MAST90104 - Lecture 5

Weichang Yu

Room 108, Old Geology South Bldg
School of Mathematics and Statistics, University of Melbourne

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

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

where $\mathbf{X}$ is $n \times p$, $n \geq p$, $r(\mathbf{X}) = p$, p equals the number of regression coefficients. Following assumptions (III) and (IV), the errors $\boldsymbol{\epsilon}$ have:

- mean $\mathbf{0}$;
- variance $\sigma^2\mathbf{I}$;

Moreover, we can enforce an even stronger distributional assumption (V) on the random errors: $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

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 predictor variables (and the intercept) have any relevance for predicting y, then all the parameters $\boldsymbol{\beta}$ will be $\mathbf{0}$.

We test for this using the null hypothesis

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

The full rank model

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

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

ANOVA

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

If $\boldsymbol{\beta} = \mathbf{0}$, then $\mathbf{y} = \boldsymbol{\varepsilon}$ 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 $\boldsymbol{\beta} \neq \mathbf{0}$, then $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. 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.

Under $\boldsymbol{\beta} = \mathbf{0}$, total error: $\mathbf{y}^T\mathbf{y}$.

Decompose this total error as:

$$\mathbf{y}^T\mathbf{y} = \underbrace{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}_{=SS_{Res}} + (\mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{X}\widehat{\boldsymbol{\beta}}) \tag{1}$$

The first term on the right of equation (1) is the residual sum of squares, SS_{Res}, which can be expressed as

$$SS_{Res} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(y - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat matrix that takes the response variable $\mathbf{y}$ to its fitted value.

The second term on the right of (1) is

$$(\mathbf{X}\widehat{\beta})^T(\mathbf{X}\widehat{\beta}) = \hat{\mathbf{y}}^T\hat{\mathbf{y}} = \mathbf{y}^T\mathbf{H}\mathbf{y} = \hat{\mathbf{y}}^T\mathbf{X}\widehat{\beta}, \tag{2}$$

the equalities using the fact that $\mathbf{H}$ is idempotent and symmetric.

This second term is called the *regression sum of squares* and denoted 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:

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

Example. Consider a very exceptional situation where $\sigma^2 = 0$. Then, $\mathbb{P}(\mathbf{y} = \mathbf{X}\beta) = 1$. Then

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

and $SS_{Res} = 0$.

On the other hand, suppose we know that there is no signal and the true intercept is zero, so that $\boldsymbol{\beta} = \mathbf{0}$ and hence $\mathbf{y} = \boldsymbol{\varepsilon}$. Then, naturally $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} = \mathbf{0}$ and hence

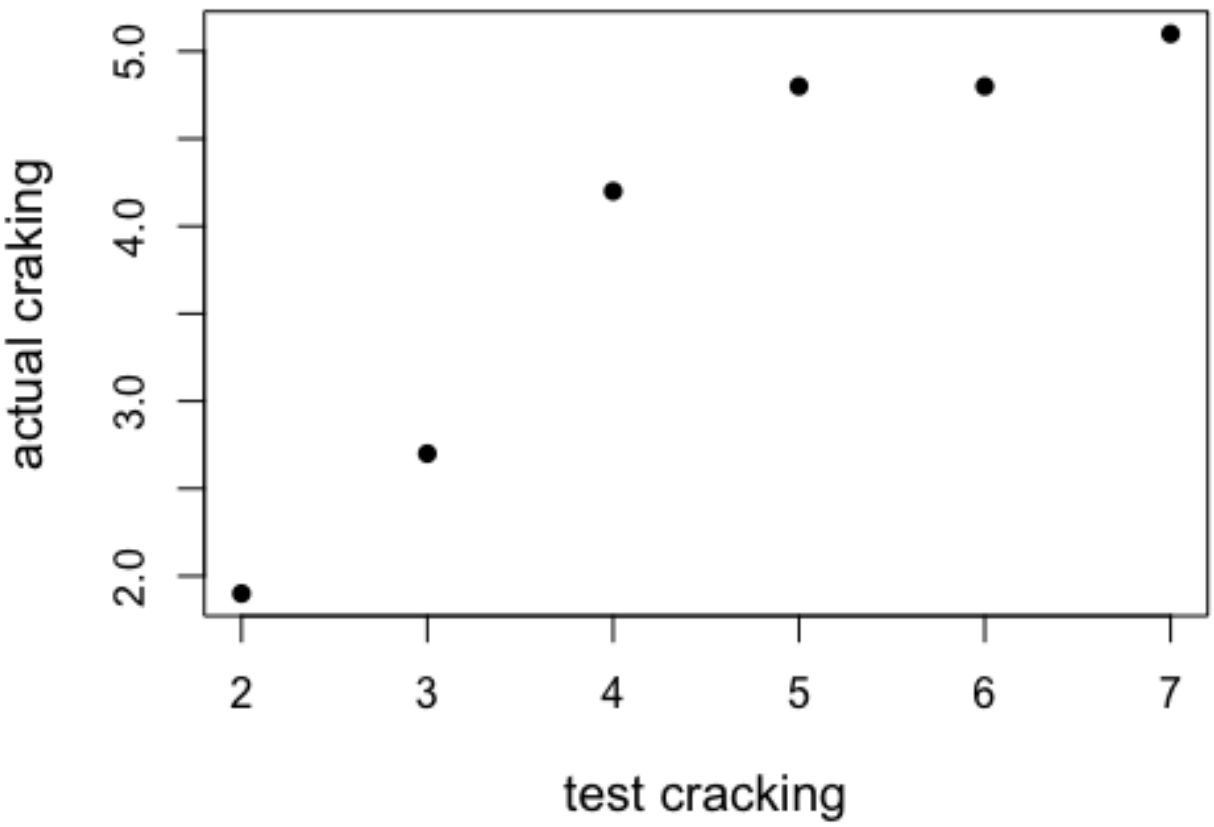
$$\begin{aligned} SS_{Res} &= (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= \mathbf{y}^T\mathbf{y} = SS_{Total} \end{aligned}$$

and $SS_{Reg} = 0$.

These two examples are opposite ends of the spectrum.

Paint Example - ANOVA

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



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

## [1] 1.096762

(SSTotal <- sum(actual^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 $H_0 : \boldsymbol{\beta} = \mathbf{0}$, we compare SS_{Res} with SS_{Reg} of the full regression model. If SS_{Reg} is large compared to SS_{Res}, then we have evidence that H_0 is not true.

To know exactly *how* large, we must first derive the distributions of SS_{Reg} and SS_{Res}. Of course, we already know the latter (which we re-state).

Theorem 5.1

- (Theorem 4.13 of Lecture 4 Part II) Assume (I), (II) and (V) holds. In the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, SS_{Res}/σ^2 *has a* χ^2 *distribution with* $n - p$ *degrees of freedom.*

Theorem 5.2

Assume (I), (II) and (V) holds. In the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, SS_{Reg}/σ^2 *has a noncentral* χ^2 *distribution with* p *degrees of freedom and noncentrality parameter*

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

Proof of theorem 5.2: A straightforward application of Theorem 3.9.

Theorem 5.3

Assume (I), (II) and (V) holds. In the general linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, SS_{Res} *and* SS_{Reg} *are independent.*

Proof: A straightforward application of Theorem 3.12.

Now how do we test $H_0 : \boldsymbol{\beta} = \mathbf{0}$? Note that the test $H_0 : \boldsymbol{\beta} = \mathbf{0}$ against $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$ is known as the *model relevance test*.

Observe that if the null is true, then noncentrality parameter of SS_{Reg}/σ^2 equals to $\frac{1}{2\sigma^2}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0$.

Thus, under H_0,

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

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

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

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

(Recall the expectation of a non-central chi-square random variable.)

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

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

Therefore, we should reject H_0 if the statistic is large and not reject it otherwise, with the critical value determined from the F distribution with p and $n - p$ degrees of freedom.

ANOVA

To lay out all the calculations, we can use an ANOVA table.

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

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

The R output to test this directly from the calculations follows.

