More Linear Algebra

Matrix Rank

A matrix can be expressed as a combination of the column vectors. The rank of a matrix corresponds to the number of linearly independent columns. If a matrix is not full rank, this will have implications on the invertibility of the matrix (or a product of the matrix $X^T X$).

```
X <- tibble(x_cat = sample(c('1','2'), 20, replace = T),
            x_contin = runif(20))

X_mat <- model.matrix(~x_cat + x_contin, data = X)
rankMatrix(X_mat)
```

```
## [1] 3
## attr(,"method")
## [1] "tolNorm2"
## attr(,"useGrad")
## [1] FALSE
## attr(,"tol")
## [1] 4.440892e-15
```

Vector space As previously mentioned, a matrix can be constructed as a linear combination of vectors. Formally, a vector space is the combination of vectors that can be constructed with a set of vectors. A vector space. V_n has the two following properties. First assume $\underline{x}$ and $\underline{y}$ are in V_n, then

1. $\alpha\underline{x} \in V_n$ for all scalar α

2. $\underline{x} + \underline{y} \in V_n/$

If V_n consists of all linear combinations of the set of vectors $v_1, v_2, \cdots, v_s$, then the set of vectors $(v_1, v_2, \cdots, v_s)$ is a spanning set of V_n.

If the spanning set are linearly independent, then the spanning set is a **basis set** and the number of vectors in the basis set is the dimension of the vector set.

For purposes of this class, and most statistical applications, we will think about a vector space as the combination of column vectors from a matrix. This is referred to as a column space. Here the dimension of the column space corresponds to the rank of the matrix, X.

Idempotent matrix Let P be an $n \times n$ matrix. Then P is idempotent if $P \times P = P$

```
I <- diag(3); I
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

```
I %*% I
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

```
P <- matrix(c(3,1,-6,-2),2,2); P
```

```
##      [,1] [,2]
## [1,]    3   -6
## [2,]    1   -2
```

```
P %*% P
```

```
##      [,1] [,2]
## [1,]    3   -6
## [2,]    1   -2
```

A square matrix P is also a projection operator if $P = P \times P$. In our settings projection operators will be used to project the column space of a matrix.

Let $\underline{y} \sim N(X\underline{\beta}, \sigma^2 I)$. We will derive this later, but consider the hat matrix $H = X(X^T X)^{-1} X^T$.

Test $H \times H$

$$X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = X(X^T X)^{-1} X^T$$

so H is idempotent and a projection operator. Formally, it will project the columns of X onto the column space of X, such that the least squares estimator is achieved.

Let $\underline{y} \sim N(X\underline{\beta}, \sigma^2 I)$. Then following a simple example from Boik, $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{bmatrix}$ and $y = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$.

The OLS estimate of $\underline{\beta}$, $\underline{\hat{\beta}}_{OLS} = \begin{bmatrix} 1 \\ \frac{3}{7} \end{bmatrix}$.

Formally, $\underline{\beta}_{OLS} = (X^T X)^{-1} X^T$, so

$$X\underline{\beta}_{OLS} = X(X^T X)^{-1} X^T \underline{y} = H\underline{y}$$

```
X <- matrix(c(1, 1, 1, 1, 2, 4),3, 2)
beta_hat <- c(1, 3/7)

X %*% beta_hat
```

```
##          [,1]
## [1,] 1.428571
## [2,] 1.857143
## [3,] 2.714286
```

```
H <- X %*% solve(t(X) %*% X) %*% t(X)
y <- c(2, 1, 3)
H %*% y
```

```
##          [,1]
## [1,] 1.428571
## [2,] 1.857143
## [3,] 2.714286
```

