REGRESSION

7/05

Training set:

$$\mathbb{X} = \begin{bmatrix} X_1 & X_2 & \cdots & X_p & Y \\ x_{11} & x_{12} & \cdots & x_{1p} & y_1 \\ x_{21} & x_{22} & & x_{2p} & y_2 \\ \vdots & & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_n \end{bmatrix}$$

$\underline{X}$ covariates — Y output variable

$\underline{x}_i \in \mathbb{R}^p$

$y_i \in \mathbb{R}$ (if y_i is a lable then this is a classification problem)

General goal: "explain" Y in terms of $\underline{X}$ (explain/predict: some models are used for prediction and don't look for the explainability, others go on the opposite way)

(≡ "explain" the distribution/variability of Y in terms of $\underline{X}$)

More specifically regression is concerned with estimating:

$$\mathbb{E}[Y \mid \underline{X} = \underline{x}] = f(\underline{x})$$

Unknown function: **REGRESSION FUNCTION**

Use $\mathbb{X}$ to estimate f by means of $\hat{f}$.

Two basic approaches:

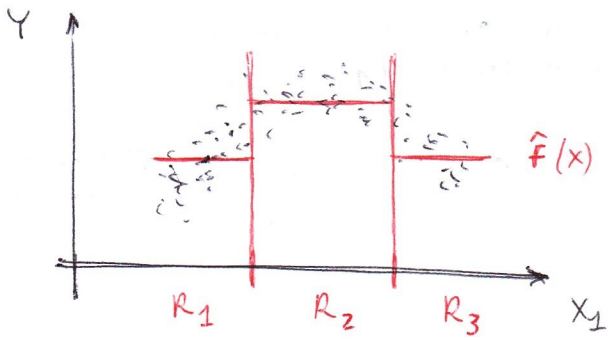
1. Totally data driven (non-parametric): good for prediction if we take good care of the bias-variance trade-off (overfitting) — it's good if we don't know much about the problem (about f)
2. Parametric approach (model based): good for interpretation and prediction, quantify the uncertainty, embedding of prior knowledge

CART

CLASSIFICATION AND REGRESSION TREES (totally data driven (1))

Basic idea: $\hat{f}$ is piecewise constant

$\Longrightarrow$ partition $\mathbb{R}^p$ (features space) in $\{R_1, \dots, R_j\}$ j finite (unknown) and assume $\hat{f}$ to be constant over the elements of the partition



Cart takes specific values for these constants, which are the means of the y we observed in the partition.

Let $\bar{y}_i = \frac{1}{\#R_i} \sum_{x_j \in R_i} y_i \qquad i = 1, \dots, j$

$$\Longrightarrow \boxed{\hat{f}(\underline{x}) = \sum_{i=1}^{j} y_i \, \mathbb{1}\{\underline{x} \in R_j\}}$$

Note: $\{R_1, \dots, R_j\}$ is a finite partition of $\mathbb{R}^p$ such that:

$R_i \cap \{\underline{x}_1, \dots, \underline{x}_n\} \neq \emptyset \quad i = 1, \dots, j$

we don't want partitions with empty regions (we wouldn't be able to compute the mean in those regions)

How do we find $\{R_1, \dots, R_j\}$ and j?

Goal: minimize $_{\{R_1, \dots, R_j\}, j} \sum_{i=1}^{j} \sum_{\underline{x} \in R_j} (y_i - \bar{y}_j)^2$

This reminds of K-means: we're trying to split the $\underline{x}$'s (to find groups among the $\underline{x}$'s) minimizing not the variability of the $\underline{x}$'s within the group but minimizing the variability of the y's (the thing we want to predict)

without overfitting the data ← this is the real problem (if we create a set for each point then the objective function will be zero, this solves the optimality problem but it's obviously overfitting the data (we have to remember bias-variance trade off))

CART → **greedy algorithm** for "solving" the problem

Iterative process: (CART)

1. Consider the cut off $s_1 \in \mathbb{R}$ such that

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\text{total variability}} - \left[\sum_{\{x_{j_1} \leq s_1\}} (y_j - \bar{y}_1)^2 + \sum_{\{x_{j_1} > s_1\}} (y_j - \bar{y}_2)^2 \right] \qquad (*)$$

is maximized, where:

$$\bar{y}_1 = \frac{1}{\#\{j : x_{j_1} \leq s_1\}} \sum_{\{x_{j_1} \leq s_1\}} y_j$$

$$\bar{y}_2 = \frac{1}{\#\{j : x_{j_1} > s_1\}} \sum_{\{x_{j_1} > s_1\}} y_i$$

= the mean of the y's for all the units for which the first variable (x_{j_1}) is less than s_1

What we are trying to do is: we have the variability of the y's, we try to find the split with respect to the first variable that (sort of) makes the larger difference in total variability (*)

Repeat for $x_2, x_3, \dots, x_p \Longrightarrow$ it generates the splits: $s_1, \dots, s_p$

Choose the cutoff s_j^* which maximizes $(*)$.

For each split we have a value of $(*)$; we choose the one that maximizes $(*)$.

(*) we're trying to find the split so that the sum of the variability within the two groups is less than the big variability that we had before the split.

Hence $\mathbb{R}^p$ is split:

$$R_1 = \{\underline{x} \in \mathbb{R}^p : x_j \leq s_j^*\}$$

$$R_2 = \{\underline{x} \in \mathbb{R}^p : x_j > s_j^*\}$$

2. Now we iterate the same procedure on the elements of the partition. (In the first case we iterate on R_1 or R_2)

Stop partitioning an element R_j if R_j contains less than 5 units (5/10/.., just a given number of units).

The problem is overfitting.

Solution for overfitting : **PRUNING** the tree (penalize large trees) ← which are the ones that overfit data
(we grow a large tree and then we prune it)

Choose $\alpha > 0$.

Grow a large tree following the previous procedure.

Prune bottom-up the branches of the tree * with the goal of minimizing :

$$\sum_{i=1}^{m} \sum_{x_j \in R_i} (y_j - \bar{y}_i)^2 + \alpha m = W(m, \alpha)$$

* which mathematically means to merge the rectangles

→ this penalize the number of branches that we got in the tree (= number of elements of the partition)

Choose α by cross-validation.

Note: CART usually doesn't beat linear models (we can always generate a linear model that reproduces exactly the CART (it's a matter of taking the right transformations)). However it's a good thing to use CART to check what type of non-linearity the model should have.

RANDOM FORESTS
(Build an ensamble of trees and average them.)
We build many trees (many CARTS) by building many different training sets (by **BOOTSTRAP**: sampling from the training set with replacing). We build an ensamble of trees more or less uncorrelated (not totally correlated because we have sampled from training set and from the variables). Each of these tree will produce a prediction. We take the average of all the predictions as final prediction.
Note: We lose some interpretability of the model with this method.

LINEAR MODELS FOR REGRESSION

$$X = \begin{array}{c} \begin{matrix} X_1 & X_2 & \cdots & X_p & Y \end{matrix} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2p} & y_2 \\ \vdots & & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_n \end{bmatrix} \end{array}$$

Design matrix :

$$Z = \begin{array}{c} \begin{matrix} 1 & Z_1 & \cdots & Z_r \end{matrix} \\ \begin{bmatrix} 1 & z_{11} & \cdots & z_{1r} \\ \vdots & z_{21} & \cdots & z_{2r} \\ & \vdots & & \\ 1 & z_{n1} & \cdots & z_{nr} \end{bmatrix} \end{array} \quad \in \mathbb{R}^{n \times r}$$

$Z_1, \dots, Z_r$ are known functions (so transformations) of $X_1, \dots, X_p$

Linear model :

$$\mathbb{E}[Y \mid Z_1, \dots, Z_r] = \beta_0 + \beta_1 Z_1 + \dots + \beta_r Z_r$$

where the unknowns are $\beta_0, \dots, \beta_r$.

The model for Y :

given $\underline{Z} = [1, Z_1, \dots, Z_r]^T$: $\boxed{Y = f(\underline{Z}) + \varepsilon}$

such that : $\mathbb{E}[\varepsilon] = 0$ and $\varepsilon \perp\!\!\!\perp \underline{Z}$. → ε is the part that is left out (that is not connected to informations expressed by $\underline{Z}$ (so by $\underline{X}$))

with $f(\underline{Z}) = \beta_0 + \beta_1 Z_1 + \dots + \beta_r Z_r$

Consider the "data":

$$\underline{Y} = [Y_1, \dots, Y_n]^T$$

$$Z = \begin{bmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1r} \\ 1 & z_{21} & z_{22} & \cdots & z_{2r} \\ \vdots & & & & \\ 1 & z_{n1} & z_{n2} & \cdots & z_{nr} \end{bmatrix}$$

$$\Rightarrow \boxed{\begin{array}{l} \underline{Y} = Z\underline{\beta} + \underline{\varepsilon} \\ \text{where } \underline{\beta} = [\beta_0, \beta_1, \dots, \beta_r]^T, \\ \underline{\varepsilon} \in \mathbb{R}^n \text{ s.t. } \begin{cases} \mathbb{E}[\underline{\varepsilon}] = \underline{0} \\ \text{Cov}(\underline{\varepsilon}) = \sigma^2 I \end{cases} \end{array}}$$

! we have to say something about how the units are connected and we're saying that they're **UN**CORRELATED (≠ independent)

i.e.

$$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \dots + \beta_r z_{ir} + \varepsilon_i$$

with $\varepsilon_1, \dots, \varepsilon_n$:
- uncorrelated
- same variance
- mean 0
- independent of the Z's

Linear models are very flexible.

Example: ANOVA one-way : ANOVA is a particular case of linear models

$$\left.\begin{array}{l} X_{11}, \dots, X_{1n_1} \overset{iid}{\sim} N(\mu_1, \sigma^2) \\ X_{21}, \dots, X_{2n_2} \overset{iid}{\sim} N(\mu_2, \sigma^2) \\ \vdots \\ X_{g1}, \dots, X_{gn_g} \overset{iid}{\sim} N(\mu_g, \sigma^2) \end{array}\right\} \perp\!\!\!\perp$$

let $\underline{Y} = [X_{11}, X_{..}, X_{1n_1}, X_{21}, \dots, X_{2n_2}, \dots, X_{g1}, \dots, X_{gn_g}]^T$

$\underline{Y} \in \mathbb{R}^{n = n_1 + n_2 + \dots + n_g}$

$$\underset{(n \times (g+1))}{\mathbb{Z}} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & \vdots & \vdots & & \vdots \\ \vdots & 1 & 0 & & \\ & 0 & 1 & & \\ & \vdots & \vdots & & \vdots \\ & & 1 & & \\ & & 0 & & 0 \\ & & \vdots & & 1 \\ 1 & 0 & 0 & & 1 \end{bmatrix} \begin{array}{l} \updownarrow n_1 \\ \\ \updownarrow n_2 \\ \dots \\ \updownarrow n_g \end{array}$$

design matrix

"dummy variables"

not full rank (it corresponds to the overparametrization of the ANOVA : that's why we add a constraint ((*)).)

$$\underline{\beta} = [\mu, \tau_1, \dots, \tau_g]^T$$

$$\underline{Y} = \mathbb{Z}\,\underline{\beta} + \underline{\varepsilon} \qquad \underline{\varepsilon} \sim N(0, \sigma^2 I)$$

i.e.

$$X_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1, \dots, g \qquad j = 1, \dots, n_i$$

$$\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

((*)) constraint : $\sum_{i=1}^{g} n_i \tau_i = 0 \implies \tau_g = -\sum \frac{n_i}{n_g} \tau_i$

$$\implies \underset{(n \times g)}{\mathbb{Z}} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & \vdots & \vdots & & \vdots \\ \vdots & 1 & 0 & & \\ & 0 & 1 & & \\ & \vdots & \vdots & & \\ & & 1 & & 0 \\ & & 0 & & 1 \\ & & & & \vdots \\ & 0 & 0 & & 1 \\ & -n_1/n_g & -n_2/n_g & & -n_{g-1}/n_g \\ \vdots & \vdots & & & \vdots \\ 1 & -n_1/n_g & -n_2/n_g & & -n_{g-1}/n_g \end{bmatrix} \begin{array}{l} \updownarrow n_1 \\ \\ \updownarrow n_2 \\ \dots \\ \updownarrow n_{g-1} \\ \\ \updownarrow n_g \end{array}$$

full rank

Fitting the linear model

(i.e. estimating $\beta_0, \beta_1, \dots, \beta_r, \sigma^2$)

OLS: Ordinary Least Square

$$\hat{\underline{\beta}} = \arg\min_{\underline{\beta} \in \mathbb{R}^{r+1}} \| \underline{Y} - \mathbb{Z}\underline{\beta} \|^2$$

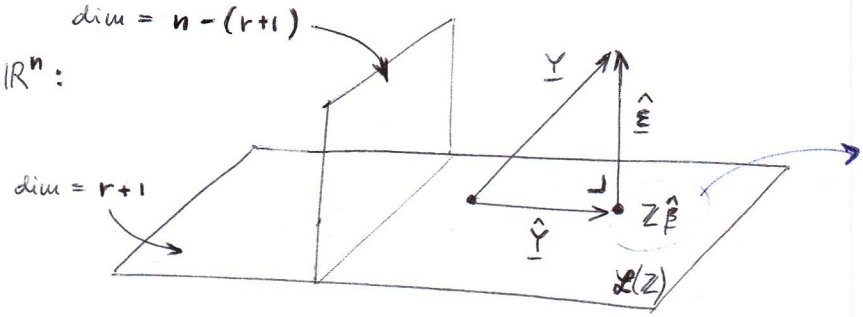
If $\mathbb{Z}$ is full rank $\Longrightarrow$ $\boxed{\hat{\underline{\beta}} = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\underline{Y}}$

(it's a linear transformation of the $\underline{Y}$ performed through a function of the design matrix)

proof. (geometrical)

$\underline{Y} = \mathbb{Z}\underline{\beta} + \underline{\varepsilon}$ model for the data

$\mathbb{Z}\underline{\beta} \in \mathcal{L}(\mathbb{Z})$ linear space spanned by the columns of $\mathbb{Z}$



we're trying to find the one vector $\in \mathcal{L}(\mathbb{Z})$ that is closest to $\underline{Y}$ in terms of Euclidean distances: this is the ortogonal projection ($\hat{\underline{Y}} = \mathbb{Z}\hat{\underline{\beta}}$)