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

```
##              [,1]
## [1,] 1.9617795
## [2,] 0.1177586
## [3,] 0.1767263
## [4,] 0.7964477
```

```r
(SSReg <- t(y)%*%X%*%betahat)
```

```
##           [,1]
## [1,] 38978.38
```

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

```
## [1] 39667.01
```

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

```
##           [,1]
## [1,] 688.6262
```

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

```
##           [,1]
## [1,] 9744.596
```

```r
(MSRes <- SSRes/(n-p))
```

```
##           [,1]
## [1,] 98.37517
```

```r
(Fstat <- MSReg/MSRes)
```

```
##           [,1]
## [1,] 99.05544
```

```r
qf(0.95,p,n-p)
```

```
## [1] 4.120312
```

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

```
##               [,1]
## [1,] 3.060186e-06
```

Intercept-only model

In most situations, the true model will contain a non-zero intercept even if there are no predictors.

We would to specify the null hypothesis as $H_0 : \beta_1 = \ldots = \beta_k = 0$. Under H_0, the model is

$$y_i = \beta_0 + \epsilon_i.$$

The above is the *intercept-only model* that is stripped of all predictor variables.

Under H_1, the model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i.$$

Intercept-only model

Under the full model, the estimate for $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Hence, residual sum of squares under H_1 is

$$SS_{Res}^{H_1} = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y}.$$

ANOVA: Corrected Sum of Squares

Matrix form of the intercept-only model:

$$\mathbf{y} = \mathbf{1}\beta_0 + \boldsymbol{\epsilon},$$

where $\mathbf{1}$ is a column vector of 1's. Hence $\mathbf{X}$ matrix under H_0 is $\mathbf{1}$.

Under H_0, our least squares estimate for β_0 is

$$\widehat{\beta}_0 = (\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{y} = \overline{y}.$$

Hence, residual sum of squares under H_0 is

$$SS_{Res}^{H_0} = \mathbf{y}^T\mathbf{y} - \tfrac{1}{n}\mathbf{y}^T\mathbf{1}\mathbf{1}^T\mathbf{y}$$

The residual sum of squares of the intercept-only model is known as corrected sum of squares.

Model parsimony and Occam's razor

Should we prefer a simpler model (H_0) over a more complicated model (H_1)?
According to Occam's razor principle: The parsimonious (simpler) model should be preferred unless there is significant evidence to choose otherwise.

General linear hypothesis

Is there significant reason (evidence) for us to choose the full model over the intercept-only model?

Example of other parsimonious models:

- $H_0 : \beta_2 = 0$
- $H_0 : \beta_1 = -1; \quad \beta_3 = 0$
- $H_0 : \beta_1 + \beta_2 - \beta_3 = 0$
- $H_0 : \beta_1 = \beta_2 = \beta_3; \quad \beta_4 = \beta_5 = 2\beta_6$

We can frame all of above as hypotheses about linear combinations of $\boldsymbol{\beta}$, i.e.,

$$H_0 : \mathbf{L}\boldsymbol{\beta} = \boldsymbol{\delta}.$$

General linear hypothesis

For the hypothesis: $\beta_2 = 0$, we have

$$\mathbf{L} = (0, 0, 1, \mathbf{0}_{k-2}^T), \quad \boldsymbol{\delta} = 0.$$

For the hypothesis: $\beta_1 = -1$ and $\beta_3 = 0$, we have

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 & \mathbf{0}_{k-3}^T \\ 0 & 0 & 0 & 1 & \mathbf{0}_{k-3}^T \end{pmatrix} \quad \boldsymbol{\delta} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

For the hypothesis: $\beta_1 + \beta_2 - \beta_3 = 0$, we have

$$\mathbf{L} = (0, 1, 1, -1, \mathbf{0}_{k-3}^T), \quad \boldsymbol{\delta} = 0.$$

General linear hypothesis

It can be shown that $SS_{res}^{H_0} \geq SS_{res}^{H_1}$ (why?)

However, if the parsimonious model (H_0) has similar accuracy to the full-rank model, then we should expect $SS_{res}^{H_0} - SS_{res}^{H_1}$ to be small.

General linear hypothesis

Theorem 5.4

Assume (I), (II) and (V) holds. For a general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, *consider the pair of hypotheses*

$$H_0 : \mathbf{L}\boldsymbol{\beta} = \boldsymbol{\delta} \text{ vs. } H_1 : \mathbf{L}\boldsymbol{\beta} \neq \boldsymbol{\delta},$$

where $\mathbf{L}$ *is an* $r \times p$ *matrix of rank* $r \leq p$ *and* $\boldsymbol{\delta}$ *is an* $r \times 1$ *vector of constants. The difference between the residual sum of squares of the two models is*

$$SS_{res}^{H_0} - SS_{res}^{H_1} = (\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})^T [\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})$$

Moreover $\left(SS_{res}^{H_0} - SS_{res}^{H_1}\right)/\sigma^2$ *follows a noncentral* χ^2 *distribution with* r *degrees of freedom and noncentrality parameter*

$$\frac{(\mathbf{L}\boldsymbol{\beta} - \boldsymbol{\delta})^T [\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]^{-1}(\mathbf{L}\boldsymbol{\beta} - \boldsymbol{\delta})}{2\sigma^2}$$

General linear hypothesis

Proof of Theorem 5.4: Note that the minimiser of the least squares criterion under the null model is

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\,:\,\mathbf{L}\boldsymbol{\beta}=\boldsymbol{\delta}}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Following the Lagrange multiplier method for constrained optimisation (beyond the scope of this course), the closed form expression for the minimiser is

$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T(\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T)^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})$$

Following the proof of theorem 4.1, we have

$$\begin{aligned} & SS_{res}^{H_0} - SS_{res}^{H_1} \\ = \; & C(\widetilde{\boldsymbol{\beta}}) - C(\widehat{\boldsymbol{\beta}}) \\ = \; & (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X} (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}) \\ = \; & (\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})^T (\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T)^{-1} \mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T (\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T)^{-1} (\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta}) \\ = \; & (\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})^T (\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T)^{-1} (\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta}) \end{aligned}$$

Now,

$$\frac{SS_{res}^{H_0} - SS_{res}^{H_1}}{\sigma^2} = (\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})^T (\sigma^2 \mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T)^{-1} (\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})$$

Observe that

$$\begin{aligned} & \mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta} \\ & = \mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - \boldsymbol{\delta} \sim \text{MVN}(\mathbf{L}\boldsymbol{\beta} - \delta, \sigma^2 \mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T). \end{aligned}$$

Also, note that

$$(\sigma^2\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T)^{-1}\sigma^2\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T = \mathbf{I}_r$$

is idempotent.

Hence, by Theorem 3.9, $\left(SS_{res}^{H_0} - SS_{res}^{H_1}\right)/\sigma^2$ follows a noncentral χ^2 distribution with degrees of freedom $r(\mathbf{I}_r) = r$ and noncentrality parameter

$$\tfrac{1}{2}(\mathbf{L}\boldsymbol{\beta} - \delta)^T(\sigma^2\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T)^{-1}(\mathbf{L}\boldsymbol{\beta} - \delta).$$

General linear hypothesis

Theorem 5.5

Assume (I), (II) and (V) holds. For a general linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, *consider the pair of hypotheses*

$$H_0 : \mathbf{L}\beta = \boldsymbol{\delta} \text{ vs. } H_1 : \mathbf{L}\beta \neq \boldsymbol{\delta},$$

where $\mathbf{L}$ *is an* $r \times p$ *matrix of rank* $r \leq p$ *and* $\boldsymbol{\delta}$ *is an* $r \times 1$ *vector of constants. Under the null hypothesis,* $SS_{res}^{H_1}$ *and* $SS_{res}^{H_0} - SS_{res}^{H_1}$ *are independent.*

The proof is left as an exercise. Hint: Express $SS_{res}^{H_0} - SS_{res}^{H_1}$ as a quadratic form in $\mathbf{y}$ plus a scalar. Then, use theorem 3.12.

General linear hypothesis

If the null hypothesis is true, then $\mathbf{L}\boldsymbol{\beta} = \boldsymbol{\delta}$ and thus $(SS_{res}^{H_0} - SS_{res}^{H_1})/\sigma^2$ has a χ^2 distribution.

