

# Linear statistical models

## Inference for the full rank model

Notes by Yao-ban Chan and Owen Jones

# The full rank model

In this section, we develop various forms of hypothesis testing on the full rank model. To recap, the full rank model is

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where  $X$  is  $n \times p$ ,  $n \geq p$ ,  $r(X) = p$ , and the errors  $\boldsymbol{\varepsilon}$  have:

- ▶ mean  $\mathbf{0}$ ;
- ▶ variance  $\sigma^2 I$ ;
- ▶ are multivariate normal.

# The full rank model

The first thing we want to test is *model relevance*: does our model contribute anything at all?

If none of the  $x$  variables have any relevance for predicting  $y$ , then all the parameters  $\beta$  will be  $\mathbf{0}$ .

We test for this using the null hypothesis

$$H_0 : \beta = \mathbf{0}.$$

# The full rank model

Alternatively, if at least some of the  $x$  variables are relevant to predicting  $y$ , then the corresponding parameters will be nonzero. So our alternative hypothesis is

$$H_1 : \beta \neq \mathbf{0}.$$

To test these hypotheses, we assume throughout this section that the errors  $\varepsilon$  are multivariate normal.

# ANOVA

The method used to test the hypotheses is analysis of variance (ANOVA).

If  $\beta = \mathbf{0}$ , then  $\mathbf{y} = \epsilon$  consists entirely of errors. In this case,  $\mathbf{y}^T \mathbf{y}$ , the sum of squares of the errors, measures the variability of the errors.

However, if  $\beta \neq \mathbf{0}$ , then  $\mathbf{y} = X\beta + \epsilon$ . In this case, some of  $\mathbf{y}^T \mathbf{y}$  will come from errors, but some will come from the model predictions.

By separating  $\mathbf{y}^T \mathbf{y}$  into these two parts, we can compare them to see how well the model is doing.

# ANOVA

More precisely, the sum of squares of the residuals (variation attributed to error) is

$$\begin{aligned}SS_{Res} &= (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}) \\&= (\mathbf{y} - H\mathbf{y})^T(\mathbf{y} - H\mathbf{y}) \\&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T H\mathbf{y} + \mathbf{y}^T H^2\mathbf{y} \\&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T H\mathbf{y} \\&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X(X^T X)^{-1}X^T\mathbf{y}\end{aligned}$$

which means that

$$\mathbf{y}^T\mathbf{y} = \mathbf{y}^T X(X^T X)^{-1}X^T\mathbf{y} + SS_{Res}.$$

# ANOVA

We call  $\mathbf{y}^T X (X^T X)^{-1} X^T \mathbf{y} = \hat{\mathbf{y}}^T \hat{\mathbf{y}}$  the *regression sum of squares* and denote it by  $SS_{Reg}$ . It reflects the variation in the response variable that is explained by the model.

We call the total variation in the response variable  $SS_{Total} = \mathbf{y}^T \mathbf{y}$ . We have divided it into:

$$SS_{Total} = SS_{Reg} + SS_{Res}.$$

# ANOVA

**Example.** Suppose that there is no error, so that  $\mathbf{y} = X\beta$ . We have

$$\begin{aligned}SS_{Reg} &= \mathbf{y}^T X (X^T X)^{-1} X^T \mathbf{y} \\&= \beta^T X^T X (X^T X)^{-1} X^T X \beta \\&= \beta^T X^T X \beta \\&= \mathbf{y}^T \mathbf{y} = SS_{Total}\end{aligned}$$

and  $SS_{Res} = 0$ .



# ANOVA

On the other hand, suppose that there is no signal, so that  $\beta = \mathbf{0}$  and  $\mathbf{y} = \epsilon$ . If we put  $\mathbf{b} = \beta = \mathbf{0}$  then

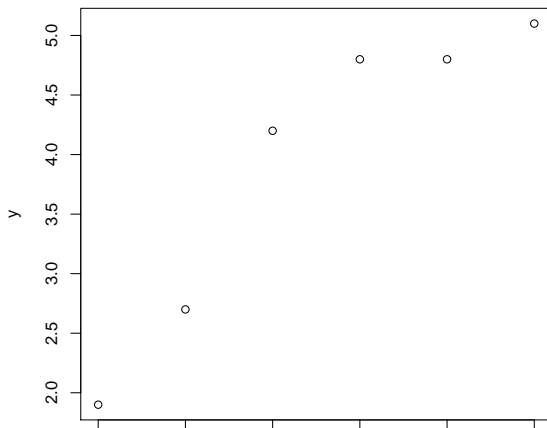
$$\begin{aligned}SS_{Res} &= (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}) \\ &= \mathbf{y}^T\mathbf{y} = SS_{Total}\end{aligned}$$

and  $SS_{Reg} = 0$ .

These are the two extremes of the spectrum.

# ANOVA

**Example.** Recall our previous paint cracking example, in which the data had a strong linear relationship.



# ANOVA

```
> y <- c(1.9,2.7,4.2,4.8,4.8,5.1)
> X <- matrix(c(rep(1,6),2:7),6,2)
> b <- solve(t(X)%*%X,t(X)%*%y)
> e <- y-X%*%b
> (SSRes <- sum(e^2))
```

```
[1] 1.096762
```

```
> (SSTotal <- sum(y^2))
```

```
[1] 100.63
```

```
> (SSReg <- SSTotal - SSRes)
```

```
[1] 99.53324
```

Since  $99.53 \gg 1.1$ , informally we would say that there is a strong linear signal in the data.

# ANOVA

To create a formal test of  $\beta = \mathbf{0}$ , we compare  $SS_{Reg}$  against  $SS_{Res}$ . If  $SS_{Reg}$  is large compared to  $SS_{Res}$ , then we have evidence that  $\beta \neq \mathbf{0}$ .

To know exactly *how* large, we must first derive the distributions of  $SS_{Reg}$  and  $SS_{Res}$ .

# ANOVA

## Theorem 5.1 (4.13)

*In the full rank general linear model  $\mathbf{y} = X\beta + \epsilon$ ,  $SS_{Res}/\sigma^2$  has a  $\chi^2$  distribution with  $n - p$  degrees of freedom.*

## Theorem 5.2

*In the full rank general linear model  $\mathbf{y} = X\beta + \epsilon$ ,  $SS_{Reg}/\sigma^2$  has a noncentral  $\chi^2$  distribution with  $p$  degrees of freedom and noncentrality parameter*

$$\lambda = \frac{1}{2\sigma^2} \beta^T X^T X \beta.$$

# ANOVA

**Proof.** By definition

$$\frac{SS_{Reg}}{\sigma^2} = \frac{1}{\sigma^2} \mathbf{y}^T H \mathbf{y}.$$

By assumption,  $\mathbf{y} \sim MVN(X\beta, \sigma^2 I)$ . We also have  $H$  symmetric and idempotent, with  $r(H) = p$ .

By Corollary 3.7,  $SS_{Reg}/\sigma^2$  has a noncentral  $\chi^2$  distribution with  $k + 1$  degrees of freedom and noncentrality parameter

$$\begin{aligned}\lambda &= \frac{1}{2\sigma^2} (X\beta)^T H (X\beta) \\ &= \frac{1}{2\sigma^2} \beta^T X^T X \beta.\end{aligned}$$

# ANOVA

## Theorem 5.3

*In the full rank general linear model  $\mathbf{y} = X\beta + \varepsilon$ ,  $SS_{Res}$  and  $SS_{Reg}$  are independent.*

This can be proved by observing that they are both quadratic forms in  $\mathbf{y}$  and applying Theorem 3.11.

Alternatively, we can write  $SS_{Reg} = \mathbf{b}^T X^T X \mathbf{b}$  and observe that  $\mathbf{b}$  and  $s^2$  are independent.

# ANOVA

Now how do we test  $\beta = \mathbf{0}$ ?

Observe that if this is true, the noncentrality parameter for  $SS_{Reg}/\sigma^2$  must be 0.

Thus, under  $H_0$ ,

$$\frac{SS_{Reg}/p\sigma^2}{SS_{Res}/(n-p)\sigma^2} = \frac{SS_{Reg}/p}{SS_{Res}/(n-p)} = \frac{MS_{Reg}}{MS_{Res}}$$

has an  $F$  distribution with  $p$  and  $n - p$  degrees of freedom.



# ANOVA

What happens if  $H_0$  is not true? The expected value of  $MS_{Reg}$  is

$$E \left[ \frac{SS_{Reg}}{p} \right] = \sigma^2 + \frac{1}{p} \beta^T X^T X \beta.$$

(Recall  $SS_{Reg} = \mathbf{y}^T H \mathbf{y}$  and  $E[\mathbf{x}^T A \mathbf{x}] = \text{tr}(A V) + \boldsymbol{\mu}^T A \boldsymbol{\mu}.$ )

The expected value of the denominator  $MS_{Res}$  is

$$E \left[ \frac{SS_{Res}}{n - p} \right] = E[s^2] = \sigma^2.$$

# ANOVA

So if  $\beta = \mathbf{0}$ ,  $E[\frac{SS_{Reg}}{p}] = \sigma^2$  and the statistic should be close to 1.

But if  $\beta \neq \mathbf{0}$ , since  $X^T X$  is positive definite, we get  $E[\frac{SS_{Reg}}{p}] > \sigma^2$  and the statistic should be bigger than 1.

