

18.650 – Fundamentals of Statistics

6. Linear Regression

Goals

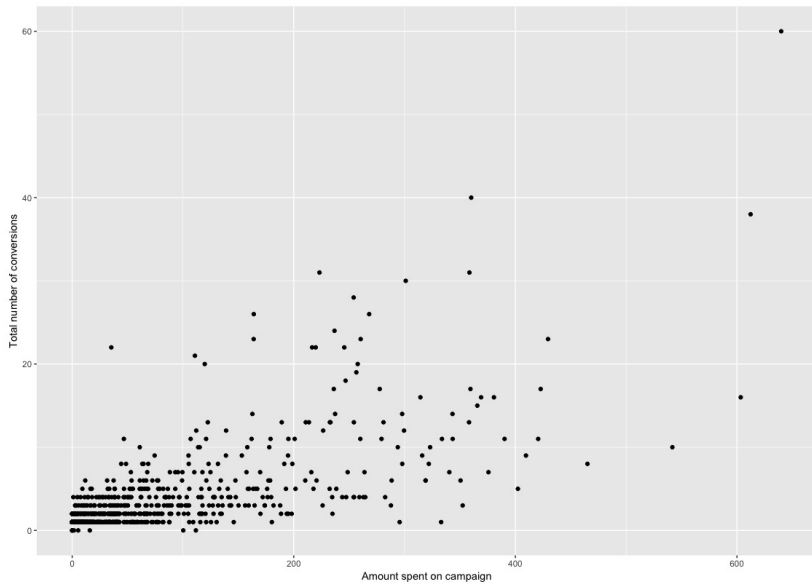
Consider two random variables X and Y . For example,

1. X is the amount of \$ spent on Facebook ads and Y is the total conversion rate
2. X is the age of the person and Y is the number of clicks

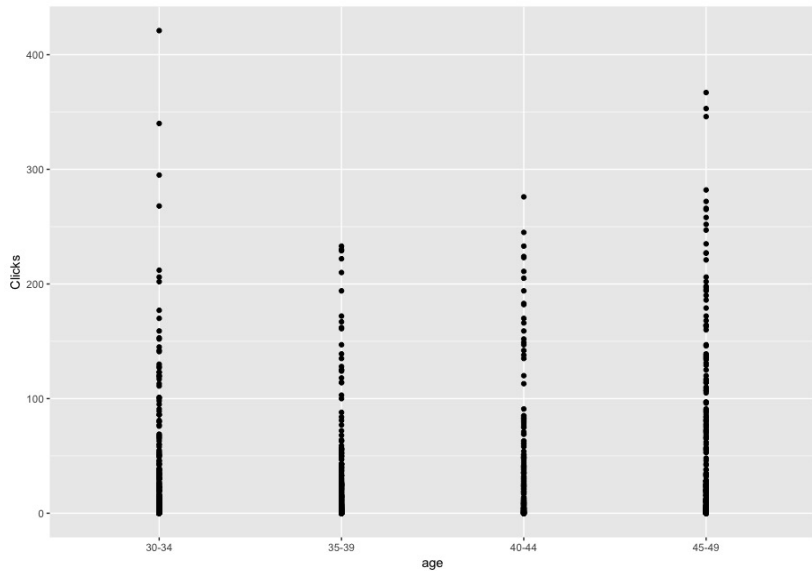
Given two random variables (X, Y) , we can ask the following questions:

- ▶ How to predict Y from X ?
- ▶ Error bars around this prediction?
- ▶ How much more conversions Y for an additional dollar?
- ▶ Does the number of clicks even depend on age?
- ▶ What if X is a random vector? For example, $X = (X_1, X_2)$ where X_1 is the amount of \$ spent on Facebook ads and X_2 is the duration in days of the campaign.

Conversions vs. amount spent



Clicks vs. age



Modeling assumptions

$(X_i, Y_i), i = 1, \dots, n$ are i.i.d from some **unknown joint distribution** \mathbb{P} .

