Lecture 14:

supervised learning → Regression → $y \in \mathbb{R}$
supervised learning → Classificatio

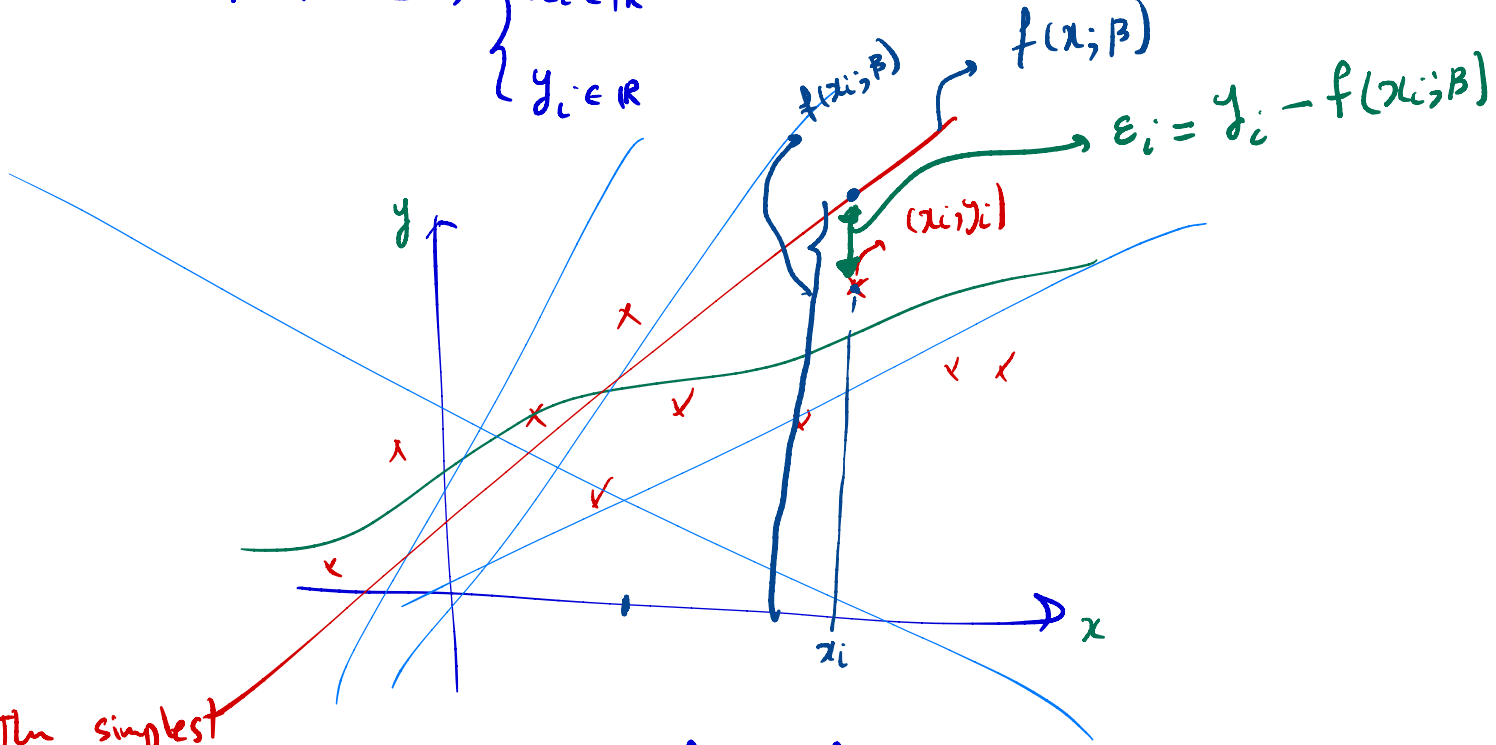
Data : $\{(x_i, y_i)\}_{i=1}^{n}$

$x_i \in \mathbb{R}^p$

$y_i \in \mathbb{R}$ → $y_i \approx f(x_i)$

f will be used for prediction on newly unseen data points

$p = 1$ → $x_i \in \mathbb{R}$, $y_i \in \mathbb{R}$



The simplest parametric family: $y = f(x;\beta) = \hat{\beta}_0 + \hat{\beta}_1 x$

our general methodology:

Come up with a parametric family of predictors

$f(x; \beta)$ ← parameters

↓

Define an objective function that quantifies how good a parameter is w.r.t the data set. Minimizing the objective will give us the best "fit" to the data set

