3.7 Reparameterization

In the previous section, we began with an already partitioned design matrix with certain orthogonality properties.

- Typically this is not the case.
- However, by reparameterization we can resolve this problem.

Definition: The linear model $\mathbf{X}\boldsymbol{\beta}$ is said to be reparameterized if

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\alpha} \text{ where } \mathbf{Sp}(\mathbf{X}) = \mathbf{Sp}(\mathbf{Z})$$

- There are certain quantities that are invariant with respect to reparameterization.
- These invariant quantities are functions of vectors such as inner products and lengths which do not depend on which particular basis is chosen with respect to which vectors are represented.

Lemma 3.5: Let $\mathbf{X}\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\alpha}$ be a reparameterization. Then

(1) If $\mathbf{L}^T\boldsymbol{\beta}$ is estimable then there exists an $\mathbf{R}^T$ such that $\mathbf{R}^T\boldsymbol{\alpha}$ is estimable and $\mathbf{L}^T\boldsymbol{\beta} = \mathbf{R}^T\boldsymbol{\alpha}$ Conversely, if $\mathbf{R}\boldsymbol{\alpha}$ is estimable, then there exists an $\mathbf{L}$ such that $\mathbf{L}^T\boldsymbol{\beta}$ is estimable and $\mathbf{R}^T\boldsymbol{\alpha} = \mathbf{L}^T\boldsymbol{\beta}$

(2) If $\mathbf{L}^T\boldsymbol{\beta} = \mathbf{R}^T\boldsymbol{\alpha}$ and they are estimable, then the LSE of $\mathbf{L}^T\boldsymbol{\beta}$ equals the LSE of $\mathbf{R}^T\boldsymbol{\alpha}$

(3) $\mathbf{Y}^T(\mathbf{I} - \mathbf{P_X})\mathbf{Y} = \mathbf{Y}^T(\mathbf{I} - \mathbf{P_Z})\mathbf{Y}$ where $\mathbf{P_X}, \mathbf{P_Z}$ are the projection operators onto $\mathbf{Sp}(\mathbf{X})$ and $\mathbf{Sp}(\mathbf{Z})$ respectively.

We now consider some reparameterizations to induce orthogonality into the design matrix.

As a first example, we consider a design matrix $\mathbf{X}$ that is already partitioned but that the partition is not orthogonal.

- Thus
$$\mathbf{Y} \sim \mathrm{WS}(\beta_0\mathbf{1} + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \sigma^2\mathbf{I})$$
but $\mathbf{X}_1^T\mathbf{D}_1\mathbf{X}_2 \neq \mathbf{0}$.
- The parameter of primary interest is $\boldsymbol{\beta}_1$ and we wish to use the results from the previous section to simplify the models considered in a data analysis.
- Consider the reparametrization in which:
$$\begin{aligned}\mathbf{D}_1\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{D}_1\mathbf{X}_2\boldsymbol{\beta}_2 &= [\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_2}]\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{D}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{P}_{\mathbf{D}_1\mathbf{X}_2}\mathbf{X}_1\boldsymbol{\beta}_1 \\ &= \mathbf{Z}_1\boldsymbol{\alpha}_1 + \mathbf{Z}_2\boldsymbol{\alpha}_2\end{aligned}$$
where
$$\begin{aligned}\mathbf{Z}_1 &= [\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_2}]\mathbf{X}_1 \\ \mathbf{Z}_2 &= \mathbf{D}_1\mathbf{X}_2 \\ \boldsymbol{\alpha}_1 &= \boldsymbol{\beta}_1 \\ \boldsymbol{\alpha}_2 &= \boldsymbol{\beta}_2 + (\mathbf{X}_2^T\mathbf{D}_1\mathbf{X}_2)^-\mathbf{X}_2^T\mathbf{D}_1\mathbf{X}_1\boldsymbol{\beta}_1\end{aligned}$$

- Since $\boldsymbol{\alpha}_1 = \boldsymbol{\beta}_1$, and we are interested only in making inferences about $\boldsymbol{\alpha}_1$ and since $\mathbf{Z}_1 \perp \mathbf{Z}_2$, we may conduct the data analysis as if

$$(\mathbf{D}_1 - \mathbf{P}_{\mathbf{D}_1\mathbf{X}_2})\mathbf{Y} \sim \mathrm{WS}(\mathbf{Z}_1\boldsymbol{\alpha}_1, \sigma^2\mathbf{I})$$

  except of course for the adjustment to the total degrees of freedom.

- Note that this is equivalent to regressing $\mathbf{Y}$ on $\mathbf{X}_2$ and using the residuals as a new reponse vector, regressing the columns of $\mathbf{X}_1$ on $\mathbf{X}_2$ and using these residuals as new covariates.

- One situation where this technique is very useful is in analyzing data from a repeated measures type experiment where the model includes a separate parameter for each individual.

  - The analysis would be simplified (computationally) if the n parameters denoting individual effects could be eliminated.
  - Applying the above results, this can be achieved by centering the responses for each individual and the independent variables for each individual about the individual's mean for these variables.

- Very often, data on the incidence of disease which are already age and gender adjusted, say, are regressed on some measure of exposure which is not adjusted.
  - In the above notation, this corresponds to regressing $(\mathbf{I}-\mathbf{P}_Z)\mathbf{Y}$ onto $\mathbf{X}$ rather than onto $(\mathbf{I}-\mathbf{P_Z})\mathbf{X}=\widetilde{\mathbf{X}}$ when the true model is

    $$E(\mathbf{Y})=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\alpha}$$

    where $\mathbf{Y}$ is disease incidence, $\mathbf{Z}$ is age and gender, and $\mathbf{X}$ is exposure.
  - In this situation, an estimate of $\boldsymbol{\beta}$ is obtained with expectation equal to (assuming $\mathbf{X},\mathbf{Z}$ full rank)

    $$\begin{aligned} E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\widetilde{\mathbf{Y}}] &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}-\mathbf{P_Z})\mathbf{X}\boldsymbol{\beta} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}-\mathbf{P_Z}\mathbf{X})\boldsymbol{\beta} \end{aligned}$$

  - If age and gender are uncorrelated with exposure, then $\mathbf{P_Z}\mathbf{X}=\mathbf{0}$ and the estimate is unbiased.
  - If however, exposure varies by age and gender, the estimator will generally be biased.
  - The stronger the association between age and gender and exposure, the closer to $\mathbf{X}$ will be the projection of $\mathbf{X}$ onto $\mathbf{Sp}(\mathbf{Z})$.
  - In the most extreme case where exposure is perfectly predicted by age and gender, $\mathbf{P_Z}\mathbf{X}=\mathbf{X}$ and $E[\mathbf{b}]=\mathbf{0}$ regardless of the magnitude of $\boldsymbol{\beta}$.

As another example of the use of reparametrization, suppose we are interested in testing the hypothesis $\mathbf{L}^T\boldsymbol{\beta} = \mathbf{0}$ for some estimable function $\mathbf{L}^T\boldsymbol{\beta} = \mathbf{0}$.