Following Theorem 5.5,

$$\frac{(\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})^T[\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})/r}{SS_{Res}^{H_1}/(n-p)}$$

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

We use this statistic to test $H_0 : \mathbf{L}\boldsymbol{\beta} = \boldsymbol{\delta}$ vs. $H_1 : \mathbf{L}\boldsymbol{\beta} \neq \boldsymbol{\delta}$

Test statistic

To justify rejecting the null hypothesis for large values of the test statistic, the expected value of the numerator can be calculated to be

$$E\left[\frac{(\mathbf{L}\widehat{\boldsymbol{\beta}}-\boldsymbol{\delta})^T[\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}}-\boldsymbol{\delta})}{r}\right]$$
$$= \sigma^2+\frac{1}{r}(\mathbf{L}\boldsymbol{\beta}-\boldsymbol{\delta})^T[\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]^{-1}(\mathbf{L}\boldsymbol{\beta}-\boldsymbol{\delta}),$$

where $[\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]^{-1}$ is positive definite.

If the null hypothesis is true, then the expectation is σ^2. However, if H_0 is false, it will be generally be greater than σ^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 $H_0 : \boldsymbol{\beta} = \begin{bmatrix} 2 & 0 & 0 & 1 \end{bmatrix}^T$.

```
delta <- c(2,0,0,1)
L <- diag(4)
r <- 4
num <- t(L%*%betahat-delta)%*%solve(L%*%solve(t(X)%*%X)%*%t(L))%*%
(L%*%betahat-delta)/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, as confirmed also by the p-value of 0.11.

This doesn't mean that H_0 is true, it just means we have insufficient evidence to reject H_0.

Example: system cost

```
library(car)
delta <- c(2,0,0,1)
L <- diag(4)
linearHypothesis(model,L,delta)

## Linear hypothesis test
## Hypothesis:
## (Intercept) = 2
## X[, - 1 = 0
## X[, - 2 = 0
## X[, - 3 = 1
## Model 1: restricted model
## Model 2: y ~ 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 $\beta_1 = \beta_2 = \beta_3$.

This is equivalent to testing $H_0 : \mathbf{L}\boldsymbol{\beta} = \boldsymbol{\delta}$, where

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad \boldsymbol{\delta} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

```r
delta <- c(0,0)
L <- matrix(c(0,0,1,0,-1,1,0,-1),2,4)
r <- 2
num <- t(L%*%betahat-delta)%*%solve(L%*%solve(t(X)%*%X)%*%t(L))%*%
(L%*%betahat-delta)/r
(Fstat <- num/(SSRes/(n-p)))
```

```
##          [,1]
## [1,] 5.785777
```

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

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

Example: system cost

The calculations can be done by hand here:

$$\begin{aligned} \mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta} &= \begin{bmatrix} -0.06 \\ -0.62 \end{bmatrix} \\ \mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T &= \begin{bmatrix} 0.013 & 0.0024 \\ 0.0024 & 0.00077 \end{bmatrix} \\ (\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})^T[\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta}) &= 1138.35. \end{aligned}$$

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 β_1, β_2, and β_3 are not identical, but not strong evidence.

The next slide has the output using the package `car` and command `linearHypothesis`.

```r
delta <- c(0,0)
L <- matrix(c(0,0,1,0,-1,1,0,-1),2,4)
linearHypothesis(model,L,delta)
```

```
## Linear hypothesis test
## Hypothesis:
## X[, - 1 - X[, - 2 = 0
## X[, - 2 - X[, - 3 = 0
##
## Model 1: restricted model
## Model 2: y ~ 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
```

Back to clover example

For the clover data, consider the null hypothesis

$$H_0 : \beta_0 = -1; \beta_1 = 0.5; \beta_2 = 1$$

Here, we have $\boldsymbol{\delta} = (-1, 0.5, 1)^T$ and $\mathbf{L} = \mathbf{I}_3$. Hence,

$$(\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})^T[\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T]^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})^T\mathbf{X}^T\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})$$

```
delta <- as.vector(c(-1, 0.5, 1))
Fstat <- ((t(betahat-delta) %*% t(X) %*% X %*% (betahat-delta)/p)/
(SSRes/(n-p)) )

##          [,1]
## [1,] 317.6183

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

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

Using anova

```
h0 <- X %*% delta
basemodel <- lm(area ~ 0, data=clover, offset=h0)
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 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
```

$H_0 : (\beta_0, \beta_1, \beta_2) = (-1.1, 0.5, 0.7)$

Now, consider the null hypothesis for the clover data

$$H_0 : \beta_0 = -1.1; \beta_1 = 0.5; \beta_2 = 0.7$$

```
delta <- as.vector(c(-1.1, 0.5, 0.7))
Fstat <- ((t(betahat-delta) %*% t(X) %*% X %*% (betahat-delta))/p)/(SSRes/(n-p))
Fstat

##          [,1]
## [1,] 21.37493

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

##              [,1]
## [1,] 2.10218e-11
```

$H_0 : (\beta_0, \beta_1, \beta_2) = (-1.1, 0.5, 0.7)$ using anova

```
h0 <- X %*% delta
basemodel <- lm(area ~ 0, data=clover, offset=h0)
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 6.9485
## 2    136 4.7221  3    2.2265 21.375 2.102e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \beta_0 = -1, \beta_1 = \beta_2$

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

```
 L <- 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(L)[1])
```

```
## [1] 2
```

```
delta <- c(-1,0)
```

