Econometrics

Andrés M. Alonso

andres.alonso@uc3m.es

Bibliography

- Creel, M. (2006) Econometrics,
  (Available at `http://pareto.uab.es/mcreel/Econometrics/econometrics.pdf`).

- Hayashi, F. (2000) Econometrics, Princeton University Press. (Library code: D 330.43 HAY).

- Wooldridge, J.M. (2000) Introductory Econometrics: A Modern Approach, South Western College Publishing (Library code: D 330.43 WOO).

Software: I recommend to use `MATLAB` (or `Octave`, its freeware version) or `S-Plus` (or `R`, its freeware version).

- Causin, M.C. (2005) MATLAB Tutorials.
  (Available at `http://halweb.uc3m.es/esp/Personal/personas/amalonso/esp/ephd.htm`).
- LeSage, J.P. (1999) Applied Econometrics using MATLAB.
  (Available at `http://www.spatial-econometrics.com/html/mbook.pdf`).

The outline for today

UNIT 1. Introduction: The regression model.

1.1 Conditional expectation.
1.2 Assumptions.
1.3 Interpretation of the model.

Addendum. Resampling methods in i.i.d. data.

- The Jackknife.
- The Bootstrap.

UNIT 2. Estimation principles.

UNIT 3: Hypothesis testing principles.

UNIT 4: Heteroscedasticity in the regression model.

UNIT 5: Endogeneity of regressors.

Introduction

Econometrics is a branch of Economy concerning the empirical study of relations among economic variables.

In econometrics is used:

- An econometric model: Based on the economic theory.
- Data: Facts.
- Econometrics techniques: Statistical inference.

Recommended readings: Chapter 2 and 3 of Creel (2006) and Chapter 1 and 6 of Greene (2000).

Objectives of an econometric study

- Structural Analysis: It consists on the use of models to measure specific economic relations. Its an important objective since we can compare different theories under the fact's (data) information.

- Prediction: It consists on the use of models to predict future values of economic variables using the fact's (data) information.

- Policies evaluation: It consists on the use of models to select among alternatives policies.

These three objectives are interrelated.

Economic and econometric models:

Example # 1: Economic theory tells us that demand functions are something like:

$$x_i = x_i(p_i, m_i, z_i)$$

- x_i is $G \times 1$ vector of quantities demanded.
- p_i is $G \times 1$ vector of prices.
- m_i is income.
- z_i is a vector of individual characteristics related to preferences.

Suppose we have a sample consisting of one observation on n individuals' demands at time period t (this is a *cross section*, where $i = 1, 2, ..., n$ indexes the individuals in the sample).

Example # 1 (cont.): This is nice economic model but it is not estimable as it stands, since:

- The form of the demand function is different for all i.
- Some components of z_i may not be observable to an outside modeler. For example, people don't eat the same lunch every day, and you can't tell what they will order just by looking at them. Suppose we can break z_i into the observable components w_i and a single unobservable component ε_i.

What can we do?

Example # 1 (cont.): An estimable (*e.g.*, econometric) model is

$$x_i = \beta_0 + p_i'\beta_p + m_i\beta_m + w_i'\beta_w + \varepsilon_i$$

Here, we impose a number of restrictions on the theoretical model:

- The functions $x_i(\cdot)$ which may differ for all i have been restricted to all belong to the same parametric family.
- Of all parametric families of functions, we have restricted the model to the class of linear in the variables functions.
- There is a single unobservable component, and we assume it is additive.

Example # 1 (cont.):

- These are **very strong restrictions**, compared to the theoretical model.
- Furthermore, **these restrictions have no theoretical basis**.
- For this reason, **specification testing** will be needed, to check that the model seems to be reasonable.

First conclusion: Only when we are convinced that the model is at least approximately correct should we use it for economic analysis.

The *classical linear model* is based upon several assumptions:

1. **Linearity**: the model is a linear function of the parameter vector β_0 :

$$y_t = \beta_1^0 x_1 + \beta_2^0 x_2 + \cdots + \beta_k^0 x_k + \varepsilon_t, \tag{1}$$

or in matrix form,

$$y = \mathbf{x}' \boldsymbol{\beta}^0 + \boldsymbol{\varepsilon},$$

where y is the dependent variable, $\mathbf{x} = (x_1, x_2, \cdots, x_k)'$, where x_t is a $k \times 1$ vector of explanatory variables and $\boldsymbol{\beta}^0$ and $\boldsymbol{\varepsilon}$ are conformable vectors.

The subscript "0" in $\boldsymbol{\beta}^0$ means this is the true value of the unknown parameter.

Suppose that we have n observations, then model (1) can be written as:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector and $\mathbf{X}$ is a $n \times k$ matrix.

2. **IID mean zero errors**:

$$\begin{aligned} \mathrm{E}(\varepsilon|\mathbf{X}) &= 0 \\ \mathrm{Var}(\varepsilon|\mathbf{X}) &= \mathrm{E}(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2\mathbf{I}_n. \end{aligned}$$

What are the implications of this second assumption?

3. **Nonstochastic, linearly independent regressors**:

   (a) $\mathbf{X}_{n\times k}$ has rank k, its number of columns.
   (b) $\mathbf{X}$ is nonstochastic.
   (c) $\lim_{n\to\infty} \frac{1}{n}\mathbf{X}'\mathbf{X} = Q_X$, a finite positive definite matrix.

What are the implications of this third assumption?

4. **Normality** (Optional): $\varepsilon|\mathbf{X}$ is normally distributed.

For economic data, the assumption of nonstochastic regressors is obviously unrealistic. Likewise, normality of errors will often be unrealistic.

Linear models are more general than they might first appear, since one can employ nonlinear transformations of the variables:

$$\varphi_0(z_t) = \begin{bmatrix} \varphi_1(w_t) & \varphi_2(w_t) & \cdots & \varphi_p(w_t) \end{bmatrix} \beta_0 + \varepsilon_t$$

where the $\phi_i()$ are known functions. Defining $y_t = \varphi_0(z_t)$, $x_{t1} = \varphi_1(w_t)$, *etc*, leads to a model in the form of equation (1).

For example, the Cobb-Douglas model:

$$z = A w_2^{\beta_2} w_3^{\beta_3} \exp(\varepsilon)$$

can be transformed logarithmically to obtain

$$\ln z = \ln A + \beta_2 \ln w_2 + \beta_3 \ln w_3 + \varepsilon.$$

Example: The Nerlove model

Theoretical background:

For a firm that takes input prices w and the output level q as given, the cost minimization problem is to choose the quantities of inputs x to solve the problem

$$\min_x w'x$$

subject to the restriction

$$f(x) = q.$$

The solution is the vector of factor demands $x(w, q)$. The *cost function* is obtained by substituting the factor demands into the criterion function:

$$C(w, q) = w'x(w, q).$$

Example: The Nerlove model

- **Monotonicity** Increasing factor prices cannot decrease cost, so $\frac{\partial C(w,q)}{\partial w} \geq 0$.

- **Homogeneity** Because the cost of production is a linear function of w, it has the property of homogeneity of degree 1 in input prices: $C(tw,q) = tC(w,q)$, where t is a scalar constant.

- **Returns to scale** The *returns to scale* parameter γ is defined as the inverse of the elasticity of cost with respect to output:

$$\gamma = \left(\frac{\partial C(w,q)}{\partial q} \frac{q}{C(w,q)} \right)^{-1}.$$

Constant returns to scale is the case where increasing production q implies that cost increases in the proportion 1:1. If this is the case, then $\gamma = 1$.