$\hat{\underline{\beta}}$ is such that: $\boxed{\mathbb{Z}\hat{\underline{\beta}} = \Pi_{\underline{Y}|\mathcal{L}(\mathbb{Z})}}$

in order to do the projection we need an orthonormal basis of $\mathcal{L}(\mathbb{Z})$

Note that:

$\mathbb{Z}^T\mathbb{Z} \in \mathbb{R}^{(r+1)\times(r+1)}$ (full rank $\Rightarrow$ spectral decomp.)

$$\mathbb{Z}^T\mathbb{Z} = \sum_{i=1}^{r+1} \lambda_i \underline{e}_i \underline{e}_i^T \qquad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{r+1} > 0$$

full rank ($\mathbb{Z}$)

$$(\mathbb{Z}^T\mathbb{Z})^{-1} = \sum_{i=1}^{r+1} \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i^T$$

(if $\mathbb{Z}$ is not full rank: ~~$\mathbb{Z}^{-1}$~~ $\rightarrow \mathbb{Z}^-$ (pseudo inverse))

let $\underline{q}_i = \sqrt{\frac{1}{\lambda_i}}\,\mathbb{Z}\underline{e}_i$ $\quad i = 1, \dots, r+1$;

- $\underline{q}_i \in \mathcal{L}(\mathbb{Z})$
- $\underline{q}_i^T\underline{q}_j = \sqrt{\frac{1}{\lambda_i\lambda_j}}\,\underline{e}_i^T\mathbb{Z}^T\mathbb{Z}\underline{e}_j = \frac{\lambda_j}{\sqrt{\lambda_i\lambda_j}}\,\underline{e}_i^T\underline{e}_j = \begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases}$

$\Longrightarrow \{\underline{q}_1, \dots, \underline{q}_{r+1}\}$ is an orthonormal basis for $\mathcal{L}(\mathbb{Z})$

$$\Pi_{\underline{Y}|\mathcal{L}(\mathbb{Z})} = \sum_{i=1}^{r+1} \Pi_{\underline{Y}|\underline{q}_i} = \sum_{i=1}^{r+1} \frac{\underline{q}_i\underline{q}_i^T}{\underline{q}_i^T\underline{q}_i}\underline{Y}$$

$$= \sum_{i=1}^{r+1} \frac{1}{\lambda_i}\mathbb{Z}\underline{e}_i\underline{e}_i^T\mathbb{Z}^T\underline{Y}$$

$$\Pi_{\underline{Y}|\mathcal{Z}(Z)} = Z\left(\sum_{i=1}^{r+2} \frac{1}{\lambda_i}\, \underline{e}_i\, \underline{e}_i^T\right) Z^T \underline{Y}$$

$$= Z\left(Z^T Z\right)^{-1} Z^T \underline{Y}$$

$$= \hat{\underline{Y}} \qquad \text{"FITTED VALUES"}$$

$$\Pi_{\underline{Y}|\mathcal{Z}(Z)} := H\underline{Y} = \hat{Y} \qquad H = \text{"HAT MATRIX"}$$

$$\Longrightarrow \hat{\beta} = (Z^T Z)^{-1} Z^T \underline{Y} \quad (\text{since } Z\hat{\beta} = \hat{\underline{Y}} = Z(\cdot\cdot))$$

■

NOTATIONS:

$\hat{\underline{Y}} = H\underline{Y}$ — fitted values

$\hat{\underline{\varepsilon}} = \underline{Y} - H\underline{Y} = (I - H)\underline{Y}$ — residuals

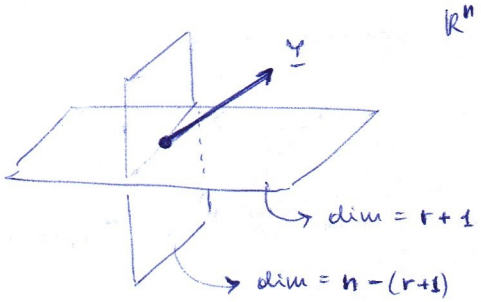
$\Longrightarrow \hat{\underline{Y}} \perp \hat{\underline{\varepsilon}}$

$\Longrightarrow \underline{Y} = \hat{\underline{Y}} + \hat{\underline{\varepsilon}}$

$\hat{\underline{Y}}$: all we can say about $\underline{Y}$ in terms of the features $\underline{X}$

$\hat{\underline{\varepsilon}}$: it's $\perp$ to $\hat{\underline{Y}}$, nothing in $\hat{\underline{\varepsilon}}$ can be captured by the informations that we have from the $\underline{X}$

Note that:



$r+1 = n$

If we take $r+1$ too large (if we take too many variables) we're filling up $\mathbb{R}^n$ with a linear space (generated by the columns of Z) that will have a perfect fitting: $\underline{Y} \in \mathcal{Y}(Z)$, no residuals. Perfect job? ABSOLUTELY NOT. Overfitting.

Linear Models:

$\underline{Y} = \mathbb{Z}\underline{\beta} + \underline{\varepsilon}$

$\underline{Y} \in \mathbb{R}^n$ (vector of dependent variables)

$\mathbb{Z} \in \mathbb{R}^{n \times (r+1)}$ design matrix

$\underline{\beta} \in \mathbb{R}^{r+1}$ (unknown)

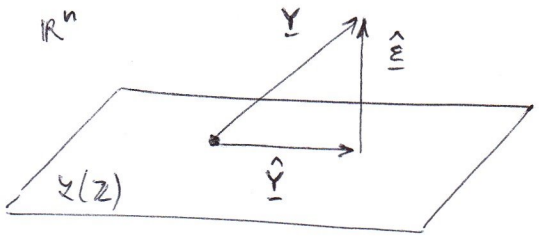
$\underline{\varepsilon}$ r.v. $\in \mathbb{R}^n$ s.t. (only random vector)
$\begin{cases} \mathbb{E}[\underline{\varepsilon}] = \underline{0} \\ \mathrm{Cov}(\underline{\varepsilon}) = \sigma^2 I \end{cases}$

OLS:

Find the vector $\hat{\underline{Y}} \in \mathcal{L}(\mathbb{Z})$ "closest" to $\underline{Y}$.

"closest" in the sense of Euclidean distance in $\mathbb{R}^n$.

Why Euclidean distance? Because $\mathrm{Cov}(\underline{\varepsilon}) = \sigma^2 I$ (otherwise Mahalanobis' distance if the $\mathrm{Cov}(\underline{\varepsilon})$ is not diagonal) !



$\hat{\underline{Y}} = \Pi_{\underline{Y}|\mathcal{L}(\mathbb{Z})}$ fitted values

$\hat{\underline{\varepsilon}} = \Pi_{\underline{Y}|\mathcal{L}(\mathbb{Z})^\perp}$ residuals

$\hat{\underline{Y}} \perp \hat{\underline{\varepsilon}}$ and $\underline{Y} = \hat{\underline{Y}} + \hat{\underline{\varepsilon}}$

($\neq \underline{\varepsilon}$, $\hat{\underline{\varepsilon}}$ is just what is left out after the projection. We would like $\hat{\underline{\varepsilon}}$ to be a realization of $\underline{\varepsilon}$ but we'll see that it's not possible)

Obs.

1. If $\mathrm{rank}(\mathbb{Z}) = r+1 \leq n$ (full rank $\mathbb{Z}$) then:

$$\hat{\underline{Y}} = \mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\underline{Y} = H\underline{Y} = \text{linear transf. of } \underline{Y}$$

where H is the orthogonal projection on $\mathcal{L}(\mathbb{Z})$, and:

$$\hat{\underline{\beta}} = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\underline{Y} = \text{linear transform. of } \underline{Y}$$

$$\hat{\underline{\varepsilon}} = (I - H)\underline{Y}$$

Note that: the geometry always remains the same, it does not depend on the rank of $\mathbb{Z}$

2. If $\mathrm{rank}(\mathbb{Z}) = k < r+1 \leq n$:

$\dim(\mathcal{L}(\mathbb{Z})) = k < r+1$

(the geometry is still the same!) we have to project $\underline{Y}$ on $\mathcal{L}(\mathbb{Z})$ but $\mathcal{L}(\mathbb{Z})$ has a dimension smaller than $r+1$: $\dim(\mathcal{L}(\mathbb{Z})) = k$

$$\mathbb{Z}^T\mathbb{Z} = \sum_{i=1}^{r+1} \lambda_i \underline{e}_i \underline{e}_i^T \qquad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k = 0 = \lambda_{k+1} = \dots = \lambda_{r+1}$$

$$(\mathbb{Z}^T\mathbb{Z})^- = \sum_{i=1}^{k} \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i^T \longrightarrow$$ we generate a basis for $\mathcal{L}(\mathbb{Z})$ which has k elements and not $r+1$

$$\Longrightarrow \hat{\underline{Y}} = \mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^-\mathbb{Z}^T\underline{Y} \qquad (\exists!\ \hat{\underline{Y}} : \text{projection})$$

$$\hat{\underline{\beta}} = (\mathbb{Z}^T\mathbb{Z})^-\mathbb{Z}^T\underline{Y}$$

↳ (this is one possible $\hat{\underline{\beta}}$)

($\not\exists!\ \hat{\underline{\beta}}$: we have $r+1$ variables and k elements of the basis

$\Longrightarrow \exists \infty$ solutions : ∞ possible representations for $\hat{\underline{\beta}}$ in terms of the original variables)

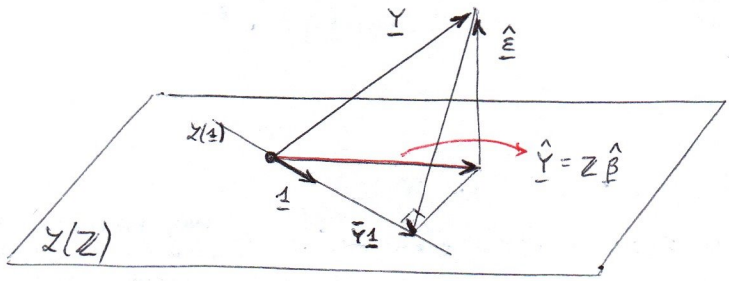
3. If $\text{rank}(Z) = r+1 = n$:

$\Longrightarrow \mathcal{L}(Z) = \mathbb{R}^n$

$\Longrightarrow \underline{Y} = \hat{\underline{Y}}$ and $\hat{\underline{\varepsilon}} = \underline{0}$

Standard way to judge how good is the fitting of the model:

COEFFICIENT OF DETERMINATION



(Assume $\text{rank}(Z) = r+1 \leq n$)

$$\|\underline{Y}\|^2 = \|\hat{\underline{Y}}\|^2 + \|\hat{\underline{\varepsilon}}\|^2$$

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} \varepsilon_i^2$$

$$\boxed{SS_{TOT} = SS_{Reg} + SS_{Res}}$$

Note that: $\underline{1} \in \mathcal{L}(Z)$ ($\underline{1}$ is the first column of Z) and:

$$\Pi_{\underline{Y}|\mathcal{L}(\underline{1})} = \bar{Y} \cdot \underline{1}$$

$$\Pi_{\hat{\underline{Y}}|\mathcal{L}(\underline{1})} = \bar{\hat{Y}} \cdot \underline{1}$$

In fact (proof.):

$$\Pi_{\hat{\underline{Y}}|\mathcal{L}(\underline{1})} = \frac{\underline{1}\cdot\underline{1}^T}{\underline{1}^T\cdot\underline{1}}\,\hat{\underline{Y}} = \frac{\underline{1}\cdot\underline{1}^T}{\underline{1}^T\cdot\underline{1}}\,H\underline{Y}$$

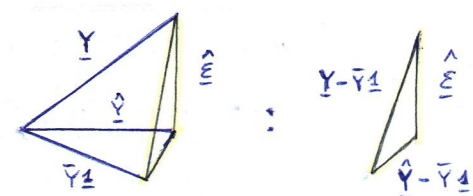
$$\underline{1}^T H = (H^T\underline{1})^T \overset{*}{=} (H\underline{1})^T \overset{*}{=} \underline{1}^T \quad \text{since } \underline{1} \in \mathcal{L}(Z)$$

* H is an orthogonal projector ($H^T = H$) on $\mathcal{L}(Z)$

$$\Longrightarrow \Pi_{\hat{\underline{Y}}|\mathcal{L}(\underline{1})} = \frac{\underline{1}\,\underline{1}^T}{\underline{1}^T\underline{1}}\,\underline{Y} = \bar{Y}\cdot\underline{1}$$

Hence:

$$\|\underline{Y} - \bar{Y}\underline{1}\|^2 = \|\hat{\underline{Y}} - \bar{Y}\cdot\underline{1}\|^2 + \|\hat{\underline{\varepsilon}}\|^2$$



$$\boxed{\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2}$$

$$\boxed{CSS_{TOT} = CSS_{Reg} + SS_{Res}}$$

CSS = centered sum of squares

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

DECOMPOSITION OF VARIANCE

the variability of $\underline{Y}$ is decomposed into 2 parts:

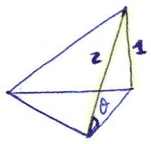
- variability accounted for by the regression
- variability that we cannot capture through the variables that we used for the model

$$\Rightarrow \quad R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \qquad (*)$$

(*) → proportion of total variability explained by the regression model

$$= 1 - \sin^2(\theta)$$

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



$$R^2 = \frac{\|1\|}{\|2\|}$$

$$R^2 = \cos^2(\theta)$$

$R^2 = 1 \Longrightarrow \vartheta = 0 \Longrightarrow \underline{\hat{\varepsilon}} = 0$ "perfect fit"

$R^2 = 0 \Longrightarrow \theta = \frac{\pi}{2} \Longrightarrow \underline{\hat{Y}} = \bar{Y} \cdot \underline{1}$ $\quad Y_i = \bar{Y} \quad i = 1, \dots, n$ =

