SL : LINEAR REGRESSION

In this chapter, we consider the problem of linear regression.

- Traditional approach: observations (X, Y) are assumed to arise from a parametric class of probability distributions, which is known up to some parameter θ :

$$(X, Y) \sim P_\theta \text{ , for } \theta \in \Theta$$

(Θ ← parameter space)

Specifically, we assume here that

$$\boxed{Y = \beta_0 + \tilde{\beta}^t x + \varepsilon}$$

where

- input variable x is conditioned on : fixed ! Non random !
  $x \in \mathbb{R}^d$, $x = (x_1, \dots, x_d)^t$
  d = number of predictors / features / covariates
  = dimension of the input space $\mathcal{X} = \mathbb{R}^d$
- $\beta_0 \in \mathbb{R}$, $\tilde{\beta} = (\beta_1, \dots, \beta_d)^t \in \mathbb{R}^d$ = unknown coefficients
- $\varepsilon \in \mathbb{R}$ = noise term ≡ what is left unexplained by the model

Typically, one assumes that $\varepsilon \sim \mathcal{N}(0, \sigma^2)$; where the noise variance is also unknown

- Consequences : – Parameter of interest is $\theta = (\beta_0, \tilde{\beta}, \sigma^2)$ (we want to estimate θ)
  – The conditional distribution of $Y \mid X = x$ is easily deduced from the distribution of ε : $Y \mid X = x \sim \mathcal{N}(\beta_0 + \tilde{\beta}^t x, \sigma^2)$

- Learning sample $\mathcal{L}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ (size n)
  ($x_i = (x_{i1}, \dots, x_{id})^t \in \mathbb{R}^d$ = i-th observation)

$$\Rightarrow \begin{cases} Y_1 = \beta_0 + \tilde{\beta}^t x_1 + \varepsilon_1 \\ \vdots \\ Y_n = \beta_0 + \tilde{\beta}^t x_n + \varepsilon_n \end{cases}$$

where it is assumed that the $\varepsilon_1, \dots, \varepsilon_n$ are independent :
Put $\underline{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^t \in \mathbb{R}^n$
$\underline{\underline{\Sigma}}_{\underline{\varepsilon}}$ = covariance matrix of $\underline{\varepsilon}$ = $\sigma^2 \underline{\underline{I}}_n$ ($(n \times n)$ identity matrix)
(also $E\,\underline{\varepsilon} = 0$)
($\underline{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \underline{\underline{I}}_n)$)

Matrix Notation ↓

$\underline{Y} = (Y_1, \dots, Y_n)^t \in \mathbb{R}^n$

$$\underline{\underline{X}} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{pmatrix}$$

($n \times (d+1)$ matrix)

$$\boxed{\underline{Y} = \underline{\underline{X}}\,\underline{\beta} + \underline{\varepsilon}}$$

($(n \times 1)$ $(n \times (d+1))$ $((d+1) \times 1)$ $(n \times 1)$)

$\underline{\beta} = (\beta_0, \beta_1, \dots, \beta_d)^t \in \mathbb{R}^{d+1}$
$= (\beta_0, \tilde{\beta}^t)^t$

- Challenges : ↘ Parameter estimation & its properties
  ↘ Model assessment
  ↘ Prediction

I – MAXIMUM LIKELIHOOD ESTIMATOR & ITS PROPERTIES

The log-likelihood is :

$$\ell(\theta) = \log \left\{ \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2} (y_i - \beta_0 - \tilde{\beta}^t x_i)^2 \right) \right\}$$

$$\ell(\theta) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \tilde{\beta}^t x_i)^2$$

⇒ Maximizing the log-likelihood with respect to $(\beta_0, \tilde{\beta})$ (we will deal with σ^2 later) is equivalent to minimize $\sum_{i=1}^{n}(y_i - \beta_0 - \tilde{\beta}^t x_i)^2$

$$\Leftrightarrow \min_{\beta_0, \tilde{\beta}} \sum_{i=1}^{n} \ell(y_i, \beta_0 + \tilde{\beta}^t x_i)$$

↑ square loss

$$\Leftrightarrow \min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

↑ $\mathcal{F} = \{ f \mid f(x) = \beta_0 + \tilde{\beta}^t x \quad \beta_0 \in \mathbb{R},\ \tilde{\beta} \in \mathbb{R}^d \}$

= class of linear functions

$$\Leftrightarrow \min_{f \in \mathcal{F}} \hat{R}_n(f)$$

↑ empirical risk

MLE under the linear gaussian model is equivalent to ERM with a square loss.

Remark: The quantity $\sum_{i=1}^{n}(y_i - \beta_0 - \tilde{\beta}^t x_i)^2$ is also known as the Residual Sum of Squares in the litterature (RSS):

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \tilde{\beta}^t x_i)^2$$
$$= (y - \underline{\underline{X}}\underline{\beta})^t (y - \underline{\underline{X}}\beta)$$
$$= \| y - X\beta \|^2$$

↳ Optimization problem is "Least Squares"

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \| y - X\beta \|^2$$

Assumption: rank $\underline{\underline{X}} = d+1$

↳ In other words, the columns of $\underline{\underline{X}}$ are linearly independent

- Minimization of the RSS:

$$\begin{cases} \dfrac{\partial RSS}{\partial \beta} = -2y^t X + 2\hat{\beta}^t X^t X = 0 & \text{to find the minimum} \\ \dfrac{\partial^2 RSS}{\partial \beta^2} = 2X^t X & \text{to check if the value of } \beta \text{ obtained by setting the first derivative of the RSS to zero is indeed a minimum.} \end{cases}$$

↑ aka the HESSIAN

Toolbox:

$$\frac{\partial y^t A \beta}{\partial \beta} = y^t A$$

$$\frac{\partial \beta^t A \beta}{\partial \beta} = \beta^t (A + A^t)$$

y, β = vectors, A = matrix of appropriate dim

And indeed,

$$RSS = (y - X\beta)^t (y - X\beta)$$
$$= y^t y - y^t X\beta - \underbrace{\beta^t X^t y}_{= y^t X \beta \in \mathbb{R}^t \text{ (so equal to its transpose)}} + \beta^t X^t X \beta$$
$$= y^t y - 2y^t X\beta + \beta^t \underbrace{X^t X}_{\text{symmetric matrix}} \beta$$

↳ $$\frac{\partial RSS}{\partial \beta} = -2y^t X + 2\beta^t X^t X.$$

Solution is such that $\hat{\beta}^t X^t X = y^t X$

$$X^t X \hat{\beta} = X^t y$$

Is $X^t X$ invertible ?

Under the assumption that rank $X = d+1$, the answer is yes !

Indeed, suppose that there exists a $\beta \in \mathbb{R}^{d+1}$, $\beta \neq 0$ such that $(X^t X)\beta = 0$ [i.e. suppose $X^t X$ not invertible]

Then $\beta^t (X^t X) \beta = 0$

$\| X\beta \|^2$

which implies that $X\beta = 0$; which contradicts the assumption that rank $X = (d+1)$.

$\Rightarrow$ The matrix $X^t X$ is positive definite.

$\Rightarrow$ The Hessian is positive definite $\Rightarrow$ the solution corresponds to a minimum indeed

$\Rightarrow$ The least squares (LS) estimate is thus

$$\boxed{\hat{\beta} = (X^t X)^{-1} X^t y}$$

We get the estimate $\hat{y} = X\hat{\beta}$,

$$\boxed{\hat{y} = X(X^t X)^{-1} X^t y =: \underset{(n \times n)}{H} y}$$

H = the 'hat' matrix : it puts a "hat" on y.

Claim: The least square solution corresponds to the orthogonal projection of $y = (y_1, \dots, y_n)^t$ onto the plane spanned by $(\underline{1}, \underline{x}_1, \dots, \underline{x}_d)$

↑ aka the column space of $\underline{\underline{X}} = \begin{bmatrix} | & | & & | \\ \underline{1} & \underline{x}_1 & \cdots & \underline{x}_d \\ | & | & & | \end{bmatrix}$,

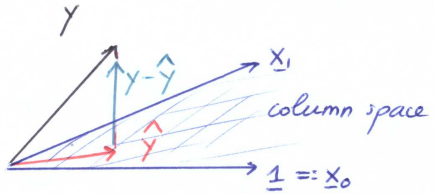
commonly denoted $\mathcal{C}(\underline{\underline{X}})$.

$\underline{x}_j \in \mathbb{R}^n$ = j-th covariate.

Indeed, let's show that $\langle \hat{y} - y, \underline{x}_j \rangle = 0$

$\langle \cdot, \cdot \rangle$ = inner product in a Euclidean space.

Recall that for $u, v \in \mathbb{R}^n$, $\langle u, v \rangle = u^t v$.



$$\langle \hat{y} - y, x_j \rangle = (\hat{y} - y)^t x_j$$

$$= \left(X(X^t X)^{-1} X^t y - y \right)^t x_j$$

$$= y^t \left(X(X^t X)^{-1} X^t - I_n \right)^t x_j$$

Next, $\underbrace{X(X^t X)^{-1} X^t}_{n \times (d+1)} \underbrace{\underline{x}_j}_{(n \times 1)} = \underbrace{\left[X(X^t X)^{-1} X^t X \right]_j}_{n \times n \text{ matrix}}$ ($\equiv j$-th column)

$n \times 1$ vector

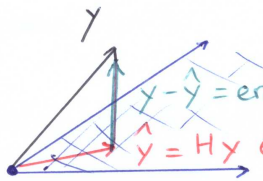
$$= [X]_j = x_j$$

$\Rightarrow \langle \hat{y} - y, x_j \rangle = 0 \iff \hat{y} - y \perp x_j$

i.e. The hat matrix H projects y onto the linear subspace spanned by the columns of $\underline{\underline{X}}$ H = projection matrix.

Toolbox: Facts about projection matrices. (7)

- A square matrix P such that $P^2 = P$ [called IDEMPOTENT] is a projection.
  If in addition P is symmetric, then it corresponds to an orthogonal projection.
  It is clear from its definition that $H = X(X^tX)^{-1}X^t$ is idempotent and symmetric.
  Then $(I-P)$ is also idempotent and symmetric:
  $(I-P)^2 = I - 2P + P^2 = I - 2P + P = I - P$
  $\Rightarrow (I-P)$ is also an orthogonal projection.



$\mathcal{C}(X)$ = column space of X; of dimension $(d+1)$
$\mathcal{C}(X)^t$ = residual space; $\perp$ to $\mathcal{C}(X)$
↳ We have the decomposition $y = y - \hat{y} + \hat{y}$
$= \underbrace{(I-H)y}_{\in \mathcal{C}(X)^t} + \underbrace{Hy}_{\in \mathcal{C}(X)}$
↑ orthogonal vectors

Fact:
True in general: we write $\mathbb{R}^n \ni y = \mathcal{C}(X) \oplus \mathcal{C}(X)^t$
↑ ie. intersection is $\{0\}$, and vectors from $\mathcal{C}(X)$ and $\mathcal{C}(X)^t$ are orthogonal

- The rank of an idempotent matrix is equal to its trace: (8)
  rank P = trace P.

  rank H = $d+1$ = trace H = $\dim\{\mathcal{C}(X)\}$
  rank $(I-H)$ = Tr $(I-H)$ = Tr I − Tr H = $n-d-1$
  $= \dim\{\mathcal{C}(X)^t\}$

- The eigenvalues of a projection matrix are equal to 0 or 1:
  Let λ = eigenvalue of P associated with x: $Px = \lambda x$.
  Then $P^2x = P(Px) = \lambda Px = \lambda^2 x$
  $\Rightarrow P^2x = \lambda x \Rightarrow \lambda^2 = \lambda \Rightarrow \lambda = 0$ or 1

end of Toolbox ■

Interpretation of the LS solution.

