Testing Linear Restrictions on Parameters via F-tests

Simon Jackman
Department of Political Science
Stanford University

January 31, 2007

References

- ALR3, §5.4
- Ruud, Paul A. 2000. *An Introduction to Classical Econometric Theory*. New York: Oxford University Press.
- Seber, George A.F. and Alan J. Lee. 2003. *Linear Regression Analysis*, 2nd edition. Hoboken, New Jersey: Wiley.

Testing Linear Restrictions on Parameters

- We use the t-test for simple hypothesis tests; e.g.,

$$H_0 : \beta_1 = 0$$

- t-test can also be used for testing hypotheses involving more than one parameter: e.g.,

$$H_0 : \beta_2 - \beta_1 = 0$$

which we would test by forming the test statistic

$$\begin{aligned} t &= \frac{\hat{\beta}_2 - \hat{\beta}_1}{\sqrt{var(\hat{\beta}_2 - \hat{\beta}_1)}} \\ &= \frac{\hat{\beta}_2 - \hat{\beta}_1}{\sqrt{var(\hat{\beta}_2) + var(\hat{\beta}_1) - 2cov(\hat{\beta}_2, \hat{\beta}_1)}} \end{aligned}$$

Testing Joint Hypotheses (Sets of Linear Restrictions on Parameters)

We use the F-test for testing *joint* or *compound* hypotheses: e.g, all slope coefficients are zero.

$$H_0 : \beta_2 = 0 \text{ AND } \beta_3 = 0 \ldots \text{ AND } \beta_k = 0$$

which can also be written as

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

For this case the test statistic

$$\boxed{F = \frac{r^2/(k - 1)}{(1 - r^2)/(n - k)} = \frac{\text{RegSS}/(k - 1)}{\text{RSS}/(n - k)}}$$

where RegSS = "regression sum of squares" and RSS = "residual sum of squares". This statistic follows the F distribution with k and $n - k$ degrees of freedom.

Some Theory: the F distribution

Proposition 1. *The ratio of two independent chi-square variables, each divided by its degrees of freedom, follows an F-distribution. That is, if*

$$s_1^2 \sim \chi^2_v,\ s_2^2 \sim \chi^2_w$$

where $p(s_1, s_2) = p(s_1)p(s_2)$, then

$$\frac{s_1^2/v}{s_2^2/w} \sim F_{v,w}$$

- We use this result for statistical tests of differences in *sums of squared residuals*. from an *unrestricted* model and a *restricted* model, testing whether the difference in these two sums of squared residuals is statistically significant.

- This comparison amounts to a test of the null hypothesis that the restrictions are true, against the alternative of the unrestricted model.

Restricted Models

We define a "restricted model" as a model with linear restrictions on the elements of $\boldsymbol{\beta}$ relative to an unrestricted model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. All linear restrictions are of the form:

$$\boxed{\mathbf{R}\boldsymbol{\beta} = \mathbf{r}}$$

where $\mathbf{R}$ is a q by k matrix and $\mathbf{r}$ is a q by 1 vector embodying the restrictions on $\boldsymbol{\beta}$, i.e., a system of q linear restrictions over the k parameters.

Linear Restrictions on $\boldsymbol{\beta}$

For instance, consider the following model

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

and the joint null hypothesis

$$H_0 : \beta_2 = \beta_3 = 0$$

Then for $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$,

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Linear Restrictions on β

Other examples:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

$$\begin{aligned} H_0: \quad \beta_1 + \beta_2 &= 2 \\ \beta_2 - 3\beta_3 &= 7 \end{aligned}$$

Then for $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$,

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & -3 \end{bmatrix} \quad \mathbf{r} = \begin{bmatrix} 2 \\ 7 \end{bmatrix}$$

Linear Restrictions on $\boldsymbol{\beta}$

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$
$$H_0 : \beta_2 + \beta_3 = 0$$

Then for $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$,

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}, \mathbf{r} = 0$$

which is **not** a joint hypothesis and could be tested using a t-test

n.b., t tests are special cases of the F test.

t-tests test just one linear restriction on $\hat{\boldsymbol{\beta}}$.

Linear Restrictions on β

The "omnibus" F test that all k - 1 slope coefficients are zero is obtained with

$$\underset{(k\text{-}1\times k)}{\mathbf{R}} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

and $\mathbf{r} = \mathbf{0}$ (a k - 1 null vector).

F tests for comparing models

- The F-test is a device for testing differences in the sum of the squared residuals obtained by estimating a restricted model and an unrestricted model.
- RSS: residual sum of squares from unrestricted model.
- RSS_r: residual sum of squares from model with q linear restrictions on $\hat{\boldsymbol{\beta}}$.
- We consider each in turn, showing how the F-test statistic actually follows the F distribution.