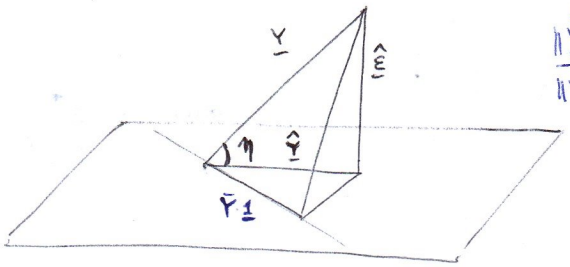
! the best explaination for $\underline{Y}$ is given by its mean, there is no contribution from the variables that we're using (basically because $\underline{Y} \perp \mathcal{Y}(\mathcal{Z})$)

Obs. If $\underline{1} \notin \mathcal{Y}(\mathcal{Z}) \Longrightarrow$ do not use R^2

(Regression through the origin : $\beta_0 = 0$)

(*) does not hold (we can compute it but it has no meaning, it may turn out negative or $+\infty$ (for instance))

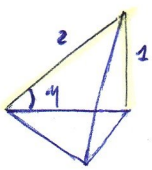
But: $\|\underline{Y}\|^2 = \|\underline{\hat{Y}}\|^2 + \|\underline{\hat{\varepsilon}}\|^2$



$$\frac{\|\underline{\hat{Y}}\|^2}{\|\underline{Y}\|^2} = \cos(\eta)^2 = \tilde{R}^2$$

! variability around 0 and not variability around the mean

: measure of fit (even if it is no longer the proportion of total variability explained by the regression model) : smaller the $\eta \to$ closer $\underline{\hat{Y}}$ to $\underline{Y}$



We can compute: $\tilde{R}^2 = 1 - \frac{\|\underline{\hat{\varepsilon}}\|^2}{\|\underline{Y}\|^2} = \frac{\|\underline{\hat{Y}}\|^2}{\|\underline{Y}\|^2}$:

Another **index**:

Adjusted R^2:

$$R^2_{adj} = 1 - \frac{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-(r+1)}}{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}}$$

! We want to take the ratio between two means spreading these lengths among the degrees of freedom that we have available

$\underline{\hat{\varepsilon}} \in \mathbb{R}^n$, $\underline{\hat{\varepsilon}} \in \mathcal{Y}^{\perp}(\mathcal{Z}) \Rightarrow \underline{\hat{\varepsilon}}$ has n components but they're not all free

$\dim(\mathcal{Y}^{\perp}(\mathcal{Z})) = n - (r+1)$

$\downarrow$

$\underline{\hat{\varepsilon}}$ cannot be the realization of our error term $\underline{\varepsilon}$ (the error has n free components)

The same for $\underline{Y} - \bar{Y}\underline{1}$ which has n components but only $n-1$ are free ($\underline{Y} - \bar{Y}\underline{1} \in \mathcal{Y}^{\perp}(\underline{1})$ and $\dim(\mathcal{Y}^{\perp}(\underline{1})) = n-1$).

(so we divide by the degrees of freedom of the vector of which we're considering the length. In this way the comparison will be fair.

Properties of $\underline{\hat{\beta}}$, $\underline{\hat{\varepsilon}}$ (Assume $\operatorname{rank}(\mathcal{Z}) = r+1 \leq n$) :

Theorem:

1. $\mathbb{E}[\underline{\hat{\beta}}] = \underline{\beta}$ $\quad \underline{\hat{\beta}}$ unbiased
2. $\operatorname{Cov}(\underline{\hat{\beta}}) = \sigma^2 (\mathcal{Z}^T \mathcal{Z})^{-1}$
3. $\mathbb{E}[\underline{\hat{\varepsilon}}] = 0$
4. $\operatorname{Cov}(\underline{\hat{\varepsilon}}) = \sigma^2 (I - H)$
5. $\mathbb{E}[\underline{\hat{\varepsilon}}^T \underline{\hat{\varepsilon}}] = \mathbb{E}[\sum_{i=1}^{n} \hat{\varepsilon}_i^2] = \sigma^2 (n - (r+1))$

proof. (just compute)

1. $\mathbb{E}[\underline{\hat{\beta}}] = \mathbb{E}[(\mathcal{Z}^T \mathcal{Z})^{-1} \mathcal{Z}^T \underline{Y}] \overset{\uparrow}{=} (\mathcal{Z}^T \mathcal{Z})^{-1} \mathcal{Z}^T \mathcal{Z} \underline{\beta} = \underline{\beta}$

$\mathbb{E}[\underline{Y}] = \mathbb{E}[\mathcal{Z}\underline{\beta} + \underline{\varepsilon}] = \mathcal{Z}\underline{\beta} + \mathbb{E}[\underline{\varepsilon}] = \mathcal{Z}\underline{\beta}$

2. $\text{Cov}(\hat{\underline{\beta}}) = \text{Cov}\left((\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\underline{Y}\right)$

$\downarrow$

$(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T(\sigma^2\mathbb{I})\mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1}$ ($\sigma^2\mathbb{I}$ = $\text{Cov}(\underline{Y})$)

$\downarrow$

$\sigma^2(\mathbb{Z}^T\mathbb{Z})^{-1}$

3. $\mathbb{E}[\hat{\underline{\varepsilon}}] = \mathbb{E}[(\mathbb{I}-H)\underline{Y}] = \mathbb{E}[\underline{Y}] - \mathbb{E}[H\underline{Y}]$

$\downarrow$

$\mathbb{Z}\underline{\beta} - \mathbb{E}[\mathbb{Z}\hat{\underline{\beta}}]$

$\downarrow$

$\mathbb{Z}\underline{\beta} - \mathbb{Z}\underline{\beta} = \underline{0}$

4. $\text{Cov}(\hat{\underline{\varepsilon}}) = \text{Cov}((\mathbb{I}-H)\underline{Y})$

$\downarrow$

$(\mathbb{I}-H)(\sigma^2\mathbb{I})(\mathbb{I}-H)^T$

$\downarrow$

$\sigma^2(\mathbb{I}-H)(\mathbb{I}-H)^T$

$\downarrow$

$\sigma^2(\mathbb{I}-H)$

$(\mathbb{I}-H)$ orthogonal projector $\Rightarrow$ idempotent and symmetric

5. $\mathbb{E}[\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}}] = \text{Tr}\,\mathbb{E}[\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}}] = \mathbb{E}[\text{Tr}(\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}})]$

$\downarrow$

$\mathbb{E}[\text{Tr}(\hat{\underline{\varepsilon}}\,\hat{\underline{\varepsilon}}^T)]$

$\downarrow$

$\mathbb{E}[\text{Tr}((\mathbb{I}-H)\underline{Y}\,\underline{Y}^T(\mathbb{I}-H)^T)]$

TRICK : $\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}}$ = number; $\hat{\underline{\varepsilon}}\,\hat{\underline{\varepsilon}}^T$ = matrix

$(\mathbb{I}-H)\underline{Y} = (\mathbb{I}-H)(\mathbb{Z}\underline{\beta}+\underline{\varepsilon}) = (\mathbb{I}-H)\underline{\varepsilon}$

$\mathbb{E}[\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}}] = \mathbb{E}[\text{Tr}((\mathbb{I}-H)\underline{\varepsilon}\,\underline{\varepsilon}^T(\mathbb{I}-H)^T)]$

$\downarrow$

$\text{Tr}((\mathbb{I}-H)\mathbb{E}[\underline{\varepsilon}\,\underline{\varepsilon}^T](\mathbb{I}-H)^T)$

$\downarrow$ ($\mathbb{E}[\underline{\varepsilon}\,\underline{\varepsilon}^T] = \text{Cov}(\underline{\varepsilon}) = \sigma^2\mathbb{I}$)

$\sigma^2\,\text{Tr}((\mathbb{I}-H)(\mathbb{I}-H)^T)$

$\downarrow$

$\sigma^2\,\text{Tr}(\mathbb{I}-H)$

$\downarrow$

$\sigma^2(n-\text{Tr}(H))$

$\text{Tr}(H) = \text{Tr}(\mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T)$

$\downarrow$

$\text{Tr}(\mathbb{Z}^T\mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1})$

$\downarrow$

$\text{Tr}(\mathbb{I}_{r+1}) = r+1$

$\mathbb{E}[\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}}] = \sigma^2(n-(r+1))$ ■

Obs.

1. $\text{Cov}(\hat{\underline{\beta}}) = \sigma^2(\mathbb{Z}^T\mathbb{Z})^{-1}$

If we have control of $\mathbb{Z}$ ($\equiv$ of $X_1, X_2, \ldots$) then design it in such a way that the variability of $\hat{\underline{\beta}}$ is small.

(DESIGN OF EXPERIMENTS)

2. $\mathbb{E}[\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}}] = \sigma^2 (n-(r+1))$

$\mathbb{E}\left[\frac{\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}}}{n-(r+1)}\right] = \sigma^2$ ← unbiased estimator for the variance

Def. $S^2 = \frac{\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}}}{n-(r+1)}$

Corollary: S^2 is unbiased for σ^2.

From now on: $\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 I)$

$\underline{Y} = Z\underline{\beta} + \underline{\varepsilon} \implies \underline{Y} \sim N_n(Z\underline{\beta}, \sigma^2 I)$

Theorem:

If $\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 I)$:

1. $\hat{\underline{\beta}}$ and $\hat{\sigma}^2 = \frac{\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}}}{n}$ ← biased (MLE estimators): since we have a model we can write the likelihood and then maximize the likelihood in terms of $\underline{\beta}$ and σ^2 (and get $\hat{\underline{\beta}}, \hat{\sigma}^2$)
2. $\hat{\underline{\beta}} \sim N_{r+1}(\underline{\beta}, \sigma^2 (Z^T Z)^{-1})$
3. $\hat{\underline{\varepsilon}} \sim N_n(\underline{0}, \sigma^2 (I-H))$
4. $\hat{\underline{\varepsilon}} \perp\!\!\!\perp \hat{\underline{\beta}}$
5. $\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \sim \sigma^2 \chi^2(n-(r+1))$

The novelty is in the distributions: $\underline{Y}$ is gaussian so $\hat{\underline{\beta}}$ and $\hat{\underline{\varepsilon}}$ and σ^2 are gaussian (since they're linear functions of $\underline{Y}$)

proof.

1. Just compute derivatives of log likelihood
2. 3. 4. Note that:

$$\begin{bmatrix} \hat{\underline{\beta}} \\ \hat{\underline{\varepsilon}} \end{bmatrix} = \begin{bmatrix} (Z^T Z)^{-1} Z^T \\ I - Z(Z^T Z)^{-1} Z^T \end{bmatrix} \underline{Y} \quad \text{and} \quad \underline{Y} \sim N_n(\underline{0}, \sigma^2 I)$$

Verify that:

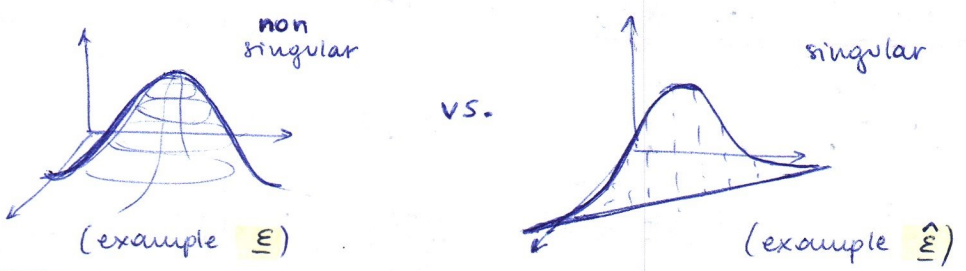
$$\mathrm{Cov}\left(\begin{bmatrix} \hat{\underline{\beta}} \\ \hat{\underline{\varepsilon}} \end{bmatrix}\right) = \sigma^2 \begin{bmatrix} (Z^T Z)^{-1} & \underline{0} \\ \underline{0} & I-H \end{bmatrix}$$

5. $\hat{\underline{\varepsilon}} \sim N_n(\underline{0}, \sigma^2 (I-H))$, $\det(I-H) = 0$ → SINGULAR GAUSSIAN DISTRIBUTION

$\mathrm{rank}(I-H) = n-(r+1)$. $\hat{\underline{\varepsilon}}$ lies in $\mathbb{R}^n$ but it has $n-(r+1)$ d.o.f.

Example of Singular Gaussian distribution: gaussian distribution defined only on a linear subspace of the entire space

$\mathbb{R}^2$:



We would like to take: $\hat{\underline{\varepsilon}}^T (I-H)^{-1} \hat{\underline{\varepsilon}}$ — for the Mahalanobis' distance between $\hat{\underline{\varepsilon}}$ and its mean (which is $\underline{0}$)

we cannot do that: $\det(I-H)=0$

$\Longrightarrow$ we consider: $\hat{\underline{\varepsilon}}^T (I-H)^{-} \hat{\underline{\varepsilon}} \sim \sigma^2 \chi^2(n-(r+1))$ (where: $\hat{\underline{\varepsilon}}^T (I-H)^{-1} \hat{\underline{\varepsilon}} \sim \sigma^2 \chi^2(n)$)

$\Longrightarrow$ (Mahalanobis' distance of Gaussian from its mean)

$\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}} \sim \sigma^2 \chi^2(n-(r+1))$

Now: we want the inference:

$$\left.\begin{array}{l} \underbrace{(\hat{\underline{\beta}}-\underline{\beta})^T (\sigma^2 (\mathbb{Z}^T\mathbb{Z})^{-1})^{-1} (\hat{\underline{\beta}}-\underline{\beta})}_{\text{Mahalanobis' distance of } \hat{\underline{\beta}} \text{ from its mean}} \sim \chi^2(r+1) \\ \hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}} \sim \sigma^2 \chi^2(n-(r+1)) \end{array}\right\} \perp\!\!\!\perp$$

$$\Longrightarrow \frac{\frac{1}{\sigma^2}(\hat{\underline{\beta}}-\underline{\beta})(\mathbb{Z}^T\mathbb{Z})(\hat{\underline{\beta}}-\underline{\beta})}{r+1} \Big/ \frac{\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}}}{\sigma^2(n-(r+1))} \sim F(r+1, n-(r+1))$$

since they're $\perp\!\!\!\perp$

$$\Longrightarrow \boxed{\frac{1}{S^2}(\hat{\underline{\beta}}-\underline{\beta})^T(\mathbb{Z}^T\mathbb{Z})(\hat{\underline{\beta}}-\underline{\beta}) \sim (r+1) F(r+1, n-(r+1))} \qquad S^2 = \frac{\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}}}{n-(r+1)}$$

pivotal quantity

Confidence regions for $\underline{\beta}$:

$$CR_{1-\alpha}(\underline{\beta}) = \left\{ \underline{\beta} \in \mathbb{R}^{r+1} : \frac{1}{S^2}(\hat{\underline{\beta}}-\underline{\beta})^T(\mathbb{Z}^T\mathbb{Z})(\hat{\underline{\beta}}-\underline{\beta}) \le (r+1) F_\alpha(r+1, n-(r+1)) \right\}$$

