Statistics for Data Science: Week 11

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch

Nested Model Selection & ANOVA

Consider the model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon.$$

This will always have higher R^2 than the sub-model:

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon.$$

- Why? (think of geometry...)
- The question is: is the first model *significantly* better than the second one?
  - ↪ i.e. does the first model explain the variation adequately enough, or should we incorporate extra explanatory variables? Need a quantitative answer.

Model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I})$. Estimator:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Interpretation: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$ is the projection of $\boldsymbol{Y}$ into the column space of $\boldsymbol{X}$, $\mathcal{M}(\boldsymbol{X})$. This subspace has dimension p, when $\boldsymbol{X}$ is of full column rank p. Now for $q < p$ write $\boldsymbol{X}$ in block notation as

$$\boldsymbol{X} = \underset{n\times q}{\boldsymbol{X}_1} \quad \underset{n\times(p-q)}{\boldsymbol{X}_2}.$$

Interpretation: $\boldsymbol{X}_1$ is built by the first q columns of $\boldsymbol{X}$ and $\boldsymbol{X}_2$ by the rest. Similarly write $\boldsymbol{\beta} = (\boldsymbol{\beta}_1 \ \boldsymbol{\beta}_2)^\top$ so that:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = (\boldsymbol{X}_1 \ \boldsymbol{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix} + \boldsymbol{\varepsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

Our question can now be stated as:

- Is $\boldsymbol{\beta}_2 = 0$?

Let $\boldsymbol{H}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$, and $\hat{\boldsymbol{Y}}_1 = \boldsymbol{H}_1\boldsymbol{Y}$, $\boldsymbol{e}_1 = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_1$.

Pythagoras tells us that:

$$\underbrace{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_1\|^2}_{RSS(\hat{\beta}_1)=\|\boldsymbol{e}_1\|^2} = \underbrace{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}_{RSS(\hat{\beta})=\|\boldsymbol{e}\|^2} + \underbrace{\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_1\|^2}_{RSS(\hat{\beta}_1)-RSS(\hat{\beta})=\|\boldsymbol{e}-\boldsymbol{e}_1\|^2}$$

Notice that $RSS(\hat{\beta}_1) \geq RSS(\hat{\beta})$ always (think why!)

So the idea is simple: to see if it is worthwhile to include β_2 we will compare how much larger $RSS(\hat{\beta}_1)$ is compared to $RSS(\hat{\beta})$.

- Equivalently, we can look at a ratio like $\{RSS(\hat{\beta}_1) - RSS(\hat{\beta})\}/RSS(\hat{\beta})$
- This is in fact the likelihood ratio test statistic for our hypothesis.
- To construct a test based on this quantity, we need to figure its sampling distributions ...

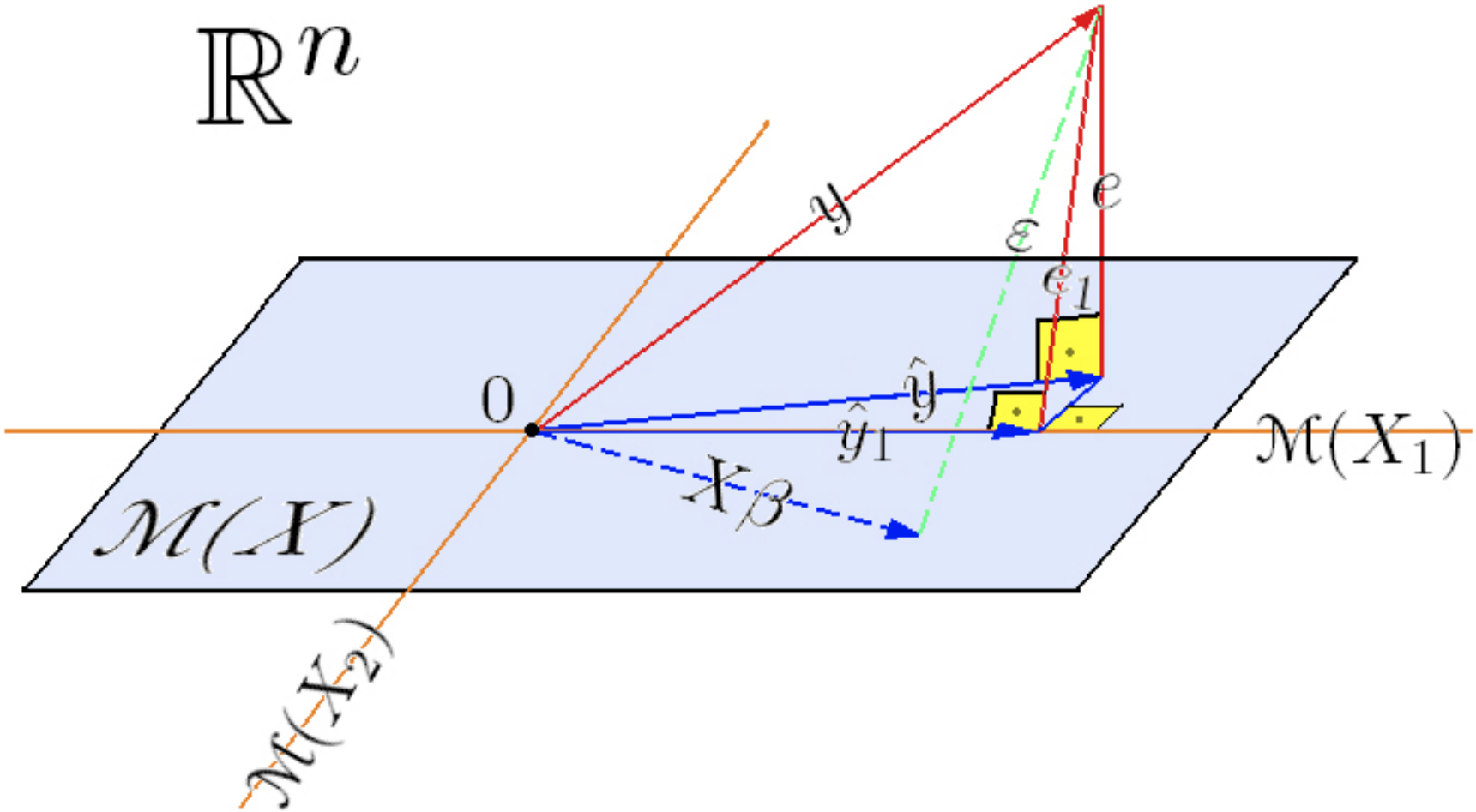


Figure: Geometry Revisited

Theorem (Sampling Distributions for Sums of Squares)

We have the following properties:

(A) $\boldsymbol{e} - \boldsymbol{e}_1 \perp \boldsymbol{e}$;

(B) $\|\boldsymbol{e}\|^2 = RSS(\hat{\beta})$ *and* $\|\boldsymbol{e}_1 - \boldsymbol{e}\|^2 = RSS(\hat{\beta}_1) - RSS(\hat{\beta})$ *are independent;*

(C) $\|\boldsymbol{e}\|^2 \sim \sigma^2 \chi^2_{n-p}$;

(D) *under the hypothesis* $H_0 : \beta_2 = 0$, $\|\boldsymbol{e}_1 - \boldsymbol{e}\|^2 \sim \sigma^2 \chi^2_{p-q}$.

Proof.