Example: The Nerlove model

Cobb-Douglas approximating model:

The Cobb-Douglas functional form is linear in the logarithms of the regressors and the dependent variable. For a cost function, if there are g factors, the Cobb-Douglas cost function has the form:

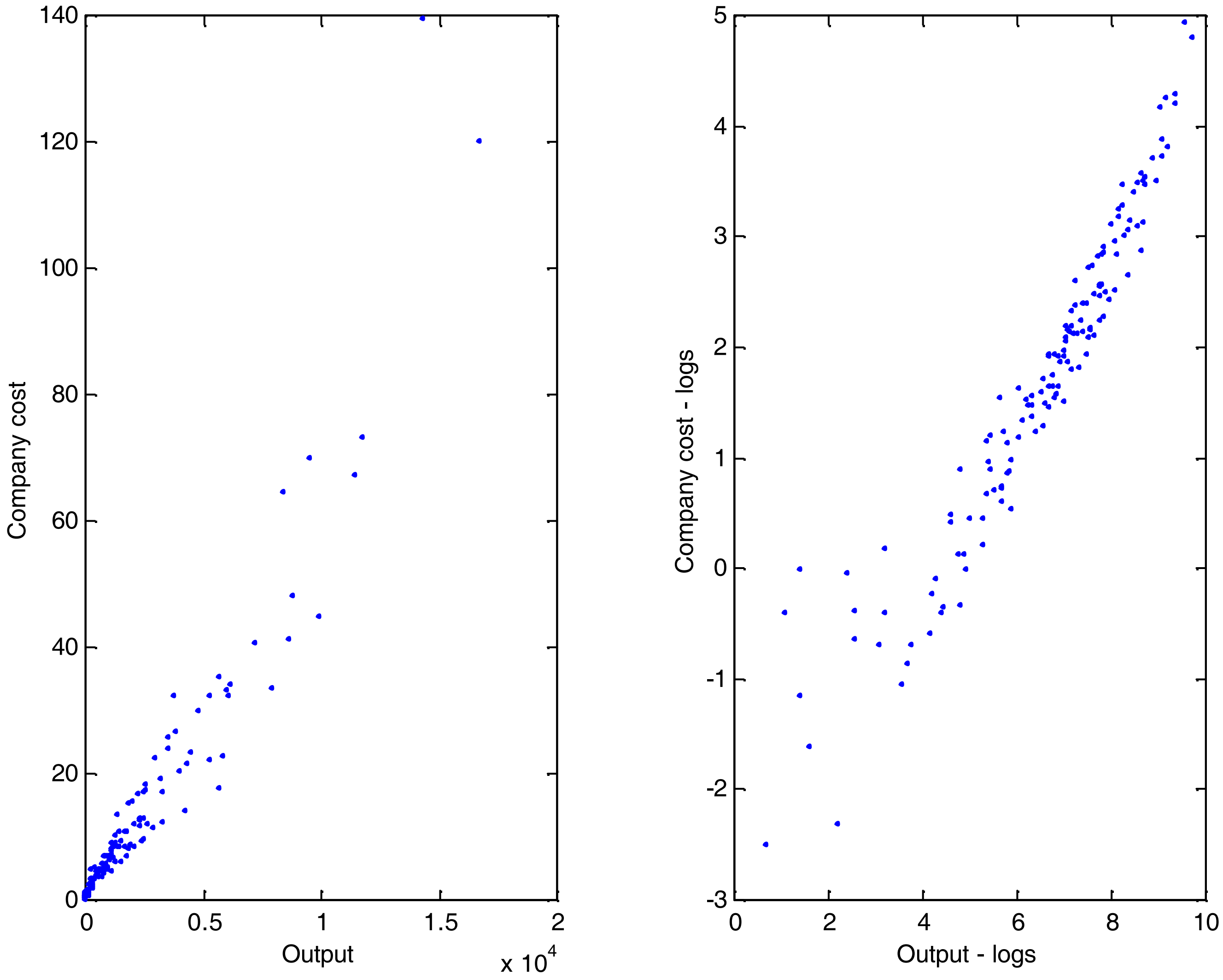
$$C = Aw_1^{\beta_1}...w_g^{\beta_g}q^{\beta_q}e^{\varepsilon}.$$

After a logarithmic transformation we obtain:

$$\ln C = \alpha + \beta_1 \ln w_1 + ... + \beta_g \ln w_g + \beta_q \ln q + \epsilon,$$

where $\alpha = \ln A$.

Example: The Nerlove model



Example: The Nerlove model

Using the Cobb-Douglas functional form we can:

- Verify that the property of HOD1, since it implies that $\sum_{i=1}^{g} \beta_g = 1$.
- Verify the hypothesis that the technology exhibits CRTS, since it implies that
$$\gamma = \frac{1}{\beta_q} = 1$$
so $\beta_q = 1$.
- Verify the hypothesis of monotonicity, since it implies that the coefficients $\beta_i \geq 0, i = 1, ..., g$.

BUT, the β's are unknowns, then we must to estimate them.

Estimation methods

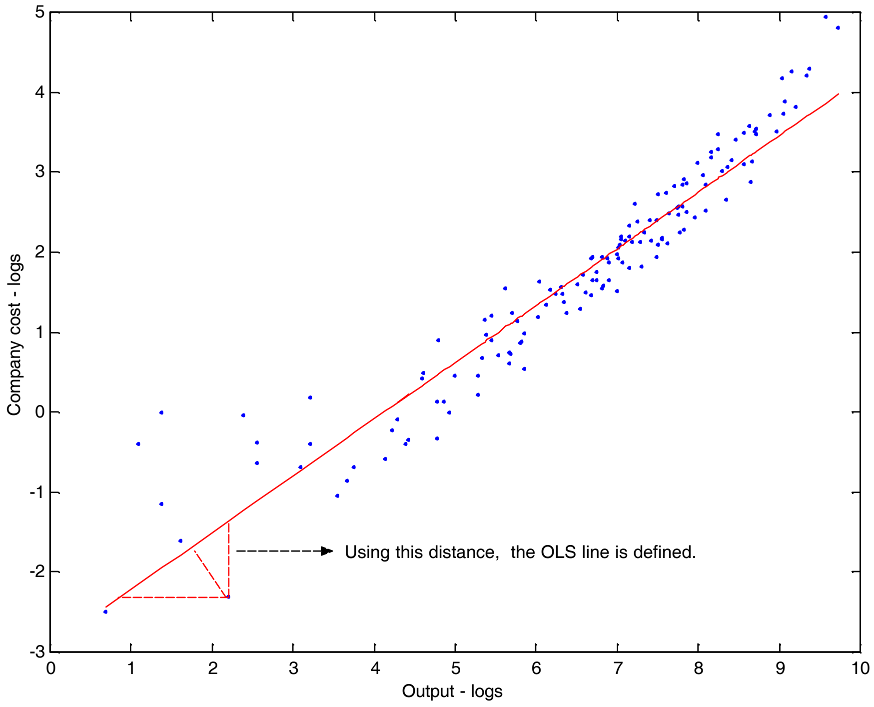
Ordinary Least Squares Method:

The *ordinary least squares* (OLS) estimator is defined as the value that minimizes the sum of the squared errors:

$$\hat{\beta} \quad = \quad \arg\min s(\beta)$$

where

$$\begin{aligned} s(\beta) &= \sum_{t=1}^{n} (y_t - \mathbf{x}_t'\beta)^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)' (\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \\ &= \| \mathbf{y} - \mathbf{X}\beta \|^2 . \end{aligned}$$



Ordinary Least Squares Method (2):

- To minimize the criterion $s(\beta)$, find the derivative with respect to β and set it to zero:

$$D_\beta s(\hat{\beta}) \quad = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} \quad = 0,$$

  so

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

- To verify that this is a minimum, check the s.o.c.:

$$D^2_\beta s(\hat{\beta}) = 2\mathbf{X}'\mathbf{X}.$$

  Since rank of $\mathbf{X}$ is equal k, this matrix is positive definite, since it's a quadratic form in a p.d. matrix (the identity matrix of order n), so $\hat{\beta}$ is in fact a minimizer.

Ordinary Least Squares Method (3):

- The *fitted values* are in the vector $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$.
- The *residuals* are in the vector $\hat{\varepsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$
- Note that
$$\mathbf{y} = \mathbf{X}\beta + \varepsilon = \mathbf{X}\hat{\beta} + \hat{\varepsilon}.$$
- The first order conditions can be written as:
$$\begin{aligned} \mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\hat{\beta} &= 0 \\ \mathbf{X}'(\mathbf{y} + \mathbf{X}\hat{\beta}) &= 0 \\ \mathbf{X}'\hat{\varepsilon} &= 0, \end{aligned}$$

  then, the OLS residuals are orthogonal to $\mathbf{X}$.

Ordinary Least Squares Method (4):

- We have that $X\hat{\beta}$ is the projection of y on the span of X, or

$$X\hat{\beta} = X\left(X'X\right)^{-1}X'y.$$

Therefore, the matrix that projects y onto the span of X is

$$P_X = X(X'X)^{-1}X',$$

since $X\hat{\beta} = P_X y$.

