Biol 206/306 – Advanced Biostatistics
Lab 5 – Multiple Regression and Analysis of Covariance
Fall 2016

By Philip J. Bergmann

0. Laboratory Objectives

1. Extend your knowledge of bivariate OLS regression to multiple regression
2. Learn how to do multiple regression in R
3. Learn how to do backward stepwise elimination
4. Learn about the different ways of doing an Analysis of Covariance (ANCOVA)
5. Learn how to do an ANCOVA with a continuous covariate in R
6. Learn how to do an ANCOVA with a categorical nuisance variable in R

1. The Multiple Regression Approach

Multiple regression is used to estimate the relationship of multiple independent or explanatory variables to a single response variable. Because there is only a single response variable, multiple regression is not generally considered a multivariate technique, but this does not negate its usefulness. Most often, all variables in a multiple regression are continuous, although we will see exceptions to this later in the lab in the case of ANCOVA. Multiple regression is simply a generalization of bivariate OLS regression, as can be seen from its linear model (presented for the two explanatory variable case):

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$$

Where y_i is the ith observation for your response variable, x_{ij} is the ith observation for the jth explanatory variable, β_k is the partial regression coefficient for each term, and ε_i is the ith residual.

The partial regression coefficients are similar to the slope of a bivariate regression, but not exactly so, because they take into account the coefficients of the other terms in the model (hence the "partial"). This is generally considered a strength of the approach because you can tease apart the effects of different explanatory variables and their interactions. There are, however, disadvantages to them as well because they cannot be estimated with accuracy if your explanatory variables are highly correlated with one another. This is termed "multicollinearity", and is assumed not to be present in a multiple regression. Fortunately, it is an easy assumption to test (although sometimes not as easy to remedy).

This brings us to the assumptions of multiple regression. Most are the same as for the bivariate case: data are random and independent, residuals are normally distributed, and the explanatory variables are measured without error. From our discussions of model I and model II regression, we can say that "without error" is in practice more of a loose requirement – the explanatory variables should be measured with substantially less error than the response variable (a rule of thumb that is often used is one third the error or less). Unfortunately, there is no implementation

of a model II multiple regression, and therefore, if this assumption is violated, there is not much that can be done about it – people pretty much ignore this assumption for multiple regression. Multiple regression also assumes that all important and no unimportant explanatory variables are included in the model. This can be difficult to ascertain. You should have a good reason for including each of your explanatory variables. Finally, multiple regression assumes that the explanatory variables do not have multicollinearity. You can evaluate this by calculating pairwise correlations among explanatory variables prior to your multiple regression, or by looking at **tolerance** values for each explanatory variable. Tolerance is a measure of how much independent variation there is in each variable, given a set of variables. Therefore, tolerance is affected by which explanatory variables are included in an analysis. The higher the tolerance, the more independent a variable is from the others in the set. Another rule of thumb is that a variable with a tolerance of less than 0.1 is highly correlated with others in the set and could confound the analysis. Simply put, variables with low tolerances need to be excluded from analysis, or else the partial regression coefficients become meaningless and unstable. If you cannot remove problem variables for some reason, then you have a problem and should try to figure out another way of analyzing the data. We will talk about one approach later in the semester (hierarchical partitioning).

2. Multiple Regression in R

You will use multiple regression to study how the characteristics of isolated patches of forest in Victoria, Australia influence forest bird abundance. The dataset was published by Loyn in 1987 (*Effects of patch area and habitat on bird abundances, species numbers and tree health in fragmented Victorian forests. In: Nature Conservation: The Role of Remnants of Native Vegetation, Saunders et al. eds., pp. 65-77, Surrey Beatty & Sons, Chipping Norton, NSW, Australia*). The dataset contains the following variables for 56 patches of forest:

- Bird abundance (individuals per hectare)
- Patch area (hectares)
- Year that the patch was isolated from nearby forest
- Distance to the nearest patch (km)
- Grazing intensity of the patch (on a scale of 1-5, 1 being light and 5 being heavy)
- Altitude of the patch above sea level (m)

Distance to nearest patch and patch area were not normally distributed, and so have also been log-transformed for your convenience (Ldist, and Larea). ***Download, convert, and load the dataset into R, as object "bdata". Then attach it for ease of reference and take a look at it by having it printed to screen and also using the str() function. Also load the package "car".***

What is the response variable and what are the explanatory variables in this example?