- Using the characterization of estimable functions requiring the existence of a $\mathbf{C}$ such that $\mathbf{C}^T\mathbf{X} = \mathbf{L}^T$, we see that the hypothesis of $\mathbf{L}^T\boldsymbol{\beta} = \mathbf{0}$ is equivalent to the hypothesis that the mean vector lies in a subspace of $\mathbf{Sp}(\mathbf{X})$.
- More specifically,

$$E[\mathbf{Y}] \subset \mathbf{Sp}(\mathbf{X}) \cap [\mathbf{Sp}(\mathbf{C})]^{\perp}$$

  - Let $\mathbf{Z}_1$ be a matrix with column vectors being a ONB of the subspace $\mathbf{Sp}(\mathbf{X}) \cap [\mathbf{Sp}(\mathbf{C})]^{\perp}$ and let $\mathbf{Z}_2$ be a matrix with column vectors being an ONB of the subspace $\left\{\mathbf{Sp}(\mathbf{X}) \cap [\mathbf{Sp}(\mathbf{C})]^{\perp}\right\}^{\perp}$.
  - Then consider the reparameterization

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{Z}_1\boldsymbol{\gamma}_1 + \mathbf{Z}\boldsymbol{\gamma}_2$$

  - Note that the null hypothesis corresponds to the hypothesis $\mathbf{Z}\boldsymbol{\gamma}_2 = \mathbf{0}$ and that $\mathbf{Z}_1 \perp \mathbf{Z}_2$.

- This reparameterization now allows use of the arguments in the previous section to show that the difference in residual sum of squares from the full and reduced models is simply $\|\mathbf{P}_{\mathbf{Z}_2}\mathbf{Y}\|^2$.
  - If we let $\mathbf{Z}_3$ be a matrix with column vectors being an ONB for $\mathbf{Sp}^{\perp}(\mathbf{X})$, and consider the transformation $\widetilde{\mathbf{Y}} = (\mathbf{Z}_1\ \mathbf{Z}_2\ \mathbf{Z}_3)\mathbf{Y}$, we note that

$$E[\widetilde{\mathbf{Y}}] = \mathbf{Z}_1\boldsymbol{\gamma}_1 + \mathbf{Z}_2\boldsymbol{\gamma}_2 \ \text{ and } \ \text{var}\,\widetilde{\mathbf{Y}} = \sigma^2\mathbf{I}$$

  - Also, we have $\|\mathbf{P}_{\mathbf{Z}_2}\mathbf{Y}\|^2 = \sum_{i=q+1}^{r} Y_i^2$ where the $q+1, q+2, \ldots, r$ columns of $(\mathbf{Z}_1\ \mathbf{Z}_2\ \mathbf{Z}_3)$ correspond to the submatrix $\mathbf{Z}_2$.
  - Thus $\|\mathbf{P}_{\mathbf{Z}_2}\mathbf{Y}\|^2$ has the representation as the sum of squares of uncorrelated random variables with variance σ^2 and mean zero under the null hypothesis.
  - This representation will prove useful in later sections.

3.8 Inference under Wide Sense Assumptions

Under wide sense assumptions, we now have results about what we can estimate, how we may estimate it and how good the estimators are. Now we discuss actually making inferences, i.e. hypothesis tests, confidence intervals, etc.

- For rank $(\mathbf{X}) = p$, we know the LSE $\mathbf{b}$ has the properties:

$$\mathbf{b} \sim \text{WS}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \text{ so that } b_i \sim \text{WS}(\beta_i, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{ii})$$

  and we can estimate var (b_i) by $s^2(\mathbf{X}^T\mathbf{X})^{-1}_{ii}$).

- To construct a confidence interval for β_i from b_i, we need the entire distribution of b_i (or at least a reasonable approximation to it).

- This approximation is available in large samples via the following result.

Theorem 3.6: In the general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim \text{WS}(\mathbf{0}, \sigma^2\mathbf{I})$$

if

- $\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ are independent, identically distributed with $\epsilon_i \sim \text{WS}(0, \sigma^2)$
- $\lim_{n\to\infty} (\max_{1\leq i\leq n}[\mathbf{P_X}]_{ii}) = 0$

Then for any estimable function $\boldsymbol{\ell}^T\boldsymbol{\beta}$, we have

$$\frac{\boldsymbol{\ell}^T\mathbf{b} - \boldsymbol{\ell}^T\boldsymbol{\beta}}{\sqrt{\sigma^2\boldsymbol{\ell}^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\ell}}} \xrightarrow{\text{d}} \text{N}(0,1)$$

The essence of the proof hinges on writing

$$\frac{\boldsymbol{\ell}^T\mathbf{b} - \boldsymbol{\ell}^T\boldsymbol{\beta}}{\sqrt{\sigma^2\boldsymbol{\ell}^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\ell}}}$$

as Z_n where

$$E(Z_n) = 0 \quad \text{and} \quad \text{var}\,(Z_n) = 1$$

and Z_n is a sum of independent random variables.

- It seems reasonable that Z_n should converge in distribution to a $N(0,1)$.
- However the usual Central Limit Theorem is not applicable since the ϵ_i's are not identically distributed.
- The Lindeberg-Feller Theorem is, however, a central limit theorem that is applicable in this situation:

  Lindeberg-Feller Theorem: Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables with distribution functions $F_1, F_2, \ldots, F_n$, let $E(Z_i) = 0$, var $(Z_i) = \sigma_i^2$ and set $S_n^2 = \sum_{i=1}^n \sigma_i^2$. If for each $t > 0$

  $$\lim_{n \to \infty} \left\{ \frac{1}{S_n^2} \sum_{i=1}^{n} \int_{E_i} y^2 f_i(y) dy \right\} = 0$$

  where $E_i = \{y : \ |y| > t\}$ then

  $$\frac{Z_1 + Z_2 + \cdots + Z_n}{S_n}$$

  converges in distribution to a $N(0,1)$ distribution.
- It can be shown that the Lindeberg-Feller Theorem is applicable. Thus the result follows.

- To provide some motivation for the requirement of $\lim[\mathbf{P_X}]_{ii} = 0$ recall that central limit theorems hinge on the distribution of the sum of many independent quantities, each of which has minimal influence on distribution of the total sum.
- Also, recall that observations that are associated with covariates far away from the others in the design space are typically very influential.
- Thus some condition preventing the covariates for any individual from being arbitrarily far from the independent variables for others must be part of a central limit theorem for estimators of regression parameters.
- This is exactly what the condition on $[\mathbf{P_X}]_{ii}$ is.
- To see this, we note, after some algebraic manipulation, that

$$[\mathbf{P_X}]_{ii} = \frac{1}{n} + \frac{1}{n}(\mathbf{X}_i - \bar{\mathbf{X}})^T[\text{var }(\mathbf{X})]^{-1}(\mathbf{X}_i - \bar{\mathbf{X}})$$

  - where $\mathbf{X}_i$ refers to the ith row of the design matrix (excluding the element from the first column corresponding the intercept term in the model.
  - $\bar{\mathbf{X}}$ refers to the mean vector of $\mathbf{X}$'s over the sample
  - var $(\mathbf{X})$ refers to the sample covariance matrix of the $\mathbf{X}$'s.
- Thus $\max(1 \leq i \leq n)[\mathbf{P_X}]_{ii}$ converging to 0 implies that no individual can have an $\mathbf{x}$ that, after standardization by var $(\mathbf{X})$, is too far removed from the others in the sample.