- $\hat{\varepsilon}$ is the projection of y *onto* the $n-k$ dimensional space orthogonal to the span of X. We have that

$$\begin{aligned}\hat{\varepsilon} &= y - X\hat{\beta} \\ &= y - X(X'X)^{-1}X'y \\ &= \left[I_n - X(X'X)^{-1}X'\right]y.\end{aligned}$$

So the matrix that projects y onto the space orthogonal to the span of X is

$$\begin{aligned} M_X &= I_n - X(X'X)^{-1}X' \\ &= I_n - P_X. \end{aligned}$$

Then $\hat{\varepsilon} = M_X y$.

- Therefore

$$\begin{aligned} y &= P_X y + M_X y \\ &= X\hat{\beta} + \hat{\varepsilon}. \end{aligned}$$

- Note that both P_X and M_X are *symmetric* and *idempotent*.
  - ◇ A symmetric matrix A is one such that $A = A'$.
  - ◇ An idempotent matrix A is one such that $A = AA$.
  - ◇ The only nonsingular idempotent matrix is the identity matrix.

Example: The Nerlove model (*estimation*)

The file <`nerlovedata.m`> contains data on 145 electric utility companies' cost of production, output and input prices. The data are for the U.S., and were collected by M. Nerlove. The observations are by row, and the columns are COMPANY CODE, COST (C), OUTPUT (Q), PRICE OF LABOR (P_L), PRICE OF FUEL (P_F) and PRICE OF CAPITAL (P_K).

Data Example:

```
101 0.082 2 2.09 17.9 183
102 0.661 3 2.05 35.1 174
103 0.990 4 2.05 35.1 171
104 0.315 4 1.83 32.2 166
105 0.197 5 2.12 28.6 233
```

Let's MATLAB works ...

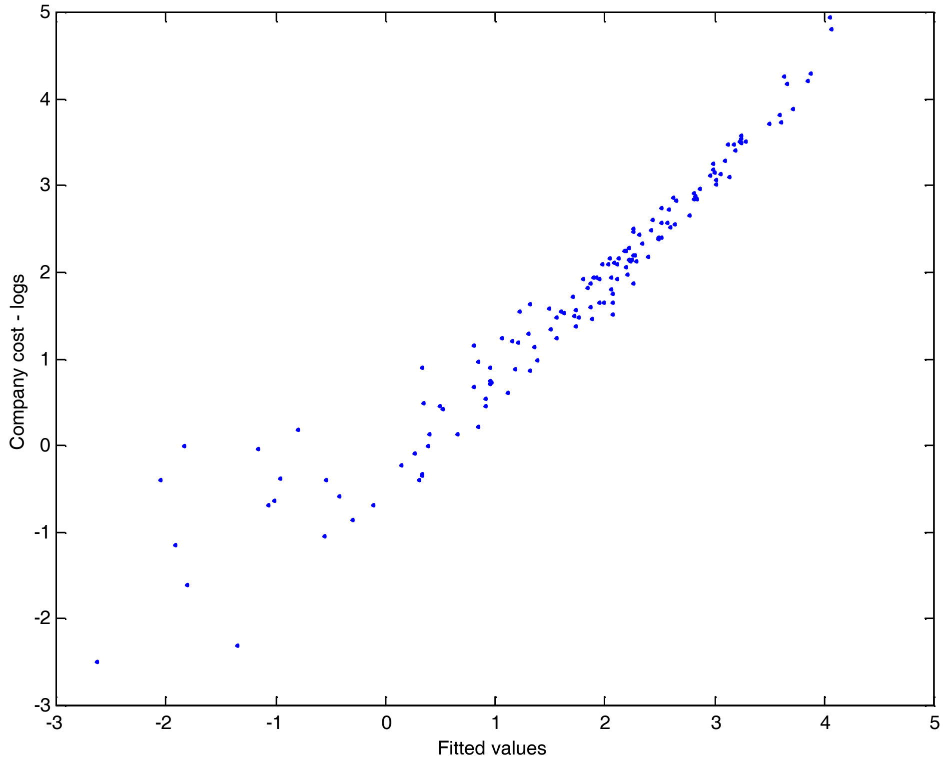
We will estimate the Cobb-Douglas model:

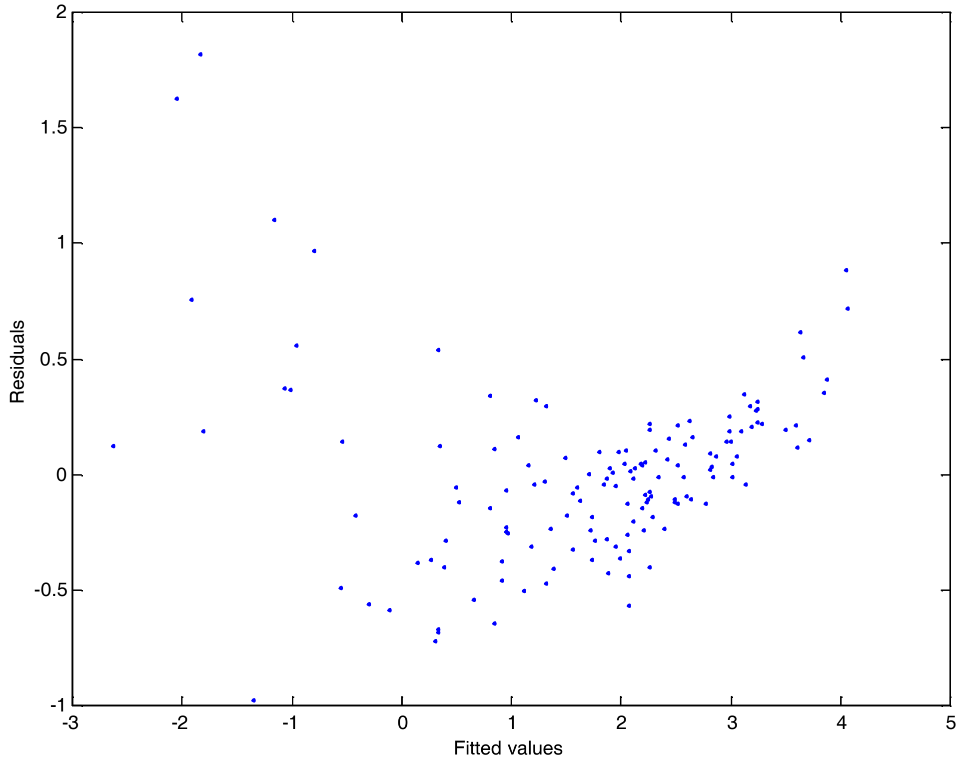
$$\ln C = \beta_1 + \beta_2 \ln Q + \beta_3 \ln P_L + \beta_4 \ln P_F + \beta_5 \ln P_K + \epsilon \tag{3}$$

using OLS.

The <`prg1sl1.m`> results are the following:

```
% OLS estimation (using the formula).

    B =
       -3.5265
        0.7204
        0.4363
        0.4265
       -0.2199
```





- Do the theoretical restrictions hold?
- Does the model fit well? $\Rrightarrow$ Influential observations or outliers?
- What do you think about CRTS?

Influential observations and outliers:

The OLS estimator of the i^{th} element of the vector β is simply:

$$
\begin{aligned}
\hat{\beta}_i &= \left[(X'X)^{-1}X'\right]_{i\cdot} y \\
&= c_i' y
\end{aligned}
$$

Then $\hat{\beta}_i$ is a linear estimator, i.e., it's a linear function of the dependent variable.

Since it's a linear combination of the observations on the dependent variable, where the weights are determined by the observations on the regressors, some observations may have more influence than others.

Influential observations and outliers (2):

Define

$$\begin{aligned} h_t &= (P_X)_{tt} \\ &= e_t' P_X e_t \\ &= \| P_X e_t \|^2 \\ &= \| e_t \|^2 = 1 \end{aligned}$$

h_t is the t^{th} element on the main diagonal of P_X and e_t is t^{th}-unit vector.

So $0 < h_t < 1$, and $\mathrm{Trace}(P_X) = k \Rightarrow \overline{h} = k/n$.

So, on average, the weight on the y_t's is k/n. If the weight is much higher, then the observation is *potentially* influential.

However, an observation may also be influential due to the value of y_t, rather than the weight it is multiplied by, which only depends on the x_t's.

