Chapter 6

Inference

In this chapter, we assume the normal linear model, $y \sim N(\mu, \sigma^2 I)$, with $\mu \in \mathcal{E}$ of dimension p. We have shown that the OLS estimate of μ is $\hat{\mu} = Py$; the BLUE of $a'y$ is $a'\hat{\mu}$; var$(a'\hat{\mu}) = \sigma^2 \| Pa \|^2$, and

$$\hat{\mu} \sim N(\mu, \sigma^2 P) \tag{6.1}$$

We will mostly discuss the case with var$(y) = \sigma^2 I$ because the somewhat more general case of var$(y) = \sigma^2 \Sigma$ with $\Sigma = \Sigma' > 0$ known requires only a change of inner product.

The density of y is

$$\begin{aligned} f(y) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\| y - \mu \|^2 / 2\sigma^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-[\| Py - \mu \|^2 + \| Qy \|^2]/2\sigma^2\right) \end{aligned} \tag{6.2}$$

By examination of the density, we see that the pair $(Py, \| Qy \|)$ is a complete minimal sufficient statistic for (μ, σ^2). We have seen in the last chapter that Py and Qy are independent since they are uncorrelated and normally distributed. In addition, by (6.1), $\hat{\mu} = Py$ has a normal distribution, and

$$\| Qy \|^2 / \sigma^2 \sim \chi^2(n - p)$$

a central chi-squared random variable.

6.1 Log-likelihood

The log-likelihood function for (μ, σ^2) is easily derived from (6.2) to be

$$L = -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{n}{2}\sigma^2\left(\| Py - \mu \|^2 + \| Qy \|^2\right)$$

For any σ^2, $\hat{\mu} = Py$ maximizes L, so the BLUE $\hat{\mu}$ is the maximum likelihood estimator given normal errors. To find the maximum likelihood estimate of σ^2 we maximize the *profile log-likelihood*, which is the log-likelihood function with the argument μ set equal to its estimate $\hat{\mu}$,

$$L^*(\sigma^2) = L(\mu = \hat{\mu}, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\sigma^2[0 + \| Qy \|^2] \quad (6.3)$$

L^* is just L with one of the parameters, here μ, fixed at its mle. In general, the estimate of the parameters fixed by substitution in the log-likelihood would depend on the remaining parameters, so we would condition on $\mu = \hat{\mu}(\sigma^2)$ to recognize this dependence. In this problem, the estimate of μ is the same for any value of σ^2, so this dependence is suppressed. The value of σ^2 that maximizes L^* must also maximize L. Differentiating L^* with respect σ^2 gives and solving gives $\hat{\sigma}^2_{mle} = \| Qy \|^2/n$, which differs slightly from the unbiased estimate $\hat{\sigma}^2 = \| Qy \|^2/(n-p)$. A third estimator of σ^2 is also plausible if we select a somewhat different criterion:

Theorem 6.1 *Let $\tilde{\sigma}^2(k) = \| Qy \|^2/k$. Then the value of k that minimizes the mean square error mse$(\tilde{\sigma}^2(k))$ = var$(\tilde{\sigma}^2(k))$ + bias^2 is $k = n - p + 2$.*

Proof. Homework.

Also, since $(Py, \| Qy \|^2)$ is a complete sufficient statistic, $a'\hat{\mu}$ is the uniform minimum variance unbiased estimator of $a'\mu$.

6.2 Coordinates

The results are similar for models in coordinate form. If we have $y = X\beta + \varepsilon$, with $\varepsilon \sim \mathrm{N}(0, \sigma^2 I)$, then any OLS estimator of β is a maximum likelihood estimate of β; it is unique if the columns of X provides a basis for $\mathcal{E}$, and then $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$. For rank-deficient choices of X, $\hat{\beta}$ will have a singular normal distribution. All the estimable functions will be normally distributed with positive variance, while the non-estimable functions will have zero variance.

6.3 Hypothesis testing

The general problem for the type of test we will consider starts with $y \in \Re^n$; $\mu = \mathrm{E}(y) \in \mathcal{E}$, a subspace of dimension p; Var$(y) = \sigma^2 I$; $\sigma^2 > 0$.

Suppose that $\mathcal{E}_0$ is a proper subspace of $\mathcal{E}$, which means that $\mathcal{E}_0 \subset \mathcal{E}$ but $\mathcal{E}_0 \neq \mathcal{E}$ and $\dim(\mathcal{E}_0) = q < p$. This does not exclude the possibility that $\mathcal{E}_0 = \{0\}$. The general linear hypothesis can be stated as

NH: $\mu \in \mathcal{E}_0$
AH: $\mu \in \mathcal{E}$ but $\mu \notin \mathcal{E}_0$

We start with $\Re^n$, which is decomposed into $\mathcal{E}$ and $\mathcal{E}^\perp$, $\mathcal{E} + \mathcal{E}^\perp = \Re^n$ and $\mathcal{E} \cap \mathcal{E}^\perp = \{0\}$. We again divide $\mathcal{E}$ into two orthogonal spaces, $\mathcal{E}_0$ and $\mathcal{E} - \mathcal{E}_0$. This is exactly the part of $\mathcal{E}$ not in $\mathcal{E}_0$ and must have dimension $p - q$. Thus,

$$\Re^n = \mathcal{E}^\perp + \mathcal{E}_0 + (\mathcal{E} - \mathcal{E}_0) \tag{6.4}$$

The hypothesis test consists of projecting the data Y onto the three subspaces in the decomposition (6.4), and then comparing lengths. In particular, if the null hypothesis is true, then $\| P_{\mathcal{E}-\mathcal{E}_0} y \|^2$ should be small relative to an appropriate reference measure of scale, such as $\| P_{\mathcal{E}^\perp} y \|^2/(n-p) = \| Qy \|^2/(n-p)$. This comparison of lengths is the basis of an F-test.

Before presenting general results, we consider a very special and important case of simple linear regression. The linear model can we written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, 2, \ldots, n$$

$X = (x_1, \ldots, x_n)'$, and $\mathcal{E} = \mathcal{R}(J_n, X)$; $\mu = (J_n, X)\beta$; $\beta' = (\beta_0, \beta_1)$. Now consider:

NH: $\beta_1 = 0$
AH: $\beta_1 \neq 0$

In terms of subspaces, this is equivalent to:

NH: $\mu \in \mathcal{R}(J_n)$
AH: $\mu \notin \mathcal{R}(J_n)$ but $\mu \in \mathcal{R}(J_n, X)$

Here, $p = 2$; $q = \dim(\mathcal{R}(J_n)) = 1$; $p - q = 1$. In this example $\mathcal{E}_0 = \mathcal{R}(J_n)$, and $\mathcal{E} - \mathcal{E}_0$ is the part of X that is orthogonal to the column of 1s, and it thus has spanning vector $(I - J_n{J_n}'/{J_n}'J_n)X = X - \bar{x}J_n = (x_i - \bar{x})$.

Next, suppose we consider the hypothesis:

NH: $\beta_1 = 3$ or $y_i = \beta_0 + 3x_i + e_i$
AH: $\beta_1 \neq 3$

We can rewrite the model as:

$$\tilde{y}_i = y_i - 3x_i = \beta_0 + \tilde{\beta}_1 x_i + \varepsilon_i$$

and one again tests NH: $\beta_1 = 0$.

Finally, suppose we wish to test:

NH: $\beta_0 = \beta_1 = 0$
AH: At least one of $\beta_j \neq 0$

In this case, $\mathcal{E} = \mathcal{R}(J_n, X); \mathcal{E}_0 = \{0\}$, so $\mathcal{E} - \mathcal{E}_0 = \mathcal{E} - \{0\}$.

6.3.1 The geometry of F tests

Before formally developing the usual tests, let's look at the geometry. Begin with $\mathcal{E}$, the projection on $\mathcal{E}, P_{\mathcal{E}}$, and the projection on $\mathcal{E}^{\perp} = Q_{\mathcal{E}}$. Consider testing:

NH: $\mu \in \mathcal{E}_0$
AH: $\mu \in \mathcal{E}$

Under NH, we find $\hat{\mu}_0 = P_{\mathcal{E}_0} y$ while under AH, $\hat{\mu} = P_{\mathcal{E}} y$. Look next at lengths of projections: $\| P_{\mathcal{E}_0} y \|^2 \leq \| P_{\mathcal{E}} y \|^2$. In fact, we can write:

$$\| P_{\mathcal{E}} y \|^2 = \| P_{\mathcal{E}_0} y \|^2 + \| P_{\mathcal{E}-\mathcal{E}_0} y \|^2$$

The basic idea of testing is this: If $\mu \in \mathcal{E}_0$, then, y should be almost as close to the smaller space $\mathcal{E}_0$ as it is to the bigger space $\mathcal{E}$. That is, $\| P_{\mathcal{E}-\mathcal{E}_0} y \|^2$ should be small. More specifically, under the null hypothesis,

$$\mathrm{E}\| P_{\mathcal{E}-\mathcal{E}_0} y \|^2 = (p-q)\sigma^2 + \| P_{\mathcal{E}-\mathcal{E}_0} \mu \|^2 \tag{6.5}$$

Since under both NH and AH

$$\mathrm{E}\| Q_{\mathcal{E}} y \|^2 = (n-p)\sigma^2$$

it follows that the ratio:

$$f = \frac{\| P_{\mathcal{E}-\mathcal{E}_0} y \|^2/(p-q)}{\| Q_{\mathcal{E}} y \|^2/(n-p)}$$

is independent of σ^2 and will be small if NH: $\mu \in \mathcal{E}_0$ is true and increase with $\| P_{\mathcal{E}-\mathcal{E}_0}\mu \|^2$, or as we move away from NH. Under normality, $\| P_{\mathcal{E}-\mathcal{E}_0}y \|^2/\sigma^2 \sim \chi^2(p-q,\delta^2)$, and the parameter is $\delta^2 = \| P_{\mathcal{E}-\mathcal{E}_0}\mu \|^2/\sigma^2$. Since $\| Q_{\mathcal{E}}y \|^2 \sim \chi^2(n-p)$, we have that f has a non-central F-distribution in general, and $f \sim F(p-q, n-p, \delta^2)$, the usual F test.

