# Statistics 3: Linear Models and Linear Regression

**Prerequisites:** Statistics I and II; Probability I and II; Linear/Matrix Algebra; Proficiency in using R Statistical Software

**References**


1. CR Rao: *Linear Statistical Inference and its Applications*, Wiley (1973)
2. SR Searle: *Linear Models*, Wiley (1971)
3. R Christensen: *Linear Models*, Marcel-Dekker (1983)
4. R Rao and P Bhimasankaram: *Linear Algebra*, 2nd edition, Hindustan Book Agency (2000)
5. R Bapat: *Linear Algebra and Linear Models*, Hindustan Book Agency (1999)


**Grading (Tentative):** 20 marks for assignments; 20 marks each for two class tests; 40 marks for the final exam

**Lectures:** Online lectures will be on Zoom; Time: Monday, Wednesday, Friday, 2-3 pm

**Lecture Notes:** Lecture notes will be posted on Moodle. Attempt will be made to post recording of lectures on Moodle too

**Assignments:** Assignments will also be posted on Moodle. Answers to these must be submitted by uploading to Moodle.

For any contingency, you may contact me at


*e-mail: mohan.delampady@gmail.com*

*mobile/whatsapp: 9880127065*

# Linear Models

It is of interest to see if a nice relationship exists between two random variables, $X$ and $Y$. Eventual objective may be either prediction of a future value or utilization of the relaionship for understanding the structure.

**Ex.** $X$ = height, $Y$ = weight of indviduals. One may ask: is there an optimal weight for a given height?
Data: $(x_i, y_i)$, observations from $n$ randomly chosen individuals, $i = 1, 2, \ldots, n$.

**Ex.** $X$ = temperature, $Y$ = pressure of a certain volume of gas.
Data: $(x_i, y_i)$, $i = 1, 2, \ldots, n$ from a *controlled experiment* where a certain volume of gas is *subjected to different temperatures* and the resulting pressure is measured.

**Ex.** In a biological assay, $Y$ = response corresponding to a dosage level of $X = x$. Again, $(x_i, y_i)$, $i = 1, 2, \ldots, n$ from $n$ laboratory subjects.

**Ex.** In an agricultural experiment, $y$ is the yield of a crop. A piece of land is divided into $I$ plots according to soil fertility; $J$ different fertilizer levels are also used. Then, if $y_{ij}$ is the yield from the $i$th plot receiving $j$th level of fertilizer, we might like to try the model:
$y_{ij} = \mu + \alpha_i + \tau_j + \epsilon_{ij}$. Why do we need $\epsilon_{ij}$? It is a random error (measurement error, noise or uncontrolled variability) needed to explain the variation in the model, which is needed in each of the other cases as well.

In general,

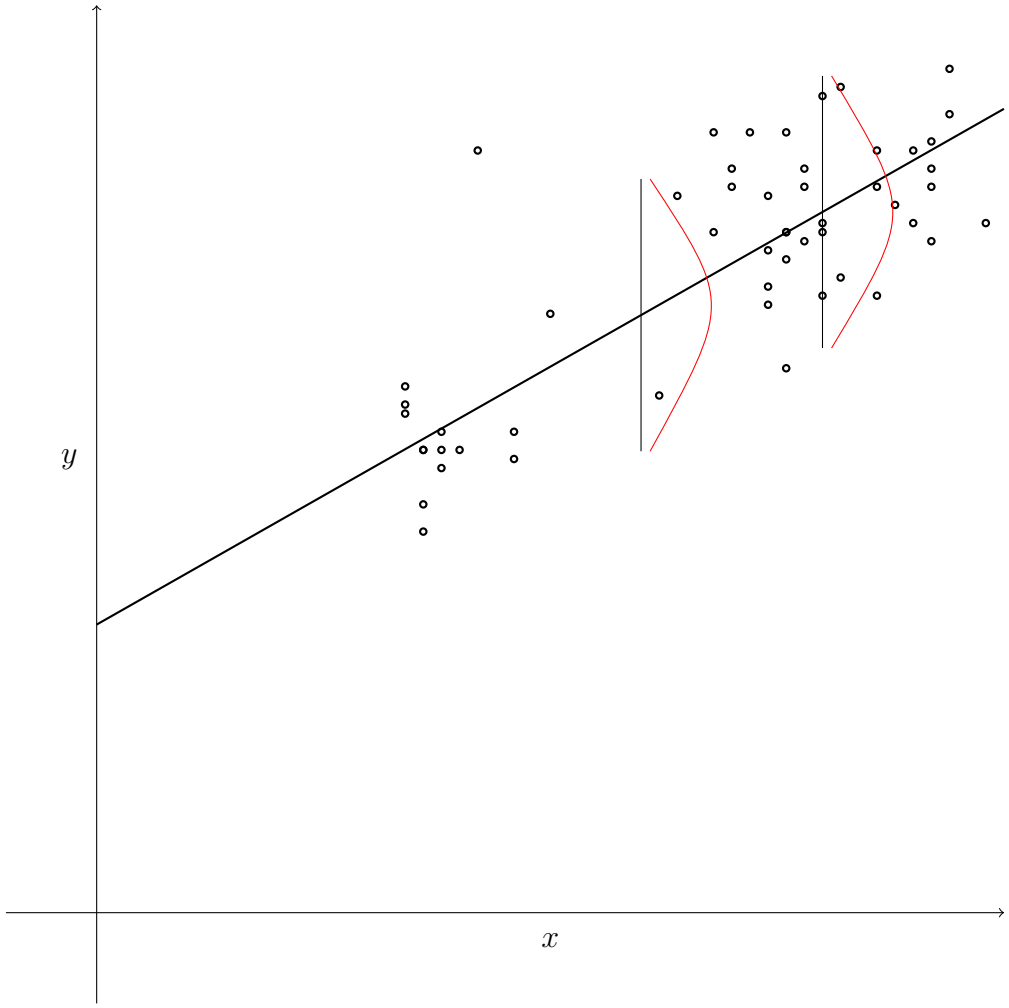
$$y_i = \alpha + \beta x_i + \epsilon_i, \tag{1}$$

where $y$ is the response variable and $x$ is the predictor variable, and $\alpha$ and $\beta$ are unknown coefficients is called a linear model. Here 'linear' stands for linear space, linear or additive in the coefficients and not for linear in $x$, as will be seen later. Equation (1) expresses the linear or additive relationship between $E(Y|X = x)$ and the influencing factors.

Observe the following data and the scatter plot of $y$ versus $x$, where $x$ = duration and $y$ = interval (both in minutes) for eruptions of Old Faithful Geyser.

| $x$ | $y$ | $x$ | $y$ | $x$ | $y$ | $x$ | $y$ | $x$ | $y$ | $x$ | $y$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 4.4 | 78 | 3.9 | 74 | 4.0 | 68 | 4.0 | 76 | 3.5 | 80 | 4.1 | 84 |
| 2.3 | 50 | 4.7 | 93 | 1.7 | 55 | 4.9 | 76 | 1.7 | 58 | 4.6 | 74 |
| 3.4 | 75 | 4.3 | 80 | 1.7 | 56 | 3.9 | 80 | 3.7 | 69 | 3.1 | 57 |
| 4.0 | 90 | 1.8 | 42 | 4.1 | 91 | 1.8 | 51 | 3.2 | 79 | 1.9 | 53 |
| 4.6 | 82 | 2.0 | 51 | 4.5 | 76 | 3.9 | 82 | 4.3 | 84 | 2.3 | 53 |
| 3.8 | 86 | 1.9 | 51 | 4.6 | 85 | 1.8 | 45 | 4.7 | 88 | 1.8 | 51 |
| 4.6 | 80 | 1.9 | 49 | 3.5 | 82 | 4.0 | 75 | 3.7 | 73 | 3.7 | 67 |
| 4.3 | 68 | 3.6 | 86 | 3.8 | 72 | 3.8 | 75 | 3.8 | 75 | 2.5 | 66 |
| 4.5 | 84 | 4.1 | 70 | 3.7 | 79 | 3.8 | 60 | 3.4 | 86 | | |

Table 1: Eruptions of Old Faithful Geyser, August 1 – 4, 1978




(1) is a linear model for $E(y|x)$, so $\epsilon$ denotes the spread or dispersion around this line. i.e., $y = E(y|x) + \epsilon$. If we let $g(x) = E(y|x)$, assuming $g$ to be smooth, we could consider the approximation:

$$\begin{aligned} g(x) &= g(0) + g'(0)x + \frac{g''(0)}{2!}x^2 + \ldots + \frac{g^{(k)}(0)}{k!}x^k \\ &= \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k. \end{aligned}$$

This is linear in the coefficients $\beta_0, \beta_1, \ldots$ but not in $x$. Also, recall Weirstrass theorem on being able to uniformly approximate by polynomials any contiuous function on a closed interval. Thus, on a reasonable range of $x$ values, such a 'linear' approximation may be quite acceptable. More importantly, special tools and techniques from linear spaces and linear algebra are available for studying linear models.

## MULTIPLE LINEAR REGRESSION MODEL

The response $y$ is often influenced by more than one predictor variable. For example, the yield of a crop may depend on the amount of nitrogen, potash, and phosphate fertilizers used. These variables are controlled by the experimenter, but the yield may also depend on uncontrollable variables such as those associated with weather. A linear model relating the response y to several predictors has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} + \epsilon. \tag{2}$$

The parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are called regression coefficients. The presence of $\epsilon$ provides for random variation in $y$ not explained by the $x$ variables. This random variation may be due partly to other variables that affect $y$ but are not known or not observed. The model in (2) is linear in the $\beta$ parameters; it is not necessarily linear in the $x$ variables. Thus models such as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_3 + \beta_4 \sin(x_2) + \epsilon$$

are included in the designation linear model. A model provides a theoretical framework for better understanding of a phenomenon of interest. Thus a model is a mathematical construct that we believe may represent the mechanism that generated the observations at hand. The postulated model may be an idealized oversimplification of the complex real-world situation, but in many such cases, empirical models provide useful approximations of the relationships among variables. These relationships may be either associative or causative.

Regression models such as (2) are used for various purposes, including the following:

**Prediction.** Estimates of the individual parameters $\beta_0, \beta_1, \ldots$ are of less importance for prediction than the overall influence of the x variables on y. However, good estimates are needed to achieve good prediction performance.

**Data Description or Explanation.** The scientist or engineer uses the estimated model to summarize or describe the observed data.

**Parameter Estimation.** The values of the estimated parameters may have theoretical implications for a postulated model.

**Variable Selection or Screening.** The emphasis is on determining the importance of each predictor variable in modeling the variation in y. The predictors that are associated with an important amount of variation in y are retained; hose that contribute little are deleted.

**Control of Output.** A cause-and-effect relationship between y and the x variables is assumed. The estimated model might then be used to control the output of a process by varying the inputs. By systematic experimentation, it may be possible to achieve the optimal output.

There is a fundamental difference between purposes 1 and 5. For prediction, we need only assume that the same correlations that prevailed when the data were collected also continue in place when the predictions are to be made. Showing that there is a significant relationship between $y$ and the $x$ variables in (2) does not necessarily prove that the relationship is causal. To establish causality in order to control output, the researcher must choose the values of the $x$ variables in the model and use randomization to avoid the effects of other possible variables unaccounted for. In other words, to ascertain the effect of the $x$ variables on $y$ when the $x$ variables are changed, it is necessary to change them.

**Vector-matrix form of linear model.**

Data is of the form: $(y_i, \mathbf{x}_i)$, $i = 1, 2, \ldots, n$, $\mathbf{x}_i = (x_{i0} = 1, x_{i1}, \ldots, x_{i(p-1)})'$. The linear model is:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i(p-1)} + \epsilon_i, \\ &= \sum_{j=0}^{p-1} \beta_j x_{ij} + \epsilon_i, i = 1, 2, \ldots, n; x_{i0} = 1. \end{aligned}$$

Equivalently,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1(p-1)} \\ 1 & x_{21} & \ldots & x_{2(p-1)} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & \ldots & x_{n(p-1)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \text{ or}$$

$$\mathbf{y} = X\beta + \epsilon.$$

$\mathbf{y}_{n\times 1}$ is the response vector, $X_{n\times p}$ is the matrix of predictors or covariates, $\beta_{p\times 1}$ is the vector of regression coefficients, and $\epsilon$ is random noise. $\mathbf{y}$ is random since $\epsilon$ is random. $X$ is treated as a fixed matrix and $\beta$ is a fixed but unknown vector of parameters. Note that the model involves random vectors and matrices, so some preliminaries on these are needed before we can proceed further.

## Multivariate Distributions

A random vector $T$ is a vector whose elements have a joint distribution. i.e., if $(\Omega, \mathcal{A}, P)$ is a probability space, $T_{p\times 1} : \Omega \to \mathcal{R}^p$ is such that $T^{-1}(B) \in \mathcal{A}$, and hence $Pr(T \in B) = P(T^{-1}(B))$.

Thus, $\mathbf{X} = (X_1, \ldots, X_p)'$ is a random vector if $X_i$'s are random variables with a joint distribution. If the joint density exists, we have $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathcal{R}^p$ such that

$$\int_{\mathcal{R}^p} f(\mathbf{x})\, d\mathbf{x} = 1 \text{ and } P(\mathbf{X} \in A) = \int_A f(\mathbf{x})\, d\mathbf{x}, \qquad A \subset \mathcal{R}^p.$$

**Example.** $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$, $-1 < \rho < 1$, if

$$f(x_1, x_2) = \frac{1}{2\pi} \frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{ \left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) \right\}}.$$

Check that $E(X_i) = \mu_i$, $Var(X_i) = \sigma_i^2$, $i = 1, 2$ and $Cov(X_1, X_2) = \rho\sigma_1\sigma_2$.

**Example.** $\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim$ Uniform on unit ball if

$$f(x_1, x_2, x_3) = \begin{cases} \frac{3}{4\pi} & \text{if } x_1^2 + x_2^2 + x_3^2 \leq 1; \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathbf{X} = (X_1, \ldots, X_p)'$ be a random vector and assume $\mu_i = E(X_i)$ exists for all $i$. Then define $E(\mathbf{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$ as the mean vector of $\mathbf{X}$. A random matrix $Z_{p\times q} = ((z_{ij}))$ is a matrix whose elements are jointly distributed random variables. If $G(Z)$ is a matrix valued function of $Z$, then $E(G(Z)) = ((E(G_{ij}(Z))))$.
If $G(Z) = AZB$, where $A$ and $B$ are constant matrices, $E(G(Z)) = AE(Z)B$.
If $(Z, T)$ has a joint distribution, and $A, B, C, D$ are constant matrices, $E(AZB + CTD) = AE(Z)B + CE(T)D$.
If $Z$ is symmetric and positive semi-definite (nnd) with probability 1, $E(Z)$ is also symmetric and positive semi-definite. i.e., show $a'E(Z)a \geq 0$ for all $a$. Note that $a'E(Z)a = E(a'Za) \geq 0$, since for all $a$, $a'Za \geq 0$ wp 1.

Suppose $Z_{p\times p}$ is p.s.d. with wp 1. Then its spectral decomposition gives $Z = \Gamma D_\lambda \Gamma'$, where $\Gamma$ is orthogonal and $D_\lambda$ is diagonal. Let $\lambda_i(Z) = i$th diagonal element of $D_\lambda$, $\lambda_1(Z) \geq \lambda_2(Z) \geq \ldots \geq \lambda_p(Z) \geq 0$ wp 1. What about $E(Z)$? Is $\lambda_i(E(Z)) = E(\lambda_i(Z))$? No. However, $E(Z)$ is p.s.d., so $\lambda_i(E(Z)) \geq 0$.

Suppose $X_{p\times 1}$ has mean $\mu$ and also $E[(X_i-\mu_i)(X_j-\mu_j)] = Cov(X_i, X_j) = \sigma_{ij}$ exists for all $i, j$. i.e., $\sigma_{ii} < \infty$ for all $i$. Then the covariance matrix (or the variance-covariance matrix or the dispersion matrix) of $X$ is defined as

$$Cov(X) = \Sigma = E\left[(X-\mu)(X-\mu)'\right] = ((E[(X_i-\mu_i)(X_j-\mu_j)])) = ((\sigma_{ij})).$$

$\Sigma$ is symmetric, $\sigma_{ii} = Var(X_i) \geq 0$ and $\Sigma$ is p.s.d.

**Theorem.** $\Sigma_{p\times p}$ is a covariance matrix (of some $X$) iff $\Sigma$ is symmetric p.s.d.

**Proof.** (i) If $\Sigma = Cov(X)$ for some $X$ and $E(X) = \mu$, then for any $\alpha \in \mathcal{R}^p$,

$$\begin{aligned}
\alpha'\Sigma\alpha &= \alpha' Cov(X)\alpha = \alpha' E\left[(X-\mu)(X-\mu)'\right]\alpha \\
&= E\left[\alpha'(X-\mu)(X-\mu)'\alpha\right] = E\left[\left\{\alpha'(X-\mu)\right\}^2\right] \\
&= E\left[(\alpha'X - \alpha'\mu)^2\right] = Var(\alpha'X) \geq 0,
\end{aligned}$$

so $\Sigma$ is p.s.d. It is actually p.d. unless there exists $\alpha \neq 0$ such that $Var(\alpha'X) = 0$ (i.e., $\alpha'X = c$ w.p.1)

(ii) Now suppose $\Sigma$ is any symmetric p.s.d matrix of rank $r \leq p$. Then $\Sigma = CC'$, $C_{p\times r}$ of rank $r$. Let $Y_1, \ldots, Y_r$ be i.i.d with $E(Y_i) = 0$, $Var(Y_i) = 1$. Let $Y = (Y_1, \ldots, Y_r)'$. Then $E(Y) = 0$, $Cov(Y) = I_r$. Let $X = CY$. Then $E(X) = 0$ and
$Cov(X) = E(XX') = E(CYY'C') = CE(YY')C' = CC' = \Sigma$.

For $a \neq 0$, $a'Cov(X)a = 0$ iff $Cov(X)a = 0$, or $Cov(X)$ has a zero eigen value.

If $X_{p\times 1}$ and $Y_{q\times 1}$ are jointly distributed with finite second moments for their elements, and with $E(X) = \mu$, $E(Y) = \nu$, then
$Cov(X_{p\times 1}, Y_{q\times 1}) = ((Cov(X_i, Y_j)))_{p\times q} = ((E(X_i-\mu_i)(Y_j-\nu_j))) = ((E(X_iY_j)-\mu_i\nu_j)) = E(XY') - \mu\nu' = E[(X-E(X))(Y-E(Y))']$.
$Cov(X) = Cov(X, X) = E[(X-E(X))(X-E(X))'] = E(XX') - E(X)(E(X))'$.
$Cov(AX, BY) = A\,Cov(X, Y)B'$,
$Cov(AX) = Cov(AX, AX) = A\,Cov(X, X)A' = A\,Cov(X)A'$.

Consider $X_{p\times 1} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ and $Y_{q\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$. Then

$$\begin{aligned} Cov(X,Y) &= \begin{pmatrix} Cov(X_1,Y_1) & Cov(X_1,Y_2) \\ Cov(X_2,Y_1) & Cov(X_2,Y_2) \end{pmatrix} \\ &\neq Cov(Y,X) = \begin{pmatrix} Cov(Y_1,X_1) & Cov(Y_1,X_2) \\ Cov(Y_2,X_1) & Cov(Y_2,X_2) \end{pmatrix} \end{aligned}$$

in general. Further, note,

$$\begin{aligned} Cov(X+Y) &= Cov(X+Y, X+Y) \\ &= Cov(X,X) + Cov(X,Y) + Cov(Y,X) + Cov(Y,Y) \\ &= Cov(X) + Cov(Y) + Cov(X,Y) + Cov(X,Y)' \\ &\neq Cov(X) + Cov(Y) + 2Cov(X,Y), \end{aligned}$$

in general. If $X$ and $Y$ are independent, we do have, $Cov(X,Y) = ((Cov(X_i,Y_j))) = 0$ since $Cov(X_i,Y_j) = 0$ for all $i$ and $j$.

**Quadratic Forms.**

$X'AX$ is called a quadratic form of $X$. Note that
$E(X'AX) = E[tr(X'AX)] = E[tr(AXX')] = tr[E(AXX')] = tr[AE(XX')] = tr[A(\Sigma + \mu\mu')] = tr(A\Sigma) + tr(A\mu\mu') = tr(A\Sigma) + \mu' A\mu$, since $Cov(X) = \Sigma = E((X-\mu)(X-\mu)') = E(XX' - X\mu' - \mu X' + \mu\mu') = E(XX') - \mu\mu'$.

The moment generating function (mgf) of $X$ at $\alpha$ is defined as $\phi_X(\alpha) = E(\exp(\alpha' X))$. This uniquely determines the probability distribution of $X$. Note that $\phi_X((t_1, 0)')E(\exp(t_1 X_1)) = \phi_{X_1}(t_1)$. If $X$ and $Y$ are independent, $\phi_{X+Y}(t) = E(\exp(t'(X+Y))) = E(\exp(t'X)\exp(t'Y))$
$= E(\exp(t'X))E(\exp(t'Y)) = \phi_X(t)\phi_Y(t)$.

**Theorem (Cramer-Wold device).** If $X$ is a random vector, its probability distribution is completely determined by the distribution of all linear functions, $\alpha' X$, $\alpha \in \mathcal{R}^p$.

**Proof.** The mgf of $\alpha' X$, for any $\alpha \in \mathcal{R}^p$ is $\phi_{\alpha' X}(t) = E(t\alpha' X)$. Suppose this is known for all $\alpha \in \mathcal{R}^p$. Now, for any $\alpha$, note $\phi_X(\alpha) = E(\exp(\alpha' X) = \phi_{\alpha' X}(1)$, which is then known.

**Remark.** To define the joint multivariate distribution of a random vector, it is enough to specify the distribution of all its linear functions.

**Multivariate Normal Distribution**

**Definition.** $X_{p\times 1}$ is $p$-variate normal if for every $\alpha \in \mathcal{R}^p$, the distribution of $\alpha' X$ is univariate normal.

**Result.** If $X$ has the $p$-variate normal distribution, then both $\mu = E(X)$ and $\Sigma = Cov(X)$ exist and the distribution of $X$ is determined by $\mu$ and $\Sigma$.

**Proof.** Let $X = (X_1, \ldots, X_p)'$. Then for each $i$, $X_i = \alpha_i' X$ where $\alpha_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$. Therefore, $X_i = \alpha_i' X \sim N(.,.)$. Hence, $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_{ii}$ exist. Also, since $|\sigma_{ij}| = |Cov(X_i, X_j)| \leq \sqrt{\sigma_{ii}\sigma_{jj}}$, $\sigma_{ij}$ exists. Set $\mu = (\mu_1, \ldots, \mu_p)'$ and $\Sigma = ((\sigma_{ij}))$. Further, $E(\alpha' X) = \alpha'\mu$ and $Var(\alpha' X) = \alpha'\Sigma\alpha$, so

$$\alpha' X \sim N(\alpha'\mu, \alpha'\Sigma\alpha), \text{ for all } \alpha \in \mathcal{R}^p.$$

Since $\{\alpha' X, \alpha \in \mathcal{R}^p\}$ determine the distribution of $X$, $\mu$ and $\Sigma$ suffice.
Notation: $X \sim N_p(\mu, \Sigma)$.

**Result.** If $X \sim N_p(\mu, \Sigma)$, then for any $A_{k\times p}$, $b_{k\times 1}$,
$Y = AX + b \sim N_k(A\mu + b, A\Sigma A')$.

**Proof.** Consider linear functions, $\alpha' Y = \alpha' AX + \alpha' b = \beta' X + c$, which are univariate normal. Therefore $Y$ is $k$-variate normal. $E(Y) = A\mu + b$, $Cov(Y) = Cov(AX) = A\Sigma A'$.

**Theorem.** $X_{p\times 1} \sim N_p(\mu, \Sigma)$ iff $X_{p\times 1} = C_{p\times r} Z_{r\times 1} + \mu$ where $Z = (Z_1, \ldots, Z_r)'$, $Z_i$ i.i.d $N(0,1)$, $\Sigma = CC'$, $r = \text{rank}(\Sigma) = \text{rank}(C)$.

**Proof.** if part: If $X = CZ + \mu$ and $Z \sim N_r(0, I_r)$, then $X \sim N_p(\mu, CC' = \Sigma)$.

$Z$ is multivariate normal since linear functions of $Z$ are linear combinations of $Z_i$'s, which are univarite normal (as can be shown using the change of variable (jacobian) formula for joint densities, or using the mgf of normal).

Only if: If $X \sim\sim N_p(\mu, \Sigma)$, and $\text{rank}(\Sigma) = r \leq p$, then consider the spectral decomposition, $\Sigma = H\Delta H'$, $H$ orthogonal, $\Delta = \begin{pmatrix} \Delta_1 & 0 \\ 0 & 0 \end{pmatrix}$, $\Delta_1 = \text{diagonal}(\delta_1, \ldots, \delta_r)$, $\delta_i > 0$. Now, $X - \mu \sim N(0, \Sigma)$, and $H'(X - \mu) \sim N(0, \Delta)$. Let $H'(X - \mu) = \begin{pmatrix} Y_{r\times 1} \\ T_{(p-r)\times 1} \end{pmatrix}$. Then,

$$\begin{pmatrix} Y_{r\times 1} \\ T_{(p-r)\times 1} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Delta_1 & 0 \\ 0 & 0 \end{pmatrix}\right).$$

Therefore, $T = 0$ w.p. 1. Let $Z = \Delta_1^{-1/2} Y$. Then $Z \sim N_r(0, I_r)$. Therefore, w.p. 1, $H'(X - \mu) = \begin{pmatrix} \Delta_1^{1/2} Z \\ 0 \end{pmatrix}$. Further, w.p. 1,

$$X - \mu = H \begin{pmatrix} \Delta_1^{1/2} Z \\ 0 \end{pmatrix} = (H_1 | H_2) \begin{pmatrix} \Delta_1^{1/2} Z \\ 0 \end{pmatrix} = H_1 \Delta_1^{1/2} Z = CZ.$$

Also, $CC' = H_1 \Delta_1^{1/2} \Delta_1^{1/2} H_1' = H_1 \Delta_1 H_1'$ and

$$\Sigma = H\Delta H' = (H_1 | H_2) \begin{pmatrix} \Delta_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} H_1' \\ H_2' \end{pmatrix} = H_1 \Delta_1 H_1'.$$

Recall that if $Z_1 \sim N(0, 1)$, its mgf is $\phi_{Z_1}(t) = E(\exp(tZ_1)) = \exp(t^2/2)$. Therefore, if $Z \sim N_r(0, I_r)$ then

$$\phi_Z(u) = E(\exp(u'Z)) = E(\exp(\sum_{j=1}^{r} u_j Z_j)) = \exp(\sum_{j=1}^{r} u_j^2/2) = \exp(\frac{1}{2} u'u).$$

Then, if $X \sim\sim N_p(\mu, \Sigma)$, its mgf is:

$$\phi_X(t) = \exp(t'\mu + \frac{1}{2} t'\Sigma t),$$

since $E(\exp(t'X)) = E(\exp(t'(CZ + \mu))) = \exp(t'\mu) E(\exp(t'CZ)) = \exp(t'\mu) \exp(t'CC't/2) = \exp(t'\mu + t'\Sigma t/2)$.

**Marginal and Conditional Distributions**

**Theorem.** If $X \sim\sim N_p(\mu, \Sigma)$, then the marginal distribution of any subset of $k$ components of $X$ is $k$-variate normal.

**Proof.** Partition as follows:

$$X = \begin{pmatrix} X^{(1)}_{k\times 1} \\ X^{(2)}_{(p-k)\times 1} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu^{(1)}_{k\times 1} \\ \mu^{(2)}_{(p-k)\times 1} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22} \end{pmatrix}.$$

Note that $X^{(1)} = (I_k|0) \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \sim N\left(\mu^{(1)}, \Sigma_{11}\right)$. Since marginals (without independence) do not determine the joint distribution, the converse is not true.

