Compulsory 4

mandag 11. april 2022 13:52

① The covariance matrix of a random vector X must be symmetric as $Cov(X_i, X_j) = Cov(X_j, X_i)$ and it must be positive semi-definite (PSD) since

$$x^T \Sigma x = \sum_{i=1}^{p}\sum_{j=1}^{p} x_i x_j Cov(X_i, X_j)$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p} x_i x_j E[(X_i-\mu_i)(X_j-\mu_j)]$$

$$= E\left[\sum_{i=1}^{p}\sum_{j=1}^{p} x_i x_j (X_i-\mu_i)(X_j-\mu_j)\right] = E\left[\left(\sum_{i=1}^{p} x_i(X_i-\mu_i)\right)^2\right] \geq 0 \text{ always positive !}$$

However, the covariance matrix of two random vectors need not have these properties.

Let X be $p\times 1$, $E[X]=\mu$

Y be $q\times 1$, $E[Y]=\gamma$ be arbitrary random vectors

Then the definition reads

$$\Sigma_{XY} = Cov(X,Y) = E[(X-\mu)(Y-\gamma)^T]$$

$$= Cov(X_i, Y_j)_{\substack{i=1,\dots,p\\ j=1,\dots,q}}$$

So we have

$$\Sigma_{XY} = \begin{pmatrix} Cov(X_1,Y_1) & Cov(X_1,Y_2) & Cov(X_1,Y_3) & \cdots \\ Cov(X_2,Y_1) & Cov(X_2,Y_2) & Cov(X_2,Y_3) & \vdots \\ Cov(X_3,Y_1) & Cov(X_3,Y_2) & Cov(X_3,Y_3) & \vdots \\ & & \ddots & \vdots \\ \cdots & \cdots & \cdots & Cov(X_p,Y_q) \end{pmatrix}$$

Which, by manipulating X and Y, may correspond to any matrix (Proof that it is not necessarily symm/PSD at bottom of the page)

Example of manipulation:

Z_q is a q-variate normal random vector with $z_i \sim iid$

$$\Sigma_{XY} = \Sigma_{XY} \cdot I_q = \Sigma_{XY}\, Cov(Z_q, Z_q) = Cov(\underset{X}{\underset{\Downarrow}{\Sigma_{XY}Z_q}}, \underset{Y}{\underset{\Downarrow}{Z_q}})$$

If Σ_{XY} is a 0 matrix, we still have a normal distribution with the Dirac-Delta function as the PDF (maybe this is not 100% correct)

② a) No

Counter example:

Let X and Y be univariate normal RV's and let $Z = |X|sign(Y)$, then Z is normal.

Consider $W = (X, Z, Z)^T$, then the linear combination $X + \frac{1}{2}Z + \frac{1}{2}Z = X + Z$ is not normal, as $P\{X+Z=0\} = \frac{1}{2}$ due to the fact that they have the same absolute value. This property cannot exist in a normal distribution as a normal distribution is continuous.

b) Yes, a vector X is multivariate normal iff $b^TX = b_1X_1 + \dots + b_nX_n = Y \sim \mathcal{N}(\mu, \sigma^2)$ has a univariate normal distribution. It follows from the fact that the sum of independent normal RV's, i.e. Y, is normal that X must be multivariate

c) First, we recall our corollary, which states that

"Any vector of components of a multivariate normal vector is multivariate normal. More exactly: Let $X = (X_1, \dots, X_p)^T$ be p-variate normal. $q < p$ and $X' = (X_{i_1}, X_{i_2}, \dots, X_{i_q})^T$ with $i_1 < i_2 < \dots < i_q$. Then X' is q-variate normal"

So in our case we need (X_1, X_2), (X_1, X_3) and (X_2, X_3) to be normal.

Consider (X_2, X_3), which is bivariate.

Let X_2, Y be normally distributed and $X_3 = |X_2|sign(Y)$

Then by a), the linear combination $X_2 + X_3$ is not univariate normal, and (X_2, X_3) is not bivariate normal. Hence X is not trivariate normal by the corollary.

