Multivariate Normal Distribution (Sections 2.11 and 5.8)

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Gesell data

Let X be the age in in months a child speaks his/her first word and let Y be the Gesell adaptive score, a measure of a child's aptitude (observed later on). Are X and Y related? How does the child's aptitude *change* with how long it takes them to speak?

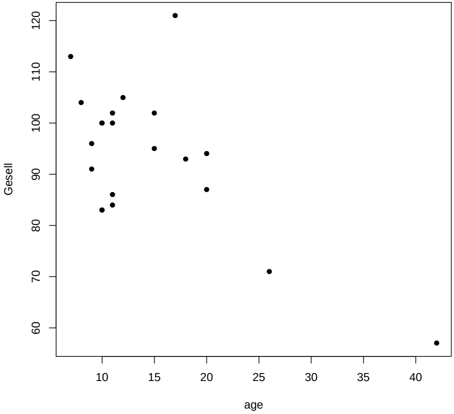
Here's the Gesell score y_i and age at first word in months x_i data, $i = 1, \ldots, 21$.

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
15	95	26	71	10	83	9	91	15	102
20	87	18	93	11	100	8	104	20	94
7	113	9	96	10	83	11	84	11	102
10	100	12	105	42	57	17	121	11	86
10	100								

In R, we compute $r = -0.640$, a moderately strong negative relationship between age at first word spoken and Gesell score.

```
> age=c(15,26,10,9,15,20,18,11,8,20,7,9,10,11,11,10,12,42,17,11,10)
> Gesell=c(95,71,83,91,102,87,93,100,104,94,113,96,83,84,102,100,105,57,121,86,100)
> plot(age,Gesell)
> cor(age,Gesell)
[1] -0.64029
```

Scatterplot of $(x_1, y_1), \ldots, (x_{21}, y_{21})$



Random vectors

A random vector $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}$ is made up of, say, k random variables.

A random vector has a joint distribution, e.g. a density $f(\mathbf{x})$, that gives probabilities

$$P(\mathbf{X} \in A) = \int_A f(\mathbf{x}) d\mathbf{x}.$$

Just as a random variable X has a mean $E(X)$ and variance $\text{var}(X)$, a random vector also has a mean *vector* $E(\mathbf{X})$ and a covariance *matrix* $\text{cov}(\mathbf{X})$.

Mean vector & covariance matrix

Let $\mathbf{X} = (X_1, \ldots, X_k)$ be a random vector with density $f(x_1, \ldots, x_k)$. The mean of $\mathbf{X}$ is the vector of marginal means

$$E(\mathbf{X}) = E\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}\right) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_k) \end{bmatrix}. \qquad (5.38)$$

The covariance matrix of $\mathbf{X}$ is given by

$$cov(\mathbf{X}) = \begin{bmatrix} cov(X_1, X_1) & cov(X_1, X_2) & \cdots & cov(X_1, X_k) \\ cov(X_2, X_1) & cov(X_2, X_2) & \cdots & cov(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_k, X_1) & cov(X_k, X_2) & \cdots & cov(X_k, X_k) \end{bmatrix}. \qquad (5.42)$$

The normal distribution generalizes to multiple dimensions. We'll first look at two jointly distributed normal random variables, then discuss three or more.

The *bivariate normal density* for (X_1, X_2) is given by $f(x_1, x_2) =$

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

There are 5 parameters: $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

Besides 5.8, also see 2.11 pp.78–83.

Bivariate normal distribution

- This density jointly defines X_1 and X_2, which live in $\mathbb{R}^2 = (-\infty, \infty) \times (-\infty, \infty)$.
- Marginally, $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ (p. 79).
- The correlation between X_1 and X_2 is given by $\text{corr}(X_1, X_2) = \rho$ (p. 80).
- For jointly normal random variables, if the correlation is zero then *they are independent*. This is not true in general for jointly defined random variables.
- $E(\mathbf{X}) = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \;\; \text{cov}(\mathbf{X}) = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}.$

Bivariate normal PDF level curves

Figure 2.15 A bivariate normal distribution with parameters $\theta_X = 1$, $\theta_Y = 2$, $\sigma_X = 3$, $\sigma_Y = 2$, $\rho = 0.5$, with expanding ellipses enclosing 5%, 25%, 50%, 75% and 95% of the probability distribution.

