Statistics for Data Science: Week 9.2

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

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

Distribution Theory of Least Squares

Gaussian Linear Model:

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_{n\times n})$$

We have derived the estimators:

- $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$
- $\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\|^2$
- $S^2 = \frac{1}{n-p}\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\|^2$

We need to study the sampling distribution of these estimators for the purpose of:

- Understanding their precision
- Building confidence intervals
- Testing hypotheses
- Comparing them to other candidate estimators
- ...

Theorem (Sampling Distribution of LSE under Gaussian Model)

Let $\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_{n\times n})$ *and assume that* $\boldsymbol{X}$ *has full rank* $p < n$. *Then,*

1. $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p\{\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\}$;
2. *the random variables* $\hat{\boldsymbol{\beta}}$ *and* S^2 *are independent; and*
3. $\dfrac{n-p}{\sigma^2} S^2 \sim \chi^2_{n-p}$, *where* χ^2_ν *denotes the chi-square distribution with* ν *degrees of freedom.*

Proof.

1. Recall our results for linear transformations of Gaussian variables:

$$\left.\begin{array}{l}\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} \\ \boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I})\end{array}\right\} \implies \hat{\boldsymbol{\beta}} \sim \mathcal{N}_p\{\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\}$$

2. If $\boldsymbol{e}$ is independent of $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, then $S^2 = \boldsymbol{e}^\top \boldsymbol{e}/(n-p)$ will be independent of $\hat{\boldsymbol{\beta}}$ (why?). Now notice that:

- $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, with $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$
- $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$
- $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$

Therefore, from the properties of the Gaussian distribution $\boldsymbol{e}$ is independent of $\hat{\boldsymbol{Y}}$ since $(\boldsymbol{I} - \boldsymbol{H})(\sigma^2 \boldsymbol{I})\boldsymbol{H} = \sigma^2(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{H} = 0$, by idempotency of $\boldsymbol{H}$.

3. For the last part recall that

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\varepsilon \implies (n-p)S^2 = (n-p)\frac{\boldsymbol{e}^\top \boldsymbol{e}}{n-p} = \varepsilon^\top(\boldsymbol{I} - \boldsymbol{H})\varepsilon$$

by idempotency of $\boldsymbol{H}$. But recall that $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I})$ so $\sigma^{-1}\varepsilon \sim \mathcal{N}_n(0, \boldsymbol{I})$. Therefore, by the properties of normal quadratic forms,

$$\frac{(n-p)}{\sigma^2} S^2 = (\sigma^{-1}\varepsilon)^\top(\boldsymbol{I} - \boldsymbol{H})(\sigma^{-1}\varepsilon) \sim \chi^2_{n-p}.$$

