Multivariate Regression

The so-called supervised learning problem is the following: we want to approximate the random variable Y with an appropriate function of the random variables $X_1, \dots, X_p$ with the method of *least squares*. That is,

$$\mathbb{E}(Y - t(X_1, \dots, X_p))^2$$

is minimized over all p-variate, measurable functions t. From probability theory it is known, that the optimum t is

$$t_{opt}(x_1, \dots, x_p) = \mathbb{E}(Y|X_1 = x_1, \dots, X_p = x_p) = \frac{\int_{-\infty}^{\infty} y f(y, x_1, \dots, x_p) dy}{\int_{-\infty}^{\infty} f(y, x_1, \dots, x_p) dy},$$

where f is the joint p.d.f. of the above random variables (usually they have an absolutely continuous distribution). t_{opt} is called regression function, and Proposition 5 of Lesson 2 guarantees that it is linear if f is a $(p+1)$-dimensional normal density. Even if our random variables do not have a multivariate normal distribution (which is very usual by the Multivariate Central Limit Theorem), a linear approximation makes sense. 0Frequently, we estimate the regression parameters from a sample, and use so-called linearizing transformations.

1 Linear Regression

Given the expectation vector and covariance matrix of the random vector $(Y, X_1, \dots, X_p)^T$, we want to approximate Y (target or response variable) with a linear combination of the predictor variable $\mathbf{X} = (X_1, \dots, X_p)^T$ in such a way that the least squares error is minimized.

The solution is the following. To minimize the function

$$g(a_1, \dots, a_p, b) = \mathbb{E}(Y - (a_1 X_1 + \dots + a_p X_p + b))^2$$

let us take its partial derivatives with respect to the parameters $a_1, \dots, a_p$ and b, then equal them to 0. Under some regularity conditions (which always hold in exponential families, especially in the multivariate Gaussian case), the differentiation with respect to the parameters results in the following system of equations:

$$\frac{\partial g}{\partial b} = -2\mathbb{E}(Y - (a_1 X_1 + \dots + a_p X_p + b)) = 0$$

which results in

$$b = \mathbb{E}Y - a_1 \mathbb{E}X_1 - \dots - a_p \mathbb{E}X_p \tag{1}$$

when a_i's are already known.

$$\frac{\partial g}{\partial a_i} = 2\mathbb{E}(Y - (a_1 X_1 + \dots + a_p X_p + b))(-X_i) = 0 \quad (i = 1, \dots, p)$$

which, using the solution for b, provides the following system of linear equations:

$$\sum_{j=1}^{p} \mathrm{Cov}(X_i, X_j) a_j = \mathrm{Cov}(Y, X_i), \qquad i = 1, \dots, p.$$

Denoting by $\mathbf{C}$ the covariance matrix of the random vector $\mathbf{X} = (X_1, \dots, X_p)^T$ and $\mathbf{d}$ the p-dimensional vector of cross-covariances between the components of $\mathbf{X}$ and Y, the above system of linear equations has the concise form:

$$\mathbf{Ca} = \mathbf{d}, \tag{2}$$

where $\mathbf{a} = (a_1, \dots, a_p)^T$ is the vector of the parameters a_j's.

The system of equations (2) has the unique solution

$$\mathbf{a} = \mathbf{C}^{-1}\mathbf{d} \tag{3}$$

whenever $|\mathbf{C}| \neq 0$, which holds if there are no linear relations between $X_1, \dots, X_p$ (in the multivariate normal case, they do not have a deformed p-variate distribution). Otherwise, we can take generalized inverse of $\mathbf{C}$, and the solution is not unique.

To show that the formulas (1) and (3) indeed give local and global minimum of our objective function, we investigate the Hessian, which is just the covariance matrix of all the $p+1$ variables, and is usually positive definite; hence, we get a unique minimum. If there were linear relations between the variables X_j's and Y, we may eliminate some of them to get a unique solution (see the forthcoming sections about the dimension reduction).

So the regression function is

$$l(\mathbf{X}) = \mathbf{a}^T\mathbf{X} = \mathbf{d}^T\mathbf{C}^{-1}\mathbf{X},$$

which is equal to $\mathbb{E}(Y \,|\, \mathbf{X})$ if $\mathbb{E}(Y) = 0$ and $\mathbb{E}(\mathbf{X}) = \mathbf{0}$, see the addendum to Lesson 6.