It is never necessary to compute $\| P_{\mathcal{E}-\mathcal{E}_0}y \|^2$ directly, since (6.5) can be used to find it by subtraction, and thus, writing $\mathit{RSS}_{AH} = \| Q_{\mathcal{E}}y \|^2$, and $\mathit{RSS}_{NH} = \| Q\mathcal{E}_0 y \|^2$,

$$f = \frac{[\mathit{RSS}_{NH} - \mathit{RSS}_{AH}]/(p-q)}{\mathit{RSS}_{NH}/(n-p)}$$

Computer programs generally obtain the needed sums of squares using the QR factorization. Start with X and obtain $X = Q_1 R$ where Q_1 has orthonormal columns that span the column space of X. Then, for example, $\| Q_{\mathcal{E}}y \|^2 = \| (I - P_{\mathcal{E}})y \|^2 = \| y - Q_1 Q_1{}'y \|^2 = y'y - (Q_1 y)'(Q_1 y)$, which is just the difference of two inner products. Computation of quantities like $\| P_{\mathcal{E}-\mathcal{E}_0}y \|^2$ are also straightforward if the columns of X can be permuted so that the first q columns span $\mathcal{E}_0$. Then computing $\| P_{\mathcal{E}_0}y \|^2$ is easy, and $\| P_{\mathcal{E}-\mathcal{E}_0}y \|^2$ can be found by subtraction.

6.4 Likelihood ratio tests

Suppose $y \sim N(\mu, \sigma^2 I), \sigma^2 > 0, \mu \in \mathcal{E}$ and consider testing the same general hypothesis as in the last section:

NH: $\mu \in \mathcal{E}_0$
AH: $\mu \notin \mathcal{E}_0$ but $\mu \in \mathcal{E}$

The likelihood function for (μ, σ^2) is:

$$L(\mu,\sigma^2;y) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp(-\| y-\mu \|^2/2\sigma^2)$$

The likelihood ratio statistic is

$$\Lambda(y) = \frac{\sup_{NH} L(\mu,\sigma^2;y)}{\sup_{AH} L(\mu,\sigma^2;y)} \tag{6.6}$$

here $\sup_{NH}$ means that $\mu \in \mathcal{E}_0$ and $\sigma^2 > 0$ while $\sup_{AH}$ means $\mu \in \mathcal{E}$ and $\sigma^2 > 0$. Evidence against NH corresponds to $\Lambda(y) < k$ for some k. We find:

$$\begin{aligned} \hat{\mu}_0 &= P_{\mathcal{E}_0} y; \\ \hat{\sigma}_0^2 &= \| Q_{\mathcal{E}_0} y \|^2 / n \\ L(\hat{\mu}_0; \hat{\sigma}_0^2; y) &= c \times (\| Q_{\mathcal{E}_0} y \|^2 / n)^{-n/2} \end{aligned} \tag{6.7}$$

where c is a constant that does not depend on the data. Under the alternative hypothesis AH, we find:

$$\begin{aligned} \hat{\mu} &= P_{\mathcal{E}} y \\ \hat{\sigma}^2 &= \| Q_{\mathcal{E}} y \|^2 / n \\ L(\hat{\mu}, \hat{\sigma}^2; y) &= c \times (\| Q_{\mathcal{E}} y \|^2 / n)^{-n/2} \end{aligned} \tag{6.8}$$

and the constant c is the same for both hypotheses. Substituting (6.7) and (6.8) into the (6.6) and simplifying gives:

$$\Lambda(y) = \left(\frac{\| Q_{\mathcal{E}_0} y \|^2 / n}{\| Q_{\mathcal{E}} y \|^2 / n} \right)^{-n/2}$$

Now, $\Lambda(y) \leq k$ if and only if $\| Q_{\mathcal{E}_0} y \|^2 / \| Q_{\mathcal{E}} y \|^2 \geq k_1$. The quadratic forms involved are easy to compute, since:

$$\begin{aligned} I = P_{\mathcal{E}} + Q_{\mathcal{E}} &= P_{\mathcal{E}_0} + P_{\mathcal{E}-\mathcal{E}_0} + Q_{\mathcal{E}} \\ Q_{\mathcal{E}_0} &= P_{\mathcal{E}-\mathcal{E}_0} + Q_{\mathcal{E}} \\ \| Q_{\mathcal{E}_0} y \|^2 &= \| P_{\mathcal{E}-\mathcal{E}_0} y \|^2 + \| Q_{\mathcal{E}} y \|^2 \end{aligned}$$

Thus:

$$\begin{aligned} \frac{\| Q_{\mathcal{E}_0} y \|^2}{\| Q_{\mathcal{E}} y \|^2} &= \frac{\| Q_{\mathcal{E}} y \|^2 + \| P_{\mathcal{E}-\mathcal{E}_0} y \|^2}{\| Q_{\mathcal{E}} y \|^2} \\ &= 1 + \frac{p-q}{n-p} f \end{aligned}$$

where f is the usual F statistic for this hypothesis,

$$f = \frac{\| P_{\mathcal{E}-\mathcal{E}_0} y \|^2 / (p-q)}{\| Q_{\mathcal{E}} y \|^2 / (n-p)}$$

which has the $F(p-q, n-p, \delta^2 = \| P_{\mathcal{E}-\mathcal{E}_0}\mu \|^2/\sigma^2)$ distribution. Thus, the likelihood ratio test is a monotonic function of f, and so the F-test is the likelihood ratio test. If the null hypothesis is true, then $\delta^2 = 0$ and $f \sim F(p-q, n-p)$. The central F is used to find significance levels of the test, and the non-central F can be used to construct power functions, as in Section 6.10.

6.5 General Coordinate Free hypotheses

In the general coordinate free approach to linear models, the general linear hypothesis is:

NH: $B\mu = 0$
AH: $B\mu \neq 0$

where B is an $r \times n$ matrix and without loss of generality, $\rho(B) = r$. We shall convert this into the format of the general linear hypothesis that we have seen previously.

Kronecker products. We introduce a bit of new notation that will (eventually) simplify some discussions. Suppose A is an $m \times n$ matrix and B is a $p \times q$ matrix. The Kronecker product of A and B, written $A \otimes B$, is defined by

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \vdots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}$$

Some useful properties of the Kronecker product are:

$$\begin{aligned} A \otimes (B \otimes C) &= (A \otimes B) \otimes C \\ (A \otimes B)(C \otimes D) &= (AC \otimes BD) \\ (A \otimes B)' &= (A' \otimes B') \\ (A+B) \otimes (C+D) &= A \otimes C + A \otimes D + B \otimes C + B \otimes D \\ a(A \otimes B) &= (aA \otimes B) = (A \otimes aB) \\ (A \otimes B)^{-1} &= A^{-1} \otimes B^{-1} \\ \text{tr}(A \otimes B) &= \text{tr}(A) \times \text{tr}(B) \end{aligned}$$

We return to a specific example of a general linear hypothesis. Suppose $y \sim N(\mu, \sigma^2 I)$ is 8×1, and let $\mu = (\mu_1,\ u_2, \mu_3, \mu_4)'$, and $J_2 = (1, 1)'$. Then suppose that $\mu = \mu_0 \otimes J_2$, so

$$\mu' = (\mu_1, \mu_1, \mu_2, \mu_2, \mu_3, \mu_3, \mu_4, \mu_4)'$$

This is the one-way model with four groups and two observations per group. The estimation space $\mathcal{E}$ is then $\mathcal{R}(I_4 \otimes J_2)$, and the columns of this matrix are an orthogonal basis for $\mathcal{E}$.