Therefore, we should use a one-tailed test and reject  $H_0$  if the statistic is large.

# ANOVA

To lay out all the calculations, we use a familiar ANOVA table.

Source of variation	Sum of squares	degrees of freedom	Mean square	F ratio
Regression	$\mathbf{y}^T X (X^T X)^{-1} X^T \mathbf{y}$	$p$	$\frac{SS_{Reg}}{p}$	$\frac{MS_{Reg}}{MS_{Res}}$
Residual	$\mathbf{y}^T \mathbf{y} - \mathbf{y}^T X (X^T X)^{-1} X^T \mathbf{y}$	$n - p$	$\frac{SS_{Res}}{n - p}$	
Total	$\mathbf{y}^T \mathbf{y}$	$n$		

## Example: system cost

A data processing system uses three types of structural elements: files, flows and processes. Files are permanent records, flows are data interfaces, and processes are logical manipulations of the data. The cost of developing software for the system is based on the number of these three elements. A study is conducted with the following results:

Cost ( $y$ )	Files ( $x_1$ )	Flows ( $x_2$ )	Processes ( $x_3$ )
22.6	4	44	18
15	2	33	15
78.1	20	80	80
28	6	24	21
80.5	6	227	50
24.5	3	20	18
20.5	4	41	13
147.6	16	187	137
4.2	4	19	15
48.2	6	50	21
20.5	5	48	17

## Example: system cost

The model we use is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i.$$

We want to test the hypothesis of model relevance, i.e.

$$H_0 : \beta = \mathbf{0} \text{ vs. } H_1 : \beta \neq \mathbf{0}.$$

```
> (b <- solve(t(X)%*%X,t(X)%*%y))
```

```
      [,1]
```

```
[1,] 1.9617795
```

```
[2,] 0.1177586
```

```
[3,] 0.1767263
```

```
[4,] 0.7964477
```

## Example: system cost

```
> (SSReg <- t(y)%*%X%*%b)
```

```
      [,1]
```

```
[1,] 38978.38
```

```
> (SSTotal <- sum(y^2))
```

```
[1] 39667.01
```

```
> (SSRes <- SSTotal - SSReg)
```

```
      [,1]
```

```
[1,] 688.6262
```

```
> (MSReg <- SSReg/p)
```

```
      [,1]
```

```
[1,] 9744.596
```

## Example: system cost

```
> (MSRes <- SSRes/(n-p))  
      [,1]  
[1,] 98.37517  
  
> (Fstat <- MSReg/MSRes)  
      [,1]  
[1,] 99.05544  
  
> qf(0.95,p,n-p)  
[1] 4.120312  
  
> pf(Fstat,p,n-p,lower.tail=FALSE)  
      [,1]  
[1,] 3.060186e-06
```

## Example: system cost

We can say that  $\beta \neq \mathbf{0}$  with a lot of confidence.

Variation	SS	d.f.	MS	F
Regression	38978.38	4	9744.60	99.06
Residual	688.63	7	98.38	
Total	39667.01	11		



## Example: system cost

```
> model <- lm(y~X[,-1])
> null <- lm(y~0)
> anova(null,model)
```

### Analysis of Variance Table

Model 1:  $y \sim 0$

Model 2:  $y \sim X[, -1]$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	11	39667				
2	7	689	4	38978	99.055	3.06e-06 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Clover example

Recall the clover example:

```
> clover <- read.csv("../data/clover.csv")
> clover <- log(clover)
> clover <- clover[-c(6,23,47,97,111,140),]
> y <- clover$area
> X <- cbind(1, clover$midrib, clover$estim)
> b <- solve(t(X) %*% X, t(X) %*% y)
> n <- length(y)
> p <- dim(X)[2]
```

## Clover example

We test for model relevance,  $H_0 : \beta = \mathbf{0}$ .

```
> (SSTotal <- sum(y^2))
```

```
[1] 381.3093
```

```
> (SSRes <- sum((y - X %*% b)^2))
```

```
[1] 4.722065
```

```
> (SSReg <- SSTotal - SSRes)
```

```
[1] 376.5873
```

```
> (Fstat <- (SSReg/p)/(SSRes/(n-p)))
```

```
[1] 3615.358
```

```
> pf(Fstat, p, n-p, lower.tail=FALSE)
```

```
[1] 1.916608e-129
```

# Clover example

```
> basemodel <- lm(area ~ 0, data=clover)
> model <- lm(area ~ midrib + estim, data=clover)
> anova(basemodel, model)
```

## Analysis of Variance Table

Model 1: area ~ 0

Model 2: area ~ midrib + estim

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	139	381.31				
2	136	4.72	3	376.59	3615.4	< 2.2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# The general linear hypothesis

We can now progress to testing the *general linear hypothesis*:

$$H_0 : C\beta = \delta^* \text{ vs. } H_1 : C\beta \neq \delta^*,$$

where  $C$  is an  $r \times p$  matrix of rank  $r \leq p$  and  $\delta^*$  is an  $r \times 1$  vector of constants.

This hypothesis makes it possible to test for many things, including relationships among the parameters, or testing the individual parameters against a constant.

# The general linear hypothesis

**Example.** Consider the null hypothesis of model relevance,  
 $H_0 : \beta = \mathbf{0}$ .

We can express this in the form of the general linear hypothesis with  $C = I_p$  (which has rank  $p$ ) and  $\delta^* = \mathbf{0}$ .

# The general linear hypothesis

**Example.** Consider a regression model with 4 parameters (3 predictors)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i.$$

Let

$$C = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad \delta^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

If we test  $H_0 : C\beta = \delta^*$ , this is equivalent to

$$\beta_1 - \beta_2 = 0$$

$$\beta_2 - \beta_3 = 0.$$

In other words, we are testing the hypothesis  $\beta_1 = \beta_2 = \beta_3$ .

# Test statistic

To develop a test statistic, we start with  $C\mathbf{b} - \delta^*$ , the least squares estimator for  $C\beta - \delta^*$ .

This is a linear combination of  $\mathbf{b}$ , a multivariate normal. Hence it is also multivariate normal.

We have

$$\begin{aligned}\mathbb{E}(C\mathbf{b} - \delta^*) &= C\beta - \delta^* \\ \text{Var}(C\mathbf{b} - \delta^*) &= C(X^T X)^{-1} C^T \sigma^2.\end{aligned}$$



# Test statistic

From Corollary 3.10, the quadratic form

$$\frac{(C\mathbf{b} - \boldsymbol{\delta}^*)^T [C(X^T X)^{-1} C^T]^{-1} (C\mathbf{b} - \boldsymbol{\delta}^*)}{\sigma^2}$$

has a noncentral  $\chi^2$  distribution with  $r$  degrees of freedom and noncentrality parameter

$$\lambda = \frac{(C\boldsymbol{\beta} - \boldsymbol{\delta}^*)^T [C(X^T X)^{-1} C^T]^{-1} (C\boldsymbol{\beta} - \boldsymbol{\delta}^*)}{2\sigma^2}.$$

(How do we know  $C(X^T X)^{-1} C^T$  has an inverse?)

# Test statistic

If the null hypothesis is true, then  $C\beta = \delta^*$  and the quadratic form has an ordinary  $\chi^2$  distribution.

Since the quadratic form depends (stochastically) only on  $\mathbf{b}$ , it is independent from  $s^2$ .

Therefore under the null hypothesis, the statistic

$$\frac{(C\mathbf{b} - \delta^*)^T [C(X^T X)^{-1} C^T]^{-1} (C\mathbf{b} - \delta^*) / r}{SS_{Res} / (n - p)}$$

has an  $F$  distribution with  $r$  and  $n - p$  degrees of freedom.

We use this statistic to test the general linear hypothesis.

# Test statistic

To justify a one-tailed test, the expected value of the numerator can be calculated to be

$$\begin{aligned} E \left[ \frac{(C\mathbf{b} - \boldsymbol{\delta}^*)^T [C(X^T X)^{-1} C^T]^{-1} (C\mathbf{b} - \boldsymbol{\delta}^*)}{r} \right] \\ = \sigma^2 + \frac{1}{r} (C\boldsymbol{\beta} - \boldsymbol{\delta}^*)^T [C(X^T X)^{-1} C^T]^{-1} (C\boldsymbol{\beta} - \boldsymbol{\delta}^*), \end{aligned}$$

where  $C(X^T X)^{-1} C^T$  is positive definite.

If the null hypothesis is true, then the expectation is  $\sigma^2$ . However, if  $H_0$  is false, it will be greater than  $\sigma^2$ .

Therefore we reject  $H_0$  when the statistic is large.

## Example: system cost

We revisit the data processing system example. We test the hypothesis  $H_0 : \beta = [2 \ 0 \ 0 \ 1]^T$ .

```
> dst <- c(2,0,0,1)
> C <- diag(4)
> r <- 4
> num <- t(C%*%b-dst)%*%solve(C%*%solve(t(X)%*%X)%*%t(C))%*%
+      (C%*%b-dst)/r
> (Fstat <- num/(SSRes/(n-p)))
```

```
      [,1]
```

```
[1,] 2.795888
```

```
> pf(Fstat, r, n-p, lower=F)
```

```
      [,1]
```

```
[1,] 0.1115939
```

## Example: system cost

The critical value of the  $F$  distribution with 4 and 7 degrees of freedom at  $\alpha = 0.05$  is 4.12, so we cannot reject the null hypothesis.

