Machine Learning

Linear Regression

Irving Gómez Méndez

August-December, 2021

UNIVERSIDAD IBEROAMERICANA
CIUDAD DE MÉXICO ®

Multivariate Normal Distribution

Let be

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}, \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \right)$$

Then

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}\left(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(\mathbf{x} - \mu_x),\ \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)$$

Proof:

Consider the matrix

$$A = \begin{bmatrix} I & -\Sigma_{yx}\Sigma_{xx}^{-1} \\ \mathbf{0} & I \end{bmatrix}$$

and let be $\mu = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}$

Thus, $A\begin{pmatrix}\mathbf{y}\\ \mathbf{x}\end{pmatrix} \sim \mathcal{N}\left(A\mu,\, A\Sigma A^T\right)$. Now, let us compute this expressions

$$A\begin{pmatrix}\mathbf{y}\\ \mathbf{x}\end{pmatrix} = \begin{pmatrix}\mathbf{y} - \Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x}\\ \mathbf{x}\end{pmatrix} \equiv \begin{pmatrix}\mathbf{u}\\ \mathbf{x}\end{pmatrix}$$

$$A\mu = \begin{pmatrix}\mu_y - \Sigma_{yx}\Sigma_{xx}^{-1}\mu_x\\ \mu_x\end{pmatrix}$$

$$\begin{aligned}
A\Sigma A^T &= \begin{pmatrix}\Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} & \Sigma_{yx} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xx}\\ \Sigma_{xy} & \Sigma_{xx}\end{pmatrix} A^T\\
&= \begin{pmatrix}\Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} & \mathbf{0}\\ \Sigma_{xy} & \Sigma_{xx}\end{pmatrix}\begin{pmatrix}I & \mathbf{0}\\ -\Sigma_{xx}^{-1}\Sigma_{xy} & I\end{pmatrix}\\
&= \begin{pmatrix}\Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} & \mathbf{0}\\ \mathbf{0} & \Sigma_{xx}\end{pmatrix}
\end{aligned}$$

Because

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix}$$

is a (multivariate) normal variable and $\mathrm{Cov}(\mathbf{u}, \mathbf{x}) = 0$, it implies that $\mathbf{u}$ and $\mathbf{x}$ are independent, and hence $\mathbf{u}|\mathbf{x}$ has the same distribution than $\mathbf{u}$, that is

$$\mathbf{u}|\mathbf{x} \sim \mathcal{N}\left(\mu_y - \Sigma_{yx}\Sigma_{xx}^{-1}\mu_x,\, \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)$$

That is,

$$\mathbf{y} - \Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x}|\mathbf{x} \sim \mathcal{N}\left(\mu_y - \Sigma_{yx}\Sigma_{xx}^{-1}\mu_x,\, \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)$$

And from here, we conclude

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}\left(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(\mathbf{x} - \mu_x),\, \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)$$

$\square$

Define

$$\begin{aligned}\mu_{y|x} &= \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(\mathbf{x} - \mu_x) \\ &= \left[\mu_y - \Sigma_{yx}\Sigma_{xx}^{-1}\mu_x\right] + \left[\Sigma_{yx}\Sigma_{xx}^{-1}\right]\mathbf{x} \\ &\equiv \beta_0 + \beta_1\mathbf{x}\end{aligned}$$

and

$$\Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$$

Remember that

$$\Sigma_{yy} = \mathbb{E}_{\mathbf{y}}\left[(\mathbf{y} - \mu_y)(\mathbf{y} - \mu_y)^T\right] \quad (\text{resp. } \Sigma_{xx})$$

and

$$\Sigma_{yx} = \mathbb{E}_{\mathbf{y},\mathbf{x}}\left[(\mathbf{y} - \mu_y)(\mathbf{x} - \mu_x)^T\right] \quad (\text{similarly } \Sigma_{xy} = \Sigma_{yx}^T)$$