**Example.** $Z \sim N(0,1)$ independent of $U$ which takes values 1 and -1 with equal probability. Then $Y = UZ \sim N(0,1)$ since

$$\begin{aligned} P(Y \le y) &= P(UZ \le y) \\ &= \frac{1}{2}P(Z \le y|U=1) + \frac{1}{2}P(-Z \le y|U=-1) \\ &= \frac{1}{2}\Phi(y) + \frac{1}{2}\Phi(y) = \Phi(y). \end{aligned}$$

Therefore, $(Z, Y)$ has a joint distribution under which the marginals are normal. However, it is not bivariate normal. Consider $Z + Y = Z + UZ = \begin{cases} 2Z & 1/2 \\ 0 & 1/2 \end{cases}$. Since $P(Z + Y = 0) = 1/2$ (i.e., a point mass at 0, and $Z + Y = 2Z \sim N(0,1)$ with probability $1/2$, it cannot be normally distributed.

**Result.** Let $X_{p\times 1} = \begin{pmatrix} X^{(1)}_{k\times 1} \\ X^{(2)}_{(p-k)\times 1} \end{pmatrix} \sim N_p\left(\begin{pmatrix} \mu^{(1)}_{k\times 1} \\ \mu^{(2)}_{(p-k)\times 1} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22} \end{pmatrix}\right)$. Then $X^{(1)}$ and $X^{(2)}$ are independent iff $\Sigma_{12} = 0$.

**Proof.** Only if: Independence implies that $Cov(X^{(1)}, X^{(2)}) = \Sigma_{12} = 0$.

If part: Suppose that $\Sigma_{12} = 0$. Then, note that

$$\begin{aligned}
& M_{(X^{(1)},X^{(2)})}(s_1, s_2) \\
&= E(\exp(s_1'X^{(1)} + s_2'X^{(2)}) = E(\exp\left(\begin{pmatrix} s_1 \\ s_2 \end{pmatrix}' X\right)) \\
&= E(\exp\left(\begin{pmatrix} s_1 \\ s_2 \end{pmatrix}' \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix} + \frac{1}{2} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}' \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}\right)) \\
&= \exp\left(s_1'\mu^{(1)} + s_2'\mu^{(2)} + \frac{1}{2}s_1'\Sigma_{11}s_1 + \frac{1}{2}s_2'\Sigma_{22}s_2 + s_1'\Sigma_{12}s_2\right) \\
&= \exp\left(s_1'\mu^{(1)} + \frac{1}{2}s_1'\Sigma_{11}s_1\right) \exp\left(s_2'\mu^{(2)} + \frac{1}{2}s_2'\Sigma_{22}s_2\right) \\
&= M_{X_1}(s_1)M_{X_2}(s_2),
\end{aligned}$$

for all $s_1$ and $s_2$ iff $\Sigma_{12} = 0$.

**Result.** Suppose $X \sim N_p(\mu, \Sigma)$ and let $U = AX$, $V = BX$. Then $U$ and $V$ are independent iff $Cov(U, V) = A\Sigma B' = 0$.

**Proof.** Same as above, since $\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} A \\ B \end{pmatrix} X \sim N(.,.)$.

**Theorem.** If $X \sim N_p(\mu, \Sigma)$ and $\Sigma$ is p.d. then

$$f_X(x) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right),\ x \in \mathcal{R}^p.$$

**Proof.** Let $\Sigma = CC'$ where $C = \Sigma^{1/2}$ is nonsingular. Then $X = CZ + \mu$, $Z \sim N(0, I_p)$. Since $Z_i$ are i.i.d $N(0, 1)$,

$$f_Z(z) = (2\pi)^{-p/2} \exp(-\frac{1}{2}\sum_{i=1}^{p} z_i^2) = (2\pi)^{-p/2} \exp(-\frac{1}{2}z'z).$$

Since $X = CZ + \mu$, $Z = C^{-1}(X - \mu)$. Jacobian of the transformation is $dz = |C|^{-1}\, dx = |\Sigma|^{-1}\, dx$. Therefore,

$$\begin{aligned}
f_X(x) &= (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)'(C')^{-1}C^{-1}(x-\mu)\right) \\
&= (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right).
\end{aligned}$$

**Note.** $f_X(x)$ is constant on the ellipsoid, $\{x : (x-\mu)'\Sigma^{-1}(x-\mu) = r^2\}$.

**Ex.** Check for $p = 2$ to see if the above results agree with those of the bivariate normal.

**Theorem.** Let $X \sim N_p(\mu, \Sigma)$, $\Sigma > 0$ (i.e., p.d.), and let

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $X_1$ and $\mu_1$ are of length $k$. Also, let $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Then $\Sigma_{11.2} > 0$ and,
(i) $X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \sim N_k(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11.2})$ and is independent of $X_2$;
(ii) The conditional distribution of $X_1$ given $X_2$ is
$N_k\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11.2}\right)$.

**Proof.** (i) Let $C = \begin{pmatrix} I_k & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{p-k} \end{pmatrix}$. Then

$$CX = \begin{pmatrix} X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \\ X_2 \end{pmatrix} \sim N_p\left(\begin{pmatrix} \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\ \mu_2 \end{pmatrix}, C\Sigma C'\right).$$

$$\begin{aligned} C\Sigma C' &= \begin{bmatrix} I_k & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{p-k} \end{bmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{bmatrix} I_k & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I_{p-k} \end{bmatrix} \\ &= \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{bmatrix} I_k & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I_{p-k} \end{bmatrix} = \begin{pmatrix} \Sigma_{11.2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}. \end{aligned}$$

Now, independence of $X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$ and $X_2$ follows from the fact that $Cov(X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2, X_2) = 0$.
(ii) Note that $X_1 = (X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2) + \Sigma_{12}\Sigma_{22}^{-1}X_2$. Therefore, from the independence of these two parts, $X_1|(X_2 = x_2) = \Sigma_{12}\Sigma_{22}^{-1}x_2 + (X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2) \sim N(\Sigma_{12}\Sigma_{22}^{-1}x_2 + \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11.2})$.

**Remark.** Under multivariate normality, the best regression is linear. If we want to predict $X_1$ based on $X_2$, the best predictor is $E(X_1|X_2)$, which is equal to $\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}x_2$. The prediction error, however, is independent of $X_2$.

**Quadratic Forms.**

Recall that, $Y'AY$ is called a quadratic form of $Y$ when $Y$ is a random vector.

**Result.** If $X \sim N_p(\mu, \Sigma)$, $\Sigma > 0$, then $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2_p$.

**Proof.** $Z = \Sigma^{-1/2}(X-\mu) \sim N_p(0, I_p)$. i.e., $Z_1, Z_2, \ldots, Z_p$ are i.i.d. $N(0,1)$. Therefore $Z'Z = \sum_{i=1}^p Z_i^2 \sim \chi^2_p$. Note that $(X-\mu)'\Sigma^{-1}(X-\mu) = Z'Z$.

**Result.** If $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, then $\bar{X}$ and $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$ are independent, $\bar{X} \sim N(\mu, \sigma^2/n)$ and $S^2/\sigma^2 \sim \chi^2_{n-1}$.

**Proof.** First note that $X = (X_1, X_2, \ldots, X_n)' \sim N_n(\mu \mathbf{1}, \sigma^2 I_n)$. Now consider an orthogonal matrix $A_{n\times n} = ((a_{ij}))$ with the first row being $\mathbf{a}_1' = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}) = \frac{1}{\sqrt{n}}\mathbf{1}'$. (Simply consider a basis for $\mathcal{R}^n$ with $\mathbf{a}_1$ as the first vector, orthogonalize the rest.) Now let $Y = AX$. i.e., $Y_i = a_i'X$, $i = 1, 2, \ldots, n$. Since $X \sim N_n(\mu\mathbf{1}, \sigma^2 I_n)$, we have that $Y \sim N_n(\mu A\mathbf{1}, \sigma^2 AA') = N_n(\mu A\mathbf{1}, \sigma^2 I_n)$. Therfeore, $Y_i$ are independent normal with variance $\sigma^2$. Further, $E(Y_i) = E(a_i'X) = \mu a_i'\mathbf{1}$. Thus, $E(Y_1) = \mu a_1'\mathbf{1} = \mu\frac{1}{\sqrt{n}}\mathbf{1}'\mathbf{1} = \sqrt{n}\mu$. For $i > 1$, $E(Y_i) = \mu a_i'\mathbf{1} = \mu\sqrt{n}a_i'a_1 = 0$. i.e., $Y_2, \ldots, Y_n$ are i.i.d $N(0, \sigma^2)$. Therefore, $\sum_{i=2}^n Y_i^2 \sim \chi^2_{n-1}$. Further, $Y_1 = a_1'X = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i = \sqrt{n}\bar{X} \sim N(\sqrt{n}\mu, \sigma^2)$ and is independent of $(Y_2, \ldots, Y_n)$. Also, $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = X'X - Y_1^2 = Y'Y - Y_1^2 = \sum_{i=2}^n Y_i^2 \sim \chi^2_{n-1}$ which is independent of $Y_1$, and therefore of $\bar{X}$.

If $X \sim N_p(0, I)$, then $X'X = \sum_{i=1}^p X_i^2 \sim \chi^2_p$. i.e., $X'IX \sim \chi^2_p$. Also, note $X'(\frac{1}{\sqrt{p}}\mathbf{1}\frac{1}{\sqrt{p}}\mathbf{1}')X = p\bar{X}^2 \sim \chi^2_1$ and $X'(I - \frac{1}{p}\mathbf{1}\mathbf{1}')X \sim \chi^2_{p-1}$.

What is the distribution of $X'AX$ for any arbitrary $A$ which is p.s.d.? Without loss of generality we can assume that $A$ is symmetric since

$$X'AX = X'(\frac{1}{2}(A+A'))X = X'BX, \text{ where } B = \frac{1}{2}(A+A') \text{ is always symmetric.}$$

Since $A$ is symmetric p.s.d., $A = \Gamma D_\lambda \Gamma'$, so $X'AX = X'\Gamma D_\lambda \Gamma' X = Y' D_\lambda Y$, where $Y = \Gamma' X \sim N_p(0, \Gamma'\Gamma = I)$. Therefore $X'AX = \sum_{i=1}^p d_i Y_i^2$, where $d_i$ are eigen values of $A$ and $Y_i$ are i.i.d $N(0,1)$. Therefore $X'AX$ has the $\chi^2$ distribution if $d_i = 1$ or 0. Equivalently, $X'AX \sim \chi^2$ if $A^2 = A$ or $A$ is symmetric idempotent or $A$ is an orthogonal projection matrix. The equivalence may be seen as follows. If $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$ are such that

$d_1 = d_2 = \ldots = d_r = 1$ and $d_{r+1} = \ldots = d_p = 0$, then

$$\begin{aligned} A &= \Gamma \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \Gamma', \\ A^2 &= \Gamma \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \Gamma' \Gamma \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \Gamma' = A. \end{aligned}$$

If $A^2 = A$ then $\Gamma D_\lambda \Gamma' \Gamma D_\lambda \Gamma' = \Gamma D_\lambda^2 \Gamma' = \Gamma D_\lambda \Gamma'$ implies that $D_\lambda^2 = D_\lambda$, or that $d_i^2 = d_i$, or that $d_i = 0$ or 1.

We will show the converse now. Suppose $X'AX \sim \chi_r^2$ and $A$ is symmetric p.s.d. Then the mgf of $X'AX$ is:

$$\begin{aligned} M_{X'AX}(t) &= \int_0^\infty \exp(tu) \frac{\exp(-u/2) u^{r/2-1}}{2^{r/2}\Gamma(r/2)}\, du \\ &= \int_0^\infty \frac{\exp(-\frac{u}{2}(1-2t)) u^{r/2-1}}{2^{r/2}\Gamma(r/2)}\, du \\ &= (1-2t)^{-r/2}, \text{ for } 1-2t > 0. \end{aligned}$$

But in distribution, $X'AX = \sum_{i=1}^p d_i Y_i^2$, $Y_i$ i.i.d. $N(0,1)$, so

$$\begin{aligned} M_{X'AX}(t) &= E\left[\exp(t \sum_{i=1}^p d_i Y_i^2)\right] = E\left[\prod_{i=1}^p \exp(t d_i Y_i^2)\right] \\ &= \prod_{i=1}^p E\left[\exp(t d_i Y_i^2)\right] = \prod_{i=1}^p (1-2td_i)^{-1/2}, \text{ for } 1-2td_i > 0. \end{aligned}$$

Now note that $X'AX \sim \chi_r^2$ implies $X'AX > 0$ wp 1. i.e., $\sum_{i=1}^p d_i Y_i^2 > 0$ wp 1, which in turn imples that $d_i \geq 0$ for all $i$. (This is because, if $d_l < 0$, since $Y_l^2 \sim \chi_1^2$ independently of $Y_i$, $i \neq l$, we would have $\sum_{i=1}^p d_i Y_i^2 < 0$ with positive probability.) Therefore, for $t < \min_i \frac{1}{2d_i}$, equating the two mgf's, we have $(1-2t)^{-r/2} = \prod_{i=1}^p (1-2td_i)^{-1/2}$, or $(1-2t)^{r/2} = \prod_{i=1}^p (1-2td_i)^{1/2}$, or $(1-2t)^r = \prod_{i=1}^p (1-2td_i)$. Equality of two polynomials mean that their roots must be the same. Check that $r$ of the $d_i$'s must be 1 and rest 0. Thus the following result follows.

**Result.** $X'AX \sim \chi_r^2$ iff $A$ is a symmetric idempotent matrix or an orthogonal projection matrix of rank $r$.

**Result.** Suppose $Y \sim N_p(0, I_p)$ and let $Y'Y = Y'AY + Y'BY$. If $Y'AY \sim \chi^2_r$, then $Y'BY \sim \chi^2_{p-r}$ independent of $Y'AY$.

**Proof.** Note that $Y'Y \sim \chi^2_p$. Since $Y'AY \sim \chi^2_r$, $A$ is symmetric idempotent of rank $r$. Therefore, $B = I - A$ is symmetric and $B^2 = (I - A)^2 = I - 2A + A^2 = I - A = B$, so that $B$ is idempotent also. Further, Rank($B$) = trace($B$) = trace($I - A$) = $p - r$. Therefore, $Y'BY \sim \chi^2_{p-r}$. Independence is shown later.

**Result.** Let $Y \sim N_p(0, I_p)$ and let $Q_1 = Y'P_1Y$, $Q_2 = Y'P_2Y$, $Q_1 \sim \chi^2_r$, and $Q_2 \sim \chi^2_s$. Then $Q_1$ and $Q_2$ are independent iff $P_1P_2 = 0$.

**Corollary.** In the result before the above one, $A(I - A) = 0$, so $Y'AY$ and $Y'(I - A)Y$ are independent.

**Proof.** $P_1$ and $P_2$ are symmetric idempotent. If $P_1P_2 = 0$ then $Cov(P_1Y, P_2Y) = 0$ so that $Q_1 = (P_1Y)'(P_1Y) = Y'P_1^2Y = Y'P_1Y$ is independent of $Q_2 = (P_2Y)'(P_2Y) = Y'P_2Y$. Conversely, if $Q_1$ and $Q_2$ are independent $\chi^2_r$ and $\chi^2_s$, then $Q_1 + Q_2 \sim \chi^2_{r+s}$. Since $Q_1 + Q_2 = Y'(P_1 + P_2)Y$, $P_1 + P_2$ is symmetric idempotent. Hence, $P_1 + P_2 = (P_1 + P_2)^2 = P_1^2 + P_2^2 + P_1P_2 + P_2P_1$, implying $P_1P_2 + P_2P_1 = 0$. Multiplying by $P_1$ on the left, we get, $P_1^2P_2 + P_1P_2P_1 = P_1P_2 + P_1P_2P_1 = 0$ $(*)$. Similarly, multiplying by $P_1$ on the right yields, $P_1P_2P_1 + P_2P_1 = 0$. Subtracting, we get, $P_1P_2 - P_2P_1 = 0$. Combining this with $(*)$ above, we get $P_1P_2 = 0$.

**Result.** Let $Q_1 = Y'P_1Y$, $Q_2 = Y'P_2Y$, $Y \sim N_p(0, I_p)$. If $Q_1 \sim \chi^2_r$, $Q_2 \sim \chi^2_s$ and $Q_1 - Q_2 \geq 0$, then $Q_1 - Q_2$ and $Q_2$ are independent, $r \geq s$ and $Q_1 - Q_2 \sim \chi^2_{r-s}$.

**Proof.** $P_1^2 = P_1$ and $P_2^2 = P_2$ are symmetric idempotent. $Q_1 - Q_2 \geq 0$ means that $Y'(P_1 - P_2)Y \geq 0$, hence $P_1 - P_2$ is p.s.d. Therefore, from Lemma shown below, $P_1 - P_2$ is a projection matrix and also $P_1P_2 = P_2P_1 = P_2$. Thus $(P_1 - P_2)P_2 = 0$. Also, Rank($P_1 - P_2$) = tr($P_1 - P_2$) = tr($P_1$) - tr($P_2$) = Rank($P_1$) - Rank($P_2$) = $r - s$. Hence, $Q_1 - Q_2 = Y'(P_1 - P_2)Y \sim \chi^2_{r-s}$, and is independent of $Q_2 = Y'P_2Y \sim \chi^2_s$.

**Lemma.** If $P_1$ and $P_2$ are projection matrices such that $P_1 - P_2$ is p.s.d., then (a) $P_1P_2 = P_2P_1 = P_2$ and (b) $P_1 - P_2$ is also a projection matrix.

**Proof.** (a) If $P_1x = 0$, then $0 \leq x'(P_1 - P_2)x = -x'P_2x \leq 0$, implying $0 = x'P_2x = x'P_2^2x = (P_2x)'P_2x$, so $P_2x = 0$. Therefore, for any $y$, $P_2(I - P_1)y = 0$ since $P_1(I - P_1)y = 0$ (Take $x = (I - P_1)y$.) Thus, for any $y$, $P_2P_1y = P_2y$ or $P_2P_1 = P_2$, and so $P_2 = P_2' = (P_2P_1)' = P_1P_2$.
(b) $(P_1 - P_2)^2 = P_1^2 + P_2^2 - P_1P_2 - P_2P_1 = P_1 + P_2 - P_2 - P_2 = P_1 - P_2$.

**Result.** Any orthogonal projection matrix (i.e., symmetric idempotent) is p.s.d.

**Proof.** If $P$ is a projection matrix, $x'Px = x'P^2x = (Px)'Px \geq 0$.

**Result.** Let $C$ be a symmetric p.s.d. matrix. If $X \sim N_p(0, I_p)$, then $AX$ and $X'CX$ are independent iff $AC = 0$.

**Proof.** (i) If part: Since $C$ is symmetric p.s.d., $C = TT'$. If $AC = 0$, then $ATT' = 0$, so $ATT'A' = (AT)(AT)' = 0$ and hence $AT = 0$. Thus $AX$ and $T'X$ are independent, so $AX$ and $(T'X)(T'X)' = X'CX$ are independent.

(ii) Only if: If $AX$ and $X'CX$ are independent, then $X'A'AX$ and $X'CX$ are independent. But the mgf of $X'BX$ for any $B$ is $E(\exp(tX'BX)) = |I - 2tB|^{-1/2}$ for an interval of values of $t$. Therefore, the joint mgf of $X'CX$ and $X'A'AX$ is $|I - 2(t_1C + t_2A'A)|^{-1/2}$, but because of independence this is given to be equal to

$$|I - 2t_1C|^{-1/2}|I - 2t_2A'A|^{-1/2} = |I - 2t_1C - 2t_2A'A + 4t_1t_2CA'A|^{-1/2}.$$

Show that, for this to hold on an open set, we must have $CA'A = 0$, implying $CA'AC' = 0$, and thus $AC' = 0$. But $C' = C$.

**Lemma.** If $X \sim N_p(\mu, \Sigma)$, then $Cov(AX, X'CX) = 2A\Sigma C\mu$.

**Proof.** Note that $(X-\mu)'C(X-\mu) = X'CX + \mu'C\mu - 2X'C\mu = X'CX - 2((X-\mu)'C\mu - \mu'C\mu$ and $E(X'CX) = tr(C\Sigma) + \mu'C\mu$. Therefore $X'CX - E(X'CX) = X'CX - \mu'C\mu - tr(C\Sigma) = (X-\mu)'C(X-\mu) + 2(X-\mu)'C\mu - tr(C\Sigma)$. Hence,

$$\begin{aligned}
&Cov(AX, X'CX) \\
&\quad = E\left[(AX - A\mu)(X'CX - E(X'CX))\right] \\
&\quad = AE\left\{(X-\mu)\left[(X-\mu)'C(X-\mu) + 2(X-\mu)'C\mu - tr(C\Sigma)\right]\right\} \\
&\quad = 2AE\left\{(X-\mu)(X-\mu)'C\mu\right\} - tr(C\Sigma)AE(X-\mu) \\
&\qquad + AE\left\{(X-\mu)(X-\mu)'C(X-\mu)\right\} \\
&\quad = 2A\Sigma C\mu,
\end{aligned}$$

since $E(X-\mu) = 0$ and $E\left\{(X-\mu)(X-\mu)'C(X-\mu)\right\} =$
$E\left\{(X-\mu)\left[\sum_i \sum_j C_{ij}(X_i-\mu_i)(X_j-\mu_j)\right]\right\} = 0$. To prove this last equality, it is enough to show that $E\left\{(X_l-\mu_l)(X_i-\mu_i)(X_j-\mu_j)\right\} = 0$ for all $i, j, l$. For this note:
(i) if $i = j = l$, $E(X_i-\mu_i)^3 = 0$.
(ii) if $i = j \neq l$, $E\left\{(X_i-\mu_i)^2(X_l-\mu_l)\right\} = 0$ since $X_l - \mu_l = \frac{\sigma_{il}}{\sigma_{ll}}(X_i-\mu_i) + \epsilon$,

where $\epsilon \sim N(0, .)$ is independent of $X_i$, so this case reduces to (i).
(iii) if $i$, $j$ and $l$ are all different, the case reduces to (i) and (ii). Alternatively, consider $Y = (Y_1, Y_2, Y_3)' \sim N_3(0, \Sigma)$. Then $Y = \Sigma^{1/2}(Z_1, Z_2, Z_3)$, where $Z_i$ are i.i.d. $N(0, 1)$. Then to show that $E(Y_1 Y_2 Y_3) = 0$, simply note that $Y_1 Y_2 Y_3$ is a linear combination of $Z_i^3$, $Z_i^2 Z_j$ and $Z_1 Z_2 Z_3$, all of which have expectation 0.

**Loynes' Lemma.** If $B$ is symmetric idempotent, $Q$ is symmetric p.s.d. and $I - B - Q$ is p.s.d., then $BQ = QB = 0$.

**Proof.** Let $x$ be any vector and $y = Bx$. Then $y'By = y'B^2x = y'Bx = y'y$, so $y'(I-B-Q)y = -y'Qy \leq 0$. But $I-B-Q$ is p.s.d., so $y'(I-B-Q)y \geq 0$, implying $-y'Qy \geq 0$. Since $Q$ is also p.s.d., we must have $y'Qy = 0$. (Note, $y$ is not arbitrary, but $Bx$ for some $x$.) In addition, since $Q$ is symmetric p.s.d., $Q = L'L$ for some $L$, and hence $y'Qy = y'L'Ly = 0$, implying $Ly = 0$. Thus $L'Ly = Qy = QBx = 0$ for all $x$. Therefore, $QB = 0$ and hence $(QB)' = B'Q' = BQ = 0$.

**Theorem.** Suppose $X_i$ are $n \times n$ symmetric matrices with rank $k_i$, $i = 1, 2, \ldots, p$. Let $X = \sum_{i=1}^{p} X_i$ have rank $k$. (It is symmetric.) Then, of the conditions
(a) $X_i$ idempotent for all $i$
(b) $X_i X_j = 0$, $i \neq j$
(c) $X$ idempotent
(d) $\sum_{i=1}^{p} k_i = k$,
it is true that
I. any two of (a), (b), and (c) imply all of (a), (b), (c) and (d)
II. (c) and (d) imply (a) and (b)
III. (c) and $\{X_1, \ldots, X_{p-1}$ idempotent, $X_p$ p.s.d.$\}$ imply that $X_p$ idempotent and hence (a), and therefore (b) and (d).

**Proof.** I (i): Show (a) and (c) imply (b) and (d). For this, note, given (c), $I - X$ is idempotent and hence p.s.d. Now, given (a), $X - X_i - X_j = \sum_{r \neq i,j} X_r$ is p.s.d, being the sum of p.s.d matrices. Therefore, $(I - X) + (X - X_i - X_j) = I - X_i - X_j$ is p.s.d., hence $X_i X_j = 0$ from Loynes' Lemma. i.e., (b). Also, given (c), $\text{Rank}(X) = \text{tr}(X) = \text{tr}(\sum X_i) = \sum \text{tr}(X_i) = \sum k_i$, if (a) is also given. i.e., (d).

(ii): Show (b) and (c) imply (a) and (d). Let $\lambda$ be an eigen value of $X_1$ and $u$ be the corresponding eigen vector. Then $X_1 u = \lambda u$. Either $\lambda = 0$, or, if $\lambda \neq 0$, $u = X_1 \frac{1}{\lambda} u$. Therefore, for $i \neq 1$, $X_i u = X_i X_1 \frac{1}{\lambda} u = 0$ given (b). Therefore, given (b), $Xu = X_1 u = \lambda u$, and so $\lambda$ is an eigen value of $X$. But given (c), $X$ is idempotent, and hence $\lambda = 0$ or 1. Therefore eigen values of $X_1$ are 0 or 1, or $X_1$ is idempotent. Similarly for the other $X_i$'s. i.e., (a).

(iii): (a) and (b) together imply (c). (Note that then they imply (d) also, since (a) and (c) give (d).) Given (b) and (a), $X^2 = (\sum X_i)^2 = \sum X_i^2 = \sum X_i = X$, which is (c).

II. Show (c) and (d) imply (a) and (b). Given (c), $I - X$ is idempotent and hence has rank $n - k$. Therefore rank of $X - I$ is also $n - k$. i.e., $X - I$ has $n - k$ linearly independent rows. i.e.,

$(X - I)x = 0$ has $n - k$ linearly independent equations. Further,
$X_2 x = 0$ has $k_2$ linearly independent equations,
$\vdots$
$X_p x = 0$ has $k_p$ linearly independent equations.

Therefore the maximum number of linearly independent equations in

$$\begin{pmatrix} X - I \\ X_2 \\ \vdots \\ X_p \end{pmatrix} x = 0 \qquad \text{is } n - k + k_2 + \ldots + k_p = n - k_1.$$

i.e., the dimension of the solution space is at least $n-(n-k_1) = k_1$. However, this space is exactly $X_1 x = x$ because the above equations reduce to that. Thus $X_1 x = 1x$ has at least $k_1$ linearly independent solutions, or 1 is an eigen value of $X_1$ with multiplicity at least $k_1$. But since the rank of $X_1$ is $k_1$, multiplicity must be exactly $k_1$. Also, the other eigen values must be 0. Therefore $X_1$ is idempotent. Similar argument for the other $X_i$'s. So, (a). Now combine it with (c) to get (b).

III. Given (c), $X$ is idempotent, so p.s.d. Therefore, $I - X$ is idempotent and hence p.s.d. If $X_1, \ldots, X_{p-1}$ are idempotent, hence p.s.d., and $X_p$ is also p.s.d., then $\sum_{r \neq i,j} X_r = X - X_i - X_j$ is p.s.d., so $(I - X) + (X - X_i - X_j) = I - X_i - X_j$ is p.s.d. Then $X_i X_j = 0$ from Loynes', giving (b). Now (b) and (c) give (a) and (d).

The above theorem in linear algebra translates into a powerful result called Fisher-Cochran theorem on the question of: when are quadratic forms independent $\chi^2$?

**Theorem.** Suppose $Y \sim N_n(0, I_n)$, $A_i$, $i = 1, \ldots, p$ are symmetric $n \times n$ matrices of rank $k_i$, and $A = \sum_{i=1}^{p} A_i$ is symmteric with rank $k$. Then
(i) $Y' A_i Y \sim \chi^2_{k_i}$, (ii) $Y' A_i Y$ are pairwise independent, and (iii) $Y' A Y \sim \chi^2_k$ iff
I. any two of (a) $A_i$ are idempotent for all $i$, (b) $A_i A_j = 0$, $i \neq j$, (c) $A$ is idempotent, are true, or
II. (c) is true and (d) $k = \sum_i k_i$, or
III. (c) is true and
(e) $A_1, \ldots, A_{p-1}$ are idempotent and $A_p$ is p.s.d. is true.