$H_0 : \beta_0 = -1, \beta_1 = \beta_2$

```
Fstat <- (t(L %*% betahat - delta) %*%
solve(L %*% solve(t(X) %*% X) %*% t(L)) %*%
(L %*% betahat - delta)/r)/(SSRes/(n-p))

##           [,1]
## [1,] 19.54309

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

##               [,1]
## [1,] 3.463526e-08
```

```r
linearHypothesis(model, L, delta)
```

```
## 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 $\boldsymbol{\beta}$ is $\mathbf{0}$

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

If a particular β_j 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 $\boldsymbol{\beta}$ is $\mathbf{0}$

We split the parameter vector

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{r-1} \\ \hline \beta_r \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \hline \boldsymbol{\gamma}_2 \end{bmatrix}$$

and test the hypotheses

$$H_0 : \boldsymbol{\gamma}_1 = \mathbf{0} \text{ vs. } H_1 : \boldsymbol{\gamma}_1 \neq \mathbf{0}.$$

We can always relabel the cofficients and rearrange the columns of $\mathbf{X}$ to consolidate the appropriate coefficient into $\boldsymbol{\gamma}_1$.

Testing if part of β is $\mathbf{0}$

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

Let $\mathbf{L} = [\mathbf{I}_r|\mathbf{0}]$ and $\boldsymbol{\delta} = \mathbf{0}$. Then $\mathbf{L}\boldsymbol{\beta} = \boldsymbol{\delta}$ iff $\boldsymbol{\gamma}_1 = \mathbf{0}$.

We define the regression sum of squares for $\boldsymbol{\gamma}_1$ in the presence of $\boldsymbol{\gamma}_2$ as

$$\begin{aligned} R(\boldsymbol{\gamma}_1|\boldsymbol{\gamma}_2) &= (\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta})^T(\mathbf{L}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{L}^T)^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\delta}) \\ &= \hat{\boldsymbol{\gamma}}_1{}^T A_{11}^{-1}\hat{\boldsymbol{\gamma}}_1, \end{aligned}$$

where $\hat{\boldsymbol{\gamma}}_1$ is the least squares estimator for $\boldsymbol{\gamma}_1$, and $\mathbf{A}_{11}$ is the upper left $r \times r$ matrix of $(\mathbf{X}^T\mathbf{X})^{-1}$.

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

Our test statistic is

$$\frac{R(\boldsymbol{\gamma}_1|\boldsymbol{\gamma}_2)/r}{SS_{Res}/(n-p)}.$$

From our previous results on the general linear hypothesis, we know that this has an $F_{r,n-p}$ distribution under the null hypothesis $\boldsymbol{\gamma}_1 = \mathbf{0}$. We reject the null when this statistic is too large.

Testing if part of β is $\mathbf{0}$

Theorem 5.6

Let X be a $n \times p$ full rank matrix. Let $\mathbf{X}$ and β be partitioned as

$$\mathbf{X} = \left[\begin{array}{c|c} \mathbf{X}_1 & \mathbf{X}_2 \end{array} \right], \beta = \left[\frac{\gamma_1}{\gamma_2} \right],$$

where $\mathbf{X}_1$ is $n \times r$ and γ_1 is $r \times 1$. Then

$$R(\gamma_1 | \gamma_2) = \hat{\gamma_1}^T A_{11}^{-1} \hat{\gamma_1}$$

where $\hat{\gamma_1}$ is the least squares estimator for γ_1, and A_{11} is the upper left $r \times r$ matrix of $(\mathbf{X}^T\mathbf{X})^{-1}$:

$$A_{11}^{-1} = \mathbf{X}_1^T\mathbf{X}_1 - \mathbf{X}_1^T\mathbf{X}_2(\mathbf{X}_2^T\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{X}_1.$$

Proof is beyond the scope of the course.

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

There is a simpler way to think about the regression sum of squares which follows after a lot of algebra:

Theorem 5.7

$$R(\boldsymbol{\gamma}_1|\boldsymbol{\gamma}_2) = R(\boldsymbol{\beta}) - R(\boldsymbol{\gamma}_2),$$

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = [\mathbf{X}_1|\mathbf{X}_2] \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{bmatrix} + \boldsymbol{\varepsilon},$$

and $R(\boldsymbol{\gamma}_2)$ is the regression sum of squares for the reduced model

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\gamma}_2 + \boldsymbol{\varepsilon}.$$

Relationship between R and SS_{res}

For a test of

$$H_0 : \boldsymbol{\gamma}_1 = \mathbf{0}, \quad H_1 : \boldsymbol{\gamma}_1 \neq \mathbf{0},$$

the following equality is true:

$$R(\boldsymbol{\gamma}_1 | \boldsymbol{\gamma}_2) = R(\boldsymbol{\beta}) - R(\boldsymbol{\gamma}_2) = SS_{res}^{H_0} - SS_{res}^{H_1}.$$

Testing if part of β is $\mathbf{0}$

We again express the test calculations in an ANOVA table. Note that SS_{res} refers to $SS_{res}^{H_1}$ in this table.

Source of variation	Sum of squares	degrees of freedom	Mean square	F ratio
Regression				
Full model	$R(\beta)$	p		
Reduced model	$R(\gamma_2)$	$p-r$		
γ_1 in presence of γ_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		

Exercise: show that $R(\gamma_2)$, $R(\gamma_1|\gamma_2)$ and SS_{Res} are all independent.

Example: system cost (revisiting the intercept-only model)

Example. Consider again the data processing system example. We rejected the hypothesis of model relevance, $\boldsymbol{\beta} = \mathbf{0}$. But that is obvious because the cost of all the systems can'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 β_1, β_2, or β_3 nonzero?

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

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \hline \beta_0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \hline \boldsymbol{\gamma}_2 \end{bmatrix}.$$

Example: system cost (revisiting the intercept-only model)

We must rearrange the columns of $\mathbf{X}$ correspondingly:

$$\mathbf{X} = \left[\begin{array}{ccc|c} 4 & 44 & 18 & 1 \\ 2 & 33 & 15 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 5 & 48 & 17 & 1 \end{array}\right] = \left[\begin{array}{c|c} \mathbf{X}_1 & \mathbf{X}_2 \end{array}\right].$$

We want to test $H_0 : \boldsymbol{\gamma}_1 = \mathbf{0}$ (only the intercept is relevant) against $H_1 : \boldsymbol{\gamma}_1 \neq \mathbf{0}$. The reduced model is the intercept-only model

$$\mathbf{y} = \mathbf{X}_2\beta_0 + \boldsymbol{\varepsilon}_2,$$

i.e independent normal errors with constant mean.

Example: system cost (revisiting the intercept-only model)

```
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 (revisiting the intercept-only model)

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
Regression				
Full	38978	4		
Reduced	21800	1		
γ_1 in presence of γ_2	17178	3	5726	58.2
Residual	689	7	98	
Total	39667	11		

Corrected sum of squares

In general, we have the following ANOVA table for the test $H_0 : \beta_1 = \cdots = \beta_k = 0$ versus the alternative that some $\beta_i \neq 0$, $i \in \{1, \ldots, k\}$.

Source of variation	Sum of squares	degrees of freedom	Mean square	F ratio
Regression				
Full model	$R(\boldsymbol{\beta}) = \mathbf{y}^T H \mathbf{y}$	$k+1$		
Reduced model	$\left(\sum_{i=1}^n y_i\right)^2 / n$	1		
$\boldsymbol{\gamma}_1$ in presence of $\boldsymbol{\gamma}_2$	$R(\boldsymbol{\gamma}_1 \vert \boldsymbol{\gamma}_2)$	k	$\frac{R(\boldsymbol{\gamma}_1 \vert \boldsymbol{\gamma}_2)}{k}$	$\frac{R(\boldsymbol{\gamma}_1) \vert \boldsymbol{\gamma}_2)/k}{MS_{Res}}$
Residual	$\mathbf{y}^T \mathbf{y} - R(\boldsymbol{\beta})$	$n-k-1$	$\frac{SS_{Res}}{n-p}$	
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

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

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	$SS_{Reg} - \left(\sum_{i=1}^{n} y_i\right)^2 / n$	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	$\mathbf{y}^T\mathbf{y} - \left(\sum_{i=1}^{n} y_i\right)^2 / n$	$n-1$		

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 $\begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \end{bmatrix}^T = \mathbf{0}$. 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.

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 β_i:

$$\widehat{\beta}_i \pm t_{\alpha/2} s \sqrt{c_{ii}},$$

where c_{ii} is the (i, i)th entry of $(\mathbf{X}^T\mathbf{X})^{-1}$, 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.

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

$$\frac{\widehat{\beta}_i}{s\sqrt{c_{ii}}}.$$

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

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

The denominator is of course s^2.

We saw previously that the numerator is

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

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

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

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

This is exactly the square of the t statistic!

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

Therefore when we square it, we get the square of a normal variable divided by a χ^2 variable. But the square of a normal variable is a χ^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.

Clover example — $H_0 : \beta_0 = 0$

We return to the clover example.

```
X2 <- X[,-1]
betahat2 <- solve(t(X2) %*% X2, t(X2) %*% y)
(SSRes2 <- sum((y - X2 %*% betahat2)^2))

## [1] 6.296435

Rg2 <- SSTotal - SSRes2

## [1] 375.0129

Rg2 <- t(y) %*% X2 %*% betahat2

##           [,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
```

Example. In Lecture 4 Part II, 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 β_1 was

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

A t test would use the statistic

$$\frac{\widehat{\beta}_1}{s\sqrt{c_{11}}} = \frac{0.31}{0.86\sqrt{0.00057}} = 14.89$$

using 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 β_1 is almost certainly positive.

t tests

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

```
(Rb <- t(y)%*%X%*%betahat)

##          [,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
```

The critical value of the F distribution with 1 and 4 degrees of freedom is $7.71 = 2.77^2$. So we can again reject the null hypothesis of $\beta_1 = 0$.

Variation	SS	d.f.	MS	F
Regression				
Full	663.77	2		
Reduced	498.68	1		
β_1 in presence of β_0	165.09	1	165.09	221.7
Residual	2.98	4	0.74	
Total	666.75	6		

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.

