

## MULTIPLE REGRESSION ANALYSIS

**Question: Along the elevational gradient we surveyed, what factor most limits the growth of ponderosa pines?**

1. State the three hypotheses your group tested.

**Include with your lab report**

2. Using Microsoft Excel, provide descriptive statistics (mean, variance, standard deviation) for the dependent variable (growth rate) and the three independent variables we examined (density, DBH, and elevation).

	mean	variance	standard deviation
<b>Ponderosa growth rate</b>			
<b>Ponderosa density</b>			
<b>DBH</b>			
<b>Elevation</b>			

Use standard deviation for regression

**Calculate these values and include them with your lab report**

To calculate descriptive statistics in Excel:

- Click the View tab. Verify the following are checked: 1) Toolbars: Standard and Formatting. 2) Formula Bar. 3) Status Bar.
- Highlight a cell below the data.
- On the Standard toolbar, click  $f_x$ .
- In the Paste Function menu, choose Statistical: AVERAGE and highlight the cells that contain the data. Alternatively, in a cell below the data type the following: =AVERAGE(data range). Repeat for variance (VAR), standard deviation (STDEV), and sample size (COUNT).
- Standard error (SE) is a useful statistic that can be used to compare sampling error between variables, populations, data sets, etc. However, Excel doesn't have a function that allows you to calculate standard error directly. To calculate SE type the following in a cell below the data: =STDEV(data range)/SQRT(n), where n=sample size. Include both n and SE in the table above.

3. Construct scatter plots in Microsoft Excel of growth rate ( $y$ ) versus each of the independent variables ( $x$ 's) in our multiple regression analysis. Provide sketches of these scatter plots here. Be sure to label the X and Y axes, and to provide a scale for your variables.

The independent variable is on the x-axis and the dependent on the y-axis. In other words, plot the dependent variable on the independent variable e.g., DBH on elevation.

To create a scatter plot in Excel:

- a. Highlight the cells that contain the dependent variable data.
- b. Click on the Chart Wizard icon on the Standard toolbar or click Insert: Chart.

In the Chart Wizard menu:

Step 1 of 4

Choose XY (Scatter), then Next.

Step 2 of 4

Click on Series. The cells in the dependent variable data range should be displayed under Y values.

Click on X values, highlight the independent variable data cells, then Next.

Step 3 of 4

Titles tab: Enter chart title and axes labels. Make sure you include units of measure.

Legend tab: Unselect "Show legend" (this step is optional).

Step 4 of 4

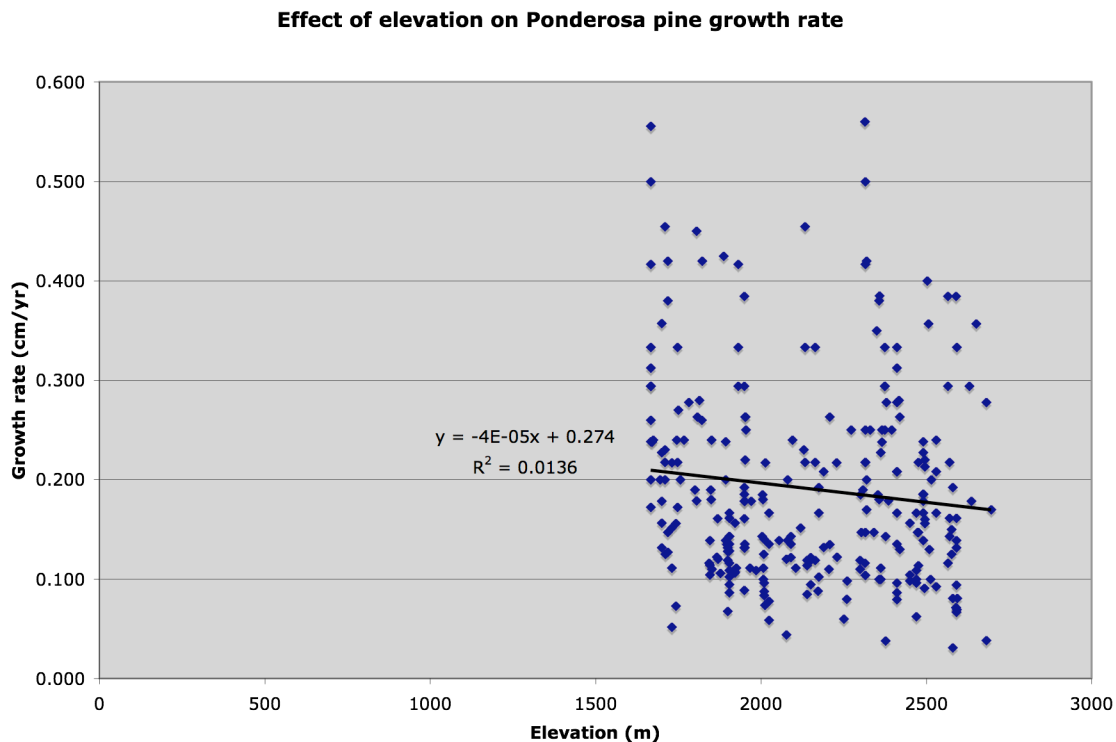
Choose a chart location.