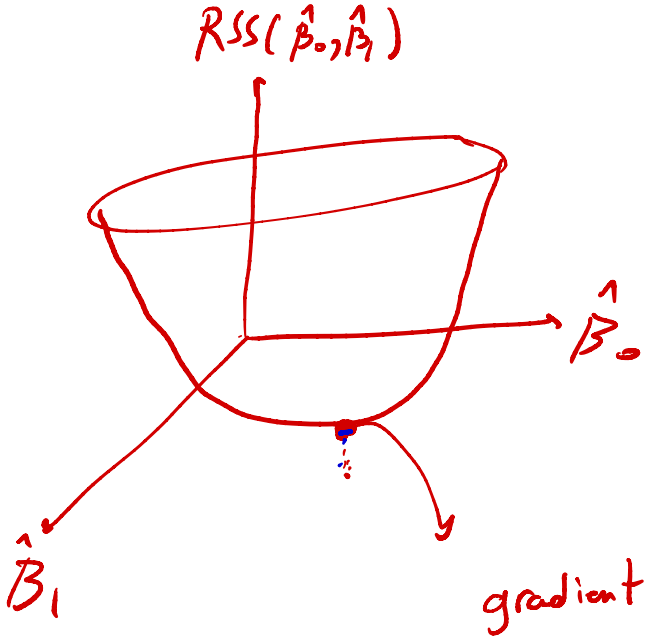
---

objective for the linear model:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \varepsilon_i^2 \triangleq \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2$$

$$RSS(\hat{\beta}_0, \hat{\beta}_1)$$

Residual Sum of Squares



gradient of RSS at the minimizer is zero.

These two equations should hold at the minimizer.

$$\frac{d\, RSS(\hat{\beta}_0, \hat{\beta}_1)}{d\hat{\beta}_0} = 0$$

$$\frac{d}{d\hat{\beta}_1} RSS(\hat{\beta}_0, \hat{\beta}_1) = 0$$

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i \right)^2$$

$$\frac{d}{d\hat{\beta}_0} RSS(\hat{\beta}_0, \hat{\beta}_1) \qquad (1)$$

$$= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{d}{d\hat{\beta}_1} RSS(\hat{\beta}_0, \hat{\beta}_1) \qquad (2)$$

$$= -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$(1) \longrightarrow \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\hookrightarrow \div n \qquad \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_0 - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} = 0$$

$$\bar{y} \qquad\qquad \bar{x}$$

$$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}} \qquad (3)$$

(equivalent to (1))

$$(2) \Rightarrow \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\div n \Rightarrow \frac{\sum x_i y_i}{n} - \underbrace{\hat{\beta}_0 \bar{x}} - \hat{\beta}_1 \frac{\sum x_i^2}{n} = 0$$

use (3)

$$\Rightarrow \frac{\sum x_i y_i}{n} - (\bar{y} - \hat{\beta}_1 \bar{x})\bar{x} - \hat{\beta}_1 \frac{\sum x_i^2}{n} = 0$$

$$\frac{\sum x_i y_i}{n} - \bar{x}\bar{y} = \hat{\beta}_1 \left(\frac{\sum x_i^2}{n} - \bar{x}^2 \right)$$

Two relations (excercise)

$$\frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \leftarrow$$

$$\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

(3) $\rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\Rightarrow \boxed{\hat{\beta}_0 = \bar{y} - \bar{x} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Question: How good are the above estimators $\hat{\beta}_0, \hat{\beta}_1$?

To evaluate the performance of our estimators of $\hat{\beta}_0, \hat{\beta}_1$, let's assume that the data is generated according to the following probabilistic model:

$$y = \beta_0^{true} + \beta_1^{true} x + \varepsilon$$

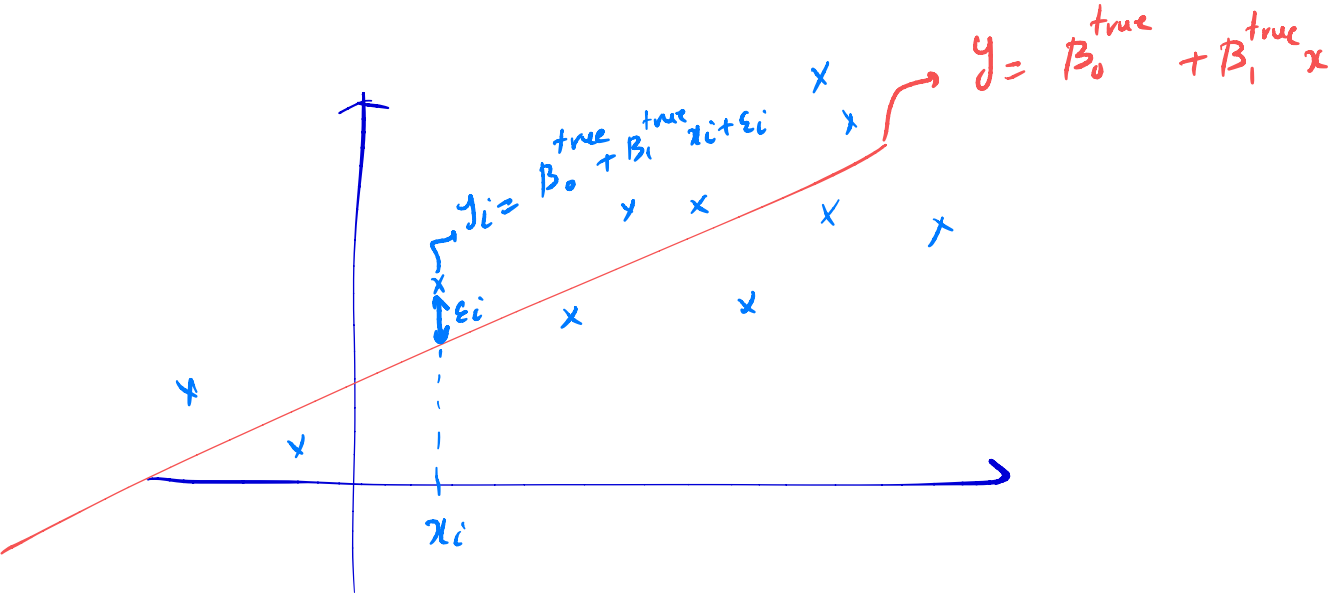
ε: noise
- zero-mean
- independent
- $\text{var}[\varepsilon^2] = \sigma^2$

Now, assume that we generate data points $\{(x_i, y_i)\}_{i=1}^{n}$ from the above model. Let $\hat{\beta}_0, \hat{\beta}_1$ be the solutions of the regression problem.

as $n \to \infty$

$\hat{\beta}_0 \to \beta_0^{true}$, $\hat{\beta}_0$ is an estimator of: β_0^{true}

$\hat{\beta}_1 \to \beta_1^{true}$, $\hat{\beta}_1$ is an estimator of: β_1^{true}



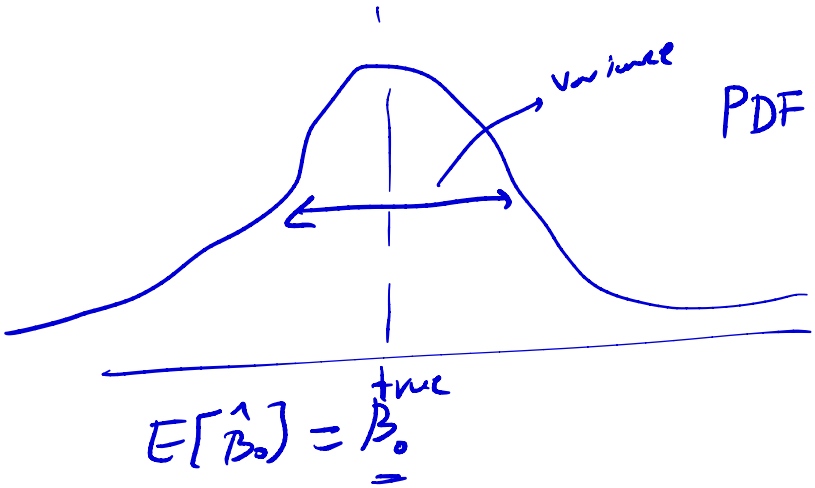
$\hat{\beta}_0 \longrightarrow \beta_0^{true}$

$\hat{\beta}_1 \longrightarrow \beta_1^{true}$

- In order to quantify the performance of the estimates $\hat{\beta}_0, \hat{\beta}_1$, we need to compute their expectation (Bias) and Variance.
- The expectation and the variance of $\hat{\beta}_0$ (and $\hat{\beta}_1$) will tell us how close $\hat{\beta}_0$ (and $\hat{\beta}_1$) are w.r.t their true value β_0^{true} (and β_1^{true})
- Note that, because of the fact that the noises ε_i's are random variables, our

data set $\{(x_i, y_i)\}$ is also random.

This means that $\hat{\beta}_0$, $\hat{\beta}_1$, which depend on the data, are random variables as well.