**Proof.** Follows from the previous theorem.

**Linear Models – Estimation**

Consider $y_i$ uncorrelated, $E(y_i) = \mu$, $Var(y_i) = \sigma^2$, $i = 1, 2, \ldots, n$. Estimate $\mu$. In the absence of distributional assumptions, an appealing approach is least squares. What is the estimate and what are its properties? Write the model as:
$y_i = \mu + \epsilon_i$, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, $Cov(\epsilon_i, \epsilon_j) = 0$, $i \neq j$. Find

$$\min_{\mu} \sum_{i=1}^{n} (y_i - \mu)^2.$$

Note that,

$$\sum_{i=1}^{n} (y_i - \mu)^2 \;=\; \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \geq \sum_{i=1}^{n} (y_i - \bar{y})^2$$

with equality iff $\hat{\mu} = \bar{y}$. Therefore, LSE of $\mu$ is $\hat{\mu}_{LS} = \bar{y}$. In vector-matrix formulation,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} = \mu\mathbf{1} + \epsilon.$$

$$||\mathbf{Y} - \mu\mathbf{1}||^2 = (\mathbf{Y} - \mu\mathbf{1})\,(\mathbf{Y} - \mu\mathbf{1})' = \sum_{i=1}^{n} (y_i - \mu)^2 = ||\epsilon||^2.$$

Therefore, least squares is equivalent to finding the multiple of $\mathbf{1}$ which minimizes $||\epsilon||$. This is achieved when we take the perpendicular or the orthogonal projection of $\mathbf{Y}$ onto the space spanned by $\mathbf{1}$. i.e.,

$$\frac{\mathbf{Y}'\mathbf{1}}{\mathbf{1}'\mathbf{1}}\mathbf{1} + (\mathbf{Y} - \frac{\mathbf{Y}'\mathbf{1}}{\mathbf{1}'\mathbf{1}}\mathbf{1}) = \mathbf{Y}$$

i.e.,

$$\hat{\mu}_{LS} = \frac{\mathbf{1}'\mathbf{Y}}{\mathbf{1}'\mathbf{1}} = \bar{y}.$$

Since $Cov(\mathbf{Y}) = \sigma^2 I_n$ and $E(\mathbf{Y}) = \mu\mathbf{1}$,

$$E(\hat{\mu}_{LS}) = \frac{1}{\mathbf{1}'\mathbf{1}}\mathbf{1}'E(\mathbf{Y}) = \frac{\mathbf{1}'\mu\mathbf{1}}{\mathbf{1}'\mathbf{1}} = \mu.$$

$$Var(\hat{\mu}_{LS}) \;=\; Cov(\frac{\mathbf{1}'\mathbf{Y}}{\mathbf{1}'\mathbf{1}}) = \frac{1}{\mathbf{1}'\mathbf{1}}\mathbf{1}'Cov(\mathbf{Y})\frac{1}{\mathbf{1}'\mathbf{1}}\mathbf{1} = \sigma^2\frac{\mathbf{1}'I_n\mathbf{1}}{(\mathbf{1}'\mathbf{1})^2} = \frac{\sigma^2}{n}.$$

Note, that $\hat{\mu}_{LS}$ is a linear unbiased estimate of $\mu$. Suppose $a'Y$ is any linear unbiased estimate of $\mu$. Then $E(a'Y) = \mu a'1 = \mu$ for all $\mu$ implies that $a'1 = 1$. What is the best linear unbiased estimator of $\mu$ (i.e., least MSE)? Note,

$$Var(a'Y) = Cov(a'Y) = a'Cov(Y)a = \sigma^2 a'a.$$

To minimize this we just need to find $a$ such that $a'1 = 1$ and $a'a$ is minimum. Simply note that $a'a = \sum_{i=1}^n a_i^2$ and

$$\frac{1}{n}\sum_{i=1}^n a_i^2 - \left(\frac{\sum_{i=1}^n a_i}{n}\right)^2 \geq 0, \text{ for all } a \text{ since } \sum_{i=1}^n (a_i - \bar{a})^2 \geq 0.$$

i.e.,

$$\frac{1}{n}\sum_{i=1}^n a_i^2 - \left(\frac{1}{n}\right)^2 \geq 0, \text{ or } \sum_{i=1}^n a_i^2 \geq \frac{1}{n}$$

with equality iff $a_i = \frac{1}{n}$ for all $i$. Therefore, $\hat{\mu}_{LS}$ is BLUE (Best Linear Unbiased Estimate) irrespective of the distribution of $\epsilon$.

**Linear models: Estimation**

Data: $(\mathbf{x}_i, y_i)$, $i = 1, 2, \ldots, n$ with multiple predictors or covariates of $y$.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i(p-1)} + \epsilon_i, i = 1, \ldots, n \\ &= \mathbf{x}_i'\beta + \epsilon, i = 1, \ldots, n \end{aligned}$$

is a model for $y|\mathbf{x}$. Let $\mathbf{Y}_{n\times 1} = (y_1, \ldots, y_n)'$, $\beta_{p\times 1} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$,
$\mathbf{X}_{n\times p} = \begin{pmatrix} x_{10} & x_{11} & \ldots & x_{1(p-1)} \\ \vdots & \vdots & \ldots & \vdots \\ x_{n0} & x_{n1} & \ldots & x_{n(p-1)} \end{pmatrix}$, $x_{i0} \equiv 1$ here but can be general also.
$\beta$ is called the vector of regression coefficients and $\mathbf{X}$ is called the regression matrix or the design matrix (especially if $x_{ij} = 0$ or 1). Quite often $y$ is called the dependent variable and $\mathbf{x}$ the set of independent variables. It is more standard to call $y$ the response and $\mathbf{x}$, the regressor or predictor. Recall from previous discussion that
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ is a linear model, but
$y_i = \beta_0 + \beta_1 x_i + x_i^{\beta_2} + \epsilon_i$ is nonlinear. i.e., linear model means linear in $\beta_j$'s.
A general $\mathbf{X}_{n\times p}$ is fine, $\mathbf{X}_0 = \mathbf{1}$ is not essential. Thus we have the linear model:
$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \epsilon$.
Since we have only $n$ observations, it does not make sense to consider $p \geq n$,

so we take $p < n$. Skip bold face for vectors and matrices unless there is ambiguity.

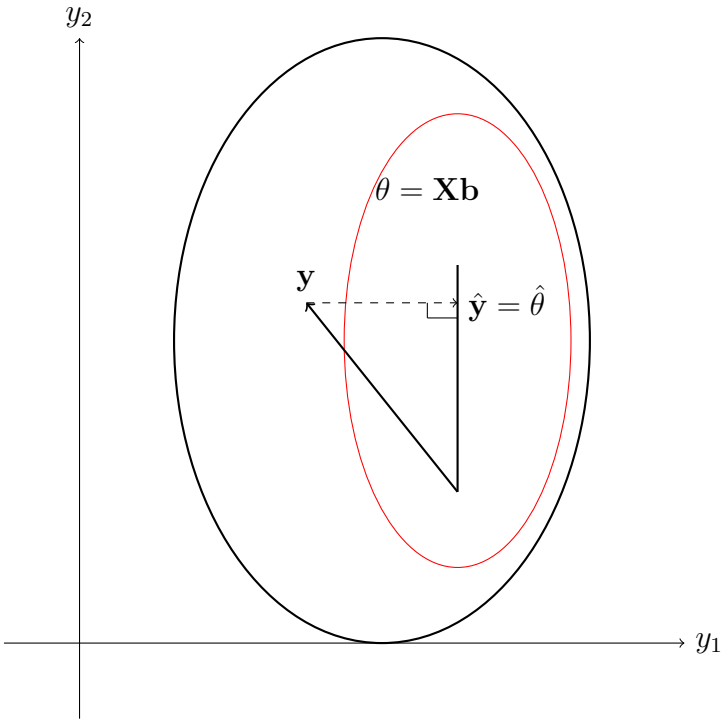
First task is to estimate $\beta$. Most common approach is to use least squares (again, in the absence of distributional assumptions on $\epsilon$). We want

$$\begin{aligned} \min_{\beta \in \mathcal{R}^p} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 &= \min ||\epsilon||^2 = \min_{\beta \in \mathcal{R}^p} ||Y - X\beta||^2 \\ &= \min_{\theta \in \mathcal{M}_C(X)} ||Y - \theta||^2, \end{aligned}$$

where $\mathcal{M}_C(X) = \{a : a = Xb \text{ for some } b \in \mathcal{R}^p\}$. Note that $Xb = b_1X_1 + b_2X_2 + \ldots + b_pX_p$ where $X_i$ are the column vectors of $X$. Now, to minimize $||Y-\theta||^2$ when $\theta \in \mathcal{M}_C(X)$, we should take $\hat{\theta}$ to be the orthogonal projection of $Y$ onto $\mathcal{M}_C(X)$. i.e., $Y - \hat{\theta}$ should be orthogonal to $\mathcal{M}_C(X)$. i.e.,

$$X'(Y - \hat{\theta}) = 0, \text{ or } X'\hat{\theta} = X'Y.$$

$\hat{\theta}$ is uniquely determined, being the unique orthogonal projection of $Y$ onto $\mathcal{M}_C(X)$. We consider the two cases, Rank$(X) = p$ and Rank$(X) < p$, separately.



**Full rank case.** Rank$(X) = p$. Since the columns of $X$ are linearly independent, there exists a unique vector $\hat{\beta}$ such that $\hat{\theta} = X\hat{\beta}$. (If the columns of $X$ are not linearly independent $\hat{\beta}$ is not unique.) Therefore,

$$X'X\hat{\beta} = X'Y.$$

Since $X$ has full column rank, $X'X$ is nonsingular. Therefore,

$$\hat{\beta}_{LS} = (X'X)^{-1}X'Y$$

is unique. One could also use calculus for this derivation:

$$||Y - X\beta||^2 = (Y - X\beta)'(Y - X\beta) = Y'Y - 2\beta'X'Y + \beta'X'X\beta,$$

so differentiating it w.r.t. $\beta$:

$$-2X'Y + 2X'X\beta = 0, \text{ or } X'X\hat{\beta} = X'Y.$$

Note that

$$\hat{\theta} = X\hat{\beta} = X(X'X)^{-1}X'Y = PY = \hat{Y},$$

where $P$ is the projection matrix onto $\mathcal{M}_C(X)$.
$\hat{\epsilon} = Y - \hat{Y} = Y - X\hat{\beta} = (I - P)Y =$ residuals.

$$\begin{aligned}
\hat{\epsilon}'\hat{\epsilon} &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'Y - \hat{\beta}'X'Y + \hat{\beta}'(X'X\hat{\beta} - X'Y) \\
&= Y'Y - \hat{\beta}'X'Y = Y'Y - \hat{\beta}'(X'X\hat{\beta} = Y'(I - P)Y \\
&= \text{ sum of squares of residuals (RSS) } = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2
\end{aligned}$$

**Example.** Find least squares estimate of $\theta_1$ and $\theta_2$ in the following:
$y_1 = \theta_1 + \theta_2 + \epsilon_1$
$y_2 = \theta_1 - \theta_2 + \epsilon_2$
$y_3 = \theta_1 + 2\theta_2 + \epsilon_3$

Obtain $X$ and $\beta$ by writing it in the vector-matrix formulation:

$$\begin{aligned}
\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} &= \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}, \text{ i.e.,} \\
Y &= X\beta + \epsilon.
\end{aligned}$$

Then, noting that

$$\begin{aligned}
X'X &= \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}, \\
(X'X)^{-1} &= \frac{1}{14} \begin{pmatrix} 6 & -2 \\ -2 & 3 \end{pmatrix}
\end{aligned}$$

we obtain

$$\begin{aligned}
\hat{\beta} &= \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} = (X'X)^{-1}X'Y \\
&= \frac{1}{14}\begin{pmatrix} 6 & -2 \\ -2 & 3 \end{pmatrix}\begin{pmatrix} y_1+y_2+y_3 \\ y_1-y_2+2y_3 \end{pmatrix} \\
&= \frac{1}{14}\begin{pmatrix} 6(y_1+y_2+y_3)-2(y_1-y_2+2y_3) \\ -2(y_1+y_2+y_3)+3(y_1-y_2+2y_3) \end{pmatrix} \\
&= \frac{1}{14}\begin{pmatrix} 4y_1+8y_2+2y_3 \\ y_1-5y_2+4y_3 \end{pmatrix} = \begin{pmatrix} \frac{2}{7}y_1+\frac{4}{7}y_2-\frac{1}{7}y_3 \\ \frac{1}{14}y_1-\frac{5}{14}y_2+\frac{2}{7}y_3 \end{pmatrix}, \\
\hat{\epsilon}'\hat{\epsilon} &= Y'Y-\hat{\beta}'X'Y = (y_1^2+y_2^2+y_3^2)-\frac{1}{14}(4y_1+8y_2+2y_3)(y_1+y_2+y_3) \\
&\quad -\frac{1}{14}(y_1-5y_2+4y_3)(y_1-y_2+2y_3).
\end{aligned}$$

**Theorem.** $P = X(X'X)^{-1}X'$ is symmetric idempotent, being the projection matrix onto $\mathcal{M}_C(X)$. $\text{Rank}(P) = \text{Rank}(X) = p$. $I - P$ is the orthogonal projection matrix. $\text{Rank}(I-P) = n-p$ and $(I-P)X = 0$.

The case of $\text{Rank}(X) = r < p$ will be discussed later.

An alternative derivation of $\hat{\beta}$:

$$\begin{aligned}
(Y-X\beta)'(Y-X\beta) &= (Y-X\hat{\beta}+X\hat{\beta}-X\beta)'(Y-X\hat{\beta}+X\hat{\beta}-X\beta) \\
&= (Y-X\hat{\beta})'(Y-X\hat{\beta})+(X\hat{\beta}-X\beta)'(X\hat{\beta}-X\beta) \\
&\quad +2(X\hat{\beta}-X\beta)'(Y-X\hat{\beta}) \\
&= (Y-X\hat{\beta})'(Y-X\hat{\beta})+(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta),
\end{aligned}$$

since

$$(X\hat{\beta}-X\beta)'(Y-X\hat{\beta}) = (\hat{\beta}-\beta)'(X'Y-X'X\hat{\beta}) = 0.$$

Therefore,

$$(Y-X\beta)'(Y-X\beta) \geq (Y-X\hat{\beta})'(Y-X\hat{\beta})$$

with equality iff $\hat{\beta}-\beta = 0$ since $X'X$ is p.d.

**Properties of least squares estimates**

If $Y = X\beta + \epsilon$, with $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I_n$. then $E(\hat{\beta}) = \beta$ since

$$\begin{aligned} E(\hat{\beta}) &= E((X'X)^{-1}X'Y) = (X'X)^{-1}X'E(Y) \\ &= (X'X)^{-1}X'X\beta = \beta, \text{ and} \\ Cov(\hat{\beta}) &= Cov((X'X)^{-1}X'Y) = (X'X)^{-1}X'Cov(Y)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}. \end{aligned}$$

**Theorem (Gauss-Markov).** Consider the Gauss-Markov model, $Y = X\beta + \epsilon$, with $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I_n$. Let $\hat{\theta}$ be the least squares estimate of $\theta = X\beta$. Fix $c \in \mathcal{R}^p$ and consider estimating $c'\theta$. Then, in the class of all linear unbiased estimates of $c'\theta$, $c'\hat{\theta}$ is the unique estimate with minimum variance. (Thus $c'\hat{\theta}$ is BLUE of $c'\theta$.)

**Proof.** $\hat{\theta} = X\hat{\beta} = PY$, where $P$ is the projection matrix onto $\mathcal{M}_C(X)$. In particular, $PX = X$. Therefore,

$$E(c'\hat{\theta}) = c'E(PY) = c'PE(Y) = c'PX\beta = c'X\beta = c'\theta,$$

so that $c'\hat{\theta} = PY$ is a linear unbiased estimate of $c'\theta$. Let $d'Y$ be any other linear unbiased estimate of $c'\theta$. Then $c'\theta = E(d'Y) = d'\theta$, or $(c-d)'\theta = 0$ for all $\theta \in \mathcal{M}_C(X)$. i.e., $(c-d)$ is orthogonal to $\mathcal{M}_C(X)$. Therefore $P(c-d) = 0$, and so $Pc = Pd$. Now,

$$\begin{aligned} Var(d'Y) - Var(c'\hat{\theta}) &= Var(d'Y) - Var(c'PY) \\ &= Var(d'Y) - Var(d'PY) \\ &= \sigma^2(d'd - d'P^2d) = \sigma^2(d'd - d'Pd) \\ &= \sigma^2 d'(I-P)d = \sigma^2 d'(I-P)(I-P)d \\ &\geq 0 \end{aligned}$$

with equality iff $(I-P)d = 0$ or $d = Pd = Pc$. i.e., $d'Y = c'PY = c'\hat{\theta}$.

**Remark.** Since we have assumed that $X$ has full column rank,
$P = X(X'X)^{-1}X'$ and so, if $\theta = X\beta$, then $X'\theta = X'X\beta$ or $\beta = (X'X)^{-1}X'\theta$. Therefore, for every $a \in \mathcal{R}^p$, $a'\beta = a'(X'X)^{-1}X'\theta = c'\theta$, where
$c = X(X'X)^{-1}a$. i.e., every linear function of $\beta$ is a linear function of $\theta$. Therefore, for every $a \in \mathcal{R}^p$, we have that $a'\hat{\beta} = a'(X'X)^{-1}X'\hat{\theta} = c'\hat{\theta}$ is BLUE of $a'\beta$. Thus, when $X$ has full column rank, all linear functions of $\beta$ have BLUE, all components of $\beta$ are estimable (BLUE exists). This will not be the case when $X$ has less than full column rank.

**Result.** In the model, $Y = X\beta + \epsilon$, $E(\epsilon) = 0$, $Cov(\epsilon) = \sigma^2 I_n$ and $X$ has full column rank $(p)$, we have that

$$E(RSS) = E((Y - X\hat{\beta})'(Y - X\hat{\beta})) = (n-p)\sigma^2,$$

so that $\text{RSS}/(n-p)$ is an unbiased estimate of $\sigma^2$.

**Proof.** Note that $Y - X\hat{\beta} = Y - PY = (I-P)Y$. Therefore,

$$\text{RSS} = (Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'(I-P)^2Y = Y'(I-P)Y,$$

where $I - P$ is symmetric idempotent with rank $n - p$.

$$\begin{aligned} E(RSS) &= E(Y'(I-P)Y) = \text{tr}(\sigma^2(I-P)) + (X\beta)'(I-P)(X\beta) \\ &= \sigma^2(n-p) + \beta' X'(I-P)X\beta \\ &= (n-p)\sigma^2. \end{aligned}$$

For confidence statements and testing we need distribution theory.

**Distribution Theory**

Suppose $\epsilon_i$ are i.i.d. $N(0, \sigma^2)$. Then $\epsilon_{n\times 1} \sim N_n(0, \sigma^2 I_n)$ and so,
$Y \sim N_n(X\beta, \sigma^2 I_n)$.

**Theorem.** If $Y \sim N_n(X\beta, \sigma^2 I_n)$ and $X$ has rank $p$, then
(i) $\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$,
(ii) $(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)/\sigma^2 \sim \chi^2_p$,
(iii) $\hat{\beta}$ is independent of $\text{RSS} = (Y - X\hat{\beta})'(Y - X\hat{\beta})$,
(iv) $\text{RSS}/\sigma^2 \sim \chi^2_{n-p}$.

**Proof.** $Y \sim N_n(X\beta, \sigma^2 I_n)$, so (i)

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \sim N_p((X'X)^{-1}X'X\beta, \sigma^2(X'X)^{-1}X'X(X'X)^{-1}) \\ &= N(\beta, \sigma^2(X'X)^{-1}). \end{aligned}$$

(ii) Since $\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$, note $(X'X)^{1/2}(\hat{\beta} - \beta) \sim N_p(0, \sigma^2 I_p)$, and hence
$(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)/\sigma^2 \sim \chi^2_p$.
(iii) $\hat{\beta} = (X'X)^{-1}X'Y = AY$ and $\text{RSS} = Y'(I-P)Y$. Since $Y \sim N_n(X\beta, \sigma^2 I_n)$, independence of $\hat{\beta}$ and $Y'(I-P)Y$ holds iff $A(I-P) = 0$. But $(I-P)A' = (I-P)X(X'X)^{-1} = 0$. Alternatively, $\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'P'Y = (X'X)^{-1}X'(PY)$, so that it is independent of $(I-P)Y$.
(iv) (a) $\text{RSS} = Y'(I-P)Y = (Y - X\beta)'(I-P)(Y - X\beta)$ since $(I-P)X = 0$. Note that since $Y - X\beta \sim N_n(0, \sigma^2 I_n)$, and $I - P$ is idempotent of rank $n-p$,

$(Y - X\beta)'(I - P)(Y - X\beta) \sim \chi^2_{n-p}$.
(b) Alternatively, note that $Q = (Y - X\beta)'(Y - X\beta) \sim \sigma^2\chi^2_n$. Now

$$\begin{aligned} Q &= (Y - X\beta)'(Y - X\beta) \\ &= (Y - X\hat{\beta} + X\hat{\beta} - X\beta)'(Y - X\hat{\beta} + X\hat{\beta} - X\beta) \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \\ &= Q_1 + Q_2, \end{aligned}$$

where $Q_2 \sim \sigma^2\chi^2_p$ and $Q_1 \geq 0$. Therefore, from a previous result, $Q_1 \sim \sigma^2\chi^2_{n-p}$ independent of $Q_2$.

**Design matrix $X$ with less than full column rank**

Consider the model,

$$y_{ij} = \mu + \alpha_i + \tau_j + \epsilon_{ij}, i = 1, 2, \ldots, I; j = 1, 2, \ldots, J,$$

for the response from the $i$th treatment in the $j$th block, say. This can be put in the usual linear model form: $Y = X\beta + \epsilon$ as follows:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1J} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2J} \\ \vdots \\ y_{I1} \\ y_{I2} \\ \vdots \\ y_{IJ} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & 1 & 0 & \ldots & 0 & 0 & 0 & 0 & \ldots & 1 \\ 1 & 0 & 1 & \ldots & 0 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & 1 & \ldots & 0 & 0 & 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & 0 & \ldots & 1 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 1 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & 0 & \ldots & 1 & 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_I \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_J \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1J} \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{2J} \\ \vdots \\ \epsilon_{I1} \\ \epsilon_{I2} \\ \vdots \\ \epsilon_{IJ} \end{pmatrix}.$$

Here, $X$ does not have full column rank. For instance, the first column is proportional to the sum of the rest. Thus $X'X$ is singular, so the previous discussion does not apply. $\beta$ itself is not estimable, but what parametric functions of $\beta$ are estimable?

**Result.** For any matrix $A$, the row space of $A$ satisfies $\mathcal{M}_C(A') = \mathcal{M}_C(A'A)$.

**Proof.** $Ax = 0$ implies $A'Ax = 0$. Also, $A'Ax = 0$ implies $x'A'Ax = 0$, so $Ax = 0$. Therefore the null space of $A$ and $A'A$ are the same. Consider the orthogonal space and note $\text{Rank}(A'A) = \text{Rank}(A) = \text{Rank}(A')$. Further, since $A'Aa = A'b$ where $b = Aa$, $\mathcal{M}_C(A'A) \subset \mathcal{M}_C(A')$. Since the ranks (or dimensions) are the same, the spaces must be the same.

**Theorem.** Let $Y = \theta + \epsilon$ where $\theta = X\beta$ and $X_{n\times p}$ has rank $r < p$. Then
(i) $\min_{\theta \in \mathcal{M}_C(X)} ||Y - \theta||^2$ is achieved (i.e., least squares is attained) when $\hat{\theta} = X\hat{\beta}$ where $\hat{\beta}$ is any solution of $X'X\beta = X'Y$;
(ii) $Y'Y - \hat{\beta}'X'Y$ is unique for all nonzero $Y$.

**Proof.** (i) $X'X\beta = X'Y$ always has some solution (for $\beta$) since $\mathcal{M}_C(X'X) = \mathcal{M}_C(X')$. However, the solution is not unique since $\text{Rank}(X'X) = r < p$.

Let $\hat{\beta}$ be any solution, and let $\hat{\theta} = X\hat{\beta}$. Then $X'(Y - \hat{\theta}) = 0$. However, given $Y \in \mathcal{R}^n$, the decomposition, $Y = \hat{\theta} \oplus (Y - \hat{\theta})$ where $Y - \hat{\theta}$ is orthogonal to $\mathcal{M}_C(X)$ is unique, and for such a $\hat{\theta}$, $||Y - \theta||^2$ is minimized. We know from previous discussion that $\min_{\theta \in \mathcal{M}_C(X)} ||Y - \theta||^2$ is achieved with $\hat{\theta} = PY$ which is unique.
(ii) Note that

$$Y'Y - \hat{\beta}'X'Y = Y'Y - \hat{\theta}'Y = (Y - \hat{\theta})'(Y - \hat{\theta}),$$

since $\hat{\theta}'(Y - \hat{\theta}) = 0$. Also, $(Y - \hat{\theta})'(Y - \hat{\theta}) = ||Y - \hat{\theta}||^2$ is the unique minimum.

**Question.** Earlier we could find $\hat{\beta}$ directly. How do we find $\hat{\theta}$ now?

**Projection matrices**

From the theory of orthogonal projections, given $X_{n \times p}$ (i.e., $p$ many $n$-vectors), there exists $P_{n \times n}$ satisfying
(i) $Px = x$ for all $x \in \mathcal{M}_C(X)$, and
(ii) if $\xi \in \mathcal{M}_C^{\perp}(X)$, then $P\xi = 0$.
What are the properties of such a $P$?

1. $P$ is unique: Suppose $P_1$ and $P_2$ satisfy (i) and (ii). Let $w \in \mathcal{R}^n$. Then $w = Xa + b$, $b \in \mathcal{M}_C^{\perp}(X)$. Then,

$$(P_1 - P_2)w \;\; = \;\; (P_1 - P_2)Xa + (P_1 - P_2)b = (Xa - Xa) + (P_1 b - P_2 b) = 0.$$

Since this is true for all $w \in \mathcal{R}^n$, we must have $P_1 - P_2 = 0$.

2. $P$ is idempotent and symmetric:

$$P^2 x = P(Px) = Px = x \text{ for all } x \in \mathcal{M}_C(X);$$

$$P^2 \xi = P(P\xi) = P0 = 0 \text{ for all } \xi \in \mathcal{M}_C^{\perp}(X).$$

Therefore $P^2$ satisfies (i) and (ii), and since $P$ is unique, $P^2 = P$. Further, $Py \perp (I - P)x$ for all $x, y$, so that $y'P'(I - P)x = 0$. i.e., $P' = P'P$, so $P = (P')' = (P'P)' = P'P = P'$.

**Result.** Let $\Omega$ be a subspace of the vector space $\mathcal{R}^n$, and let $P_\Omega$ be its projection matrix. Then $\mathcal{M}_C(P_\Omega) = \Omega$.

**Proof.** Note that $\mathcal{M}_C(P_\Omega) \subset \Omega$. For this, take $y \in \mathcal{M}_C(P_\Omega)$. Then $y$ is a linear combination of columns of $P_\Omega$, or $y = P_\Omega u$ for some $u$. Since $u = w \oplus v$, $w \in \Omega$, $v \in \Omega^{\perp}$, we have, $y = P_\Omega u = P_\Omega(w \oplus v) = P_\Omega w = w \in \Omega$. Conversely, if $x \in \Omega$, then $x = P_\Omega x \in \mathcal{M}_C(P_\Omega)$.

$I_n - P_\Omega$ represents the orthogonal projection. i.e., $\mathcal{R}^n = \Omega \oplus \Omega^\perp$. Thus for any $y \in \mathcal{R}^n$, we have $y = P_\Omega y \oplus (I - P_\Omega)y$.

