**http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm**

**Lecture 1—Wednesday, August 25, 2010**

**Topics**

* [Simple linear regression](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#simple)
* [Multiple regression](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#multiple)
* [Interactions in multiple regression](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#interactions)
  + [When should interactions be included in models?](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#when)
* [Categorical predictors in regression](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#categorical)
  + [Coding categorical predictors](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#coding)
  + [Interactions of categorical predictors: two-factor interactions](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#twofactor)
* [R code used in lecture 1](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#Rcode)

**Terminology**

* [additive model](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#additive)
* [analysis of variance](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#anova)
* [baseline level](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#baseline)
* [dummy variable](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#dummy)
* [F test](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#Ftest)
* [indicator variable](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#dummy)
* [main effects](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#additive)
* [observational study](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#when)
* [partial F test](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#partial)
* [principle of marginality](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#interactions)
* [profile (interaction) plot](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#twofactor)
* [reference level](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#baseline)
* [regressor](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#dummy)
* [R2](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#R2)
* [two-factor interaction](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#twofactor)
* [treatment levels](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#categorical)

**Simple linear regression**

In simple linear regression we attempt to model the relationship between a single response variable *y* and a predictor *x* with an equation of the following form.

reg1

This is just the equation a line with intercept β0, slope β1 in which we allow that the fit won't be perfect so that there is error represented by ε. Typically we assume that reg2, i.e., that the errors are independent and are drawn from a common normal distribution. Equivalently we could also write

reg3

From this formulation we see that the regression line is the mean of the response variable *y*, i.e., reg4, so that the mean changes depending on the value of *x*. The parameters of the regression model are typically estimated using least squares.

To illustrate these ideas and to examine the typical linear regression output from statistical packages, I generate some random data, then use it to construct response variables, and finally estimate some simple linear regression models using R.

**#generate data for example**

set.seed(10)

x1<-runif(90)

x2<-rbinom(90,10,.5)

x3<-rgamma(90,.1,.1)

**#organize predictors in data frame**

mydata<-data.frame(x1,x2,x3)

**#create noise**

epsilon<-rnorm(90,0,3)

**#generate response: additive model plus noise, intercept=0**

mydata$y<-2\*x1+x2+3\*x3+epsilon

**#simple linear regression with x1 as predictor**

out0<-lm(y~x1,data=mydata)

**#plot regression line and mean line**

plot(y~x1,data=mydata)

abline(h=mean(mydata$y),col='pink',lwd=3)

abline(out0,lty=2)

**#simple linear regression with x3 as a predictor**

out1<-lm(y~x3,data=mydata)

**#graph regression line and mean line**

plot(y~x3,data=mydata)

abline(out1)

abline(h=mean(mydata$y),col=2,lwd=2)

Individual regression lines of y versus x1 and y versus x3 are shown in Fig. 1. A horizontal line located at the mean of the response variable is included for reference. Without additional information or predictors, the sample mean of the response is the best predictor of a new value of y.

|  |  |
| --- | --- |
| (a) fig1a | (b) fig1b |
| **Fig. 1** Individual regression lines for y versus (a) x1 and (b) y versus x2. The pink horizontal line denotes the response mean. | |

From the graphs it would appear that y is linearly related to x3 but perhaps not to x1. (Observe also that it may be the case that a single point in Fig. 1b is having a profound effect on the location of the regression line.) To assess the fit of the lines we can examine the output from the regression.

summary(out0)

Call:  
lm(formula = y ~ x1, data = mydata)

Residuals:  
Min 1Q Median 3Q Max   
-12.178 -4.564 -1.986 1.389 97.676

Coefficients:  
Estimate Std. Error t value Pr(>|t|)   
(Intercept) 9.6623 2.4493 3.945 0.00016 \*\*\*  
x1 -0.5975 4.8060 -0.124 0.90134   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.86 on 88 degrees of freedom  
Multiple R-squared: 0.0001756, Adjusted R-squared: -0.01119   
F-statistic: 0.01546 on 1 and 88 DF, p-value: 0.9013

The most useful features of the output are highlighted in yellow. We see that the estimated regression line is 9.66 – 0.60x1. The highlighted p-value in the column Pr(>|t|) is from a test of the null hypothesis that β1 = 0, i.e., that the slope of the regression line is zero. Because the reported p-value is large, p = 0.90, we fail to reject the null hypothesis. Hence we fail to find evidence that the response y is linearly related to the predictor x1. This is also confirmed by the statistic labeled "Multiple R-squared" a commonly-used goodness of fit statistic. It measures the proportion of the original variability in the response that has been explained by the linear regression. In this case that proportion is almost zero telling us that the regression model is essentially worthless.