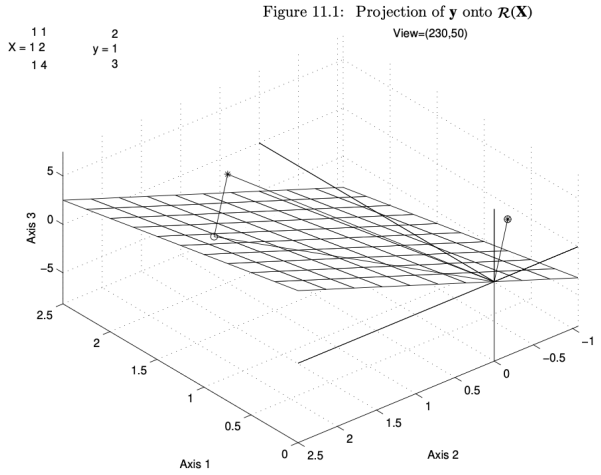


Figure 1: Projection, obtained from Boik 505 notes

Properties of random vectors and matrices

Expectation , Variance, Covariance Expectation of a vector is fairly simple, formally

$$E[\underline{y}] = \begin{pmatrix} E[y_1] \\ E[y_2] \\ \vdots \\ E[y_n] \end{pmatrix} = \begin{pmatrix} X_1\underline{\beta} \\ X_2\underline{\beta} \\ \vdots \\ X_n\underline{\beta} \end{pmatrix}$$

The same idea holds for a random matrix.

If $\underline{y}$ is an $n \times 1$ vector and $\underline{x}$ is an $r \times 1$ vector, then.

1. $\text{Cov}(\underline{y}, \underline{x})$ is an $n \times r$ covariance matrix, (Σ_{yx}) where the individual elements are covariance terms. Specifically $\Sigma_{yx}[i, j]$ is the covariance between the i^{th} element in y and the j^{th} element in x.

2. $\text{Var}(\underline{y})$ is a variance matrix, such as Σ.

If $\underline{y} \sim N(\underline{\mu}, \Sigma)$ then the pdf of $\underline{y}$ is

$$p(\underline{y}) = |\Sigma|^{-1/2}(2\pi)^{n/2} \exp\left[-\frac{1}{2}\left(\underline{y} - \underline{\mu}\right)^T \Sigma^{-1}\left(\underline{y} - \underline{\mu}\right)\right]$$

Partitioned Matrices

Now consider splitting the sampling units into two partitions such that $\underline{y} = \begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix}$. Then,

$$\begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \underline{\beta}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix} \right)$$

Fundamentally, there is no change to the model, we have just created "groups" by partitioning the model. Do note that Σ_{11} is an $n_1 \times n_1$ covariance matrix.

$$\Sigma_{11} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n_1} \\ \sigma_{22} & \sigma_2^2 & \cdots & \sigma_{2n_1} \\ \sigma_{31} & \sigma_{32} & \ddots & \sigma_{3n_1} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_1 1} & \sigma_{n_1 2} & \ddots & \sigma_{n_1}^2 \end{bmatrix}$$

However, while $\Sigma_{12} = \Sigma_{21}^T$, neither of these are necessarily symmetric matrices. They also do not have any variance components, but rather just covariance terms. Σ_{12} will be an $n_1 \times n_2$ matrix.

$$\Sigma_{11} = \begin{bmatrix} \sigma_{1,n_1+1} & \sigma_{1,n_1+2} & \cdots & \sigma_{1,n_1+n_2} \\ \sigma_{2,n_1+1} & \sigma_{2,n1+2} & \cdots & \sigma_{2,n_1+n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_1,n_1+1} & \sigma_{n_1,n_1+2} & \ddots & \sigma_{n_1,n_1+n_2} \end{bmatrix}$$

Conditional Multivariate Normal

Here is where the magic happens with correlated data. Let $\underline{y_1}|\underline{Y_2} = \underline{y_2}$ be a conditional distribution for $\underline{y_1}$ given that $\underline{y_2}$ is known. Then

$$\underline{y_1}|\underline{y_2} \sim N\left(X_1\beta + \Sigma_{12}\Sigma_{22}^{-1}\left(\underline{y_2} - X_2\beta\right), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

Now let's consider a few special cases (in the context of the DC housing dataset.)

1. Let $\Sigma = \sigma^2 I$, then the batch of houses in group 1 are conditionally dependent from the houses in group 2 and

$$\underline{y_1}|\underline{y_2} \sim N\left(X_1\beta, \Sigma_{11}\right)$$

2. Otherwise, let $\Sigma = \sigma^2 H$ and we'll assume Σ_{12} has some non-zero elements. Then we have a more precise estimate of $\underline{y_1}$ as $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ will be "less than" Σ_{11} (that positive definite thing). Furthermore, the mean will shift such that highly correlated observations such as houses in close proximity (local model structure) will tend to differ from the global mean in the same fashion.

