smfl5.tex
Lecture 5. 28.1.2011

5. CONDITIONING AND REGRESSION

Recall that the *conditional* density of Y *given* $X = x$ is

$$f_{Y|X}(y|x) := f_{X,Y}(x,y) / \int f_{X,Y}(x,y)dy.$$

Conditional means.

The conditional mean of Y given $X = x$ is

$$E(Y|X = x),$$

a function of x called the *regression* function (of Y on x). So, if we do not specify the value x, we get $E(Y|X)$. This is *random*, because X is random (until we observe its value, x; then we get the regression function of x as above). As $E(Y|X)$ is random, we can look at its mean and variance.

Recall (SP, Ch. II)

THEOREM (Conditional Mean Formula). $E[E(Y|X)] = EY$.

Interpretation. EY takes the random variable Y, and averages out all the randomness to give a number, EY.
$E(Y|X)$ takes the random variable Y, and averages out all the randomness in Y NOT accounted for by knowledge of X.
$E[E(Y|X)]$ then averages out the remaining randomness, which IS accounted for by knowledge of X, to give EY as above.
Example: Bivariate normal distribution, $N(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2; \rho)$, or $N(\mu, \sigma)$,

$$\mu = (\mu_1, \mu_2)^T, \qquad \sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

Then

$$E(Y|X = x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \qquad \text{so} \qquad E(Y|X) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X - \mu_1).$$

So

$$E[E(Y|X)] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(EX - \mu_1) = \mu_2 = EY, \qquad \text{as} \qquad EX = \mu_1.$$

As with the bivariate normal, we should keep some concrete instance in mind as a motivating example, e.g.:
X = incoming score of student [in medical school or university, say], Y = graduating score;
X = child's height at 2 years (say), Y = child's eventual adult height,
or X = mid-parent height, Y = child's adult height, as in Galton's study.

Recall also (SP, Ch. II)

THEOREM (Conditional Variance Formula).

$$varY = E_X var(Y|X) + var_X E(Y|X).$$

Interpretation.

$$varY = \text{total variability in } Y,$$

$$E_X var(Y|X) = \text{variability in } Y \text{ not accounted for by knowledge of } X,$$

$$var_X E(Y|X) = \text{variability in } Y \text{ accounted for by knowledge of } X.$$

Example: the bivariate normal.

$$Y|X = x \text{ is } N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \sigma_2^2(1-\rho^2)), \qquad varY = \sigma_2^2,$$

$$E(Y|X=x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \qquad E(Y|X) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X-\mu_1),$$

which has variance $(\rho\sigma_2/\sigma_1)^2 varX = (\rho\sigma_2/\sigma_1)^2\sigma_1^2 = \rho^2\sigma_2^2$;

$$var(Y|X=x) = \sigma_2^2(1-\rho^2), \quad E_X var(Y|X) = \sigma_2^2(1-\rho^2).$$

COROLLARY. $E(Y|X)$ has the same mean as Y and smaller variance (if anything) than Y.

Proof. From the Conditional Mean Formula, $E[E(Y|X)] = EY$. Since $var(Y|X) \geq 0$, $E_X var(Y|X) \geq 0$, so

$$varE[Y|X] \leq varY$$

from the Conditional Variance Formula. //

This result has important applications in estimation theory. Suppose we are to estimate a parameter θ, and are considering a statistic X as a possible estimator (or basis for an estimator) of θ. We would naturally want

X to contain all the information on θ contained within the entire sample. What (if anything) does this mean in precise terms? The answer lies in the concept of *sufficiency* ('data reduction') - one of the most important contributions to statistics of the great English statistician R. A. (Sir Ronald) Fisher (1880-1962) in 1920. In the language of sufficiency, the Conditional Variance Formula is seen as (essentially) the Rao-Blackwell Theorem, a key result in the area (see the index in your favourite Statistics book if you want more here).

Regression.

In the bivariate normal, with X = mid-parent height, Y = child's height, $E(Y|X = x)$ is linear in x (*regression line*). In a more detailed analysis, with U = father's height, V = mother's height, Y = child's height, one would expect $E(Y|U = u, V = v)$ to be linear in u and v (*regression plane*), etc.

In an n-variate normal distribution $N_n(\mu, \sigma)$, suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is partitioned into $\mathbf{X}_1 := (X_1, \cdots, X_r)^T$ and $\mathbf{X}_2 := (X_{r+1}, \cdots, X_n)^T$. Let the corresponding partition of the mean vector and the covariance matrix be

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

where $E\mathbf{X}_i = \mu_i$, σ_{11} is the covariance matrix of $\mathbf{X}_1$, σ_{22} that of $\mathbf{X}_2$, $\sigma_{12} = \sigma_{21}^T$ the covariance matrix of $\mathbf{X}_1$ with $\mathbf{X}_2$.

We restrict attention, for simplicity, to the non-singular case, where σ is positive definite.

LEMMA. If σ is positive definite, so is σ_{11}.