When we turn to the linear regression of y on x3, the results appear more promising.

summary(out1)

Call:  
lm(formula = y ~ x3, data = mydata)

Residuals:  
Min 1Q Median 3Q Max   
-9.0089 -2.4768 -0.3318 1.7561 10.3645

Coefficients:  
Estimate Std. Error t value Pr(>|t|)   
(Intercept) 6.3972 0.3965 16.13 <2e-16 \*\*\*  
x3 3.1489 0.1080 29.16 <2e-16 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.632 on 88 degrees of freedom  
Multiple R-squared: 0.9062, Adjusted R-squared: 0.9051   
F-statistic: 850.3 on 1 and 88 DF, p-value: < 2.2e-16

The reported regression line is 6.40 – 3.15x3. So, a one unit change in x3 yields a 3.15 unit change in the response. The reported p-value for the test of whether the slope is equal to zero is extremely small, reported as zero to 16 decimal places. So, we reject the null hypothesis and conclude that there is a significant linear relationship between y and x3. The reported R2 = 0.91 is also exceedingly high. It tells us that 91% of the original variability of the response is explained by its linear relationship with x3.

The reason that R2 is so high here is largely because of the outlier we observed in Fig. 1b. That point is a long way from the mean response and hence is a major contributor to the variability of the response. The fitted regression line passes almost exactly through this point and so that it contributes nothing to the variability about the regression line. Not surprisingly if we remove this point and refit the regression line to the remaining points, the reported R2 is much reduced.

**#remove outlier in x3 space**

mydata1<-mydata[mydata$x3<25,]

**#verify that one observation was removed**

dim(mydata)

dim(mydata1)

**#refit model to reduced data**

out1a<-lm(y~x3,data=mydata1)

summary(out1a)

Call:  
lm(formula = y ~ x3, data = mydata1)

Residuals:  
Min 1Q Median 3Q Max   
-9.0012 -2.4907 -0.3944 1.7376 10.3448

Coefficients:  
Estimate Std. Error t value Pr(>|t|)   
(Intercept) 6.4170 0.4250 15.10 <2e-16 \*\*\*  
x3 3.1128 0.2894 10.76 <2e-16 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.653 on 87 degrees of freedom  
Multiple R-squared: 0.5708, Adjusted R-squared: 0.5658   
F-statistic: 115.7 on 1 and 87 DF, p-value: < 2.2e-16

R2 is now reported to be 0.57. This tells us that R2 as a goodness of fit statistic is highly sensitive to outliers.

Of course just because a point is an outlier doesn't mean that it's also influential. If we compare the parameter estimates reported after removing the outlier to those that were obtained from fitting the model to the full data set, we see that they've barely changed.

coef(out1)

(Intercept) x3   
6.397223 3.148869

coef(out1a)

(Intercept) x3   
6.416984 3.112826

When we include both lines in the same plot we see that they are nearly indistinguishable.

plot(y~x3, data=mydata1)

**#regression line without outlier**

abline(out1, lwd=3, col='grey70')

abline(h=mean(mydata1$y), col='pink', lwd=3)

**#regression line with outlier**

abline(out1a, col=2, lty=2)

legend('topleft', c('regression with full data', 'regression with reduced data', 'mean line'), col=c('grey70',2,'pink'), lty=c(1,2,1), lwd=c(3,1,3), bty='n', cex=.9)

|  |
| --- |
| fig2 |
| **Fig. 2** Comparison of the regression lines of y on x3 with and without an observation that is an outlier in x3-space. |

**Multiple regression**

Multiple regression is really just a trivial extension of simple linear regression in which one predictor is replaced with multiple predictors. Excluding those situations where the additional predictor is just a function of the original predictor (such as in a quadratic model), a regression model with two predictors defines a plane in three-dimensional space. The geometric object defined by regression models with three or more predictors is called a hyperplane. Without a satisfactory way to visualize multiple regression models geometrically, we're forced to explore them analytically.

As an illustration I fit a linear regression model with three predictors. The regression model assumes the mean is given by

multiple regression model

The output of special interest from this model is highlighted in yellow.

out2<-lm(y~x1+x2+x3,data=mydata)

summary(out2)

Call:  
lm(formula = y ~ x1 + x2 + x3, data = mydata)

Residuals:  
Min 1Q Median 3Q Max   
-8.3681 -2.0349 -0.2121 1.6987 10.0899

Coefficients:  
Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.3468 1.5363 1.528 0.1303   
x1 2.6665 1.4303 1.864 0.0657 .   
x2 0.5756 0.2646 2.175 0.0324 \*   
x3 3.1128 0.1073 29.014 <2e-16 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.517 on 86 degrees of freedom  
Multiple R-squared: 0.9141, Adjusted R-squared: 0.9111   
F-statistic: 305 on 3 and 86 DF, p-value: <2.2e-16

The Estimate column tells us that estimated equation for the mean response is given by μ = 2.35 + 2.67x1 + 0.58x2 + 3.11x3. Each coefficient in this equation is referred to as a partial regression coefficient (to distinguish it from the coefficient in a simple linear regression model). A partial regression coefficient is a regression coefficient obtained after controlling for the effects of other variables. So, for instance, having controlled for the effect of x2 and x3, we see that a one unit increase in x1 yields a 2.67 unit change in the response. Observe that is very different from the regression coefficient that was estimated in the simple linear regression model in which x1was the only predictor. There the estimated coefficient was negative (although not significantly different from zero).