The above minimization is also equivalent to some correlation maximization problem as follows. The above minimization task is equivalent to minimizing $\mathrm{Var}(\varepsilon)$ in the model

$$Y = l(\mathbf{X}) + \varepsilon,$$

where $l(\mathbf{X}) = \sum_{i=1}^{p} a_i X_i + b = \mathbf{a}^T\mathbf{X} + b$ is linear function of the coordinates of $\mathbf{X}$.

In view of (1), $\mathbb{E}(\varepsilon) = 0$, and because the covariance is a bilinear function, not affected by constant shifts of its variables, we get that

$$\begin{aligned} \mathrm{Cov}(l(\mathbf{X}), \varepsilon) &= \mathrm{Cov}(\mathbf{a}^T\mathbf{X}, Y - \mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mathbf{d} - \mathbf{a}^T\mathbf{Ca} \\ &= \mathbf{d}^T\mathbf{C}^{-1}\mathbf{d} - \mathbf{d}^T\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1}\mathbf{d} = 0, \end{aligned} \tag{4}$$

and consequently,

$$\mathrm{Var}(Y) = \mathrm{Var}(l(\mathbf{X})) + \mathrm{Var}(\varepsilon). \tag{5}$$

Further,

$$\mathrm{Cov}(l(\mathbf{X}), Y) = \mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}$$

which is the first term on the right hand side of (4), and

$$\mathrm{Var}(l(\mathbf{X})) = \mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}$$

which is the second term on the right hand side of (4), are the same.

Definition 1 *The multiple correlation between the target variable Y and the predictor variables $X_1, \dots, X_p$ is*

$$\mathrm{Corr}(Y, l(\mathbf{X})) = \frac{\mathrm{Cov}(l(\mathbf{X}), Y)}{\sqrt{\mathrm{Var}(Y)\mathrm{Var}(l(\mathbf{X})}} = \frac{\mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}}{\sqrt{\mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}}\sqrt{\mathrm{Var}(Y)}} = \frac{\sqrt{\mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}}}{\sqrt{\mathrm{Var}(Y)}}$$

which is nonnegative and denoted by $r_{Y(X_1,\dots,X_p)} = r_{Y\mathbf{X}}$.

It is easy to see that in the $p = 1$ case this is the absolute value of the usual correlation coefficient between Y and the only predictor X.

The square of the multiple correlation coefficient can be written in the following form:

$$r^2_{Y\mathbf{X}} = \frac{\mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(l(\mathbf{X}))}{\mathrm{Var}(Y)}.$$

Therefore, the equation (5) gives rise to the following decomposition of the variance of Y:

$$\mathrm{Var}(Y) = r^2_{Y\mathbf{X}}\mathrm{Var}(Y) + (1 - r^2_{Y\mathbf{X}})\mathrm{Var}(Y). \tag{6}$$

Here the first term is the variance of Y explained by the predictor variables, and the second term is the so-called *residual variance*, that is, the variance of the error term ε. Observe that $r^2_{Y\mathbf{X}} = 1$ is equivalent to $\mathrm{Var}(\varepsilon) = 0$, i.e., there is a linear relation between Y and the components of $\mathbf{X}$ with probability 1. The other extreme case $r^2_{Y\mathbf{X}} = 0$ means that $\mathrm{Var}(l(\mathbf{X})) = 0$, i.e., the best linear approximation is constant with probability 1, consequently $a_1 = \dots = a_p = 0$, or equivalently, $\mathbf{a} = \mathbf{0}$ and $\mathbf{d} = \mathbf{0}$; in other words, Y is uncorrelated with all the X_j's, and hence, its best linear approximation is its own expectation.

Without proof we state that the above $l(\mathbf{X})$ has the maximal possible correlation with Y among all possible linear combinations of the components of $\mathbf{X}$.

Proposition 1 *For any linear combination $h(\mathbf{X})$ of $X_1, \dots, X_p$, the following relation holds true:*

$$r_{Y(X_1,\dots,X_p)} = \mathrm{Corr}(Y, l(\mathbf{X}) \geq |\mathrm{Corr}(Y, h(\mathbf{X})|.$$

Consequently, when subtracting $l(\mathbf{X})$ from Y, ε can be considered as the residual after eliminating the effect of the variables $X_1, \dots, X_p$ from Y.