We have $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \underline{x}_1 + \cdots + \hat{\beta}_d \underline{x}_d \in \mathbb{R}^n$
Coefficient $\hat{\beta}_j$ = coordinate of $\hat{y}$ in the base $(\underline{1}, \underline{x}_1, \cdots, \underline{x}_d)$ of the column space of $\underline{\underline{X}}$.

OCTopo *H10

> $\hat{\beta}_j$ are not the coordinates of the projection of y onto x_j. This is true only if the base $(\underline{1}, \underline{x}_1, \cdots, \underline{x}_d)$ is orthogonal. Look:

- Consider =
  $\underset{(n\times 1)}{\underline{y}} = \underset{(n\times d)}{\underline{\underline{X}}} \underset{(d\times 1)}{\underline{\beta}} + \underset{(n\times 1)}{\underline{\varepsilon}}$, $\underline{\underline{X}} = (\underline{1}\ \underline{x}_0 \cdots \underline{x}_d)$
  The LS fit is $\hat{\beta} = (X^tX)^{-1}X^ty$.
  Then
  $X^ty = \underset{(d\times n)}{\begin{pmatrix} 1 & 1 \\ x_0 & \cdots & x_d \\ 1 & 1 \end{pmatrix}^t} \underset{(n\times 1)}{y} = \underset{(d\times 1)}{\begin{pmatrix} x_0^t y \\ \vdots \\ x_d^t y \end{pmatrix}} = \begin{pmatrix} \langle x_0, y \rangle \\ \vdots \\ \langle x_d, y \rangle \end{pmatrix}$

$$X^t X = \begin{pmatrix} - & x_0^t & - \\ & \vdots & \\ - & x_d^t & - \end{pmatrix} \begin{pmatrix} | & & | \\ x_0 & \cdots & x_d \\ | & & | \end{pmatrix} = \begin{pmatrix} \langle x_0, x_0 \rangle & \cdots & \langle x_0, x_d \rangle \\ \vdots & & \\ \langle x_d, x_0 \rangle & \cdots & \langle x_d, x_d \rangle \end{pmatrix}$$

If the columns of X are orthogonal, then $\langle x_i, x_j \rangle = 0$ for $i \neq j$; and the matrix $X^t X$ is diagonal:

$$(X^t X)^{-1} = \begin{pmatrix} 1/\langle x_0, x_0 \rangle & & 0 \\ & \ddots & \\ 0 & & 1/\langle x_d, x_d \rangle \end{pmatrix}$$

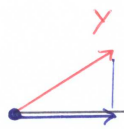
We obtain

$$\hat{\beta} = (X^t X)^{-1} X^t y = \begin{pmatrix} \vdots \\ \dfrac{\langle x_j, y \rangle}{\langle x_j, x_j \rangle} \\ \vdots \end{pmatrix} \Rightarrow \boxed{\hat{\beta}_j = \frac{\langle x_j, y \rangle}{\langle x_j, x_j \rangle}}$$

$\hat{\beta}_j$ = coordinate of the projection of y onto x_j.

= contribution of input j to output y, independently of other inputs.

Recall:



$$\text{proj}_{x_j} y = \frac{\langle x_j, y \rangle}{\langle x_j, x_j \rangle} x_j$$

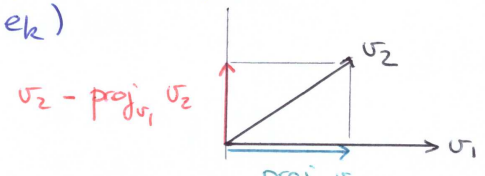
But then, what is the meaning of $\hat{\beta}_j$ when the columns of $\underline{\underline{X}}$ are not orthogonal?

→ Use Gram-Schmidt orthogonalization procedure

Toolbox: Consider $\mathcal{M} = \text{span}\{v_1, \cdots, v_k\}$, where $v_1, \cdots, v_k$ are not orthogonal

What does "span" mean? Well, it is the set of vectors that can be expressed as a linear combination of $v_1, \cdots, v_k$.

The goal is to derive a new basis for $\mathcal{M}$, which is orthogonal, or, even better, orthonormal. We will call it $(e_1, \cdots, e_k)$



$$= \frac{\langle v_2, v_1 \rangle}{\langle v_1, v_1 \rangle} v_1$$

Then

$$\begin{aligned} u_1 &= v_1 & e_1 &= u_1 / \|u_1\| \\ u_2 &= v_2 - \text{proj}_{u_1} v_2 & e_2 &= u_2 / \|u_2\| \\ u_3 &= v_3 - \text{proj}_{u_1} v_3 - \text{proj}_{u_2} v_3 & e_3 &= \frac{u_3}{\|u_3\|} \\ &\vdots & & \\ u_k &= v_k - \sum_{j=1}^{k-1} \text{proj}_{u_j} v_k & e_k &= \frac{u_k}{\|u_k\|} \end{aligned}$$

Back to our linear regression problem. To fix ideas, let's consider the case of a single predictor with intercept:

$\underline{\underline{X}} = (\underline{1}, \underline{x})$ (n×2)

$$\rightarrow X^t X = \begin{pmatrix} 1^t \\ x^t \end{pmatrix} (1 \; x) = \begin{pmatrix} n & \Sigma x_i \\ \Sigma x_i & \langle x, x \rangle \end{pmatrix}$$

$$\rightarrow (X^t X)^{-1} = \frac{1}{n\langle x, x \rangle - (\Sigma x_i)^2} \begin{pmatrix} \langle x, x \rangle & -\Sigma x_i \\ -\Sigma x_i & n \end{pmatrix}$$

$$\rightarrow X^t y = \begin{pmatrix} 1^t \\ x^t \end{pmatrix} y = \begin{pmatrix} \Sigma y_i \\ \langle x, y \rangle \end{pmatrix}$$

$$\Rightarrow \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X^t X)^{-1} X^t y$$

We get

$$\hat{\beta}_1 = \frac{1}{n\langle x,x\rangle - (\Sigma x_i)^2}\left\{-(\Sigma x_i)(\Sigma y_i) + n\langle x,y\rangle\right\}$$

$$= \frac{\langle x,y\rangle - \bar{x}(\Sigma y_i)}{\langle x,x\rangle - n^{-1}(\Sigma x_i)^2} \qquad \bar{x} = \frac{1}{n}\Sigma x_i$$

$$= \frac{\langle \underline{x},\underline{y}\rangle - \langle \bar{x}\underline{1},\underline{y}\rangle}{\langle \underline{x},\underline{x}\rangle - n(\bar{x})^2}$$

Next, since $\langle x-\bar{x}1, x-\bar{x}1\rangle = \langle x,x\rangle - \bar{x}\langle 1,x\rangle - \bar{x}\langle x,1\rangle + (\bar{x})^2\langle 1,1\rangle$

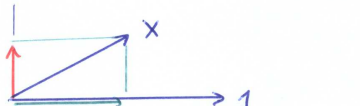
$= \langle x,x\rangle - n(\bar{x})^2$,

we obtain

$$\hat{\beta}_1 = \frac{\langle x-\bar{x}1,\ y\rangle}{\langle x-\bar{x}1,\ x-\bar{x}1\rangle}$$

$\equiv$ Regress y on residual $z_1 = x - \bar{x}1$ to get coefficient $\hat{\beta}_1$.

$z_1 := x - \text{proj}_1 x$ (residual)



$\text{proj}_1 x = \frac{\langle x,1\rangle}{\langle 1,1\rangle} 1 = \bar{x}1$

$\hat{\beta}_1 \equiv$ added contribution coming from x ("what is new") after x has been adjusted from 1.

More generally:

GRAM-SCHMIDT

(i) Initialize $z_0 := x_0 := 1$

(ii) For $j = 1, \cdots, d$

- Regress x_j on $z_0, \cdots, z_{j-1}$ to produce coef:
  $\hat{\gamma}_{\ell j} = \langle x_j, z_\ell\rangle / \langle z_\ell, z_\ell\rangle$, $\ell = 0, \cdots, j-1$
- and $z_j = x_j - \sum_{\ell=0}^{j-1} \hat{\gamma}_{\ell j} z_\ell$ = residual vector

(iii) Regress y on residual z_d to get $\hat{\beta}_d$

Remark: the z_j do not have unit length in general.

In words: the regression coefficient $\hat{\beta}_j$ represents the additional contribution of x_j on y, after x_j has been adjusted for $x_0, \cdots, x_{j-1}, x_{j+1}, \cdots, x_d$ (what is left in x_j is non-informative since redundant with the information contained in the remaining covariates)

$\Rightarrow$ Linear model is highly interpretable.

- The matrix form of the Gram-Schmidt orthogonalization procedure yields the QR decomposition of X:

$$X = Z\Gamma$$

$n \times (d+1)$ $\quad$ $n \times (d+1)$ $\quad$ $(d+1) \times (d+1)$

- Z has orthogonal columns & the column space spanned by the first ℓ columns of X and Z are the same.

$$Z = \begin{pmatrix} | & | & & | \\ z_0 & z_1 & \cdots & z_d \\ | & | & & | \end{pmatrix}$$

- Γ is upper triangular: rewrite step (ii) in the GS procedure page 11 as

$$x_j = z_j + \sum_{\ell=0}^{j-1} \hat{\gamma}_{\ell j} z_\ell$$

$$\Gamma = \begin{pmatrix} 1 & \hat{\gamma}_{01} & \hat{\gamma}_{02} & \hat{\gamma}_{03} & \cdots \\ & 1 & \hat{\gamma}_{12} & \hat{\gamma}_{13} & \cdots \\ & & 1 & \hat{\gamma}_{23} & \cdots \\ & 0 & & 1 & \\ & & & & \ddots \end{pmatrix}$$

- Put $D = \begin{pmatrix} \sqrt{\langle z_0, z_0\rangle} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\langle z_d, z_d\rangle} \end{pmatrix}$ $(d+1) \times (d+1)$ = diagonal matrix

Put $Q = Z D^{-1}$ ($n \times (d+1)$ · $n \times (d+1)$ · $(d+1) \times (d+1)$) and $R = D\Gamma$ ($(d+1) \times (d+1)$ · $(d+1) \times (d+1)$); R is upper triangular

Why?

Because $Q^t Q = D^{-1} Z^t Z D^{-1} = D^{-1} \begin{pmatrix} \langle z_0, z_0 \rangle & & 0 \\ & \ddots & \\ 0 & & \langle z_d, z_d \rangle \end{pmatrix} D^{-1}$

$= I_{d+1}$

$\Rightarrow$ Q has orthogonal columns that form a basis for the column space of X.

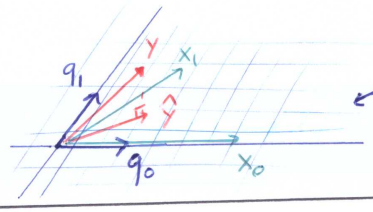
We write $X = QR$ ($n \times (d+1)$ = $n \times (d+1)$ · $(d+1) \times (d+1)$) "the QR decomposition of X".

This decomposition is very useful. Let $Q = (q_0, q_1, \dots, q_d)$. The LS solution is the projection of y onto the column space of X. The solution can be represented in an orthonormal basis $(q_0, \dots, q_d)$ for $\mathcal{C}(X)$ since

$$\rightarrow (X^t X)^{-1} X^t y = R^{-1} (R^t)^{-1} R^t Q^t y = R^{-1} Q^t y$$

($X^t X = R^t \underbrace{Q^t Q}_{I} R = R^t R$)

$$\rightarrow \hat{y} = X (X^t X)^{-1} X^t y = (QR)(R^{-1} Q^t y) = QQ^t y = \sum_{j=0}^{d} \langle y, q_j \rangle q_j$$



← x_0 and x_1 span the same plane as q_0 and q_1, this should be clear.