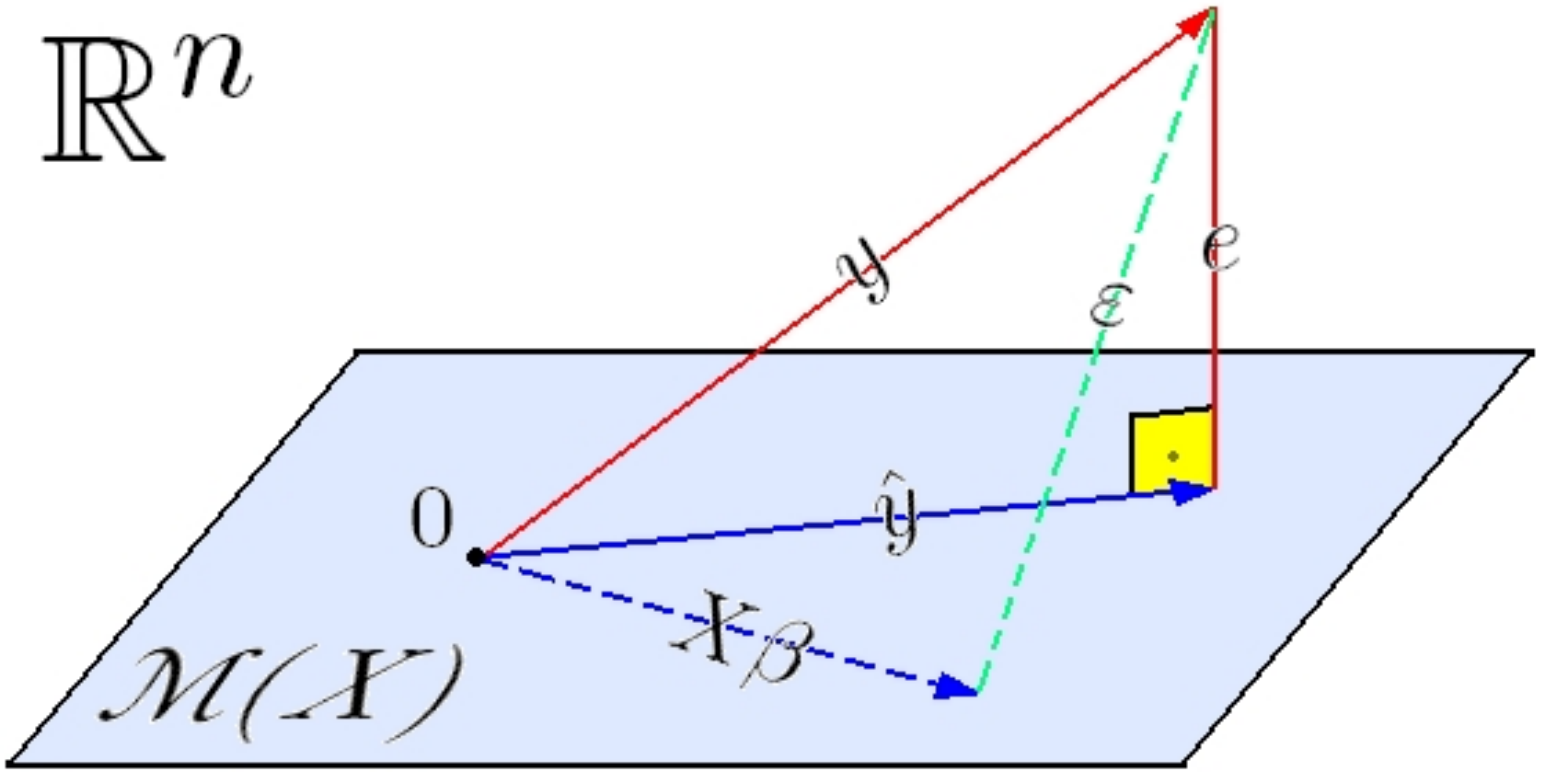
Recall that $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

Corollary

Let $\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_{n\times n})$. The statistic $\boldsymbol{HY}$ is sufficient for the parameter $\boldsymbol{\beta}$. If $\boldsymbol{X}$ has full rank $p < n$, then $\hat{\boldsymbol{\beta}}$ is also sufficient for $\boldsymbol{\beta}$.

Corollary

S^2 is unbiased whereas $\hat{\sigma}^2$ is biased (so we prefer S^2).


$\mathbb{R}^n$
y
ε
e
0
$\hat{y}$
$X\beta$
$\mathcal{M}(X)$

Proof of the first Corollary.

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

If we can show that the conditional distribution of the $2n$-dimensional vector $\boldsymbol{W} = (\hat{\boldsymbol{Y}}, \boldsymbol{e})^\top$ given $\hat{\boldsymbol{Y}}$ does not depend on β, then we will also know that the conditional distribution of $\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \boldsymbol{e}$ given $\hat{\boldsymbol{Y}}$ does not depend on β either, proving the proposition.

