

Instructions

Due Week 4 · Friday, September 18

Prof. Jack Reilly

Fall 2026

For this assignment, you must turn in two documents:

1. Your answers (as a PDF), and
2. A plain-text .do file containing the Stata code you used to arrive at your answers.

Please note: in this problem set, you may calculate answers to questions 1–5 by hand or using Stata. For questions 6–12, you must use Stata.

The DMV

The Commissioner of the NYS Department of Motor Vehicles is analyzing the impact of new software for processing new Drivers License applications. To do so, he divided DMV employees at the Syracuse branch into two groups: GROUP A processes Drivers License applications using the current system, whereas GROUP B processes Drivers License applications using the new software. Below is a table showing the time it takes (in minutes) for people in each group to process applications.

GROUP A	14	18	16	22	25	12	32	16	15	18
GROUP B	12	10	13	14	9	17	11	10	8	11

1. What is the mean, median, first quartile, third quartile, and range for each group?
2. Based on the above data, which software should the DMV decide to use?

The ACT and the SAT

There are two major tests of readiness for college, the ACT and the SAT. ACT scores are reported on a scale from 1 to 36. The distribution of ACT scores is approximately Normal with mean = 21.5 and standard deviation = 5.4. SAT scores are reported on a scale from 600 to 2400. The SAT scores are approximately Normal with mean = 1509 and standard deviation = 321. Tonya scores 1820 on the SAT. Jermaine scores 29 on the ACT.

3. If both tests measure the same thing, who has the higher score? Report the z-scores for both students.

Age at First Birth

The following questions are based on information from the National Survey of Family Growth (NSFG) from 2011–2013. This is the nationally representative survey with information on fertility in the United States. From this dataset, you learn that the age of a woman at the time of her first birth is $N(24, 12)$.

4. What is the probability that a randomly drawn person is older than 30 years old at the time of her first birth?
5. What percent of women are teenagers (i.e., 19 or younger) when they have their first birth?

Madison County Jail

For the following questions, use the dataset `Madison county jail data.dta` located on Blackboard. The unit of analysis is prisoner (each row in the dataset represents a different prisoner). The dataset currently has the following two variables:

- **Number of days:** This variable captures the number of days that a person has spent in jail.
- **Education level:** This variable captures the highest education level achieved by the individual imprisoned. The numbers in the value represent categories in the following way:
 - 1 = “Elementary School”
 - 2 = “Middle School”
 - 3 = “High School”
 - 4 = “College”

The director of the Madison County jail is planning to testify before the Wisconsin state legislature with the goal of obtaining more state funds for the county jail system. In preparing her remarks, the director plans to present data on the average number of days that prisoners remain in jail as well as data on the educational background of the individuals. The director asks you, as chief operations officer for the jail, to calculate descriptive statistics for a random sample of 250 prisoners processed in the last 9 months.

6. Obtain one appropriate measure of central tendency (mean, median, and/or mode) for “Number of days” and for “Education level” variables. (*Hint: remember that not all measures work for all types of variables.*) Explain what the measures mean in the context of these variables and why you chose this measure.
7. Create a plot for the variable “Number of days”. Make sure to use a type of plot that is adequate for the variable you are plotting. Paste the plot in your word document. Would you say that the distribution is symmetrical, skewed left, or skewed right? How do you know?

8. Create a plot for the variable "Education level". Make sure to use a type of plot that is adequate for the variable you are plotting. Paste the plot in your word document.
9. Under the current system, the costs for Madison County are the following: \$86 for processing a new inmate, and \$35 for every day a person is imprisoned. Using a linear transformation, create a new variable called "cost" that reflects this quantity for every prisoner in the dataset. Obtain a summary of that variable and paste it in your word document.
10. Regress your new "cost" variable (y) on "number of days" (x). Does the regression command accurately recover the linear transformation that you used in question 9?
11. Regress "number of days" (y) on "education level" (x). What is the intercept coefficient? What does this mean? What is the slope coefficient? What does this mean?
12. Use the results from the prior question. How many days of jail do we predict for those who have an elementary level of education (1), a middle school level of education (2), a high school level of education (3), and a college level of education (4)?
13. Can you think of any reservation you might have with the regression you calculated in question (11) and the predictions in question (12)? What might they be?