Consider the NH: $B_1\mu = 0$ with B_1 given by:

$$B_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & -2 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

which picks out contrasts $\mu_1 - \mu_3$ and $\mu_1 - 2\mu_2 + \mu_3$ to be equal to zero while completely ignoring group four. There are other matrices that would pick out the same pair of restrictions, such as

$$B_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -2 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Since $\mu \in \mathcal{E}$, the hypotheses $B_1\mu = 0$ and $B_2\mu = 0$ must be equivalent to $B_1(P_{\mathcal{E}}\mu) = B_2(P_{\mathcal{E}}\mu) = 0$, or equivalently if B is any matrix so that $B\mu = B_1\mu$,

$$(P_{\mathcal{E}}B')' = 0$$

We next compute $P_{\mathcal{E}}$. If e_i is the i-th canonical basis vector (the ith column of I_4, then

$$P_{\mathcal{E}} = \sum_{i=1}^{4} \frac{(e_i \otimes J_2)(e_i \otimes J_2)'}{(e_i \otimes J_2)'(e_i \otimes J_2)} = (I_4 \otimes J_2 J_2'/2)$$

and an easy calculation gives

$$\begin{aligned} (P_{\mathcal{E}}B_1')' &= \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 \\ 1 & 1 & -2 & -2 & 1 & 1 & 0 & 0 \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} 1 & 0 & -1 & 0 \\ 1 & -2 & 1 & 0 \end{pmatrix} \otimes (1,1) \end{aligned}$$

We can now describe the subspaces involved. Under the null hypothesis, $\mu \notin \mathcal{R}(P_{\mathcal{E}}B_1')$, and hence $\mathcal{E} - \mathcal{E}_0 = \mathcal{R}(P_{\mathcal{E}}B_1')$. Although we don't really need to find $\mathcal{E}_0$ for the F-test, we can compute $\mathcal{E}_0$ to be the span of any A matrix whose columns provide a completion of the rows of $P_{\mathcal{E}}B_1{}'$ as a basis for $\mathcal{E}$. Then A provides a basis for $\mathcal{E}_0$. One such basis for $\mathcal{E}_0$ is

$$A' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \otimes (1, 1)$$

A general expression for $\mathcal{R}(A)$ is $\mathcal{R}(P_{\mathcal{E}}Q_{P_{\mathcal{E}}B'})$ Thus the general coordinate free hypothesis is equivalent to

NH: $\mu \in \mathcal{E}_0$
AH: $\mu \notin \mathcal{E}_0$ but $\mu \in \mathcal{E}$

where

$$\mathcal{E} - \mathcal{E}_0 = \mathcal{R}(P_{\mathcal{E}}B') \text{ of dimension } r$$

and

$$\mathcal{E}_0 = \mathcal{R}(P_{\mathcal{E}}Q_{P_{\mathcal{E}}B'}) \text{ of dimension } \rho(\mathcal{E}) - r$$

The statistic for testing the general coordinate free hypothesis is then

$$f = \frac{\parallel P_{\mathcal{E}-\mathcal{E}_0}y \parallel^2 / r}{\parallel Q_{\mathcal{E}}y \parallel^2 / (n - \rho(\mathcal{E}))}$$

which is distributed as a non-central $F(r, n - \rho(\mathcal{E}), \delta^2)$, with

$$\delta^2 = \frac{\parallel P_{\mathcal{E}-\mathcal{E}_0}\mu \parallel^2}{\sigma^2}$$

If the hypothesis is NH: $B\mu = h$, for some known vector h, we can proceed by translation. Suppose we can find an $\alpha \in \mathcal{E}$ such that $B\alpha = h$. Then $B\mu = h\ B\alpha$ or $B(\mu - \alpha) = 0$. That is, translate the problem $y \sim N(\mu, \sigma^2 I)$ to $y \sim N(\mu - \alpha, \sigma^2 I)$. We can always find a specific α as follows. Let $B = VDU'$ be the singular value factorization of B, so V and D are $r \times r$, and U is $n \times r$. Then one solution is given by $\alpha = UD^{-1}V'h$.

6.6 Parametric hypotheses

We return to a parameterized linear model,

$$y = X\beta + \varepsilon;\ \mathrm{E}(\varepsilon) = 0;\ \mathrm{Var}(\varepsilon) = \sigma^2 I;\ \mathcal{E} = \mathcal{R}(X);\ X : n \times r \text{ rank } p \leq r.$$

Suppose we partition $\beta' = (\beta_1{}', \beta_2{}')$, where β_1 is $q \times 1$ and consider testing

NH: $\beta_2 = 0$
AH: $\beta_2 \neq 0$

assuming, of course, that β_2 is estimable. We can partition the model to conform to the hypothesis by writing $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$. We can identify $\mathcal{E}_0 = \mathcal{R}(X_1)$ and $P_{\mathcal{E}_0}$ = Projection on columns of X_1. The space $\mathcal{E} - \mathcal{E}_0$ is just the part of the column space of X_2 that is orthogonal to X_1, given by $\mathcal{E} - \mathcal{E}_0 = \mathcal{R}((I - P_1)X_2)$ All the computations are thus quite straightforward. If the NH were: $\beta_2 = \beta_{20}$, we can proceed by translation.

This is equivalent to the testing situation:

NH: $y = X_1\beta_1 + \varepsilon$
AH: $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$

Letting P_1 be the projection on the column space of X_1, we can rewrite the AH model as:

$$AH^* : y = X_1\beta_1^* + (I - P_1)X_2\beta_2 + \varepsilon = X_1\beta_1^* + X_{2.1}\beta_2$$

where $X_{2.1}$ is the part of X_2 orthogonal to X_1. The two representations are the same only if $\mathcal{R}(X_2) = \mathcal{R}(X_{2.1})$, which is guaranteed if X has full rank, but not otherwise. Even in the full rank case, is it legitimate to use the same symbol β_2 in AH and AH*? Is it necessary to use a different symbol β_1 and β_1^*? Is the notation for β_1 under NH appropriate?

In the orthogonalized version, the F-test is immediate, since *SSreg* is just the length of the projection onto the column space of $X_{2.1}$, and, since X_1 and $X_{2.1}$ are orthogonal, this is just $Y'P_{2.1}Y$, or, assume $X_{2.1}$ is of full rank,

$$Y'X_{2.1}(X_{2.1}'X_{2.1})^{-1}X_{2.1}'Y$$

with df equal to the number of columns in $X_{2.1}$. The Analysis of Variance table is given by Table 6.6.

Finally, we turn to the general parametric hypothesis. Let $A_1 : r \times p$ be a rank r matrix so that the columns of A_1' are all in $\mathcal{R}(X')$, insuring that $A_1'\beta$ is a vector of estimable functions of β. Suppose we were interested in the hypothesis:

NH: $\Psi_1 = A_1\beta = 0$
AH: $\Psi_1 = A_1\beta \neq 0$

Table 6.1: Analysis of Variance for a parametric hypothesis

Source	*df*	*SS*	*MS*	F	*E(MS)*
$X_1 + X_{2.1}$	p	$\parallel P_X y \parallel^2$			
X_1	q	$\parallel P_1 y \parallel^2$	$\parallel P_1 y \parallel^2 / q$		
$X_{2.1}$	$p - q$	$\parallel P_{2.1} y \parallel^2$	$\parallel P_{2.1} y \parallel^2 / (p-q)$		
$\mathcal{E}^\perp$	$n - p$	$\parallel Q_X y \parallel^2$	$\hat{\sigma}^2 = \parallel Q_{\mathcal{E}} y \parallel^2 / (n-p)$		

The structure of this hypothesis indicates that the coordinates β are not of primary interest, rather interest centers on the set of estimable functions Ψ_1. This is generally true for over-parameterized models, since the elements of β are usually not estimable.

Since A_1 is of full row rank, we can always find a matrix A_0 so that the square matrix

$$A = \begin{pmatrix} A_0 \\ A_1 \end{pmatrix}$$

is of full rank so A^{-1} exists. Then:

$$y = X\beta + \varepsilon = XA^{-1}A\beta + \varepsilon = Z\Psi + \varepsilon = Z_0\Psi_0 + Z_1\Psi_1 + \varepsilon$$

where $Z = XA^{-1}$, so we can now proceed as in the multiple regression case just discussed. This 'trick' of reparameterizing to get the parameters reduces a new problem to an old one. If we have constructed A so that $A_0{}'A_1 = 0$, then $Z_0{}'Z_1 = 0$, so the F test becomes particularly simple, since then $(I - P_{Z_0})Z_1 = Z_1$.

We can also get the same result, but without finding A_0. Since $\Psi_1 = A_1\beta$ is estimable, the columns of $A_1{}' \in \mathcal{R}(X')$. Thus there is a $B_1 : r \times n$ such that

$$A_1{}' = X'B_1{}'$$

or

$$A_1 = B_1 X$$

Thus, we can write

$$A_1\beta = 0 \Leftrightarrow B_1 X\beta = B_1\mu = 0$$

and this is exactly the same as the general coordinate free hypothesis test derived in the last section.