Claim 1: distribution of unrestricted residual sum of squares

- RSS = residual sum of squares = $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$. We want to know its distribution.
- Start by assuming $\varepsilon_i \sim N$, and recalling that by assumption, $E(\varepsilon_i) = 0$ and $\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \forall\, i \neq j$.
- Then $z_i = \varepsilon_i/\sigma \sim N(0,1)$ and $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/\sigma = \sum_{i=1}^n z_i^2 \sim \chi_n^2$
- Now recall that $\hat{\boldsymbol{\varepsilon}} = \mathbf{My}$, where $\mathbf{M} = \mathbf{I}_n$ - $\mathbf{H}$, and $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
- Hence, $\text{var}(\hat{\boldsymbol{\varepsilon}}|\mathbf{X}) = \text{var}(\mathbf{My}|\mathbf{X}) = \mathbf{M}\sigma^2\mathbf{I}_n\mathbf{M} = \sigma^2\mathbf{M}$ (i.e., $\mathbf{M}$ is a symmetric idempotent matrix). But $\mathbf{M}$ is not a diagonal matrix, and so $\text{cov}(\hat{\varepsilon}_i\hat{\varepsilon}_j|\mathbf{X}) \neq 0$.
- So, even though $\hat{\varepsilon}_i|\mathbf{X} \sim N(0, \sigma^2)$, we can't assert that $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/\sigma = \sum_{i=1}^n \hat{\varepsilon}_i^2/\sigma \sim \chi_n^2$ (i.e., we need mutual independence of the z_i for the claim $\sum_{i=1}^n z_i^2 \sim \chi_n^2$ to be true).

Claim 1: distribution of unrestricted residual sum of squares

- We note that $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$, where $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$, and $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{M}$ and $\mathbf{H}$ both symmetric, idempotent, n-by-n matrices.

- Note that $\mathbf{M}$ is not full rank, but has rank $n - k$. Proof to come later; see Theorem 6 and the discussion.

- We use the following useful result on the distribution of a *quadratic form*:

- **Theorem 1. [Seber and Lee (2003), Theorem 2.7]** *If $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_n)$, and $\mathbf{A}$ is a symmetric n-by-n matrix, then $\mathbf{z}'\mathbf{A}\mathbf{z} \sim \chi^2_p$ if and only if $\mathbf{A}$ is idempotent with rank p.*

- $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ and so $\mathbf{z} = \boldsymbol{\varepsilon}/\sigma \sim N(\mathbf{0}, \mathbf{I}_n)$ and by Theorem 1, $\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}/\sigma^2 = \mathbf{z}'\mathbf{M}\mathbf{z} \sim \chi^2_{n-k}$.

- In turn, since RSS $= \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$, we have RSS$/\sigma^2 \sim \chi^2_{n-k}$.

Claim 2: distribution of RSS_r - RSS

We begin by stating (without proof):

$$\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{r} - \mathbf{R}\hat{\boldsymbol{\beta}})$$

The proof requires solving the constrained optimization problem

$$\min_{\hat{\boldsymbol{\beta}}} ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2$$

subject to

$$\mathbf{R}\hat{\boldsymbol{\beta}} = \mathbf{r}.$$

See an advanced text for details...

Seber and Lee, *Linear Regression Analysis* p60

3.8.1 Method of Lagrange Multipliers

Let $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{X}$ is $n \times p$ of full rank p. Suppose that we wish to find the minimum of $\varepsilon'\varepsilon$ subject to the linear restrictions $\mathbf{A}\beta = \mathbf{c}$, where $\mathbf{A}$ is a known $q \times p$ matrix of rank q and $\mathbf{c}$ is a known $q \times 1$ vector. One method of solving this problem is to use Lagrange multipliers, one for each linear constraint $\mathbf{a}_i'\beta = c_i$ $(i = 1, 2, \ldots, q)$, where $\mathbf{a}_i'$ is the ith row of $\mathbf{A}$. As a first step we note that

$$\begin{aligned} \sum_{i=1}^{q} \lambda_i(\mathbf{a}_i'\beta - c_i) &= \boldsymbol{\lambda}'(\mathbf{A}\beta - \mathbf{c}) \\ &= (\beta'\mathbf{A}' - \mathbf{c}')\boldsymbol{\lambda} \end{aligned}$$

(since the transpose of a 1×1 matrix is itself). To apply the method of Lagrange multipliers, we consider the expression $r = \varepsilon'\varepsilon + (\beta'\mathbf{A}' - \mathbf{c}')\boldsymbol{\lambda}$ and solve the equations

$$\mathbf{A}\beta = \mathbf{c} \tag{3.35}$$

and $\partial r/\partial\beta = \mathbf{0}$; that is (from A.8),

$$-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta + \mathbf{A}'\boldsymbol{\lambda} = \mathbf{0}. \tag{3.36}$$

For future reference we denote the solutions of these two equations by $\hat{\beta}_H$ and $\hat{\boldsymbol{\lambda}}_H$. Then, from (3.36),

$$\begin{aligned} \hat{\beta}_H &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \tfrac{1}{2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\hat{\boldsymbol{\lambda}}_H \\ &= \hat{\beta} - \tfrac{1}{2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\hat{\boldsymbol{\lambda}}_H, \end{aligned} \tag{3.37}$$

and from (3.35),

$$\begin{aligned} \mathbf{c} &= \mathbf{A}\hat{\beta}_H \\ &= \mathbf{A}\hat{\beta} - \tfrac{1}{2}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\hat{\boldsymbol{\lambda}}_H. \end{aligned}$$