Influential observations and outliers (3):

Consider the estimation of β without using the t^{th} observation (denote this estimator by $\hat{\beta}^{(t)}$). Then,

$$\hat{\beta}^{(t)} = \hat{\beta} - \left(\frac{1}{1-h_t}\right)(X'X)^{-1}X_t'\hat{\varepsilon}_t$$

so the change in the t^{th} observations fitted value is

$$X_t\hat{\beta} - X_t\hat{\beta}^{(t)} = \left(\frac{h_t}{1-h_t}\right)\hat{\varepsilon}_t$$

While an observation may be influential if it doesn't affect its own fitted value, it certainly *is* influential if it does.

A fast means of identifying influential observations is to plot $\left(\frac{h_t}{1-h_t}\right)\hat{\varepsilon}_t$ as a function of t.

Influential observations and outliers (4):

The following example is based on Table 13.8 of Page 423 in Peña, D. (1995) "Estadística. Modelos y Métodos". It consists on three data sets that differ in only one observation (see file <`penadata.m`>).

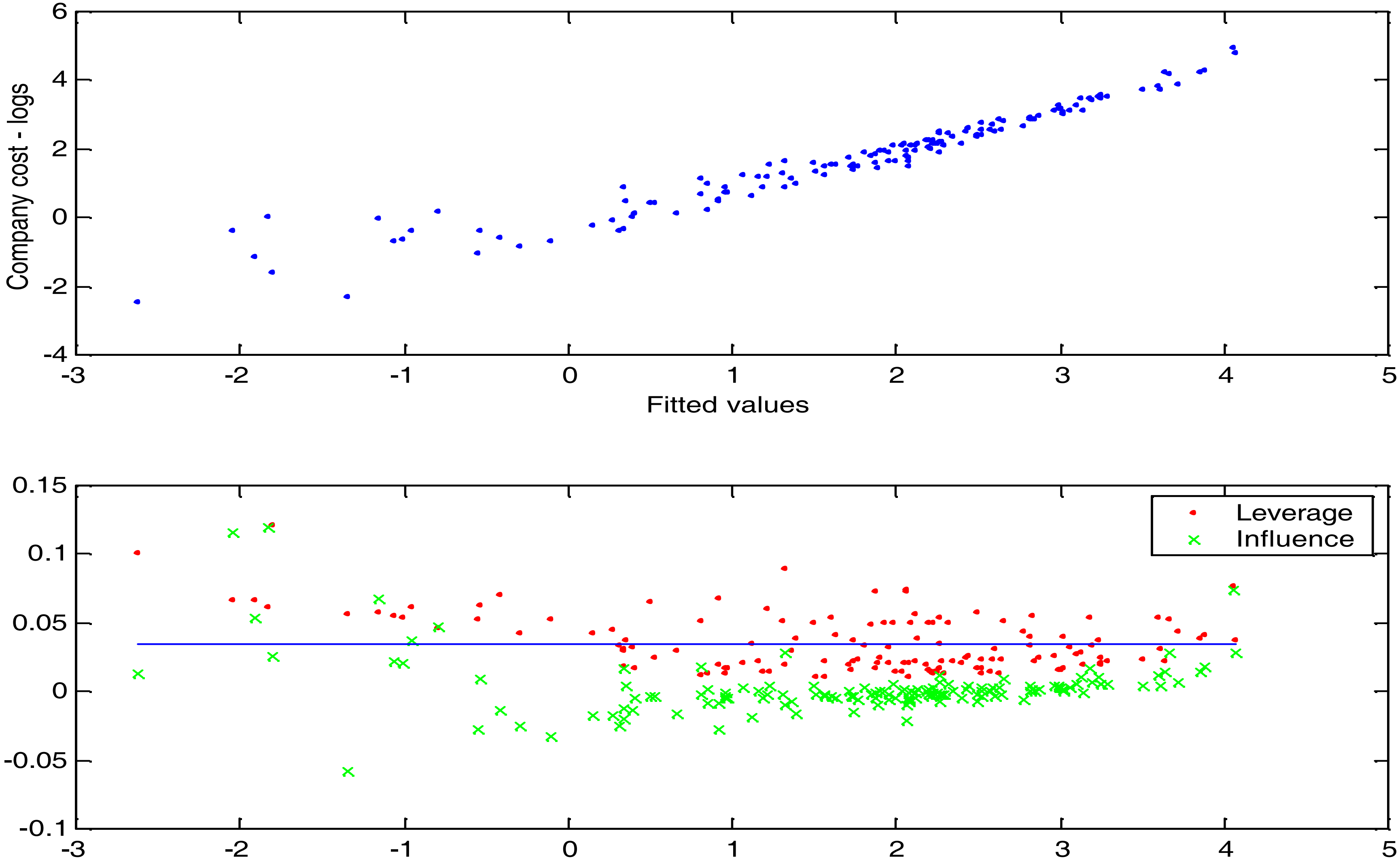
Data Set	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{s}_\varepsilon$	R^2	h_t	$\hat{\varepsilon}_t$	$\left(\frac{h_t}{1-h_t}\right)\hat{\varepsilon}_t$
Original	2.38	1.12	-0.30	0.35	0.98	-	-	-
A	13.12	1.77	-1.72	0.97	0.88	0.11	4.26	0.54
B	-2.74	0.80	0.38	0.37	0.98	0.91	0.19	2.01
C	-25.36	-0.62	3.43	0.91	0.89	0.65	2.48	4.70

A is an outlier but is not an influential observation.

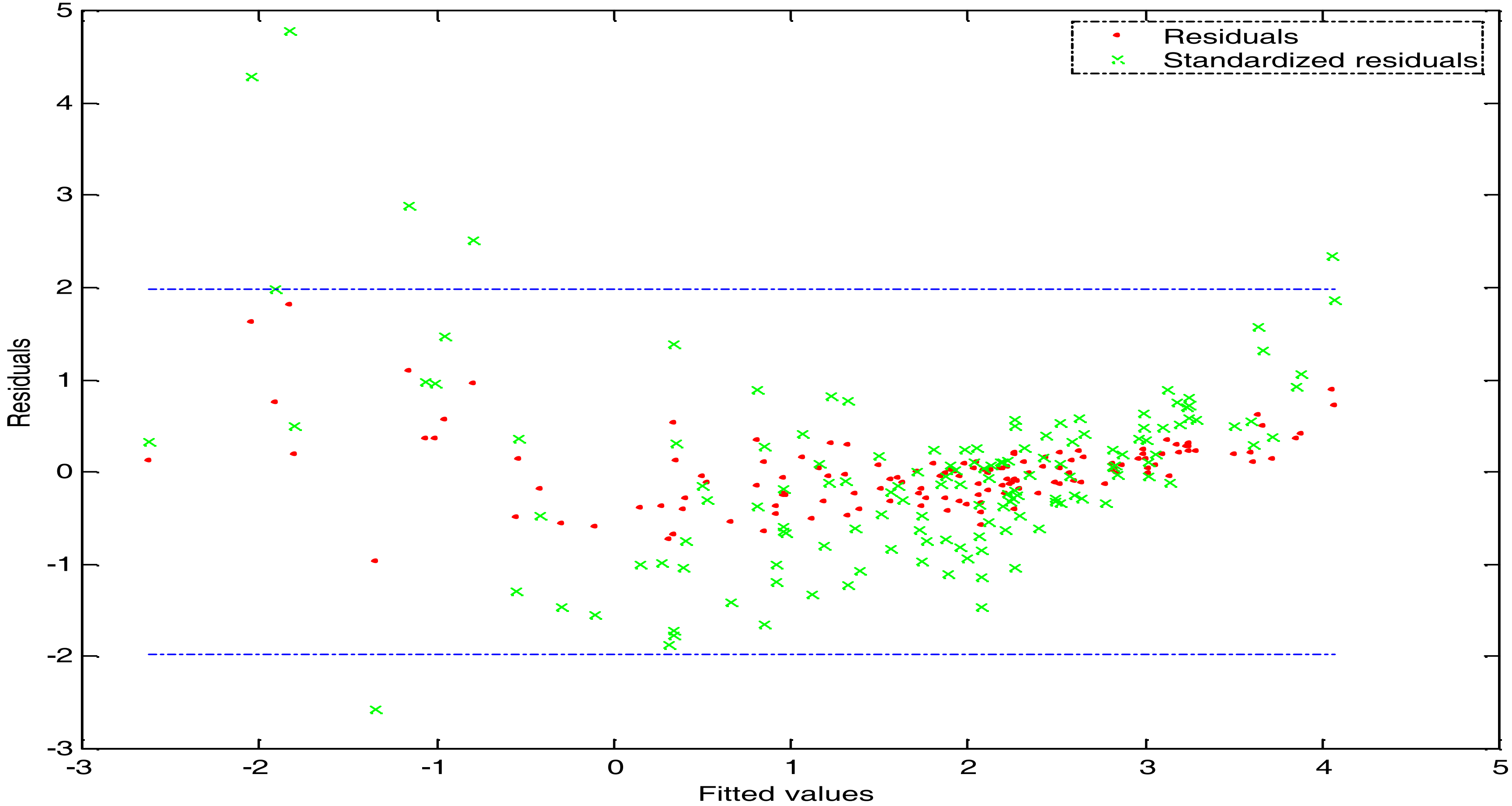
B is an influential but is not an outlier observation.

