Biol 206/306 – Advanced Biostatistics
Lab 4 – Bivariate Regression
Fall 2016

By Philip J. Bergmann

0. Laboratory Objectives

1. Review Ordinary Least Squares (OLS) regression
2. Place ANOVA and Regression techniques in a common model framework
3. Learn how to do an OLS regression in R
4. Learn to test the assumptions of a regression in R
5. Learn how Reduced Major Axis (RMA) and Major Axis (MA) regression compare to OLS, and when each should be used.
6. Learn how to do RMA and MA regression in R

1. Review of Ordinary Least Squares Regression

OLS regression is the go-to technique for fitting a line of best fit to continuous data. Typically, you have one continuous response variable and one continuous independent variable. OLS regression assumes, like many statistical techniques, that the data are randomly sampled and independent. There are also assumptions that the residuals are normally distributed and homoscedastic. Finally, OLS regression assumes that the *x* (independent) variable is measured without error (or with very little error) – we will get back to this later.

OLS regression is very similar to a single-factor ANOVA, with the main difference being that the regression uses a continuous independent variable, while the ANOVA uses a categorical independent. As such, the linear model for a regression is very similar to that of the ANOVA. For an OLS regression, the model is:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

This is the equation of a line, with y_i and x_i being the response and explanatory variables, respectively, α being the intercept, β being the slope, and ε_i being the residual or error term. ***How does this compare to the model equation for a single-factor ANOVA?***

2. OLS Regression in R

You have already learned the functions needed to do an OLS regression, but have used them for ANOVA instead. OLS regression is done by fitting a linear model to your data. The function is as follows:

```
> lm(formula, data, weights, …)
```

You can specify your formula in the usual form: Y~X and can specify a dataset if needed, as well as weights for the observations. There are many other possible arguments that could be used as well and things can get quite complicated. All we need is the formula argument.

We will fit a linear model to data that you have seen before. Last week you worked with a dataset that examined how the density of mosquito larvae influenced mosquito egg hatching success. You considered density of larvae, species of eggs, and species of larvae as factors in an ANOVA. You will use a subset of the same dataset this week to do an OLS regression. ***Start by loading the dataset into R. You need not go back to the Excel file, as you should have the dataset in tab-delimited text already. Call the resulting object "mdata" again. Now let's simplify the dataset to only consider eggs of*** **Aedes aegypti*:***

```
> crp_mdata <- mdata[1:54,]
```

We will also ignore the effects of larval species – it is sufficient that they all eat bacteria from egg surfaces. ***Attach the cropped dataset and take a look at it and the variables in it. Now try fitting a linear model with "Hatch" as your response, and "Density" as your independent, saving your model as "mosq_lm". Examine your linear model with the str() function to reinforce that his is a pretty complex object class, with many things in it. To get your OLS regression results, simply use the summary() function. What is the intercept and the slope? Is each significantly different from zero?***

What is the R^2 value, your measure of the strength of the relationship between Hatch and Density?

The linear model produced by the lm function is useful beyond just getting a table of regression coefficients and tests of their significance (although these are useful parts). Other functions, similar to summary() give you other information as well, or allow you to extract important information. ***Try the following functions with your "mosq_lm" object:***

```
> coefficients()      # Returns a vector with the slope and intercept.
> residuals()         # Returns a vector with the residuals of the model.
> df.residual()       # Returns the error/residual degrees of freedom, n-2 when bivariate.
```

One other useful trick to getting more information out of your lm object is "str(summary(mosq_lm))". This shows you the structure of the summary object and shows you how you can refer to specific information in the summary. What would you type to have the R^2 value for the model printed to the screen?

Having done an OLS regression, we can now test the assumptions that the technique makes. This is typically done using two tools. First, if one plots the residuals against the independent variable, they can gain a lot of information. Examining this plot helps to ensure that there are no

strange data qualities that would indicate non-independence of the data or non-randomness. Although this is not a perfect approach, it is a useful check. Your plot of the residuals should appear as a random scatter of points with no odd patterns. Note that if the independent variable is discretely sampled (as in the case of larval density) then, there will be columns of dots in the plot of residuals. Within each column, there should be a seemingly random distribution of points. The plot is also a reasonable test of the assumption of homoscedasticity. There should be approximately equal variation in residuals across the range of the independent variable (no funnel-shaped patterns, for example). The homoscedasticity assumption is not testable with regression using a Bartlett test, or similar test, if the independent variable is not categorical (note that the mosquito example doesn't quite fit into this description, but imagine if the larval density varied with sample instead of being fixed at 4, 12, and 24). ***Plot the residuals against the independent variable, either using the GUI of Deducer and JGR, or with the function "plot(x, y)". What can you say about the plot of residuals and whether the regression assumptions are met?***