This doesn't mean that it is true, just that it is close!

Exercise: show that for  $C = I$ ,

$$(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X (\mathbf{b} - \boldsymbol{\beta}^*) = (\mathbf{y} - X\boldsymbol{\beta}^*)^T (\mathbf{y} - X\boldsymbol{\beta}^*) - (\mathbf{y} - X\mathbf{b})^T (\mathbf{y} - X\mathbf{b}).$$

That is, the quadratic form in the numerator is the  $SS_{Res}$  for the model under  $H_0$  minus the  $SS_{Res}$  for the full model.

## Example: system cost

```
> library(car)
> dst <- c(2,0,0,1)
> C <- diag(4)
> linearHypothesis(model,C,dst)
```

Linear hypothesis test

```
Hypothesis:
(Intercept) = 2
X[, - 1 = 0
X[, - 2 = 0
X[, - 3 = 1
```

Model 1: restricted model

Model 2:  $y \sim X[, -1]$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	11	1788.81				
2	7	688.63	4	1100.2	2.7959	0.1116

## Example: system cost

Now we test the hypothesis  $H_0 : C\beta = \delta^*$ , where

$$C = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad \delta^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

```
> dst <- c(0,0)
> C <- matrix(c(0,0,1,0,-1,1,0,-1),2,4)
> r <- 2
> num <- t(C%*%b-dst)%*%solve(C%*%solve(t(X)%*%X)%*%t(C))%*%
+      (C%*%b-dst)/r
> (Fstat <- num/(SSRes/(n-p)))

      [,1]
[1,] 5.785777

> pf(Fstat, r, n-p, lower=F)

      [,1]
[1,] 0.03287564
```

## Example: system cost

Thus we can reject the null hypothesis at the 5% level, but not at the 1% level.

That is, there is evidence that the parameters  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  are not identical, but not strong evidence.



## Example: system cost

```
> dst <- c(0,0)
> C <- matrix(c(0,0,1,0,-1,1,0,-1),2,4)
> linearHypothesis(model,C,dst)
```

Linear hypothesis test

Hypothesis:

```
X[, - 1 - X[, - 2 = 0
X[, - 2 - X[, - 3 = 0
```

Model 1: restricted model

Model 2:  $y \sim X[, -1]$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	9	1826.98				
2	7	688.63	2	1138.3	5.7858	0.03288 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Clover example

For the clover data, consider the null hypothesis

$$H_0 : (\beta_0, \beta_1, \beta_2) = (-1, 0.5, 1).$$

```
> bst <- as.vector(c(-1, 0.5, 1))
> ( Fstat <- ((t(b-bst) %*% t(X) %*% X %*% (b-bst))/p)/
+   (SSRes/(n-p)) )
```

```
      [,1]
[1,] 317.6183
```

```
> pf(Fstat, p, n-p, lower.tail=FALSE)
```

```
      [,1]
[1,] 3.230366e-61
```

# Clover example

```
> model <- lm(area ~ midrib + estim, data=clover)
> linearHypothesis(model, diag(3), bst)
```

Linear hypothesis test

Hypothesis:

(Intercept) = - 1

midrib = 0.5

estim = 1

Model 1: restricted model

Model 2: area ~ midrib + estim

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	139	37.806				
2	136	4.722	3	33.084	317.62	< 2.2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Clover example

Let's try the null hypothesis  $H_0 : \beta_0 = -1, \beta_1 = \beta_2$ .

```
> ( C <- matrix(c(1,0,0,1,0,-1),2,3) )
```

```
      [,1] [,2] [,3]  
[1,]     1     0     0  
[2,]     0     1    -1
```

```
> library(Matrix)
```

```
> (r <- rankMatrix(C)[1])
```

```
[1] 2
```

```
> dst <- c(-1,0)
```

## Clover example

```
> ( Fstat <- (t(C %*% b - dst) %*%  
+   solve(C %*% solve(t(X) %*% X) %*% t(C)) %*%  
+   (C %*% b - dst)/r)/(SSRes/(n-p)) )  
  
      [,1]  
[1,] 19.54309  
  
> pf(Fstat, r, n-p, lower=FALSE)  
  
      [,1]  
[1,] 3.463526e-08
```

# Clover example

```
> linearHypothesis(model, C, dst)
```

Linear hypothesis test

Hypothesis:

(Intercept) = - 1

midrib - estim = 0

Model 1: restricted model

Model 2: area ~ midrib + estim

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	138	6.0792				
2	136	4.7221	2	1.3571	19.543	3.464e-08 ***
---						

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Testing if part of $\beta$ is $\mathbf{0}$

If we find that  $\beta \neq \mathbf{0}$ , we cannot say which  $\beta_i$  are nonzero, only that at least one is not.

If a particular  $\beta_i$  is zero, then it is best to remove it from the model. Otherwise it will only fit noise, and reduce the ability of the model to predict.

Thus, we need to find a way of testing whether *parts* of the parameter vector are  $\mathbf{0}$  or not.

## Testing if part of $\beta$ is $\mathbf{0}$

We split the parameter vector

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{r-1} \\ \beta_r \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}$$

and test the hypotheses

$$H_0 : \gamma_1 = \mathbf{0} \text{ vs. } H_1 : \gamma_1 \neq \mathbf{0}.$$

By relabelling the indices, we can test the relevance of any subset of the parameters.



## Testing if part of $\beta$ is 0

An important thing to note is that we are testing  $\gamma_1 = \mathbf{0}$  *in the presence of the other parameters*, not by itself.

In other words, we are comparing two models: in  $H_1$ , the full model

$$\mathbf{y} = X\beta + \varepsilon,$$

and in  $H_0$ , the *reduced* model

$$\mathbf{y} = X_2\gamma_2 + \varepsilon_2$$

where  $X_2$  contains the last  $p - r$  columns of  $X = [X_1|X_2]$  (suitably re-ordered).

## Testing if part of $\beta$ is 0

We can do this in the framework of the general linear hypothesis.

Let  $C = [I_r | 0]$  and  $\delta^* = \mathbf{0}$ . Then  $C\beta = \delta^*$  iff  $\gamma_1 = \mathbf{0}$ .

We define the regression sum of squares for  $\gamma_1$  in the presence of  $\gamma_2$  as

$$\begin{aligned} R(\gamma_1 | \gamma_2) &= (C\mathbf{b} - \delta^*)^T (C(X^T X)^{-1} C^T)^{-1} (C\mathbf{b} - \delta^*) \\ &= \hat{\gamma}_1^T A_{11}^{-1} \hat{\gamma}_1, \end{aligned}$$

where  $\hat{\gamma}_1$  is the least squares estimator for  $\gamma_1$ , and  $A_{11}$  is the  $r \times r$  principal minor of  $(X^T X)^{-1}$ .

# Testing if part of $\beta$ is 0

Our test statistic is

$$\frac{R(\gamma_1|\gamma_2)/r}{SS_{Res}/(n-p)}.$$

From our previous results on the general linear hypothesis, we know that this has an  $F_{r,n-p}$  distribution under the null hypothesis  $\gamma_1 = \mathbf{0}$ . We reject the null when this statistic is too large.

# Testing if part of $\beta$ is 0

## Theorem 5.4

$$R(\gamma_1|\gamma_2) = R(\beta) - R(\gamma_2),$$

where  $R(\beta)$  is the regression sum of squares for the full model

$$\mathbf{y} = X\beta + \epsilon = [X_1|X_2] \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \epsilon,$$

and  $R(\gamma_2)$  is the regression sum of squares for the reduced model

$$\mathbf{y} = X_2\gamma_2 + \epsilon.$$

## Testing if part of $\beta$ is 0

We will content ourselves with showing that

$$\begin{aligned}\mathbb{E}\hat{\gamma}_1^T A_{11}^{-1} \hat{\gamma}_1 &= \mathbb{E}(R(\beta) - R(\gamma_2)) \\ &= \mathbb{E}\mathbf{y}^T [X(X^T X)^{-1} X^T - X_2(X_2^T X_2)^{-1} X_2^T] \mathbf{y}.\end{aligned}$$

### Lemma 5.5

*Suppose that*

$$A = \left[ \begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array} \right], A^{-1} = B = \left[ \begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array} \right],$$

*and  $B_{22}^{-1}$  exists. Then*

$$A_{11}^{-1} = B_{11} - B_{12} B_{22}^{-1} B_{21}.$$

# Testing if part of $\beta$ is 0

$$\begin{aligned}
 & \mathbb{E} \mathbf{y}^T [X(X^T X)^{-1} X^T - X_2(X_2^T X_2)^{-1} X_2^T] \mathbf{y} \\
 &= \sigma^2 \text{tr}(X(X^T X)^{-1} X^T - X_2(X_2^T X_2)^{-1} X_2^T) \\
 &\quad + \beta^T X^T [X(X^T X)^{-1} X^T - X_2(X_2^T X_2)^{-1} X_2^T] X \beta \\
 &= \sigma^2(p - (p - r)) + \beta^T [X^T X - X^T X_2(X_2^T X_2)^{-1} X_2^T X] \beta \\
 &= \sigma^2 r \\
 &\quad + \begin{bmatrix} \gamma_1^T & \gamma_2^T \end{bmatrix} \left[ \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix} \begin{bmatrix} X_1 & X_2 \end{bmatrix} - \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix} X_2(X_2^T X_2)^{-1} X_2^T \begin{bmatrix} X_1 & X_2 \end{bmatrix} \right] \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \\
 &= \sigma^2 r + \begin{bmatrix} \gamma_1^T & \gamma_2^T \end{bmatrix} \left[ \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} - \begin{bmatrix} X_1^T X_2 (X_2^T X_2)^{-1} X_2^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} \right] \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \\
 &= \sigma^2 r + \gamma_1^T [X_1^T X_1 - X_1^T X_2 (X_2^T X_2)^{-1} X_2^T X_1] \gamma_1 \\
 &= \sigma^2 \text{tr}(A_{11}^{-1} A_{11}) + \gamma_1^T A_{11}^{-1} \gamma_1 \\
 &= \mathbb{E} \hat{\gamma}_1^T A_{11}^{-1} \hat{\gamma}_1.
 \end{aligned}$$

Here we have applied the lemma with

$$A = (X^T X)^{-1}, \quad B = X^T X = \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix}.$$

# Testing if part of $\beta$ is 0

We again express the test calculations in an ANOVA table.