We will proceed with the computations as if X has full column rank. If the less than full rank case, simply replace $(X'X)^{-1}$ with any generalized inverse to get the same result. We find

$$\begin{aligned}
\mathcal{E} - \mathcal{E}_0 &= \mathcal{R}(P_{\mathcal{E}} B_1{}') \\
&= \mathcal{R}(X(X'X)^{-1}X'B_1{}') \\
&= \mathcal{R}(X(X'X)^{-1}A_1{}') \\
\mathcal{E}_0 &= \mathcal{R}(P_{\mathcal{E}} Q_{\mathcal{R}(P_{\mathcal{E}} B_1{}')}) \\
&= \mathcal{R}(X(X'X)^{-1}X' - X(X'X)^{-1}A_1{}')
\end{aligned}$$

A bit of straightforward algebra will give expressions both for the projection on $\mathcal{E} - \mathcal{E}_0$ and for its length. We find

$$\begin{aligned}
P_{\mathcal{E}-\mathcal{E}_0} y &= X(X'X)^{-1}A_1{}'[A_1(X'X)^{-1}X'X(X'X)^{-1}A_1{}']^{-1}A_1(X'X)^{-1}X'y \\
&= X(X'X)^{-1}A_1{}'[A_1(X'X)^{-1}A_1{}']^{-1}A_1(X'X)^{-1}X'Y
\end{aligned}$$

and

$$\begin{aligned}
\| P_{\mathcal{E}-\mathcal{E}_0} y \|^2 &= Y'X(X'X)^{-1}A_1{}'[A_1(X'X)^{-1}A_1{}']^{-1}A_1(X'X)^{-1}X'Y \\
&= \hat{\beta}'A_1{}'[A_1(X'X)^{-1}A_1{}']^{-1}A_1\hat{\beta}
\end{aligned}$$

Since $\hat{\psi}_1 = A_1\hat{\beta}$, we can write $\text{Var}(\hat{\psi}_1) = \sigma^2[A_1(X'X)^{-1}A_1{}']$, we can rewrite the last result as:

$$\| P_{\mathcal{E}-\mathcal{E}_0} y \|^2 = \sigma^2 \hat{\psi}_1{}'[\text{Var}(\hat{\psi}_1)]^{-1}\hat{\psi}_1$$

and the F test is

$$\begin{aligned}
F &= \frac{\| P_{\mathcal{E}-\mathcal{E}_0} y \|^2}{(p-q)\hat{\sigma}^2} \\
&= \frac{\hat{\psi}_1[\widehat{\text{Var}}(\hat{\psi}_1)]^{-1}\hat{\psi}_1}{p-q}
\end{aligned}$$

and $F \sim F(p-q, n-p, \delta^2)$. The non-centrality parameter as a function of A_1 and β is

$$\delta^2 = \frac{\beta'A_1{}'[A_1(X'X)^{-1}A_1{}']^{-1}A_1\beta^2}{\sigma^2}$$

and $\delta^2 = 0$ only if $A_1\beta = \psi_1 = 0$.

6.7 Relation of least squares estimators under NH and AH

In general regression situations, it is often of interest to compare estimates of parameters under different models/hypotheses, particularly when the Xs are covariates. We can proceed as follows to make this comparison in the full rank case:

$$\begin{aligned}\hat{\mu}_{NH} &= X\hat{\beta}_{NH} = P_{\mathcal{E}}y - P_{\mathcal{E}-\mathcal{E}_0}y \\ &= X(X'X)^{-1}X'y - X(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A(X'X)^{-1}X'Y\end{aligned}$$

while

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

or

$$X\hat{\beta}_{NH} = X\hat{\beta}_{AH} - X(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A(X'X)^{-1}X'Y$$

Since X is of full rank,

$$\hat{\beta}_{NH} = (I - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A)\hat{\beta}_{AH}$$

$$\mathrm{E}(\hat{\beta}_{NH}) = (I - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A)\beta_{AH}$$

which can be used to explore biases, etc. $A(X'X)^{-1}A'$ is called the *alias matrix*.

Example. The one-way anova model can be written as $y_{ij} = \beta_i + \varepsilon_{ij},\ i = 1, \ldots, b; j = 1, \ldots, n_i;\ \sum n_i = n; \varepsilon \sim N(0, \sigma^2 I)$. Then, as usual, $\mathcal{E} = \mathcal{R}(X_1, \ldots, X_p)$, where X_i is a vector of zeroes, except for the rows from group i where it is one. Now $\dim(\mathcal{E}) = p$ and $P_{\mathcal{E}}y$ is a vector with value $\bar{y}_{i+}$ for all observations at level i; $Q_{\mathcal{E}}y = (y_{ij} - \bar{y}_{i+})$ and $\hat{\sigma}^2 = \parallel Q_{\mathcal{E}}y \parallel^2/(n-p) = \sum\sum(y_{ij} - \bar{y}_{i+})^2/(n-p)$. Suppose that the p-th treatment is a control and we wish to test:

NH: $\beta_p = \frac{1}{p-1}\sum_{i=1}^{p-1}\beta_i$
AH: NH not true

Under NH, $\hat{\beta}' = (\bar{y}_{1+}, \ldots \bar{y}_{p+})$, $\widehat{\mathrm{Var}(\hat{\beta})} = \hat{\sigma}^2\mathrm{diag}(1/n_i) = \hat{\sigma}^2(X'X)^{-1}$. To apply the previous full-rank set-up, write the null hypothesis in the form: NH: $A\beta = 0;\ A = (1/(p-1), \ldots, 1/(p-1), -1)$ so that

$$A\hat{\beta} = \frac{1}{p-1}\sum_{i=1}^{p-1}\bar{y}_{i+} - \bar{y}_{p+}$$

and

$$A(X'X)^{-1}A' = \left(\frac{1}{p-1}\right)^2 \sum_{i=1}^{p-1} n_i^{-1} + n_p^{-1}$$

Then $p - q = 1$ and:

$$\begin{aligned} F &= (A\hat{\beta})'[A(X'X)^{-1}A']^{-1}(A\hat{\beta})/\hat{\sigma}^2 \\ &= (A\hat{\beta})'[\widehat{\text{Var}}(A\hat{\beta})]^{-1}(A\hat{\beta}) \\ &= \frac{\frac{1}{p-1}\left[\sum_{i=1}^{p-1} \bar{y}_{i+} - \bar{y}_{p+}\right]^2}{\hat{\sigma}^2[\frac{1}{(p-1)^2}\sum_{i=1}^{p-1} n_i^{-1} + n_p^{-1}]} \sim F(1, n-p, \sigma^2) \end{aligned}$$

To find δ^2, simply substitute μ_i and σ^2 for corresponding statistics into F. Finally, we can examine the aliases under the NH. We get:

$$\begin{aligned} \hat{\beta}_{NH} &= (I - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A)\hat{\beta}_{AH} \\ &= \hat{\beta}_{AH} - \frac{(X'X)^{-1}A'A\hat{\beta}_{AH}}{[A(X'X)^{-1}A']^{-1}} \\ &= \hat{\beta}_{AH} - (X'X)^{-1}A'\frac{(p-1)^{-1}\sum \bar{y}_{i+} - \bar{y}_{p+}}{(p-1)^{-2}\sum n_i^{-1} + n_p^{-1}} \end{aligned}$$

The j-th element of this vector is:

$$\bar{y}_{j+} - c_j(\frac{1}{p-1}\sum \bar{y}_{i+} - \bar{y}_{p+})$$

where

$$c_j^{-1} = \frac{n_j}{p-1}\sum_{i=1}^{p-1} n_i^{-1} + \frac{n_j(p-1)}{n_p}$$

The adjustment c_j decreases with n_j so we get more adjustment on the means that are relatively more variable.

For this testing situation, the earlier ANOVA can be subdivided into Table 6.7 The "Remainder" line in the ANOVA can be used to test for differences between the first $p - 1$ treatments (ignoring treatment p).

6.8 Analysis of Variance Tables

The computations for the division of sums of squares into components due to various sources are usually combined into an "Analysis of Variance" table, which has the canonical form given in Table 6.8. Mean squares are of course SS/df.

Table 6.2: Analysis of variance table

Source	*df*	*SS*	*MS*	F	*E(MS)*
Mean= $\mathcal{E}_0$	1	$n\bar{y}_{++}^2$			$\sigma^2 + n\beta^2$
Treatments=$\mathcal{E} - \mathcal{E}_0$	$p-1$	$\sum n_i(\bar{y}_{i+} - \bar{y}_{++})^2$			$\sigma^2 + \sigma^2\delta^2/(p-1)$
$\mathcal{R}(X(X'X)^{-1}A')$	1				
Remainder	$p-2$				
Error	$n-p$	$\sum\sum(y_{ij} - \bar{y}_{i+})^2$			σ^2

Table 6.3: Canonical Analysis of Variance Table

Source	*df*	*SS*	*MS*	F	*E(MS)*
$\mathcal{E}$	p	$\parallel P_{\mathcal{E}} y \parallel^2$			
$\mathcal{E}_0$	q	$\parallel P_{\mathcal{E}_0} y \parallel^2$	$\parallel P_{\mathcal{E}_0} y \parallel^2/q$		
$\mathcal{E} - \mathcal{E}_0$	$p-q$	$\parallel P_{\mathcal{E}-\mathcal{E}_0} y \parallel^2$	$\parallel P_{\mathcal{E}_0^\perp \cap \mathcal{E}} y \parallel^2/(p-q)$		
$\mathcal{E}^\perp$	$n-p$	$\parallel Q_{\mathcal{E}} y \parallel^2$	$\hat{\sigma}^2 = \parallel Q_{\mathcal{E}} y \parallel^2/(n-p)$		

To calculate expected mean squares, recall that if $\text{Var}(y) = \sigma^2 I$, $\text{E}(y) = \mu$, then $\text{E}(\| P_M y \|^2) = \dim(M)\sigma^2 + \| P_M \mu \|^2$, so

$$\text{E}\left(\frac{\| P_M y \|^2}{\dim(M)}\right) = \sigma^2 + \frac{1}{\dim(M)}\| P_M \mu \|^2$$

so we simply substitute μ for y in the expression for the mean square and add σ^2.

