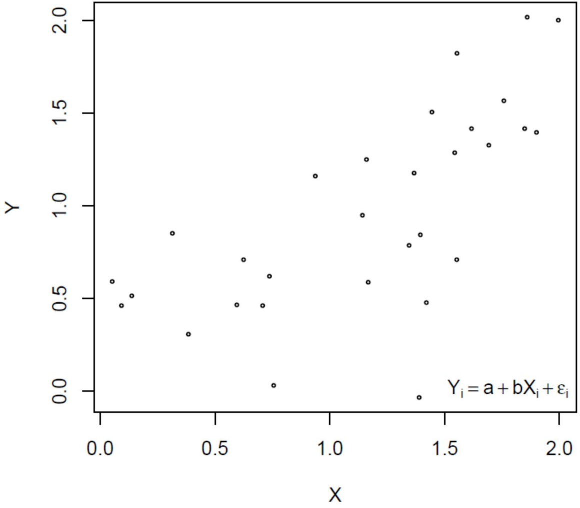
Statistics for Applications

Chapter 6: Linear regression

Heuristics of the linear regression (1)

Consider a cloud of i.i.d. random points $(X_i, Y_i), i = 1, \ldots, n$ **:**



Heuristics of the linear regression (2)

- **Idea:** Fit the *best* line fitting the data.
- Approximation: $Y_i \approx a + bX_i, i = 1, \ldots, n$, for some (unknown) $a, b \in \mathbb{R}$.
- Find $\hat{a}, \hat{b}$ that approach a and b.
- More generally: $Y_i \in \mathbb{R}, X_i \in \mathbb{R}^d$,

$$Y_i \approx a + X_i' b, \quad a \in \mathbb{R}, b \in \mathbb{R}^d.$$

- **Goal:** Write a rigorous model and estimate a and b.

Heuristics of the linear regression (3)

Examples:

- **Economics:** Demand and price,

$$D_i \approx a + bp_i, \quad i = 1, \ldots, n.$$

- **Ideal gas law:** $PV = nRT$,

$$\ln P_i \approx a + b \ln V_i + c \ln T_i, \quad i = 1, \ldots, n.$$

Linear regression of a r.v. Y on a r.v. X (1)

- Let X and Y be two real r.v. (non necessarily independent) with two moments and such that $Var(X) \neq 0$.

- The *theoretical linear regression* of Y on X is the *best approximation in quadratic means* of Y by a linear function of X, i.e. the r.v. $a + bX$, where a and b are the two real numbers minimizing $\mathbb{E}\left[(Y - a - bX)^2\right]$.

- By some simple algebra:
  - $b = \dfrac{cov(X, Y)}{Var(X)}$,
  - $a = \mathbb{E}[Y] - b\mathbb{E}[X] = \mathbb{E}[Y] - \dfrac{cov(X, Y)}{Var(X)}\mathbb{E}[X]$.

Linear regression of a r.v. Y on a r.v. X (2)

- If $\varepsilon = Y - (a + bX)$, then

$$Y = a + bX + \varepsilon,$$

  with $\mathbb{E}[\varepsilon] = 0$ and $cov(X, \varepsilon) = 0$.

- Conversely: Assume that $Y = a + bX + \varepsilon$ for some $a, b \in \mathbb{R}$ and some centered r.v. ε that satisfies $cov(X, \varepsilon) = 0$.

- E.g., if $X \perp\!\!\!\perp \varepsilon$ or if $\mathbb{E}[\varepsilon | X] = 0$, then $cov(X, \varepsilon) = 0$.

- Then, $a + bX$ is the theoretical linear regression of Y on X.

