# Quadratic Forms and Geometry of Least Squares

## 1 Basic Geometry

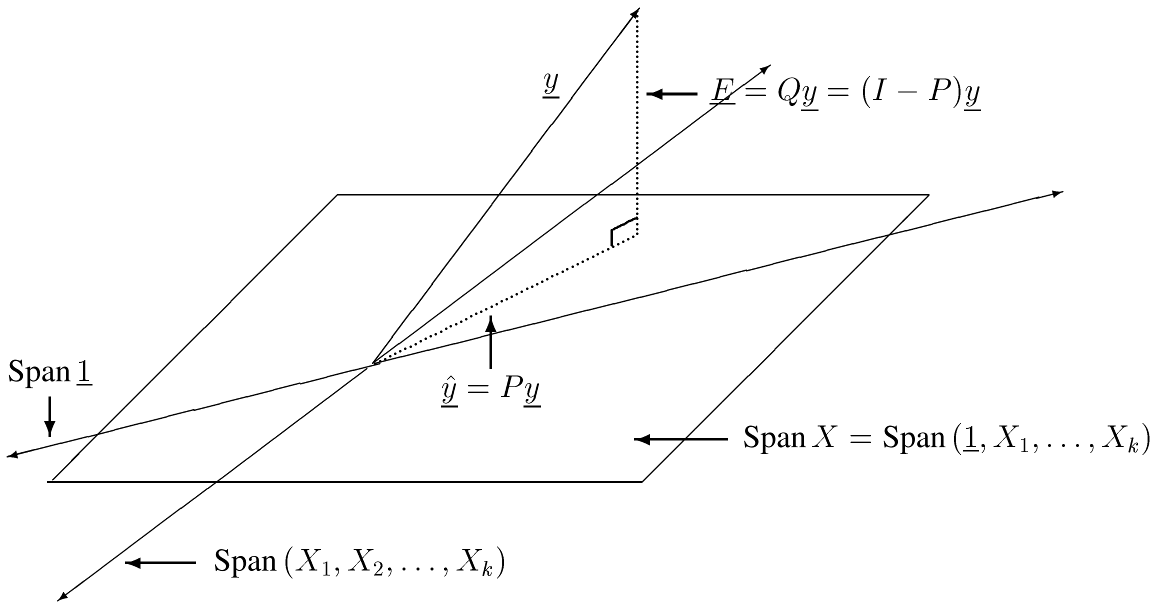


Figure 1: Geometry of Linear Regression

We have seen that fitting the linear regression model

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \quad \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I_{n\times n}),$$

where

$$X = \begin{bmatrix} | & | & | & & | \\ \underline{1} & \underline{X}_1 & \underline{X}_2 & \cdots & \underline{X}_k \\ | & | & | & & | \end{bmatrix}, \quad \underline{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_k),$$

by least squares was equivalent to decomposing the vector $\underline{y} \in \mathcal{R}^n$ into orthogonal components

$$\underline{y} = P\underline{y} + Q\underline{y} = \hat{\underline{y}} + \underline{E}$$

where

$$\begin{aligned} \hat{\underline{y}} &= P\underline{y} \in ColSpace(P) = ColSpace(X) \\ \underline{E} &= Q\underline{y} \perp ColSpace(P) \end{aligned}$$

and

$$\begin{aligned}
P &= X(X^TX)^{-1}X^T \\
Q &= I_{n\times n} - P \\
\underline{\hat\beta} &= (X^TX)^{-1}X^T\underline{y} \\
\underline{\hat y} &= X\underline{\hat\beta} && = P\underline{y} \\
\underline{E} &= \underline{y} - \underline{\hat y} && = Q\underline{y}
\end{aligned}$$

since

$$\begin{aligned}
P^2 &= P;\ P^T = P && (idempotent, symmetric) \\
Q^2 &= Q;\ Q^T = Q && (idempotent, symmetric) \\
PQ &= QP = \mathbf{0} && (orthogonal) \\
P + Q &= I && (complete) \\
\text{and } PX &= X && (projection\ onto\ ColSpace(X))
\end{aligned}$$

This gives rise to the picture in Figure 1. It is not hard to show for example that the Pythagorean Theorem holds,

$$||\underline{y}||^2 = \underline{y}^T\underline{y} = [(P+Q)\underline{y}]^T[(P+Q)\underline{y}] = (\underline{y}^TP+\underline{y}^TQ)(Py+Qy) = \underline{y}^TP\underline{y}+\underline{y}^TQ\underline{y} = ||\underline{\hat y}||^2+||\underline{E}||^2$$

and it turns out, according to a theorem called "Cochran's Theorem" which we shall see below, that $\underline{y}^TP\underline{y}$ and $\underline{y}^TQ\underline{y}$ are statistically independent of one another. Much of the theory of $F$-statistic testing in linear models is built on elaborations of this Pythagorean Theorem, and on the corresponding statistical indepedence facts about the component sums of squares.

# 2 Some Results on Quadratic Forms

Let $\underline{y} \in \mathcal{R}^n$, and suppose $A$ is an $n \times n$ matrix.

**Definition 1** *$\underline{y}^TA\underline{y}$ is a* quadratic form *if*

1. $A^T = A$ *(symmetric)*
2. $\underline{y}^TA\underline{y} \geq 0,\ \forall \underline{y} \in \mathcal{R}^n$ *(positive semidefinite)*

*In this case, $A$ is the* defining matrix *of $\underline{y}^TA\underline{y}$.*

**Definition 2** *$A$ is a* projection matrix *if $A^2 = A$ (we say $A$ is* idempotent*).*

Note that if $A$ is a projection matrix then $A(I_{n\times n} - A) = \underline{0}$, and every $\underline{y} \in \mathcal{R}^n$ has a unique decomposition

$$\underline{y} = A\underline{y} + (I - A)\underline{y} = \underline{y}_A + \underline{y}_{A^\perp}$$

into a vector $\underline{y}_A \in ColSpace(A)$ and a vector $\underline{y}_{A^\perp} \perp ColSpace(A)$. If $A$ is also symmetric, then

$$\text{trace}(A) = \text{rank}(A) = \text{dim}[ColSpace(A)].$$

Note that, therefore, trace$(A)$ is the "number of dimensions" that $A\underline{y}$ can wander around in, and $\underline{y}^TA\underline{y} = (A\underline{y})^T(A\underline{y}) = ||A\underline{y}||^2$ is the squared length of $A\underline{y}$.

