stats545.2 545.2 The multivariate normal distribution and projections in the linear model

Gunnar Stefansson

5. september 2022

Efnisyfirlit

1 Addding distributional assumptions: The multivariate normal and related distributions

1.1 A theorem from calculus

1.1.1 Handout

To find a multivariate density of a transformed variable, recall from calculus that if g is a $1-1$ function $g : \mathbf{R}^n \to \mathbf{R}^n$

$$\int f(\mathbf{z})d\mathbf{z} = \int f(g(\mathbf{y}))|J|d\mathbf{y}\ (*)$$

$whereJistheJacobianofthetransformationJ = \left|\frac{d\mathbf{z}}{d\mathbf{y}}\right| = \left|\frac{\partial g(\mathbf{y})}{\partial \mathbf{y}}\right|$ and the integrals are over corresponding regions.
If f is the density of $\mathbf{Z}$, then the left-hand integral over a set A is $P[\mathbf{Z} \in A]$, and if $\mathbf{Y} = g(\mathbf{Z})$ we also know that

$$P[\mathbf{Y} \in B] = P[g(\mathbf{Z}) \in g(A)],$$

but this left-hand side is the integral of the joint p.d.f. of $\mathbf{Y}$ over B, which must now be equal to the r.h.s. og (*).
It follows that the joint pdf of $\mathbf{Y}$ is h with $h(y) = f(g(y))|J|$.

1.2 The multivariate normal distribution

1.2.1 Handout

Suppose $Z_1,...,Z_n$ are independent Gaussian with mean zero and variance one (i.e. $Z_1,...,Z_n \sim n(0,1)$, i.i.d.) so their joint density is

$$f(\mathbf{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-z_i^2/2\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-(1/2)\mathbf{z}^T\mathbf{z}\right)$$

and this is the density of the multivariate random variable $\mathbf{Z} = (Z_1,\ldots,Z_n)'$.
Let A be an invertible $n \times n$ matrix and $\mu \in \check{}\mathbf{R}^n$ and define a new multivariate random variable, $\mathbf{Y} = A\mathbf{Z} + \mu$.
Some linear algebra gives

$$h(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)\right)$$

where $\Sigma = AA^T$.
This leads to a natural definition of the multivariate normal distribution.
The n-dimensional random vector, $\mathbf{Y}$ is **defined** to have a multivariate normal distribution, denoted $\mathbf{Y} \sim n(\mu,\Sigma)$ if the density of $\mathbf{Y}$ is of the form

$$h(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)\right)$$

where $\mu \in \mathbb{R}^n$ and Σ is a symmetric positive definite $n \times n$ matrix.
It is left to the reader to prove that if $\mathbf{Y} \sim n(\mu,\Sigma)$ and B is an $p \times n$ matrix of full rank p ($p < n$), then $B\mathbf{Y}$ also has a multivariate normal distribution.

1.3 Related distributions

1.3.1 Handout

If $Z \sim n(0,1)$ is standard normal, then we **define** the chi-squared distribution on one degree of freedom, χ_1^2 to be the distribution of $U := Z^2$ and write

$$U \sim \chi_1^2.$$

If $U_1, \dots, U_p$ are i.i.d. χ_1, then we **define** χ_p^2 to be the distribution of their sum and write

$$\sum_{i=i}^{p} U_i \sim \chi_p^2.$$

Finally, if $U \sim \chi_{\nu_1}^2$ and $V \sim \chi_{\nu_2}^2$ are independent, then we **define** the **F distribution on** ν_1 **and** ν_2 **degrees of freedom** to be the distribution of the ratio $\frac{U/\nu_1}{V/\nu_2}$ and write

$$\frac{U/\nu_1}{V/\nu_2} \sim F_{\nu_1,\nu_2}.$$

2 Orthogonal projections in multiple regression

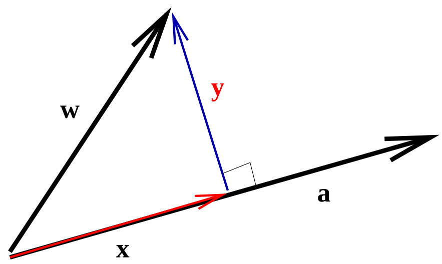
2.1 Background to projections

If $\mathbf{a}$ is a vector then we can write a general vector $\mathbf{w}$ in the form $\mathbf{w} = \mathbf{x} + \mathbf{y}$ where $\mathbf{x} = k\mathbf{a}$ and $\mathbf{a}'\mathbf{y} = \mathbf{a} \cdot \mathbf{y} = 0$.
In the general case,