Power considerations usually require calculation of the non-centrality parameter $\delta^2 = \| P_{\mathcal{E}-\mathcal{E}_0} \mu \|^2$. Typically, δ^2 is a function of the unknown μ and of sample sizes.

6.9 F tests and t tests

Recall that we have justified F-test as the likelihood ratio tests for hypotheses of interest assuming normality. The t-tests are usually justified by looking at ratios like

$$\frac{\text{Estimate} - \text{hypothesized value}}{\text{Standard error of the estimate}}$$

where the numerator is normally distributed, and the denominator is an independent χ^2-distributed estimate of its standard deviation. This type of test, comparing an estimate to an estimate of its error, is called in general a *Wald test*. Wald tests and likelihood ratio tests are generally asymptotically equivalent, but for some tests in the normal linear model they are in fact identical.

Suppose $y \sim N(\mu, \sigma^2 I), \mu \in \mathcal{E}$, and we want to test

NH: $c'\mu = 0, c \in \Re^n$,
AH: $c'\mu \neq 0$

which is a specialization of the general coordinate free hypothesis test, with $r = 1$. Then:

$$\begin{aligned}\mathcal{E}_0 = \{\mu | \mu \in \mathcal{E}, \mu \perp c\} &= \{P_{\mathcal{E}} z | z' P_{\mathcal{E}} c = 0, z \in \Re^n\} \\ &= \{P_{\mathcal{E}} z | z \in \mathcal{R}^{\perp}(P_{\mathcal{E}} c)\}\end{aligned}$$

If $c \in \mathcal{E}^{\perp}$, then $\mathcal{E} = \mathcal{E}_0$ and there is nothing to test. Thus, assume $P_{\mathcal{E}} c \neq 0$, and $\mathcal{E} - \mathcal{E}_0 = \mathcal{R}(P_{\mathcal{E}} c)$.

Let $z = P_{\mathcal{E}}c$. Then:

$$\begin{aligned} F &= \frac{\| P_{\mathcal{E}-\mathcal{E}_0} y \|^2}{\hat{\sigma}^2} \\ &= \frac{\| \frac{(z,y)}{(z,z)} z \|^2}{\hat{\sigma}^2} \\ &= \frac{(z,y)^2}{(z,z)\hat{\sigma}^2} \\ &= \frac{(c,\hat{\mu})^2}{\| P_{\mathcal{E}}c \|^2 \hat{\sigma}^2} \\ &= \frac{(\text{BLUE of } c'\mu)^2}{(\text{SE of } c'\mu)^2} \end{aligned}$$

since $(z, y) = (P_{\mathcal{E}}c, y) = c(\hat{\mu})$ and $\text{Var}(c'\hat{\mu}) = \hat{\sigma}^2 \| P_{\mathcal{E}}c \|^2$.

Under normality,

$$\frac{c'\hat{\mu} - c'\mu}{\sigma \| P_{\mathcal{E}}c \|} \sim N(0, 1)$$

and

$$\hat{\sigma}^2 \sim \sigma^2 \chi^2(n-p)/(n-p)$$

independent of $\hat{\mu}$. So,

$$\begin{aligned} t &= \frac{(c'\hat{\mu} - c'\mu)/(\sigma \| P_{\mathcal{E}}c \|)}{\sqrt{\hat{\sigma}^2/\sigma^2}} \\ &= \frac{c'\hat{\mu}}{\hat{\sigma} \| P_{\mathcal{E}}c \|} \\ &= \frac{\mathrm{N}(0,1)}{\sqrt{\chi^2(n-p)/(n-p)}} \sim t(n-p) \end{aligned}$$

and $F = t^2$.

Now suppose that the hypothesis is

NH: $c'\mu = k \neq 0$
AH: $c'\mu \neq k$

Under NH, we can find (many) vectors $\alpha \in \mathcal{E}$ such that $c'\alpha = k$. We can always do this if $P_{\mathcal{E}}c \neq 0$ (if $P_{\mathcal{E}}c \neq 0$, then there is at least one vector $\alpha^* \in \mathcal{E}$ such that $(c, \alpha^*) \neq 0$. Let $\alpha = k\alpha^*/(c, \alpha^*))$. The hypothesis can then we written as NH: $c'\mu = c'\alpha$ or NH: $c'(\mu - \alpha) = 0$. This suggests translating y to $y - \alpha$ and proceeding as before. This yields:

$$\begin{aligned} t &= \frac{c'\hat{\mu}_\alpha}{\hat{\sigma}^2_\alpha \| P_{\mathcal{E}} c \|^2} \\ &= \frac{c' P_{\mathcal{E}}(y - \alpha)}{\hat{\sigma}^2_\alpha \| P_{\mathcal{E}} c \|^2} \end{aligned}$$

But,

$$\hat{\sigma}^2_\alpha = \frac{\| Q_{\mathcal{E}}(y - \alpha) \|^2}{n - p} = \frac{\| Q_{\mathcal{E}} y \|^2}{n - p} = \hat{\sigma}^2$$

because $\alpha \in \mathcal{E}$. Thus,

$$\begin{aligned} t &= \frac{c'\hat{\mu}_\alpha}{\hat{\sigma}^2_\alpha \| P_{\mathcal{E}} c \|^2} \\ &= \frac{c' P_{\mathcal{E}} y - c'\alpha}{\hat{\sigma} \| P_{\mathcal{E}} c \|} \\ &= \frac{c'\hat{\mu} - k}{\hat{\sigma} \| P_{\mathcal{E}} c \|} \end{aligned}$$

as expected. Finally it is clear that we can construct confidence intervals for $c'\mu$ in the usual way: A $1 - \alpha \times 100\%$ confidence interval is the set of points in the range

$$c'\hat{\mu} \pm t(1 - \alpha/2; n - p)\hat{\sigma} \| P_{\mathcal{E}} c \|$$

6.10 Power and Sample Size

In the fixed effects linear model $y \sim \mathrm{N}(\mu, \sigma^2 I)$, with $\mu \in \mathcal{E} \subset \Re^n$, the general linear hypothesis is a test of $\mu \in \mathcal{E}_0 \subset \mathcal{E}$ versus $\mu \in \mathcal{E}$. The test statistic is given by

$$f = \frac{\| P_{\mathcal{E}-\mathcal{E}_0} y \|^2 / \nu_1}{\| Q_{\mathcal{E}} y \|^2 / \nu_2}$$

where $\nu_1 = \rho(\mathcal{E} - \mathcal{E}_0)$ and $\nu_2 = n - \rho(\mathcal{E})$. and $f \sim F(\nu_1, \nu_2, \delta^2)$. The non-centrality parameter δ^2 is

$$\delta^2 = \| P_{\mathcal{E}-\mathcal{E}_0} \mu \|^2 / \sigma^2$$

Although not completely clear from the definition, δ^2 depends on the sample size, or in design problems, on the number of replications.

We define the *power* of a statistical test, denoted as $1-\beta$, to be the probability of detecting a false null hypothesis, or, more precisely, the probability of rejecting the null hypothesis at a given level α, for a given value of δ^2, or

$$1-\beta = \Pr\left[F(\nu_1,\nu_2,\delta^2) > F(\alpha;\nu_1,\nu_2,\delta^2=0)\right]$$

Example. Consider a one-way model with t groups and m observations per group, for a total of $n = tm$ observations. The estimation space $\mathcal{E}$ is spanned by $\mathcal{R}(I_t \otimes J_m)$. Suppose that $\mathcal{E}_0 = \mathcal{R}(J_n)$, and consider the test of $\mu \in \mathcal{E}_0$ versus the alternative $\mu \in \mathcal{E}$, the usual test for the equality of group means. For this test, the non-centrality parameter is

$$\begin{aligned}\delta^2 &= \parallel P_{\mathcal{E}-\mathcal{E}_0}\mu \parallel^2/\sigma^2 \\ &= \sum_{i=1}^{t} m(\mu_i - \bar{\mu}_+)^2/\sigma^2\end{aligned}$$

For example, suppose $t = 5$, $m = 10$. To compute the power of the F test, we need to specify the significance level, say $\alpha = 0.05$, and a particular value of δ^2 by specifying a pattern for the group means. For this example, suppose we consider the alternative hypothesis to be $\mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$ and $\mu_5 = k\sigma$. Thus four of the group means are equal but only group five may be different. Then δ^2 can be computed to be $4mk^2/5$. To get the power, we first need to get the critical value. Using R,

```
> qf(.95,4,45)   get the upper tail of the central F distribution
[1] 2.578739
```

At $k = 1, 2, 3$, the power is:

```
> pf(qf(.95,4,45),4,45,4*10*c(1,2,3)^2/5, lower.tail=FALSE)
[1]  0.5540384 0.9959826 0.9999999
```

A graph of the power as a function of k is shown in Figure 6.1 The following generates this graph:

```
> kvals <- seq(0,3,length=41)
> pow <- pf(qf(.95,4,45),4,45,4*10*kvals^2/5, lower.tail=FALSE)
> plot(kvals,pow,type="l",xlab="k",ylab="Power")
> pow1 <- pf(qf(.95,4,20),4,20,4*5*kvals^2/5, lower.tail=FALSE)
> lines(kvals,pow1)
```

