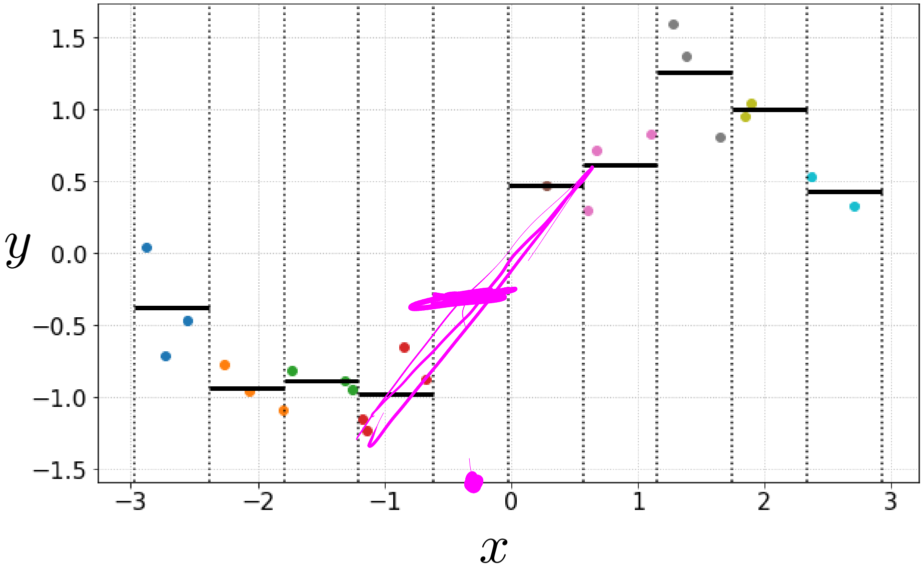
Statistics and Data Science for Engineers
E178 / ME276DS

Linear regression
Part 1

$$\mathcal{D} = \{(x_i, y_i)\}_{24}$$

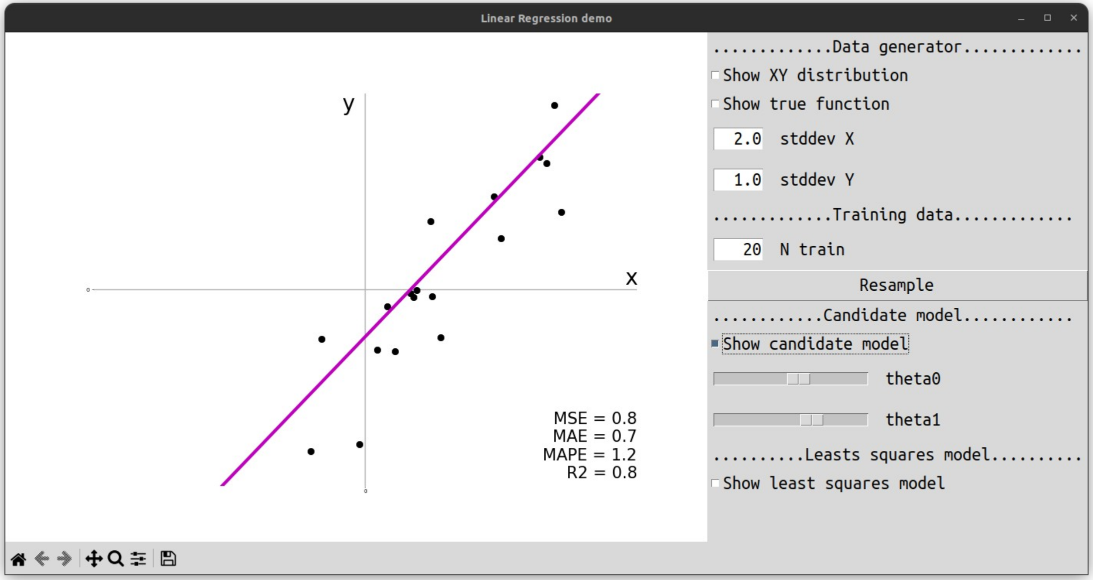
x	y
2.704286	0.323832
-2.063888	-0.954098
0.606690	0.300653
-2.876493	0.043129
-1.725965	-0.816896
-1.174547	-1.156533
0.671117	0.720568
-1.247132	-0.946372
-1.801957	-1.087639
1.850384	0.952331
1.105398	0.829523
-2.267771	-0.778359
0.280262	0.472226
1.650797	0.808036
2.368964	0.530195
-2.728636	-0.707794
-0.667936	-0.880451
-1.314393	-0.891488
-2.552696	-0.468816
1.892769	1.047608
-0.849206	-0.653065
-1.134106	-1.231050
1.377637	1.367601
1.279469	1.590128

