- likelihood
- dichotomous responses (logistic regression) → GLM

In the previous lectures, we obtain an estimate of β, by minimizing the least squares criterion

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \underbrace{x_i^T \beta}_{f(x_i;\beta)})^2$$

It really works well $\varepsilon \sim N(0, \sigma^2)$.

We need a more general criterion to estimate the parameters
↳ likelihood criterion
↳ maximize the likelihood

Define a parametric family of probability distributions,

E.g., Gaussian $N(\mu, \sigma^2)$

$$\mathcal{F} = \{ p_Y(y;\vartheta), \ \vartheta \in \Theta \}$$

parametric space

$\vartheta = (\mu, \sigma^2)$ $\Theta = \mathbb{R} \times \mathbb{R}^+$

parameter

probability density function y cont
probability function y discr.

depends on a p-dimensional parameter, which we need to estimate

in the Gaussian example, $p_Y(y; \vartheta = (\mu, \sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (y - \mu)^2 \right\}$

$X\beta \to \vartheta = (\beta, \sigma^2)$

The likelihood function is defined as

$$L(\vartheta) = c \, p_Y(y;\vartheta) \quad \to \text{for } n \text{ independent obs,} \quad L(\vartheta) = c \prod_{i=1}^{n} p_{Y_i}(y_i;\vartheta)$$

where:
- c is a constant (everything that can be dropped-out in the computations)
- y is fixed (are the data, so given), therefore $L(\vartheta)$ is a function of only ϑ

Since $p_Y(y;\vartheta) \in [0, +\infty)$, it is possible (and more convenient) to work with the log-likelihood

$$\ell(\vartheta) = \log L(\vartheta)$$

if $p_Y(y;\vartheta) = 0$, $\ell(\vartheta) = -\infty$

Then, to estimate ϑ, we maximize $L(\vartheta)$ or $\ell(\vartheta)$

$$\hat{\vartheta} = \arg\max_{\vartheta} L(\vartheta) = \arg\max_{\vartheta} \ell(\vartheta)$$

e.g., in the Gaussian example (given σ^2)

$$\hat{\beta} = \arg\max_{\beta} \left(\prod_{i=1}^{n} \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\to c} \exp\left\{ -\frac{1}{2\sigma^2} (y_i - x_i^T\beta)^2 \right\} \right)$$

$$= \arg\max_{\beta} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T\beta)^2 \right\}$$

$$= \arg\max_{\beta} \left(-\sum_{i=1}^{n} (y_i - x_i^T\beta)^2 \right)$$

$$= \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - x_i^T\beta)^2 \right)$$

In the special case of Gaussian distribution, the maximum likelihood estimate is the same of least squares one

How to find $\hat{\vartheta}$ in practice?

$$\hat{\vartheta} = \arg\max_{\vartheta} \ell(\vartheta)$$

$\hat{\vartheta}$

$\to$ find a stationary point $\dfrac{\partial \ell(\vartheta)}{\partial \vartheta} = 0$

verify that is a global maximum $\to \dfrac{\partial^2 \ell(\vartheta)}{\partial \vartheta^2} < 0$

theoretically more tricky, usually not a problem in practice

$\to$ useful to compute a measure of uncertainty around our estimai

(related to the variance of the maximum likelihood estimator)

$$j(\vartheta) = -\frac{\partial^2 \ell(\vartheta)}{\partial \vartheta^2} \qquad \to j(\hat{\vartheta}) = -\left.\frac{\partial^2 \ell(\vartheta)}{\partial \vartheta^2}\right|_{\vartheta=\hat{\vartheta}}$$

Fisher observed information, it is the inverse of the variance of $\hat{\vartheta}$

$$s.e.(\hat{\vartheta}) = \mathrm{diag}\left(j(\hat{\vartheta})^{-1}\right)^{1/2} \qquad i(\vartheta) = E[j(\vartheta)]$$

expected information

$$\hat{\vartheta} \,\dot\sim\, N(\vartheta, j^{-1}(\hat{\vartheta}))$$

approximately distributed

$$\hat{\vartheta} \sim \mathcal{N}\left(\vartheta, j^{-1}(\hat{\vartheta})\right)$$

95%

We can then construct a confidence interval around $\hat{\vartheta}$, with level $1-\alpha$

For the r-th component of $\hat{\vartheta}$

$$CI(1-\alpha) = \hat{\vartheta}_r \pm z_{1-\frac{\alpha}{2}} \sqrt{j^{-1}(\hat{\vartheta})_{[r,r]}}$$