Each of the reported p-values in the coefficients table is a variables-added-last test. Thus p = 0.0657 for x1 is a test of H0: β1 = 0 given that x2 and x3 are already in the regression model. A similar interpretation holds for the reported p-values for the tests of the coefficients of x2 and x3. Notice that the coefficient of x1 is reported to be almost significant (at α = .05), p = 0.066, whereas in the simple linear regression model it wasn't even close to being significant, p = 0.901. A possible explanation for this is that most of the variability in the response is due its linear relationship with x3 . In the simple linear regression model the variability in the response due to x3 swamped the variability in the response due to x1 (in part because of the very different scales that x1 and x3 are measured on). Having removed most of the variability due to x3 by including x3 in the multiple regression model, the linear relationship between y and x1 was allowed to emerge.

The p-value that appears at the bottom of the output is for the reported **F-test** that tests the overall statistical significance of the regression model. Formally the F-test tests the following hypothesis.

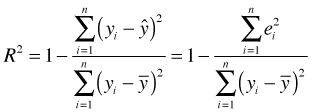
mult hypothesis

Because the p-value for this test is very small, we reject the null hypothesis and conclude that the response is linearly related to one or more of the predictors (when controlling for the rest).

The reported Multiple R-squared statistic is also an intuitive goodness of fit statistic in multiple regression. When we fit a regression model we replace the sample mean as our best guess for the value of the response variable with the estimated regression surface. Before we fit the regression model the variability of the response is measured with respect to the sample mean. After fitting a regression model we replace the sample mean with the regression surface. In the regression model the mean is no longer assumed to be constant but instead varies with the values of the predictors. Consequently variability is measured with respect to the regression surface. The R2 statistic compares these two measures of variability, the variability after we fit the model to the variability before we fit the model (Fig. 3).

|  |  |
| --- | --- |
| http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture1/fig3a.jpg | http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture1/fig3b.jpg |
| **Fig. 3** The two measures of variability that are being compared in the R2 statistic. (a) illustrates the variability about the sample mean and (b) illustrates the variability about the regression line. The amount that the variability has decreased in going from (a) to (b) divided by how much variability there originally was in (a) defines R2. | |

Instead of using average variability, as measured by the variance for example, R2 uses total variability, which is the variability not normalized by sample size. If we let yhatdenote the estimate of the response from the regression equation then R2 can be formulated as follows.



Here *n* is the number of observations and ei is the residual, i.e., the deviation of the ith observed response value from its estimate on the regression surface. As a demonstration of the formula I calculate the R2 using the formula and compare it to the number that is reported in the regression output.

SStotal <- sum((mydata$y-mean(mydata$y))^2)

SSE <- sum(out2$residual^2)

1-SSE/SStotal

[1] 0.9140759

summary(out2)$r.squared

[1] 0.9140759

The statistic labeled "adjusted R-squared" in the output is obtained by modifying the R2 formula so that the two sums of squares terms are replaced by their averages in which unbiased estimates are used for each average. The "adjusted R-squared" can be interpreted as a penalized R2. Adjusted R-squared is not necessarily better than R2 but it has a certain appeal . Ordinary R2 can never decrease when variables are added to a model. It must go up or at worse stay the same. Adjusted R-squared on the other hand can decrease if a newly added variable doesn't decrease the total variability enough. Thus adjusted R-squared can be used in model selection, whereas ordinary R2 is not particularly good for this purpose except in very simple cases.

**Interactions in multiple regression**

Interaction terms in multiple regression models with continuous predictors are formed by taking the products of predictors. The inclusion of such cross-product terms has the geometric effect of changing a planar surface into a curved surface, often referred to as a response surface. The general rule is that any time an interaction is included in a model, then the individual terms that comprise the interaction should also be included in the model. This is called the **principle of marginality**. Although models that violate the principle of marginality are legitimate, they are nearly always silly and hence not worth considering.

With three predictors there are three possible two-factor interactions to consider. There is also one three-factor interaction. The following R code fits a model that contains all two-factor interactions and a second model that contains the three-factor interaction. In each case I adhere to the principle of marginality and include all appropriate lower level effects.

**#3-factor: only two-factor interactions plus main effects**

out3<-lm(y~x1+x2+x3+x1:x2+x2:x3+x1:x3,data=mydata)

**#3-factor and all 2-factor**

out4<-lm(y~x1+x2+x3+x1:x2+x2:x3+x1:x3+x1:x2:x3,data=mydata)

To simultaneously test whether a group of terms is needed in a model we can carry out what's called a **partial F-test**. A partial F-test compares two nested models to determine if the terms that are present in one model, but not the contained in the other, are needed. I start by carrying out a partial test for the 3-factor interaction. When only one term is involved (as is the case here for the three-factor interaction), a partial F-test is the same as the variables added last t-test that appears in the summary table.

