HW 04: Multiple linear regression

Jwalin Patel

03/24/2021

```r
library(tidyverse)
library(broom)
library(knitr)
library(patchwork)
```

```r
sitting <- read.csv("data/sitting.csv")
```

Part 1

Question 1

We will use SLR to model MET and sitting as predictor and response variables, respectively.

$$Simple\ Linear\ Regression\ Model: sitting = \beta_0 + \beta_1 MET$$

```r
m1 <- lm(sitting ~ MET, data = sitting)
m1 %>%
  tidy(conf.int = TRUE) %>%
  kable (digits = 3, caption = "Prediction of the Reported hours per day spent sitting for subjects give
```

Table 1: Prediction of the Reported hours per day spent sitting for subjects given reported metabolic equivalent unit minutes per week

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	7.502	0.915	8.203	0.000	5.641	9.362
MET	0.000	0.000	-0.421	0.676	-0.001	0.001

Now we calculate R^2 to consider how well the model fits the relationship between reported hours per day spent sitting and the reported metabolic equivalent unit minutes per week.

```r
anova(m1) %>%
  kable(digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
MET	1	2.008	2.008	0.177	0.676
Residuals	33	373.592	11.321	NA	NA

We calculate R^2 using the formula:

$$R^2 = \frac{SS_{Model}}{SS_{Total}}$$

```r
2.008/(2.008+373.592)
```

```
## [1] 0.005346113
```

Thus, $R^2 = 0.005346113$. This means only around 0.53 percent of the variation in the reported hours per day spent sitting is explained by the reported metabolic equivalent unit minutes per week. Since the model only considers one predictor variable to explain variation in sitting hours, this is not a good representation of sitting hours.

Part 2

Question 2

```r
m2 <- lm(MTL ~ sitting, data = sitting)
m2 %>%
  tidy(conf.int = TRUE) %>%
  kable (digits = 3, caption = "Prediction of the Medial temporal lobe thickness in mm for subjects give
```

Table 3: Prediction of the Medial temporal lobe thickness in mm for subjects given Reported hours per day spent sitting

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.700	0.073	36.933	0.000	2.551	2.848
sitting	-0.023	0.009	-2.476	0.019	-0.042	-0.004

The coefficient is estimated as -0.023, with expected value ranging from lower bound of -0.042 and upper bound of -0.004. This means we are 95% confident that with every 1 hour increase in Reported hours per day spent sitting for subjects, we can expect the Medial temporal lobe thickness to reduce by 0.004 to 0.042 mm, and reduce by 0.023 mm on average.

Question 3

```r
sitting %>%
  summarize (mean = mean(age))
```

```
##       mean
## 1 60.37143
```

```r
sitting <- sitting %>%
  mutate(age_cent = age - 60.37143)
m3 <- lm(MTL ~ sitting + MET + age_cent, data = sitting)
m3 %>%
  tidy(conf.int = TRUE) %>%
  kable (digits = 3, caption = "Prediction of the Medial temporal lobe thickness in mm for subjects give
```

Table 4: Prediction of the Medial temporal lobe thickness in mm for subjects given Reported hours per day spent sitting, Reported metabolic equivalent unit minutes per week and Mean centered age

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.682	0.088	30.551	0.000	2.503	2.861
sitting	-0.021	0.010	-2.189	0.036	-0.040	-0.001
MET	0.000	0.000	0.078	0.939	0.000	0.000

term	estimate	std.error	statistic	p.value	conf.low	conf.high
age_cent	0.004	0.004	1.113	0.274	-0.004	0.012

Sitting: The coefficient of sitting is estimated as -0.021, with expected value ranging from lower bound of -0.040 and upper bound of -0.001. This means we are 95% confident that with every 1 hour increase in Reported hours per day spent sitting for subjects, we can expect the Medial temporal lobe thickness to reduce by 0.001 to 0.040 mm, and reduce by 0.021 mm on average, ceteris paribus. Age:

Question 4

```r
glance(m2) %>%
  select(r.squared, adj.r.squared, AIC, BIC)
```

```
## # A tibble: 1 x 4
##   r.squared adj.r.squared   AIC   BIC
##       <dbl>         <dbl> <dbl> <dbl>
## 1     0.157         0.131 -17.1 -12.5
```

```r
glance(m3) %>%
  select(r.squared, adj.r.squared, AIC, BIC)
```

```
## # A tibble: 1 x 4
##   r.squared adj.r.squared   AIC   BIC
##       <dbl>         <dbl> <dbl> <dbl>
## 1     0.189         0.111 -14.5 -6.73
```

1. Based on adjusted R^2 we would choose the first model. Even though the second model has higher R^2, it has a lower adjusted R^2 implying that there the tradeoff/penalty of adding MET and age_cent to our model is greater than their contribution to explanatory power.
2. Based on AIC, we choose the model with the smaller value of AIC which indicates better fit, which is the first model. This means adding MET and age_cent to our model did not reduce the sum of squares error enough to give us a lower AIC.
3. Based on BIC, we choose the model with the smaller value of BIC which indicates better fit, which is again, the first model. This means adding MET and age_cent to our model did not reduce the sum of squares error enough to give us a lower BIC. Note, BIC value is much higher in the second model also because BIC has a greater penalty for the addition of predictor variables to the regression model.

...