③ We find the estimators in the usual way, by minimizing the RSS

$$RSS = \sum_{i=1}^{n}(X_i - \hat{X}_i)^2 + \sum_{j=1}^{m}(Y_j - \hat{Y}_j)^2$$

$$= \sum_{i=1}^{n}(X_i - E[X_i])^2 + \sum_{j=1}^{m}(Y_j - E[Y_j])^2$$

$$= \sum_{i=1}^{n}(X_i - \beta_0)^2 + \sum_{j=1}^{m}(Y_j - (\beta_0 + \beta_1 y_j))^2$$

Differentiate w.r.t. β_0 and β_1 and equate to 0

$$\frac{\partial RSS}{\partial \beta_0} = \sum_{i=1}^{n} -2(X_i - \beta_0) + \sum_{j=1}^{m} -2(Y_j - (\beta_0 + \beta_1 y_j)) = 0$$

$$= \sum_{i=1}^{n} -2X_i + 2n\beta_0 + \sum_{j=1}^{m} -2Y_j + 2\beta_1 y_j + 2m\beta_0 = 0$$

$\Updownarrow$ solve for β_0 to obtain estimator

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n}X_i + \sum_{j=1}^{m}(Y_j - \hat{\beta}_1 y_j)}{n+m}$$

$$\frac{\partial RSS}{\partial \beta_1} = \sum_{j=1}^{m} -2y_j(Y_j - (\beta_0 + \beta_1 y_j)) = 0$$

$$= \sum_{j=1}^{m} -2Y_j y_j + 2\beta_0 \sum_{j=1}^{m} y_j + 2\beta_1 \sum_{j=1}^{m} y_j^2 = 0$$

$\Updownarrow$ solve for β_1 to obtain estimator

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{m} Y_j y_j - \hat{\beta}_0 \sum_{j=1}^{m} y_j}{\sum_{j=1}^{m} y_j^2}$$

Remove dependencies:

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{m} Y_j y_j - \frac{\sum_{i=1}^{n}X_i + \sum_{j=1}^{m}(Y_j - \hat{\beta}_1 y_j)}{n+m}\sum_{j=1}^{m} y_j}{\sum_{j=1}^{m} y_j^2} = \frac{\sum_{j=1}^{m} Y_j y_j - \frac{1}{n+m}\cdot\left(\sum_{i=1}^{n}X_i\sum_{j=1}^{m}y_j + \sum_{j=1}^{m}Y_j\sum_{j=1}^{m}y_j - \hat{\beta}_1\left(\sum_{j=1}^{m}y_j\right)^2\right)}{\sum_{j=1}^{m} y_j^2}$$

$$\Rightarrow \hat{\beta}_1\left(1 - \frac{\left(\sum_{j=1}^{m}y_j\right)^2}{(n+m)\sum_{j=1}^{m}y_j^2}\right) = \frac{\sum_{j=1}^{m} Y_j y_j - \frac{1}{n+m}\cdot\left(\sum_{i=1}^{n}X_i\sum_{j=1}^{m}y_j + \sum_{j=1}^{m}Y_j\sum_{j=1}^{m}y_j\right)}{\sum_{j=1}^{m} y_j^2}$$

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{m} Y_j y_j - \frac{1}{n+m}\cdot\left(\sum_{i=1}^{n}X_i\sum_{j=1}^{m}y_j + \sum_{j=1}^{m}Y_j\sum_{j=1}^{m}y_j\right)}{\left(1 - \frac{\left(\sum_{j=1}^{m}y_j\right)^2}{(n+m)\sum_{j=1}^{m}y_j^2}\right)\sum_{j=1}^{m} y_j^2} = \frac{(n+m)\sum_{j=1}^{m}Y_j y_j - \left(\sum_{i=1}^{n}X_i\sum_{j=1}^{m}y_j + \sum_{j=1}^{m}Y_j\sum_{j=1}^{m}y_j\right)}{(n+m)\sum_{j=1}^{m}y_j^2 - \left(\sum_{j=1}^{m}y_j\right)^2} \quad ✓$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n}X_i + \sum_{j=1}^{m}Y_j - \hat{\beta}_1\sum_{j=1}^{m}y_j}{n+m} = \frac{\sum_{i=1}^{n}X_i + \sum_{j=1}^{m}Y_j - \frac{(n+m)\sum_{j=1}^{m}Y_j y_j - \left(\sum_{i=1}^{n}X_i\sum_{j=1}^{m}y_j + \sum_{j=1}^{m}Y_j\sum_{j=1}^{m}y_j\right)}{(n+m)\sum_{j=1}^{m}y_j^2 - \left(\sum_{j=1}^{m}y_j\right)^2}\sum_{j=1}^{m}y_j}{n+m} \quad ✓$$