(A) holds since $\boldsymbol{e} - \boldsymbol{e}_1 = \boldsymbol{Y} - \hat{\boldsymbol{Y}} - \boldsymbol{Y} + \hat{\boldsymbol{Y}}_1 = -\hat{\boldsymbol{Y}} + \hat{\boldsymbol{Y}}_1 \in \mathcal{M}(\boldsymbol{X}_1, \boldsymbol{X}_2)$ but $\boldsymbol{e} \in [\mathcal{M}(\boldsymbol{X}_1, \boldsymbol{X}_2)]^\perp$.

To show (B), we notice that

$$\boldsymbol{e}_1 = (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H}_1\boldsymbol{H})\boldsymbol{Y}$$

because $\mathcal{M}(\boldsymbol{X}_1) \subset \mathcal{M}(\boldsymbol{X}_1, \boldsymbol{X}_2)$.

Therefore,

$$\boldsymbol{e} - \boldsymbol{e}_1 = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} - (\boldsymbol{I} - \boldsymbol{H}_1\boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} - \boldsymbol{Y} + \boldsymbol{H}_1\boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{H}\boldsymbol{Y}.$$

But recall that $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{1}\beta, \sigma^2\boldsymbol{I})$. Therefore, to prove independence of $\boldsymbol{e} - \boldsymbol{e}_1 = (\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, we need to show that

$$(\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{H}[\sigma^2\boldsymbol{I}](\boldsymbol{I} - \boldsymbol{H})^\top = 0.$$

This is clearly the case since $\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H}) = 0$, proving (B).

(C) follows immediately, since we have already proven last time that $\forall \boldsymbol{\beta}$ (even when $\boldsymbol{\beta}_2 = 0$)

$$RSS(\hat{\boldsymbol{\beta}}) \sim \sigma^2\chi^2_{n-p}$$

To prove (D), we note that

$$\boldsymbol{e} - \boldsymbol{e}_1 = (\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{H}\boldsymbol{Y} \sim \mathcal{N}\{(\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{H}\boldsymbol{X}\beta, \sigma^2 \underbrace{(\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{H}\boldsymbol{H}^\top(\boldsymbol{H}_1 - \boldsymbol{I})^\top}_{=\boldsymbol{H}-\boldsymbol{H}_1}\}.$$

But $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{X}$. So, in block notation,

$$\boldsymbol{e} - \boldsymbol{e}_1 \sim \mathcal{N}((\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{X}_1\beta_1 + (\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{X}_2\beta_2, \sigma^2(\boldsymbol{H} - \boldsymbol{H}_1)).$$

Now $(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_1\beta_1 = 0$ always, since $\boldsymbol{I} - \boldsymbol{H}_1$ projects onto $\mathcal{M}^\perp(\boldsymbol{X}_1)$. Therefore,

$$\boldsymbol{e} - \boldsymbol{e}_1 \sim \mathcal{N}(0, \sigma^2(\boldsymbol{H} - \boldsymbol{H}_1)), \qquad \text{when } \beta_2 = 0.$$

Now observe that $(\boldsymbol{H} - \boldsymbol{H}_1)^\top = (\boldsymbol{H} - \boldsymbol{H}_1)$ and $(\boldsymbol{H} - \boldsymbol{H}_1)^2 = (\boldsymbol{H} - \boldsymbol{H}_1)$ (because $\mathcal{M}(\boldsymbol{X}_1) \subset \mathcal{M}(\boldsymbol{X}_1, \boldsymbol{X}_2)$). Thus,

$$\begin{aligned} \boldsymbol{e} - \boldsymbol{e}_1 \sim \mathcal{N}(0, \sigma^2(\boldsymbol{H} - \boldsymbol{H}_1)^2) \implies \boldsymbol{e} - \boldsymbol{e}_1 &\overset{d}{=} (\boldsymbol{H} - \boldsymbol{H}_1)\varepsilon \\ \implies RSS(\hat{\beta}_1) - RSS(\hat{\beta}) = \|\boldsymbol{e} - \boldsymbol{e}_1\|^2 &\overset{d}{=} \varepsilon^\top(\boldsymbol{H} - \boldsymbol{H}_1)\varepsilon \sim \sigma^2\chi^2_{p-q}. \end{aligned}$$

since $(\boldsymbol{H} - \boldsymbol{H}_1)$ is symmetric idempotent with trace $p - q$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_n)$.

□

Corollary

We conclude that, under the hypothesis $\beta_2 = 0$,

$$\frac{\left(\frac{RSS(\hat{\beta}_1) - RSS(\hat{\beta})}{p-q}\right)}{\left(\frac{RSS(\hat{\beta})}{n-p}\right)} \sim F_{p-q,n-p}$$

Distributional results suggest the following test:

- Have $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2, \sigma^2\boldsymbol{I})$
- $H_0 : \beta_2 = 0$
- Data: $(\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{X}_2)$.
- Test statistic: $T = \dfrac{\left(\dfrac{RSS(\hat{\beta}_1) - RSS(\hat{\beta})}{p-q}\right)}{\left(\dfrac{RSS(\hat{\beta})}{n-p}\right)}$

Then, under H_0, it holds that $T \sim F_{p-q,n-p}$. Suppose we observe $T = \tau$. Then,

$$p = \mathbb{P}_{H_0}[T(Y) \geq \tau] = \mathbb{P}[F_{p-q,n-p} \geq \tau]$$

Reject the null hypothesis if $p < \alpha$, some small α, usually 0.05.

Example (Nested Models in Cement Data)

- We fitted the model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

- But would the following simpler model be in fact adequate?

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- Intuitively: is the extra explanatory power of the "larger" model significant enough in order to justify its use instead of a simpler model? (i.e., is the residual vector for the "larger" model significantly smaller than that of the simpler model?)
- In this case, $n = 13$, $p = 5$, $q = 2$ and

$$RSS(\hat{\beta}) = 47.86, \qquad RSS(\hat{\beta}_1) = 1265.7$$

yielding

$$\tau = \frac{(1265.7 - 47.86)/(5-2)}{(47.86)/(13-5)} = 67.86$$

- $p = \mathbb{P}[F_{3,8} \geq 67.86] = 4.95 \times 10^{-6}$, so we reject the hypothesis $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$.

- Let $\mathbf{1}$, $\boldsymbol{X}_1$, ..., $\boldsymbol{X}_r$ be groups of columns of $\boldsymbol{X}$ (the "terms"), such that

$$\boldsymbol{X} = (\underset{n\times 1}{\mathbf{1}} \ \underset{n\times q_1}{\boldsymbol{X}_1} \ \underset{n\times q_2}{\boldsymbol{X}_2} \ \dots \ \underset{n\times q_r}{\boldsymbol{X}_r}), \quad \beta = (\underset{1\times 1}{\beta_0} \ \underset{1\times q_1}{\boldsymbol{\beta}_1} \ \underset{1\times q_2}{\boldsymbol{\beta}_2} \ \dots \ \underset{1\times q_r}{\boldsymbol{\beta}_r})^\top$$

We have

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{1}\beta_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \cdots + \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}$$

- Would like to do the same "F-test investigation", but this time do it term-by-term. That is, we want to look at the following sequence of nested models:
  - $\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{\varepsilon}$
  - $\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$
  - $\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$

    $\vdots$
  - $\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \cdots + \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}$

Proceed similarly as before. Define:

- $\boldsymbol{X}_0 := \mathbf{1}$ and $\mathcal{X}_k = (\boldsymbol{X}_0\ \boldsymbol{X}_1\ \boldsymbol{X}_2\ \ldots\ \boldsymbol{X}_k), \quad k \in \{0, \ldots, r\}$
- $\mathcal{H}_k := \mathcal{X}_k(\mathcal{X}_k^\top \mathcal{X}_k)^{-1}\mathcal{X}_k^\top, \quad k \in \{0, \ldots, r\}$
- $\hat{\boldsymbol{Y}}_k := \mathcal{H}_k \boldsymbol{Y}, \quad k \in \{0, \ldots, r\}$
- $\boldsymbol{e}_k = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_k, \quad k \in \{0, \ldots, r\}$
- Note that $\hat{\boldsymbol{Y}}_0 = \bar{Y}\mathbf{1}$.

