**Week 6 - CPaT Stats: Correlation and Linear Regression**

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Please turn in a hardcopy of just the questions with your answers inserted in **BOLD** text. Bring this completed lab assignment to next week’s lab (week 7).

You will upload the associated Excel fine to the fileshare - in

Workspace/\_StatsLabReports/week\_6/

wk6\_stats\_data\_cushingSkomraWeiss.xlsx (or .xls)

where you insert your last names in place of “cushingSkomraWeiss”

Part 1: Correlation & Regression

To compare two continuous variables (either when you think one variable depends on the other, or if you think they might just be correlated), you run either a regression or a correlation analysis. This lab will walk you through both Simple Linear Regression (SLR) and Pearson’s correlation coefficient calculations. As in the other labs, we will show you how to do these analyses in a traditional, parametric way, and also in a Resampling framework.

In this lab we will use a dataset that we have drawn from the 1kcs study. The data are in the handouts folder on Orca:

Handouts/Stats/week-6-regressionAndCorrelation/

For the first part of this lab, use the **worksheet**: PSME-dbh-height

The research question is: How do DBH and Height relate? Can we reliably use Height to “predict” DBH? *(It is more complicated to ask if there is a causal relationship between Height and DBH. Think about it: similarly to the relationship between Person Shoe Size and Person Height, is it reasonable to talk about one causing the other? Or talk about how having smaller than necessary feet might cause a tall person to die off?)*

That said, for the sake of this lab, we will assume it does make sense with trees to talk about causation between Height and DBH.

**Q1.** Write null and alternative hypotheses that would involve these two variables: What are the null and alternative hypotheses if we run a correlation analysis?

Ho:

Ha:

**Q2**. Now, if we run a regression analysis, what would the null and alternative hypotheses be? Remember – try to determine which variable would be dependent on the other.

Ho:

Ha:

***Correlation:***

First, let’s pretend that neither variable is dependent on the other (for the sake of practicing a correlation analysis).

In Excel, calculate a correlation coefficient (using Excel’s “CORREL” function) between DBH and Height. You will need to highlight the values for one variable as array 1 and for the other variable as array 2.

**Q3**. What is the value of Pearson’s *r* for the correlation?

**Q4**. What is the *r2* value for the correlation? What percentage of the variation in the dataset is explained by the linear relationship between X and Y?

Now, reset all variables and arrays. Using Resampling Stats, determine the number of times (out of 1000) that we would randomly find correlation coefficients (*r* – using Excel’s “CORREL” function) larger than or equal to the *r* value we found from our original dataset due to chance alone. You will do this by shuffling one of the variables in relation to the other and re-calculating the “CORREL” function 1000 or 10,000 times.

**Q5**. Interpret the results of your resampling test (remember to include your Pearson’s *r* and p-value in parentheses following your results statement). You might also consider writing about your *r2* value because it tells you something about the strength of the correlation. Hint – use this statement as a template

when interpreting your results:

“There is a significant positive linear relationship between leaf area index of maple trees and *Rhytisma punctatum* infection rate (Pearson’s *r* = 0.87, p = 0.0034). In fact, over 64% of the variation in the dataset is explained by the correlation between these two variables.”

**Correlation in JMP**

Copy and paste your data into JMP (use the “Paste with Column Headings” option). Make sure JMP categorizes each variable correctly as continuous (blue ramp) (look under Columns on the left hand navigation bar). Correlation and Regression approaches analyze continuous data only.

Click “Analyze” and go to “Multivariate Methods” options. Choose “Multivariate.” For this analysis, choose both of your Y variables (because both variables are considered dependent in this case). Under the Multivariate pull-down menu choose “Pairwise correlations” to get the p-value for the correlation.

**Q6**. What is the *r* value for this correlation from JMP?

**Q7**. What is the p-value for this correlation from JMP?

**Q8**. Are your JMP and resampling results similar?

***Regression:***

Now, we actually want to do more than know that Height and DBH are correlated. We want equations that will allow us to take one and predict the other, and we want to know under what circumstances we can do this, and how accurate our estimate is.

Note that sometimes it is easier to get the DBH measurement, and sometimes the Height. DBH is easier if we are measuring by hand. Height is easier if we are using remote sensing devices such as LiDAR.

First we’ll do (DBH = Y, Height = X).

In Excel, use the “SLOPE” function to determine the slope (*β1*) of the line for this same regression question.

**Q9:** What is *β1?*

Using Resampling Stats, determine the number of times (out of 1000) that the slope (*β1*) (determined randomly) would be larger than or equal to the *β1* that we saw in our original dataset. This time instead of shuffling your data, “**Resample**” them within columns – NOT a normal shuffle! You are running a bootstrap analysis! You will need to know how many values are in your list (823).

**Q10**. What is your p-value (you will probably have to count the number of slopes both larger than your value and the number smaller than the negative of your number – or you might use the absolute value of your slope…)….

**Q11**. Would you reject the null hypothesis? Interpret your results as in Q6, but this time say something about how increasing your x-variable led to an increase/decrease in Y (slope, p-value).