Linear regression of a r.v. Y on a r.v. X (3)

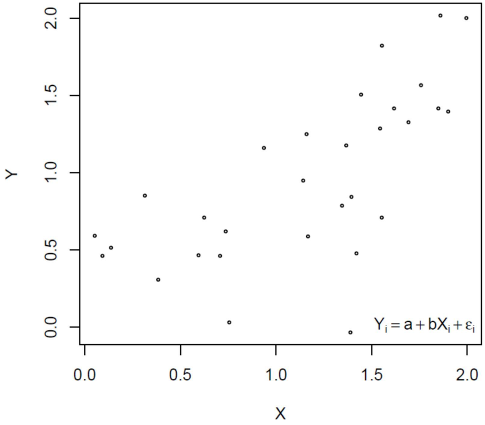
- A sample of n i.i.d. random pairs $(X_1, \ldots, X_n)$ with same distribution as (X, Y) is available.

- We want to estimate a and b.

Linear regression of a r.v. Y on a r.v. X (3)

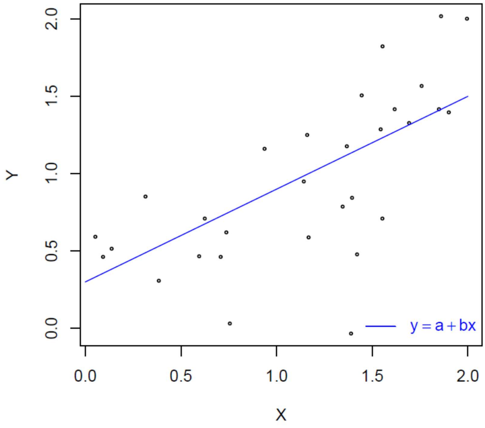
- A sample of n i.i.d. random pairs $(X_1, \ldots, X_n)$ with same distribution as (X, Y) is available.

- We want to estimate a and b.



Linear regression of a r.v. Y on a r.v. X (3)

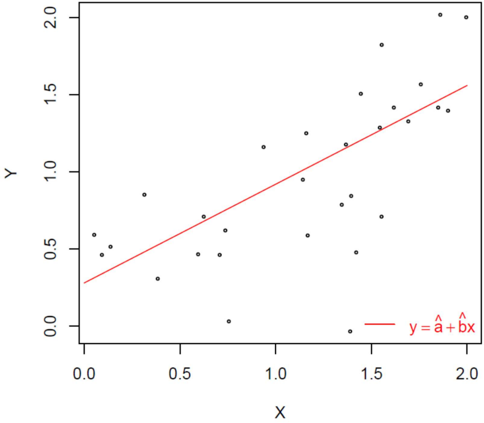
- A sample of n i.i.d. random pairs $(X_1, \ldots, X_n)$ with same distribution as (X, Y) is available.
- We want to estimate a and b.



Linear regression of a r.v. Y on a r.v. X (3)

- A sample of n i.i.d. random pairs $(X_1, \ldots, X_n)$ with same distribution as (X, Y) is available.

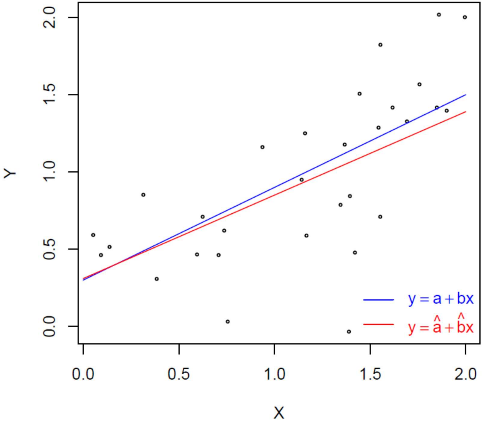
- We want to estimate a and b.



Linear regression of a r.v. Y on a r.v. X (3)

- A sample of n i.i.d. random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ with same distribution as (X, Y) is available.

- We want to estimate a and b.