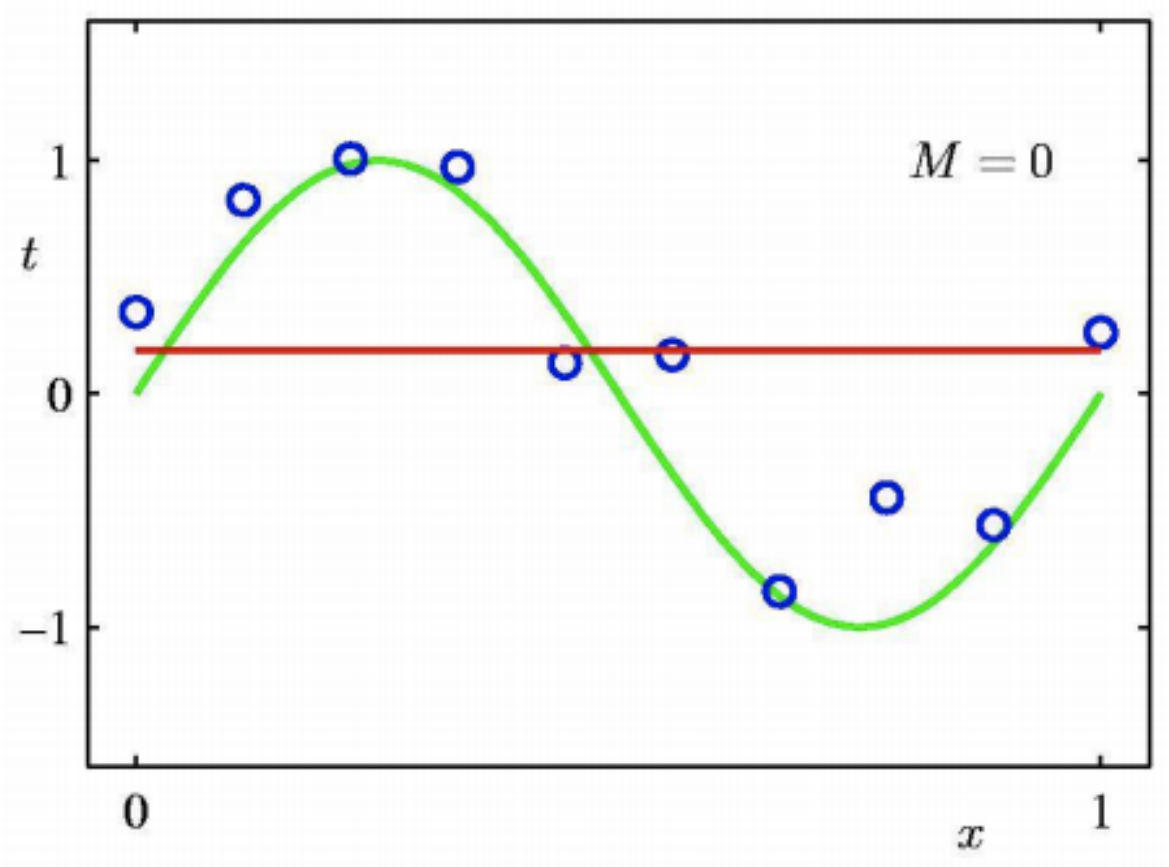
We could fit a model using all of them, but this runs the risk of *overfitting*: using irrelevant variables to explain noise by coincidence.

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.

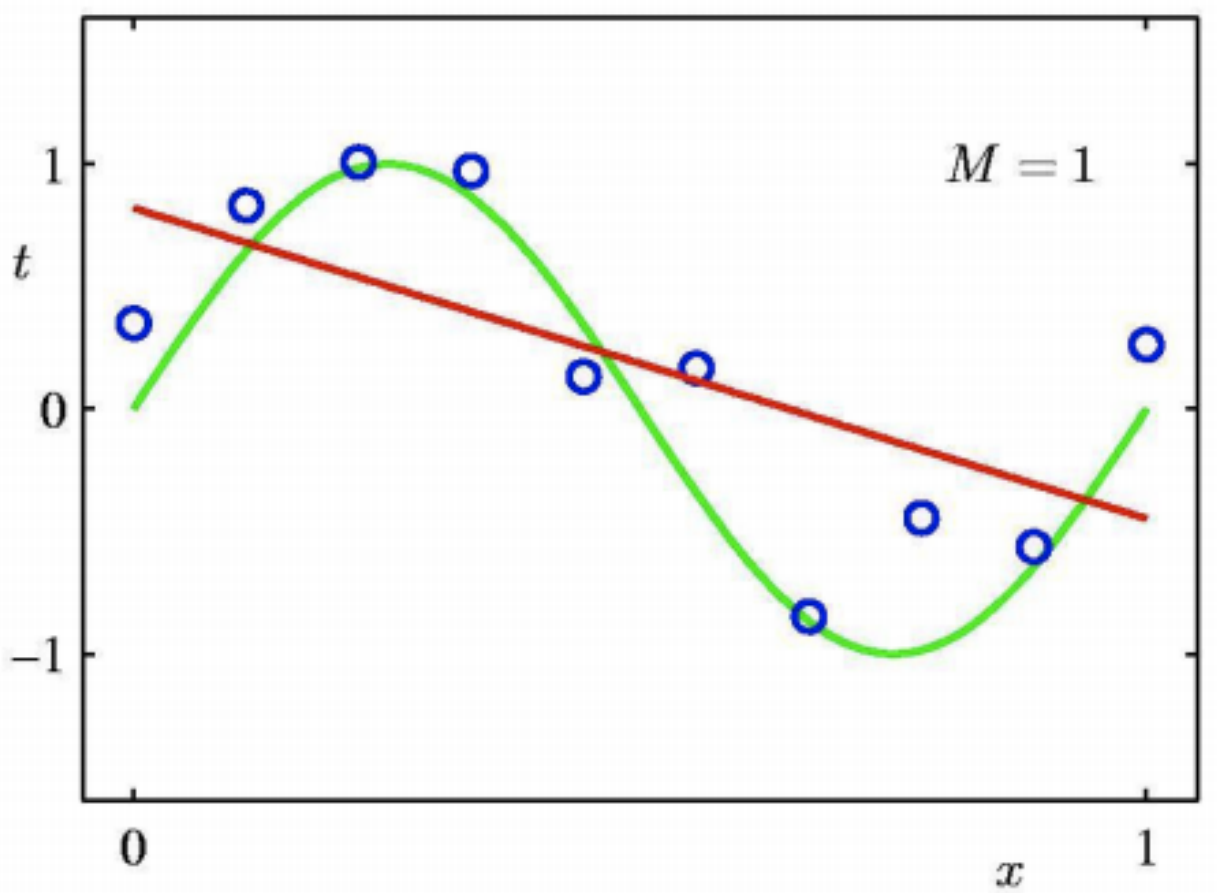
Sequential testing

Predictors $= \{x^0, x^1, \ldots, x^M\}$.