$$k = \frac{\mathbf{w} \cdot \mathbf{a}}{||\mathbf{a}||^2},$$

and for unit vectors $\mathbf{a}$ we obtain

$$k = \mathbf{w} \cdot \mathbf{a}.$$



2.1.1 Handout

If $\mathbf{a}$ is a vector then we can write a general vector $\mathbf{w}$ in the form $\mathbf{w} = \mathbf{x} + \mathbf{y}$ where $\mathbf{x} = k\mathbf{a}$ and $\mathbf{a} \cdot \mathbf{y} = 0$.
With this

$$\begin{aligned} \mathbf{w} \cdot \mathbf{a} &= (\mathbf{x} + \mathbf{y}) \cdot \mathbf{a} \\ &= (k\mathbf{a} + \mathbf{y}) \cdot \mathbf{a} = k\mathbf{a} \cdot \mathbf{a} + \underbrace{\mathbf{y} \cdot \mathbf{a}}_{=0} \\ &= k \cdot ||\mathbf{a}||^2 \end{aligned}$$

i.e.

$$k = \frac{\mathbf{w} \cdot \mathbf{a}}{||\mathbf{a}||^2},$$

and therefore $\mathbf{w} = \mathbf{x} + \mathbf{y}$ with

$$\mathbf{x} = \frac{\mathbf{w} \cdot \mathbf{a}}{||\mathbf{a}||^2}\mathbf{a}$$

and residual:

$$\mathbf{y} = \mathbf{w} - \frac{\mathbf{u} \cdot \mathbf{a}}{||\mathbf{a}||^2}\mathbf{a}.$$

Note that we have shown that this is the **only possible** solution to writing $\mathbf{w} = \mathbf{x} + \mathbf{y}$ where $\mathbf{x} = k\mathbf{a}$ and $\mathbf{a} \cdot \mathbf{y} = 0$ but not that it is indeed such a solution. Obviously $\mathbf{x}$ is of the stated form and it is not hard to see that $\mathbf{a} \cdot \mathbf{y} = 0$ is indeed true for this solution. This orthogonal decomposition therefore both exists and is unique.

2.2 Projections and bases

The Gram-Schmidt technique uses projections to iteratively build an orthonormal basis, $\mathbf{u}_1, \ldots \mathbf{u}_r$ which spans the same space as a sequence of arbitrary starting vectors, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_p$.
In linear regression, the starting vectors are typically the columns of the $\mathbf{X}$-matrix. r above is then the rank of the matrix.

2.2.1 Handout

The Gram-Schmidt technique uses the projections of the previous section to iteratively build an orthonormal basis, $\mathbf{u}_1, \ldots \mathbf{u}_r$ which spans the same space as a sequence of arbitrary starting vectors, $\mathbf{a}_1,\ \mathbf{a}_2, \ldots \mathbf{a}_p$:

$$\mathbf{u}_1 := \frac{1}{||\mathbf{a}||}\mathbf{a}_1$$

then for $i = 1, \ldots, p-1$

$$\mathbf{v}_{i+1} := (\mathbf{a}_{i+1} \cdot \mathbf{a}_1)\mathbf{a}_1 + \ldots + (\mathbf{a}_{i+1} \cdot \mathbf{a}_i)\mathbf{a}_i$$

with residual

$$\mathbf{e}_{i+1} := \mathbf{a}_{i+1} - \mathbf{v}_{i+1}$$

and next vector

$$\mathbf{u}_{i+1} := \frac{1}{||\mathbf{e}_{i+1}||}\mathbf{e}_{i+1}$$

If the starting vectors are linearly independent, then $r = p$, otherwise $r < p$ (some of the proposed basis vectors have turned out to be zero and are omitted).
It is often useful to expand the set of starting vectors, to $\mathbf{a}_1,\ \mathbf{a}_2, \ldots \mathbf{a}_p, \mathbf{e}_1,\ \mathbf{e}_2, \ldots \mathbf{e}_n$, where the $\mathbf{e}_i$ are the usual unit vectors. The method will then result in a full basis for $\mathbb{R}^n$, the first r of which span the same space as the starting vectors.

3 A basis for V=span(X)

3.1 Subspaces

In the model $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, assume $rank(\mathbf{X}) = r$ where $\mathbf{X}$ is $n \times p$ and $r \leq p$

Recall $\mathbf{X}\hat{\beta}$ is a project so $\mathbf{y} - \mathbf{X}\hat{\beta} \perp \mathbf{X}\hat{\beta}$ so that $\mathbf{y} - \mathbf{X}\hat{\beta} \in \mathbf{V}^{\perp} = \{v : v \perp sp(\mathbf{X})\}$ and $dim(\mathbf{V}^{\perp}) = n - r$.
If $r = p$, then:

$$\underbrace{\hat{\mathbf{e}}}_{n\times 1} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$$
$$= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')y = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

and $rank(\mathbf{I} - \mathbf{H}) = dim(\mathbf{V}^{\perp}) = n - p$

3.1.1 Details

Assume $rank(\mathbf{X}) = r \leq p$ ($\mathbf{X}$ is $n \times p$).