**Facts:**

1. If $\underline{y}$ is a random vector with $E[\underline{y}] = \underline{\mu}$, Var $(\underline{y}) = V$, and if $\underline{y}^T A \underline{y}$ is a quadratic form, then

$$E[\underline{y}^T A \underline{y}] = \text{ trace } AV + \underline{\mu}^T A \underline{\mu}$$

2. If $y_i \sim N(\mu_i, \sigma^2)$, independent, and $A$ is symmetric and idempotent, then $\underline{y}^T A \underline{y}/\sigma^2$ has a *noncentral* $\chi^2$ *distribution on* $k$ *degrees of freedom*, where $k =$ rank $A$. The quantity $\underline{\mu}^T A \underline{\mu}/2\sigma^2$ is called the *noncentrality parameter* of the distribution. Note: here, $V =$ Var $\underline{y} = \sigma^2 I$.

**Theorem 1** (Cochran's Theorem, Scheffé, 1959, pp 419–421)
*Let $y_i \sim N(\mu_i, 1), i = 1 \ldots n$, independent, and suppose*

$$\sum_{i=1}^{n} y_i^2 = \underline{y}^T A_1 \underline{y} + \underline{y}^T A_2 \underline{y} + \ldots + \underline{y}^T A_s \underline{y}$$

*where each $\underline{y}^T A_j \underline{y}$ is a* quadratic form, *with rank $A_j = n_j$. <u>Then:</u>*

*(a) Each $A_j^2 = A_j$* (<u>idempotent</u>, so rank $A_j =$ trace $A_j$)*; and $A_i A_j = \mathbf{0}$ for all $i \neq j$ ($A_i$ and $A_j$ are* orthogonal*).*

*(b) The quadratic forms $\underline{y}^T A_j \underline{y}$ are distributed as independent, noncentral $\chi^2$'s on $n_j$ degrees of freedom.*

*<u>if and only if</u> $n_1 + n_2 + \ldots + n_s = n$.*

**Corollary 1** *If $y_i \sim N(\mu_i, \sigma^2)$, the theorem still holds but now $\underline{y}^T A_j \underline{y}/\sigma^2$ will be independent noncentral $\chi^2_{n_j}$'s.*

Cochran's Theorem or its corollary tell us immediately, for example, that since

$$\sum_i y_i^2 = ||\underline{y}||^2 = \underline{y}^T P \underline{y} + \underline{y}^T Q \underline{y}$$

and since trace$(P)$ + trace$(Q)$ = trace$(P + Q)$ = trace$(I) = n$, then $\underline{y}^T P \underline{y}$ and $\underline{y}^T Q \underline{y}$ are (proportional to) independent $\chi^2$'s on trace$(P)$ and trace$(Q)$ degrees of freedom.

# 3 Overall ANOVA for regression with an intercept

Let's continue now with the model

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \quad \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I_{n\times n}), \tag{1}$$

where

$$X = \begin{bmatrix} | & | & | & & | \\ \underline{1} & \underline{X}_1 & \underline{X}_2 & \cdots & \underline{X}_k \\ | & | & | & & | \end{bmatrix}, \quad \underline{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_k), \quad k' = k+1,$$

with projection matrix $P$ and $Q = I - P$.

We wish to compare this model with the simpler model

$$\underline{y} = \underline{1}\beta_0 + \underline{\varepsilon} \tag{2}$$

of regression onto just the intercept $\beta_0$. You can verify that the projection matrix for this regression is $J/n$ where $J$ is the $n \times n$ matrix of all 1's. It is easy to see that

$$\begin{array}{rcccccc} \sum_{i=1}^n y_i^2 & = & \|\underline{y}\|^2 & = & \underline{y}^T I \, \underline{y} & & \\ & = & \underline{y}^T(I-P)\underline{y} & + & \underline{y}^T(P-J/n)\underline{y} & + & \underline{y}^T J/n \, \underline{y} \\ & = & \|Q\underline{y}\|^2 & + & \|(P-J/n)\underline{y}\|^2 & + & \|(J/n)\underline{y}\|^2 \\ & = & RSS & + & RegSS & + & n(\bar{y})^2 \end{array}$$

and we can readily verify[1] that the matrices $A_1 = (I - P)$, $A_2 = (P - J/n)$, and $A_3 = J/n$, are symmetric ($A_i^T = A$), idempotent ($A_i^2 = A_i$) and orthogonal ($A_i A_j = \mathbf{0}$), and $\text{trace}(A_1) + \text{trace}(A_2) + \text{trace}(A_3) = n$. So, by *Cochran's Theorem* and its corollary, the quadratic forms

$$\underline{y}^T(J/n)\underline{y}/\sigma^2, \quad \underline{y}^T(P-J/n)\underline{y}/\sigma^2, \quad \underline{y}^T Q\underline{y}/\sigma^2$$

are independent noncentral $\chi^2$'s. Moreover you can easily show that

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum y_i^2 - n(\bar{y})^2 = \underline{y}^T(I - J/n)\underline{y}$$

so the RegSS and RSS terms are in fact independent $\chi^2$'s in the overall ANOVA table

| Source | df | SS |
|---|---|---|
| Regression | $k'-1$ | $RegSS = \underline{y}^T(P-J/n)\underline{y}$ |
| Residual | $n-k'$ | $RSS = \underline{y}^T Q\underline{y}$ |
| Total | $n-1$ | TSS = $\underline{y}^T(I-J/n)\underline{y}$ |

The geometry for this ANOVA table is given in Figure 2.

The overall $F$-statistic for the regression is

$$F = \frac{(TSS - RSS)([(n-1)-(n-k')])}{RSS/(n-k')} = \frac{RegSS/(k'-1)}{RSS/(n-k')} = \frac{RegSS/k}{RSS/(n-(k+1))},$$

and tests $H_0$ : eq. (2) holds, vs. $H_1$ : equation (1) holds. We note that

[1] Use the facts that $P$ and $J/n$ are the projection matrices for the regressions (1) and (2) and the fact that $\underline{1}$ is a column of $X$, so that $(I-P)(J/n) = \mathbf{0}$.

