject the null hypothesis that β_{age} and $\beta_{bmi} = 0$ at the 0.05 significance value. Age and bmi are both significant predictor variables in this regression model.

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.

```r
chol_in <- cholesterol
chol_in$age_bmi <- chol_in$age * chol_in$bmi
bmi_int <- lm(chol ~ age + bmi + age_bmi, data=chol_in)
summary(bmi_int)
```

```
Call:
lm(formula = chol ~ age + bmi + age_bmi, data = chol_in)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.56312 -0.72399 -0.05217  0.40839  2.40946 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.546427   8.947853  -0.732    0.471
age          0.154186   0.170975   0.902    0.375
bmi          0.457127   0.395281   1.156    0.258
age_bmi     -0.004933   0.007426  -0.664    0.512

Residual standard error: 1.002 on 26 degrees of freedom
Multiple R-squared:  0.4743,	Adjusted R-squared:  0.4137 
F-statistic:  7.82 on 3 and 26 DF,  p-value: 0.0007002
```

Intercept: -6.546427. Interpretation: the mean cholesterol level for female at age 0 and bmi 0 is -6.546427.

β_{age} = 0.154186. Interpretation: the mean difference in cholesterol level per 1 year in age difference is 0.154186 for female holding bmi at 0.

β_{bmi} = 0.457127. Interpretation: the mean difference in cholesterol level per 1 unit difference in bmi is 0.457127 for female holding age at 0.

$\beta_{age*bmi}$ = -0.004933. Interpretation: the interaction term shows the effect of the combined action of bmi and age in this regression model. The negative value means the effect of the interaction is less then the sum of the individual effects.

Conclusion: The p-value of the coefficients age, bmi, and age*bmi are bigger than the 0.05 significance value. Therefore, there is inadequate evidence to reject the null hypothesis that β_{age}, β_{bmi} and $\beta_{age*bmi}$= 0 at the 0.05 significance value. Age, bmi and age*bmi are not significant predictor variables in this regression model.

2.4

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

```r
mean(cholesterol$bmi)
```

```
[1] 23.18
```

```r
chol_cen_bmi <- cholesterol
chol_cen_bmi$bmi_cen <- chol_cen_bmi$bmi - 23.18
chol_cen_bmi$age_bmi_cen <- chol_cen_bmi$age * chol_cen_bmi$bmi_cen
cen_int <- lm(chol ~ age + bmi_cen + age_bmi_cen, data=chol_cen_bmi)
summary(cen_int)
```

```

Call:
lm(formula = chol ~ age + bmi_cen + age_bmi_cen, data = chol_cen_bmi)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.56312 -0.72399 -0.05217  0.40839  2.40946 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.049775   0.744817   5.437 1.06e-05 ***
age          0.039835   0.013878   2.870  0.00804 ** 
bmi_cen      0.457127   0.395281   1.156  0.25801    
age_bmi_cen -0.004933   0.007426  -0.664  0.51231    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.002 on 26 degrees of freedom
Multiple R-squared:  0.4743,	Adjusted R-squared:  0.4137 
F-statistic:  7.82 on 3 and 26 DF,  p-value: 0.0007002
```

I centered the bmi by minus the mean bmi.

Intercept: 4.049775. Interpretation: the mean cholesterol level for female at age 0 and at mean bmi 23.18 is 4.049775.

$\beta_{age} = 0.039835$. Interpretation: the mean difference in cholesterol level per 1 year in age difference is 0.039835 for female holding bmi at 23.18.

β_{bmi} = 0.457127. Interpretation: the mean difference in cholesterol level per 1 unit difference in bmi is 0.457127 for female holding age at 0.

$\beta_{age*bmi}$ = -0.004933. Interpretation: the interaction term shows the effect of the combined action of bmi and age in this regression model. The negative value means the effect of the interaction is less then the sum of the individual effects.

Conclusion: The p-value of the coefficients age is smaller than 0.05 significance level. Therefore, there is adequate evidence to reject the null hypothesis that $\beta_{age} = 0$, but we don't have adequate evidence to reject β_{bmi} and $\beta_{age*bmi}$= 0 at the 0.05 significance level. Age is a significant predictor variable in this regression model.