\mathbb{P} can be described entirely by (assuming all exist)

- ▶ Either a joint PDF $h(x, y)$
- ▶ The marginal density of X $h(x) = \int h(x, y)dy$ **and** the conditional density

$$h(y|x) = \frac{h(x, y)}{h(x)}$$

$h(y|x)$ answers all our questions. It contains all the information about Y given X

Partial modeling

We can also describe the distribution only partially, e.g., using

- ▶ The expectation of Y : $\mathbb{E}[Y]$
- ▶ The conditional expectation of Y given $X = x$: $\mathbb{E}[Y|X = x]$
The function

$$x \mapsto f(x) := \mathbb{E}[Y|X = x] = \int y h(y|x) dy$$

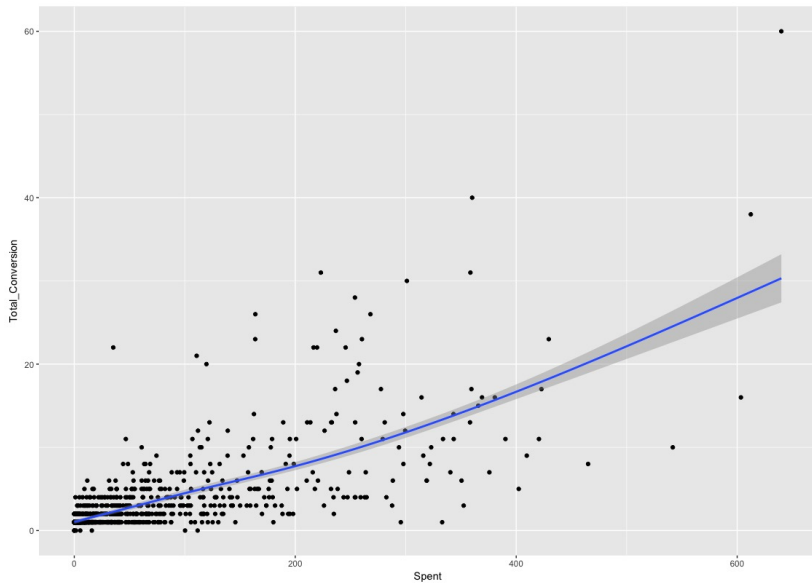
is called **regression function**

- ▶ Other possibilities:
 - ▶ The conditional median: $m(x)$ such that

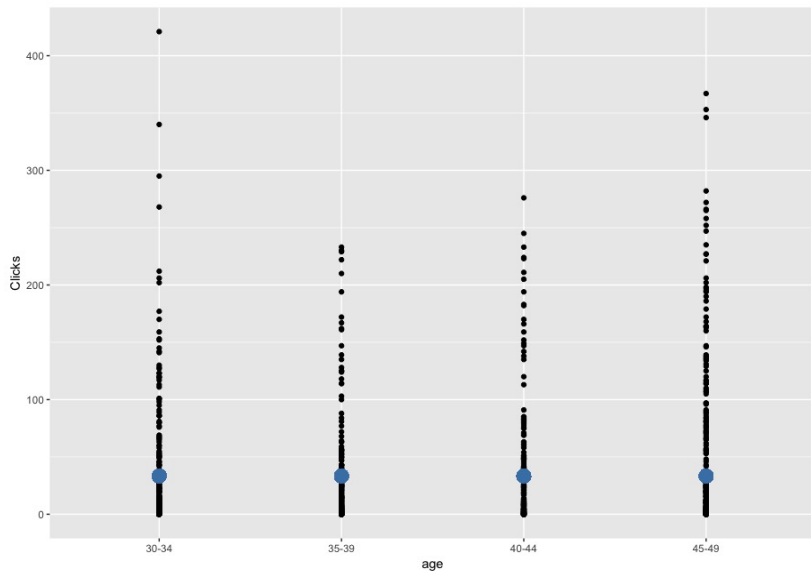
$$\int_{-\infty}^m h(y|x) dy = \frac{1}{2}$$