Source of variation	Sum of squares	degrees of freedom	Mean square	F ratio
Regression				
Full model	$R(\beta)$	$p$		
Reduced model	$R(\gamma_2)$	$p - r$		
$\gamma_1$ in presence of $\gamma_2$	$R(\gamma_1 \gamma_2)$	$r$	$\frac{R(\gamma_1 \gamma_2)}{r}$	$\frac{R(\gamma_1 \gamma_2)/r}{MS_{Res}}$
Residual	$\mathbf{y}^T \mathbf{y} - R(\beta)$	$n - p$	$\frac{SS_{Res}}{n - p}$	
Total	$\mathbf{y}^T \mathbf{y}$	$n$		

## Example: system cost

**Example.** Consider again the data processing system example. We rejected the hypothesis of model relevance,  $\beta = \mathbf{0}$ . But that is obvious because the cost of a system doesn't have average 0.

The question we want to test is, does the cost depend on the files, flows or processes? In other words, is one of  $\beta_1$ ,  $\beta_2$ , or  $\beta_3$  nonzero?

To do this, we re-arrange the parameter vector as

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_0 \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}.$$



## Example: system cost

We must rearrange the columns of  $X$  correspondingly:

$$X = \left[ \begin{array}{ccc|c} 4 & 44 & 18 & 1 \\ 2 & 33 & 15 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 5 & 48 & 17 & 1 \end{array} \right] = [X_1 \mid X_2].$$

We want to test  $H_0 : \gamma_1 = \mathbf{0}$  (only the intercept is relevant) against  $H_1 : \gamma_1 \neq \mathbf{0}$ . The reduced model is

$$\mathbf{y} = X_2\beta_0 + \varepsilon_2.$$

## Example: system cost

```
> X2 <- X[,1]
> (Rg2 <- t(y)%*%X2)%*%solve(t(X2)%*%X2)%*%t(X2)%*%y)
      [,1]
[1,] 21800.55
> (Rg1g2 <- SSReg - Rg2)
      [,1]
[1,] 17177.83
> (Fstat <- (Rg1g2/3)/(SSRes/(n-p)))
      [,1]
[1,] 58.20517
> pf(Fstat,3,n-p,lower=F)
      [,1]
[1,] 2.577615e-05
```

## Example: system cost

The intercept alone does not explain the variation in the response variable adequately, and we are (reasonably) certain that we need at least one of the terms in the model.

Variation	SS	d.f.	MS	F
<hr/>				
Regression				
Full	38978	4		
Reduced	21800	1		
$\gamma_1$ in presence of $\gamma_2$	17178	3	5726	58.2
Residual	689	7	98	
<hr/>				
Total	39667	11		

# Corrected sum of squares

In general, we have the following ANOVA table for the test

$H_0 : \beta_1 = \dots = \beta_k = 0$  versus the alternative that some  $\beta_i \neq 0$ ,  
 $i \in \{1, \dots, k\}$ .

Source of variation	Sum of squares	degrees of freedom	Mean square	F ratio
Regression				
Full model	$R(\beta) = \mathbf{y}^T H \mathbf{y}$	$k + 1$		
Reduced model	$(\sum_{i=1}^n y_i)^2 / n$	1		
$\gamma_1$ in presence of $\gamma_2$	$R(\gamma_1   \gamma_2)$	$k$	$\frac{R(\gamma_1   \gamma_2)}{k}$	$\frac{R(\gamma_1   \gamma_2) / k}{MS_{Res}}$
Residual	$\mathbf{y}^T \mathbf{y} - R(\beta)$	$n - k - 1$	$\frac{SS_{Res}}{n - p}$	
Total	$\mathbf{y}^T \mathbf{y}$	$n$		

## Corrected sum of squares

The  $SS_{Reg}$  for the reduced model comes from

$$\mathbf{y}^T \mathbf{1} (\mathbf{1}^T \mathbf{1})^{-1} \mathbf{1}^T \mathbf{y} = \left( \sum_{i=1}^n y_i \right) \frac{1}{n} \left( \sum_{i=1}^n y_i \right).$$

This ANOVA table is sometimes presented differently. Observe that

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} = \mathbf{y}^T \mathbf{y} - R(\gamma_2).$$

This is called the *corrected sum of squares*, and  $R(\gamma_2)$  the *correction factor*.

# Corrected sum of squares

We break down the corrected sum of squares into  $R(\gamma_1|\gamma_2)$  and  $SS_{Res}$ , and test using an  $F$  statistic ratio. The end result is the same as before, but the table looks slightly different.

Source of variation	Sum of squares	degrees of freedom	Mean square	F ratio
Regression	$SS_{Reg} - (\sum_{i=1}^n y_i)^2 / n$	$k$	$\frac{R(\gamma_1 \gamma_2)}{k}$	$\frac{R(\gamma_1 \gamma_2)/k}{MS_{Res}}$
Residual	$SS_{Res}$	$n - k - 1$	$\frac{SS_{Res}}{n - k - 1}$	
Total	$\mathbf{y}^T \mathbf{y} - (\sum_{i=1}^n y_i)^2 / n$	$n - 1$		

Some computer software (including R!) will use a corrected sum of squares layout instead of an uncorrected sum, so you should be familiar with both.

## Example: system cost

**Example.** In the data processing example, we rejected the hypothesis that  $[\beta_1 \ \beta_2 \ \beta_3]^T = \mathbf{0}$ . The ANOVA table for a corrected sum of squares test is

Variation	SS	d.f.	MS	F
Regression	17178	3	5726	58.2
Residual	689	7	98	
Total	17867	10		

The actual test does not change: the  $F$  statistic and degrees of freedom are the same.

## Example: system cost

```
> model <- lm(y~X[,-1])
> null <- lm(y~1)
> anova(null,model)
```

### Analysis of Variance Table

Model 1: y ~ 1

Model 2: y ~ X[, -1]

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	10	17866.5				
2	7	688.6	3	17178	58.205	2.578e-05 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



## Clover example — $H_0 : \beta_0 = 0$

We return to the clover example.

```
> X2 <- X[,-1]
> b2 <- solve(t(X2) %*% X2, t(X2) %*% y)
> (SSRes2 <- sum((y - X2 %*% b2)^2))
```

```
[1] 6.296435
```

```
> (Rg2 <- SSTotal - SSRes2)
```

```
[1] 375.0129
```

```
> (Rg2 <- t(y) %*% X2 %*% b2)
```

```
      [,1]
```

```
[1,] 375.0129
```

```
> (Rg1g2 <- SSReg - Rg2)
```

```
      [,1]
```

```
[1,] 1.57437
```

## Clover example — $H_0 : \beta_0 = 0$

```
> r <- 1
> (Fstat <- (Rg1g2/r)/(SSRes/(n-p)))
      [,1]
[1,] 45.34336
> pf(Fstat, r, n-p, lower.tail=FALSE)
      [,1]
[1,] 4.255185e-10
```

# Clover example — $H_0 : \beta_0 = 0$

```
> null <- lm(area ~ 0 + midrib + estim, data=clover)
> anova(null, model)
```

## Analysis of Variance Table

Model 1: area ~ 0 + midrib + estim

Model 2: area ~ midrib + estim

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	137	6.2964				
2	136	4.7221	1	1.5744	45.343	4.255e-10 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Clover example — $H_0 : \beta_1 = 0$

```
> X2 <- X[, -2]
> (Rg2 <- t(y) %*% X2 %*% solve(t(X2) %*% X2) %*% t(X2) %*% y)

      [,1]
[1,] 375.2721

> (Rg1g2 <- SSReg - Rg2)

      [,1]
[1,] 1.315149

> r <- 1
> (Fstat <- (Rg1g2/r)/(SSRes/(n-p)))

      [,1]
[1,] 37.87756

> pf(Fstat, r, n-p, lower.tail=FALSE)

      [,1]
[1,] 7.920166e-09
```

# Clover example — $H_0 : \beta_1 = 0$

```
> null <- lm(area ~ estim, data=clover)
> anova(null, model)
```

Analysis of Variance Table

Model 1: area ~ estim

Model 2: area ~ midrib + estim

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	137	6.0372				
2	136	4.7221	1	1.3152	37.878	7.92e-09 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Clover example — $H_0 : \beta_2 = 0$

```
> X2 <- X[, -3]
> (Rg2 <- t(y) %*% X2 %*% solve(t(X2) %*% X2) %*% t(X2) %*% y)

      [,1]
[1,] 371.9034

> (Rg1g2 <- SSReg - Rg2)

      [,1]
[1,] 4.683866

> r <- 1
> (Fstat <- (Rg1g2/r)/(SSRes/(n-p)))

      [,1]
[1,] 134.8998

> pf(Fstat, r, n-p, lower.tail=FALSE)

      [,1]
[1,] 4.288499e-22
```