Linear regression of a r.v. Y on a r.v. X (4)

Definition

The *least squared error (LSE)* estimator of (a, b) is the minimiser of the sum of squared errors:

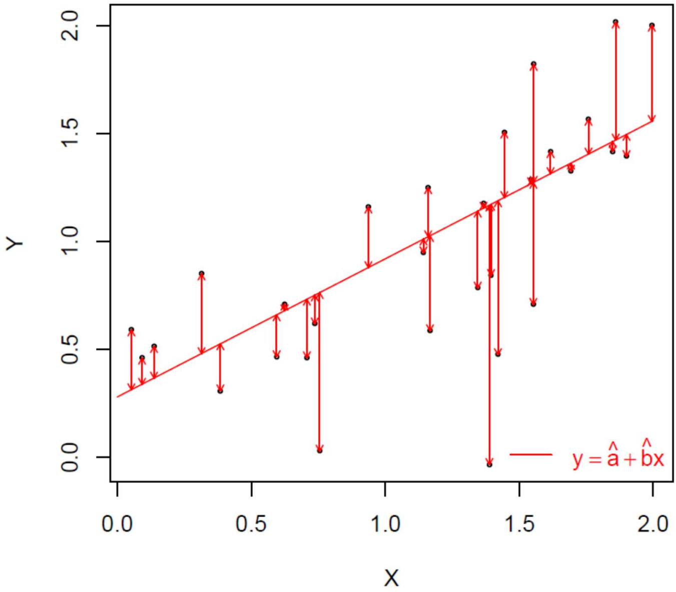
$$\sum_{i=1}^{n}(Y_i - a - bX_i)^2.$$

$(\hat{a}, \hat{b})$ is an M-estimator, and:

$$\hat{b} = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2},$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}.$$

Linear regression of a r.v. Y on a r.v. X (5)



Multivariate case (1)

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n.$$

- Vector of *explanatory variables* or *covariates*: $\mathbf{X}_i \in \mathbb{R}^p$ (wlog, assume its first coordinate is 1).
- *Dependent variable*: Y_i.
- $\boldsymbol{\beta} = (a, \mathbf{b}')'$; $\beta_1 (= a)$ is called the *intercept*.
- $\{\varepsilon_i\}_{i=1,\ldots,n}$: noise terms satisfying $cov(\mathbf{X}_i, \varepsilon_i) = \mathbf{0}$.

Definition

The *least squared error (LSE)* estimator of $\boldsymbol{\beta}$ is the minimiser of the sum of square errors:

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{t} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i'\mathbf{t})^2$$

Multivariate case (2)

LSE in matrix form

- Let $\mathbf{Y} = (Y_1, \ldots, Y_n)' \in \mathbb{R}^n$.
- Let $\mathbf{X}$ be the $n \times p$ matrix whose rows are $\mathbf{X}_1', \ldots, \mathbf{X}_n'$ ($\mathbf{X}$ is called the *design*).
- Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)' \in \mathbb{R}^n$ (unobserved noise)
- $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
- The LSE $\hat{\boldsymbol{\beta}}$ satisfies:

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{t} \in \mathbb{R}^p}{\operatorname{argmin}} \; \|\mathbf{Y} - \mathbf{X}\mathbf{t}\|_2^2.$$

Multivariate case (3)

- Assume that $rank(\mathbf{X}) = p$.
- Analytic computation of the LSE:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

- Geometric interpretation of the LSE
- $\mathbf{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of $\mathbf{Y}$ onto the subspace spanned by the columns of $\mathbf{X}$:

$$\mathbf{X}\hat{\boldsymbol{\beta}} = P\mathbf{Y},$$

  where $P = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Linear regression with deterministic design and Gaussian noise (1)

Assumptions:

- The design matrix $\mathbf{X}$ is deterministic and $rank(\mathbf{X}) = p$.
- The model is *homoscedastic*: $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d.
- The noise vector $\boldsymbol{\varepsilon}$ is Gaussian:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I_n),$$

for some known or unknown $\sigma^2 > 0$.

Linear regression with deterministic design and Gaussian noise (2)

- LSE = MLE: $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$.

- Quadratic risk of $\hat{\boldsymbol{\beta}}$: $\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] = \sigma^2 \mathrm{tr}\left((\mathbf{X}'\mathbf{X})^{-1}\right)$.

- Prediction error: $\mathbb{E}\left[\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_2^2\right] = \sigma^2(n-p)$.