But we have proven that $\hat{\boldsymbol{Y}}$ is independent of $\boldsymbol{e}$. Therefore, conditional on $\hat{\boldsymbol{Y}}$, $\boldsymbol{e}$ always has the same distribution $\mathcal{N}(0, (\boldsymbol{I} - \boldsymbol{H})\sigma^2)$. It follows that, conditional on $\hat{\boldsymbol{Y}}$, the vector $\boldsymbol{W}$ has a distribution whose first n coordinates equal $\hat{\boldsymbol{Y}}$ almost surely, and whose last n coordinates are $\mathcal{N}(0, (\boldsymbol{I} - \boldsymbol{H})\sigma^2)$. Neither of those two depend on β, and the proof is complete.

When $\boldsymbol{X}$ has full rank, $\hat{\beta}$ is a 1-1 function of $\boldsymbol{H}\boldsymbol{y}$, and is also sufficient for β.

□

Proof of the second Corollary.

Recall that if $Q \sim \chi^2_d$, then $\mathbb{E}[Q] = d$. □

We have characterized various properties of LSE,
now what about characterizing whether the resulting linear model is *good*?

How to construct $1-\alpha$ CI for a linear combination of the parameters, $\boldsymbol{c}^\top\boldsymbol{\beta}$?

- Have $\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} \sim \mathcal{N}_1(\boldsymbol{c}^\top\boldsymbol{\beta}, \sigma^2\boldsymbol{c}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{c}) = \mathcal{N}_1(\boldsymbol{c}^\top\boldsymbol{\beta}, \sigma^2\delta)$
- Therefore $Q = (\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} - \boldsymbol{c}^\top\boldsymbol{\beta})/(\sigma\sqrt{\delta}) \sim \mathcal{N}_1(0,1)$
- Hence $Q^2 \sim \chi^2_1$
- and Q^2 is independent of S^2 (since $\hat{\beta}$ is independent of S^2)
- while $\frac{n-p}{\sigma^2}S^2 \sim \chi^2_{n-p}$.

In conclusion:

$$\frac{\frac{Q^2}{1}}{\frac{\frac{(n-p)}{\sigma^2}S^2}{n-p}} \sim F_{1,n-p} \Rightarrow \frac{\frac{(\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}-\boldsymbol{c}^\top\boldsymbol{\beta})^2}{\sigma^2\delta}}{\frac{S^2}{\sigma^2}} = \left(\frac{\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} - \boldsymbol{c}^\top\boldsymbol{\beta}}{\sqrt{S^2\boldsymbol{c}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{c}}}\right)^2 \sim F_{1,n-p}$$

- But for real W, $W^2 \sim F_{1,n-p} \iff W \sim t_{n-p}$, so base CI on:

$$\frac{\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} - \boldsymbol{c}^\top\boldsymbol{\beta}}{\sqrt{S^2\boldsymbol{c}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim t_{n-p}$$

- We obtain $(1-\alpha) \times 100\%$ CI:

$$\boldsymbol{c}^\top \hat{\boldsymbol{\beta}} \pm t_{n-p}(\alpha/2)\sqrt{S^2 \boldsymbol{c}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{c}}.$$

- What about a $(1-\alpha)$ CI for β_r? (rth coordinate)
- Let $\boldsymbol{c}_r = (0, 0, \ldots, 0, \underbrace{1}_{r^{th} \text{ position}}, 0, \ldots, 0)$
- Then $\beta_r = \boldsymbol{c}_r^\top \boldsymbol{\beta}$
- Therefore, base CI on

$$\frac{\boldsymbol{c}_r^\top \hat{\boldsymbol{\beta}} - \boldsymbol{c}_r^\top \boldsymbol{\beta}}{\sqrt{S^2 \boldsymbol{c}_r^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{c}_r}} = \frac{\hat{\beta}_r - \beta_r}{\sqrt{S^2 v_{r,r}}} \sim t_{n-p},$$

  where $v_{r,s}$ is the r, s element of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.
- Obtain $(1-\alpha) \times 100\%$ CI:

$$\hat{\beta}_r \pm t_{n-p}(\alpha/2)\sqrt{S^2 v_{rr}}.$$

A prediction interval aims to give confidence bounds on a potential response.

