Regression

The idea of regression is to model a variable y (called response variable), as a function of a certain number of variables $x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$ (called explanatory variables):

$$y = g(x) = g(x_1, \dots, x_p).$$

Data takes the form of a sample of n $(p+1)$-tuples (x, y), and the goal is to estimate g.
The simplest model consists in taking g to be linear, which means there exists a vector of coefficients $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ such that

x' stands for transpose

$$y = x'\beta = \beta_1 x_1 + \dots + \beta_p x_p.$$

In practice, this does not work exactly, and one has to take into account measurement errors.

The idea is then to see y as the realization (outcome) of a random variable Y. More precisely, the model writes as

$$Y = x'\beta + \varepsilon$$

$$= \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon,$$

where the random variable ε is centered with unknown variance σ^2

$$\mathbb{E}[\varepsilon] = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2.$$

The goal is to estimate the parameter β and the variance σ^2 of the error ε. ($\rightarrow$ Back to parametrics!)

Example 1: (Ozone concentration)

O_3 = daily maximum of ozone concentration ($\mu g/m^3$)

is to be explained by

T = temperature at noon.

We write the model $O_3 = \beta_1 + \beta_2 T + \varepsilon$.

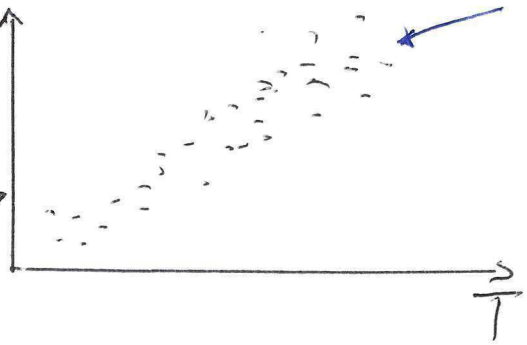
In such a case where there is only one "true" explanatory variable (temperature), we talk about <u>simple linear regression</u>.

Bk: Notation!! ⚠

The textbook uses "$Y = \beta_0 + \beta_1 x + \varepsilon$".

Data would look like: O_3

<u>Scatterplot</u> →



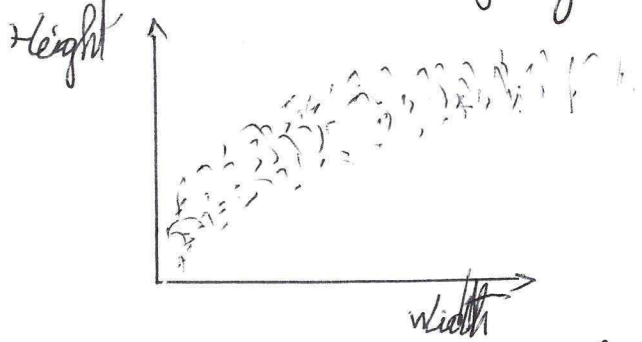
1 point = one piece of data (x_i, y_i)

One can refine this model by taking into account a cloud index C and the wind speed W to get

$$O_3 = \beta_1 + \beta_2 T + \beta_3 C + \beta_4 W + \varepsilon$$

<u>Example 2</u> (Height of Eucalyptus trees)

Data of tree width x_i and their height y_i look like this (for eucalyptus)



Seeing this, we could try to write a model of the form

$$Y = \beta_1 + \beta_2 x + \beta_3 \sqrt{x} + \varepsilon$$

Rk this example shows that linear regression is only linear in the parameter β, not in the variable x!

Example 3: (Cobb - Douglas model)

Source: "A theory of production", 1928

We study the U.S. economy, and write:

P : Production

K : Capital

W : Number of Workers

The model would write as

$$P = \alpha_1 K^{\alpha_2} W^{\alpha_3}.$$

Taking the logarithm, writing $(\beta_1, \beta_2, \beta_3) = (\log \alpha_1, \alpha_2, \alpha_3)$, and taking into account modelling errors, we end up getting

$$\log P = \beta_1 + \beta_2 \log K + \beta_3 \log W + \varepsilon.$$

Rk: Here, coming from a nonlinear model, we "linearized" it using the logarithm.

Rk: (Linearizing in general)

If the true regression model is $Y = g(x'\beta + \varepsilon)$, then using the inverse transformation g^{-1}, it becomes linear (in β!). In practice, we don't know g, so we can just "try" some classical functions (exp, log, ...) and see if the scatterplot gets linearized.

The linear Model

Modeling

We assume that the collected data follow the model

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon$$

$$(1 \leq i \leq n)$$

where

- are random variables, of which we observe outcomes y_i's
- x_{ij}'s are known, non-random (x_{i1} being often $= 1$ for all i)
- Parameters β_j are unknown, but non-random
- ε_i's are random variables and unknown (not observed, as opposed to Y_i's)

Rk: As mentionned above, the "constant" (= intercept) often belongs to the model, so that the textbook writes the model as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad (1 \leq i \leq n)$$

so that p still corresponds to the number of "true" explanatory variables. With our convention of notation, if $x_{i1} = 1$ for all $1 \leq i \leq n$, p is the number of parameters to estimate, while the number of explanatory variables is, properly speaking, $(p-1)$.