If $P_{n\times n}$ is any symmetric idempotent matrix, it represents a projection onto $\mathcal{M}_C(P)$: if $y \in \mathcal{R}^n$, then $y = Py + (I - P)y = u + v$. Note

$$u'v = (Py)'(I - P)y = y'P(I - P)y = y'(P - P^2)y = 0,$$

so that we get $y = u \oplus v$, $u \in \mathcal{M}_C(P)$, $v \in \mathcal{M}_C^\perp(P)$.

**Question.** Given $X$, how to find $P$ such that $\mathcal{M}_C(X) = \mathcal{M}_C(P)$?

**Result.** If $\Omega = \mathcal{M}_C(X)$, then $P_\Omega = X(X'X)^- X'$, where $X'X)^-$ is any generalized inverse of $X'X$.

**Definition.** If $B_{m\times n}$ is any matrix, *a* generalized inverse of $B$ is any $n \times m$ matrix $B^-$ satisfying $BB^-B = B$.

**Existence.** From singular value decomposition of $B$, there exist orthogonal matrices $P_{m\times m}$ and $Q_{n\times n}$ such that

$$P_{m\times m}B_{m\times n}Q_{n\times n} = \Delta_{m\times n} = \begin{pmatrix} D_{r\times r} & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{pmatrix},$$

where $r = \text{Rank}(B)$. Define $\Delta^-_{n\times m} = \begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}$ and let $B^- = Q\Delta^- P$. First,

$$\Delta\Delta^-\Delta = \begin{pmatrix} D_r & 0 \\ 0 & 0 \end{pmatrix}\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}\begin{pmatrix} D_r & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} D_r & 0 \\ 0 & 0 \end{pmatrix} = \Delta.$$

Further, $B = P'\Delta Q'$, so that

$$BB^-B = P'\Delta Q'Q\Delta^- PP'\Delta Q' = P'\Delta\Delta^-\Delta Q' = P'\Delta Q' = B.$$

**Proof of Result.** Let $B = X'X$. Find $B^-$ such that $BB^-B = B$. For any $Y \in \mathcal{R}^n$, let $c = X'Y$, and let $\tilde{\beta}$ be any solution of $X'X\beta = X'Y$, or that of $B\beta = c$. Then

$$B(B^-c) = BB^-B\tilde{\beta} = B\tilde{\beta} = c,$$

so that $\hat{\beta} = B^-c$ is a particular solution of $B\beta = c$. Let $\hat{\theta} = X\hat{\beta} = XB^-c$. Then, $Y = \hat{\theta} + (Y - \hat{\theta})$, where

$$\hat{\theta}'(Y - \hat{\theta}) = \hat{\beta}'X'(Y - X\hat{\beta}) = \hat{\beta}'(X'Y - X'X\hat{\beta}) = 0.$$

Therefore we have an orthogonal decomposition of $Y$ such that $\hat{\theta} \in \mathcal{M}_C(X)$ and $(Y - \hat{\theta})\perp\mathcal{M}_C(X)$. Now note that $\hat{\theta} = X\hat{\beta} = X(X'X)^-X'Y$. i.e., for $Y$, its projection onto $\mathcal{M}_C(X)$ is given by $X(X'X)^-X'Y$. Therefore, $P_\Omega = X(X'X)^-X'$ since $P_\Omega$ is unique.

Techniques for finding $B^-$ are needed: if $B = X'X$, then $P_\Omega = X(X'X)^-X'$; if we want to solve $X'X\beta = X'Y$, or $B\beta = c$, then $\hat{\beta} = B^-c$.

For $B_{p\times m}$ with rank $r < p$ and $B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$, where $B_{11}$ (which is $r \times r$ of rank $r$) is nonsingular, if we take $B^- = \begin{pmatrix} B_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix}$, then note that

$$\begin{aligned} BB^-B &= \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}\begin{pmatrix} B_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \\ &= \begin{pmatrix} I_r & 0 \\ B_{21}B_{11}^{-1} & 0 \end{pmatrix}\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{21}B_{11}^{-1}B_{12} \end{pmatrix}. \end{aligned}$$

Now note that $(B_{21}|B_{22})$ is a linear function of $(B_{11}|B_{12})$, or $(B_{21}|B_{22}) = K(B_{11}|B_{12}) = (KB_{11}|KB_{12})$ for some matrix $K$. Therefore, $KB_{11} = B_{21}$, or $K = B_{21}B_{11}^{-1}$, so $B_{22} = KB_{12} = B_{21}B_{11}^{-1}B_{12}$.

**Example.** Let $B = \begin{pmatrix} 1 & 2 & 5 & 2 \\ 3 & 7 & 12 & 4 \\ 0 & 1 & -3 & -2 \end{pmatrix}$. Then rank of $B$ is 2 since 2nd row - 3× 1st row = 3rd row. Partition $B$ as: $B = \left(\begin{array}{cc|cc} 1 & 2 & 5 & 2 \\ 3 & 7 & 12 & 4 \\ \hline 0 & 1 & -3 & -2 \end{array}\right) = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$. Take

$$B^- = \begin{pmatrix} B_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 7 & -2 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

**Example.** Consider the model:
$y_1 = \beta_1 + \beta_2 + \epsilon_1$
$y_2 = \beta_1 + \beta_2 + \epsilon_2$
$y_3 = \beta_1 + \beta_2 + \epsilon_3$

This is equivalent to

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}.$$

$X$ has rank 1; $X'X = \begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} = \left(\begin{array}{c|c} 3 & 3 \\ \hline 3 & 3 \end{array}\right)$, so choose $(X'X)^- = \begin{pmatrix} 1/3 & 0 \\ 0 & 0 \end{pmatrix}$.

Then check that $(X'X)(X'X)^{-}(X'X) = X'X$. We have then

$$\begin{aligned} X\hat{\beta} &= \hat{\theta} = X(X'X)^{-}X'Y \\ &= \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1/3 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \\ &= \begin{pmatrix} 1/3 & 0 \\ 1/3 & 0 \\ 1/3 & 0 \end{pmatrix} \begin{pmatrix} y_1 + y_2 + y_3 \\ y_1 + y_2 + y_3 \end{pmatrix} = \begin{pmatrix} (y_1 + y_2 + y_3)/3 \\ \bar{y} \\ \bar{y} \end{pmatrix} \\ &= \begin{pmatrix} \widehat{\beta_1 + \beta_2} \\ \widehat{\beta_1 + \beta_2} \\ \widehat{\beta_1 + \beta_2} \end{pmatrix}. \end{aligned}$$

Only $\beta_1 + \beta_2$ can be estimated? Note $\beta_1 + \beta_2 = (1 \;\; 1) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ and $\mathcal{M}_C \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \mathcal{M}_C(X')$. More on this later.

**Theorem.** If $Y \sim N_n(X\beta, \sigma^2 I_n)$, where $X_{n\times p}$ has rank $r$ and $\hat{\beta} = (X'X)^- X'Y$ is a least squares solution of $\beta$,
(i) $X\hat{\beta} \sim N_n(X\beta, \sigma^2 P)$,
(ii) $(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta) \sim \sigma^2\chi^2_r$
(iii) $X\hat{\beta}$ is independent of RSS $= (Y - X\hat{\beta})'(Y - X\hat{\beta})$. and
(iv) RSS$/\sigma^2 \sim \chi^2_{n-r}$ (independent of $X\hat{\beta}$)

**Proof.** (i) Since $X\hat{\beta} = X(X'X)^- X'Y = PY$, we have

$$X\hat{\beta} \sim N_n(PX\beta, \sigma^2 P^2) = N_n(X\beta, \sigma^2 P).$$

(ii) Since $X\hat{\beta} = PY$ and $X\beta = PX\beta$,

$$\begin{aligned}(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta) &= (X\hat{\beta}-X\beta)'(X\hat{\beta}-X\beta)\\ &= (Y-X\beta)'P(Y-X\beta) \sim \sigma^2\chi^2_r,\end{aligned}$$

$P$ being symmetric idempotent of rank $r$.
(iii) We have $X\hat{\beta} = PY$, RSS $= (Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'(I-P)Y$ and $P(I-P) = 0$. Therefore independence of $X\hat{\beta}$ and RSS follows.
(iv) Note again that

$$\text{RSS} = Y'(I-P)Y = (Y-X\beta)'(I-P)(Y-X\beta) \sim \sigma^2\chi^2_{n-r},$$

$I - P$ being a projection matrix of rank $n - r$.

**Estimability**

Consider the Gauss-Markov model again: $Y = X\beta + \epsilon$, with $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I_n$. Now suppose rank of $X$ is $r < p$.

**Definition.** A linear parametric function $a'\beta$ is said to be estimable if it has a linear unbiased estimate $b'Y$.

**Theorem.** $a'\beta$ is estimable iff $a \in \mathcal{M}_C(X') = \mathcal{M}_C(X'X)$.

**Proof.** $a'\beta$ is estimable iff there exists $b$ such that $E(b'Y) = a'\beta$ for all $\beta \in \mathcal{R}^p$. i.e., $b'X\beta = a'\beta$ for all $\beta \in \mathcal{R}^p$. i.e., $b'X = a'$ or $a = X'b$ for some $b \in \mathcal{R}^n$.

**Theorem (Gauss-Markov).** If $a'\beta$ is estimable, and $\hat{\beta}$ is any least squares solution (i.e., solution of $X'X\beta = X'Y$),
(i) $a'\hat{\beta}$ is unique,
(ii) $a'\hat{\beta}$ is the BLUE of $a'\beta$.

**Proof.** (i) If $a'\beta$ is estimable, $a'\beta = b'X\beta = b'\theta$ for some $b \in \mathcal{R}^n$. Since $\hat{\theta}$ is the unique projection of $Y$ onto $\mathcal{M}_C(X)$, we note $b'\hat{\theta} = b'X\hat{\beta} = a'\hat{\beta}$ is

unique. i.e., if $\tilde{\beta}$ is any other LS solution, then also $b'X\tilde{\beta} = b'X\hat{\beta} = a'\hat{\beta}$.
(ii) If $d'Y$ is any other linear unbiased estimate of $a'\beta$, then
$E(d'Y) = d'X\beta = d'\theta = a'\beta = b'X\beta = b'\theta$ for all $\beta \in \mathcal{R}^p$.
i.e., $d'\theta = b'\theta$ for all $\theta \in \mathcal{M}_C(X)$.
i.e., $(d-b)'\theta = 0$ for all $\theta \in \mathcal{M}_C(X)$, or $(d-b) \perp \mathcal{M}_C(X)$. Consider $P = P_{\mathcal{M}_C(X)} = X(X'X)^-X'$. Then $P(d-b) = 0$ or $Pd = Pb$. Therefore,

$$\begin{aligned} Var(d'Y) - Var(a'\hat{\beta}) &= Var(d'Y) - Var(b'\hat{\theta}) \\ &= Var(d'Y) - Var(b'PY) = Var(d'Y) - Var(d'PY) \\ &= \sigma^2(d'd - d'Pd) = \sigma^2 d'(I-P)d \geq 0, \end{aligned}$$

with equality iff $(I-P)d = 0$ or $d = Pd = Pb$. i.e., $d'Y = b'PY = b'\hat{\theta} = a'\hat{\beta}$.

**Remark.** Parametric functions $a'\beta$ are estimable when $a \in \mathcal{M}_C(X') =$ Row space of $X$.

**Example.** Consider again the model:
$y_{ij} = \mu + \alpha_i + \tau_j + \epsilon_{ij},\ i = 1,2,3,4;\ j = 1,2.$
Suppose comparing $\tau_1$ and $\tau_2$ is of interest. Since

$$Y = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{41} \\ y_{42} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \tau_1 \\ \tau_2 \end{pmatrix} + \epsilon,$$

$\mu+\alpha_i+\tau_j$ is estimable for all $i$ and $j$. Therefore, $(\mu+\alpha_i+\tau_1)-(\mu+\alpha_i+\tau_2) = \tau_1 - \tau_2$ is estimable.
$(\mu+\alpha_i+\tau_1) - (\mu+\alpha_j+\tau_1) = \alpha_i - \alpha_j$ is estimable.
What else is estimable, apart from linear combinations of these?

**Result.** If $a'\beta$ is estimable, and $Y \sim N_n(X\beta, \sigma^2 I_n)$, a $100(1-\alpha)\%$ confidence interval for $a'\beta$ is given by

$$a'\hat{\beta} \pm t_{n-r}(1-\alpha/2)\sqrt{a'(X'X)^-a}\sqrt{\text{RSS}/(n-r)}.$$

**Proof.** Note that $a'\beta = c'X\beta = c'\theta$ for some $c$. Therefore, $a'\hat{\beta} = c'\hat{\theta} = c'PY \sim N(a'\beta, \sigma^2 c'Pc)$. Now $c'Pc = c'X(X'X)^-X'c = a'(X'X)^-a$. Therefore,

$$\frac{a'\hat{\beta} - a'\beta}{\sqrt{\sigma^2 a'(X'X)^-a}} \sim N(0,1).$$

Further, since $\mathrm{RSS}/\sigma^2 \sim \chi^2_{n-r}$ independent of $X\hat{\beta}$, and hence of $c'X\hat{\beta} = c'\hat{\theta} = a'\hat{\beta}$,

$$\frac{a'\hat{\beta} - a'\beta}{\sqrt{\sigma^2 a'(X'X)^- a}\sqrt{\mathrm{RSS}/(\sigma^2(n-r))}} \sim t_{n-r}.$$

Hence,

$$P\left(|a'\hat{\beta} - a'\beta| \leq t_{n-r}(1-\alpha/2)\sqrt{a'(X'X)^- a}\sqrt{\frac{\mathrm{RSS}}{n-r}}\right) = 1-\alpha.$$

**Maximum likelihood estimation**

Does LS estimate have other optimality properties?

Since we have assumed that $Y \sim N_n(X\beta, \sigma^2 I_n)$ to derive distributional properties of $\hat{\beta}$, let us derive the maximum likelihood estimates of $\beta$ and $\sigma^2$ under this assumption. $\hat{\beta}_{mle}$ and $\hat{\sigma}^2$ are values of $\beta$ and $\sigma^2$ which maximize the likelihood,

$$(2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right).$$

Equivalently, we may maximize the loglikelihood,

$$-\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta).$$

Fix $\sigma^2$ and maximize over $\beta$, then maximize over $\sigma^2$. Now note that maximizing over $\beta$ (for any fixed $\sigma^2$) is equivalent to minimizing $(Y-X\beta)'(Y-X\beta) = ||Y - X\beta||^2$, which yields the same estimate as the least squares. i.e., $\hat{\beta}_{mle} = \hat{\beta}_{ls}$. However, $\hat{\sigma^2} = \text{RSS}/n$, which is not unbiased.

**Estimation under linear restrictions or constraints**

Consider the following examples.
(i) $y_{ij} = \mu + \alpha_i + \tau_j + \epsilon_{ij}$. Test $H_0 : \tau_1 = \tau_2$. i.e., test whether there is any difference between treatments 1 and 2. Under $H_0$, $\tau_1 - \tau_2 = 0$, or $A\beta = c$ where $A = a' = (0, 0, \ldots, 0, 1, -1, 0, \ldots, 0)$, $\beta = (\mu, \alpha_1, \ldots, \alpha_I, \tau_1, \tau_2, \ldots)'$.
(ii) $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i(p-1)} + \epsilon_i$. Test $H_0 : X_1, \ldots, X_{p-1}$ are not useful.

Recall that, to derive the GLRT, we need to estimate the parameters of the model, both with and without restrictions. While testing linear hypotheses in a linear model, we need to estimate $\beta$ under the linear constraint $A\beta = c$.

Consider $Y = X\beta + \epsilon$, $X_{n\times p}$ of rank $p$, first. We will consider the deficient rank case later. Let us see how we can find the least squares estimate of $\beta$ subject to $H : A\beta = c$, where $A_{q\times p}$ of rank $q$ and $c$ is given. We can use the Lagrange multiplier method of calculus for this as follows.

$$\begin{aligned} &\min_{\beta} ||Y - X\beta||^2 + \lambda'(A\beta - c) \\ &= \min_{\beta} \left\{Y'Y - 2\beta'X'Y + \beta'X'X\beta + \lambda'A\beta - \lambda'c\right\}, \end{aligned} \tag{1}$$

differentiating which (w.r.t. $\beta$) and setting equal to 0, we get,

$$-2X'Y + 2X'X\beta + A'\lambda = 0 \text{ or } X'X\beta = X'Y - \frac{1}{2}A'\lambda_H.$$

Therefore,

$$\hat{\beta}_H = (X'X)^{-1}\left\{X'Y - \frac{1}{2}A'\lambda_H\right\} = \hat{\beta} - \frac{1}{2}(X'X)^{-1}A'\lambda_H \quad (*).$$

Differentiating (1) w.r.t. $\lambda$, we get $A\beta - c = 0$. Since

$$\begin{aligned}
c = A\hat{\beta}_H &= A\hat{\beta} - \frac{1}{2}A(X'X)^{-1}A'\lambda_H, \\
c - A\hat{\beta} &= -\frac{1}{2}A(X'X)^{-1}A'\lambda_H, \text{ and hence} \\
-\frac{1}{2}\lambda_H &= \left[A(X'X)^{-1}A'\right]^{-1}(c - A\hat{\beta}), \text{ and therefore} \\
\hat{\beta}_H &= \hat{\beta} + (X'X)^{-1}A'\left[A(X'X)^{-1}A'\right]^{-1}(c - A\hat{\beta}).
\end{aligned}$$

To establish minimization subject to $A\beta = c$, note that

$$\begin{aligned}
||X(\hat{\beta} - \beta)||^2 &= (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \\
&= (\hat{\beta} - \hat{\beta}_H + \hat{\beta}_H - \beta)'X'X(\hat{\beta} - \hat{\beta}_H + \hat{\beta}_H - \beta) \\
&= (\hat{\beta} - \hat{\beta}_H)'X'X(\hat{\beta} - \hat{\beta}_H) + (\hat{\beta}_H - \beta)'X'X(\hat{\beta}_H - \beta) \\
&\quad +2(\hat{\beta} - \hat{\beta}_H)'X'X(\hat{\beta}_H - \beta) \\
&= ||X(\hat{\beta} - \hat{\beta}_H)||^2 + ||X(\hat{\beta}_H - \beta)||^2,
\end{aligned}$$

since, from (*) above, and subject to $A\beta = c$,

$$\begin{aligned}
(\hat{\beta} - \hat{\beta}_H)'X'X(\hat{\beta}_H - \beta) &= \frac{1}{2}\lambda_H'A(X'X)^{-1}X'X(\hat{\beta}_H - \beta) \\
&= \frac{1}{2}\lambda_H'A(\hat{\beta}_H - \beta) = \frac{1}{2}\lambda_H'(A\hat{\beta}_H - A\beta) = 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
||Y - X\beta||^2 &= ||Y - X\hat{\beta}||^2 + ||X(\hat{\beta} - \beta)||^2 \\
&= ||Y - X\hat{\beta}||^2 + ||X(\hat{\beta} - \hat{\beta}_H)||^2 + ||X(\hat{\beta}_H - \beta)||^2 \\
&\geq ||Y - X\hat{\beta}||^2 + ||X(\hat{\beta} - \hat{\beta}_H)||^2,
\end{aligned}$$

and is a minimum when $\beta = \hat{\beta}_H$. (Note, $X(\hat{\beta}_H - \beta) = 0$ implies $X'X(\hat{\beta}_H - \beta) = 0$, so $\hat{\beta}_H - \beta = 0$ since columns of $X$ are linearly independent.) Also, from above, we get,

$$||Y - X\hat{\beta}_H||^2 = ||Y - X\hat{\beta}||^2 + ||X(\hat{\beta} - \hat{\beta}_H)||^2.$$

If we let $\hat{Y} = X\hat{\beta}$ and $\hat{Y}_H = X\hat{\beta}_H$, then

$$||Y - \hat{Y}_H||^2 = ||Y - \hat{Y}||^2 + ||\hat{Y} - \hat{Y}_H||^2.$$

Note that this can also be established using projection matrices, and not just for the full column rank case. Let us first establish it for the case $\text{Rank}(X_{n\times p}) = p$ again, and next extend it.

Let $\beta_0$ be a solution of $A\beta = c$. Then $Y - X\beta_0 = X(\beta - \beta_0) + \epsilon$ or $\tilde{Y} = X\gamma + \epsilon$ with $A\gamma = A(\beta - \beta_0) = 0$. i.e.,

$$\tilde{Y} = \theta + \epsilon, \quad \theta \in \mathcal{M}_C(X) = \Omega, \quad \text{and}$$

$$A(X'X)^{-1}X'\theta = A(X'X)^{-1}X'X(\beta - \beta_0) = A(\beta - \beta_0) = A\gamma = 0.$$

Set $A_1 = A(X'X)^{-1}X'$ and $\omega = \mathcal{N}(A_1) \cap \Omega$. Then $A_1\theta = A\gamma = 0$ and we want the projection of $\tilde{Y}$ onto $\omega$ since we want:

$$\min_{\theta \in \mathcal{M}_C(X)} ||\tilde{Y} - \theta||^2 \text{ subject to } A_1\theta = 0.$$

We need the following series of results to solve this.

**Result A.** If $\mathcal{N}(C)$ is the null space of $C$, then $\mathcal{N}(C) = \mathcal{M}^{\perp}(C')$.

**Proof.** If $x \in \mathcal{N}(C)$, then $Cx = 0$ so that $x$ is orthogonal to each row of $C$. i.e., $x \perp \mathcal{M}(C')$. Conversely, if $x \perp \mathcal{M}(C')$, then $x'C' = (Cx)' = 0$, or $Cx = 0$, hence $x \in \mathcal{N}(C)$.

**Result B.** $(\Omega_1 \cap \Omega_2)^{\perp} = \Omega_1^{\perp} + \Omega_2^{\perp}$.

**Proof.** Let $\Omega_i = \mathcal{N}(C_i), i = 1, 2$. Then,

$$(\Omega_1 \cap \Omega_2)^{\perp} = \left(\mathcal{N}\begin{pmatrix} C_1 \\ C_2 \end{pmatrix}\right)^{\perp} = \mathcal{M}\left(C_1' | C_2'\right) = \mathcal{M}(C_1') + \mathcal{M}(C_2') = \Omega_1^{\perp} + \Omega_2^{\perp}.$$

**Result C.** If $\omega \subset \Omega$, then $P_\Omega P_\omega = P_\omega P_\Omega = P_\omega$.

**Proof.** Show that $P_\Omega P_\omega$ and $P_\omega P_\Omega$ both satisfy the defining properties of $P_\omega$: If $x \in \omega \subset \Omega$, then $P_\Omega P_\omega x = P_\Omega x = x$; if $\xi \in \omega^{\perp}$, $P_\Omega P_\omega \xi = P_\Omega 0 = 0$. Similar is the other case.

**Result D.** If $\omega \subset \Omega$, then $P_\Omega - P_\omega = P_{\omega^{\perp} \cap \Omega}$.

**Proof.** $\Omega = \mathcal{M}_C(P_\Omega)$, so each $x \in \Omega$ can be written $x = P_\Omega y$. Consider the decomposition, $P_\Omega y = P_\omega y + (P_\Omega - P_\Omega)y$. Now $P_\omega y \in \omega \subset \Omega$, and already $P_\Omega y \in \Omega$, so $(P_\Omega - P_\Omega)y = P_\Omega y - P_\omega y \in \Omega$. Further, $P_\omega(P_\Omega - P_\omega) = P_\omega P_\Omega - P_\omega = P_\omega - P_\omega = 0$, so that $(P_\omega y)'(P_\Omega - P_\omega)y = y'P_\omega(P_\Omega - P_\omega)y = 0$. Therefore, $P_\Omega y = P_\omega y \oplus (P_\Omega - P_\omega)y$ is the orthogonal decomposition of $\Omega$ into $\omega \oplus (\omega^{\perp} \cap \Omega)$.

**Result E.** If $A_1$ is any matrix such that $\omega = \mathcal{N}(A_1) \cap \Omega$, then $\omega^{\perp} \cap \Omega = \mathcal{M}_C(P_\Omega A_1')$.

**Proof.** Note that

$$\omega^{\perp} \cap \Omega = (\Omega \cap \mathcal{N}(A_1))^{\perp} \cap \Omega = \left(\Omega^{\perp} \oplus \mathcal{N}^{\perp}(A_1)\right) \cap \Omega = \left(\Omega^{\perp} \oplus \mathcal{M}_C(A_1')\right) \cap \Omega.$$

Now, let $x \in \omega^\perp \cap \Omega \left(= \left(\Omega^\perp \oplus \mathcal{M}_C(A_1')\right) \cap \Omega.\right)$. Then $x \in \Omega$, so $x = P_\Omega x$. Also, $x \in \Omega^\perp \oplus \mathcal{M}_C(A_1')$, so $x = (I - P_\Omega)\alpha + A_1'\beta$. Therefore,

$$x = P_\Omega x = P_\Omega \left\{(I - P_\Omega)\alpha + A_1'\beta\right\} = P_\Omega A_1'\beta \in \mathcal{M}_C(A_1').$$

Conversely, if $x \in \mathcal{M}_C(P_\Omega A_1')$, then $x = P_\Omega A_1'\beta = P_\Omega(A_1'\beta) \in \mathcal{M}_C(P_\Omega) = \Omega$. For any $\xi \in \omega (\subset \Omega)$, we have $x'\xi = \beta' A_1 P_\Omega \xi = \beta' A_1 \xi = 0$ since $\omega = \mathcal{N}(A_1) \cap \Omega$. Therefore, $x \in \omega^\perp$.

**Result F.** If $A_1$ is a $q \times n$ matrix of rank $q$, then $\text{Rank}(P_\Omega A_1') = q$ iff $\mathcal{M}_C(A_1') \cap \Omega^\perp = \{0\}$.

**Proof.** $\text{Rank}(P_\Omega A_1') \leq \text{Rank}(A_1') = \text{Rank}(A_1) = q$. Suppose $\text{Rank}(P_\Omega A_1') < q$. Let the rows of $A_1$ (i.e., columns of $A_1'$) be $a_1', \ldots, a_q'$. Columns of $P_\Omega A_1'$ are linearly dependent, so $\sum_{i=1}^q c_i P_\Omega a_i = P_\Omega(\sum_{i=1}^q c_i a_i) = 0$ for some $\mathbf{c} \neq \mathbf{0}$. Then there exists a vector $\sum_{i=1}^q c_i a_i \in \mathcal{M}_C(A_1')$ ( $\neq 0$ since rank of $A_1$ is $q$) such that $\sum_{i=1}^q c_i a_i \perp \Omega$. i.e., $\mathcal{M}_C(A_1') \cap \Omega^\perp \neq \{0\}$. If $\text{Rank}(P_\Omega A_1') = q = \text{Rank}(A_1')$ then $\mathcal{M}_C(A_1') = \mathcal{M}_C(P_\Omega A_1') = \omega^\perp \cap \Omega \subset \Omega$.