Remark: Alternatively, you may use an SVD decomposition of $X = U \Lambda V^t$ ($n \times (d+1)$ = $n \times (d+1)$ · $(d+1) \times (d+1)$ · $(d+1) \times n$), rank $X = d+1$, where the columns of U provide an orthonormal basis for $\mathcal{C}(X)$. The LS solution can then be expressed as $\hat{y} = \sum \langle y, u_j \rangle u_j$, where $u_j = j$-th column of U.

Note that the orthonormal basis provided by U and Q are not necessarily the same ones.

Further geometrical considerations.

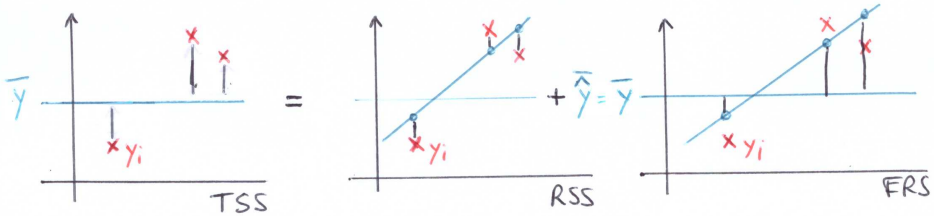
If the intercept is included in the model, we have the decomposition:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{\hat{y}})^2$$

Total Sum of Squares (TSS) = Residual Sum of Squares (RSS) + Explained Sum of Squares (ESS)

$$\| y - \bar{y} 1 \|^2 = \| y - \hat{y} \|^2 + \| \hat{y} - \bar{\hat{y}} 1 \|^2$$

(where $\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$)



Note that if the intercept is in the model, then $\bar{y} = \bar{\hat{y}}$.

Indeed, making use of the first equation in $(X^t X)\hat{\beta} = X^t y$,

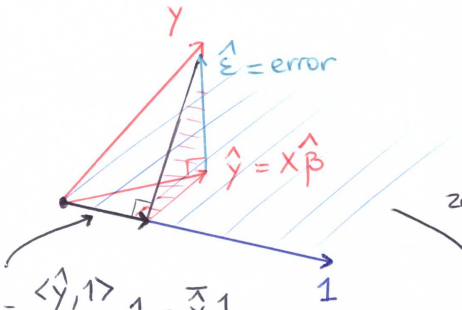
rewritten $X^t(y - X\hat{\beta}) = 0$, we obtain

$$\sum_{i=1}^{n} y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_d x_{id}) = 0$$

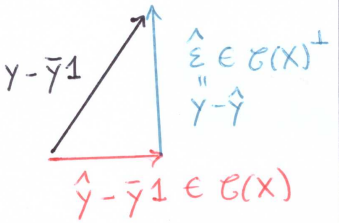
$$\Leftrightarrow \sum_{i=1}^{n} y_i - \hat{y}_i = 0$$

$$\Leftrightarrow \bar{y} = \bar{\hat{y}}. \quad \blacksquare$$

The decomposition $TSS = ESS + RSS$ can be derived by direct computation, or can be visualized geometrically as follows:



zoom on the triangle with green / black / red sides:



$$\text{proj}_1 \hat{y} = \frac{\langle \hat{y}, 1 \rangle}{\langle 1, 1 \rangle} 1 = \bar{\hat{y}} 1$$

$$\text{proj}_1 y = \frac{\langle y, 1 \rangle}{\langle 1, 1 \rangle} 1 = \bar{y} 1$$

Since $\bar{\hat{y}} = \bar{y}$, we conclude that $\text{proj}_1 \hat{y} = \text{proj}_1 y$.

Pythagoras theorem yiels indeed

$$\| y - \bar{y}1 \|^2 = \| y - \hat{y} \|^2 + \| \hat{y} - \bar{y}1 \|^2$$

$\Rightarrow$ Introduce $\boxed{R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}}$

Interpretation: R^2 represents the proportion of variability in y that is explained using the linear model.

$\approx$ measure of linear association between X and Y. No wonder that for simple linear regression ($d=1$), we have $R^2 = \hat{\rho}^2$,

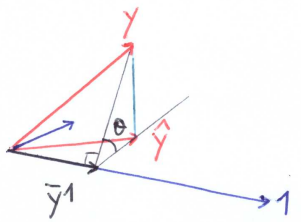
where $\hat{\rho}$ denotes the empirical correlation coefficient between the predictor and the response variable:

$$\hat{\rho} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Geometrically, $R^2 = \cos^2 \theta = \frac{\| \hat{y} - \bar{y}1 \|^2}{\| y - \bar{y}1 \|^2} = \frac{ESS}{TSS}$;

θ = angle between y and $\hat{y}$, taken at $\bar{y}1$.

↳ When $\theta = 0$, $R^2 = 1$, and all the variance is explained: $y \in \mathcal{C}(X)$; so that y is exactly a linear combination of the covariates



Remarks: (i) R^2 does not necessarily lie between 0 and 1 if you do not include the constant in the model.

(ii) R^2 is not comparable across different response variables ys

Eg: In one model consider y, and in another $\ln y$. Since the scale is changed, the sum of squares as well, and R^2 is affected.

(iii) R^2 almost always increases when you add a new predictor (proof below)

Consequence: you might be tempted to add more and more variables into the model.

Alternative: Use the adjusted R^2 square,

$$\boxed{R^2_a = \frac{\| \hat{\varepsilon} \|^2 / (n-d-1)}{\| y - \bar{y}1 \| / (n-1)} = 1 - \frac{n-1}{n-d-1} \frac{RSS}{TSS}}$$

↳ As the number of features increases, the RSS decreases, but the adjustment term $n-1/(n-d-1)$ increases.

- Proof that R^2 (almost) always increases when adding a new predictor.

Step I: Partitioned Regression.

Suppose that the regression model can be written

→ $y = X\beta + \varepsilon$

$= X_1\beta_1 + X_2\beta_2 + \varepsilon$

$= (X_1 \mid X_2)\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + \varepsilon$

sets of variables ; β_1/β_2 = vectors

→ LS solution satisfies $(X^tX)\beta = X^ty$, with

$$X^tX = \begin{pmatrix} X_1^tX_1 & X_1^tX_2 \\ X_2^tX_1 & X_2^tX_2 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} X_1^tX_1 & X_1^tX_2 \\ X_2^tX_1 & X_2^tX_2 \end{pmatrix}\begin{pmatrix}\hat\beta_1\\ \hat\beta_2\end{pmatrix} = \begin{pmatrix} X_1^ty \\ X_2^ty\end{pmatrix} \quad \text{——— (A)}$$

→ Solve for $\hat\beta_1$: $(X_1^tX_1)\hat\beta_1 + (X_1^tX_2)\hat\beta_2 = X_1^ty$

$$\hat\beta_1 = (X_1^tX_1)^{-1}X_1^ty - (X_1^tX_1)^{-1}X_1^tX_2\hat\beta_2$$

$$\hat\beta_1 = (X_1^tX_1)^{-1}X_1^t(y - X_2\hat\beta_2) \quad \text{——— (B)}$$

Note that if $X_1 \perp X_2$, $X_1^tX_2 = 0$ and the estimated coefficient $\hat\beta_1$ remains the same, whether the variable X_2 is included in the model or not. This is not true if X_1 and X_2 are not orthogonal.

→ Substitute expression (B) for $\hat\beta_1$ back into the second equation in (A):

$$(X_2^tX_1)\hat\beta_1 + (X_2^tX_2)\hat\beta_2 = X_2^ty$$

$$(X_2^tX_1)\left\{(X_1^tX_1)^{-1}X_1^ty - (X_1^tX_1)^{-1}X_1^tX_2\hat\beta_2\right\} + (X_2^tX_2)\hat\beta_2 = X_2^ty$$

$$\left\{X_2^tX_2 - X_2^tX_1(X_1^tX_1)^{-1}X_1^tX_2\right\}\hat\beta_2 = X_2^ty - (X_2^tX_1)(X_1^tX_1)^{-1}X_1^ty$$

$$X_2^t\left\{I - \underbrace{X_1(X_1^tX_1)^{-1}X_1^t}_{H_1}\right\}X_2\hat\beta_2 = X_2^t\left\{I - X_1(X_1^tX_1)^{-1}X_1^t\right\}y$$

H_1: projection matrix onto $\mathcal{C}(X_1)$.

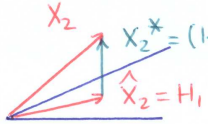
→ $\hat\beta_2 = \left(X_2^t(I-H_1)X_2\right)^{-1}\left(X_2^t(I-H_1)y\right)$

Since $(I-H_1)$ is idempotent, introducing variables

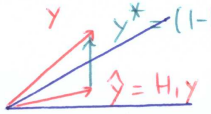
$$\begin{cases} X_2^* = (I-H_1)X_2 \\ y^* = (I-H_1)y \end{cases}$$

yields $\hat\beta_2 = (X_2^{*t}X_2^*)^{-1}X_2^{*t}y^*$ ——— (*)

→ Interpretation:



$X_2^* = (I-H_1)X_2$ = vector of residuals from regressing X_2 on X_1



$y^* = (I-H_1)y$ = vector of residuals from regressing y on X_1

⇒ To estimate $\hat{\beta}_2$, first regress X_2 and y on X_1, and then regress residuals y^* on residuals X_2^*.

Step II : R^2 and the addition of a new variable.

→ First, start with the model $y = X\beta + \varepsilon$, and the fit $y = X\hat{\beta} + \hat{\varepsilon}$
($\hat{\varepsilon}$ ↑ vector of residuals $\hat{\varepsilon} = (1-H)y$; $\hat{\beta}$ ↑ Least square fit $(X^tX)^{-1}X^ty$)

Next, add a single extra variable z, so that

$$y = X\hat{\beta}' + \hat{\gamma} z + \hat{\varepsilon}'$$

($\hat{\beta}'$, $\hat{\gamma}$ ↑ LS coefficients; $\hat{\varepsilon}'$ ↑ new vector of residuals)

Unless $z \perp X$, $\hat{\beta}$ and $\hat{\beta}'$ differ; see the remark bottom of page 17.

→ The RSS in the original model is $\|\hat{\varepsilon}\|^2$, and in the second model $\|\hat{\varepsilon}'\|^2$ ⇒ the R^2 will increase if $\|\hat{\varepsilon}'\|^2 < \|\hat{\varepsilon}\|^2$.

→ LS solution $(\hat{\beta}', \hat{\gamma})$ satisfies

$$\begin{pmatrix} X^tX & X^tz \\ z^tX & z^tz \end{pmatrix} \begin{pmatrix} \hat{\beta}' \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} X^ty \\ z^ty \end{pmatrix}$$

↳ Proceeding a in Step I, we see that the first line yield the estimate

$$\hat{\beta}' = (X^tX)^{-1}X^t(y - \hat{\gamma} z)$$

(compare with expression (B) page 17)

$$\boxed{\hat{\beta}' = \hat{\beta} - \hat{\gamma}(X^tX)^{-1}X^tz} \quad \text{(C)}$$

→ Moreover, $\hat{\varepsilon}' = y - X\hat{\beta}' - \hat{\gamma} z$
↑ Insert expression (C) here.