- c. Insert trendline:

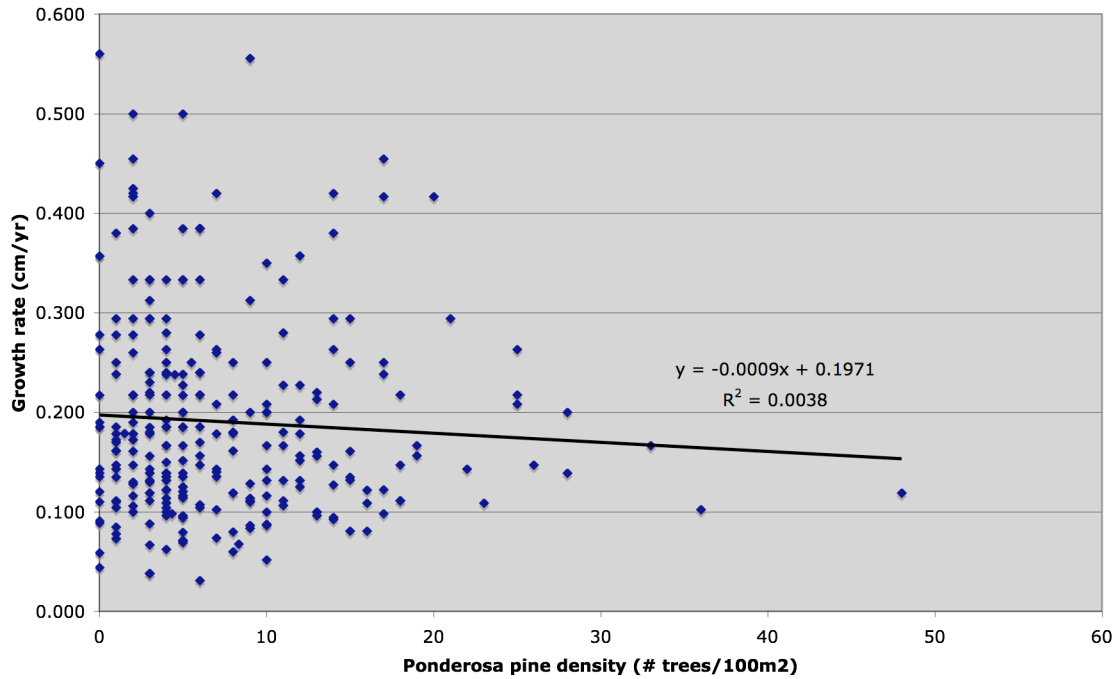
In the Add Trendline menu:

Type: Linear

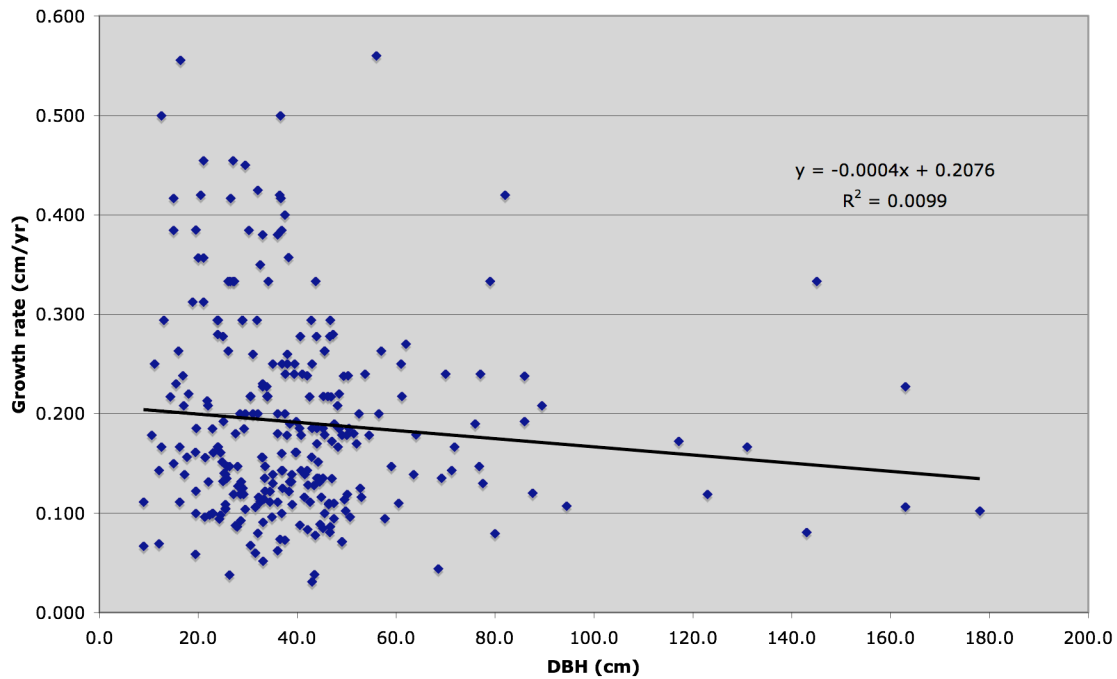
Option: Display both "equation on chart" and "R-squared value on chart", then OK



**Effect of conspecific competition on Ponderosa pine growth rate**



**Effect of DBH on Ponderosa pine growth rate**



4. Using JMP IN, provide the results of your multiple regression analysis. Include the  $R^2$ , the F-ratio for the overall model, and the p-value for the overall model. Write a brief interpretation

of your results, where you address whether your data analysis supports any of the three hypotheses you stated above; be sure to reference results for each of your independent variables. Which independent variable is the strongest predictor of Ponderosa pine growth rate? How do you know?

I've included an image of the JMP output below. Please familiarize yourself with the location of the values that I mention in the following interpretation of the results. You may need to find these values on your own on the exam.

Summary of Fit:

$$R^2 = 0.031019 \text{ (RSquare)}$$

Analysis of Variance:

$$F_{(4, 272)} = 2.1768 \text{ (F Ratio, DF Model, DF Error)}$$

$$p = 0.0719$$

Parameter Estimates (values multiple regression equation):

$$\text{GROWTH RATE} = 0.3123348 + (-0.000045)\text{ELEVATION} + (-0.000465)\text{DBH} + (-0.000989)\text{DENSITY} + (-2.63\text{e-}7)\text{ELEVATION*DBH}.$$

The first number is the intercept.

Effect Tests:

$$\text{Elevation } F_{(1, 272)} = 4.9156, p = 0.0274$$

$$\text{DBH Ponderosa } F_{(1, 272)} = 3.3485, p = 0.0684$$

$$\text{Ponderosa Density } F_{(1, 272)} = 1.2532, 0.2639$$

$$\text{DBH Ponderosa*Elevation } F_{(1, 272)} = 0.0967, 0.7560$$

Interpretation:

Do growing season length, precipitation, congeneric competition, or tree age affect the growth of ponderosa pines along an elevational gradient in the Boulder foothills? In order to answer this question, we conducted a multiple linear regression of elevation, DBH, ponderosa pine neighbor density, and the interaction between elevation and DBH on growth rate. When considered as a group, we fail to reject the null hypothesis and conclude that there is not a significant effect of elevation, DBH, neighbor density and elevation\*DBH on growth rate ( $F_{(4, 272)} = 2.1768, p = 0.0719$ ). Together these variables only explain three percent of the observed variance in growth rate.

**CAUTION:** If you test a multiple linear regression model and determine that it is not a significant predictor of the dependent variable, stop right there! You cannot continue to interpret the individual effects of the variables included in the model, but rather should continue testing other biologically sound models until you find one that fits the data better. It is also important to remember that with each additional variable added to a model, the power to detect an effect is decreased. In other words, you are less likely to find evidence of a significant effect, even if an effect really does exist. However, for the sake of instructing you in how to interpret the individual effects of a multiple regression model, I'm going to do so here.