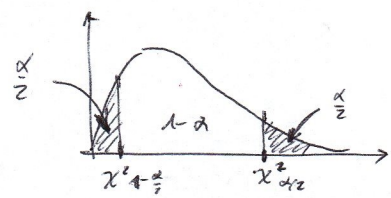
Confidence intervals for σ^2:

$\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}} \sim \sigma^2\chi^2(n-(r+1)) \Longrightarrow \frac{1}{\sigma^2}\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}} \sim \chi^2(n-(r+1))$

$$\boxed{\frac{(n-(r+1))S^2}{\sigma^2} \sim \chi^2(n-(r+1))}$$

pivotal quantity

$$CI_{1-\alpha}(\sigma^2) = \left\{ \sigma^2 \in \mathbb{R} : \chi^2_{1-\frac{\alpha}{2}}(n-(r+1)) \le \frac{(n-(r+1))S^2}{\sigma^2} \le \chi^2_{\frac{\alpha}{2}}(n-(r+1)) \right\}$$



$$= \left\{ \sigma^2 \in \mathbb{R} : \frac{(n-(r+1))S^2}{\chi^2_{\frac{\alpha}{2}}(n-(r+1))} \le \sigma^2 \le \frac{(n-(r+1))S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-(r+1))} \right\}$$

Simultaneous CI for $\underline{a}^T\underline{\beta}$, $\underline{a} \in \mathbb{R}^{r+1}$ (for linear combinations of β_i)

$$\max_{\underline{a}\in\mathbb{R}^{r+1}} \frac{1}{S^2} \frac{(\underline{a}^T(\hat{\underline{\beta}}-\underline{\beta}))^2}{\underline{a}^T(\mathbb{Z}^T\mathbb{Z})^{-1}\underline{a}} = \frac{1}{S^2}(\hat{\underline{\beta}}-\underline{\beta})^T(\mathbb{Z}^T\mathbb{Z})(\hat{\underline{\beta}}-\underline{\beta}) \sim (r+1)F(r+1, n-(r+1))$$

$$\Longrightarrow \text{Sim } CI_{1-\alpha}(\underline{a}^T\underline{\beta}) = \left[\underline{a}^T\hat{\underline{\beta}} \pm \sqrt{\underline{a}^T(\mathbb{Z}^T\mathbb{Z})^{-1}\underline{a}} \sqrt{S^2(r+1)F_{1-\alpha}(r+1, n-(r+1))} \right]$$

11/05

$\underline{Y} = \underline{\beta}\,\mathbb{Z} + \underline{\varepsilon}$ $\quad \mathbb{Z} \in \mathbb{R}^{n\times(r+1)}$ design matrix

$\underline{\beta} \in \mathbb{R}^{r+1}$

$\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 I) \Rightarrow \underline{Y}\ (\in \mathbb{R}^n) \sim N_n(\mathbb{Z}\beta, \sigma^2 I)$

(+ assume $\operatorname{rank}(\mathbb{Z}) = r+1 < n$)

$\hat{\underline{\beta}} \sim N_{r+1}(\underline{\beta}, \sigma^2(\mathbb{Z}^T\mathbb{Z})^{-1})$

$\hat{\underline{\varepsilon}} \sim N_n(\underline{0}, \sigma^2(I-H))$ $\qquad \hat{\underline{\varepsilon}} = (I-H)\underline{Y}$

$\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}} \sim \sigma^2 \chi^2(n-(r+1))$

$\Rightarrow \frac{1}{S^2}(\hat{\underline{\beta}} - \underline{\beta})^T(\mathbb{Z}^T\mathbb{Z})(\hat{\underline{\beta}} - \underline{\beta}) \sim (r+1)\,F(r+1, n-(r+1))$

$\Rightarrow CR_{1-\alpha}(\underline{\beta})$, $\quad \text{Sim } CI_{1-\alpha}(\underline{a}^T\underline{\beta}) \quad \forall \underline{a} \in \mathbb{R}^{r+1}$:

$$\text{Sim } CI_{1-\alpha}(\underline{a}^T\underline{\beta}) = \left[\underline{a}^T\hat{\underline{\beta}} \pm \sqrt{\underline{a}^T(\mathbb{Z}^T\mathbb{Z})^{-1}\underline{a}}\ \sqrt{S^2(r+1)\,F_\alpha(r+1, n-(r+1))}\ \right]$$

Simultaneous $CI_{1-\alpha}$ for any linear comb. of the β parameters

Special case:

$\text{Sim } CI_{1-\alpha}(\beta_i) \qquad i = 0, 1, \dots, r$ $\quad$ (it's like $\underline{a} = [0 \dots \underset{i}{1} \dots 0]^T$)

$$\text{Sim } CI_{1-\alpha}(\beta_i) = \left[\hat{\beta}_i \pm \sqrt{\text{diag}_i(\mathbb{Z}^T\mathbb{Z})^{-1}}\ \sqrt{S^2(r+1)\,F_\alpha(r+1, n-(r+1))}\ \right]$$

Obs.! All the software packages for regression compute one-at-the-time confidence intervals (not simultaneous w.r.t. $\forall \underline{a}$) for β_i :

$$\left[\hat{\beta}_i \pm \sqrt{\text{diag}_i(\mathbb{Z}^T\mathbb{Z})^{-1}}\ \sqrt{S^2}\ t_{\frac{\alpha}{2}}(n-(r+1))\ \right]$$ ← (the CI that we get after fixing $\underline{a}$, so $CI(\underline{a}^T\underline{\beta})$ for a fixed $\underline{a}$)

Obvious compromise: use Bonferroni CI's

$\alpha \longrightarrow$ ask for α/k (Bonferroni's correction)

$k = \#$ of β's for which we want CI's

Testing the β's (test for a set of the parameters)

$$\begin{cases} H_0: C\underline{\beta} = \underline{0} \\ H_1: C\underline{\beta} \neq \underline{0} \end{cases}$$

(more generally $\begin{cases} H_0: C\underline{\beta} = \underline{k}_0 \\ H_1: C\underline{\beta} \neq \underline{k}_0 \end{cases}$)

where $C \in \mathbb{R}^{p\times(r+1)}$ matrix given.

We're testing p different linear combinations

$$C\underline{\beta} = \begin{bmatrix} c_{11}\beta_0 + c_{12}\beta_1 + \dots + c_{1(r+1)}\beta_r \\ c_{21}\beta_0 + c_{22}\beta_1 + \dots + c_{2(r+1)}\beta_r \\ \vdots \\ c_{p1}\beta_0 + c_{p2}\beta_1 + \dots + c_{p(r+1)}\beta_r \end{bmatrix}$$

Meaning of β_i:
Every β is the derivative of our model wrt the variable:

$\beta_1 = \frac{\partial Y}{\partial z_1}$

= how much $\mathbb{E}[Y]$ is increasing by increasing in 1 unit in z_1

$C\hat{\beta}$ estimator for $C\beta$.

Under H_0:

$$\left.\begin{aligned} C\hat{\beta} &\sim N_p\left(\underline{0},\ \sigma^2 C\left(Z^TZ\right)^{-1}C\right) \\ \hat{\varepsilon}^T\hat{\varepsilon} &\sim \sigma^2\chi^2(n-(r+1)) \end{aligned}\right\} \perp\!\!\!\perp$$

(since the first depends on $\hat{\beta}$ and the second depends on $\hat{\varepsilon}$ and the two vectors are stochastically independent)

$$\Rightarrow \frac{\dfrac{(C\hat{\beta})^T\left(\sigma^2 C(Z^TZ)^{-1}C\right)^{-1}(C\hat{\beta})}{p}}{\dfrac{\hat{\varepsilon}^T\hat{\varepsilon}}{\sigma^2(n-(r+1))}} \sim F(p,\ n-(r+1)) \qquad \left(\sim \chi^2\right)$$

$$\Rightarrow \boxed{\frac{1}{s^2}(C\hat{\beta})^T\left(C(Z^TZ)^{-1}C^T\right)^{-1}(C\hat{\beta}) \sim p\,F(p,\ n-(r+1))}$$

we've basically taking the Mahalanobis' distance of $C\hat{\beta}$ from its mean (0) and H_0 assumes that if this distance is **too big** we have evidence **against** H_0

Reject H_0 at level α if:

$$\frac{1}{s^2}(C\hat{\beta})^T\left(C(Z^TZ)^{-1}C^T\right)^{-1}(C\hat{\beta}) > p\,F_\alpha(p,\ n-(r+1))$$

Special case:

$$\begin{cases} H_0: & \beta_r = \beta_{r-1} = \dots = \beta_{r-(p-1)} = 0 \\ H_1: & \exists\ \beta_j \neq 0 \qquad j = r-(p-1),\dots,r-1,r \end{cases}$$

: the last regressors of the model can be selected out? (Taken out of the model) We would get a simpler model

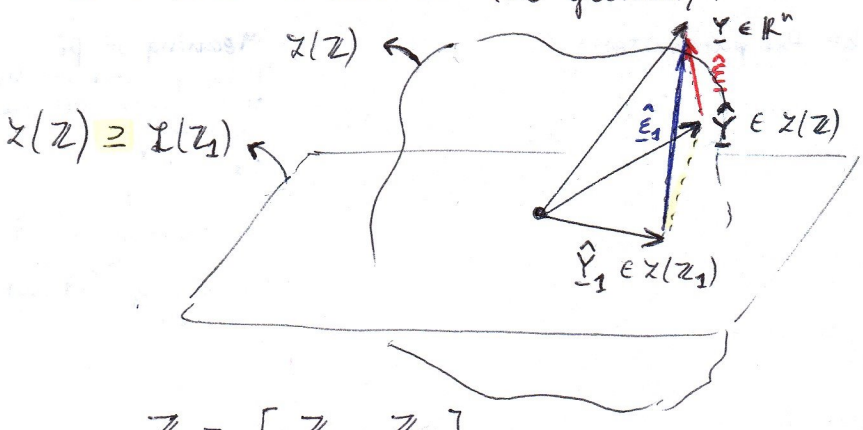
(i.e. Can we select out of the model the regressors $Z_{r-(p-1)},\dots,Z_r$?)

Why? We would gain degrees of freedom for estimating the variability (bias-variance trade off). We would gain degrees of freedom for the space where $\hat{\varepsilon}$ lives ($\Rightarrow$ less uncertainty about σ^2).

$\Longrightarrow$ take C such that:

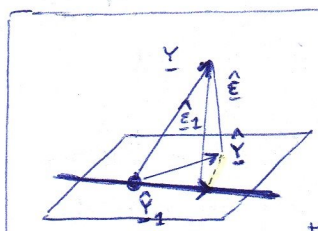
$$C = \begin{bmatrix} 0 & \dots & 0 & 1 & & \\ \vdots & & & & \ddots & \\ 0 & \dots & 0 & & & 1 \end{bmatrix} = \begin{bmatrix} 0 & I_p \end{bmatrix} \in \mathbb{R}^{p\times(r+1)}$$

Let's have a look at the geometry:



$$Z = \underbrace{\begin{bmatrix} Z_1 & Z_2 \end{bmatrix}}_{p}$$

When the two models are the same? When $\dots$ is small (= the residuals $\hat{\varepsilon}$ and $\hat{\varepsilon}_1$ are basically the same)



here we have:
$\underline{Y} \in \mathbb{R}^3$
$\hat{\underline{Y}} \in \mathbb{R}^2$
$\hat{\underline{Y}}_1 \in \mathbb{R}^1$
If $\dots$ is small then we can use $\hat{\underline{Y}}_1$

that's why $\mathcal{Y}(Z_1) \subseteq \mathcal{Y}(Z)$

Comparing:

$$\begin{cases} \text{model} & \underline{Y} = Z\beta + \underline{\varepsilon} \\ \text{model} & \underline{Y} = Z_1\beta_1 + \underline{\varepsilon}_1 \end{cases}$$

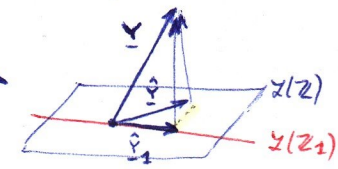
we're comparing 2 models for explaining the same $\underline{Y}$ (the two are nested: Z_1 comes from Z, obtained by setting to 0 some of the parameters)

Reject $H_0 : \beta_r = \beta_{r-1} = \dots = \beta_{r-(p-1)} = 0$ if:

$$SS_{res}(Z_1) - SS_{res}(Z) = \hat{\underline{\varepsilon}}_1^T \hat{\underline{\varepsilon}}_1 - \hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}} \quad \text{is big.}$$

In fact:

$$\frac{SS_{res}(Z_1) - SS_{res}(Z)}{S^2 \cdot p} \sim F(p, n-(r+1))$$



the two models are different ($\hat{\underline{Y}}$ and $\hat{\underline{Y}}_1$) if ---- is big.

$$S^2 = \frac{\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}}}{n-(r+1)} = \frac{SS_{res}(Z)}{n-(r+1)}$$

Very special case:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \dots = \beta_r = 0 \\ H_1 : \exists \beta_j \neq 0 \quad j = 1, \dots, r \end{cases} \longrightarrow$$

is it worth doing this regression or if we take the mean of Y it'll be enough? (β_0)

$$\rightarrow Z_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \rightarrow \hat{\underline{\varepsilon}}_1^T \hat{\underline{\varepsilon}}_1 = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$\frac{SS_{res}(Z_1) - SS_{res}(Z)}{S^2 r} = \frac{\dfrac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n \hat{\varepsilon}_i^2}{r}}{\dfrac{\sum_{i=1}^n \varepsilon_i^2}{n-(r+1)}} = \frac{\dfrac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n \hat{\varepsilon}_i^2}{r}}{\dfrac{\sum_{i=1}^n \varepsilon_i^2}{n-(r+1)}} \sim F(r, n-(r+1))$$

$p = r$ (we're taking r regressors)

(classical F test produced by any software for regression) → that's the test that comes out with the summary:

it checks if the variability explained by the model wrt the variability that is residual is big enough to guarantee that the model has some meaning or not

PREDICTION

Which, btw, is not the basic goal: the basic goal of regression is to understand the link between variables (covariates) and the dependent variables ($\underline{Y}$) (which we want to understand how to control w.r.t. variables)

$\underline{Y} = Z\underline{\beta} + \underline{\varepsilon}$ model for the training set

Assumptions: ....