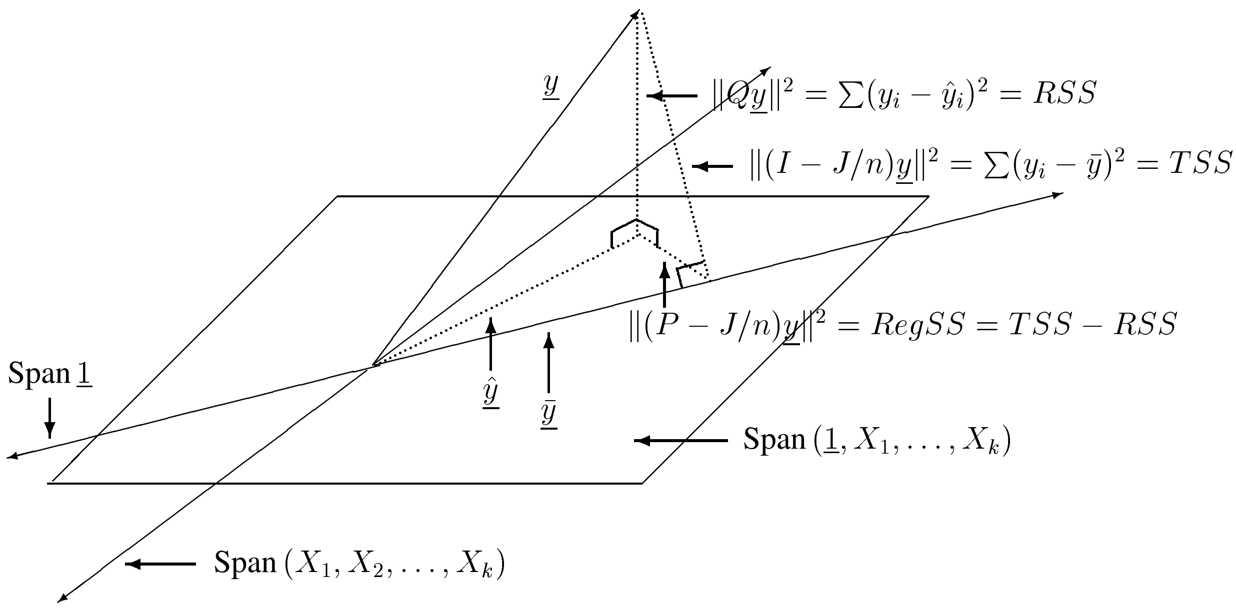


Figure 2: Geometry of the overall ANOVA table in linear regression

- $RegSS/\sigma^2 = \underline{y}^T(P - J/n)\underline{y}/\sigma^2 \;\sim$ noncentral $\chi^2_{k-1}$ :

$$\begin{aligned} df &= \text{rank } (P - J/n) = \text{ trace } (P - J/n) = \text{ trace } P - \text{ trace } (J/n) \\ &= \text{trace } \left[X(X^TX)^{-1}X^T\right] - 1 = \text{ trace } (X^TX(X^TX)^{-1}) - 1 \\ &= \text{trace } I_{k' \times k'} - 1 \;=\; k' - 1 \;=\; k \end{aligned}$$

$$\begin{aligned} E\left[\underline{y}^T(P - J/n)\underline{y}\right] &= \text{trace } \left[(P - J/n)\sigma^2 I\right] + (X\underline{\beta})^T(P - J/n)(X\underline{\beta}) \\ &= \sigma^2(k' - 1) + \underline{\beta}^T X^T(P - J/n)X\underline{\beta} \\ &= \sigma^2(k' - 1) + \underline{\beta}^T X^T(I - J/n)X\underline{\beta} \\ &\quad \text{since } PX = X. \end{aligned}$$

- $RSS/\sigma^2 = \underline{y}^T Q \underline{y} \;\sim$ noncentral $\chi^2_{n-k'}$ :

$$\begin{aligned} df &= \text{rank } Q = \text{ trace } Q = \text{ trace } (I - P) = \text{ trace } I - \text{ trace } P \\ &= n - k' \;=\; n - (k+1) \end{aligned}$$

$$\begin{aligned} E\left[\underline{y}^T Q \underline{y}\right] &= \text{trace } \left[Q(\sigma^2 I)\right] + (X\underline{\beta})^T Q(X\underline{\beta}) \\ &= \sigma^2(n - k') + \underline{\beta}^T X^T Q X \underline{\beta} \\ &= \sigma^2(n - k') \;=\; \sigma^2(n - (k+1)) \end{aligned}$$

(Note: $\underline{\beta}^T X^T Q X \underline{\beta} = 0$ since $QX = 0$.)

After some algebra we see that the ratio of the expected values in the $F$ statistic is

$$\frac{E\left[RegSS/(k'-1)\right]}{E\left[RSS/(n-k')\right]} = 1 + \left[\underline{\beta}^T X^T (I - J/n) X \underline{\beta}\right] / (k'-1)\sigma^2 \geq 1$$

The thing that prevents this from being exactly 1 is $\underline{\beta}^T X^T (I - J/n) X \underline{\beta}$. If you multiply out $X^T (I - J/n) X$ you will see[2] that

$$\begin{aligned}
\underline{\beta}^T X^T (I - J/n) X \underline{\beta} &= (\beta_0, \beta_1 \cdots \beta_k) \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & \text{stuff} & \ldots & \text{stuff} \\ \vdots & \vdots & & \vdots \\ 0 & \text{stuff} & \ldots & \text{stuff} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \\
&= (\beta_1 \cdots \beta_k) \begin{bmatrix} \text{stuff} & \ldots & \text{stuff} \\ \vdots & & \vdots \\ \text{stuff} & \ldots & \text{stuff} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}
\end{aligned}$$

which will equal zero when $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is true.

**Distribution of $F$ = $\frac{RegSS/(k'-1)}{RSS/(n-k')}$ under $H_0$:**

According to Cochran's theorem, $F$ is a ratio of independent *noncentral* $\chi^2$'s divided by their respective degrees of freedom. If they were *central* $\chi^2$'s (noncentrality parameter $= 0$), $F$ would follow the $F$ distribution we always use.

But they *are* central $\chi^2$'s:

- Noncentrality param. of $RSS$ is $\frac{1}{2\sigma^2}(X\underline{\beta})^T Q (X\underline{\beta}) = \underline{\beta}^T X^T Q X \underline{\beta} = 0$ since $QX = \mathbf{0}$
- Noncentrality param. of $RegSS$ is $\frac{1}{2\sigma^2}\underline{\beta}^T X^T (I - J/n) X \underline{\beta} = 0$ under $H_0$ (See above!).

**Distribution of $F$ under** $H_1$:

The numerator $\chi^2$, $RegSS/\sigma^2$, now has noncentrality parameter

$$\frac{1}{2\sigma^2}\,\underline{\beta}^T X^T (I - J/n) X \underline{\beta} \geq 0$$

The larger this parameter, the easier $H_1$ is to detect.

[2] *Hint*: Apply the formula $\sum_{i=1}^n x_i^2 - n(\bar{x})^2 = \underline{x}^T (I - J/n) \underline{x}$ to situations where $\underline{x}^T$ and $\underline{x}$ represent a column of $X$, and recall that the first column of $X$ is all 1's.