We'll now compute the expectation of $\hat{\beta}_0$, $\hat{\beta}_1$ and show that they are unbiased estimators of β_0^{true} and β_1^{true}, respectively.

$$E[\hat{\beta}_1] \stackrel{?}{=} \beta_1^{true}$$

$\downarrow$

$$E\left[\hat{\beta}_1\right] = E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$$

(note: $y_i = \beta_0^{true} + \beta_1^{true} x_i + \varepsilon_i$)

ε_i: zero-mean, independent of everything else

$\forall i$: $y_i = \beta_0^{true} + \beta_1^{true} x_i + \varepsilon_i$

$$\Downarrow$$

$$\bar{y} = \beta_0^{true} + \beta_1^{true} \bar{x} + \bar{\varepsilon} \quad (1) \qquad \left(\bar{\varepsilon} = \frac{\varepsilon_1 + \cdots + \varepsilon_n}{n}\right)$$

$$E\left[\hat{\beta}_1\right] = E\left[\frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}\right]$$

$$\overset{(1)}{=} E\left[\frac{\sum_i (x_i-\bar{x})\left(\beta_1^{true}(x_i-\bar{x})+\varepsilon_i-\bar{\varepsilon}\right)}{\sum_i (x_i-\bar{x})^2}\right]$$

$$(y_i - \bar{y})$$

$$(\beta_0^{true} + \beta_1^{true} x_i + \varepsilon_i) - (\beta_0^{true} + \beta_1^{true} \bar{x} + \bar{\varepsilon})$$

$$= \beta_1^{true} (x_i - \bar{x}) + \varepsilon_i - \bar{\varepsilon}$$

$$E[\hat{\beta}_1] = E\left[\frac{\beta_1^{true} \sum_i (x_i - \bar{x})^2 + \sum (\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right]$$

$$= \beta_1^{true} + E\left[\frac{\sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right]$$

noises were independent of everything else

$$= \beta_1^{true} + \sum_{i=1}^{n} E\left[\frac{(\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right]$$

$$= \beta_1^{true} + \sum_{i=1}^{n} E[\varepsilon_i - \bar{\varepsilon}] \, E\left[\frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right]$$

$E[\varepsilon_i] - E[\bar{\varepsilon}]$ → 0 (each)

$$\Rightarrow \quad E[\hat{\beta}_1] = \beta_1^{true}$$

In general:

(1) $E[\hat{\beta}_1] = \beta_1^{true}$

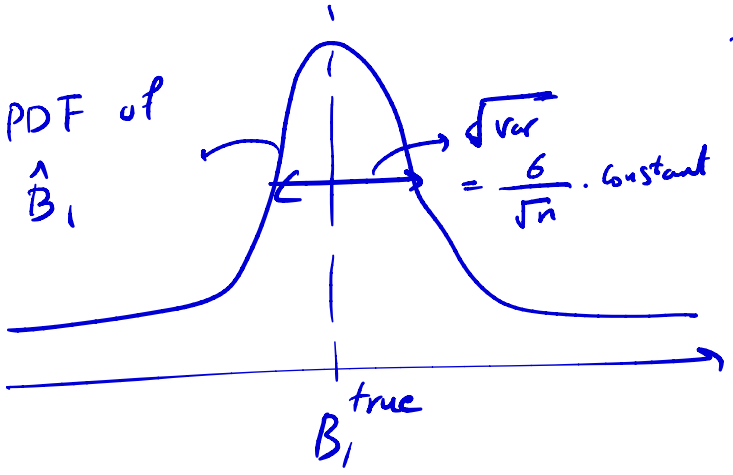
(2) $E[\hat{\beta}_0] = \beta_0^{true}$

(3) $Var(\hat{\beta}_0) = E\left[(\hat{\beta}_0 - \beta_1^{true})^2\right]$

$$\left(Var(\varepsilon_i) = \sigma^2\right) \qquad = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

(4) $Var(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \dfrac{\sigma^2}{n\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)}$

$$= \frac{\sigma^2}{n \cdot (\text{empirical variance of } \{x_i\})}$$



So far we've only been talking about the 1-dim case ($p=1$). Let's now consider the general case:

$$\{(x_i, y_i)\}_{i=1}^{n} \qquad x_i \in \mathbb{R}^p, \; y_i \in \mathbb{R}$$

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^p, \; y_i \in \mathbb{R}$$

simplest parametric family (linear):

$$y_i \simeq \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}}_{f(x_i, \vec{\beta})}$$