Now let us return to the problem of finding the projection of $\tilde{Y}$ onto $\omega = \mathcal{N}(A_1) \cap \Omega$ which achieves:

$$\min_{\theta \in \mathcal{M}_C(X)} ||\tilde{Y} - \theta||^2 \text{ subject to } A_1\theta = 0.$$

From Results A and B, $\omega^\perp \cap \Omega = (\mathcal{N}(A_1) \cap \Omega)^\perp \cap \Omega = (\mathcal{M}_C(A_1') + \Omega^\perp) \cap \Omega$ and from Result E, $\omega^\perp \cap \Omega = \mathcal{M}_C(P_\Omega A_1')$. Now note that

$$P_\Omega A_1' = (X(X'X)^{-1}X')X(X'X)^{-1}A' = X(X'X)^{-1}A' = A_1'.$$

Therefore, $\text{Rank}(P_\Omega A_1') = \text{Rank}(A_1') \leq q$. However, since $\text{Rank}(P_\Omega A_1') = \text{Rank}(X(X'X)^{-1}A') \geq \text{Rank}(X'X(X'X)^{-1}A') = \text{Rank}(A') = q$, we must have $\text{Rank}(P_\Omega A_1') = q$. Therefore, from Result D,

$$\begin{aligned} P_\Omega - P_\omega &= P_{\omega^\perp \cap \Omega} = P_{\mathcal{M}_C(P_\Omega A_1')} \\ &= P_\Omega A_1'(A_1 P_\Omega^2 A_1')^{-1}(P_\Omega A_1')' \\ &= X(X'X)^{-1}A' \left[A(X'X)^{-1}X'X(X'X)^{-1}A'\right]^{-1} A(X'X)^{-1}X' \\ &= X(X'X)^{-1}A' \left(A(X'X)^{-1}A'\right)^{-1} A(X'X)^{-1}X'. \end{aligned}$$

Therefore,

$$\begin{aligned} &X\hat{\beta}_H - X\beta_0 = X\hat{\gamma}_H = P_\omega \tilde{Y} = P_\Omega \tilde{Y} - P_{\omega^\perp \cap \Omega}\tilde{Y} \\ &= P_\Omega Y - X\beta_0 - X(X'X)^{-1}A' \left(A(X'X)^{-1}A'\right)^{-1} A(X'X)^{-1}X'(Y - X\beta_0) \\ &= P_\Omega Y - X\beta_0 - X(X'X)^{-1}A' \left(A(X'X)^{-1}A'\right)^{-1} A\left((X'X)^{-1}X'Y - \beta_0\right) \\ &= P_\Omega Y - X\beta_0 - X(X'X)^{-1}A' \left(A(X'X)^{-1}A'\right)^{-1} \left(A\hat{\beta} - c\right). \end{aligned}$$

Therefore,

$$X\hat{\beta}_H = X\hat{\beta} - X(X'X)^{-1}A' \left(A(X'X)^{-1}A'\right)^{-1} \left(A\hat{\beta} - c\right).$$

Multiplying by $(X'X)^{-1}X'$ on the left, we get,

$$\hat{\beta}_H = \hat{\beta} - (X'X)^{-1}A' \left(A(X'X)^{-1}A'\right)^{-1} \left(A\hat{\beta} - c\right).$$

This yields the minimum since $||Y - X\hat{\beta}_H||^2 = ||\tilde{Y} - X\hat{\gamma}_H||^2$.

**Case of $X$ having less than full column rank**

$\text{Rank}(X_{n\times p}) = r < p$. Since only estimable linear functions $a'\beta$ can be estimated, assume $a_i'\beta$, $i = 1, 2, \ldots, q$ are estimable and $A_{q\times p} = \begin{pmatrix} a_1' \\ \vdots \\ a_q' \end{pmatrix}$. However, since $a_i' = m_i'X$ for some $m_i'$, we have $A = M_{q\times n}X_{n\times p}$. Since $A$ has rank $q$, $M$ also has rank $q$ ($\leq r$). Proceeding as before, let $\beta_0$ be any solution of $A\beta = c$. Then consider: $\tilde{Y} = Y - X\beta_0 = X(\beta - \beta_0) + \epsilon$ or $\tilde{Y} = X\gamma + \epsilon$ or

$$\tilde{Y} = \theta + \epsilon, \theta \in \mathcal{M}_C(X) = \Omega, \text{ and}$$

$M\theta = MX\gamma = A\gamma = 0$. We want to find $\hat{\beta}_H$, the least squares solution subject to $H : A\beta = c$. If $\omega = \Omega \cap \mathcal{N}(M)$, then $\omega^\perp \cap \Omega = \mathcal{M}_C(P_\Omega M')$, and $P_\Omega M' = X(X'X)^- X'M' = X(X'X)^- A'$. Further,
$MP_\Omega M' = MX(X'X)^- X'M' = A(X'X)^- A'$ is nonsingular. This is because, (since $X'P_\Omega = X'$)

$$\begin{aligned} q &= \text{Rank}(M') \geq \text{Rank}(P_\Omega M') \geq \text{Rank}(X'P_\Omega M') \\ &= \text{Rank}(X'M') = \text{Rank}(A') = q. \end{aligned}$$

Therefore

$$\begin{aligned} P_\Omega - P_\omega &= P_{\omega^\perp \cap \Omega} = P_{\mathcal{M}_C(P_\Omega M')} \\ &= P_\Omega M'(MP_\Omega M')^{-1} MP_\Omega \\ &= X(X'X)^- A' \left(A(X'X)^- A'\right)^{-1} A(X'X)^- X'. \end{aligned}$$

Hence,

$$\begin{aligned} X\hat{\beta}_H - X\beta_0 = X\hat{\gamma}_H = P_\omega \tilde{Y} = P_\Omega \tilde{Y} - P_{\omega^\perp \cap \Omega}\tilde{Y} \\ = P_\Omega Y - X\beta_0 - P_\Omega M'(MP_\Omega M')^{-1} MP_\Omega (Y - X\beta_0), \text{ so that} \end{aligned}$$

$$X'X\hat{\beta}_H - X'X\beta_0 = X'P_\Omega Y - X'X\beta_0 - X'P_\Omega M'(MP_\Omega M')^{-1} MP_\Omega (Y - X\beta_0).$$

Thus,

$$\begin{aligned} X'X\hat{\beta}_H &= X'Y - X'M'(MP_\Omega M')^{-1} \{MP_\Omega Y - MP_\Omega X\beta_0\} \\ &= X'Y - X'M'(MP_\Omega M')^{-1} \left\{MX(X'X)^- X'Y - MX\beta_0\right\} \\ &= X'Y - X'M'(MP_\Omega M')^{-1} \left\{A(X'X)^- X'Y - A\beta_0\right\} \\ &= X'Y - A' \left(A(X'X)^- A'\right)^{-1} \left\{A\hat{\beta} - c\right\} \\ &= X'X\hat{\beta} - A' \left(A(X'X)^- A'\right)^{-1} \left\{A\hat{\beta} - c\right\}. \end{aligned}$$

Now recall, a solution of $Bu = d$ is $\hat{u} = B^{-}d$. Therefore, from above, since

$$X'X(\hat{\beta}_H - \hat{\beta}) = -A' \left(A(X'X)^{-}A'\right)^{-1} \left\{A\hat{\beta} - c\right\},$$

we have that

$$\hat{\beta}_H = \hat{\beta} - (X'X)^{-}A' \left(A(X'X)^{-}A'\right)^{-1} \left\{A\hat{\beta} - c\right\}.$$

Also, these two together yield,

$$\begin{aligned}
&(\hat{\beta}_H - \hat{\beta})'X'X(\hat{\beta}_H - \hat{\beta}) \\
&\quad = \quad (A\hat{\beta} - c)' \left(A(X'X)^{-}A'\right)^{-1} A(X'X)^{-}A' \left(A(X'X)^{-}A'\right)^{-1} (A\hat{\beta} - c) \\
&\quad = \quad (A\hat{\beta} - c)' \left(A(X'X)^{-}A'\right)^{-1} (A\hat{\beta} - c).
\end{aligned}$$

## Linear Regression

Consider the model:
$Y = X\beta + \epsilon$, with $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I_n$. Then $\hat{\beta} = (X'X)^- X'Y$ is a least squares solution. If $X_{n\times p}$ has rank $p$, it is the least squares estimate of $\beta$. It is an optimal estimate in the sense that for all $a \in \mathcal{R}^p$, $a'\hat{\beta}$ is the BLUE of $a'\beta$. Also, $E(\hat{\beta}) = \beta$ and $Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$. If $X$ has rank $r < p$, $\hat{\beta}$ is still optimal in the sense that for all estimable $a'\beta$ (i.e., $a = X'b$), we still have that $a'\hat{\beta}$ is the BLUE of $a'\beta$.

If $Y \sim N_n(X\beta, \sigma^2 I_n)$, then $a'\hat{\beta} \sim N(a'\beta, \sigma^2 a'(X'X)^- a)$ and hence
$a'\hat{\beta} \pm t_{n-r}(1-\alpha/2)\sqrt{\frac{\text{RSS}}{n-r}a'(X'X)^- a}$
is a $100(1-\alpha)\%$ confidence interval for $a'\beta$ for any estimable $a'\beta$.

Now we want to explore the question: how good is the model $Y = X\beta + \epsilon$ for the given data?

### Analysis of Variance (ANOVA) for Regression

Given $\mathbf{Y}_{n\times 1}$, we look at $Y'Y = \sum_{i=1}^n y_i^2$ as its variation around 0, in the absence of any other assumptions. It has $n$ degrees of freedom. If a centre (or intercept) is considered useful, (i.e., $y_i = \beta_0 + \epsilon_i$) then we can decompose it as $\sum_{i=1}^n y_i^2 = n\bar{y}^2 + \sum_{i=1}^n (y_i - \bar{y})^2$ and check how much is the reduction in variation. If we think that the predictor set $X$ is relevant (i.e., $Y = X\beta + \epsilon$), the sum of squares $\text{SST} = Y'Y$ can be decomposed as follows:

$$\begin{aligned}
\text{SST} = Y'Y &= (Y - \hat{Y})'(Y - \hat{Y}) + \hat{Y}'\hat{Y} \\
&= Y'(I - P)Y + Y'PY \\
&= Y'(I - P)Y + \hat{\beta}'X'X\hat{\beta} \\
&= Y'Y - \hat{\beta}'X'Y + \hat{\beta}'X'Y \\
&= \text{RSS} + \text{SSR},
\end{aligned}$$

where RSS is the residual sum of squares and SSR is the sum of squares due to regression. If $X_{n\times p}$ has rank $r \le p$, then $n = (n-r) + r$ is the corresponding decomposition of the degrees of freedom. Thus, analysis of variance is simply the decomposition of total sum of squares into components which can be attributed to different factors. Then this simple minded ANOVA for $Y = X\beta + \epsilon$ will look as follows.

| source of variation | sum of squares | d.f. | mean squares | $F$-ratio |
|---|---|---|---|---|
| model: $Y = X\beta + \epsilon$ | SSR = $\hat{\beta}'X'Y$ | $r$ = Rank($X$) | MSR = SSR/$r$ | $F$ = MSR/MSE |
| residual error | SSE = $Y'Y - \hat{\beta}'X'Y$ | $n - r$ | MSE = SSE/$n - r$ | |
| Total | SST = $Y'Y$ | $n$ | | |

If $Y \sim N_n(X\beta, \sigma^2 I_n)$,
(i) $X\hat{\beta}$ is independent of SSE = RSS = $(Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'Y - \hat{\beta}'X'Y$, and
(ii) SSE = RSS $\sim \sigma^2\chi^2_{n-r}$;
(iii) if indeed the linear model is not useful, then $\beta = 0$ so that $\hat{\beta}'X'X\hat{\beta} = (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \sim \sigma^2\chi^2_r$.
Therefore, to check usefulness of the linear model, use
F = MSR/MSE $\sim F_{r,n-r}$ (if $\beta = 0$).

If $\beta \neq 0$, then $\hat{\beta}'X'X\hat{\beta} \sim$ non-central $\chi^2$ and $E(\hat{\beta}'X'X\hat{\beta}) = r\sigma^2 + \beta'X'X\beta > r\sigma^2$, so large values of F-ratio indicate evidence for $\beta \neq 0$.

However, this ANOVA is not particularly useful since (usually) the first column of $X$ is $\mathbf{1}$ indicating that the model includes an intercept or centre. This constant term is generally useful, and we only want to check $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ to check the usefulness the actual regressors, $X_1, \ldots, X_{p-1}$ (not $X_0 = \mathbf{1}$). Before discussing this, let us recall a result in probability on decomposing the variance:
If $X$ and $Y$ are jointly distributed (with finite second moments), then

$$Var(Y) = E\left[Var(Y|X)\right] + Var\left[E(Y|X)\right].$$

The first term on RHS is the 'within variation': if $Y$ is partitioned according to values of $X$, how much is left to be explained in Y for given $X$. The second term is the variation between $\hat{Y}(X)$ values, and is the 'between variation'. In a study, $Var(Y)$ may be large, but if $Var(Y|X)$ is small, it makes sense to use $X$ to predict $Y$ using $X$. This result is known as the Analysis of Variance formula, and the ANOVA for regression is based on it. Some more results are needed to derive it.

**The F-test (to check the goodness of linear models)**

We have the model, $Y = X\beta+\epsilon$, $X_{n\times p}$ of rank $r \le p$ and with $\epsilon \sim N_n(0, \sigma^2 I_n)$. Suppose we want to test $H_0 : A\beta = c$, $A_{q\times p}$ of rank $q \le r$, and $c$ is given. Then

$$\begin{aligned} \text{RSS} = \text{SSE} &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'(I - P)Y \\ \text{RSS}_{H_0} &= (Y - X\hat{\beta}_{H_0})'(Y - X\hat{\beta}_{H_0}), \text{ where} \\ \hat{\beta}_{H_0} &= \hat{\beta} + (X'X)^- A' \left(A(X'X)^- A'\right)^{-1} \left\{c - A\hat{\beta}\right\}. \end{aligned}$$

**Theorem.** Under the above mentioned assumptions, we have,
(i) $\text{RSS} \sim \sigma^2\chi^2_{n-r}$;
(ii) $\text{RSS}_{H_0} - \text{RSS} = (A\hat{\beta} - c)' \left(A(X'X)^- A'\right)^{-1} (A\hat{\beta} - c)$;
(iii) $E\left(\text{RSS}_{H_0} - \text{RSS}\right) = q\sigma^2 + (A\beta - c)' \left(A(X'X)^- A'\right)^{-1} (A\beta - c)$;
(iv) under $H_0 : A\beta = c$,

$$F = \frac{(\ \text{RSS}_{H_0}\ \text{-}\ \text{RSS}\ )/q}{\text{RSS}\ /(n-r)} \sim F_{q,n-r};$$

(v) when $c = 0$,

$$F = \left(\frac{n-r}{q}\right) \frac{Y'(P - P_{H_0})Y}{Y'(I_n - P)Y},$$

where $P_{H_0}$ is symmetric idempotent and $P_{H_0}P = PP_{H_0} = P_{H_0}$.

**Proof.** (i) Already known.
(ii) Note that

$$\begin{aligned} \text{RSS}_{H_0} &= (Y - X\hat{\beta}_{H_0})'(Y - X\hat{\beta}_{H_0}) \\ &= (Y - X\hat{\beta} + X\hat{\beta} - X\hat{\beta}_{H_0})'(Y - X\hat{\beta} + X\hat{\beta} - X\hat{\beta}_{H_0}) \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (X\hat{\beta} - X\hat{\beta}_{H_0})'(X\hat{\beta} - X\hat{\beta}_{H_0}) \\ &\quad +2(X\hat{\beta} - X\hat{\beta}_{H_0})'(Y - X\hat{\beta}) \\ &= \text{RSS} + (\hat{\beta} - \hat{\beta}_{H_0})'X'X(\hat{\beta} - \hat{\beta}_{H_0}), \end{aligned}$$

since $(X\hat{\beta} - X\hat{\beta}_{H_0})'(Y - X\hat{\beta}) = (\hat{\beta} - \hat{\beta}_{H_0})'(X'Y - X'X\hat{\beta}) = 0$. Now from an earlier result, $(\hat{\beta} - \hat{\beta}_{H_0})'X'X(\hat{\beta} - \hat{\beta}_{H_0}) = (A\hat{\beta} - c)' \left(A(X'X)^- A'\right)^{-1} (A\hat{\beta} - c)$.
(iii) $A\hat{\beta} = MX\hat{\beta} = MPY \sim N_q(A\beta, \sigma^2(A(X'X)^- A'))$, so that $E(A\hat{\beta} - c) = A\beta - c$ and $Cov(A\hat{\beta}) = \sigma^2 A(X'X)^- A'$. Therefore,

$$\begin{aligned} &E(\text{RSS}_{H_0} - \text{RSS}) \\ &= E\left\{(A\hat{\beta} - c)' \left(A(X'X)^- A'\right)^{-1} (A\hat{\beta} - c)\right\} \\ &= (A\beta - c)' \left(A(X'X)^- A'\right)^{-1} (A\beta - c) + \text{tr}\left\{\sigma^2 A(X'X)^- A' \left(A(X'X)^- A'\right)^{-1}\right\} \\ &= q\sigma^2 + (A\beta - c)' \left(A(X'X)^- A'\right)^{-1} (A\beta - c), \end{aligned}$$

which is large if $A\beta$ is far from $c$.
(iv) Note that

$$\text{RSS}_{H_0} - \text{RSS} = (A\hat{\beta} - c)' \left(A(X'X)^- A'\right)^{-1} (A\hat{\beta} - c) \sim \sigma^2 \chi^2_q,$$

under $H_0$ since $A\hat{\beta} - c \sim N_q(A\beta - c, \sigma^2(A(X'X)^- A')) = N_q(0, \sigma^2 A(X'X)^- A')$. Also, RSS $\sim \sigma^2 \chi^2_{n-r}$ from (i). Further, RSS is independent of $X\hat{\beta} = PY$. Since $A\beta$ is estimable, $A = MX$, so that $A\hat{\beta} = MX\hat{\beta} = MPY$, which is independent of RSS.
(v) If $c = 0$, we have,

$$\begin{aligned} X\hat{\beta}_{H_0} &= X\left\{\hat{\beta} - (X'X)^- A' \left(A(X'X)^- A'\right)^{-1} A\hat{\beta}\right\} \\ &= X\left\{(X'X)^- X'Y - (X'X)^- A' \left(A(X'X)^- A'\right)^{-1} A(X'X)^- X'Y\right\} \\ &= \left\{X(X'X)^- X' - X(X'X)^- A' \left(A(X'X)^- A'\right)^{-1} A(X'X)^- X'\right\} Y \\ &= (P - P_1)Y = P_{H_0} Y. \end{aligned}$$

Clearly, $P_{H_0}$ is symmetric. Further, $P_1$ is symmetric, $P_1^2 =$
$X(X'X)^- A' \left(A(X'X)^- A'\right)^{-1} A(X'X)^- X'X(X'X)^- A' \left(A(X'X)^- A'\right)^{-1} A(X'X)^- X' =$
$X(X'X)^- A' \left(A(X'X)^- A'\right)^{-1} \{A(X'X)^- X'X(X'X)^- A'\} \left(A(X'X)^- A'\right)^{-1} A(X'X)^- X'$
$= X(X'X)^- A' \left(A(X'X)^- A'\right)^{-1} \{A(X'X)^- A'\} \left(A(X'X)^- A'\right)^{-1} A(X'X)^- X'$
$= X(X'X)^- A' \left(A(X'X)^- A'\right)^{-1} A(X'X)^- X' = P_1$, since the term in the middle of the expression,
$A(X'X)^- X'X(X'X)^- A' = MX(X'X)^- X'X(X'X)^- X'M' = MP^2M'$
$= MPM' = A(X'X)^- A'$. Also,
$P_1 P = X(X'X)^- A' \left(A(X'X)^- A'\right)^{-1} A(X'X)^- X'X(X'X)^- X'$
$= X(X'X)^- A' \left(A(X'X)^- A'\right)^{-1} A(X'X)^- X' = P_1$,
since $X'X(X'X)^- X' = X'P = X'P' = (PX)' = X'$. Note, $P_1 = (P_1)' = (P_1P)' = PP_1$. Therefore,
$P_{H_0}^2 = (P - P_1)^2 = P^2 - PP_1 - P_1P + P_1^2 = P - 2P_1 + P_1 = P - P_1 = P_{H_0}$
and $P_{H_0}P = (P - P_1)P = P - P_1 = P_{H_0} = PP_{H_0}$. Therefore,

$$\begin{aligned} \text{RSS}_{H_0} &= ||Y - X\hat{\beta}_{H_0}||^2 = (Y - X\hat{\beta}_{H_0})'(Y - X\hat{\beta}_{H_0}) \\ &= (Y - P_{H_0}Y)'(Y - P_{H_0}Y) = Y'(I - P_{H_0})Y \end{aligned}$$

and

$$\text{RSS}_{H_0} - \text{RSS} = Y'(I - P_{H_0})Y - Y'(I - P)Y = Y'(P - P_{H_0})Y.$$

Now we use the above result for checking the goodness of the linear fit. ANOVA for checking the goodness of $Y = X\beta + \epsilon$, or $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{i(p-1)} + \epsilon_i$, or equivalently for testing $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ is what is needed. Intuitively, if $X_1, \ldots, X_{p-1}$ provide no useful information, then the appropriate model is $y_i = \beta_0 + \epsilon_i$, so $\bar{y}$ is the only quantity that can help in predicting $y$. Then $\mathrm{RSS}_{H_0} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sum of squares unexplained, and it has $n-1$ d.f. If $X_1, \ldots, X_{p-1}$ are also used in the model, then $(Y - X\hat{\beta})'(Y - X\hat{\beta}) = \mathrm{RSS}$ is the unexplained part with $n - r$ d.f. How much better is RSS compared to $\mathrm{RSS}_{H_0}$? Let $\mathrm{SS}_{reg}$ denote the sum of squares due to $X_1, \ldots, X_{p-1}$ and without an intercept. Then,

$$\begin{aligned} \mathrm{RSS}_{H_0} &= \mathrm{RSS} + \mathrm{SS}_{reg} \\ \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \mathrm{SS}_{reg} \end{aligned}$$

In other words,

$$\begin{aligned} Y'Y - \frac{1}{n}Y'1'1Y &= Y'(I-P)Y + \mathrm{SS}_{reg}, \text{ or} \\ Y'Y &= Y'(I-P)Y + \left(\mathrm{SS}_{reg} + \frac{1}{n}Y'1'1Y\right), \text{ or} \\ \mathrm{SSR} &= \hat{\beta}'X'X\hat{\beta} = \hat{\beta}'X'Y = \left(\mathrm{SS}_{reg} + \frac{1}{n}Y'1'1Y\right), \end{aligned}$$

since $Y'Y = Y'(I-P)Y + Y'PY = Y'(I-P)Y + \hat{\beta}'X'X\hat{\beta}$. Now, $\mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' = \frac{1}{n}\mathbf{1}\mathbf{1}' = P_{\mathcal{M}(\mathbf{1})} = P_{\mathcal{M}(X_0)}$, so that $\mathrm{SSR} = n\bar{y}^2 + \mathrm{SS}_{reg}$ is the orthogonal decomposition of SSR into components attributed to $\mathcal{M}(\mathbf{1})$ and $\mathcal{M}(X_1, \ldots, X_{p-1})$. Therefore $\mathrm{SS}_{reg}$ with $r-1$ d.f. is the quantity to measure the merit of the regressors, $X_1, \ldots, X_{p-1}$.

ANOVA with mean

| source of variation | d.f. | sum of squares | mean squares | $F$-ratio |
|---|---|---|---|---|
| mean | 1 | $\mathrm{SSM} = n\bar{y}^2$ | $\mathrm{MSM} = \mathrm{SSM}/1$ | $F_{mean} = \mathrm{MSM}/\mathrm{MSE}$ |
| regression on $X_1, \ldots, X_{p-1}$ | $r-1$ | $\mathrm{SS}_{reg} = \hat{\beta}'X'Y - n\bar{y}^2$ | $\mathrm{MS}_{reg} = \mathrm{SS}_{reg}/(r-1)$ | $F_{reg} = \mathrm{MS}_{reg}/\mathrm{MSE}$ |
| residual error | $n-r$ | $\mathrm{SSE} = \mathrm{RSS} = Y'Y - \hat{\beta}'X'Y$ | $\mathrm{MSE} = \mathrm{SSE}/(n-r)$ | |
| Total | $n$ | $\mathrm{SST} = Y'Y$ | | |

ANOVA for regression (corrected for mean)

| source of variation | d.f. | sum of squares | mean squares | $F$-ratio |
|---|---|---|---|---|
| regression (corrected) | $r-1$ | $\text{SS}_{reg} = \hat{\beta}'X'Y - n\bar{y}^2$ | $\text{MS}_{reg} = \text{SS}_{reg}/(r-1)$ | $F_{reg} = \text{MS}_{reg}/\text{MSE}$ |
| residual error | $n-r$ | $\text{SSE} = \text{RSS} = Y'Y - \hat{\beta}'X'Y$ | $\text{MSE} = \text{SSE}/(n-r)$ | |
| Total (corrected) | $n-1$ | $\text{SST(Corrected)} = \sum(y_i - \bar{y})^2$ | | |

How good is the linear fit? There are two things to consider here.
(i) The ANOVA F-test: Under $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$, the F-ratio, $F_{reg} \sim F_{r-1,n-r}$ and large values of the statistic provide evidence against $H_0$, or equivalently indicate that the regressors are useful.
(ii) The proportion of variability in $y$ not explained by the actual regressors is: RSS/SST (corrected), so the proportion of variability in $y$ around its mean, explained by the actual regressors is

$$1 - \frac{\text{RSS}}{\text{SST (corrected)}} \equiv R^2 = \text{ Coefficient of determination.}$$

In other words,

$$\begin{aligned}
R^2 &= 1 - \frac{\text{RSS}}{\text{SST (corrected)}} = 1 - \frac{Y'(I-P)Y}{Y'(I - \frac{1}{n}\mathbf{1}\mathbf{1}')Y} \\
&= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - Y'(I-P)Y}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n y_i^2 - n\bar{y}^2 - Y'(I-P)Y}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{Y'Y - n\bar{y}^2 - Y'Y + Y'PY}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{Y'PY - n\bar{y}^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\text{SSR } - n\bar{y}^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\text{SS}_{reg}}{\text{SST (corrected)}} \\
&= \text{ proportion of variability explained by regressors.}
\end{aligned}$$

Also,

$$\begin{aligned}
R^2 &= \frac{\text{SS}_{reg}}{\text{SST (corrected)}} = \frac{\text{SS}_{reg}}{\text{RSS} + \text{SS}_{reg}} \\
&= \frac{\text{SS}_{reg}/\text{RSS}}{1 + \text{ SS}_{reg}/\text{RSS}} = \frac{(\frac{r-1}{n-r})F_{reg}}{1 + (\frac{r-1}{n-r})F_{reg}}
\end{aligned}$$

is an increasing function of the F-ratio.

Note that to interpret the F-ratio, normality of $\epsilon_i$ is needed. $R^2$, however, is a percentage with a straightforward interpretation.

**Example 1 (socio-economic study).** The demand for a consumer product is affected by many factors. In one study, measurements on the relative urbanization ($X_1$), educational level ($X_2$), and relative income ($X_3$) of 9 randomly chosen geographic regions were obtained in an attempt to determine their effect on the product usage ($Y$). The data were:

| $X_1$ | $X_2$ | $X_3$ | $Y$ |
|---|---|---|---|
| 42.2 | 11.2 | 31.9 | 167.1 |
| 48.6 | 10.6 | 13.2 | 174.4 |
| 42.6 | 10.6 | 28.7 | 160.8 |
| 39.0 | 10.4 | 26.1 | 162.0 |
| 34.7 | 9.3 | 30.1 | 140.8 |
| 44.5 | 10.8 | 8.5 | 174.6 |
| 39.1 | 10.7 | 24.3 | 163.7 |
| 40.1 | 10.0 | 18.6 | 174.5 |
| 45.9 | 12.0 | 20.4 | 185.7 |

We fit the model: $Y = X\beta + \epsilon$, with $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I_n$. In this case, $n = 9$, $p = 4$. $\bar{y} = 167.07$ and the model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$.

We get $\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} 60.0 \\ 0.24 \\ 10.72 \\ -0.75 \end{pmatrix}$. The detailed ANOVA (with mean) is

| source | d.f. | SS | MS | $F$-ratio |
|---|---|---|---|---|
| mean | 1 | SSM $= n\bar{y}^2 =$ 251201.44 | MSM $=$ 251201.44 | |
| regression $(X_1, X_2, X_3)$ | 3 | $\text{SS}_{reg} =$ 1081.35 | $\text{MS}_{reg} =$ 360.45 | $F_{reg} = \frac{360.45}{39.57} = 9.11$ |
| residual error | 5 | SSE $=$ RSS $=$ 197.85 | MSE $=$ 39.57 | |
| Total (corrected) | 8 | 1279.20 | | |
| Total | 9 | 252480.64 | | |

From this note that $s^2 = \text{RSS}/(n-r) = \text{MSE} = 39.57$, so $s = 6.29 = \hat{\sigma}$, and $R^2 = 1081.35/1279.20 = 84.5\%$. Abridged ANOVA is

| source | d.f. | SS | MS | $F$-ratio |
|---|---|---|---|---|
| regression $(X_1, X_2, X_3)$ | 3 | $\text{SS}_{reg} =$ 1081.35 | $\text{MS}_{reg} =$ 360.45 | $F_{reg} = \frac{360.45}{39.57} = 9.11$ |
| residual error | 5 | SSE = RSS = 197.85 | MSE = 39.57 | |
| Total (corrected) | 8 | 1279.20 | | |

$R^2 = 84.5\%$ is substantial. What about $F = 9.11$? $F_{3,5}(.95) = 5.41$ and $F_{3,5}(.99) = 12.06$, so there is some evidence against the null and justifying the linear fit.

