Multiple Regression

The simple linear regression model can be easily extended to include more than one predictor/covariate. This analytic approach inherits all of the issues of simple regression, plus adds a few of its own.

The intuition mostly carries over. Instead of fitting a straight line to data, we are fitting a plane (or hyper-plane). This makes visualization more difficult, and there are various tools to help. This is a rich subject, and we won't do justice to it here. Instead, we'll glance on the major points.

The model is

\[ Y_i = a_0 + a_1 X_{1i} + a_2 X_{2i} + \ldots + a_p X_{pi} + e_i \]

Where the X's represent different covariates. As before, the errors are independent and normally distributed.

The method of Maximum Likelihood produces the same estimators as does Least Squares, and there are no surprises from a simple-regression perspective.

One important new assumption is that the X's must be linearly independent. (No set of observations can be a linear combination of the other.) When this fails, predictors are said to be collinear. For the most part, in this context, the concept of linear independence maps onto that of statistical independence rather well. This means that each variable added to the model must bring entirely new information. For example, if you wished to predict academic success, you can use GPA, and maybe then add socio-economic status, but then to add parents' income might cause problems because you would expect income to predict socio-economic status rather well.

In practice this is a difficult assumption to satisfy. In most problems there is some collinearity, but the trick is to assess how bad it is. Often, nothing needs to be done. If the problem is bad, then there are some less than satisfying remedies available.

What's it Good For?

One possible framework for using multiple-regression is as follows: There's a variable you want to model or predict. Call it Y. You know that Y depends on a set of known factors X1, through Xp, but also there are some unknown or unmeasurable influences. If you knew them all, you would be able to predict Y exactly (you would have a deterministic system), but the fact that you don't means that you observe different values for Y even when the factors X are set to the same values. Multiple regression allows you to quantify the resulting uncertainty in your estimates.

Another use for multiple regression is that frequently we're interested in the relationship between Y and a single predictor X (or a small number of predictors.) However, there
are a variety of other factors that "get in the way" of our studying this relationship. By including these other terms in the model we can "tease out" the relationship between Y and X by "statistically controlling" for the other factors.

*Caution:* Statistical Control is different from experimental control. In an experimental control, the researcher fixes the values of these "nuisance" predictors at a certain value and then observes the relation between X and Y. If one wants to study the relationship between atmospheric pressure and volume of a gas, one holds ambient temperature constant while varying atmospheric pressure and recording volume, for example. In a statistical control, values of x are determined for us by nature. We simply take our gas up and down a mountain and record changes in volume with respect to the recorded changes in pressure, and we record temperature since we know it also affects things. If we are lucky, we have recorded varying pressures at the same temperature, and can then statistically "control" for temperature. If we are not lucky, we can still produce something resembling such a control.

This leads to a familiar warning: correlation is not causation. When one uses statistical controls, one can not conclude with certainty that "x causes y" because one cannot logically rule out the existence of other unrecorded variables that might have played a role. (Technically, a "confounder" is an unrecorded variable that affects both the predictor and the response and therefore could mislead one into believe that x and y.)

**Interpretation**

Once the model is fit, and assuming that the assumptions hold, we need to interpret the output.

Y-hat: the fitted plane gives the mean value y when the x-values are set to a particular value.

Intercept: the mean value of y when all predictors are set to 0.

Slope: there are, of course, several slopes -- one for each predictor. The interpretation is that the slope measures the change in the mean value of y with respect to that predictor while all other predictors are "held" constant.

To visualize this geometrically, imagine just two predictors. We can plot their values on the "floor", and plot the response value using a vertical z-axis. Thus, for any given point on the floor, the response is a point floating somewhere in the room. We fit a plane, that therefore floats across the room at some orientation. The slope with respect to a single predictor is the slope of the line we would get if we walked parallel to the fixed axis. Difficult to describe in words, but a picture will make it clear. (Which I can't do here. Hope you came to class.)

R-squared: the percent of variation in y explained by the entire set of predictors included in the model. Another way of looking at this is a measure that compares two models: the
"full" model that you fit, and the very simple model in which \( Y = a \). In other words, the response is best modelled in terms of its mean and depends on none of the predictors.

In simple regression, R-squared was also the square of the correlation between \( x \) and \( y \). It's no longer that simple. Now it's the correlation between the observed \( y \)'s and the predicted (or fitted) \( y \)'s.

**Inference:**

The first step, usually, is to check to see whether any of the slopes are 0. If you are interested in one in particular, this is usually the one you want to check. The null hypothesis is:

- \( H_0: \) slope = 0 when all other predictors are included in the model.
- \( H_a: \) slope \( \neq \) 0 when all other predictors are included in the model.

Prediction intervals are the same as before, although of course now not just one value, but several, must be provided.

One can also do null hypotheses on subsets of predictors, as well as linear combinations. (The sum of the first three slopes is 0, for example.) But this is a topic for a more in-depth course.

**Helpful Tools**

One of the most helpful is the "pairs" command. It accepts a matrix of variables (in which each column is a variable) or a data object and returns all possible scatterplots. This helps you to see, at a glance, how each predictor relates to \( y \). It also helps you see how the predictors relate to each other.

**Model Building**

This is an issue now. First, it gets complicated deciding which variables are related in which way. Worse, because of (even slight) collinearity, you might make a slight change to one variable and see big changes result in the others! Some guidelines:

a) as much as possible, use any theoretical knowledge to build a functional (linear) relation between the predictors and the response.

b) many people recommend fitting each variable separately (doing a series of simple regressions), and using this to decide about whether or not to transform the predictor or use a quadratic relationship.

c) never remove/add more than one variable at a time.

e) the order in which variables are added or removed often makes a difference. Not much can be done about this, which has not prevented a very large literature from springing up
around the matter. Be guided by these thoughts: you can have different models that fit equally well. Let theory be your guide.

Example

This is the Ozone data discussed in Brieman and Friedman (JASA, 1985, p. 580). These data are for 330 days in 1976. All measurements are in the area of Upland, CA, east of Los Angeles.

Name   Type  n   Info
Height  Variate 330  Vandenberg 500 millibar height (m)
Humidity  Variate 330  humidity, percent
InversionHt  Variate 330  Inversion base height, feet
Ozone  Variate 330  Ozone conc., ppm, at Sandbug AFB.
Pressure  Variate 330  Daggett pressure gradient (mm Hg)
Temp2  Variate 330  inversion base temperature, degrees F.
Temperature  Variate 330  Temperature F. (max?).
Visibility  Variate 330  Visibility (miles)
WindSpeed  Variate 330  wind speed, mph

I'm not sure what all of these mean, and perhaps we can work it out as a class.

The goal of this example is to produce a model that predicts ozone levels.