Adopting a matrix notation, we get the following notation:

Def: (Linear Regression Model)

A linear regression model is defined by

$$Y = X\beta + \varepsilon,$$

where

- Y is a random vector of dimension n (known)
- X is a (known) $n \times p$ matrix (called design matrix)
- β is the vector of parameters (unknown) and has dimension p.
- ε is the (unknown) vector of errors

Rk: Expanding the matrix notation, we can write

$$\underset{n\times 1}{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}} = \underset{n\times p}{\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{pmatrix}} \underset{p\times 1}{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}} + \underset{n\times 1}{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}$$

$$= \begin{pmatrix} x_{11}\beta_1 + \cdots + x_{1p}\beta_p + \varepsilon_1 \\ \vdots \\ x_{n1}\beta_1 + \cdots + x_{np}\beta_p + \varepsilon_n \end{pmatrix}$$

Assumptions on the model

$$(A)\begin{cases} (A_1): \operatorname{rank}(X) = p \\ (A_2): \text{The } \varepsilon_i\text{'s are iid with } \mathbb{E}[\varepsilon] = 0 \\ \qquad\quad \text{and } \operatorname{Var}(\varepsilon) = \sigma^2 I_n \end{cases}$$

Rk:

- About (A_1): It ensures that the model is identifiable.

Note that (A_1) is equivalent to requiring that the matrix $X'X$ is invertible

Indeed, if there exists $\beta \in \mathbb{R}^p$ with $X'X\beta = 0$, then $\|X\beta\|^2 = \beta' X' X \beta = 0$, so $X\beta = 0$ and hence $\beta = 0$ since $\operatorname{rank}(X) = p$.

In other words, the symmetric matrix $X'X$ is positive definite.

. About (A_2): Assuming errors to be centered is very natural: if it was not the case, their mean m would go into the deterministic part of the model, up to add a parameter $\beta_0 = m$ if the constant is not already taken into account. Furthermore, we'll see in the proofs that we could actually drop the "iid" assumption and work with $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I_m$ only → Such an assumption is referred to as homoscedasticity

---

Notation: . We'll write $X = (X_1 | \cdots | X_p)$, where X_j is the column vector of dimension m corresponding to the j^{th} variable.

. The i^{th} row of X will be denoted by $x_i = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,p} \end{pmatrix}$, and corresponds to the i^{th} "individual", with $Y_i = x_i'\beta + \varepsilon_i$.

Least Square Estimator

Our goal is first to estimate β. Roughly speaking, we want to pick one $\hat{\beta}$ that has $Y_i \simeq \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$ for all $1 \le i \le n$ simultaneously, or equivalently, $\left(Y_i - \sum \hat{\beta}_j x_{ij}\right)$ small for all $1 \le i \le n$. The simplest way (mathematically speaking) to encode this idea is to minimize the sum of the squares of the $\left(Y_i - \sum \hat{\beta}_j x_{ij}\right)$'s.

Def: (Least Square Estimator)

The least square estimator $\hat{\beta}$ is defined by:

$$\hat{\beta} = \underset{\alpha \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \alpha_j x_{ij}\right)^2$$

$$= \underset{\alpha \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - x_i' \alpha\right)^2$$

$$= \underset{\alpha \in \mathbb{R}^p}{\operatorname{argmin}} \left\| Y - \sum_{j=1}^{p} \alpha_j X_j \right\|^2$$

$$= \underset{\alpha \in \mathbb{R}^p}{\operatorname{argmin}} \left\| Y - X\alpha \right\|^2$$

Let think in a geometric way to find what $\hat{\beta}$ is. The design matrix $X = (X_1 | \dots | X_p)$ consists of p column vectors of $\mathbb{R}^n$ (the first one usually being composed of 1's). The subspace spanned by these vectors is

$$\mathcal{M}_X = \mathrm{Im}(X) = \mathrm{Span}(X_1, \dots, X_p).$$

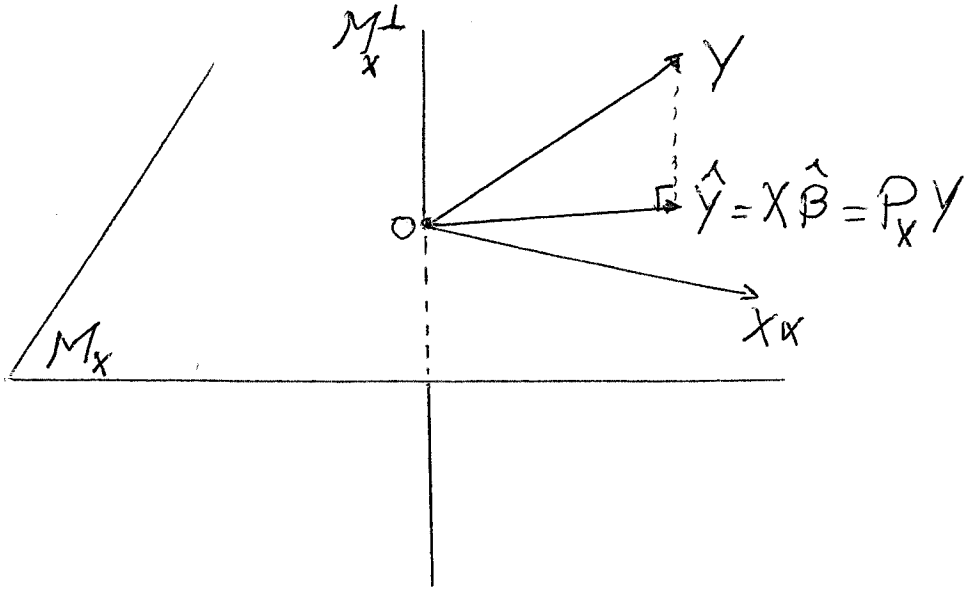
It has dimension $\dim(\mathcal{M}_X) = p$ from assumption (A_1), and all the vectors of this space can be written as

$$X\alpha = \alpha_1 X_1 + \dots + \alpha_p X_p. \qquad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix}$$

According to the model, Y is the sum of an element $X\beta \in \mathcal{M}_X$ with an error term ε, which has no reason to belong to $\mathcal{M}_X$.