How many explanatory variables are there?

If you used a completely crossed model, including all possible interactions among your explanatory variables, how many terms would there be in the model (not counting intercept)?

When there are many explanatory variables, it becomes both impractical and uninteresting to include all possible interactions (how would you interpret a significant interaction among four main effects?). An alternative approach is to either include only interactions that you find compelling from a biological standpoint, or just include the main effects (no interactions). In R, a model without interactions must first be fit if one wants to calculate the tolerances of the explanatory variables.

Test for multicollinearity in two ways: first calculate all pairwise Pearson correlations among the explanatory variables (use the cor(x) function). Record your results in the table below (place the variable names in the top row and first column, leaving cell [1,1] blank.

	1.0				
		1.0			
			1.0		
				1.0	
					1.0

What do they tell you? Which variables are highly correlated with one another?

Second, calculate tolerance values for the explanatory variables. This is a two step process: a linear model must be fit to the data and then tolerances extracted and calculated. The car package allows extraction of "Variance Inflation Factors" (VIFs), which are the inverse of tolerance. Tolerance is more useful than VIF because it ranges from zero to one and is what is provided in most statistical software. ***Fit a model using the lm(formula) function, where the formula does not include interactions (use '+' symbols between the explanatory variables), and save the output to an object called "bird_reg1". Now use the vif function to calculate tolerance for your explanatory variables as follows:***

```
> bird_tol <- 1/vif(bird_reg1)
```

Take a look at the output. How do your conclusions from looking at the tolerances compare to those from looking at the correlations? The tolerance data tends to be more informative in this situation because it takes into account the set of variables you are studying, which the correlation analysis does not.

Now use the summary() function to take a look at the multiple regression output for your linear model. Complete the table below.

Adjusted R^2 = **$df_{residual}$ =**

Effect	Coefficient	t	p
Intercept			

Test the assumption that the residuals are normally distributed using a Shapiro-Wilk test, and provide the results. Is the assumption met?

W = **p =**

Finally, how would you interpret your multiple regression analysis above (biologically)?

3. Backward Stepwise Elimination

In having calculated the number of terms that a fully crossed model including five main effects would have, you may be wondering how one can simplify the model without just ignoring interactions, as we did above. There are multiple ways to do this, and some are controversial. First you should check the tolerance and/or intercorrelation of main effects to see if some of them should be eliminated. An approach that is then often taken is called **"backward stepwise elimination"**. Some researchers view this as an integral part of finding the best model that fits your data, while others view it as "data mining", which is doing multiple analyses to just see what happens. Critics of this approach argue that it violates the principle that you should plan your analyses prior to doing the experiment. Repeating different forms of an analysis can be interpreted as "doing things until you get a significant result". An alternative approach would be to think carefully about which interactions may be meaningful and include only those. We will take such an approach later in the semester as well.

To do backward stepwise elimination, you would take the following steps:

1. Do a multiple regression with all the terms included
2. Examine the highest order interactions and eliminate those that are not significant
3. Redo your multiple regression excluding those terms
4. Repeat until the highest order terms are significant

In taking this approach, there are two important rules to adhere to. First, if an interaction term is significant, then you have to keep all of its main effects in the model, *even if they are not significant*. Second, in your final model, all terms (including interactions) should have tolerances >0.1 for reasons discussed above.

We will use the bird abundance data from the previous section to do a backward stepwise elimination. To keep things manageable, only use the explanatory variables Larea, grazing, and years. ***Attach the "bdata" object, if not already done. Fit the fully crossed model as follows:***

```
> bmr1 <- lm(Abund ~ Larea + Graze + Year + Larea:Graze +
        Larea:Year + Graze:Year + Larea:Graze:Year)
```

You could specify this same model formula much more efficiently by using other operators than '+', but typing all of the above will make it easier to modify so as to eliminate terms that are not significant. ***How would you specify the model formula above most efficiently?***

Use summary(bmr1) to view the regression results. Starting with just looking at the highest order interactions, which terms (if any) would you eliminate?