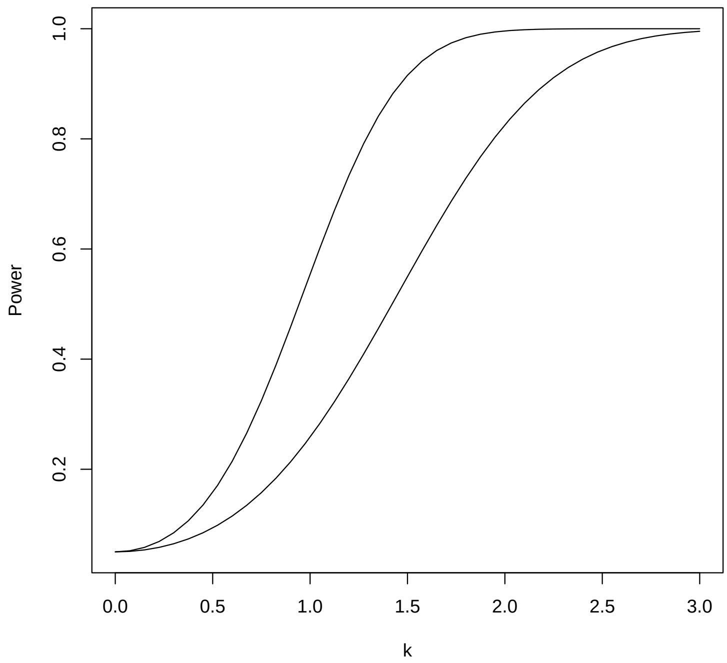

Figure 6.1: Power for one-way anova with the alternative as specified in the text.

This graph shows the power function both for the case $m = 10$ and for the case $m = 5$; the latter has uniformly lower power. For example, at $k = 1$, the power with $m = 5$ is 25% while for $m = 10$ it is 55%.

6.11 Simple linear regression

The simple linear regression model is $\mathrm{E}(y) = \beta_0 1 + \beta_1 x + \varepsilon$, with $\varepsilon \sim N(0, \sigma^2 I)$ and $x = (x_1, \dots, x_n)'$, and at least two of the x_i are distinct. Then $\mathcal{E} = \mathcal{R}(J_n, X)$, and $p = \dim(\mathcal{E}) = 2$. Consider the test of:

NH: $\beta_1 = 0$
AH: $\beta_1 \neq 0$

Under NH, $\mathcal{E}_0 = \mathcal{R}(J_n)$ and $\mathcal{E} - \mathcal{E}_0 = \mathcal{R}(X - \bar{x}J_n)$. Thus, the F test is just

$$\begin{aligned} F &= \frac{\| P_{\mathcal{E}-\mathcal{E}_0} y \|^2/1}{\| Q_{\mathcal{E}} y \|^2/(n-2)} \\ &= \frac{\| Q_{\mathcal{E}} y \|^2 - \| Q_{\mathcal{E}_0} y \|^2}{\| Q_{\mathcal{E}} y \|^2/(n-2)} \\ &= \frac{\| P_{\mathcal{E}_0} y \|^2 - \| P_{\mathcal{E}} y \|^2}{\| Q_{\mathcal{E}} y \|^2/(n-2)} \end{aligned}$$

Each of these projections are very easy to evaluate for simple linear regression. We find:

$$\begin{aligned} \| Q_{\mathcal{E}_0} y \|^2 &= \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = \| y \|^2 - \| P_{\mathcal{E}_0} y \|^2 \\ \| Q_{\mathcal{E}} y \|^2 &= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum (y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x}))^2 \end{aligned}$$

$P_{\mathcal{E}_0} y$ has components $y_i - \bar{y}$ and $P_{\mathcal{E}-\mathcal{E}_0} y$ has components $\hat{\beta}_1 (x_i - \bar{x})$. Continuing,

$$\begin{aligned} \| Q_{\mathcal{E}} y \|^2 &= \sum (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum (x_i - \bar{x}))^2 \\ &= \| Q_{\mathcal{E}} y \|^2 - \| P_{\mathcal{E}-\mathcal{E}_0} y \|^2 \\ &= \| y \|^2 - \| P_{\mathcal{E}_0} y \|^2 - \| P_{\mathcal{E}-\mathcal{E}_0} y \|^2 \end{aligned}$$

It is easy to see that $\hat{\beta}_1^2 \| x - \bar{x}J_n \|^2 = \| P_{\mathcal{E}-\mathcal{E}_0} y \|^2$, so the F-statistic is just

$$F = \frac{\hat{\beta}_1^2 \| x - \bar{x}J_n \|^2}{\hat{\sigma}^2} = \frac{\hat{\beta}_1^2}{\widehat{\text{var}}(\hat{\beta}_1)} = t^2$$

and $F = t^2 \sim F(1, n-2, \delta^2)$ with

$$\delta^2 = \frac{\| P_{\mathcal{E}-\mathcal{E}_0} y \|^2}{\sigma^2} = \frac{\beta_1^2 \| X - \bar{x}J_n \|^2}{\sigma^2}$$

The power depends on δ^2, and hence on the size of the slope, in units of σ, and on the dispersion of the x_is. To increase power, the only factor under the control of the experimenter is the $\| X - \bar{x}J_n \|^2$, which should be made as large as possible. Suppose each $x_i \in [-1, 1]$; else, take $x_i = \infty$ or $-\infty$. First, argue that $\bar{x} = 0$, by symmetry, since the power of the test does not depend on sign. Then: take $n/2$

Table 6.4: Analysis of Variance for Simple linear regression

Source	*df*	*SS*	*MS*	F	*E(MS)*
$\mathcal{R}(J_n)$	1	$n\bar{y}^2$			$\sigma^2 + n(\beta_0 + \beta_1\bar{x})^2$
$\mathcal{R}(X - \bar{x}1)$	1	$\hat{\beta}_1^2 \parallel x - \bar{x}1 \parallel^2$			$\sigma^2 + \beta_1^2 \parallel x - \bar{x}1 \parallel^2$
Error $= \mathcal{E}^\perp$	$n-2$	$\parallel Q_{\mathcal{E}} y \parallel^2$			σ^2

observations at -1 and $n/2$ observations at $+1$ (show that any other arrangement has smaller value for the norm $\parallel x - \bar{x} \parallel^2$.

The analysis of variance table for simple regression is given in Table 6.11. For β_0, the EMS is obtained from

$$\sigma^2 + \parallel P_{\mathcal{R}(J_n)} y \parallel^2 = \sigma^2 + \parallel \mu 1 \parallel^2 = \sigma^2 + n\mu^2$$

All the subspaces in the anova are orthogonal so we get a decomposition of the sum of squares.

Next, consider the test of:

NH: $\beta_0 = 0$
AH: $\beta_0 \neq 0$

Why is the F-test NOT equal to $n\bar{y}^2/\hat{\sigma}^2$? Order matters! Here, $\mathcal{E} = \mathcal{R}(J_n, X)$ as before, but $\mathcal{E}_0 = \mathcal{R}(X)$ and $\mathcal{E} - \mathcal{E}_0 = \mathcal{R}(Q_{\mathcal{E}_0} J_n) = \mathcal{R}(I - \frac{XX'}{X'X} J_n)$, and

$$P_{\mathcal{E}-\mathcal{E}_0} y = \frac{(I - \frac{XX'}{X'X}) J_n J_n{}' (I - \frac{XX'}{X'X})}{J_n{}'(I - \frac{XX'}{X'X}) J_n} Y$$

and

$$\begin{aligned}
\parallel P_{\mathcal{E}-\mathcal{E}_0} y \parallel^2 &= \frac{(\sum y_i - \sum y_i x_i \sum x_i / \sum x_i^2)^2}{n - \frac{(\sum x_i)^2}{\sum x_i^2}} \\
&= \frac{(\bar{y} - \tilde{\beta}_1 \bar{x})^2}{1/n - \bar{x}^2 / \sum x_i^2} \\
&= \frac{n \sum x_i^2 (\bar{y} - \tilde{\beta}_1 \bar{x})^2}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

where $\tilde{\beta}_1$ is estimated under NH that $\beta_0 = 0$, $\tilde{\beta}_1 = \sum x_i y_i / \sum x_i^2$.

Since $\sum x_i \mu_i / \sum x_i^2 = \sum x_i(\beta_0 + \beta_1 x_i)/\sum x_i^2 = n\beta_0 \bar{x}/\sum x_i^2 + \beta_1$, the non-centrality parameter for this F- test is (after some algebra):

$$\delta^2 = \frac{\| P_{\mathcal{E}-\mathcal{E}_0}\mu \|^2}{\sigma^2} = n\beta_0^2 \frac{\sum(x_i - \bar{x})^2}{\sum x_i^2}$$

Now $\sum(x_i - \bar{x})^2 / \sum x_i^2 \leq 1$, and it equals 1 if $\bar{x} = 0$. Thus, the design that maximizes the power of this test is: any $\{x_1, \ldots, x_n\}$ with $\bar{x} = 0$.