Parameters in the model $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ are estimated with $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$ if the inverse exists or in general with any $\hat{\beta}$ which is such that $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ is a projection onto the subspace $sp(\mathbf{X})$.

By definition, a projection $\hat{\mathbf{y}}$ simply corresponds to a decomposition of the original vector into two orthogonal components, i.e. writing $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$. We have $\hat{\mathbf{e}} = y - \mathbf{X}\hat{\beta} \perp \hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ so that $y - \mathbf{X}\hat{\beta} \in \mathbf{V}^{\perp} = \{v : v \perp sp(\mathbf{X})\}$ and $dim(\mathbf{V}^{\perp}) = n - r$.

$$\underbrace{\hat{\mathbf{e}}}_{n\times 1} = y - \mathbf{X}\hat{\beta} = y - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$$
$$= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')y = (\mathbf{I} - \mathbf{H})y$$

and $rank(\mathbf{I} - \mathbf{H}) = dim(\mathbf{V}^{\perp}) = n - r$

3.2 A basis for the span of X

Orthonormal basis, $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ for $\mathbf{R}^n$:

Using Gram-Schmidt, first generate $\mathbf{u}_1, \ldots, \mathbf{u}_r$ which span $sp\{\mathbf{X}\}$, with $rank\{\mathbf{X}\} = r$ and the rest, $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_n$ are chosen so that the entire set, $\mathbf{u}_1, \ldots, \mathbf{u}_n$ spans $\mathbf{R}^n$.

$$\begin{aligned} \mathbf{X}\hat{\beta} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_r\mathbf{u}_r \\ \mathbf{y} &= \hat{\zeta}_1\mathbf{u}_1 + \ldots \hat{\zeta}_r\mathbf{u}_r + \hat{\zeta}_{r+1}\mathbf{u}_{r+1} + \ldots + \hat{\zeta}_n\mathbf{u}_n \end{aligned}$$

3.2.1 Details

The probability distributions can best be viewed by defining a new orthonormal basis, $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ for $\mathbf{R}^n$.

This basis is defined by first generating a set of r vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ which span the space defined by $sp\{\mathbf{X}\}$, and the rest, $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_n$ are chosen so that the entire set, $\mathbf{u}_1, \ldots, \mathbf{u}_n$

spans $\mathbf{R}^n$. This is obviously always possible using the method of Gram-Schmidt. This gives the following sequence of spaces and spans:

$$\begin{aligned} sp\{\mathbf{X}\} &= sp\{\mathbf{u}_1,\ldots,\mathbf{u}_r\} \\ \mathbf{R}^n &= sp\{\mathbf{u}_1,\ldots,\mathbf{u}_r,\mathbf{u}_{r+1},\ldots\mathbf{u}_n\} \end{aligned}$$

One can then write each of $\mathbf{X}\hat{\beta}$ and $\mathbf{y}$ in terms of the new basis as follows:

$$\begin{aligned} \mathbf{X}\hat{\beta} &= \hat{\zeta}_1\mathbf{u}_1+\ldots\hat{\zeta}_r\mathbf{u}_r \\ \mathbf{y} &= \hat{\zeta}_1\mathbf{u}_1+\ldots\hat{\zeta}_r\mathbf{u}_r+\hat{\zeta}_{r+1}\mathbf{u}_{r+1}+\ldots\hat{\zeta}_n\mathbf{u}_n \end{aligned}$$

where it is well-known that $\hat{\zeta}_i = \mathbf{u}_i\cdot\mathbf{y}$.

It is important to note that the same coefficients $\hat{\zeta}_i$ are obtained for $1\le i\le r$. This follows from considering the coefficient of $\mathbf{y}$ in the basis and noting that $\mathbf{y} = \mathbf{X}\hat{\beta}+\hat{\mathbf{e}}$ where the residual vector $\hat{\mathbf{e}}$ is orthogonal to all column vectors of $\mathbf{X}$ and therefore also to $\mathbf{u}_i$ for $1\le i\le r$. Therefore,

$$\hat{\zeta}_i = \mathbf{u}_i\cdot\mathbf{y} = \mathbf{u}_i\cdot\mathbf{X}\hat{\beta}$$

3.3 Q-R decomposition

$$\mathbf{Q} := \left[\mathbf{u}_1 \vdots \mathbf{u}_2 \vdots \ldots \vdots \mathbf{u}_n\right]$$

is the $\mathbf{Q}$ in the Q-R decomposition of $\mathbf{X} = \mathbf{Q}\mathbf{R}$.