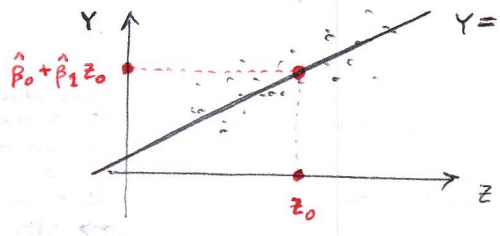
We take a new case (new statistical unit) for which we now the Z's but we want to use the model to predict Y or at least the mean of Y:

let $\underline{z}_0 = [1 \quad z_{01} \quad \dots \quad z_{0r}]^T$ is a given vector capturing the values of the regressors for which we want the prediction of Y.

new Y (not used in the training) $Y_0 = z_0^T \underline{\beta} + \varepsilon_0$ (new statistical unit)

$\varepsilon_0 \perp\!\!\!\perp \underline{\varepsilon}$

$\varepsilon_0 \sim N(0, \sigma^2)$



what is the uncertainty about this prediction?

Recall: $\mathbb{E}[Y_0 \mid \underline{z}_0] = \underline{z}_0^T \underline{\beta}$ → we're predicting the **MEAN** of **Y**, **not the Y** using the line

Natural estimator for $\underline{z}_0^T \underline{\beta}$ is : $\underline{z}_0^T \hat{\underline{\beta}}$ (Why? Gauss-Markov theorem)

Gauss-Markov theorem:

$\underline{z}_0^T \hat{\underline{\beta}}$ is the best estimator (minimum variance) among those:

- linear
- unbiased

$\underline{z}_0^T \hat{\underline{\beta}}$ is **BLUE** for $\underline{z}_0^T \underline{\beta}$.

$$\underline{z}_0^T \hat{\underline{\beta}} \sim N_1\left(\underline{z}_0^T \underline{\beta},\ \sigma^2 \underline{z}_0^T (\mathbb{Z}^T \mathbb{Z})^{-1} \underline{z}_0\right)$$

because: $\hat{\underline{\beta}} \sim N_{r+1}\left(\underline{\beta},\ \sigma^2 (\mathbb{Z}^T \mathbb{Z})^{-1}\right)$ (*)

$\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}} \sim \sigma^2 \chi^2(n-(r+1))$

$\Big\} \perp\!\!\!\perp \left(\underline{z}_0^T \hat{\underline{\beta}} \perp\!\!\!\perp \hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}}\right)$

(*) and $\underline{z}_0^T \hat{\underline{\beta}}$ is a linear transformation of $\hat{\underline{\beta}}$

$$\frac{\dfrac{\underline{z}_0^T \hat{\underline{\beta}} - \underline{z}_0 \underline{\beta}}{\sqrt{\sigma^2 \underline{z}_0^T (\mathbb{Z}^T \mathbb{Z})^{-1} \underline{z}_0}} \sim N(0,1)}{\sqrt{\dfrac{\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}}}{\sigma^2 (n-(r+1))}}} \sim t(n-(r+1))$$

(*) So we'll have confidence intervals but also a region that covers the linear model for all possible values of $\underline{z}_0$

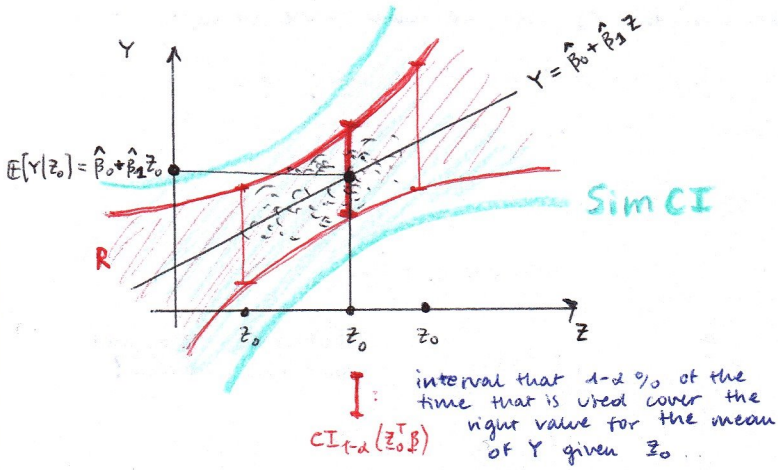
$$\frac{\underline{z}_0^T \hat{\underline{\beta}} - \underline{z}_0 \underline{\beta}}{S \sqrt{\underline{z}_0 (\mathbb{Z}^T \mathbb{Z})^{-1} \underline{z}_0}} \sim t(n-(r+1))$$

→ we can use this for building up confidence intervals for the **MEAN** of the **Y** corresponding to a new setting of our regressors (it's like a special case of CI for linear combinations of the β's)

$$CI_{1-\alpha}(\underline{z}_0^T \underline{\beta}) = CI_{1-\alpha}(\mathbb{E}[Y_0 \mid \underline{z}_0])$$

$$\downarrow$$

$$\left[\underbrace{\underline{z}_0^T \hat{\underline{\beta}}}_{\text{prediction}} \pm \underbrace{S \sqrt{\underline{z}_0^T (\mathbb{Z}^T \mathbb{Z})^{-1} \underline{z}_0}\ t_{\frac{\alpha}{2}}(n-(r+1))}_{\text{uncertainty}}\right] \qquad \alpha \in (0,1)$$



: interval that $1-\alpha$ % of the time that is used cover the right value for the mean of Y given z_0

- We generate an $CI_{1-\alpha}(\underline{z}_0^T \underline{\beta})$ for any possible z_0. Note that: the closer we are to the baricenter, the smaller are these intervals (I). That's because of "$\underline{z}_0^T (\mathbb{Z}^T \mathbb{Z})^{-1} \underline{z}_0$" which makes the uncertainty longer when we go further away from the data.

- Note that the band R is the region of the CI for our predictions when z_0 varies but it doesn't mean that $1-\alpha$ % of the time the regression line will fall in R because THE IC ARE NOT SIMULTANEOUS! (They're only one-at-the-time!) Each one of them is generated by an algorithm that $1-\alpha$ % of the time covers the right value for the prediction of the mean but we're taking an ∞ number of them here so we cannot expect that that band will cover with confidence $1-\alpha$ the true regression line.

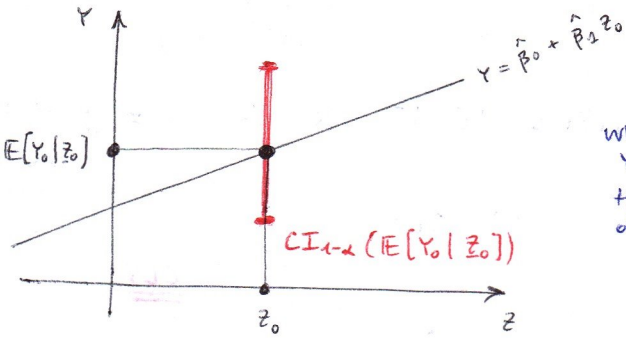
→ SIMULTANEOUS CI (*)

$$\text{Sim } CI_{1-\alpha}\left(z_0^T\beta\right) = \text{Sim } CI_{1-\alpha}\left(\mathbb{E}[Y_0 \mid z_0]\right)$$

$$\downarrow$$

$$\left[z_0^T\hat{\beta} \pm S\sqrt{z_0^T(Z^TZ)^{-1}z_0}\sqrt{(r+1)\,F_\alpha(r+1, n-(r+1))} \right]$$

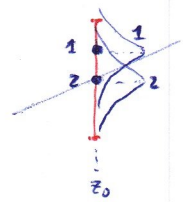
This is prediction for $\mathbb{E}[Y_0 \mid z_0]$, meaning:



$$Y_0 \sim N\left(z_0^T\beta, \sigma^2\right)$$

pro:
- we guarantee that, no matter how many predictions, even 1 billion of z_0, we're sure that the confidence level is still $1-\alpha$ %
- we know that within the region ( ) will be the TRUE linear model with confidence $1-\alpha$ % of the time

what about Y_0? Not its mean. What do we know about Y_0? $Y_0 \sim N(z_0^T\beta, \sigma^2)$ so:
the $CI_{1-\alpha}(\mathbb{E}[Y_0 \mid z_0])$ is predicting the mean of the distribution of Y_0, we're not predicting Y_0.



If the mean (the real mean) is in 1 at z_0 then the distribution of Y_0 will be the 1, otherwise if the mean is 2 the distribution of Y_0 will be the 2.

→ predicting the mean is not equivalent to predicting the Y (the true observation that we would make for a statistical unit).
what this $CI_{1-\alpha}$ is capturing is the uncertainty about the mean of the distribution generating that Y, not about Y.

Can we find an interval s.t. $\mathbb{P}\left(Y_0 \in I \mid z_0\right) = 1-\alpha$?
(:= **prediction interval**) : we know already (looking at both of pictures) that this interval will be longer than CI. (we're already uncertain about where the center of this distribution is, then Y is distributed around the center ⇒ extra variability)

$$\left.\begin{aligned} Y_0 &\sim N_1\left(z_0^T\beta, \sigma^2\right) \\ z_0^T\hat{\beta} &\sim N_1\left(z_0^T\beta, \sigma^2 z_0^T(Z^TZ)^{-1}z_0\right) \end{aligned}\right\} \perp\!\!\!\perp \text{ (since } \varepsilon_0 \perp\!\!\!\perp \varepsilon)$$

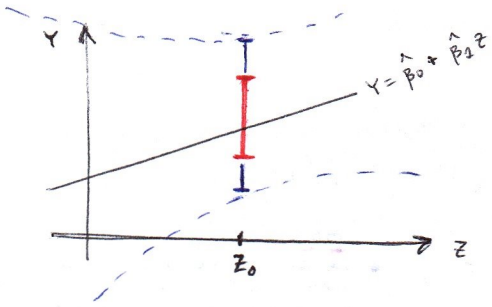
$$\Rightarrow \left.\begin{aligned} Y_0 - z_0^T\hat{\beta} &\sim N\left(0, \sigma^2\left(1 + z_0^T(Z^TZ)^{-1}z_0\right)\right) \\ \hat{\varepsilon}^T\hat{\varepsilon} &\sim \sigma^2\chi^2(n-(r+1)) \end{aligned}\right\} \perp\!\!\!\perp \text{ (since } \hat{\varepsilon} \perp\!\!\!\perp \hat{\beta})$$

$$\frac{\dfrac{Y_0 - z_0^T\hat{\beta}}{\sqrt{\sigma^2\left(1+z_0^T(Z^TZ)^{-1}z_0\right)}}}{\sqrt{\dfrac{\hat{\varepsilon}^T\hat{\varepsilon}}{\sigma^2(n-(r+1))}}} \sim t(n-(r+1)) \qquad \left(\frac{Y_0 - z_0^T\hat{\beta}}{\sqrt{\sigma^2\left(1+z_0^T(Z^TZ)^{-1}z_0\right)}} \sim N(0,1)\right)$$

$$\boxed{\frac{Y_0 - z_0^T\hat{\beta}}{S\sqrt{1+z_0^T(Z^TZ)^{-1}z_0}} \sim t(n-(r+1))}$$

$$PI_{1-\alpha} = \left[z_0^T\hat{\beta} \pm S\sqrt{1 + z_0^T(Z^TZ)^{-1}z_0}\; t_{\frac{\alpha}{2}}(n-(r+1)) \right]$$

prediction interval of **PROBABILITY** $1-\alpha$
($1-\alpha$ is the probability that Y_0 belongs to $PI_{1-\alpha}$).



$I: CI_{1-\alpha}(\mathbb{E}[Y_0|z_0])$

$I: PI_{1-\alpha}(Y_0)$ → it takes care of the fact that Y has a variability around its mean ⇒ the length of the prediction interval is bigger than the confidence interval (one is for Y_0, the other for $\mathbb{E}[Y_0|z_0]$)

Are these simultaneous prediction intervals? NO. If we plot all of them (----) we will not find a prediction region for Y. If we want to find a prediction region for Y: **SIMULTANEOUS PREDICTION INTERVALS**

Exercise: find the simultaneous prediction intervals (not easy)

A short excursion on **GLS** (GENERALIZED LEAST SQUARES) = what happen if we have correlated error

$$\underline{Y} = \mathbb{Z}\underline{\beta} + \underline{\varepsilon} \qquad \left.\begin{array}{l}\mathbb{E}[\underline{\varepsilon}] = \underline{0} \\ \mathrm{Cov}(\underline{\varepsilon}) = \sigma^2 I\end{array}\right\} \Rightarrow \text{OLS}$$

OLS

$$\hat{\underline{\beta}} = \arg\min_{\underline{\beta}} \|\underline{Y} - \mathbb{Z}\underline{\beta}\|^2 = \arg\min_{\underline{\beta}} (\underline{Y} - \mathbb{Z}\underline{\beta})^T(\underline{Y} - \mathbb{Z}\underline{\beta})$$

GLS: Assume $W \in \mathbb{R}^{n\times n}$ positive definite (that sort of control the metric that we want to use in the space of β's instead of taking the euclidean metric)

$$\hat{\underline{\beta}} = \arg\min_{\underline{\beta}} (\underline{Y} - \mathbb{Z}\underline{\beta})^T W^{-1} (\underline{Y} - \mathbb{Z}\underline{\beta})$$

$$(\underline{Y} - \mathbb{Z}\underline{\beta})^T W^{-1}(\underline{Y} - \mathbb{Z}\underline{\beta}) = (W^{-1/2}\underline{Y} - W^{-1/2}\mathbb{Z}\underline{\beta})^T (W^{-1/2}\underline{Y} - W^{-1/2}\mathbb{Z}\underline{\beta})$$
$$= \|W^{-1/2}\underline{Y} - W^{-1/2}\mathbb{Z}\underline{\beta}\|^2 \sim \|\tilde{\underline{Y}} - \tilde{\mathbb{Z}}\underline{\beta}\|^2$$

(**GLS**) it's like transforming our data and then apply OLS on transformed data

$$\Rightarrow \boxed{\hat{\underline{\beta}} = (\mathbb{Z}^T W^{-1}\mathbb{Z})^{-1}\mathbb{Z}^T W^{-1}\underline{Y}}$$

In fact: notice that:

$$W^{-1/2}\underline{Y} = W^{-1/2}\mathbb{Z}\underline{\beta} + W^{-1/2}\underline{\varepsilon}$$

$$\mathrm{Cov}(W^{-1/2}\underline{\varepsilon}) = \sigma^2 W^{-1}$$

Example: Assume that $\underline{\varepsilon}$ is such that $\mathrm{Cov}(\underline{\varepsilon}) = \sigma^2\Sigma'$ (σ^2 unknown and Σ' known).