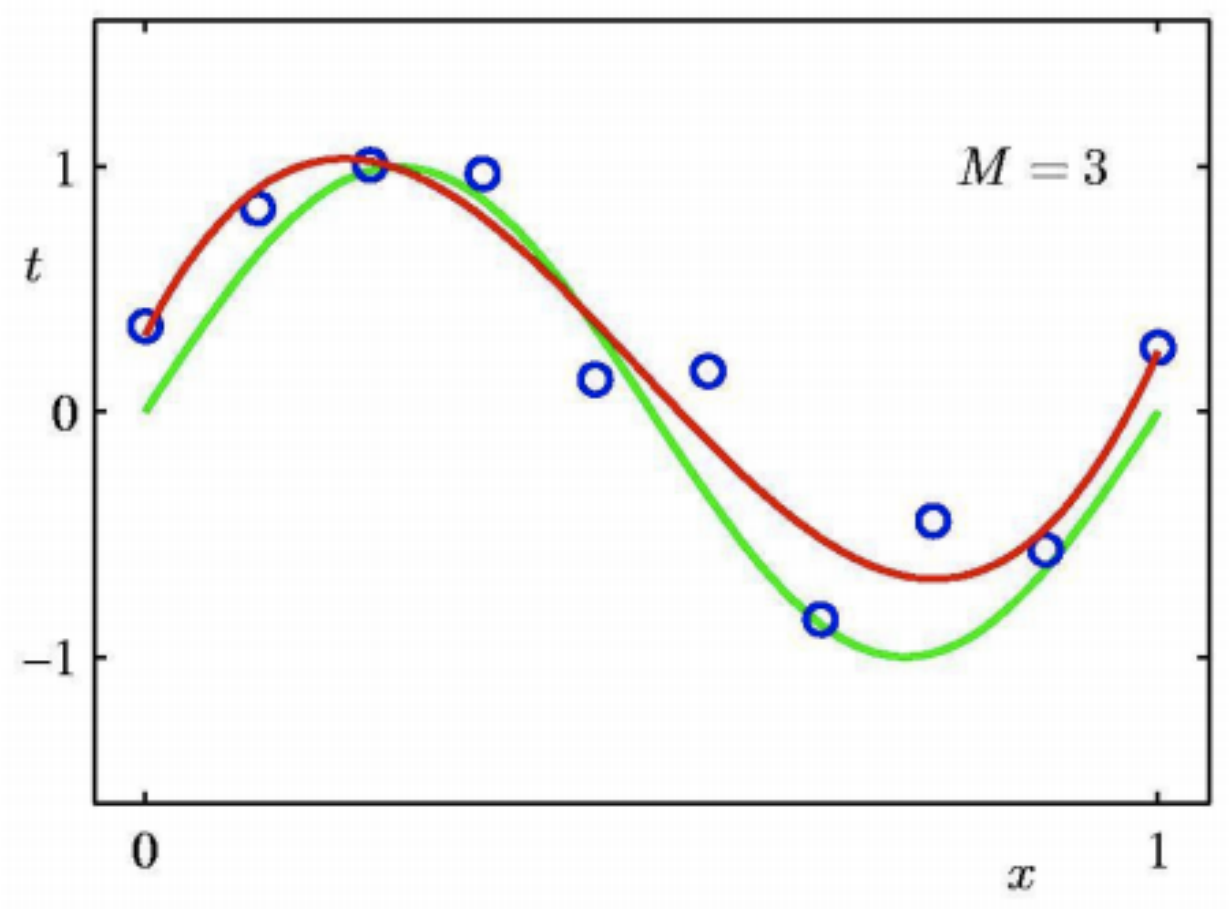
Sequential testing

Predictors $= \{x^0, x^1, \ldots, x^M\}$.



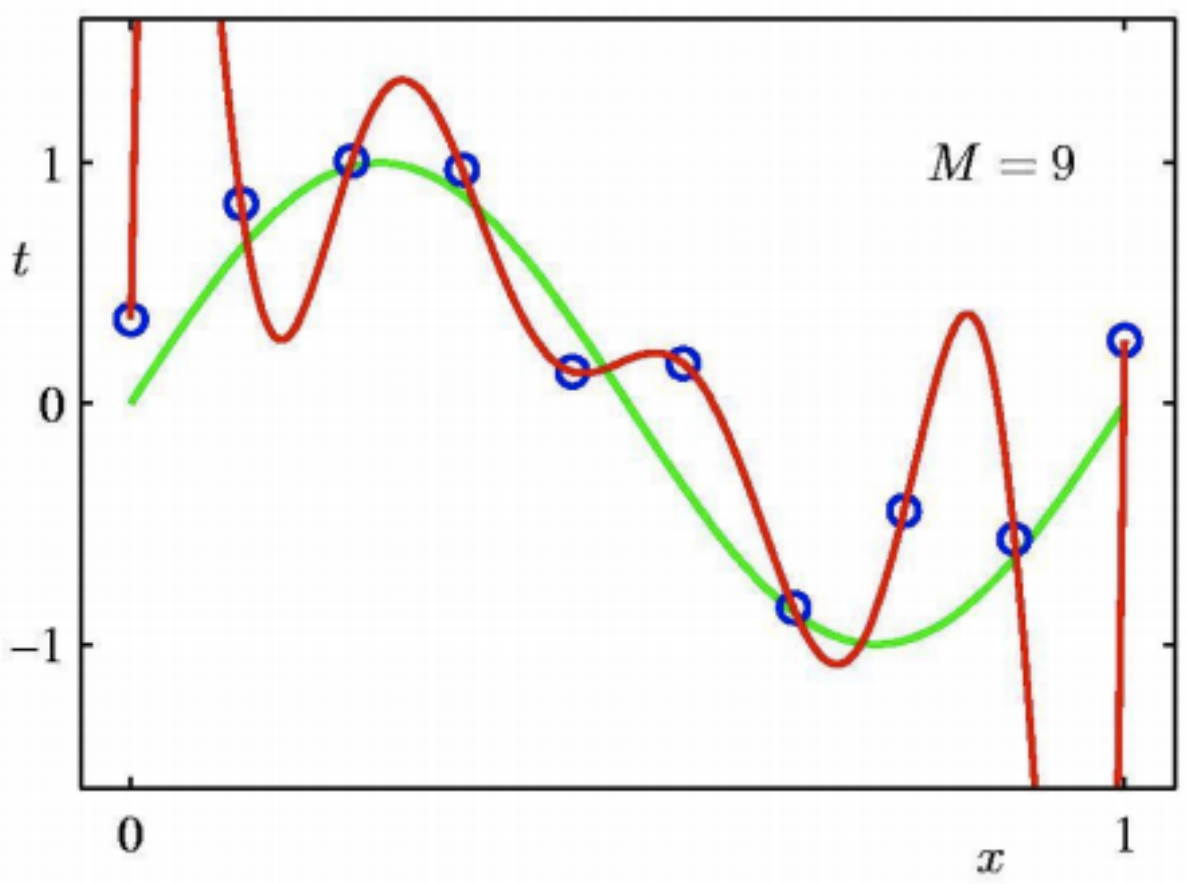
Sequential testing

Predictors $= \{x^0, x^1, \ldots, x^M\}$.



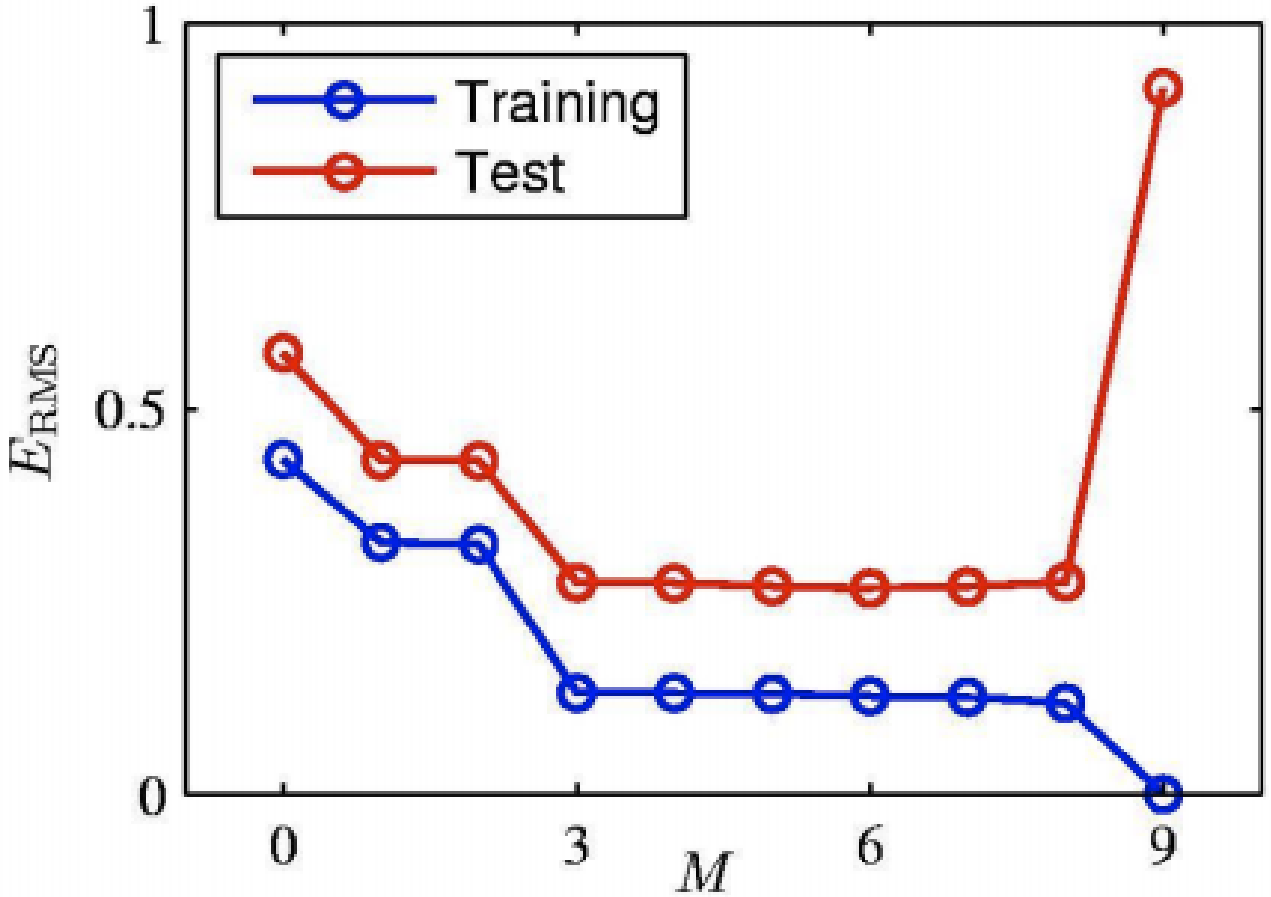
Sequential testing

Predictors $= \{x^0, x^1, \ldots, x^M\}$.