K-bins

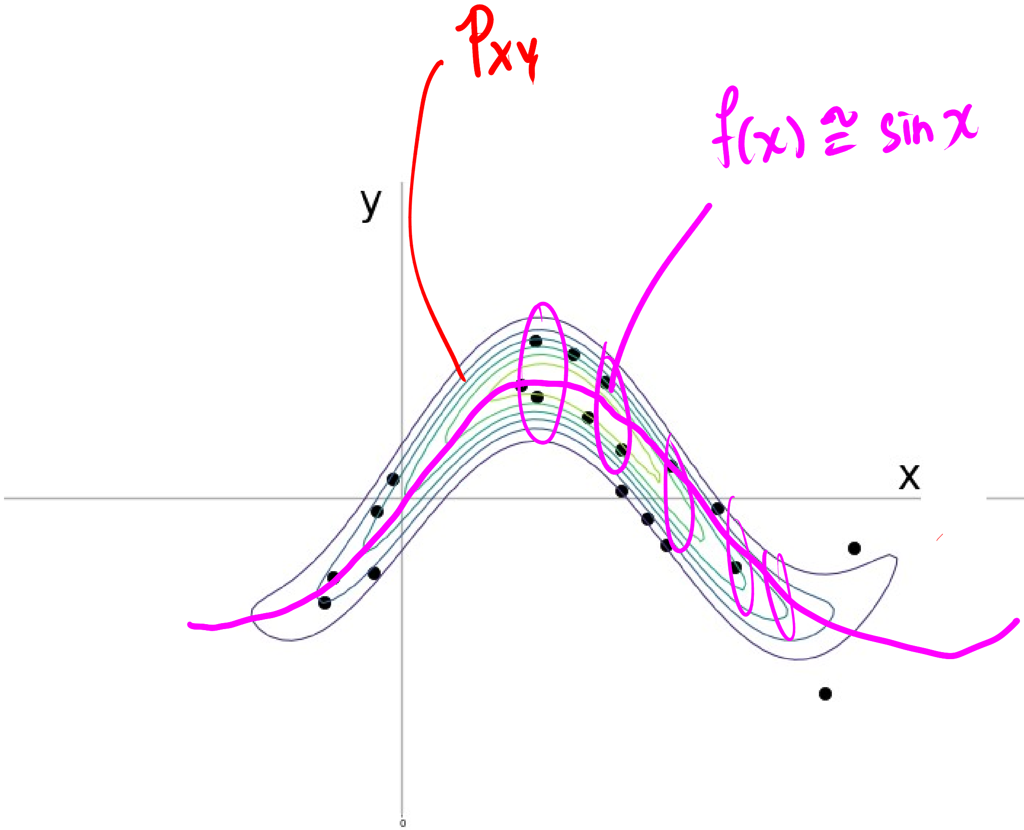


Interpolations.

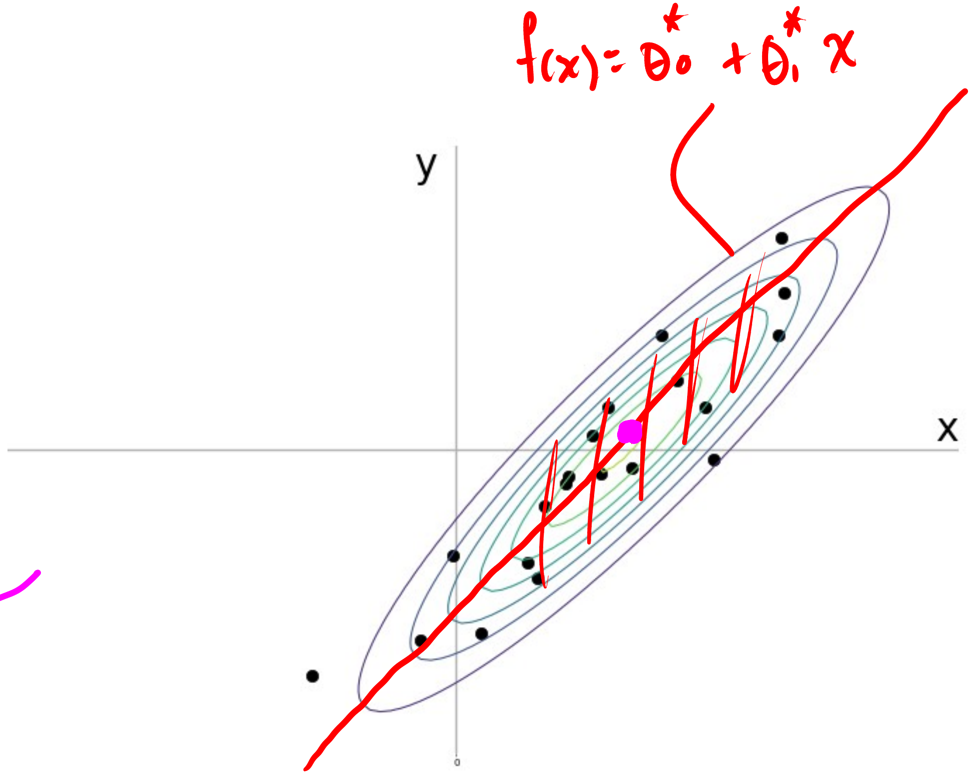
Simple linear regression demo



$$y = \theta_0 + x\theta_1$$



$$y_i = \sin x + \varepsilon_i$$



$$y = \theta_0 + \theta_1 x + \varepsilon_i$$

$x_i \in \mathbb{R}$, $y_i \in \mathbb{R}$.

$\mathcal{D} = \{(x_i, y_i)\}_N$

Sample (point) estimates

$x \in \mathbb{R}^D$ $D=1$.