# Clover example — $H_0 : \beta_2 = 0$

```
> null <- lm(area ~ midrib, data=clover)
> anova(null, model)
```

Analysis of Variance Table

Model 1: area ~ midrib

Model 2: area ~ midrib + estim

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	137	9.4059				
2	136	4.7221	1	4.6839	134.9	< 2.2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Clover example — $H_0 : \beta_1 = \beta_2 = 0$

```
> X2 <- X[,1]
> (Rg2 <- t(y) %*% X2 %*% solve(t(X2) %*% X2) %*% t(X2) %*% y)

      [,1]
[1,] 310.708

> (Rg1g2 <- SSReg - Rg2)

      [,1]
[1,] 65.87925

> r <- 2
> (Fstat <- (Rg1g2/r)/(SSRes/(n-p)))

      [,1]
[1,] 948.6927

> pf(Fstat, r, n-p, lower.tail=FALSE)

      [,1]
[1,] 1.323441e-80
```



# Clover example — $H_0 : \beta_1 = \beta_2 = 0$

```
> null <- lm(area ~ 1, data=clover)
> anova(null, model)
```

Analysis of Variance Table

Model 1: area ~ 1

Model 2: area ~ midrib + estim

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	138	70.601				
2	136	4.722	2	65.879	948.69	< 2.2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Can you find the $p$ -values?

```
> summary(model)
```

Call:

```
lm(formula = area ~ midrib + estim, data = clover)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.57403	-0.10000	0.00737	0.11681	0.49398

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.38148	0.20516	-6.734	4.26e-10 ***
midrib	0.65037	0.10567	6.154	7.92e-09 ***
estim	0.69199	0.05958	11.615	< 2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1863 on 136 degrees of freedom

Multiple R-squared: 0.9331, Adjusted R-squared: 0.9321

F-statistic: 948.7 on 2 and 136 DF, p-value: < 2.2e-16

# Sequential testing

Suppose that we have a number of explanatory variables in a model, but it's not obvious if all of them are relevant.

We could fit a model using all of them, but this runs the risk of *overfitting*: using irrelevant variables to explain noise by coincidence.

Ideally, we prefer to fit a parsimonious model, i.e. using a minimal number of explanatory variables.

A parsimonious model is less likely to suffer from overfitting.

# Sequential testing

In a parsimonious model, if we were to test if the parameter  $\beta_i$  is 0, in the presence of the other model parameters, we should always reject the null.

We can use the tests we have developed to tell if a model is parsimonious or not.

How do we find such a minimal set of parameters?

# Sequential testing

Conceivably, with the help of a computer, we could test all the possible parameter sets to find the largest  $\gamma_1$  such that the hypothesis  $\gamma_1 = 0$  is not rejected.

The problem with this approach (apart from the time required) is that it can give inconsistent results. For example we might reject  $\beta_1 = \beta_2 = 0$  given  $\beta_3$ , but not reject  $\beta_1 = 0$  given  $\beta_2$  and  $\beta_3$ , and also not reject  $\beta_2 = 0$  given  $\beta_1$  and  $\beta_3$ .

This can happen when  $x_1$  and  $x_2$  are strongly correlated, so that given one of them the other isn't needed, but you need to have at least one of them.

## Partial testing

If we have  $p = k + 1$  parameters  $\beta_0, \dots, \beta_k$  we could consider  $p$  tests of the form  $H_0 : \beta_i = 0$ , given all the other parameters are in the model. Such tests are called *partial* tests.

Then we could remove all parameters where we do not reject  $\beta_i = 0$ .

The discussion above suggests that this could lead us to remove too many variables, *because the partial tests are not independent*.

In a partial test, acceptance or rejection of  $H_0$  does not mean that the parameter is useful or useless in the *best* model, just useful or useless in the *full* model.

# Sequential testing

To avoid the problem of dependence between partial tests we can consider a nested sequence of models.

That is, we can start with a simple model and sequentially add parameters until adding parameters does not significantly improve the fit. Then we have a parsimonious model.

Alternatively we can start with a full model and sequentially remove parameters until removing parameters significantly worsens the fit. Then we again have a parsimonious model.

# Sequential testing

Consider the series of models (subject to relabelling)

$$\begin{aligned}y &= \beta_0 + \varepsilon^{(0)} \\y &= \beta_0 + \beta_1 x_1 + \varepsilon^{(1)} \\&\vdots \\y &= \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon^{(k)}.\end{aligned}$$

We denote the corresponding  $X$  matrices by  $X^{(j)}$ , which are the first  $j + 1$  columns of  $X$ .

The regression sum of squares for each of these models is calculated in the usual way:

$$R(\beta_0, \beta_1, \dots, \beta_j) = \mathbf{y}^T X^{(j)} ((X^{(j)})^T X^{(j)})^{-1} (X^{(j)})^T \mathbf{y}.$$



# Sequential testing

Note that these are 'full' regression sums of squares, i.e. we are looking at the total variation explained by the model in the presence of no other parameters.

Now by taking the difference between the sums of squares, we get the extra variation explained as we add variables to the model one at a time:

$$\begin{aligned}R(\beta_1|\beta_0) &= R(\beta_0, \beta_1) - R(\beta_0) \\R(\beta_2|\beta_0, \beta_1) &= R(\beta_0, \beta_1, \beta_2) - R(\beta_0, \beta_1) \\&\vdots \\R(\beta_k|\beta_0, \beta_1, \dots, \beta_{k-1}) &= R(\boldsymbol{\beta}) - R(\beta_0, \beta_1, \dots, \beta_{k-1}).\end{aligned}$$

# Sequential testing

## Theorem 5.6

*In the full rank general linear model  $\mathbf{y} = X\beta + \varepsilon$ , assume  $\varepsilon \sim N(0, \sigma^2 I)$ . Then*

$$\begin{aligned} \frac{1}{\sigma^2} \mathbf{y}^T \mathbf{y} &= \frac{1}{\sigma^2} SS_{Res} + \frac{1}{\sigma^2} R(\beta_0) + \frac{1}{\sigma^2} R(\beta_1 | \beta_0) + \frac{1}{\sigma^2} R(\beta_2 | \beta_0, \beta_1) + \\ &\quad \cdots + \frac{1}{\sigma^2} R(\beta_k | \beta_0, \beta_1, \dots, \beta_{k-1}), \end{aligned}$$

*and the quadratic forms on the right are all independent with noncentral  $\chi^2$  distributions.  $SS_{Res}$  has  $n - p$  d.f. and the rest have 1 d.f. each.*

# Sequential testing

To prove this theorem, we first prove the following lemma.

## Lemma 5.7

*Let  $A = [A_1 | A_2]$  be a matrix of full rank. Then the matrix  $A(A^T A)^{-1} A^T - A_1(A_1^T A_1)^{-1} A_1^T$  is idempotent.*

### Proof.

First we note that

$$\begin{aligned} A^T [I - A(A^T A)^{-1} A^T] &= A^T - A^T A(A^T A)^{-1} A^T \\ &= 0. \end{aligned}$$

# Sequential testing

Partitioning the first factor,

$$\begin{aligned}\begin{bmatrix} A_1^T \\ A_2^T \end{bmatrix} [I - A(A^T A)^{-1} A^T] &= 0 \\ A_1^T [I - A(A^T A)^{-1} A^T] &= 0 \\ A_1^T &= A_1^T A(A^T A)^{-1} A^T \\ A_1 &= A(A^T A)^{-1} A^T A_1.\end{aligned}$$

# Sequential testing

So

$$\begin{aligned} & (A(A^T A)^{-1} A^T - A_1(A_1^T A_1)^{-1} A_1^T)^2 \\ &= A(A^T A)^{-1} A^T - A(A^T A)^{-1} A^T A_1(A_1^T A_1)^{-1} A_1^T \\ &\quad - A_1(A_1^T A_1)^{-1} A_1^T A(A^T A)^{-1} A^T + A_1(A_1^T A_1)^{-1} A_1^T \\ &= A(A^T A)^{-1} A^T - A_1(A_1^T A_1)^{-1} A_1^T \\ &\quad - A_1(A_1^T A_1)^{-1} A_1^T + A_1(A_1^T A_1)^{-1} A_1^T \\ &= A(A^T A)^{-1} A^T - A_1(A_1^T A_1)^{-1} A_1^T. \end{aligned}$$

# Sequential testing

**Proof of theorem.** The sum follows from the definition. To prove the rest, we use Theorem 3.14.

Let  $X^{(j)}$  be the first  $j + 1$  columns of  $X$ , and

$$\begin{aligned}H_j &= X^{(j)}((X^{(j)})^T X^{(j)})^{-1}(X^{(j)})^T, \\R_j &= \mathbf{y}^T H_j \mathbf{y} = R(\beta_0, \dots, \beta_j).\end{aligned}$$

Then

$$R(\beta_j | \beta_0, \dots, \beta_{j-1}) = R_j - R_{j-1} = \mathbf{y}^T (H_j - H_{j-1}) \mathbf{y}.$$

# Sequential testing

From the lemma,  $H_j - H_{j-1}$  is idempotent. Furthermore both  $H_0$  and  $I - H_p$  are idempotent.

Thus we have a set of idempotent matrices with sum  $I$  (also idempotent). Thus Theorem 3.14 applies and the quadratic forms all have independent noncentral  $\chi^2$  distributions.

To show the degrees of freedom, observe that the sum of the ranks is  $n$  and  $r(I - H_p) = n - p$ . Now there are  $p$  remaining terms with total rank  $p$ . Hence each term has rank 1.

# Sequential testing

Each sequential regression sum of squares has 1 degree of freedom. Therefore under the hypothesis  $\beta_j = 0$ , the test statistic

$$\frac{R(\beta_j | \beta_0, \beta_1, \dots, \beta_{j-1})}{SS_{Res} / (n - p)}$$

has an  $F$  distribution with 1 and  $n - p$  degrees of freedom.

Note that this is still not entirely satisfactory, because the result will depend heavily on the order of the parameters considered. Different orderings can result in different sets of parameters being included in the final model.



# Squid example

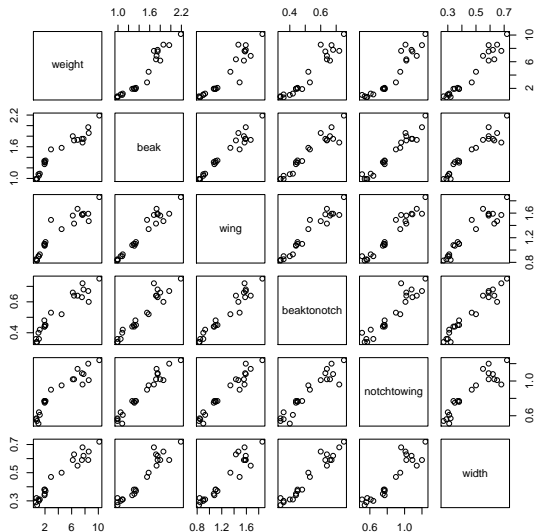
**Example.** An experiment is conducted to study the size of squid. The response is the weight of the squid, and the predictors are

- ▶  $x_1$ : Beak length
- ▶  $x_2$ : Wing length
- ▶  $x_3$ : Beak to notch length
- ▶  $x_4$ : Notch to wing length
- ▶  $x_5$ : Width

A total of 22 squid are sampled.