To justify, we can check if the estimators are unbiased:

$$E[\hat{\beta}_1] = E\left[\frac{(n+m)\sum_{j=1}^{m}Y_j y_j - \left(\sum_{i=1}^{n}X_i\sum_{j=1}^{m}y_j + \sum_{j=1}^{m}Y_j\sum_{j=1}^{m}y_j\right)}{(n+m)\sum_{j=1}^{m}y_j^2 - \left(\sum_{j=1}^{m}y_j\right)^2}\right] = \frac{E\left[(n+m)\sum_{j=1}^{m}Y_j y_j\right] - E\left[\left(\sum_{i=1}^{n}X_i\sum_{j=1}^{m}y_j + \sum_{j=1}^{m}Y_j\sum_{j=1}^{m}y_j\right)\right]}{(n+m)\sum_{j=1}^{m}y_j^2 - \left(\sum_{j=1}^{m}y_j\right)^2} = \frac{(n+m)\sum_{j=1}^{m}E[Y_j]y_j - \left(\sum_{i=1}^{n}E[X_i]\sum_{j=1}^{m}y_j + \sum_{j=1}^{m}E[Y_j]\sum_{j=1}^{m}y_j\right)}{(n+m)\sum_{j=1}^{m}y_j^2 - \left(\sum_{j=1}^{m}y_j\right)^2}$$

$$= \frac{\cancel{(n+m)\beta_0\sum_{j=1}^{m}y_j} + (n+m)\beta_1\sum_{j=1}^{m}y_j^2 - \left(\cancel{n\beta_0\sum_{j=1}^{m}y_j} + \cancel{m\beta_0\sum_{j=1}^{m}y_j} + \beta_1\left(\sum_{j=1}^{m}y_j\right)^2\right)}{(n+m)\sum_{j=1}^{m}y_j^2 - \left(\sum_{j=1}^{m}y_j\right)^2} = \frac{(n+m)\beta_1\sum_{j=1}^{m}y_j^2 - \beta_1\left(\sum_{j=1}^{m}y_j\right)^2}{(n+m)\sum_{j=1}^{m}y_j^2 - \left(\sum_{j=1}^{m}y_j\right)^2} = \beta_1 \quad \text{Unbiased!}$$

$$E[\hat{\beta}_0] = E\left[\frac{\sum_{i=1}^{n}X_i + \sum_{j=1}^{m}(Y_j - \hat{\beta}_1 y_j)}{n+m}\right] = \frac{1}{n+m}E\left[\sum_{i=1}^{n}X_i + \sum_{j=1}^{m}(Y_j - \hat{\beta}_1 y_j)\right]$$

$$= \frac{1}{n+m}\left(\sum_{i=1}^{n}E[X_i] + \sum_{j=1}^{m}E[Y_j] - \sum_{j=1}^{m}E[\hat{\beta}_1]y_j\right)$$

$$= \frac{1}{n+m}\left(n\beta_0 + m\beta_0 + \beta_1\sum_{j=1}^{m}y_j - \beta_1\sum_{j=1}^{m}y_j\right)$$

$$= \frac{n\beta_0 + m\beta_0}{n+m} = \beta_0 \Rightarrow \text{Unbiased!}$$

These estimators are found by minimizing RSS, i.e. they are the ordinary least squares estimators. From the Gauss-Markov theorem, these estimators thus have the lowest sampling variance among all linear unbiased estimators. Therefore they are a reasonable choice.