3.9 Multivariate Normal Distribution Theory

In this section we investigate the distribution theory of estimators associated with general linear models.

3.9.1 Distribution Theory of LSE under MVN

Let $\mathbf{Y} \sim \text{MVN}\,(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Then for any linearly estimable function $\mathbf{L}^T\boldsymbol{\beta}$, we know the unique LSE is $\mathbf{L}^T\mathbf{b}$, and

$$E(\mathbf{L}^T\mathbf{b} = \mathbf{L}^T\boldsymbol{\beta} \quad \text{and} \quad \text{var}\,(\mathbf{L}^T\mathbf{b} = \sigma^2\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L}$$

- Note that
$$\mathbf{L}^T\mathbf{b} = \mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{Y}$$
which is a linear combination of MVN random variables.
- Thus
$$\mathbf{L}^T\mathbf{b} \sim \text{MVN}\,(\mathbf{L}^T\boldsymbol{\beta}, \sigma^2\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L}$$
- In a previous section, we decomposed $\mathbf{Y}$ into two vectors, $\mathbf{P_X Y}$, the LSE of $E(\mathbf{Y})$ and $(\mathbf{I} - \mathbf{P_X})\mathbf{Y}$, the residual vector.
    - Since these are linear combinations of $\mathbf{Y}$, they are MVN and, in addition, are independent because
    $$\text{cov}\,(\mathbf{P_X Y}, (\mathbf{I} - \mathbf{P_X})\mathbf{Y}) = \sigma^2\mathbf{P_X}(\mathbf{I} - \mathbf{P_X}) = \mathbf{0}$$
    - Thus, the LSE of any estimable function, since it can be written as a function of $\mathbf{P_X Y}$, is independent of any function of the residual vector.

- More specifically $\mathbf{P_X Y}$ and $\mathbf{Y}^T(\mathbf{I} - \mathbf{P_X})\mathbf{Y}$ are independent.

The main theoretical result on the distribution theory of LSE's under MVN is the following theorem on the distribution of quadratic forms.

Theorem 3.7: (Fisher-Cochran Theorem). Let $\mathbf{Y} \sim$ MVN $(\boldsymbol{\mu}, \mathbf{I})$ and $Q_1, \ldots, Q_k$ be quadratic forms in $\mathbf{Y}$ with $Q_i = \mathbf{Y}^T \mathbf{A}_i \mathbf{Y}; 1 \leq i \leq k$ for $\mathbf{A}_i$ an $n \times n$ matrix such that rank $(\mathbf{A}_i) = n_i$ and

$$\mathbf{Y}^T \mathbf{Y} = Q_1 + Q_2 + \cdots + Q_k$$

Then the Q_i's are independent with distribution $\chi^2(n_i, \delta_i)$ where $\delta_i = (\boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu})$ if and only if $\sum_{i=1}^k n_i = n$. In this case, $\|\boldsymbol{\mu}^T\|^2 = \sum_{i=1}^k \delta_i^2$.

- The most important application of the Fisher-Cochran Theorem is to the analysis of variance table.
- In this table, the total sum of squares, $\mathbf{Y}^T \mathbf{Y}$, is partitioned into sums of squares due to different factors and an error sum of squares.
- The terms in this partition correspond to the quadratic forms $Q_1, \ldots, Q_k$ in the theorem.
- The Fisher-Cochran Theorem provides the theoretical result required to justify F-tests used in the ANOVA table.

- The ANOVA table for regression sometimes takes on a simple form with only two terms in the partition: the error sum of squares and the sum of squares due to regression.
  - This partition corresponds to the decomposition

$$\mathbf{Y} = \mathbf{P_X Y} + (\mathbf{I} - \mathbf{P_X})\mathbf{Y}$$

    in that

$$\begin{aligned} \mathbf{Y}^T\mathbf{Y} &= \|\mathbf{Y}\|^2 \\ &= \|\mathbf{P_X Y}\|^2 + \|(\mathbf{I} - \mathbf{P_X Y}\|^2 \\ &= \mathbf{Y}^T\mathbf{P_X Y} + \mathbf{Y}^T(\mathbf{I} - \mathbf{P_X})\mathbf{Y} \end{aligned}$$

    by the Pythagorean theorem
  - Note that

$$\begin{aligned} \text{rank}\,(\mathbf{P_X}) + \text{rank}\,(\mathbf{I} - \mathbf{P_X}) &= \text{tr}\,(\mathbf{P_X}) + \text{tr}\,(\mathbf{I} - \mathbf{P_X}) \\ &= \text{tr}\,(\mathbf{I}) \\ &= n \end{aligned}$$

    by idempotence.

- Thus we have by the Fisher-Cochran theorem that

$$\frac{1}{\sigma^2}\|\mathbf{P_X Y}\|^2 \text{ and } \frac{1}{\sigma^2}\|(\mathbf{I}-\mathbf{P_X Y}|^2$$

are independent chi-square variables with degrees of freedom equal to $\text{rank}(\mathbf{P_X}) = r$ and $n-r$, respectively, and non-centrality parameters

$$\delta_1 = [\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{P_X}\mathbf{X}\boldsymbol{\beta}]^{\frac{1}{2}} = \|\mathbf{X}\boldsymbol{\beta}\| \text{ and } \delta_2 = [\boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I}-\mathbf{P_X})\mathbf{X}\boldsymbol{\beta}]^{\frac{1}{2}} = 0$$

- Thus, by definition of the F distribution, we have that

$$\frac{\|\mathbf{P_X Y}\|^2/r}{\|(\mathbf{I}-\mathbf{P_X Y}\|^2/n-r} \sim F_{r,n-r}(\delta_1)$$

- This last result is the F test used in the analysis of the general linear model.

There are a variety of corollaries to the Fisher-Cochran Theorem that are very useful. Several of these results are:

Corollary 3.8: Let $\mathbf{Y} \sim \text{MVN}(\boldsymbol{\mu}, \mathbf{I})$. Then $\mathbf{Y}^T\mathbf{A}\mathbf{Y} \sim \chi^2_\nu(\delta)$ with $\nu = \text{rank}(\mathbf{A})$ and $\delta^2 = \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$ if and only if A is idempotent.

