Random vectors and the multivariate normal distribution,

conditional expectation

1.1. Random vectors

For the purposes of statistics we will understand random vectors as columns, or $n \times 1$ matrices. We will write

$$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

Definition: The expected value of a random vector is the vector of expected values of components. In symbols

$$E(\underline{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}$$

Definition: An $m \times n$ array of random variables $X_{i,j}$ is called a random matrix.

Similarly as before we define the expectation of a random matrix as the matrix of expected values. In symbols

$$E(\underline{X}) = \begin{pmatrix} E(X_{11}) & \dots & E(X_{1n}) \\ \vdots & & \\ E(X_{m1}) & \dots & E(X_{mn}) \end{pmatrix}$$

Theorem 1.1: Let $\underline{A}, \underline{B}$ be fixed matrices and $\underline{X}$ a random matrix. We have

$$E(\underline{A}\,\underline{X}\,\underline{B}) = \underline{A}\; E(\underline{X}) \cdot \underline{B}$$

Comment: we assume $\underline{A}, \underline{B}$ are of right dimension.

Proof: We have

$$(\underline{A}\,\underline{X}\,\underline{B})_{i,j} = \sum_{k,l} a_{ik} \cdot X_{k,l}\, b_{lj}.$$

The proof follows from linearity of expected value.

We need to extend the notion of variance and covariance to random vectors. In one dimension we have

$$\operatorname{cov}(X, Y) = E(X \cdot Y) - E(X) \cdot E(Y).$$

The analogy for vectors would be

$$\operatorname{cov}(\underline{X}, \underline{Y}) = E(\underline{X} \cdot \underline{Y}^T) - E(\underline{X}) \cdot E(\underline{Y})^T$$

Comment: obviously

$$E(\underline{Y}^T) = E(\underline{Y})^T.$$

If we write it componentwise and say $\underline{c} = \text{cov}(\underline{X}, \underline{Y})$ we get

$$\underline{c}_{ij} = E(X_i \cdot Y_j) - E(X_i)E(Y_j)$$

$$= \text{cov}(X_i, Y_j)$$

The covariance of $\underline{X}, \underline{Y}$ is an array or matrix of covariances between components of $\underline{X}$ and $\underline{Y}$.

$$\text{cov}(\underline{X}, \underline{Y}) = \begin{pmatrix} \text{cov}(X_1, Y_1), \ldots, \text{cov}(X_1, Y_n) \\ \vdots \\ \text{cov}(X_m, Y_1) \ldots \text{cov}(X_m, Y_n) \end{pmatrix}$$

Why is this definition good?
It makes many computations elegant.

Theorem 1.2: Let $\underline{A}, \underline{B}$ be fixed matrices. We have

$$\boxed{\operatorname{cov}(\underline{A}\underline{X}, \underline{B}\underline{Y}) = \underline{A} \operatorname{cov}(\underline{X}, \underline{Y}) \underline{B}^T}$$

Proof: We compute

$$E\left((\underline{A}\underline{X})(\underline{B}\underline{Y})^T\right)$$

$$= E\left(\underline{A}\,\underline{X} \cdot \underline{Y}^T \underline{B}^T\right)$$

$$= \underline{A}\, E(\underline{X}\,\underline{Y}^T)\, \underline{B}^T$$

and we know

$$E(\underline{A}\underline{X}) = \underline{A} \cdot E(\underline{X})$$

$$E(\underline{Y}^T \cdot \underline{B}^T) = E(\underline{Y}^T)\, \underline{B}^T$$

We need to define $\operatorname{var}(\underline{X})$. By analogy we have

$$\operatorname{var}(\underline{X}) = \operatorname{cov}(\underline{X}, \underline{X}).$$

For variances Theorem 1.2 becomes

$$\mathrm{var}(\underline{A}\,\underline{X}) = \underline{A}\;\mathrm{var}(\underline{X})\,\underline{B}^T$$

<u>Example</u>: Let $\underline{X}^T = (X_1, X_2, \ldots, X_n)$ where all X_k are independent and $X_k \sim N(\mu, \sigma^2)$ for $k = 1, 2, \ldots, n$. What is the variance of

$$\underline{Y} = \begin{pmatrix} X_1 - \overline{X} \\ X_2 - \overline{X} \\ \vdots \\ X_n - \overline{X} \end{pmatrix}.$$

Let $\underline{H} = \underline{\underline{I}} - \frac{1}{n}\,\underline{1}\cdot\underline{1}^T$ where

$$\underline{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$ We have

$$\underline{Y} = \underline{H}\cdot\underline{X}$$ and

$$\underline{H}^T = \underline{H},$$ and.

$$\underline{H}^2 = \left(\underline{I} - \tfrac{1}{n}\cdot\underline{1}\cdot\underline{1}^T\right)\left(\underline{I} - \tfrac{1}{n}\underline{1}\cdot\underline{1}^T\right)$$

$$= \underline{I} - \tfrac{2}{n}\underline{1}\cdot\underline{1}^T + \tfrac{1}{n^2}\underline{1}\underbrace{\underline{1}^T\underline{1}}_{=n}\cdot\underline{1}^T$$

$$= \underline{I} - \tfrac{1}{n}\underline{1}\cdot\underline{1}^T$$

$$= \underline{H}.$$

It follows that

$$\operatorname{var}(\underline{Y}) = \operatorname{var}(\underline{H}\cdot\underline{X})$$

$$= \underline{H}\cdot\operatorname{var}(\underline{X})\cdot\underline{H}^T$$

$$= \underline{H}\cdot\sigma^2\underline{I}\cdot\underline{H}^T$$

$$= \sigma^2\,\underline{H}\cdot\underline{H}^T$$

$$= \sigma^2\cdot\underline{H}^2$$

$$= \sigma^2\left(\underline{I} - \tfrac{1}{n}\underline{1}\cdot\underline{1}^T\right).$$