↳ quantile of level $1-\frac{\alpha}{2}$ of a standard normal distribution $\mathcal{N}(0;1)$

$$j(\vartheta) = \frac{\partial^2 \ell(\vartheta)}{\partial \vartheta^2} = \begin{array}{c c} & p \quad r \\ r & \begin{bmatrix} \frac{\partial^2 \ell(\vartheta)}{\partial\vartheta_{[1]}\partial\vartheta_{[1]}} & \cdots & \frac{\partial^2 \ell(\vartheta)}{\partial\vartheta_{[1]}\partial\vartheta_{[p]}} \\ \vdots & & \\ \frac{\partial^2 \ell(\vartheta)}{\partial\vartheta_{[p]}\partial\vartheta_{[1]}} & \cdots & \frac{\partial^2 \ell(\vartheta)}{\partial\vartheta_{[p]}\partial\vartheta_{[p]}} \end{bmatrix} \\ p & \end{array}$$

dimension of ϑ ↗ p

$$\frac{\partial^2 \ell(\vartheta)}{\partial\vartheta_{[r]}\partial\vartheta_{[r]}}$$

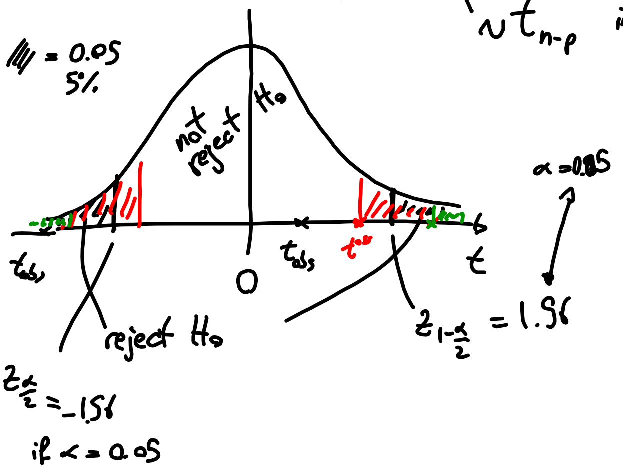
Confidence intervals and hypothesis testing are closely connected

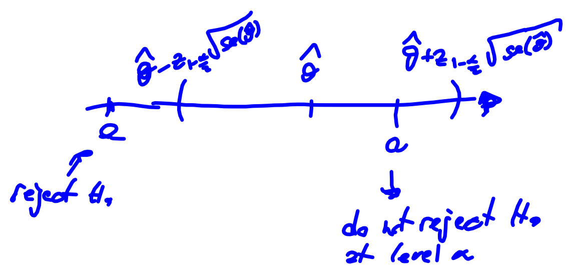
$$H_0 : \vartheta_r = a$$

For a fixed statistical significance level α, the Wald test is used for testing the hypothesis, and it is based on the t-statistic

$$t = \frac{\vartheta_r - a}{\widehat{s.e.}(\hat{\vartheta}_r)} \overset{H_0}{\underset{a}{\sim}} \mathcal{N}(0;1)$$

$\sim t_{n-p}$ in the case of Gaussian distribution
$n > 30$, t is very close to $\mathcal{N}(0;1)$





Alternatively, one can compute the p-value : $2 \min\left(\Phi(t), 1-\Phi(t)\right)$

$2\Phi(-|t|)$ → p-value that we contrast with $\alpha (=0.05)$

More generally, for

$\vartheta_{[p-q+1:p]} = a_{[1:q]}$

we can use the likelihood ratio test

$$w = 2\left(\ell(\hat\vartheta) - \ell(\vartheta_0)\right)$$

where $\vartheta_0 = (\hat\vartheta_{1:q}, a_{[1:q]})$

$q=2 \qquad p=4$

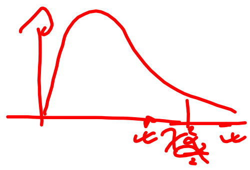
$\vartheta_1\ \vartheta_2\ \vartheta_3\ \vartheta_4$

$\hat\vartheta = (\hat\vartheta_1, \hat\vartheta_2, \hat\vartheta_3, \hat\vartheta_4)$

$\vartheta_0 = (\hat\vartheta_1, \hat\vartheta_2, a_1, a_2)$

Since $w \sim \chi^2_q$, the p-value is computed as

$$\text{p-value} = \Pr(\chi^2 \geq w) \qquad \chi^2 \sim \chi^2_q$$



Gaussian example (with only one variable x) $\qquad \beta$ - 1-dimensional

$Y \sim \mathcal{N}(x\beta, \sigma^2)$

$$L(\beta,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i\beta)^2\right\}$$

$$= \cancel{c}\,(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2\right\}$$