Assume for now that $\mathbf{x} \equiv x \in \mathbb{R}$ and $\mathbf{y} \equiv y \in \mathbb{R}$, and that we count with a sample $\mathcal{D}_n = ((x_1, y_1), \ldots, (x_n, y_n))$ from independent and identically distributed random variables with the same distribution than the generic vector (x, y).
Thus (intuitively) good estimators of the previous quantities would be given by

- $\widehat{\Sigma}_{yy} \equiv S_{yy} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$ (resp. S_{xx})
- $\widehat{\Sigma}_{yx} \equiv S_{yx} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = S_{xy}$
- $\hat{\mu}_{y|x} \equiv \hat{y}|x = \hat{\beta}_0 + \hat{\beta}_1 x$ where

$$\hat{\beta}_0 = \bar{y} - S_{yx} S_{xx}^{-1} \bar{x}$$

$$\hat{\beta}_1 = S_{yx} S_{xx}^{-1}$$

- $\widehat{\Sigma}_{y|x} \equiv \hat{\sigma}^2 = S_{yy} - S_{yx} S_{xx}^{-1} S_{xy}$

Let be $\hat{y}_i = \hat{y}_i | x_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, \ldots, n$, we define the Sum of Squared Estimate Errors (SSE) (Suma de Cuadrados del Error (SCE), in Spanish) also known as Sum of Squared Residuals (SSR) or Residual Sum of Squares (RSS)

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Remember that $\hat{y}_i = \bar{y} + S_{yx} S_{xx}^{-1} (x_i - \bar{x})$, thus applying simple algebra we get

$$
\begin{aligned}
SSR &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 + S_{yx} S_{xx}^{-1} \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right) S_{xx}^{-1} S_{xy} \\
&\quad - 2 \sum_{i=1}^{n} (y_i - \bar{y}) S_{yx} S_{xx}^{-1} (x_i - \bar{x}) \\
\frac{1}{n} SSR &= S_{yy} + S_{yx} \cancel{S_{xx}^{-1}} \cancel{S_{xx}} S_{xx}^{-1} S_{xy} - 2 S_{yx} S_{xx}^{-1} S_{xy} \\
&= S_{yy} - S_{yx} S_{xx}^{-1} S_{xy}
\end{aligned}
$$

Linear Regression

Mathematical modeling refers to the construction of mathematical expressions that describes the behavior of a variable of interest Y. Frequently we want to add to the model some variables (features) X, which give information about the variable of interest Y denoted as response.

In regression analysis one considers (X, Y) as random vector, where X is $\mathbb{R}^p$-valued ($X \in \mathcal{X} \subseteq \mathbb{R}^p$) and Y is $\mathbb{R}$-valued ($Y \in \mathcal{Y} \subset \mathbb{R}$). We are interested on how the variable Y depends on the value of the observation vector X. This means that we want to find a function $f : \mathcal{X} \to \mathcal{Y}$, such that $f(X)$ is a good approximation of Y, that is, $f(X)$ should be close to Y in some sense, which is equivalent to making $|f(X) - Y|$ "small". Since X and Y are random vectors, $|f(X) - Y|$ is random as well, therefore it is not clear what "small $|f(X) - Y|$" means.

We can resolve this problem by introducing the so-called L_2 risk or mean squared error of f,

$$\mathbb{E}_{X,Y}\left[f(X) - Y\right]^2,$$

and requiring it to be as small as possible. So we are interested in a (measurable) function $m : \mathcal{X} \to \mathcal{Y}$ such that

$$m = \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\min}\, \mathbb{E}_{X,Y}\left[f(X) - Y\right]^2$$

Such function that minimizes the mean squared error is given by the regression function

$$m(X) = \mathbb{E}[Y|X]$$

Proof:

For any arbitrary function $f : \mathcal{X} \to \mathcal{Y}$,

$$
\begin{aligned}
\mathbb{E}_{X,Y}\left[f(X) - Y\right]^2 &= \mathbb{E}_{X,Y}\left[f(X) - m(X) + m(X) - Y\right]^2 \\
&= \mathbb{E}_{X,Y}\left[f(X) - m(X)\right]^2 + \mathbb{E}_{X,Y}\left[m(X) - Y\right]^2,
\end{aligned}
$$

where we have used

$$
\begin{aligned}
&\mathbb{E}_{X,Y}\left[(f(X) - m(X))(m(X) - Y)\right] \\
&= \mathbb{E}_X\left\{\mathbb{E}_{Y|X}\left[(f(X) - m(X))(m(X) - Y)\right]\right\} \\
&= \mathbb{E}_X\left\{(f(X) - m(X))\mathbb{E}_{Y|X}\left[(m(X) - Y)\right]\right\} \\
&= \mathbb{E}_X\left\{(f(X) - m(X))(m(X) - m(X))\right\} \\
&= 0
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\min}\, \mathbb{E}_{X,Y}\left[f(X) - Y\right]^2 \\
&= \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\min}\, \mathbb{E}_{X,Y}\left[f(X) - m(X)\right]^2 + \mathbb{E}_{X,Y}\left[m(X) - Y\right]^2 \\
&= \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\min}\, \mathbb{E}_{X}\left[f(X) - m(X)\right]^2
\end{aligned}
$$

Note that $\mathbb{E}_X\left[f(X) - m(X)\right]^2$, called the L_2 error of f is nonnegative and is zero if $f(X) = m(X)$. Therefore

$$
m = \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\min}\, \mathbb{E}_{X,Y}\left[f(X) - Y\right]^2
$$

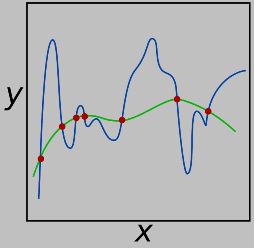
$\square$

For practical problems, the distribution of (X, Y) is unknown and hence, the regression function is unknown as well. However, in our framework, we have access to a training set $\mathcal{D}_n = (X_i, Y_i)_{i=1,\dots,n}$ where the collected data has the same distribution than (X, Y) and are considered independent. The goal is to use the data $\mathcal{D}_n$ to construct a learning model, also called learner or predictor, $m_n : \mathcal{X} \to \mathcal{Y}$ which estimates the function m, and enables us to predict the outcome for new unseen objects.

Thus, instead of minimizing the L_2 risk we minimize the empirical L_2 risk

$$\mathbb{E}_{\mathcal{D}_n}[f(X) - Y]^2 = \frac{1}{n}\sum_{i=1}^{n}[f(X_i) - Y_i]^2$$

Note that minimizing the above expression over all the functions $f : \mathcal{X} \to \mathcal{Y}$ is not well-defined, since every function which takes the value Y_i for every X_i would have zero empirical risk.



[1]

[1] picture taken from Wikipedia: https://en.wikipedia.org/wiki/Regularization_(mathematics)

We can resolve this problem restricting the search of the function that minimizes the empirical risk into a pre-defined set of functions $\mathcal{F}$. Moreover, the parametric estimation uses a model belonging to a set of functions $\mathcal{F}_\Theta$ determined by a finite number of parameters Θ, then the estimation is made through the inference of this set of parameters that minimize the empirical risk,

$$m_n = m_n(\cdot, \hat{\theta}) = \underset{f_\theta \in \mathcal{F}_\Theta}{\arg\min}\, \mathbb{E}_{\mathcal{D}_n}[f_\theta(X) - Y]^2,$$

where

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min}\, \mathbb{E}_{\mathcal{D}_n}[f_\theta(X) - Y]^2$$