Theorem 1.3: If $\underline{X}$ and $\underline{Y}$ are random vectors of the same dimension then

$$\operatorname{var}(\underline{X}+\underline{Y}) = \operatorname{var}(\underline{X}) + \operatorname{var}(\underline{Y}) + \operatorname{cov}(\underline{X},\underline{Y}) + \operatorname{cov}(\underline{Y},\underline{X})$$

Proof: This follows from definitions:

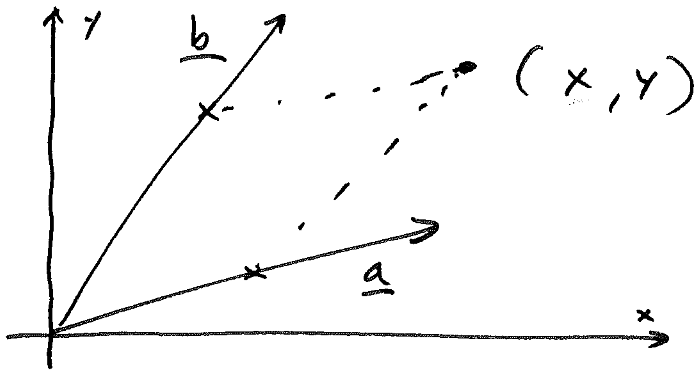
$$(\underline{X}+\underline{Y})(\underline{X}+\underline{Y})^T = \underline{X}\cdot\underline{X}^T + \underline{X}\cdot\underline{Y}^T + \underline{Y}\cdot\underline{X}^T + \underline{Y}\cdot\underline{Y}^T$$

Take expectations and collect terms.

1.2. Multivariate normal distribution

Idea: We want to create a random point in the plane or a random point in space. Take two vectors $\underline{a}$ and $\underline{b}$ and create a linear combination with random components

Figure:



We can take only one vector as well but then all the points will be on a line.

One possibility is to take coefficients as independent random variables. We can specialize and say the coefficients $z_1, z_2, \ldots$ are independent normal random variables. Denote $\underline{z}^T = (z_1, \ldots, z_p)$. We can create a point in $\mathbb{R}^n$ by taking p vectors $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_p$ and say

$$\underline{x} = z_1 \cdot \underline{a}_1 + \cdots + z_p \underline{a}_p + \underline{\mu}.$$

If we collect the vectors in a matrix

$$\underline{A} = \left(\underline{a}_1; \underline{a}_2, \ldots, \underline{a}_p \right)$$

in matrix notation we have

$$\underline{X} = \underline{A} \cdot \underline{Z} + \underline{\mu}$$

<u>Comment</u>: If $p < n$ the random points created will be in a hyperplane.

From 1.1 we have that

$$E(\underline{X}) = \underline{A} \cdot E(\underline{Z}) + \underline{\mu} = \mu$$

and

$$\begin{aligned} \operatorname{var}(\underline{X}) &= \underline{A} \cdot \operatorname{var}(\underline{Z}) \cdot A^T \\ &= \underline{A} \cdot \underline{I} \cdot \underline{A}^T \\ &= \underline{A} \cdot \underline{A}^T. \end{aligned}$$

We will say later that $\underline{X}$ has multivariate normal distribution with parameters $\underline{\mu}$ and $\underline{\Sigma} = \underline{A}\,\underline{A}^T$. But we must answer another question first. Suppose

$$\underline{Y} = \underline{B} \cdot \underline{Z} + \underline{\mu}$$

and $\underline{A}\,\underline{A}^T = \underline{B}\,\underline{B}^T$. Do $\underline{X}$ and $\underline{Y}$ have the same distribution?

This means $P(\underline{X} \in \mathcal{U}) = P(\underline{Y} \in \underline{\mathcal{U}})$ for all reasonable sets $\mathcal{U}$.

The idea of the proof is the following: we will show that

$$\underline{X} = \underline{M}\,\underline{Z} + \underline{\mu} \quad \text{and} \quad \underline{Y} = \underline{M}\,\underline{\tilde{Z}} + \underline{\mu}$$

where $\underline{z}$ and $\tilde{\underline{z}}$ will have the same distribution. Once we have that we will know that $\underline{A}\underline{A}^T$ and $\underline{\mu}$ uniquely determine the distribution of $\underline{X}$.

We need the singular value decomposition for matrices:

if $\underline{A}$ is a $p \times q$ matrix there are orthogonal matrices $\underline{S}$ $(p \times p)$ and T $(q \times q)$ such that

$$\underline{A} = \underline{S} \cdot \underline{D} \cdot T^T$$

where

(i) $\underline{D}$ is of the form

$$\underline{D} = \begin{pmatrix} \sqrt{\lambda_1} & & & 0 \\ 0 & \ddots & \sqrt{\lambda_r} & \\ & & & 0 \end{pmatrix}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_r$ are non-zero eigenvalues of matrices $\underline{A}\underline{A}^T$ or $\underline{A}^T\underline{A}$ (they are the same).