Since $(\mathbf{X}'\mathbf{X})^{-1}$ is positive-definite, being the inverse of a positive-definite matrix, $\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'$ is also positive-definite (A.4.5) and therefore nonsingular. Hence

$$-\tfrac{1}{2}\hat{\boldsymbol{\lambda}}_H = \left[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\right]^{-1}(\mathbf{c} - \mathbf{A}\hat{\beta})$$

and substituting in (3.37), we have

$$\hat{\beta}_H = \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\left[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\right]^{-1}(\mathbf{c} - \mathbf{A}\hat{\beta}). \tag{3.38}$$

Linear Transformations of Normals

The following theorem will prove handy:

Theorem 2. *Let*

$$\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathbf{z} \in \mathbb{R}^n.$$

Let $\mathbf{A}$ *be a m by n matrix, and* $\mathbf{b}$ *be a m by 1 vector. Then*

$$\mathbf{Az} + \mathbf{b} \sim N(\mathbf{A}\boldsymbol{\mu} + \mathbf{b},\ \ \mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{A}').$$

Claim 2: distribution of RSS_r - RSS

So if

$$\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{r} - \mathbf{R}\hat{\boldsymbol{\beta}})$$

then

$$\begin{aligned}
\mathrm{RSS}_r - \mathrm{RSS} &= (\mathbf{y} - \hat{\mathbf{y}}_r)'(\mathbf{y} - \hat{\mathbf{y}}_r) - (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) \\
&= (\hat{\mathbf{y}} - \hat{\mathbf{y}}_r)'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_r) \\
&= (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_r)' \, \mathbf{X}'\mathbf{X} \, (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_r) \\
&= (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})
\end{aligned}$$

Ordinarily, $\hat{\boldsymbol{\beta}}|\sigma^2 \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. But when H_0 is true (i.e., the restrictions are true), Theorem 2 tell us that $\mathbf{R}\hat{\boldsymbol{\beta}}|\sigma^2 \sim N(\mathbf{r}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')$, i.e., $\mathrm{var}(\mathbf{R}\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$. Thus

$$\frac{\mathrm{RSS}_r - \mathrm{RSS}}{\sigma^2} = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \left[\mathrm{var}(\mathbf{R}\hat{\boldsymbol{\beta}})\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})$$

Claim 2: distribution of RSS_r - RSS

We have

$$\frac{\mathrm{RSS}_r - \mathrm{RSS}}{\sigma^2} = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \left[\mathrm{var}(\mathbf{R}\hat{\boldsymbol{\beta}})\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})$$

- Note that under H_0, $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_r)$, where $\boldsymbol{\Sigma}_r = \mathrm{var}(\mathbf{R}\hat{\boldsymbol{\beta}})$.
- Let $\boldsymbol{\Sigma}_r^{1/2}$ be the q-by-q, positive-definite "square-root" matrix such that $\boldsymbol{\Sigma}_r^{1/2}\boldsymbol{\Sigma}_r^{1/2'} = \boldsymbol{\Sigma}_r$, and similarly $\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\Sigma}_r^{-1/2'} = \boldsymbol{\Sigma}_r^{-1}$.
- By Theorem 2, $\mathbf{z} = \boldsymbol{\Sigma}_r^{-1/2}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \sim N(\mathbf{0}, \mathbf{I}_q)$, since $\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_r^{-1/2'} = \mathbf{I}_q$.
- Thus, by Theorem 1,

$$\begin{aligned}\frac{\mathrm{RSS}_r - \mathrm{RSS}}{\sigma^2} &= (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \left[\mathrm{var}(\mathbf{R}\hat{\boldsymbol{\beta}})\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \\ &= \mathbf{z}'\mathbf{I}_q\mathbf{z} \sim \chi^2_q\end{aligned}$$

Claim 3: Conditional Independence of RSS and RSS_r - RSS

- RSS_r - RSS is a function of $\hat{\boldsymbol{\beta}}$; see previous slides.

- RSS is a function of $\hat{\boldsymbol{\varepsilon}}$, i.e., RSS = $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$.

- We now show that $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are independent, by showing

  1. $\text{cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}} \mid \mathbf{X}) = \mathbf{0}$. (a necessary condition for conditional independence)
  2. Since both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ have normal distributions, zero conditional covariance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ implies conditional independence of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$.

- And thus our main claim is true: conditional on $\mathbf{X}$, RSS and RSS_r - RSS are independent.

- I state some theorems to help us prove these assertions.

Zero Covariance Implies Independence for Normals

Theorem 3. *Suppose*

$$\begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}'_{12} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right),$$

then $\mathbf{z}_1$ *and* $\mathbf{z}_2$ *are independent if and only if* $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

Corollary: If $\mathbf{z}_1$ and $\mathbf{z}_2$ follow normal distributions, then a necessary and sufficient condition for the independence of $\mathbf{z}_1$ and $\mathbf{z}_2$ is to show that their covariance is zero.

Independence of Linear Transforms of Normals

Theorem 4. *If* $\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\mathbf{U} = \mathbf{A}\mathbf{z}$ *and* $\mathbf{V} = \mathbf{B}\mathbf{z}$. *Then* $\mathbf{U}$ *and* $\mathbf{V}$ *are independent if and only if* $cov(\mathbf{U}, \mathbf{V}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' = \mathbf{0}$.