Take $W = \Sigma'$:

$$\hat{\underline{\beta}} = \arg\min_{\underline{\beta}} (\underline{Y} - \mathbb{Z}\underline{\beta})^T \Sigma'^{-1}(\underline{Y} - \mathbb{Z}\underline{\beta}) = \text{minimize wrt the Mahalanobis' distance}$$
$$= (\mathbb{Z}^T\Sigma'^{-1}\mathbb{Z})^{-1}\mathbb{Z}^T\Sigma'^{-1}\underline{Y}$$

Note: $\mathrm{Cov}(\Sigma'^{-1/2}\underline{\varepsilon}) = \sigma^2\Sigma'^{-1/2}\Sigma'\Sigma'^{-1/2} = \sigma^2 I$ → we have transformed the variables s.t. the error term is again uncorrelated with the same variance along the components

Special cases:

$$\mathrm{Cov}(\underline{\varepsilon}) = \sigma^2\,\mathrm{diag}(w_1, .., w_n), \text{ i.e. } \Sigma = \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{bmatrix}\sigma^2$$

⇒ WEIGHTED LEAST SQUARES (special case of GLS)

we believe that the statistical units do not share the same variance (they have different variances). σ^2 has a different weight along units.

Example: • Suppose Y_i $i = 1, \dots, n$ is a **mean** of n_i observations.

$$Var(Y_i) = \frac{\sigma^2}{n_i}$$

$$\Longrightarrow W = diag\left(\frac{1}{n_1}, \dots, \frac{1}{n_n}\right)$$

For instance: we observe the average income but the unit is a municipality: average income for Milano, for Torino, ... • We know that the number of people in those units is different so the averages have different variabilities.

• Suppose Y_i $i = 1, \dots, n$ is the **sum** of n_i observations

$$Var(Y_i) = n_i \sigma^2$$

$$\Longrightarrow W = diag\left(n_1, \dots, n_n\right)$$

For instance: in every municipality we observe a number of people having a certain disease. It's not the average for person, is the number of people having the disease (so it's a sum of Bernoulli's variables, each one may have the same variability but we're observing a different number of people (in Milano, lecco, Torino, ...))

(# Covid in Italy cannot be directly compared with # Covid in USA (different sizes))

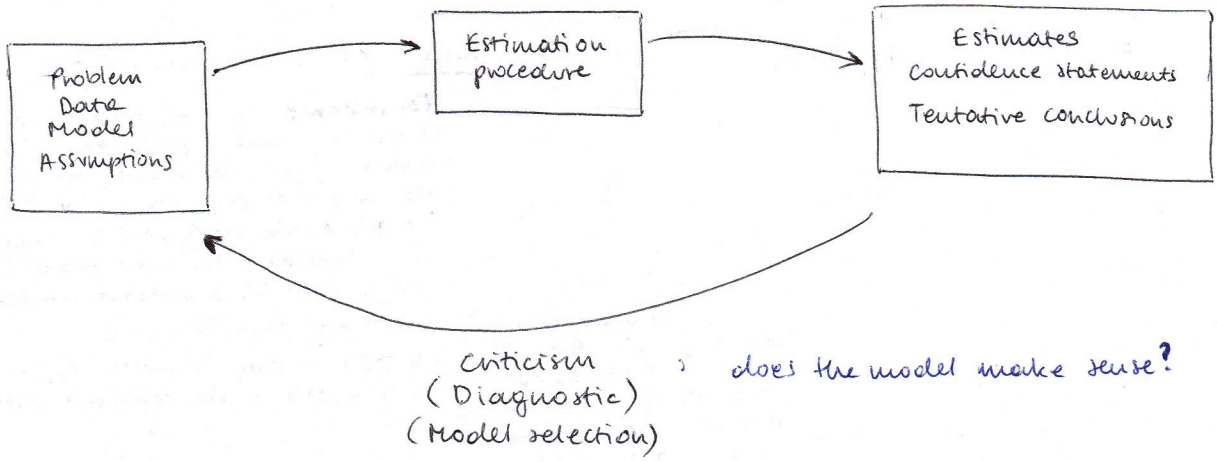
it's not just a matter of rescaling wrt. the number of people, but also taking into account that the uncertainty that we have about these estimates is pretty different (because the variability is different)

DIAGNOSTIC : what we do after the analysis to check how good is the model

12/05

(R^2 is not enough)

Box :



Criticism : does the model make sense?
(Diagnostic)
(Model selection)

Diagnostic for Linear Models

- Residual analysis : outliers, non-constant variance (heteroschedasticity), normality, autocorrelation, ...
- Influencial cases or statistical units
- Collinearity among the regressors

Intro: $\underline{\varepsilon}$ vs. $\underline{\hat{\varepsilon}}$

$$\underline{Y} = \mathbb{Z}\underline{\beta} + \underline{\varepsilon} \qquad \underline{\varepsilon} \in \mathbb{R}^n : \begin{cases} \mathbb{E}[\underline{\varepsilon}] = \underline{0} \\ \mathrm{Cov}(\underline{\varepsilon}) = \sigma^2 I \end{cases}$$

$$\left(\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 I)\right) \text{ not necessary}$$

What do we know about $\underline{\hat{\varepsilon}}$? (what changes between $\underline{\varepsilon}$ and $\underline{\hat{\varepsilon}}$?)

$$\underline{\hat{\varepsilon}} = (I-H)\underline{Y} = \Pi_{\mathcal{Y}|\mathcal{Y}^\perp(\mathbb{Z})} \cdot \qquad \underline{\hat{\varepsilon}} \in \mathcal{Y}^\perp(\mathbb{Z}) \subseteq \mathbb{R}^n$$

($\underline{\hat{\varepsilon}}$ has n components but only $\dim(\mathcal{Y}^\perp(\mathbb{Z}))$ are independent)

$$\begin{cases} \mathbb{E}[\underline{\hat{\varepsilon}}] = \underline{0} \\ \mathrm{Cov}(\underline{\hat{\varepsilon}}) = \sigma^2(I-H) \end{cases} \longrightarrow$$ the residuals are correlated even if the components of the error ($\underline{\varepsilon}$) are not correlated

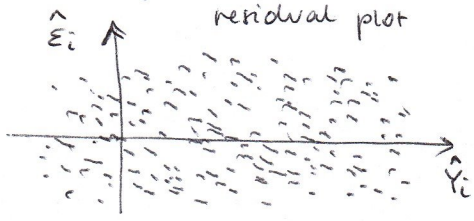
$$\left(\underline{\hat{\varepsilon}} \sim N_n(\underline{0}, \sigma^2(I-H))\right) \text{ not necessary}$$

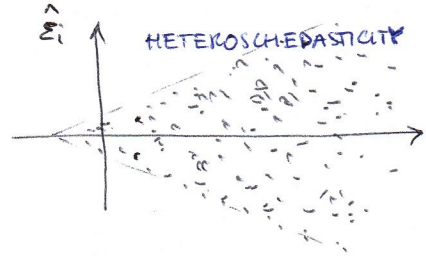
$\underline{\hat{\varepsilon}}$ is gaussian on a subspace of $\mathbb{R}^n$ (singular)

$\underline{\hat{\varepsilon}}$ cannot be the realizations of $\underline{\varepsilon}$ but $\underline{\hat{\varepsilon}}$ is the only thing to check the assumptions on $\underline{\varepsilon}$ ($\equiv$ model)

- Residual analysis

Searching for large $\hat{\varepsilon}_i$ (= the fitting of the model is very poor)





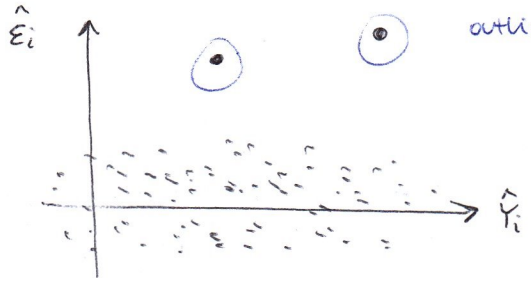
In this case there is a variability which is non constant w.r.t. the fitted values (in this case large fitted values implies large variance for $\hat{\varepsilon}_i$)

1. we want to see they're not big
2. we want to check that there's no shape (if there is information (shape) we should capture it in the model)

How can we fix it? (the second plot) ($\Rightarrow$ the heteroscedasticity)

- weighted linear regression
- transform the Y's
  (variance Stabilizing transformations)

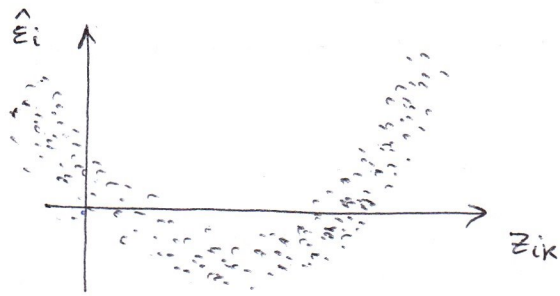
Another possibility for residuals:



outliers?

They shouldn't be easily eliminated. We have to understand what they represent (why they've outliers). It might be that these outliers are a signal about existence of an other population that we didn't sampled (and sometimes that's exactly what we've looking for)

We can plot $\hat{\varepsilon}_i$ vs. one regressor (instead of fitted values)



$k = 1, \dots, r$

Again: we don't want to see patterns (the error should not be related to regressors)

$\longrightarrow$ in this case the errors are not independent with the k-regressor.

quadratic in z_{ik}? $\Rightarrow$ introduce z_{ik}^2

cubic transformation?
sinusoidal transformation? (sin/cos)

Note that:

$$Cov(\hat{\underline{\varepsilon}}) = \sigma^2 (I - H) \quad \Longrightarrow \quad Var(\hat{\varepsilon}_i) = \sigma^2 (1 - h_{ii}) \qquad i = 1, \dots, n$$

$$(h_{ii} = diag_i (H))$$

$\Rightarrow$ $\hat{\varepsilon}_i$ have different variances ($i = 1, \dots$)

Repeat the residual analysis with:

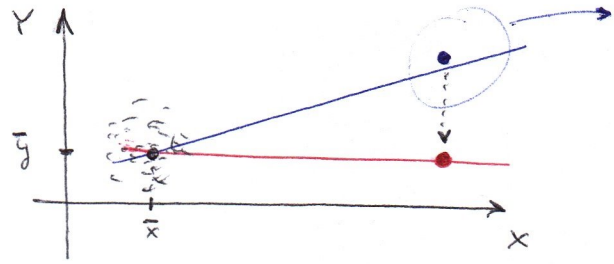
$$\frac{\hat{\varepsilon}_i}{\sqrt{S^2 (1 - h_{ii})}}$$

studentized residuals
(in this way the residuals will have the same variance)

- Check Gaussianity: QQ-plot of residuals or studentized residuals
- Test for autocorrelation using (for instance) Durbin-Watson test ($Cov(\hat{\underline{\varepsilon}}) = \sigma^2 (I - H)$) but how large is this

(Of course the residuals are correlated autocorrelation?)

- **Influencial cases:** are there cases that drives all the analysis no matter how many data we ha ?



LEVERAGING EFFECT
if we move this point the linear regression changes drastically ⟹ what we read from the data (the model) is entirely determined basically by just **1** point

Note that: $Var(\hat{\varepsilon}_i) = \sigma^2(1-h_{ii}) \quad i = 1, \dots, n$

h_{ii} is called leveraging

$h_{ii} = diag_i\left(\mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\right)$: it doesn't even depend on Y, before performing the experiment we can chose $\mathbb{Z}$ s.t. we have small residuals

$0 \le h_{ii} \le 1$ (prove it, use: $H = H^T$ and $H \cdot H = H$)

for a specific case that is driving the entire model

$h_{ii} \uparrow 1 \implies Var(\hat{\varepsilon}_i) \to 0 \implies \hat{\varepsilon}_i \to 0$

Another way to identify influencial cases is to work out what happens if a specific case (a specific unit) is taken out of the dataset:

Holding out unit i to check if it is influencial

$\mathbb{X}$ data frame $\longrightarrow$ $\mathbb{Z} \in \mathbb{R}^{n \times (r+1)}$

$\mathbb{X}_{-i}$ data set obtained by taking out the case i $\longrightarrow$ $\mathbb{Z}_{-i} \in \mathbb{R}^{(n-1)\times(r+1)}$

$\underline{Y} = \mathbb{Z}\underline{\beta} + \underline{\varepsilon} \longrightarrow \hat{\underline{\beta}} \in \mathbb{R}^{r+1}$

$\underline{Y}_{-i} = \mathbb{Z}_{-i}\underline{\beta} + \underline{\varepsilon}_{-i} \longrightarrow \hat{\underline{\beta}}^{(i)} \in \mathbb{R}^{r+1}$

If $\hat{\underline{\beta}}^{(i)}$ and $\hat{\underline{\beta}}$ are very different ⟹ case i is influencial

Compute:

$$\frac{(\hat{\underline{\beta}}^{(i)} - \hat{\underline{\beta}})^T(\mathbb{Z}^T\mathbb{Z})(\hat{\underline{\beta}}^{(i)} - \hat{\underline{\beta}})}{S^2(r+1)} := D_i$$

Cook's distance
how far $\hat{\underline{\beta}}^{(i)}$ is from $\hat{\underline{\beta}}$ in terms of the Mahalanobis distance

We conclude that case i is influencial if D_i is large: ?