- ▶ Conditional quantiles
- ▶ Conditional variance (not informative about location)

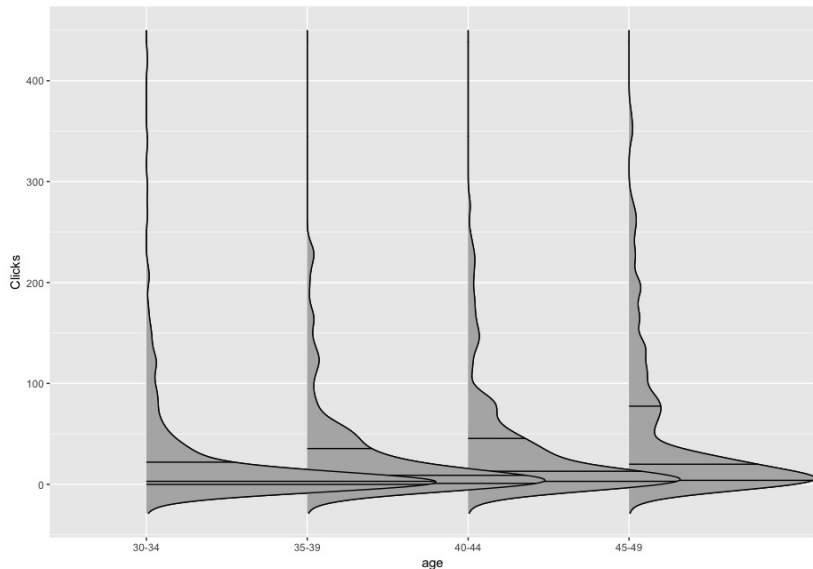
Conditional expectation and standard deviation