```
> squid <- read.csv('../data/squid.csv')  
> pairs(squid)
```

# Squid example



## Squid example

Let's first test if any parameters should be in the model, i.e. if  $\beta = \mathbf{0}$ .

```
> n <- dim(squid)[1]
> p <- dim(squid)[2]
> y <- squid$weight
> X <- as.matrix(cbind(rep(1,n),squid[, -1]))
> b <- solve(t(X)%*%X,t(X)%*%y)
> SSRes <- sum((y-X%*%b)^2)
> SSReg <- sum(y^2) - SSRes
> (Fstat <- (SSReg/p)/(SSRes/(n-p)))

[1] 200.4545

> pf(Fstat,p,n-p,lower=F)

[1] 3.879047e-14
```

# Squid example

```
> sqmodel <- lm(weight ~ ., data=squid)
> sqnull <- lm(weight ~ 0, data=squid)
> anova(sqnull, sqmodel)
```

Analysis of Variance Table

Model 1: weight ~ 0

Model 2: weight ~ beak + wing + beaktonotch + notchtowing +

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	22	603.08				
2	16	7.92	6	595.16	200.45	3.879e-14 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The null hypothesis  $\beta = 0$  is rejected strongly.

## Squid example

Next we test to see which parameters should be included in the model.

```
> R <- c()
> for (i in 1:p) {
+   Xi <- X[,1:i]
+   R[i] <- t(y)%*%Xi%%solve(t(Xi)%*%Xi,t(Xi)%*%y)
+ }
> R

[1] 387.1566 586.3019 586.4285 590.5481 590.8116 595.1638
> R - c(0,R[-length(R)])

[1] 387.1565500 199.1453356    0.1266641    4.1195388    0.263
```

## Squid example

Thus the sequential sums of squares are:

$$R(\beta_0) = 387.16$$

$$R(\beta_1|\beta_0) = 199.15$$

$$R(\beta_2|\beta_0, \beta_1) = 0.127$$

$$R(\beta_3|\beta_0, \beta_1, \beta_2) = 4.12$$

$$R(\beta_4|\beta_0, \beta_1, \beta_2, \beta_3) = 0.263$$

$$R(\beta_5|\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = 4.35$$

## Squid example

These sum to the regression sum of squares for the full model:

```
> sum(R - c(0,R[-length(R)]))
```

```
[1] 595.1638
```

```
> SSReg
```

```
[1] 595.1638
```

Each of these sums of squares should be compared against the critical  $F$  value with 1 and  $n - p$  degrees of freedom, multiplied by  $SS_{Res}/(n - p)$ . With  $\alpha = 0.05$ , this is:

```
> SSRes/(n-p)*qf(0.95,1,n-p)
```

```
[1] 2.223833
```

## Squid example

So starting with a model with no parameters, we should definitely add  $\beta_0$  and then  $\beta_1$ , but not  $\beta_2$ .

The subsequent tests are harder to interpret. For example, if  $\beta_0, \beta_1, \beta_2$ , and  $\beta_3$  are in the model, we should not add  $\beta_4$ . But  $\beta_2$  is not in the model!

The tests for  $\beta_3, \beta_4$  and  $\beta_5$  need to be repeated, supposing only that  $\beta_0$  and  $\beta_1$  are in the model.



# Squid example

```
> anova(sqmodel)
```

Analysis of Variance Table

Response: weight

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
beak	1	199.145	199.145	402.4397	9.131e-13	***
wing	1	0.127	0.127	0.2560	0.619804	
beaktonotch	1	4.120	4.120	8.3249	0.010765	*
notchtowing	1	0.263	0.263	0.5325	0.476114	
width	1	4.352	4.352	8.7951	0.009109	**
Residuals	16	7.918	0.495			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Squid example

Note that we use the  $SS_{Res}$  (and residual degrees of freedom) of the *full* model in the denominator of our  $F$  statistics.

This is because we cannot assume that variables that are not in the model are irrelevant. If there are relevant variables,  $SS_{Res}$  of a reduced model may be disproportionately large, and more importantly not conform to our distributional assumptions.

The only way to be safe about this is to use the  $SS_{Res}$  of the full model, even if it means losing a few degrees of freedom to truly irrelevant variables.

## Clover example

We try some sequential tests on the clover example. We test in the order  $\beta_0 \rightarrow \beta_1 \rightarrow \beta_2$ .

```
> R <- c()
> for (i in 1:p) {
+   Xi <- X[,1:i]
+   R[i] <- t(y)%*%Xi%%solve(t(Xi)%*%Xi,t(Xi)%*%y)
+ }
> R - c(0,R[-length(R)])

[1] 310.708028  61.195381   4.683866

> (R - c(0,R[-length(R)]))/(SSRes/(n-p))

[1] 8948.6892 1762.4857  134.8998

> qf(0.95, 1, n-p)

[1] 3.910747
```

# Clover example

```
> model <- lm(area ~ midrib + estim, data=clover)
> nm1 <- lm(area ~ 0, data=clover)
> nm2 <- lm(area ~ 1, data=clover)
> nm3 <- lm(area ~ midrib, data=clover)
```

# Clover example

```
> anova(nm1, nm2, nm3, model)
```

Analysis of Variance Table

Model 1: area ~ 0

Model 2: area ~ 1

Model 3: area ~ midrib

Model 4: area ~ midrib + estim

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	139	381.31				
2	138	70.60	1	310.708	8948.7	< 2.2e-16 ***
3	137	9.41	1	61.195	1762.5	< 2.2e-16 ***
4	136	4.72	1	4.684	134.9	< 2.2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Forward selection

Let's look at some strategies for finding a parsimonious model, using sequential tests.

Forward selection starts off with an empty model, and adds the variable which is found to be most significant.

Significance is measured in relation to the current model, so all tests are conducted in the presence of already included parameters, but not the other parameters.

When no variables are significant enough to add, we stop and take the current model as the final model.

# Forward selection

1. Start with an empty model.
2. Calculate the  $F$ -values for the tests  $H_0 : \beta_i = 0$ , for all parameters not in the model, in the presence of parameters already in the model.
3. If none of the tests are significant (we do not reject any null hypotheses), then stop.
4. Otherwise add the most significant parameter (i.e. parameter with the largest  $F$ -value).
5. Return to step 2.

## Cement example: forward selection

We model the hardening of cement.

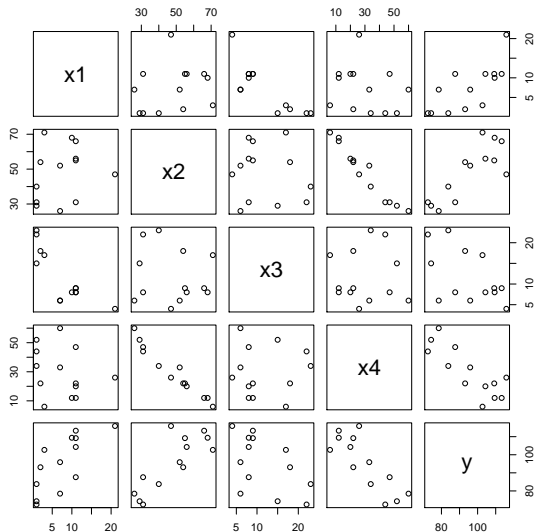
```
> heat <- read.csv("../data/heat.csv")
> str(heat)

'data.frame': 13 obs. of  5 variables:
 $ x1: int  7 1 11 11 7 11 3 1 2 21 ...
 $ x2: int  26 29 56 31 52 55 71 31 54 47 ...
 $ x3: int  6 15 8 8 6 9 17 22 18 4 ...
 $ x4: int  60 52 20 47 33 22 6 44 22 26 ...
 $ y : num  78.5 74.3 104.3 87.6 95.9 ...