▶ As before, Pythagoras implies

$$\underbrace{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_0\|^2}_{\|\boldsymbol{e}_0\|^2} = \underbrace{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_r\|^2}_{\|\boldsymbol{e}_r\|^2} + \underbrace{\underbrace{\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{r-1}\|^2}_{\|\boldsymbol{e}_r - \boldsymbol{e}_{r-1}\|^2} + \cdots + \underbrace{\|\hat{\boldsymbol{Y}}_1 - \hat{\boldsymbol{Y}}_0\|^2}_{\|\boldsymbol{e}_1 - \boldsymbol{e}_0\|^2}}_{\|\boldsymbol{e}_r - \boldsymbol{e}_0\|^2}$$

$$= \underbrace{\|\boldsymbol{e}_r\|^2}_{RSS_r} + \sum_{k=0}^{r-1} \underbrace{\|\boldsymbol{e}_{k+1} - \boldsymbol{e}_k\|^2}_{RSS_k - RSS_{k+1}}$$

with RSS_k the residual sum of squares for $\hat{\boldsymbol{Y}}_k$, with ν_k degrees of freedom.

Some observations:

- $RSS_k - RSS_{k+1}$ is the reduction in residual sum of squares caused by adding $\boldsymbol{X}_{k+1}$, when the model already contains $\boldsymbol{X}_0, \ldots, \boldsymbol{X}_k$.
- RSS_r and $\{RSS_k - RSS_{k+1}\}_{k=0}^{r-1}$ are all mutually independent.
- Obviously, $\nu_0 \geq \nu_1 \geq \nu_2 \geq \cdots \geq \nu_r$
- $\nu_{k+1} = \nu_k$ if $\boldsymbol{X}_{k+1} \in \mathcal{M}(\mathcal{X}_k)$.

▶Given this information, we want to see how adding each term in the model sequentially, affects the explanatory capacity of the model.

↪ In other words, we want to investigate the reduction in the residual sum of squares (RSS) achieved by adding each term to the model. Is this significant?

Terms	df	Residual RSS	Terms added	df	Reduction in RSS	F-test
$\mathbf{1}$	$n-1$	RSS_0				
$\mathbf{1}, \boldsymbol{X}_1$	ν_1	RSS_1	$\boldsymbol{X}_1$	$n-1-\nu_1$	$RSS_0 - RSS_1$	
$\mathbf{1}, \boldsymbol{X}_1, \boldsymbol{X}_2$	ν_2	RSS_2	$\boldsymbol{X}_2$	$\nu_1 - \nu_2$	$RSS_1 - RSS_2$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$\mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_r$	ν_r	RSS_r	$\boldsymbol{X}_r$	$\nu_{r-1} - \nu_r$	$RSS_{r-1} - RSS_r$	

The F-statistic for testing the significance of the reduction in RSS when $\boldsymbol{X}_k$ is added to the model containing terms $1, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ is

$$F_k = \frac{(RSS_{k-1} - RSS_k)/(\nu_{k-1} - \nu_k)}{RSS_r/\nu_r},$$

and $F_k \sim F_{\nu_{k-1}-\nu_k, \nu_r}$ under the null hypothesis $H_0 : \beta_k = 0$.

Recall that large values of F_k relative to the null distribution are evidence against H_0.

Example

Nested Sequence in Cement Data

- Reductions in overall sum of squares when sequentially entering terms x_1, x_2, x_3 and x_4.
- Does adding extra variables improve model significantly?

	Df	Red Sum Sq	F value (τ)	p-value
x_1	1	1450.08	242.37	2.88×10^{-7}
x_2	1	1207.78	201.87	5.86×10^{-7}
x_3	1	9.79	1.64	0.2366
x_4	1	0.25	0.04	0.8441
Residual SSq	8	47.86		

- In this case, each term is a single column (variable).

Warning!

- Significance of entering a term depends on how the sequence is defined: when entering terms in different order get different results! (why?)
- When a term is entered "early" and is significant, this does not tell us much (why?)
- When a term is entered "late" is significant, then this is quite informative (why?)

▶ Why is this true? Are there special cases when the order of entering terms doesn't matter?

- Consider terms $\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2$ from $\boldsymbol{X}$, so

$$\underset{n\times 1}{\boldsymbol{X}} = (\underset{n\times 1}{\boldsymbol{X}_0} \ \underset{n\times q_1}{\boldsymbol{X}_1} \ \underset{n\times q_2}{\boldsymbol{X}_2}), \quad \underset{1\times 1}{\boldsymbol{\beta}} = (\underset{1\times 1}{\beta_0} \ \underset{1\times q_1}{\boldsymbol{\beta}_1} \ \underset{1\times q_2}{\boldsymbol{\beta}_2})^\top$$

- Assume orthogonality of terms, i.e. $\boldsymbol{X}_i^\top \boldsymbol{X}_j = 0, \quad i \neq j$

Notice that in this case

$$\hat{\beta} = \begin{pmatrix} \boldsymbol{X}_0^\top \boldsymbol{X}_0 & 0 & 0 \\ 0 & \boldsymbol{X}_1^\top \boldsymbol{X}_1 & 0 \\ 0 & 0 & \boldsymbol{X}_2^\top \boldsymbol{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}_0 & \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{pmatrix}^\top \boldsymbol{Y}$$

$$\implies \hat{\beta}_0 = \bar{Y}, \ \hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{Y}, \ \hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{Y}$$

It follows that the reductions of sums of squares are unique, in the sense that they do not depend upon the order of entry of the terms in the model. (show this!)

Intuition: $\boldsymbol{X}_i$ contains completely linearly independent information from $\boldsymbol{X}_j$ for $\boldsymbol{Y}$, $i \neq j$

Model Selection / Collinearity / Regularisation

- **Theory:** We are given a relationship

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

and asked to provide estimators, tests, confidence intervals, optimality properties ...

...and we can do it with complete success!

- **Practice**: We are given data $(\boldsymbol{Y}, \boldsymbol{X})$ and suspect a linear relationship between $\boldsymbol{Y}$ and some of the columns of $\boldsymbol{X}$. We don't know a priori which exactly!
  - ↪ Need to select a "most appropriate" subset of the columns of $\boldsymbol{X}$
  - ↪ General principle: parsimony (Latin *parsimōnia*: sparingness; simplicity and least number of requisites and assumptions; economy or frugality of components and associations).