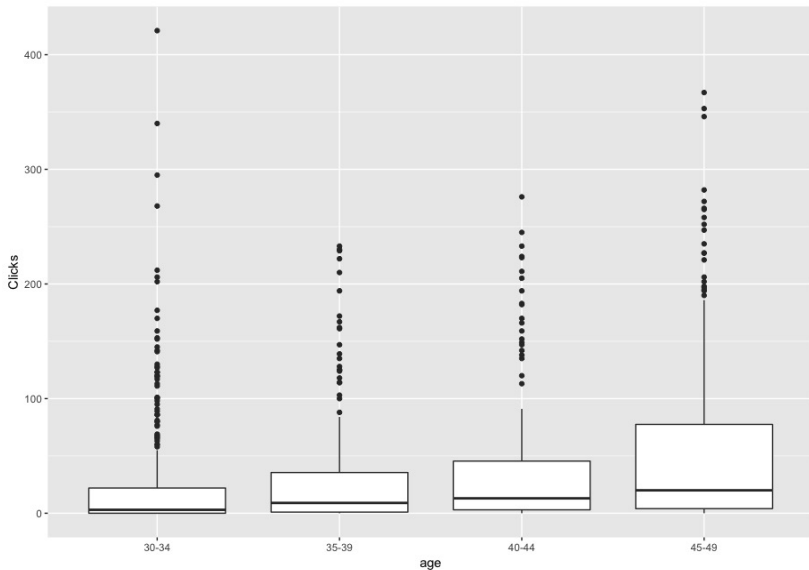
Conditional expectation



Conditional density and conditional quantiles



Conditional distribution: boxplots



Linear regression

We first focus on modeling the regression function

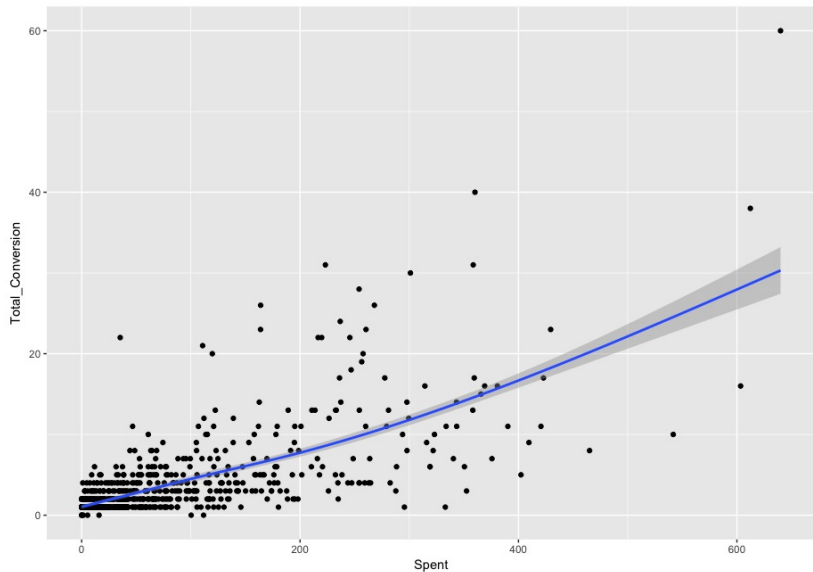
$$f(x) = \mathbb{E}[Y|X = x]$$

- ▶ Too many possible regression functions f (nonparametric)
- ▶ Useful to restrict to **simple** functions that are described by a few parameters
- ▶ Simplest:

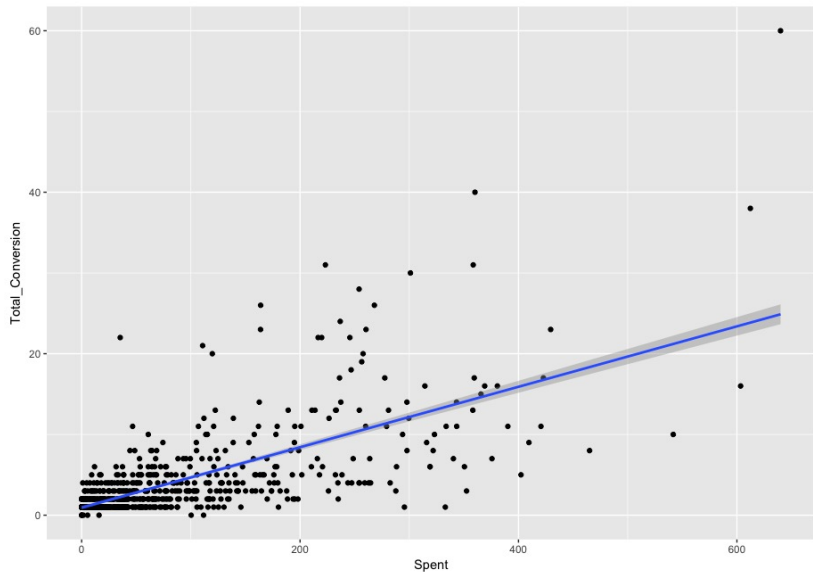
$$f(x) = a + bx \quad \text{linear (or affine) functions}$$

Under this assumption, we talk about **linear regression**

Nonparametric regression



Linear regression



Probabilistic analysis

- ▶ Let X and Y be two real r.v. (not necessarily independent) with two moments and such that $\text{var}(X) > 0$.
- ▶ The **theoretical linear regression** of Y on X is the line $x \mapsto a^* + b^*x$ where

$$(a^*, b^*) = \underset{(a,b) \in \mathbb{R}^2}{\operatorname{argmin}} \mathbb{E} \left[(Y - a - bX)^2 \right]$$

- ▶ Setting partial derivatives to zero gives
 - ▶ $b^* = \frac{\text{cov}(X, Y)}{\text{var}(X)},$
 - ▶ $a^* = \mathbb{E}[Y] - b^* \mathbb{E}[X] = \mathbb{E}[Y] - \frac{\text{cov}(X, Y)}{\text{var}(X)} \mathbb{E}[X].$

Noise

Clearly the points are not exactly on the line $x \mapsto a^* + b^*x$ if $\text{var}(Y|X = x) > 0$. The random variable $\varepsilon = Y - (a^* + b^*X)$ is called *noise* and satisfies

$$Y = a^* + b^*X + \varepsilon,$$

with

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

Statistical problem

In practice a^*, b^* need to be estimated from data.