Corollary 3.9: Let $\mathbf{Y}^T\mathbf{A}_1\mathbf{Y} \sim \chi^2_{\nu_1}(\delta_1)$ and $\mathbf{Y}^T\mathbf{A}_2\mathbf{Y} \sim \chi^2_{\nu_2}(\delta_2)$. Then $\mathbf{Y}^T\mathbf{A}_1\mathbf{Y}$ and $\mathbf{Y}^T\mathbf{A}_2\mathbf{Y}$ are independent if and only if $\mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$.

Corollary 3.10: Let Let $\mathbf{Y} \sim \text{MVN}(\boldsymbol{\mu}, \mathbf{I})$ and

$$\mathbf{Y}^T\mathbf{Y} = Q_1 + Q_2 + \cdots + Q_k$$

where $Q_i = \mathbf{Y}^T\mathbf{A}_i\mathbf{Y}$ and $\mathbf{A}_i$ is an $n \times n$ matrix. Then either of the following conditions are necessary and sufficient for

$$Q_i \sim \chi^2_{n_i}(\delta_i) \text{ where } \text{rank}(\mathbf{A}_i) = n_i \text{ and } \delta_i^2 = \boldsymbol{\mu}^T\mathbf{A}_i\boldsymbol{\mu}$$

and for $Q_1, Q_2, \ldots, Q_k$ to be independent.

- $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ are each idempotent matrices
- $\mathbf{A}_i\mathbf{A}_j = \mathbf{0}$ for all $i \neq j$.

Corollary 3.11: Let $\mathbf{Y} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. with $\boldsymbol{\Sigma}$ full rank. Then $\mathbf{Y}^T\mathbf{A}\mathbf{Y} \sim \chi^2_\nu(\delta)$ with $\nu = \text{rank}(A)$ and $\delta^2 = \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$ and only if $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{A}$.

Corollary 3.12: Let $\mathbf{Y} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. with $\boldsymbol{\Sigma}$ full rank. Then

- $\mathbf{P}^T\mathbf{Y}$ and $\mathbf{Y}^T\mathbf{A}\mathbf{Y}$ are independent if and only if $\mathbf{P}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{0}$.
- $\mathbf{Y}^T\mathbf{A}\mathbf{Y}$ and $\mathbf{Y}^T\mathbf{B}\mathbf{Y}$ are independent if and only if $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B} = \mathbf{0}$.

3.9.2 Confidence Ellipsoids for Estimable Functions

Let $\mathbf{Y} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ with $r = \text{rank}\,(\mathbf{X})$.

- Let $\boldsymbol{\Psi} = \mathbf{L}^T\boldsymbol{\beta}$ be an estimable function of $\boldsymbol{\beta}$ where $\mathbf{L}$ is $p \times q$ $(q \leq p)$ of rank q.
- Let $\widehat{\boldsymbol{\Psi}} = \mathbf{L}^T\mathbf{b}$ denote the (unique) LSE of $\boldsymbol{\Psi}$.
- Then we know from previous results that

$$\widehat{\boldsymbol{\Psi}} \sim \text{MVN}(\boldsymbol{\Psi}, \sigma^2\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L})$$

- We also note that $\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L}$ is of full rank.
- Therefore by Corollary 3.11, we have

$$\frac{1}{\sigma^2}(\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi})T(\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L})^{-1}(\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}) \sim \chi^2_q(0)$$

- Using previous results on the independence of SSE and estimable functions and on the distribution of SSE, it follows that

$$\frac{(\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi})T(\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L})^{-1}(\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi})/q}{\text{SSE}/(n-r)} \sim F_{q,n-r}(0)$$

Confidence sets are generalizations of the familiar notion of confidence intervals.

- A confidence set for a vector valued parameter $\mathbf{\Psi} \in \mathbf{R}^p$ with confidence coefficient $1-\alpha$ is defined as a random region, $\mathbf{R}(\mathbf{Y}) \subset \mathbf{R}^p$, depending on the observations with the property

$$P_{\mathbf{\Psi}}(\mathbf{\Psi} \in \mathbf{R}(\mathbf{Y})) = 1 - \alpha$$

  for any $\mathbf{\Psi} \in \mathbf{R}^p$.

- This probability statement is to be interpreted as the long run proportion of sets $\mathbf{R}(\mathbf{Y})$, obtained by repeated sampling of $\mathbf{Y}$, that cover the true parameter value $\mathbf{\Psi}$, is $1-\alpha$.

- With this definition, we see that

$$\left\{\mathbf{\Psi} \in \mathbf{R}^p \ : \ (\widehat{\mathbf{\Psi}} - \mathbf{\Psi})^T (\mathbf{L}^T (\mathbf{X}^T\mathbf{X})^- \mathbf{L})^{-1} (\widehat{\mathbf{\Psi}} - \mathbf{\Psi}) \leq q s^2 F^{1-\alpha}_{q,n-r}(0)\right\}$$

  is a confidence set for $\mathbf{\Psi}$ with confidence coefficient $1-\alpha$ where $F^{1-\alpha}_{q,n-r}(0)$ is the $1-\alpha$ quantile of a central F distribution with q and $n-r$ degrees of freedom.

- Because this region is an ellipsoid in $\mathbf{R}^q$ with center at $\widehat{\mathbf{\Psi}}$, we call this a $100(1-\alpha)\%$ confidence ellipsoid for $\mathbf{\Psi}$.

- Recalling the relationship between Student's t distribution with $n-r$ degrees of freedom and the F distribution with 1 and $n-r$ degrees of freedom, we note that when $q = 1$, the $100(1-\alpha)\%$ confidence ellipsoid given above, reduces to the familiar confidence interval based on the Student's t distribution:

$$\left\{ \boldsymbol{\Psi} \in \mathbf{R} \ : \ \boldsymbol{\Psi} \in \ \widehat{\boldsymbol{\Psi}} - t_{n-r}^{1-\alpha/2} \sigma_{\widehat{\boldsymbol{\Psi}}} \ , \ \widehat{\boldsymbol{\Psi}} + t_{n-r}^{1-\alpha/2} \sigma_{\widehat{\boldsymbol{\Psi}}} \right\}$$

where $t_{n-r}^{1-\alpha/2}$ is the $1-\alpha/2$ quantile of Student's t distribution with $n-r$ degrees of freedom and

$$\sigma^2_{\widehat{\boldsymbol{\Psi}}} = \frac{s^2}{\boldsymbol{\ell}^T (\mathbf{X}^T \mathbf{X})^- \boldsymbol{\ell}}$$

3.9.3 Hypotheis Tests for LInearly Estimable Functions

Tests of hypotheses of the form $H_0: \boldsymbol{\Psi} = \mathbf{L}^T\boldsymbol{\beta} = \mathbf{0}$ may be constructed from the distribution theory used to derive confidence intervals for $\boldsymbol{\Psi}$.