$$\left\{\begin{array}{l}\hat{\mu}_X = \dfrac{1}{N}\sum x_i \quad \ldots \quad E[X] \\ \hat{\mu}_Y = \dfrac{1}{N}\sum y_i \quad \ldots \quad E[Y]. \end{array}\right.$$

$$\hat{\sigma}_X^2 = \frac{1}{N-1}\sum(x_i - \hat{\mu}_X)^2 \ldots Var[X]$$

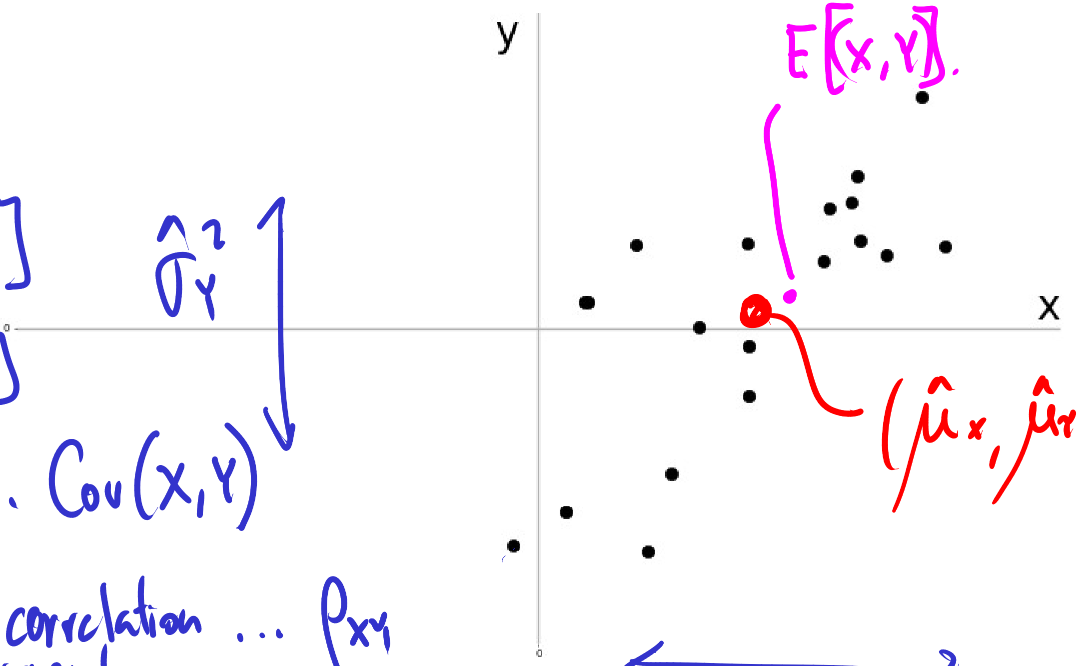
$$\hat{\sigma}_Y^2 = \frac{1}{N-1}\sum(y_i - \hat{\mu}_Y)^2 \ldots Var[Y]$$

$$\hat{\sigma}_{XY} = \frac{1}{N-1}\sum_{i}^{N}(x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y) \ldots Cov(X,Y)$$

$$r = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} \in [-1, 1] \ldots \text{sample correlation coefficient} \ldots \rho_{XY}$$

$\hat{\sigma}_Y^2$

$E[(X,Y)]$.



$(\hat{\mu}_X, \hat{\mu}_Y)$

$\hat{\sigma}_X^2$

$\therefore \rho_{XY} = \dfrac{Cov(X,Y)}{Std(X)\, Std(Y)}$

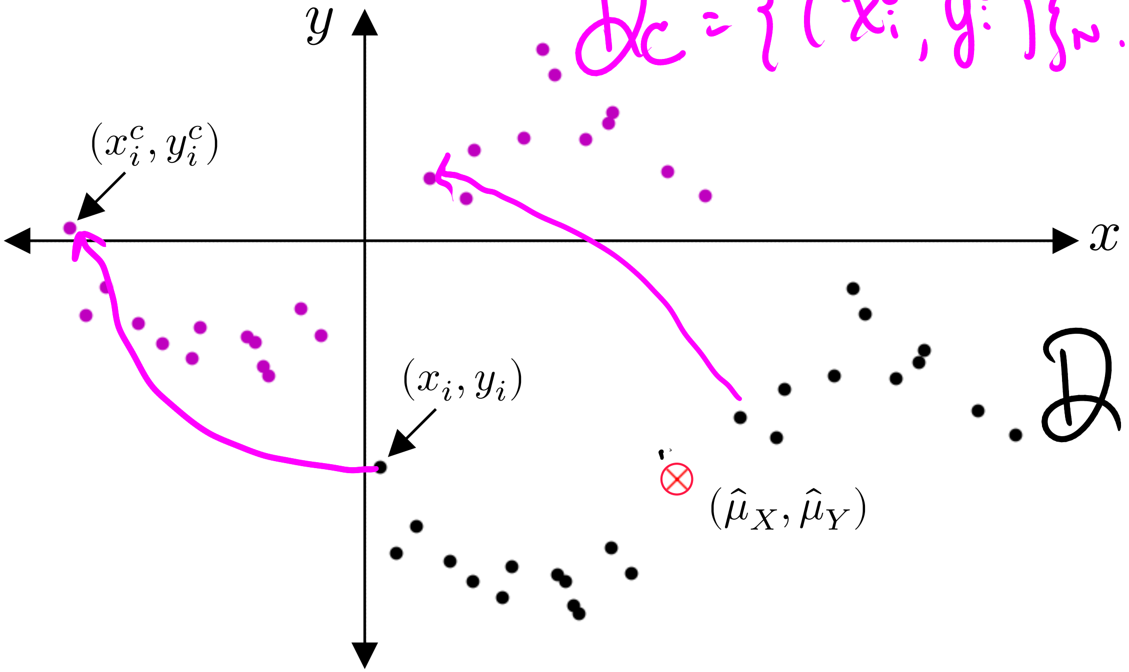
All sums are over $i = 1, \ldots, N$.