First a quick interlude about matrix inversion. The inverse of a symmetric matrix is defined such that $E \times E^{-1} = I$. We can calculate the inverse of a matrix for a 1×1 matrix, perhaps as 2×2, matrix and maybe even a 3×3 matrix. However, beyond that it is quite challenging and time consuming. Furthermore, it is also (relatively) time intensive for your computer.

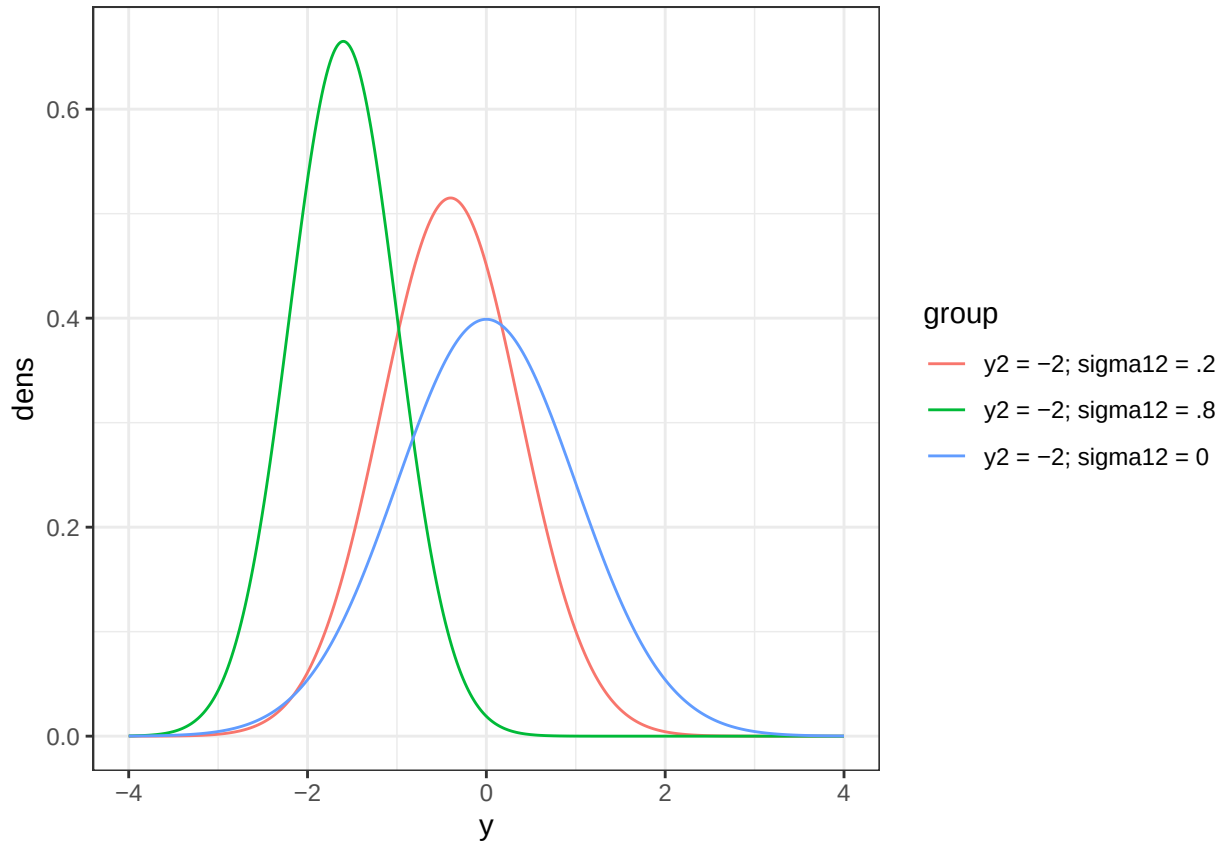
3. Let $n_1 = 1$ and $n_2 = 1$, then

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right)$$

and

$$y_1 | y_2 \sim N\left(\mu_1 + \sigma_{12}(\sigma_2^2)^{-1}\left(y_2 - \mu_2\right), \sigma_1^2 - \sigma_{12}(\sigma_2^2)^{-1}\sigma_{21} \right)$$

Now consider an illustration for a couple simple scenarios. Let $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$. Now assume $y_2 = -2$ and we compare the conditional distribution for a few values of σ_{12}.