For example let be $\mathcal{F}_\Theta = \{f : \mathcal{X} \to \mathcal{Y} : f(X) = X^T\beta,\ \beta \in \mathbb{R}^p\}$ $(\Theta = \{\beta : \beta \in \mathbb{R}^p\})$,

$$m_n(X) = X^T\hat{\beta} = \underset{f_\theta \in \mathcal{F}_\Theta}{\arg\min}\, \mathbb{E}_{\mathcal{D}_n}[f_\theta(X) - Y]^2$$

where

$$\begin{aligned}
\hat{\beta} &= \underset{\beta \in \mathbb{R}^p}{\arg\min}\, \mathbb{E}_{\mathcal{D}_n}[X^T\beta - Y]^2 \\
&= \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} [X_i^T\beta - Y_i]^2
\end{aligned}$$

This is known as Ordinary Leas Squares (OLS).

Let be $\mathbf{X} = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$ and $\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $\mathbf{X}$ is known as the design matrix while $\mathbf{Y}$ is known as the response vector. Then

$$\sum_{i=1}^{n} [X_i^T \beta - Y_i]^2$$

can be written as

$$\begin{aligned} \sum_{i=1}^{n} [X_i^T \beta - Y_i]^2 &= [\mathbf{X}\beta - \mathbf{Y}]^T [\mathbf{X}\beta - \mathbf{Y}] \\ &= [\beta^T \mathbf{X}^T - \mathbf{Y}^T][\mathbf{X}\beta - \mathbf{Y}] \\ &= \beta^T \mathbf{X}^T \mathbf{X} \beta - 2\beta^T \mathbf{X}^T \mathbf{Y} + \mathbf{Y}^T \mathbf{Y}, \end{aligned}$$

$\hat{\beta}$ can be obtained from the right-hand side of the above expression.

$\hat{\beta}$ satisfies

$$\frac{\partial}{\partial \beta}(\beta^T \mathbf{X}^T \mathbf{X} \beta - 2\beta^T \mathbf{X}^T \mathbf{Y} + \mathbf{Y}^T \mathbf{Y})\Big|_{\hat{\beta}} = 0$$

That is,

$$\begin{aligned} 2\mathbf{X}^T \mathbf{X} \hat{\beta} - 2\mathbf{X}^T \mathbf{Y} = 0 \\ \Leftrightarrow \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\hat{\beta}) = 0 \end{aligned} \qquad (1)$$

Equation (1) is known as the **normal equations**. It is easy to see that $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \mathbf{Y}$.

Hence, under this model, the best prediction $\widehat{\mathbf{Y}}$ for the vector of response $\mathbf{Y}$ is given by

$$\widehat{\mathbf{Y}} = \begin{bmatrix} \widehat{Y}_1 \\ \vdots \\ \widehat{Y}_n \end{bmatrix} = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} \hat{\beta}$$
$$= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Let be $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, so $\widehat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}$.
We will see that $\widehat{\mathbf{Y}}$ is the orthogonal projection of $\mathbf{Y}$ over the span of the the columns of $\mathbf{X}$ (What does this mean?).

Projections

Let $\mathbf{X} = \left[\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)}\right]$ be an $n \times p$ matrix, let $W = \mathrm{Col}(\mathbf{X})$, and let $\mathbf{Y}$ be a vector in $\mathbb{R}^n$.
Let $\mathbf{Y} = \mathbf{Y}_W + \mathbf{Y}_{W^\perp}$ be the orthogonal decomposition with respect to W. By definition $\mathbf{Y}_W$ lies in $W = \mathrm{Col}(\mathbf{X})$ so there is a vector $\hat{\beta} \in \mathbb{R}^p$ with $\mathbf{Y}_W = \mathbf{X}\beta$, that is

$$\begin{aligned}
\mathbf{Y}_W &= \mathbf{X}\hat{\beta} \\
&= \left[\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)}\right] \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} \\
&= \hat{\beta}_1 \mathbf{X}^{(1)} + \cdots + \hat{\beta}_p \mathbf{X}^{(p)}
\end{aligned}$$