**#partial F-test for the three-factor interaction**

anova(out3,out4)

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x1:x2 + x2:x3 + x1:x3  
Model 2: y ~ x1 + x2 + x3 + x1:x2 + x2:x3 + x1:x3 + x1:x2:x3  
Res.Df RSS Df Sum of Sq F Pr(>F)  
1 83 1030.0   
2 82 1013.7 1 16.315 1.3198 0.2540

**#we get an identical result with a variables-added-last t-test**

summary(out4)

Call:  
lm(formula = y ~ x1 + x2 + x3 + x1:x2 + x2:x3 + x1:x3 + x1:x2:x3,   
data = mydata)

Residuals:  
Min 1Q Median 3Q Max   
-7.9378 -1.9017 -0.3867 1.5942 10.0821

Coefficients:  
Estimate Std. Error t value Pr(>|t|)   
(Intercept) 3.9131 2.9114 1.344 0.18264   
x1 -0.9381 5.9513 -0.158 0.87513   
x2 0.2030 0.5351 0.379 0.70535   
x3 5.5606 1.8483 3.009 0.00349 \*\*  
x1:x2 0.8836 1.1137 0.793 0.42983   
x2:x3 -0.3052 0.2494 -1.224 0.22445   
x1:x3 -6.3719 4.6214 -1.379 0.17171   
x1:x2:x3 0.8125 0.7073 1.149 0.25396   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.516 on 82 degrees of freedom  
Multiple R-squared: 0.9181, Adjusted R-squared: 0.9111   
F-statistic: 131.4 on 7 and 82 DF, p-value: < 2.2e-16

From the output we conclude first that the three-factor interaction is unnecessary, so we can drop it from the model.

There are three two-factor interactions and we can use a partial F-test to determine if we need to bother with any interactions at all. The partial F-test simultaneously tests whether any of the coefficients of the two-factor interactions are different from zero. This partial F-test is not equivalent to the summary table output because there each individual test is a test for that specific two-factor interaction given that the other two two-factor interactions are already in the model. To carry out the partial F-test I compare a main effects model with a model that contains the main effects plus all of the two-factor interactions.

**#test two-factor interactions**

anova(out2,out3)

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3  
Model 2: y ~ x1 + x2 + x3 + x1:x2 + x2:x3 + x1:x3  
Res.Df RSS Df Sum of Sq F Pr(>F)  
1 86 1063.7   
2 83 1030.0 3 33.752 0.9066 0.4415

So, the high p-value tells us that we don't need to include two-factor interactions in the model.

**When should interactions be included in models?**

As a general rule, interactions should always be examined with experimental data, and rarely examined for observational data. **Observational studies** are quasi-experimental designs that fall short of being true experiments for various reasons. In a typical observational study treatments are imposed by nature rather than the experimenter. As a result there is no guarantee that treatments have been randomly assigned to subjects and rarely any balance causing some treatment combinations to be under-represented. All of this makes assessing interaction in observational studies dangerous. Main effects are hard enough to assess in such studies; interactions are truly pushing the envelope.

Based on these considerations I approach the statistical analysis of experiments and observational studies quite differently. In an experiment in which all relevant factors have been assiduously controlled and in which subjects have been randomly assigned to treatments I typically start with the most complicated interaction model possible and try to simplify it. On the other hand when I analyze observational data I start with main effects and maybe tentatively examine a few interactions that have a theoretical basis.

The problem is that it's too easy to obtain significant interactions with observational data that are not real, but arise because of sampling bias. By this I mean cases where important control variables may have been left out of the analysis (especially if they have a non-random distribution in the sample) or where this is just data imbalance in the predictors leading to an effect on the response analogous to genetic drift. The basic problem is that observational data are often not random samples and without random assignments to treatment there are just way too many uncertainties in inference.

If one of the measured variables is categorical then invariably with observational data there will be unequal numbers of observations in the various categories. The imbalance may not be severe enough to cause any inference problems when the variable occurs only as a main effect, but when you interact it with another variable (say also categorical to keep things simple), this imbalance gets carved up even further. The distribution of the first variable may start to look entirely different across the levels of the second variable just by luck. At this point genetic drift kicks in and the distribution of the response also looks different. All of this gets interpreted as an interaction when in fact it's a consequence of differences in the distributions of the predictors. Simpson's paradox is not exactly an example of this, but it is an inferential problem that is in the same spirit.

**Categorical predictors in regression**

Categorical predictors pose special problems in regression models. By a categorical predictor I mean a variable that is measured on a nominal (or perhaps ordinal) scale whose values serve only to label the categories. The categories of a categorical variable are referred to as its **levels**.

The common situation with experimental data is that the response variable is continuous but all of the predictors are categorical. In this situation the categorical predictors are often referred to as treatments. The treatments may be generated artificially by dividing the scale of a continuous variable into discrete categories. Using categories makes it possible to obtain results with fewer observations and avoids the problem of determining the proper functional form for the relationship between the response and continuous treatment. Categorization is also useful in preliminary work where the primary interest is to determine whether a treatment has any effect at all. Typical choices for the discretization are:

* presence, absence (corresponding to treatment and control)
* high, low
* high, medium, low

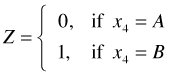
It's also common for the treatment to be intrinsically categorical. For instance in a competition experiment the "treatment" may be the identity of the species that is introduced as the competitor. In this case the categories would be the different species.

The analysis of the relationship between a continuous response and a set of categorical variables (predictors) is easily handled by standard regression techniques. Historically though the methodology developed to handle this situation was called **analysis of variance**, or ANOVA. In its modern implementation ANOVA is just regression although its usual presentation may not look like it. As an organizational tool ANOVA turns out to be a very clever way of extracting the maximal amount of information from a regression problem in which the response is continuous and the predictors are categorical.