- Suppose we want to predict the value of Y_+ for an $\boldsymbol{x}_+ \in \mathbb{R}^p$
- Our model predicts Y_+ by $\boldsymbol{x}_+^\top \hat{\boldsymbol{\beta}}$.
- But $Y_+ = \boldsymbol{x}_+^\top \boldsymbol{\beta} + \varepsilon_+$ so a prediction interval is different from an interval for a linear combination $\boldsymbol{c}^\top \boldsymbol{\beta}$ (extra uncertainty due to ε_+):
  - $\mathbb{E}[\boldsymbol{x}_+^\top \hat{\boldsymbol{\beta}} + \varepsilon_+] = \boldsymbol{x}_+^\top \boldsymbol{\beta}$
  - $\mathrm{var}[\boldsymbol{x}_+^\top \hat{\boldsymbol{\beta}} + \varepsilon_+] = \mathrm{var}[\boldsymbol{x}_+^\top \hat{\boldsymbol{\beta}}] + \mathrm{var}[\varepsilon_+] = \sigma^2[\boldsymbol{x}_+^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_+ + 1]$
- Base prediction interval on:

$$\frac{\boldsymbol{x}_+^\top \hat{\boldsymbol{\beta}} - Y_+}{\sqrt{S^2\{1 + \boldsymbol{x}_+^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_+\}}} \sim t_{n-p}.$$

- Obtain $(1 - \alpha)$ prediction interval:

$$\boldsymbol{x}_+^\top \hat{\boldsymbol{\beta}} \pm t_{n-p}(\alpha/2) \sqrt{S^2\{1 + \boldsymbol{x}_+^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_+\}}.$$

R^2 is a *measure of fit* of the model to the data.

- We are trying to best approximate $\boldsymbol{Y}$ through an element of the column-space of $\boldsymbol{X}$.
- How successful are we? Squared error is $\boldsymbol{e}^\top \boldsymbol{e}$.
- How large is this, relative to data variation? Look at