Definition 2 *If two target variables Y_1 and Y_2 are expressed as (different) linear combinations of the same predictor $\mathbf{X}$:*

$$Y_1 = l_1(\mathbf{X}) + \varepsilon_1 \quad \textit{and} \quad Y_2 = l_2(\mathbf{X}) + \varepsilon_2,$$

then the partial correlation between Y_1 and Y_2 after eliminating the effect of $\mathbf{X}$ is the usual Pearson correlation coefficient between the error terms ε_1 and ε_2. We use the notation

$$r_{Y_1Y_2|\mathbf{X}} = \mathrm{Corr}(\varepsilon_1, \varepsilon_2).$$

Note that in the $p = 1$ case, when the only predictor is denoted by X, the following formula is used to calculate the partial correlation:

$$r_{Y_1Y_2|X} = \frac{\mathrm{Corr}(Y_1, Y_2) - \mathrm{Corr}(Y_1, X) \cdot \mathrm{Corr}(Y_2, X)}{\sqrt{(1 - \mathrm{Corr}^2(Y_1, X)) \cdot (1 - \mathrm{Corr}^2(Y_2, X))}}.$$

In fact, the partial correlation measures the correlation between two random variables after eliminating the effect of some nuisance variables. Indeed, in multivariate data structures, it can happen that the Pearson correlation coefficient between two variables does not reflect the pure association between them. Because the correlations are highly interlaced through the correlation matrix, we cannot just pull out two of them. Other variables, which are strongly intercorrelated with both, will disturb their relation; therefore, first we have to eliminate the effect of these nuisance variables. For example, if we consider the correlation between the kiwi consumption and the number of registered cancer cases during the years (in the US), we experience a high correlation between them. However, it does not mean that kiwi causes cancer, it just means that there are other variables (the time and the increasing living standard, which makes rise to buy more tobacco and alcoholic drinks for example, that may cause cancer).

In case of an i.i.d. sample $(Y_1, \mathbf{X}^1)$, . . . ,$(Y_n, \mathbf{X}^n)$ we estimate the parameters by the formulas (1), (3), where we substitute the empirical quantities (ML-estimators) for $\mathbf{C}$ and $\mathbf{d}$. The squared multiple correlation coefficient R^2 is also estimated from the sample, and it gives the proportion of the total variation of Y which is explained by the predictor variables. Therefore, we call it *coefficient of determination.* In the next section we will use hypothesis testing for the significance of R^2.

Note, that with some *linearization* formulas we can use linear regression in the following models:

- *Multiplicative model*:
$$Y \sim b{X_1}^{a_1} \dots {X_p}^{a_p}.$$
After taking the logarithms, one gets
$$\ln Y \sim \ln b + a_1 \ln X_1 + \cdots + a_p \ln X_p,$$
therefore, we can use linear regression for the log-log data. While, the linear regression performs well for data from a multivariate normal distribution, this model favors lognormally distributed data (for example, chemical concentrations).

- *Polynomial regression*: Now we want to approximate Y with a given degree polynomial of X:
$$Y \sim a_1X + a_2X^2 + \cdots + a_pX^p + b.$$
The solution is obtained by applying multivariate linear regression for Y with the predictor variables $X_j = X^j$, $j = 1, \dots, p$.

2 The linear model (with deterministic predictors)

Now our model is the following.

$$Y_i = \sum_{j=1}^{p} a_j x_{ij} + \varepsilon_i \qquad (i = 1, \dots, n),$$

where x_{ij} is the prescribed value of the j-th predictor in the i-th measurement. Since the measurement is burdened with the random noise ε_i, the measured value Y_i of the target variable in the i-th measurement is a random variable. For simplicity, the constant term is zero (in fact, it would be $\bar{Y} - \sum_{j=1}^{p} a_j \bar{x}_j$, but we assume that it has already been subtracted from the left hand sides). We also assume that $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ $(i = 1, \dots, n)$, and the measurement errors are uncorrelated. Because of their equal (but unknown) variance they are called *homoscedastic errors.* Therefore, $\mathbb{E}(Y_i) = \sum_{j=1}^{p} a_j x_{ij}$ and $\mathrm{Var}(Y_i) = \sigma^2$ $(i = 1, \dots, n)$, and the Y_i's are also uncorrelated. Very frequently, the measurement errors are Gaussian, and thus, the random variables $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are also independent, akin to the Y_i's.