**Coding categorical predictors**

To say analysis of variance is just regression with categorical predictors avoids an obvious question, namely how do you include categorical predictors in a regression model? As noted above categorical predictors are nominal variables, i.e., their values serve merely as labels for categories. So even if the categories happen to be assigned numerical values such as 1, 2, 3, …, those values don't mean anything. They're just labels. They could just as well be 'a', 'b', 'c', etc.

Suppose we have a nominal variable x4 with two categories *A* and *B*. The usual way to include a categorical variable in a regression model is by creating a set of **dummy (indicator) variables**. The following dummy variable could be used to represent this variable in a regression model.



To include the predictor x4 in a regression model we use the variable Z (called now a **regressor**) instead. As usual the regression equation gives the mean of the response variable Y.

regdummy

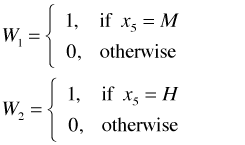
Because Z can take only two values, the regression model returns only two different values for the mean, one for each category of the original variable x4. Table 1 summarizes the relationships between the model predictions and the original categories.

|  |
| --- |
| **Table 1** Regression model for a categorical predictor with two levels |

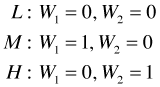
| x4 | Z | regdummy |
| --- | --- | --- |
|  |  |  |
| *A* | 0 | muA |
| *B* | 1 | muB |

The regression coefficient β1 represents the amount the mean response changes when we switch from category *A* to category *B*. As a result, a test of H0: β1 = 0 is a test of whether the means of categories *A* and *B* are the same.

The true essence of the dummy variable method becomes clear when we consider a nominal variable that has more than two categories. Suppose x5 is a nominal variable with categories *L*, *M*, and *H* (corresponding to the low, medium, and high levels of a treatment). Converting this predictor into a form suitable for use in a regression model requires the creation of two dummy variables as shown below.



This coding scheme uniquely identifies each of the three categories as follows.



(Observe that the combination *W*1 = 1 and *W*2 = 1 is a logical impossibility.) The category we obtain when all of the dummy variables are set equal to zero is called the **reference** or **baseline category**. So in the above coding scheme the "*L*" category is the reference category.

To include the predictor x5 in a regression model we need to include both of the regressors *W*1 and *W*2 . Once again the regression equation is for the mean of the response variable Y.

reg dummy2

In this case the regression model returns three different values for the mean, one for each category of the original variable x5. Table 2 summarizes the relationships between categories, dummy variables, and model predictions.

|  |
| --- |
| **Table 2** Regression model with a categorical predictor with three levels |

| x5 | W1 | W2 | reg dummy2 |
| --- | --- | --- | --- |
|  |  |  |  |
| L | 0 | 0 | muA |
| M | 1 | 0 | muB |
| H | 0 | 1 | regdummy3 |

The regression coefficient β1 represents the amount the mean response changes when we switch from category *L* to category *M*. A test of H0: β1 = 0 then is a test of whether the means of categories *L* and *M* are the same. The regression coefficient β2 represents the amount the mean response changes when we switch from category *L* to category *H*. A test of H0: β2 = 0 then is a test of whether the means of categories *L* and *H* are the same.

To illustrate these ideas I create categorical variables x4 and x5 and generate a response variable that varies according to the categories. The **factor** function in R is used to generate dummy regressors from categorical predictors.

**#categorical predictors**

x4<-rep(c('A','B'), c(45,45))

x5<-rep(c('L','M','H'),30)

**#noise for response**

set.seed(20)

epsilon<-rnorm(90,0,2)

**#create dummy variables to generate response**

z1<-(x4=='B')

w1<-(x5=='M')

w2<-(x5=='H')

**#response generated from two-factor interaction plus error**

y<- 10+5\*z1+2\*w1+4\*w2-4\*z1\*w1+z1\*w2+epsilon

mydata1<-data.frame(x4cat=factor(x4), x5cat=factor(x5, levels=c('L','M','H')), y)

**#categorical predictors**

x4

[1] "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A"  
[16] "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A"  
[31] "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "A"  
[46] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"  
[61] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"  
[76] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"

x5

[1] "L" "M" "H" "L" "M" "H" "L" "M" "H" "L" "M" "H" "L" "M" "H"  
[16] "L" "M" "H" "L" "M" "H" "L" "M" "H" "L" "M" "H" "L" "M" "H"  
[31] "L" "M" "H" "L" "M" "H" "L" "M" "H" "L" "M" "H" "L" "M" "H"  
[46] "L" "M" "H" "L" "M" "H" "L" "M" "H" "L" "M" "H" "L" "M" "H"  
[61] "L" "M" "H" "L" "M" "H" "L" "M" "H" "L" "M" "H" "L" "M" "H"  
[76] "L" "M" "H" "L" "M" "H" "L" "M" "H" "L" "M" "H" "L" "M" "H"

The **contrasts** function allows us to view the dummy coding scheme that was created by R. Observe that the dummy coding displayed by R is identical to that shown in Tables 1 and 2 above.