C is an outlier influential observation.

Example: The Nerlove model (*influential observations*)

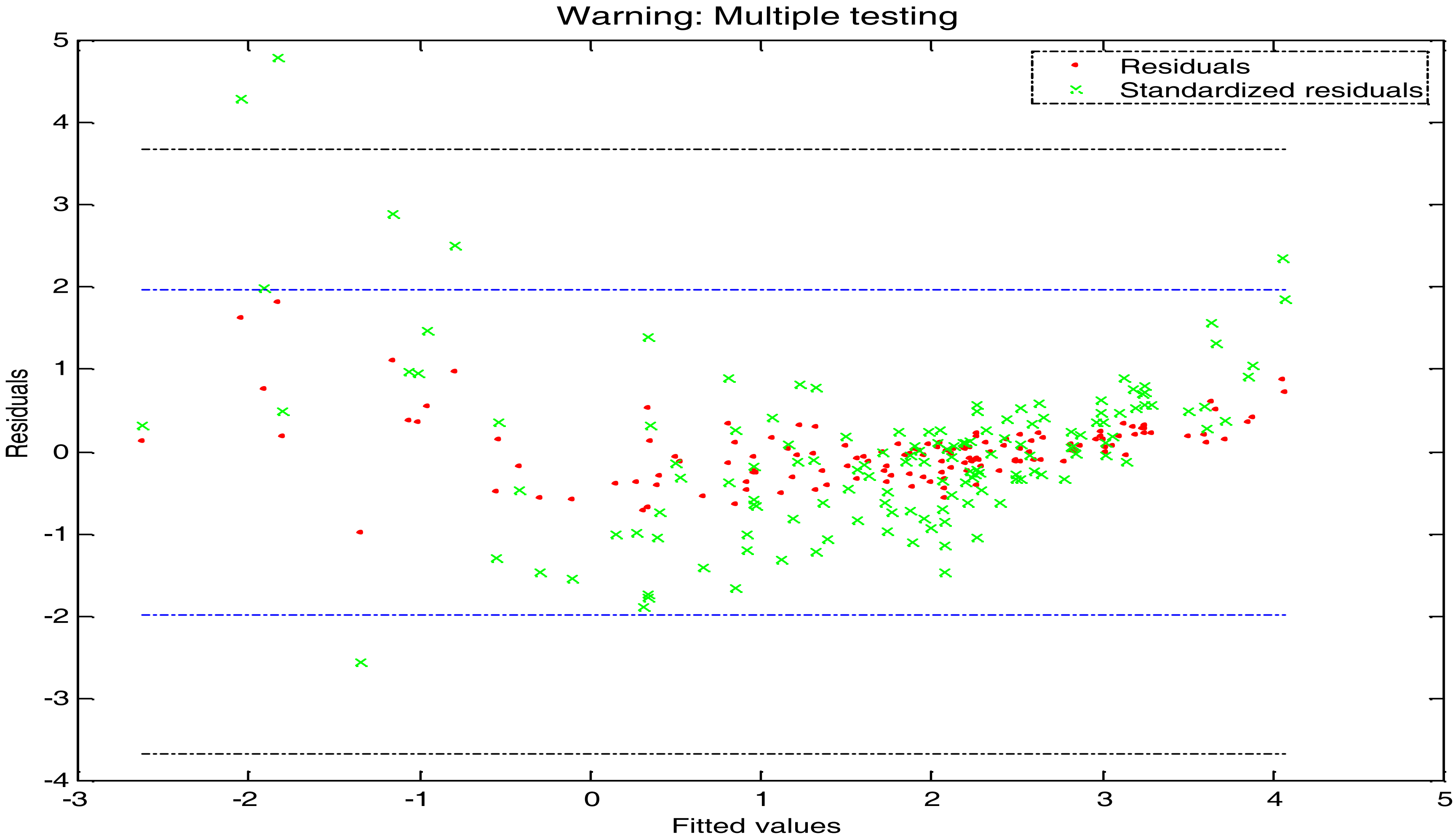


Example: The Nerlove model (*outliers observations*)



Standardized residuals: $r_t = \frac{\hat{\varepsilon}_t}{\hat{s}_\varepsilon \sqrt{1-h_t}} \approx t_{n-k-2}$.

Example: The Nerlove model (*outliers observations*)



Instead of using the α-critical value, we will use α/n-critical value.

Properties of OLS estimator

Small sample properties of the least squares estimator:

- *Unbiasedness*: For $\hat{\beta}$ we have

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'(X\beta + \varepsilon) \\ &= \beta + (X'X)^{-1}X'\varepsilon, \end{aligned}$$

then applying the strong exogeneity assumption and the law of iterated expectation, we have

$$\begin{aligned} \mathrm{E}[(X'X)^{-1}X'\varepsilon] &= \mathrm{E}[\mathrm{E}[(X'X)^{-1}X'\varepsilon|X]] \\ &= \mathrm{E}[(X'X)^{-1}X'\mathrm{E}[\varepsilon|X]] \\ &= 0, \end{aligned}$$

So the OLS estimator is unbiased.

Small sample properties of the least squares estimator (2):

- *Normality*: $\hat{\beta}|X \sim \mathcal{N}(\beta, (X'X)^{-1}\sigma^2)$, where $\sigma^2 = E[\varepsilon_t^2]$.

Proof: Note that $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$, then $\hat{\beta}$ is a linear function of ε.

It only rest to obtain the conditional variance.

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}|X) &= \mathrm{E}[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] \\
&= \mathrm{E}[(X'X)^{-1}X'\varepsilon\varepsilon' XX'X)^{-1}|X] \\
&= (X'X)^{-1}X'\,\mathrm{E}[\varepsilon\varepsilon'|X]\; XX'X)^{-1} \\
&= (X'X)^{-1}\sigma^2.
\end{aligned}$$

Notice that the above results is true regardless the distribution of ε.

Small sample properties of the least squares estimator (3):

- *Efficiency* (Gauss-Markov theorem): In the classical linear regression model, the OLS estimator, $\hat{\beta}$, is the minimum variance linear unbiased estimator of β. Moreover, for any vector of constants $\mathbf{w}$, the minimum variance linear unbiased estimator of $\mathbf{w}'\beta$ is $\mathbf{w}'\hat{\beta}$.

Proof: The OLS estimator is a linear estimator, i.e., it is a linear function of the dependent variable. It is also *unbiased*, as we proved two slides ago.

Now, consider a new linear estimator $\tilde{\beta} = \mathbf{C}y$. If this estimator is unbiased, then we must have $\mathbf{C}X = \mathbf{I}$ since

$$\begin{aligned} \mathrm{E}[\mathbf{C}y] &= \mathrm{E}(\mathbf{C}X\beta + \mathbf{C}\varepsilon) \\ &= \mathbf{C}X\beta \end{aligned}$$

Therefore, $\mathbf{C}X = \mathbf{I}$.

The variance of an unbiased $\tilde{\beta}$ is

$$V(\tilde{\beta}) = \mathbf{C}\mathbf{C}'\sigma^2.$$

Define

$$\mathbf{D} = \mathbf{C} - (X'X)^{-1}X'$$

Since $\mathbf{C}X = \mathbf{I}$, then $\mathbf{D}X = 0$ so

$$\begin{aligned} \operatorname{Var}(\tilde{\beta}|X) &= \left(\mathbf{D} + (X'X)^{-1}X'\right)\left(\mathbf{D} + (X'X)^{-1}X'\right)'\sigma^2 \\ &= \left(\mathbf{D}\mathbf{D}' + (X'X)^{-1}\right)\sigma^2 \;=\; \mathbf{D}\mathbf{D}'\sigma^2 + \operatorname{Var}(\hat{\beta}|X) \end{aligned}$$

Finally,

$$\operatorname{Var}(\tilde{\beta}) \geq V(\hat{\beta}).$$

Notice that the above results is true regardless the distribution of ε.

