Lab 6: Multiple linear regression and ANCOVA

Entomology 5126 - Spring 2026

Jeremy Hemberger and Brian Aukema

February 23, 2026

Table of contents

1 Introduction

Last week, we refreshed our memory of simple linear regression. This week, we will do multiple regression, i.e., examining the effect of multiple continuous variables on some response. Actually, in today's data set, there is a factor included as well. Such analyses are known as Analysis of Covariance, or ANCOVA. ANCOVA is simply a hybrid of regression and ANOVA.

1.1 Factors vs. continuous or discrete variables

As a reminder, if you have a numeric variable (like trials 1 and 2) and you would like to convert it to a factor, use the `as.factor()` command:

```
> myniftydata$trial <- as.factor(myniftydata$trial)
```

If you ever need to turn an appropriate categorical variable (such as above) into a numeric variable, use this command:

```
> myniftydata$trial <- as.numeric(as.character(myniftydata$trial))
```

2 Multiple Regression

The basic formula for a multiple regression model is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \epsilon_i$$

The β_0 is the intercept, or where the line crosses the y-axis. The β_1 is the slope of the first covariate, or the increase in y per unit change in x_1 *holding all other covariates constant.* The β_2 is the slope of the second covariate, or the expected increase in y per unit change of x_2 *holding all other covariates constant.* The ϵ_i refers to the "errors" about the line. Of course, if the model was perfect (never the case), there would be no errors and your data would lie along a perfectly straight line.

There are four assumptions for a multiple linear regression model:

1. The model is correct
2. The observations are independent
3. The variances are equal
4. The errors are normally distributed

As you might expect, you can check the assumptions of equal variances and normally distributed errors in your residual plot. As I mentioned in lecture, you can also detect curvature in the data (first noted when you plotted the data) that might require a special techniques beyond a straight-line fit. I'll introduce some of those techniques next week.

3 ANCOVA

ANCOVA mixes factors (i.e., categories) with covariates (i.e., continuous variables). The basic formula for an ANCOVA model is:

$$y_{ij} = \beta_0 + \alpha_i + ... + \beta_1 x_{1ij} + ... + \epsilon_{ij}$$

As you can see, the model has an intercept, or where the line crosses the y-axis. For levels of a factor, there are α_i values that adjust the intercept level up or down. Then, the response is adjusted by the different slopes of covariates. The ϵ_i refers to the "errors" about the line. If the model was perfect (never the case), there would be no errors and your data would lie along a perfectly straight line.
Again, there are four assumptions:

1. The model is correct
2. The observations are independent
3. The variances are equal
4. The errors are normally distributed

You have seen this before, yes? And you know how to judge these assumptions, yes?

4 Useful commands for fitting ANCOVA models

- The linear model command, `lm()`, as we have used the past two weeks. Specify the response variable, and then the covariates or factors. A + sign is used to specify multiple independent variables. *If* you were examining the interaction terms between your covariates (or factors), a : can be used to fit an interaction between two variables. If you want to fit two variables of interest and their interaction, use an asterisk $*$ as shorthand.

  I will note that interactions are typically left to ANOVA designs. It is not conventional to fit or study interaction terms in multiple regression models, although I suppose you could if you had some really good reasons.

- The anova command, `anova()`. This is not commonly examined in multiple regression since the effect of each variable can be examined as easily using `summary()` (since each slope estimate uses one degree of freedom). The `anova()` command is often used to examine the overall effect of factors in ANOVA and/or ANCOVA (a group with more than 2 levels will need more than one degree of freedom, so the significance test for the variable as a whole cannot simply be recovered from one line in the `summary()` table).

- The summary command, `summary()`. Examine the F-statistic of the overall model, which you can find at the bottom of the summary. This tells you whether there is/are *any* significant relationships between your response variable and your predictor(s). If the p-value is greater than 0.05, then there is nothing significant.

- Remember, while the p-value from the overall model is good to examine, you should also simultaneously examine how well the model fits and if a transformation might be necessary!

5 Sequential vs. Marginal fitting

We have covered in class the differences between sequential and marginal fits in a model; the inferential tests for the variables being examined can be quite different. We'll tease apart some of these details in lab today. One key thing to keep in mind: `summary()` uses marginal fitting, while `anova()` uses sequential.

6 Today's data

Today's data set, in `Lab6data.xls`, is a dataset from D.G. Altman (1991) *Practical Statistics for Medical Research*, Table 12.11, Chapman & Hall. It is a data set of lung function for cystic fibrosis patients between 7-23 years old. There are 25 patients and 10 columns:

Variable Description

age age in years sex 0: male, 1: female height height in cm weight weight in kg bmp body mass (percent of normal) fev1 forced expiratory volume rv residual volume frc functional residual capacity tlc total lung capacity pemax maximum expiratory pressure

7 Today's Assignment

1. Load the data, convert all necessary variables to factors, and examine a scatterplot matrix. The latter step is especially important when doing multiple regression. Make a mental note (not for graded assignment) of which variables are highly correlated with each other, and also which variables might be linearly (or curvily) related to the response variable of interest.