Choose any such vector $\hat{\beta}$. We know that $\mathbf{Y} - \mathbf{Y}_W = \mathbf{Y} - \mathbf{X}\hat{\beta}$ lies in $W^\perp$, which is equal to $\mathrm{Null}(\mathbf{X}^T)$. We thus have

$$0 = \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{X}\hat{\beta}$$

and so

$$\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{Y}$$

Hence,

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

Remember that $\mathbf{Y}_W = \mathbf{X}\hat{\beta}$, so it can be written as

$$\begin{aligned}\mathbf{Y}_W &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{P}\mathbf{Y}\end{aligned}$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.
Thus, $\mathbf{P}$ is a projection matrix over the columns of $\mathbf{X}$. Actually is the orthogonal projection onto $\mathrm{Col}(\mathbf{X})$.

Properties of a Projection Matrix

- If $\mathbf{P}$ is a projection matrix in a space W, then $\mathbf{P}^2 = \mathbf{P}$. Remember that a vector that has been projected onto W belongs to that space, thus projecting again over W would led the same result.
- If $\mathbf{P} = \mathbf{P}^T$, then $\mathbf{P}$ is the creates orthogonal projections onto W.

Suppose that $\mathbf{P}$ satisfies both conditions, and consider its SVD decomposition, so

$$\mathbf{P} = USV^T$$

where U and V are orthogonal matrices. (An orthogonal matrix satisfies: $Q^T Q = QQ^T = I$).

Actually, U and V are rotation or reflection matrices. So, we might think as if the projection is "computed" by S.
Because $\mathbf{P}^2 = \mathbf{P}$,

$$USV^T USV^T = USV^T$$

which implies $SV^T US = S$.
Using the fact that $\mathbf{P} = \mathbf{P}^T$, we get that $USV^T = VSU^T$. So $U = V$.
Therefore, it is satisfied

$$US^2U^T = USU^T$$

Since

$$S = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix}$$

Then $\lambda_i \in \{0, 1\}$

Those places where $\lambda_i = 1$ represent the coordinates (in the rotated space) where the projection is perform, the basis of W. On the other hand, the places where $\lambda_i = 0$ would lead to a basis for $W^{\perp}$.

Example 1

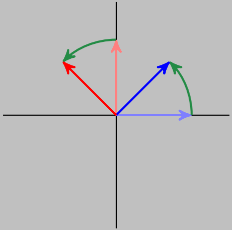
Consider the matrix

$$\mathbf{P} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}$$

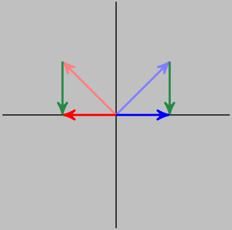
which can be written as

$$\mathbf{P} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

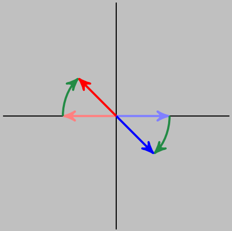
$$\mathbf{P} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$



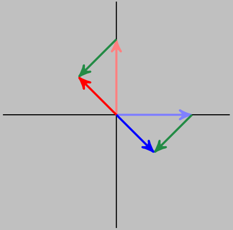
$$\mathbf{P} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$
$$= \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 \end{bmatrix}$$



$$\mathbf{P} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}$$



$$\mathbf{P} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}$$



Example 2

Let $\mathbf{X} = [\mathbf{X}^{(1)}] = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$

It can be shown that

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}$$

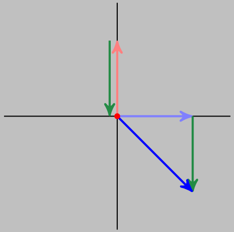
so $\mathbf{P}$ is the orthogonal projection onto Span $\left(\mathbf{X}^{(1)}\right)$

Now, consider the matrix

$$\mathbf{P'} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}$$

note that $\mathbf{P'}^2 = \mathbf{P'}$, hence $\mathbf{P'}$ is a projection matrix.