6.12 One Way layout

The model can be written as $y_{ij} = \beta_i + \varepsilon_{ij}, i = 1, \ldots, b;\ j = 1, \ldots, n_i;\ \sum n_i = n;\ \varepsilon \sim N(0, \sigma^2 I)$. Then, as usual, $\mathcal{E} = \mathcal{R}(X_1, \ldots, X_p)$, where X_i is a vector of zeroes, except for the rows from group i where it is one (or any common nonzero number). As shown previously, $\dim(\mathcal{E}) = p;\ P_{\mathcal{E}}y = (\bar{y}_{i+})$; $Q_{\mathcal{E}}y = (y_{ij} - \bar{y}_{i+})$; $\hat{\sigma}^2 = \| Q_{\mathcal{E}}y \|^2/(n-p) = \sum\sum(y_{ij} - \bar{y}_{i+})^2/(n-p)$. Consider some tests of hypotheses.

6.12.1 Overall test

This hypothesis can be stated as

NH: $\beta_i = \beta,\ i = 1, 2, \ldots, p$
AH: β_i not all equal

Again all the computations are easy. Under NH:

$$P_{\mathcal{E}_0}y = \bar{y}_{++}J_n = \frac{\sum n_i \bar{y}_{i+}}{n} J_n$$

$$\| P_{\mathcal{E}_0}y \|^2 = n\bar{y}_{++}^2$$

Projections and lengths on $\mathcal{E} - \mathcal{E}_0$ are easily then obtained by subtraction:

$$P_{\mathcal{E}-\mathcal{E}_0}y = P_{\mathcal{E}}y - P_{\mathcal{E}_0}y = \bar{y}_{i+} - \bar{y}_{++}$$

$$\begin{aligned} \| P_{\mathcal{E}-\mathcal{E}_0}y \|^2 &= \| P_{\mathcal{E}}y \|^2 - \| P_{\mathcal{E}_0}y \|^2 \\ &= \sum_{i=1}^{p} n_i(\bar{y}_{i+} - \bar{y}_{++})^2 \end{aligned}$$

Table 6.5: Analysis of Variance for the one-way design.

Source	*df*	*SS*	*MS*	F	*E(MS)*
Mean = $\mathcal{E}_0 = \mathcal{R}(1)$	1	$n\bar{y}_{++}^2$			$\sigma^2 + n\bar{\beta}^2$
Groups = $\mathcal{E} - \mathcal{E}_0$	$p-1$	$\sum n_i(\bar{y}_{i+} - \bar{y}_{++})^2$			$\sigma^2 + \sigma^2\delta^2/(p-1)$
Error = $\mathcal{E}^\perp$	$n-p$	$\parallel Q_{\mathcal{E}} y \parallel^2$			σ^2

and the last sum of squares has $p-1$ d.f. The F-test is then given by

$$F = \frac{\parallel P_{\mathcal{E}-\mathcal{E}_0} y \parallel^2}{(p-1)\hat{\sigma}^2} \sim F(p-1, n-p, \delta^2)$$

and the non-centrality parameter is given by

$$\delta^2 = \parallel P_{\mathcal{E}-\mathcal{E}_0}\mu \parallel^2/\sigma^2 = \sum_{i=1}^{p} n_i(\beta_i - \bar{\beta}_+)^2)/\sigma^2.$$

These results can be summarized in the analysis of variance table given in Table 6.5.

6.12.2 Orthogonal Contrasts

In most linear models, we will want to make multiple inferences concerning ${c_j}'\mu, j = 1, \ldots, m$. One setting for this uses orthogonal contrasts. Consider the coordinate free linear model $y \sim N(\mu, \sigma^2 I)$.

Definition 6.1 (Contrast) *For $c \in \Re^n$, $c'\mu$ is a contrast if $c \perp J_n$; that is $(c, J_n) = 0$.*

Several comments are in order here. First, contrasts depend on the inner product. In the usual inner product (which is the one generally used when the covariance matrix is proportional to the identity), $(c, J_n) = c'J_n$, so a contrast is any linear combination of the elements of μ with sum of the multipliers equal to zero. When any other inner product is used, then the notion of orthogonality changes with the inner product. For example, if $y \sim \mathrm{N}(0, \Sigma)$, the natural inner product is $(a, y) = a'\Sigma^{-1}y$, so the definition requires that $a'\Sigma^{-1}J_n = 0$. Second, all contrasts are estimable, since they are just linear combinations of the elements of μ.

Finally, this definition of a contrast differs from the usual definition, which is typically defined as a linear combination of the elements of a parameter vector. Using this definition avoids any problems caused by changes in parameterization.

Definition 6.2 (Orthogonal contrasts) *Let $c_j, j = 1, \ldots, m$ be m vectors such that $(c_j, c_k) = 0, j \neq k$ and $(c_j, J_n) = 0, j = 1, \ldots, m$. Then $(c_j, \mu), j = 1, \ldots, m$, are a set of m orthogonal contrasts.*

Of course, orthogonality depends on the inner product.

Consider testing:

NH: $(c_j, \mu) = 0$
AH: $(c_j, \mu) \neq 0$

for $j = 1, \ldots, m$. This corresponds to conducting m separate tests and is clearly not the same as testing the hypothesis $C\mu = 0$, where C is $m \times n$ with rows $c_j{}'$.

Theorem 6.2 *If $c_j{}'\mu$ is a contrast for $\mathcal{E}$, then the* BLUE *of $c_j{}'\mu$ is $c_j{}'\hat{\mu}$.*

Proof. The BLUE of (c, μ) is $(Pc, y) = (c, \hat{\mu})$.

Let $c_{m+1}, \ldots, c_p$ be any completion of the c- basis for $\mathcal{E}$. Clearly, we can write:

$$\Re^n = \mathcal{R}(1) + \sum_{j=1}^{m} \mathcal{R}(c_j) + \mathcal{R}(c_{m+1}, \ldots, c_p) + \mathcal{E}^{\perp}$$

and, in terms of lengths,

$$\| \, y \, \|^2 = \| \, P_1 y \, \|^2 + \sum_{j=1}^{m} \| \, P_{c_j} y \, \|^2 + \| \, P_{[c_{m+1},\ldots c_p]} y \, \|^2 + (n-p)\hat{\sigma}^2$$

Now we know that $P_{c_j} y = [(c_j, y)/(c_j, c_j)]c_j$ so that $\| \, P_{c_j} y \, \|^2 = (c_j, y)^2/(c_j, c_j)$ is trivial to compute. The test statistic for testing NH: $c_j{}'\mu = 0$ is just:

$$F = \frac{\| \, P_{c_j} y \, \|^2}{\hat{\sigma}^2} \sim F(1, n-p, \delta^2)$$

with $\delta^2 = \| \, P_{c_j} \mu \, \|^2 / \sigma^2$. Since the c_j are mutually orthogonal, they must be linearly independent, and then the statistic for simultaneously testing the NH that all the $c_j{}'\mu = 0, j = 1, \ldots, m$ will then be given by

$$F = \frac{\sum_{i=1}^{m} \| \, P_{c_j} y \, \|^2}{m\hat{\sigma}^2} \sim F(1, n-p, \delta^2)$$

with $\delta^2 = \sum_{i=1}^m \parallel P_{c_j}\mu \parallel^2/m\sigma^2$.

We now turn to the parametric version. In the full rank parametric case, still with Var$(y) = \sigma^2 I$, $\mu = X\beta$, a contrast in the βs is a linear combination $a'\beta$ such that

$$a'\beta = a'(X'X)^{-1}X'\mu$$

so that if $c = X(X'X)^{-1}a$, then $a'\beta$ is the same as $c'\mu$. Since c is just a linear combination of the columns of $X, c \in \mathcal{E}$.

Look at the conditions of orthogonality to J_n and mutual orthogonality. We need impose $c'J_n = 0$ to get a contrast, equivalent to $a'(X'X)^{-1}X'J_n = 0$. This may be a bit unexpected, since one might expect the conditions $a'J_n = 0$.

If ${c_i}'c_j = 0$, then

$${a_i}'(X'X)^{-1}X'X(X'X)^{-1}a_j = {a_i}'(X'X)^{-1}a_j = 0$$

Thus, with respect to the inner product $[a, b] = a'(X'X)^{-1}b$, we have the conditions $[a_i, a_j] = 0$ and $[a_j, X'J_n] = 0$, for all i and j.

Let's look at one-way anova with n_i observations per group, $\mathcal{E}(y_{ij}) = \beta_i$. Then compute

$$c = X(X'X)^{-1}a = (X_1, \ldots, X_p)\text{diag}(n_j^{-1})a = \begin{pmatrix} J_{n_1}a_1/n_1 \\ \vdots \\ J_{n_p}a_p/n_p \end{pmatrix}$$

Two contrasts are orthogonal if ${c_i}'c_j = 0 = {a_i}'\text{diag}(n_j^{-1})a_j$. Also, ${J_n}'c = 0$ if and only if ${J_n}'X(X'X)^{-1}a = 0$, if and only if $(n_1, \ldots, n_p)\text{diag}(n_j^{-1})a = 0$, or if and only if ${J_n}'a = 0$.

Here is a little numerical example, $p = 3$. Suppose that the first contrast is ${c_1}' = (-1, 0, 1)$, and the second contrast is ${c_2}' = (a_1, a_2, a_3)$ For the second orthogonal contrast, we must have $a_1+a_2+a_3 = 0$ and also $-a_1/n1+a_3/n_3 = 0$. There are of course lots of solutions to these equations, and they depend on n_1 and n_3. One choice is $a_1 = n_1/(n_1 + n_3)$, $a_2 = 1$ and $a_3 = n_3/(n_1 + n_3)$. No one would typically be interested in such a contrast, since it depends on sample size as well as on parameters. Consequently, one would usually not use orthogonal contrasts unless $X'X = kI$.