Centering transformation

$$x_i^c = x_i - \hat{\mu}_X$$

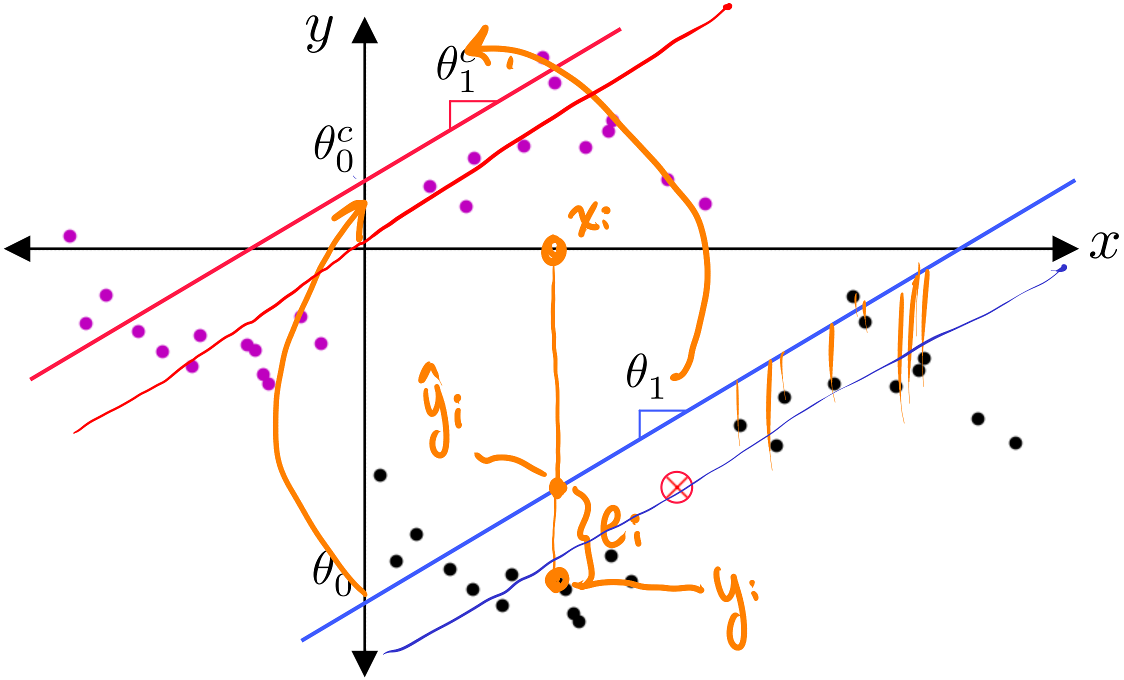
$$y_i^c = y_i - \hat{\mu}_Y$$

$\mathcal{D}_c = \{(x_i^c, y_i^c)\}_n.$



Linear model

$$\hat{y}_i = h(x_i; \theta_0, \theta_1) = \theta_0 + x_i\theta_1 \qquad i = 1 \dots N$$



Centered linear model

$$\hat{y}_i^c = \theta_0^c + x_i^c\theta_1^c$$

$$\begin{cases} \theta_1^c = \theta_1 \\ \theta_0^c = \theta_0 - \hat{\mu}_Y + \hat{\mu}_X \theta_1 \end{cases}$$

optimal parameters.

Optimization problem

squared — $\hat{y}_i$ — "least squares"

$$(\hat{\theta}_0, \hat{\theta}_1) = \underset{(\theta_0,\theta_1)\in\mathbb{R}^2}{\text{argmin}} \sum_{i=1}^{N} L_2(y_i, h(x_i;\theta_0,\theta_1))$$

$$= \underset{(\theta_0,\theta_1)\in\mathbb{R}^2}{\text{argmin}} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

$$= \underset{(\theta_0,\theta_1)\in\mathbb{R}^2}{\text{argmin}} \sum_{i=1}^{N} (\theta_0 + x_i\theta_1 - y_i)^2$$

Centered problem: $$(\hat{\theta}_0^c, \hat{\theta}_1^c) = \underset{(\theta_0,\theta_1)\in\mathbb{R}^2}{\text{argmin}} \boxed{\sum_{i=1}^{N} (\theta_0 + x_i^c\theta_1 - y_i^c)^2}$$

stationary $\Longleftrightarrow$ global optimum.

Centered problem: Stationary points

$$J(\theta_0, \theta_1) = \sum_{i=1}^{N} (\theta_0 + x_i^c \theta_1 - y_i^c)^2 .$$

$\frac{1}{N} \sum x_i^c = \hat{\mu}_x^c = 0.$

$\frac{1}{N} \sum y_i^c = \hat{\mu}_y^c = 0.$

- $$\frac{\partial J}{\partial \theta_0} = 2 \sum (\theta_0 + x_i^c \theta_1 - y_i^c)$$
$$= 2 \left(N\theta_0 + \left(\cancelto{0}{\sum x_i^c} \right) \theta_1 - \cancelto{0}{\sum y_i^c} \right)$$
$$= 2N\theta_0 = 0$$

$\hat{\theta}_0^c = 0 \Rightarrow$ L.S. regression line goes through C.O.M $(\hat{\mu}_x, \hat{\mu}_y)$.