# 4 Using sums of squares to test nested models

*Model:* $\underline{y} = X\underline{\beta} + \underline{\varepsilon}\,, \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I)$ with some small changes of notation: Let $k' = k_1 + k_2$, and

$$X = \begin{bmatrix} | & | & & | \\ \underline{X}_1 & \underline{X}_2 & \cdots & \underline{X}_{k'} \\ | & | & & | \end{bmatrix} = \Bigg[\underbrace{\begin{matrix} | & | & & | \\ \underline{X}_1 & \underline{X}_2 & \cdots & \underline{X}_{k_1} \\ | & | & & | \end{matrix}}_{X_1} \underbrace{\begin{matrix} | & | & & | \\ \underline{X}_{k_1+1} & \underline{X}_{k_1+2} & \cdots & \underline{X}_{k_1+k_2} \\ | & | & & | \end{matrix}}_{X_2}\Bigg] = [\; X_1 \quad X_2 \;]$$

and

$$\underline{\beta}^T = (\beta_1 \cdots \beta_{k'}) = \Big(\underbrace{\beta_1, \ \beta_2, \ \cdots, \ \beta_{k_1},}_{\underline{\beta}_1^T,} \underbrace{\beta_{k_1+1}, \ \beta_{k_1+2}, \ \cdots, \ \beta_{k_1+k_2}}_{\underline{\beta}_2^T}\Big) = (\;\underline{\beta}_1^T, \quad \underline{\beta}_2^T\;)$$

so that

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} = X_1\underline{\beta}_1 + X_2\underline{\beta}_2 + \underline{\varepsilon}$$

(We could treat "regression with an intercept" again by making the column $\underline{X}_1$ just a column of 1's.)

We want to compare the two models

| | |
|---|---|
| SMALL: $\underline{y} = X_1\underline{\beta}_1 + \underline{\varepsilon}$ | Hat matrix $P_1 = X_1(X_1^T X_1)^{-1}X_1^T$ |
| LARGE: $\underline{y} = X_1\underline{\beta}_1 + X_2\underline{\beta}_2 + \varepsilon$ | Hat matrix $P_{12} = X(X^T X)^{-1}X^T$ |

$$\begin{aligned} RSS(\underline{\beta}_1, \underline{\beta}_2) = RSS_12 &= \textstyle\sum_{i=1}^n (y_i - \hat{y}_{i(12)})^2 \\ &= (\underline{y} - \underline{\hat{y}}_{(12)})^T(\underline{y} - \underline{\hat{y}}_{(12)}) \\ &= [(I - P_{12})\underline{y}]^T[(I - P_{12})\underline{y}] \\ &= \underline{y}^T(I - P_{12})\,\underline{y} = \underline{y}^T Q_{12}\underline{y} && (Q_{12} = I - P_{12}) \\ \text{Similarly} \\ RSS(\underline{\beta}_1) = RSS_1 &= \textstyle\sum_{i=1}^n (y_i - y_{i(1)})^2 = \underline{y}^T Q_1 \underline{y} && (Q_1 = I - P_1) \end{aligned}$$

We want a test for $H_0 : \underline{\beta}_2 = \underline{0}$ vs. $H_1 : \underline{\beta_2}$ free. [See Figure 3.] This is a test for

$$\begin{aligned} H_0 : \underline{y} &= X_1\underline{\beta}_1 + \underline{\varepsilon} \\ \text{vs.}\quad H_1 : \underline{y} &= X_1\underline{\beta}_1 + X_2\underline{\beta}_2 + \underline{\varepsilon} \end{aligned}$$

Clearly we can write, thinking of the smaller model,

$$\begin{aligned} \textstyle\sum y_i^2 &= \textstyle\sum \hat{y}_{i(1)}^2 + \left(\sum y_i^2 - \sum \hat{y}_{i(1)}^2\right) \\ &= \underline{y}^T P_1 \underline{y} + \underline{y}^T I \underline{y} - \underline{y}^T P_1 \underline{y} \\ &= \underline{y}^T P_1 \underline{y} + \underline{y}^T (I - P_1)\underline{y} \\ &= \underline{y}^T P_1 \underline{y} + \underline{y}^T Q_1 \underline{y} \\ &= \underline{y}^T P_1 \underline{y} + \underline{y}^T (Q_1 - Q_{12})\underline{y} + \underline{y}^T Q_{12}\underline{y} \\ &= \underline{y}^T P_1 \underline{y} + (RSS_1 - RSS_{12}) + RSS_{12} \\ &= \underline{y}^T P_1 \underline{y} + SS(\underline{\beta}_2 | \underline{\beta}_1) + RSS_{12} \end{aligned}$$

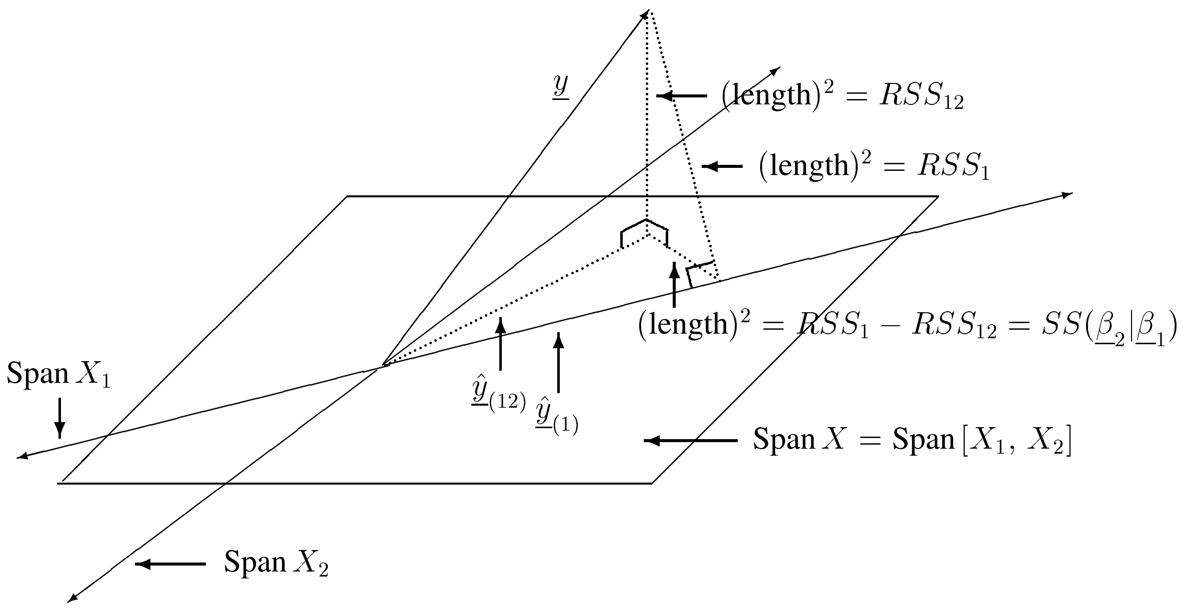


Figure 3: Geometry of the $F$-test comparing the nested models $X = [X_1,\, X_2]$ and $X_1$

where the third-to-last inequality follows from $Q_1 = (Q_1 - Q_{12}) + Q_{12}$ and the fact that (check!) $Q_{12}(Q_1 - Q_{12}) = 0$.