Proof: e.g., Seber and Lee p25. By Theorem 2,

$$\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \mathbf{z} \sim N\left(\begin{bmatrix} \mathbf{A}\boldsymbol{\mu} \\ \mathbf{B}\boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' & \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' \\ \mathbf{B}\boldsymbol{\Sigma}\mathbf{A}' & \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}' \end{bmatrix} \right)$$

and by Theorem 3, $\mathbf{U}$ and $\mathbf{V}$ are independent if and only if $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' = \mathbf{0}$.

- We use the theorem by setting $\mathbf{z} = \mathbf{y}$, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$, $\mathbf{U} = \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\mathbf{V} = \hat{\boldsymbol{\varepsilon}} = \mathbf{M}\mathbf{y}$, so $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{B} = \mathbf{M} = \mathbf{I}_n - \mathbf{H}$.

- Thus, $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_n(\mathbf{I}_n - \mathbf{H})' = \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathbf{0}$.

- And so $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are independent.

Claim 3: Conditional Independence of RSS and RSS_r - RSS

- Since $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are independent, so too are $\hat{\boldsymbol{\beta}}$ and RSS = $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$.

- Finally, since $\hat{\boldsymbol{\beta}}$ and RSS are independent, so too are RSS - RSS_r (a continuous function of $\hat{\boldsymbol{\beta}}$) and RSS.

The F test statistic

We have established that

1. $\text{RSS}/\sigma^2 \sim \chi^2_{n-k}$

2. $(\text{RSS}_r - \text{RSS})/\sigma^2 \sim \chi^2_q$

3. RSS_r - RSS and RSS are independent.

Accordingly,

$$\boxed{\frac{(\text{RSS}_r - \text{RSS})/q}{\text{RSS}/(n-k)} = \frac{(\text{RSS}_r - \text{RSS})/q}{\hat{\sigma}^2} \sim F_{q,n-k}}$$

remembering that q is the number of linear restrictions being tested.

A More Compact Proof, using Properties of Orthogonal Projections

- Seber and Lee, Theorem 4.1(iv), or more generally, Theorem 4.3; some definitions of these terms appear at the end of these slides.

- Regression is an *orthogonal decomposition* of $\mathbf{y}$. That is, we have $\mathbf{y} = \hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}$. But $\hat{\mathbf{y}} \perp \hat{\boldsymbol{\varepsilon}}$, where the symbol "$\perp$" means "orthogonal to". How so?

- We know that $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. And $\hat{\boldsymbol{\varepsilon}} = \mathbf{M}\mathbf{y}$. We also know that $\hat{\mathbf{y}}$ and $\hat{\boldsymbol{\varepsilon}}$ are orthogonal: $(\mathbf{H}\mathbf{y})'(\mathbf{I}_n - \mathbf{H})\mathbf{y} = \mathbf{0}$.

- Hence, $\mathbf{H}$ is an *orthogonal projector*, decomposing $\mathbf{y}$ into two orthogonal components; or, more formally, decomposing $\mathbf{y} \in \mathbb{R}^n$ into two vectors that lie in orthogonal subspaces of $\mathbb{R}^n$.

- That is, $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \in \mathcal{C}(\mathbf{X})$, the *column space* of $\mathbf{X}$. $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{M}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y} \in \mathcal{C}(\mathbf{X})^{\perp}$, the *orthogonal complement* of $\mathcal{C}(\mathbf{X})$.

A More Compact Proof, using Properties of Orthogonal Projections

- An unrestricted model has the "hat matrix" $\mathbf{H}$ projecting $\mathbf{y} \in \mathbb{R}^n$ to $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \in \mathcal{C}(\mathbf{X})$.

- And $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$ projects $\mathbf{y} \in \mathbb{R}^n$ into the orthogonal complement $\mathbb{S}^{\perp}$.

- A *restricted* model imposes q linear restrictions on $\hat{\boldsymbol{\beta}}$ relative to the unrestricted model such that the "restricted hat matrix", $\mathbf{H}_r$, projects from $\mathbb{R}^n$ into $\mathbb{S}_r \subset \mathbb{S}$.

- Example: the simplest case is where the restricted model drops a predictor from the unrestricted model (i.e., imposes the constraint the corresponding element of $\hat{\boldsymbol{\beta}}$ is 0). The restrictive model is thus projecting $\mathbf{y}$ into a column space $\mathcal{C}(\mathbf{X}_r) \subset \mathcal{C}(\mathbf{X})$.

- And $\mathbf{M}_r = \mathbf{I}_n - \mathbf{H}_r$ projects from $\mathbb{S}_r^{\perp}$, the orthogonal complement of $\mathbb{S}_r$.

A More Compact Proof, using Properties of Orthogonal Projections

We also state the following properties of orthogonal projections:

1. Orthogonal projection matrices are symmetric and idempotent. For example, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}\mathbf{H} = \mathbf{H}'\mathbf{H} = \mathbf{H}'$.

2. Orthogonal projection matrices project from their image into their image (a consequence of idempotency). For example, $\mathbf{Hx} = \mathbf{x}, \forall\, \mathbf{x} \in \mathcal{C}(\mathbf{X})$.