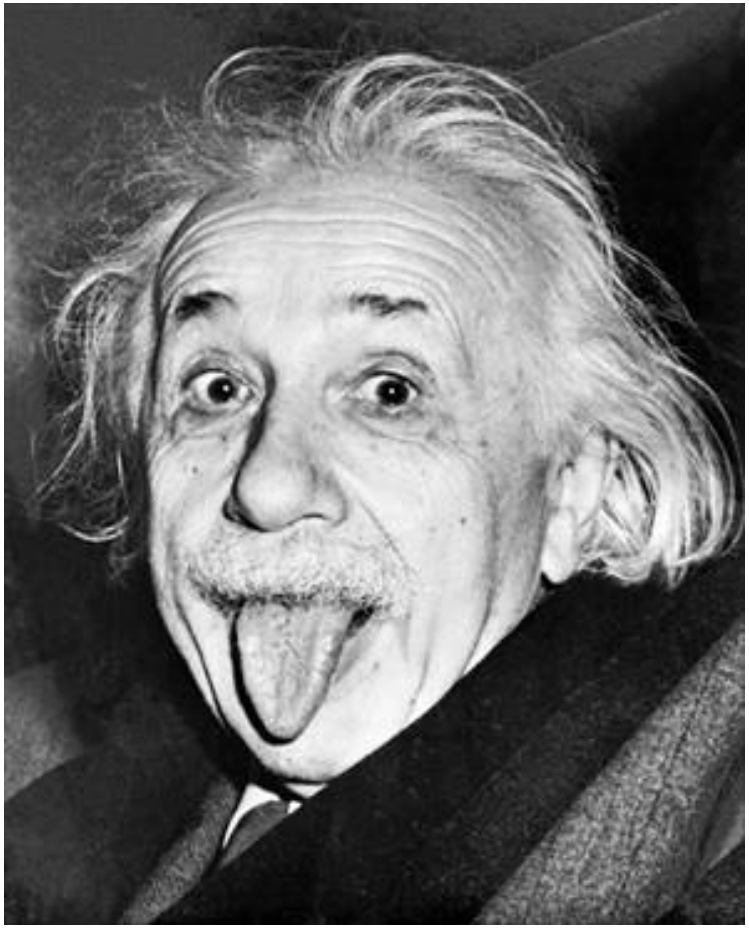


Figure: Albert Einstein (1879–1955): *'Everything should be made as simple as possible, but no simpler.'*

William of Ockham (?1285–1347): '*It is vain to do with more what can be done with fewer*' (Occam's razor: Given several explanations of the same phenomenon, we should prefer the simplest.)

Graphical exploration $\rightsquigarrow$ provides initial picture:

- plots of $\boldsymbol{Y}$ against candidate variables;
- plots of transformations of $\boldsymbol{Y}$ against candidate variables;
- plots of transformations of certain variables against $\boldsymbol{Y}$;
- plots of pairs of candidate variables.

This will often provide a starting point, but:

- Automatic Model Selection: Need objective model comparison criteria, as a screening device.
  - $\hookrightarrow$ We saw how to do an F-test, but what if models to be compared are not nested?
- Automatic Model Building: Situations when p large, so there are *lots* of possible models.
  - $\hookrightarrow$ Automatic methods for building a model? We saw that ANOVA depends on the order of entry of variables in the model ...

Consider design matrix $\boldsymbol{X}$ with p variables.

- 2^p possible models!
- Denote set of all models generated by $\boldsymbol{X}$ by $2^{\boldsymbol{X}}$ (model powerset)
- If wish to consider k different transformations of each variable, then p becomes $(1 + k)p$
- Fast algorithms (branch and bound, `leaps` in `R`) exist to fit them, but they don't work for *large* p, and anyway ...
- ... need criterion for comparison.

So given a collection of models, we need an automatic (objective) way to pick out a "best" one (unfortunately cannot look carefully at all of them, but nothing can replace careful scrutiny of the final model by an experienced researcher).

Many possible choices, none universally accepted. Some (classical) possibilities:

- Prediction error based criteria (CV)
- Information criteria (AIC, BIC, ...)
- Mallow's C_p statistic

Before looking at these, let's introduce terminology: Suppose that the truth is

$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ but with $\beta_r = 0$ for some subset β_r of β.

- The true model contains only the columns for which $\beta_r \neq 0$
  - ↪ Equivalently, the true model uses $\boldsymbol{X}_\heartsuit$ as the design matrix, the latter being the matrix of columns of $\boldsymbol{X}$ corresponding to non-zero coefficients.
- A correct model is the true model plus extra columns.
  - ↪ Equivalently, a correct model has a design matrix $\boldsymbol{X}_\diamondsuit$, such that $\mathcal{M}(\boldsymbol{X}_\heartsuit) \subset \mathcal{M}(\boldsymbol{X}_\diamondsuit)$.
- A wrong model is a model that does not contain all the columns of the true model.
  - ↪ Equivalently, a wrong model has a design matrix $\boldsymbol{X}_\diamondsuit$, such that $\mathcal{M}(\boldsymbol{X}_\heartsuit) \cap \mathcal{M}(\boldsymbol{X}_\diamondsuit) \neq \mathcal{M}(\boldsymbol{X}_\heartsuit)$.

▶ We may wish to choose a model by minimising the error we make on average, when predicting a future observation given our model.
Our "experiment" is:

- Design matrix $\boldsymbol{X}$
- response $\boldsymbol{Y}$ at $\boldsymbol{X}$

Every model $f \in 2^{\boldsymbol{X}}$, will yield fitted values $\hat{\boldsymbol{Y}}(f) = \boldsymbol{H}_f \boldsymbol{Y}$. And suppose we now obtain new independent responses $\boldsymbol{Y}_+$ for the same "experimental setup" $\boldsymbol{X}$.
Then, one approach is to select the model

$$f^* = \arg\min_{f \in 2^{\boldsymbol{X}}} \underbrace{\frac{1}{n} \mathbb{E} \left\{ \| \boldsymbol{Y}_+ - \hat{\boldsymbol{Y}}(f) \|^2 \right\}}_{\Delta(f)},$$

where expectation is taken over both $\boldsymbol{Y}$ and $\boldsymbol{Y}_+$.

Let $\boldsymbol{X}$ be a design matrix, and let $\boldsymbol{X}_\diamondsuit$ ($n \times p$) and $\boldsymbol{X}_\heartsuit$ ($n \times q$) be matrices built using columns of $\boldsymbol{X}$. Suppose that the **true** relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$ is

$$\boldsymbol{Y} = \underbrace{\boldsymbol{X}_\heartsuit \beta}_{\mu} + \varepsilon$$

but we use the matrix $\boldsymbol{X}_\diamondsuit$ instead of $\boldsymbol{X}_\heartsuit$ (i.e., we fit a different model). Therefore our fitted values are

$$\hat{\boldsymbol{Y}} = (\boldsymbol{X}_\diamondsuit^\top \boldsymbol{X}_\diamondsuit)^{-1} \boldsymbol{X}_\diamondsuit^\top \boldsymbol{Y} = \boldsymbol{H}_\diamondsuit \boldsymbol{Y}.$$

Now suppose that we obtain new observations $\boldsymbol{Y}_+$ corresponding to the same design $\boldsymbol{X}$

$$\boldsymbol{Y}_+ = \boldsymbol{X}_\heartsuit \boldsymbol{\beta} + \varepsilon_+ = \boldsymbol{\mu} + \varepsilon_+.$$

Then, observe that

$$\begin{aligned} \boldsymbol{Y}_+ - \hat{\boldsymbol{Y}} &= \boldsymbol{\mu} + \varepsilon_+ - \boldsymbol{H}_\diamondsuit(\boldsymbol{\mu} + \varepsilon) \\ &= (\boldsymbol{I} - \boldsymbol{H}_\diamondsuit)\boldsymbol{\mu} + \varepsilon_+ - \boldsymbol{H}_\diamondsuit \varepsilon. \end{aligned}$$

It follows that

$$\begin{aligned} \|\boldsymbol{Y}_+ - \hat{\boldsymbol{Y}}\|^2 &= (\boldsymbol{Y}_+ - \hat{\boldsymbol{Y}})^\top (\boldsymbol{Y}_+ - \hat{\boldsymbol{Y}}) \\ &= \boldsymbol{\mu}^\top (\boldsymbol{I} - \boldsymbol{H}_\diamondsuit)\boldsymbol{\mu} + \boldsymbol{\varepsilon}^\top \boldsymbol{H}_\diamondsuit \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}_+^\top \boldsymbol{\varepsilon}_+ + [\text{cross terms}]. \end{aligned}$$

