

Biol 458 – Biometry

Final Exam - Fall 2011

Due December 15 - Open book, open note. Pledge that you neither gave nor received help on this test. Try to minimize the length of your exam paper by making the graphs and tables small, and only including the essential tables.

1. (20 pts) Design an experiment to test the null hypothesis that acid precipitation has no effect on needle production in Monterey pine. Outline how you would use preliminary data to design an experiment with low probabilities of both Type I and Type II errors.

There are many facets of the interaction of acid precipitation and its effects on needle production in white pine that could be experimentally addressed. As a first step in testing the hypothesis that acid rain has no effect needle production, I would pot white pine saplings of equal size and age in identical containers with identical soil mixtures and place them in a greenhouse. I would then assign these trees independently and at random to at least 2 treatments. Watering with neutral pH water or with water of pH 4. I could then estimate needle production on each tree over time. This experiment would be a 2-factor repeated measures ANOVA (treatment x time) with repeated measures on the time factor. Prior to performing this experiment I would go into the field and sample a number of white pine saplings to determine the the amount of needle production per branch on each tree and the variance among trees. I would then use this preliminary estimate of variance needle production in conjunction with various values of alpha, and the effect I wish to detect to determine the sample size I would need to have a low probability of a Type II error. I choose this particular experiment since under greenhouse conditions I could control or eliminate many extraneous sources of variation in needle production and perform a strong test of the hypothesis with high power. However, by doing a highly controlled greenhouse experiment and limiting my observations to the sapling stage of white pine I sacrifice both the realism and generality of the results obtained. A field based experiment should be performed to follow up on any positive results obtained in the greenhouse.

2. (10pts) When choosing between two alternative regression models for the same data, what criteria should be used to select the best model?

Two criteria can be used to determine the best fit model: adequacy and precision. An adequate model fits the data with no pattern in the residuals. The model neither systematically over or underestimates the mean value of $y|x$. Two approaches can be used to determine adequacy. If multiple observations of y at each value of x are available, one can calculate and test for the "lack-of-fit" of the model. An F-test can be performed to

determine if there is significant lack of fit. Alternatively, if multiple observations of y at each value of x are not available, a visual inspection of the residuals can be used to determine if the model is adequate. If two or more models are determined to be "adequate," the precision of the models can be used to determine which of the adequate models fit the data best. The R^2 or the MSE can be used for this purpose. The adequate model with the highest R^2 or the lowest MSE would be considered the best fit model.

3. (5 pts) What is the difference between parametric and non-parametric tests?

Parametric statistical tests involve the estimation and use of the parameters of an assumed underlying distribution of the sampled population. Non-parametric tests do not require the assumption of any particular underlying population distribution nor the estimation of any population parameter.

4. (10pts) Under what circumstances would a repeated measures ANOVA be preferable to a factorial ANOVA? What is a repeated measures ANOVA designed to achieve?

A. Power - A repeated measures ANOVA would be preferable to a factorial ANOVA if initial differences between subjects or differences in the levels of responsiveness of subjects are large. If this is true, then the reduction in error variation resulting from removing the variance due to initial differences between subjects will offset the loss in degrees of freedom associated with the repeated measures design.

B. Purpose - Repeated measures ANOVA allows the experimenter to remove variation caused by initial differences between subjects from the estimated error variation so that more powerful tests of treatment effects can be performed. A repeated measures design may also be preferable when subjects are scarce or if it is difficult or expensive to apply experimental treatments.

5. (5pts) What can and what cannot be inferred from a significant correlation between two variables?

One can infer that variable y and variable x are associated. One cannot infer that either variable y causes variation in variable x or vice versa.

6. (20 pts) When would you use nonparametric correlation? Calculate the correlation between the following data using both parametric and non-parametric approaches.

X	Y
2.5	4.7
3.1	4.3
5.4	7.4

5.3	6.8
8.9	12.6
6.2	9.5
11.3	12.4
16.9	32.3
25.7	21.4
28.5	35.9

Nonparametric correlation could be used if the distribution of (x,y) is not bivariate normal. However, since correlation is robust with respect to the violation of the assumption of bivariate normality, the product moment correlation coefficient is most often used.