**#examine regressors created by factor function**

levels(mydata1$x4cat)

[1] "A" "B"

contrasts(mydata1$x4cat)

B  
A 0  
B 1

levels(mydata1$x5cat)

[1] "L" "M" "H"

contrasts(mydata1$x5cat)

M H  
L 0 0  
M 1 0  
H 0 1

To fit a regression model with both x4 and x5, I use their factor versions, x4cat and x5cat. The model shown below is called the **additive** or **main effects model**.

additive

**#main effects model without interaction**

out5<-lm(y~x4cat+x5cat, data=mydata1)

**#test main effects model assuming no interaction**

anova(out5)

Analysis of Variance Table

Response: y  
Df Sum Sq Mean Sq F value Pr(>F)   
x4cat 1 524.09 524.09 97.761 7.704e-16 \*\*\*  
x5cat 2 455.25 227.62 42.460 1.490e-13 \*\*\*  
Residuals 86 461.04 5.36   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The anova output reports what are called sequential partial F-tests. In terms of the regression model given above, the first reported p-value for x4cat is a test of H0: β1 = 0 but without the variable x5 being in the model. It compare a model that includes x4 to a model that contains no predictors, only an intercept. The second p-value for x5cat is a test of H0: β2 = β3 = 0 given that the variable x4 is already in the model. It compares a model with x4 to a model that contains both x4 and x5. If we wanted to test x5 in isolation from x4, we would need to reverse the order of the variables when we fit the model.

From the output we can conclude that both x4 and x5 have significant effects on the mean response. To understand the nature of the effects we need to look at the summary table.

**#examine individual categories**

summary(out5)

Call:  
lm(formula = y ~ x4cat + x5cat, data = mydata1)

Residuals:  
Min 1Q Median 3Q Max   
-5.68420 -1.68756 -0.06405 1.61572 5.19102

Coefficients:  
Estimate Std. Error t value Pr(>|t|)   
(Intercept) 9.9048 0.4881 20.292 < 2e-16 \*\*\*  
x4catB 4.8263 0.4881 9.887 7.70e-16 \*\*\*  
x5catM 0.0845 0.5978 0.141 0.888   
x5catH 4.8127 0.5978 8.050 4.18e-12 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.315 on 86 degrees of freedom  
Multiple R-squared: 0.6799, Adjusted R-squared: 0.6688   
F-statistic: 60.89 on 3 and 86 DF, p-value: < 2.2e-16

* The p-value for x4catB is a test of H0: β1 = 0 and is the same as the partial F-test result in the anova output. Referring back to Table 1, the reported significant p-value for x4catB tells us that the means for the *A* and *B* groups of x4 are significantly different. From the estimate we conclude that the mean of the *B* group is 4.83 units higher than the mean for the *A* group.
* The p-value for x5catM is a test of H0: β2 = 0. Referring back to Table 2, the fact that it is not significant tells us that the means of the *M* and *L* groups are not significantly different.
* The p-value for x5catH is a test of H0: β3 = 0. The fact that it is not significant tells us that the means of the *H* and *L* groups are not significantly different.

If we want to obtain a test comparing group *H* with group *M*, we should define the factor variable for x5 using a different reference group and refit the model.

**#rearrange x5 levels so M is the baseline**

mydata1<-data.frame(x4cat=factor(x4), x5cat=factor(x5, levels=c('L','M','H')), x6cat=factor(x5,levels=c('M','L','H')),y)

**#test M is different from H in main effects model**

out5a<-lm(y~x4cat+x6cat, data=mydata1)

summary(out5a)

Call:  
lm(formula = y ~ x4cat + x6cat, data = mydata1)Residuals:  
Min 1Q Median 3Q Max   
-5.68420 -1.68756 -0.06405 1.61572 5.19102

Coefficients:  
Estimate Std. Error t value Pr(>|t|)   
(Intercept) 9.9893 0.4881 20.465 < 2e-16 \*\*\*  
x4catB 4.8263 0.4881 9.887 7.70e-16 \*\*\*  
x6catL -0.0845 0.5978 -0.141 0.888   
x6catH 4.7282 0.5978 7.909 8.08e-12 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.315 on 86 degrees of freedom  
Multiple R-squared: 0.6799, Adjusted R-squared: 0.6688   
F-statistic: 60.89 on 3 and 86 DF, p-value: < 2.2e-16

The mean of the *M* group is significantly different from the mean of the *H* group.

