**Doing Inference for Regression**

This lab should look familiar. We’ve played with all of these problems before when we saw regression for the first time, but now we’ll also explore the residuals from the fit and use them to assess the assumptions of Normality and constant variance. If we can verify those assumptions, we can also calculate CIs for the slope parameters.

In general, the R code we’ll use to achieve this looks like:

model=lm(*y\_var* ~ *x\_var*, data=*dataset\_name*) ##fit the model

par(mfrow=c(2,2)) ##setting up plotting window

plot(model) ##plots to assess Normality and constant variance

confint(model) ##if we verify the assumptions, calculate CIs

In this code above, you need to actually type in the *y\_var*, *x\_var*, and *dataset\_name.* For example, if we wanted to use a dataset called LafGPA to predict a student’s GPA based on his/her hours of studying (HrsStudy), we would type:

model=lm(GPA~HrsStudy,data=LafGPA)

**#1** At the course website, access the folder for today’s lab, where I have placed a dataset called introReg\_ex3.R, which, on a Mac, you can double-click directly to load it into R (with the load command produced in the console). On a Windows machine, use the load command again (if the data are saved in the same location). For a statistics class project at a large northeastern university (Penn State), a student examined the relationship between *x =*  body weight (in pounds) and *y =*  time to chug a 12-ounce beverage (in seconds). The student collected data from 13 individuals.

**a.** Produce a scatterplot of the measurements. The y-variable is “chug time” and the x-variable is weight. In R, we can use:

with(ex3,plot(x=Weight,y=ChugTime, main='Relationship between Weight and \nTime to Chug a 12oz. Beverage')) ## Wanna split that title into 2 lines??

- Describe the main features of the graph. Specifically, is there a negative or a positive association? Does the pattern look to be linear or curved? Are there any outliers? If there is an outlier, describe where it’s located on the graph.

**b.** In general, outliers should not be thrown out unless there’s a good reason, but there are several reasons why it may be legitimate to conduct an analysis without them. In this case, let’s ignore the data point for the heaviest person and then determine a regression line for the remainder of the data.

ex3\_tmp=subset(ex3,ex3$Weight<max(ex3$Weight))

with(ex3\_tmp,summary(lm(ChugTime~Weight)))

Write the estimated regression equation (look in the Estimates column of the Coefficients table for the intercept and slope, in that order).

**c.** Write a sentence that interprets what this slope says about the relationship between chug time and body weight.

**d.** Now, this is just the estimated slope based on these 12 individuals. If we did this experiment with a different sample of n=12, we would get a different slope. What are plausible values for the true population slope? First, check the assumptions with the appropriate plots. The plot in the top left corner can be used to assess the assumption of constant variance, and the QQ-plot in the top right corner can be used to assess the assumption of Normality. Discuss these two plots and the features thereof that lead you to make these assumptions or not.

**e.** If you were satisfied with the assumptions in part **d.** above, you would then be able to calculate CIs for the intercept and slope parameters. Let’s calculate, report and interpret a 95% CI for the population slope parameter.

**#2** At the course website, access the folder for today’s lab, where I have placed a dataset called introReg\_ex4.R, which, on a Mac, you can double-click directly to load it into R (with the load command produced in the console). On a Windows machine, use the load command again (if the data are saved in the same location). The dataset includes teenage mother birth rates and poverty rates for the 50 states of the U.S. and the District of Columbia. The variable ***PovPct***is the percent of a state’s population in 2000 living in households with incomes below the federally defined poverty level. The variable ***Brth15to17***is the birth rate for females 15 to 17 years old in 2002, calculated as births per 1000 persons in this age group.

**a.** Plot ***Brth15to17***(as y-variable) versus ***PovPct*** (as x-variable).

with(ex4,plot(PovPct,Brth15to17))

Describe the direction of the relationship, comment on whether the pattern appears to be linear or curved, and comment on whether there are any outliers.

**b.** Determine a regression line for these data with ***Brth15to17***as the y-variable and ***PovPct*** as the x-variable. Write the equation.

**c.** Write a sentence that interprets what this slope says about the relationship between ***PovPct*** and ***Brth15to17*** .

**d.** Justifying calculating a CI for the slope here is more difficult. This isn’t a sample of states – this is data from all states. So in a sense this is a census, but we might think of this data in 2000 as a sample in time. That is, we can use this to tell us about past or future slopes. Let’s do that, and still try to figure out the “true” relationship between Poverty rate and Teen Birth rate. First, check the assumptions with the appropriate plots. The plot in the top left corner can be used to assess the assumption of constant variance, and the QQ-plot in the top right corner can be used to assess the assumption of Normality. Discuss these two plots and the features thereof that lead you to make these assumptions or not.

**e.** If you were satisfied with the assumptions in part **d.** above, you would then be able to calculate CIs for the intercept and slope parameters. Let’s calculate, report and interpret a 95% CI for the population slope parameter.