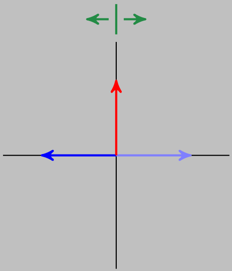
$$\mathbf{P}' = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}$$



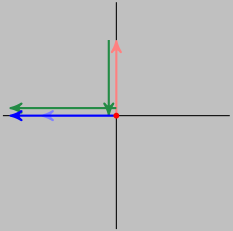
$\mathbf{P}'$ is an oblique projection onto Span $\left(\mathbf{X}^{(1)}\right)$.
The SVD decomposition of $\mathbf{P}'$ is

$$\mathbf{P}' = \underbrace{\begin{bmatrix} -1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}}_{U'} \underbrace{\begin{bmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{bmatrix}}_{S'} \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}}_{V'^T}$$

$$\mathbf{P}' = \begin{bmatrix} -1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}$$

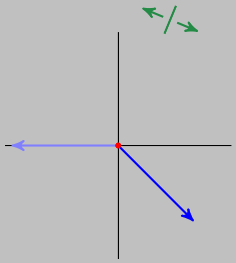


$$\mathbf{P}' = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}$$
$$= \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} -\sqrt{2} & 0 \\ 0 & 0 \end{bmatrix}$$



$$\mathbf{P}' = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} -\sqrt{2} & 0 \\ 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}$$



Let be $\mathbf{Z} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and

$$\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Moreover, let $\mathbf{B} = \mathbf{P_Z X}$ be the orthogonal projection of $\mathbf{X}$ onto $\mathrm{Col}(\mathbf{Z})$,

$$\mathbf{B} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

Simple calculations show $\mathbf{P'} = \mathbf{X}(\mathbf{B}^T\mathbf{X})^{-1}\mathbf{B}^T$.
Define $\mathbf{Y}_{\mathrm{IV}} = (\mathbf{X}(\mathbf{B}^T\mathbf{X})^{-1}\mathbf{B}^T)\mathbf{Y}$, $\mathbf{Y}_{\mathrm{IV}}$ is an oblique projection over $\mathrm{Col}(\mathbf{X})$.
This method is called **Two-Stage Least Squares (2SLS)**.
In the first stage we get the orthogonal projection of $\mathbf{X}$ onto $\mathrm{Col}(\mathbf{Z})$, where $\mathbf{Z}$ is called **instrumental variables**.

Linear Regression (2)

Consider the model

$$Y_i = X_i^T \beta + \varepsilon_i,\ i = 1, \dots, n$$

where $\mathbb{E}[\varepsilon_i | X_i] = 0, \quad \forall i = 1, \dots, n.$

So $\mathbb{E}[Y_i | X_i] = X_i^T \beta$, and that $\varepsilon_i = Y_i - X_i^T \beta$. $\varepsilon_i,\ i = 1, \dots, n$ are called the **errors of the model**.

Denote by $\boldsymbol{\varepsilon}$ the vector with these errors,

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Consider $\hat{\beta} = \mathbf{P}\mathbf{Y}$, where $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

Let be $\widehat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ ($\widehat{Y}_i = X_i^T\hat{\beta}$). Then $Y_i - \widehat{Y}_i$, $i = 1\dots, n$, called the **residuals of the model**, estimate the errors ε_i, $i = 1, \dots, n$ and $\mathbf{Y} - \widehat{\mathbf{Y}}$ estimates $\boldsymbol{\varepsilon}$.

Denote by $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$ the vector of residuals, note that $\mathbf{e} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$.