2.5

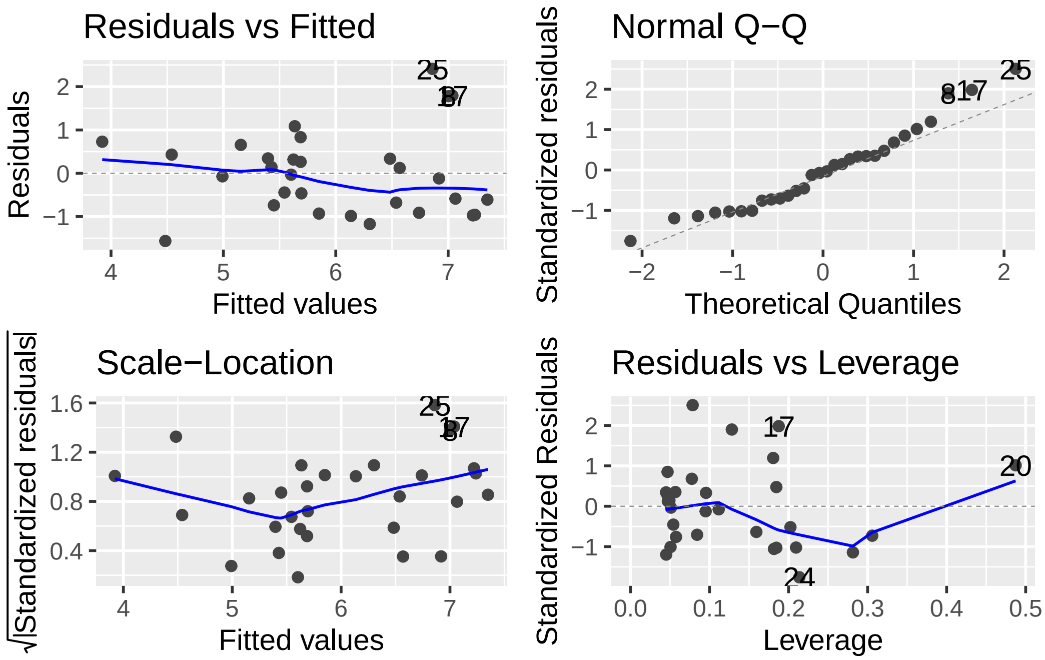
How did centering change your results?

The intercept changed from negative to positive, which is more meaningful compared to the previous model after centering. The standard errors of the intercept and β_{age} is smaller after centering, which means the coefficients are more centered now. And in the new model, age is a significant predictor variable.

2.6

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

```r
library(ggfortify)
autoplot(cen_int)
```



1. Normality: The qq plot shows that the model is not perfectly normal. There are some deviation from the ideal normal curve, but acceptable.
2. Correct functional form: The model seems correct.
3. Homoskedasticity: the variance seems constant, linear, and centered at 0.
4. Independence: we assume the observations are independent.

   Overall, it seems like a good model to use.

2.7

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

I rescaled age - minimum of age. There seems to be not a lot of changes. The standard errors of intercept and bmi appear to be smaller.

```
min(cholesterol$age)
```