(ii) The columns of $\underline{S}$ are orthogonal eigenvectors of $\underline{A}\underline{A}^T$, the first r belonging to eigenvalues $\lambda_1, \ldots, \lambda_r$

(iii) The columns of $\underline{T}$ are eigenvectors of $\underline{A}^T\underline{A}$; the first r belong to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_r$.

If $\underline{A}\,\underline{A}^T = \underline{B}\,\underline{B}^T$ this means that we can write

$$\underline{A} = \underline{S} \cdot \underline{D}_1 \cdot \underline{T}_1^T$$

$$\underline{B} = \underline{S} \cdot \underline{D}_2 \cdot \underline{T}_2^T$$

Fact from probability: if $\underline{X}$ has density $f_{\underline{X}}(\underline{x})$ and $\underline{A}$ is an invertible matrix the vector

$$\underline{Y} = \underline{A} \cdot \underline{X} + \underline{\nu}$$ has

density

$$f_{\underline{Y}}(\underline{y}) = \frac{1}{|\det(A)|} \times$$

$$\times f_{\underline{X}}\left(\underline{A}^{-1}(\underline{y} - \underline{\nu})\right).$$

If $\underline{z}^T = (z_1, z_2, \dots, z_n)$ has independent normal components the density is the product i.e.

$$f_{\underline{z}}(\underline{z}) = \prod_{k=1}^{n} f_{z_k}(z_k)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot e^{-\frac{z_1^2 + \quad + z_n^2}{2}}$$

$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{\underline{z}^T \cdot \underline{z}}{2}}$$

If $\underline{A}$ is invertible the density of $\underline{w} = \underline{A}\,\underline{z} + \underline{\nu}$ is

$$f_{\underline{w}}(\underline{w}) = \frac{1}{(2\pi)^{n/2} |\det A|} \times$$

$$\times\ e^{-\frac{1}{2}[\underline{A}^{-1}(w - \underline{\nu})]^T [\underline{A}^{-1}(w - \underline{\nu})]}$$

$$= (*)$$

$$(*) = \frac{1}{(2\pi)^{n/2}\,|\det(\underline{A})|} \times e^{-\frac{1}{2}(\underline{w}-\underline{\nu})^T(\underline{A}^{-1})^T A^{-1}(\underline{w}-\underline{\nu})}$$

$$= \frac{1}{(2\pi)^{n/2}\,|\det(\underline{A})|} \times e^{-\frac{1}{2}(\underline{w}-\underline{\nu})^T(\underline{A}\,\underline{A}^T)^{-1}(\underline{w}-\underline{\nu})}$$

This means that

$\underline{T}_1^T \underline{z}_1 = \underline{w}_1$ has independent normal $N(0,1)$ components because $\underline{\nu} = 0$ and $\underline{T}_1^T \underline{T}_1 = \underline{I}$. The same is true for $\underline{w}_2 = \underline{T}_2^T \cdot \underline{z}_2$.

Putting these facts together gives

$$\underline{X} = \underline{S} \cdot \underline{D}_1 \underline{W}_1 + \underline{\mu}$$

$$\underline{Y} = \underline{S} \cdot \underline{D}_2 \underline{W}_2 + \underline{\mu}$$

But $\underline{W}_1$ and $\underline{W}_2$ have the $N(0,I)$ density. This implies $\underline{X}$ and $\underline{Y}$ have the same distribution. because the components are the same linear combinations of indep. normals.

Theorem 1.4 : Let $\underline{Z}^T = (Z_1, \dots, Z_n)$ have independent standard normal distribution and let $\underline{A}$ be a matrix and $\underline{\mu} \in R^m$ a vector. The distribution of

$$\underline{X} = \underline{A}\,\underline{Z} + \underline{\mu}$$

is uniquely determined by $\underline{\mu}$ and $\underline{\Sigma} = \underline{A}\,\underline{A}^T$.

Proof: Done already.

This justifies the following definition:

Definition: Let $\underline{z}^T = (z_1, .., z_n)$ be a vector of independent standard normal variables. The distribution of the vector

$$\underline{X} = \underline{A}\,\underline{z} + \underline{\mu}$$

is called multivariate normal with parameters $\underline{\mu}$ and $\underline{\Sigma} = \underline{A}\,\underline{A}^T$. Symbol: $\underline{X} \sim N(\underline{\mu}, \underline{\Sigma})$

Comments:

(i) just as in the one-dimensional case we have

$$\underline{\mu} = E(\underline{X}) \quad \text{and} \quad \underline{\Sigma} = var(\underline{X}).$$

(ii) If $\underline{A}$ is invertible $\underline{X}$ has a density

<u>Theorem A.5</u>: If $\underline{X} \sim N(\underline{\mu}, \underline{\Sigma})$ then:

(i) all marginal distributions are normal.

(ii) If $\underline{Y} = \underline{B}\,\underline{X} + \underline{\nu}$ then $\underline{Y}$ is multivariate normal with parameter $\underline{\Sigma}' = \underline{B}\,\underline{\Sigma}\,\underline{B}^T$ and $\underline{\mu}' = \underline{B}\,\underline{\mu} + \underline{\nu}$.

<u>Proof</u>: Follows directly from definitions.