3. Suppose $\mathbf{H}$ and $\mathbf{H}_r$ are both orthogonal projectors such that $\mathbf{H} : \mathbb{R}^n \to \mathbb{S}$ and $\mathbf{H}_r : \mathbb{R}^n \to \mathbb{S}_r$, with $\mathbb{S}_r \subset \mathbb{S}$. Then $\mathbf{H}\mathbf{H}_r = \mathbf{H}_r\mathbf{H} = \mathbf{H}_r$.
   The intuition here is that since image($\mathbf{H}_r$) = $\mathbb{S}_r \subset$ image($\mathbf{H}$) = $\mathbb{S}$, applying both projections always puts us in the "smaller" of the two spaces, $\mathbb{S}_r$, and the order in which we apply the projections doesn't matter.

Independence of Quadratic Forms

Another useful theorem:

Theorem 5. [Example 2.12, Seber and Lee (2003)] *Suppose* $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_n)$ *and* $\mathbf{A}$ *and* $\mathbf{B}$ *are symmetric, idempotent matrices, such that* $\mathbf{z}'\mathbf{A}\mathbf{z} \sim \chi^2$ *and* $\mathbf{z}'\mathbf{B}\mathbf{z} \sim \chi^2$ *(see Theorem 1). Then* $\mathbf{z}'\mathbf{A}\mathbf{z}$ *and* $\mathbf{z}'\mathbf{B}\mathbf{z}$ *are independent if and only if* $\mathbf{AB} = \mathbf{0}$.

Proof: Since $\mathbf{A}$ and $\mathbf{B}$ are symmetric and idempotent, $\mathbf{z}'\mathbf{A}\mathbf{z} = \mathbf{z}'\mathbf{A}'\mathbf{A}\mathbf{z}$ and similarly for $\mathbf{z}'\mathbf{B}\mathbf{z}$. By Theorem 2,

$$\begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \mathbf{z} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{A}'\mathbf{A} & \mathbf{A}'\mathbf{B} \\ \mathbf{B}'\mathbf{A} & \mathbf{B}'\mathbf{B} \end{bmatrix} \right).$$

Thus, by Theorem 4, $\mathbf{Az}$ and $\mathbf{Bz}$ are independent if and only if $\mathbf{A}'\mathbf{B} = \mathbf{AB} = \mathbf{0}$.

and finally...

- $\text{RSS} = \boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}$, $\text{RSS}/\sigma^2 \sim \chi^2_{n-k}$.

- $\text{RSS}_r - \text{RSS} = \boldsymbol{\varepsilon}'\mathbf{M}_r\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'(\mathbf{M}_r - \mathbf{M})\boldsymbol{\varepsilon}$. By the rank nullity theorem (Theorem 6), $\mathbf{M}_r - \mathbf{M}$ has rank $(n - k) - (n - k - q) = q$. In addition, $\mathbf{M}_r - \mathbf{M}$ is an orthogonal projector and so is symmetric and idempotent, and hence by Theorem 1, $(\text{RSS}_r - \text{RSS})/\sigma^2 \sim \chi^2_q$.

- By Theorem 5, RSS and RSS_r - RSS are independent if and only if $(\mathbf{I}_n - \mathbf{H})(\mathbf{H} - \mathbf{H}_r) = \mathbf{0}$. Checking this, we have $(\mathbf{I}_n - \mathbf{H})(\mathbf{H} - \mathbf{H}_r) = \mathbf{H} - \mathbf{H}_r - \mathbf{H}\mathbf{H} + \mathbf{H}\mathbf{H}_r = \mathbf{0}$, because $\mathbf{H}\mathbf{H}_r = \mathbf{H}_r$ (which we showed a few slides earlier).

- All this means that (again) we can state that

$$\boxed{\frac{(\text{RSS}_r - \text{RSS})/q}{\text{RSS}/(n-k)} = \frac{(\text{RSS}_r - \text{RSS})/q}{\hat{\sigma}^2} \sim F_{q,n-k}}$$

remembering that q is the number of linear restrictions being tested.

Interpreting F test statistics

- Under H_0, the restrictions are true, and the two models are the same.
- Thus, under H_0: $\text{RSS}_r = \text{RSS}$ and the numerator of the F test is zero.
- To the extent the restricted and unrestricted models diverge, $\text{RSS}_r > \text{RSS}$ and the numerator of the F test is positive.
- The further away F is from zero, the less plausible is H_0.
- We reject H_0 is favor of the unrestricted model when F crosses a (pre-specified) critical value, the 1 - α quantile of the $F_{q,n\text{-}k}$ distribution.
- In this sense, a lot like a χ^2 test, and indeed, if σ^2 was known, $(\text{RSS}_r - \text{RSS})/\sigma^2 \sim \chi^2_q$.

Testing Linear Restrictions on Parameters

Usually the linear restrictions we seek to test are simple *exclusion* restrictions: e.g.,

- that *all* $\mathbf{X}$ variables don't belong in the model (i.e., their coefficients are all jointly zero)
- that a particular *subset* of the parameters are zero or, in other words, that a subset of the independent variables don't belong in the model.
- that a single parameter is zero (n.b., the t test is a special case of the F test)

Testing Linear Restrictions on Parameters

$$\boxed{\frac{(\text{RSS}_r - \text{RSS})/q}{\text{RSS}/(n-k)} = \frac{(\text{RSS}_r - \text{RSS})/q}{\hat{\sigma}^2} \sim F_{q,n-k}}$$

- A large value of the F statistic means that the **change** in the goodness-of-fit between the two specifications is statistically significant.