- $$\frac{\partial J}{\partial \theta_1} = 2 \sum (\theta_0 + x_i^c \theta_1 - y_i^c) x_i^c$$
$$= 2 \left(\theta_0 \cancel{\sum x_i^c} + \theta_1 \sum (x_i^c)^2 - \sum x_i^c y_i^c \right)$$

$$(N-1)\,\hat{\sigma}_{xy}$$

$$\sum (x_i^c)^2 = \sum (x_i - \hat{\mu}_x)^2 = (N-1)\,\hat{\sigma}_x^2$$

$$= \theta_1 \hat{\sigma}_x^2 - \hat{\sigma}_{xy} = 0 \quad \longleftrightarrow \quad \hat{\theta}_1^c = \boxed{\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} = \hat{\theta}_1}$$

Simple linear regression: Solution

$$\hat{\theta}_1 = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}$$

$$\hat{\theta}_0 = \hat{\mu}_Y - \hat{\mu}_X \hat{\theta}_1$$

$$r = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} = \hat{\theta}_1 \cdot \frac{\hat{\sigma}_X}{\hat{\sigma}_Y} \longrightarrow \boxed{\hat{\theta}_1 = r \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}}$$

Simple linear regression: Performance

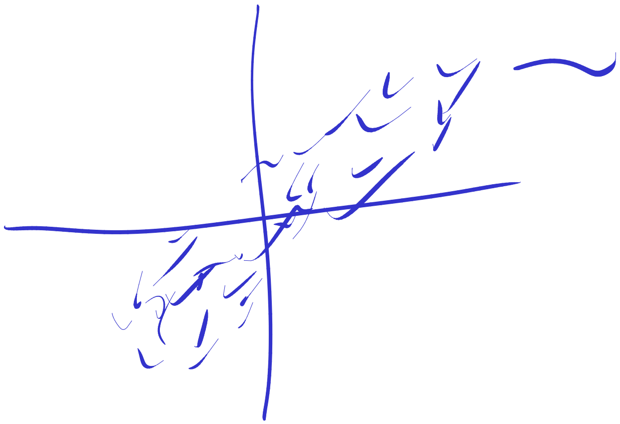
coefficient of determination: ....

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \hat{\mu}_Y)^2}$$

$$= \quad \vdots$$

$$= r^2$$

... square of the correlation.



$r \cong 0.7$

$R^2 = 0.49.$

Statistical properties of simple linear regression

x's (inputs) are fixed.

Assumed data generating process:

$$\underbrace{\theta_0^* + x_i\theta_1^*}_{\text{linear function}} + \varepsilon_i \quad (y_i = \ldots)$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

linear function

Gaussian noise.

$$\left\{\begin{aligned} \hat\theta_0 &= \hat\mu_Y - \hat\mu_X \hat\theta_1 \\ \hat\theta_1 &= r\frac{\hat\sigma_Y}{\hat\sigma_X} \end{aligned}\right.$$

$$\hat y_i = \hat\theta_0 + \hat\theta_1 \cdot x_i$$

point estimators of θ_0^*, θ_1^*, y_i^*

- unbiased?
- variance?

$\hat{\theta}_1$ as an estimate of θ_1^*

least squares.

$$\hat{\theta}_1 = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}$$

$$= \frac{1}{\hat{\sigma}_X^2} \frac{1}{N-1} \sum x_i^c y_i^c$$

$$= \frac{1}{(N-1)\hat{\sigma}_X^2} \sum x_i^c \left(\theta_1^* x_i^c + \varepsilon_i - \bar{\varepsilon}\right)$$

$$= \frac{1}{(N-1)\hat{\sigma}_X^2} \left(\theta_1^* \sum (x_i^c)^2 + \sum x_i^c \varepsilon_i - \bar{\varepsilon} \sum x_i^c\right)$$

$$= \frac{\theta_1^*}{\hat{\sigma}_X^2} \frac{\sum (x_i^c)^2}{(N-1)} + \frac{\sum x_i^c \varepsilon_i}{(N-1)\hat{\sigma}_X^2}$$

$$= \theta_1^* + \frac{\sum x_i^c \varepsilon_i}{(N-1)\hat{\sigma}_X^2}$$

everything is deterministic except the ε_i.

Define: $\bar{\varepsilon} = \frac{1}{N} \sum_{j=1}^{N} \varepsilon_j$

Use: $y_i^c = x_i^c \theta_1^* + \varepsilon_i - \bar{\varepsilon}$

(proved in the reader)

$\hat{\theta}_1$ as an estimate of θ_1^*

$\hat{\theta}_1 = \hat{\Theta}_1$

$\varepsilon_i = \mathcal{E}_i$

substitute in R.V.s.

$$\hat{\theta}_1 = \theta_1^* + \frac{\sum x_i^c \varepsilon_i}{(N-1)\hat{\sigma}_X^2} \quad \Rightarrow \quad \hat{\Theta}_1 = \theta_1^* + \frac{\sum x_i^c \mathcal{E}_i}{(N-1)\hat{\sigma}_X^2}$$

$$\begin{aligned} E\left[\hat{\Theta}_1\right] &= E\left[\theta_1^* + \frac{\sum x_i^c \mathcal{E}_i}{(N-1)\hat{\sigma}_X^2}\right] \\ &= \theta_1^* + \cancel{\frac{\sum x_i^c E\left[\mathcal{E}_i\right]}{(N-1)\hat{\sigma}_X^2}} \\ &= \theta_1^* \end{aligned}$$

$\therefore$ $\hat{\Theta}_1$ is an unbiased estimate of θ_1^*

$\hat{\theta}_1$ as an estimate of θ_1^*

$$Var[\hat{\Theta}_1] = Var\left[\theta_1^* + \frac{\sum x_i^c \mathcal{E}_i}{(N-1)\hat{\sigma}_X^2}\right]$$

$$= \sum \left(\frac{(x_i^c)^2}{(N-1)^2 \hat{\sigma}_X^4} Var[\mathcal{E}_i] \right)$$

$$= \frac{\sigma^2}{(N-1)^2 \hat{\sigma}_X^4} \sum (x_i^c)^2 \qquad \hat{\sigma}_X^2$$

$$= \frac{\sigma^2}{(N-1)\hat{\sigma}_X^2}$$

vertical variance

horizontal spread.