Small sample properties of the least squares estimator (3):

- *Unbiasedness* of $\hat{\sigma}^2 = \frac{1}{n-k}\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{n-k}\varepsilon' M_X \varepsilon$, where M_X is idempotent matrix that projects onto the space orthogonal to the span of X and $\hat{\varepsilon} = M_X y$.

Proof:

$$\begin{aligned} \mathrm{E}[\hat{\sigma}^2] &= \frac{1}{n-k}\mathrm{E}[\mathrm{Trace}(\varepsilon' M_X \varepsilon)] \\ &= \frac{1}{n-k}\mathrm{E}(\mathrm{Trace}(M_X \varepsilon\varepsilon')] \\ &= \frac{1}{n-k}\mathrm{Trace}(\mathrm{E}[M_X \varepsilon\varepsilon']) \\ &= \frac{1}{n-k}\sigma^2\mathrm{Trace}(M_X) \\ &= \sigma^2. \end{aligned}$$

Resampling methods - Introduction

Problem: Let $\boldsymbol{X} = (X_1, \ldots, X_N)$ be observations generated by a model $\mathcal{P}$, and let $T(\boldsymbol{X})$ be the statistics of interest that estimate the parameter θ. We want to know:

- Bias: $b_T = \mathrm{E}\left[T(\boldsymbol{X})\right] - \theta$,
- Variance: v_T,
- Distribution: $\mathcal{L}(T, \mathcal{P})$.

Resampling methods in i.i.d. data

Jackknife: Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$ be a sample of size N and let $T_N = T_N(\boldsymbol{X})$ be an estimator of θ.

In the *jackknife samples* an observation of $\boldsymbol{X}$ is excluded each time, i.e. $\boldsymbol{X}_{(i)} = (X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N)$ for $i = 1, 2, \ldots, N$, and we calculate the i-th jackknife statistic $T_{N-1,i} = T_{N-1}\left(\boldsymbol{X}_{(i)}\right)$.

Jackknife bias estimator:

$$b_{Jack} = (N-1)(\bar{T}_N - T_N)$$

Jackknife variance estimator:

$$v_{Jack} = \frac{N-1}{N} \sum_{i=1}^{N} \left(T_{N-1,i} - \bar{T}_N\right)^2,$$

where $\bar{T}_N = N^{-1} \sum_{i=1}^{N} T_{N-1,i}$

Jackknife

Outline of the resampling procedure:

$$(X_1,\ldots,X_N) \Rrightarrow \begin{cases} \boldsymbol{X}\backslash X_1 & \Rrightarrow T_{N-1,1} \\ \boldsymbol{X}\backslash X_2 & \Rrightarrow T_{N-1,2} \\ \vdots & \vdots \\ \boldsymbol{X}\backslash X_N & \Rrightarrow T_{N-1,N} \end{cases} \Rrightarrow \begin{cases} b_{Jack} = (N-1)(\bar{T}_N - T_N) \\ v_{Jack} = \frac{N-1}{N}\sum_{i=1}^{N}\left(T_{N-1,i} - \bar{T}_N\right)^2 \end{cases}$$

Possible uses:

- Bias reduction: $T_{Jack} = T_N - b_{Jack}$.
- If $\sqrt{N}(T_N - \theta)$ is asymptotically normal distributed, then we can obtain an estimate of the asymptotic variance by v_{Jack}.

Example # 1: Let $X_1, \ldots, X_N$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ observations and $T_N = \bar{X}$ be the statistic of interest. In this case, we know that:

$$b_T = 0, \quad v_T = \frac{\sigma^2}{N}, \quad \mathcal{L}(T) = \mathcal{N}(\mu, \frac{\sigma^2}{N}).$$

Jackknife results: $N = 100$, $\mu = 0$ and $\sigma^2 = 1$.

```
jackknife(data = x, statistic = mean)

Number of Replications: 100

Summary Statistics:
      Observed         Bias    Mean     SE
mean  -0.1081 -1.374e-015 -0.1081 0.1094
```

$$b_{jack} = -1.374e - 015 \approx 0 \text{ and } v_{jack} = 0,01196836 \approx 0.01 = \frac{1}{100}.$$

d-**jackknife**: Let $S_{n,r}$ be the subsets of $\{1, 2, \ldots, N\}$ of size r. For any $s = \{i_1, i_2, \ldots, i_r\} \in S_{N,r}$ we obtain the *d-jackknife replica* by $T_{r,s^c} = T_r(X_{i_1}, X_{i_2}, \ldots, X_{i_r})$.

d-jackknife variance estimator:

$$v_{Jack-d} = \frac{r}{dC} \sum_{s \in S_{N,r}} \left(T_{r,s^c} - C^{-1} \sum_{s \in S_{N,r}} T_{r,s^c} \right)^2,$$

where $C = \binom{N}{d}$.

d-jackknife as estimator of the distribution of T_N: Let $H_n(x) = \Pr\{\sqrt{n}(T_n - \theta) \leq x\}$ be the distribution we want to estimate. We define the *jackknife histogram* by:

$$H_{Jack}(x) = \frac{1}{C} \sum_{s \in S_{N,d}} I\left(\sqrt{Nr/d}(T_{r,s} - T_N) \leq x \right).$$

To learn more: Shao and Tu (1995).

Example # 2: We consider the following statistics: $T_n = \bar{X}_n^2$ and the sampling median $T_n = F_n^{-1}(\frac{1}{2}) = \widehat{Q}_2$:

$T_n = \bar{X}_n^2$				
	$F =$ normal,	$\sigma_F^2 = 6.25$	$F =$ exponential,	$\sigma_F^2 = 6.256$
d	RB	RM	RB	RM
1	1.2%	2.125	7.2%	4.992
10	1.2%	2.198	9.8%	5.152
20	1.6%	2.250	13.0%	5.520
$T_n = F_n^{-1}(\frac{1}{2})$				
	$F =$ normal,	$\sigma_F^2 = 6.28$	$F =$ Cauchy,	$\sigma_F^2 = 4.93$
d	RB	RM	RB	RM
1	92.2%	25.56	106.8%	24.02
10	21.8%	6.140	42.1%	6.327
20	8.9%	3.911	29.0%	4.030

$$RB = \widehat{\mathrm{E}}\left[\frac{v_{Jack} - v_T}{v_T}\right] \text{ and } RM = \widehat{\mathrm{E}}\left[(v_{Jack} - v_T)^2\right].$$

Resampling methods in i.i.d. data

Bootstrap: Let $\boldsymbol{X} = (X_1, \ldots, X_N)$ be observations generated by model P and let $T(\boldsymbol{X})$ be the statistic whose distribution $\mathcal{L}(T, P)$ we want to estimate.

The bootstrap proposes the distribution $\mathcal{L}^*(T^*; \widehat{P}_N)$ of $T^* = T(\boldsymbol{X}^*)$ as estimator of $\mathcal{L}(T, P)$, where $\boldsymbol{X}^*$ are observations generated by the estimated model $\widehat{\mathcal{P}}_N$.

- Possible estimators of P in the i.i.d. case:
  - ◇ Standard bootstrap: $F_N(x) = N^{-1} \sum_{i=1}^{N} I\left(X_i \leq x\right)$.
  - ◇ Parametric bootstrap: $F_{\widehat{\vartheta}}$ (Its assumed that $P = F_\vartheta$).
  - ◇ Smoothed bootstrap: $F_{n,h}$ is a kernel estimator of F.

To learn more: Efron y Tibshirani (1993), Shao and Tu (1995), and Davison and Hinkley (1997).

Standard bootstrap

Outline of the resampling procedure:

$$(X_1, \ldots, X_N) \Rrightarrow F_N \Rrightarrow \begin{cases} (X_1^{*\,(1)}, \ldots, X_N^{*\,(1)}) & \Rrightarrow T_N^{*\,(1)} \\ (X_1^{*\,(2)}, \ldots, X_N^{*\,(2)}) & \Rrightarrow T_N^{*\,(2)} \\ \vdots & \vdots \\ (X_1^{*\,(B)}, \ldots, X_N^{*\,(B)}) & \Rrightarrow T_N^{*\,(B)} \end{cases}$$

$$\Rrightarrow \begin{cases} b_{Boot} = \mathrm{E}^* \left[T_N^*\right] - T_N \\ v_{Boot} = \mathrm{E}^* \left[(T_N^* - \mathrm{E}^*[T_N^*])^2\right] \\ \mathcal{L}_T^*(x) = \mathrm{Pr}^* \left\{(T_N^* - T_N) \leq x\right\} \end{cases}$$

where E^* and Pr^* are the bootstrap expectation and the bootstrap probability.

