Use the Height tab. This dataset has 4 variables and *n* = 56 observations:

*female*: has values ‘1’ for females and ‘0’ for males

*sleep*: the ‘typical’ amount of sleep for the person (in hours)

*shoe*: the American shoe size of the person

*height*: the height of the person (in inches) Use this dataset to answer the following questions:

1. Run a simple regression of *Y* = *height* vs. *X* = *female*. Perform a hypothesis test to see if the variables are associated. What is the interpretation of the slope estimate, *b*1?
2. Run a simple regression of *Y* = *height* vs. *X* = sleep. Perform a hypothesis test to see if the variables are associated.
3. Run a simple regression of *Y* = *height* vs. *X* = *shoe*. Perform a hypothesis test to see if the variables are associated.
4. Run a multiple regression of *Y* = *height* vs. *X*1 = *female* and *X*2 = *shoe*. Perform a single hypothesis test to see if either of the *X*-variables are associated to *height*.
5. Run a multiple regression of *Y* = *height* vs. *X*1 = *female, X*2 = *shoe,* and *X.*= *sleep*. Perform a single hypothesis test to see if any of the *X*-variables are associated to *height*.
6. Using you multiple regression model in part (e), interpret the coefficient, *b*1, for the variable *female*.
7. Using you multiple regression model in part (e), make a prediction of your height based on the values of the predictors for you (sex, shoe size, and typical amount of sleep). Be sure to clearly label what your *x*-values are.
8. Create two plots for the multiple regression model from part (e): 1) the histogram of the residuals and
9. the residuals-versus-fitted scatterplot (remember: you can create the residuals variable using the “predict residuals, resid” command in Stata). Use these plots to comment on the validity of the assumptions for this model (be sure to include these plots).
   1. Compare your results in parts (a) through (e) [**using a table to organize your answer is fine**]. Make sure to compare R2 values and note any major changes in significance and/or signs of slopes, especially that related to the variable *female*. Which model would be your choice as a ‘best predictive model’?
10. A study was designed to determine whether different types of writers die at different ages on average. Three categories of writers were examined: novelists, poets and nonfiction writers.

The age of death for these North American writers are found in the dataset ‘writers.csv.’ For this dataset, run a binary multiple regression using the variables *novelist* and *poet* as predictors of age of death.

1. Write out the regression model statement (in terms of *Y*, *β*’s, and *X*’s). Be sure to label what each of the variables represents.
2. What is the reference group? What is the estimated difference in age at death between poets and non-fiction writers? How about between poets and novelists?
3. Perform the appropriate *F*-test to determine whether there is a difference in average age of death for the 3 types of writers.
4. Poet William Butler Yeats once wrote, “She is the Gaelic muse, for she gives inspiration to those she persecutes. The Gaelic poets die young, for she is restless, and will not let them remain long on earth.” Using the “test” command in Stata, perform an appropriate test to determine whether the average age of death for poets is different than the other 2 groups combined (you can weigh the other two groups equally or based on their sample sizes).
5. For parts (a) through (d )below, **set-up** a test of hypothesis using a 0.05 level of significance. Your tests of hypothesis set-up should include: 1) a statement of the type of procedure (one- sample t-test for means, chi-squared test for association, etc…), 2) a statement of your null and alternative hypothesis, 3) whether you are performing a one or two-sided test, and 4) the general formula for the test statistic (**no need to calculate it or plug in the numbers**). If you are having difficulty deciding the type of procedure for a problem, it’s recommended to consult the *Roadmap to Inference* document on the course website.
6. The nutrition label on Kellogg’s Rice Krispies cereal claims that their product contains 3 grams of sugar per serving. To test this claim a consumer research organization conducted tests on a random sample of 16 boxes of Kellogg’s Rice Krispies cereal from grocery stores around the US. The tests showed an average of 3.25 grams of sugar with a standard deviation of 0.4 grams for the 16 servings tested. Does this consumer research organization study support the claim by Kellogg’s?
7. Carbon monoxide emissions are measured every 2 years for automobiles in Massachusetts during mandatory vehicle inspections. Carbon Monoxide levels (CO grams per mile) were collected for 15 Fords, 20 Hondas, and 30 Toyotas. The mean and standard deviation of the Ford CO levels were 14.2 and 4.2, for the Hondas were 11.9 and 4.1, and for the Toyotas were 12.3 and 4.3. Is this sufficient evidence of a difference in CO emissions between these three types of autos?
8. The Civil War Battle at Antietam was a one-day battle on September 17, 1862

near Sharpsburg, Maryland. Antietam was the bloodiest single day of battle in American history. It resulted in more than 22,000 American casualties, more than three times as many casualties than on June 6, 1944 – D-Day, the so-called “longest day” of World War

II. Battle casualties are classified into three categories: 1) dead, 2) wounded, and 3) missing or captured. The approximate number of casualties at Antietam on September 17, 1862 is given below. Is there evidence that casualty type is associated with Army?

|  |  |  |  |
| --- | --- | --- | --- |
|  | Union (North) | Confederate (South) | Total |
| Killed | 2100 | 1530 | 3630 |
| Wounded | 9550 | 7750 | 17300 |
| Missing/captured | 750 | 1020 | 1770 |
| Total | 12400 | 10300 | 22700 |

1. Government agents in Massachusetts wanted to see if older cars have worse carbon monoxide emissions than newer cars. During mandatory vehicle inspections, carbon monoxide levels (CO grams per mile) were measured for 50 cars along with the ages of the cars (in years). The mean and standard deviation of CO levels were 13.2 and 4.2, and 4.5 and 2.3 for the age of the cars in years, with a correlation of 0.32 between CO and age. Is there evidence that older cars have increased CO emissions?