Minimizing $\|Y - X\alpha\|^2$ amounts to search for the element of $\mathcal{M}_X$ which is closest to Y for the Euclidean norm.

This (unique) element is, by definition, the orthogonal projection of Y onto $\mathcal{M}_X$. It is denoted by $\boxed{\hat{Y} = P_X Y}$, where P_X is the orthogonal projection matrix on $\mathcal{M}_X$. It can also be written as $\hat{Y} = X\hat{\beta}$, where $\hat{\beta}$ is the least square estimator.



The orthogonal space of $\mathcal{M}_x$, denoted by $\mathcal{M}_x^\perp$, is sometimes called the "space of residuals". As the orthogonal complement of $\mathcal{M}_x$, we has

$$\dim(\mathcal{M}_x^\perp) = \dim(\mathbb{R}^p) - \dim(\mathcal{M}_x) = n - p.$$

The following expressions of $\hat{\beta}$ and P_x may be the most important of this chapter, because one can find (back) all the results from them

Prop The least square estimator $\hat{\beta}$ writes as

$$\hat{\beta} = (X'X)^{-1}X'y,$$

and P_x, the orthogonal projection matrice onto $\mathcal{M}_x = \text{Im}(X)$ is

$$P_x = X(X'X)^{-1}X'.$$

(12)

Proof (1): We can show the result by different manners:

1) By differentiation: We are looking for $\alpha \in \mathbb{R}^p$ that minimizes the function

$$S(\alpha) = \|Y - X\alpha\|^2$$
$$= \alpha'(X'X)\alpha - 2Y'X\alpha + \|Y\|^2.$$

Since S is quadratic in α, with $X'X$ symmetric and positive definite, the problem has a unique solution $\hat{\beta}$: this is the point where its gradient vanishes, so we solve

$$\nabla S(\hat{\beta}) = 2\hat{\beta}'X'X - 2Y'X = 0$$

$$\Leftrightarrow (X'X)\hat{\beta} = X'Y.$$

Since $X'X$ is invertible (from (A_1)), we get $\hat{\beta} = (X'X)^{-1}X'Y$.

Furthermore, by definition, $\hat{Y} = P_X Y = X\hat{\beta} = X(X'X)^{-1}X'Y$ and that this relation is valid for all $Y \in \mathbb{R}^n$, we deduce that $P_X = X(X'X)^{-1}X'$.

2) By projection: $\hat{Y} = X\hat{\beta}$ is the unique vector such that $(Y - \hat{Y})$ is orthogonal to $\mathcal{M}_X$. Since $\mathcal{M}_X$ is spanned by $X_1, \dots, X_p$, this amounts to say that $(Y - \hat{Y})$ is orthogonal to all the X_i's.

This translates to the system $\begin{cases} X_1'(Y - X\hat{\beta}) = 0 \\ \vdots \\ X_p'(Y - X\hat{\beta}) = 0 \end{cases}$, that we can compactify into the matrix form $X'(Y - X\hat{\beta}) = 0$, from which we deduce the expression of $\hat{\beta}$, and then P_X].

Notation: From now on, $P_X = X(X'X)^{-1}X'$ is the orthogonal projection matrix onto $\mathcal{M}_X = Im(X)$, and $P_{X^\perp} = (I_n - P_X)$ is the orthogonal projection matrix onto $\mathcal{M}_X^\perp$. In particular, the decomposition

$$Y = \hat{Y} + (Y - \hat{Y}) = P_X Y + (I_n - P_X) Y$$

$$= P_X Y + P_{X^\perp} Y$$

is nothing but the orthogonal decomposition of Y on $\mathcal{M}_X$ and $\mathcal{M}_X^\perp$.

Reminder on projections:

Let P be a square matrix of size n. We say that P is a projection matrix when $P^2 = P$. This name comes from the fact that Px is the projection of $x \in \mathbb{R}^n$ onto $Im(P)$ parallelly to $Ker(P)$.

If, in addition to satisfying $P^2 = P$, P is symmetric ($P' = P$), Px is the orthogonal projection of x on $Im(P)$ parallelly to $Ker(P) = Im(P)^\perp$.

This means that we have

$$x = Px + (x - Px) \quad \text{with} \quad Px \perp (x - Px).$$

In this case, one can find $Q \in \mathbb{R}^{n \times n}$, an orthogonal matrix ($QQ' = I_n$) and $\Delta \in \mathbb{R}^{n \times n}$, a diagonal matrix composed of 0's and 1's only, such that $P = Q \Delta Q'$. One easily sees that Δ has exactly p 1's, since $Trace(P) = Trace(\Delta)$.