Since $\mathbb{E}[\text{cross terms}] = 0$ (why?), we observe that

$$\Delta = \begin{cases} n^{-1}\boldsymbol{\mu}^\top (\boldsymbol{I} - \boldsymbol{H}_\diamondsuit)\boldsymbol{\mu} + (1 + p/n)\sigma^2, & \text{if model wrong,} \\ (1 + p/n)\sigma^2, & \text{if model correct,} \\ (1 + q/n)\sigma^2, & \text{if model true.} \end{cases}$$

- Selecting a correct model instead of the true model brings in additional variance, because $q < p$.
- Selecting a wrong model instead of the true model results in bias, since $(\boldsymbol{I} - \boldsymbol{H}_\diamondsuit)\boldsymbol{\mu} \neq 0$ when $\boldsymbol{\mu}$ is not in the column space of $\boldsymbol{X}_\diamondsuit$.
- Must find a balance between small variance (few columns in the model) and small bias (all columns in the model).

- Impossible to calculate Δ (depends on unknown $\boldsymbol{\mu}$ and σ^2), so we must find a proxy (estimator) $\widehat{\Delta}$.

Suppose that n is large so that we can split the data in two pieces:

- $\boldsymbol{X}^*$, $\boldsymbol{Y}^*$ used to estimate the model
- $\boldsymbol{X}'$, $\boldsymbol{Y}'$ used to estimate the prediction error for the model

The estimator of the prediction error will be

$$\widehat{\Delta} = (n')^{-1} \| \boldsymbol{Y}' - \boldsymbol{X}' \hat{\boldsymbol{\beta}}^* \|^2.$$

In practice n might be small and we often cannot afford to split the data (variance of $\hat{\Delta}$ is too large).
Instead we use the leave-one-out cross validation:

$$n\widehat{\Delta}_{CV} = CV = \sum_{j=1}^{n} (Y_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_{-j})^2,$$

where $\hat{\boldsymbol{\beta}}_{-j}$ is the estimate produced when dropping the jth case (line).

Does this mean we need to fit n different models?

No! Rank 1 perturbation theory proves that

$$CV = \sum_{j=1}^{n} \frac{(Y_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}})^2}{(1 - h_{jj})^2},$$

so the full regression may be used! Alternatively one may use a more stable version:

$$GCV = \sum_{j=1}^{n} \frac{(Y_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}})^2}{(1 - \text{trace}(\boldsymbol{H})/n)^2},$$

where "G" stands for "generalised", and we guard against any $h_{jj} \approx 1$.
It can be shown that:

$$\mathbb{E}[GCV] = \frac{\boldsymbol{\mu}^\top (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu}}{(1 - p/n)^2} + \frac{n\sigma^2}{1 - p/n} \approx n\Delta.$$

▷ Suggests strategy: pick variables to minimise (G)CV.

Criteria can be obtained based on the notion of *relative entropy (KL divergence)*.

- Same basic idea as for prediction error: aim to choose candidate model $f(\boldsymbol{y})$ to minimise *information distance*:

$$\int \log\left\{\frac{g(\boldsymbol{y})}{f(\boldsymbol{y})}\right\} g(\boldsymbol{y})dy \geq 0,$$

  where $g(\boldsymbol{y})$ represents true model—equivalent to maximising expected log likelihood

$$\int \log f(\boldsymbol{y})g(\boldsymbol{y})dy.$$

- Can show that (apart from constants) information distance is estimated by

$$\text{AIC} = -2\hat{\ell} + 2p \quad (\equiv n \log \hat{\sigma}^2 + 2p \text{ in linear model})$$

  where $\hat{\ell}$ is maximised log likelihood for given model, and p is number of parameters.

There are many flavours of such critera:

- Improved (corrected) version of AIC for regression problems:

$$\text{AIC}_c \equiv \text{AIC} + \frac{2p(p+1)}{n-p-1}.$$

- Also can use *Bayes' information criterion*

$$\text{BIC} = -2\hat{\ell} + p \log n.$$

- Mallows suggested

$$C_p = \frac{SS_p}{s^2} + 2p - n,$$

  where SS_p is RSS for fitted model and s^2 estimates σ^2.
- Comments:
  - AIC tends to choose models that are too complicated, buts AIC_c cures this somewhat;
  - BIC is *model selection consistent*—if the true model is among those fitted, BIC chooses it with probability $\to 1$ as $n \to \infty$ (for fixed p).

Example (Simulation Experiment)

For each $n \in \{10, 20, 40\}$ we construct 20 $n \times 7$ design matrices. We multiply each of these design matrices from the right with $\beta = (1, 2, 3, 0, 0, 0, 0)^\top$ and we add a $n \times 1$ Gaussian error. We do this independently 50 times, obtaining 1000 regressions with $p = 3$. Selected models with 1 or 2 covariates have a bias term, and those with 4 or more covariates have excess variance.

n		Number of covariates						
		1	2	3	4	5	6	7
10	C_p		131	504	91	63	83	128
	BIC		72	373	97	83	109	266
	AIC		52	329	97	91	125	306
	AIC_c	15	398	565	18	4		
20	C_p		4	673	121	88	61	53
	BIC		6	781	104	52	30	27
	AIC		2	577	144	104	76	97
	AIC_c		8	859	94	30	8	1
40	C_p			712	107	73	66	42
	BIC			904	56	20	15	5
	AIC			673	114	90	69	54
	AIC_c			786	105	52	41	16

- We saw so far:

Automatic Model Selection: build a set of models and select the "best" one.

- Now look at different philosophy:

Automatic Model Building: construct a single model in a way that would hopefully provide a good one.

There golden oldies for doing this are:

- Forward Selection
- Backward Elimination
- Stepwise Selection

Caution: Although widely used, these have little theoretical basis. Element of arbitrariness ...

- *Forward selection*: starting from the model with constant only,
  1. add each remaining term separately to the current model;
  2. if none of these terms is significant, stop; otherwise
  3. update the current model to include the most significant new term; go to step 1.

- *Backward elimination*: starting from the model with all terms,
  1. if all terms are significant, stop; otherwise
  2. update current model by dropping the term with the smallest F statistic; go to step 1.

- *Stepwise*: starting from an arbitary model,
  1. consider three options—add a term, delete a term, swap a term in the model for one not in the model, and choose the most significant option;
  2. if model unchanged, stop; otherwise go to step 1.

Some thoughts:

- Each procedure may produce a different model.
- Systematic search minimising Prediction Error, AIC or similar over all possible models is preferable— BUT not always feasible (e.g., when p large).
- Stepwise methods can fit 'highly significant' models to purely random data! Main problem is lack of objective function.
- Can be improved by comparing Prediction Error/AIC for different models at each step — uses objective function, but no systematic search.