Controlling for the effects of DBH, neighbor density, and elevation\*DBH, we find that growth rate significantly decreases as elevation increases ( $F_{(1, 272)} = 4.9156$ ,  $p = 0.0274$ ). Therefore, we conclude that growing season length has a greater effect on growth rate than precipitation levels. We find no evidence of a significant effect of DBH ( $F_{(1, 272)} = 3.3485$ ,  $p = 0.0684$ ), neighbor density ( $F_{(1, 272)} = 1.2532$ ,  $0.2639$ ), or the interaction term ( $F_{(1, 272)} = 0.0967$ ,  $0.7560$ ) on growth rate. Based on these results, we find support for the hypothesis that growing season length limits growth of Ponderosa pine in the Boulder foothills.

Response Growth Rate (cm/yr)						
▼ Whole Model						
▼ Summary of Fit						
RSquare		0.031019				
RSquare Adj		0.016769				
Root Mean Square Error		0.09903				
Mean of Response		0.190859				
Observations (or Sum Wgts)		277				
▼ Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Ratio		
Model	4	0.0853896	0.021347	2.1768		
Error	272	2.6674699	0.009807		Prob > F	
C. Total	276	2.7528595				0.0719
▼ Lack Of Fit						
Source	DF	Sum of Squares	Mean Square	F Ratio		
Lack Of Fit	271	2.6674699	0.009843			.
Pure Error	1	0.0000000	0.000000		Prob > F	
Total Error	272	2.6674699				.
				Max RSq		
				1.0000		
▼ Parameter Estimates						
Term		Estimate	Std Error	t Ratio	Prob> t	
Intercept		0.3123348	0.0465	6.72	<.0001	
Elevation (meters)		-0.000045	0.00002	-2.22	0.0274	
DBH Ponderosa (cm)		-0.000465	0.000254	-1.83	0.0684	
Ponderosa Density (#/100m <sup>2</sup> )		-0.000989	0.000883	-1.12	0.2639	
(DBH Ponderosa (cm)-40.6747)*(Elevation (meters)-2147.32)		-2.63e-7	8.4e-7	-0.31	0.7560	
▼ Effect Tests						
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F	
Elevation (meters)	1	1	0.04820692	4.9156	0.0274	
DBH Ponderosa (cm)	1	1	0.03283792	3.3485	0.0684	
Ponderosa Density (#/100m <sup>2</sup> )	1	1	0.01228970	1.2532	0.2639	
DBH Ponderosa (cm)*Elevation (meters)	1	1	0.00094853	0.0967	0.7560	

The multiple linear regression model was found to be a bad predictor of growth rate. Moreover, in hopes of showing you how to interpret the results of a multiple regression I gave you the answers to this section. For that reason, I would like you to consider the effects of elevation and neighbor density alone using a simple linear regression (single independent variable). I know, thanks Erin for giving us more work. But the more practice and feedback you get now, the better prepared you will be for your independent projects.

## SIMPLE REGRESSION ANALYSIS

5. Using Excel, provide the results of your simple regression analyses. Include the  $R^2$ , the F-ratio, and the p-value for each model. Write a brief interpretation of your results, where you address whether your data analysis supports any of the three hypotheses you stated above; be sure to reference results for each of your independent variables. Which independent variable is the strongest predictor of Ponderosa pine growth rate? How do you know?

Because the effect of DBH on growth rate was not one of the hypotheses explicitly stated above, I will provide an example of how to use Excel to test a simple linear regression of growth rate against DBH.

**Question: Do older trees grow more slowly?**

Hypothesis/Prediction: If older trees grow more slowly, then larger DBH trees have lower growth rates.

To conduct a simple linear regression in Excel:

- Under the Tools tab, choose Data Analysis. If the Data Analysis tool does not appear under your Tools tab, you'll need to add it. Under the Tools tab, choose Add-ins. Check Analysis ToolPak (don't check Analysis ToolPak – VBA) and click OK. The Data Analysis tool should now be available.
- In the Data Analysis menu, highlight Regression and click OK.
- In the Regression menu, click on Input Y range: select the cells that contain dependent variable data (growth rate). Repeat for the Input X range: independent variable data (DBH).
- Choose a location for the output.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.099437572
<b>R Square</b>	<b>0.009887831</b>
Adjusted R Square	0.006287423
Standard Error	0.099548969
Observations	277

ANOVA					
	df	SS	MS	F	Significance F
<b>Regression</b>	<b>1</b>	0.027215909	0.027215909	<b>2.746308465</b>	<b>0.098619607</b>
<b>Residual</b>	<b>275</b>	2.725249218	0.009909997		
Total	276	2.752465127			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
<b>Intercept</b>	<b>0.207584318</b>	0.011729489	17.69764325	2.60166E-47	0.184493319	0.230675317	0.184493319	0.230675317
<b>DBH</b>	<b>-0.000411098</b>	0.000248068	-1.657198982	0.098619607	-0.000899451	7.72555E-05	-0.000899451	7.72555E-05

I've highlighted those values that you need to report in any regression analysis for this class.