Suppose $\underline{A}$ is invertible

and

$$\underline{A}\,\underline{A}^T = \begin{pmatrix} \underline{\Sigma}_{11} & 0 \\ 0 & \underline{\Sigma}_{22} \end{pmatrix} \begin{matrix} \}\, p \\ \}\, q \end{matrix}$$

We assume $\underline{\Sigma}_{11}$ and $\underline{\Sigma}_{22}$ are invertible and square. If we write $\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix} \begin{matrix} \}\, p \\ \}\, q \end{matrix}$

then the density has the form

$$f_{\underline{X}}(\underline{x}) = f_{\underline{X}_1}(\underline{x}_1) \cdot f_{\underline{X}_2}(\underline{x}_2).$$

This means that $\underline{X}_1$ and $\underline{X}_2$ are independent. But we have

$$\mathrm{cov}(\underline{X}_1, \underline{X}_2) = 0.$$

In this particular case

we have that uncorrelated vectors are independent. This is not generally true.

<u>Theorem 1.6</u>: Let $\begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix} \sim N(\underline{\mu}, \underline{\Sigma})$.
If $\mathrm{cov}(\underline{X}_1, \underline{X}_2) = 0$ then $\underline{X}_1$ and $\underline{X}_2$ are independent.

<u>Proof</u>: By Cholesky we can find a matrix $\underline{A}$ such that $\underline{\Sigma} = \underline{A} \cdot \underline{A}^T$. Write

$$\underline{X} = \underline{A} \cdot \underline{Z} + \underline{\mu}$$

$$= \underline{S}\, \underline{D} \cdot \underline{T}^T \cdot \underline{Z} + \underline{\mu} \quad \text{by SVD.}$$

$$= \underline{S} \cdot \underline{D} \cdot \widehat{\underline{Z}} + \underline{\mu}$$

But linear algebra gives us that $\underline{S}$ must be of the form $\underline{S} = \begin{pmatrix} \underline{S}_{11} & 0 \\ 0 & \underline{S}_{22} \end{pmatrix} \begin{matrix} \}p \\ \}q \end{matrix}$

This means that $\underline{X}_1$ depends on the first p components of $\underline{\tilde{Z}}$ and $\underline{X}_2$ on the last q components of $\underline{\tilde{Z}}$. This implies independence.

<u>Example</u>: Let $\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix} \sim N(\underline{\mu}, \underline{\Sigma})$ with $\underline{\mu} = \begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}$ and $\underline{\Sigma} = \begin{pmatrix} \underline{\Sigma}_{11} & \underline{\Sigma}_{12} \\ \underline{\Sigma}_{21} & \underline{\Sigma}_{22} \end{pmatrix}$.

Assume that $\underline{\Sigma}_{11}$ is invertible.

Is there a matrix $\underline{C}$ such that

$\underline{X}_1$ and $\underline{X}_2 - \underline{C}\underline{X}_1$ are independent?

We just need to compute

$\mathrm{cov}(\underline{X}_1, \underline{X}_2 - \underline{C}\underline{X}_1)$ because we know that $\begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 - \underline{C}\underline{X}_1 \end{pmatrix}$

is multivariate normal.

We have

$$\operatorname{cov}(\underline{X}_1, \underline{X}_2 - \underline{C}\underline{X}_1)$$

$$= \operatorname{cov}(\underline{X}_1, \underline{X}_2) - \operatorname{cov}(\underline{X}_1, \underline{C}\underline{X}_1)$$

$$= \underline{\Sigma}_{12} - \underline{\Sigma}_{11} \cdot \underline{C}^T$$

If we want this to be 0

we need

$$\underline{C}^T = \underline{\Sigma}_{11}^{-1} \underline{\Sigma}_{12} \quad \text{or}$$

$$\underline{C} = \underline{\Sigma}_{21} \underline{\Sigma}_{11}^{-1}.$$

1.3. Quadratic forms

Let us recall a few classical definitions.

(i) Let $Z_1, \ldots, Z_r$ be independent standard normal.

Let

$$U = Z_1^2 + Z_2^2 + \cdots + Z_r^2.$$

We say that U has the χ^2-distribution with r degrees of freedom.

Symbol: $U \sim \chi^2(r)$.

<u>Comment</u>: We know that

$$\chi^2(r) = \Gamma\left(\tfrac{r}{2}, \tfrac{1}{2}\right).$$

(ii) Let $Z \sim N(0,1)$ independent of $U \sim \chi^2(r)$. The random variable

$$T = \frac{Z}{\sqrt{U/r}}$$

has the t-distribution with

r degrees of freedom.

Symbol: $T \sim t_r$

(iii) Let $U \sim \chi^2(m)$ independent of $V \sim \chi^2(n)$. Then

$$F = \frac{U/m}{V/n}$$

has the F-distribution with m, n degrees of freedom.

Symbol: $F \sim F_{m,n}$.

In linear algebra a quadratic form is the expression

$$Q = \sum_{k,l=1}^{n} a_{kl} \, x_k x_l .$$

If we collect the constants $a_{k,l}$ into a matrix $\underline{A}$ and define $\underline{x}^T = (x_1, x_2, .., x_n)$ we can write

$$Q = \underline{a}^T \underline{A} \, \underline{a}$$

Without loss of generality $\underline{A}$ can be assumed to be symmetric.

In statistics $\underline{x}$ will be replaced by a random vector $\underline{X}$ so that Q will be a random variable. Such expressions arise frequently. We will be interested in such random variables and their distributions.

We will need some facts from linear algebra.