$2\pi^{-n/2}$

$$\ell(\beta,\sigma^2) = \log L(\vartheta) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2$$

$$\ell_\beta(\beta,\sigma^2) = \frac{\partial \log L(\vartheta)}{\partial\beta} = +\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)x_i$$

$$\ell_{\sigma^2}(\beta,\sigma^2) = \frac{\partial \log L(\vartheta)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2$$

$$\ell_\beta(\beta,\sigma^2) = 0 \to \cancel{\frac{1}{2\sigma^2}}\sum_{i=1}^{n}(x_i y_i - x_i^2\beta) = 0 \qquad \hat\beta = \frac{\sum x_i y_i}{\sum x_i^2}$$

if x_i is p-dimensional $\quad = (X^T X)^{-1} X^T y$

$$\ell_{\sigma^2}(\hat\beta,\sigma^2) = 0 \to -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - x_i\hat\beta)^2 = 0$$

$$\hat\sigma^2 = \frac{\sum_{i=1}^{n}(y_i - x_i\hat\beta)^2}{n} = \frac{D(\hat\beta)}{n}$$

$\leftarrow n-p$ because $\hat\sigma^2$ is biased, but they are equal as $n\to\infty$

To evaluate the hypothesis

$\vartheta_{p-q+1} = 0$

we contrast the deviance of the constrained model and the full model

When the error is Gaussian, we use the F

$$F = \frac{[D(\vartheta_0) - D(\hat\vartheta)]/q}{D(\hat\vartheta)/(n-p)} \sim F_{q,\,n-p}$$

it is an F distribution with q degrees of freedom at the numerator and n-p at the denominator

Binomial distribution

$Y_i \in \{0, 1\}$ $\quad Y_i \sim Bi(1, \pi)$ $\quad$ Bernulli distribution
π: probability of success

$Y = \sum_{i=1}^{n} Y_i \sim Bi(n, \pi)$ $\quad$ number of successes in # trials

$$P_Y(y;\pi) = \binom{n}{y} \pi^{y} (1-\pi)^{n-y}$$

$$L(\pi) = \underbrace{\binom{n}{y}}_{=c} \pi^{y} (1-\pi)^{n-y}$$

$$\ell(\pi) = y \log \pi + (n-y) \log(1-\pi)$$

$$\ell_\pi(\pi) = \frac{y}{\pi} + \frac{n-y}{1-\pi}(-1) \to \frac{y}{\hat\pi} - \frac{n-y}{1-\hat\pi} = 0 \qquad y - \cancel{\hat\pi y} - n\hat\pi + \cancel{\hat\pi y} = 0 \qquad \hat\pi = \frac{y}{n}$$

$$\ell_{\pi\pi}(\pi) = -\frac{y}{\pi^2} - \frac{n-y}{(1-\pi)^2}$$

$$j(\hat\pi) = \frac{y}{\hat\pi^2} + \frac{n-y}{(1-\hat\pi)^2} \qquad j(\hat\pi) = \frac{y\, n^2}{y^2} + \frac{n-y}{(1-\frac{y}{n})^2} = \frac{n}{\hat\pi} + \frac{(n-y)n^2}{(n-y)^2} = \frac{n}{\hat\pi} + \frac{n}{1-\hat\pi}$$

$$se(\hat\pi) = \sqrt{\left(\frac{n}{\hat\pi} + \frac{n}{(1-\hat\pi)}\right)^{-1}} = \sqrt{\left(\frac{n - n\hat\pi + n\hat\pi}{\hat\pi(1-\hat\pi)}\right)^{-1}} = \sqrt{\frac{\hat\pi(1-\hat\pi)}{n}}$$

Comparison of two groups

Brasilian bank example $\quad Y$ = satisfaction $\in \{\text{low}, \text{high}\}$
X = age $\in \{\text{'young'}_1, \text{'old'}_2\}$ $\quad$ young < 45

$$\ell(\pi_1, \pi_2) = c + y_1 \log \pi_1 + (n_1 - y_1)\log(1-\pi_1) + y_2 \log \pi_2 + (n_2 - y_2)\log(1-\pi_2)$$

$H_0: \pi_1 = \pi_2$ $\qquad H_0: \pi_1 - \pi_2 = 0$

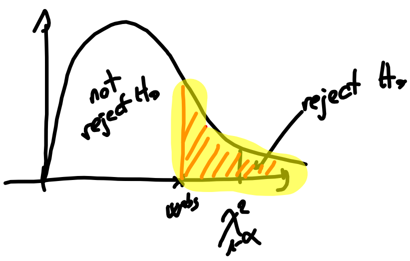
$$W = 2\left(\ell(\hat\pi_1, \hat\pi_2) - \ell(\hat\pi, \hat\pi)\right)$$

where: $\hat\pi_1 = \frac{y_1}{n_1} \simeq 0.78$

$\hat\pi_2 = \frac{y_2}{n_2} \simeq 0.82$

$\hat\pi = \frac{y_1 + y_2}{n_1 + n_2} \simeq 0.76$

$W \sim \chi^2_1$



$Pr(X < w^{obs})$ = pchisq(wobs, 1)