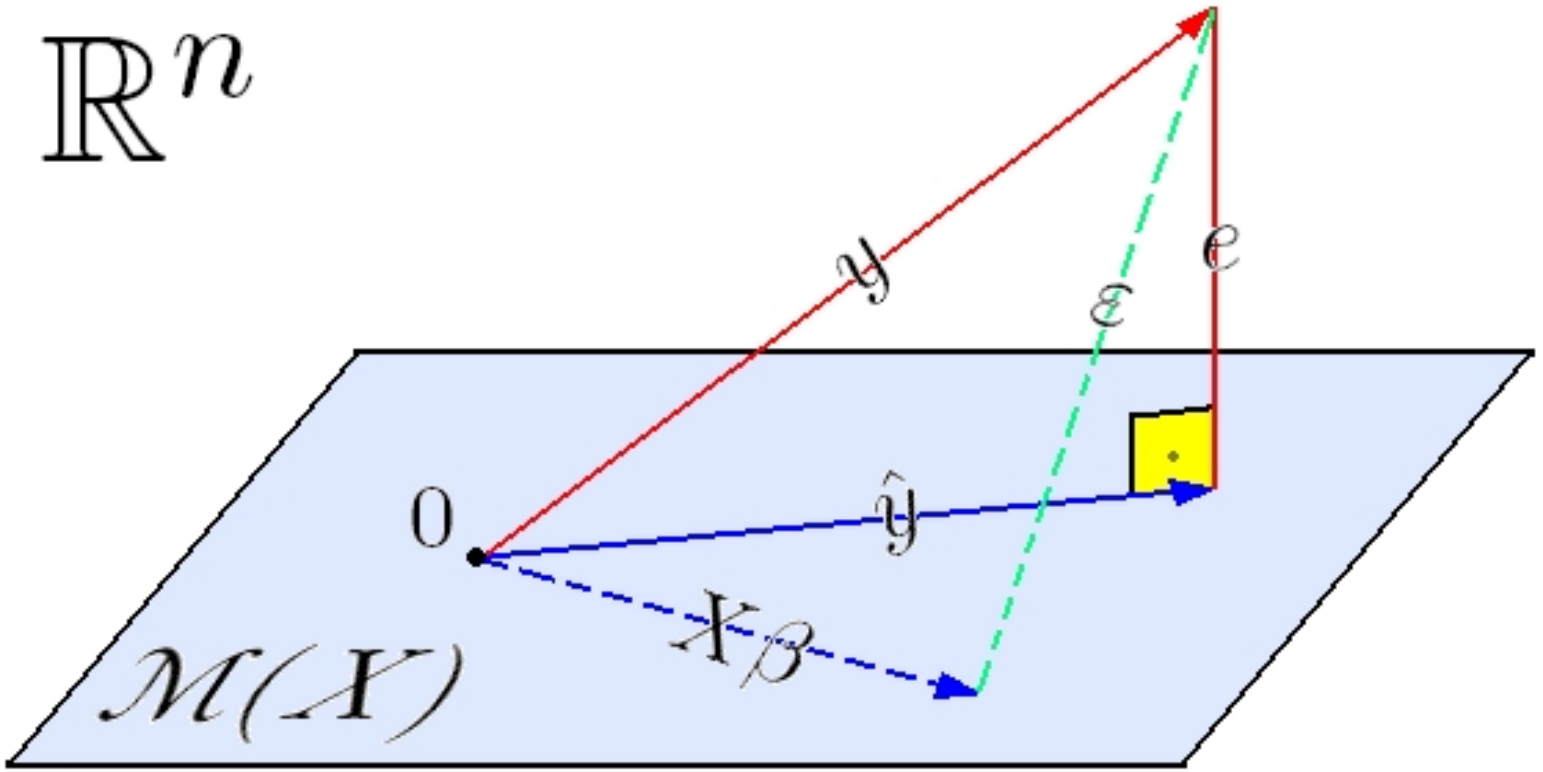
$$\frac{\|\boldsymbol{e}\|^2}{\|\boldsymbol{Y}\|^2} = \frac{\boldsymbol{e}^\top \boldsymbol{e}}{\boldsymbol{Y}^\top \boldsymbol{Y}} = \frac{\boldsymbol{Y}^\top (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}}{\boldsymbol{Y}^\top \boldsymbol{Y}} = 1 - \frac{\hat{\boldsymbol{Y}}^\top \hat{\boldsymbol{Y}}}{\boldsymbol{Y}^\top \boldsymbol{Y}}$$

- Define

$$R_0^2 = \frac{\hat{\boldsymbol{Y}}^\top \hat{\boldsymbol{Y}}}{\boldsymbol{Y}^\top \boldsymbol{Y}} = \frac{\|\hat{\boldsymbol{Y}}\|^2}{\|\boldsymbol{Y}\|^2}$$

- Note that $0 \leq R_0^2 \leq 1$

Interpretation: what proportion of the squared norm of $\boldsymbol{Y}$ does our fitted value $\hat{\boldsymbol{Y}}$ explain?


$\mathbb{R}^n$
y
ϵ
e
0
$\hat{y}$
$X\beta$
$\mathcal{M}(X)$

"**Centred (in fact, usual)** R^2". Compares empirical variance of $\hat{\boldsymbol{Y}}$ to empirical variance of $\boldsymbol{Y}$, instead of the empirical norms. In other words:

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n \hat{Y}_i^2 - n\bar{Y}^2}{\sum_{i=1}^n Y_i^2 - n\bar{Y}^2}.$$

(note that $\frac{1}{n}\sum_{i=1}^n \hat{Y}_i = \frac{1}{n}\sum_{i=1}^n (Y_i - e_i) = \bar{Y}$ because $\boldsymbol{e} \perp \mathbf{1}$ (here $\mathbf{1}$ is the vector of 1's = first column of design matrix $\boldsymbol{X}$) so $\sum_i e_i = 0$.