Exercise: $P_X = X(X'X)^{-1}X'$ is an orthogonal projector. (Show this!)

$$Trace(P_X) = p, \quad Trace(P_{X^\perp}) = n - p$$

Vocabulary: P_X is sometimes denoted by H (for "Hat"), because it "puts a hat" on the vector Y: $P_X Y = HY = \hat{Y}$

"Hat matrix"

Let us now focus on the bias and variance of $\hat{\beta}$. Recall that

$$\text{Cov}(\hat{\beta}) = \mathbb{E}\left[(\hat{\beta} - \mathbb{E}(\hat{\beta}))(\hat{\beta} - \mathbb{E}(\hat{\beta}))'\right]$$

$$= \mathbb{E}\left[\hat{\beta}\hat{\beta}'\right] - \mathbb{E}\left[\hat{\beta}\right]\mathbb{E}\left[\hat{\beta}\right]' \in \mathbb{R}^{p\times p}$$

Furthermore, for any matrix $A \in \mathbb{R}^{m\times p}$,

$$\mathbb{E}\left[A\hat{\beta} + b\right] = A\mathbb{E}\left[\hat{\beta}\right] + b \quad \text{and} \quad \text{Cov}(A\hat{\beta} + b) = A\,\text{Cov}(\hat{\beta})A'$$

Prop: The least square estimator is unbiased, i.e. $\mathbb{E}(\hat{\beta}) = \beta$, and has covariance matrix

$$\text{Cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

Proof

Bias: $\mathbb{E}(\hat{\beta}) = \mathbb{E}\left((X'X)^{-1}X'Y\right)$

$= (X'X)^{-1}X'\mathbb{E}(Y)$

$= (X'X)^{-1}X'\mathbb{E}(X\beta + \varepsilon)$

$= (X'X)^{-1}X'\left(X\beta + \underbrace{\mathbb{E}(\varepsilon)}_{=0}\right)$

$= (X'X)^{-1}X'X\beta$

$= \beta$ ∎

Covariance : $\mathrm{Cov}(\hat{\beta}) = \mathrm{Cov}\left((X'X)^{-1}X'Y\right)$

$$= (X'X)^{-1}X' \underbrace{\mathrm{Cov}(Y)}_{= \mathrm{Cov}(X\beta+\varepsilon) = \mathrm{Cov}(\varepsilon) = \sigma^2 I_n} X(X'X)^{-1},$$

So that $\mathrm{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}X'X(X'X)^{-1}$

$$= \sigma^2(X'X)^{-1}$$

The residuals are defined by

$$\hat{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = Y - X\hat{\beta}$$

$$= Y - \hat{Y}$$

$$= (I_n - P_X)Y$$

$$= P_{X^\perp} Y$$

$$= P_{X^\perp} \varepsilon,$$

since $Y = X\beta + \varepsilon$ and $X\beta \in \mathcal{M}_X$.

If $\hat{\beta}$ estimates β well, then the residuals $\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta}$ estimate the errors well. Hence, a natural estimator of the residual variance σ^2 is

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

$$= \frac{\|\hat{\varepsilon}\|^2}{n}$$

$$= \frac{SSR}{n},$$

where $\boxed{SSR = \|\hat{\varepsilon}\|^2}$ is the Sum of Squared Residuals.

Actually, this estimator is biased, according to the following result

Prop The statistic

$$\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-p} = \frac{SSR}{n-p}$$

is an unbiased estimator of σ^2

Proof: $\mathbb{E}(\|\hat{\varepsilon}\|^2) = \mathbb{E}\left[\|P_{X^\perp}\varepsilon\|^2\right]$

$$= \mathbb{E}\left[\varepsilon' P_{X^\perp}' P_{X^\perp}\varepsilon\right]$$

$$= \mathbb{E}\left[\varepsilon' P_{X^\perp}\varepsilon\right].$$

(18) Hence, $\mathbb{E}(\|\hat{\varepsilon}\|^2) = \mathbb{E}\left(\sum_{i,j=1}^{n} P_{X^\perp}(i,j)\, \varepsilon_i \varepsilon_j \right)$

$$= \sum_{i,j=1}^{n} P_{X^\perp}(i,j)\, \mathbb{E}(\varepsilon_i \varepsilon_j) \quad \leftarrow \begin{cases} = 0 \text{ if } i \neq j \\ = \sigma^2 \text{ if } i = j \end{cases}$$

$$= \sum_{i=1}^{n} P_{X^\perp}(i,i)\, \sigma^2$$

$$= \sigma^2 \, \mathrm{Trace}(P_{X^\perp})$$

$$= \sigma^2 (n-p)$$

Rk: We deduce an unbiased estimator of $\mathrm{Cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$:

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$$

$$= \frac{\|\hat{\varepsilon}\|^2}{n-p} (X'X)^{-1}$$

$$= \frac{SSR}{n-p} (X'X)^{-1}$$

In particular, we get an estimator of the standard deviation of $\hat{\beta}_j$, the j^{th} coefficient of $\hat{\beta}$:

$$\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma} \sqrt{\left[(X'X)^{-1} \right]_{jj}}$$

The Gaussian linear Model

So far, we have assumed that $Y = XB + \varepsilon$, with

$$(A)\begin{cases}(A_1) & \text{rank}(X) = p \\ (A_2) & \varepsilon_i \text{ iid with } \mathbb{E}(\varepsilon) = 0 \text{ and } \text{Cov}(\varepsilon) = \sigma^2 I_n\end{cases}$$

In what follows, we will do a stronger assumption, namely the Gaussianity of the residuals. From now on, we'll assume that

$$\boxed{(A)\begin{cases}(A_1) & \text{rank}(p) \\ (A_2) & \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)\end{cases}}$$

The benefit of this distributional assumption is that we are going to be able to derive the distributions of our estimators, and therefore, to build confidence intervals and design testing procedures.

