MORE ON MULTIVARIATE DISTRIBUTIONS

Recall from handout "Independence of More Than Two Random Variables":

With more than two random variables that covary (e.g., the Big Mac data), we have various types of conditional distributions.

Similarly, we have various types of *marginal* distributions:

- Marginal (univariate) distributions of single variables.
- Marginal (bivariate) distributions of two variables at a time.
- Etc.

Example: Big Mac data, with response Big Mac and explanatory variables Bread, TeachSal, TeachTax, and BusFare. There are 30 marginal distributions:

- Marginal (univariate) distributions of each variable separately [5 total]
- Marginal (bivariate) distributions of pairs of variables [5x4/2 = 10 total]
- Marginal (joint) distributions of 3 variables at a time [(5x4x3)/(3x2) = 10 total]
- Marginal (joint) distributions of 4 variables at a time [5 total]

But we can only easily plot and see 1 and 2 variable marginal plots.

Most statistical software allows us to see all of these at one time with a *scatterplot matrix*.

In arc, use the command on the Graph and Fit menu.

Example: Big Mac

Note:

- The order in which variables are entered determines the order in which they appear in the scatterplot matrix.
- The variable on the vertical axis is the row label; the variable on the horizontal axis is the column label.
- Shift-Control-Click (or some variation depending on platform) blows up an individual plot.

Plots of the response vs a single explanatory variable are called *marginal response plots*.

Example: The scatterplot matrix for the Big Mac data shows that some of the marginal response plots do not appear to show a linear mean function.

Question: Does this imply that the mean function for Big Mac conditioned on all four explanatory variables is not linear?

Special Case: *Multivariate normal distribution*.

$f(x_1, x_2, \ldots, x_m) =$

$$\frac{1}{(2\pi)^{m/2}[\det(\Sigma)]^{1/2}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1}(\underline{x}-\underline{\mu})\right],$$

where

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_m \end{bmatrix},$$

$$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_m \end{bmatrix}$$ is the vector of means of the X_i's,

and $\sum$ is an m x m matrix called the *covariance matrix*.

(Superscript T denotes matrix transpose.)

Case m = 2: If $\sum = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$, we get

f(x_1, x_2) =

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2}{2(1-\rho^2)}\right]$$

Note that $\rho\sigma_1\sigma_2$ is the covariance of X_1 and X_2; in the general case, the i,jth entry of the covariance matrix will be the covariance of X_i and X_j.

Properties of multivariate normal distributions:

Recall: If X_1 and X_2 are bivariate normal, then:

- Each X_i is normal
- $E(X_1 | X_2) = a + bX_2$.

These properties generalize:

If $X_1, X_2, \ldots, X_m$ are (jointly) multivariate normal, then:

1. Any subset of these variables is also (multivariate) normal.

2. Each conditional mean obtained by conditioning one variable on a subset of the other variables is a linear function of the remaining variables -- e.g.,

$E(X_1 | X_2, \ldots, X_m) = \alpha_0 + \alpha_2 X_2 + \ldots + \alpha_m X_m$.

Consequences for Regression:

1. If $X_1, X_2, \ldots, X_p$, Y are multivariate normal, then each subset of $X_1, X_2, \ldots, X_p$, Y is also (multivariate) normal.

2. For each subset of $X_1, X_2, \ldots, X_p$, the conditional mean of Y conditioned on those variables is a linear function of those variables. In particular:

- $E(Y| X_1, X_2, \ldots, X_p)$ is a linear function of $X_1, X_2, \ldots, X_p$ (i.e., a linear model fits)
- Even if we drop some predictors, a linear model fits.
- For a single j, $E(Y|x_j) = a + bx_j$.

Practical consequence: If even one marginal response plot clearly indicates that the corresponding mean function is *not* linear, then $X_1, X_2, \ldots, X_p$, Y are *not* multivariate normal.

Caution: The converse is *not* true -- the marginal response plots might all be linear, without having the variables be multivariate normal.

Examples:

1. n = 2: Y = X, X uniform on [0,1]

2. n = 3: Z = Y = X uniform on [0,1]

Nonetheless, it is often <u>useful</u> to have marginals "linear" (i.e., with linear mean) and response Y normal.

- It's a little more reassuring that we might be able to drop predictors and still have a linear model.
- Also, inference will assume conditionals of Y are normal.

Thus, transforming variables more to this state can be helpful. Arc facilitates this by putting transformation slidebars on the scatterplot matrix.