> basemodel <- lm(y ~ 1, data=heat)
```



# Cement example: forward selection



# Cement example: forward selection

```
> add1(basemodel, scope= ~ . + x1 + x2 + x3 + x4, test="F")
```

Single term additions

Model:

```
y ~ 1
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)	
<none>			2715.76	71.444			
x1	1	1450.08	1265.69	63.519	12.6025	0.0045520	**
x2	1	1809.43	906.34	59.178	21.9606	0.0006648	***
x3	1	776.36	1939.40	69.067	4.4034	0.0597623	.
x4	1	1831.90	883.87	58.852	22.7985	0.0005762	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> model2 <- lm(y ~ x4, data=heat)
```

# Cement example: forward selection

```
> add1(model2, scope= ~ . + x1 + x2 + x3, test="F")
```

Single term additions

Model:

```
y ~ x4
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)	
<none>			883.87	58.852			
x1	1	809.10	74.76	28.742	108.2239	1.105e-06	***
x2	1	14.99	868.88	60.629	0.1725	0.6867	
x3	1	708.13	175.74	39.853	40.2946	8.375e-05	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> model3 <- lm(y ~ x1 + x4, data=heat)
```

# Cement example: forward selection

```
> add1(model3, scope= ~ . + x2 + x3, test="F")
```

Single term additions

Model:

```
y ~ x1 + x4
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			74.762	28.742		
x2	1	26.789	47.973	24.974	5.0259	0.05169 .
x3	1	23.926	50.836	25.728	4.2358	0.06969 .

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We use variables  $x_1$  and  $x_4$  in our final model.

# Backward elimination

A conceptually similar method is *backward elimination*:

1. Start with the full model.
2. Calculate the  $F$ -values for the tests  $H_0 : \beta_i = 0$ , for all parameters in the model, in the presence of the other parameters in the model.
3. If all of the tests are significant (we reject all null hypotheses), then stop.
4. Otherwise, remove the least significant parameter (i.e. parameter with smallest  $F$ -value).
5. Return to step 2.

# Backward elimination

Backward elimination is complementary to forward selection, i.e. starts from the full model and removes the least important variable until all variables are important.

Forward selection and backward elimination are easy to understand and to apply, but do not always produce the optimal results.

One reason this is so is the inability to remove an already added variable (or add an already removed variable). This inflexibility is often limiting.

# Cement example: backward elimination

```
> fullmodel <- lm(y ~ ., data=heat)
> drop1(fullmodel, scope= ~ ., test="F")
```

Single term deletions

Model:

```
y ~ x1 + x2 + x3 + x4
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			47.864	26.944		
x1	1	25.9509	73.815	30.576	4.3375	0.07082 .
x2	1	2.9725	50.836	25.728	0.4968	0.50090
x3	1	0.1091	47.973	24.974	0.0182	0.89592
x4	1	0.2470	48.111	25.011	0.0413	0.84407

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> model2 <- lm(y ~ x1 + x2 + x4, data=heat)
```

# Cement example: backward elimination

```
> drop1(model2, scope= ~ ., test="F")
```

Single term deletions

Model:

```
y ~ x1 + x2 + x4
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)	
<none>			47.97	24.974			
x1	1	820.91	868.88	60.629	154.0076	5.781e-07	***
x2	1	26.79	74.76	28.742	5.0259	0.05169	.
x4	1	9.93	57.90	25.420	1.8633	0.20540	

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> model3 <- lm(y ~ x1 + x2, data=heat)
```



# Cement example: backward elimination

```
> drop1(model3, scope = ~ ., test="F")
```

Single term deletions

Model:

```
y ~ x1 + x2
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)	
<none>			57.90	25.420			
x1	1	848.43	906.34	59.178	146.52	2.692e-07	***
x2	1	1207.78	1265.69	63.519	208.58	5.029e-08	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We use variables  $x_1$  and  $x_2$  in our final model.

## Stepwise selection

Stepwise selection functions similarly to forward or backward selection, but with the possibility of either adding or eliminating a variable at each step.

In order to assess the appropriateness of a model, we use a *goodness-of-fit* measure.

We give a procedure using a goodness-of-fit measure called Akaike's information criterion (*AIC*), but it is trivial to adjust for any other goodness-of-fit statistic.

# Stepwise selection

1. Start with any model.
2. Compute the  $AIC$  of all models which either have one extra variable or one less variable than the current model.
3. If the  $AIC$  of all such models is more than the  $AIC$  of the current model, stop.
4. Otherwise, change to the model with the lowest  $AIC$ .
5. Return to step 2.

## Stepwise selection

Stepwise selection is generally better than forward or backward selection, because it avoids the problem that an already added variable can never be removed (or the opposite).

However the final model depends on the starting model, so it does not necessarily find a global optimum for the goodness-of-fit statistic. Instead it finds a local optimum.

It is possible for small numbers of variables to find a global minimum through an exhaustive search of all possible combinations. However, as the number of variables increases, this will take too long.

## Goodness-of-fit measures

The  $F$  test is used to compare *nested* models, that is, it requires the variable set of one model to be fully contained in the variable set of the other model.

We cannot use an  $F$  test to compare models which, for example, have replaced one variable with another variable.

Also, use of the  $F$  test requires the somewhat arbitrary choice of a significance level.

To overcome these problems many authors have proposed *goodness-of-fit* measures, which try to give a measure of how good a model is, independently of other models.

## Residual sum of squares

The residual sum of squares,  $SS_{Res}$ , measures how well the model fits the (training) data. However, it is not a good goodness-of-fit measure, as it does not take into account model complexity, and thus can not prevent overfitting.

We can overcome this by using  $s^2$  as a goodness-of-fit statistic. When we add a variable to the model,  $SS_{Res}$  always decreases. However, the degrees of freedom  $n - p$  also decreases, so  $s^2$  will decrease only if the variable is “good”.

Unfortunately, in practice using  $s^2$  for goodness-of-fit does not discourage overfitting enough.

$R^2$ 

A commonly reported goodness-of-fit statistic is the proportion of (corrected) total sums of squares that is explained by the model:

$$R^2 = 1 - \frac{SS_{Res}}{SS_{Total} - (\sum_i y_i)^2 / n}.$$

$R^2$  lies between 0 and 1, and the larger it is, the more variation in  $y$  is explained by the model.

However  $R^2$  can never decrease when we add a variable to a model, as even an irrelevant variable will 'explain' a small extra amount of variation. We would like to remove irrelevant variables, so, like  $SS_{Res}$ ,  $R^2$  is not appropriate for model selection.

## Adjusted $R^2$

The adjusted  $R^2$  tries to account for model complexity by introducing a penalty based on the number of parameters in the model:

$$\text{adj } R^2 = 1 - \frac{n-1}{n-1-k}(1-R^2).$$

Here we assume that  $\beta_0$  is in the model, and  $k$  is the number of other parameters in the model.

The adjusted  $R^2$  is better for model selection than  $s^2$ , but there are other more sophisticated goodness-of-fit measures that we can use, such as the AIC, BIC or Mallows'  $C_p$  statistic.



## Akaike's information criterion

A very popular goodness-of-fit statistic is *Akaike's information criterion*, or AIC. This is based on the likelihood of the observed values of the response.

$$\begin{aligned}AIC &= -2 \ln(\text{likelihood}) + 2p \\&= n \ln \left( \frac{SS_{Res}}{n} \right) + 2p + \text{const.}\end{aligned}$$

(Here the likelihood is the maximised likelihood.) A smaller value of *AIC* indicates a better model.

The form of the AIC can be justified using information theory.

## Bayesian information criterion

Another very similar goodness-of-fit measure is the *Bayesian information criterion*, or BIC. This is calculated as

$$\begin{aligned} BIC &= -2 \ln(\text{likelihood}) + p \ln n \\ &= n \ln \left( \frac{SS_{Res}}{n} \right) + p \ln n + \text{const.} \end{aligned}$$

Again, a smaller value of *BIC* indicates a better model.

The BIC penalises extra parameters more harshly than the AIC, and so promotes a model with fewer variables.

# Mallows' $C_p$

Another goodness-of-fit statistic is *Mallows'  $C_p$  statistic*. This statistic compares the residual sum of squares of an intermediate model against the the residual sum of squares for a full model:

$$C_p = \frac{SS_{Res}(\text{model})}{s^2(\text{full model})} + 2p - n,$$

where  $p$  is the number of parameters in the (intermediate) model.

The smaller  $C_p$  is, the better the model.

# Goodness-of-fit measures

Note that any goodness-of-fit statistic should only to be used to compare various models for the same data. There is no absolute measure of how good a model is, for any of them.

# Cement example: stepwise selection