Second, there are at least two implemented tests of normality in R that you can use to test the assumption of normality:

```
> shapiro.test(x)
> ks.test(x, "pnorm", mean=mean(x), sd=sqrt(var(x)))
```

Both the Shapiro-Wilks and the Kolmogorov-Smirnov tests are perfectly valid tests of normality, and which is used often depends on what an individual statistician is taught. The Shapiro-Wilks test is typically viewed as being more powerful (more likely to find a significant difference), while the KS test is more conservative. Note that the Kolmogorov-Smirnov test requires more arguments because it is a more flexible test, designed to compare two distributions in general. The above arguments implement a KS test that compares the distribution of variable x to a normal distribution with mean and standard deviation equal to those of the variable x. ***Do both tests on the residuals to your regression. How do the results compare? Which test is more stringent?***

Note: you may get a warning about ties being present for the KS test. Ties (the same value for two data points) can decrease the robustness of the test, but you still get output. Also remember that you can use these tests of normality to test the assumptions of ANOVA.

3. RMA and MA Regression

OLS regression is also called Model I regression, similar to an ANOVA with fixed factors, because the independent variable is assumed to be measured without error (and so is fixed). Model II regression relaxes this assumption and is appropriate when both variables being regressed have error associated with them. *Error* refers to not only a scientist making a mistake

or not doing a measurement as well as they could have, but also to imprecision in the instrument used to take the measurement, and also random or unmodeled biological effects associated with a trait. For example, in measuring a trait on a series of clones (so individuals that are genetically-identical), various biological processes can result in non-identical traits.

Reduced Major Axis (RMA) and Major Axis (MA) regressions are model II techniques, and so they are more appropriate to use than OLS regression any time when there is a reasonable amount of uncertainty in the *x* variable. Also note that OLS regression implies causality – that changes in the independent variable result in variation in the response variable. In model II regression, there is no independent and response variable, they are just *x* and *y* (or *y* and *x*). Consider doing a regression on leg length and arm length in a sample of humans. ***Which variable causes the other to change?***

Although both RMA and MA regression take into account that both variables are measured with error, they are not created equal: how their residuals are calculated differs, and since a regression is optimized by minimizing the sum of squared residuals, they give slightly different lines (often almost the same). RMA regression calculates residuals as the area of the triangle made by a data point and the line, while MA regression calculates them as the shortest distance from the line to the point (this distance is perpendicular to the regression line). Although the MA residual is more intuitive than the RMA residual, it has been shown by simulation that MA regression is outperformed by RMA when there is comparable error in both *x* and *y*, and that it is outperformed by OLS when the error in *x* is much smaller than the error in *y*. Hence, we will focus on RMA regression here.

The slope of the RMA line can be easily calculated from the OLS line because $\beta_{RMA} = \beta_{OLS}/r$. Since all regression lines (OLS, RMA, and MA) pass through the point $(\bar{x}, \bar{y})$, the intercept of an RMA regression can simply be calculated from the equation of a line once the slope is known. ***What is the equation of a line? Rearrange the equation to solve for the intercept.***

4. Doing RMA (and MA) Regression in R

Despite the importance of model II regression techniques and their ease of calculation, most software packages do not do them. Even in R, model II regression is not implemented as part of the standard "stats" package. We will use two functions to conduct model II regressions: the package **lmodel2** available online, and the function **rma**, written by Dr. Bergmann. ***Install and load the package "lmodel2" using the standard approach in the "Packages" menu (or "Packages and Data" menu on a Mac). Also download the "RMA_function.txt" file from the course website. Then load the RMA function script as follows: File > Source R Code…, and select the file.***

Type "ls()" and notice that "rma" is now an object in memory. This is because the function isn't part of a package; it is only a text file with the R code. I have used comment lines

(designated with the # symbol) that tell you what each set of lines does. You can try to read through it and reason through the steps taken to do the RMA regression.