Regression Statistics:

$$R^2 = 0.0099$$

ANOVA:

$$F_{(1, 275)} = 2.75, p=0.099$$

Equation of the line:

$$\text{GROWTH RATE} = 0.20 + (-0.00097)\text{DBH}$$

Interpretation:

Do ponderosa pines grow more slowly as they age? As ponderosa pines age, they add wood to their girth. Consequently, we used DBH as a proxy for age. We conducted a simple linear regression of growth rate against DBH in ponderosa pines. We fail to reject the null hypothesis and find no evidence that DBH has a statistically significant effect on growth rate ( $F_{(1, 275)} = 2.75, p=0.099$ ). Furthermore, DBH only explains 0.9% of the variation in growth rate observed. Therefore, we conclude that tree age does not have a significant effect on ponderosa pine growth rates in the areas sampled.

Note: Because of limitations imposed by Excel, most serious scientists and statisticians use different software (SAS, SPSS, R) to run their analyses. However, in the interest with providing you with the tools to run the analyses for your projects on your own, here is the information. Please be aware that Excel 2008 does not have the Data Analysis add-on. If you only have access to a computer with Microsoft 2008, please let me know and I can provide you with assistance.

Also, I'd like to give you a heads up with regard to the p-value returned for the Elevation data, which is  $p=0.05214$ . Different disciplines and scientists deal with p-values just slightly above the level of 0.05 differently. It is common in the ecological literature to refer to this level as "marginally significant at the 0.05 level." Please interpret your results in this way.

## T-TEST Analysis

**Question: Do Kittredge and University ponds, on average, differ in their pH?**

1. Using Microsoft Excel, provide the descriptive statistics (mean, variance, standard error) for the pH of Kittredge and University ponds.

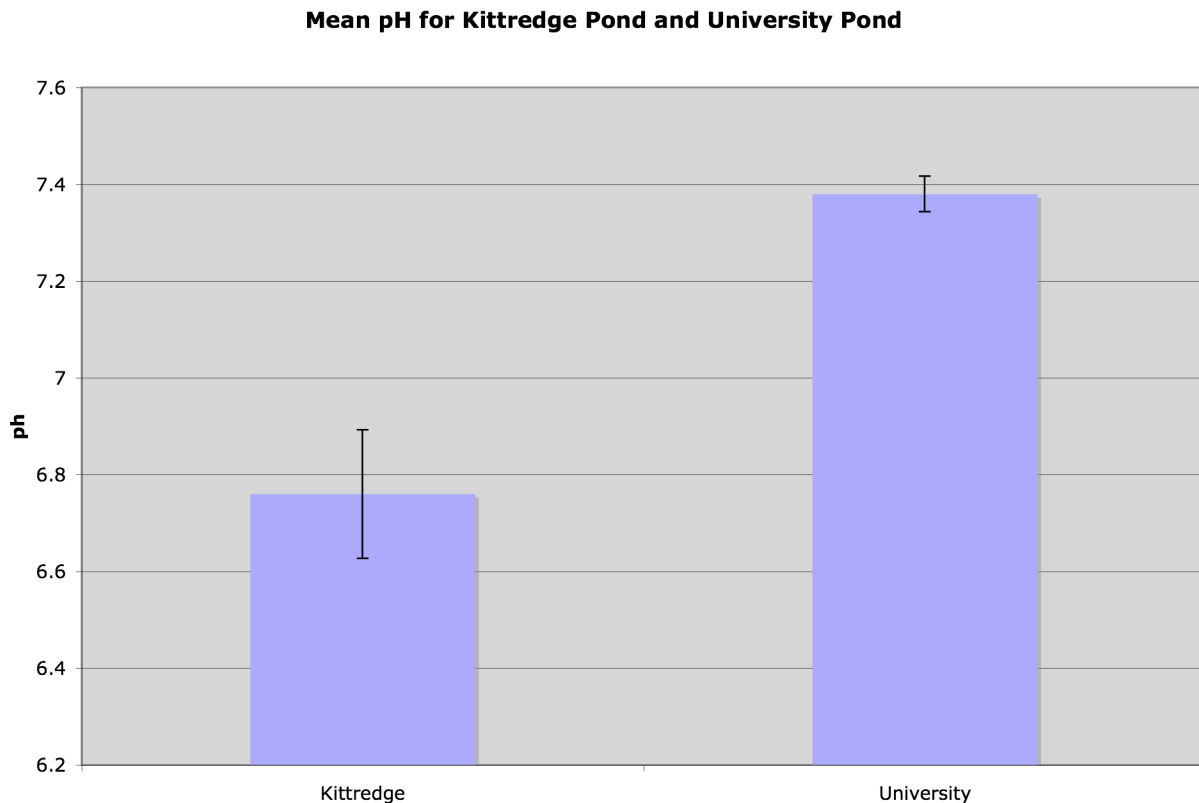
	mean	variance	standard error
Kittredge	6.670	0.088	0.133
University	7.380	0.007	0.037