We get $\hat{\varepsilon}' = y - X\hat{\beta} + X(X^tX)^{-1}X^t(\hat{\gamma} z) - (\hat{\gamma} z)$

$= y - X\hat{\beta} - (I - \underbrace{X(X^tX)^{-1}X^t}_{=: H})(\hat{\gamma} z)$

$= \hat{\varepsilon} - (1-H)(\hat{\gamma} z)$

Put
$$\boxed{\begin{aligned} z^* &= (1-H)z \\ \hat{\varepsilon}' &= \hat{\varepsilon} - \hat{\gamma} z^* \end{aligned}}$$

→ From the last written expression, we conclude that

$$\boxed{\|\hat{\varepsilon}'\|^2 = \|\hat{\varepsilon}\|^2 + \hat{\gamma}^2\|z^*\|^2 - 2\hat{\gamma}\hat{\varepsilon}^t z^*} \quad \text{(D)}$$

We turn our attention to this term

→ $\hat{\varepsilon} = (1-H)y = y^*$ = residual from regressing y on X
$z^* = (1-H)z$ = residual from regressing z on X

⇒ Applying formula (*) gives

$$\boxed{\hat{\gamma} = (z^{*t}z^*)^{-1}(z^{*t}y^*)}$$

(with $X_1 = X$ and $X_2 = z$, in the notation of page 18)

⇒ It follows that $(z^{*t}z^*)\hat{\gamma} = z^{*t}y^*$
$= z^{*t}\hat{\varepsilon}$
$= \hat{\varepsilon}^t z^*$

↑ Plug this expression of $\hat{\varepsilon}^t z^*$ back into (D):

$$\|\hat{\varepsilon}'\|^2 = \|\hat{\varepsilon}\|^2 + \hat{\gamma}^2\|z^*\|^2 - 2\hat{\gamma}^2\|z^*\|^2$$

$$\boxed{\|\hat{\varepsilon}'\|^2 = \|\hat{\varepsilon}\|^2 - \hat{\gamma}^2\|z^*\|^2}$$

⇒ $\|\hat{\varepsilon}'\|^2 < \|\hat{\varepsilon}\|^2$, unless $\hat{\gamma} = 0$

⇒ Unless the LS on the additional variable is zero, adding an extra variable will decrease R^2. ∎

Properties of the LS solution.

(1) $\hat\beta$ is unbiased for β.

Indeed, since $\hat\beta = (X^tX)^{-1}X^ty$, we have

$$\mathbb{E}\hat\beta = (X^tX)^{-1}X^t\underbrace{\mathbb{E}y}_{=X\beta \text{ since } y=X\beta+\varepsilon} = \beta, \text{ as required}$$

Remarks (i) Unbiasedness of $\hat\beta$ holds independently of the distribution of ε, provided $\mathbb{E}(\varepsilon|X)=0$ and $\mathbb{E}(\varepsilon\varepsilon^t|X) = \sigma^2 I$.
⇒ Normal distribution is not needed.

(ii) In a random design (X random), $\hat\beta$ is still unbiased since

$$\mathbb{E}\hat\beta = \mathbb{E}_X\, \mathbb{E}\hat\beta|X = \mathbb{E}_X\beta = \beta.$$

(2) The covariance matrix of $\hat\beta$ is $\Sigma_{\hat\beta} = \mathbb{E}(\hat\beta-\beta)(\hat\beta-\beta)^t$, with

$$\hat\beta-\beta = (X^tX)^{-1}X^ty - \beta = (X^tX)^{-1}X^t(y-X\beta),$$

so that

$$\Sigma_{\hat\beta} = (X^tX)^{-1}X^t\underbrace{\mathbb{E}\{(y-X\beta)(y-X\beta)^t\}}_{=\sigma^2 I_n} \times (X^tX)^{-1}$$

$$\boxed{\Sigma_{\hat\beta} = \sigma^2(X^tX)^{-1}}$$

Remark: In a random design, use

$$\Sigma_{\hat\beta} = \mathbb{E}_X\left(\Sigma_{\hat\beta|X}\right) + \underbrace{\Sigma_{\mathbb{E}(\hat\beta|X)}}_{=0} = \sigma^2\mathbb{E}(X^tX)^{-1}.$$

(3) $\hat\sigma^2 := \frac{1}{n-d-1}\sum_{i=1}^n (y_i-\hat y_i)^2$ is unbiased for σ^2.

Preliminary calculations:

- $\sum(y_i-\hat y_i)^2 = \|y-\hat y\|^2 = \|y-X\hat\beta\|^2 = \|\hat\varepsilon\|^2$

  where we defined $\hat\varepsilon = y - X\hat\beta = (I-H)y$ = vector of residuals

- $\mathbb{E}\hat\varepsilon = \mathbb{E}(I-H)y = (I-H)\mathbb{E}y = 0$ since $\mathbb{E}y \in \mathcal{C}(X)$

- Since $y = \underbrace{X\beta}_{\in\mathcal{C}(X)} + \varepsilon \Rightarrow \underbrace{(I-H)}_{\text{projects onto } \mathcal{C}(X)^\perp} y = (I-H)\varepsilon$

Thus

$$\begin{aligned}\Sigma_{\hat\varepsilon} &= \mathbb{E}(\hat\varepsilon\,\hat\varepsilon^t) \quad \text{since } \mathbb{E}\hat\varepsilon = 0\\ &= \mathbb{E}\left[(I-H)y\right]\left[y^t(I-H)\right]\\ &= \mathbb{E}\left[(I-H)\varepsilon\right]\left[\varepsilon^t(I-H)\right] \quad \text{since } (I-H)y=(I-H)\varepsilon\\ &= (I-H)\underbrace{\mathbb{E}(\varepsilon\varepsilon^t)}_{=\sigma^2 I_n}(I-H)\\ &= \sigma^2(I-H) \quad \text{since } (I-H) \text{ is idempotent.}\end{aligned}$$

Back to $\mathbb{E}\sum(y_i-\hat y_i)^2 = \mathbb{E}\|\hat\varepsilon\|^2$ (since a scalar is equal to its trace)

$$\begin{aligned}&= \mathbb{E}\,\mathrm{Tr}\,\|\hat\varepsilon\|^2\\ &= \mathbb{E}\,\mathrm{Tr}(\hat\varepsilon^t\hat\varepsilon)\\ &= \mathbb{E}\,\mathrm{Tr}(\hat\varepsilon\,\hat\varepsilon^t)\\ &= \mathrm{Tr}\,\mathbb{E}(\hat\varepsilon\,\hat\varepsilon^t)\\ &= \mathrm{Tr}\,\Sigma_{\hat\varepsilon}\\ &= \sigma^2\,\mathrm{Tr}(I-H)\\ &= \sigma^2\,\mathrm{Rk}(I-H) = \sigma^2(n-d-1)\end{aligned}$$

Since $\forall$ matrice $\mathrm{Tr}(AA^t) = \mathrm{Tr}(A^tA)$

Remarks: (i) $\hat{\beta} = (X^tX)^{-1}X^ty$ is called a LINEAR estimator, in the sense that it is computed as a linear combination of the entries in y.

Any estimator of the form By is called a linear estimator (of β).

$\hat{\beta}$ satisfies an interesting property:

> **Theorem:** (GAUSS-MARKOV THEOREM)
> $\hat{\beta}$ has minimum variance amongst all unbiased linear estimators of β

OK, $\hat{\beta}$ is a vector; so what do we mean by minimum variance?

We mean that if → $\hat{\beta}$ has covariance matrix $\Sigma_{\hat{\beta}}$

→ $\tilde{\beta} = By$ is another estimator of β such that $E\tilde{\beta} = \beta$; with covariance matrix $\Sigma_{\tilde{\beta}}$,

Then $\Sigma_{\tilde{\beta}} \geq \Sigma_{\hat{\beta}}$; in the sense that $\Sigma_{\tilde{\beta}} - \Sigma_{\hat{\beta}}$ is a positive semi-definite matrix.

(ii) Since $\hat{\beta}$ is a linear estimator and y (conditionally on $X = x$) is normally distributed, it follows that $\hat{\beta}$ is normally distributed and

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

(iii) Unbiased estimator $\hat{\sigma}^2$ of σ^2 defined on page 15 does not correspond to the maximum likelihood estimator of σ^2. → biased estimator

MLE is $s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

(log lik. given on page 3)

(iv) What about the distribution of $\hat{\sigma}^2$?

To answer this question, we make use of COCHRAN theorem; which stipulates that the decomposition of a multivariate Gaussian vector onto orthogonal subspaces produces independant random variables, whose laws are explicit.

> **Theorem** = (COCHRAN)
> Let $\cdot\ Z \sim \mathcal{N}(\mu, \sigma^2 I_n)$
> $\cdot\ \mathcal{M}$ = a subspace of $\mathbb{R}^n$ of dimension p
> $\cdot\ P$ = matrix of orthogonal projection on $\mathcal{M}$
> $\cdot\ P_\perp$ = matrix of orthogonal projection on $\mathcal{M}^\perp$, where $\mathcal{M}^\perp$ = orthogonal complement of $\mathcal{M}$
>
> Then
> (a) $PZ \sim \mathcal{N}(P\mu, \sigma^2 P)$ & $P_\perp Z \sim \mathcal{N}(P_\perp\mu, \sigma^2 P_\perp)$
> (b) Vectors PZ and $P_\perp Z = (I-P)Z$ are indpt.
> (c) $\frac{1}{\sigma^2}\|P(Z-\mu)\|^2 \sim \chi^2_p$
> and $\frac{1}{\sigma^2}\|P_\perp(Z-\mu)\|^2 \sim \chi^2_{n-p}$

Consequences:

→ $\hat{\beta}$ and $\hat{\sigma}^2$ are independent since applying (b):

$\cdot\ \hat{\beta} = (X^tX)^{-1}X^ty$
$= (X^tX)^{-1}X^t(X(X^tX)^{-1}X^t)y$
$= (X^tX)^{-1}X^tHy$
$=$ function of Hy

$\cdot\ \hat{\sigma}^2 = \frac{1}{n-d-1}\|\hat{\varepsilon}\|^2 = \frac{1}{n-d-1}\|(I-H)y\|^2$
$=$ function of $(I-H)y$

↘ $\frac{(n-d-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-d-1}$ since, in the notation of the theorem,

with $Z = \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, we have that

$$\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-d-1} = \frac{1}{n-d-1}\|(I-H)\varepsilon\|^2$$

$$= \frac{1}{n-d-1}\|H_\perp(\varepsilon - E\varepsilon)\|^2 \quad (E\varepsilon = 0)$$

so that

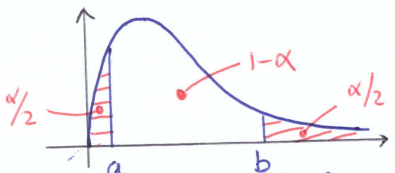
$$\frac{(n-d-1)\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\|H_\perp(\varepsilon - E\varepsilon)\|^2 \sim \chi^2_{n-d-1}$$

since $H_\perp = I-H$ projects onto a subspace of dimension $n-d-1$.

- Since the distributions of $\hat{\beta}$ and $\hat{\sigma}^2$ are known, we can make good use of them to construct confidence intervals:

↘ For σ, we have that $P\left(a \leq \frac{(n-d-1)\hat{\sigma}^2}{\sigma^2} \leq b\right) = 1-\alpha$,

where



$a = \chi^2_{\alpha/2}(n-d-1)$, $b = \chi^2_{1-\frac{\alpha}{2}}(n-d-1)$ ← quantiles of the $\chi^2(n-d-1)$ distribution

Re-arranging terms yields

$$\left[\frac{(n-d-1)\hat{\sigma}^2}{b}, \frac{(n-d-1)\hat{\sigma}^2}{a}\right]$$

↑ a $100(1-\alpha)\%$ confidence interval for σ^2.

↘ For β, we have the choice to construct $(d+1)$ separate confidence intervals; each with nominal coverage $(1-\alpha)$; or construct a global confidence interval, so that all components are taken into account simultaneously (≡ aka confidence regions)

Put $v_j = [(X^tX)^{-1}]_{jj}$

= j-th diagonal element of the square matrix $(X^tX)^{-1}$ → $\mathrm{Var}\,\hat{\beta}_j = \sigma^2 v_{jth}$.