```
> model2 <- step(basemodel, scope=~.+x1+x2+x3+x4, steps=1)
```

Start: AIC=71.44

y ~ 1

	Df	Sum of Sq	RSS	AIC
+ x4	1	1831.90	883.87	58.852
+ x2	1	1809.43	906.34	59.178
+ x1	1	1450.08	1265.69	63.519
+ x3	1	776.36	1939.40	69.067
<none>			2715.76	71.444

Step: AIC=58.85

y ~ x4

# Cement example: stepwise selection

```
> model3 <- step(model2, scope=~.+x1+x2+x3, steps=1)
```

Start: AIC=58.85

y ~ x4

	Df	Sum of Sq	RSS	AIC
+ x1	1	809.10	74.76	28.742
+ x3	1	708.13	175.74	39.853
<none>			883.87	58.852
+ x2	1	14.99	868.88	60.629
- x4	1	1831.90	2715.76	71.444

Step: AIC=28.74

y ~ x4 + x1

# Cement example: stepwise selection

```
> model4 <- step(model3, scope=~. + x2 + x3, steps=1)
```

Start: AIC=28.74

y ~ x4 + x1

	Df	Sum of Sq	RSS	AIC
+ x2	1	26.79	47.97	24.974
+ x3	1	23.93	50.84	25.728
<none>			74.76	28.742
- x1	1	809.10	883.87	58.852
- x4	1	1190.92	1265.69	63.519

Step: AIC=24.97

y ~ x4 + x1 + x2

# Cement example: stepwise selection

```
> step(model4, scope=~.+x3)
```

Start: AIC=24.97

```
y ~ x4 + x1 + x2
```

	Df	Sum of Sq	RSS	AIC
<none>			47.97	24.974
- x4	1	9.93	57.90	25.420
+ x3	1	0.11	47.86	26.944
- x2	1	26.79	74.76	28.742
- x1	1	820.91	868.88	60.629

Call:

```
lm(formula = y ~ x4 + x1 + x2, data = heat)
```

Coefficients:

(Intercept)	x4	x1	x2
71.6483	-0.2365	1.4519	0.4161



# Cement example: stepwise selection

```
> model2 <- step(fullmodel, scope=~., steps=1)
```

Start: AIC=26.94

```
y ~ x1 + x2 + x3 + x4
```

	Df	Sum of Sq	RSS	AIC
- x3	1	0.1091	47.973	24.974
- x4	1	0.2470	48.111	25.011
- x2	1	2.9725	50.836	25.728
<none>			47.864	26.944
- x1	1	25.9509	73.815	30.576

Step: AIC=24.97

```
y ~ x1 + x2 + x4
```

# Cement example: stepwise selection

```
> step(model2, scope=~.+x3)
```

Start: AIC=24.97

```
y ~ x1 + x2 + x4
```

	Df	Sum of Sq	RSS	AIC
<none>			47.97	24.974
- x4	1	9.93	57.90	25.420
+ x3	1	0.11	47.86	26.944
- x2	1	26.79	74.76	28.742
- x1	1	820.91	868.88	60.629

Call:

```
lm(formula = y ~ x1 + x2 + x4, data = heat)
```

Coefficients:

(Intercept)	x1	x2	x4
71.6483	1.4519	0.4161	-0.2365

## $t$ tests

We can also use a  $t$  test for a partial test of one parameter. That is, to test  $H_0 : \beta_i = 0$  against  $H_1 : \beta_i \neq 0$  in the presence of all the other parameters.

Recall our confidence interval for  $\beta_i$ :

$$b_i \pm t_{\alpha/2} s \sqrt{c_{ii}},$$

where  $c_{ii}$  is the  $(i, i)$ th entry of  $(X^T X)^{-1}$ , and we use a  $t$  distribution with  $n - p$  degrees of freedom.

If this confidence interval includes 0, we do not reject  $H_0$ ; otherwise, we can reject it.

## $t$ tests

In other words, we use the  $t$  statistic (with  $n - p$  degrees of freedom)

$$\frac{b_i}{s\sqrt{c_{ii}}}.$$

Let us compare this with our partial  $F$  test. The statistic we use for this is

$$\frac{R(\beta_i|\beta_0, \beta_1, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_k)}{SS_{Res}/(n - p)}.$$

The denominator is of course  $s^2$ .

## $t$ tests

We saw previously that the numerator is

$$R(\beta_i | \beta_0, \beta_1, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_k) = \hat{\gamma}_1^T A_{11}^{-1} \hat{\gamma}_1$$

where  $\hat{\gamma}_1 = b_i$ , and  $A_{11}$  is the top left element of  $(X^T X)^{-1}$  after the columns have been re-arranged so that the  $i$ th column comes first.

In other words,  $A_{11} = c_{ii}$  and

$$\begin{aligned} R(\beta_i | \beta_0, \beta_1, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_k) &= b_i (c_{ii})^{-1} b_i = \frac{b_i^2}{c_{ii}} \\ \frac{R(\beta_i | \beta_0, \beta_1, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_k)}{s^2} &= \frac{b_i^2}{c_{ii} s^2}. \end{aligned}$$

This is exactly the square of the  $t$  statistic!

## $t$ tests

This is actually not too surprising. The  $t$  distribution can be expressed as a normal variable divided by the square root of a  $\chi^2$  variable.

Therefore when we square it, we get the square of a normal variable divided by a  $\chi^2$  variable. But the square of a normal variable is a  $\chi^2$  variable with 1 d.f.

Therefore the square of a  $t$  variable with  $n$  d.f. is an  $F$  variable with 1 and  $n$  d.f.

This means that the  $t$  test and the  $F$  test are (nearly) identical; the  $t$  test is actually slightly more useful, because it also gives an indication of the sign of the parameter.

## $t$ tests

**Example.** In the previous section, we modelled the amount of a chemical which dissolves in water, when held at a certain temperature. We found that the 95% confidence interval for  $\beta_1$  was

$$0.31 \pm 2.78 \times 0.86\sqrt{0.00057} = 0.31 \pm 2.78 \times 0.021 = [0.25, 0.36].$$

A  $t$  test would use the statistic

$$\frac{b_1}{s\sqrt{c_{11}}} = \frac{0.31}{0.021} = 14.89$$

against a  $t$  distribution with  $n - p = 6 - 2 = 4$  degrees of freedom.

This rejects the hypothesis  $\beta_1 = 0$  at the 0.05 level (critical value 2.78). We can also say that  $\beta_1$  is almost certainly positive.

## $t$ tests

On the other hand, if we use an  $F$  test:

```
> (Rb <- t(y)%*%X)%*%b)
      [,1]
[1,] 663.771

> (Rb0 <- t(y)%*%X[,1]%*%solve(t(X[,1])%*%X[,1],t(X[,1])%*%y))
      [,1]
[1,] 498.6817

> (Rb1_b0 <- Rb - Rb0)
      [,1]
[1,] 165.0893

> (Fstat <- Rb1_b0/s^2)
      [,1]
[1,] 221.6672
```



## $t$ tests

```
> pf(Fstat,1,df,lower=F)
```

```
      [,1]
```

```
[1,] 0.0001185219
```

```
> sqrt(Fstat)
```

```
      [,1]
```

```
[1,] 14.88849
```

```
> pt(sqrt(Fstat),df,lower=F)*2
```

```
      [,1]
```

```
[1,] 0.0001185219
```

# $t$ tests

```
> summary(lm(y~X[, -1]))
```

```
Call:
```

```
lm(formula = y ~ X[, -1])
```

```
Residuals:
```

1	2	3	4	5	6
0.661905	-0.009524	-1.480952	0.547619	0.076190	0.204762

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.43810	0.62459	2.302	0.082715 .
X[, -1]	0.30714	0.02063	14.888	0.000119 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.863 on 4 degrees of freedom
```

```
Multiple R-squared:  0.9823, Adjusted R-squared:  0.9778
```

```
F-statistic: 221.7 on 1 and 4 DF,  p-value: 0.0001185
```

# Shrinkage

Not all selection procedures employ sequential addition and/or deletion of variables. Some go for a more holistic approach.

A common approach is to try and 'shrink' all fitted parameters toward 0, so that irrelevant variables have little or no effect on the model.

Some of the fitted parameters might actually become 0, and the associated variables can then be removed.

# Shrinkage

For example, *ridge regression* uses a penalised least squares approach. Here we minimise the residual sum of squares, but include a term which penalises the size of the parameters. We choose  $\mathbf{b}$  to minimise

$$\sum_{i=1}^n e_i^2 + \lambda \sum_{j=0}^k b_j^2.$$

The  $\lambda$  term controls the amount of 'shrinkage' of the parameters. The penalized least squares estimators can be calculated to be

$$\mathbf{b} = (X^T X + \lambda I)^{-1} X^T \mathbf{y}.$$

This approach will never shrink parameters to 0.

# Shrinkage

Another approach is the LASSO (Least Absolute Shrinkage and Selection Operator), which minimises

$$\sum_{i=1}^n e_i^2 + \lambda \sum_{j=0}^k |b_j|.$$

The LASSO actually shrinks small parameters to 0, and can be used for variable selection by removing those variables.

Choosing an appropriate shrinking parameter  $\lambda$  is quite involved. A common method is *cross-validation*, which estimates the predictive power of the model by removing parts of the dataset and using them as test sets.