Rk: . We are back to a well defined parametric model on the distribution of Y:

$$\{P_\theta\}_{\theta \in \Theta} = \left\{\mathcal{N}(XB, \sigma^2 I_n)\right\}_{B \in \mathbb{R}^p, \sigma > 0}$$

This model is identifiable when, by definition, $(B, \sigma) \longmapsto \mathcal{N}(XB, \sigma^2 I_n)$ is one-to-one. But this is true only when $\text{rank}(X) = p$ (A_1).

- As opposed to what we have done in the other chapters, we are not in a sampling model, since all the variables Y_i do not have the same distribution:

$$Y_i \sim N(x_i'\beta, \sigma^2).$$

An important reminder before we carry on:

Thm: (Cochran)

Let $Y \sim N(\mu, \sigma^2 I_n)$, and $M \subset \mathbb{R}^n$ be a linear subspace of dimension p. Write P for the orthogonal projection onto M, and $P_\perp = (I_n - P)$ for the orthogonal projection onto $M^\perp$. Then

(i) $PY \sim N(P\mu, \sigma^2 P)$ and $P_\perp Y \sim N(P_\perp \mu, \sigma^2 P_\perp)$

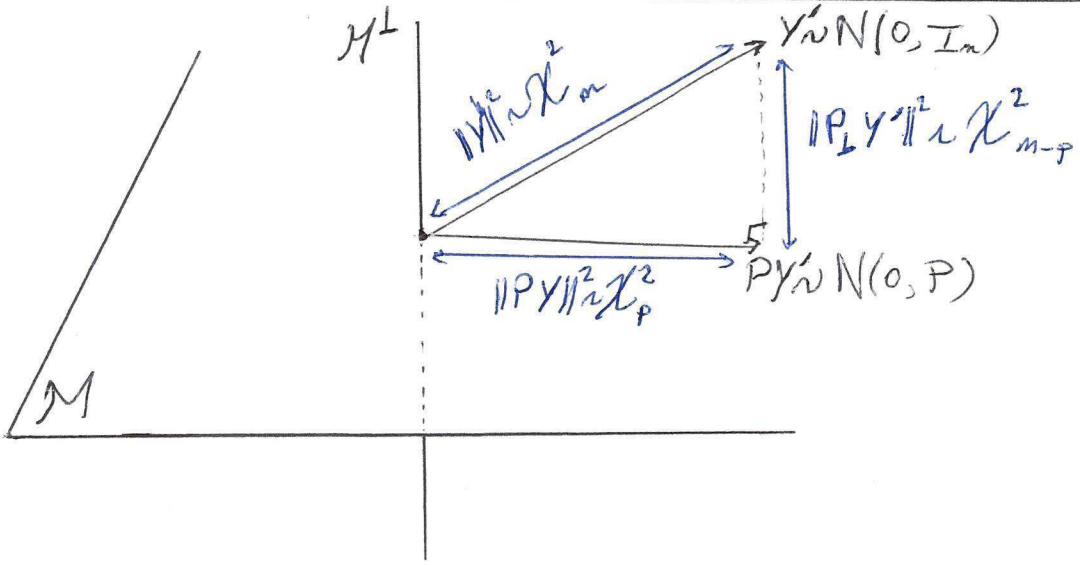
(ii) PY and $P_\perp Y = (Y - PY)$ are independent

(iii) $\dfrac{\|P(Y-\mu)\|^2}{\sigma^2} \sim \chi^2_p$ and $\dfrac{\|P_\perp(Y-\mu)\|^2}{\sigma^2} \sim \chi^2_{n-p}$

Illustration:

for $\begin{cases} \mu = 0 \\ \sigma^2 = 1 \end{cases}$

(or by writing $Y' = \dfrac{Y-\mu}{\sigma}$)



Distributions of Estimators and Confidence Domains

We will apply Cochaan to $Y = X\beta + \varepsilon \sim N(X\beta, \sigma^2 I_n)$, since $\hat{\beta}$ and $\hat{\sigma}^2$ can be seen as projections of it on orthogonal subspaces.

Prop. (Distribution of Estimators with known Variance)

Under assumptions (A),

(i) $\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$

(ii) $\hat{\beta}$ and $\hat{\sigma}^2$ are independent

(iii) $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$

Proof: (i) $\hat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$ with $\varepsilon \sim N(0, \sigma^2 I_n)$

(ii) We have $\hat{\beta} = (X'X)^{-1}X'Y$ $\quad = P_X$

$$= (X'X)^{-1}X'\left(X(X'X)^{-1}X'\right)Y$$

$$= (X'X)^{-1}X'P_X Y,$$

$\Rightarrow \hat{\beta}$ is a function of the orthogonal projection $P_X Y$ of Y onto M_X. On the other hand,

$$\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-p} = \frac{\|P_{X^\perp} Y\|^2}{n-p}.$$

From Cochan, we get the result

(iii) $\hat{\varepsilon} = P_{X^\perp}\varepsilon$ with $\begin{cases} \dim(\mathcal{M}_X^\perp) = n-p \\ \varepsilon \sim N(0, \sigma^2 I_n) \end{cases}$. Therefore, from Cochran,

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\|P_{X^\perp}\varepsilon\|^2}{\sigma^2} = \frac{\|P_{X^\perp}(\varepsilon - \mathbb{E}(\varepsilon))\|^2}{\sigma^2} \sim \chi^2_{n-p}$$

Of course, the previous proposition is not satisfactory to derive confidence regions for β since it assumes σ^2 to be known. The following proposition overcomes this issue.