<u>Definition</u>: A symmetric matrix $\underline{H}$ is called idempotent if

$$\underline{H}^2 = \underline{H}.$$

If $\underline{x}$ is an eigenvector of $\underline{H}$ we have

$$\underline{H}^2 \underline{x} = \underline{H} \cdot (\underline{H} \underline{x})$$
$$= \underline{H} \cdot \lambda \underline{x}$$
$$= \lambda^2 \cdot \underline{x}$$
$$= \lambda \underline{x}$$

If $\underline{x} \neq 0$ this means $\lambda^2 = \lambda$. The eigenvalues of $\underline{H}$ can only be in $\{0, 1\}$. Because every symmetric matrix can be diagonalized we have

$$\underline{H} = \underline{Q} \operatorname{diag}\{1, \dots, 1, 0, \dots, 0\} \underline{Q}^T$$

for an orthogonal matrix. It is clear that the rank of $\underline{H}$ is equal to the number of non-zero eigenvalues.

Definition: If $\underline{A}$ is a square $(n \times n)$ matrix its trace is

$$\mathrm{Tr}(\underline{A}) = \sum_{k=1}^{n} a_{kk}.$$

From linear algebra we will borrow the fact that for matrices $A\ (p \times q)$ and $B\ (q \times p)$ we have

$$\mathrm{Tr}(\underline{A} \cdot \underline{B}) = \mathrm{Tr}(\underline{B} \cdot \underline{A}).$$

From this we have for idempotent $\underline{H}$:

$$\begin{aligned} \mathrm{Tr}(\underline{H}) &= \mathrm{Tr}\left(\underline{Q}\, \mathrm{diag}(1, \ldots, 1, 0, 0)\, \underline{Q}^T\right) \\ &= \mathrm{Tr}\Big(\mathrm{diag}(1, \ldots, 1, 0 \ldots 0)\, \underbrace{\underline{Q}^T \underline{Q}}_{= I}\Big) \\ &= \mathrm{rank}(\underline{H}). \end{aligned}$$

Let $\underline{z}$ be a vector with independent standard normal components. In the notation of the multivariate normal random vectors we have $\underline{z} \sim N(\underline{0}, \underline{I})$. Let $\underline{H}$ be idempotent. Define

$$U = \underline{z}^T \underline{H} \underline{z}$$

Rewrite

$$\underline{U} = \underline{z}^T \underline{Q}^T \operatorname{diag}\{1, 1, \ldots, 1, 0, 0\} \underline{Q} \underline{z}.$$

We know that $\tilde{\underline{z}} = \underline{Q} \underline{z} \sim N(\underline{0}, \underline{Q} \underline{I} \underline{Q}^T)$.
But $\underline{Q} \underline{I} \underline{Q}^T = \underline{I}$ so $\tilde{\underline{z}} \sim N(0, \underline{I})$.
It follows

$$U = \tilde{z}_1^2 + \cdots + \tilde{z}_r^2$$

where $r = \operatorname{rank}(\underline{H})$.

The result is still true if $\underline{z} \sim N(\underline{\mu}, \underline{I})$ and $\underline{H}\,\underline{\mu} = 0$.

We simply rewrite

$$u = \underline{z}^T \underline{H}\, \underline{z}$$

$$= (\underline{z} - \underline{\mu})^T \cdot \underline{H}\, \underbrace{(\underline{z} - \underline{\mu})}_{N(\underline{0}, I)}.$$

We can take this further.

<u>Theorem 1.7</u> (Cochran) Let $\underline{H}_1, \dots, \underline{H}_s$ be idempotent matrices such that $\underline{H}_k \cdot \underline{H}_\ell = 0$ for $k \neq \ell$. Then the random variables

$u_k = \underline{z}^T \underline{H}_k \underline{z}$ are independent

and $u_k \sim \chi^2(\mathrm{rank}(\underline{H}_k))$.

<u>Proof</u>: The vector

$$\begin{pmatrix} \underline{H}_1 \underline{z} \\ \underline{H}_2 \underline{z} \\ \\ H_s z \end{pmatrix}$$

is multivariate normal because its components are linear combinations of a multivariate normal vector.

But

$$\begin{aligned} \operatorname{cov}(\underline{H}_k \underline{z}, \underline{H}_\ell \underline{z}) &= \underline{H}_k \cdot H_\ell^T \\ &= \underline{H}_k \cdot \underline{H}_\ell \\ &= 0 \end{aligned}$$

by assumption. So all $\underline{H}_1 \underline{z}, \ldots, \underline{H}_s \underline{z}$ are independent.

But

$$\begin{aligned} u_k &= \underline{z}^T \underline{H}_k \underline{z} \\ &= \underline{z}^T \underline{H}_k^T \cdot \underline{H}_k z \\ &= (\underline{H}_k \underline{z})^T (\underline{H}_k \cdot \underline{z}) \end{aligned}$$

This implies that all U_k are independent. The fact that $U_k \sim \chi^2(\text{rank}(\underline{H}_k))$ we have already proved.

<u>Example</u>: If $\underline{H}$ is idempotent so is $\underline{I} - \underline{H}$. But

$$\underline{H}(\underline{I} - \underline{H}) = \underline{H} - \underline{H}^2$$
$$= \underline{H} - \underline{H}$$
$$= 0.$$

So for $\underline{z} \sim N(0, \underline{I})$ we have that $U = \underline{z}^T \underline{H}\, \underline{z}$ and $V = \underline{z}^T(\underline{I} - \underline{H})\underline{z}$ are independent and χ^2-distributed.

<u>Comment</u>: The independence assertion of Theorem 1.7 is true even if $\underline{z} \sim N(\underline{\mu}, \underline{I})$.