**Interactions of categorical predictors: two-factor interactions**

Interactions can be understood graphically by displaying treatment means in what is sometimes called a **mean profile** or **interaction plot**. Suppose we have two factors of interest labeled A and B. The mean response is plotted on the *y*-axis and the levels of one of the factors, say factor A, are displayed on the *x*-axis. The mean values of the response at all combinations of the levels of A and B are plotted and those values that correspond to the same level of B are connected by line segments. The connected B line segments define what are called the B profiles. Fig. 4 illustrates this for two factors A and B in which each has two levels designated "low" and "high".

|  |  |
| --- | --- |
| (a) http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture1/fig4a.png | (b) http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture1/fig4b.png |
| **Fig. 4** Understanding two-factor interactions. (a) A two-factor interaction between A and B is absent. The B profiles are parallel and so the A main effect and B main effect are well-defined. Increasing either A or B from its low to high level has the effect of increasing the response. The change is the same regardless of the value of the other variable. (b) A substantive two-factor interaction A × B is present. The effect of each factor depends upon the level of the other. Here when B is at its low level the effect of increasing A from low to high is to increase the value of the response. When B is at its high level the effect of increasing A from low to high is to decrease the value of the response. | |

If two factors fail to interact then they have the same effect on the response variable regardless of the value of the other factor. These unchanging effects are referred to as the **main effects** of the factor. The way an absence of interaction manifests itself in an interaction plot is that the profiles are all parallel. Fig. 4a displays this situation. The magnitudes of the main effects of A and B are indicated on the graph.

If on the other hand there is a two-factor interaction between factors A and B, denoted A × B, then one can't speak of the main effects of A and B per se. Instead the effect of A on the response depends on the level of B. Similarly the effect of B on the response depends on the level of A. For clarity let's suppose the two-factor interaction is such that when B is at a high level the effect of changing A from low to high is negative (decreases the response), but when B is at a low level, the effect of changing A from low to high is positive (increases the response). Fig. 4b depicts this situation.

Assessing interaction graphically is only slightly more complicated when one of the variables has three levels. The key again is to create mean profiles for one variable stratified by the values of the second and assess whether the profiles are sufficiently parallel. Below I use R functions from the effects package to produce the interaction plot.

**#two-factor interaction model**

out4<-lm(y~x4cat+x5cat+x4cat:x5cat, data=mydata1)

**#create a new response without interaction: additive model**

y2<- 10+5\*z1+2\*w1+4\*w2+epsilon

mydata2<-data.frame(x4cat=factor(x4), x5cat=factor(x5, levels=c('L','M','H')),y2)

**#fit 2-factor interaction model to new response**

lm(y2~x4cat\*x5cat, data=mydata2)-> out6

**#create interaction plot: install effects package first**

library(effects)

**#compare models for response with and without interaction**

plot(effect('x4cat:x5cat', out6), multiline=T, row=1, col=1, nrow=1, ncol=2, more=T, ylab='y')

plot(effect('x4cat:x5cat' ,out4), multiline=T, row=1, col=2, nrow=1, ncol=2, more=F)

|  |
| --- |
| fig5 |
| **Fig. 5** Graph (a) depicts a situation in which there is no interaction between x4 and x5. Graph (b) depicts a situation in which a significant two-factor interaction is present. As was the case in Fig. 4b, the hallmark of an interaction is the fact that the mean profiles for one of the variables (x5 here) are not parallel for different levels of the second variable (x4). |

Fig. 5a shows a situation where the variables x4 and x5 do not interact (but both x4 and x5 separately affect the response). In Fig. 5b I deliberately created the response so that an interaction between x4 and x5 is present. The obvious difference in the graphs is that in Fig. 5a the two mean profiles (one for x4 = *A* and the other for x4 = *B*) are essentially parallel while in Fig. 5b the two profiles are clearly not parallel. Because the profiles are parallel in Fig. 5a we can unambiguously talk about an x4 main effect. The vertical distance between the two mean profiles is the difference in the mean response when x4 = *B* versus when x4 = *A*. Because this distance is more or less the same no matter what value we choose for x5, the main effect of x4 is well-defined. We don't need to know the value of x5 to know the effect that x4 has on the response. x4 and x5 do not interact. This is not the case in Fig. 5b. Here the x4 effect changes depending on the value of x5. When x5 = *L* or *H* the effect of switching from x4 = *A* to x4 = *B* is quite large, but when x5 = *M*, the effect of switching from x4 = *A* to x4 = *B* is much smaller. Geometrically, the x4 mean profiles are closer together when x5 = *M* than when x5 = *L* or *H*.

To formally test for the presence of an interaction we can carry out a partial F-test in which we compare the full model, containing the two main effects and a two-factor interaction, to a model that contains only the main effects (and no interaction). Alternatively we can examine the sequential partial F-tests that are returned by the anova function of R.

**#test interaction**

anova(out4)

Analysis of Variance Table

Response: y  
Df Sum Sq Mean Sq F value Pr(>F)   
x4cat 1 524.09 524.09 138.338 < 2.2e-16 \*\*\*  
x5cat 2 455.25 227.62 60.084 < 2.2e-16 \*\*\*  
x4cat:x5cat 2 142.81 71.40 18.847 1.731e-07 \*\*\*  
Residuals 84 318.23 3.79   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The first line is test of the x4 effect. The second line is a test of the x5 effect controlling for the x4 effect. The last line is a test of the interaction of x4 and x5 given that the main effects of x4 and x5 are already in the model. As we can see the reported p-value is quite small indicating the interaction is statistically significant confirming our impressions gleaned from Fig. 5.

Having decided that there is a significant interaction occurring, we can quantify the exact nature of the interaction by looking at the summary output of the model. This output should hopefully confirm the pattern we see in Fig. 5.