$\hat{\theta}_0$ as an estimate of θ_0^*

$\bar{\varepsilon} = \frac{1}{N} \sum \varepsilon_i$

$E[\bar{\varepsilon}] = 0 \quad Var[\bar{\varepsilon}] = \frac{\sigma^2}{N}$

Express in terms of well described quantities:

$$\hat{\theta}_0 = \theta_0^* + \hat{\mu}_X \theta_1^* + \bar{\varepsilon} - \hat{\mu}_X \hat{\theta}_1$$

(proved in the reader)

Substitute RVs:

$$\hat{\Theta}_0 = \theta_0^* + \hat{\mu}_X \theta_1^* + \bar{\mathcal{E}} - \hat{\mu}_X \hat{\Theta}_1$$

Statistical analysis:

$$E[\hat{\Theta}_0] = E\left[\theta_0^* + \hat{\mu}_X \theta_1^* + \bar{\mathcal{E}} - \hat{\mu}_X \hat{\Theta}_1\right]$$

(proved in the reader)

$$= \theta_0^*$$

... unbiased!

$$Var\left[\hat{\Theta}_0\right] = \frac{\sigma^2}{N} + \frac{\sigma^2 \hat{\mu}_X^2}{(N-1)\hat{\sigma}_X^2}$$

(proved in the reader)

$\hat{y}_i$ as an estimate of y_i^*

Express in terms of well described quantities:

$$\hat{y}_i = \theta_0^* + \hat{\mu}_X \theta_1^* + \bar{\varepsilon} + x_i^c \hat{\theta}_1$$

(proved in the reader)

RV.

Substitute RVs:

$$\hat{Y}_i = \theta_0^* + \hat{\mu}_X \theta_1^* + \bar{\mathcal{E}} + x_i^c \hat{\Theta}_1$$

Statistical analysis:

$$E[\hat{Y}_i] = \theta_0^* + x_i \theta_1^*$$

... also unbiased!

(proved in the reader)

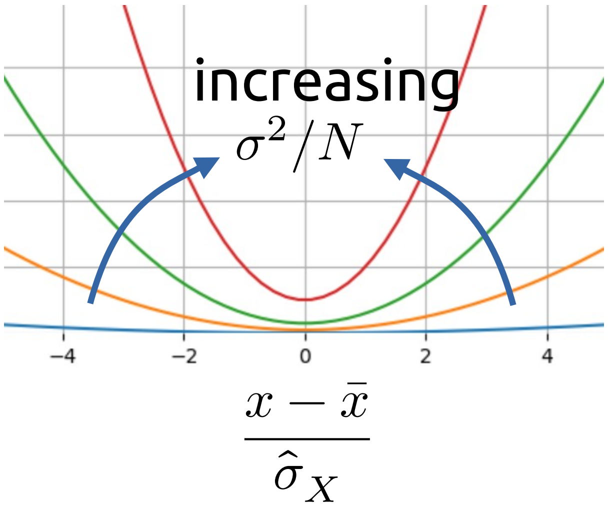
$$Var[\hat{Y}_i] = \frac{\sigma^2}{N} + \frac{\sigma^2 (x_i - \hat{\mu}_X)^2}{\underbrace{\sum_{j=1}^N (x_j - \hat{\mu}_X)^2}_{(N-1)\hat{\sigma}_x^2}}$$

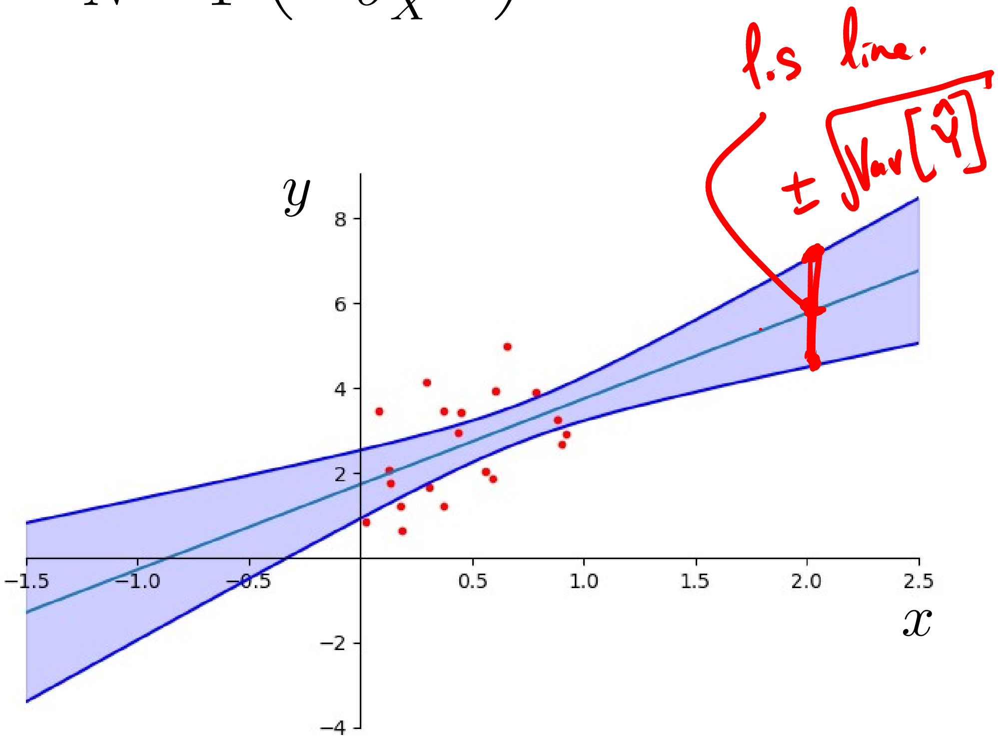
(proved in the reader)

rewrite

Prediction uncertainty in simple linear regression

$$\text{Std dev}[\hat{Y}] = \sqrt{Var[\hat{Y}]} = \sqrt{\frac{\sigma^2}{N} + \frac{\sigma^2}{N-1}\left(\frac{x-\hat{\mu}_X}{\hat{\sigma}_X}\right)^2}$$





$\hat{\theta}_0$: Know expected value and variance

$\hat{\theta}_1$ Gaussian!

