Lab 4: Scatterplots and Regression

In this lab we examine in greater detail a variety of scatterplots, draw regression lines in them, and discuss correlation.

Loading a new dataset

We will start our explorations by loading a new dataset, and refreshing that process.

- Find the datasets section in the web site, and find a link called drivingAll. Do NOT click it.
- Right-click the link, and choose "Save link as" or "Save target as", and save the file somewhere on your computer.
- Start SPSS, and choose "File -> Read Text Data". Find the file, and click Open.
- Now we go through the 6 steps of the wizzard.
    1. You should be seeing the first line with variable names, then lots of rows of data. Click Next to go to next step.
    2. In the second step you must switch the checkbox about "variable names included at top of the file". It should be at Yes. Click Next.
    3. Third page is always OK.
    4. On step 4, you will see that some things are out of place. This is because it has chosen to use Space as a delimiter in addition to Comma. Uncheck the "Space" checkbox, and it should all line up. Click Next.
    5. This time we have work to do on step 5. Some of the variable columns are special, they are dates or time. We have to tell it about them.
        - You will see right now that the "Day" column is selected. It currently thinks the format is "String". Choose instead "Date/Time". It shows you a list of formats. Find the format that says "yyyy/mm/dd".
        - Then click on the second column, "WeekDay". Change that to "Date/-Time" as well, and choose "Day of Week" from the list of formats to the right.
        - Next up is the "LeaveTime" variable. It should be "Date/Time" again, and with format "hh:mm".
        - Do the same for the "ArrTime" variable.
        - Click Next.
    6. 6th step is always OK. Click Finish
- Your data should now be set! SAVE IT!

This dataset is data I recorded over an entire semester. I resided in Louisville at the time, and I recorded for every day information about my driving from Louisville to Hanover and back. The "direction" column indicates where I was headed, work or home.

We will come back to this dataset later. For now save it to a location that you can find it later, and we will bring up another dataset to work with.

- Find a dataset called “Cigarettes and cancer”. Right-click and save it.
- Start the Read Text Data wizard like before.
- Tell it about the included Variable names in step 2.
- In step 4 you will need to again uncheck the “Space” checkbox.
- Everything else should be fine. Finish the wizard.
- Save!

These are measurements relating the amount of per capita sales of cigarettes in each state vs the incidence rate of certain types of cancer in the state.

We will start by considering scatterplots of cigarette sales and other types of cancer.

Doing scatterplots with regression lines

Go to the Chart Builder, choose Scatter/dot, choose the first graph, then put cigarettes on the x axis, and “blad” (cancer of the bladder) in the y axis. You should be seeing what seems like a somewhat weak positive relationship.

Double-click the graph to edit it, and under elements choose “Fit Line at Total”

When you do this, it should give you a regression line, with an equation on it, “y=1.09 + 0.12*x“. You should also see that it says what”r-squared" is, 0.495. so this relation “explains” only about 50% of the variation in bladder cancer amounts.

Another way to “assess” how good a line fit is is to choose the “Loess” option for Fit Line, instead of “Linear”. Do that now, and it should show you a curve that fits the data as best as a curve could do. If that deviates a fair bit from being linear, then it might be an indication that a linear regression is not the best tool.

Do the same to compare between cigarettes and the other types of cancer.

- Scatterplot, assess the relation.
- Add linear line, look at r-squared.
- Switch it to Loess curve.

1. Which types of Cancer appear to have the strongest relation to Cigarette sales?
2. There are some states that have a very large number of cigarette sales per capita, but correspondingly low incidence of cancer. Sort the data to find out who those states are (note that the “coast guard” prefixes[1] are used). Try to find a reasonable explanation for the large number of per-capita cigarette sales (i.e. cigarette sales per person residing in the state).

Computing Correlations

Let us compute the correlation coefficients for each pair of variables. In fact we will have SPSS create a table of all pairs of correlations.

[1] http://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations#Coast_Guard_vessel_prefixes

- Go to “Analyze -> Correlate -> Bivariate”
- Select all variables and move them to the right
- Click OK.

You should now be seeing a table of all pairwise correlations. Look especially at those that are close to or more than 0.7. Which pairs of variables have a correlation of around 0.7 or more?

Working with a subset of the Cases

It seems that there are a couple of states that are unusual, we should try to remove them. We do not want to delete them of course, we just want to tell the system to ignore them for a while. This goes under the name of “**Select Cases**”.

- Go to “Data -> Select Cases”. It’s near the end.
- Right now it should say “All cases”. Switch it to “If condition is satisfied”.
- Click on the “If” button. This brings up an “equation editor”.
- We need to write in the box on the top right which cases we want to have included. The formula would be “CIG < 39”. You can either type it in or drag and click on things.
- Click on Continue, then OK.
- You should now see in your data that the row numbers for those two cases have been crossed out. Everything we do from now on will ignore those rows.
- When you want to get back to looking at all values, go to “Data -> Select Cases” again and switch to “All cases”.

Go ahead and do a comparison of cigarette sales and lung cancer now. Does the fit look any better?

Let’s identify some of those points that are outliers by deviating from the pattern. Go back to Chart Builder, but this time use “Groups/Point ID” and activate the “Point ID label”, and put “STATE” in the Point ID spot in the graph. You should be able to see a number of outliers now, identified by state name. Try to think of reasons for those.

- Some states have large cancer incidence compared to per-capita cigarette sales.
- Some states have small cancer incidence compared to per-capita cigarette sales.

Think about explanations for both scenarios.

Further explorations

The “cereal” dataset has a wealth of scalar variables to look at, looking at dietary features of various cereal products. Use any remaining time you have to download it, import it using Read Text Data, and explore relations between its scalar variables.