Note that

$$R^2 = \frac{\|\hat{\boldsymbol{Y}}\|^2 - \|\bar{Y}\mathbf{1}\|^2}{\|\boldsymbol{Y}\|^2 - \|\bar{Y}\mathbf{1}\|^2}.$$

- R_0^2 mathematically more natural (does not treat first column of $\boldsymbol{X}$ as special).
- R^2 statistically more relevant (expresses variance—the first column of $\boldsymbol{X}$ usually *is* special, in statistical terms!).
- R_0^2 and R^2 may differ a lot when $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ large.

Geometrical interpretation of R^2: project $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ on orthogonal complement of $\mathbf{1}$, then compare the norms (of the projections):

- $\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\boldsymbol{Y} = \mathbf{1}n^{-1}\sum_{i=1}^n Y_i = \mathbf{1}\bar{Y}$.
- $\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\hat{\boldsymbol{Y}} = \mathbf{1}n^{-1}\sum_{i=1}^n \hat{Y}_i = \mathbf{1}\bar{Y}$.

So

$$R^2 = \frac{\|\hat{\boldsymbol{Y}}\|^2 - \|\bar{Y}\mathbf{1}\|^2}{\|\boldsymbol{Y}\|^2 - \|\bar{Y}\mathbf{1}\|^2} = \frac{\|(I - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1})\hat{\boldsymbol{Y}}\|^2}{\|(I - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1})\boldsymbol{Y}\|^2}$$

Intuition: Should not take into account the part of $\|\boldsymbol{Y}\|$ that is explained by a constant, we only want to see the effect of the explanatory variables.

Note: Statistical packages (e.g., R) provide R^2 (and/or R^2_a, see below), not R^2_0.

Exercise: Show that $R^2 = [\text{corr}(\{\hat{Y}_i\}_{i=1}^n, \{Y_i\}_{i=1}^n)]^2$.
Exercise: Show that $R^2 \leq R^2_0$.

The adjusted R^2 takes into account the number of variables employed. It is defined as:

$$R_a^2 = 1 + (1 - R^2)\frac{n-1}{n-p}.$$

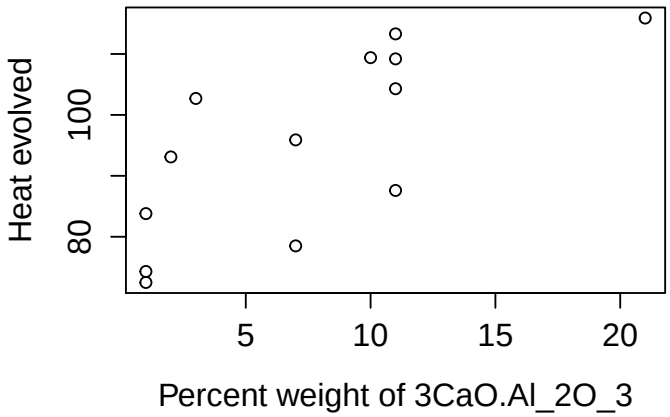
Corrects for the fact that we can always increase R^2 by adding variables. One can also correct the un-centred R_0^2 and take into account instead

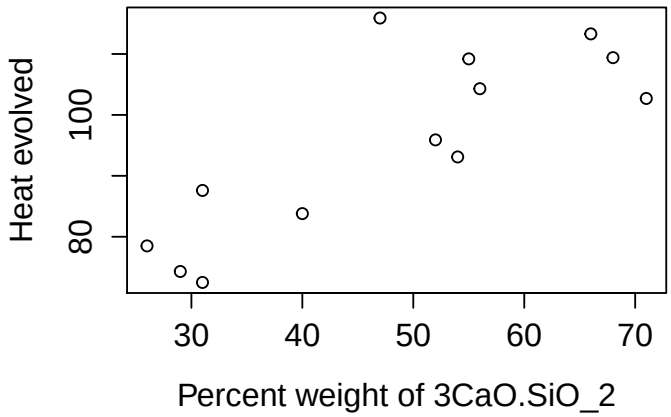
$$R_{a0}^2 = 1 + (1 - R_0^2)\frac{n}{n-p}.$$