Now, if $SS(\underline{\beta}_2|\underline{\beta}_1) = \underline{y}^T(Q_1 - Q_{12})\underline{y}$:

(1) is a quadratic form (positive-semidefinite)?

(2) has noncentrality parameter $= 0$ under $H_0$?

then according to Cochran's theorem, the $F$-statistic

$$F = \frac{SS(\underline{\beta}_2|\beta_1)/\ \text{trace}\ [Q_1 - Q_{12}]}{RSS(\underline{\beta}_1, \underline{\beta}_2)/\ \text{trace}\ Q_{12}}$$

can be used to test $H_0 : \underline{\beta}_2 = 0$ vs. $H_1 : (H_0)^c$. In order to verify (1) and (2) for $SS(\underline{\beta}_2|\underline{\beta}_1) = \underline{y}^T(Q_1 - Q_{12})\underline{y}$ we will *reparametrize* the *larger* models to make $Q_{12}$ easier to calculate. The idea is similar to the idea that $\hat{\underline{\beta}}_2$ can be obtained by regressing the residuals $Q_1\underline{y}$ onto the residuals $Q_1X_2$ that we saw empirically earlier in the course.

Define

$$X_{2\cdot 1} \;=\; Q_1X_2 \;\;=\; X_2 - \underbrace{\overbrace{X_1\left(X_1^TX_1\right)^{-1}X_1^T}^{P_1}X_2}_{A}$$

Clearly $X_{2\cdot 1}$ is the *residuals* of (each column of) $X_2$ after regressing on $X_1$.

- $X_{2\cdot 1}$ is the part of $X_2$ unexplained by regression on $X_1$.
- $X_{2\cdot 1}$ is orthogonal to $X_1 : X_1^T X_{2\cdot 1} = \mathbf{0}$.

With $X_{2\cdot 1}$ and $A = (X_1^T X_1)^{-1} X_1^T X_2$ we can write

$$
\begin{aligned}
\underline{y} &= X_1 \underline{\beta}_1 + X_2 \underline{\beta}_2 + \underline{\varepsilon} \;=\; X_1 \underline{\beta}_1 + (X_1 A + X_2 - X_1 A)\underline{\beta}_2 + \underline{\varepsilon} \\
&= X_1 \underbrace{(\underline{\beta}_1 + A\underline{\beta}_2)} \;+\; \underbrace{(X_2 - X_1 A)} \underline{\beta}_2 \;+\; \underline{\varepsilon} \\
&= \quad X_1 \quad \underline{\alpha}_1 \qquad + \qquad X_{2\cdot 1} \quad \underline{\beta}_2 \quad + \; \underline{\varepsilon}
\end{aligned}
$$

So the larger model is *identical* to

$$
\underline{y} = X_1 \underline{\alpha}_1 + X_{2\cdot 1} \underline{\beta}_2 + \underline{\varepsilon}
$$

with model matrix $\tilde{X} = [X_1 \; X_{2\cdot 1}]$ instead of $X = [X_1 \; X_2]$. Therefore

$$
\begin{aligned}
\underline{y}^T Q_{12} \underline{y} &= RSS\left(\underline{\beta}_1, \underline{\beta}_2\right) = RSS\left(\underline{\alpha}_1, \underline{\beta}_2\right) = \underline{y}^T \left[ I - \tilde{X} \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \right] \underline{y} \\
&= \underline{y}^T \left\{ I - [X_1 \; X_{2\cdot 1}] \begin{bmatrix} X_1^T X_1 & X_1^T X_{2\cdot 1} \\ X_{2\cdot 1}^T X_1 & X_{2\cdot 1}^T X_{2\cdot 1} \end{bmatrix}^{-1} \begin{bmatrix} X_1^T \\ X_{2\cdot 1}^T \end{bmatrix} \right\} \underline{y} \\
&= \underline{y}^T \left\{ I - [X_1 \; X_{2\cdot 1}] \begin{bmatrix} \left( X_1^T X_1 \right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left( X_{2\cdot 1}^T X_{2\cdot 1} \right)^{-1} \end{bmatrix} \begin{bmatrix} X_1^T \\ X_{2\cdot 1}^T \end{bmatrix} \right\} \underline{y} \\
&= \underline{y}^T \left( I - P_1 - P_{2\cdot 1} \right) \underline{y}
\end{aligned}
$$

Hence

$$
\begin{aligned}
\underline{y}^T Q_{12} \underline{y} &= \underline{y}^T \left( I - P_1 - P_{2\cdot 1} \right) \underline{y} \\
&= \underline{y}^T Q_1 \underline{y} - \underline{y}^T P_{2\cdot 1} \underline{y}
\end{aligned}
$$

Therefore

$$
SS\left(\underline{\beta}_2 | \underline{\beta}_1\right) = \underline{y}^T \left( Q_1 - Q_{12} \right) \underline{y} = \underline{y}^T P_{2\cdot 1} \underline{y}
$$

is a quadratic form with defining matrix $P_{2\cdot1} = X_{2\cdot1}\left(X_{2\cdot1}^T X_{2\cdot1}\right)^{-1} X_{2\cdot1}^T$. (It is positive definite because

$$\underline{y}_1^T P_{2\cdot1}\underline{y} = \sum \hat{y}_{i(2\cdot1)}^2 \geq 0$$

from the model $\underline{y} = X_{2\cdot1}\tilde{\underline{\beta}} + \underline{\varepsilon}$).

Now, since $P_{2\cdot1}$ projects onto $X_{2\cdot1}$ which is *orthogonal* to $X_1$, it follows that $P_{2\cdot1}X_1 = \mathbf{0}$. Therefore,

$$\begin{aligned}
& E[SS(\underline{\beta}_2|\underline{\beta}_1)] = E\left[\underline{y}^T P_{2\cdot1}\underline{y}\right] \\
= \quad & \text{trace } (P_{2\cdot1}\sigma^2 I) + \left([X_1\ X_2]\,\underline{\beta}\right)^T P_{2\cdot1}\left([X_1\ X_2]\,\underline{\beta}\right) \\
= \quad & \sigma^2 \text{ trace } \left[X_{2\cdot1}^T X_{2\cdot1}\left(X_{2\cdot1}^T X_{2\cdot1}\right)^{-1}\right] + \left[\underline{\beta}_1^T\ \underline{\beta}_2^T\right]\begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix} P_{2\cdot1}\,[X_1\ X_2]\begin{bmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \end{bmatrix} \\
= \quad & \sigma^2 k_2 + \underline{\beta}_2^T X_2^T P_{2\cdot1} X_2 \underline{\beta}_2
\end{aligned}$$

This last term controls the noncentrality parameter, and equals 0 under $H_0 : \underline{\beta}_2 = \underline{0}$.