2. T-tests in *all* regression models return a test using a technique known as *marginal fitting.* Marginal fitting examines the effect of each variable as if it was fit into the model *last*, as if all other variables specified were already in the model and accounted for.

   Frequently, a variable associated with a high p-value from a t-test in a multiple regression may exhibit a high (i.e., non-significant) p-value because it is highly correlated with something else already in the analysis. In other words, if you have a included a variable in your analysis that looks an awful lot like another one, chances are it will be non-significant: who needs two copies of the same thing?

   In contrast, F-tests in the `anova()` command return a test using a technique known as *sequential fitting.* The p-value associated with a term examines the effect of the term as fit *in the order specified in the model.* The first term listed in the ANOVA table displays

the p-value for just that term, with nothing else in the model. The second term listed takes into account the variation already explained by the first term, and so on.
Fit a multiple regression model of `pemax` on `height` and `weight` and examine the summary. Why do you get different p-values for one of the terms between the `summary()` and the `anova()` outputs? What is happening?

3. There is one variable, `sex`, in today's dataset that is a factor with two levels. Binary variables are often coded 0/1, which can be advantageous in model fitting. How many degrees of freedom are used in estimating a coefficient for a factor with two levels? How many degrees of freedom are used in estimating a slope for a continuous variable? (If you see where this is going, yell *eureka* now ...).

   - We have covered how factors are treated in the `summary()` output before: one level becomes incorporated into the intercept, the other(s) are differences from that (with associated t-tests). Fit a model of forced expiratory volume as a function of gender and examine the summary. What does the t-test test? Do boys vs. girls blow more hot air? Report the levels for each.

   - Now, convert the `sex` variable to a continuous variable and refit the model. Examine the `summary()` output. Now what does the associated t-test test? Report the y values of the regression line for values of $x = 0$ and then $x = 1$. Where have you seen these values?

4. It is most intuitive to examine *marginal effects* when selecting candidate variables for models. (Think: does this variable explain significant variation, after I have taken into account the effects of everything else specified?) Remember, in `anova()`, factor tests are *sequential* and not necessarily as informative (although statisticians can fight holy wars over the virtues of sequential vs. marginal fitting).
   There is a small challenge when using factors with more than two levels: the output of `summary()` provides an estimate for each of the different levels from the baseline/intercept (all of the α_i). Of course, we can't just eliminate one level of a factor; we have to either keep or retain the entire variable! To get the overall effect of the factor in question, we of course have to use the `anova()` command, which yields a sequential test.
   There are ways to coerce statistics programs to provide automatic marginal fits for factors (sometimes called Type III SS). But let's say all we had is the two commands `anova()` and `summary()`. What is an easy way to find the marginal fit for a specific factor using `anova()`?

5. Back to our data set. Let's build a good model to explain maximum expiratory pressure, or `pemax`.
   In general, the simplest model is the best. We should try to select models with covariates

that contribute significant information to the model (by convention, we always leave the intercept in), and remove extraneous variables. If a variable is not significant after having everything else in the model, we might as well remove it.

One popular method of model building is known as "backward elimination." Start with a plausible "full model", examine the summary, and remove the variable with the highest non-significant p-value. Fit the slimmer model and repeat. Continue until all variables left in the model are significant (using $\alpha = 0.05$).

Now, it is important to realize that (1) *you are the boss* and (2) *you do not always have to remove the variable with the highest non-significant* p-value. In practice, there may be good *ecological* reasons to leave a particular explanatory variable in. Let's say we were examining the effect of multiple variables on a response y, plant growth. If faced with a choice between eliminating either

1. the price of tea on successive Tuesdays ($p = 0.6216$), or
2. precipitation ($p = 0.6826$),

we may chose to retain the latter for an elimination round or two even if it does have the highest p-value. Such decisions are part of the art of doing statistics.

In practice, we often check model diagnostics through residual plots, transformations, etc. at the beginning and end of our work, as it would be highly tedious to check every step of model building (e.g., pretend we had 20 or 200 candidate variables).

If we did have 30 candidate predictor variables, could we include them in a first model all at the same time for this data set? Why or why not? (Hint: think of the ANOVA table).

6. Our goal is to predict the maximum expiratory pressure of a child with cystic fibrosis based on biological indicators. Using backward selection procedures, find a model of `pemax` on any or all of `sex`, `fev1`, `age`, `height`, `weight`, `bmp`, `frc`, and `tlc`. **Do *not* simply turn in 10 pages of printout.** Instead, for each variable elimination step, report which variable you removed (or retained) and why. Also, for each step, report both the multiple and adjusted R-squared values. What do you notice about the trend in the *difference* between these two values as you proceed?

   Present your results of a sensible regression model in a brief paragraph. Provide the final model equation, with F and P values for the overall model. (Note: there may be a variety of acceptable models). Be sure to explain any necessary transformations and the range of data over which your model might be tractable.