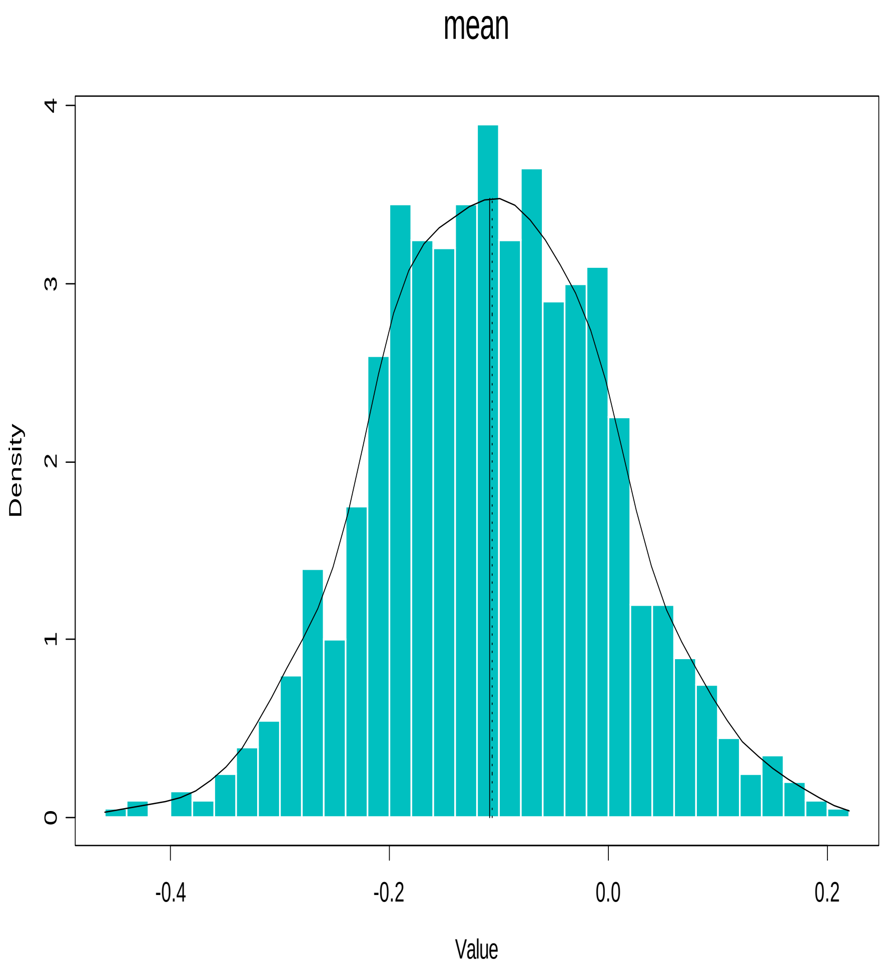
Remark: The X_i^* are i.i.d. F_N observations, therefore the resampling procedure could be interpreted as a sampling with replacement from the original sample $(X_1, \ldots, X_N)$.

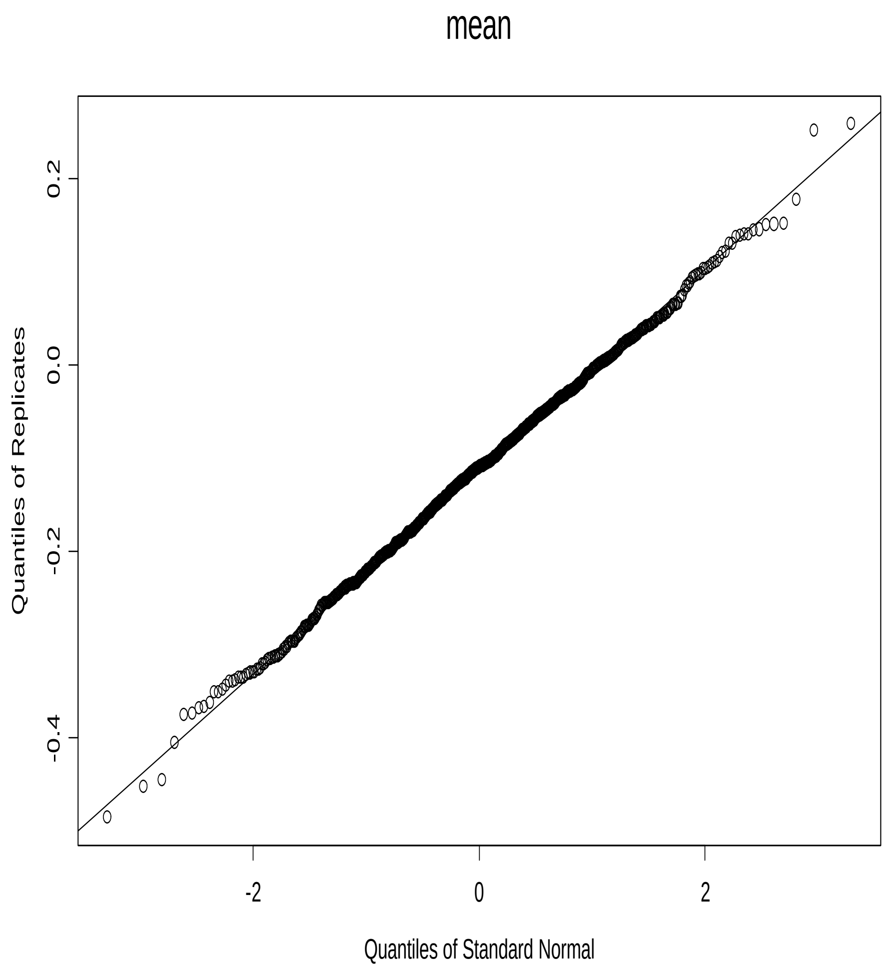
Example # 1 (cont.): $T_N = \bar{X}$ and B $=$ 1000 resamples of size N.

```
bootstrap(data = x.f, statistic = mean)

Number of Replications: 1000

Summary Statistics:
     Observed      Bias    Mean     SE
mean  -0.1081 -0.004746 -0.1129 0.1071

Empirical Percentiles:
        2.5%       5%      95%  97.5%
mean -0.3271 -0.2966 0.05394 0.0976
```

$$b_{boot} = -0.004746 \approx 0 \text{ y } v_{boot} = 0,01147041 \approx 0.01 = \tfrac{1}{100}.$$

Example # 1 (cont.):





Example # 3: Let $X_1, \ldots, X_N$ be i.i.d. observations $\mathcal{N}(\mu = 0, \sigma^2 = 1)$ and $T_N = \widehat{Q}_2$. In this case, we know that the asymptotic distribution of $\sqrt{N}(\widehat{Q}_2 - Q_2)$ is $\mathcal{N}(0, 1/4\phi(0)^2)$ and $v_T \approx 1/(4 \times 100 \times 0.3989423^2) = 0.01570796$.

