# Assignment 2

*Note:* Unless otherwise specified, data sets comes from Wooldridge's econometrics textbook. To use them, use the R package "wooldridge".

**Problem 1: Covariance and Correlation.** For this problem, use the Excel file called "Uninsured.xslx" in the "Datasets" folder on GitHub. This data set contains information on 20 municipalities in Massachusetts. For each municipality, the fraction of people without health insurance (frac uninsured) and the fraction of people declaring bankruptcy (frac bankrupt) are reported. I recommend using the read_excel() command from the tidyverse to import this data into R.

a. What is the covariance between these two variables? Make a nice-looking scatterplot of the variables' relationship. How does the covariance you reported jibe with the graph? Why do you think this is?

b. Create new variables for both bankruptcy and (un-)insurance that is measured in people (rather than percentages). Use the population variable to do so. Does this change the linear relationship? What is the new covariance? What does this teach you about covariance and data viz?

c. As discussed in class, the correlation is a unitless measure that resolves some of the problems discussed above. What is the correlation between the two original variables? Does this correlation change when you use the new variables (based on people, not percentages) instead? Why (or why not)?

d. What is the correlation between frac uninsured and your count of uninsured? What explains this? Why doesn't this cause a problem in calculating the correlation between your new variables? (A scatter plot—or multiple plots—may be useful here.)

e. However, even looking at the correlation can be misleading. Create a data set of 1000 observations and 2 variables: $X$ that ranges continuously over the interval $[0, 5]$ and $Y$ given by $Y = -X * (X - 5)$. Create a scatter plot of this relationship. What economic variables may have this relationship? What is the correlation between $X$ and $Y$, and what drives this result? When should I be careful of looking only at the correlation coefficient $\rho$?

- To create $X$, I recommend looking at the seq() command.

**Problem 2: Confidence Intervals in R.** Suppose we sample $n$ data points from a distribution $N(\theta, 36)$, where the value of the central mean $\theta$ is unknown.

a. Suppose that $n = 100$ and the sample mean is estimated to be $\bar{x} = 25$. What is the 90% confidence interval for $\theta$? The 95%? The 99%?

b. Repeat the exercise for $n = 1{,}000$. Why are the confidence intervals always narrower?

**Problem 3: Confidence Intervals and Probabilities.** People may mistake a confidence interval to mean there is a certain probability that the true parameter value lies in the interval. Let's explore more why this is incorrect.

a. **Simulation.** Suppose the true mean is $\mu = 10$. Write a loop that performs and stores this output 100 times:
   - Sample 25 observations from $N(\mu, 25)$ distribution.
   - Compute the 95% confidence interval given your sample.

How many times is $\mu$ in your interval? If you repeat the experiment infinitely, what should this probability be exactly?

b. **An updated simulation.** Suppose instead we have the following information about possible means and their (prior) probabilities:

| $\mu$ | 0 | 1 | 2 |
|---|---|---|---|
| $p(\mu)$ | 0.94 | 0.04 | 0.02 |

Suppose that I pick $\mu$ at random from this distribution, and sample $N(\mu, 25)$ 64 times. I tell you that the sample mean $\bar{x} = 1.6$. (Note: this means that the confidence interval is [0.375,2.825], can you verify this?)

- Compute the prior probability that $\mu$ is contained in that interval (using only the probability table)
- Compute the posterior probability that $\mu$ is in the interval, by filling in the following table. Are either the prior or posterior probabilities close to 95%?

| Hypothesis ($\mu$) | Prior ($P(\mu)$) | Likelihood given data $f(1.6\|\mu)$ | Numerator (prior multiplied by likelihood) | Posterior Probability $P(\mu\|\bar{x} = 1.6)$ |
|---|---|---|---|---|
| $\mu = 0$ | 0.94 | | | |
| $\mu = 1$ | 0.04 | | | |
| $\mu = 2$ | 0.02 | | | |
| **Total** | 1 | | | 1 |

*Notes:* to calculate the likelihood function, use the command `dnorm(1.6, μ, 5/8)` for each cell. The numerator is then the product of the prior and the likelihood. Finally, the posterior probability for each row is the numerator divided by the sum of all numerators (the total row).

c. Suppose we have data that a new cancer treatment increases patient survival by an average of 15 months longer than the conventional treatment, where the 95% confidence interval is [10,20]. Say whether each of the following statements is true or false. False means that the statement does not follow logically from the confidence-interval result.

- The true increase in survival is in the range [10,20] with 95% probability.
- The treatment increases survival time at all with at least 95% probability.
- [10,20] is an estimate of the true average increase in survival.
- The null hypothesis that the treatment does not affect survival time is likely incorrect.
- After 100 experiments, in approximately 95 of them the 95% confidence intervals would contain the true value of survival increases.
- We reject the null hypothesis of no improvement in survival at 5% significance.