- Assume that we observe n i.i.d. random pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ with same distribution as (X, Y) :

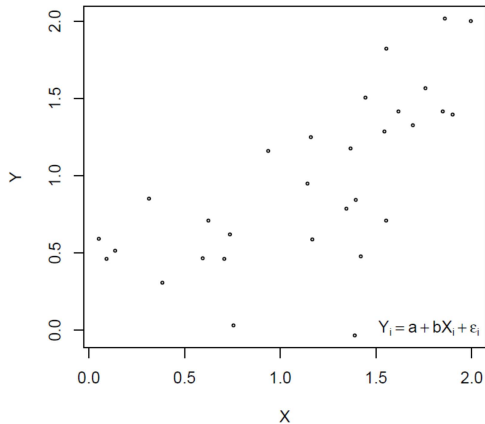
$$Y_i = a^* + b^* X_i + \varepsilon_i$$

- We want to estimate a^* and b^* .



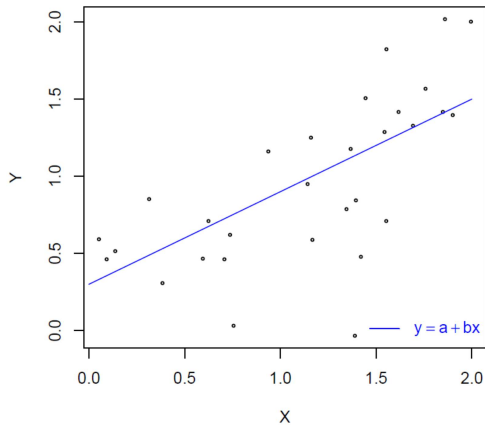
Statistical problem

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



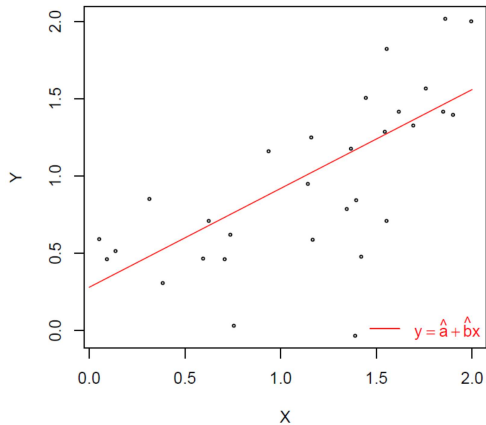
Statistical problem

$$Y_i = a^* + b^*X_i + \varepsilon_i$$



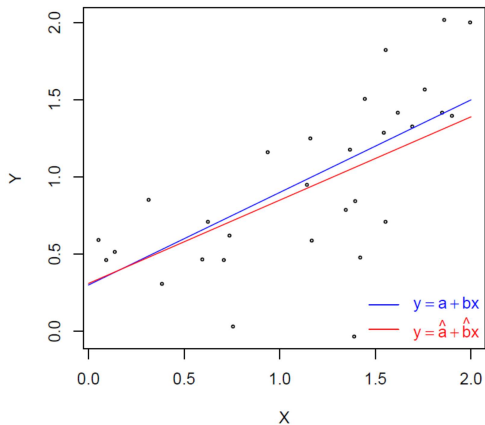
Statistical problem

$$Y_i = a^* + b^*X_i + \varepsilon_i$$



Statistical problem

$$Y_i = a^* + b^* X_i + \varepsilon_i$$



Least squares

Definition

The **least squares estimator (LSE)** of (a^*, b^*) is the minimizer of the sum of squared errors:

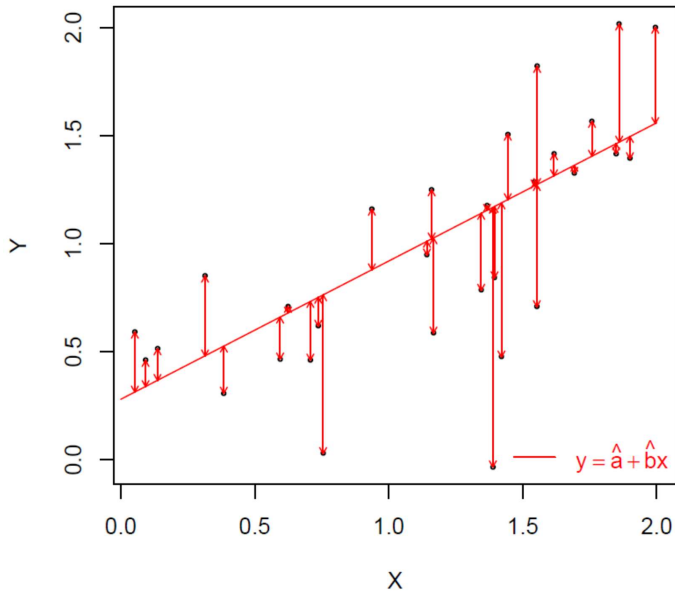
$$\sum_{i=1}^n (Y_i - a - bX_i)^2.$$

(\hat{a}, \hat{b}) is given by

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

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

Residuals



Multivariate regression

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta}^* + \varepsilon_i, \quad i = 1, \dots, n.$$

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

Definition

The **least squares estimator (LSE)** of $\boldsymbol{\beta}^*$ is the minimizer of the sum of square errors:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2$$

LSE in matrix form

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

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

Closed form solution

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

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

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

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

where $P = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top$.

Statistical inference

To make inference (confidence regions, tests) we need more assumptions.

Assumptions:

- ▶ The design matrix \mathbb{X} is deterministic and $\text{rank}(\mathbb{X}) = p$.
- ▶ The model is **homoscedastic**: $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d.
- ▶ The noise vector ε is Gaussian:

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

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

Properties of LSE

- ▶ LSE = MLE
- ▶ Distribution of $\hat{\beta}$: $\hat{\beta} \sim \mathcal{N}_p(\beta^*, \sigma^2(\mathbb{X}^\top \mathbb{X})^{-1})$.
- ▶ Quadratic risk of $\hat{\beta}$: $\mathbb{E} [\|\hat{\beta} - \beta^*\|_2^2] = \sigma^2 \text{tr}((\mathbb{X}^\top \mathbb{X})^{-1})$.
- ▶ Prediction error: $\mathbb{E} [\|\mathbf{Y} - \mathbb{X}\hat{\beta}\|_2^2] = \sigma^2(n - p)$.
- ▶ Unbiased estimator of σ^2 : $\hat{\sigma}^2 = \frac{1}{n - p} \|\mathbf{Y} - \mathbb{X}\hat{\beta}\|_2^2$.

Theorem

- ▶ $(n - p) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}^2$.
- ▶ $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$.

Significance tests

- ▶ 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 $(\mathbb{X}^\top \mathbb{X})^{-1}$ ($\gamma_j > 0$):

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

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

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

$$R_{j,\alpha} = \{|T_n^{(j)}| > q_{\frac{\alpha}{2}}(t_{n-p})\}$$

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

- ▶ We can also compute p-values.

Bonferroni's test

- ▶ 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 $S \subseteq \{1, \dots, p\}$.
- ▶ *Bonferroni's test*: $R_{S,\alpha} = \bigcup_{j \in S} R_{j,\alpha/k}$, where $k = |S|$.
- ▶ This test has nonasymptotic level at most α .

Remarks

- ▶ Linear regression exhibits correlations, **NOT** causality
- ▶ Normality of the noise: One can use goodness of fit tests to test whether the residuals $\hat{\varepsilon}_i = Y_i - \mathbb{X}_i^\top \hat{\beta}$ are Gaussian.
- ▶ Deterministic design: If \mathbb{X} is not deterministic, all the above can be understood conditionally on \mathbb{X} , if the noise is assumed to be Gaussian, conditionally on X .