Example (Nuclear Power Station Data)

Data on light water reactors (LWR) constructed in the USA. The covariates are `date` (date construction permit issued), `T1` (time between application for and issue of permit), `T2` (time between issue of operating license and construction permit), `capacity` (power plant capacity in MWe), `PR` (=1 if LWR already present on site), `NE` (=1 if constructed in north-east region of USA), `CT` (=1 if cooling tower used), `BW` (=1 if nuclear steam supply system manufactured by Babcock–Wilcox), `N` (cumulative number of power plants constructed by each architect-engineer), `PT` (=1 if partial turnkey plant).

	cost	date	T_1	T_2	capacity	PR	NE	CT	BW	N	PT
1	460.05	68.58	14	46	687	0	1	0	0	14	0
2	452.99	67.33	10	73	1065	0	0	1	0	1	0
3	443.22	67.33	10	85	1065	1	0	1	0	1	0
4	652.32	68.00	11	67	1065	0	1	1	0	12	0
5	642.23	68.00	11	78	1065	1	1	1	0	12	0
6	345.39	67.92	13	51	514	0	1	1	0	3	0
7	272.37	68.17	12	50	822	0	0	0	0	5	0
8	317.21	68.42	14	59	457	0	0	0	0	1	0
⋮											
32	270.71	67.83	7	80	886	1	0	0	1	11	1

Example (Nuclear Power Station Data, continued)

	Full model		Backward		Forward	
	Est	t	Est	t	Est	t
Int.	−14.24	−3.37	−13.26	−4.22	−7.62	−2.66
`date`	0.2	3.21	0.21	4.91	0.13	3.38
`logT1`	0.092	0.38				
`logT2`	0.29	1.05				
`logcap`	0.694	5.10	0.72	6.09	0.67	4.75
`PR`	−0.092	−1.20				
`NE`	0.25	3.35	0.24	3.36		
`CT`	0.12	1.82	0.14			
`BW`	0.033	0.33				
`log(N)`	−0.08	−1.74	−0.08	−2.11		
`PT`	−0.22	−1.83	−0.22	−1.99	−0.49	−4.77
s (df)	0.164 (21)		0.159 (25)		0.195 (28)	

Recall: $\hat{\boldsymbol{Y}}$ is projection of $\boldsymbol{Y}$ onto $\mathcal{M}(\boldsymbol{X})$

$\hookrightarrow$ Adding more variables (columns) into $\boldsymbol{X}$ "enlarges" $\mathcal{M}(\boldsymbol{X})$

. . . if the rank increases by the # of new variables

Consider two extremes

- Adding a new variable (column) $\boldsymbol{X}_{p+1} \in \mathcal{M}^{\perp}(\boldsymbol{X})$
  - $\hookrightarrow$ Gives us completely "new" information.
- Adding a new variable (column) $\boldsymbol{X}_{p+1} \in \mathcal{M}(\boldsymbol{X})$
  - $\hookrightarrow$ Gives no "new" information — cannot even do least squares (why not?)

What if we are between the two extremes? What if

$$\boldsymbol{X}_{p+1} \notin \mathcal{M}(\boldsymbol{X}) \quad \text{but} \quad \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}_{p+1} = \boldsymbol{H}\boldsymbol{X}_{p+1} \simeq \boldsymbol{X}_{p+1}?$$

We can certainly fit the regression, but what will happen?

Using block matrix properties, have

$$\mathrm{var}(\hat{\beta}) = \sigma^2 \left[(\boldsymbol{X}\ \boldsymbol{X}_{p+1})^\top (\boldsymbol{X}\ \boldsymbol{X}_{p+1}) \right]^{-1}$$

with

$$\left[(\boldsymbol{X}\ \boldsymbol{X}_{p+1})^\top (\boldsymbol{X}\ \boldsymbol{X}_{p+1}) \right]^{-1} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$$

where

$$\begin{aligned}
\boldsymbol{A} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} X^\top \boldsymbol{X}_{p+1} \\
&\qquad \times (\boldsymbol{X}_{p+1}^\top \boldsymbol{X}_{p+1} - \boldsymbol{X}_{p+1}^\top \boldsymbol{H} \boldsymbol{X}_{p+1})^{-1} \boldsymbol{X}_{p+1}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \\
\boldsymbol{B} &= -(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}_{p+1} (\boldsymbol{X}_{p+1}^\top \boldsymbol{X}_{p+1} - \boldsymbol{X}_{p+1}^\top \boldsymbol{H} \boldsymbol{X}_{p+1})^{-1}, \\
\boldsymbol{C} &= -(\boldsymbol{X}_{p+1}^\top \boldsymbol{X}_{p+1} - \boldsymbol{X}_{p+1}^\top \boldsymbol{H} \boldsymbol{X}_{p+1})^{-1} \boldsymbol{X}_{p+1}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \\
\boldsymbol{D} &= (\boldsymbol{X}_{p+1}^\top \boldsymbol{X}_{p+1} - \boldsymbol{X}_{p+1}^\top \boldsymbol{H} \boldsymbol{X}_{p+1})^{-1}.
\end{aligned}$$

Multicollinearity: when p covariates concentrate around a subspace of dimension $q < p$

[simplest case: pairs of variables that are correlated]

But: might exist even if pairs of variables appear uncorrelated!

Can be caused by:

- Poor design [can try designing again],
- Inherent relationships [other remedies needed].

So what are the results?

- Huge variances of the estimators!
  - ↪ Can even flip signs for different data, to give the impression of inverse effects.
- Individual coefficients insignificant:
  - ↪ t-test p-values inflated.
- But global F-test might give significant result!