Use standard error for T-tests

2. Provide a sketch of an Excel bar graph showing the average pH of Kittredge Pond and University Pond. Be sure to label the x- and y-axes, to provide a scale for your dependent variable and to add error bars (standard error).

The directions in the lab manual should be adequate. You need to input the summary statistics (mean and SE) in the same format indicated in the lab manual. Please remember to print your graphs rather than sketch them.

Note: You need to use standard error bars on your graphs. The formula for calculating SE is in the descriptive statistics directions above (multiple regression).



3. Provide the results of your Excel statistical analyses. Provide the  $t$  and  $P$  values and your sample size. State your null hypothesis and write a brief interpretation of your results.

I've provided instructions of how to run a t-test in Excel below, as well as a sample of how to interpret the results. You are welcome to use JMP IN if you prefer.

Question: Do Kittredge and University ponds differ in their pH?

Note: Pay attention to the way this question is worded. For example, the in class assignment for Question 2 reads: are ponderosa pines LARGER than other species. You'll need to decide whether to use the p-value from the one- or two-tailed t-test to answer this question. In the case of comparing pH between ponds, there is nothing in the way the question is worded that predicts pH of one pond will be greater than the other. Therefore, I will interpret the results of the two-tailed t-test.

To conduct a t-test in Excel:

- Under the Tools tab, choose Data Analysis.
- In the Data Analysis menu, highlight t-test: Two-Sample Assuming Equal Variances and click OK. Assume equal variances for all t-tests you run in this class.
- In the t-test menu, click on Variable 1 range: select the cells that contain the first column of data (Kittredge). Repeat for the Variable 2 range: second column of data (University).
- Choose a location for the output.

	Kittredge	University
	6.8	7.4
	7.1	7.5
	6.3	7.3
	6.9	7.3
	6.7	7.4
MEAN	6.760	7.380
VAR	0.088	0.007
STDEV	0.297	0.084
SE	0.133	0.037
N	5	5

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	Variable 2
Mean	6.76	7.38
Variance	0.088	0.007
Observations	5	5
Pooled Variance	0.0475	
Hypothesized Mean Difference	0	
df	8	
t Stat	-4.497952751	
P(T<=t) one-tail	0.001003759	
t Critical one-tail	1.859548033	
P(T<=t) two-tail	0.002007519	
t Critical two-tail	2.306004133	

Species	Mean	SE
Kittredge	6.76	0.133
University	7.38	0.037

t-value = -4.498

p = 0.002

d.f. = 8

Interpretation:

Do Kittredge and University ponds differ in their pH? Kittredge Pond has a mean pH of 6.8 while University Pond has a mean pH of 7.4. We conducted a two-tailed t-test assuming equal variances. We reject the null hypothesis and find that mean pH is statistically different between the two ponds ( $t = -4.498$ ,  $p=0.002$ ,  $d.f. = 8$ ).

Note: If we had performed a one-tailed t-test, i.e. is the pH of University Pond greater than that of Kittredge Pond, the interpretation would be slightly different. For example: We conducted an one-tailed t-test assuming equal variances. We reject the null hypothesis and find that mean pH of University Pond is greater than that of Kittredge Pond ( $t = -4.498$ ,  $p=0.002$ ,  $d.f. = 8$ ).

You need to be able to do the following for the exam.

1. Given a question, determine which type of statistical analysis is most appropriate.
2. Given the statistical output, identify and report the parameters indicated above for a t-test, simple linear regression, and chi-square test (we'll cover this test later in the semester).
3. Provide a biologically meaningful interpretation of your results.
4. You do not need to know how to conduct a multiple linear regression, but you should be familiar with the output and be able to interpret the results.