- Note that $\text{RSS}_r \geq \text{RSS}$; by corollary, $r_r^2 \leq r^2$ (i.e., we never fit any worse by adding variables, and we never fit any better by dropping variables).

- RSS and RSS_r are random quantities (they vary in repeated sampling)

- *Intuition:* the F-distribution is how we assess whether the improved fit of the unrestricted model over the restricted model is statistically significant.

Testing Linear Restrictions on Parameters

The F test statistic can also be computed using the r^2 of the restricted (r) and unrestricted models (ur):

$$\boxed{F = \frac{(r^2_{ur} - r^2_r)/(\mathrm{df}_r - \mathrm{df}_{ur})}{(1 - r^2_{ur})/\mathrm{df}_{ur}}}$$

Testing Linear Restrictions on Parameters

Typical example: Testing for conditioning effects in a regression: e.g., *is the relationship between age and salary different for men and women*

Restricted:	$\text{salary}_i = \alpha_0 + \beta_0 \text{age}_i + \varepsilon_i$
Unrestricted:	$\text{salary}_i = \alpha_0 + \alpha_1 D_i + \beta_0 \text{age}_i + \beta_1 [D_i \times \text{age}_i] + \varepsilon_i$
D_i:	1 if ith observation is female, 0 otherwise
H_0:	$\alpha_1 = 0$ AND $\beta_1 = 0$

Note that $H_0 : \alpha_1 = 0$ can be tested with a t-statistic, as can $H_0 : \beta_1 = 0$. i.e., the possibility that there is merely a different intercept or a difference slope for females can be tested with a t-statistic, but we need the F-test to examine whether **both** are simultaneously true.

Implementation

- `anova()` function in `R`
- `linear.hypothesis` function in `library(car)`
- `ellipse` function in `library(car)`; fun teaching tool, but not very practical.

Examples

- "default" F-test produced by `summary.lm` in R; i.e., H_0: all slopes zero.
- faculty salary example, see homework 2 from 2004

“Odd” Examples

- It is possible to run a regression and have the slope coefficients be *individually* statistically significant, but to fail to reject the joint null hypothesis that the coefficients are jointly zero.

- Likewise, the converse: slope coefficients not statistically significant *individually*, but the F test lets us reject the null hypothesis that the coefficients are jointly zero.

Joint Confidence Regions for $\hat{\boldsymbol{\beta}}$; ALR3 §5.5

A joint confidence region for $\hat{\boldsymbol{\beta}}$ with confidence level α is a hyper-ellipsoid in $\mathbb{R}^k$ with surface

$$\hat{\boldsymbol{\beta}}' \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = k\,\hat{\sigma}^2\, F^{\alpha}_{k,n\text{-}k}$$

where

- $\hat{\boldsymbol{\beta}}$ is the k-by-1 vector of least squares estimates of $\boldsymbol{\beta}$

- $F^{\alpha}_{k,n\text{-}k}$ is the α critical values of the F distribution with k and n - k degrees of freedom (i.e., the 1 - α quantile of the $F_{k,n\text{-}k}$ distribution).

- $\hat{\boldsymbol{\beta}}$ lies at the center of the confidence region

- Not very practical (i.e., almost never reported in published research); but help illuminate some important conceptual issues. I.e., about the only time you'll ever see a joint confidence region for $\hat{\boldsymbol{\beta}}$ is in a statistics class.

Joint Confidence Regions for $\hat{\boldsymbol{\beta}}$; ALR3 §5.5

Consider the simple case of $k = 2$ (so we can visualize the confidence ellispe); see ALR3 Figure 5.3 (p109). With $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)'$, we have

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \begin{bmatrix}\hat{\boldsymbol{\beta}}_1 \ \hat{\boldsymbol{\beta}}_2\end{bmatrix} \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} \\
& \quad \text{(i.e., denoting } \mathbf{X}'\mathbf{X} \text{ as } \mathbf{L}\text{)} \\
&= \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 L_{11} + \hat{\boldsymbol{\beta}}_2 L_{21} & \hat{\boldsymbol{\beta}}_1 L_{12} + \hat{\boldsymbol{\beta}}_2 L_{22} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} \\
&= \left(\hat{\boldsymbol{\beta}}_1^2 L_{11} + \hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_2 L_{21} + \hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_2 L_{12} + \hat{\boldsymbol{\beta}}_2^2 L_{22}\right) \\
&= \hat{\boldsymbol{\beta}}_1^2 L_{11} + \hat{\boldsymbol{\beta}}_2^2 L_{22} + 2L_{12}\hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_2 \\
& \quad \text{(exploiting the fact that } \mathbf{X}'\mathbf{X} \text{ is a symmetric matrix and so } L_{12} = L_{21}\text{).}
\end{aligned}
$$

Joint Confidence Regions for $\hat{\boldsymbol{\beta}}$

Then the boundary of the α-level confidence ellipse for $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}}_1^2 L_{11} + \hat{\boldsymbol{\beta}}_2^2 L_{22} + 2L_{12}\hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_2 = 2\hat{\sigma}^2 F^{\alpha}_{2,n\text{-}k}.$$

- The quadratic form in $\hat{\beta}_1$ and $\hat{\beta}_2$ are why we get an ellipse: recall high school geometry definition of an ellipse as $ax^2 + by^2 + cxy = d$.