Example 2×2 tables. Suppose $y_{ijk} = \mu_{ij} + \varepsilon_{ijk},$ Var$(\varepsilon) = \sigma^2 I$. The usual model is:

$$\begin{aligned} y_{ijk} &= \mu_{ij} + \varepsilon_{ijk} \\ &= \mu + (\mu_{i+} - \mu) + (\mu{+}j - \mu) + (\mu_{ij} - \mu_{i+} - \mu{+}j + \mu) + \varepsilon_{ijk} \\ &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \end{aligned}$$

where:

$$\begin{aligned} \mu &= \bar{\mu}_{++} \\ \alpha_i &= (\bar{\mu}_{i+} - \bar{\mu}_{++}) \\ \beta_j &= (\bar{\mu}_{+j} - \bar{\mu}_{++}) \\ (\alpha\beta)_{ij} &= \mu_{ij} - \bar{\mu}_{i+} - \bar{\mu}_{+j} + \bar{(}\mu)_{++} \end{aligned}$$

Now, the μ_{ij} are the coordinates (parameters) relative to the implicit columns of X (with a separate column for each cell). Are the contrasts that correspond to the "usual" parameters orthogonal? Suppose the sample size in cell (i, j) is n_{ij}. The usual contrasts are given by:

Cell:	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
$a(\alpha_1)$	1	1	-1	-1
$a(\beta_1)$	1	-1	1	-1
$a(\alpha\beta)_{11}$	1	-1	-1	1

Are these contrasts orthogonal? For example, is $a(\alpha_1)'(X'X)^{-1}a(\beta_1) = 0$? Since $(X'X)^{-1} = \text{diag}(1/n_{11}, \ldots, 1/n_{22})$,

$$a(\alpha_1)'(X'X)^{-1}a(\beta_1) = 1/n_{11} - 1/n_{22} - 1/n_{21} + 1/n_{22}$$

which can of course be zero, even if all the n_{ij} are not equal. But, one can easily show that the 3 contrasts above (all orthogonal to the overall mean vector) are all orthogonal if and only if $n_{ij} = n, i = 1, 2, j = 1, 2$.

$\Sigma \neq \sigma^2 I$. Let's look at contrasts for the general linear model $y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \Sigma)$, with Σ positive definite. By the usual method, this is equivalent to $\Sigma^{-1/2}2y = Z \sim N(W\beta, \sigma^2 I)$, so we can proceed as before, using the Zs and Ws. A test of NH: $B\mu = 0$ is equivalent to a test of NH: $B\Sigma^{-1/2}\mu = 0$. Orthogonality changes in the same way: one must account for the inner product and the criterion for a_i and a_j to be orthogonal is ${a_i}'(X'\Sigma^{-1}X)^{-1}a_j = 0$.

6.13 Confidence Regions

Let $\{y_1, \ldots, y_n\}$ be independent identically distributed with distribution $F(\theta)$, with $\theta \in S_\theta \subset \Re^k$. The set S_θ need not be a subspace. A *confidence region* $C_\alpha(\{y_1, \ldots, y_n\})$ for θ is a map, $\{y_1, \ldots, y_n\} \to S_\theta$ with the property that, for all $\theta \in S_\theta$,

$$\Pr_\theta(\theta \in C_\alpha(\{y_1, \ldots, y_n\})) \geq 1 - \alpha \tag{6.9}$$

C_α is called *conservative* if inequality holds in (6.9) for some θ; otherwise, C_α is called exact.

Example. Suppose $\{y_1, \ldots, y_n\} \sim \mathrm{N}(\mu, \sigma^2)$, with σ^2 known, and $\mu \in \Re$. Then

$$z = \sqrt{n}\left(\frac{\bar{y} - \mu}{\sigma}\right) \sim N(0, 1)$$

and, for all $\mu,\ \sigma^2$,

$$\Pr(-z_{\alpha/2} \leq z \leq z_{\alpha/2}) = 1 - \alpha$$

and thus, for all μ,

$$\Pr(\bar{y} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \bar{y} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}) = 1 - \alpha$$

The random interval given by $\bar{y} \pm \sigma z_{\alpha/2}/\sqrt{n}$ will contain μ with probability $1-\alpha$, and is therefore a $1 - \alpha \times 100\%$ confidence region for μ.

The previous example is standard and it serves to illustrate the basic idea of a confidence region, but is provides little help as a general method. A usual method for constructing confidence regions is to invert test procedures:

$$C_\alpha = \{\theta_0 | \text{NH} : \theta = \theta_0 \text{ is not rejected at level } \alpha\}$$

When likelihood ratio tests are used, such regions will have useful properties, such as being based on minimal sufficient statistics.

It follows that:

$$\Pr(\theta \in C_\alpha) = \Pr(\text{NH} : \theta = \theta_0 \text{ is not rejected at level } \alpha) = 1 - \alpha$$

Example. Suppose that $\{y_1, \ldots, y_n\} \sim \mathrm{N}_2(\mu, \sigma^2 I_2)$. The likelihood ratio test for NH : $\mu = \mu_0$ is

$$\begin{aligned} \Lambda &= \frac{\exp\{-\frac{1}{2}\sum(y_i - \mu_0)'(y_i - \mu_0)\}}{\exp\{-\frac{1}{2}\sum(y_i - \bar{y})'(y_i - \bar{y})\}} \\ &\propto n \| \bar{y} J_n - \mu_0 \|^2 \end{aligned}$$

Under the null hypothesis, $\bar{y} - \mu_0 \sim N_2(0, (\sigma^2/n) I_2)$, so that $n \| \bar{y} - \mu_0 \|^2/\sigma^2 \sim \chi^2(2)$.

To turn this into a confidence statement, we will not reject the null hypothesis if

$$\| \bar{y} - \mu_0 \|^2 \leq \chi^2(\alpha, 2) \times \frac{\sigma^2}{n}$$

As a function of μ_0, the confidence region is a circle of radius $(\chi^2(\alpha, 2)\sigma^2/n)^{1/2}$.

Confidence region for any estimable function of β. Suppose we have a parametric linear model $y \sim \mathrm{N}(X\beta, \sigma^2 I)$, where X is an $n \times k$ matrix of rank p. Suppose that $\psi = A\beta = BX\beta$ is any estimable function of β (the second equality follows from estimability), with rank $A = q \leq p$. Then the hypothesis test

NH: $\psi = \psi_0$
AH: $\psi \neq \psi_0$

is rejected at level α if

$$\begin{aligned} f &= \frac{1}{p-q}(\hat{\psi} - \psi_0)' \widehat{\mathrm{Var}(\hat{\psi})}(\hat{\psi} - \psi_0) \\ &= \frac{(\hat{\psi} - \psi_0)'(A(X'X)^+A')^{-1}(\hat{\psi} - \psi_0)}{(p-q)\hat{\sigma}^2} \end{aligned}$$

This is not the same formula for the F-test we have seen before, although it is equivalent. If $F(1-\alpha, p-q, n-p)$ is the $1-\alpha$ percentile of the $F(p-q, n-p)$ distribution, then the confidence region for ψ is

$$\left\{\psi_0 | (\hat{\psi} - \psi_0)'(A(X'X)^+A')^{-1}(\hat{\psi} - \psi_0) \leq (p-q)\hat{\sigma}^2 F(1-\alpha, p-q, n-p)\right\} \tag{6.10}$$

This is an ellipsoid centered at $\hat{\psi} = A\hat{\beta}$, with contours determined by the eigenstructure of $(A(X'X)^+A')^{-1}$.

For the full-rank parameterization case, $k = p$, here are some special cases:

- Confidence region for β:

$$\left\{\beta_0 | (\hat{\beta} - \beta_0)'(X'X)^{-1}(\hat{\beta} - \beta_0) \leq p\hat{\sigma}^2 F(1-\alpha, p-q, n-p)\right\}$$

- If $X = (J_n, X_1)$, Then $\psi = (0, I_{p-1})\beta$ picks out all the coefficients except for the intercept. Substituting into (6.10) gives the confidence region. f $\mathcal{X}_1 = (I - P_{J_n})X_1$, then the confidence region is

$$\left\{\psi_0 | (\hat{\psi} - \psi_0)'(\mathcal{X}_1'\mathcal{X}_1)^{-1}(\hat{\psi} - \psi_0) \leq (p-1)\hat{\sigma}^2 F(1-\alpha, p-q, n-p)\right\}$$

The matrix $\mathcal{X}_1$ is like X_1, except that column means have been subtracted off.

- If $\psi = A\beta$ picks out the last q components of β, then the confidence region is given by

$$\left\{\psi_0 | (\hat{\psi} - \psi_0)' C^{-1} (\hat{\psi} - \psi_0) \leq (p-q)\hat{\sigma}^2 F(1-\alpha, p-q, n-p)\right\}$$

where C is the lower-right $q \times q$ submatrix of $X'X)^{-1}$.

The R package `car`, written by John Fox, contains functions for computing confidence regions for pairs of regression coefficients. Similar routines, for pairs and for triples, are available in `Arc`.