Proof. $\mathbf{x}^T\sigma\mathbf{x} > \mathbf{0}$ as σ is positive definite. Take $\mathbf{x} = (\mathbf{x}_1, \mathbf{0})^T$, where $\mathbf{x}_1$ has the same number of components as the order of σ_{11} [i.e., in matrix language, so that the partition of $\mathbf{x}$ is conformable with those of μ and σ above]. Then $\mathbf{x}_1\sigma_{11}\mathbf{x}_1 > 0$ for all $\mathbf{x}_1$. This says that σ_{11} is positive definite, as required. //

THEOREM. The conditional distribution of $\mathbf{X}_2$ given $\mathbf{X}_1 = \mathbf{x}_1$ is

$$\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1 \sim N(\mu_2 + \sigma_{21}\sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1), \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12}).$$

COROLLARY. The regression of $\mathbf{X}_2$ on $\mathbf{X}_1$ is linear:

$$E(\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1) = \mu_2 + \sigma_{21}\sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1).$$

Proof. Recall that $\mathbf{AX}, \mathbf{BX}$ are independent iff $\mathbf{A}\sigma\mathbf{B}^T = \mathbf{0}$, or as σ is symmetric, $\mathbf{B}\sigma\mathbf{A}^T = \mathbf{0}$. Now

$$\mathbf{X}_1 = \mathbf{AX} \text{ where } \mathbf{A} = (\mathbf{I}, \mathbf{0}),$$

$$\mathbf{X}_2 - \sigma_{21}\sigma_{11}^{-1}\mathbf{X}_1 = \begin{pmatrix} -\sigma_{21}\sigma_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \mathbf{BX}, \text{ where } \mathbf{B} = \begin{pmatrix} -\sigma_{21}\sigma_{11}^{-1} & \mathbf{I} \end{pmatrix}.$$

Now

$$\mathbf{B}\sigma\mathbf{A}^T = \begin{pmatrix} -\sigma_{21}\sigma_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} -\sigma_{21}\sigma_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \sigma_{11} \\ \sigma_{21} \end{pmatrix}$$

$$= -\sigma_{21}\sigma_{11}^{-1}\sigma_{11} + \sigma_{21} = \mathbf{0},$$

so $\mathbf{X}_1$ and $\mathbf{X}_2 - \sigma_{21}\sigma_{11}^{-1}\mathbf{X}_1$ are *independent*. Since both are linear transformations of $\mathbf{X}$, which is multinormal, both are *multinormal*. Also,

$$E(\mathbf{BX}) = \mathbf{B}E\mathbf{X} = \begin{pmatrix} -\sigma_{21}\sigma_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \mu_2 - \sigma_{21}\sigma_{11}^{-1}\mu_1.$$

To calculate the covariance matrix, introduce $\mathbf{C} := -\sigma_{21}\sigma_{11}^{-1}$, so $\mathbf{B} = (\mathbf{C}\ \mathbf{I})$, and recall $\sigma_{12}^T = \sigma_{21}$, so $\mathbf{C}^T = -\sigma_{11}^{-1}\sigma_{12}$:

$$var(\mathbf{BX}) = \mathbf{B}\sigma\mathbf{B}^T = \begin{pmatrix} \mathbf{C} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \mathbf{C}^T \\ \mathbf{I} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{C} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \sigma_{11}\mathbf{C}^T + \sigma_{12} \\ \sigma_{21}\mathbf{C}^T + \sigma_{22} \end{pmatrix} = \mathbf{C}\sigma_{11}\mathbf{C}^T + \mathbf{C}\sigma_{12} + \sigma_{21}\mathbf{C}^T + \sigma_{22}$$

$$= \sigma_{21}\sigma_{11}^{-1}\sigma_{11}\sigma_{11}^{-1}\sigma_{12} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12} + \sigma_{22}$$

$$= \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12}.$$

By independence, the conditional distribution of $\mathbf{BX}$ given $\mathbf{X}_1 = \mathbf{AX}$ is the same as its marginal distribution, which by above is $N(\mu_2 - \sigma_{21}\sigma_{11}^{-1}\mu_1, \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12})$. So given $\mathbf{X}_1$, $\mathbf{X}_2 - \sigma_{21}\sigma_{11}^{-1}\mathbf{X}_1$ is $N(\mu_2 - \sigma_{21}\sigma_{11}^{-1}\mu_1, \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12})$.

To pass from the conditional distribution of $\mathbf{X}_2 - \sigma_{21}\sigma_{11}^{-1}\mathbf{X}_1$ given $\mathbf{X}_1$ to that of $\mathbf{X}_2$ given $\mathbf{X}_1$: just add $\sigma_{21}\sigma_{11}^{-1}\mathbf{X}_1$. Then

$$\mathbf{X}_2|\mathbf{X}_1 \sim N(\mu_2 + \sigma_{21}\sigma_{11}^{-1}(\mathbf{X}_1 - \mu_1), \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12}). \quad //$$

(here $\sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12}$ is called the *partial covariance matrix* of $\mathbf{X}_2$ given $\mathbf{X}_1$).