Then

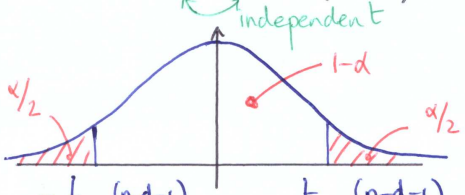
$$\frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 v_{j+1}}}}{\sqrt{\frac{(n-d-1)\hat{\sigma}^2}{\sigma^2}/(n-d-1)}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_{jth}}} \sim t_{n-d-1}$$

$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 v_{j+1}}} \sim \mathcal{N}(0,1)$

$\frac{(n-d-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-d-1}$

↑ Student-t distribution with $(n-d-1)$ degrees of freedom.

(This follows directly from the definition of the Student-t distribution: $T = \frac{Z}{\sqrt{S/n}} \sim t_n$, where $Z \sim \mathcal{N}(0,1)$ and $S \sim \chi^2_n$ independent)



$-t_{1-\frac{\alpha}{2}}(n-d-1)$, $t_{1-\frac{\alpha}{2}}(n-d-1)$ = quantiles of the t_{n-d-1} distribution.

$$\Rightarrow \left[\hat{\beta}_j - t_{1-\frac{\alpha}{2}}(n-d-1)\hat{\sigma}\sqrt{v_{jth}}, \hat{\beta}_j + t_{1-\frac{\alpha}{2}}(n-d-1)\hat{\sigma}\sqrt{v_{jth}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for β_j.

Alternatively, you may want to test for $H_0: \beta_j = 0$ ("absence of effect" = typically what you put inside a null hypothesis)

$\rightarrow := \hat{\beta}_j / \hat{\sigma}\sqrt{v_{j+1}}$ "z-score"

⇒ A large absolute value of z_j leads to the rejection of the null hypothesis; equivalently, a small p-value indicates the absence of statistical evidence that there is no association between the corresponding predictor x_j and the response.

(we never accept the null, we fail to reject the null)

OCTOPUS HO If we have a large number of predictors, say 100, then testing for the 100 coefficients individually not only is a waste of time, but also is inappropriate. Why? Because of statistical regularity: at the 95% nominal level for example, and assuming that all $\beta_j \neq 0$, 5 of the z-scores will appear in the destitution tails just by chance, leading to the rejection of the null.

→ Use a global test aka FISHER test.

In other words, test for several coefficients simultaneously:

Consider two candidate models, one being the reduced version of the other one, by removing some of the features of the larger model (we say that the models are nested)

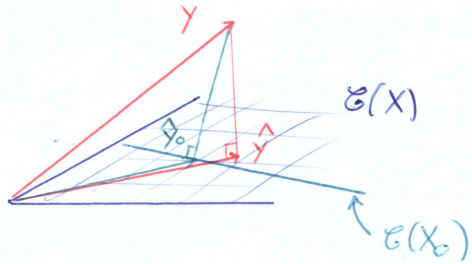
⇔ a subset of the β_j are equal to zero.

- Full model: (d_1+1) parameters
- Reduced model: (d_0+1) parameters, $d_0 \leq d_1$

↳ Corresponds to $y = X_0\beta_0 + \varepsilon_0$, $\operatorname{rank} X_0 = (d_0+1)$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

where the columns of X_0 are a subset of the columns of X.

If the reduced model is correct, then $\hat{y}_0$ = orthogonal projection of y onto the column space of X_0, is "close" to $\hat{y}$ = ⊥ projection of y onto $\mathcal{C}(X)$.



⇒ Check the size of $\|\hat{y}_0 - \hat{y}\|^2$ = square distance between $\hat{y}_0$ and $\hat{y}$.

It turns out that we are in a better position if we remove the scale effect on the output variable (square distance depends on the measurement unit, right?)

⇒ Consider $\dfrac{\|\hat{y}_0 - \hat{y}\|^2}{\|y - \hat{y}\|^2}$ $\left(= \dfrac{\|(\hat{y}_0 - y) + (y - \hat{y})\|^2}{\|y - \hat{y}\|^2} \right)$

= relative error (see p. 30)

In fact, $(\hat{y}_0 - \hat{y})$ and $(y - \hat{y})$ live in subspaces of different dimension, and we need to readjust for the dimension to be able to derive the distribution of the relative error:

$$F := \frac{\|\hat{y}_0 - \hat{y}\|^2 / (d_1 - d_0)}{\|y - \hat{y}\|^2 / (n - d_1 - 1)}$$

Facts: the numerator and denominator are independent,

and $\frac{1}{\sigma^2}\|y-\hat{y}\|^2 \sim \chi^2_{n-d_1-1}$ &

$$\frac{1}{\sigma^2}\|\hat{y}-\hat{y}_0\|^2 \sim \chi^2_{d_1-d_0}$$

under the null hypothesis that the reduced model is correct.

⇒ Under H_0: $\boxed{F \sim F_{d_1-d_0,\, n-d_1-1}}$

"large" value, reject, bla bla bla...

F-distribution with parameters (d_1-d_0) and $(n-d_1-1)$.

Remark: Why are the num & den independent, intuitively?
Because they live in orthogonal subspaces.

Remarks: (i) An alternative notation for F is widely used:

$$\boxed{F = \frac{(RSS_0 - RSS_1)/(d_1-d_0)}{RSS_1/(n-d_1-1)}}, \text{ where}$$

↘ RSS_1 = Residual Sum of Squares of the larger model with (d_1+1) parameters

↘ RSS_0 = Residual Sum of Squares of the reduced model with (d_0+1) parameters, $(d_0<d_1)$

Indeed,

$$\|y-\hat{y}_0\|^2 = \|y-\hat{y}+\hat{y}-\hat{y}_0\|^2$$
$$= \|y - Hy + Hy - H_0 y\|^2$$

where H_0 = projection matrix onto $\mathcal{C}(X_0)$

$$\|y-\hat{y}_0\|^2 = \|(I-H)y + Hy - H_0 H y\|^2$$

since $H_0 y = H_0 H y$ as $\mathcal{C}(X_0) \subset \mathcal{C}(X)$

$$= \|\underbrace{(I-H)y}_{\in \mathcal{C}(X)^\perp} + \underbrace{(I-H_0)Hy}_{\in \mathcal{C}(X)\cap\mathcal{C}(X_0)^\perp}\|^2$$

perpendicular subspaces ⇒ Pythagora

$$= \|(I-H)y\|^2 + \|(I-H_0)Hy\|^2$$
$$= \|y-\hat{y}\|^2 + \|\hat{y}-\hat{y}_0\|^2$$

Thus

$$\|\hat{y}-\hat{y}_0\|^2 = \|y-\hat{y}_0\|^2 - \|y-\hat{y}\|^2$$
$$= RSS_0 - RSS_1$$

(ii) If the simpler model drops only one coefficient, the F-test is testing for the nullity of this coefficient. But this is exactly what the Student-t test was designed for! So, which test shall we be using? It turns out that the two tests are equivalent in this case, and it is possible to show that $F = z_j^2$ (the z-score associated with the corresponding coefficient)

(iii) What about confidence regions for β?
Since $\hat{\beta} \sim \mathcal{N}(\beta,\ \sigma^2(X^tX)^{-1})$, $\Sigma_{\hat{\beta}}$, you may want to use $(\hat{\beta}-\beta)^t \Sigma_{\hat{\beta}}^{-1} (\hat{\beta}-\beta) \sim \chi^2_{d+1}$

but unfortunately $\Sigma_{\hat\beta} = \sigma^2 (X^t X)^{-1}$, with σ unknown. Big deal: replace σ^2 by $\hat\sigma^2$:

→ Let R be an $r \times (d+1)$ matrix of rank r

Typically, we will take R of the form

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \end{pmatrix} \quad \uparrow r \downarrow$$

$\leftarrow (d+1) \rightarrow$

Role of R: select a subset of the original coefficients $\beta_0, \beta_1, \ldots, \beta_d$

↓ Consider $R(\hat\beta - \beta)$.

We know that $\underset{r\times 1}{R\hat\beta} \sim \mathcal{N}\left(\underset{r\times 1}{R\beta}, \sigma^2 \underset{r\times r}{R(X^tX)^{-1}R^t}\right)$

$$\Rightarrow \frac{1}{\sigma^2}\left[R(\hat\beta-\beta)\right]^t \left[R(X^tX)^{-1}R^t\right]^{-1}\left[R(\hat\beta-\beta)\right] \sim \chi^2_r$$

→ Also, $\dfrac{(n-d-1)\hat\sigma^2}{\sigma^2} \sim \chi^2_{n-d-1}$

→ Finally, $\hat\sigma^2$ and $\hat\beta$ are independent (cosq of Cochran theorem)

Thus:

$$\frac{\frac{1}{\sigma^2}\left[R(\hat\beta-\beta)\right]^t \left[R(X^tX)^{-1}R^t\right]^{-1}\left[R(\hat\beta-\beta)\right] \,/\, r}{\frac{(n-d-1)\hat\sigma^2}{\sigma^2} \,/\, (n-d-1)} \sim F_{r,n-d-1}$$

Simplifying terms yields:

$$\boxed{\frac{1}{r\hat\sigma^2}\left[R(\hat\beta-\beta)\right]^t \left[R(X^tX)^{-1}R^t\right]^{-1}\left[R(\hat\beta-\beta)\right] \sim F_{r,n-d-1}}$$

↑ From this, we can construct a confidence region for simultaneously r parameters of the model, $\beta_{j_1} \ldots \beta_{j_r}$:

$$\left\{(\beta_{j_1}, \ldots, \beta_{j_r}) \;\middle|\; \frac{1}{r\hat\sigma^2}\left[R(\hat\beta-\beta)\right]^t \left[R(X^tX)^{-1}R^t\right]^{-1}\left[R(\hat\beta-\beta)\right] \leq f_{1-\alpha}(r, n-d-1)\right\}$$

↑ confidence region, with nominal coverage $(1-\alpha)$.

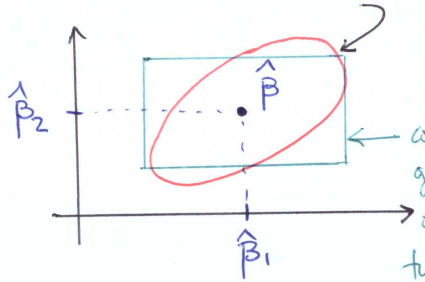
↑ $(1-\alpha)$ quantile of the $F_{r,n-d-1}$ distribution

↘ Example: taking $r=2$ and $R = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \end{bmatrix}$,

$R(\hat\beta - \beta) = \begin{pmatrix} \hat\beta_0 - \beta_0 \\ \hat\beta_1 - \beta_1 \end{pmatrix}$ ≡ keep one coefficient + intercept

$$\left\{(\beta_0, \beta_1) \;\middle|\; \frac{v_{22}(\hat\beta_0-\beta_0)^2 - 2v_{12}(\hat\beta_0-\beta_0)(\hat\beta_1-\beta_1) + v_{11}(\hat\beta_1-\beta_1)^2}{2\hat\sigma^2(v_{11}v_{22}-v_{12}^2)} \leq f_{1-\alpha}(2, n-d-1)\right\}$$

= ellipsoid = confidence region which takes into account the correlation between $\hat\beta_1$ and $\hat\beta_2$.



← confidence region you would get by ignoring the correlation of $\hat\beta_1$ and $\hat\beta_2$; and combining two separately constructed confidence intervals.

II. MODEL ASSESSMENT

The linear model $y = X\beta + \varepsilon$ assumes that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, and rank $X = (d+1)$. Once we have an estimate of β, it is necessary to check these assumptions. We study:

- ↘ Residuals $\hat{\varepsilon} = y - \hat{y} = y - X\hat{\beta}$, and see how we can use them to check that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ is legitimate.
- ↘ Projection Matrix $H = X(X^tX)^{-1}X^t$
- ↘ Matrix of Observation X and tools to detect colinearity.

II.1. Study of the residuals.

