Getting the Data Ready

Keep in mind that the linear model

$$Y = X\beta + e$$

represents columns of Y and X_i values, so the raw input data for a linear model usually resembles a spreadsheet.

Getting the Data Ready

Keep in mind that the linear model

$$Y = X\beta + e$$

represents columns of Y and X_i values, so the raw input data for a linear model usually resembles a spreadsheet.

Although the data lends itself to storage in a spreadsheet, there are often problems using a spreadsheet directly as input to a statistical program:

- There are many different spreadsheet programs
- Even if we only consider one "brand" of spreadsheet, there are different versions
- Statistical software packages are often behind the curve in terms of reading the latest versions
- Internal structures of spreadsheets are complicated and/or proprietary

Comma Delimited Files

Spreadsheets, while familiar, tend not to be very "portable" across platforms (Mac vs PC) and time.

A lightweight alternative is a *comma delimted* (.csv) file.

Comma Delimited Files

Spreadsheets, while familiar, tend not to be very "portable" across platforms (Mac vs PC) and time.

A lightweight alternative is a *comma delimted* (.csv) file. Spreadsheets usually offer ".csv" format as a SAVE AS option.

Comma Delimited Files

Spreadsheets, while familiar, tend not to be very "portable" across platforms (Mac vs PC) and time.

A lightweight alternative is a *comma delimted* (.csv) file. Spreadsheets usually offer ".csv" format as a SAVE AS option.

Comma delimited files are inherently simple, just columns of values separated by commas on each row, optionally with the first row containing a column name.

They also have the advantage that the format does not change over time, i.e., there are no "versions"

Comma Delimited files

Most statistical packages have the ability to read comma delimited files directly.

The simplicity and portability of comma delimited files often makes them a good choice.

Comma Delimited files

Most statistical packages have the ability to read comma delimited files directly.

The simplicity and portability of comma delimited files often makes them a good choice.

An alternative that is rapidly gaining popularity is XML.

Comma Delimited files

Most statistical packages have the ability to read comma delimited files directly.

The simplicity and portability of comma delimited files often makes them a good choice.

An alternative that is rapidly gaining popularity is XML.

XML is a "markup" language like HTML, the language of the world wide web.

Like comma delimited files, XML files can be self-describing and very portable. Software to decipher XML files is built in to most packages today, and many software products that collect data store it as XML.

Beginning in 2007, all microsoft office documents (.xlsx, .docx, .pptx) are compressed folders containing XML files.