From SPSS:

Parametric: Pearson's $r = 0.895$, $n = 10$, $p < 0.001$

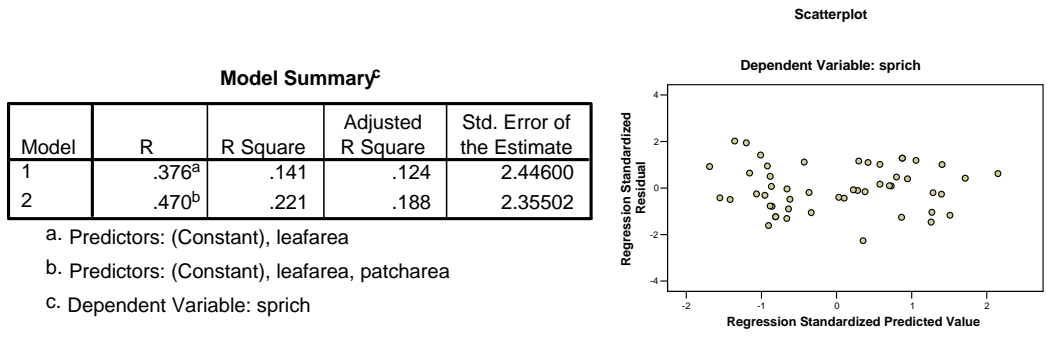
Non-parametric: Spearman's $\rho = 0.964$, $n = 10$, $p < 0.001$

7. (25pts) It has been hypothesized that urbanization affects the abundance and species richness of animals occupying remnant patches of habitat embedded in an urban matrix. Urban habitat patches are thought to have fewer species and that species occupying such patches would be lower in abundance than in similar habitat patches not surrounded by an urban matrix. To test this hypothesis for a group of oak feeding insects, a graduate student sampled insect communities on live oak (*Quercus agrifolia*) in oak forests scattered throughout the San Francisco Bay Area. Besides measuring the richness and abundance of this insect group in each patch, she collected data on patch size, isolation, patch shape, the number of distinct oak habitat fragments and the nature of the habitat matrix surrounding each study site (amount of several habitats with 1 km radius of the study site: agriculture, grassland, non-oak forest, urban, non-urban). Build multiple regression models (one each for richness and abundance) to determine if the amount of urban or other habitat types is useful in explaining the richness and abundance of this insect community (see data file: oakdata.sav). Do not use abundance as an independent variable to explain species richness and vice versa. Treat leafarea as a nuisance parameter and force it into each model as the first variable. Leafarea varied from site to site and is an index of sampling effort. Forcing leafarea into each model allows you to test the effects of other variables after correcting the data for site to site variation in sampling effort. Explain the process you used to arrive at the model you chose as the best model.

Model for Species Richness - Since species richness was approximately normally distributed (based on checking skewness and kurtosis), I did not transform the dependent variable. I inspected scatter plots of the relationship between species richness and each potential independent variable and all plots were approximately linear, so I did not transform any variables. I inspected the correlation matrix of independent variables, but

only one correlation was moderately high (urban and non-urban were negatively correlated at the -0.85 level). However, I decided that this would not likely generate problems with collinearity.

I performed a multiple regression with species richness as the dependent variable. I first forced leafarea into that model and then used the forward algorithm and all the remaining candidate independent variables. The model produced had 2 independent variables and accounted for only 0.22% of the variation in the data.



However, the residuals were homoscedastic and there were no data points with large Cook's D values, so I chose this as the final model (species richness = 1.412 + 0.000428 leaf area + 3.23 x 10⁻⁵ patch area). The overall model was significant ($F_{2,47} = 6.656$, $p = 0.002$). I concluded that matrix type was not useful in explaining variation in species richness.

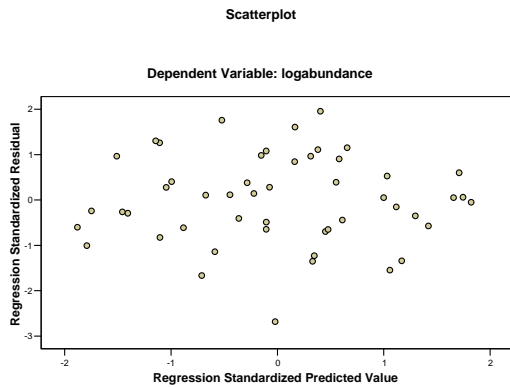
Model for Abundance - The abundance of insects was not normally distributed so I sought transformations that normalized the dependent variable. I decided to use the log base 10 transformation which made the data approximately normal. Since from the analysis of the species richness data I already knew that none of the independent variables were so strongly correlated as to cause problems of collinearity ($r > 0.9$), I made no effort to exclude variables to prevent problems with collinearity. I inspected scatter plots of the relationship between logabundance and the independent variables to check for linearity of the relationships. Based on the plots, I decided not to transform any of the independent variables.

I performed a multiple regression with lgabund (log 10 abundance) as the dependent variable. I first forced leafarea into that model and then used the forward algorithm and all the remaining candidate independent variables. The model included 3 independent variables. Since there were no data points with large Cook's D values, and the residuals were homoscedastic. I selected this as my final model (logabundance = 0.678 + 8.622 x 10⁻⁵ leafarea + 7.547 x 10⁻⁶ patcharea - 0.075 oakfragments). The overall model was significant ($F_{3,46} = 6.068$, $p = 0.001$). Based on this analysis I would conclude the abundance of these insects is unrelated to the matrix of surrounding habitat.

