Lecture 8.

The multivariate normal distribution with mean $\mathbf{a} = (a_1, \ldots, a_d)$ and covariance $C = \{C_{i,j}\}$, a symmetric positive definite matrix, is given by the density

$$(\frac{1}{\sqrt{2\pi}})^d \frac{1}{|C|^{\frac{1}{2}}} \exp[-\frac{1}{2} < \mathbf{x} - \mathbf{a}, C^{-1}\mathbf{x} - \mathbf{a} >] d\mathbf{x}$$

where $\mathbf{x} = (x_1, \ldots, x_d)$. By completing the square, it is easy to check that

$$\int_{R^d} (\frac{1}{\sqrt{2\pi}})^d \frac{1}{|C|^{\frac{1}{2}}} \exp[< \theta, \mathbf{x} > -\frac{1}{2} < \mathbf{x} - \mathbf{a}, C^{-1}\mathbf{x} - \mathbf{a} >] d\mathbf{x}$$
$$= \exp[< \mathbf{a}, \theta > +\frac{1}{2} < \theta, C\theta >]$$

When C is degenerate the normal distribution does not have a density, but lives on suitable hyperplane, where it has a density with respect to the Lebesgue measure on that hyperplane. If $\mathbf{y} = T\mathbf{x} + \mathbf{c}$ then

$$E[\mathbf{y}] = \mathbf{b} = T\mathbf{a} + \mathbf{c}$$

and

$$E[(\mathbf{y} - \mathbf{b})(\mathbf{y} - \mathbf{b})'] = TCT'$$

and the distribution is again normal. The sum of two independent normal random variable is again normal with both the mean and the variance adding up. Of special interest is the two dimensional situation. If the means are 0,

$$f(x, y) = \frac{1}{2\pi} \frac{1}{\sigma_1 \sigma_2 \sqrt{(1 - \rho^2)}} \exp\big[-\frac{1}{2(1 - \rho^2)}[\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1 \sigma_2} + \frac{y^2}{\sigma_2^2}]\big]$$

ρ is the correlation coefficient given by $\rho = \frac{E[xy]}{\sigma - 1\sigma_2}$. $-1 \leq \rho \leq 1$.

$$y|x \simeq N\big[\frac{\rho\sigma_2 x}{\sigma_1}, \sigma_2^2(1 - \rho^2)\big]$$

Regression. Fitting a straight line for the data. Given N points $\{(x_i, y_i)\}$ minimize

$$\sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2$$

The minimizers are

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$s_x^2 = \text{var}\,(x) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2 \quad s_y^2 = \text{var}\,(y) = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \bar{y}^2$$

$$\begin{aligned} \text{cov}\,(x, y) &= \frac{1}{N}\sum_{i}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{N}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \\ &= \overline{xy} - \bar{x}\bar{y} \end{aligned}$$

$$\beta = \frac{\text{cov}\,(xy)}{\text{var}\,(x)}$$

$$\alpha = \bar{y} - \beta\bar{x}$$

In terms of the correlation coefficient

$$r = \frac{\text{cov}\ (x, y)}{s_x s_y}$$

β can be expressed as

$$\beta = \frac{r s_y}{r_x}$$

Testing that the correlation is 0 in a bivariate (normal) data.

$$r = \frac{\text{cov}\,(x, y)}{\sqrt{\text{var}\,(x)\,\text{var}\,(y)}}$$

If x and y are independent what is the distribution of r?

There is a theme that comes up often in linear modeling. Let $x_1, x_2, \ldots, x_n$ be n in dependent Gaussian random variables with mean 0 and variance 1. Let

$$\{0\} = V_0 \subset V_1 \subset V_2 \cdots \subset V_k \subset V_{k+1} = R^n$$

be an increasing family of subspaces of R^n. Then $R_n = \oplus_{j=0}^{k} W_j$ where for $j = 0, 1, \ldots, k$

$$W_j = V_j \cap V_{j-1}^{\perp}$$

Then the quadratic form

$$\|x\|^2 = Q(x) = \sum_{i=1}^{n} x_i^2$$

can be written as

$$Q(x) = \sum_{i=0}^{k} Q_i(x) = \sum_{i=0}^{k} \|P_i x\|^2$$

where $\{P_i\}$ are orthogonal projections on to the subspaces $\{W_i\}$. Then $\{Q_i(x)\}$ are mutually independent and for every i, $Q_i(x)$ is distributed as a χ^2 with $d_{i+1} - d_i$ degrees of freedom, where $d_i = \dim V_i$.

