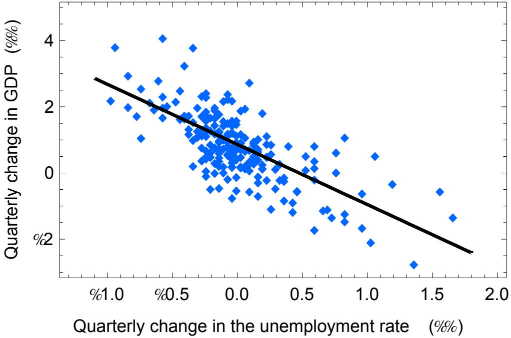
Linear regression

Eléments de Statistiques pour les Data Sciences

Simple linear regression



In economics, Okun's law is an empirical relationship between the increase in unemployment rate x and the increase in GDP y.

In statistics x are y are called

y the *response* (or *dependent variable*)

x the *explanatory* (or *independent*) *variable*

- What is the "best" linear function of x, so of the form $ax + b$, to approximate y?
- We will define "the best" as the one which minimizes the *mean squared error* (MSE).

This is the problem of linear regression. We talk about *simple* linear regression when there is a single explanatory variable.

Simple linear regression from a sample : statement

We consider a collection of observations (x_i, y_i) and consider the question: What is the linear (or more precisely affine) transformation of x that best approximates y in the least square sense?

$$\min_{a,b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (y_i - a\, x_i - b)^2$$

$$\min_{a,b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (y_i - a\, x_i)^2 - 2b\, \frac{1}{n} \sum_{i=1}^{n} (y_i - a\, x_i) + b^2$$

so that $\hat{b} = \bar{y} - a\bar{x}$ for a given value of a. Replacing b by its optimal value we get:

$$\min_{a,b\in\mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (\breve{y}_i - a\breve{x}_i)^2 \qquad \text{with} \quad \breve{x}_i := x_i - \bar{x},\ \breve{y}_i := y_i - \bar{y}.$$

$$\min_{a\in\mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \breve{y}_i^2 - 2a\, \frac{1}{n} \sum_{i=1}^{n} \breve{x}_i \breve{y}_i + a^2 \frac{1}{n} \sum_{i=1}^{n} \breve{x}_i^2.$$

Simple linear regression from a sample : statement

$$\min_{a\in\mathbb{R}} \tfrac{1}{n-1}\sum_{i=1}^{n} \breve{y}_i^2 - 2a\,\tfrac{1}{n-1}\sum_{i=1}^{n} \breve{x}_i\,\breve{y}_i + a^2\,\tfrac{1}{n-1}\sum_{i=1}^{n} \breve{x}_i^2$$

We consider the sample variances and covariance

$$s_x^2 := \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2, \quad s_y^2 := \tfrac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2, \quad s_{xy} := \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

and the correlation coefficient $\quad r := \dfrac{s_{xy}}{s_x\,s_y}.$

We can rewrite the optimization problem as

$$\min_{a\in\mathbb{R}} s_y^2 - 2a\,s_{xy} + a^2 s_x^2, \qquad \text{so that} \quad \hat{a} = \frac{s_{xy}}{s_x^2} = \frac{r\,s_x s_y}{s_x^2} = r\,\frac{s_y}{s_x}.$$

Simple linear regression in the sample case: solution

We found that the best affine function of X to approximate Y in the least square sense is of the form

$$\hat{y} := \hat{f}(x) := \hat{a}x + \hat{b} \qquad \text{with} \qquad \hat{a} = r\,\frac{s_y}{s_x}, \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

So that for any value x we have the *estimated response* $\hat{y}$.

$$\hat{y} = \hat{f}(x) = \bar{y} + r\,\frac{s_y}{s_x}\,(x - \bar{x}).$$

If we consider the residual $e_i := y_i - f^*(x_i)$, then we can write

$$y_i = \hat{y}_i + e_i = \bar{y} + r\,s_y\,\frac{x_i - \bar{x}}{s_x} + e_i.$$

Properties of the residuals $e_i = y_i - \hat{y}_i$

Given that $\hat{b} = \bar{y} - \hat{a}\bar{x}$, we have $\hat{y}_i - \bar{y} = \hat{a}x_i + \hat{b} - \bar{y} = \hat{a}(x_i - \bar{x})$, we have

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (\hat{y}_i - y_i) = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} \hat{a}(x_i - \bar{x}) = 0.$$

$$\begin{aligned} \sum_{i=1}^{n} e_i x_i &= \sum_{i=1}^{n} e_i(x_i - \bar{x}) = \sum_{i=1}^{n} (\hat{y}_i - y_i)(x_i - \bar{x}) = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})(x_i - \bar{x}) - \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \\ &= \sum_{i=1}^{n} \hat{a}(x_i - \bar{x})(x_i - \bar{x}) - (n-1) r s_x s_y = (n-1)\hat{a}^2 s_x^2 - (n-1) r s_x s_y = 0. \end{aligned}$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y} + e_i)^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} e_i^2,$$

but $$\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) = \hat{a}\sum_{i=1}^{n} e_i(x_i - \bar{x}) = 0$$

Summary: properties of the residuals $e_i = y_i - \hat{y}_i$

Let $(x_1, y_1), \ldots, (x_n, y_y)$ be a collection of datapoints for linear regression.