$\hat{y}_i$

Know the distributions of these estimators!

Confidence intervals and hypothesis tests

90% confidence

$\hat{\theta}_0 \in [0.2, 0.3]$

$\hat{\theta}_1 \in [\cancel{1}.7, 2.2]$

- Find confidence interval for the parameters
- Find confidence interval for the output

$\hat{y}_i = \_\_\_\_ \pm \_\_\_\_$

- Hypothesis test:

$$H_0 : \quad \theta_d = 0$$
$$H_1 : \quad \theta_d \neq 0$$

$\cancel{d = 1 \ldots D}$ D = 1

"y does not depend on x"

"y does depend on x"

- Unknown σ^2:

If unknown:

$$\boxed{\hat{\sigma}^2 = \frac{1}{N - D} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$

1

D>1

and use t-test.

Multiple linear regression $(D > 1)$

$$\mathcal{D} = \{(x_i, y_i)\}_N = \{(x_i^1, \dots, x_i^D, y_i)\}_N$$

$$x_i = [x_i^1 \quad \dots \quad x_i^D] \in \mathbb{R}^{1\times D}$$

$$\mathbf{X} = \begin{bmatrix} x_1^1 & \dots & x_1^D \\ \vdots & & \vdots \\ x_N^1 & \dots & x_N^D \end{bmatrix} \in \mathbb{R}^{N\times D}$$

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^{N\times 1}$$

	x^1	x^2	x^3	...	x^D	y
0	0.247746	36.0	266.0	0.247746	88.019654	57.000823
1	0.179340	37.0	365.0	0.179340	27.211935	22.471227
2	0.956807	21.0	151.0	0.956807	97.456012	56.357366
3	0.869653	43.0	437.0	0.869653	43.203221	34.[illegible]475
4	0.825345	47.0	160.0	0.825345	98.933930	64.426163
5	0.331114	321.0	371.0	0.331114	8.257917	14.218720
6	0.765523	17.0	364.0	0.765523	96.696783	59.123769
7	0.956807	21.0	151.0	0.956807	97.456012	52.983470

$\mathbf{X}$ $\quad$ $\mathbf{Y}$

$\hat{\mu}_X$ $\quad$ $\hat{\mu}_Y$

$$\hat{\mu}_X = \frac{1}{N}\mathbf{1}_N^T\mathbf{X}$$

$$\hat{\mu}_Y = \frac{1}{N}\mathbf{1}_N^T\mathbf{Y}$$

Data-centering transformation:

$$\mathbf{X}^c = \mathbf{X} - \mathbf{1}_N \hat{\mu}_X \qquad \mathbf{1}_N^T \mathbf{X}^c = \mathbf{0}_D$$

$$\mathbf{Y}^c = \mathbf{Y} - \mathbf{1}_N \hat{\mu}_Y \qquad \mathbf{1}_N^T \mathbf{Y}^c = 0$$

Linear model:

$$\begin{aligned} \hat{y}_i &= \theta_0 + x_i^1 \theta_1 + \ldots + x_i^D \theta_D \\ &= \theta_0 + x_i \, \underline{\theta}_1 \end{aligned}$$

$$\underline{\theta}_1 = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta^D \end{bmatrix} \in \mathbb{R}^{D \times 1}$$

Matrix form: $\hat{\mathbf{Y}} = \mathbf{1}_N \theta_0 + \mathbf{X} \underline{\theta}_1$

Centered: $\hat{\mathbf{Y}}^c = \mathbf{1}_N \theta_0^c + \mathbf{X}^c \underline{\theta}_1^c$

Optimization problem

Original: $$(\widehat{\theta}_0, \widehat{\underline{\theta}}_1) = \underset{(\theta_0, \underline{\theta}_1) \in \mathbb{R}^{D+1}}{\operatorname{argmin}} \quad \sum_{i=1}^{N} (\theta_0 + x_i \, \underline{\theta}_1 - y_i)^2$$

Matrix: $$(\widehat{\theta}_0, \widehat{\underline{\theta}}_1) = \underset{(\theta_0, \underline{\theta}_1) \in \mathbb{R}^{D+1}}{\operatorname{argmin}} \left(\mathbf{1}_N \theta_0 + \mathbf{X} \, \underline{\theta}_1 - \mathbf{Y}\right)^T \left(\mathbf{1}_N \theta_0 + \mathbf{X} \, \underline{\theta}_1 - \mathbf{Y}\right)$$

Centered: $$(\widehat{\theta}_0^c, \widehat{\underline{\theta}}_1^c) = \underset{(\theta_0, \underline{\theta}_1) \in \mathbb{R}^{D+1}}{\operatorname{argmin}} \left(\mathbf{1}_N \theta_0 + \mathbf{X}^c \, \underline{\theta}_1 - \mathbf{Y}^c\right)^T \left(\mathbf{1}_N \theta_0 + \mathbf{X}^c \, \underline{\theta}_1 - \mathbf{Y}^c\right)$$