- Shape and orientation of a joint confidence ellipse depends on $\mathbf{X}'\mathbf{X}$ (sum of squares and cross-products for $\mathbf{X}$).

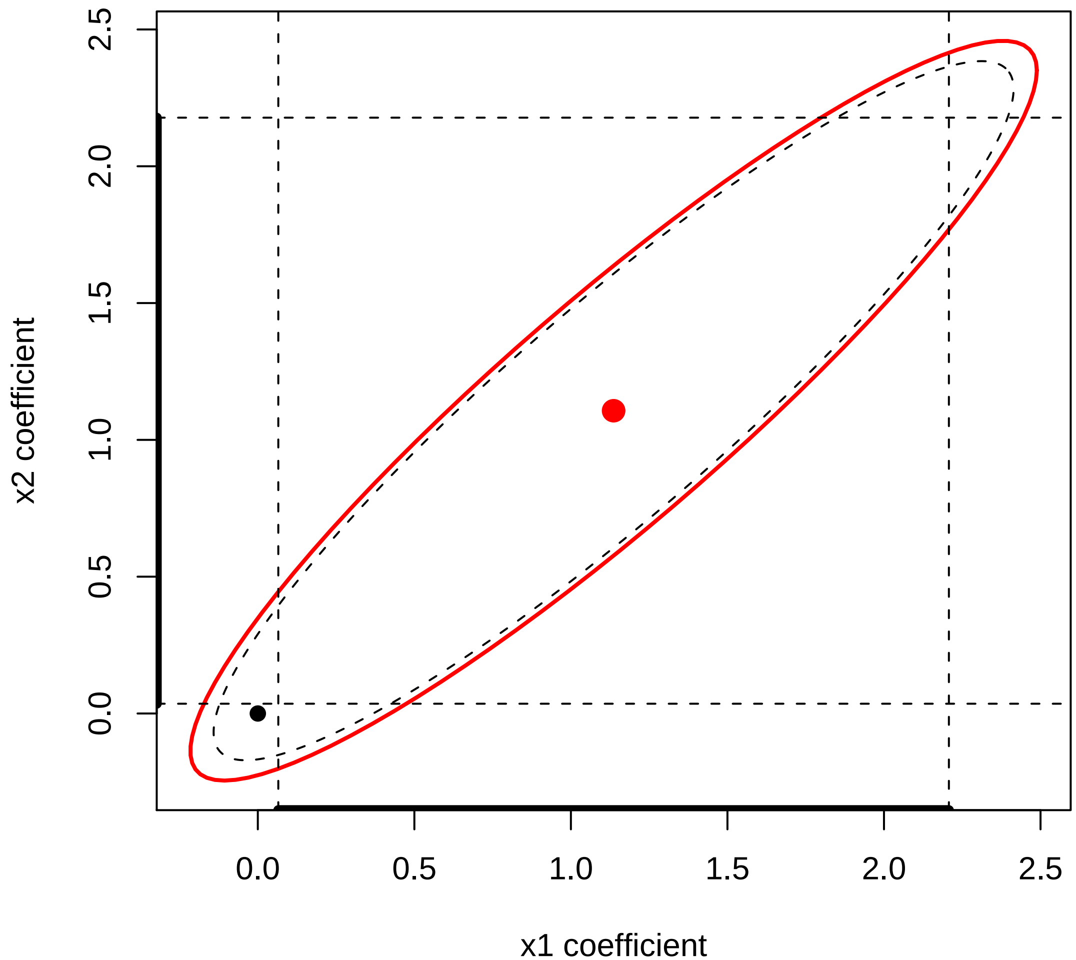
Joint Confidence Regions for $\hat{\boldsymbol{\beta}}$

- Positively correlated $\mathbf{X}$ imply negatively correlated $\hat{\boldsymbol{\beta}}$ and a joint confidence ellipse for $\hat{\boldsymbol{\beta}}$ that "points down" (the principal axis of the ellipse has a negative slope); negatively correlated $\mathbf{X}$ imply positively correlated $\hat{\boldsymbol{\beta}}$ and a joint confidence ellipse for $\hat{\boldsymbol{\beta}}$ that "points up" (the principal axis of the ellipse has a positive slope).

- Projections of a k-dimensional confidence hyper-ellipsoid onto the j-th reference axis will not equal the confidence interval for $\hat{\beta}_j$. See text.

Contrived, Unusual Case

```
Call:
lm(formula = y ~ x, x = T, y = T)
Residuals:
    Min      1Q  Median      3Q     Max
-1.9362 -0.5938  0.0459  0.4798  2.3378
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1777     0.2011   0.884   0.3847
x1            1.1363     0.5219   2.177   0.0384 *
x2            1.1066     0.5219   2.120   0.0433 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.102 on 27 degrees of freedom
Multiple R-Squared: 0.1517,     Adjusted R-squared: 0.0889
F-statistic: 2.415 on 2 and 27 DF,  p-value: 0.1084
```



2.5
2.0
1.5
1.0
0.5
0.0
x2 coefficient
0.0
0.5
1.0
1.5
2.0
2.5
x1 coefficient

Joint Confidence Ellipses and Multicollinearity: a connection

Highly correlated **X** variables produce

- at least in two dimensions, confidence ellispes that are quite elongated and tilted either up or down

- regression results that tend not to be informative about the coefficients on correlated **X** variables (i.e., large estimated standard errors)

- regression results that tend to be informative about the effects of a *linear combinations* of correlated **X** variables

- In the previous graph, x_1 and x_2 are highly negatively correlated (at -.92). These data are much more informative about β_1 - β_2 than it is about β_1, β_2 or $\beta_1 + \beta_2$.