The dataset that we will be using to do various model II regressions consists of a series of measurements of various body parts of the horned lizard, *Phrynosoma blainvilli,* from Baja California. The dataset contains measurements of head length, snout-vent length, body width, and the lengths of four horns found on the posterior and lateral margins of the head (P1, P2, S1, S2) for 77 specimens belonging to both sexes. The horns are listed in order from postero-medial to lateral and large to small. The unlabeled column contains specimen numbers. The specimens range from small to large, forming an ontogenetic series, so we can study how the proportions of the horns and body parts change as the animals grow. ***Open the dataset in Excel, save it as a text file, and import it into R as an object named "horns". Attach the object horns so that you can refer to its variables simply.***

We will use the *Phrynosoma* horn dataset to study the allometry or scaling of the horns. Allometry is the study of how body part proportions change during ontogeny. If we consider the allometry of horns relative to body length (SVL), we could see one of three patterns. If the horn maintains the same proportions relative to the body through ontogeny, it is **isometric**. If it gets longer relative to the body as the animal grows, it is **positively allometric**. If it gets relatively shorter, it is **negatively allometric**. A two-tailed one-sample t-test can be used to test whether the slope of the regression differs significantly from one (isometry). Since scaling or allometric relationships follow an exponential equation, in these studies we always log-transform the data so that it is more normal and linear.

Let's start by using the lmodel2 package and function to calculate model II regressions. The package computes slope and intercept for OLS, MA, and RMA regression, but it calls RMA regression "SMA", or Standard Major Axis. The difference in terminology is unfortunate, but the reason we refer to it as RMA in this course is because this name is more prevalent, at least in the biological literature. What is nice about this function is that it also calculates 95% confidence intervals for the slope and intercept, which you can then use to test hypotheses, like the hypothesis of isometry – if the 95% CIs for the slope cross one, then the slope is not significantly different from one. The function also returns the sample size, R, and R^2. It also gives the angle between OLS slopes of X on Y and Y on X, and some other numbers, all of which are not frequently used. The function is as follows:

```
> lmodel2(formula)
```

Where the formula is of the form **Y~X**. Note that you can have only one x variable because this is bivariate regression. Also note that you can log-transform variables in the formula. ***Try:***

```
> lmodel2(log(P1)~log(SVL))
```

Repeat this procedure for head length and the other horns (P2, S1, and S2). Then complete the table on the next page with the information you get for the RMA (=SMA) regression.

The disadvantage to the lmodel2 function is that it does not calculate regression residuals, and so they cannot be tested for the assumption of normality. The **rma** function does this. Because there is no good guidance about how residuals that are tested for normality should be calculated, the rma function calculates them in three ways for the RMA regression: as the area of the

triangle formed by each point and the regression line (the actual RMA residuals), as the square root of the first method (note that areas are in squared units, while all other residuals are Euclidean distances in the same units as the original measurements, hence the square root), and as the shortest distance from the line to the point (as would be done for MA residuals). The function then tests for normality of each set of residuals using the Shapiro-Wilks test, and the Kolmogorov-Smirnov test. Due to the lack of guidance in the literature as to what residuals to use, and different preferences of researchers for the test for normality, anyone using this function can make their own decision as to what type of residuals and what test they want to use.

The syntax for the rma function is:

```
> rma(x, y)
```

Where x and y are the variables that you wish to regress on one another. ***Use the rma function to repeat your regression analysis for HL and the four horns (don't forget to use the logs of the variables. Do you get the same slopes and intercepts?***

Assignment (10 points):

In this sort of study, why is using model II regression superior to using model I regression?

How would a model I regression (OLS) bias the slope estimate? (Hint: think about how you calculate an RMA slope from an OLS slope.)

Complete the table below with information for each RMA regression. In the "Scaling" column note whether each variable is isometric, negatively allometric, or positively allometric relative to SVL.

Y var	R	Intercept	Slope	Low CI	High CI	Scaling
HL						
P1						
P2						
S1						
S2						

What are your biological conclusions from this regression analysis?

Is the KS or the SW test more sensitive to departures from normality?

If you use the KS test for normality for the square rooted RMA residuals, which regressions meet the assumption of residual normality and which ones do not?

What about if you consider the KS test of regular RMA residuals?

For the regression of the horn log(P1) on log(SVL), make a scatterplot of the two variables, and a scatterplot of the square rooted RMA residuals against log(SVL). Also make a histogram of the residuals (Hint: all of these data are in the RMA function output). Insert the two figures with residuals at the end of this document and resize them so that they fit on a page along with your question answers. From examining the figures, why do you think that the SW test suggests that the residuals for this regression are not normal?