With the notation

$$\mathbf{Y} := (Y_1, \dots, Y_n)^T, \quad \underline{\varepsilon} := (\varepsilon_1, \dots, \varepsilon_n)^T$$

and $\boldsymbol{X} = (x_{ij})$ $(i = 1, \dots, n;\ j = 1, \dots, p)$, our model equation can be put into the following matrix form:

$$\mathbf{Y} = \boldsymbol{X}\mathbf{a} + \underline{\varepsilon},$$

where the parameter vector $\mathbf{a} = (a_1, \dots, a_p)^T$ is estimated by the method of least squares, i.e.,

$$\sum_{i=1}^{n} \varepsilon_i^2 = \|\mathbf{Y} - \boldsymbol{X}\mathbf{a}\|^2$$

is minimized with respect to $\mathbf{a}$.

Here $(\mathbf{Y}, \boldsymbol{X})^T = (\mathbf{Y}, \mathbf{x}_1, \dots, \mathbf{x}_p)^T$ is the data matrix, where $\mathbf{x}_j$ denotes the j-th column of the matrix $\boldsymbol{X}$. If the solution is denoted by $\hat{\mathbf{a}}$, a simple linear algebra guarantees that $\boldsymbol{X}\hat{\mathbf{a}}$ is the projection of the random vector $\mathbf{Y}$ onto $F = \mathrm{Span}\{\mathbf{x}_1, \dots, \mathbf{x}_p\} \subset \mathbb{R}^n$. Let us denote the $n \times n$ matrix of this projection by $\mathbf{P}$. Consequently, $\boldsymbol{X}\hat{\mathbf{a}} = \mathbf{P}\mathbf{Y}$ and $\mathbf{Y} - \boldsymbol{X}\hat{\mathbf{a}} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$ are orthogonal, and latter vector is also orthogonal to any vector $\boldsymbol{X}\mathbf{b} \in F$. Therefore,

$$(\boldsymbol{X}\mathbf{b})^T \cdot (\mathbf{Y} - \boldsymbol{X}\mathbf{a}) = 0, \quad \forall \mathbf{b} \in \mathbb{R}^p.$$

From this,

$$\mathbf{b}^T \boldsymbol{X}^T (\mathbf{Y} - \boldsymbol{X}\mathbf{a}) = 0, \quad \forall \mathbf{b} \in \mathbb{R}^p$$

holds, which implies that

$$\boldsymbol{X}^T (\mathbf{Y} - \boldsymbol{X}\mathbf{a}) = \mathbf{0}.$$

In summary, $\hat{\mathbf{a}}$ is the solution of the so-called *Gauss normal equation*

$$\boldsymbol{X}^T \boldsymbol{X}\mathbf{a} = \boldsymbol{X}^T \mathbf{Y}.$$

This equation is always consistent, since $\boldsymbol{X}^T\mathbf{Y}$ is in F, which is also spanned by the column vectors of $\boldsymbol{X}^T\boldsymbol{X}$. In the rank r of F (this is also the rank of $\boldsymbol{X}$) is equal to $p(\leq n)$, then we have a unique solution:

$$\hat{\mathbf{a}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{Y}.$$

From here,

$$\mathbf{PY} = \boldsymbol{X}\hat{\mathbf{a}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{Y},$$

therefore

$$\mathbf{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

If the rank of $\boldsymbol{X}$ is less than p, then there are infinitely many solutions, including the one, obtained by the Moore–Pen rose inverse:

$$\hat{\mathbf{a}} = (\boldsymbol{X}^T\boldsymbol{X})^{+}\boldsymbol{X}^T\mathbf{Y}.$$

Proposition 2 *If* $\text{rank}(\boldsymbol{X}) = p \leq n$ *and* $\underline{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$*, then* $\hat{\mathbf{a}} \sim \mathcal{N}_p(\mathbf{a}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$*.*

Therefore $\hat{\mathbf{a}}$ is an unbiased estimator of $\mathbf{a}$. In this case, $\hat{\mathbf{a}}$ is also ML-estimate of $\mathbf{a}$. Further, the ML-estimate of σ^2 is

$$\hat{\sigma^2} = \frac{SSE}{n}$$

that is biased, see the forthcoming definition of SSE. The unbiased estimate of σ^2 is $\frac{SSE}{n-p-1}$ or $\frac{SSE}{n-p}$ depending on, whether there is or there is no constant term (intercept) in the model (when the variables are previously transformed to have zero mean, there is no constant term).