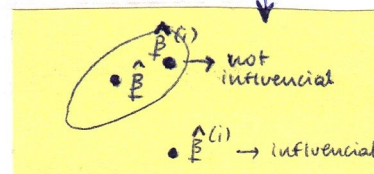
"large" ?:

- compare with the quantiles of $F(r+1, n-(r+1))$:
  if we're very far in terms of this quantile means that we're out of any reasonable confident region around $\hat{\underline{\beta}}$
- Rule of thumb:
  check cases for which $D_i > 1$: the median for this distribution (for reasonable values of r and n) is always centered on 1 (or close to 1). By checking that $D_i > 1$ we're saying that the Cook's distance is larger than the median that we've expecting

Note:

$$D_i = \underbrace{\left(\frac{\hat{\varepsilon}_i}{S\sqrt{1-h_{ii}}}\right)^2}_{\text{studentized residuals}} \underbrace{\frac{h_{ii}}{1-h_{ii}}}_{\text{monotone} \uparrow \text{ function of } h_{ii}} \cdot \frac{1}{r+1}$$

: takes care of both outliers and leveraging cases



- **Colinearity**

One or more regressors can be expressed as a linear combination of the others, i.e. "close" to "columns of $\mathbb{Z}$ are not linearly independent."

it's not a problem for the geometrical point of view (we can always project Y on $\mathcal{L}(\mathbb{Z})$). the problem is the formula: $\hat{\underline{\beta}} = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}\underline{Y}$

$(\mathbb{Z}^T\mathbb{Z})$ is "close" to be singular

$\Longrightarrow \text{Cov}(\hat{\underline{\beta}}) = \sigma^2(\mathbb{Z}^T\mathbb{Z})^{-1}$: it would be a "crazy" matrix (something that jumps from $-\infty$ to $+\infty$) $\Rightarrow$ the variability of $\hat{\underline{\beta}}$ will be very very large $\Rightarrow$ the uncertainty about estimating the parameters will be very large

Recall that: $\hat{\underline{Y}} = \Pi_{\underline{Y}|\mathcal{L}(\mathbb{Z})}$; we need an orthogonal basis for $\mathcal{L}(\mathbb{Z})$, then we project.

Use Grahm-Schmidt to build up an orthogonal basis for $\mathcal{L}(\mathbb{Z})$.

$$\mathbb{Z} = \begin{bmatrix} \underline{1} & \underline{z}_1 & \cdots & \underline{z}_r \\ 1 & z_{11} & \cdots & z_{1r} \\ \vdots & & & \\ 1 & z_{n1} & \cdots & z_{nr} \end{bmatrix} = \begin{bmatrix} \underline{1} & \underline{z}_1 & \cdots & \underline{z}_r \end{bmatrix}$$

G-S:

$$\underline{q}_1 = \underline{1}$$

$$\underline{q}_2 = \underline{z}_1 - \Pi_{\underline{z}_1|\underline{1}}$$

$$\underline{q}_3 = \underline{z}_2 - \Pi_{\underline{z}_2|\underline{q}_1,\underline{q}_2}$$

$$\vdots$$

$$\underline{q}_r = \underline{z}_{r-1} - \Pi_{\underline{z}_r|\underline{z}_1,\dots,\underline{z}_{r-1},\underline{1}}$$

residuals when we regress $\underline{z}_r$ on $(\underline{1}, \underline{z}_1, \dots, \underline{z}_{r-1})$ = what is left out after we project $\underline{z}_r$ on the linear space generated by $\underline{1}, \dots, \underline{z}_{r-1}$

$$\Longrightarrow \hat{\beta}_r = \frac{\underline{q}_r^T \underline{Y}}{\underline{q}_r^T \underline{q}_r}$$

(how do we get $\hat{\beta}_r$? By projecting Y on $\underline{q}_r$)

$$\text{Var}(\hat{\beta}_r) = \frac{1}{(\underline{q}_r^T \underline{q}_r)^2}\,\sigma^2\,\underline{q}_r^T\underline{q}_r = \frac{\sigma^2}{\underline{q}_r^T\underline{q}_r}$$

Decomposition of variance formula applied to the regression of $\underline{z}_r$ on $(\underline{1}, \underline{z}_1, \dots, \underline{z}_{r-1})$:

$$\sum_{i=1}^{n}(z_{ir} - \bar{z}_r)^2 = \sum_{i=1}^{n}(\hat{z}_{ir} - \bar{z}_r)^2 + \underline{q}_r^T\underline{q}_r$$

$\underline{q}_r^T\underline{q}_r$: sum of the residuals^2

$$\underline{q}_r^T\underline{q}_r = \sum_{i=1}^{n}(z_{ir} - \bar{z}_r)(1 - R_r^2)$$

R_r^2 for the regression of $\underline{z}_r$ on $(\underline{1}, \underline{z}_1, \dots, \underline{z}_{r-1})$

$$\text{Var}(\hat{\beta}_r) = \frac{\sigma^2}{\sum_{i=1}^{n}(z_{ir} - \bar{z}_r)^2} \cdot \frac{1}{1 - R_r^2}$$

$$\downarrow$$

$$\doteq \frac{\sigma^2}{(n-1)\,S_r^2} \cdot \frac{1}{1 - R_r^2}$$

→ this works not only for the last regressor (r^{th}), it's true for any regressor (since we can change the order for the regressors)

True for $j = 1, \dots, r$

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n}(z_{ij} - \bar{z}_j)^2} \cdot \frac{1}{1 - R_j^2}$$

R_j^2: coefficient of determination that we get when we regress variable j against the other regressors

$\text{Var}(\hat{\beta}_j) \downarrow$ if $\sum_{i=1}^{n}(z_{ij} - \bar{z}_j)^2 \uparrow$ → this gives an indication of how we should design our matrix if we're in control of it

$\text{Var}(\hat{\beta}_j) \uparrow$ if $R_j^2 \uparrow$ → $R_j^2 \uparrow 1$ means that the variable z_j could be expressed as a linear combination of the other regressors

(that is **collinearity**)

$$\boxed{VIF_j = \frac{1}{1 - R_j^2}}$$ **Variance Inflation Factor** $\left(VIF(\hat{\beta}_j)\right)$

Red Alarm: $VIF(\hat{\beta}_j) > 5$ or > 10

Model Selection – Variable Selection

maybe we want to extract some variable out of the model (we keep the model and take out variables, we're not changing completely the model)

Training data: $\underline{Z} = [\underline{1}\ \underline{z}_1 \cdots \underline{z}_r]$ $\quad \underline{z}_j \in \mathbb{R}^n$

We can build 2^r models. (for every $\underline{z}_j$ we have two options: **in** or **out** the model)

1st line of attack: **check them all** (fails with big r)

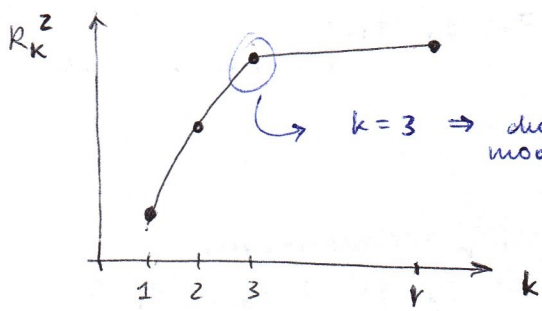
For $k = \dots, r$ fit the $\binom{r}{k}$ model with k regressors and compute R^2.

Choose the model with **k** regressors with highest R^2: R_k^2

Then, next k.

→ $R_1^2 \le R_2^2 \le \cdots \le R_r^2$

monotonic sequence. Why?
Because the best model with 1 regressor cannot beat the best model with 2 regressors (because in the model with 2 regr. we can fit one regressor to be 0 and then we go down to 1)



or plot: R^2_{adj}, AIC, BIC, ...

2nd line of attack: iterative procedures
(forward and backward)

Forward:

Start with the best model with 1 regressor.

Until convergence:

add 1 variable which increases the fit of the model.

Use **F** test to decide when to stop:

consider two models: k, k' : $k' > k$

$$\left[\frac{SS_{res}(Z_k) - SS_{res}(Z_{k'})}{\frac{SS_{res}(Z_k)}{n-(k'+1)}} \cdot \frac{1}{k'-k} \right]$$

stop if the p-value is **large**. (= it doesn't pay to move to a larger model)

(The **backward** selection goes the way around:
we start from the complete model and then we take out one variable, we stop taking out variables as soon as the p-values becomes to small (= the two models are different))

14/05

Linear models:

COLLINEARITY & VARIABLE SELECTION

Let's make an observation about the fitted model that we get through OLS.

$\underline{Y} = \mathbb{Z}\underline{\beta} + \underline{\varepsilon}$ model for the observed data

OLS: $\hat{\underline{\beta}} = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\underline{Y}$ (assuming rank($\mathbb{Z}$) = full)

Fitted model: $Y_0 = \underline{z}_0^T\hat{\underline{\beta}}$ = the model that we use for $\mathbb{E}[Y_0|\underline{z}_0]$
(for the mean of Y)

The fitted model goes through the **baricenter** of the observed data.

Let $\underline{\bar{z}}_0 = \dfrac{\mathbb{Z}^T\underline{1}}{\underline{1}^T\underline{1}} = [\,1\ \bar{z}_1\ \dots\ \bar{z}_r\,]^T$ = vector of the means for the regressors (vector of the sample means for the columns of $\mathbb{Z}$)

$$Y_0 = \underline{z}_0^T\hat{\underline{\beta}}$$
$$= \frac{\underline{1}^T\mathbb{Z}}{n}(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\underline{Y}$$
$$= \frac{1}{n}\underline{1}^T H\underline{Y}$$
$$= \frac{1}{n}(H\underline{1})^T\underline{Y}$$
$$= \frac{\underline{1}^T\underline{Y}}{n} = \bar{Y}$$

$\underline{1} \in \mathcal{V}(\mathbb{Z}) \Rightarrow H\underline{1} = \underline{1}$

So corresponding (for the fitted model) to the mean values of the regressors we get as output always the mean value for the dependent variable ($\bar{Y}$).

Hence: $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1\bar{z}_1 + \dots + \hat{\beta}_r\bar{z}_r$

$\Longrightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{z}_1 + \dots + \hat{\beta}_r\bar{z}_r$

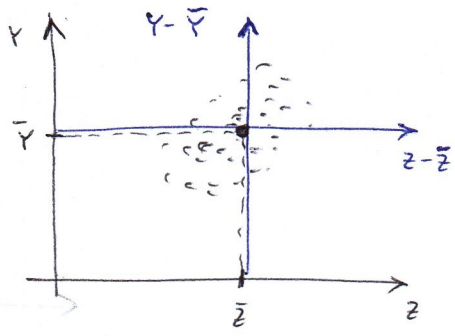
Different representation for the fitted model: (we substitute $\hat{\beta}_0$)

$$Y_0 - \bar{Y} = \hat{\beta}_1(z_1 - \bar{z}_1) + \dots + \hat{\beta}_r(z_r - \bar{z}_r)$$

$\Rightarrow$ the only question is to find $\hat{\beta}_1, \dots, \hat{\beta}_r$: $\hat{\beta}_0$ is automatically obtained knowing that the model will go through the origin (OLS)

$\Rightarrow$ we can first center the variables in the baricenter and then fit the regression model:

we know that the model will pass through $(\bar{z}, \bar{Y})$ so we move directly the system there and we fit the model w.r.t. the new reference system



Centering:

$$\underline{Y} \longrightarrow \begin{bmatrix} Y_1 - \bar{Y} \\ \vdots \\ Y_n - \bar{Y} \end{bmatrix} = \underline{Y}^*$$

$$\mathbb{Z} \longrightarrow \begin{bmatrix} z_{11}-\bar{z}_1 & z_{12}-\bar{z}_2 & \cdots & z_{1r}-\bar{z}_r \\ z_{21}-\bar{z}_1 & \cdots & & z_{2r}-\bar{z}_r \\ \vdots & & & \\ z_{n1}-\bar{z}_1 & \cdots & & z_{nr}-\bar{z}_r \end{bmatrix} = \mathbb{Z}^* \quad \in \mathbb{R}^{n \times r}$$

$\Longrightarrow$ OLS becomes:

$$\begin{cases} \underset{\underline{\beta} \in \mathbb{R}^r}{\operatorname{argmin}} \; \| \underline{Y}^* - \mathbb{Z}^* \underline{\beta} \|^2 = \hat{\underline{\beta}}^* \\ \begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1^* \bar{z}_1 - \cdots - \hat{\beta}_r^* \bar{z}_r \\ \hat{\beta}_i = \hat{\beta}_i^* \end{cases} \quad (*) \end{cases}$$

← Instead of solving the original problem we solve this and then we fix the β's with (*) in order to get the problem expressed in the original system

From now on we assume $\underline{Y}$ and $\mathbb{Z}$ to be centered (after getting estimates of $\hat{\underline{\beta}}$, fix $\hat{\beta}_0$)

→ so to go back to the original system (something that is almost always asked)

PCA REGRESSION

: the problem of collinearity is due to the fact that the regressors are not orthogonal, they can be correlated. How can we fix it? We can transform the regressors so that we have new regressors that are orthogonal. One possibility is the principal component analysis of the independent variables.

$$\mathbb{Z} = [\underset{z_1}{\underline{z}_1}, \ldots, \underset{z_r}{\underline{z}_r}] \qquad \underline{z}_i \in \mathbb{R}^n$$

$\Longrightarrow$ PCA of $\mathbb{Z}$ (just the dependent variables)

$\Longrightarrow$ $PC_1, \ldots, PC_r$

Reduce dimensionality: $PC_1, \ldots, PC_k, \quad k \leq r$ (the last ones probably do not account much of the variability of the data)

$$\mathbb{Z}^* = [\underline{PC}_1 \; \ldots \; \underline{PC}_k]$$

Fit: $\underline{Y} = \mathbb{Z}^* \underline{\beta} + \underline{\varepsilon}$ $\qquad \underline{\beta} \in \mathbb{R}^k$ → we reduced dimensionality