We derived page 22 that

$$\boxed{\begin{aligned} E\,\hat{\varepsilon} &= 0 \\ \Sigma_{\hat{\varepsilon}} &= \sigma^2(I-H) \end{aligned}}$$

RESIDUALS

these should be compared with the model assumptions

$$\boxed{\begin{aligned} E\,\varepsilon &= 0 \\ \Sigma_{\varepsilon} &= \sigma^2 I_n \end{aligned}}$$

ERRORS

↘ Residuals are correlated, while we assume uncorrelated errors. Even worse,

$\mathrm{Var}\,\hat{\varepsilon}_i = \sigma^2(1-h_{ii})$ ← $h_{ij} := [H]_{ij}$

= non-constant variance.

We get rid of this extra variability, and consider the normalized residuals $r_i := \dfrac{\hat{\varepsilon}_i}{\sigma\sqrt{1-h_{ii}}}$.

However, σ is unknown, and we shall make use of the standardized residuals instead: $t_i := \dfrac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$.

ratio of a normal RV with the square root of a χ^2 RV (up to a renormalization factor)
⇒ Expect t_i to have a Student distribution.
Almost! Consider instead the studentized residuals

$$\boxed{t_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{-i}\sqrt{1-h_{ii}}}}$$

where $\hat{\sigma}_{-i}$ = estimator of σ when observation i is removed from the learning sample

We are going to show that t_i^* has a Student distribution.

Derivation of the distribution of t_i^*.

Notations:

$$X_{-i} = \begin{pmatrix} — & x_1^t & — \\ — & x_{i-1}^t & — \\ — & x_{i+1}^t & — \\ — & x_n^t & — \end{pmatrix} = \text{matrix } X \text{ without the } i\text{-th raw}$$

$(n-1)\times(d+1)$

$\hat{\beta}_{-i}$ = LS estimate of β obtained from the data set $\mathcal{L}_n^{-i} = \{(x_1,y_1)\dots(x_{i-1},y_{i-1}),(x_{i+1},y_{i+1}),\dots,(x_n,y_n)\}$

$\tilde{y}_i = x_i^t\hat{\beta}_{-i}$ = prediction of y_i

A few facts:

→ $\hat{\beta}_{-i} \sim \mathcal{N}(\beta,\ \sigma^2(X_{-i}^tX_{-i})^{-1})$

→ $y_i - \tilde{y}_i = x_i^t(\beta - \hat{\beta}_{-i}) + \varepsilon_i$

$\rightarrow \mathrm{Var}(y_i - \tilde{y}_i) = \mathrm{Var}\left(x_i^t(\beta - \hat{\beta}_{-i}) + \varepsilon_i\right)$

where $\hat{\beta}_{-i} = (X_{-i}^t X_{-i})^{-1} X_{-i}^t y$, which depends on $\varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_{i+1}, \ldots, \varepsilon_n$ only

$\Rightarrow \hat{\beta}_{-i}$ is uncorrelated with ε_i.

$= \mathrm{Var}\{x_i^t(\beta - \hat{\beta}_{-i})\} + \mathrm{Var}\,\varepsilon_i$

$= \mathrm{Var}\{x_i^t \hat{\beta}_{-i}\} + \sigma^2$

$= \sigma^2\left(1 + x_i^t (X_{-i}^t X_{-i})^{-1} x_i\right)$

$\rightarrow y_i - \tilde{y}_i$ is normally distributed.

Thus

$$\frac{\dfrac{y_i - \tilde{y}_i}{\sigma\sqrt{1 + x_i^t (X_{-i}^t X_{-i})^{-1} x_i}}}{\sqrt{\dfrac{(n-d-2)\hat{\sigma}_{-i}^2}{\sigma^2} \Big/ (n-d-2)}} = \frac{y_i - \tilde{y}_i}{\hat{\sigma}_{-i}\sqrt{1 + x_i^t (X_{-i}^t X_{-i})^{-1} x_i}} \sim t_{n-d-2}$$

$\sim N(0,1)$ in the numerator; $\sim \chi^2_{n-d-2}$ in the denominator

(since we have $n-1$ observations and $d+1$ predictors)

We derive an alternative expression for this term.

We make use of the Sherman-Morrison-Woodbury theorem, which ensures that for a $(d \times d)$ nonsingular matrix A and $(d \times 1)$ column vectors u and v, we have

$$(A + uv^t)^{-1} = A^{-1} - \frac{A^{-1} u v^t A^{-1}}{1 + v^t A^{-1} u}$$

Apply SMW theorem to

$X_{-i}^t X_{-i} = X^t X - x_i x_i^t$ (since $X^t X = \sum_{i=1}^n x_i x_i^t$)

$\Downarrow$

$$(X_{-i}^t X_{-i})^{-1} = (X^t X - x_i x_i^t)^{-1} = (X^t X)^{-1} + \frac{(X^t X)^{-1} x_i x_i^t (X^t X)^{-1}}{1 - x_i^t (X^t X)^{-1} x_i}$$

Thus

$$x_i^t (X_{-i}^t X_{-i})^{-1} x_i = \underbrace{x_i^t (X^t X)^{-1} x_i}_{h_{ii}} + \frac{(x_i^t (X^t X)^{-1} x_i)(x_i^t (X^t X)^{-1} x_i)}{1 - x_i^t (X^t X)^{-1} x_i}$$

$$= h_{ii} + \frac{h_{ii}^2}{1 - h_{ii}} = \frac{h_{ii}}{1 - h_{ii}}$$

We obtain

$$1 + x_i^t (X_{-i}^t X_{-i})^{-1} x_i = 1 + \frac{h_{ii}}{1 - h_{ii}} = \frac{1}{1 - h_{ii}} \quad (1)$$

Moreover,

$$y_i - \hat{y}_i = (1 - h_{ii})(y_i - \tilde{y}_i) \quad (2)$$

↳ Indeed, using $X_{-i}^t y_i = X^t y - x_i y_i$, we get

$$\begin{pmatrix} | & | & | & | \\ x_1 & x_{i-1} & x_{i+1} & x_n \\ | & | & | & | \end{pmatrix}_{(d+1)\times(n-1)} \begin{pmatrix} y_1 \\ y_{i-1} \\ y_{i+1} \\ y_n \end{pmatrix}_{(n-1)\times 1}$$

$$\hat{\beta}_{-i} = (X_{-i}^t X_{-i})^{-1} X_{-i}^t y_i = (X_{-i}^t X_{-i})^{-1}(X^t y - x_i y_i)$$

$$\hat{\beta}_{-i} = \left[(X^tX)^{-1} + \frac{(X^tX)^{-1} x_i x_i^t (X^tX)^{-1}}{1 - x_i^t (X^tX)^{-1} x_i} \right] (X^ty - x_i y_i)$$

$$= \hat{\beta} - (X^tX)^{-1} x_i y_i + \frac{(X^tX)^{-1} x_i x_i^t (X^tX)^{-1}}{1 - x_i^t (X^tX)^{-1} x_i} (X^ty - x_i y_i)$$

$$= \hat{\beta} - \frac{(X^tX)^{-1} x_i y_i (1 - x_i^t (X^tX)^{-1} x_i) + (X^tX)^{-1} x_i x_i^t (X^tX)^{-1} (X^ty - x_i y_i)}{1 - x_i^t (X^tX)^{-1} x_i}$$

$$= \hat{\beta} - \frac{(X^tX)^{-1} x_i y_i - (X^tX)^{-1} x_i y_i x_i^t (X^tX)^{-1} x_i + (X^tX)^{-1} x_i x_i^t (X^tX)^{-1} (X^ty - x_i y_i)}{1 - x_i^t (X^tX)^{-1} x_i}$$

$$= \hat{\beta} - \frac{(X^tX)^{-1} x_i y_i - (X^tX)^{-1} x_i x_i^t \underbrace{(X^tX)^{-1} X^ty}_{=\hat{\beta}}}{1 - x_i^t (X^tX)^{-1} x_i}$$

$$= \hat{\beta} - \frac{(X^tX)^{-1} x_i \underbrace{(y_i - x_i^t \hat{\beta})}_{= y_i - \hat{y}_i}}{1 - \underbrace{x_i^t (X^tX)^{-1} x_i}_{= h_{ii}}}$$

$$\Rightarrow \boxed{\hat{\beta}_{-i} = \hat{\beta} - \frac{(X^tX)^{-1} x_i}{1 - h_{ii}} (y_i - \hat{y}_i)} \qquad (\bullet)$$

We obtain

$$y_i - x_i^t \hat{\beta}_{-i} = y_i - x_i^t \hat{\beta} + \frac{\overbrace{x_i^t (X^tX)^{-1} x_i}^{= h_{ii}}}{1 - h_{ii}} (y_i - \hat{y}_i)$$

$$(y_i - \tilde{y}_i) = (y_i - \hat{y}_i) + \frac{h_{ii}}{1 - h_{ii}} (y_i - \hat{y}_i)$$

$$= \frac{1}{1 - h_{ii}} (y_i - \hat{y}_i), \text{ as required } \blacksquare$$

Putting (1) and (2) together,

$$\frac{\tilde{y}_i - y_i}{\hat{\sigma}_{-i} \sqrt{1 + x_i^t (X_i^t X_i)^{-1} x_i}} = \frac{(y_i - \hat{y}_i)/(1 - h_{ii})}{\hat{\sigma}_{-i} \sqrt{1/(1 - h_{ii})}} = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{-i} \sqrt{1 - h_{ii}}} \sim t_{n-d-2}.$$

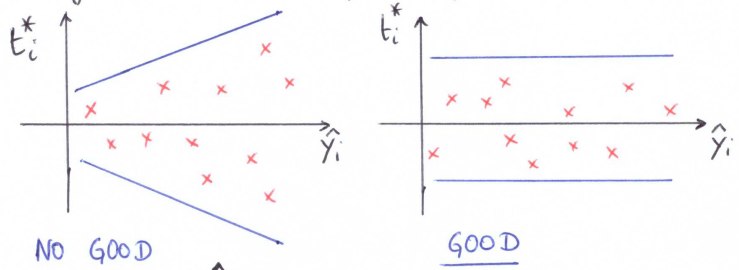
page 35

good.

The studentized residuals can be plotted in various ways to detect possible anomalies, and check the model assumptions.

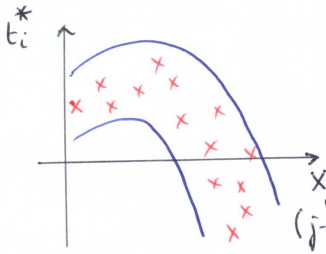
(a) Check constant variance.

By plotting t_i^* as a function of $\hat{y}_i$, you may visually check the assumption of constant variance:



You might consider transforming the response variable: $\sqrt{y}$, $\log y$ (if appropriate).

(b) Plot t_i^* against a predictor variable.



If you observe something like this, you should include terms like x_j^2, x_j^3, ... in the model.

(c) Check normality.

The residuals are not normally distributed & are correlated. However, $t_i^* \sim$ Student, which is almost normal if $n \gg d$. You may use a QQ plot to check visually if the studentized residuals look normal indeed.

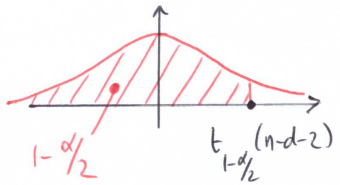
(d) Assumption of independence.

This assumption is technically hard to check. If the data is chronological (or spacial), a plot of the residuals versus time (or space) may reveal a systematic pattern.

(e) Use residuals to dectect outliers.

= points for which the response variable y_i is far from the predicted value.

For example, (x_i, y_i) = outlier if $|t_i^*| \gg t_{1-\alpha/2}(n-d-2)$



In practice, if $\alpha = 0.05$ and $n-d-2 \geq 30$, then $t_{1-\frac{\alpha}{2}}(n-d-2) \simeq 2$, and

(x_i, y_i) = outlier if $|t_i^*| \gg 2$

Outliers are BAD and should be omitted from the analysis, if this makes sense to do so.
Outliers can have a dramatic effect on R^2, by artificially increasing RSS, thus decreasing R^2.