Example: In regression we assume

$$\underline{Y} = \underline{X} \cdot \underline{\beta} + \underline{\varepsilon}$$

with $E(\underline{\varepsilon}) = 0$ and $\mathrm{Var}(\underline{\varepsilon}) = \sigma^2 \cdot \underline{I}$.

Assume further that $\underline{\varepsilon} \sim N(0, \sigma^2 \cdot \underline{I})$.

The best unbiased linear estimator of $\underline{\beta}$ is

$$\hat{\beta} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \cdot \underline{Y}$$

We define residuals as

$$\underline{\hat{\varepsilon}} = (\underline{I} - \underline{X} (\underline{X}^T \underline{X})^{-1} \underline{X}^T) \underline{Y}$$

Let $\underline{H} = \underline{X} (\underline{X}^T \underline{X})^{-1} \underline{X}^T$. We have that $\underline{H}$ is symmetric and

$$\underline{H}^2 = \underline{X} (\underline{X}^T \underline{X})^{-1} \cancel{\underline{X}^T \cdot \underline{X} (\underline{X}^T \underline{X})^{-1}} \underline{X}^T = \underline{H}.$$

The matrix $\underline{H}$ is idempotent. In fact it is the projection onto the subspace spanned by the columns of $\underline{X}$ and hence

$$\operatorname{rank}(\underline{H}) = \operatorname{rank}(\underline{X}) = m.$$

Since $\begin{pmatrix} \hat{\underline{\beta}} \\ \hat{\underline{\varepsilon}} \end{pmatrix}$ is multivariate normal we have that

$$\operatorname{cov}(\hat{\beta}, \hat{\varepsilon}) = \operatorname{cov}\left((\underline{X}^T\underline{X})^{-1}\underline{X}^T Y, (\underline{I} - \underline{H})\underline{Y}\right)$$

$$= (\underline{X}^T\underline{X})^{-1}\underline{X}^T \cdot \sigma^2 \cdot \underline{I}\left(\underline{I} - \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T\right)$$

$$= \sigma^2\left((\underline{X}^T X)^{-1}\underline{X}^T - \cancel{(\underline{X}^T\underline{X})^{-1}}\cancel{(\underline{X}^T\underline{X})}(X^T\underline{X})^{-1}\underline{X}^T\right)$$

$$= 0.$$

This means that $\hat{\beta}$ and $\hat{\varepsilon}$ are independent.

Further

$$\underline{\hat{\varepsilon}}^T \underline{\hat{\varepsilon}} = [(\underline{I} - \underline{H})\underline{Y}]^T [(\underline{I} - \underline{H})\underline{Y}]$$

$$= \underline{Y}^T (\underline{I} - \underline{H}) \underline{Y}$$

$$= (\underline{Y} - \underline{X}\beta)^T (\underline{I} - \underline{H}) (\underline{Y} - \underline{X}\beta)$$

because $(\underline{I} - \underline{H}) \underline{X} = (\underline{I} - \underline{X} (\underline{X}^T\underline{X})^{-1} \underline{X}^T\underline{X})$

$$= \underline{I} - \underline{I}$$

$$= 0.$$

But $\underline{Y} - \underline{X}\beta \sim N(0, \sigma^2 \underline{I})$. This implies that

$$\underline{\hat{\varepsilon}}^T \underline{\hat{\varepsilon}} \sim \chi^2(n-m)$$

where $m = \text{rank}(\underline{X})$.

If $\hat{\beta}_k$ is the k-th component of $\hat{\beta}$ it is independent of $\hat{\underline{\varepsilon}}$.

We know that $\hat{\beta}_k \sim N(\beta_k, \sigma^2 \cdot c_{kk})$ where $\underline{C} = (\underline{X}^T \underline{X})^{-1}$ and c_{kk} is the diagonal element. The expression

$$T = \frac{\dfrac{\hat{\beta}_k - \beta_k}{\sqrt{c_{kk}}}}{\sqrt{\hat{\underline{\varepsilon}}^T \underline{\varepsilon} \,/\, (n-m)}}$$

is a quotient of a $N(0, \sigma^2)$ random variable and the square root of a $\sigma^2 \chi^2(n-m)/n-m$ random variable hence

$$T \sim t_m.$$

This is what you see on regression printouts.

2. Abstract expected values.

Example: Players A and B each get 5 cards from a well shuffled deck of cards. Let X be the number of aces of player A and Y be the number of aces of player B. From elementary probability we know that

$$E(Y \mid X = k) = 5 \cdot \frac{4-k}{47}$$

for $k = 0, 1, 2, 3, 4$. The right side is a function of k. Call it $\psi(k)$.

Suppose we ask about the

conditional expectation <u>before</u> the cards are dealt. At that time k is unknown and is a random variable. But then the conditional expectation is also a random variable! Which one? Obviously $\psi(X)$. This is Kolmogorov's idea of a random variable that plays the role of abstract conditional expectation

$$E(Y \mid X) = \psi(X).$$

Let X, Y be a pair of discrete random variables and denote

$$E(Y \mid X = x) = \psi(x).$$

Assume $E|Y| < \infty$ and let g be a bounded function. We compute

$$E[\psi(X)g(X)]$$

$$= \sum_x \psi(x) g(x) P(X=x)$$

$$= \sum_x \underbrace{\sum_y y \cdot P(Y=y \mid X=x)}_{\psi(x)} g(x) P(X=x)$$

$$= \sum_{x,y} y\, g(x)\, P(Y=y \mid X=x) P(X=x)$$

$$= \sum_{x,y} y\, g(x)\, P(X=x, Y=y)$$

$$= E[Y \cdot g(X)].$$

This expression uniquely determines the function ψ because we can take $g(x) = I_{\{x\}}$.