**Example 2.** $X$ = height (cm) and $Y$ = weight (kg) for a sample of $n = 10$ eighteen-year-old American girls:

| $X$ | $Y$ |
|---|---|
| 169.6 | 71.2 |
| 166.8 | 58.2 |
| 157.1 | 56.0 |
| 181.1 | 64.5 |
| 158.4 | 53.0 |
| 165.6 | 52.4 |
| 166.7 | 56.8 |
| 156.5 | 49.2 |
| 168.1 | 55.6 |
| 165.3 | 77.8 |

Upon fitting the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, we get $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} -36.9 \\ 0.582 \end{pmatrix}$, $s^2 = MSE = 71.50$, $s = 8.456$, $R^2 = 21.9\%$, $\bar{y} = 59.47$. ANOVA is

| source | d.f. | SS | MS | F | $R^2$ |
|---|---|---|---|---|---|
| $X$ | 1 | 159.95 | 159.95 | 2.24 | 21.9% |
| error | 8 | 512.01 | 71.50 | | |
| Total (C) | 9 | 731.96 | | | |

Note the following. (i) $X$ is expected to be a useful predictor of $Y$, but the relationship may not be simple. (ii) $F_{1,8}(.90) = 3.46 = (1.86)^2 = t_8^2(.95)$, so is there a connection between the ANOVA F-test and a t-test?

Consider simple linear regression again: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$, $\epsilon_i$ i.i.d. $N(0, \sigma^2)$. Then the F-ratio is the $F$ statistic for testing the goodness of fit of the linear model, or for testing $H_0 : \beta_1 = 0$. Writing the linear model

in the standard form, we have

$$X_{n\times 2} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \quad X'X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}, \text{ and}$$

$$(X'X)^{-1} = \frac{1}{n\sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}.$$

Therefore

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X'X)^{-1}X'Y = \frac{1}{n\sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

Letting $S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2$, $S_{XY} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$, and extracting the least squares equations, we get,

$$\begin{aligned} \hat{\beta}_1 &= \frac{1}{S_{XX}} \left\{ -n\bar{x}\bar{y} + \sum_{i=1}^n x_i y_i \right\} = \frac{S_{XY}}{S_{XX}}, \\ \hat{\beta}_0 &= \frac{1}{S_{XX}} \left\{ \bar{y} \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i y_i \right\} = \frac{1}{S_{XX}} \left\{ \bar{y} S_{XX} + n\bar{y}\bar{x}^2 - \bar{x} \sum_{i=1}^n x_i y_i \right\} \\ &= \frac{1}{S_{XX}} \left\{ \bar{y} S_{XX} - \bar{x} \left( \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \right) \right\} = \bar{y} - \bar{x}\hat{\beta}_1. \end{aligned}$$

Now, $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{XX})$, so that, to test $H_0 : \beta_1 = 0$, use the test statistic,

$$\frac{\sqrt{S_{XX}}\hat{\beta}_1}{\sqrt{\text{RSS}/(n-2)}} \sim t_{n-2}, \qquad \text{or} \qquad \frac{\hat{\beta}_1^2 S_{XX}}{\text{MSE}} \sim F_{1,n-2},$$

if $H_0$ is true. The ANOVA table shows that

$$\begin{aligned} \sum_{i=1}^n y_i^2 &= n\bar{y}^2 + \sum_{i=1}^n (y_i - \bar{y})^2 = n\bar{y}^2 + \text{RSS} + \text{SS}_{reg}, \text{ so} \\ \text{SS}_{reg} &= \sum_{i=1}^n (y_i - \bar{y})^2 - \text{RSS}. \end{aligned}$$

However,

$$\begin{aligned}
\text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n}\Big(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})\Big)^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 + \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 + \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 - 2\hat{\beta}_1\hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2.
\end{aligned}$$

Therefore, $\text{SS}_{reg} = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$, so that

$$t^2 = \frac{\hat{\beta}_1^2 S_{XX}}{\text{RSS}/(n-2)} = \text{ F-ratio of ANOVA.}$$

In Example 1, F-ratio tests $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. What if we want to test only $\beta_1 = \beta_3 = 0$? Then we have $H_0 : A\beta = 0$, where $A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}_{2\times 4}$ if of rank 2. Then apply the theorem: $\text{RSS}_{H_0} = (Y - X\hat{\beta}_{H_0})'(Y - X\hat{\beta}_{H_0})$ where $\hat{\beta}_{H_0} = \hat{\beta} + (X'X)^{-}A'(A(X'X)^{-}A')^{-1}(c - A\hat{\beta})$ and the test statistic is

$$F = \frac{(\text{RSS}_{H_0} - \text{RSS})/q}{\text{RSS}/(n-r)} \sim F_{q,n-r} \text{ under } H_0.$$

## Multiple Correlation

As seen earlier, the proportion of variation explained by the linear regression of $Y$ on the regressors $X_1, \ldots, X_{p-1}$ is given by

$$R^2 = \frac{\text{SS}_{reg}}{\text{SST (corrected)}} = 1 - \frac{\text{RSS}}{\text{SST (corrected)}} = 1 - \frac{Y'(I-P)Y}{Y'(I - \frac{1}{n}\mathbf{1}\mathbf{1}')Y}.$$

Consider simple linear regression: Then $p = 2$ and $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\text{RSS} = \sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2,$$

so that

$$\text{SS}_{reg} = \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\left\{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right\}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Therefore,

$$\begin{aligned} R^2 &= \frac{\text{SS}_{reg}}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\left\{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right\}^2}{\left\{\sum_{i=1}^n (x_i - \bar{x})^2\right\}\left\{\sum_{i=1}^n (y_i - \bar{y})^2\right\}} \\ &= \left\{\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left\{\sum_{i=1}^n (x_i - \bar{x})^2\right\}\left\{\sum_{i=1}^n (y_i - \bar{y})^2\right\}}}\right\}^2 = r_{XY}^2, \end{aligned}$$

where

$$\begin{aligned} r_{XY} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \text{sample correlation coefficient between } X \text{ and } Y. \end{aligned}$$

This connection between $R^2$ and $r^2$ is intuitively meaningful since a good linear fit is related to a good linear association between $X$ and $Y$. What happens when there are multiple regressors, $X_1, X_2, \ldots, X_{p-1}$?

We define the *multiple correlation coefficient* between $Y$ and $X_1, \ldots, X_{p-1}$ as the *maximum* correlation coefficient between $Y$ and any linear function of $X_1, \ldots, X_{p-1} = \max_{\mathbf{a}} Corr(Y, a_0 + a_1 X_1 + \cdots + a_{p-1} X_{p-1}) = R^*$ (say).

If $Cov\left(\begin{pmatrix} Y \\ \mathbf{X} \end{pmatrix}\right) = \begin{pmatrix} \sigma_{YY} & \sigma'_{XY} \\ \sigma_{XY} & \Sigma_X \end{pmatrix}$, then

$$Corr^2(Y, a'X) = \frac{Cov^2(Y, a'X)}{Var(Y)Var(a'X)} = \frac{\{a'Cov(Y,X)\}^2}{Var(Y)Var(a'X)} = \frac{\{a'\sigma_{XY}\}^2}{\sigma_{YY}a'\Sigma_X a}.$$

Further, taking $u' = a'\Sigma_X^{1/2}$ and $v = \Sigma_X^{-1/2}\sigma_{XY}$,

$$\begin{aligned} \frac{a'\sigma_{XY}}{(\sigma_{YY}a'\Sigma_X a)^{1/2}} &= \frac{a'\Sigma_X^{1/2}\Sigma_X^{-1/2}\sigma_{XY}}{(\sigma_{YY}a'\Sigma_X a)^{1/2}} = \frac{u'v}{(\sigma_{YY}a'\Sigma_X a)^{1/2}} \\ &\leq \frac{(u'u)^{1/2}(v'v)^{1/2}}{(\sigma_{YY}a'\Sigma_X a)^{1/2}} = \frac{(a'\Sigma_X a)^{1/2}\left(\sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}\right)^{1/2}}{(\sigma_{YY}a'\Sigma_X a)^{1/2}} \\ &= \left(\frac{\sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}}{\sigma_{YY}}\right)^{1/2}, \end{aligned}$$

with equality if we take $u \propto v$ or $a = \Sigma_X^{-1}\sigma_{XY}$. Since $R^* = \sqrt{\sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}/\sigma_{YY}}$, $0 \leq R^* \leq 1$ unlike the ordinary correlation coefficient. Now let us see why $(R^*)^2$ (square of multiple correlation coefficient) is the same as the coefficient of determination, $R^2$ (proportion of variability explained by the regressors). Suppose

$$\begin{pmatrix} Y \\ \mathbf{X} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_{YY} & \sigma'_{XY} \\ \sigma_{XY} & \Sigma_X \end{pmatrix}\right).$$

Then,

$$Y|\mathbf{X} \sim N\left(\mu_Y + \sigma'_{XY}\Sigma_X^{-1}(\mathbf{X} - \mu_X), \sigma_{YY} - \sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}\right).$$

Thus, $E(Y|\mathbf{X}) = \mu_Y - \sigma'_{XY}\Sigma_X^{-1}\mu_X + \sigma'_{XY}\Sigma_X^{-1}\mathbf{X}$ and
$Var(Y|\mathbf{X}) = \sigma_{YY} - \sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}$). Therefore,

$$\begin{aligned} Corr(Y, E(Y|\mathbf{X})) &= \frac{Cov(Y, \sigma'_{XY}\Sigma_X^{-1}\mathbf{X})}{\sqrt{\sigma_{YY}\sigma'_{XY}\Sigma_X^{-1}\Sigma_X\Sigma_X^{-1}\sigma_{XY}}} \\ &= \frac{\sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}}{\sqrt{\sigma_{YY}}\sqrt{\sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}}} = R^*. \end{aligned}$$

i.e., $R^*$ = correlation coefficient between $Y$ and the conditional expectation of $Y|\mathbf{X}$ (or the regression of $Y$ on $\mathbf{X}$, when the conditional expectation is linear). Further, $Var(Y) - E\left(Var(Y|\mathbf{X})\right) = \sigma_{YY} - \left(\sigma_{YY} - \sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}\right) = \sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}$, so that the proportion of variation in $Y$ explained by the regression on $\mathbf{X}$ is equal to

$$R^2 = \frac{Var(Y) - E\left(Var(Y|\mathbf{X})\right)}{Var(Y)} = \frac{\sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}}{\sigma_{YY}} = (R^*)^2.$$

## Partial Correlation Coefficients

**Example.** In a study, $X_1$ = weekly amount of coffee/tea sold by a refreshment stand at a summer resort, and $X_2$ = weekly number of visitors to the resort. If $X_2$ is large, so should $X_1$ be, right? Actually no! With a certain resort, $r_{12} = -0.3$. Why? Consider $X_3$ = average weekly temperature at the resort. Both $X_1$ and $X_2$ are related to $X_3$. If temperature is high, there will be more visitors, but they will prefer cold drinks to coffee/tea. If temperature is low, there will be fewer visitors, but they will prefer coffee/tea. Say, $r_{13} = -0.7$, $r_{23} = .8$. It is then more meaningful to investigate the relationship between $X_1$ and $X_2$ conditional on $X_3$ (i.e., when $X_3$ is kept fixed) to eliminate the effect of $X_3$.

Partial correlation coefficient between $X_1$ and $X_2$ when $X_3$ is fixed is

$$r_{12.3} = Corr(X_1|X_3, X_2|X_3) = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}.$$

Suppose $\mathbf{X} \sim N_m(\mu, \Sigma)$ and partition $\mathbf{X}$, $\mu$ and $\Sigma$ as:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix},$$

where $\mathbf{X}_1$ is $k$-dimensional. Then $\mathbf{X}_1|\mathbf{X}_2 \sim N_k(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_2 - \mu_2), \Sigma_{11.2})$, where $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}' = ((\sigma_{ij.k+1,\ldots,m}))$. Note that $\sigma_{ij.k+1,\ldots,m}$ = partial covariance between $X_i$ and $X_j$ conditional on $\mathbf{X}_2 = (X_{k+1}, \ldots, X_m)'$. Therefore the partial correlation coefficient between $X_i$ and $X_j$ given $\mathbf{X}_2$ is

$$\rho_{ij.k+1,\ldots,m} = \frac{\sigma_{ij.k+1,\ldots,m}}{\sqrt{\sigma_{ii.k+1,\ldots,m}}\sqrt{\sigma_{jj.k+1,\ldots,m}}}.$$

Recall the notation, $\rho$ for the population and $r$ for a sample. From the expression for $\Sigma_{11.2}$ note that $\sigma_{ij.l} = \sigma_{ij} - \sigma_{il}\sigma_{jl}/\sigma_{ll}$. Thus,

$$\begin{aligned} \rho_{ij.l} &= \frac{\sigma_{ij.l}}{\sqrt{\sigma_{ii.l}}\sqrt{\sigma_{jj.l}}} = \frac{\sigma_{ij} - \frac{\sigma_{il}\sigma_{jl}}{\sigma_{ll}}}{\sqrt{\left(\sigma_{ii} - \frac{\sigma_{il}^2}{\sigma_{ll}}\right)\left(\sigma_{jj} - \frac{\sigma_{jl}^2}{\sigma_{ll}}\right)}} \\ &= \frac{\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} - \frac{\sigma_{il}\sigma_{jl}}{\sigma_{ll}\sqrt{\sigma_{ii}\sigma_{jj}}}}{\sqrt{\left(1 - \frac{\sigma_{il}^2}{\sigma_{ii}\sigma_{ll}}\right)\left(1 - \frac{\sigma_{jl}^2}{\sigma_{jj}\sigma_{ll}}\right)}} = \frac{\rho_{ij} - \rho_{il}\rho_{jl}}{\sqrt{\left(1-\rho_{il}^2\right)\left(1-\rho_{jl}^2\right)}}. \end{aligned}$$

## Simultaneous confidence sets

When we have a scalar parameter, such as the mean $\mu$ of $X$, we can construct a confidence interval for it using a sample of observations: $\bar{X} \pm \frac{s}{\sqrt{n}} t_{n-1}(1-\alpha/2)$. What about the vector $\beta$ of regression coefficients? We know that if $Y = X\beta + \epsilon$, where $\epsilon \sim N_n(0, \sigma^2 I_n)$, then $(\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta) \sim \sigma^2 \chi^2_r$ independent of $\text{RSS} = Y'(I - P)Y \sim \sigma^2 \chi^2_{n-r}$, so that

$$\frac{(\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta)/r}{Y'(I - P)Y/(n - r)} \sim F_{r,n-r},$$

and hence

$$P\left( (\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta) \leq \frac{r}{n-r} \;\; Y'(I-P)Y \;\; F_{r,n-r}(1-\alpha) \right) = 1 - \alpha.$$

Therefore,

$$C = \left\{ \beta : (\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) \leq \frac{r}{n-r} \;\; \text{RSS} \;\; F_{r,n-r}(1-\alpha) \right\}$$

is a $100(1-\alpha)\%$ confidence set for $\beta$. This is an ellipsoid, and if $p$ is not small (1 or 2), a set which is difficult to appreciate.

Suppose we are only interested in $a'\beta$ for some fixed $a$. Then $a'\hat{\beta} \pm t_{n-r}(1-\alpha/2)\sqrt{\text{RSS}/(n-r)}\sqrt{a'(X'X)^{-}a}$ is a $100(1-\alpha)\%$ confidence interval for $a'\beta$. Let us see if we can extend this when we are interested in deriving a simutaneous confidence set of coefficient $1-\alpha$ for $a_1'\beta, a_2'\beta, \ldots, a_k'\beta$.

**Scheffe's method.**

Let $A'_{p\times d} = (a_1, a_2, \ldots, a_d)$ where $a_1, a_2, \ldots, a_d$ are linearly independent and $a_{d+1}, \ldots, a_k$ are linearly dependent on them. Then $d \leq \min\{k, r\}$. Let $\phi = A\beta$ and $\hat{\phi} = A\hat{\beta}$. Then

$$F(\beta) = \frac{(\hat{\phi} - \phi)' \left(A(X'X)^- A'\right)^{-1} (\hat{\phi} - \phi)/d}{\text{RSS}/(n-r)} \sim F_{d,n-r}.$$

Therefore,

$$\begin{aligned} 1-\alpha &= P[F(\beta) \leq F_{d,n-r}(1-\alpha)] \\ &= P\left\{(\hat{\phi} - \phi)' \left(A(X'X)^- A'\right)^{-1} (\hat{\phi} - \phi) \leq d\frac{\text{RSS}}{n-r} F_{d,n-r}(1-\alpha)\right\}. \end{aligned}$$

This gives an ellipsoid as before, but consider the following result.

**Result.** If $L$ is positive definite,

$$b'L^{-1}b = \sup_{h\neq 0} \frac{(h'b)^2}{h'Lh}.$$

**Proof.** Note that

$$\frac{(h'b)^2}{h'Lh} = \frac{(h'L^{1/2}L^{-1/2}b)^2}{h'Lh} \leq \frac{h'Lhb'L^{-1}b}{h'Lh} = b'L^{-1}b.$$

Therefore,

$$1-\alpha = P\left\{\sup_{h\neq 0} \frac{\left\{h'(\phi - \hat{\phi})\right\}^2}{h'\left(A(X'X)^- A'\right)h} \leq \frac{d}{n-r}\text{RSS}\, F_{d,n-r}(1-\alpha)\right\}$$

$$= P\left\{\frac{\left\{h'(\phi - \hat{\phi})\right\}^2}{h'\left(A(X'X)^- A'\right)h} \leq \frac{d}{n-r}\text{RSS}\, F_{d,n-r}(1-\alpha) \text{ for all } h \neq 0.\right\}$$

$$= P\left\{\frac{|h'(\phi - \hat{\phi})|}{\sqrt{\frac{\text{RSS}}{n-r}}\sqrt{h'\left(A(X'X)^- A'\right)h}} \leq \{dF_{d,n-r}(1-\alpha)\}^{1/2} \text{ for all } h \neq 0.\right\}$$

$$= P\left\{|h'(\phi - \hat{\phi})| \leq \{dF_{d,n-r}(1-\alpha)\}^{1/2} \text{ s.e.}(h'\hat{\phi}) \text{ for all } h \neq 0.\right\},$$

where $\text{s.e.}(h'\hat{\phi}) = \sqrt{\frac{\text{RSS}}{n-r}}\sqrt{h'\left(A(X'X)^- A'\right)h}$. Therefore,

$$a_i'\hat{\beta} \pm \{dF_{d,n-r}(1-\alpha)\}^{1/2} \sqrt{\frac{\text{RSS}}{n-r}}\sqrt{a_i'(X'X)^- a_i}, i = 1, 2, \ldots, k$$

is a simultaneous $100(1-\alpha)\%$ confidence set for $a_1'\beta, a_2'\beta, \ldots, a_k'\beta$, by noting that

$$P\left(a_i'\beta \in a_i'\hat{\beta} \pm \{dF_{d,n-r}(1-\alpha)\}^{1/2} \text{ s.e.}(a_i'\hat{\beta}, i = 1, 2, \ldots, k\right) \geq$$

$$P\left\{|h'(\phi - \hat{\phi})| \leq \{dF_{d,n-r}(1-\alpha)\}^{1/2} \text{ s.e.}(h'\hat{\phi}) \text{ for all } h \neq 0.\right\} = 1 - \alpha.$$

Many other methods are also available.

## Regression diagnostics

**Lack of fit.** Suppose the true model is $Y = f(X) + \epsilon$, $\epsilon \sim N_n(0, \sigma^2 I_n)$, whereas we fit $Y = X\beta + \epsilon$. We do get $\hat{\beta} = (X'X)^{-1}X'Y$ and $\hat{\sigma}^2 = \text{RSS}/(n-r)$. $\sigma^2$ is supposed to account for only the statistical errors $(\epsilon_i)$, and not model misspecification. Therefore, if $f(X) \neq X\beta$, we have statistical errors, $\epsilon_i$, as well as the bias, $f(X) - X\beta$. Then, $\hat{\sigma}^2 = \text{RSS}/(n-r)$ will estimate a quantity which includes $\sigma^2$ as well as (bias)$^2$. If $\sigma^2$ is known, then comparing $\hat{\sigma}^2$ with $\sigma^2$ can act as a check for lack of fit. In other words,
$\text{RSS}/\sigma^2 \sim \chi^2_{n-r}$ if the model, $Y = X\beta + \epsilon$, $\epsilon \sim N_n(0, \sigma^2 I_n)$ is true. Therefore to test
$H_0 : Y = X\beta + \epsilon, \epsilon \sim N_n(0, \sigma^2 I_n)$ versus $H_1 : Y$ has some other model, use $\text{RSS}/\sigma^2$ as the test statistic. If the observed value is too large compared to $\chi^2_{n-r}$, there is evidence against $H_0$.

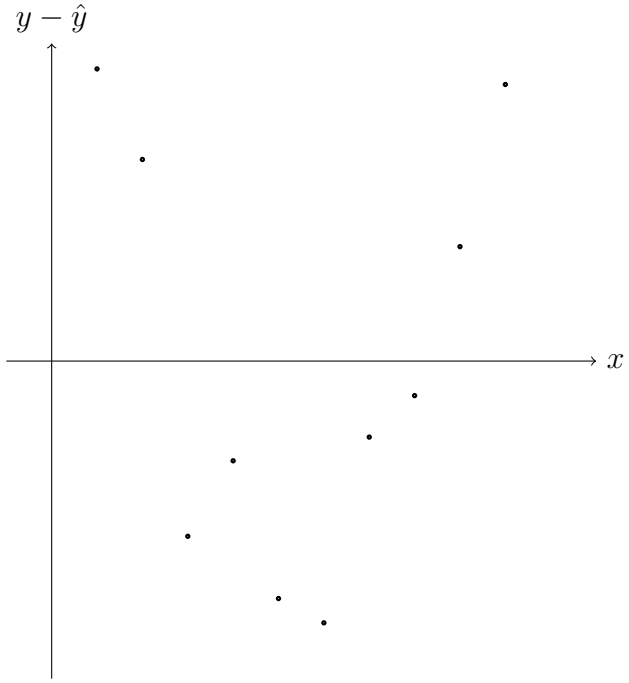
Consider a simulation study where data are generated from $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, with $\beta_0 = 5$, $\beta_1 = \beta_2 = 2$ and $\sigma^2 = 2^2$:

| $x$ | .5 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 8.68 | 12.85 | 10.71 | 18.54 | 21.67 | 27.3 | 37.56 | 44.64 | 54.09 | 63.83 |

Regress $Y$ on $X$. i.e., fit $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Then we get $\hat{\beta}_0 = -3.925$, $\hat{\beta}_1 = 12.33$ and the ANOVA table:

| source | d.f | SS | MS | F | $R^2$ |
|---|---|---|---|---|---|
| Regression | 1 | 3134.2 | 3134.2 | 130.76 | 94.2% |
| Error | 8 | 191.7 | 24.0 | | |
| Total | 9 | 3325.9 | | | |

These are very good results, but $\text{RSS}/\sigma^2 = 191.7/4 = 47.925 >> \chi^2_8(.99) = 20.08$. $R^2 = 94.2\%$ is high, and F-ratio of 130.76 at (1, 8) d.f. is very high, indicating that $X$ is a very useful predictor of $Y$. However this does not mean that the fitted model is the correct one. Check the residual plot:




Now regress $Y$ on $X$ and $X^2$.

| source | d.f | SS | MS | F | $R^2$ |
|---|---|---|---|---|---|
| Regression | 2 | 3305.7 | 1652.8 | 572.28 | 99.4% |
| Error | 7 | 20.2 | 2.9 | | |
| Total | 9 | 3325.9 | | | |

$\text{RSS}/\sigma^2 = 20.2/4 = 5.5 << \chi^2_7(.90) = 12.02.$

$\sigma^2$ is usually unknown, so this test is difficult, but what this indicates is that residual plots are useful for checking lack of fit (see plot above). Another possibility is to check for any pattern between fitted values and residuals. Yet another reason to explore this is the following.
$\hat{\epsilon} = Y - \hat{Y} = Y - X\hat{\beta} = (I - P)Y$ and $\hat{Y} = X\hat{\beta} = PY$ are uncorrelated (since $(I - P)P = 0$) if $Cov(Y) = \sigma^2 I_n$. If one sees significant correlation and some trend, then the model is suspect. What if $Var(y_i) = \sigma_i^2$, not a constant? This is called heteroscedasticity (as against homoscedastcity), a problem discussed in Sanford Weisberg: *Applied Linear Regression* in the context of regression diagnostics.

With the model: $Y = X\beta + \epsilon$, with $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I_n$, normality of $\epsilon$ is essential for hypothesis testing and confidence statements. How does one check this?

**Normal probability plot or Q-Q plot.**

This is a graphical technique to check for normality. Suppose we have a random sample $T_1, T_2, \ldots, T_n$ from some population, and we want to check whether the population has the normal distribution with some mean $\mu$ and some variance $\sigma^2$. The method described here depends on examining the order statistics, $T_{(1)}, \ldots, T_{(n)}$. Let us recall a few facts about order statistics from a continuous distribution. Since

$$\begin{aligned} f_{T_1,\ldots,T_n}(t_1,\ldots,t_n) &= \prod_{i=1}^{n} f(t_i), \quad (t_1,\ldots,t_n) \in \mathcal{R}^n, \\ f_{T_{(1)},\ldots,T_{(n)}}(t_{(1)},\ldots,t_{(n)}) &= n! \prod_{i=1}^{n} f(t_{(i)}), \quad t_{(1)} < t_{(2)} < \cdots t_{(n)}, \\ f_{T_{(i)}} &= \frac{n!}{(i-1)!(n-i)!} \left(1 - F(t_{(i)})\right)^{n-i} F^{i-1}(t_{(i)}) f(t_{(i)}). \end{aligned}$$

If $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$ are o.s. from $U(0,1)$, then

$$\begin{aligned} E(U_{(k)}) &= \int_0^1 u f_{U_{(k)}}(u)\, du = \frac{n!}{(k-1)!(n-k)!} \int_0^1 u^{k+1-1}(1-u)^{n-k+1-1}\, du \\ &= \frac{n!}{(k-1)!(n-k)!} \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} = \frac{k}{n+1}. \end{aligned}$$

An additional result needed is the following. If $X$ is a random variable which is continuous on an interval $I$ with c.d.f. $F$ strictly increasing on $I$, then $V = F(X) \sim U(0,1)$. For this, note that $0 \le V \le 1$ and for $0 \le v \le 1$, $P(V \le v) = P(F(X) \le v) = P(X \le F^{-1}(v)) = F(F^{-1}(v)) = v$.

Now argue as follows. If $T_1, T_2, \ldots, T_n$ are i.i.d. from $N(\mu, \sigma^2)$, then

$$E\left(\Phi\left(\frac{T_{(i)} - \mu}{\sigma}\right)\right) \approx \frac{i - 0.5}{n}, i = 1, 2, \ldots, n.$$

Therefore, plot of $\Phi\left(\frac{T_{(i)}-\mu}{\sigma}\right)$ versus $\frac{i-0.5}{n}$ is on the line $y = x$. Equivalently, the plot of $\frac{T_{(i)}-\mu}{\sigma}$ versus $\Phi^{-1}(\frac{i-0.5}{n})$ is on the line $y = x$. In other words, the plot of $T_{(i)}$ versus $\Phi^{-1}(\frac{i-0.5}{n})$ is linear. To check this, $\mu$ and $\sigma^2$ are not needed. Since $T_{(i)}$ is the quantile of order $i/n$ and $\Phi^{-1}(\frac{i-0.5}{n})$ is the standard normal

quantile of order $\frac{i-0.5}{n}$, this plot is called the Quantile - Quantile plot. One looks for nonlinearity in the plot to check for non-normality.