Moreover, it is easy to show that $\mathbf{I} - \mathbf{P}$ is the projection matrix onto $\mathrm{Col}(\mathbf{X})^{\perp}$, hence $\mathbf{X}^T\mathbf{e} = \mathbf{0}$

Writing $\mathbf{X} = \left[\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)}\right]$,

$$\mathbf{X}^T\mathbf{e} = \begin{bmatrix} \mathbf{X}^{(1)T}\mathbf{e} \\ \vdots \\ \mathbf{X}^{(p)T}\mathbf{e} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n X_i^{(1)} e_i \\ \vdots \\ \sum_{i=1}^n X_i^{(p)} e_i \end{bmatrix}$$

We can conclude that $\sum_{i=1}^n X_i^{(h)} e_i = 0$, for all $h \in \{1, \ldots, p\}$.
For example, if

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

We have proved that $\sum_{i=1}^n e_i = 0$ and $\sum_{i=1}^n x_i e_i = 0$

Distributional Properties

Let be $y \sim \mathcal{N}(\mu, \sigma^2\mathbf{I})$, then

1. $Ay \perp By \iff AB^T = 0$.
2. $Ay \perp y^T Cy \iff AC = 0$, where C is non-negative definite.
3. $y^T Cy \perp y^T Dy \iff CD = 0$, where C and D are non-negative definite.

Let be $y \sim \mathcal{N}(\mu, \Sigma)$, then $y^T Ay \sim \chi^2_{k,\lambda}$ if and only if $A\Sigma$ is symmetric and idempotent of range k, where $\lambda = \frac{1}{2}\mu^T A\mu$.

Assume $\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I})$, which is equivalent to the assumption $\varepsilon_i|X_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.
Consider the least squares estimate $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, then

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

Remember that
$SSE = (\mathbf{Y}-\widehat{\mathbf{Y}})^T(\mathbf{Y}-\widehat{\mathbf{Y}}) = (\mathbf{Y}-\mathbf{PY})^T(\mathbf{Y}-\mathbf{PY}) = \mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y}$,
then

- $$\frac{SSE}{\sigma^2} = \mathbf{Y}^T\frac{\mathbf{I}-\mathbf{P}}{\sigma^2}\mathbf{Y}$$

  and

  $$\frac{SSE}{\sigma^2}|\mathbf{X} \sim \chi^2_{n-p}$$

Denote by $\hat{\sigma}^2$ the unbiased estimate of σ, $\hat{\sigma}^2 = \frac{SSE}{n-p}$, then

- $$\hat{\beta} \perp \hat{\sigma}^2 | \mathbf{X}$$

To see this, remember that $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and $\hat{\sigma}^2 = \frac{1}{n-p}\mathbf{Y}^T(\mathbf{I} - \mathbf{P})\mathbf{Y}$. Hence $\hat{\beta}$ and $\hat{\sigma}^2$ are independent if and only if

$$\frac{1}{n-p}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{P}) = \mathbf{0}$$

As a consequence of the previous results, we have that

- $$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} | \mathbf{X} \sim \chi^2_{n-p}$$
- $$\frac{a^T\hat{\beta} - a^T\beta}{\sqrt{\sigma^2 a^T(\mathbf{X}^T\mathbf{X})^{-1}a}} | \mathbf{X} \sim \mathcal{N}_1(0,1), \quad \forall a \in \mathbb{R}^p, a \neq 0$$
- $$\frac{a^T\hat{\beta} - a^T\beta}{\sqrt{\hat{\sigma}^2 a^T(\mathbf{X}^T\mathbf{X})^{-1}a}} | \mathbf{X} \sim t_{n-p}, \quad \forall a \in \mathbb{R}^p, a \neq 0$$

- $$(\hat{\beta}-\beta)^T \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2}(\hat{\beta}-\beta)|\mathbf{X} \sim \chi^2_p$$
- $$(\hat{\beta}-\beta)^T \frac{\mathbf{X}^T\mathbf{X}}{p\hat{\sigma}^2}(\hat{\beta}-\beta)|\mathbf{X} \sim F^p_{n-p}$$
- Let be $\mathbf{K}$ a $q \times p$ matrix of range q,

$$(\mathbf{K}\hat{\beta}-\mathbf{K}\beta)^T \frac{\left[\mathbf{K}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{K}^T\right]^{-1}}{q\hat{\sigma}^2}(\mathbf{K}\hat{\beta}-\mathbf{K}\beta)|\mathbf{X} \sim F^q_{n-p}$$