Proof that X_1 indpendent X_2 when $\rho = 0$

When $\rho = 0$ the joint density for (X_1, X_2) simplifies to

$$
\begin{aligned}
f(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{1}{2}\left[\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2\right]\right\} \\
&= \left[\frac{1}{\sqrt{2\pi}\sigma_1} e^{-0.5\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2}\right]\left[\frac{1}{\sqrt{2\pi}\sigma_2} e^{-0.5\left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2}\right].
\end{aligned}
$$

Since these are each respectively functions of x_1 and x_2 only, and the range of (X_1, X_2) factors into the produce of two sets, X_1 and X_2 are independent and in fact $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$.

Conditional distributions $[X_1|X_2 = x_2]$ and $[X_2|X_1 = x_1]$ (pp. 80–81)

The conditional distribution of X_1 given $X_2 = x_2$ is

$$[X_1|X_2 = x_2] \sim N\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2), \sigma_1^2(1-\rho^2)\right).$$

Similarly,

$$[X_2|X_1 = x_1] \sim N\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho(x_1 - \mu_1), \sigma_2^2(1-\rho^2)\right).$$

This ties directly to linear regression:

To predict $X_2|X_1 = x_1$, we have

$$E(X_2|X_1 = x_1) = \left[\mu_2 - \frac{\sigma_2}{\sigma_1}\rho\mu_1\right] + \left[\frac{\sigma_2}{\sigma_1}\rho\right]x_1 = \beta_0 + \beta_1 x_1.$$

Bivariate normal distribution as data model

Here we assume

$$\begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix} \overset{iid}{\sim} N_2\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}\right),$$

or succinctly,

$$\mathbf{X}_i \overset{iid}{\sim} N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

If the bivariate normal model is appropriate for paired outcomes, it provides a convenient probability model with some nice properties.

Say n outcome pairs are to be recorded:
$\{(X_{11}, X_{12}), (X_{21}, X_{22}), \ldots, (X_{n1}, X_{n2})\}$. The i^{th} pair is (X_{i1}, X_{i2}).

Sample mean vector & covariance matrix

The *sample mean vector* is given elementwise by

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} X_{i1} \\ \frac{1}{n}\sum_{i=1}^{n} X_{i2} \end{bmatrix},$$

and the *sample covariance matrix* is given elementwise by

$$\mathbf{S} = \begin{bmatrix} \frac{1}{n-1}\sum_{i=1}^{n}(X_{i1}-\bar{X}_1)^2 & \frac{1}{n-1}\sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2) \\ \frac{1}{n-1}\sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2) & \frac{1}{n-1}\sum_{i=1}^{n}(X_{i2}-\bar{X}_2)^2 \end{bmatrix}.$$

Sample mean vector & covariance matrix

The sample mean $\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$ is the MLE of $\boldsymbol{\mu}$ and the sample covariance matrix $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$ is unbiased for $\boldsymbol{\Sigma}$.

It can be shown that

$$\bar{\mathbf{X}} \sim N_2\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right).$$

The matrix $(n-1)\mathbf{S}$ has a "Wishart" distribution (generalizes χ^2).

Estimation

The sample mean vector $\bar{\mathbf{X}}$ estimates $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and the sample covariance matrix $\mathbf{S}$ estimates
$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$.
We will place hats on parameter *estimators* based on the data. So

$$\hat{\mu}_1 = \bar{X}_1,\ \hat{\mu}_2 = \bar{X}_2,\ \hat{\sigma}_1^2 = s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2,$$

$$\hat{\sigma}_2^2 = s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{i2} - \bar{X}_2)^2.$$

Also,

$$\widehat{cov}(X_1, X_2) = \frac{1}{n-1}\sum_{i=1}^{n}(X_{i1} - \bar{X})(X_{i2} - \bar{X}_2).$$

Correlation coefficient r

So a natural estimate of ρ is then

$$\hat{\rho} = \frac{\widehat{cov}(X_1, X_2)}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2}}.$$

This is in fact the MLE estimate based on the bivariate normal model. It is also a “plug-in” estimator based on the method-of-moments as well as the now-familiar Pearson correlation coefficient.