How is this plot to be used in regression? We want to check the normality of $\epsilon_i$, but they are not observable. Instead $y_i$ are observable, but they have different means. We consider the residuals. $\hat{\epsilon} = Y - \hat{Y} = (I-P) \sim N_n(0, \sigma^2(I-P))$ if normality holds. i.e., $\hat{\epsilon}_i \sim N(0, \sigma^2(1-P_{ii}))$ if $Y \sim N(X\beta, \sigma^2 I_n)$. For a fixed number of regressors $(p-1)$, as $n$ increases, $P_{ii} \to 0$ (Weisberg), so the residuals can be used in the Q-Q plot.

## Stepwise regression (forward selection)

Consider a situation where there are a large number of predictors. A model including all of them is not desirable since it will be unweildy and there may be difficulties involving multicollinearity and computational complexities. There are many such situations in weather forecasting, economics, finance, agriculture and medicine.

Consider the approach where one variable is added at a time until a good model is available, or equivalently, a stopping rule is met. Possible rules are
(i) $r$ many predictors are chosen ($r$ is pre-dertmined)
(ii) $R^2$ is large enough.

**Procedure.** (i) Calculate the correlation coefficient between $Y$ and $X_i$ for all $i$, say $r_{iy}$. Select as the first variable to enter the regression model the one most highly correlated with $Y$.
(ii) Regress $Y$ on the chosen predictor, say $X_l$, and compute $R^2 = r_{ly}^2$. This is the maximum possible $R^2$ with one predictor.
(iii) Calculate the partial correlation coefficients given $X_l$ of all the predictors not yet in the regression model, with the response $Y$. Choose as the next predictor to enter the model, the one with the highest (in magnitude) partial correlation coefficient $r_{iy.l}$: the idea is to add a factor which is most useful given that $X_l$ is already in.
(iv) Regress $Y$ on $X_l$ as well as the one chosen next, say $X_m$, and find if $X_m$ should be added or not. Compute $R^2$.
(v) Calculate $r_{iy.lm}$ and proceed similarly.

**Example.** Data on breeding success of the common Puffin in different habitats at Great Island, Newfoundland:
$y$ = nesting frequency (burrows/$9m^2$)
$x_1$ = grass cover (%), $x_2$ = mean soil depth ($cm$)
$x_3$ = angle of slope (degrees), $x_4$ = distance from cliff edge ($m$)

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $Y$ |
|---|---|---|---|---|
| 45 | 39.2 | 38 | 3 | 16 |
| 65 | 47.0 | 36 | 12 | 15 |
| 40 | 24.3 | 14 | 18 | 10 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

Correlation matrix:

| | $Y$ | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|---|
| $X_1$ | 0.158 | | | |
| $X_2$ | 0.022 | 0.069 | | |
| $X_3$ | 0.836 | -0.017 | 0.066 | |
| $X_4$ | -0.908* | -0.205 | 0.212 | -0.815 |

Choose $X_4$ first, since $r_{4y}$ = -0.908 is the highest in magnitude. Then $R^2 = (-0.908)^2 = 82.4\%$. $F = 168.79 >> F_{1,36}(.99)$. Now compute

$$r_{iy.4} = \begin{cases} -0.07 & i = 1; \\ 0.518 & i = 2; \\ 0.398 & i = 3. \end{cases}$$

Choose $X_2$ next and note $R^2 = 87.2\%$. Also, $X_2$ is a useful predictor. Compute

$$r_{iy.42} = \begin{cases} -0.152 & i = 1; \\ 0.233 & i = 3. \end{cases}$$

The formula for this is

$$r_{iy.42} = \frac{r_{iy.4} - r_{i2.4} r_{y2.4}}{\sqrt{(1 - r_{i2.4}^2)(1 - r_{y2.4}^2)}}.$$

If we pick $X_3$ now, $R^2 = 87.9\%$, not very different from the previous regression. Also, $X_3$ is not particularly useful in regression.

## Basics of Design of Experiments and ANOVA

So far we concentrated on analysis of a given experiment or data. Structure of the experiment is now explored. Design of experiments is a study of construction and analysis of experiments where purposeful changes are made to input variables of a process or system so as to observe and identify the reasons for changes in the output response. A cause-effect mechanism is of interest here.

**Example.** Different or different amounts of fertilizers versus yield of a crop

Experimental designs are used mostly for comparative experiments:
Comparing treatments in a clinical trial
Comparing factors (fertilizers, crop patterns etc.) in agricultural experiments

Randomization, replication, blocking and confounding of effects are some important concepts in this context. Randomization means that each subject has the same chance of being placed in any given experimental group. Then factors which cannot be controlled need not be considered since their effects are averaged out.
Replication means having multiple subjects in all experimental groups, ensuring that 'within group' variation can be estimated.
Blocking and confounding of effects will be considered later.

Consider the following example of a completely randomized design.

**Example.** Monosodium glutamate (MSG), a common ingredient of preserved food is known to cause brain damage in various mammals. In a study of the other effects, weight of ovaries ($mg$), both for a sample of rats treated with MSG and for an independent control sample of similar but untreated rats were obtained:

| | sample size ($n_i$) | sample mean | sample s.d. |
|---|---|---|---|
| MSG | 10 | 29.35 | 4.55 |
| Control | 12 | 21.86 | 10.09 |

Consider the linear model,

$$y_i = \begin{cases} \mu_m + \epsilon_i & i = 1, 2, \ldots, n_1; \\ \mu_c + \epsilon_i & i = n_1 + 1, n_1 + 2, \ldots, n_1 + n_2, \end{cases}$$

$\epsilon_i$ i.i.d $N(0, \sigma^2)$. Write it in the vector/matrix form, $Y = X\beta + \epsilon$:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_m \\ \mu_c \end{pmatrix} + \epsilon,$$

$\epsilon \sim N_{n_1+n_2}(0, \sigma^2 I)$. Then

$$X'X = \begin{Bmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}, \text{ so } (X'X)^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix}.$$

Therefore,

$$\begin{pmatrix} \hat{\mu}_m \\ \hat{\mu}_c \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} \sim N_2 \left( \begin{pmatrix} \mu_m \\ \mu_c \end{pmatrix}, \sigma^2 \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix} \right),$$

independent of

$$\text{RSS} = \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2 + \sum_{i=n_1+1}^{n_1+n_2} (y_i - \bar{y}_2)^2 \sim \sigma^2 \chi^2_{n_1+n_2-2}.$$

We want to compare $\mu_m$ with $\mu_c$. $H_0 : \mu_m = \mu_c = 0$ is meaningless; $H_0 : \mu_m = \mu_c$ is of interest. i.e., $H_0 : (1 - 1) \begin{pmatrix} \mu_m \\ \mu_c \end{pmatrix} = 0$. Note,

$$\frac{(\bar{y}_1 - \bar{y}_2 - (\mu_m - \mu_c)) / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{\text{RSS}/(n_1 + n_2 - 2)}} \sim t_{n_1+n_2-2}.$$

If $H_0$ is true, then

$$\frac{(\bar{y}_1 - \bar{y}_2) / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{\text{RSS}/(n_1 + n_2 - 2)}} \sim t_{n_1+n_2-2},$$

or equivalently,

$$\frac{(\bar{y}_1 - \bar{y}_2)^2 / \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}{\text{RSS}/(n_1 + n_2 - 2)} \sim F_{1,n_1+n_2-2}.$$

The design in this experiment has complete randomization. The observations inside the groups are independent, and also the two samples are independent.

For this reason, the design is called a completely randomized design. We can generalize this procedure if we want to compare $k$ means, as will be done later.

**Paired differences - example of a block design**

Sometimes independent samples, such as the ones in a completely randomized design, from two (or $k > 2$) populations is not an efficient way for comparisons. Consider the following example.

**Example.** It is of interest to compare an enriched formula with a standard formula for baby food. Weights of infants vary significantly and this influences weight gain more than the difference in food quality. Therefore, independent samples (with infants having very different weights) for the two formulas will not be very efficient in detecting the difference. Instead, pair babies of similar weight and feed one of them the standard formula, and the other the enriched formula. Then observe the gain in weight:

| pair | 1 | 2 | 3 | ... | $n$ |
|---|---|---|---|---|---|
| enriched | $e_1$ | $e_2$ | $e_3$ | ... | $e_n$ |
| standard | $s_1$ | $s_2$ | $s_3$ | ... | $s_n$ |

However, the samples may not be treated as independent but correlated. The $n$ pairs of observations, $(e_1, s_1), \ldots, (e_n, s_n)$ may still be treated to be uncorrelated (or even independent). These $n$ pairs are like $n$ independent blocks, inside each of which we can compare enriched with standard. This is the idea of blocking and block designs. Blocks are supposed to be homogeneous inside, so comparison of treatments within blocks becomes efficient.

**Paired differences - example of a block design**

Sometimes independent samples, such as the ones in a completely randomized design, from two (or $k > 2$) populations is not an efficient way for comparisons. Consider the following example.

**Example.** It is of interest to compare an enriched formula with a standard formula for baby food. Weights of infants vary significantly and this influences weight gain more than the difference in food quality. Therefore, independent samples (with infants having very different weights) for the two formulas will not be very efficient in detecting the difference. Instead, pair babies of similar weight and feed one of them the standard formula, and the other the enriched formula. Then observe the gain in weight:

| pair | 1 | 2 | 3 | ... | $n$ |
|---|---|---|---|---|---|
| enriched | $e_1$ | $e_2$ | $e_3$ | ... | $e_n$ |
| standard | $s_1$ | $s_2$ | $s_3$ | ... | $s_n$ |

However, the samples may not be treated as independent but correlated. The $n$ pairs of observations, $(e_1, s_1), \ldots, (e_n, s_n)$ may still be treated to be uncorrelated (or even independent). These $n$ pairs are like $n$ independent blocks, inside each of which we can compare enriched with standard. This is the idea of blocking and block designs. Blocks are supposed to be homogeneous inside, so comparison of treatments within blocks becomes efficient. We assume that $\begin{pmatrix} e_i \\ s_i \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$. In the above example, we want to test $H_0 : \mu_D \equiv \mu_1 - \mu_2 = 0$, so consider $y_i = e_i - s_i$. Then, $y_i = \mu_D + \epsilon_i$, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 = Var(y_i)$. If normality is assumed, then we have, $y_1, \ldots, y_n$ i.i.d. $N(\mu_D, \sigma_D^2)$ and we want to test $H_0 : \mu_D = 0$. Consider the test statistic,

$$\frac{\sqrt{n}\bar{y}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2/(n-1)}} \sim t_{n-1},$$

if $H_0$ is true, or equivalently,

$$\frac{n\bar{y}^2}{\sum_{i=1}^n (y_i - \bar{y})^2/(n-1)} \sim F_{1,n-1}.$$

Note that,

$$\begin{aligned}
&\frac{n\bar{y}^2}{\sum_{i=1}^n (y_i-\bar{y})^2/(n-1)} \\
&= \frac{n(\bar{e}-\bar{s})^2}{\frac{1}{n-1}\sum_{i=1}^n \left[(e_i-\bar{e})-(s_i-\bar{s})\right]^2} \\
&= \frac{n(\bar{e}-\bar{s})^2}{\frac{1}{n-1}\left[\sum_{i=1}^n (e_i-\bar{e})^2+\sum_{i=1}^n (s_i-\bar{s})^2-2\sum_{i=1}^n (e_i-\bar{e})(s_i-\bar{s})\right]} \\
&= \frac{(\bar{e}-\bar{s})^2/(\frac{1}{n}+\frac{1}{n})}{\frac{1}{2(n-1)}\left[\sum_{i=1}^n (e_i-\bar{e})^2+\sum_{i=1}^n (s_i-\bar{s})^2-2\sum_{i=1}^n (e_i-\bar{e})(s_i-\bar{s})\right]}.
\end{aligned}$$

Compare this test statistic with the one used for independent samples. $Cov(e,s)$ is expected to be positive (due to blocking), so the variance in the denominator above is typically less than $\frac{1}{2(n-1)}\left[\sum_{i=1}^n (e_i-\bar{e})^2+\sum_{i=1}^n (s_i-\bar{s})^2\right]$, which appears there. This is the positive effect due to blocking.

**Confounding of effects.**

**Example.** Consider two groups of similar students and two teachers. It is of interest to compare two different training methods. Consider the design where teacher A teaches one group using method I, whereas teacher B teaches the other group using method II. Later the results are analyzed. The problem with this design is that, if one group performs better it may be due to teacher effect or due to method effect, but it is not possible to separate the effects. We say then that the two effects are confounded. Sometimes we may not be interested in certain effects, in which case we may actually look for designs that will confound their effects. This will reduce the number of parameters to be estimated.

## Experiments with a single factor – One-way ANOVA

We want to compare $k > 2$ treatments. Treatment $i$ produces a population of $y$ values with mean $\mu_i$, $i = 1, 2, \ldots, k$. Or, if treatment $i$ is applied, then the response $Y \sim N(\mu_i, \sigma^2)$, $i = 1, 2, \ldots, k$. Are these $k$ populations different?

**Design.** $n_i$ observations are made independently from population $i$, so the $k$ samples are independent. Equivalently, we may look at this experiment as a design where $N$ subjects are available to study the $k$ treatments. $n_1$ of these are randomly selected and assigned to a group which will get treatment 1, $n_2$ of the remaining for treatment 2, and so on. Such a design is called a *completely randomized design* (as mentioned previously). Model for such a

design is as follows.
Let $y_{ij}$ = response of the $j$th individual in the $i$th group ($i$th treatment), $j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, k$. Then,
$y_{ij} = \mu_i + \epsilon_{ij},\ j = 1, 2, \ldots, n_i;\ i = 1, 2, \ldots, k.\ E(\epsilon_{ij}) = 0,\ Var(\epsilon_{ij}) = \sigma^2$, uncorrelated errors; $\epsilon_{ij} \sim N(0, \sigma^2)$ i.i.d. for testing and confidence statements.
In the usual linear model formulation:

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} + \epsilon.$$

Since $(X'X)^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \cdots & 0 \\ 0 & 0 & \cdots & \frac{1}{n_k} \end{pmatrix}$ and $X'Y = \begin{pmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \vdots \\ \sum_{j=1}^{n_1} y_{kj} \end{pmatrix}$, we get

$\begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_k \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_k \end{pmatrix}$ and

RSS $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum \sum \hat{\epsilon}_{ij}^2 = \sum \sum (y_{ij} - \hat{\mu}_i)^2$.

**Questions.**

(i) Are the group means $\mu_i$ equal? i.e., test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.
(ii) If not, how are they different?

$y_{ij} = \mu_i + \epsilon_{ij}, j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, k$ $E(\epsilon_{ij}) = 0$, $Var(\epsilon_{ij}) = \sigma^2$.

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} + \epsilon.$$

$$\begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_k \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_k \end{pmatrix}$$

RSS $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum \sum \hat{\epsilon}_{ij}^2 = \sum \sum (y_{ij} - \hat{\mu}_i)^2$.

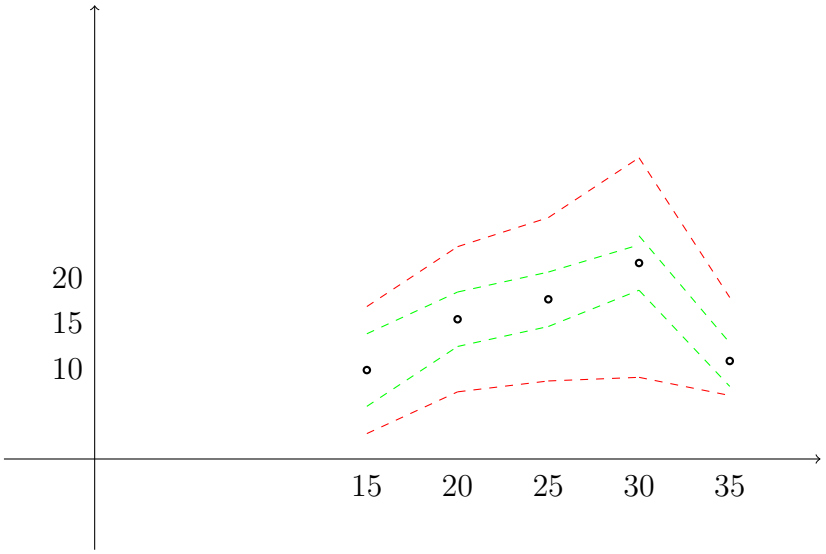
**Questions.**

(i) Are the group means $\mu_i$ equal? i.e., test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.
(ii) If not, how are they different?

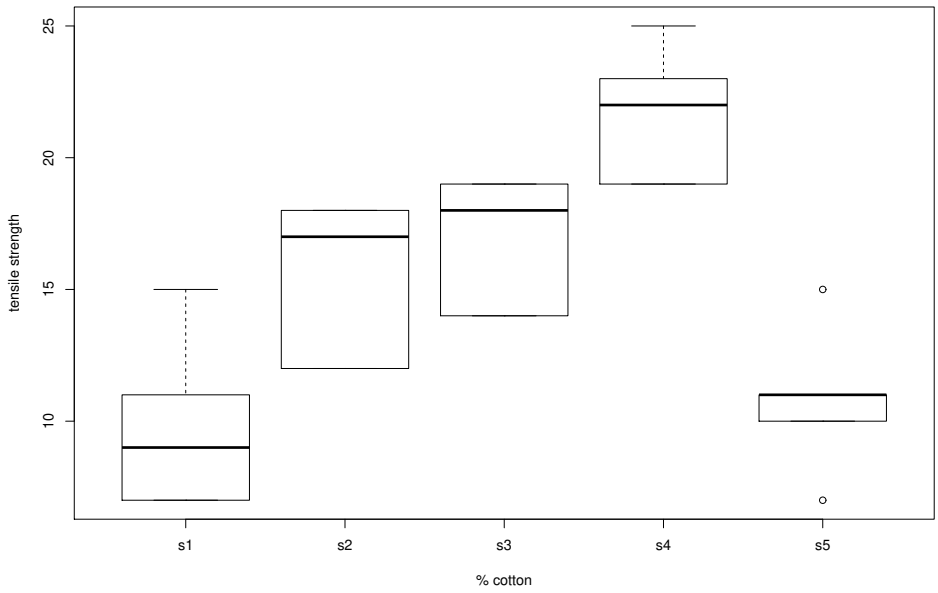
**Example.** It is believed that the tensile (breaking) strength of synthetic fibre is affected by the %age of cotton in fibre:

| % cotton | tensile strength ($lb/inch^2$) | sample mean |
|---|---|---|
| 15 | 7, 7, 15, 11, 9 | $\bar{y}_1 = 9.8$ |
| 20 | 12, 17, 12, 18, 18 | $\bar{y}_2 = 15.4$ |
| 25 | 14, 18, 18, 19, 19 | $\bar{y}_3 = 17.6$ |
| 30 | 19, 25, 22, 19, 23 | $\bar{y}_4 = 21.6$ |
| 35 | 7, 10, 11, 15, 11 | $\bar{y}_5 = 10.8$ |

Are there substantial differences in the mean breaking strength?
(i) Plot the sample means:




But sample means do not tell the whole story, especially for small samples. One must look at variation within samples and between samples. In the plot above, the conclusions would be different according to whether the error bands are green or red.




It is easier to do this investigation of variations using box-plots, as shown above. Variation within samples is not too large or different, but between

sample variation is large. Note that, if within sample variation is large compared to between sample variation (like the red error bands in the plot), then the different samples can be considered to be from a single population. However, if within sample variation is small compared to between sample variation (like the green error bands in the plot, i.e., $|\bar{y}_i - \bar{y}_j|$ are large compared to the error) then there is reason to believe that the groups differ.

To formalize this, we return to linear models:
$y_{ij} = \mu_i + \epsilon_{ij}$, $j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, k$, $\epsilon_{ij} \sim N(0, \sigma^2)$ i.i.d. Are the group means different?

$$\begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_k \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_k \end{pmatrix} \text{ so that RSS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

To test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, consider
$A_{(k-1)\times k} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$. Then we test $H_0 : A\mu = 0$ where $A$ has rank $k - 1$. To test $H_0$, we obtain $\hat{\mu}_{H_0}$, $\text{RSS}_{H_0}$ and consider

$$F = \frac{(\text{RSS}_{H_0} - \text{RSS})/(k-1)}{\text{RSS}/(\sum_{i=1}^{k} n_i - k)}, \text{ which} \sim F_{k-1, \sum_{i=1}^{k} n_i - k} \text{ under } H_0.$$

To find $\hat{\mu}_{H_0}$, $\text{RSS}_{H_0}$, note that, under $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, these means are equal, and so it is enough to find

$$\min_{\mu_1 = \mu_2 = \cdots = \mu_k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 = \min_{\mu} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu)^2.$$

Therefore,

$$\hat{\mu}_{H_0} = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \equiv \bar{y}_{..}, \text{ and hence } \text{RSS}_{H_0} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2.$$

$y_{ij} = \mu_i + \epsilon_{ij}$, $j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, k$, $\epsilon_{ij} \sim N(0, \sigma^2)$ i.i.d. Are the group means different?

$$\begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_k \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_k \end{pmatrix} \text{ so that RSS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

To test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, consider

$A_{(k-1)\times k} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -1 \end{pmatrix}$. Then we test $H_0 : A\mu = 0$ where $A$ has rank $k - 1$. To test $H_0$, we obtain $\hat{\mu}_{H_0}$, $\text{RSS}_{H_0}$ and consider

$$F = \frac{(\text{RSS}_{H_0} - \text{RSS})/(k-1)}{\text{RSS}/(\sum_{i=1}^{k} n_i - k)}, \text{ which} \sim F_{k-1, \sum_{i=1}^{k} n_i - k} \text{ under } H_0.$$

To find $\hat{\mu}_{H_0}$, $\text{RSS}_{H_0}$, note that, under $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, these means are equal, and so it is enough to find

$$\min_{\mu_1 = \mu_2 = \cdots = \mu_k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 = \min_{\mu} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu)^2.$$

Therefore,

$$\hat{\mu}_{H_0} = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \equiv \bar{y}_{..}, \text{ and hence } \text{RSS}_{H_0} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2.$$

Introduce further notation: $\bar{y}_{i.} = \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$, $i = 1, 2, \ldots, k$. Note, further, that

$$\begin{aligned} &\text{RSS}_{H_0} \\ = \; & \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})^2 \\ = \; & \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2 + 2 \sum_{i=1}^{k} \left\{ (\bar{y}_{i.} - \bar{y}_{..}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.}) \right\} \\ = \; & \text{RSS} + \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2, \end{aligned}$$

since $\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.}) = 0$ for all $i$. Therefore,

$$\text{RSS}_{H_0} - \text{RSS} = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

and therefore,

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2/(\sum_{i=1}^{k} n_i - k)} \sim F_{k-1,\sum_{i=1}^{k} n_i - k} \text{ under } H_0.$$

It is instructive to consider these sum of squares.
$\text{RSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$
= the sum total of all the sum of squares of deviations from the sample means
= within groups or within treatments sum of squares, $\text{SS}_W$.

$\text{RSS}_{H_0} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$
= total sum of squares of deviations assuming no treatment effect
= total variability (corrected) in the $k$ samples, $\text{SS}_T$.

Therefore, $\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$ = $\text{SS}_T$ - $\text{SS}_W$ = between groups or between treatments sum of squares = $\text{SS}_B$. Thus,
$\text{SS}_T = \text{SS}_W + \text{SS}_B$ is the decomposition of sum of squares along with
$\sum_{i=1}^{k} n_i - 1 = (\sum_{i=1}^{k} n_i - k) + (k-1)$, decomposition of d.f.

**ANOVA for One-way classification**

| source | d.f. | SS | MS | $F$ |
|---|---|---|---|---|
| Treatments (groups) | $k-1$ | $\text{SS}_B = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$ | $\text{MS}_B = \frac{\text{SS}_B}{k-1}$ | $\frac{\text{MS}_B}{\text{MSE}} \sim$ (under $H_0$) $F_{k-1,\sum_{i=1}^{k} n_i - k}$ |
| Error | $\sum n_i - k$ | $\text{SS}_W = \sum\sum(y_{ij} - \bar{y}_{i.})^2$ | $\text{MSE} = \frac{\text{SS}_W}{\sum_{i=1}^{k} n_i - k}$ | |
| Total (corrected) | $\sum n_i - 1$ | $\text{SS}_T = \sum\sum(y_{ij} - \bar{y}_{..})^2$ | | |
| Mean | 1 | $(\sum_{i=1}^{k} n_i)\bar{y}_{..}^2$ | | |
| Total | $\sum n_i$ | $\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2$ | | |

**Example.** Tensile strength data. $k = 5$, $n_i = 5$. ANOVA is as follows.

| source | d.f. | SS | MS | $F$ |
|---|---|---|---|---|
| Factor levels (% cotton) | 4 | 475.76 | 118.94 | $14.76 >> 4.43 = F_{4,20}(.99)$ |
| Error | 20 | 161.20 | 8.06 | |
| Total(corrected) | 24 | 636.96 | | |

$R^2 = \frac{475.76}{636.96} \approx 75\%$

Now that the ANOVA $H_0$ has been rejected, we should look at the group means (estimates) closely. Suppose we want to compare $\mu_r$ and $\mu_s$ either with $H_0 : \mu_r = \mu_s$ or using a confidence interval for $\mu_r - \mu_s$.

$$\hat{\mu}_r - \hat{\mu}_s = \bar{y}_{r.} - \bar{y}_{s.} \sim N\left(\mu_r - \mu_s, \sigma^2\left(\frac{1}{n_r} + \frac{1}{n_s}\right)\right)$$

independently of

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 \sim \sigma^2\chi^2_{\sum_{i=1}^k n_i - k}.$$

Therefore,

$$\frac{\{(\bar{y}_{r.} - \bar{y}_{s.}) - (\mu_r - \mu_s)\}/\sqrt{\frac{1}{n_r} + \frac{1}{n_s}}}{\sqrt{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2/\left(\sum_{i=1}^{k} n_i - k\right)}} \sim t_{\sum_{i=1}^k n_i - k}.$$

$100(1-\alpha)\%$ confidence interval for $\mu_r - \mu_s$ is

$$\bar{y}_{r.} - \bar{y}_{s.} \pm t_{\sum_{i=1}^k n_i - k}(1-\alpha/2)\sqrt{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2/\left(\sum_{i=1}^{k} n_i - k\right)}\sqrt{\frac{1}{n_r} + \frac{1}{n_s}}.$$

Further, test statistic for testing $H_0 : \mu_r = \mu_s$ is

$$T = \frac{(\bar{y}_{r.} - \bar{y}_{s.})/\sqrt{\frac{1}{n_r} + \frac{1}{n_s}}}{\sqrt{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2/\left(\sum_{i=1}^{k} n_i - k\right)}} \sim t_{\sum_{i=1}^k n_i - k},$$

if $H_0$ is true.

**Multiple comparison of group means**

$y_{ij} = \mu_i + \epsilon_{ij}$, $j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, k$, $\epsilon_{ij} \sim N(0, \sigma^2)$ i.i.d.

The classic ANOVA test is the test of $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, which is uninteresting and the hypothesis is usually not true. What an experimentor usually wants to find out is which treatments are better, so rejection of $H_0$ is usually not the end of the analysis. Once it is rejected, further work is needed to find out why it was rejected.

**Definition.** A linear parametric function $\sum_{i=1}^k a_i\mu_i = a'\mu$ with known constants $a_1, \ldots, a_k$ satisfying $\sum_{i=1}^k a_i = a'\mathbf{1} = 0$ is called a contrast (linear contrast).

**Example.** If $a = (1, -1, 0, \ldots, 0)'$, then $a'\mu = \mu_1 - \mu_2$.

**Result.** $\mu_1 = \mu_2 = \cdots = \mu_k$ if and only if $a'\mu = 0$ for all $a \in \mathcal{A} = \left\{a = (a_1, \ldots, a_k)' : \sum_{i=1}^k a_i = 0\right\}$.

**Remark.** $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is true iff $H_a : a'\mu = 0$ for all $a \in \mathcal{A}$, or all linear contrasts are zero.

**Proof.** $\mu_1 = \mu_2 = \cdots = \mu_k$ iff $\mu = \alpha\mathbf{1}$ for some $\alpha$, or $\mu \in \mathcal{M}_C(\mathbf{1})$. Note, $\mathcal{A} = \mathcal{M}_C^{\perp}(\mathbf{1})$.