Repeat the multiple regression with the appropriate term(s) eliminated and assign it to object "bmr2". Again, view the summary and check the tolerances. If an interaction has very low tolerance, even if it is significant, it needs to be eliminated. Eliminate highest order interactions with the lowest tolerances first. Continue eliminating terms until you are left with significant terms with acceptable tolerances. For each step of your backward stepwise elimination, list the terms that you eliminate below, and mention why (not significant and/or low tolerance). For your final model, write the terms that remain, marking significant ones with an asterisk. You may find that some terms that were previously significant, no longer are. This frequently happens, showing you how low tolerances can make the coefficients unstable.

Does your biological interpretation of the bird abundance data change from when you did not consider any interactions? If so, how would you now interpret your analysis?

4. Two Ways of Doing Analysis of Covariance

Now that you have learned how to do various forms of ANOVA and multiple regression in R, it is time to build on these techniques by learning Analysis of Covariance, or ANCOVA. ANCOVA is very useful when you have nuisance variables – either continuous or categorical – that you need to account for in your analysis. We will start by expanding the ANOVA design to include a continuous covariate. We will then expand on our multiple regression skills by adding a categorical nuisance variable. This may seem counter-intuitive because ANOVA is normally used for categorical variables and vice versa for multiple regression, but it is correct. To further reinforce this, ***complete the following table, specify whether each variable is categorical or continuous:***

Variable	ANCOVA building on ANOVA	ANCOVA building on Regression
Response		
Explanatory		
Nuisance		

5. ANCOVA with Continuous Nuisance Variables

Using ANCOVA by building on ANOVA when you have a continuous nuisance variable, also called a **covariate**, is simpler than the other form of ANCOVA, so we will cover it first. You are accustomed to ANOVA having one or more categorical explanatory variables, called factors. For this type of ANCOVA, you also have a continuous explanatory variable, and it just so happens that in many instances, this is some sort of nuisance variable. Consider the linear model for this type of ANCOVA:

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij}$$

The response, grand mean, and residual should all be familiar from all the linear models we have studied. the α_i term is just taken from an ANOVA and is your categorical factor. The remaining term refers to the covariate, where β is the partial regression coefficient, x_{ij} is each observation of the covariate, and then you have the covariate mean. Subtracting the covariate mean from the covariate essentially factors out the effects of the covariate from your ANOVA. You can also add other factors and covariates – the above model is the simplest (and most useful) case, with one factor and one covariate.

Recall that when you learned how to do ANOVA in R, if you had a numerical factor, you had to ensure that it was treated as a categorical variable (e.g., three levels of density of mosquito larvae). If you neglected to do that, R would simply treat it as a continuous covariate. At that time, you did not want to do that because you were comparing specific density levels to one another, which cost *p-1* degrees of freedom (where *p* is the number of levels in the factor), and

resulted in a maximally powerful test. But what if you do not have your potentially continuous variable distributed so cleanly? ANCOVA is the answer.

Load the stickleback dataset that you used in lab 1, and assign it to an object called "sdata". Use "str(sdata)" to refresh your memory of the dataset. Data included sex, standard length and mass for a sample of stickleback fish, as well as the number of parasites and total parasite mass in each fish. Do an ANCOVA to see if total parasite mass in each fish differs between the sexes, while taking the standard length of the fish into account as a covariate (you might expect larger parasites to fit in a larger fish). Use the following syntax, and then take the summary of it:

```
> aov(Response~Factor+Covariate)
```

What are your conclusions from your analysis?

Test for normality of the residuals using a KS test. Are they normally distributed? Please report the test statistic and p-value.

One limitation to the ANCOVA that builds on an ANOVA design is that it assumes that there is equality of slopes among factor levels. What this means is that it is assumed that the slope between the response and covariate is the same for each level of the factor. The factor then tests for a difference in intercept. If the slope is different, then comparing intercepts is not very useful (since they will most likely differ). A good way to test this assumption is by fitting a model with the interaction between factor and covariate and seeing if the interaction is significant. ***What would a significant interaction mean?***

Repeat your ANCOVA, but this time include the interaction term. Complete the table below. Is the equality of slopes assumption met? Explain why or why not.

