This course will involve five homework/project assignments. For three of these five, you will be using data that you have selected.

For the three projects in which you obtain your own data . . .

Do not take data from a textbook. You should obtain your data from original sources. Please provide complete source information for your data. If the data come from a printed source, include a photocopy, as well as the book name and page information. If the data come from the World Wide Web, provide the complete URL and the data of download. Many Web sites are complicated and a reference like “espn.com” is not enough to locate the data.

Data sets should not be taken from university archives or statistical libraries. Data sets should not be taken from a source that provides a similar analysis.

You are encouraged to use data that you find potentially interesting. If you have a culinary enthusiasm for turnips, you might want to seek data relating the weights of turnips to fertilizer and watering practices. If you are a fan of television, you might want to seek data relating the viewer counts for various shows to network, air time, and production budgets.

The “potentially” in the paragraph above is a reminder that not all analyses end up with fascinating conclusions. Your objective here is only performing a competent analysis.

The five projects are based on these analysis techniques:

1. Simple linear regression
2. Multiple linear regression
3. Time series regression
4. Analysis of variance/analysis of covariance
5. Logistic regression

You will supply your own data for 1, 3, and 5.

In all of these, there will be one variable that is plausibly dependent on other variables.

The word “variable” refers to a column of a spreadsheet.
Variables that are obtained by measuring are called \textit{continuous}.

Variables that are obtained by counting are called \textit{discrete}.

Variables that are obtained by classifying are called \textit{categorical}.
   If the categories are ordered, as with bond ratings, the variable is \textit{ordinal}.
   If the categories are not ordered, as with color preferences, the variable is \textit{nominal}.

A categorical variable with two categories is called \textit{binary}. For binary variables, it does not matter whether the categories are ordered or not ordered.

1. For simple linear regression, you will need a spreadsheet with at least two columns. You will try to explain one of the variables in terms of the other. The variable that is “explained by” is called the dependent variable, and the explaining variable is called the independent variable. For example, you might to explain the rates of sales growth for $n$ retail chains in terms of the rates of growth of their numbers of outlets.

   The variables can be discrete or continuous.
   Please avoid discrete random variables with only a few values; “cars in household” would be a poor choices, as the values are almost certainly in set \{0, 1, 2, 3\}.

   The value of $n$ should be at least 50, although we can handle smaller values for certain exotic problems.

   Please do not prune down a data set to get to a smaller $n$. If the data source has $n = 32,810$, you can take all of it.

   Please do not use a time series for this project. Time series will be used in project 3.

2. For multiple linear regression, you will receive a spreadsheet in which one variable is plausibly explained by other variables. The variable that is “explained by” is called the dependent variable, and the explaining variables are called independent variables. Your task will include, among other things, exactly which of the independent variables are really needed.

3. For time series regression, you will need data in which one of the columns is time itself. This column might have values JAN1990, FEB1990, MAR1990, …., DEC2008. The data should have one dependent variable, a column with time, and three other independent variables. Business data is very often provided as time series. The role of time will be critical in this analysis.
4. For analysis of variance/analysis of covariance, you will receive a spreadsheet in which there is
   a non-categorical dependent variable
   a categorical independent variable with three or more levels
   at least one non-categorical independent variable

5. For logistic regression, you will need data in which the dependent variable is binary. This dependent variable could be dead/alive, bankrupt/solvent, or fresh/spoiled. There should be at least three independent variables. The independent variables may be continuous, ordinal categorical, or nominal.
This chart outlines the structures for the projects:

<table>
<thead>
<tr>
<th>Project</th>
<th>Description</th>
<th>Dependent variable</th>
<th>Dependent variable(s)</th>
<th>Time Independent Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Simple linear regression</td>
<td>Continuous or discrete*</td>
<td>One; continuous or discrete*</td>
<td>Not permitted</td>
</tr>
<tr>
<td>2</td>
<td>Multiple linear regression</td>
<td>Continuous or discrete*</td>
<td>Three or more; at least one should be continuous or discrete*</td>
<td>Not permitted</td>
</tr>
<tr>
<td>3</td>
<td>Time series regression</td>
<td>Continuous or discrete*</td>
<td>At least one (other than time); this should be continuous or discrete*</td>
<td>Required</td>
</tr>
<tr>
<td>4</td>
<td>Analysis of variance / covariance</td>
<td>Continuous or discrete*</td>
<td>Two nominal categorical (binary OK) and at least one continuous</td>
<td>Not permitted</td>
</tr>
<tr>
<td>5</td>
<td>Logistic regression</td>
<td>Binary</td>
<td>Three or more</td>
<td>Not permitted</td>
</tr>
</tbody>
</table>

The use of discrete is that of *ordinal* discrete. This is meant to exclude nominal discrete. Variables in the nominal discrete category are color, make of automobile, breed of dog, favorite TV show, and so on.

Nominal discrete variables are also called categorical variables.

The * on discrete is used here to note that the discrete variable has to have many levels. For example, please do not use a variable like “number of family cars” or “number of living parents.” These variables are just not rich enough for our purposes.