- We know that

$$\frac{(n-r)}{qSSE}\widehat{\boldsymbol{\Psi}}^T(\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L})^{-1}\widehat{\boldsymbol{\Psi}} \sim F_{q,n-r}(\delta)$$

  where

$$\delta^2 = \boldsymbol{\Psi}^T(\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L})^{-1}\boldsymbol{\Psi}$$

  and $\boldsymbol{\Psi}$ is the true parameter value.

- In particular, when the null hypothesis, is true, we have

$$\frac{(n-r)}{qSSE}\widehat{\boldsymbol{\Psi}}^T(\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L})^{-1}\widehat{\boldsymbol{\Psi}} \sim F_{q,n-r}(0)$$

- Thus a test with size α of the hypothesis $H_0: \boldsymbol{\Psi} = \mathbf{0}$ is given by the rule:

$$\text{reject } H_0 \text{ if } \frac{(n-r)}{qSSE}\widehat{\boldsymbol{\Psi}}^T(\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L})^{-1}\widehat{\boldsymbol{\Psi}} > F^{1-\alpha}_{q,n-r}(0)$$

- The power of this test against alternatives $\mathbf{\Psi} \neq \mathbf{0}$ is found by computing the probability that a random variable with distribution $F_{n-r}(\delta)$ where $\delta^2 = \mathbf{\Psi}^T(\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^{-}\mathbf{L})^{-1}\mathbf{\Psi}$ exceeds the critical value $F_{n-r}^{1-\alpha}(0)$.
- Tables of non central F distributions can be found in Scheffe's book, The Analysis of Variance.
- Most software packages contain the non central F as a part of their routines.

3.9.4 Likelihood Ratio Tests

In deriving the likelihood ratio test for the hypothesis

$$H_0 : \ \boldsymbol{\Psi} = \mathbf{L}^T\boldsymbol{\beta} = \mathbf{0}$$

it is convenient to express H_0 in an alternate form used in a previous section.

- Specifically, H_0 may be expressed in tems of a restriction on $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ to lie in a $(r-q)$ dimensional subspace of $\mathbf{Sp}(\mathbf{X})$.
- In particular, write $\mathbf{L} = \mathbf{CX}$ (since $\mathbf{L}^T\boldsymbol{\beta}$ linearly estimable)
- Note that $\mathbf{L}^T\boldsymbol{\beta} = \mathbf{0}$ if and only if $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \in [\mathbf{Sp}\,(\mathbf{C})]^{\perp}$.
- Thus, in the "full" (non-restricted) model, $E(\mathbf{Y}) \in \mathbf{Sp}(\mathbf{X})$ and in the "null" (restricted) model, $E(\mathbf{Y}) \in \mathbf{Sp}(\mathbf{X}) \cap [\mathbf{Sp}(\mathbf{C})]^{\perp}$.

The likelihood ratio test is constructed by maximizing the likelihood under the null model and comparing this maximized likelihood to the maximized likelihood under the full model.

- If the maximized likelihood under the null model is not too much less than that under the full model (i.e., if the null model provides about as good a fit to the data as the full model), then we do not reject the hypothesis.
- If the full model fits the data much better than the null model, we reject the hypothesis.

Let $(\mathbf{Xb})_F$ and $(\mathbf{Xb})_R$ denote the LSE's of $E(\mathbf{Y})$ under the full and reduced models respectively.

- That is, $(\mathbf{Xb})_F$ is the projection of $\mathbf{Y}$ onto $\mathbf{Sp}(\mathbf{X})$ and $(\mathbf{Xb})_R$ is the projection of $\mathbf{Y}$ onto $\mathbf{Sp}(\mathbf{X}) \cap [\mathbf{S}(\mathbf{C})]^{\perp}$.
- Then, recalling the form of MVN p.d.f., we have that

$$(2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}\|\mathbf{y} - (\mathbf{Xb})_F\|^2\}$$

$$(2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}\|\mathbf{y} - (\mathbf{Xb})_R\|^2\}$$

  are the likelihoods under the full and null models that have been maximized with respect to the parameter $\boldsymbol{\beta}$.
- Recall that the maximization over $\boldsymbol{\beta}$ may be done independently of σ^2 by virtue of the orthogonality of σ^2 and $\boldsymbol{\beta}$.
- Now, maximizing these functions with respect to σ^2, we obtain the (fully) maximized likelihoods under the full and null models, respectively:

$$\left(2\pi\frac{\|\mathbf{y} - (\mathbf{Xb})_F\|^2}{n}\right)^{-\frac{n}{2}} \exp\{-\frac{n}{2}\}$$

$$\left(2\pi\frac{\|\mathbf{y} - (\mathbf{Xb})_R\|^2}{n}\right)^{-\frac{n}{2}} \exp\{-\frac{n}{2}\}$$

- Thus the ratio of the maximized likelihoods is

$$\begin{aligned}\lambda &= \left\{\frac{\|\mathbf{y}-(\mathbf{Xb})_R\|^2}{\|\mathbf{y}-(\mathbf{Xb})_F\|^2}\right\}^{-\frac{n}{2}} \\ &= \left\{\frac{\|\mathbf{y}-(\mathbf{Xb})_F\|^2}{\|\mathbf{y}-(\mathbf{Xb})_R\|^2}\right\}^{\frac{n}{2}}\end{aligned}$$

  and the likelihood ratio test is the rule: reject H_0 : if$\lambda < c_\alpha$.

- To determine the critical value, c_α , we need the distribution of λ (or some monotone function of λ).

- Recall from orthogonality relationships,

$$\|\mathbf{Y}-(\mathbf{Xb})_R\|^2 = \|\mathbf{Y}-(\mathbf{Xb})_F\|^2 + \|(\mathbf{Xb})_F-(\mathbf{Xb})_R\|^2$$

- Thus

$$\begin{aligned}\lambda &= \left\{\frac{\|\mathbf{Y}-(\mathbf{Xb})_F\|^2}{\|\mathbf{Y}-(\mathbf{Xb})_F\|^2+\|(\mathbf{Xb})_F-(\mathbf{Xb})_R\|^2}\right\}^{\frac{n}{2}} \\ &= \left\{\frac{1}{1+\frac{\|(\mathbf{Xb})_F-(\mathbf{Xb})_R\|^2}{\|(\mathbf{Y}-(\mathbf{Xb})_F\|^2}}\right\}^{\frac{n}{2}}\end{aligned}$$

- Thus the rejection role for the likelihood ratio test may be expressed:

$$\text{reject } H_0 \text{ if } \frac{\|(\mathbf{Xb})_F-(\mathbf{Xb})_R\|^2}{\|\mathbf{Y}-(\mathbf{Xb})_F\|^2} > \tilde{c}_\alpha$$

  for some critical value $\tilde{c}_\alpha$.