Gesell data

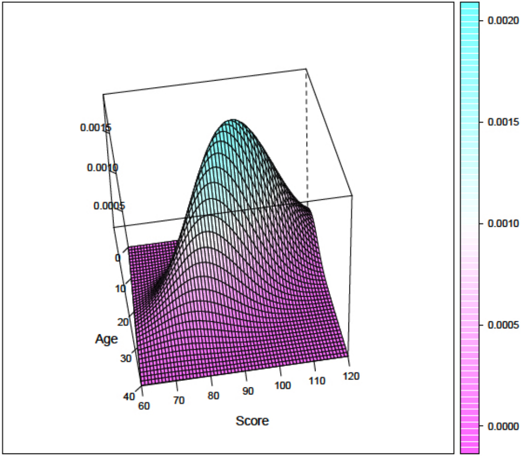
Recall: X is age in months a child speaks his/her first word and let Y is Gesell adaptive score, a measure of a child's aptitude.
Question: how does the child's aptitude *change* with how long it takes them to speak? Here, $n = 21$.
In R we find $\bar{\mathbf{X}} = \begin{bmatrix} 14.38 \\ 93.67 \end{bmatrix}$. Also, $\mathbf{S} = \begin{bmatrix} 60.14 & -67.78 \\ -67.78 & 186.32 \end{bmatrix}$.
Assuming a bivariate model, we plug in the estimates and obtain the estimated PDF for (X, Y):

$$f(x,y) = \exp(-60.22+1.3006x-0.0134x^2+0.9520y-0.0098xy-0.0043y^2).$$

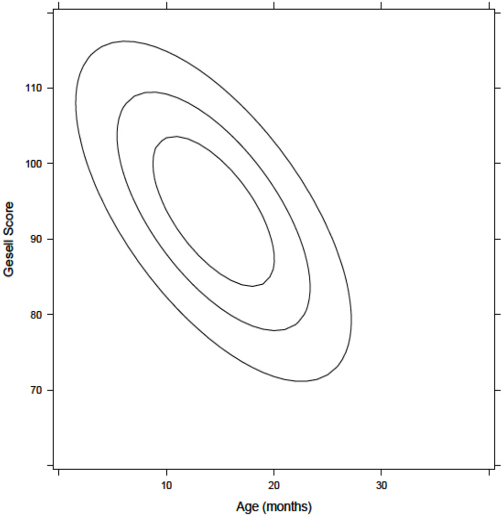
We can further find from $Y \stackrel{\bullet}{\sim} N(93.67, 186.32)$,

$$f_Y(y) = \exp(-3.557 - 0.00256(y - 93.67)^2).$$

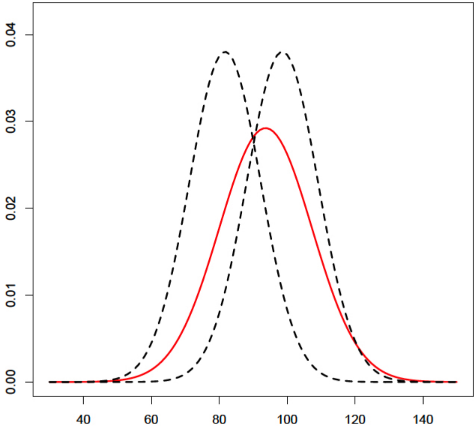
3D plot of $f(x, y)$ for (X, Y) estimated from data



Density estimate with actual data



Gesell conditional distribution



Solid is $f_Y(y)$; left dashed is $f_{Y|X}(y|25)$ the right dashed is $f_{Y|X}(y|10)$. As the age in months of first words $X = x$ increases, the distribution of Gesell Adaptive Scores Y decreases.

Multivariate normal distribution

In general, a k-variate normal is defined through the mean and covariance matrix:

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} \sim N_k \left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{bmatrix} \right).$$

Succinctly,

$$\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Recall that if $Z \sim N(0,1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$. The definition of the multivariate normal distribution just extends this idea.

Multivariate normal made from independent normals

Instead of one standard normal, we have a list of k independent standard normals $\mathbf{Z} = (Z_1, \ldots, Z_k)$, and consider the same sort of transformation in the multivariate case using matrices and vectors.

Let $Z_1, \ldots, Z_k \overset{iid}{\sim} N(0,1)$. The joint pdf of $(Z_1, \ldots, Z_k)$ is given by

$$f(z_1, \ldots, z_k) = \prod_{i=1}^{k} \exp(-0.5 z_i^2)/\sqrt{2\pi}.$$

Let

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{bmatrix},$$

where $\boldsymbol{\Sigma}$ is symmetric (i.e. $\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}$, which implies $\sigma_{ij} = \sigma_{ji}$ for all $1 \le i, j \le k$).

