BIOL 417: Biostatistics
Laboratory #10
26 April 2011

Using MYSTAT for ANCOVA and multiple regression

ANCOVA: Analysis of Covariance

To test whether phenotypic plasticity is present in the feeding apparatus in bluegill sunfish in response to available prey, a researcher raised 15 bluegill on 3 different diets. The researcher raised 5 bluegill on krill, 5 bluegill on worms, and 5 bluegill on fish food flakes. The researcher then measured the lower jaw of each fish to determine if there was a significant relationship between length of the lower jaw and the type of diet the fish were fed. The data from the experiment are in the table below.

Krill		Worms		Flakes	
Standard Length	Jaw Length	Standard Length	Jaw Length	Standard Length	Jaw Length
8.5	1.2	13	2.2	7.6	0.5
8.4	1.5	11	1.5	9.6	0.9
7	0.8	10.5	1.2	11.5	1.2
9.2	2.1	8	0.9	8.9	0.8
6.8	0.5	13.2	2	10.9	1.3

First, use ANOVA to determine if there is a difference between the 3 different groups in mean jaw length, ignoring any possible effects of standard length for now. You should obtain the ANOVA shown below:

Source of variation	Sum of squares	df	Mean squares	F-ratio	P
Groups	0.964	2	0.482	1.8467	0.1999
Error	3.132	12	0.261		

You would end up concluding that diet had no significant effect on lower jaw length (assuming α = 0.05).

This experiment, however, has a confounding variable that could affect interpretation of the results of this experiment. Jaw length likely varies among fish of different sizes simply because of those size differences; thus, the size of each fish also needs to be accounted for when comparing jaw lengths between the different diet groups. If you don’t do this, then any variation in jaw length induced by differences in standard length would get included in MSE.

To do this you need to analyze the data using analysis of covariance (ANCOVA) to take into account the size of each fish.

If you ran the preliminary analysis correctly, you should already have the Jaw length variable in the "Dependent" box and the Diet variable in the "Factor" box. Click on "Analyze" → "Analysis of Variance" and simply add the Standard length variable to the Covariate box and re-run the analysis. You should obtain the table below:

Source of variation	Sum of squares	df	Mean squares	F-ratio	P
Groups (Diet)	1.1522	2	0.5761	8.8138	0.0052
Covariate (SL)	2.413	1	2.413	36.9162	0.0001
Error	0.719	11	0.0654		
Total	4.2842	14			

Note that the effect of diet is now very clear; fishes raised on different diets had significantly different jaw lengths, after accounting for the effect of standard length on jaw length. You can also infer from this table that there is a significant relationship between standard length and jaw length (i.e., the H_o: $\beta = 0$ in this relationship can be rejected). Thus, two quite distinct null hypotheses are being tested in one analysis, one comparing means (the ANOVA part, with $k – 1 = 2$ df) and one determining if a relationship exists between X and Y (the regression or covariate part, with 1 df). Note also that even the order of means had changed compared to when you ran the ANOVA with no covariate.

Multiple Regression: Relationship between Y and more than one X

The ABC Corporation is opening new retail sales outlets and they want to staff these stores with employees most likely to be successful at selling the products. To meet this goal, ABC decides to study the sales staff at existing stores to determine if intelligence and extroversion (i.e., a friendly and outgoing personality) predict sales performance of current employees. ABC's logic is that if intelligence and extroversion predicts sales performance, then a good strategy for new stores is to hire intelligent extroverts for the sales positions.

To conduct the study, all current retail sales employees at existing stores take psychological tests designed to measure intelligence and extroversion. Also, past sales performance data are checked for each employee. In the end, there are three scores for each sales person:

1. an intelligence score (on a scale of 50-low intelligence to 150-high intelligence),
2. an extroversion score (on a scale of 15-low extroversion to 30-high extroversion), and
3. sales performance expressed as the average dollar amount sold per week.

The data are presented below for the 20 current sales employees of the ABC corporation.

Sales Person	**Intelligence**	**Extroversion**	**$ Sales/ Week**
1	89	21	2625
2	93	24	2700
3	91	21	3100
4	122	23	3150
5	115	27	3175
6	100	18	3100
7	98	19	2700
8	105	16	2475
9	112	23	3625
10	109	28	3525
11	130	20	3525
12	104	25	3450
13	104	20	2425
14	111	26	3025
15	97	28	3625
16	115	29	2750
17	113	25	3150
18	88	23	2600
19	108	19	2525
20	101	16	2650

Using the ABC data from above, first determine if intelligence is a good predictor of sales performance using a simple linear regression as you've done before. You should obtain the table below:

Source of variation	Sum of squares	df	Mean squares	F-ratio	P
Regression	510589.89	1	510589.89	3.511	0.0773
Residual	2617660.11	18	145425.56		
Total		19			

This is not quite a significant relationship, so ABC analysts decide to also include the extroversion data to see if they could improve the predictability of the model….

Now analyze the data using both intelligence and extroversion to predict sales performance using a multiple regression. You should already have the Intelligence variable in the "Independent" box; simply add the Extroversion variable to the same box below the Intelligence variable and re-run the analysis. You should get the ANVOA table below:

Source of variation	Sum of squares	df	Mean squares	F-ratio	P
Regression	1096526.78	2	548263.39	4.5875	0.0255
Residual	2031723.22	17	119513.13		
Total	3128250	19			

Note that you now have a significant regression model (assuming α = 0.05) by which you can predict sales; ABC executives need only know an applicant’s intelligence and extroversion scores to be able to predict how productive that employee will be. Be sure you understand why the df change between these two tables.

Look back at the t-test (Coefficient) tables from these two analyses; notice how the P value for the significance of the Intelligence variable changes when you include the Extroversion variable versus when it is not present in the model. Do these tables suggest an even simpler linear model that ABC executives could use???

Assume that during a three-hour period spent outside, a person recorded the temperature, the amount of time they mowed the grass, and their water consumption. The experiment was conducted on 7 randomly selected days during the summer. The data are shown in the table below with the temperature placed in increasing order.

Temperature (F)	**Water Consumption (oz)**	**Time mowing the grass (hrs)**
75	16	1.85
83	20	1.25
85	25	1.5
85	27	1.75
92	32	1.15
97	48	1.75
99	48	1.6

Analyze the data using both temperature and time to predict water consumption using a multiple regression and record your results below. Enter the above data in an Excel spreadsheet and copy and paste the data in the next available columns in Mystat. Rename the variables “TEMP,” “WATER,” and “TIME,” respectively. “WATER” is the dependent variable and “TIME” and “TEMP” are the independent variables.

Source of variation	Sum of squares	df	Mean squares	F-ratio	P
Regression					
Residual					
Total					

Interpret the results:

Source of variation	Sum of squares	df	Mean squares	F-ratio	P
Regression	970.6583	2	495.3291	313.1723	0.0000
Residual	6.1989	4	1.5497		

Total