If

$$\mathbf{z} = \left(\hat{\zeta}_1,\hat{\zeta}_2,\ldots,\hat{\zeta}_n\right)$$

then

$$\mathbf{z} = \mathbf{Q}'\mathbf{y}$$

and hence

$$E[\mathbf{z}] = \mathbf{Q}'\mathbf{X}\beta$$

$$V[\mathbf{z}] = \mathbf{Q}'\sigma^2\mathbf{I}\mathbf{Q} = \sigma^2\mathbf{I}$$

3.3.1 Details

$\mathbf{Q} := [\mathbf{u}_1\mathbf{u}_2\ldots\mathbf{u}_n]$ is the $\mathbf{Q}$ in the Q-R decomposition of $\mathbf{X}$.

$\mathbf{Q}$ has important properties, e.g. $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$ so $\mathbf{Q}^{-1} = \mathbf{Q}'$.

If

$$\mathbf{z} = \left(\hat{\zeta}_1,\hat{\zeta}_2,\ldots,\hat{\zeta}_n\right)$$

then

$$\mathbf{z} = \mathbf{Q}'\mathbf{y} \text{ and } \mathbf{y} = \mathbf{Q}\mathbf{z}$$

and hence

$$E[\mathbf{z}] = \mathbf{Q}'\mathbf{X}\beta$$

$$V[\mathbf{z}] = \mathbf{Q}'\sigma^2\mathbf{I}\mathbf{Q} = \sigma^2\mathbf{I}$$

3.4 Variances of coefficients

For each i we obtain

$$V\left[\hat{\zeta}_i\right] = \sigma^2$$

3.4.1 Details

For each i we trivially obtain

$$V\left[\hat{\zeta}_i\right] = \sigma^2$$

3.5 Expected values of coefficients

For $i = r+1, \ldots, n$ we obtain

$$E\left[\hat{\zeta}_i\right] = 0$$

3.5.1 Details

The expected values of the coefficients, $\hat{\zeta}_i$ depend on which space these correspond to. Define

$$\zeta_i = E\left[\hat{\zeta}_i\right]$$

and by linearity we obtain

$$\zeta_i = E\left[\mathbf{u}_i \cdot \mathbf{y}\right] = \mathbf{u}_i \cdot (\mathbf{X}\beta).$$

Now note that we have defined the basis vectors in three sets. The first is such that they span the same space as the columns of $\mathbf{Z}$. The second set complements the first to span the $\mathbf{X}$ and the last set complements the set to span all of $\mathbf{R}^n$. The basis vectors are of course all orthogonal and each basis vector is orthogonal to all vectors in spaces spanned by preceding vectors.

For $i = r+1, \ldots, n$ we obtain

$$E\left[\hat{\zeta}_i\right] = \mathbf{u}_i \cdot (\mathbf{X}\beta) = 0$$

since $\mathbf{X}\beta$ is trivially in the space spanned by the column vectors of $\mathbf{X}$ and is therefore a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_r$ and $\mathbf{u}_i$ is orthogonal to all of these.

3.6 Sums of squares and norms

$$SSE(F) = ||\mathbf{y} - \mathbf{X}\hat{\beta}||^2 \quad = \sum_{i=p+1}^{n} \hat{\zeta}_i^2$$

3.6.1 Details

It is now quite easy to see how to form sums of squared deviations based on the new orthonormal basis, since each set of deviations corresponds to a specific portion of the space.

$$SSE(F) = ||\mathbf{y} - \mathbf{X}\hat{\beta}||^2 \quad = \sum_{i=r+1}^{n} \hat{\zeta}_i^2$$

3.7 Normality and independence of coeffients

Note that $\hat{\zeta}_i$ are linear combinations of the various y_j since $\hat{\zeta}_i = \mathbf{u}_i \cdot \mathbf{y}$.

When the y_i are independent Gaussian random variables, $\hat{\zeta}_i$ have zero covariance and are thus also independent.

3.7.1 Details

Note that $\hat{\zeta}_i$ are linear combinations of the various y_j since $\hat{\zeta}_i = \mathbf{u}_i \cdot \mathbf{y}$. The $\hat{\zeta}_i$ have zero covariance and when the y_i are independent Gaussian random variables, the $\hat{\zeta}_i$ are also independent.

This final result uses the fact that Gaussian random variables which have zero covariance are also independent. The fact that they have zero covariance is easy to establish, but the corrollary of independence is a result from multivariate normal theory.

The normal theory is fairly simple in this case:

$$\mathbf{z} = \left(\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_n\right) = \mathbf{Q}'\mathbf{y}$$

and

$$\mathbf{y} \sim n\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$$

.
It follows that $\mathbf{z}$ is multivariate normal and from the earlier derivations of the mean and variance we have

$$\mathbf{z} \sim n\left(\mathbf{Q}'\mathbf{X}\beta, \sigma^2\mathbf{I}\right).$$

3.8 Degrees of freedom

$SSE(F)$has $n-r$ degrees of freedom.

3.8.1 Details

$SSE(F)$ has $n-r$ degrees of freedom.