Confidence Intervals

- A $(1-\alpha)\times 100\%$ confidence interval for $a^T\beta$ is given by

$$a^T\hat{\beta} \pm t_{n-p,\alpha/2}\sqrt{\hat{\sigma}^2 a^T(\mathbf{X}^T\mathbf{X})^{-1}a}$$

- A $(1-\alpha)\times 100\%$ confidence interval for β_j is given by

$$\hat{\beta}_j \pm t_{n-p,\alpha/2}\sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{j,j}}$$

where $\beta = [\beta_1,\ldots,\beta_p]^T$

Confidence Regions

- A $(1-\alpha) \times 100\%$ confidence region for β is given by

$$\left\{\beta : (\hat{\beta}-\beta)^T \frac{\mathbf{X}^T\mathbf{X}}{p\hat{\sigma}^2}(\hat{\beta}-\beta) \leq F^p_{n-p,1-\alpha}\right\}$$

- **Scheffé Intervals**. A $(1-\alpha) \times 100\%$ confidence region for $\mathbf{K}\beta$ is given by

$$\left\{\beta : (\mathbf{K}\hat{\beta}-\mathbf{K}\beta)^T \frac{\left[\mathbf{K}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{K}^T\right]^{-1}}{q\hat{\sigma}^2}(\mathbf{K}\hat{\beta}-\mathbf{K}\beta) \leq F^q_{n-p,1-\alpha}\right\}$$

where $\mathbf{K}$ is a $q \times p$ matrix of range q.

Prediction Interval

Remember that $Y = X^T\beta + \varepsilon$,

$$Y|X \sim \mathcal{N}(X^T\beta, \sigma^2)$$

and $\widehat{Y} = X^T\hat{\beta}$,

$$\widehat{Y}|(X, \mathbf{X}) \sim \mathcal{N}(X^T\beta, \sigma^2 X^T(\mathbf{X}^T\mathbf{X})^{-1}X)$$

Because $Y \perp \hat{Y}|\mathbf{X}$, then

$$Y - \widehat{Y}|(X, \mathbf{X}) \sim \mathcal{N}(0, \sigma^2(1 + X^T(\mathbf{X}^T\mathbf{X})^{-1}X))$$

Therefore,

$$\frac{Y - \widehat{Y}}{\sqrt{\hat{\sigma}^2(1 + X^T(\mathbf{X}^T\mathbf{X})^{-1}X)}}|(X, \mathbf{X}) \sim t_{n-p}$$

- A $(1-\alpha) \times 100\%$ prediction interval for Y is given by

$$X^T\hat{\beta} \pm t_{n-p,\alpha/2}\sqrt{\hat{\sigma}^2(1 + X^T(\mathbf{X}^T\mathbf{X})^{-1}X)}$$

Hypothesis Test

- Reject $H : \beta_j = m$ if

$$\frac{\left|\hat{\beta}_j - m\right|}{\sqrt{\hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}_{j,j}}} > t_{n-p,1-\alpha/2}$$

- Reject $H : a^T \beta_j = m$ if

$$\frac{\left|a^T \hat{\beta} - m\right|}{\sqrt{\hat{\sigma}^2 a^T (\mathbf{X}^T \mathbf{X})^{-1} a}} > t_{n-p,1-\alpha/2}$$

- Reject $H : \mathbf{K}\beta = \mathbf{m}$ if

$$(\mathbf{K}\hat{\beta} - \mathbf{m})^T \frac{\left[\mathbf{K}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{K}^T\right]^{-1}}{q \hat{\sigma}^2} (\mathbf{K}\hat{\beta} - \mathbf{m}) > F^q_{n-p,1-\alpha}$$