Let

- s_x^2 and s_y^2 be the *sample variances*.
- $s_{xy} := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ the *sample covariance*.
- $r := \frac{s_{xy}}{s_x s_y}$ the *sample correlation*.

Let $\hat{a} := r\,\frac{s_y}{s_x}, \quad \hat{b} := \bar{y} - \hat{a}\bar{x}, \quad \hat{y}_i := \hat{a} x_i + \hat{b}.$

Then the *residuals* $e_i := y_i - \hat{y}_i$ satisfy the following properties:

- They are *centered*: $\bar{e} = \frac{1}{n} \displaystyle\sum_{i=1}^{n} e_i = 0.$
- They are *empirically decorrelated* from the x_i: $\displaystyle\sum_{i=1}^{n} e_i (x_i - \bar{x}) = 0$

Empirical variances of the estimated $\hat{y}_i$ and of the residuals e_i

We have $\bar{\hat{y}} = \bar{y} - \bar{e} = \bar{y}$. So the sample variance of $\hat{y}_i$ is

$$\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\hat{a}^2(x_i - \bar{x})^2 = r^2\frac{s_y^2}{s_x^2}s_x^2 = r^2 s_y^2.$$

But we have proven that

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \frac{1}{n-1}\sum_{i=1}^{n}e_i^2,$$

so that

$$\frac{1}{n-1}\sum_{i=1}^{n}e_i^2 = s_y^2 - r^2 s_y^2 = s_y^2\left(1 - r^2\right).$$

Pythagoras and a decomposition between explained and residual variance

$$\overbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}^{(n-1)s_y^2} = \overbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}^{(n-1)r^2\,s_y^2} + \overbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}^{(n-1)(1-r^2)\,s_y^2}$$

$$\text{TSS} = \text{ESS} + \text{RSS}$$

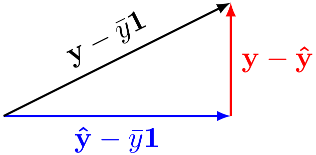
$$\text{Total SoS} = \text{Explained SoS} + \text{Residual SoS}$$

with SoS=sum of squares.

Let

- $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$
- $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^\top \in \mathbb{R}^n$,
- $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$.

The decomposition of the TSS corresponds to the Pythagorean triangle:



Note that

$$\hat{\mathbf{y}} - \bar{y}\mathbf{1} = \hat{a}(\mathbf{x} - \bar{x}\mathbf{1}) = r\frac{s_y}{s_x}(\mathbf{x} - \bar{x}\mathbf{1}).$$

Coefficient of determination

The *coefficient of determination* noted R^2 is defined as the fraction of the variance explained by the *explanatory variable*

$$R^2 := \frac{\mathsf{ESS}}{\mathsf{TSS}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{r^2 s_y^2}{s_y^2} = r^2.$$

So the coefficient of determination is the *square of the correlation coefficient* between x and y in the data.

Linear regression with a vector of explanatory variables

Given a dataset

$$\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\},$$

with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, we consider

- the *design matrix* $\boldsymbol{X}$ and
- the vector of *responses* $\boldsymbol{y}$

defined as

$$\boldsymbol{X} = \begin{bmatrix} \text{---} & \mathbf{x}_1^\top & \text{---} \\ \text{---} & \mathbf{x}_2^\top & \text{---} \\ & \vdots & \\ \text{---} & \mathbf{x}_n^\top & \text{---} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Remark: most of the time it is relevant to

- center the data: $\mathbf{x}_i^{\mathsf{c}} = \mathbf{x}_i - \bar{\mathbf{x}}$
- normalize via e.g. $x_{ij}^{\mathsf{s}} = x_{ij}^{\mathsf{c}} / \widehat{\sigma}_j$ or mapping $\mathbf{x}_{ij}^{\mathsf{c}}$ to $[0, 1]$, etc

Linear regression *aka* ordinary least square regression (OLS)

Given a dataset $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ we have

$$\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} — & \mathbf{x}_1^\top & — \\ — & \mathbf{x}_2^\top & — \\ — & \vdots & — \\ — & \mathbf{x}_n^\top & — \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} y_1 - \mathbf{x}_1^\top \boldsymbol{\beta} \\ y_2 - \mathbf{x}_2^\top \boldsymbol{\beta} \\ \vdots \\ y_n - \mathbf{x}_n^\top \boldsymbol{\beta} \end{bmatrix}$$

So that we have

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 = \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

with

- the vector of responses $\boldsymbol{y}^\top = (y_1, \ldots, y_n) \in \mathbb{R}^n$
- the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ whose ith row is equal to $\mathbf{x}_i^\top$.