Once again, $E\left[RSS\left(\underline{\beta}_1, \underline{\beta}_2\right)\right] = E\left[\underline{y}^T Q_{12}\underline{y}\right] = \sigma^2\left(n - k_1 - k_2\right)$, (verify this!) so

$$\frac{E\left[SS\left(\underline{\beta}_2|\underline{\beta}_1\right)/k_2\right]}{E\left[RSS\left(\underline{\beta}_1, \underline{\beta}_2\right)/\left(n - k_1 - k_2\right)\right]} = 1 + \underline{\beta}_2^T X_2^T P_{2\cdot1} X_2 \underline{\beta}_2$$

equals 1 when $H_0 : \underline{\beta}_2 = \underline{0}$ is *true*; and $\geq 1$ when $H_0$ is *false*. This is suggestive of the behavior of the $F$ statistic

$$F = \frac{SS\left(\underline{\beta}_2|\underline{\beta}_1\right)/k_2}{RSS\left(\underline{\beta}_1, \underline{\beta}_2\right)/\left(n - k_1 - k_2\right)}$$

# 5 Testing a General Linear Hypothesis

Let us return to the general linear regression model

$$\underline{y}_{n\times1} = X_{n\times k'}\underline{\beta}_{k'\times1} + \varepsilon_{n\times1}$$

and now consider inference for a set of general linear combinations of the parameters

$$\underline{\psi}_{q\times 1} = L_{q\times k'}\beta_{k'\times 1}$$

Some examples of $L$'s that might be of interest include

- $L$ selects the last $k_2$ parameters (as in the previous section):

$$L_{k_2\times(k_1+k_2)} = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}; \quad L\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{k_1+k_2} \end{bmatrix} = \begin{bmatrix} \beta_{k_1+1} \\ \vdots \\ \beta_{k_1+k_2} \end{bmatrix}$$

$$\underbrace{\qquad\qquad\qquad}_{k_1 \text{ columns}}\underbrace{\qquad\qquad\qquad}_{k_2 \text{ columns}}$$

- $L$ codes the contrasts to test the hypotheses:

$$H_0 : \ \mu_1 = (\mu_2 + \mu_3)/2 \ vs. \ H_1 : \ \mu_1 \neq (\mu_2 + \mu_3)/2$$

$$H_0 : \ \mu_2 = \mu_3 \qquad\qquad vs. \ H_1 : \ \mu_2 \neq \mu_3$$

  in a one-way ANOVA model with three cells, i.e.

$$L_{2\times 3} = \begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & 1 & -1 \end{bmatrix}; \quad L\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1 - (\mu_2 + \mu_3)/2 \\ \mu_2 - \mu_3 \end{bmatrix}$$

  (note that $L$ is the transpose of the "contrast matrix" that we have dealt with previously).

Now, since $\underline{\psi}_{q\times 1} = L_{q\times k'}\underline{\beta}_{k'\times 1}$, it follows that

$$\underline{\hat{\psi}} = L\underline{\hat{\beta}} = L(X^TX)^{-1}X^Ty = Ay \sim N(\psi, \Sigma_{\hat{\psi}})$$

(either by least-squares, or more simply from a maximum likelihood argument), where

$$\Sigma_{\hat{\psi}} = \text{Var}\, Ay = A\text{Var}\, yA^T = AA^T\sigma_\epsilon^2$$

Now let

$$B = \Sigma_{\hat{\psi}}/\sigma_\epsilon^2 = AA^T = L(X^TX)^{-1}X^TX(X^TX)^{-1}L^T = L(X^TX)^{-1}L^T$$

It then follows (see for example Scheffé pp. 1–30) that

$$(\hat{\underline{\psi}} - \underline{\psi})^T B^{-1}(\hat{\underline{\psi}} - \underline{\psi}) \sim \sigma_\epsilon^2 \chi_q^2, \quad \text{independent of } RSS$$

*Intuition:*

**Chi-squared:** Let $R$ be a "square root" of $B^{-1}$; then $R(\hat{\underline{\psi}}-\underline{\psi})$ uncorrelates $(\hat{\underline{\psi}}-\underline{\psi})$ and the result is a vector of $q$ independent $N(0,1)'s$; the quadratic form $(\hat{\underline{\psi}}-\underline{\psi})^T B^{-1}(\hat{\underline{\psi}}-\underline{\psi}) = \|R(\hat{\underline{\psi}}-\underline{\psi})\|^2$ is just the sum of squares of these $q$ $N(0,1)$'s;

**Indep. of RSS:** Like our earlier arguments based on Cochran's Theorem.

Thus, when the true value of $\underline{\psi} = \underline{\psi}_0$,

$$F = \frac{(\hat{\underline{\psi}} - \underline{\psi}_0)^T B^{-1}(\hat{\underline{\psi}} - \underline{\psi}_0)/q}{RSS/(n-k')} = \frac{(\hat{\underline{\psi}} - \underline{\psi}_0)^T [L(X^TX)^{-1}L^T]^{-1}(\hat{\underline{\psi}} - \underline{\psi}_0)/q}{S_E^2} \sim F_{q,n-k'}$$

This can immediately be "inverted" to provide direct tests of hypotheses:

- Under $H_0: \ \underline{\psi} = \underline{\psi}_0$, the $F$ statistic above is $F_{q,n-k'}$;
- Under $H_1 : \underline{\psi}$ free, the $F$ statistic tends to be too large.

One can show that when $L$ selects $k_2$ parameters and we test for $\underline{\psi} = \underline{0}$, we get exactly the same test that we derived in connection with least-squares geometry in the last section. In this case $[L(X^TX)^{-1}L^T]^{-1}$ has the effect of "selecting out" the rows and columns of $(X^TX)^{-1}$ corresponding to the parameters $\beta_{k_1+1}, \ldots, \beta_{k_1+k_2}$ and inverting the resulting submatrix.

# 6 Confidence Regions

The same basic fact, that

$$F = \frac{(\hat{\underline{\psi}} - \underline{\psi}_0)^T [L(X^TX)^{-1}L^T]^{-1}(\hat{\underline{\psi}} - \underline{\psi}_0)/q}{S_E^2} \sim F_{q,n-k'}$$

when $\underline{\psi} = \underline{\psi}_0$, can be used to generate confidence regions for parameters or combinations of parameters. The idea of a confidence region is the same as that of a confidence interval, except that it is a region in $q$-space for all $q$ parameters $\underline{\psi} = (\psi_1, \ldots, \psi_q)$.

**Definition 3** *A 100$\times(1-\alpha)$% confidence ellipsoid for the parameters $\underline{\psi} = (\psi_1, \ldots, \psi_q)$ is given by*

$$\{\underline{\psi} : \ (\hat{\underline{\psi}} - \underline{\psi})^T [L(X^TX)^{-1}L^T]^{-1}(\hat{\underline{\psi}} - \underline{\psi}) \le qS_E^2 F_{\alpha,q,n-k'}\}$$

Once again, $L$ may be designed so that it selects a particular group of parameters $\beta_j$ of interest, or a particular group of linear combinations of $\beta_j$'s, etc.