We note that r can be represented as $\frac{Z}{\sqrt{Z^2+\chi_{n-2}^2}}$

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\text{var (y)}}} = Z = \sum b_i x_i$$

with $\sum_i b_i = 0$ and $\sum_i b_i^2 = 1$. Treating $\{y_i\}$ as just constants, not all of them equal so that $s_y > 0$, we have three mutually orthogonal subspaces in R^n

$$\begin{aligned}
W_1 &= \{c(1, 1, \ldots, 1)\} \\
W_2 &= \{c(y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y})\} \\
W_3 &= (W_1 \cup W_2)^{\perp}
\end{aligned}$$

of dimensions $1, 1, n-2$ and projections P_1, P_2, P_3 respectively.

$$\begin{aligned}
\sum_{i=1}^{n} x_i^2 &= [\sqrt{n}\bar{x}]^2 + ns^2 \\
&= [\sum_{i=1}^{n} a_i x_i]^2 + [\sum_i b_i x_i]^2 + Q(x) \\
&= Z_1^2 + Z_2^2 + Q(x)
\end{aligned}$$

with $a_i \equiv \frac{1}{\sqrt{n}}$, $b_i = \frac{y_i - \bar{y}}{s_y}$. Z_1 and Z_2 are normal with mean 0 and some variance σ^2 and $Q(x) = ns_x^2 - Z_2^2$ is a $\sigma^2\chi^2_{n-2}$ and they are independent . Finally

$$r = \frac{Z_2}{\sqrt{ns_x^2}}$$

$$\frac{r}{\sqrt{1-r^2}}\sqrt{n-2} = \frac{Z_2}{\sqrt{Q(x)}}\sqrt{n-2} =\simeq t_{n-2}$$

General linear regression models. Suppose we have a variable Y that we want to predict using $\{X_j\}$, $j = 1, 2, \ldots, k$. We make a model of the form

$$Y = a_0 + a_1 X_1 + \cdots + a_k X_k + Z$$

We can dispense with a_0 by adding X_0 which is always 1. We have data $\{y_i; x_{i,1}, x_{i,2}, \ldots, x_{i,k}\}$ for $i = 1, 2, \ldots, N$

$$y_i = \sum_{j=1}^{k} a_j x_{i,j} + z_j$$

z_j are assumed to be independent normal random variables with mean 0 and some variance σ^2. So the general linear model in matrix notation is

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{z}$$

$\{y_i\}$ and $\{x_{i,j}\}$ are given. We need to estimate $\{a_j\}$ and σ^2. We do least square approximation.

$$\sum_{i=1}^{N} [y_i - \sum_{j=1}^{k} a_j x_{i,j}]^2$$

is to be minimized. Or minimize

$$\|\mathbf{y} - \mathbf{X}\mathbf{a}\|^2$$

over choices of $\mathbf{a}$. Leads to

$$\langle \mathbf{y} - \mathbf{X}\mathbf{a}, \mathbf{X}\mathbf{c} \rangle = 0$$

for all $\mathbf{c}$.

$$\mathbf{X}^*\mathbf{X}\mathbf{a} = \mathbf{X}^*\mathbf{y}$$

$$\widehat{\mathbf{a}} = [\mathbf{X}^*\mathbf{X}]^{-1}\mathbf{X}^*\mathbf{y}$$

The residual sum of squares is

$$\|\mathbf{y} - \mathbf{X}\widehat{\mathbf{a}}\|^2 = \langle \mathbf{y}, \mathbf{y} \rangle - 2\langle \mathbf{y}, \mathbf{X}\widehat{\mathbf{a}} \rangle + \langle \mathbf{X}\widehat{\mathbf{a}}, \mathbf{X}\widehat{\mathbf{a}} \rangle$$

$$\langle \mathbf{y}, \mathbf{X}\widehat{\mathbf{a}} \rangle = \langle \mathbf{X}^*\mathbf{y}, [\mathbf{X}^*\mathbf{X}]^{-1}\mathbf{X}^*\mathbf{y} \rangle$$

