Noah Tigner

GEOG 495

Quiz 4

For a) regression analysis, b) analysis of variance, c) nonparameteric regression, and d) principal components or factor analysis, describe the particular version of that common conceptual model that applies, and why that conceptual model makes sense given the goals of the analysis.

Regression Analysis	data = predictable component + unpredictable component
Analysis of Variance (ANOVA)	data = common variation + unique variation
Nonparametric Regression	data = smooth function determined by data + noise
Principal Component Analsyis (PCA) / Factor Analysis (FA)	data = common factors + unique factors

a. The aim of regression is to move information into the predictable component while reducing the information/pattern of the unpredictable component (residuals).
b. For ANOVA we essentially compare the variability of observations among groups to the variability within groups.
c. This model makes sense because unlike linear regression analysis, no assumption is made that the relationship between response variable and predictor variable(s) is represented by a straight line. Local relationships are “smoothed”, similar to scatterplot smoothing.
d. PCA looks at the maximum variance and maximum similarities while in FA variables are assembled into unique factors and common factors. PCA can be thought of as a special case of FA, where PCA looks at the covariations among a group of variables.

Describe the general context in which multiple linear regression analysis is applicable. (What is it used for? Are there any assumptions that underlie its use? How is it implemented in practice?)

Multiple Regression is used to estimate the influence of multiple predictor variables on the response. There are four main assumptions associated with its use:

- Predicted errors or residuals are assumed to be independent and normally distributed
- The predictors are known without error
- The predictors are not correlated
- The correct predictors are used in the model

If any of these assumptions are violated, then there may be more bias in the estimated coefficients, and they make have larger variances.

In practice it is implemented by use of matrices and matrix algebra to simplify the calculations involved.

For as many statistics you can think of, characterize in words what the quantities in the numerator and denominator represent, and then comment why that particular arrangement (something in the numerator, something in the denominator) makes sense in general.

Among-Group Variance	$MS_A = SS_A/df_A$	Variance among groups is represented as the sum-of-squares among-groups over the degrees-of-freedom among-groups
Within-Group Variance	$MS_W = SS_W/df_W$	Variance within groups is represented as the sum-of-squares within-group over the degrees-of-freedom within-group
Test Statistic (ANOVA)	$F = MS_A/MS_W$	This is used in ANOVA, which compares the among group's variability relative to the within group's variability
Test Statistic (Difference of means)	$t = (X - \mu)/\sigma_X$ $t = (X_1-X_2)/(\sigma_{X1-X2})$	For the one-sample statistic we have the difference between the sample mean and the hypothesized mean relative to the std error of the mean (below) For the two-sample statistic we have the difference in means between the two samples and the measure of variability of the differences in sample means. This gives us the scaled difference between two means
Standard Error of the Mean	$\sigma_X = \sigma/\text{sqrt}(n)$	The standard deviation over the square root of the population size makes sense because it increases as the variability increases, and decreases as the sample size (n) increases

Suppose you are in charge of the data-analysis component of a project that generates one or more data sets (or data frames), that include the following kinds of variables (i.e. columns):

1. **some kind of text identification label (like the abbreviations of the weather station names in the Oregon climate data set);**
2. **locational information (e.g. latitude and longitude, or x and y);**
3. **one more response variables (i.e. variables that you would like to "explain" or predict);**
4. **several candidate predictor variables;**
5. **one or more factor (or group membership) variables (like the Reach variable in the Summit Cr. data set) that identify which group a particular observation comes from or is assigned to.**

Describe an overall strategy for making sense of this data set. What kind of plots or visualizations might you apply (*and why*)? What kind of analyses (*and why*)?

As an example, I will be using data from the American Community Survey (ACS), a survey conducted by the US Census Bureau. Unlike the census, this data is collected yearly. A common task might be looking at the correlation between proximity to a college or university and the level of school enrollment (table B14007 has observations for the level of school enrollment by location). This data is broken down by the US Census block group, with the block group's idea as the primary key or identifier in the table. Each row then has the block group's name, the associated census tract, county, and state (group membership factors), and fields for the percentage of that block group that are enrolled in each level of education (middle school, high school, undergraduate, graduate or professional school, etc.). The level of education is the response variable, while the block group (and it's location) are the predictor variables. This predictor variable, however, does not come directly with location data. Instead, the table would need to be joined with a table or shapefile containing the shapes and locations of the block groups, by joining on the common field of the block group id's. Once this is done, we could perhaps plot the percentage of those pursuing higher education per block group (perhaps using ggplot). We could obtain coordinate points for colleges and universities and calculate the distance between each block group and one of these locations, likely using Euclidean distance (we would first need to find the centroid of each block group). We could then plot the relationship between percentage enrollment in higher education and proximity to colleges. My hypothesis is that there is a strong positive relationship, with a majority of those living near colleges being enrolled in those colleges. For example, there is likely much higher college enrollment on the edge of the University of Oregon campus and in downtown Eugene compared to in Cottage Grove. I imagine that this relationship is fairly linear (although perhaps less so for "commuter-schools"), so we could easily fit a regression line to this relationship, which would allow us to predict the level of enrollment for other block groups.