Joint Confidence Ellipses and Multicollinearity: a connection

```
> vcov(badfoo)
              (Intercept)            x1            x2
(Intercept)  4.044851e-02 -7.219998e-18 -7.501402e-18
x1          -7.219998e-18  2.724172e-01  2.506239e-01
x2          -7.501402e-18  2.506239e-01  2.724172e-01
> v <- vcov(badfoo)
> sqrt(v[2,2] + v[3,3] - 2*v[2,3])   ## std err for beta2 - beta3
[1] 0.2087744
> sqrt(v[2,2] + v[3,3] + 2*v[2,3])   ## std err for beta2 + beta3
[1] 1.022782
```

n.b., the large covariance term between $\hat{\beta}_2$ and $\hat{\beta}_3$

Some Definitions

Definition 1. [Vector Space] *A vector space is a nonempty set $\mathcal{V}$ of vectors closed under addition and scalar multiplication.*

Definition 2. [Span] *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are vectors in $\mathbb{R}^n$. The span of $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is the set of linear combinations of these vectors:*

$$span(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \{\mathbf{y} \in \mathbb{R}^n : \mathbf{y} = \sum_i \alpha_i \mathbf{x}_i\}.$$

Definition 3. [Linear Dependence] *A set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is said to be linearly dependent if there exists a non-zero linear combination*

$$\alpha_1 \mathbf{x}_1 + \cdots + \alpha_k \mathbf{x}_k = \mathbf{0}.$$

If a set of vectors is not linearly dependent, it is said to be linearly independent.

Some Definitions

Definition 4. [Basis of a Vector Space] *A basis for a vector space $\mathcal{V}$ is a set of linearly independent vectors that span $\mathcal{V}$.*

Definition 5. [Dimension of a Vector Space] *The dimension of $\mathcal{V}$, denoted $\dim(\mathcal{V})$, is the number of vectors in any basis for $\mathcal{V}$.*

Definition 6. [Column Space] *The column space of a matrix $\mathbf{X}$, $\mathcal{C}(\mathbf{X})$, is the vector space spanned by the columns of $\mathbf{X}$.*

Definition 7. [Rank of a Matrix] *The rank of a matrix $\mathbf{X}$ is the dimension of its column space, $\mathcal{C}(\mathbf{X})$*

Definition 8. [Orthogonal Complement] *If $\mathbf{X}$ is a $n \times p$ matrix, the set*

$$\mathbf{X}^{\perp} = \{\mathbf{y} \in \mathbb{R}^n : \mathbf{X}'\mathbf{y} = \mathbf{0}\}$$

is the orthogonal complement of $\mathbf{X}$.

Some Definitions

Definition 9. [Null Space] *The null space of a matrix* $\mathbf{A}$ *is the set of vectors*

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\}$$

Definition 10. [Nullity] *The nullity of a matrix* $\mathbf{A}$ *is the dimension of* $\mathcal{N}(\mathbf{A})$.

Rank Nullity Theorem

Theorem 6. [Rank Nullity] *If* $\mathbf{A}$ *is a m-by-n matrix with rank r and nullity s then* $r + s = n$.

Corollary: If $\mathbf{A}^{\perp}$ is the null space of $\mathbf{A}$, and $\mathbf{A}$ has rank p and n columns, then $\dim(\mathbf{A}^{\perp}) = n - p$.

Example: $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ projects from $\mathcal{C}(\mathbf{X})$ to $\mathcal{C}(\mathbf{X})$ and so $\mathcal{C}(\mathbf{H}) = \mathcal{C}(\mathbf{X})$, implying that $\text{rank}(\mathbf{H}) = \text{rank}(\mathbf{X}) = k$. Conversely, $\mathbf{M} = \mathbf{I} - \mathbf{H}$ projects to $\mathcal{C}(\mathbf{X})^{\perp} = \mathcal{C}(\mathbf{H})^{\perp} = \mathcal{N}(\mathbf{H}')$ (Seber and Lee, Proposition B.2.1), but since $\mathbf{H}' = \mathbf{H}$, we have $\mathcal{C}(\mathbf{X})^{\perp} = \mathcal{N}(\mathbf{H})$. Theorem 6 says that the dimension of the null space of $\mathbf{H}$ is $n - k$; since (by definition) $\mathbf{M}$ projects from $\mathcal{C}(\mathbf{X})^{\perp}$ to $\mathcal{C}(\mathbf{X})^{\perp}$, $\mathcal{C}(\mathbf{M}) = \mathcal{N}(\mathbf{H})$, and so we deduce that the rank of $\mathbf{M}$ is $n - k$.