Prop (Distribution of Estimators with unknown variance)

Under assumptions (A),

(i) for all $i \in \{1, \dots, p\}$,

$$T_j \overset{\text{def.}}{=} \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim T_{n-p}.$$

T-distribution

(ii) $$F \overset{\text{def.}}{=} \frac{1}{p\hat{\sigma}^2}(\hat{\beta}-\beta)(X'X)(\hat{\beta}-\beta) \sim F_{p,n-p}$$

F-distribution

Proof: Applying Cochran again!

(i) $\hat{\beta}_j \sim N(\beta_j, \sigma^2[(X'X)^{-1}]_{jj}) \perp\!\!\!\perp \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$

$$\Rightarrow T_j = \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{[(X'X)^{-1}]_{jj}}} \Big/ \frac{\hat{\sigma}}{\sigma} \sim T_{n-p}$$

(ii) $\frac{1}{\sigma}(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta) \sim \chi^2_p \perp\!\!\!\perp (n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$

$= \|X(\beta-\hat{\beta})\|^2$

$$\Rightarrow F \sim F_{p,n-p}$$

Example: Consider the case $p=2$, so that $(\hat{\beta}-\beta) = \begin{pmatrix} \hat{\beta}_1 - \beta_1 \\ \hat{\beta}_2 - \beta_2 \end{pmatrix}$ in the model $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$. We have

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

so that $X'X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = \begin{pmatrix} n & n\bar{x}_n \\ n\bar{x}_n & \sum x_i^2 \end{pmatrix}$.

Point (ii) of the previous proposition writes as

$$\frac{1}{2\hat{\sigma}^2}\left(n\left(\hat{\beta}_1-\beta_1\right)^2 + 2n\bar{x}_n\left(\hat{\beta}_1-\beta_1\right)\left(\hat{\beta}_2-\beta_2\right) + \left(\sum x_i^2\right)\left(\hat{\beta}_2-\beta_2\right)^2 \right) \sim F_{2,n-2},$$

which is going to help us build an <u>confidence ellipse</u> for $\beta=(\beta_1,\beta_2)'$.

More generally, for $p>2$, it yields a <u>confidence ellipsoid</u> for β centered at $\hat{\beta}$.

<u>Thm</u>: (Confidence Intervals and Confidence Regions)

(i) For all $j\in\{1,\dots,p\}$, a confidence interval of level $(1-\alpha)$ for β_j is

$$\left[\hat{\beta}_j - t_{1-\frac{\alpha}{2},n-p}\,\hat{\sigma}\sqrt{\left[(X'X)^{-1}\right]_{jj}}\ ,\ \hat{\beta}_j + t_{1-\frac{\alpha}{2},n-p}\,\hat{\sigma}\sqrt{\left[(X'X)^{-1}\right]_{jj}}\right]$$

$1-\frac{\alpha}{2}$ quantile of T_{n-p}

(ii) A confidence interval of level $(1-\alpha)$ for σ^2 is

$$\left[\frac{(n-p)\hat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2},n-p}}\ ,\ \frac{(n-p)\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2},n-p}}\right]$$

$1-\frac{\alpha}{2}$ and $\frac{\alpha}{2}$ quantiles of χ^2_{n-p}

(iii) A confidence region of level $(1-\alpha)$ for β is the ellipsoid

$$\left\{\beta\in\mathbb{R}^p \;\middle|\; \frac{1}{p\hat{\sigma}^2}(\hat{\beta}-\beta)(X'X)(\hat{\beta}-\beta) \le F_{1-\alpha,p,n-p}\right\}$$

$1-\alpha$ - quantile of $F_{p,n-p}$

Proof: Just apply the previous result. ∎

Reminder: If $(x_0, y_0) \in \mathbb{R}^2$, $c^2 > 0$ and $S = \begin{pmatrix} s_{11} & s_{21} \\ s_{12} & s_{22} \end{pmatrix} \in \mathbb{R}^{2\times 2}$ be a symmetric matrix ($s_{12} = s_{21}$) the set of points $(x, y) \in \mathbb{R}^2$ such that

$$(x - x_0, y - y_0)\, S \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \leq c^2 \quad \left(\Leftrightarrow s_{11}(x-x_0)^2 + 2 s_{12}(x-x_0)(y-y_0) + s_{22}(y-y_0)^2 \leq c^2 \right)$$

is an ellipse centered at (x_0, y_0) with axes given by the eigenvectors of S.