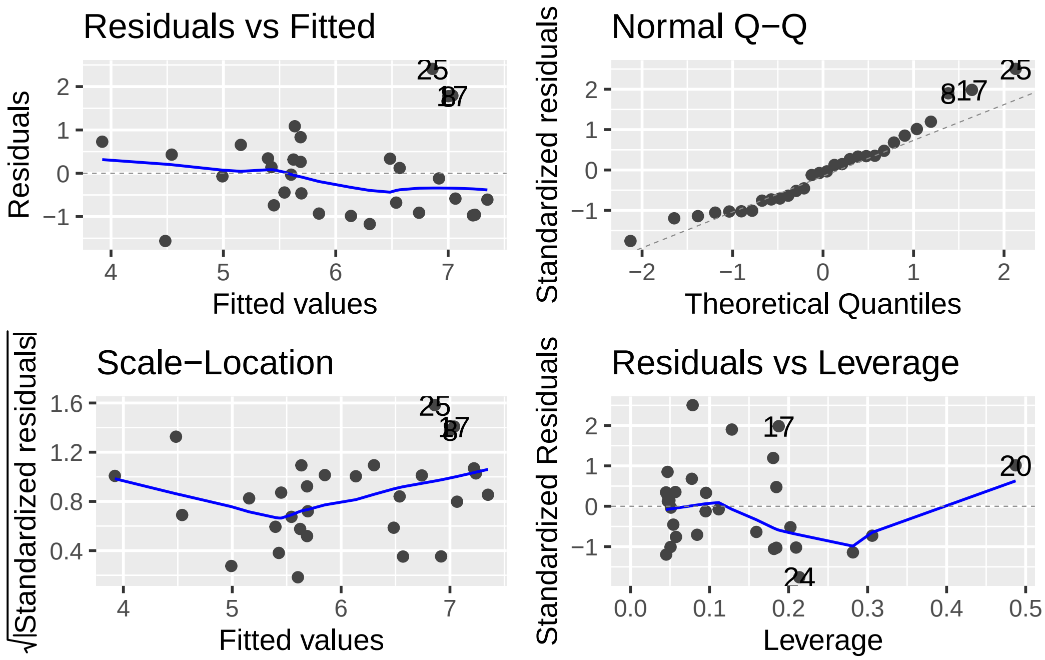
```
[1] 21
```

```
chol_age_cen <- chol_cen_bmi
chol_age_cen$age_cen <- chol_age_cen$age - 21
chol_age_cen$age_bmi_cen2 <- chol_age_cen$age_cen * chol_age_cen$bmi_cen
```

```r
cen_int_2 <- lm(chol ~ age_cen + bmi_cen + age_bmi_cen2, data=chol_age_cen)
summary(cen_int_2)
```

```

Call:
lm(formula = chol ~ age_cen + bmi_cen + age_bmi_cen2, data = chol_age_cen)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.56312 -0.72399 -0.05217  0.40839  2.40946 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.886319   0.471821  10.356 1.02e-10 ***
age_cen       0.039835   0.013878   2.870  0.00804 ** 
bmi_cen       0.353530   0.245970   1.437  0.16256    
age_bmi_cen2 -0.004933   0.007426  -0.664  0.51231    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.002 on 26 degrees of freedom
Multiple R-squared:  0.4743,	Adjusted R-squared:  0.4137 
F-statistic:  7.82 on 3 and 26 DF,  p-value: 0.0007002
```

```r
autoplot(cen_int_2)
```



3 Stratification

In the lecture notes, we fit the following stratified model for infant birthweights, with different slopes and intercepts for each sex:

```r
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

Parameter	Coefficient	SE	95% CI	t(20)	p
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

Alternatively, we could have fit two separate models, 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?

The standard error of the unstratified model is 180, and the standard error of the stratified model for male is 200, the stratified standard error of the model for female is 158. They are all different from each other, and the standard error of the unstratified model is approximately in the middle of the two standard errors of the stratified models.

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

```
[1] 180.6135
```

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

```
[1] 200.9225
```

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

```
[1] 157.7105
```

```r
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 4: 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	$< .001$
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	$< .001$

Note that the last two coefficients are the logs of ˆ parameters for females and males, respectively.We can get out the exponentiated versions like so:

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

```
  female     male 
143.9693 183.4163 
```

```
#> female male
#> 144.0 183.4
```

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
```

```
#> female male
#> 157.7 200.9
```

3.2

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

They are the same.

3.3

This survreg() approach has given us some extra information - namely, SEs and confidenceintervals for the logarithms of the ˆ estimates. If you exponentiate the CIs, you'll have get 95% confidence intervals for ˆ.Do that, and state your scientific conclusions about and .

95% CI of σ_{female}: $(e^{4.57}, e^{5.37}) = (96.544, 214.8629)$

95% CI of σ_{male}: $(e^{4.81}, e^{5.61}) = (211.7316, 273.144)$

Both σ_{female} and σ_{male} 95% confidence intervals tells that we are 95% confident that the true σ should be included by the the confidence interval.