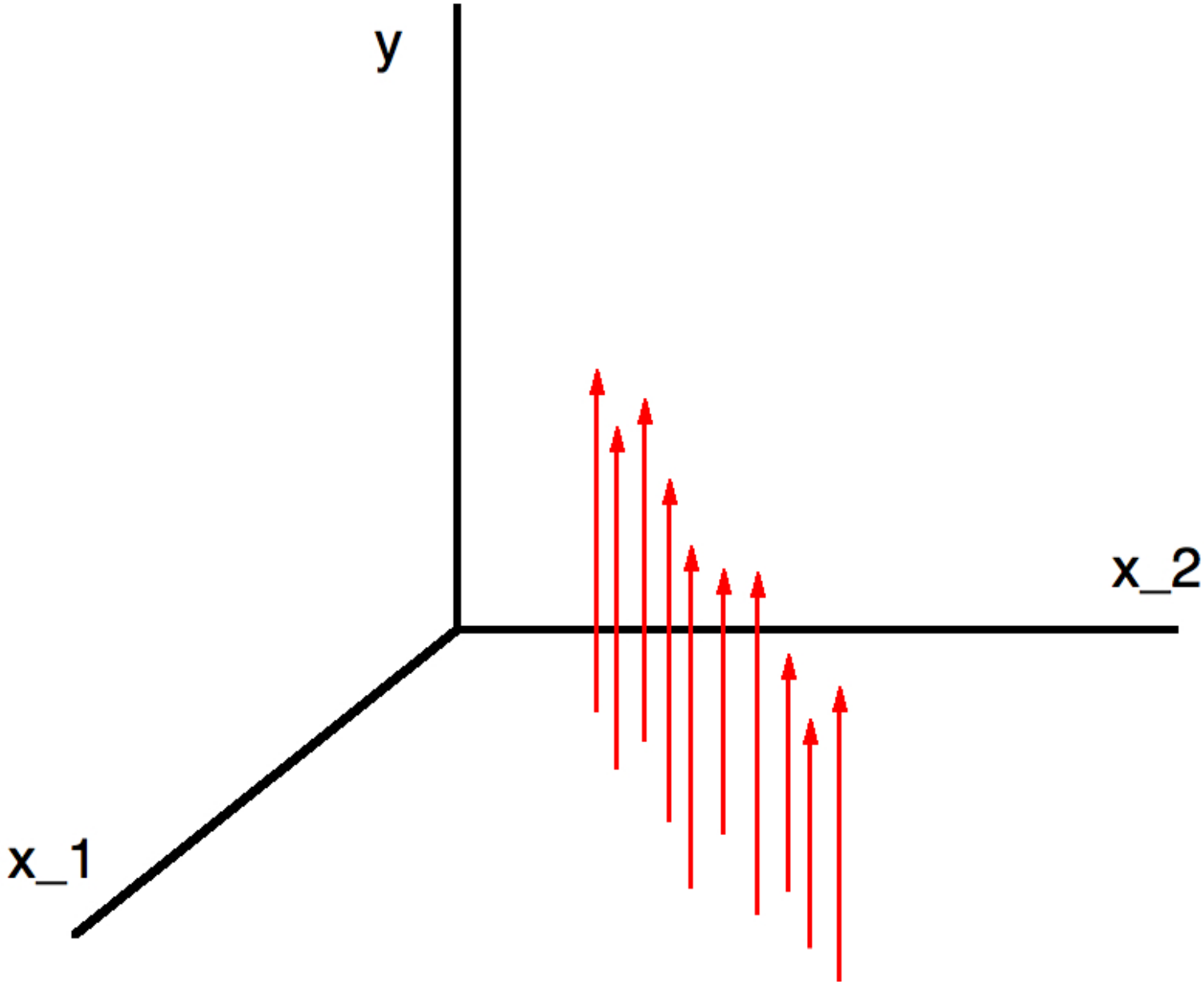


Figure: The Picket-Fence (Hocking & Pendleton)

Simple first steps:

- Look at scatterplots,
- Look at correlation matrix of covariates,

Might not reveal more complex linear constraints, though.

- Look at the *variance inflation factors*:

$$VIF_j = \frac{\mathrm{var}(\hat{\beta}_j)\|\boldsymbol{X}_j\|^2}{\sigma^2} = \|\boldsymbol{X}_j\|^2 \left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right]_{jj}.$$

Can show that

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination for the regression

$$\text{of } \boldsymbol{X}_j \text{ on } \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, \ldots, \boldsymbol{X}_p\},$$

measuring linear dependence of $\boldsymbol{X}_j$ on the other columns of $\boldsymbol{X}$.

Let $\boldsymbol{X}_{-j}$ be the design matrix without the j-th variable. Then

$$R_j^2 = \frac{\|\boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^\top \boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^\top \boldsymbol{X}_j\|^2}{\|\boldsymbol{X}_j\|^2} \in [0,1]$$

is close to 1 if $\underbrace{\boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^\top \boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}}_{\boldsymbol{H}_{-j}}\boldsymbol{X}_j \simeq \boldsymbol{X}_j$.

Large values of VIF_j indicate that $\boldsymbol{X}_j$ is linearly dependent on the other columns of the design matrix.

Interpretation: how much the variance is inflated when including variable j as compared to the variance we would obtain if $\boldsymbol{X}_j$ were orthogonal to the other variables—how much worse are we doing as compared to the ideal case.

Rule of thumb: $VIF_j > 5$ or $VIF_j > 10$ considered to be "large".

Consider the spectral decomposition of $\boldsymbol{X}^\top \boldsymbol{X}$, $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ with $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_p\}$ and $\boldsymbol{U}^\top \boldsymbol{U} = I$. Then

$$\text{rank}(\boldsymbol{X}^\top \boldsymbol{X}) = \#\{j : \lambda_j \neq 0\}, \qquad \det(\boldsymbol{X}^\top \boldsymbol{X}) = \prod_{j=1}^{p} \lambda_j.$$

Hence "small" λ_j's mean "almost" reduced rank, revealing the effect of collinearity. Measure using condition index:

$$CI_j(\boldsymbol{X}^\top \boldsymbol{X}) := \sqrt{\lambda_{\max}/\lambda_j}$$

Global "instability" measured by the condition number,

$$CN(\boldsymbol{X}^\top \boldsymbol{X}) = \sqrt{\lambda_{\max}/\lambda_{\min}}$$

Rule of thumb: $CN > 30$ indicates moderate to significant collinearity, $CN > 100$ indicates severe collinearity (choices vary).

Remedies?

If design faulty, may redesign.

$\hookrightarrow$ Otherwise? Inherent relationships between covariates.

- Variable deletion - attempt to remove problematic variables
  - $\rightarrow$ E.g., by backward elimination.

- Choose an orthogonal basis for $\mathcal{M}(\boldsymbol{X})$ and use its elements as covariates
  - $\rightarrow$ Use columns of $\boldsymbol{U}$ from spectrum, $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top$
  - $\rightarrow$ OK for prediction
  - $\rightarrow$ Problem: lose interpretability

Other approaches?

Example (Body Fat Data)

Body fat is measure of health $\rightarrow$ not easy to measure!
Collect 252 measurements on body fat and some explanatory variables.

Can we use measuring tape and scales only to find body fat?

Explanatory variables:

- age
- weight
- height
- biceps
- neck
- chest
- abdomen
- forearm
- hip
- thigh
- knee a
- ankle
- wrist

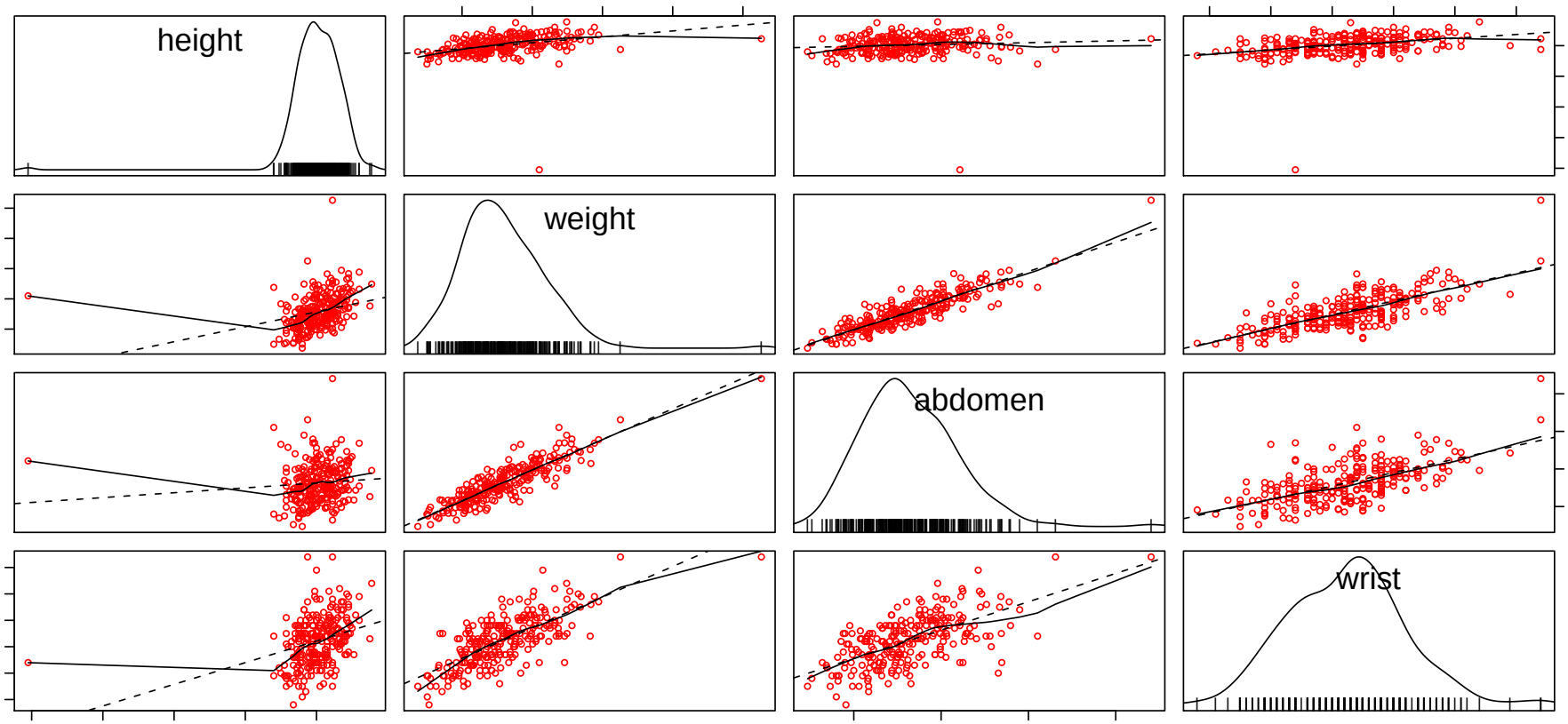


Figure: Some Scatterplots [library(car);scatterplot.matrix(. . .)]. Looks like we're in trouble. Let's go ahead and fit anyway . . .