Example: Back to $p = 2$, we have derived the a confidence region of level $(1-\alpha)$ is the ellipse (with $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$)

"Interior of an ellipse"

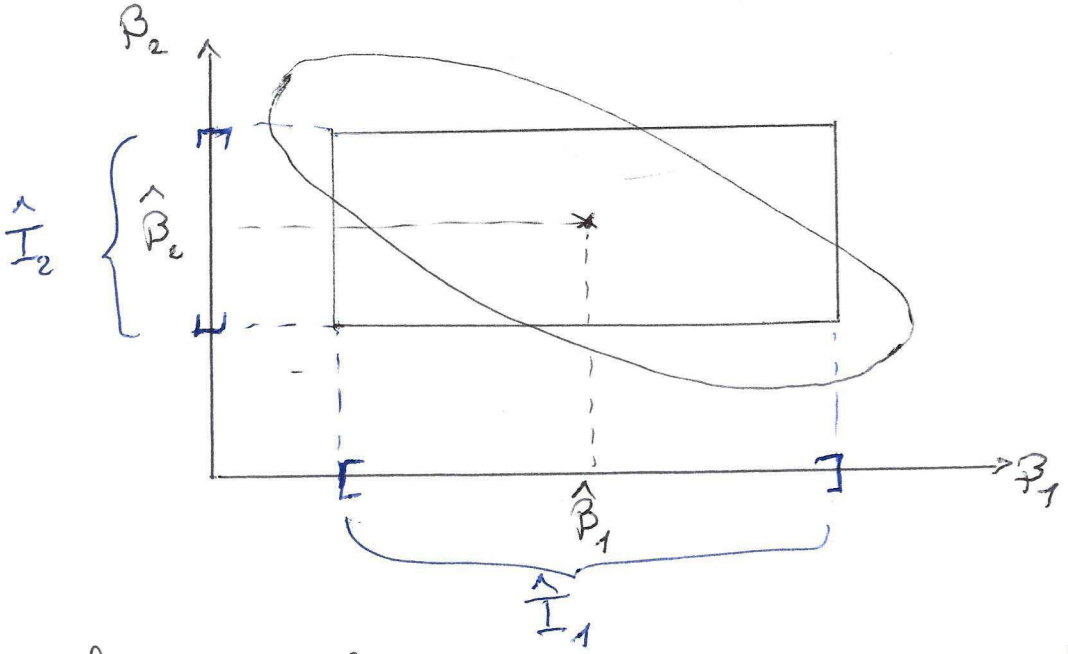
$$\left\{ (\beta_1, \beta_2) \in \mathbb{R}^2 \;\middle|\; \frac{1}{2\hat{\sigma}^2} \left(n(\beta_1 - \hat{\beta}_1)^2 + 2n\bar{x}(\beta_1 - \hat{\beta}_1)(\beta_2 - \hat{\beta}_2) + \left(\sum x_i^2\right)(\beta_2 - \hat{\beta}_2)^2 \right) \leq F_{1-\alpha,\, 2,\, n-2} \right\}.$$

Let us now consider the confidence intervals $\hat{I}_1$ and $\hat{I}_2$ of level $(1-\alpha)$ for β_1 and β_2 given by point (i). Defining the rectangle $\hat{R} = \hat{I}_1 \times \hat{I}_2$ (cuboid), we have

$$\begin{aligned} \mathbb{P}\left((\beta_1, \beta_2) \notin \hat{R} \right) &= \mathbb{P}\left(\beta_1 \notin \hat{I}_1 \text{ or } \beta_2 \notin \hat{I}_2 \right) \\ &\leq \mathbb{P}\left(\beta_1 \notin \hat{I}_1 \right) + \mathbb{P}\left(\beta_2 \notin \hat{I}_2 \right) \\ &\leq 2\alpha, \end{aligned}$$

so that $\hat{R}$ is only a confidence region of level $1 - 2\alpha$ ...

To obtain a confidence rectangle $\hat{I}_1 \times \hat{I}_2 = \hat{R}$ of level $(1-\alpha)$, we need to start from confidence intervals of levels $\left(1 - \frac{\alpha}{2}\right)$.



The above figure allows to distinguish between confidence intervals considered separately for β_1 and β_2, and a confidence region for $\beta = (\beta_1, \beta_2)$

simultaneous

Prediction

Once the model has been built, meaning that the parameters β and σ^2 have been estimated, one can of course use them to make prediction.

Let then $x'_{n+1} = (x_{n+1,1} \dots x_{n+1,p})$ be a new value for which we would like to predict Y_{n+1}. This response variable is defined by

$$Y_{n+1} = x'_{n+1}\beta + \varepsilon_{n+1} \quad \text{with } \varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2)$$
independent from $(\varepsilon_i)_i$.

A natural method consists in predicting the associated value corresponding to the adjusted model

$$\hat{Y}_{n+1} = x'_{n+1}\hat{\beta}$$

$\hat{\beta}$ computed using the first n data points $(X_1, Y_1), \dots, (X_n, Y_n)$.

The prediction error is then defined by

$$\hat{\varepsilon}_{n+1} = Y_{n+1} - \hat{Y}_{n+1}$$
$$= x'_{n+1}(\beta - \hat{\beta}) + \varepsilon_{n+1}.$$

Two different kinds of error are corrupting this prediction:

- Randomness of ε_{n+1},
- Uncertainty on $\hat{\beta}$.

Prop: (Prediction Error)

The prediction error $\hat{\varepsilon}_{n+1} = (Y_{n+1} - \hat{Y}_{n+1})$ has distribution

$$\hat{\varepsilon}_{n+1} \sim N\left(0, \sigma^2\left(1 + x'_{n+1}(X'X)^{-1}x_{n+1}\right)\right).$$

comes from ε_{n+1} — comes from uncertainty on $\hat{\beta}$.