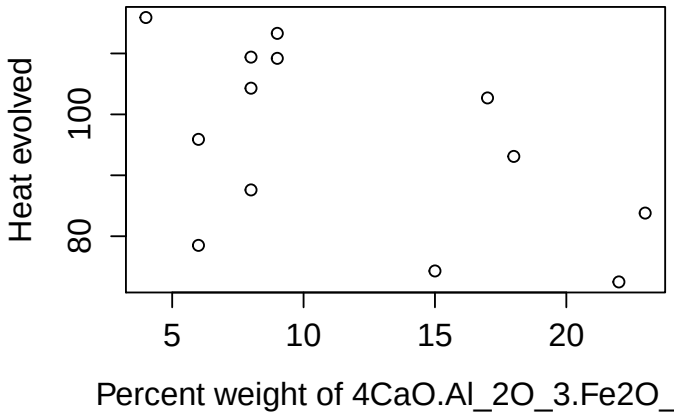
Example: Cement Heat Evolution

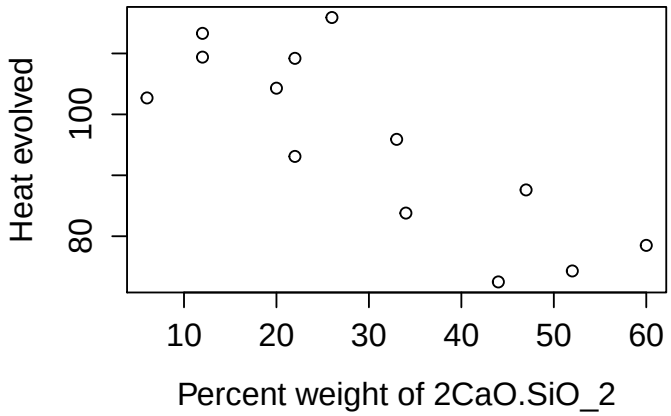
Case	$3CaO.Al_2O_3$	$3CaO.SiO_2$	$4Cao.Al_2O_3.Fe_2O_3$	$2CaO.SiO_2$	Heat
1	7.00	26.00	6.00	60.00	78.50
2	1.00	29.00	15.00	52.00	74.30
3	11.00	56.00	8.00	20.00	104.30
4	11.00	31.00	8.00	47.00	87.60
5	7.00	52.00	6.00	33.00	95.90
6	11.00	55.00	9.00	22.00	109.20
7	3.00	71.00	17.00	6.00	102.70
8	1.00	31.00	22.00	44.00	72.50
9	2.00	54.00	18.00	22.00	93.10
10	21.00	47.00	4.00	26.00	115.90
11	1.00	40.00	23.00	34.00	83.80
12	11.00	66.00	9.00	12.00	113.30
13	10.00	68.00	8.00	12.00	109.40

Example: Cement Heat Evolution









```
> cement.lm<-lm(y~1+x1+x2+x3+x4,data=cement)
> summary(cement.lm)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	62.4054	70.0710	0.89	0.3991
x1	1.5511	0.7448	2.08	0.0708
x2	0.5102	0.7238	0.70	0.5009
x3	0.1019	0.7547	0.14	0.8959
x4	−0.1441	0.7091	−0.20	0.8441

```
Residual standard error:  2.446 on 8 degrees of freedom
R-Squared:  0.9824
```

```
> x.plus
[1] 25 25 25 25
predict(cement.lm,x.plus,interval="confidence",
se.fit=T,level=0.95)
```

Fit	Lower	Upper
112.8	97.5	128.2

```
predict(cement.lm,x.plus,interval="prediction",
se.fit=T,level=0.95)
```

Fit	Lower	Upper
112.8	96.5	129.2