Model Summary^d

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.313 ^a	.098	.079	.55348
2	.456 ^b	.208	.174	.52414
3	.532 ^c	.284	.237	.50383

- a. Predictors: (Constant), leafarea
- b. Predictors: (Constant), leafarea, patcharea
- c. Predictors: (Constant), leafarea, patcharea, oakfragments
- d. Dependent Variable: lgabund



Extra Credit

9. (10 pts) The file “parasitoid” linked to the class web page contains information about the number of leaf mines (nummines) of *Cameraria hamdryadella* on an individual white oak leaf (*Quercus alba*) and the proportion of those mines observed to be visited by a foraging parastoid wasp (*Closterocerus tricinctus*) (pvis). Use regression to determine if the proportion of mines visited depends on the number of mines per leaf.

I first examined the dependent variable, pvis, to determine if it was normally distributed. Since approximate 95% confidence intervals on the skewness and kurtosis did not contain 0, I attempted to transform pvis. I tried, the arcsin(square root) transformation often used for proportion, the log, and the square root transformation. Only the square root transformation (sqpvis) was able to normalize pvis. I then examined a scatter plot of sqpvis and observed that the relationship between the dependent variable and independent variable was not linear. I attempted several transformations of the independent variable (nummines), but none linearized the relationship. I used the log number of mines (lgnum), since it appeared to do the best job at linearizing the relationship. I then performed the linear regression of sqpvis on lgnum.

Model Summary^a

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.292 ^a	.085	.079	.24351

- a. Predictors: (Constant), lgnum
 b. Dependent Variable: sqpvis

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.810	1	.810	13.655	.000 ^a
	Residual	8.717	147	.059		
	Total	9.526	148			

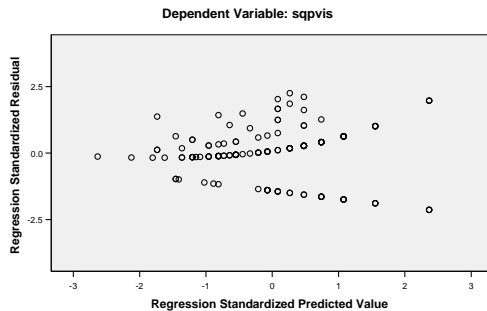
- a. Predictors: (Constant), lgnum
 b. Dependent Variable: sqpvis

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.520	.051		10.125	.000
	lgnum	-.200	.054	-.292	-3.695	.000

- a. Dependent Variable: sqpvis

Scatterplot



This model is statistically significant but accounted for only 8.5% of the variation in sqpvis. Furthermore, based on the residual plot is not an adequate model. Clearly the relationship between sqpvis or pvvis and lgnum or nummines is curvilinear, and some other approach than linear regression (like curvilinear regression) must be used to describe this relationship. The final, inadequate, linear regression model I chose was $\text{sqpvis} = 0.520 - 0.2(\text{lgnum})$.

The proportion of mines visited seems to decline as the number of mines on a leaf decreases. However, a curvilinear regression might fit the data much better.

10. (10 pts) Calculate a 95% prediction interval on a novel observation of the proportion of mines visited for a leaf with 22 mines based on the regression model you selected as best in problem 9.

Using the model I chose in problem 5 where I transformed the x variable by a log base 10 transformation and the y variable by a square root transformation, the predicted value of y for 22 mines on a leaf is

$$\hat{y}_{22} = 0.52 - (0.2 * 1.342) = 0.2516.$$

The formula for the prediction interval is

$$\hat{y}_i \pm S_{y|x} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_x^2(n-1)}}.$$

With $n = 149$, mean $x = 0.8757$, S^2 of the x 's $= 0.137$, $x_{22} = 1.342$, and $S_{y|x} = 0.24351$ the upper and lower confidence interval on a novel prediction of a leaf with 22 mines would be that its expected square root proportion visited is

$$0.2516 \pm 0.24351 \sqrt{1 + \frac{1}{149} + \frac{(1.342 - 0.8757)^2}{0.137(149 - 1)}}$$

$$0.2516 \pm 0.24351 \sqrt{1 + 0.0067 + \frac{0.21744}{20.276}}$$

11. (10 pts) When is it appropriate to use transformations in Ordinary Least Squares regression?

Transformations are appropriate to normalize the dependent variable, to make the variance of the residuals more homoscedastic, and to linearize the relationship between the dependent and independent variable. In the latter two cases, transformation of either or both the dependent or independent variable may be appropriate.