R-solution: Exercises Day 1

PSY8003

Matthias Mittner

spring 2022

Exercise 2: Simple regression

Interpret the Intercept as the level where the predictor is zero, i.e., a person with no work experience earned about 198.8 thousand NOK. The slope is the increase with the predictor, i.e., for each year of work experience, someone will earn 0.88 thousand NOK more.

```
loenn <- haven::read_dta("../data/loenn.dta")
summary(lm(loenn ~ erfaring, data=loenn))
```

```
Call:
lm(formula = loenn ~ erfaring, data = loenn)

Residuals:
    Min      1Q  Median      3Q     Max
-140.27  -79.80  -26.30   49.55  779.36

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 198.7606     9.2682  21.445   <2e-16 ***
erfaring      0.8807     0.4310   2.043   0.0416 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 110.3 on 462 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.008957,  Adjusted R-squared:  0.006812
F-statistic: 4.176 on 1 and 462 DF,  p-value: 0.04157
```

```r
summary(lm(loenn ~ kvinne, data=loenn))
```

```
Call:
lm(formula = loenn ~ kvinne, data = loenn)

Residuals:
    Min      1Q  Median      3Q     Max
-142.84  -80.14  -25.30   46.06  788.70

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  234.136      6.694  34.975  < 2e-16 ***
kvinne       -43.832     10.074  -4.351 1.66e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 108.6 on 469 degrees of freedom
Multiple R-squared:  0.0388,    Adjusted R-squared:  0.03675
F-statistic: 18.93 on 1 and 469 DF,  p-value: 1.663e-05
```

Exercise 3: Multiple regression

When interpreting, it is important to include the "keeping constant" qualifier for all included variables. I.e., For each year in work experience, someone would earn 2400 NOK more **controlling for gender and education**.

```r
loenn <- haven::read_dta("../data/loenn.dta")
summary(lm(loenn ~ erfaring + kvinne + utdann, data=loenn))
```

```
Call:
lm(formula = loenn ~ erfaring + kvinne + utdann, data = loenn)

Residuals:
    Min      1Q  Median      3Q     Max
-212.47  -55.40  -12.92   38.44  818.33

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -58.8169    28.7160  -2.048   0.0411 *
erfaring      2.4172     0.4131   5.851 9.41e-09 ***
kvinne      -50.1182     9.2419  -5.423 9.58e-08 ***
utdann       19.0848     1.8722  10.194  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 97.68 on 451 degrees of freedom
  (16 observations deleted due to missingness)
Multiple R-squared:  0.2294,    Adjusted R-squared:  0.2243
F-statistic: 44.75 on 3 and 451 DF,  p-value: < 2.2e-16
```

```
mod <- lm(loenn ~ kvinne + alder + fagfor + gift, data=loenn)

predict(mod, newdata=data.frame(kvinne=0, alder=40, fagfor=1, gift=1))
```

```
       1
256.9429
```

Exercise 4: Brain weight and total sleep across species

The two "weight" variables are highly correlated (collinear). Individually, they are both related to the total sleep. However, when inputting them simultaneously, the effect disappears.

```
sleep <- haven::read_dta("../data/total_sleep.dta")
cor(sleep[2:4])
```

```
              brainwt     bodywt totalsleep
brainwt     1.0000000  0.9558487 -0.3370815
bodywt      0.9558487  1.0000000 -0.3428373
totalsleep -0.3370815 -0.3428373  1.0000000
```

```
lm(totalsleep ~ brainwt, data=sleep)
```

```
Call:
lm(formula = totalsleep ~ brainwt, data = sleep)

Coefficients:
(Intercept)      brainwt
  11.116558    -0.002166
```

```
lm(totalsleep ~ bodywt, data=sleep)
```

```
Call:
lm(formula = totalsleep ~ bodywt, data = sleep)

Coefficients:
(Intercept)       bodywt
  11.047676    -0.004016
```

Exercise 5: Random data

When running enough of the regressions, some will be significant (approximately 5% when using the alpha-level of 0.05). The correlation matrix has some pretty large correlations. This is because of the low number of observations (N=20).

```
randomdata <- haven::read_dta("../data/random_data.dta")
lm(x1 ~ x43+x21, data=randomdata)
```

```
Call:
lm(formula = x1 ~ x43 + x21, data = randomdata)

Coefficients:
(Intercept)          x43          x21
   -0.05434      0.49716      0.30401
```

```
cor(randomdata[,1:10])
```

```
              x1          x2          x3          x4           x5          x6
x1   1.000000000  0.11695293 -0.07624683  0.04819686 -0.339109954 -0.09408270
x2   0.116952932  1.00000000  0.25331670 -0.29523441 -0.457247484 -0.06376500
x3  -0.076246825  0.25331670  1.00000000  0.01284676 -0.420154967 -0.01864681
x4   0.048196864 -0.29523441  0.01284676  1.00000000 -0.114017442  0.12023231
x5  -0.339109954 -0.45724748 -0.42015497 -0.11401744  1.000000000  0.29563117
x6  -0.094082703 -0.06376500 -0.01864681  0.12023231  0.295631170  1.00000000
x7  -0.152702852  0.09020149 -0.10225600 -0.23092627  0.132059502  0.06113374
x8   0.030415870 -0.31572654  0.10557823  0.15043456 -0.090024403  0.13995050
x9   0.019132528  0.13002219  0.27760513  0.29558188  0.001988955  0.13341252
x10  0.009460436  0.02845329 -0.04259574  0.25292105  0.137002912  0.06930351
             x7          x8          x9          x10
x1  -0.15270285  0.03041587  0.019132528  0.009460436
x2   0.09020149 -0.31572654  0.130022193  0.028453286
x3  -0.10225600  0.10557823  0.277605126 -0.042595739
x4  -0.23092627  0.15043456  0.295581877  0.252921046
x5   0.13205950 -0.09002440  0.001988955  0.137002912
x6   0.06113374  0.13995050  0.133412520  0.069303512
x7   1.00000000 -0.53114608 -0.081560057  0.333877580
x8  -0.53114608  1.00000000  0.184580080 -0.102162005
x9  -0.08156006  0.18458008  1.000000000  0.204044226
x10  0.33387758 -0.10216200  0.204044226  1.000000000
```