Problem sheet 5

Quantitative Methods

Download the data set named `smoking.csv` and read it into R. The data set contains data on a sample of 1,000 people, with the following three variables:

- ⋆ `age`: the age of the person.
- ⋆ `smoker`: a binary variable indicating whether or not the person smokes.
- ⋆ `years`: the expected number of years the person has left to live.

1. Explore the data set.

2. Plot the association between `smoker` and `years` within age groups (e.g., split the `age` variable into quartiles).

3. Regress `years` on `smoker`. What do you find? How do you interpret this, especially in light of your answer to the previous question?

4. Now regress `years` on `smoker`, whilst also controlling for `age`. In R, this is done by using '`+ age`' inside the regression formula. The parameter estimate for `smoker` is now interpreted as the association between `smoker` and `years` when holding `age` constant at its mean. In other words, the new parameter estimate corresponds to

$$\mathbb{E}(\texttt{years} \mid \texttt{smoker} = 1,\, \overline{\texttt{age}}) - \mathbb{E}(\texttt{years} \mid \texttt{smoker} = 0,\, \overline{\texttt{age}}).$$

What do the new results reveal? What is your interpretation? What does this say about your earlier findings?

Thus far, we have ignored the fact that our models rely on sample data. Sample regression estimates — just like sample means or any other function of the data — are subject to variability: they can differ from one sample to another. Although we might believe that there is a fixed, "true" association between a group of variables, our estimate of this association is likely to vary depending on which data sample we happen to observe. In short, our parameter estimates can be understood as random variables that have a *sampling variance*. This variance quantifies how much our parameter estimates might change if we were to draw repeated hypothetical samples from the same population. In practice, we often work with the standard deviation of the sampling distribution rather than the variance. The standard deviation of a sampling distribution is called its *standard error*.

5. Run the same model as in the previous question, but this time save it as an object named `model`. Then type `summary(model)`. What do you see?

Let β be our parameter of interest and let $\hat{\beta}$ denote our *estimate* of that parameter. The standard error associated with $\hat{\beta}$ can be used to construct what is known as a *confidence interval* (CI), which measures the range of values of $\hat{\beta}$ within which (the true) β is likely to be located. A commonly used version of this is a 95% CI, which is defined as an interval I such that $\mathbb{P}(\beta \in I) = 0.95$. In other words, over a hypothetically repeated data generating process, a 95% CI "captures" or "traps" the true value β with probability 0.95. As a rough rule of thumb, it is constructed as $\hat{\beta} \pm 2 \times \mathsf{SE}(\hat{\beta})$, where $\mathsf{SE}(\hat{\beta})$ is the standard error associated with $\hat{\beta}$.

6. Use the results from the previous question to construct a 95% CI for each parameter estimate in `model`. Using the Central Limit Theorem (Lecture 4) and the 68-95-99% rule (Lecture 5), try to explain why we are 95% confident that β lies within 2 standard deviations from $\hat{\beta}$. How would you construct a 99% CI?

7. Use the `sample_n()` function to draw four random subsamples from the data set, with 20, 50, 100, and 500 observations, respectively. Redo the previous exercises for each subsample. What happens to your parameter estimates and confidence intervals as the sample size increases? (Tip: you can also use `confint(model)`.)

8. Watch this video, then this video. If you have time, skim this article. How would you interpret the rest of the output from `summary(model)`?