Multivariate normal made from independent normals

Let $\mathbf{\Sigma}^{1/2}$ be any matrix such that $\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}$. Then $\mathbf{X} = \boldsymbol{\mu} + \mathbf{\Sigma}^{1/2}\mathbf{Z}$ is said to have a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$, written

$$\mathbf{X} \sim N_k(\boldsymbol{\mu}, \mathbf{\Sigma}).$$

Written in terms of matrices

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} + \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{bmatrix}^{1/2} \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_k \end{bmatrix}.$$

Joint PDF

Using some math, it can be shown that the pdf of the new vector $\mathbf{X} = (X_1, \ldots, X_k)$ is given by

$$f(x_1, \ldots, x_k | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\{-0.5(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\}.$$

In the one-dimensional case, this simplifies to our old friend

$$f(x_1 | \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\{-0.5(x - \mu)(\sigma^2)^{-1}(x - \mu)\},$$

the pdf of a $N(\mu, \sigma^2)$ random variable X.

$|\mathbf{A}|$ is the determinant of the matrix $\mathbf{A}$, and is a function of the elements of $\mathbf{A}$, but beyond this course.

Properties of multivariate normal vectors

Let

$$\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Then

1. For each X_i in $\mathbf{X} = (X_1, \ldots, X_k)$, $E(X_i) = \mu_i$ and $var(X_i) = \sigma_{ii}$. That is, marginally, $X_i \sim N(\mu_i, \sigma_{ii})$.
2. For any two (X_i, X_j) where $1 \leq i < j \leq k$, $cov(X_i, X_j) = \sigma_{ij}$. The off-diagonal elements of $\boldsymbol{\Sigma}$ give the covariance between two elements of $(X_1, \ldots, X_k)$. Note then $\rho(X_i, X_j) = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$.

Properties of multivariate normal vectors

Let

$$\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Then

1. For any $r \times k$ matrix $\mathbf{M}$,

$$\mathbf{MX} \sim N_r(\mathbf{M}\boldsymbol{\mu}, \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}').$$

2. For any $k \times 1$ vector $\mathbf{m} = (m_1, \ldots, m_k)$, $\mathbf{m} + \mathbf{X} \sim N_k(\mathbf{m} + \boldsymbol{\mu}, \boldsymbol{\Sigma})$.
3. For $r_1 \times k$ matrix $\mathbf{M_1}$ and $r_2 \times k$ matrix $\mathbf{M_2}$, the joint distribution of $\mathbf{M_1Y}$ and $\mathbf{M_2Y}$ can be found as

$$\begin{pmatrix} \mathbf{M_1} \\ \mathbf{M_2} \end{pmatrix} \mathbf{Y} \sim N_{r_1+r_2}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \mathbf{M_1}\Sigma\mathbf{M}_1' & \mathbf{M_1}\Sigma\mathbf{M}_2' \\ \mathbf{M_2}\Sigma\mathbf{M}_1' & \mathbf{M_2}\Sigma\mathbf{M}_2' \end{pmatrix} \right)$$

Example

Let

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N_3\left(\begin{bmatrix} -2 \\ 5 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 4 \end{bmatrix} \right).$$

E.g., $X_2 \sim N(5, 3)$ and $cov(X_2, X_3) = -1$.

Define

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & -1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \text{ and } \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \mathbf{MX} = \begin{bmatrix} 1 & 0 & -1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}.$$

Example

Then

$$\begin{bmatrix} 1 & 0 & -1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim$$

$$N_2\left(\begin{bmatrix} 1 & 0 & -1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} -2 \\ 5 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & -1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 4 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{3} \\ 0 & \frac{1}{3} \\ -1 & \frac{1}{3} \end{bmatrix} \right),$$

or simplifying,

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N_2\left(\begin{bmatrix} -2 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & \frac{11}{9} \end{bmatrix} \right).$$

Note that for the transformed vector $\mathbf{Y} = (Y_1, Y_2)$, $cov(Y_1, Y_2) = 0$ and therefore Y_1 and Y_2 are uncorrelated, i.e. $\rho(Y_1, Y_2) = 0$.

For the linear model (e.g. simple linear regression or the two-sample model) $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the error vector is assumed (pp. 222–223)

$$\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \mathbf{I}_{n\times n}\sigma^2).$$

Then the least squares estimators have a multivariate normal distribution

$$\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2).$$

$p = 2$ is the number of mean parameters. (The MSE has a gamma distribution).