One last note, the marginal distributions for any partition $\underline{y_1}$ are quite simple.

$$\underline{y_1} \sim N\left(X_1\beta, \Sigma_{11}\right)$$

or just

$$y_1 \sim N\left(X_1\beta, \sigma_1^2\right)$$

if y_1 is scalar.

OLS / WLS Let $\underline{y} \sim N(X\underline{\beta}, \Sigma)$, then the least squares estimate of $\underline{\beta}$ is the minimizer of

$$SSE(\underline{\beta}) = (\underline{y} - X\underline{\beta})^T \Sigma^{-1} (\underline{y} - X\underline{\beta})$$

$$\frac{\delta SSE(\underline{\beta})}{\delta \underline{\beta}} = 2X^T\Sigma^{-1}X\underline{\beta} - 2X^T\Sigma^{-1}\underline{y},$$

setting this equal to zero results in

$$\hat{\underline{\beta}} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\underline{y}$$

If $\Sigma = \sigma^2 I$, then $\hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y}$ and this is referred to as the ordinary least squares estimator.

In this case, $\hat{\sigma^2} = \frac{\underline{e}^T\underline{e}}{n-p} = \frac{\underline{y}^T(I-H)\underline{y}}{n-p}$, where $\underline{e} = \underline{y} - X\underline{\beta}$.

If Σ is a diagonal matrix with elements equal to $d_1, d_2, \cdots, d_n$, then $\hat{\underline{\beta}}$ is referred to as weighted least squares. As $SSE(\underline{\beta}) = \sum_i \frac{(y_i - X_i\underline{\beta})^2}{d_i} = (\underline{y} - X\underline{\beta})^T\Sigma^{-1}(\underline{y} - X\underline{\beta})$

The more general case, when Σ is not a diagonal matrix, is referred to as Generalized Least Squares (GLS).

Note, these estimators are equivalent to maximum likelihood estimation.

Linear Functions and Contrasts Suppose we are interest in a linear combination of the model coefficients, such as $\psi = \beta_0 + \beta_1$ or $\psi = \beta_2 - \beta_3$. These can be constructed with a matrix C, s.t. $C = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \end{bmatrix}$ or $C = \begin{bmatrix} 0 & 0 & 1 & -1 & 0 & \cdots & 0 \end{bmatrix}$

If ψ contains linear combinations that sum to zero, $\underline{1}_p^T \underline{c} = O$, then ψ is called a contrast.

Contrasts are common in ANOVA models, where $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, where τ_i is the effect for treatment i

A linear function is **estimable** if a linear unbiased estimator of the function exists. Mathematically, $\psi = \underline{c}^T \underline{\beta}$ is estimable if there exists a statistic, $\hat{\psi} = A^T \underline{y} + \underline{k}$ such that $E[\hat{\psi}] = \psi$.

Best Linear Unbiased Estimator (BLUE) Suppose that $C^T \underline{\beta}$ is an estimable function. Let $\hat{\psi}_c$ be any unbiased estimator of $C^T \underline{\beta}$. Meaning

1. $\hat{\psi}_c = A^t \underline{y} + \underline{l}$ and
2. $E[\hat{\psi}_c] = \underline{c}^t \underline{\beta} \ \ \forall \beta$

An estimator is the Best Linear Unbiased Estimator (BLUE) if it is unbiased and has the minimum variance among all unbiased estimators (or in the matrix case if all elements are BLUE).

Gauss - Markov Theorem The Gauss-Markov theorem typically states that if $\underline{y} \sim N(X\underline{\beta}, \sigma^2 I)$, then the OLS estimator is the BLUE. It can also be extended to the GLS setting, see Boik.

It is important to mention a lesser known theorem (particularly in the classical linear models setting), **Stein's Paradox** The associated James–Stein estimator dominates the "ordinary" least squares approach, meaning the James-Stein estimator has a lower or equal mean squared error than OLS. So biased, but lower overall MSE. **more later in the context of shrinkage and hierarchical models**.