Now let $\mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}$ denote the projection operator onto the subspace

$$\mathbf{Sp}(\mathbf{X}) \cap [\mathbf{Sp}(\mathbf{C})]^{\perp}$$

- Note that $\mathbf{X}\mathbf{b}_R = \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}\mathbf{Y}$.
- Also let $\mathbf{P}_{\mathbf{X}}$ denote the projection operator onto $\mathbf{Sp}(\mathbf{X})$ so that $(\mathbf{X}\mathbf{b})_F = \mathbf{P}_{\mathbf{X}}\mathbf{Y}$.
- Because

  $$\mathbf{Sp}(\mathbf{X}) \cap [\mathbf{Sp}(\mathbf{C})]^{\perp} \subset \mathbf{Sp}(\mathbf{X})$$

  we have

  $$\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}\mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}\mathbf{P}_{\mathbf{X}} = \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}$$

- Therefore

  $$\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}$$

  is a projection operator onto the subspace $\mathbf{Sp}(\mathbf{X}) \cap [\mathbf{Sp}(\mathbf{C})]^{\perp}$) since

  - It is obviously symmetric.
  - It is also idempotent since

$$\begin{aligned}\left(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}\right) &= \mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}} - \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}\mathbf{P}_{\mathbf{X}} + \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}\mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}} \\ &= \mathbf{P}_{\mathbf{X}} - 2\mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}} + \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}} \\ &= \mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}\end{aligned}$$

- We therefore have the following orthogonal decomposition of $\mathbf{Y}$

$$\begin{aligned}\mathbf{Y} &= \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}\mathbf{Y} + (\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}})\mathbf{Y} + (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y} \\ &= (\mathbf{X}\mathbf{b}_R + [(\mathbf{X}\mathbf{b})_F - (\mathbf{X}\mathbf{b})_R)] + (\mathbf{Y} - (\mathbf{X}\mathbf{b})_F)\end{aligned}$$

- and the corresponding decomposition of the sum of squares:

$$\begin{aligned}\mathbf{Y}^T\mathbf{Y} &= \mathbf{Y}^T\mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}\mathbf{Y} + \mathbf{Y}^T(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}})\mathbf{Y} + \mathbf{Y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y} \\ &= \|(\mathbf{X}\mathbf{b})_R\|^2 + \|(\mathbf{X}\mathbf{b})_F(\mathbf{X}\mathbf{b})_R\|^2 + \|(\mathbf{Y} - (\mathbf{X}\mathbf{b})_F\|^2\end{aligned}$$

- By Corollary 3.10 (1) we have that $\|(\mathbf{X}\mathbf{b})_F - (\mathbf{X}\mathbf{b})_R\|^2$ and $\|(\mathbf{Y} - \mathbf{X}\mathbf{b})_F\|^2$ are independent chi square random variables on q and $n-r$ degrees of freedom, respectively.
- The error sum of squares, $\|(\mathbf{Y} - \mathbf{X}\mathbf{b})_F\|^2$, is a central chi-square
- The non centrality parameter for $\|(\mathbf{X}\mathbf{b})_F - (\mathbf{X}\mathbf{b})_R\|^2$, δ, is given by

$$\begin{aligned}\delta^2 &= \boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}})\mathbf{X}\boldsymbol{\beta} \\ &= \|\mathbf{X}\boldsymbol{\beta} - \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}\mathbf{X}\boldsymbol{\beta}\|^2\end{aligned}$$

- Thus δ may be interpreted as the distance that the true mean, $\mathbf{X}\boldsymbol{\beta}$, is from the subspace to which it is constrained under the null hypothesis. If the null hypothesis is true, then $\delta = 0$.

- Therefore, by normalizing the ratio of quadratic forms, we have the likelihood ratio test given by the rule:

$$\text{reject } H_0 if \frac{\|(\mathbf{Xb})_F - (\mathbf{Xb})_R\|^2/q}{\|\mathbf{Y} - (\mathbf{Xb})_F\|^2/(n-r)} > F_{q,n-r}^{1-\alpha}(0)$$

- Note that because of the orthogonality of

$$\mathbf{Y} - (\mathbf{Xb})_F \quad \text{and} \quad (\mathbf{Xb})_F - (\mathbf{Xb})_R$$

we have

$$\|(\mathbf{Xb})_F - (\mathbf{Xb})_R\|^2 = \|\mathbf{Y} - (\mathbf{Xb})_F\|^2 - \|\mathbf{Y} - (\mathbf{Xb})_R\|^2$$

- Therefore the likelihood ratio test statistic may also be computed by the formula

$$\left\{ \frac{\|\mathbf{Y} - (\mathbf{Xb})_F\|^2 - \|\mathbf{Y} - (\mathbf{Xb})_R\|^2}{\|\mathbf{Y} - (\mathbf{Xb})_F\|^2} \right\} \frac{n-r}{q} = \left\{ \frac{\text{SSE}_R - \text{SSE}_F}{\text{SSE}_F} \right\} \frac{n-r}{q}$$

where SSE_F and SSE_R are the error sums of squares or deviances under the full and reduced models respectively.

3.9.5 Likelihood Ratio Tests and Confidence Ellipsoids

In a previous section , we derived a test for the hypothesis

$$H_0 : \ \mathbf{\Psi} = \mathbf{L}^T \boldsymbol{\beta} = \mathbf{0}$$

based on the distribution of the LSE $\widehat{\mathbf{\Psi}}$, which was of the form:

$$\text{reject } H_0 \text{ if } \frac{\widehat{\mathbf{\Psi}}^T [\mathbf{L}^T (\mathbf{X}^T \mathbf{X})^- \mathbf{L}]^{-1} \widehat{\mathbf{\Psi}}}{q s^2} > F^{1-\alpha}_{q,(n-r)}(0)$$

In the previous section we found that the likelihood ratio test of

$$H_0 : \ \mathbf{\Psi} = \mathbf{L}^T \boldsymbol{\beta} = \mathbf{0}$$

was of the form:

$$\text{reject } H_0 \text{ if } \frac{\|(\mathbf{Xb})_F - (\mathbf{Xb})_R\|^2}{q s^2} > F^{1-\alpha}_{q,(n-r)}(0)$$

To show the equivalence of these tests, we need only show that

$$\widehat{\boldsymbol{\Psi}}^T[\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L}]^{-1}\widehat{\boldsymbol{\Psi}} = \mathrm{SSE}_R - \mathrm{SSE}_F$$

To accomplish this, we consider the reparameterization

$$\mathbf{X} = \mathbf{X}\mathbf{W}^{-1}\mathbf{W}\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\gamma}$$

where $\mathbf{W}$ is a $p \times p$ matrix of full rank written as

$$\mathbf{W} = \begin{bmatrix} \mathbf{M}^T \\ \mathbf{L}^T \end{bmatrix}$$

where $\mathbf{M}$ is a $p \times (p-q)$ matrix with column vectors spanning $\mathbf{Sp}^{\perp}(\mathbf{L})$ and $\mathbf{L}$ is the $p \times q$ matrix defining $\boldsymbol{\Psi} = \mathbf{L}^T\boldsymbol{\beta}$.