Thus, if $H_0$ fails, atleast one of the $H_a$ must fail for $a \in \mathcal{A}$. i.e., $a'\mu \neq 0$. The experimenter may be interested in this contrast, and its inference. Consider inference of any linear parametric function, $a'\mu = \sum_{i=1}^k a_i\mu_i$. We have the model,
$y_{ij} \sim N(\mu_i, \sigma^2), j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, k$ independent. Then, $\bar{y}_{i\cdot} \sim N(\mu_i, \sigma^2/n_i)$, $i = 1, 2, \ldots, k$ independent, and

$$E(\sum_{i=1}^k a_i\bar{y}_{i\cdot}) = \sum_{i=1}^k a_i\mu_i = a'\mu, \quad Var(\sum_{i=1}^k a_i\bar{y}_{i\cdot}) = \sigma^2 \sum_{i=1}^k \frac{a_i^2}{n_i},$$

so that

$$\frac{\sum_{i=1}^k a_i\bar{y}_{i\cdot} - \sum_{i=1}^k a_i\mu_i}{\sqrt{\sigma^2 \sum_{i=1}^k \frac{a_i^2}{n_i}}} \sim N(0, 1).$$

Let $S_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$, $i = 1, 2, \ldots, k$. Then $S_i^2 \sim \sigma^2\chi^2_{n_i-1}$ independent of $\bar{y}_{i\cdot}$, $i = 1, 2, \ldots, k$. Also, $(S_1^2, \ldots, S_k^2)$ is independent of $\bar{\mathbf{y}} = (\bar{y}_{1\cdot}, \ldots, \bar{y}_{k\cdot})$. Let $S_p^2 = \sum_{i=1}^k S_i^2$. Then $S_p^2 \sim \sigma^2\chi^2_{\sum_{i=1}^k n_i - k}$ independent of $\bar{\mathbf{y}}$. Note that this is just a repeat of our old result that RSS $= \sum_{i=1}^k \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 = S_p^2$ is

independent of $\hat{\beta} = \hat{\mu}$. Thus, as discussed previously,

$$\frac{a'\bar{\mathbf{y}} - a'\mu}{\sqrt{S_p^2 \left(\sum_{i=1}^k \frac{a_i^2}{n_i}\right) / (\sum_{i=1}^k n_i - k)}} \sim t_{\sum_{i=1}^k n_i - k},$$

so that

$$a'\bar{\mathbf{y}} \pm t_{\sum_{i=1}^k n_i - k}(1-\alpha/2)\sqrt{S_p^2 \left(\sum_{i=1}^k \frac{a_i^2}{n_i}\right) / (\sum_{i=1}^k n_i - k)}$$

is a $100(1-\alpha)\%$ confidence interval for $a'\mu$. Also, reject $H_{a,0} : a'\mu = 0$ in favour of $H_{a,1} : a'\mu \neq 0$ if

$$\left| \frac{a'\bar{\mathbf{y}}}{\sqrt{S_p^2 \left(\sum_{i=1}^k \frac{a_i^2}{n_i}\right) / (\sum_{i=1}^k n_i - k)}} \right| > t_{\sum_{i=1}^k n_i - k}(1-\alpha/2).$$

What if we want investigate a set of contrasts simultaneously? From Boole's Inequality,
$P\left(\cup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$, so $P\left(\cup_{i=1}^{\infty} A_i^c\right) \leq \sum_{i=1}^{\infty} P(A_i^c)$.
Since $\cup_{i=1}^{\infty} A_i^c = \left(\cap_{i=1}^{\infty} A_i\right)^c$,

$$1 - P(\cap_{i=1}^n A_i) \leq \sum_{i=1}^n (1 - P(A_i)) = n - \sum_{i=1}^n P(A_i), \text{ or}$$

$$P\left(\cap_{i=1}^n A_i\right) \geq \sum_{i=1}^n P(A_i) - (n-1).$$

This is known as the Bonferroni Inequality. Apply this to the above problem.
If we want a simultaneous confidence set for $a^{(1)'}\mu, \ldots, a^{(d)'}\mu$, consider

$$C = \left\{ a^{(j)'}\bar{\mathbf{y}} \pm t_{\sum_{i=1}^k n_i - k}(1 - \frac{\alpha}{2d})\sqrt{S_p^2 \left(\sum_{i=1}^k \frac{(a_i^{(j)})^2}{n_i}\right) / (\sum_{i=1}^k n_i - k)}, j = 1, 2, \ldots, d \right\}.$$

Then

$$P(C) = P(\cap_{l=1}^d A_l) \geq \sum_{l=1}^d P(A_l) - (d-1) = \sum_{l=1}^d (1 - \frac{\alpha}{d}) - (d-1) = d - \alpha - d + 1 = 1 - \alpha.$$

This procedure is useful when $d$ is not too large.

**Reparametrization of the one-way model.**

Suppose $n_i$ are all equal, and equal to $J$. Also, let the number of groups be $k = I$. Then $\sum_{i=1}^{k} n_i = IJ$, and $\bar{y}_{i.} = \sum_{j=1}^{J} y_{ij}/J$, for $i = 1, \ldots, I$.
$\bar{y_{..}} = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}/(IJ)$. Further,
$\mathrm{SS}_W = \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \bar{y}_{i.})^2$ has d.f. $(IJ - I)$;
$\mathrm{SS}_B = \sum_{i=1}^{I} n_i (y_{i.} - \bar{y}_{..})^2 = J \sum_{i=1}^{I} (y_{i.} - \bar{y}_{..})^2$ has d.f. $I - 1$.

We can rewrite the model, $y_{ij} = \mu_i + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, \sigma^2)$ i.i.d. as follows.
$\mu_i = \bar{\mu}_{.} + (\mu_i - \bar{\mu}_{.}) = \mu + \alpha_i$, where $\bar{\mu}_{.} = \sum_{i=1}^{I} \mu_i / I$ and $\alpha_i = \mu_i - \bar{\mu}_{.}$. Then, $\sum_{i=1}^{I} \alpha_i = \alpha_{.} = \sum_{i=1}^{I} (\mu_i - \bar{\mu}_{.}) = 0$. Further, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ is the same as $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_{I-1} = 0$ ($\alpha_{.} = 0$ implies that $\alpha_I = -\sum_{i=1}^{I-1} \alpha_i = 0$ also.)
Similarly write
$\bar{\epsilon}_{i.} = \bar{\epsilon}_{..} + \bar{\epsilon}_{i.} - \bar{\epsilon}_{..}$, so that
$\epsilon_{ij} = \bar{\epsilon}_{..} + (\bar{\epsilon}_{i.} - \bar{\epsilon}_{..}) + (\epsilon_{ij} - \bar{\epsilon}_{i.})$. Therefore

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \epsilon_{ij}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \epsilon_{..}^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} (\epsilon_{ij} - \bar{\epsilon}_{i.})^2,$$

since $\bar{\epsilon}_{..} \sum_{i=1}^{I} (\bar{\epsilon}_{i.} - \bar{\epsilon}_{..}) = 0$, $\bar{\epsilon}_{..} \sum_{i=1}^{I} \sum_{j=1}^{J} (\epsilon_{ij} - \bar{\epsilon}_{i.}) = 0$ and
$\sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})(\epsilon_{ij} - \bar{\epsilon}_{i.}) = \sum_{i=1}^{I} (\bar{\epsilon}_{i.} - \bar{\epsilon}_{..}) \sum_{j=1}^{J} (\epsilon_{ij} - \bar{\epsilon}_{i.}) = 0$.

Now, since $\epsilon_{ij} = y_{ij} - \mu - \alpha_i$, we get $\bar{\epsilon}_{i.} = \bar{y}_{i.} - \mu - \alpha_i$, $\bar{\epsilon}_{..} = \bar{y}_{..} - \mu$, and further, from above,

$$\begin{aligned} &\sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \mu - \alpha_i)^2 \\ &= \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{y}_{..} - \mu)^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{y}_{i.} - \bar{y}_{..} - \alpha_i)^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \bar{y}_{i.})^2. \end{aligned}$$

Least squares estimates subject to $\sum_{i=1}^{I} \alpha_i = 0$ may be obtained simply by examination of the above, and they are:

$$\hat{\mu} = \bar{y}_{..}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..},$$

and hence RSS $= \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \bar{y}_{i.})^2$.

Under $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_{I-1} = 0$, we have

$$\begin{aligned}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\mu-\alpha_i)^2 &\equiv \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\mu)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{..}-\mu)^2+\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{i.}-\bar{y}_{..})^2+\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y}_{i.})^2,\end{aligned}$$

so that, then, $\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\mu)^2$ is minimized when $\hat{\mu}=\bar{y}_{..}$ (with $\alpha_i = 0$). We then get

$$\begin{aligned}\mathrm{RSS}_{H_0} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{i.}-\bar{y}_{..})^2+\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y}_{i.})^2 \\ &= J\sum_{i=1}^{I}(\bar{y}_{i.}-\bar{y}_{..})^2+\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y}_{i.})^2.\end{aligned}$$

Therefore,

$$\mathrm{RSS}_{H_0}-\mathrm{RSS} = J\sum_{i=1}^{I}(\bar{y}_{i.}-\bar{y}_{..})^2.$$

Note that all these can be done by just inspection, even though we have derived these previously using other methods. The simplicity of this approach, however, is very useful for higher-way classification models.

**One-way ANOVA with equal number of observations per group.**

| source | d.f | SS | MS | F |
|---|---|---|---|---|
| Treatments | $I-1$ | $J\sum(\bar{y}_{i.}-\bar{y}_{..})^2$ | $\mathrm{SS}_B/(I-1)$ | $\frac{J\sum(\bar{y}_{i.}-\bar{y}_{..})^2/(I-1)}{\sum\sum(y_{ij}-\bar{y}_{i.})^2/(IJ-I)}$ |
| Error | $IJ-I$ | $\sum\sum(y_{ij}-\bar{y}_{i.})^2$ | $\mathrm{SS}_W/(IJ-I)$ | |
| Total (C) | $IJ-1$ | $\sum\sum(y_{ij}-\bar{y}_{..})^2$ | | |

This approach of reparametrization and decomposition generalizes to higher-way classification where there are substantial simplifications.

## 2-factor Analysis or 2-way ANOVA

**Example.** An engineer is designing a battery for use in a device that will be subjected to some extreme temperature variations. The only design parameter that he can select at this time is the plate material for the battery, and he has three possible choices. When the device is manufactured and shipped

to the field, the engineer has no control over the temperature extremes that the device will encounter, and he knows from past experience that temperature may impact the effective battery life. However, temperature can be controlled in the product development laboratory for the purposes of testing.

The engineer decides to test all three plate materials at three different temperature levels, 15°F, 70°F and 125°F (-10, 21 and 51 degree C), as these temperature levels are consistent with the product end-use environment. Four batteries are tested at each combination of plate material and temperature, and the 36 tests are run in random order.

Question 1. What effects do material type and temperature have on the life of the battery?
Question 2. Is there a choice of material that would give uniformly long life regardless of temperature? (Robust product design?)

Question 1. What effects do material type and temperature have on the life of the battery?
Question 2. Is there a choice of material that would give uniformly long life regardless of temperature? (Robust product design?)

Life (in hrs) data for the battery design experiment:

| material type | temperature (°F) | | | | | |
|---|---|---|---|---|---|---|
| | 15 | | 70 | | 125 | |
| 1 | 130 | 155 | 34 | 40 | 20 | 70 |
| | 74 | 180 | 80 | 75 | 82 | 58 |
| 2 | 150 | 188 | 126 | 122 | 25 | 70 |
| | 159 | 126 | 106 | 115 | 58 | 45 |
| 3 | 138 | 110 | 174 | 120 | 96 | 104 |
| | 168 | 160 | 150 | 139 | 82 | 60 |

Both factors, material type and temperature are important and there may be interaction between the two also. Let us denote the row factor as factor A and column factor as factor B (in general). Then the model for the data may be developed as follows.

Let $y_{ijk}$ be the observed response when factor A is at the $i$th level ($i = 1, 2, \ldots, I$) and factor B is at the $j$th level ($j = 1, 2, \ldots, J$) for the $k$th replicate ($k = 1, 2, \ldots, K$). In the example, $I = 3, J = 3, K = 4$. This design is like having $IJ$ different cells each of which has $K$ observations, and one wants to see if the $IJ$ cell means are different or not (in various ways).

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, i = 1, 2, \ldots, I; j = 1, 2, \ldots, J; k = 1, 2, \ldots, K.$$

Therefore it is also called a completely randomized 2-factor design. We assume, $\epsilon_{ijk}$ are i.i.d. $N(0, \sigma^2)$. As before, this is a linear model, and hence various linear hypotheses can be tested. Let
$\bar{y}_{ij.} = \frac{1}{K}\sum_{k=1}^{K} y_{ijk}, i = 1, 2, \ldots, I; j = 1, 2, \ldots, J$
$\bar{y}_{i..} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk} = \frac{1}{J}\sum_{j=1}^{J} \bar{y}_{ij.}, i = 1, 2, \ldots, I$
$\bar{y}_{.j.} = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} y_{ijk} = \frac{1}{I}\sum_{i=1}^{I} \bar{y}_{ij.}, j = 1, 2, \ldots, J$
$\bar{y}_{...} = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} \bar{y}_{ij.} = \frac{1}{I}\sum_{i=1}^{I} \bar{y}_{i..} = \frac{1}{J}\sum_{j=1}^{J} \bar{y}_{.j.}$

Now, $\hat{\mu}_{ij} = \bar{y}_{ij.}$ under no constraints, and hence
RSS $= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{ij.})^2$ has $IJ(K-1)$ d.f. To consider interesting questions, it is best to adopt the reparametrization,
$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, where
$\mu = \bar{\mu}_{..} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij},\ \alpha_i = \bar{\mu}_{i.} - \bar{\mu}_{..},\ \beta_j = \bar{\mu}_{.j} - \bar{\mu}_{..}$ and
$(\alpha\beta)_{ij} = \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}$.

Then note that $\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$, $\sum_{i=1}^{I} (\alpha\beta)_{ij} = 0$ for all $j$ and $\sum_{j=1}^{J} (\alpha\beta)_{ij} = 0$ for all $i$.
(Note, $\sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{i=1}^{I} \mu_{ij} - \sum_{i=1}^{I} \bar{\mu}_{i.} - I\bar{\mu}_{.j} + I\bar{\mu}_{..} = \sum_{i=1}^{I} (\mu_{ij} - \bar{\mu}_{.j}) = 0$.)
These are the conditions required for identifiability of the parameters under reparametrization.

Now consider the interpretation of these parameters. $\mu = \bar{\mu}_{..}$ is the overall effect. $\alpha_i = \bar{\mu}_{i.} - \bar{\mu}_{..}$ = main effect of factor A at level $i$ since eliminating the effect of level $j$ by averaging over it leaves the departure of effect $i$ (of factor A) from overall, and similarly, $\beta_j = \bar{\mu}_{.j} - \bar{\mu}_{..}$ = main effect of factor B at level $j$. What does $(\alpha\beta)_{ij} = \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}$ measure?

Suppose we want to see if the effect of factor A at level $i$ depends on the level of factor B. If there were no such interaction, we would expect the difference in means $\mu_{i_1 j} - \mu_{i_2 j}$ depend on $i_1$ and $i_2$ and not on $j$. i.e.,

$$\begin{aligned} \mu_{i_1 j} - \mu_{i_2 j} &= \phi(i_1, i_2) = \frac{1}{J} \sum_{j=1}^{J} \phi(i_1, i_2) \\ &= \frac{1}{J} \sum_{j'=1}^{J} (\mu_{i_1 j'} - \mu_{i_2 j'}) = \bar{\mu}_{i_1 .} - \bar{\mu}_{i_2 .}, \end{aligned}$$

for all $i_1, i_2$. Or, equivalently, $\mu_{i_1 j} - \bar{\mu}_{i_1 .} = \mu_{i_2 j} - \bar{\mu}_{i_2 .}$ for all $i_1, i_2$. i.e.,

$$\begin{aligned} \mu_{ij} - \bar{\mu}_{i.} &= \Phi(j) \text{ (independent of } i) \\ &= \frac{1}{I} \sum_{i'=1}^{I} \Phi(j) = \frac{1}{I} \sum_{i'=1}^{I} (\mu_{i' j} - \bar{\mu}_{i' .}) \\ &= \bar{\mu}_{.j} - \bar{\mu}_{..}, \text{ for all } i, j. \end{aligned}$$

i.e., $\mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..} = 0$ for all $i, j$. Because of symmetry, we could have begun with $\mu_{ij_1} - \mu_{ij_2}$ depending on $j_1, j_2$, but not on $i$. Thus, we see that $(\alpha\beta)_{ij} = \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}$ measures the interaction of $i$ and $j$. Therefore, to investigate the existence of interaction, we should test,
$H_{AB} : (\alpha\beta)_{ij} = 0$ $(i = 1, 2, \ldots, I; j = 1, 2, \ldots, J)$ as the restricted model without interaction. Estimation of $(\alpha\beta)_{ij}$ can also be considered. Now, consider the main effects of factors A and B.
To test for lack of difference in levels of factor A, use, $H_A : \alpha_i = 0$ for all $i$.
To test for lack of difference in levels of factor B, use, $H_B : \beta_j = 0$ for all $j$. If $H_{AB} : (\alpha\beta)_{ij} = 0$ has been rejected, there is evidence for significant interaction, so main effects cannot be non-existent.

Life (in hrs) data for the battery design experiment:

| material | temperature (°F) | | | | | |
|---|---|---|---|---|---|---|
| type | 15 | | 70 | | 125 | |
| 1 | 130 | 155 | 34 | 40 | 20 | 70 |
| | 74 | 180 | 80 | 75 | 82 | 58 |
| 2 | 150 | 188 | 126 | 122 | 25 | 70 |
| | 159 | 126 | 106 | 115 | 58 | 45 |
| 3 | 138 | 110 | 174 | 120 | 96 | 104 |
| | 168 | 160 | 150 | 139 | 82 | 60 |

Let $y_{ijk}$ be the observed response when factor A is at the $i$th level ($i = 1, 2, \ldots, I$) and factor B is at the $j$th level ($j = 1, 2, \ldots, J$) for the $k$th replicate ($k = 1, 2, \ldots, K$).

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, i = 1, 2, \ldots, I; j = 1, 2, \ldots, J; k = 1, 2, \ldots, K.$$

Now, $\hat{\mu}_{ij} = \bar{y}_{ij.}$ under no constraints, and hence
RSS $= \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (y_{ijk} - \bar{y}_{ij.})^2$ has $IJ(K-1)$ d.f.
Reparametrization: $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, where
$\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$, $\sum_{i=1}^{I} (\alpha\beta)_{ij} = 0$ for all $j$ and $\sum_{j=1}^{J} (\alpha\beta)_{ij} = 0$ for all $i$ are the identifiability conditions.

To investigate the existence of interaction, we should test,
$H_{AB} : (\alpha\beta)_{ij} = 0 (i = 1, 2, \ldots, I; j = 1, 2, \ldots, J)$ as the restricted model without interaction. Estimation of $(\alpha\beta)_{ij}$ can also be considered. Now, consider the main effects of factors A and B.
To test for lack of difference in levels of factor A, use, $H_A : \alpha_i = 0$ for all $i$.
To test for lack of difference in levels of factor B, use, $H_B : \beta_j = 0$ for all $j$. If $H_{AB} : (\alpha\beta)_{ij} = 0$ has been rejected, there is evidence for significant interaction, so main effects cannot be non-existent.

To find estimates, confidence intervals and to conduct tests, we proceed as follows. Since
$\mu_{ij} = \bar{\mu}_{..} + (\bar{\mu}_{i.} - \bar{\mu}_{..}) + (\bar{\mu}_{.j} - \bar{\mu}_{..}) + (\mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$,
we use a similar representation for $\epsilon_{ijk}$:

$$\epsilon_{ijk} = \bar{\epsilon}_{...} + (\bar{\epsilon}_{i..} - \bar{\epsilon}_{...}) + (\bar{\epsilon}_{.j.} - \bar{\epsilon}_{...}) + (\bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..} - \bar{\epsilon}_{.j.} + \bar{\epsilon}_{...}) + (\epsilon_{ijk} - \bar{\epsilon}_{ij.}).$$

Therefore, as in one-way classification,

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\epsilon_{ijk}^2 = IJK\bar{\epsilon}_{...}^2 + JK\sum_{i=1}^{I}(\bar{\epsilon}_{i.}-\bar{\epsilon}_{..})^2 + IK\sum_{j=1}^{J}(\bar{\epsilon}_{.j}-\bar{\epsilon}_{..})^2$$
$$+K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{\epsilon}_{ij.}-\bar{\epsilon}_{i..}-\bar{\epsilon}_{.j.}+\bar{\epsilon}_{...})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\epsilon_{ijk}-\bar{\epsilon}_{ij.})^2,$$

since cross products vanish. Noting that $\epsilon_{ijk} = y_{ijk}-\mu-\alpha_i-\beta_j-(\alpha\beta)_{ij}$, with $\sum_{i=1}^{I}\alpha_i = 0$, $\sum_{j=1}^{J}\beta_j = 0$, $\sum_{i=1}^{I}(\alpha\beta)_{ij} = 0$ for all $j$ and $\sum_{j=1}^{J}(\alpha\beta)_{ij} = 0$ for all $i$, we get $\bar{\epsilon}_{...} = \bar{y}_{...}-\mu$, $\bar{\epsilon}_{i..} = \bar{y}_{i..}-\mu-\alpha_i$, $\bar{\epsilon}_{.j.} = \bar{y}_{.j.}-\mu-\beta_j$, $\bar{\epsilon}_{ij.} = \bar{y}_{ij.}-\mu-\alpha_i-\beta_j-(\alpha\beta)_{ij}$. Hence,

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\mu-\alpha_i-\beta_j-(\alpha\beta)_{ij})^2$$
$$= IJK(\bar{y}_{...}-\mu)^2 + JK\sum_{i=1}^{I}(\bar{y}_{i..}-\bar{y}_{...}-\alpha_i)^2 + IK\sum_{j=1}^{J}(\bar{y}_{.j.}-\bar{y}_{...}-\beta_j)^2$$
$$+K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}-(\alpha\beta)_{ij})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{ij.})^2.$$

Subject to the identifiability conditions, we obtain the least squares estimates:
$\hat{\mu} = \bar{y}_{...}$, $\hat{\alpha}_i = \bar{y}_{i..}-\bar{y}_{...}$, $\hat{\beta}_j = \bar{y}_{.j.}-\bar{y}_{...}$ and $\hat{(\alpha\beta)}_{ij} = \bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}$.
Therefore, RSS $= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{ij.})^2$, as seen earlier.

Consider $H_{AB} : (\alpha\beta)_{ij} = 0$ for all $i, j$. Due to the identiability constraints on these parameters, namely, $0 = \sum_{i=1}^{I}(\alpha\beta)_{ij} = \sum_{j=1}^{J}(\alpha\beta)_{ij} = \sum_{i=1}^{I}\sum_{j=1}^{J}(\alpha\beta)_{ij}$, there are $IJ-I-J+1 = (I-1)(J-1)$ linearly independent equations, so the $A$ matrix used to express this as a linear hypothesis has rank $IJ-I-J+1 = (I-1)(J-1)$. Further, by inspection,

$$\text{RSS}_{H_{AB}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{ij.})^2 + K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2,$$

since $\hat{\mu}, \hat{\alpha}_i$ and $\hat{\beta}_j$ remain as before. Hence

$$\text{RSS}_{H_{AB}} - \text{RSS} = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2 = K\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{(\alpha\beta)}_{ij}^2,$$

which has d.f. $(I-1)(J-1)$. To test $H_{AB}$, use

$$F_{AB} = \frac{\left(\mathrm{RSS}_{H_{AB}} - \mathrm{RSS}\right)/\{(I-1)(J-1)\}}{\mathrm{RSS}/\{IJ(K-1)\}} \sim F_{(I-1)(J-1),IJ(K-1)}$$

under $H_{AB}$. Now consider $H_A : \alpha_i = 0$ for all $i$. There are $I-1$ linearly independent equations here, so the rank of $A$ matrix is $I-1$. Again, by inspection, note that estimates of the remaining parameters, $\mu$, $\beta_j$ and $(\alpha\beta)_{ij}$ remain unchanged, so

$$\mathrm{RSS}_{H_A} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{ij.})^2 + JK\sum_{i=1}^{I}(\bar{y}_{i..} - \bar{y}_{...})^2, \text{ so}$$

$$\mathrm{RSS}_{H_A} - \mathrm{RSS} = JK\sum_{i=1}^{I}(\bar{y}_{i..} - \bar{y}_{...})^2 = JK\sum_{i=1}^{I}\hat{\alpha}_i^2$$

with d.f. $I-1$. Similarly,

$$\mathrm{RSS}_{H_B} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{ij.})^2 + IK\sum_{j=1}^{J}(\bar{y}_{.j.} - \bar{y}_{...})^2, \text{ so}$$

$$\mathrm{RSS}_{H_B} - \mathrm{RSS} = IK\sum_{j=1}^{J}(\bar{y}_{.j.} - \bar{y}_{...})^2 = IK\sum_{j=1}^{J}\hat{\beta}_j^2$$

with d.f. $J-1$. Therefore, for the respective tests use,

$$F_A = \frac{\left(\mathrm{RSS}_{H_A} - \mathrm{RSS}\right)/(I-1)}{\mathrm{RSS}/\{IJ(K-1)\}} \sim F_{I-1,IJ(K-1)}$$

under $H_A$ and

$$F_B = \frac{\left(\mathrm{RSS}_{H_B} - \mathrm{RSS}\right)/(J-1)}{\mathrm{RSS}/\{IJ(K-1)\}} \sim F_{J-1,IJ(K-1)}$$

under $H_B$. The decomposition of the total sum of squares along with its d.f. is as follows.

$$\begin{aligned}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}y_{ijk}^2 &= IJK\bar{y}_{...}^2 + JK\sum_{i=1}^{I}(\bar{y}_{i..} - \bar{y}_{...})^2 + IK\sum_{j=1}^{J}(\bar{y}_{.j.} - \bar{y}_{...})^2 \\ &\quad + K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{ij.})^2. \\ IJK &= 1 + (I-1) + (J-1) + (IJ - I - J + 1) + (IJK - IJ).\end{aligned}$$

ANOVA table for 2-factor analysis:

| source | d.f | SS | MS | F |
|---|---|---|---|---|
| A main effects | $I-1$ | $\text{SS}_A = JK\sum_{i=1}^{I}\hat{\alpha}_i^2$ | $\text{MS}_A =$ | $F_A = \text{MS}_A/\text{MSE}$ |
| B main effects | $J-1$ | $\text{SS}_B = IK\sum_{j=1}^{J}\hat{\beta}_j^2$ | $\text{MS}_B =$ | $F_B = \text{MS}_B/\text{MSE}$ |
| AB interactions | $(I-1)(J-1)$ | $\text{SS}_{AB} = K\sum\sum\left(\hat{\alpha\beta}\right)_{ij}^2$ | $\text{MS}_{AB} =$ | $F_{AB} = \text{MS}_{AB}/\text{MSE}$ |
| Error | $IJ(K-1)$ | $\text{RSS} = \sum\sum\sum(y_{ijk}-\bar{y}_{ij.})^2$ | $\text{MSE} = \frac{\text{RSS}}{IJ(K-1)}$ | |
| Total (c) | $IJK-1$ | $\sum\sum\sum(y_{ijk}-\bar{y}_{...})^2$ | | |
| Mean | 1 | $IJK\bar{y}_{...}^2$ | | |
| Total | $IJK$ | $\sum\sum\sum y_{ijk}^2$ | | |

ANOVA for the battery example:

| source | d.f | SS | MS | F |
|---|---|---|---|---|
| plate | 2 | 10684 | 5342 | 7.91 (2, 27) |
| temperature | 2 | 39119 | 19559 | 28.97 (2, 27) |
| interactions | 4 | 9614 | 2413 | 3.56 (4, 27) |
| error | 27 | 18231 | 675 | |
| total (c) | 35 | 77647 | | |