Sequential testing

Predictors $= \{x^0, x^1, \ldots, x^M\}$.



Sequential testing

In a parsimonious model, if we were to test if the parameter β_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 γ_1 such that the hypothesis $\gamma_1 = \mathbf{0}$ 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 β_3, but not reject $\beta_1 = 0$ given β_2 and β_3, and also not reject $\beta_2 = 0$ given β_1 and β_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 $\beta_0, \ldots, \beta_k$ we could consider p tests of the form $H_{0j} : \beta_j = 0$, 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 $\beta_j = 0$.

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

Acceptance or rejection of H_{0j} does not mean that the model under H_{0j} or the full model is the best model, it just informs us about the usefulness of x_j 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)

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

We denote the corresponding $\mathbf{X}$ matrices by $\mathbf{X}_{0:j}$, which are the first $j+1$ columns of $\mathbf{X}$. Let $\mathbf{H}_j = H(\mathbf{X}_{0:j})$, where for any matrix $\mathbf{B}$ of full rank, $H(\mathbf{B}) = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ is the hat matrix for $\mathbf{B}$.

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

$$R_j = R(\beta_0, \beta_1, \ldots, \beta_j) = \mathbf{y}^T\mathbf{H}_j\mathbf{y},$$

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:

$$\begin{aligned} R(\beta_1|\beta_0) &= R_1 - R_0 \\ R(\beta_2|\beta_0,\beta_1) &= R_2 - R_1 \\ &\vdots \\ R(\beta_k|\beta_0,\beta_1,\ldots,\beta_{k-1}) &= R(\boldsymbol{\beta}) - R_{k-1}. \end{aligned}$$

Sequential testing

Theorem 5.8

Assume (I), (II), and (V) holds for the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. *Then,*

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

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

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

Lemma 5.9

Let $\mathbf{A} = [\mathbf{A}_1|\mathbf{A}_2]$ *be a matrix of full rank with* n *rows and* m *columns, where* $n > m$. *Then the matrix* $H(\mathbf{A}) - H(\mathbf{A}_1)$ *is idempotent.*

Proof.

First we note that

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

Sequential testing

Partitioning the first factor and noting that both $\mathbf{A}_1$ and $\mathbf{A}_2$ have n rows,

$$\begin{aligned}
\left[\begin{array}{c}\mathbf{A}_1^T \\ \hline \mathbf{A}_2^T\end{array}\right][\mathbf{I} - H(\mathbf{A})] &= \mathbf{0} \\
\mathbf{A}_1^T[\mathbf{I} - H(\mathbf{A})] &= \mathbf{0} \\
\mathbf{A}_1^T &= \mathbf{A}_1^T H(\mathbf{A}) \\
\mathbf{A}_1 &= H(\mathbf{A})\mathbf{A}_1.
\end{aligned}$$

since $H(\mathbf{A})$ is symmetric.

So, using the fact that $H(\mathbf{B})$ is idempotent for any matrix $\mathbf{B}$,

$$\begin{aligned}
&(H(\mathbf{A}) - H(\mathbf{A}_1))^2 \\
&= H(\mathbf{A}) - H(\mathbf{A})\mathbf{A}_1(\mathbf{A}_1^T\mathbf{A}_1)^{-1}\mathbf{A}_1^T \\
&\qquad -\mathbf{A}_1(\mathbf{A}_1^T\mathbf{A}_1)^{-1}\mathbf{A}_1^T H(\mathbf{A}) + H(\mathbf{A}_1) \\
&= H(\mathbf{A}) - H(\mathbf{A}_1) - H(\mathbf{A}_1) + H(\mathbf{A}_1) \\
&= H(\mathbf{A}) - H(\mathbf{A}_1).
\end{aligned}$$

m quadratic forms

Let's recap Theorem 3.15 from Lecture 3.....

Sequential testing

Proof of Theorem 5.8: The sum follows from the definitions. To prove the rest, we use Theorem 3.15 with the hypothesis that we have idempotent matrices whose sum is idempotent.

The conclusion is that quadratic forms in a multivariate normal random vector, based on the components of the sum, have independent non-central chi-square distributions with degrees of freedom equal to the rank of the matrices.

From the lemma, $H_j - H_{j-1}$ is idempotent. Furthermore both $\mathbf{H}_0$ and $\mathbf{I} - \mathbf{H}_k$ are idempotent.

Thus we have a set of idempotent matrices $H_1, H_2 - H_1, \cdots H_p - H_{p-1}, I - H_p$ whose sum I also idempotent. Theorem 3.14 thus gives that the quadratic forms all have independent noncentral χ^2 distributions.

Sequential testing

To show the degrees of freedom, observe that the sum of the ranks is n and $r(\mathbf{I} - \mathbf{H}_k) = n - p$ (use Theorem 2.3)

Each sequential regression sum of squares, $\mathbf{H}_j - \mathbf{H}_{j-1}$ has 1 degree of freedom because, being idempotent,

$$r(\mathbf{H}_j - \mathbf{H}_{j-1}) = tr(\mathbf{H}_j - \mathbf{H}_{j-1}) = tr(\mathbf{H}_j) - tr(\mathbf{H}_{j-1}) = j - (j-1) = 1.$$

Sequential testing

Therefore under the hypothesis $\beta_j = 0$, the test statistic

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

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

This is still not entirely satisfactory.

Unfortunately, the order in which the parameters are tested can heavily influence the results. Thus, 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. Figure 103 shows pairwise plots of the data.