$$\langle \mathbf{X}\widehat{\mathbf{a}}, \mathbf{X}\widehat{\mathbf{a}} \rangle = \langle \mathbf{X}[\mathbf{X}^*\mathbf{X}]^{-1}\mathbf{X}^*\mathbf{y}, \mathbf{X}[\mathbf{X}^*\mathbf{X}]^{-1}\mathbf{X}^*\mathbf{y} \rangle = \langle \mathbf{X}^*\mathbf{y}, [\mathbf{X}^*\mathbf{X}]^{-1}\mathbf{X}^*\mathbf{y} \rangle$$

$$\|\mathbf{y} - \mathbf{X}\widehat{\mathbf{a}}\|^2 = \langle \mathbf{y}, \mathbf{y} \rangle - \langle \mathbf{X}^*\mathbf{y}, [\mathbf{X}^*\mathbf{X}]^{-1}\mathbf{X}^*\mathbf{y} \rangle$$

$$\frac{1}{n-k}\|\mathbf{y} - \mathbf{X}\widehat{\mathbf{a}}\|^2 = \widehat{\sigma}^2$$

F Test. In general in order test if a particular linear model is valid, we need to compare the error we get from fitting the model to the intrinsic error or noise level σ^2. If σ^2 is large it is impossible to say anything because the data is corrupted by very high noise. If the model is correctly specified with k linearly independent parameters $a_1, a_2, \ldots, a_k$, then we can fit the model and the residual sum of squares $\|\mathbf{x} - \mathbf{X}\widehat{\mathbf{a}}\|^2$ provides an estimate $\widehat{\sigma}^2$ of σ^2. If we want to test whether $\mathbf{a} = (a_1, a_2, \cdots, a_k)$ is from a subspace $S \subset R^k$ of dimension d can be tested by examining

$$R_1(x) = \inf_{\mathbf{a}\in S} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|^2 \geq \inf_{\mathbf{a}\in R^k} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|^2 = R(x)$$

If

$$R_1(x) - R(x) = Q_1(x)$$

Then Q_1 and R_1 are independent χ^2 and

$$\frac{\frac{Q_1(x)}{k-d}}{\frac{R(x)}{n-k}} \simeq F_{k-d,n-k}$$

Example 1. Testing all means are 0. We have a collection of Normal variables $\{x_{i,j}\}$, $1 \leq j \leq n_i$ and for $i = 1, 2, \ldots, k$ they have $E[x_{i,j}] = \mu_i$ and $\operatorname{var}(x_{i,j}) = \sigma^2$ for all i and j. Then

$$X_{n\times k} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

$n = n_1 + \cdots + n_k$, $x_i = \mu_i + z_i$. $\mu_1 = \cdots = \mu_{n_1} = a_1, \mu_{n_1+1} = \cdots = \mu_{n_1+n_2} = a_2$ etc. The rows come in blocks of sizes $n_1, n_2, \ldots, n_k$. If we minimize over H

$$\inf_{a_1,\ldots,a_k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - a_i)^2 = \sum_i n_i s_i^2$$

where for each i, s_i^2 is the variance of n_i variables $\{x_{i,j}\}$. If all the a_i are 0 the infimum is just $\sum_{i,j} x_{i,j}^2 = Q(x)$.

$$Q(x) = \sum_i n_i s_i^2 + R(x)$$

and

$$\frac{\frac{1}{k}\sum_i n_i s_i^2}{\frac{1}{n-k} R(x)} \simeq F_{k,n-k}$$

Example 2. Testing all means are equal. $\mu_j = \mu + a_j$, $\sum_j a_j = 0$.

$$X_{n\times k} = \begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ 1 & 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & 0 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \end{pmatrix}$$

$$\sum_i n_i \bar{x}_i^2 = n\bar{x}^2 + \sum_i n_i (\bar{x}_i - \bar{x})^2$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{x}_i$

$$\frac{n\bar{x}^2}{\frac{1}{n-k} R(x)} \simeq F_{1,n-k}$$

or

$$\frac{\bar{x}}{\sqrt{\frac{R(x)}{n}}} \sqrt{n-k} \simeq t_{n-k}$$