- Thus, writing our reparameterized model in partitioned form, we have
$$\mathbf{X}\boldsymbol{\beta} = \mathbf{Z}_1\boldsymbol{\gamma}_1 + \mathbf{Z}_2\boldsymbol{\gamma}_2 \text{ where } \boldsymbol{\gamma}_2 = \boldsymbol{\Psi}$$
- By previous results on reparameterizations, we know that γ_2 is linearly estimable and that the LSE's of estimable functions and error sum of squares are invariant under reparameterization.

- Let $\mathbf{P}_{\mathbf{Z_1}}$ denote the projection operator onto the subspace $\mathbf{Sp}(\mathbf{Z}_1)$. Then the LSE of $\boldsymbol{\gamma}_2$ is given by

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_2 &= [\mathbf{Z}_2^T(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_1})\mathbf{Z}_2]^{-1}\mathbf{Z}_2^T(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_1})\mathbf{Y} \\ \text{and } \text{var}\,(\hat{\boldsymbol{\gamma}}_2) &= \sigma^2[\mathbf{Z}_2^T(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_1})\mathbf{Z}_2]^{-1} \end{aligned}$$

- By our invariance result under reparameterization, we have

$$\widehat{\boldsymbol{\Psi}}^T \text{var}\,(\widehat{\boldsymbol{\Psi}})\widehat{\boldsymbol{\Psi}} = \hat{\boldsymbol{\gamma}}_2^T \text{var}\,(\hat{\boldsymbol{\gamma}}_2)\hat{\boldsymbol{\gamma}}_2$$

- Therefore

$$\begin{aligned} \widehat{\boldsymbol{\Psi}}^T[\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L}]^{-1}\widehat{\boldsymbol{\Psi}} &= \hat{\boldsymbol{\gamma}}_2^T[\mathbf{Z}_2^T(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_1})\mathbf{Z}_2]\hat{\boldsymbol{\gamma}}_2 \\ &= \mathbf{Y}^T(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_1})\mathbf{Z}_2[\mathbf{Z}_2^T(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_1})\mathbf{Z}_2]^{-1}\mathbf{Z}_2^T(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_1})\mathbf{Y} \\ &= \|\mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_1})\mathbf{Z}_2}\mathbf{Y}\|^2 \end{aligned}$$

- Note that we have the orthogonal decomposition of $\mathbf{Y}$,

$$\begin{aligned} \mathbf{Y} &= \mathbf{P}_{\mathbf{Z}}\mathbf{Y} + (\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{Y} \\ &= \mathbf{P}_{\mathbf{Z}_1}\mathbf{Y} + \mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_1})\mathbf{Z}_2}\mathbf{Y} + (\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{Y} \end{aligned}$$

  and a corresponding decomposition of $\mathbf{Y}^T\mathbf{Y}$ as

$$\mathbf{Y}^T\mathbf{Y} = \|\mathbf{P}_{\mathbf{Z}_1}\mathbf{Y}\|^2 + \|\mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_1})\mathbf{Z}_2}\mathbf{Y}\|^2 + \|(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{Y}\|^2$$

- Since error sums of squares are invariant under reparameterization, we have

$$\begin{aligned} \mathrm{SSE}_F &= \|(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{Y}\|^2 \\ \mathrm{SSE}_R &= \mathbf{Y}^T\mathbf{Y} - \|\mathbf{P}_{\mathbf{Z}_1}\mathbf{Y}\|^2 \end{aligned}$$

- Therefore

$$\begin{aligned} \widehat{\mathbf{\Psi}}^T[\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L}]^{-1}\widehat{\mathbf{\Psi}} &= \|\mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_1})\mathbf{Z}_2}\mathbf{Y}\|^2 \\ &= \mathrm{SSE}_R - \mathrm{SSE}_F \end{aligned}$$

  and the likelihood ratio test and the F-test based on the LSE are equivalent.

This simple principle for constructing hypothesis tests (sometimes called the **Principle of Conditional Error** may be summarized as follows:

1. Fit the full model $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ to get SSE .
2. Use the hypothesis $\mathbf{L}^T\boldsymbol{\beta} = \mathbf{0}$ to obtain a reduced or conditional model (i.e. conditional on $\mathbf{L}^T\boldsymbol{\beta} = \mathbf{0}$) and obtain SSE from a fit of this reduced model. Reparameterization may be required for this step.
3. Compute the sum of squares due to the hypothesis $\mathbf{L}^T\boldsymbol{\beta} = \mathbf{0}$ by subtraction of SSE_F from SSE_R.
4. Compute error degrees of freedom under the full and reduced model $(\text{df}_F, \text{df}_R)$, respectively.
5. Construct the test statistic (F-test) by

$$\frac{(\text{SSE}_R - \text{SSE}_F)/(\text{df}_R - \text{df}_F)}{\text{SSE}_F/\text{df}_F}$$

Optimality of the F Test

- Among all tests of size α of the hypothesis $\boldsymbol{\Psi} = \mathbf{L}^T\boldsymbol{\beta} = \mathbf{0}$, with the property that the power of the test depends on $\boldsymbol{\beta}$ through the intermediary

$$\|\mathbf{X}\boldsymbol{\beta} - \mathbf{P}_{\mathbf{X}|\boldsymbol{\Psi}=\mathbf{0}}\mathbf{P}_{\mathbf{X}}\boldsymbol{\beta}\|^2 = \boldsymbol{\Psi}^T[\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L}]^{-1}\boldsymbol{\Psi}$$

  - The F test is uniformly most powerful (UMP).
  - i.e. the F test has higher power than any other test (in the class described above) uniformly over all possible alternatives.

3.9.6 ANOVA Tables

A convenient way to present results from a least squares analysis of a linear model is the analysis of variance (ANOVA) table.

- The ANOVA table presents information about hypothesis tests (sums of squares, degrees of freedom) but does not include information about the LSE of $\boldsymbol{\beta}$ itself.
- The essential requirement in constructing an ANOVA table is to determine an orthogonal decomposition of $\mathbf{Y}$, or equivalently, an orthogonal decomposition of the total sum of squares $\mathbf{Y}^T\mathbf{Y}$.
- We consider the case of

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = \sum_{i=1}^{k} \mathbf{X}_i\boldsymbol{\beta}_i$$

  where the hypotheses of interest are

$$H_{01} : \boldsymbol{\beta}_1 = \mathbf{0};\ H_{02} : \boldsymbol{\beta}_2 = \mathbf{0};\ \ldots;\ H_{0k} : \boldsymbol{\beta}_k = \mathbf{0}$$

- The usual way that an orthogonal decomposition is determined (that is relevant to $H_{01}, H_{02}, \ldots, H_{0k}$ is to decide upon an ordering of the hypotheses to be tested. Without loss of generality, let this ordering be $H_{01}, H_{02}, \ldots, H_{0k}$

- By imposing this ordering, the k hypotheses are now tested in the sense that our first test will be of H_{0k} , our second test will be $H_{0(k-1)}$ given H_{0k} is true, etc.