Let’s now run a **parametric Simple Linear Regression** using JMP.

We will keep things simple for this lab and just use the Fit Y by X dialog box. Using the Fit Y by X dialog box, enter “Height” as your independent variable and “DBH” as your dependent variable.

**Q12**. You will see a cloud of points – does it look at the outset like Height explains a significant proportion of the variation in DBH?

Next, using the red arrow to the left of the Bivariate fit box, click on “Fit Line” to fit a line through your cloud of points.

**Q13.** What is the equation of the line?

**Q14**. What is your R2 value?

**Q15**. What is your F-value?

**Q16**. What are your degrees of freedom? Model \_\_\_\_\_\_\_\_\_ Error \_\_\_\_\_\_\_\_

**Q17**. What is the p-value associated with this test? (keep this window open!)

**Q18**. Interpret your results as if in a scientific paper (write a couple sentences and cite your *R2=*, *Fdf1,df2=*, and p= in parentheses at the end):

**Q19**. Do the results from your resampling regression approach correspond to the results from your simple linear regression analysis in JMP? Explain why you say “yes” or “no”.

**Q20**. To make a scatterplot in JMP, you can just alter the graph from the Fit Y by X output. Change the color of the regression line under the linear fit box red arrow, alter the axes and annotate the figure to add the R2, F(dfs) and p-value (go up to the hidden toolbar and find the annotate tool). Try this – turn the line black, increase the size of each point to “Large” (under the Bivariate Fit red arrow) and copy and paste that scatterplot below (using the fat cross selection tool). Also, paste the equation just below the graph – titled “Linear Fit”.

**Now, switch your X and Y variables, to** do (DBH = X, Height = Y), and perform a Regression analysis using JMP.

**Q21.** Insert your scatterplot for this second regression below.

In the table below, show your results for both:

|  |  |  |
| --- | --- | --- |
| Regression Analyses | Height = X, DBH = Y | DBH = X, Height = Y |
| Equation of the line |  |  |
| *R2* |  |  |
| F-value |  |  |
| p-value |  |  |

**Q22.**  What differences between the two regressions (if any) do you observe?

To explore the residual patterns in your data, you need to use the more complicated “Fit Model” option in JMP. Try this (going back to our original problem of predicting DBH using Height. “Go to Analyze – Fit Model” – again, Height as X and DBH as Y. This time you will get a bit more information, but the results should be the same as those from the Fit Y by X option. /set Y to DBH, in “Construct Model Effects” Add Height. Click Run. At the bottom of the read-out will be a residual plot.

Another way to do this is to expand on your regression run (Analysis -> Fit Y by X -> Fit Line) Expand “Linear Fit”, Click **Plot Residuals.**

**Q23**. Paste your residual plot below and describe any patterns you see in the plot. Compare your residual plot to the one in Fig. 9.5 in the Primer. Do you see any problems with these residuals?

**Q24**. In this lab we have performed four different statistical analyses with Height as X and DBH as Y (two to examine linear relationships between two variables: resampling correlation and parametric correlation; and two to examine whether a change in X results in a corresponding change in Y: resampling regression and parametric regression) on a subset of the 1kcs data. Is one of these statistical methods more appropriate than the others? Say which (or say “it doesn’t matter”). In either case, say why or why not, i.e., which analysis method is the most appropriate and why?

**Optional Extension 1:**

It is likely that the “allometric” Height/DBH equations that we derived above can be refined with separate regressions for species (or age class, or both).

Propose one or more hypotheses about whether the relationship differs depending on species or age class. Then propose (and run) the regression analyses. Do they look different to you? How might you test to see if the different regressions are different and hence better than one? Can you think of another kind of analyses that might be better than using different regression equations? You can use the full dataset I’ve prepared for you in the worksheet 1kcsStem, or use the simpler dataset below.

To make this easy for you, I’ve set up a worksheet to run regressions on 3 different species (ABAM, PSME, and TSHE). See the worksheet: Compare3regressions.

**Q25**: What is your scientific hypothesis?

What is/are your null hypothesis/hypotheses?

What is your alternative hypothesis?

Report your results below:

**Optional Extension 2:**

You can export the residual values for each experimental unit – go to the red drop down arrow and choose “Save Columns – Residuals” – this will paste a column of residual values into your original dataset. Now you can run additional tests on the residuals. Try this.

Create a new X-variable Height-Class, based on your X-variable Height; it should be categorical. Propose categories for Height Class, and figure out which Height values should below to which Height-Class (e.g., <20, <30. <40, <50, 50+. Make sure JMP recognizes these as categorical – a red histogram icon will tell you. Now, using ANOVA, test for variation among Height-Classes in the residuals [hint: Y = residuals, X = categorical Height-Class variable; another hint – make sure your data meet the assumptions of ANOVA first!].

**Q26**. If you are able, run a parametric ANOVA and interpret your results including your statistical results in parentheses (remember what Y is!). If you can’t run a parametric ANOVA, do a resampling ANOVA.