Los siguientes problemas están en Wooldridge, J.M, *Introductory Econometrics: A Modern Approach. 6th Edition. South-Western. CENGAGE Learning.* Las bases de datos se encuentran en www.cengagebrain.com (como se indica en la página XV del prefacio del libro).

```
library(wooldridge)
```

3. Problema 3, capítulo 3, pág. 94

The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u$$

where *sleep* and *totwrk* (total work) are measured in minutes per week and *educ* and *age* are measured in years.

i) If adults trade off sleep for work, what is the sign of β_1?

- It should be negative, to reflect the fact that when total work increases, sleep decreases, hence resulting in a tradeoff.

ii) What signs do you think β_2 and β_3 will have?

- The more educated a person is, the more likely it is that this person is older and has to work for longer, which means they give less importance to sleep; that is to say, β_2 should be negative Also, the older a person is, the less that person sleeps because of work and children, hence β_3 should be negative; however, it is possible that this coefficient turns positive after some threshold, once people retire so that they have more time to sleep and rest.

iii) Using the data in SLEEP75, the estimated equation is

$$\widehat{sleep} = 3{,}638.25 - .148\, totwork - 11.13\, educ + 2.20\, age$$
$$n = 706, R^2 = .113$$

If someone works five more hours per week, by how many minutes is sleep predicted to fall? Is this a large tradeoff?

- Recall *totwrk* is measured in minutes; this means 5 hours of extra work would be 300 minutes of extra work. Then, given that $\hat{\beta}_1 = -.148$, sleep is predicted to change by $300 * (-.148) = -44.4$ minutes. That is, when total work increases five hours, predicted sleep falls by 44.4 minutes.

  Is this a large tradeoff? We can try to provide an answer to this question through a rather heuristic fashion. Start by realizing, with the help from the code that's below this paragraph that, in average, people in the sample work 2,122.92 minutes. Also, in average they sleep 3,266.36 minutes.

```
# average work
mean(sleep75$totwrk)
```

```
## [1] 2122.921
```

```r
# average sleep
mean(sleep75$sleep)
```

```
## [1] 3266.356
```

With these numbers in mind, it is straighforward to see that five hours of extra work means increasing total work by about 14.13% for the average person in the sample; on the other hand, and according to the calculation previously made for the effect that this has in total sleep, the average person would decrease her minutes of sleep by just 1.36%. Hence, we could say the tradeoff isn't that large.

iv) Discuss the sign and magnitude of the estimated coefficient on educ.

- The fact that $\hat{\beta}_2 = -11.13$ tells us that by every extra year of education, in average, people sleep 11.13 less minutes. The negative sign of this is according to what was expected in (ii); the magnitude, on the other hand, is rather small. Following the same fashion, an extra year of education means increasing education by 7.82% for the average person, while the predicted decrease of sleep is only 0.34%. Again, this seems to be a small effect.

v) Would you say *totwrk*, *educ*, and *age* explain much of the variation in sleep? What other factors might affect the time spent sleeping? Are these likely to be correlated with totwrk?

- Considering what we concluded in previous exercises, I could not say that these variables explain much of the variation in sleep, given the figures and percentage changes we calculated. This is reinforced by the small value of the R-squared measure: $R^2 = 11.3\%$. That is to say, the proportion of the sample variation in the dependent variable explained by the independent variables is just about 11.3%.

4. Problema 5, capítulo 3, pág. 94

In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168.

i) In the model

$$GPA = \beta_0 + \beta_1\, study + \beta_2\, sleep + \beta_3\, work + \beta_3\, leisure + u$$

does it make sense to hold *sleep*, *work*, and *leisure* fixed, while changing *study*?

- Not really, since changing *study* changes the relative importance of the other activities within the structure considered (supposing the student only spends her time in those four activities: $study + sleep + work + leisure = 168$), so that hours spent in other activities should also change when *study* changes.

ii) Explain why this model violates Assumption MLR.3.

- It is a violation due to the fact that, by definition, there is a linear relationship among the independent variables: $study + sleep + work + leisure = 168$.

iii) How could you reformulate the model so that its parameters have a useful interpretation and it satisfies Assumption MLR.3?

- Omit one of the variables, say *leisure*, so that the model becomes $GPA = \beta_0 + \beta_1\, study + \beta_2\, sleep + \beta_3\, work + u$. Note that there is now no exact linear relationship among those variables. This can be done with any of the four variables. In this new model, $\hat{\beta}_j$ would give us the predicted effect on GPA that has replacing an hour of *leisure* (following our example) for an extra hour of x_j (the equality from (i) and (ii) still must hold).

8. Problema C1, capítulo 3, pág. 97

A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth weight that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother. An equation that recognizes this is

$$bwght = \beta_0 + \beta_1\, cigs + \beta_2\, faminc + u.$$

i) What is the most likely sign for β_2?