The forthcoming Gauss–Markov theorem states that $\hat{\mathbf{a}}$ is also efficient among the linear, unbiased estimators.

Theorem 1 (Gauss –Markov Theorem) *For any other unbiased linear estimator* $\tilde{\mathbf{a}}$ *of* $\mathbf{a}$*:*

$$\text{Var}(\hat{\mathbf{a}}) \leq \text{Var}(\tilde{\mathbf{a}}).$$

This means that the difference of the right-hand and left-hand side $p \times p$ *covariance matrices is positive semidefinite. Shortly,* $\hat{\mathbf{a}}$ *provides a* **BLUE** *(Best, Linear, Unbiased Estimate) for* $\mathbf{a}$.

In view of $\mathbf{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{X}^T$, the minimum of our objective function is

$$SSE := \|\mathbf{Y} - \boldsymbol{X}\hat{\mathbf{a}}\|^2 = (\mathbf{Y} - \boldsymbol{X}\hat{\mathbf{a}})^T(\mathbf{Y} - \boldsymbol{X}\hat{\mathbf{a}}),$$

called *residual variance*. It can also be written as

$$\begin{aligned} SSE &= (\mathbf{Y} - \mathbf{PY})^T(\mathbf{Y} - \mathbf{PY}) = ((\mathbf{I} - \mathbf{P})\mathbf{Y})^T((\mathbf{I} - \mathbf{P})\mathbf{Y}) = \\ &= \mathbf{Y}^T(\mathbf{I} - \mathbf{P})^2\mathbf{Y} = \mathbf{Y}^T(\mathbf{I} - \mathbf{P})\mathbf{Y}. \end{aligned}$$

Since $\mathbf{I} - \mathbf{P}$ is a projection of rank $n - p$, SSE has $\sigma^2\chi^2(n-p)$-distribution, see also the Fisher–Cochran theorem of the next lesson.

To make inference on the significance of the regression, we will intensively use the sample counterpart of the variance decomposition (6):

$$SST = SSR + SSE = R^2 \cdot SST + (1 - R^2) \cdot SST,$$

where $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is the total variation of the measurements (sum of squares total),

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} \hat{a}_j x_{ij})^2$$

is the *residual sum of squares* (sum of squares due to error), and $SSR = SST - SSE$ is the part of the total variation explained by the regression (sum of squares due to regression). Further, R is the sample estimate of the multiple correlation coefficient.

To investigate the quality of the regression, we pose the alternative

$$H_0 : \mathbf{a} = \mathbf{0} \quad \text{versus} \quad H_1 : \mathbf{a} \neq \mathbf{0}.$$

Under H_0, SSR has $\sigma^2\chi(p)$-distribution, and independent of SSE (see also the ANOVA setup of Lesson 7). Therefore,

$$F = \frac{SSR/p}{SSE/(n-p)} = \frac{R^2}{1-R^2} \cdot \frac{n-p}{p} \sim \mathcal{F}(p, n-p) \tag{7}$$

has Fisher F-distribution with degrees of freedom p and $n-p$ (in fact, $n-p-1$ if there is a constant term as well). If this $F \geq F_\alpha(p, n-p)$ (the upper α-point, or equivalently, the $(1-\alpha)$-quantile value) of this F-distribution, then we reject H_0 with significance α. This means that the regression is significant, and it makes sense to approximate the target variable with the predictors.