Fitted model: $Y_0 = \underline{z}_0^T \hat{\underline{\beta}}$ $\qquad \underline{z}_0 = [PC_1, \ldots, PC_k]^T$

$$= \hat{\beta}_1 PC_1 + \ldots + \hat{\beta}_k PC_k = \hat{\beta}_1 z_1 + \ldots + \hat{\beta}_k z_k$$

with $z_j \in \mathbb{Z}^*$, to $z_j = PC_j$

$$PC_1 = e_{11} z_1 + e_{21} z_2 + \ldots + e_{r1} z_r$$
$$PC_2 = e_{12} z_1 + e_{22} z_2 + \ldots + e_{r2} z_r$$
$$\vdots$$
$$PC_k = e_{1k} z_1 + e_{2k} z_2 + \ldots + e_{rk} z_r$$

Substituting:

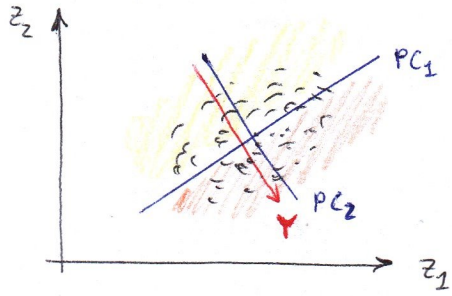
$$Y_0 = z_1 (e_{11} \hat{\beta}_1 + e_{12} \hat{\beta}_2 + \ldots + e_{1k} \hat{\beta}_k) + z_2 (e_{21} \hat{\beta}_1 + e_{22} \hat{\beta}_2 + \ldots + e_{2k} \hat{\beta}_k) + \ldots + z_r (e_{r1} \hat{\beta}_1 + e_{r2} \hat{\beta}_2 + \ldots + e_{rk} \hat{\beta}_k)$$

$\Longrightarrow$ $\boxed{Y_0 := \hat{\gamma}_1 Z_1 + \dots + \hat{\gamma}_r Z_r}$ $\qquad \hat{\gamma}_j = e_{j1}\hat{\beta}_1 + \dots + e_{jk}\hat{\beta}_k$

we'll be a fitted model with the same r variables $(Z_1, \dots, Z_r)$ with coefficients but these coefficients are not the OLS coefficients: these are linear combinations of the k parameters so those are k coefficients that we got after PCA.

Criticism:

- solution is not sparse in terms of $Z_1, \dots, Z_r$ (no variables selection) : we still got all the r variables
- we might have thrown away the information of Z correlated with Y: (doing the dimensional reduction in PCA not guarantee that the information we need to predict Y is not contained in the components that we're throwing away)



Suppose that Y increase along PC_2 (Y is higher in the zone). If we consider just PC_1 (and we throw away PC_2 for dimensional reduction) we're throwing away the direction where Y is varying as a function of Z_1 and Z_2 (we're keeping a direction so that there is no increase or decrease, Y is basically constant among PC_1).

RIDGE REGRESSION

(sort of a smoother version of PCA regression)
(Hoerl & Kennard)

Collinearity $\longrightarrow$ variability of $\hat{\beta}_{OLS}$ is large (which make predictions and interpretations all incertain)

Ridge optimization problem:

$$\begin{cases} \arg\min_{\beta} \|Y - Z\beta\|^2 \\ \|\beta\|^2 \leq s \end{cases}$$

s: a new parameter that we enter to the problem (without this constraint it's OLS)

Note that:

$$\|Y - Z\beta\|^2 = \|\underbrace{Y - \hat{Y}}_{\hat{\varepsilon}} - Z(\beta - \hat{\beta})\|^2$$

with $\hat{Y} = HY$

$\hat{\beta} = (Z^T Z)^{-1} Z^T Y$

$$= \|\underbrace{\hat{\varepsilon}}_{\in \mathcal{Z}^\perp(Z)} - \underbrace{Z(\beta - \hat{\beta})}_{\in \mathcal{Z}(Z)}\|^2$$

$$= \|\hat{\varepsilon}\|^2 + \|Z(\beta - \hat{\beta})\|^2$$

Ridge problem:

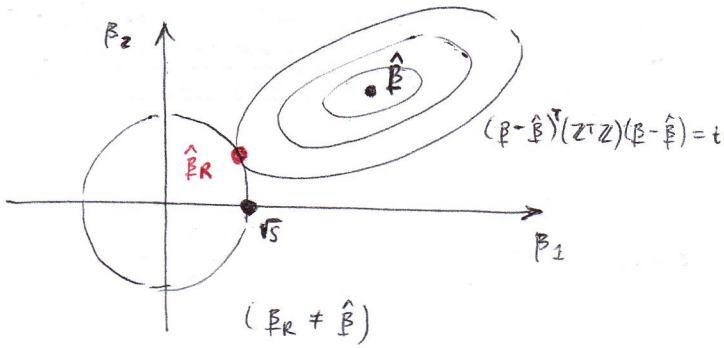
$$\begin{cases} \arg\min_{\beta} \|Z(\beta - \hat{\beta})\|^2 \\ \|\beta\|^2 \leq s \\ \hat{\beta} = (Z^T Z)^{-1} Z^T Y \end{cases}$$

(since $\|\hat{\varepsilon}\|^2 \perp\!\!\!\perp \beta$) ("it's constant w.r.t. β")

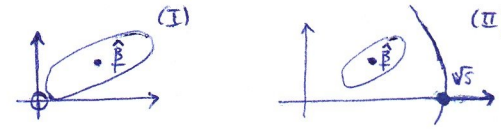
(OLS solution)

$$\begin{cases} \underset{\beta}{\arg\min} \; (\beta - \hat{\beta})^T Z^T Z (\beta - \hat{\beta}) \\ \beta^T \beta \le S \end{cases}$$

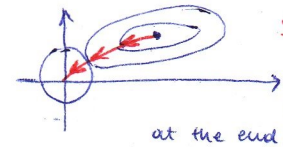
= we don't want solutions that are too big in terms of norm



Since S is a parameter (that control the variability of β's) we can have:



For very large S (II) the old OLS solution ($\hat{\beta}$) is still good because satisfies the constraints. For very small S, $S \to 0$ (I) happens that the solution is being shrink toward zero:



the smaller to zero the closer we get to the origin at the end we say β's = 0 so no regressors

$\hat{\beta}_R$ is a "restringimento" "**shrinked**" version of $\hat{\beta}_{OLS}$.

The solution is obtained via Lagrange method, which implies to solve:

$$\underset{\beta}{\arg\min} \; \|Y - Z\beta\|^2 + \lambda \|\beta\|^2 \qquad (\lambda = 0 \Rightarrow OLS)$$

where λ is a function of S: $\lambda \uparrow$ if $S \downarrow$ (we penalize solutions, the larger we penalize, the closer we get to zero)

Differentiate wrt β and get:

$$\hat{\beta}_R = (Z^TZ + \lambda I)^{-1} Z^T Y$$

we're free to chose any solution but we're paying λ dollars for every kilometer that we're far from zero (so we also want to minimize the dollours that we pay at the end).

this solves the colinearity problem: $(Z^TZ + \lambda I)$ is always invertible if $\lambda \neq 0$ and the larger the λ is the easiest is to invert

Obs.

- $\hat{\beta}_R$ is biased estimator of β
- $\exists \lambda$ s.t. $\mathbb{E}\left[\|\hat{\beta}_k - \beta\|^2\right] < \mathbb{E}\left[\|\hat{\beta}_{OLS} - \beta\|^2\right]$

we've paying in terms of bias but gaining in terms of variability (bias - variance trade off)

- Find the best λ by cross validation using prediction error for Y

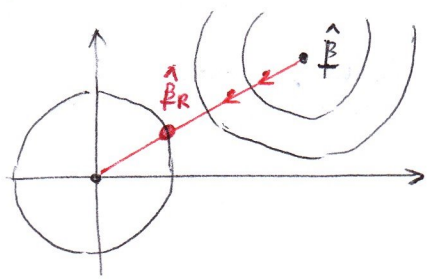
we cross validate our predictions through many possible choices on a grid of possible λ's and then we find the λ that minimize the prediction error for the Y

Obs.

- Standardize the regressors **and** the Y before fitting Ridge Regression since $\hat{\beta}_R$ **is not** scale invariant ($\hat{\beta}_{OLS}$ **is** scale invariant)

with RIDGE: if we measures in meters and kilograms and then we decide to switch into centimeters and grams, we cannot simply multiply our coefficients for appropriate factor in order to use the same model with a different scale

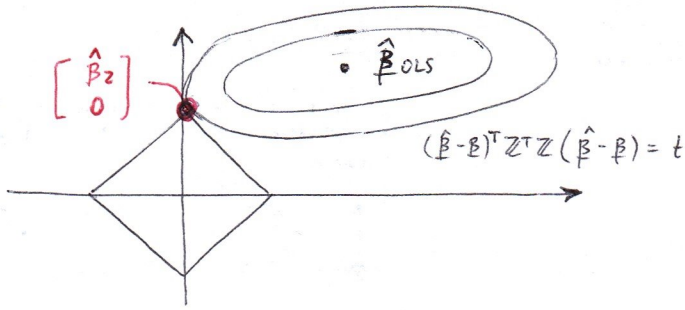
- if regressors are standardized and are orthogonal:



(!) it's not a curve anymore, it's a line:

$\hat{\beta}_R$ is $\hat{\beta}$ scaled.

LASSO : let's change the constraint



we want a constraint with angles instead of a constraint spherical.
Why? Because we hope that the tangent point is gonna be on one of the vertices.
if it's there, this means that $\beta_1 = 0$
$\Rightarrow$ the solution will be $[\hat{\beta}_2 \; 0]^T$, meaning that we're not only shrinking but also selecting the variables

Controlling the variability of $\hat{\beta}$ **and** selecting variables.

Lasso optimization problem:

$$\begin{cases} \underset{\beta}{\arg\min} \; (\beta - \hat{\beta})^T (Z^T Z)(\beta - \hat{\beta}) \\ \|\beta\|_1 \le s \qquad \left(\sum_{i=1}^{r} |\beta_i| \le s \right) \end{cases}$$

There is **no** analitical solution.

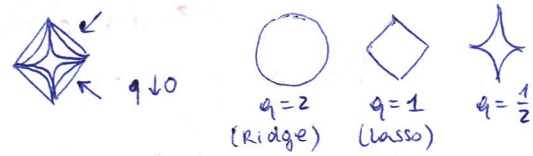
lagrangian (lagrange method):

$$\underset{\beta}{\arg\min} \; \|Y - Z\beta\| + \lambda \|\beta\|_1$$

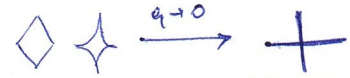
Find the optimal λ through cross validation and pred. error for Y.

Variations of Lasso: we've consider even more spiky constraints:

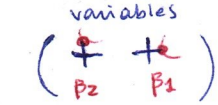
$$\begin{cases} \underset{\beta}{\arg\min} \; \|Y - Z\beta\|^2 \\ \|\beta\|_q \le s \end{cases}$$



$$\|\beta\|_q = \left(\sum_{i=1}^{r} |\beta_i|^q \right)^{1/q} \qquad (q<1)$$

For $q \downarrow 0$ $\Longrightarrow$ model selection problem because: 

it's pure selection of the variables



Extension: (**Elastic nets**) (take both Ridge & lasso)

$$\underset{\beta}{\arg\min} \; \|Y - Z\beta\| + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

General idea: (for all these problems)

fit a regression with a penalization.

This becomes super good when we already know something about the solution (in terms of model) $\rightarrow$ we penalize solutions that are far from the ideal solutions.

$\Rightarrow$ these methods take care of both: model selection part and the colinearity (= variability of β's).
Ridge is not very good in selecting the variables. lasso and variations go in the direction of model selection

Connection with **GLM**
(Generalized Linear Models)

$$Y = f(\underline{z}) + \varepsilon \qquad \text{basic model}$$

$$\mathbb{E}[Y|\underline{z}] = f(\underline{z})$$

Linear regression: $f(\underline{z}) = \beta_0 + \beta_1 z_1 + \dots + \beta_r z_r$

In GLM: we think that (↑) holds but not for $f(\underline{z})$ but for a transformation of $f(\underline{z})$: $g(f(\underline{z}))$

$$g(\mathbb{E}[Y|\underline{z}]) = \beta_0 + \beta_1 z_1 + \dots + \beta_r z_r \quad := \textbf{LINK FUNCTION}$$

Linear Regression:

- $\left.\begin{array}{l}\mathbb{E}[\varepsilon] = 0 \\ \text{Var}(\varepsilon) = \sigma^2\end{array}\right\} \Rightarrow Y \sim \begin{array}{l}\mathbb{E}[Y|\underline{z}] = f(\underline{z}) \\ \text{Var}(Y|\underline{z}) = \sigma^2\end{array}$
- $\varepsilon \sim N(0, \sigma^2) \Rightarrow Y|\underline{z} \sim N(f(\underline{z}), \sigma^2)$

GLM:

$$Y|\underline{z} \sim F \left\{\begin{array}{l}\text{Bernoulli} \\ \text{Poisson} \\ \text{Gaussian} \\ \vdots\end{array}\right. ; \quad \begin{array}{l}((*)) \\ \text{the parameters } \underline{\beta} \text{ are estimated by ML (maximum likelihood) and not by least squares}\end{array}$$

the model is not only an extension of the linear model in the sense that we're taking a linear representation for a transformation of the mean (and not for the mean itself) but also we allow the error structure to be far away from the gaussian. what do we pay? least squares is no longer working. ((*))

Link functions:

- Y Gaussian $\Rightarrow$ $g(\mu) = \mu$ — **Linear** Regression
- Y Bernoulli (p) $\Rightarrow$ $g(p) = \log\left(\frac{p}{1-p}\right)$ — **Logit** Regression
- Y Poisson (λ) $\Rightarrow$ $g(\lambda) = \log(\lambda)$ — **Poisson** Regression
- Y Bernoulli (p) $\Rightarrow$ $g(p) = \Phi^{-1}(p)$ — **Probit** Regression

Φ cumulative distr. of $N(0,1)$

Everything we said about Lasso, Ridge, ..., we can apply here.
(Ex. Lasso Logistic Regression, ...)