**Introduction to Biostatistics (171:161)**
**Breheny**

# Lab #5

In lab # 4 we were introduced to the "Tips" dataset and explored some its variables by looking at summary statistics, plots and histograms. In this lab we will look more closely at the relationships between different continuous variables using regression. We will also learn how to create and manipulate variables to further explore a data set.

## 1 Regression

When you are interested in predicting one variable from another variable you can create a regression line. The procedure in SAS that calculates the intercept and slope of this line is called `PROC REG`. As mentioned in lecture, regression is a big topic, and `PROC REG` comes with hundreds of options, we will only use a few of them.

In `R`, the function will use is called `lm`, which stands for "linear model". The reason regression is so useful is that it can be generalized to all kinds of settings by the notion of a model, as alluded to be its name in `R` (we'll discuss models further at the end of this course).

So, even though it may seem a little grandiose calling a straight line a "model" for tipping, in an abstract sense, that's what it is. So, in `PROC REG`, you specify the regression line with a `MODEL` statement:

SAS:

```
PROC REG DATA = Tips;
  MODEL Tip = TotBill;
RUN;
```

`R`:

```
lm(Tip ~ TotBill)
summary(lm(Tip ~ TotBill))
```

The SAS output tells you the estimated values of the intercept and slope ("Totbill"), along with plenty of other information that probably doesn't mean anything to you (yet). The `R` output just gives you the intercept and slope (if you want more information, you have to ask for it using the summary function). The output tells us that for every additional dollar that a meal costs, the waiter can expect to get 10 and a half cents more on his tip. Note that we could have calculated this from the output `PROC CORR/cor` gave us as well:

$$.68\left(\frac{1.38}{8.90}\right) = .105$$

So, what tip should the waiter expect on a $25 meal?

$$\begin{aligned}\alpha + \beta(25) &= .92 + .105(25) \\ &= \$3.55\end{aligned}$$

Of course, perhaps our prediction of the waiter's tip should depend on whether the table is in the smoking section or not. One approach would be to fit regression lines separately and make separate predictions for the two groups (which I would encourage you to do for practice). However, for a lot of reasons, making predictions based on multiple variables gets complicated – this is the kind of thing you would explore in the Design and Analysis of Biomedical Studies course, should you decide to take it.

## 2 Creating New Variables

Many times it is useful create new variables out of existing ones. This allows us to do things like create variables showing rates or percentages or create variables for the sum or difference of existing variables.

For example in the "Tips" dataset we might be interested in the tipping rate. In the United States is customary to tip 10-20%. However this depends on many factors including the waiter, the customer, the restaurant, etc. This dataset is particularly interesting because all the data was collected using the same waiter and restaurant; meaning that it should provide a good look at the variability in tipping primarily due to the customers.

So to analyze the tipping rate we first need to create it. In SAS creating a new variable requires a data step. In R will just create a new variable seperate of our data set.

SAS:

```
DATA tips;
SET tips;
TipRate = 100*Tip/TotBill;
RUN;
```

R:

```
tip_rate <- 100*Tip/TotBill
```

Note that the SAS data step shown above will start with the dataset "Tips", add the new variable "TipRate" and then overwrite the old "Tips". Use caution when doing this; if you had called the new variable the same name as an existing variable the old variable will be lost.

Now lets take a look at what TipRate looks like:

SAS:

```
PROC SGPLOT DATA=tips;
HISTOGRAM TipRate;
RUN;
```

R:

```
histogram(tip_rate)
```

There is no limit to the number of new variables you can create, and you can build upon new variables that you have already created. For example, if you wanted to create a variable called BigTip that records whether the tip was above 20%, you could do so as follows:

SAS:

```
DATA tips;
SET tips;
BigTip = TipRate > 20;
RUN;
```

R:

```
BigTip <- tip_rate > 20
```

Notice that the variable "BigTip" takes on values of either "0" or "1". This type of variable is called an indicator variable, and is commonly used in statistics. You can easily calculate the number of "successes" of an indicator variable simply by taking its sum.

For the remainder of lab you will answer some questions about the "Tips" data set. Below is a list of some questions that one might wonder about the relationship between the data set's variables. This is not an exhaustive list, but some examples. Choose a few of them and solve them in a way that allows you to see the relationship graphically, and also in way that you could numerically report the answer to someone.

- How does tip rate change with total bill? Do small bills have more variation in tip rate than large bills? Are people proportionally more generous with smaller bills?
- Do smokers tip differently than nonsmokers?
- Suppose that an equal number of men and women dine at the restaurant. Are men more likely to pick up the check than women? Does this depend on whether the meal is lunch or dinner?
- Does tipping behavior change at lunch versus dinner?
- Does tipping behavior differ by days of the week?

This concludes lab #5.