- Positive; it is highly likely that with a larger family income, the mother has a better access to nutrition and medical care.

ii) Do you think *cigs* and $faminc$ are likely to be correlated? Explain why the correlation might be positive or negative

- Yes, they could be correlated. If we suppose that with higher income comes better education and health habits, cigarettes consumption could be lower, hence finding a negative correlation. Nonetheless, with higher income, smoking could also increase, if we suppose that cigarettes are not a "bad" good (as in the opposite case) but rather a normal good, whose consumption increases with increased income.

iii) Now, estimate the equation with and without $faminc$, using the data in BWGHT. Report the results in equation form, including the sample size and R-squared. Discuss your results, focusing on whether adding $faminc$ substantially changes the estimated effect of *cigs* on *bwght*.

- We'll do this with the next chunk of code:

```
# model with family income
model_Faminc <- lm(bwght ~ cigs + faminc, data = bwght)
coefs1 <- round(
  model_Faminc$coefficients,
  digits = 2
)
# model without family income
model_NoFaminc <- lm(bwght ~ cigs, data = bwght)
coefs2 <- round(
  model_NoFaminc$coefficients,
  digits = 2
)
```

Results for the model with family income:

$$\widehat{bwght} = 116.97 - 0.46\, cigs + 0.09\, faminc$$
$$n = 1388, R^2 = 2.84\%$$

Results for the model without family income:

$$\widehat{bwght} = 119.77 - 0.51\, cigs$$
$$n = 1388, R^2 = 2.2\%$$

Adding $faminc$ changes slightly the estimator for *cigs*, and also increases very little the R-squared value.

9. Problema C7, capítulo 3, pág. 99

Use the data in MEAP93 to answer this question.

i) Estimate the model

$$math10 = \beta_0 + \beta_1 \, log(expend) + \beta_2 \, lnchprg + u,$$

and report the results in the usual form, including the sample size and R-squared. Are the signs of the slope coefficients what you expected? Explain.

- Code:

```
mathModel <- lm(math10 ~ lexpend + lnchprg, data = meap93)
coefs <- round(
  mathModel$coefficients,
  digits = 2)
mathModel <- summary(mathModel)
```

Equation, sample size and R-squared:

$$\widehat{math10} = -20.36 + 6.23 \, log(expend) - 0.3 \, lnchprg$$
$$n = 408, R^2 = 17.59\%$$

The model predicts that if spending per student increases by just 1%, the math pass rate (percentage of students passing the math exam) increases by 6.23%. Although the sign of this coefficient is as expected, its magnitude turned out impressively high. Now, turning our attention to the estimated slope of $lnchprg$, we certainly do not expect it to be negative, althogh some would say that the higher the elegibility for lunch program, the more likely it is that students come from low-income households; if this was the case, then we could say that u is correlated with $lnchprg$, which means that $\mathbb{E}[U|X] \neq 0$.

ii) What do you make of the intercept you estimated in part (i)? In particular, does it make sense to set the two explanatory variables to zero? [*Hint*: Recall that $log(1) = 0$.]

- It does not make sense to set $log(expend) = 0$, which is basically saying that spending per student is of only one dollar, which is unrealistic. Notwithstanding the case of $log(expend)$, it makes sense indeed to set $lnchprg = 0$, which means stating that elegibility in the lunch program is zero; this could be the case of schools in high-income zones. Setting only $lnchprg = 0$, gives an expression that, despite the negative value of the incercept, there is no value for $log(expend)$ in the sample such that $\widehat{math10}$ is below 30. Finally, it is worth noting that, in fact, the estimated intercept is not statistically different from zero (i.e., it is not significant); actually, estimating the model without an intercept results in a much better model if we judge by the value of the R-squared.

iii) Now run the simple regression of $math10$ on $log(expend)$, and compare the slope coefficient with the estimate obtained in part (i). Is the estimated spending effect now larger or smaller than in part (i)?

- Code:

```
mathModel2 <- lm(math10 ~ lexpend, data = meap93)
coefs2 <- round(
  mathModel2$coefficients,
  digits = 2)
mathModel2 <- summary(mathModel2)
```

Equation, sample size and R-squared:

$$\widehat{math10} = -69.34 + 11.16\, log(expend)$$
$$n = 408, R^2 = 2.73\%$$

The slope becomes larger if we stop consider *lnchprg* in the model.

iv) Find the correlation between $lexpend = log(expend)$ and *lnchprg*. Does its sign make sense to you?

- Code:

```
cor(meap93$lexpend, meap93$lnchprg)
```

```
## [1] -0.1927042
```

It does make sense: a higher level of spending per student helps increase the quality of services provided to students, including food services. This in turn reduces the eligibility of students to lunch programs, since they stop needing them as much. In other words, those schools were students come from lower income households (and where *lnchprg* is higher), likely spend less per student (*lexpend* is lower).

v) Use part (iv) to explain your findings in part (iii).

- It seems that, failing to consider *lnchprg* in the model, makes us overestimate the effect that *lexpend* has on *math*10.