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

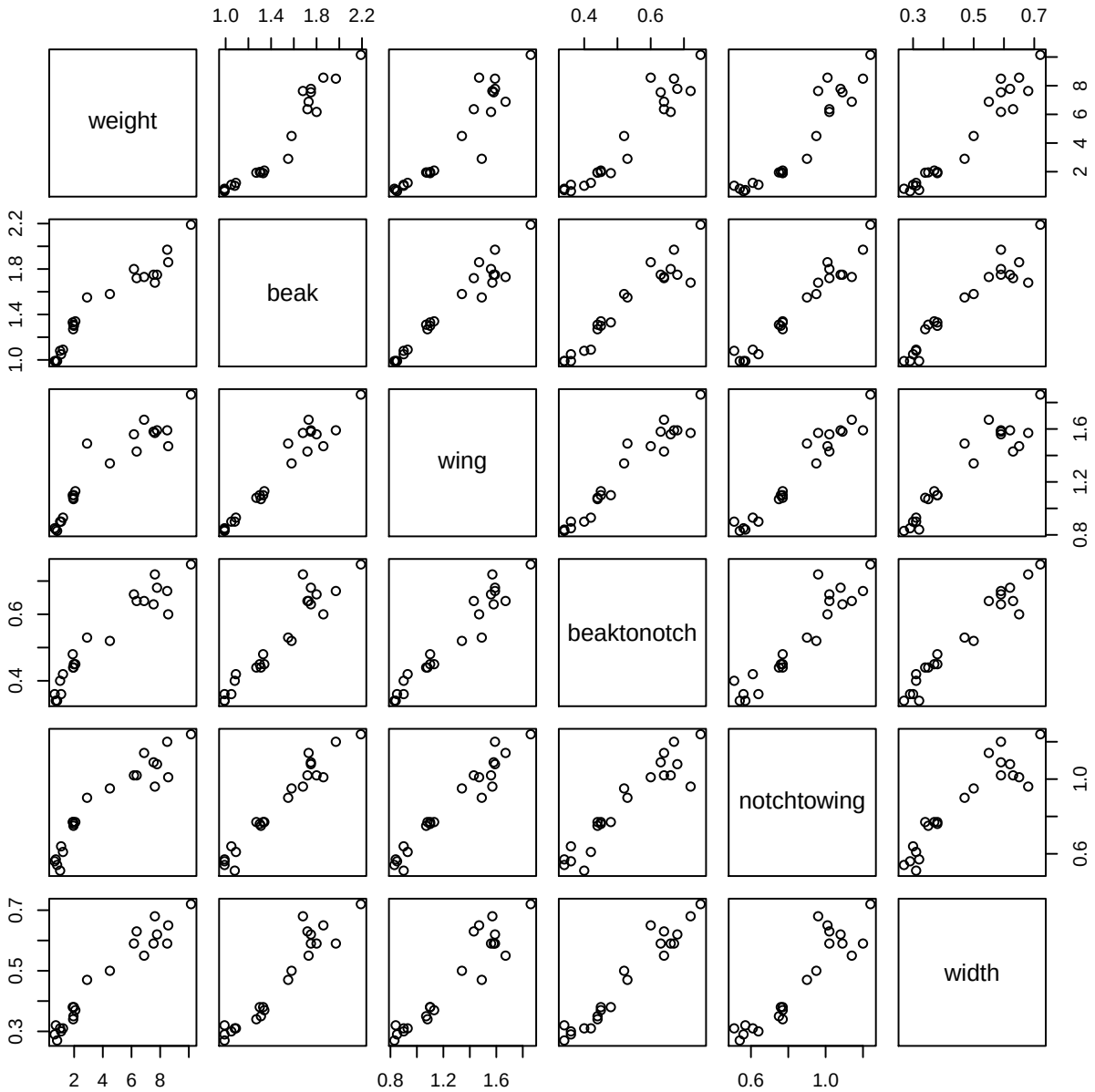


Figure: Pairwise plots of the variables in the squid data set

Squid example

Let's first test if any parameters should be in the model, i.e. if $H_0 : \boldsymbol{\beta} = \mathbf{0}$.

```
n <- dim(squid)[1]
p <- dim(squid)[2]
y <- squid$weight
X <- as.matrix(cbind(rep(1,n),squid[,-1]))
betahat <- solve(t(X)%*%X,t(X)%*%y)
SSRes <- sum((y-X%*%betahat)^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
+ width
##   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 $H_0 : \boldsymbol{\beta} = \mathbf{0}$ 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   0.2634957   4.3521933
```

Squid example

Thus the sequential sums of squares are:

$$
\begin{aligned}
R(\beta_0) &= 387.16 \\
R(\beta_1|\beta_0) &= 199.15 \\
R(\beta_2|\beta_0,\beta_1) &= 0.127 \\
R(\beta_3|\beta_0,\beta_1,\beta_2) &= 4.12 \\
R(\beta_4|\beta_0,\beta_1,\beta_2,\beta_3) &= 0.263 \\
R(\beta_5|\beta_0,\beta_1,\beta_2,\beta_3,\beta_4) &= 4.35
\end{aligned}
$$

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 β_0 and then β_1, but not β_2.

The subsequent tests are harder to interpret. For example, if $\beta_0, \beta_1, \beta_2$, and β_3 are in the model, we should not add β_4. But β_2 is not in the model!

The tests for β_3, β_4 and β_5 need to be repeated, supposing only that β_0 and β_1 are in the model.

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.

Squid example

Note: R cannot do this unless all the models are presented at once (see the clover example below)! To test for β_i in the presence of $\beta_0, \ldots, \beta_{i-1}$ it uses the residual sum of squares from the model using $\beta_0, \ldots, \beta_i$. This still gives a valid test though.

Clover example

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

```
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 '
```

Forward selection

Reference for this Section is Faraway (2005)
To resolve the difficulties of different final models depending on parameter ordering, there are strategies to find 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 hypothesis $H_{0j} : \beta_j = 0$, 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 heat expended y (in calories) in cement during 180 days hardening, depending on percentages of 4 different chemicals in the mixture. Note that 1 to 5 percent of the cement is not in these chemicals. Figure 118 shows the pairwise scatter plots.

```
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)
```

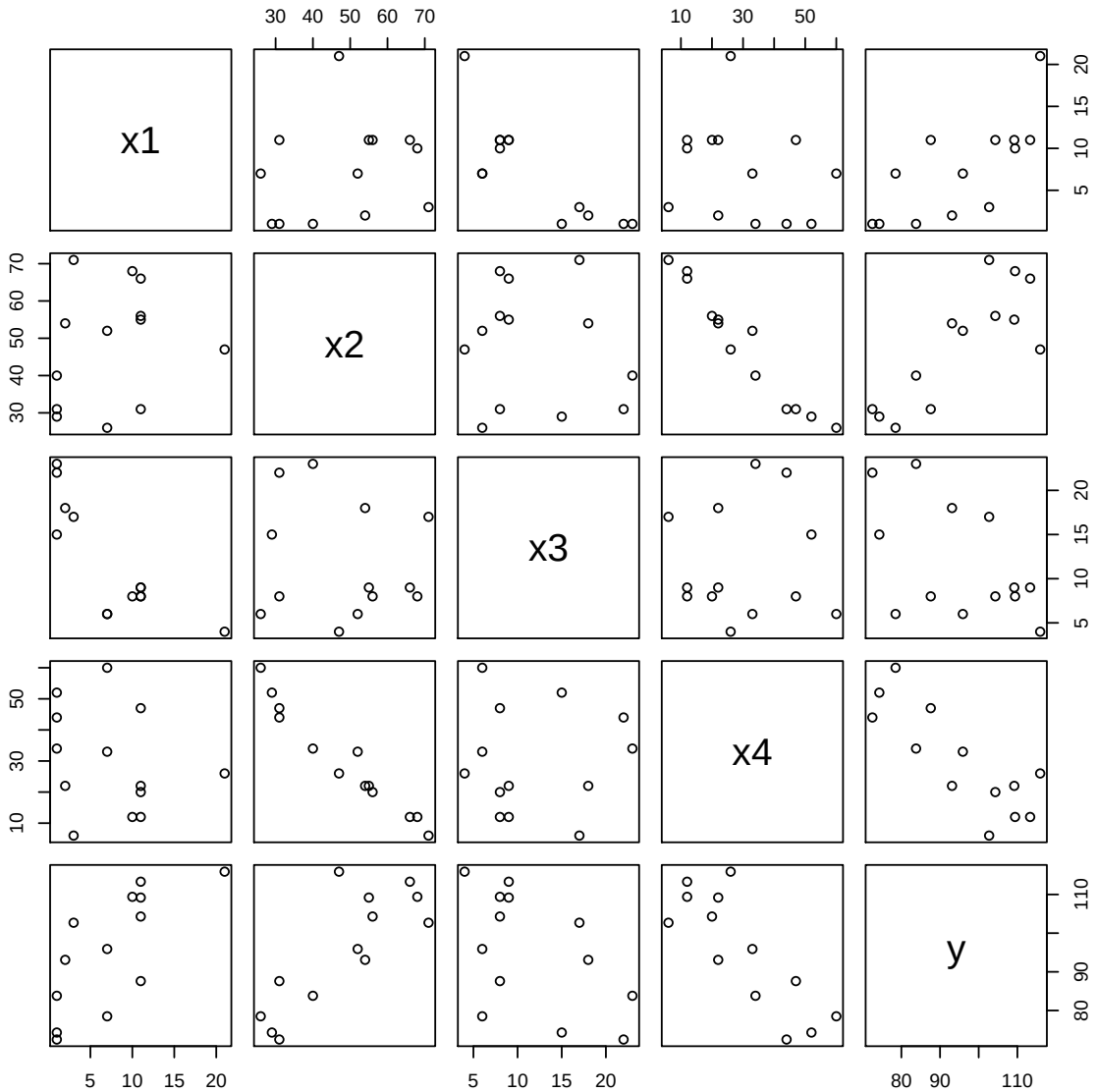


Figure: Pairwise plots of cement data

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 inflexibilty 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 easy to adjust the procedure for any other goodness-of-fit statistic. The smaller is the AIC, the better is the model - details follow after describing stepwise selection.

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:

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

R^2 lies between 0 and 1, and the larger it is, the more variation in y is explained by the model. It is also equal to the square of the correlation between y and $\hat{y}$.

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 β_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.

AIC

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.

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. For none of them is there an absolute measure of how good a model is.

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
```