**Lab 5: Regression: Internet Use and Babies**

The World Bank collects data on many variables related to world development for countries throughout the world. Two of these are Internet use, in number of users per 100 people, and birthrate, in births per 1000 people, on 106 countries that have data available for both variables. (These data were collected in 2011; This lab is an adaptation of textbook exercises from Chapter 2.)

Read in the .csv file INBIRTH, which you will find in the Data folder. To avoid having to refer to variables by data\_frame\_name$variable\_name, use the attach command to attach the data frame:

* attach(INBIRTH)

After attaching the data frame, you can refer to the variables simply by their names; for example, BirthRate and Users (without first specifying the name of the data frame). **However, at the end of your session, remember to detach the data frame:**

* detach(INBIRTH)

Include all graphs and output on which your answers are based as part of your solution.

1. a. Make a scatterplot of BirthRate versus Users. Would you describe the relationship as positive or negative? Explain.

b. Determine the correlation between BirthRate and Users. Based on the correlation, would you interpret the strength of the linear relationship as strong, moderate, or weak?

c. Fit a least-squares line to predict BirthRate from number of Internet Users. Save the results from the lm command as LinearModel.

LinearModel <- lm(BirthRate ~ Users)

Write the equation of the least-squares line. Interpret the slope and y-intercept in the context of these data.

d. Determine the coefficient of determination and interpret its value in the given context.

2. In question 1, you calculated a correlation and coefficient of determination. **You did this before you knew whether a linear model was adequate to describe the pattern of these data, which is never a good idea!** Correlation measures the strength of a *linear relationship* – so, first you need to establish that a *linear model is adequate* to describe the pattern between BirthRate and Users. If not, you’ll need to move on to a different model. One good way to judge the adequacy of a linear model, is with a residual plot.

a. Make a residual plot of residuals versus Users. Superimpose the line *y* = 0 is on your plot.

plot(LinearModel$residuals ~ Users, ylab = "Residuals", xlab = "Users",

pch = 16, col = “red”)

abline(0,0)

b. Make a residual plot of residuals versus fitted values. How does the pattern in this plot differ from the pattern in (a)?

plot(LinearModel$residuals ~ LinearModel$fitted.values, ylab = "Residuals",

xlab = "Fitted Values", pch = 16, col = "blue")

abline(0,0)

c. Based on your plots in (a) and (b), is the least-squares line adequate to describe the pattern in the data? Explain.

d. Use the command below to produce 4 graphs involving the residuals. Just look at the first graph. Does this plot support your answer to (c)? Explain.

layout(matrix(1:4,2,2))

plot(LinearModel)

e. Based on your answers above, should you have computed a correlation and interpreted its value for these data? Explain.

3. a. Create a new variable by taking the natural log of BirthRate. Attach this to the data frame INBIRTH.Then use the head(INBIRTH) to print out the first six cases of the modified INBIRTH data frame.

logBirthRate <- log(BirthRate) **Note:** The command log( ) takes the natural log of the

input; it calculates ln(*x*).

b. Make a scatterplot of logBirthRate versus Users. If your plot only shows up in the upper quarter of your plots window, use the following command and then rerun your plot command:

layout(matrix(1:1))

c. Describe the relationship shown in your scatterplot from (b).

d. Fit a linear model to logBirthRate versus Users. Store the results as LogModel. Write the equation using the variable names in the equation (as opposed to *x* and *y*).

LogModel <- lm(logBirthRate ~ Users)

LogModel

e. Make a residual plot to check the adequacy of this model. Based on your residual plot, is this model adequate to describe the pattern in the logBirthRate-Users data? In particular, does this plot look better than the one from question 2(a)? Explain.

plot(LogModel$residuals~Users, pch = 16, ylab = "Residuals",

xlab = "Users", col = "purple")

abline(0,0)

4. In question 3(d) your model should look something like:

ln(BirthRate) = 3.52249 – 0.01489(Users)

OR

ln(*y*) = 3.52249 – 0.01489*x*

a. Transform the equation above into the form below (using values for *a* and *b* from the equations above). Show your work.



OR



b. Make a scatterplot of the data BirthRate versus Users. Overlay a graph of your model from (a).

Step 1: Create some *x* values (Users) in small increments from the minimum User value to the maximum User value. Then, calculate the predicted values from model fit in question 3(d) (and stored as LogModel).

Uservalues <-seq(min(Users), max(Users), 1)

Predictedvalues <- predict(LogModel, list(Users = Uservalues))

Step 2: The predicted values in Step 1 are log(BirthRate) or ln(BirthRate) (since log(*x*) in *R* is log base *e*, or the natural log, ln(*x*)). Transform the predicted values from the model (log(BirthRate) into Birthrates. (In other words, undo the *ln* using the inverse function exp(*x*))

PredBirthrates <- exp(Predictedvalues)

Step 3: Plot the data: BirthRate versus Users and then the points generated in Steps 1 and 2 connected by line segments. (Since there are a lot of points in Steps 1 and 2, the result will appear as a smooth curve.)

plot(BirthRate ~ Users, xlab = "Internet Users per 1000 people",

ylab = "Birthrate per 100 people", pch = 16, col = "blue")

lines(Uservalues,PredBirthrates, col = "red")

c. Based on your model in (a), predict the birthrate for a country in which the number of users in 1000 people is 35. You can use your calculator for this. (Can you figure out how to get *R* to do this?) Is this an example of interpolation or extrapolation?