II.2. Analysis of the projection matrix.

Equation (2) page 36 states that

$$\underbrace{y_i - \hat{y}_i}_{\text{error made when using the full dataset}} = (1 - h_{ii}) \underbrace{(y_i - \tilde{y}_i)}_{\text{error made when predicting } y_i \text{ without information contained in } x_i}$$

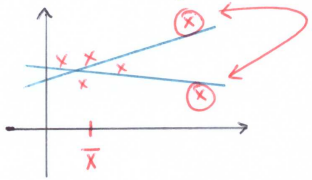
⇒ When h_{ii} is small (close to 0), the error we make,

whether we include x_i in the dataset or not, is similar. In this case, x_i does not have a big impact on the prediction of y_i. The situation is reversed if h_{ii} is close to 1.

⇒ h_{ii} may be used to quantify which points in the learning sample influence the most the prediction.

Example: For simple linear regression ($d=1$) with intercept, we can show that

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \in \left[\frac{1}{n}, 1\right]$$



far away points influence the slope of the linear fit a lot. These points are away from the center of mass $\bar{x}$
⇒ They have a large value of h_{ii}, close to 1.

When • $x_i = \bar{x}$, $h_{ii} = \frac{1}{n} = h_{min}$

• x_i far from $\bar{x}$, then $h_{ii} \nearrow 1 = h_{max}$.

Lemma:
Let H = projection matrix onto $\mathcal{C}(X)$; $\dim \mathcal{C}(X) = d+1$.
Then

(i) $\mathrm{Tr}\, H = d+1$

(ii) $\forall i \in \{1, \dots, n\}$, $0 \leq h_{ii} \leq 1$

(iii) If $h_{ii} = 0$ or 1, then $h_{ij} = 0 \quad \forall j \neq i$

(iv) $\forall j \neq i$, $-\frac{1}{2} \leq h_{ij} \leq \frac{1}{2}$.

Since $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$, we deduce from the lemma that

↘ If $h_{ii} = 1$, then $h_{ij} = 0$ for any $i \neq j$, and $\hat{y}_i$ is completely determined by y_i since $\hat{y}_i = y_i$

↘ If $h_{ii} = 0$, then, then $h_{ij} = 0 \quad \forall i \neq j$ and $\hat{y}_i = 0$; y_i has absolutely no influence on $\hat{y}_i$.

↘ Keeping all other values of y_j for $j \neq i$ fixed,

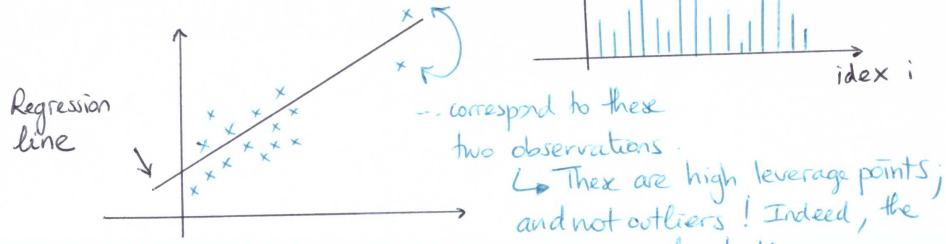
$$(\text{Change in } \hat{y}_i) = h_{ii} \times (\text{Change in } y_i)$$

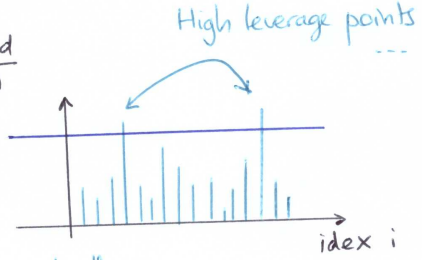
↑ A "high" leverage indicates that y_i provides a major contribution to the predicted value $\hat{y}_i$.

Note that $\text{Tr } H = d+1 = \sum h_{ii}$.
⇒ the average value of the h_{ii} is $\frac{d+1}{n}$.

Use this to quantify what we mean by "high" leverage:

Example: • If $h_{ii} > \frac{2d}{n}$
• If $h_{ii} > \frac{3d}{n}$





... correspond to these two observations.
↳ These are high leverage points; and not outliers! Indeed, the points are close to the regression line.

#Take Away Message

- Analysis of the (studentized) residuals allows us to detect observations (x_i, y_i) with unusual values of y_i (≡ far from the regression line)
- Analysis of the projection matrix allows us to detect points (x_i, y_i) located away from the center of mass $\bar{x}$ (≡ points with a high influence on the position of the regression line)

↳ Cook (77) introduced a measure which combines information about the residuals (outliers) and about the projection matrix (high leverage points) ↘

II.3. Cook's distance.

To evaluate the impact of observation (x_i, y_i) on the estimation of $\hat{\beta}$, we can naturally remove it from the learning sample, compute the least square estimate (denoted $\hat{\beta}_{-i}$), and evaluate the distance between $\hat{\beta}$ and $\hat{\beta}_{-i}$.

→ If the distance is "small", then (x_i, y_i) has little impact, while if de distance is "large", observation (x_i, y_i) influences the estimation of $\hat{\beta}$ significantly.

→ Since $\hat{\beta}$ and $\hat{\beta}_{-i} \in \mathbb{R}^{d+1}$, a Euclidean distance between $\hat{\beta}$ and $\hat{\beta}_{-i}$ based on an inner product can be written $d(\hat{\beta}, \hat{\beta}_{-i}) = \sqrt{(\beta - \hat{\beta}_{-i})^t Q (\beta - \hat{\beta}_{-i})}$, where Q is a positive definite and symmetrical matrix.

Many choices for Q are possible.

In view of the confidence region established on page 32, we may take $Q = \frac{1}{(d+1)\hat{\sigma}^2}(X^tX)$, with $R = I_{d+1}$, so that

$$\left\{ \beta \in \mathbb{R}^{d+1} \;\middle|\; (\hat{\beta}-\beta)^t \frac{(X^tX)}{(d+1)\hat{\sigma}^2} (\hat{\beta}-\beta) \leq f_{1-\alpha}(d+1, n-d-1) \right\}$$

is a $(1-\alpha)$ confidence region for β.

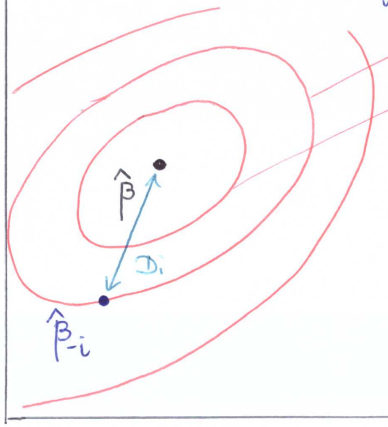
↑ Take this as a measure of the influence of each data point. Specifically, set

$$D_i = (\hat{\beta}-\hat{\beta}_{-i})^t \frac{(X^tX)}{(d+1)\hat{\sigma}^2} (\hat{\beta}-\hat{\beta}_{-i})$$

(COOK'S DISTANCE)

⇒ Evaluate Cook's distance for each observation.

If for example $D_i \approx f_{0.5}(d+1, n-d-1)$, then the removal of the i-th observation moves the LS estimate to the edge of the 50% confidence region. That's not good.



↘ Cook suggested that each $\hat{\beta}_{-i}$ to stay within a 10% confidence region to cause no concern.

↘ At first, it seems like computing $(n+1)$ regressions is costly. In fact,

we can show that Cook's distance admits a simple form:

$$D_i = \frac{1}{d+1} \frac{h_{ii}}{1-h_{ii}} t_i^2 \qquad (*)$$

Two terms contribute to Cook's distance:

→ t_i^2 = large value for outliers (p. 34)

→ $\frac{h_{ii}}{1-h_{ii}}$ = large for 'high' leverage points (corresponding to large values of h_{ii})

⇒ Cook's distance naturally combine these two elements in a unique coefficient.

Proof of (*)

Expression (•) page 37: $\hat{\beta}_{-i} - \hat{\beta} = -(y_i - \hat{y}_i) \frac{(X^tX)^{-1}x_i}{1-h_{ii}}$

Expression (2) page 36: $y_i - \hat{y}_i = (1-h_{ii})(y_i - \tilde{y}_i)$

$$\Rightarrow \hat{\beta}_{-i} - \hat{\beta} = -(y_i - \tilde{y}_i)(X^tX)^{-1}x_i$$

Thus

$$D_i = (y_i - \tilde{y}_i)^2 \, x_i^t (X^tX)^{-1} \frac{(X^tX)}{(d+1)\hat{\sigma}^2} (X^tX)^{-1} x_i$$

$$= (y_i - \tilde{y}_i)^2 \frac{1}{(d+1)\hat{\sigma}^2} \underbrace{x_i^t (X^tX)^{-1} x_i}_{=h_{ii}}$$

$$= \frac{h_{ii}}{(d+1)\hat{\sigma}^2} (y_i - \tilde{y}_i)^2$$

$$= \frac{h_{ii}}{(d+1)\hat{\sigma}^2} \frac{1}{(1-h_{ii})^2} \underbrace{(1-h_{ii})^2 (y_i - \tilde{y}_i)^2}_{(y_i - \hat{y}_i)^2}$$

$$D_i = \frac{h_{ii}}{(1-h_{ii})^2} \frac{1}{(d+1)\hat{\sigma}^2} (y_i - \tilde{y}_i)^2$$

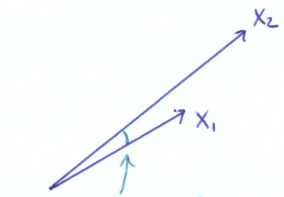
$$= \frac{h_{ii}}{1-h_{ii}} \frac{1}{d+1} \left(\frac{y_i - \tilde{y}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \right)^2$$

$\approx t_i$ (top of page 34)

$$= \frac{h_{ii}}{1-h_{ii}} \frac{t_i^2}{d+1}, \text{ as required} \quad \blacksquare$$

II.4. Detecting colinearity.

The derivation of the LS estimate requires that rank $X = d+1$, otherwise X^tX is not invertible. Even if this assumption is rarely violated in practice, it often happens that the columns of X are "nearly" linearly dependent. The consequences are rather bad.



small angle between x_1 and x_2.

If the angle between x_1 and x_2 is small, the plane spanned by x_1 and x_2 changes dramatically if one of the x_1 or x_2 changes slightly.
Since $\hat{y}$ = $\perp$ projection on $\mathcal{C}(X)$, the predicted value will also be greatly affected.
⇒ Prediction is unreliable and the estimation of β is unstable.

Mathematically, computing the inverse of X^tX is numerically unstable: the matrix X^tX is ill-conditioned; the determinant of X^tX will be close to 0, and thus $(X^tX)^{-1}$ has large values.

But remember that the covariance matrix of $\hat{\beta}$ is precisely $\sigma^2 (X^tX)^{-1}$. We see that colinearity amongst the predictors implies an increase in the variance of the coefficient estimates.

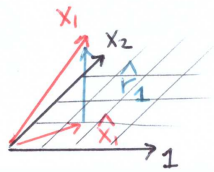
↳ This impacts any hypothesis test of the form $H_0: \beta_i = 0$, since the z-score $z_i = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{v_{i+1}}}$ (page 21) is smaller than it should be (due to an artificial increase in v_{i+1}), hence failing to reject the null when in fact you should. In other words, you conclude that there is no statistical evidence of association of a given predictor with the response variable, when in fact there is. The POWER of the test is reduced.

↳ $\mathbb{P}$(reject H_0 | H_1 is true)

⇒ We need ways to detect colinearity. We derive a simple coefficient (VIF) to assess colinearity. We proceed in two steps.