This is the idea for the general matematical definition of $E(Y|X)$.

Definition :

(i) Let Y be a random variable with $E|Y| < \infty$. The conditional expectation of Y with respect of X is the function $\psi(X)$ such that for any bounded g we have

$$E(Y g(X)) = E(\psi(X) g(X)).$$

(ii) Let Y be a random variable with $E(|Y|) < \infty$. The conditional expectation of Y given $X_1, X_2, \ldots, X_n$ is a function $\psi(X_1, X_2, .., X_n)$

such that for any bounded $g: \mathbb{R}^n \to \mathbb{R}$ we have

$$E(Y g(X_1, \dots, X_n)) = E[\psi(X_1 \cdots X_n) g(X_1, \dots, X_n)]$$

<u>Theorem 2.1</u> (Radon - Kolmogorov)
The conditional expectation of Y given $X_1, X_2, \dots, X_n$ for Y with $E|Y| < \infty$ exists and is uniquely determined.

<u>Proof</u>: R. Durrett, Probability: Theory and Examples, 2nd Ed., Duxbury 1995

Let us look at elementary properties of conditional expectation. We will write

$$E(Y | X_1, \dots, X_n) = E(Y | \underline{X}).$$

<u>Theorem 2.2</u> (Linearity) If Y_1, Y_2 are random variables with $E|Y_1| < \infty$ and $E|Y_2| < \infty$. Then

$$E(\alpha Y_1 + \beta Y_2 \mid \underline{X})$$

$$= \alpha E(Y_1 \mid \underline{X}) + \beta E(Y_2 \mid \underline{X})$$

<u>Proof</u>: The right side should satisfy the definition. We compute

$$E[(\alpha E(Y_1 \mid \underline{X}) + \beta E(Y_2 \mid \underline{X})) \cdot g(\underline{X})]$$

$$= \alpha E[E(Y_1 \mid \underline{X}) g(\underline{X})]$$

$$+ \beta E[E(Y_2 \mid \underline{X}) g(\underline{X})]$$

$$= \alpha E(Y_1 g(\underline{X})) + \beta E(Y_2 g(X))$$

$$= E[(\alpha Y_1 + \beta Y_2) g(\underline{X})].$$

Theorem 2.3 : (Tower property).

Let $E|Y| < \infty$ and $m < n$.

Then

$$E[Y \mid X_1, \dots, X_m]$$

$$= E[\, E(Y \mid X_1, \dots, X_n) \mid X_1, \dots, X_m].$$

Proof : We compute

$$E[\ E[\,E(Y \mid X_1, \dots, X_n) \mid X_1, \dots, X_m]\, g(X_1, \dots, X_m)]$$

$$\overset{\text{def.}}{=} E[\,E(Y \mid X_1, \dots X_n)\, g(X_1, \dots, X_m)]$$

$$\overset{\text{def}}{=} E(Y\, g(X_1, \dots, X_m))$$

This concludes the proof.

Theorem 2.4: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function such that $E|f(\underline{x})Y| < \infty$. Then

$$E(f(\underline{x})Y \mid \underline{x}) = f(\underline{x}) E(Y \mid \underline{x}).$$

Proof: Left to the reader.

In a similar way we can think about variances. They are functions of $\underline{x}$. To get a formal definition we just replace expectations by conditional expectation.

<u>Definitions</u>: The (abstract) conditional variance of Y given $\underline{X}$ is defined by

$$\operatorname{var}(Y \mid \underline{X}) = E(Y^2 \mid \underline{X}) - E(Y \mid \underline{X})^2.$$

<u>Examples</u>:

(i) Let $X_1, X_2, \ldots, X_n$ be independent and equally distributed. Let $S_n = X_1 + X_2 + \cdots + X_n$. What is $E(X_1 \mid S_n)$?

By symmetry (X_k, S_n) have the same distribution and so $E(X_k\, g(S_n)) = E[\psi(S_n)\, g(S_n)]$ for the same ψ.

By linearity

$$E(X_1 | S_n) + \cdots + E(X_n | S_n)$$

$$= E(S_n | S_n)$$

$$= S_n$$

But all the terms on the left are equal to $\psi(S_n)$. So

$$E(X_1 | S_n) = \frac{S_n}{n}$$

(ii) Let X, Z be discrete and independent. Let f be a function such that $E|f(X,Z)| < \infty$. Define $\psi(x) = E[f(x,Z)]$. We claim that

$$E[f(X,Z) | X] = \psi(X).$$

We compute

$$E[\psi(X) g(X)]$$

$$= \sum_x \psi(x) g(x) P(X=x)$$

$$= \sum_x \sum_z f(x,z) P(Z=z) P(X=x) g(x)$$

(indep)

$$= \sum_{x,z} f(x,z) P(X=x, Z=z) g(x)$$

$$= E[f(X,Z) g(X)].$$

In general this is also true but slightly more difficult to prove. A more general version for vectors is also true. If $\underline{X}, \underline{Z}$ are independent we have

$$E[f(\underline{X},\underline{Z}) \mid \underline{Z}] = \psi(\underline{X}) \quad \text{where}$$