Condensed: $$(\hat{\theta}_0^c, \hat{\underline{\theta}}_1^c) = \underset{\underline{\theta}^c \in \mathbb{R}^{D+1}}{\operatorname{argmin}} \quad \left(\mathbb{X}\,\underline{\theta}^c - \mathbf{Y}^c\right)^T \left(\mathbb{X}\,\underline{\theta}^c - \mathbf{Y}^c\right)$$

where $\mathbb{X} = \begin{bmatrix} \mathbf{1}_N & \mathbf{X}^c \end{bmatrix} \qquad \underline{\theta}^c = \begin{bmatrix} \theta_0^c \\ \underline{\theta}_1^c \end{bmatrix}$

Cost function: $$J(\underline{\theta}^c) = \left(\mathbb{X}\,\underline{\theta}^c - \mathbf{Y}^c\right)^T \left(\mathbb{X}\,\underline{\theta}^c - \mathbf{Y}^c\right)$$

$$= \left((\underline{\theta}^c)^T \mathbb{X}^T - (\mathbf{Y}^c)^T\right) \left(\mathbb{X}\,\underline{\theta}^c - \mathbf{Y}^c\right)$$

$$= (\underline{\theta}^c)^T \mathbb{X}^T \mathbb{X} \underline{\theta}^c - 2(\mathbf{Y}^c)^T \mathbb{X}\,\underline{\theta}^c + (\mathbf{Y}^c)^T \mathbf{Y}^c$$

Gradient: $$\nabla J = 2\mathbb{X}^T \mathbb{X} \underline{\theta}^c - 2\mathbb{X}^T \mathbf{Y}^c$$

Stationary points: $$\hat{\underline{\theta}}^c = \left(\mathbb{X}^T \mathbb{X}\right)^{-1} \mathbb{X}^T \mathbf{Y}^c$$

Unpack:

$$\begin{bmatrix}\hat{\theta}_0^c \\ \underline{\hat{\theta}}_1^c\end{bmatrix} = \left(\begin{bmatrix}\mathbf{1}_N^T \\ (\mathbf{X}^c)^T\end{bmatrix}\begin{bmatrix}\mathbf{1}_N & \mathbf{X}^c\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{1}_N^T \\ (\mathbf{X}^c)^T\end{bmatrix}\mathbf{Y}^c$$

$$= \quad \vdots$$

$$= \begin{bmatrix} 0 \\ \left((\mathbf{X}^c)^T\mathbf{X}^c\right)^{-1}(\mathbf{X}^c)^T\mathbf{Y}^c\end{bmatrix}$$

Un-center:

$$\boxed{\begin{aligned}\hat{\theta}_0 &= \hat{\mu}_Y - \hat{\mu}_X\underline{\hat{\theta}}_1 \\ \underline{\hat{\theta}}_1 &= \left((\mathbf{X}^c)^T\mathbf{X}^c\right)^{-1}(\mathbf{X}^c)^T\mathbf{Y}^c\end{aligned}}$$

Statistical properties of multiple linear regression

Assumed data generating process:

$$y_i = \theta_0^* + x_i \underline{\theta}_1^* + \varepsilon_i$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Centered process:

$$y_i^c = (\theta_0^*)^c + x_i^c \, (\underline{\theta}_1^*)^c + \varepsilon_i^c$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$\underline{\hat{\theta}}_1$ as an estimate of $\underline{\theta}_1^*$

$$\underline{\hat{\theta}}_1 = \underline{\theta}_1^* + \Sigma_X^{-1}(\mathbf{X}^c)^T\underline{\varepsilon}^c$$

where $\Sigma_X^{-1} = \left((\mathbf{X}^c)^T\mathbf{X}^c\right)^{-1}$

$$\underline{\hat{\Theta}}_1 = \underline{\theta}_1^* + \Sigma_X^{-1}(\mathbf{X}^c)^T\underline{\mathcal{E}}^c$$

(proved in the reader)

$$E\left[\underline{\hat{\Theta}}_1\right] = \underline{\theta}_1^*$$

$$Var\left[\underline{\hat{\Theta}}_1\right] = \sigma^2\Sigma_X^{-1}$$

$\underline{\hat{\theta}}_0$ as an estimate of $\underline{\theta}_0^*$

(proved in the reader)

$$\hat{\theta}_0 = \theta_0^* + \hat{\mu}_X \underline{\theta}_1^* + \bar{\varepsilon} - \hat{\mu}_X \underline{\hat{\theta}}_1$$

$$\hat{\Theta}_0 = \theta_0^* + \hat{\mu}_X \underline{\theta}_1^* + \bar{\mathcal{E}} - \hat{\mu}_X \underline{\hat{\Theta}}_1$$

$$E\left[\hat{\Theta}_0\right] = \theta_0^*$$

$$Var\left[\hat{\Theta}_0\right] = \frac{\sigma^2}{N} + \sigma^2 \hat{\mu}_X \Sigma_X^{-1} \hat{\mu}_X^T$$

$\hat{y}_i$ as an estimate of y_i^*

Same as the scalar case:

$$\hat{Y}_i = \theta_0^* + \hat{\mu}_X \underline{\theta}_1^* + \bar{\mathcal{E}} + x_i^c \underline{\hat{\Theta}}_1$$

$$E[\hat{Y}_i] = \theta_0^* + \theta_1^* x_i$$

$$Var[\hat{Y}_i] = \frac{\sigma^2}{N} + \sigma^2 (x_i - \hat{\mu}_X) \Sigma_X^{-1} (x_i - \hat{\mu}_X)^T$$