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−18.1885	17.3486	−1.05	0.2955
age	0.0621	0.0323	1.92	0.0562
weight	−0.0884	0.0535	−1.65	0.0998
height	−0.0696	0.0960	−0.72	0.4693
neck	−0.4706	0.2325	−2.02	0.0440
chest	−0.0239	0.0991	−0.24	0.8100
abdomen	0.9548	0.0864	11.04	0.0000
hip	−0.2075	0.1459	−1.42	0.1562
thigh	0.2361	0.1444	1.64	0.1033
knee	0.0153	0.2420	0.06	0.9497
ankle	0.1740	0.2215	0.79	0.4329
biceps	0.1816	0.1711	1.06	0.2897
forearm	0.4520	0.1991	2.27	0.0241
wrist	−1.6206	0.5349	−3.03	0.0027

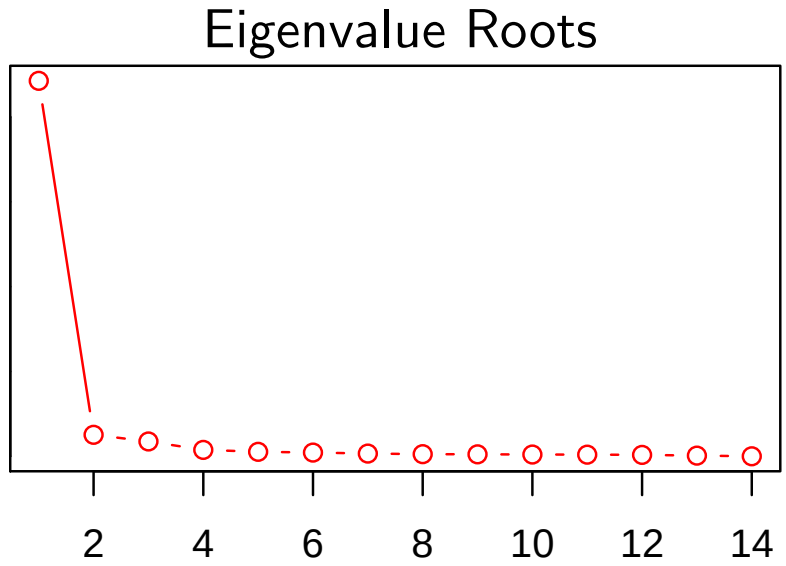
$R^2 = 0.749$, F-test: $p < 2.2 \times 10^{-16}$.

Split Data in Two and Fit Separately (Picket Fence)

	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)
(Intercept)	−32.6564	0.1393	−1.2221	0.9730
age	0.1048	0.0153	0.0256	0.6252
weight	−0.1285	0.0502	−0.0237	0.8223
height	−0.0666	0.5207	−0.1005	0.7284
neck	−0.5086	0.0721	−0.4619	0.2635
chest	0.0168	0.9002	−0.0910	0.5877
abdomen	0.9750	0.0000	0.8924	0.0000
hip	−0.2891	0.1265	−0.0265	0.9130
thigh	0.3850	0.0565	0.0334	0.8793
knee	0.2218	0.5111	−0.1310	0.7366
ankle	0.4377	0.0694	−0.5037	0.3516
biceps	−0.1297	0.5485	0.4458	0.1179
forearm	0.8871	0.0174	0.2247	0.3750
wrist	−1.7378	0.0309	−1.5902	0.0560

	VIF
age	2.25
weight	33.51
height	1.67
neck	4.32
chest	9.46
abdomen	11.77
hip	14.80
thigh	7.78
knee	4.61
ankle	1.91
biceps	3.62
forearm	2.19
wrist	3.38

	CI
1	1.00
2	17.47
3	25.30
4	58.61
5	83.59
6	100.63
7	137.90
8	175.29
9	192.62
10	213.01
11	228.16
12	268.21
13	555.67



Condition Number $\simeq$ 556 !

```
 Multiple R-Squared:  0.7466,
F-statistic p-value:  < 2.2e-16
```

	Estimate	Std. Error	t value	Pr(>|t|)	VIF
(Intercept)	−22.6564	11.7139	−1.93	0.0543	
age	0.0658	0.0308	2.14	0.0336	2.05
weight	−0.0899	0.0399	−2.25	0.0252	18.82
neck	−0.4666	0.2246	−2.08	0.0388	4.08
abdomen	0.9448	0.0719	13.13	0.0000	8.23
hip	−0.1954	0.1385	−1.41	0.1594	13.47
thigh	0.3024	0.1290	2.34	0.0199	6.28
forearm	0.5157	0.1863	2.77	0.0061	1.94
wrist	−1.5367	0.5094	−3.02	0.0028	3.09

Define $\boldsymbol{Z} = \boldsymbol{XU}$ as design matrix. R^2=0.749, F-test p-value<2.2×10^{-16}

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−18.1885	17.3486	−1.05	0.2955
Z[, 1]	−0.1353	0.0619	−2.19	0.0297
Z[, 2]	−0.0168	0.0916	−0.18	0.8546
Z[, 3]	0.2372	0.1070	2.22	0.0276
Z[, 4]	−0.7188	0.0571	−12.58	0.0000
Z[, 5]	0.0248	0.0827	0.30	0.7649
Z[, 6]	0.4546	0.1001	4.54	0.0000
Z[, 7]	0.5903	0.1366	4.32	0.0000
Z[, 8]	−0.1207	0.1742	−0.69	0.4890
Z[, 9]	−0.0836	0.1914	−0.44	0.6627
Z[, 10]	0.5043	0.2082	2.42	0.0162
Z[, 11]	−0.5735	0.2254	−2.54	0.0116
Z[, 12]	0.3007	0.2628	1.14	0.2536
Z[, 13]	1.5168	0.5447	2.78	0.0058

- Eigenvector approach rotates space so as to "free" the dependence of one coefficient β_j on others $\{\beta_i\}_{i \neq j}$

  $\hookrightarrow$ Imposes constraint on $\boldsymbol{X}$ (orthogonal columns)

Problem: lose interpretability! (prediction OK)

  - Example: most significant "rotated" term in fat data: `Z[,4]=-0.01*age -0.058*weight -0.011*height +0.46*neck -0.144*chest -0.441*abdomen +0.586*hip +0.22*thigh -0.197*knee -0.044*ankle -0.07*biceps -0.33*forearm -0.249*wrist`

- Other approach to reduce this strong dependence?

  $\hookrightarrow$ Impose constraint on $\boldsymbol{\beta}$! How? (introduces bias)