- **Step I**: Alternative expression for $\hat{\beta}_1$
  (we treat the case $\hat{\beta}_1$ without loss of generality. All coefficients can be treated the same way).

Idea: Since we worry about colinearity amongst the predictors, let's regress one of them (here x_1) on the remaining ones $(1, x_2, \dots, x_d)$; $x_1 = \hat{x}_1 + \hat{r}_1$



$\hat{r}_1$: residual error.
$\hat{x}_1$: our output variable, predicted using $(1, x_2, \dots, x_d)$.

⇒ $\forall i = 1, \dots, n$ $\quad x_{i1} = \hat{x}_{i1} + \hat{r}_1$.

- Next, recall that $X^t(y - X\hat{\beta}) = 0$; the second row of this equation is

$$\sum_{i=1}^{n} x_{i1}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_d x_{id}\right) = 0$$

(x_{i1} ↓ replace with $\hat{x}_{i1} + \hat{r}_{i1}$)

$$(\diamond) \quad \sum_{i=1}^{n} (\hat{x}_{i1} + \hat{r}_{i1})(y_i - \hat{\beta}_0 - \cdots - \hat{\beta}_d x_{id}) = 0$$

- Likewise, rows 2 to (d+1) of $X^t(y - X\hat{\beta})$ yield

$$\sum_{i=1}^{n} x_{ik}(y_i - \hat{y}_i) = 0$$

- Also, since $\hat{x}_{i1}$ is a linear function of $1, x_{i2}, \cdots, x_{id}$, we immediately see that $\sum_{i=1}^{n} (y_i - \hat{y}_i)\hat{x}_{i1} = 0$

($\langle y - \hat{y}, \hat{x}_1 \rangle = 0$ since vectors $(y - \hat{y})$ and $\hat{x}_1$ are $\perp$)

⇓

We get from $(\diamond)$ that

$$\sum_{i=1}^{n} (\hat{x}_{i1} + \hat{r}_{i1})(y_i - \hat{y}_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{y}_i) = 0$$

⇓

$$(\square) \quad \sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_d x_{id}) = 0$$

In addition, ↘ $\sum_{i=1}^{n} \hat{r}_{i1} = 0$ since the intercept is included in the model.

↘ $\sum_{i=1}^{n} \hat{r}_{i1} x_{ik} = \sum_{i=1}^{n} x_{ik}(x_{i1} - \hat{x}_{i1}) = 0$

($k = 2, \cdots, d$; LS equation when regressing x_1 on $(1, x_2, \cdots, x_d)$)

Equation $(\square)$ becomes

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1 x_{i1}) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1 \hat{x}_{i1} - \hat{\beta}_1 \hat{r}_{i1}) = 0$$

Next, notice that

$$\sum_{i=1}^{n} \hat{r}_{i1}\hat{x}_{i1} = \sum_{i=1}^{n} \hat{r}_{i1}(\hat{\alpha}_0 + \hat{\alpha}_2 x_{i2} + \cdots + \hat{\alpha}_d x_{id})$$

($\hat{\alpha}_0, \hat{\alpha}_2, \cdots, \hat{\alpha}_d$: regression coefficients when regressing x_1 on $(1, x_2, \cdots, x_d)$)

$$= \hat{\alpha}_0 \underbrace{\sum_{i=1}^{n} \hat{r}_{i1}}_{=0} + \hat{\alpha}_2 \underbrace{\sum_{i=1}^{n} \hat{r}_{i1} x_{i2}}_{=0} + \cdots + \hat{\alpha}_d \underbrace{\sum_{i=1}^{n} \hat{r}_{id} x_{id}}_{=0}$$

$$= 0$$

Thus $\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1 \hat{r}_{i1}) = 0$, which yields

$$\hat{\beta}_1 = \frac{\sum \hat{r}_{i1} y_i}{\sum \hat{r}_{i1}^2}$$

(> 0 as long as x_1 is not an exact linear combination of $1, x_2, \cdots, x_d$, which is ok if rank $X = (d+1)$.)

We can go further since $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id} + \varepsilon_i$

$$\sum \hat{r}_{i1} y_i = \beta_0 \underbrace{\sum \hat{r}_{i1}}_{=0} + \beta_1 \sum \hat{r}_{i1} x_{i1} + \underbrace{\cdots}_{=0} + \beta_d \sum \hat{r}_{i1} x_{id} + \sum \hat{r}_{i1} \varepsilon_i$$

$$= \sum \hat{r}_{i1}(\hat{x}_{i1} + \hat{r}_{i1})$$

$$= \underbrace{\sum \hat{r}_{i1} \hat{x}_{i1}}_{=0 \text{ since } \hat{x}_1 \perp \hat{r}_1} + \sum \hat{r}_{i1}^2 = \sum \hat{r}_{i1}^2$$

$$\Rightarrow \sum \hat{r}_{i1} y_i = \beta_1 \sum \hat{r}_{i1}^2 + \sum \hat{r}_{i1} \varepsilon_i \quad \Rightarrow \quad \hat{\beta}_1 = \beta_1 + \frac{\sum \hat{r}_{i1} \varepsilon_i}{\sum \hat{r}_{i1}^2}$$

(*)

- **Step I.** Alternative expression for $\text{Var}\,\hat{\beta}_1$.

Making use of (*) at bottom of page 48, we get

$$\text{Var}\,\hat{\beta}_1 = \frac{\sum \hat{r}_{i1}^2 \,\text{Var}\,\varepsilon_i}{\left(\sum \hat{r}_{i1}^2\right)^2} = \frac{\sigma^2}{\sum \hat{r}_{i1}^2}$$

Also, $R_1^2 = 1 - \frac{RSS_1}{TSS_1}$ ← $\sum \hat{r}_{i1}^2$ ← $\sum (x_{i1} - \bar{x}_1)^2$

↑ R^2 coefficient obtained by regressing x_1 on all other predictors.

Thus $\sum \hat{r}_{i1}^2 = TSS_1 (1 - R_1^2)$, and $\text{Var}\,\hat{\beta}_1 = \frac{\sigma^2}{TSS_1} \frac{1}{1-R_1^2}$.

Generalizing this expression for other coefficients,

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

where R_j^2 = R^2 coefficient obtained by regressing x_j on all other predictors.

A few observations:

- A larger σ^2 ⇒ a larger variance of the coefficient estimates
- More variability in x_j ⇒ Better accuracy in the estimation of $\hat{\beta}_j$.
- Increasing n decreases $\text{Var}\,\hat{\beta}_j$.
- Smallest variance is obtained for $R_j^2 = 0$; i.e. when x_j has 0 sample correlation with every other predictor.
  Worst case scenario when $R_j^2 = 1$, i.e. when x_j is an exact linear combination of other predictors (colinearity !)

⇒ We define the Variance Inflation Factor (VIF) to be the ratio of the variance of $\hat{\beta}_j$ with its smallest possible value, given by $\sigma^2 / \sum (x_{ij} - \bar{x}_j)^2$.

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R_j^2} = \frac{\text{Var}\,\hat{\beta}_j}{\min \text{var}\,\hat{\beta}_j}$$

↑ A large VIF is a sign of colinearity of x_j with other predictors.
Rule of thumb: a VIF > 5 or 10 is problematic.

What shall we do when we detect colinearity ?

- drop one or more variables
- combine them in some way (if meaningful to do so)
- use principal components instead.

III. MAKING PREDICTIONS.

Last step in the inference procedure: making predictions.
There are two kinds of uncertainty associated with the prediction task:

↳ uncertainty about the coefficient estimates $\hat{\beta}_j$.

To quantify the source of uncertainty associated with the estimation of β_j, we construct CONFIDENCE INTERVALS using $\text{Var}\,(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_d x_d)$

↑ new observation $x = (x_1, \cdots, x_d)$

↳ Even if we knew the true parameters β_j, the response variable would not be predicted perfectly because of the additive error ε. Taking this source of uncertainty into account leads to a PREDICTION INTERVAL: the response variable is

$$Y_{n+1} = \beta_0 + \beta_1 x_{n+1,1} + \cdots + \beta_d x_{n+1,d} + \varepsilon_{n+1}$$

New input point

with
- $E\,\varepsilon_{n+1} = 0$
- $\mathrm{Var}\,\varepsilon_{n+1} = \sigma^2$
- $\mathrm{Cov}(\varepsilon_j, \varepsilon_{n+1}) = 0$, $j = 1, \dots, n$

→ Prediction is $\hat{y}_{n+1} = \hat{\beta}_0 + \cdots + \hat{\beta}_d x_{n+1,d}$

→ Prediction error is

$$\hat{\varepsilon}_{n+1} := Y_{n+1} - \hat{y}_{n+1} = x_{n+1}^t (\beta - \hat{\beta}) + \varepsilon_{n+1}$$

(where $x_{n+1} = (1, x_{n+1,1}, \dots, x_{n+1,d})^t$.)

$$E\,\hat{\varepsilon}_{n+1} = x_{n+1}^t (\beta - E\hat{\beta}) + E\,\varepsilon_{n+1} = 0$$

$$\mathrm{Var}\,\hat{\varepsilon}_{n+1} = \mathrm{Var}\left\{ x_{n+1}^t (\beta - \hat{\beta}) + \varepsilon_{n+1} \right\}$$

Since $\hat{\beta}$ is computed from the first n observations, it is uncorrelated with ε_{n+1}

$$= \mathrm{Var}\left\{ x_{n+1}^t (\beta - \hat{\beta}) \right\} + \mathrm{Var}\,\varepsilon_{n+1}$$

$$= x_{n+1}^t \Sigma_{\hat{\beta}} x_{n+1} + \sigma^2$$

$$= \sigma^2 \left(1 + x_{n+1}^t (X^t X)^{-1} x_{n+1} \right)$$

Then, business as usual:

$$\frac{\dfrac{\hat{\varepsilon}_{n+1}}{\sigma\sqrt{1 + x_{n+1}^t (X^t X)^{-1} x_{n+1}}}}{\sqrt{\dfrac{(n-d-1)\,\hat{\sigma}^2}{\sigma^2}} \Big/ (n-d-1)} = \frac{Y_{n+1} - \hat{y}_{n+1}}{\hat{\sigma}\sqrt{1 + x_{n+1}^t (X^t X)^{-1} x_{n+1}}} \sim t_{n-d-1}$$

$\hat{\varepsilon}_{n+1} \sim \mathcal{N}(0,1)$

$\sim \chi^2_{n-d-1}$

Denoting by $t_\alpha(n-d-1)$ the α-quantile of the t_{n-d-1} distribution, a $100(1-\alpha)\%$ prediction interval for Y_{n+1} is

$$x_{n+1}^t \hat{\beta} = \hat{y}_{n+1} \pm t_{1-\alpha/2}(n-d-1)\,\hat{\sigma}\sqrt{1 + x_{n+1}^t (X^t X)^{-1} x_{n+1}}$$

PREDICTION INTERVAL (PI)

Remarks: (i) The confidence interval for $x_{n+1}^t \beta$ is narrower than the PI, as $x_{n+1}^t \hat{\beta} \sim \mathcal{N}\left(x_{n+1}^t \beta,\ \sigma^2 x_{n+1}^t (X^t X)^{-1} x_{n+1} \right)$, and is given by

$$x_{n+1}^t \hat{\beta} = \hat{y}_{n+1} \pm t_{1-\alpha/2}(n-d-1)\,\hat{\sigma}\sqrt{x_{n+1}^t (X^t X)^{-1} x_{n+1}}$$

CONFIDENCE INTERVAL (CI)

(ii) For simple linear regression ($d=1$), the PI simplifies to

$$\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \pm t_{1-\alpha/2}(n-2)\,\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

⇒ Prediction error has minimum variance at the center of mass $x_{n+1} = \bar{x}$. This can be shown to be true when $d > 1$ as well.

(iii) For known β_js, CI has width 0, but not PI.