Effect	**DF**	**MS**	**F**	**P**	**partial η^2**
Sex					
SL					
Sex*SL					
Residual					

6. ANCOVA with Categorical Nuisance Variables

As the last part of this lab, we will learn to do ANCOVA in a different way, by building on what we learned about multiple regression. This is a little more involved than the previous version of ANCOVA, but possibly even more useful. In this approach, you have a nuisance variable that is categorical and are studying relationships between two continuous variables. Consider that you could be studying the scaling relationships of body parts in some species and have samples from different populations. ***In this hypothetical example, what is the nuisance variable?***

Normally, to do an ANCOVA in this way, you first need to create a series of variables, called **dummy variables**, that recode your nuisance variable in a binary fashion. If you have *p* groups in your nuisance variable, you will make *p-1* dummy variables, where one of the groups takes a zero for all dummy variables, and each of the remaining groups takes a one for exactly one of the dummy variables. In this way, each observation (row) will have a maximum of a single one, with all other dummy variables having a zero.

Assignment (1 point)

Let's give dummy variable coding a try. If you refer back to the age-at-first-feeding in stickleback dataset from Lab 3, we had samples of fish from four lakes: Bearpaw, Cornelius, Kalmbach, and Willow. Since there are four lakes, you would have p-1 = 3 dummy variables. Complete the table below for this example, where "D#" is each of the dummy variables:

Lake	D1	D2	D3
Bearpaw			
Cornelius			
Kalmach			
Willow			

Now that you understand dummy variables, we can take a look at the linear model for this type of ANCOVA and learn the technique:

$$y_{ij} = \alpha + \beta_0 x_i + \beta_1 D1 + \beta_2 D1 x_i + \cdots + \varepsilon_i$$

Here, y_{ij} is the response variable for individual *i* from nuisance category *j*, α is the intercept, and β_0 is the slope for the relationship between *x* and *y* for the nuisance category that has all dummy variables set to zero. The next two terms are the adjustment in intercept and slope, respectively, for the nuisance category that has D1 = 1, relative to the category that has all dummy variables equal to zero. So, the analysis uses the regression for the category with all dummies set to zero as a reference line. Subsequent terms for intercept (βD#) and slope ($\beta D\# x_i$) are differences from the reference line. The tests of whether these β-values (partial regression coefficients) are significant actually test whether they are significantly different from the reference line, not from zero. Because you include terms for the intercept for each line, the assumption we made earlier, that slopes are equal, is not made in this type of ANCOVA. Finally, to clarify, each line (whether reference or otherwise) is represented by two terms.

Assignment (2 points)

For the following linear model, that corresponds to the lake example above, label each term. Is it a slope or intercept term? Which lake does it correspond to? Also identify what terms represent the reference line, and which ones are differences from the reference line.

$$y_{ij} = \alpha + \beta_0 x_i + \beta_1 D1 + \beta_2 D1 x_i + \beta_3 D2 + \beta_4 D2 x_i + \beta_5 D3 + \beta_6 D3 x_i + \varepsilon_i$$

As you can see, understanding dummy variables is key to understanding how to do this version of ANCOVA. Coding dummy variables and including them in the model formula in R is a perfectly acceptable way of doing this. However, if you fit a model using lm and include a categorical variable as one of your explanatory variables, R automatically implements this sort of ANCOVA without you having to code any dummy variables. Note that most other statistical software require that you code and include dummy variables. Let's do the ANCOVA, letting R make its own dummy variables instead of coding them ourselves.

Assignment (7 points)

Use the Stickleback parasite data from Lab 1, which you should already have in your workspace as "sdata". Attach the data frame so that you can refer to it easily.

Biologically, you are interested in how mass increases with size in threespine stickleback fish. This could occur in different ways for males and females, and so this is your nuisance variable. Use the lm(model) function to fit your ANCOVA model to the data, and assign this model to an object (named whatever you choose). Write the model formula that you specified in R notation below.

Complete the table, generated from the summary of your model below. Instead of using the usual α=0.05, use α=0.10. Under these conditions, what do you conclude from your ANCOVA? Make sure you address whether slope and intercept are different between the sexes, and whether the values for males or females are greater. Include both a statistical interpretation and a biological one.

Effect	Coefficient	t-value	P

Make a scatter plot of mass and SL with the sexes represented using different symbols or colors and a regression line provided for each sex. Add this figure here.

Finally, are the residuals of your ANCOVA normally distributed? Use a Shapiro-Wilk test.