When we reject the null-hypothesis, we may further investigate whether the coefficients a_j's significantly differ from zero. For $j = 1, \dots, p$ we investigate the alternative

$$H_{0j} : a_j = 0 \quad \text{versus} \quad H_{1j} : a_j \neq 0.$$

Under H_{0j}, $\hat{a}_j$ has zero expectation, and standardizing by its standard error

$$s_j = \sqrt{\frac{SSE/(n-p)}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

(which is based on Proposition 2), then under H_{0j}, the test statistic

$$t_j = \frac{\hat{a}_j - 0}{s_j} \sim t(n-p)$$

follows Student's t-distribution with degrees of freedom $n-p$; in fact, $n-p-1$ if there is a constant term (intercept) as well in the model.

If this $t_j \geq t_{\alpha/2}(n-p)$ (the upper $\alpha/2$-point, or equivalently, the $(1-\alpha/2)$-quantile value) of this t-distribution, then we reject H_{0j} with significance α, and conclude that the predictor variable j significantly influences the response.

AN IMPORTANT CONSEQUENCE OF THE GAUSS-MARKOV THEOREM

An easy consequence of the Gauss-Markov theorem is the following. Sometimes this is called Gauss–Markov theorem.

Proposition: If $r = p$, then for any $\mathbf{b} \in \mathbb{R}^p$, the statistic $\mathbf{b}^T \hat{\mathbf{a}}$ is a linear unbiased estimate of the univariate parameter function $\mathbf{b}^T \mathbf{a}$, and it has minimum variance among all of such (linear, unbiased) estimates of $\mathbf{b}^T \mathbf{a}$ (BLUE).
Proof. Unbiasedness is obvious; for any $\mathbf{b} \in \mathbb{R}^p$:

$$\mathrm{Var}(\mathbf{b}^T \hat{\mathbf{a}}) = \mathbf{b}^T \mathrm{Var}(\hat{\mathbf{a}})\mathbf{b} \quad \text{and} \quad \mathrm{Var}(\mathbf{b}^T \tilde{\mathbf{a}}) = \mathbf{b}^T \mathrm{Var}(\tilde{\mathbf{a}})\mathbf{b}.$$

In view of the Gauss–Markov theorem, $\mathrm{Var}(\tilde{\mathbf{a}}) - \mathrm{Var}(\hat{\mathbf{a}})$ is a positive semidefinite matrix, which means that for any vector $\mathbf{b} \in \mathbb{R}^p$:

$$\mathrm{Var}(\mathbf{b}^T \tilde{\mathbf{a}}) - \mathrm{Var}(\mathbf{b}^T \hat{\mathbf{a}}) \geq 0.$$

Definition The parameter function $\mathbf{b}^T \mathbf{a}$ (with $\mathbf{b} \in \mathbb{R}^p$) is said to be *estimable* (in a linear and unbiased way) if there exists a vector $\mathbf{c} \in \mathbb{R}^n$ such that $\mathbb{E}(\mathbf{c}^T \mathbf{Y}) = \mathbf{b}^T \mathbf{a}$.
Proposition The parameter function $\mathbf{b}^T \mathbf{a}$ is estimable if and only if $\mathbf{b}$ is within the linear subspace of $\mathbb{R}^p$ spanned by the row vectors of $\mathbf{X}$.
Proof. The following are equivalent steps:

$$\begin{aligned} \mathbf{c}^T \mathbb{E}(\mathbf{Y}) &= \mathbf{b}^T \mathbf{a} \qquad (\forall\, \mathbf{a} \in \mathbb{R}^p)\,, \\ \mathbf{c}^T \mathbf{X}\mathbf{a} &= \mathbf{b}^T \mathbf{a} \qquad (\forall\, \mathbf{a} \in \mathbb{R}^p)\,, \\ \mathbf{c}^T \mathbf{X} &= \mathbf{b}^T\,, \\ \mathbf{b} &= \mathbf{X}^T \mathbf{c}, \end{aligned}$$

which means that the vector $\mathbf{b}$ is within the subspace spanned by the column vectors of $\mathbf{X}^T$, which is the same as the subspace spanned by the row vectors of $\mathbf{X}$.

If $r = p$, this holds for any $\mathbf{b} \in \mathbb{R}^p$; therefore, any parameter function $\mathbf{b}^T \mathbf{a}$ is estimable. However, if $r < p$, then the BLUE estimate of the first proposition can be extended only for an estimable $\mathbf{b}^T \mathbf{a}$.

Only for the Multivariate Statistics course

Prove that the above test based on the F statistic (7) is a likelihood ratio test.