Some examples will help to fix ideas here. The `prestige` data contains various variables that were thought to be predictive of occupational prestige in Canada, taken largely from 1971 Canadian Census archives. The variables are fairly self-explanatory

```
> names(prestige)
[1] "education"    "income"       "pct.female"   "occ.prestige" "occ.code"
[6] "occ.type"
> dim(prestige)
[1] 102   6
```

Here we review a couple of the fits that we did when we first encountered this data set.

```
> summary(fit.ed <- lm(occ.prestige ~ education, data = prestige))

Call: lm(formula = occ.prestige ~ education, data = prestige)
Residuals:
    Min      1Q Median    3Q   Max
 -26.04 -6.523 0.6611 6.743 18.16

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept) -10.7320   3.6771    -2.9186   0.0043
  education   5.3609   0.3320    16.1478   0.0000

Residual standard error: 9.103 on 100 degrees of freedom
Multiple R-Squared: 0.7228
F-statistic: 260.8 on 1 and 100 degrees of freedom, the p-value is 0

Correlation of Coefficients:
          (Intercept)
education -0.9695
```

We see that neither the slope nor the intercept are likely to be zero, according to their own $t$-tests; we could also see this with 95% $t$-intervals:

```
> qt((1-.05/2),100)
[1] 1.983972
> -10.732+c(-1,1)*1.983972*3.6771    # for beta_0
[1] -18.027263  -3.436737
> 5.3609  +c(-1,1)*1.983972*0.332    # for beta_1
[1] 4.702221 6.019579
```

on the other hand, the parameter estimates are very highly correlated, and these intervals are not corrected for doing two inferences at once.

We can fix the multiple inference problem with the *Bonferroni* correction, replacing $\alpha$ with $\alpha$/(number of tests),

```
> -10.732+c(-1,1)*qt((1-(.05/2)/2),100)*3.6771    # for beta_0
[1] -19.099801  -2.364199
> 5.3609  +c(-1,1)*qt((1-(.05/2)/2),100)*0.332    # for beta_1
[1] 4.605383 6.116417
```

but we still don't learn how the parameters covary. Instead we calculate the confidence ellipse

```
> X _ model.matrix(fit.ed)
> dim(X)
[1] 102   2
> X[1:5,]
                    (Intercept) education
 GOV_ADMINISTRATORS           1     13.11
   GENERAL_MANAGERS           1     12.26
        ACCOUNTANTS           1     12.77
PURCHASING_OFFICERS           1     11.42
           CHEMISTS           1     14.62
> SE2 _ sum(residuals(fit.ed)^2)/100
> sqrt(SE2) # check against earlier output
[1] 9.103291
> M _ t(X)%*%X/(2*SE2)
> b0 _ seq(-20,-2,length=50)
> b1 _ seq(3,8,length=50)
> bhat0 _ coefficients(fit.ed)[1]
> bhat1 _ coefficients(fit.ed)[2]
> qform _ function(psi, psi0, M)
+ {
+ #
+ # qform
+ #
+ # computes the quadratic form (psi - psi0)' %*% M %*% (Psi - Psi0)
+ # on every row of the matrix psi, using the single row in psi0
+ #
+ #
+       qfn <- function(x, M)
+       {
+               x %*% M %*% x
+       }
+       n <- dim(psi)[1]
+       delta <- psi - matrix(rep(psi0, n), byrow = T, ncol = 2)
+       result <- apply(delta, 1, qfn, M)
+       return(result)
+ }
> qform(cbind(0,0),c(bhat0,bhat1),M)
```

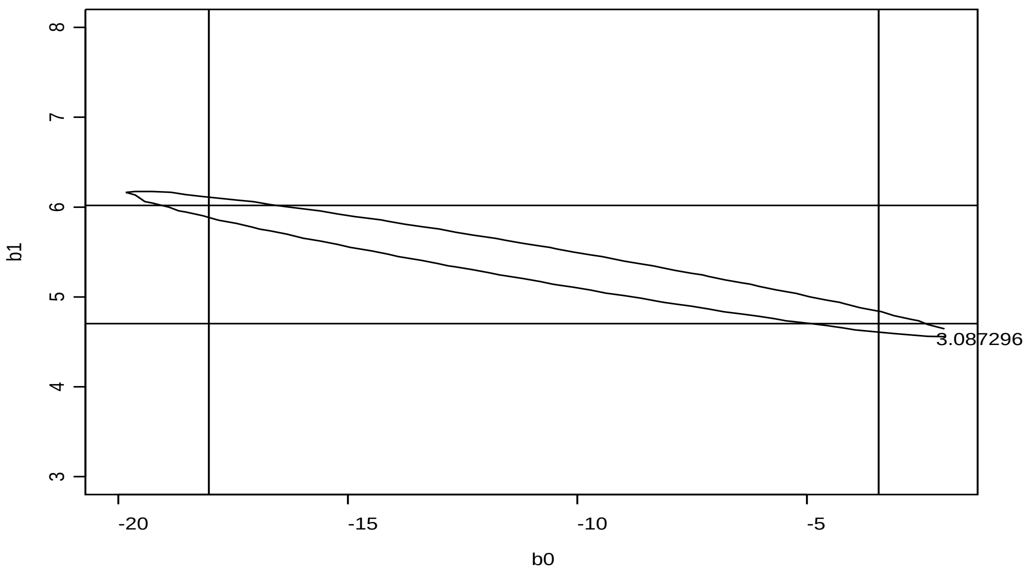


Figure 4: 95% Confidence Ellipse for $\beta_0$ and $\beta_1$ from `fit.ed`

```
[1] 1480.219
> qlev _ qform(expand.grid(b0,b1),c(bhat0,bhat1),M) # wait a while...
> contour(b0,b1,matrix(qlev,nrow=length(b0),ncol=length(b1)),
+ levels=qf(.95,2,100))
> abline(h=c(4.702221, 6.019579),v=c(-18.027263, -3.436737))
```

The resulting confidence ellipse is shown in Figure 4. Even though the individual confidence intervals are fairly wide, the confidence ellipse shows that $\hat{\beta}_1$ can vary only a little, for example, at each fixed value of $\hat{\beta}_0$. Horizontal and vertical lines indicating the original uncorrected $t$-based confidence intervals for $\beta_0$ and $\beta_1$ are also overlaid on the plot; one can see that the "box" formed by these two intervals both misses some of the area covered by the ellipse—reflecting the fact that the intervals are uncorrected for multiple inferences—and covers much more area than the ellipse covers—reflecting the inefficiency of using individual intervals.