④

$$X = \overset{n \diagdown p}{\begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & \vdots \\ \vdots & 1 & \vdots \\ \vdots & \vdots & \vdots \\ 1 & 1 & \vdots \\ \vdots & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix}} \begin{matrix} \left.\right\} 1 \leq i \leq 10 \\ \left.\right\} 11 \leq i \leq 20 \\ \left.\right\} 21 \leq i \leq 30 \end{matrix}$$

$$R^2 = 0{,}4246 = 1 - \frac{SSE}{SST}$$

$$SST = 42{,}6397$$

General hypothesis test

$H_0: \beta_1 - \beta_2 = 1$ $H_1: \beta_1 - \beta_2 \neq 1$ $\alpha = 0.05$

$A\beta = d$ $A\beta \neq d$ $p = 3$ $n = 30$

In our case:

$A = (0, 1, -1)$ $d = 1$

$r = \text{rank } A = 1$

$$\hat{\beta} = (X^TX)^{-1}X^TY = [\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3]^T$$

By a calculator we find

$$(X^TX)^{-1} = \begin{pmatrix} 0.1 & -0.1 & -0.1 \\ -0.1 & 0.2 & 0.1 \\ -0.1 & 0.1 & 0.2 \end{pmatrix}$$

Our test statistic:

$$F = \frac{(SSE^R - SSE)/r}{SSE/(n-p)}$$

$$= \frac{(A\hat{\beta} - d)(A(X^TX)^{-1}A^T)^{-1}(A\hat{\beta} - d)}{SSE/(n-p)} \quad \text{where } SSE = SST(1-R^2)$$

which under H_0 follows a Fisher distribution, $F \sim F_{r,n-p}$ and if $F \geq f_{\alpha,r,n-p}$ we reject H_0

Now we must find $\hat{\beta}$. We use the given values and the regression model.

This is an over-determined system so we can find the $\hat{\beta}$ values from the given table

$\hat{\beta}_0$ is easily found as $x_{1j} = \dots = x_{10j} = 0$ $j = 1,2$

So

$$\hat{\beta}_0 = Y_2 - \beta_1 X_{21} - \beta_2 X_{22} - \varepsilon_2$$

$$= Y_2 - \varepsilon_2$$

$$= 1.4 - (-1.1) = 2.5$$

$Y_9, \hat{\varepsilon}_9$ gives the same values

We use $\hat{\beta}_0$ to find $\hat{\beta}_1$ as $x_{11,2} = \dots = x_{20,2} = 0$

$$\hat{\beta}_1 = Y_{17} - \hat{\beta}_0 - \hat{\beta}_2 x_{17,2} - \varepsilon_{17}$$

$$= Y_{17} - \hat{\beta}_0 - \varepsilon_{17}$$

$$= 3.8 - 2{,}5 - 0.5 = 0.8$$

$Y_{12}, \hat{\varepsilon}_{12}$ gives the same values

and lastly

$$\hat{\beta}_2 = Y_{25} - \hat{\beta}_0 - \beta_1 x_{25,1} - \varepsilon_{25}$$

$$= Y_{25} - \hat{\beta}_0 - \varepsilon_{25}$$

$$= 5{,}7 - 2{,}5 - 1.4 = 1.8$$

So $\hat{\beta} = \begin{pmatrix} 2{,}5 \\ 0{,}8 \\ 1{,}8 \end{pmatrix}$

and so we can find F by plugging in all values

$F = 22{,}01$

From a table we find $f_{0{,}05,1,27} = 4{,}21$

and hence we must reject H_0 as $F > f_{0{,}05,1,27}$, concluding that $\beta_1 - \beta_2 \neq 1$

1) Proof

Quickly, we prove it is not symm/PSD:

$$Cov(X,Y)^T = E[(X - E[X])(Y - E[Y])^T]^T$$

$$= E[(Y - E[Y])(X - E[X])^T]$$

$$= Cov(Y, X)$$

So $Cov(X,Y) \neq Cov(Y,X)$

Furthermore

$$y^T\Sigma x = \sum_{i=1}^{p}\sum_{j=1}^{q} x_i y_j Cov(X_i, Y_j)$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q} x_i y_j E[(X_i - \mu_i)(Y_j - \gamma_j)]$$

$$= E\left[\sum_{i=1}^{p}\sum_{j=1}^{q} x_i y_j (X_i - \mu_i)(Y_j - \gamma_j)\right] \quad \text{which can be negative, eg if } p=1=q, \text{ and } x_i y_j (X_i - \mu_i)(Y_i - \gamma_j) < 0$$