Homework 5

Gather data having at least 30 observations where there is a numerical target variable of interest, and there are two predicting variables that are used to model the target. One of the predictors should be a categorical variable with at least three levels. The other predictor can have any form you want — numerical or categorical (if categorical, with any number of levels you want). The groups defined by the categorical variable should come naturally from the context of the question of interest. The variable can be based on the discretization of some underlying continuous measure if you wish, but only if that discretization is justifiable from the nature of the problem, and not just done arbitrarily for convenience (that is, there’s a good reason to think of using this measure in grouped form, rather than in its inherent continuous form). If you choose to take a continuous numerical predictor and turn it into a categorical one you must provide specific and detailed evidence from an outside source that this is a natural thing to do in the context of your problem, and has been done by others before. If you are doing an analysis based on two categorical predictors (call them rows and columns), it must not be the case that you have one and only observation for each combination of row category and column category.

Analyze the data using appropriate models and methods, and attempting to address model violations using the methods we have discussed in class. Discuss what you find. Even if you believe that there are violations of assumptions in your final model, you should still discuss the implications of that model, while also (of course) noting the potential limitations of those implications.

It is important to not make the data gathering for this homework more difficult than it should be. This is just a regression problem, where (at least) one predictor happens to be categorical. You should be thinking of data possibilities the same way you’ve been thinking all semester, just making sure that at least one predictor is (or can be viewed as) categorical. In particular, you need to be careful dealing with data that come in the form of a table, as such data often are not ANOVA/ANCOVA data (in particular, tables of counts of observations are better analyzed as contingency tables using loglinear models than using ANOVA). In fact, a good bit of advice is that if the data come in the form of a table, don’t use them — they are probably inappropriate for ANOVA modeling.

You should be aware of a potential pitfall when choosing a data set. You should not choose data where the observations represent repeated measurements on the same unit over time, with time then being a potential predictor variable (whether it is as a numerical or categorical variable, or even if it is not included as a predictor at all). An example of this would be a data set where you are examining the proportion of adults who are married for five consecutive years for each of 50 different countries, and you take the continent of the country and the year as your two predicting variables. The problem is that if you knew that a particular country’s marriage rate was higher than expected in one year, chances are that it will also be higher than expected in other years; that is, you can view each country’s data as a short time series, and there is the potential for autocorrelation. This, of course, violates the assumption of independence of errors, and cannot be handled with the time series methods that we have talked about in class (since you have a set of 50 time series here, each of length five, rather than one time series of length 250). This is an example of panel data (also called longitudinal data), and its analysis requires special methodologies.

A similar mistake to avoid is confusing multiple response variables with the existence of a
categorical predictor. Say you’re interested in birth rates for a sample of countries, and you have categorized the countries by continent. If you have birth rates for mothers under the age of 15, ages 15 to 21, and over the age of 21, you might imagine that you could analyze this as a two-way ANOVA, with continent and age level as the two categorical predictors. This is not correct. What you actually have is a multivariate ANOVA (MANOVA) problem; a situation with multiple response variables (in this case three of them: birth rates for mothers under the age of 15, ages 15 to 21, and over the age of 21) and one categorical predictor (continent). The problem with handling this as a two-way ANOVA is that the errors in your model will be correlated (violating the ANOVA assumptions), since if a country has a higher-than-expected birth rate for one age level, they are likely to also have a higher-than-expected rate for other levels. The easiest way to be sure that you have legitimate ANOVA data is to imagine a structure similar to that described in the “Two-way ANOVA” handout: data where each observation falls independently into one (and only one) cell of a cross-classification based on the predictors.

To summarize: there is no difference between an ANOVA or ANCOVA data set and a regression data set, other than (at least) one of the predicting variables is categorical. Your data should take the form of a data matrix with n rows (corresponding to the n ≥ 30 observations in your sample), and three columns, where one of the columns is the numerical response variable and two are the predicting variables, and (at least) one of the predictors is not numerical, but rather categorical with at least three categories. The observations should be a sample from some population – for example, a sample of countries (each row is a different country), a sample of cities (each row is a different city), a sample of companies (each row is a different company), a sample of individual people (each row is a different person). You then have a numerical response variable for that observation – for example, birth rate for the country, murder rate for the city, market capitalization for the company, grade point average for the person. You also have two predicting variables, at least one of which is categorical – for example, continent of the country, geographic region of the city, industry group of the company, highest educational degree of the person. This is exactly the same as any other regression problem, except that a predictor is categorical rather than numerical. If there is something that ties any rows together (they refer to different characteristics or measures of the same country or company, for example), these are not appropriate data for this assignment. Please ask me if a data set is appropriate if you are not sure, as you will be penalized for using an inappropriate one.

A reminder: get your data from an original data source. As was stated in the syllabus, you should not take your data from a textbook, a journal article that includes ANOVA or ANCOVA analysis of the data, an online digest of data sets that have been put together for teaching or expository purposes, or a data analysis competition (see Homeworks 2 and 3 for lists of the kinds of sites this refers to). This also includes digests of data sets from textbooks and articles specifically gathered together to be examples of ANOVA or ANCOVA (that is, don’t try to find data by doing a Google search of “ANOVA data” or use data from a web page with the title “Analysis of covariance data sets”).

Please be sure to include a cover page for your homework that has your name on it. It needn’t have anything else on it, but it should definitely not contain any of the text of your homework.

Due date: May 2