Proof: We have $\hat{\varepsilon}_{n+1} = x'_{n+1}(\beta - \hat{\beta}) + \varepsilon_{n+1}$, so it is normal because it is the sum of two independent normals. Furthermore,

$$\begin{aligned} \mathbb{E}(\hat{\varepsilon}_{n+1}) &= x'_{n+1}\mathbb{E}(\beta - \hat{\beta}) + \mathbb{E}(\varepsilon_{n+1}) \\ &= 0 + 0 \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(\hat{\varepsilon}_{n+1}) &= \mathrm{Var}\left(x'_{n+1}(\beta - \hat{\beta})\right) + \mathrm{Var}(\varepsilon_{n+1}) + \mathrm{Cov}\left(x'_{n+1}(\beta - \hat{\beta}), \varepsilon_{n+1}\right) \\ &= x'_{n+1}\,\mathrm{Var}(\beta - \hat{\beta})\,x_{n+1} + \sigma^2 + 0 \\ &= x'_{n+1}\left(\sigma^2(X'X)^{-1}\right)x_{n+1} + \sigma^2 \end{aligned}$$

Rk: One can show that the variance of the prediction error is minimal for $x'_{n+1} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$ as soon as the intercept (= constant) is one of the variables. This is actually easy to see for simple linear regression: indeed, in this case we have, for $x'_{n+1} = (1 \; x)$,

$$\mathrm{Var}(\hat{\varepsilon}_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

$$\geq \sigma^2 \left(1 + \frac{1}{n}\right).$$

→ equality (only) for $x = \bar{x}$

In other words, it is harder to make a prediction for a point x_{n+1} which is far from the center of the point cloud $\bar{x}$. Intuitively, this is explained by the fact that the farther x_{n+1} from $\bar{x}$, the less information we have on it.

This idea leads to the notion of <u>leverage</u> of a point (not covered here)

As usual, the knowledge of the distribution of $\hat{\varepsilon}_{n+1}$ leads to an interval where the true target Y_{n+1} lies with high probability. In this context, we talk about a <u>prediction interval</u>.

Prop: (Prediction Interval)

A confidence interval, referred to as Prediction Interval, of level $(1-\alpha)$ for Y_{n+1} is

$$\left[\underbrace{x'_{n+1}\hat{\beta}}_{=\hat{Y}_{n+1}} \pm t_{1-\frac{\alpha}{2},\, n-p}\, \hat{\sigma}\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}} \right]$$

Proof: We notice that

$$\frac{Y_{n+1} - \hat{Y}_{n+1}}{\hat{\sigma}\sqrt{1+x'_{n+1}(X'X)^{-1}x_{n+1}}} = \frac{\left.\dfrac{Y_{n+1} - \hat{Y}_{n+1}}{\sigma\sqrt{1+x'_{n+1}(X'X)^{-1}x_{n+1}}}\right\} \sim N(0,1)}{\left.\dfrac{\hat{\sigma}}{\sigma}\right\} \sim \dfrac{\chi^2_{n-p}}{n-p}}$$

And:
- $Y_{n+1} - \hat{Y}_{n+1} = x'_{n+1}(\beta - \hat{\beta}) + \varepsilon_{n+1}$
- with $\hat{\sigma} \perp\!\!\!\perp \hat{\beta}$ and $\hat{\varepsilon}_{n+1}$

so numerator and denominator are independent

$$\sim T_{n-p}.$$

The prediction interval follows straightforwardly.

Rk: Check by yourself that for simple linear regression $Y = \beta_1 + \beta_2 x + \varepsilon$, the prediction interval writes as

$$\left[\hat{Y}_{n+1} \pm t_{1-\frac{\alpha}{2},\, n-2}\, \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right].$$

Maximum Likelihood Estimators

In the Gaussian linear model, one can make the connection between the least squares estimators $\hat{\beta}$ and $\hat{\sigma}^2$, and the MLE's $\hat{\beta}_{MLE}$ and $\hat{\sigma}^2_{MLE}$

Here, the parameter to estimate is $\theta = (\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_{>0}$. Recall that in the Gaussian setting, the variables Y_i are independent and

$$Y_i = x_i'\beta + \varepsilon_i \sim \mathcal{N}(x_i'\beta, \sigma^2).$$

As a consequence, the likelihood of the observation $Y = (Y_1, \dots, Y_n)'$ is

$$L_n(\beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - x_i'\beta)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - x_i'\beta)^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right).$$

Hence, the log-likelihood writes as

$$\ell_n(\beta, \sigma^2) = \log(L_n(\beta, \sigma^2))$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|Y - X\beta\|^2.$$

We looks for estimators $\hat{\beta}_{MLE}$ and $\hat{\sigma}^2_{MLE}$ that maximize the log-likelihood. From the previous formula, it is clear that we must minimize $\|Y - X\beta\|^2$ (which is exactly the least squares method!), so

$$\boxed{\hat{\beta}_{MLE} = \hat{\beta} = (X'X)^{-1}X'y.}$$

Once this is done we maximize $\sigma^2 \mapsto l_n(\hat{\beta}, \sigma^2)$, or equivalently solve

$$\frac{\partial l_n}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\|Y - X\hat{\beta}\|^2 = 0,$$

which yields $\boxed{\hat{\sigma}^2_{MLE} = \frac{\|Y - X\hat{\beta}\|^2}{n}}$. We notice that

$$\hat{\sigma}^2_{MLE} = \frac{n-p}{n}\hat{\sigma}^2, \quad \left(= \left(1 - \frac{p}{n}\right)\hat{\sigma}^2\right)$$

from which we deduce that $\hat{\sigma}^2_{MLE}$ is biased (since $\hat{\sigma}^2$ is not), but gets less and less biased as the number of variables p is small compared to the sample size n.