- Unbiased estimator of σ^2: $\hat{\sigma}^2 = \dfrac{1}{n-p}\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_2^2$.

Theorem

- $(n-p)\dfrac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.

- $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp \hat{\sigma}^2$.

Significance tests (1)

- Test whether the j-th explanatory variable is significant in the linear regression ($1 \leq j \leq p$).

- H_0 : "$\beta_j = 0$" v.s. H_1 : "$\beta_j \neq 0$".

- If γ_j is the j-th diagonal coefficient of $(\mathbf{X}'\mathbf{X})^{-1}$ ($\gamma_j > 0$):

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 \gamma_j}} \sim t_{n-p}.$$

- Let $T_n^{(j)} = \dfrac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \gamma_j}}$.

- Test with non asymptotic level $\alpha \in (0,1)$:

$$\delta_\alpha^{(j)} = \mathbb{1}_{|T_n^{(j)}| > q_{1-\frac{\alpha}{2}}},$$

where $q_{1-\frac{\alpha}{2}}$ is the $(1-\alpha/2)$-quantile of t_{n-p}.

Significance tests (2)

- Test whether a group of explanatory variables is significant in the linear regression.

- H_0 : "$\beta_j = 0, \forall j \in S$" v.s. H_1 : "$\exists j \in S, \beta_j \neq 0$", where $S \subseteq \{1, \ldots, p\}$.

- *Bonferroni's test*: $\delta_\alpha^B = \max_{j \in S} \delta_{\alpha/k}^{(j)}$, where $k = |S|$.

- δ_α has non asymptotic level at most α.

More tests (1)

Let G be a $k \times p$ matrix with $rank(G) = k$ $(k \leq p)$ and $\boldsymbol{\lambda} \in \mathbb{R}^k$.

- Consider the hypotheses:

$$H_0 : \text{"} G\boldsymbol{\beta} = \boldsymbol{\lambda} \text{"} \text{ v.s. } H_1 : \text{"} G\boldsymbol{\beta} \neq \boldsymbol{\lambda} \text{"}.$$

- The setup of the previous slide is a particular case.

- If H_0 is true, then:

$$G\hat{\boldsymbol{\beta}} - \boldsymbol{\lambda} \sim \mathcal{N}_k \left(0, \sigma^2 G(\mathbf{X}'\mathbf{X})^{-1} G'\right),$$

and

$$\sigma^{-2}(G\hat{\boldsymbol{\beta}} - \boldsymbol{\lambda})' \left(G(\mathbf{X}'\mathbf{X})^{-1} G'\right)^{-1} (G\boldsymbol{\beta} - \boldsymbol{\lambda}) \sim \chi^2_k.$$

More tests (2)

- Let $S_n = \frac{1}{\hat{\sigma}^2} \frac{(G\hat{\boldsymbol{\beta}} - \boldsymbol{\lambda})' \left(G(\mathbf{X}'\mathbf{X})^{-1}G'\right)^{-1} (G\boldsymbol{\beta} - \boldsymbol{\lambda})}{k}$.
- If H_0 is true, then $S_n \sim F_{k,n-p}$.
- Test with non asymptotic level *alpha* $\in (0,1)$:

$$\delta_\alpha = \mathbb{1}_{S_n > q_{1-\alpha}},$$

where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of $F_{k,n-p}$.

Definition

The *Fisher distribution with p and q degrees of freedom*, denoted by $F_{p,q}$, is the distribution of $\frac{U/p}{V/q}$, where:

- $U \sim \chi^2_p$, $V \sim \chi^2_q$,
- $U \perp\!\!\!\perp V$.

Concluding remarks

- Linear regression exhibits correlations, **NOT** causality

- Normality of the noise: One can use goodness of fit test to test whether the residuals $\hat{\varepsilon}_i = Y_i - \mathbf{X}_i'\hat{\beta}$ are Gaussian.

- Deterministic design: If $\mathbf{X}$ is not deterministic, all the above can be understood conditional on $\mathbf{X}$, if the noise is assumed to be Gaussian, conditionally on X.