$$\psi(\underline{x}) = E[f(\underline{x}, \underline{Z})].$$

(iii) Compute

$$E[E(Y|\underline{x})]$$

$$= E[E(Y|\underline{x}) \cdot g(\underline{x})] \qquad g \equiv 1$$

$$= E(Y)$$

So

$$\boxed{E[E(Y|\underline{x})] = E(Y)}$$

For variances we get

$$E[\text{var}(Y|\underline{x})]$$

$$= E[E(Y^2|\underline{x}) - E(Y|\underline{x})^2]$$

$$= E(Y^2) - E[E(Y|\underline{x})^2]$$

$$= E(Y^2) - E(Y)^2$$

$$- \left(E[E(Y|\underline{x})^2] - E(Y)^2\right)$$

$$= \operatorname{var}(y) - \operatorname{var}(E(y \mid \underline{x}))$$

Rearrange to get

$$\boxed{\operatorname{var}(y) = E[\operatorname{var}(y \mid \underline{x})] + \operatorname{var}(E(y \mid x))}$$

This is a well known <u>variance decomposition</u> formula.

(iv) Let $\underline{x}$ be multivariate normal. Write

$$\underline{x} \sim N\left(\begin{pmatrix} \mu_1 \\ \underline{\mu}_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \underline{\Sigma}_{12} \\ \underline{\Sigma}_{22} & \underline{\Sigma}_{22} \end{pmatrix} \right).$$

What is $E(x_1 \mid x_2, \dots, x_n)$?

$\operatorname{var}(x_1 \mid x_2, \dots, x_n)$.

Preliminary calculation: Let $\underline{X}, Y$ be independent. In this case $E(Y \mid X) = E(Y)$ and $\mathrm{var}(Y \mid \underline{X}) = \mathrm{var}(Y)$. We know that $X_1 - \underline{\Sigma}_{12}\underline{\Sigma}_{22}^{-1}\underline{X}^2$ is independent of $\underline{X}^2$, so

$$E\left(X_1 - \underline{\Sigma}_{12}\underline{\Sigma}_{22}^{-1}\underline{X}^2 \mid \underline{X}^2\right)$$

$$= E\left(X_1 - \underline{\Sigma}_{12}\underline{\Sigma}_{22}^{-1}\underline{X}^2\right)$$

$$= \mu_1 - \underline{\Sigma}_{12}\underline{\Sigma}_{22}\,\underline{\mu}^2$$

By linearity, however,

$$E\left(\underline{\Sigma}_{12}\underline{\Sigma}_{22}^{-1}\underline{X}^2 \mid \underline{X}^2\right) = \underline{\Sigma}_{12}\underline{\Sigma}_{22}^{-1}\underline{X}^2.$$

Putting all the pieces together

$$E(X_1 \mid \underline{X}^2) = \mu_1 + \underline{\Sigma}_{12}\Sigma_{11}^{-1}\left(\underline{X}^2 - \underline{\mu}^2\right)$$

For any function h we have

$$\mathrm{var}(Y + h(\underline{X}) \mid \underline{X}) = \mathrm{var}(Y \mid \underline{X}).$$

The reader can check that. So

$$\mathrm{var}(X_1 \mid \underline{X}^2)$$

$$= \mathrm{var}(X_1 - \underline{\Sigma}_{12} \underline{\Sigma}_{22} \underline{X}^2 \mid \underline{X}^2)$$

$$\overset{\text{(indep)}}{=} \mathrm{var}(X_1 - \Sigma_{12} \underline{\Sigma}_{22}^{-1} \underline{X}^2)$$

$$= \mathrm{var}(X_1) + \mathrm{var}(\Sigma_{12} \underline{\Sigma}_{22}^{-1} \underline{X}^2)$$

$$+ 2\,\mathrm{cov}(X_1, \underline{\Sigma}_{12} \underline{\Sigma}_{22}^{-1} \underline{X}^2)$$

$$= \underline{\Sigma}_{11} + \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{22} \Sigma_{22}^{-1} \Sigma_{21}$$

$$+ 2\, \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$$

$$= \underline{\Sigma}_{11} - \underline{\Sigma}_{12} \underline{\Sigma}_{22}^{-1} \underline{\Sigma}_{21}$$

(iv) What is the best predictor of Y given $X_1, X_2, \dots, X_n$?

A predictor is a function $f(X_1, X_2, \dots, X_n)$. The best predictor is the conditional expectation. We prove this by direct calculation :

$$E\left[\left(f(X_1, \dots, X_n) - Y\right)^2\right]$$

$$= E\left[\left(f(\underline{X}) - \psi(\underline{X}) + \psi(\underline{X}) - Y\right)^2\right]$$

$$= \underbrace{E\left[\left(f(\underline{X}) - \psi(\underline{X})\right)^2\right]}_{\geq 0.}$$

$$+ E\left[\left(\psi(\underline{X}) - Y\right)^2\right]$$

$$+ 2 E\left[\left(f(\underline{X}) - \psi(\underline{X})\right)\left(\psi(\underline{X}) - \underline{Y}\right)\right]$$

Side calculation:

$$E\left[\left(f(\underline{x})-\psi(\underline{x})\right)\left(\psi(\underline{x})-y\right)\right]$$

$$= E\left[E\left[\text{''} \mid \underline{x}\right]\right]$$

$$\overset{\text{Thm 2.3}}{=} E\left[\left(f(\underline{x})-\psi(\underline{x})\right)\underbrace{E\left[\psi(\underline{x})-y \mid \underline{x}\right]}_{\substack{=0 \\ \text{by def.}}}\right]$$

$$= 0.$$

So $\psi(\underline{x})$ is the best predictor!