Solving linear regression

We can rewrite the MSE as $\frac{1}{n}Q(\boldsymbol{\beta})$ with

$$Q(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - 2\,\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \|\boldsymbol{y}\|^2.$$

A minimum has to be stationary point, i.e., such that $\nabla Q(\boldsymbol{\beta}) = 0$. To compute the gradient, we can use the property that for Q differentiable, we have

$$Q(\boldsymbol{\beta} + \boldsymbol{h}) = Q(\boldsymbol{\beta}) + \nabla Q(\boldsymbol{\beta})^\top \boldsymbol{h} + o(\|\boldsymbol{h}\|),$$

where $o(\|\boldsymbol{h}\|)$ is a higher order term in $\boldsymbol{h}$. In our case we have

$$Q(\boldsymbol{\beta} + \boldsymbol{h}) = Q(\boldsymbol{\beta}) + \boldsymbol{h}^\top \boldsymbol{X}^\top \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\,\boldsymbol{h} + \boldsymbol{h}^\top \boldsymbol{X}^\top \boldsymbol{X}\,\boldsymbol{h} - 2\,\boldsymbol{h}^\top \boldsymbol{X}^\top \boldsymbol{y}$$

from which we deduce that $\nabla Q(\boldsymbol{\beta}) = 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - 2\,\boldsymbol{X}^\top \boldsymbol{y}$.

Normal equations

We have thus established that the stationary points of Q satisfy the

Normal equations: $\boxed{\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{0}}$

Given that $\boldsymbol{X}^\top \boldsymbol{X}$ is a positive semi-definite matrix, the curvature of the function is non-negative everywhere and so all stationary points are global minima. The normal equation thus characterize exactly the vectors $\boldsymbol{\beta}$ which solutions of the linear regression problem.

If $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, then there is a unique solution to the normal equations and and $\widehat{\boldsymbol{\beta}}$ is given by:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Remarks:

- $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible iff the columns of $\boldsymbol{X}$ are linearly independent
- they are linearly dependent that one of them is a linear combination of the others.
- $\boldsymbol{X}^\top \boldsymbol{X}$ is never invertible for $p > n$.

Linear or affine regression?

Compare the linear vs affine functions of $\mathbf{x}$

$$f_{\boldsymbol{\beta}}(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x} \qquad \text{vs} \qquad f_{\boldsymbol{\beta},b}(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x} + b = \widetilde{\boldsymbol{\beta}}^\top \widetilde{\mathbf{x}}$$

With a new definition of the variables

$$\widetilde{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{\beta} \\ b \end{bmatrix} \quad \text{and} \quad \widetilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$$

we can rewrite an affine model in dimension p as a linear model in dimension $p+1$, in which the last column of the design matrix is $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$.

Exercise: What is the value of $\hat{b}$ if the data is centered?

Gaussian conditional model and linear regression

We decide to model the conditional distribution of Y given X by

$$Y \mid X \sim \mathcal{N}(\boldsymbol{\beta}^\top X + b, \sigma^2)$$

or equivalently $Y = \boldsymbol{\beta}^\top X + \varepsilon \quad$ with $\quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Given a dataset $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ we can consider the likelihood of $\boldsymbol{\beta}$ in the conditional model of Y given X and estimate $\boldsymbol{\beta}$ using the maximum likelihood principle.

Likelihood for one pair

$$p(y_i \mid \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2}{\sigma^2}\right)$$

Negative log-likelihood

$$-\ell(\boldsymbol{\beta}, \sigma^2) = -\sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2}{\sigma^2}.$$

Gaussian conditional model and linear regression

$$\min_{\sigma^2,\boldsymbol{\beta}} \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2}{\sigma^2}$$

The minimization problem in $\boldsymbol{\beta}$

$$\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

that we recognize as the usual linear regression.
Optimizing over σ^2, we find:

$$\widehat{\sigma}^2_{\mathsf{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\boldsymbol{\beta}}_{\mathsf{MLE}}^\top \mathbf{x}_i)^2$$

Properties if the data is actually Gaussian

Assume that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with

Full column rank *fixed* design: $\mathrm{rank}(\boldsymbol{X}) = p$ (which implies $n \geq p$).

I.i.d. centered **Gaussian** noise: $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$

then

- $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\beta}^*, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1})$
- $S^2 = \frac{1}{n-p} \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_2^2 \sim \frac{\sigma^2}{n-p} \chi^2_{n-p}$
- $\widehat{\boldsymbol{\beta}}$ and S^2 are independent

All of these are used for

- t-test and to construct confidence intervals
- Only valid if the data is Gaussian (= model is well-specified)