If one did want individual intervals for the $\beta$'s that were guaranteed to cover the $\beta$'s with 95% confidence, a natural choice would be the "shadow" of the ellipse on each axis. This generates a larger box (not shown in Figure 4); the two intervals would be guaranteed to contain $\beta_0$ and $\beta_1$ in at least (and in fact more than) 95% of repeated samples from the "model" that generated the data.

**Scheffé's method.** An extension of the "shadow" idea is Scheffé's method of generating simultaneous confidence intervals for *all possible* linear combinations of the original parameters $\underline{\beta}$.

Let $\psi = \sum_{j=1}^{k'} w_j \beta_j = \underline{w}^T \underline{\beta}$ be any linear combination of the original $k'$ model parameters; then as usual

$$\hat{\psi} = \underline{w}^T \hat{\underline{\beta}} = \underline{w}^T (X^T X)^{-1} X^T \underline{y}$$

and

$$\text{Var}\,(\hat{\psi}) = \underline{w}^T \text{Var}\,(\underline{\hat{\beta}})\underline{w} = \sigma_\epsilon^2 \underline{w}^T (X^T X)^{-1} \underline{w}$$

and hence

$$SE_{\hat{\psi}} = S_E \sqrt{\underline{w}^T (X^T X)^{-1} \underline{w}}$$

A simple version[3] and of Scheffé's method replaces the $t$ cutoff in an uncorrected confidence interval

$$\hat{\psi} \pm t_{1-\alpha/2, n-k'} \cdot SE_{\hat{\psi}}$$

with a scaled $F$ cutoff:

$$\hat{\psi} \pm \sqrt{k' F(1-\alpha, k', n-k')} \cdot SE_{\hat{\psi}}$$

The $F$ cutoff $k'F(1-\alpha, k', n-k')$ comes from calculating a confidence ellipsoid for the original set of $k'$ parameters $\underline{\beta}$; the essential idea of Scheffé's method is that once you have controlled these, you automatically control any linear combination of them. For a proof that these are really simultaneous $100(1-\alpha)\%$ intervals, see Scheffé, 1959, section 3.5.

Scheffé's intervals are very useful for

- Data snooping, or any situation in which you are evaluating a lot of linear contrasts or other linear combinations of the original parameters, and you want to control the lengths of your intervals; and
- Graphing simultanous confidence envelopes around regression lines or surfaces.

We will illustrate the latter with the `prestige` example. We wish to graph

- The data
- The fitted regression line $\hat{y} = \hat{\beta}_0 + x\hat{\beta}_1$
- A simultaneous 95% confidence envelope computed at every fitted value

We recall that $\underline{\hat{y}} = X\underline{\hat{\beta}} = P\underline{y} = H\underline{y}$ and $\text{Var}\,(y_i) = h_{ii}\sigma_\epsilon^2$, so that $SE_{\hat{y}_i} = S_E\sqrt{h_{ii}}$. Also $n = 102$ and $k' = 2$ for the simple regression of `occ.prestige` on `education`.

```
> attach(prestige)
> plot(education,occ.prestige)
> abline(fit.ed)
> SEy _ sqrt(lm.hat(fit.ed)*sum(residuals(fit.ed)^2)/100)
> l _ fitted(fit.ed) - sqrt(2*qf(.95,2,100))*SEy
> h _ fitted(fit.ed) + sqrt(2*qf(.95,2,100))*SEy
> ord _ order(education)
> lines(education[ord],l[ord],lty=3)
> lines(education[ord],h[ord],lty=3)
> detach()
```

[3] If you know that you will only be considering $\psi$'s that are linear combinations of a subset of $q$ of the $\beta$'s then you may replace $(X^T X)^{-1}$ with $L(X^T X)^{-1} L^T$ where $L$ selects the subset, and use the cutoff $qF(1-\alpha, q, n-k')$ instead.

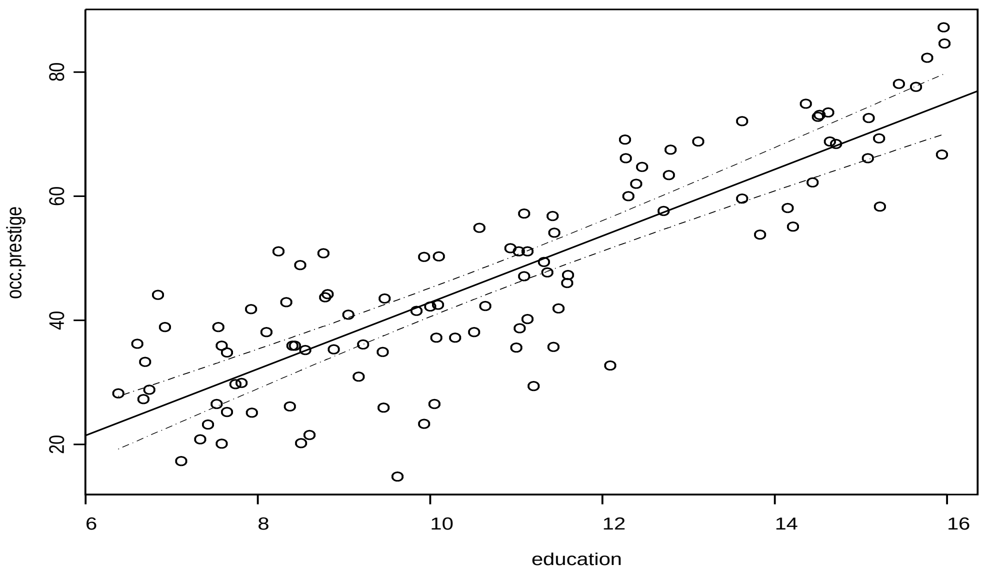


Figure 5: 95% Envelope for Fitted Regression Line from from `fit.ed`

The resulting graph is displayed in Figure 5. Comparing the $t$ cutoff with the $F$ cutoff in Scheffé's method,

```
> sqrt(2*qf(.95,2,100))
[1] 2.484873
> qt(1-.05/2,100)
[1] 1.983972
```

we see that there is a substantial difference between the corrected and the uncorrected envelopes for the regression line here.

**Remark:** There are other methods to generate confidence envelopes that may do better. For example, at each $x_i$ value, we could calculate and plot the values

$$\begin{aligned}\tilde{u}_i &= \max\{\beta_0 + x_i\beta_1 : (\beta_0, \beta_1) \in \text{ the 95\% confidence ellipsoid in Fig 4}\} \\ \tilde{\ell}_i &= \min\{\beta_0 + x_i\beta_1 : (\beta_0, \beta_1) \in \text{ the 95\% confidence ellipsoid in Fig 4}\}\end{aligned}$$

along with the $\hat{y}_i$'s.