**#examine parameter estimates**

summary(out4)

Call:  
lm(formula = y ~ x4cat + x5cat + x4cat:x5cat, data = mydata1)

Residuals:  
Min 1Q Median 3Q Max   
-5.06720 -1.29256 -0.04495 1.54850 3.99223

Coefficients:  
Estimate Std. Error t value Pr(>|t|)   
(Intercept) 9.2878 0.5026 18.481 < 2e-16 \*\*\*  
x4catB 6.0603 0.7107 8.527 5.32e-13 \*\*\*  
x5catM 2.4573 0.7107 3.457 0.000858 \*\*\*  
x5catH 4.2909 0.7107 6.037 4.10e-08 \*\*\*  
x4catB:x5catM -4.7455 1.0051 -4.721 9.27e-06 \*\*\*  
x4catB:x5catH 1.0435 1.0051 1.038 0.302161   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.946 on 84 degrees of freedom  
Multiple R-squared: 0.7791, Adjusted R-squared: 0.7659   
F-statistic: 59.24 on 5 and 84 DF, p-value: < 2.2e-16

The output shows coefficient estimates for the individual dummy variables comprising the interaction term along with tests of their statistical significance. From the reported p-values we conclude that one of these coefficients is significantly different from zero and one is not. To understand what this result means and how it relates to Fig. 5b, we need to write out the equation of the model we've fit. The regression equation shown below lists the terms of the model in the same order shown in the summary table above.

http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture1/interaction.gif

From this equation we can fill in the means for each treatment combination. For the dummy variable interaction terms we just multiply the corresponding values of the individual dummy variables

|  |
| --- |
| **Table 3** Cell means expressed in terms of model parameters |

| **x4** | **x5** | **Z** | **W1** | **W2** | **Z × W1** | **Z × W2** | **μ** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| A | L | 0 | 0 | 0 | 0 | 0 | beta0 |
| B | L | 1 | 0 | 0 | 0 | 0 | BL |
| A | M | 0 | 1 | 0 | 0 | 0 | AM |
| B | M | 1 | 1 | 0 | 1 | 0 | BM |
| A | H | 0 | 0 | 1 | 0 | 0 | AH |
| B | H | 1 | 0 | 1 | 0 | 1 | BH |

In Fig. 6 I repeat the interaction plot of Fig 5b but this time I label each mean using the model parameter relationships shown in Table 3.

|  |
| --- |
| fig 6 |
| **Fig. 6** Mean profile plot illustrating the x4 × x5 interaction. Means are labeled using parameters of the interaction model (Table 3). As the figure shows, the parameters β4 and β5 control how much the mean profiles deviate from being parallel. |

Fig. 6 reveals the meaning of the parameters β4 and β5, the coefficients of the two dummy interaction terms in the model.

* If we have a purely additive model (i.e., one without interaction) then the distance between the *A* mean profile and the *B* mean profile should be β1 for all three levels of x5. In the interaction model when x5 = *L*, the distance between the profiles is β1.
* According to the interaction model, the distance between the x4 mean profiles when x5 = *M* is β1 + β4. If β4 is not significantly different from zero, then the distance between the profiles at x5 = *M* is also approximately β1. So additivity would hold for x5 = *L* and x5 = *M*.
* According to the interaction model, the distance between the x4 mean profiles when x5 = *H* is β1 + β5. If β5 is not significantly different from zero, then the distance between the profiles at x5 = *H* is also approximately β1. Once again additivity holds.

From the summary output we see that β5 is estimated to be 1.04, meaning that the difference in means between x4 = *B* and x4 = *A* is 1.04 units higher when x5 = *H* than is the difference in means when x5 = *L*. But we also see from the output that this difference in mean differences is not deemed to be significantly different from zero (p = 0.30). So we can safely conclude that the distances between the mean profiles at x5 = *L* and x5 = *H* are roughly the same, approximately β1.

On the other hand the other interaction coefficient β4 is reported to be significantly different from zero (p < .001). Its point estimate is negative, beta4hat= –4.75, meaning that the two x4 mean profiles are significantly closer to each other when x5 = *M* than when x5 = *L*. Based on the anova output we conclude there is a significant interaction in this model and based on the summary output we see that this significant interaction is essentially being driven by the category x5 = *M*.

**R Code**

* A compact collection of all the R code displayed in this document appears [here](http://www.unc.edu/courses/2010fall/ecol/563/001/notes/lecture1%20Rcode.txt).
* Figures 3, 4, and 6 were also produced using R. The R code for these figures, albeit with only minimal explanation, appears [here](http://www.unc.edu/courses/2010fall/ecol/563/001/notes/lecture1%20notes%20Rcode.txt).

[Course Home Page](http://www.unc.edu/courses/2010fall/ecol/563/001/index.html)

|  |
| --- |
| Jack Weiss *Phone:* (919) 962-5930 *E-Mail:* jack\_weiss@unc.edu *Address:* Curriculum in Ecology, Box 3275, University of North Carolina, Chapel Hill, 27516 Copyright © 2010 Last Revised--Aug 31, 2010 URL: [http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#lecture1) |