- Thus, for the ith test, our full model will be

$$E(\mathbf{Y}) = \sum_{j=1}^{i} \mathbf{X}_j \boldsymbol{\beta}_j$$

and our reduced model will be

$$E(\mathbf{Y}) = \sum_{j=1}^{i-1} \mathbf{X}_j \boldsymbol{\beta}_j$$

- We introduce the following notation

$$\begin{aligned} \mathbf{D}_0 &= \mathbf{I} \\ \mathbf{D}_1 &= \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^-\mathbf{X}_1^T \\ \mathbf{D}_{12} &= \mathbf{D}_1 - \mathbf{D}_1\mathbf{X}_2(\mathbf{X}_2^T\mathbf{D}_1\mathbf{X}_2)^-\mathbf{X}_1^T\mathbf{D}_1 \end{aligned}$$

and in general

$$\mathbf{D}_{\pi(n)} = \mathbf{D}_{\pi(n-1)} - \mathbf{D}_{\pi(n-1)}\mathbf{X}_n(\mathbf{X}_n^T\mathbf{D}_{\pi(n-1)}\mathbf{X}_n)^-\mathbf{X}_n^T\mathbf{D}_{\pi(n-1)}$$

where $\pi(n)$ is any permutation of $\{1, 2, \ldots, n\}$, for $n = 1, 2, \ldots, k$ and $\pi(0) = 1$.

- Note that

$$\begin{aligned}\mathbf{P}_1 &= \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^-\mathbf{X}_1^T \\ &= \mathbf{D}_0 - \mathbf{D}_1\end{aligned}$$

$$\begin{aligned}\mathbf{P}_2 &= \mathbf{D}_1\mathbf{X}_1(\mathbf{X}_1^T\mathbf{D}_1\mathbf{X}_1)^-\mathbf{X}_1^T\mathbf{D}_1 \\ &= \mathbf{D}_1 - \mathbf{D}_{12}\end{aligned}$$

  and in general

$$\mathbf{P}_n\mathbf{D}_{\pi(n-1)}\mathbf{X}_n(\mathbf{X}_n^T\mathbf{D}_{\pi(n-1)}\mathbf{X}_n)^-\mathbf{X}_n^T\mathbf{D}_{\pi(n-1)} = \mathbf{D}_{\pi(n-1)} - \mathbf{D}_{\pi(n)}$$

  for $n = 1, 2, \ldots, k$.

- Note that $\mathbf{P}_i\mathbf{P}_j = \mathbf{0}$ for $i \neq j$ and that

$$\mathbf{y}^T\mathbf{y} = \sum_{j=1}^{k} \mathbf{y}^T\mathbf{P}_j\mathbf{y} + \mathbf{Y}^T\mathbf{D}_{\pi(k)}\mathbf{y}$$

- $\mathbf{y}^T\mathbf{P}_j\mathbf{y}$ is the sum of squares for $\mathbf{X}_j$ in the presence of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{j-1}$ and in the absence of $\mathbf{X}_{j+1}, \mathbf{X}_{j+2}, \ldots, \mathbf{X}_k$.

- This sum of squares is often called the sum of squares for $\mathbf{X}_j$ adjusted for $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{j-1}$ and ignoring $\mathbf{X}_{j+1}, \mathbf{X}_{j+2}, \ldots, \mathbf{X}_k$.

The ANOVA table provides a convenient way to present the corresponding partition of $\mathbf{y}^T\mathbf{y}$:

Source	df	SS
$\mathbf{X}_1$	$\text{tr}(\mathbf{P}_1)$	$\|\mathbf{P}_1\mathbf{y}\|^2$
$\mathbf{X}_2$ adjusted for $\mathbf{X}_1$	$\text{tr}(\mathbf{P}_2)$	$\|\mathbf{P}_2\mathbf{y}\|^2$
$\vdots$	$\vdots$	$\vdots$
$\mathbf{X}_k$ adjusted for $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{k-1}$	$\text{tr}(\mathbf{P}_k)$	$\|\mathbf{P}_k\mathbf{y}\|^2$
Error	$n - r$	$\|\mathbf{D}_{\pi(k)}\mathbf{y}\|^2$
Total	n	$\|\mathbf{y}^T\mathbf{y}\|^2$

where $r = \text{rank}\,[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k]$.

Sometimes, when there are two components to the linear model,

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$$

and only $\mathbf{X}$ is of primary interest, a reduced ANOVA is presented.

- This table represents the second stage of the two stage least squares procedure described before, in which both $\mathbf{X}$ and $\mathbf{Y}$ were first made orthogonal to $\mathbf{Z}$.
- Let $\mathbf{D_Z}$ denote $\mathbf{I} - \mathbf{P_Z}$.
- Then the "reduced" ANOVA table may be written

Source	df	SS
$\mathbf{X}_1$ adjusted for $\mathbf{Z}$	$\mathrm{tr}(\mathbf{P}_{\mathbf{D_Z}}\mathbf{X}_1)$	$\|\mathbf{P}_{\mathbf{D_Z}\mathbf{X}_1}\mathbf{y}\|^2$
$\mathbf{X}_2$ adjusted for $\mathbf{Z}, \mathbf{X}_1$	$\mathrm{tr}(\mathbf{P}_{\mathbf{D_Z}}\mathbf{P}_2)$	$\|\mathbf{P}_{\mathbf{D_Z}\mathbf{X}_2}\mathbf{y}\|^2$
$\vdots$	$\vdots$	$\vdots$
$\mathbf{X}_k$ adjusted for $\mathbf{Z}, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{k-1}$	$\mathrm{tr}(\mathbf{P}_{\mathbf{D_Z}}\mathbf{P}_{\mathbf{D}_k})$	$\|\mathbf{P}_{\mathbf{D_Z}\mathbf{D}_{\pi(k-1)}\mathbf{X}_k}\mathbf{y}\|^2$
Error	$\mathrm{tr}(\mathbf{I} - \mathbf{D_Z})$	$\|(\mathbf{I} - \mathbf{P}_{\mathbf{D_Z}\mathbf{X}})\mathbf{y}\|^2$
Total	$\mathrm{tr}(\mathbf{D_Z})$	$\|\mathbf{D_Z}\mathbf{y}\|^2$

- Examples of this "reduced" ANOVA table are:
  - $\mathbf{Z} = \mathbf{1}$ (the mean is removed) so (sub)total degrees of freedom is $n - 1$, (sub)total sum of squares is $\mathbf{y}^T\mathbf{y} - n\bar{y}^2$.
  - ANCOVA where $\mathbf{X}$ is the experimental design varible and $\mathbf{Z}$ are covariates.

- A very important special case of the ANOVA table is when $\mathbf{X}_1, \ldots, \mathbf{X}_k$ are mutually orthogonal.
- In this case, the sum of squares associated with the hypotheses are independent of the ordering.
- Interpretation of the hypothesis tests is much easier in this case and, for this reason, much effort has been expended to identify economical experimental designs with this orthogonality property.
- This is a major part of the study of experimental designs.