Jackknife results:

```
jackknife(data = x.f, statistic = median)

Number of Replications: 100

Summary Statistics:
       Observed        Bias   Mean    SE
median   -0.251 -1.099e-014 -0.251 0.153

Empirical Percentiles:
          2.5%      5%     95%   97.5%
median -0.2664 -0.2664 -0.2356 -0.2356
```

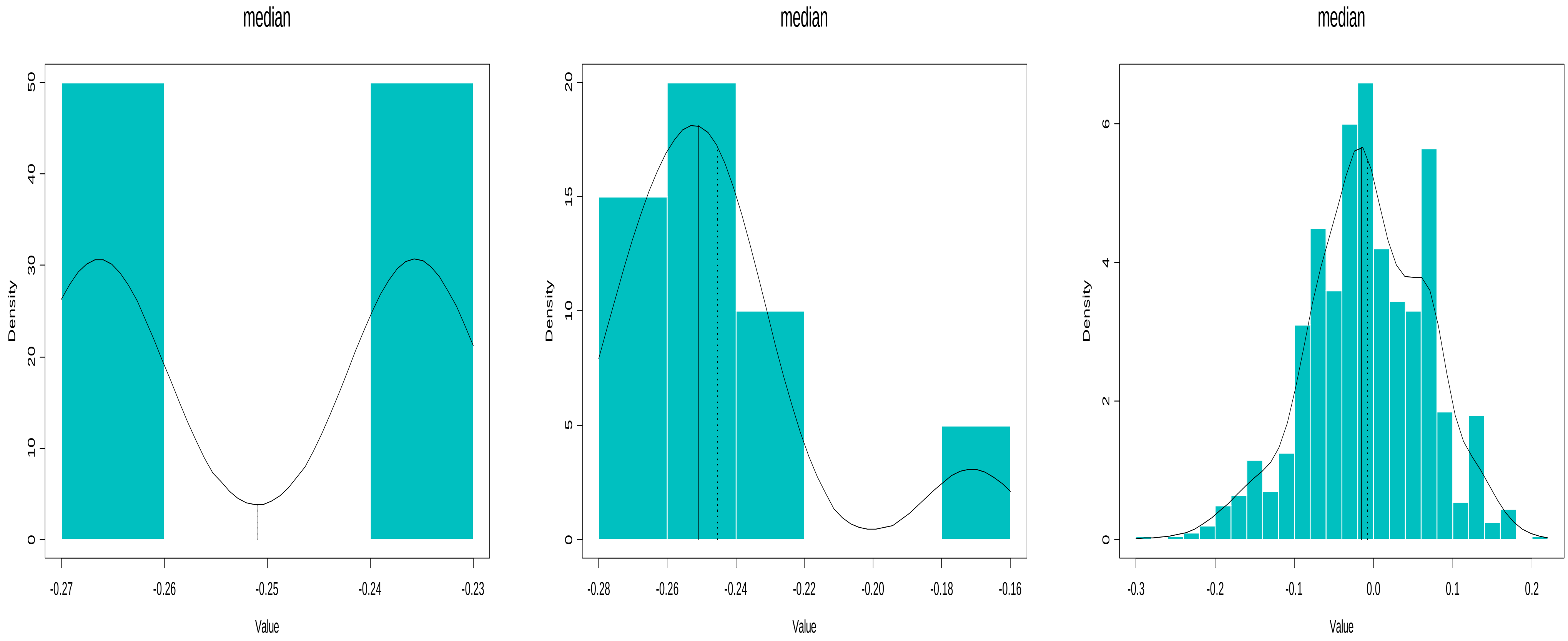
Example # 3 (cont.):

Bootstrap results:

```
bootstrap(data = x.f, statistic = median)

Number of Replications: 1000

Summary Statistics:
       Observed    Bias    Mean     SE
median   -0.251 -0,0027 -0.2537 0.1168

Empirical Percentiles:
          2.5%      5%     95%   97.5%
median -0.4614 -0.4333 0.03895 0.06271
```

$$v_{jack} = 0.023409 \text{ and } v_{boot} = 0.01364224.$$

Example # 3 (cont.):



Bootstrap always?

One example where the bootstrap fails: Let $X_1, \ldots, X_N$ be i.i.d. $\mathcal{U}(0, \theta)$. We know that the m.l.e. of θ is $T_N = \widehat{\theta} = \max_{1 \leq i \leq N} X_i$ and that the density function of T_N is:

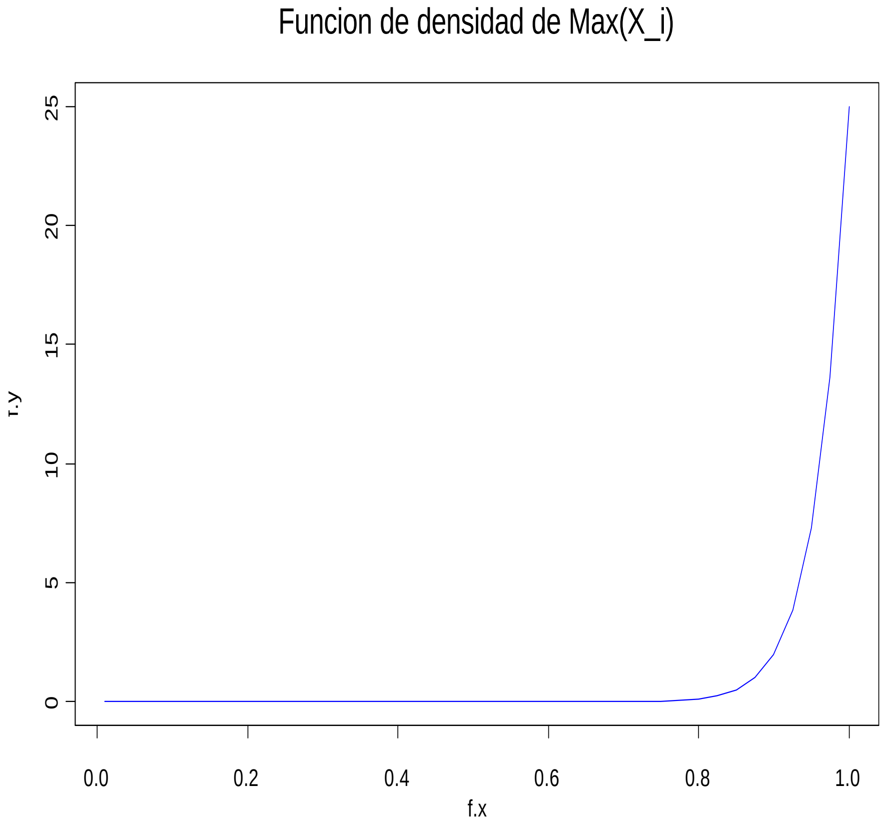
$$f_{\widehat{\theta}}(x) = \begin{cases} \frac{nx^{n-1}}{\theta^n} & \text{if} \quad 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

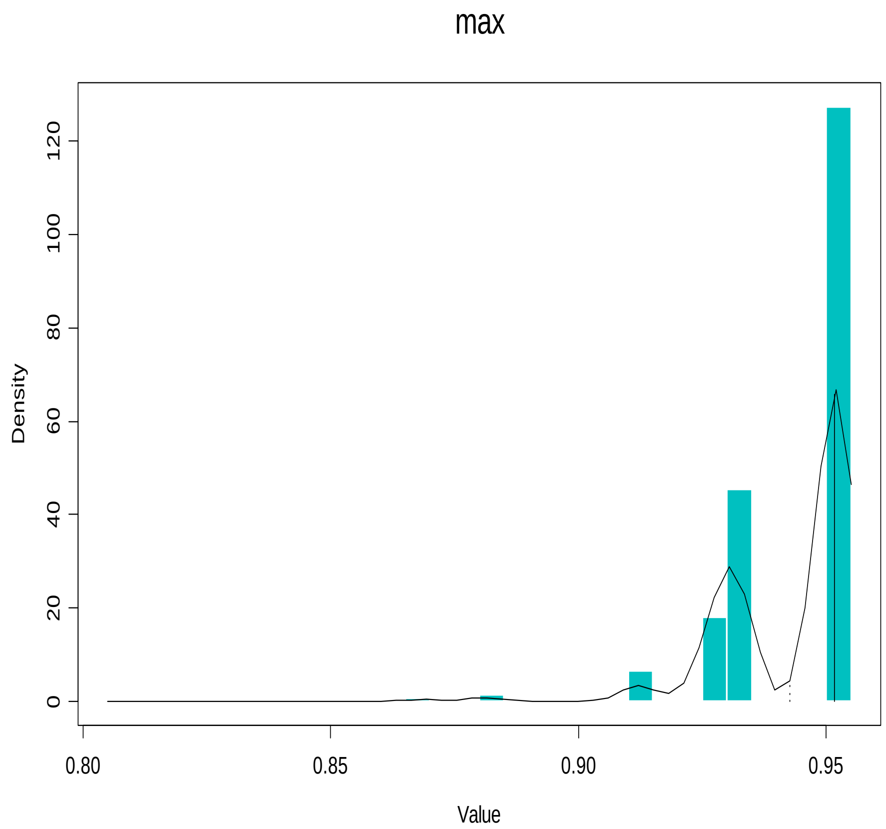
Bootstrap results:

```
bootstrap(data = x.f2, statistic = max)

Summary Statistics:
    Observed      Bias   Mean     SE
max   0.9517 -0.009143 0.9426 0.0143

Empirical Percentiles:
      2.5%     5%    95%  97.5%
max 0.9123 0.9269 0.9517 0.9517
```

Bootstrap always?

One example where the bootstrap fails (cont.):

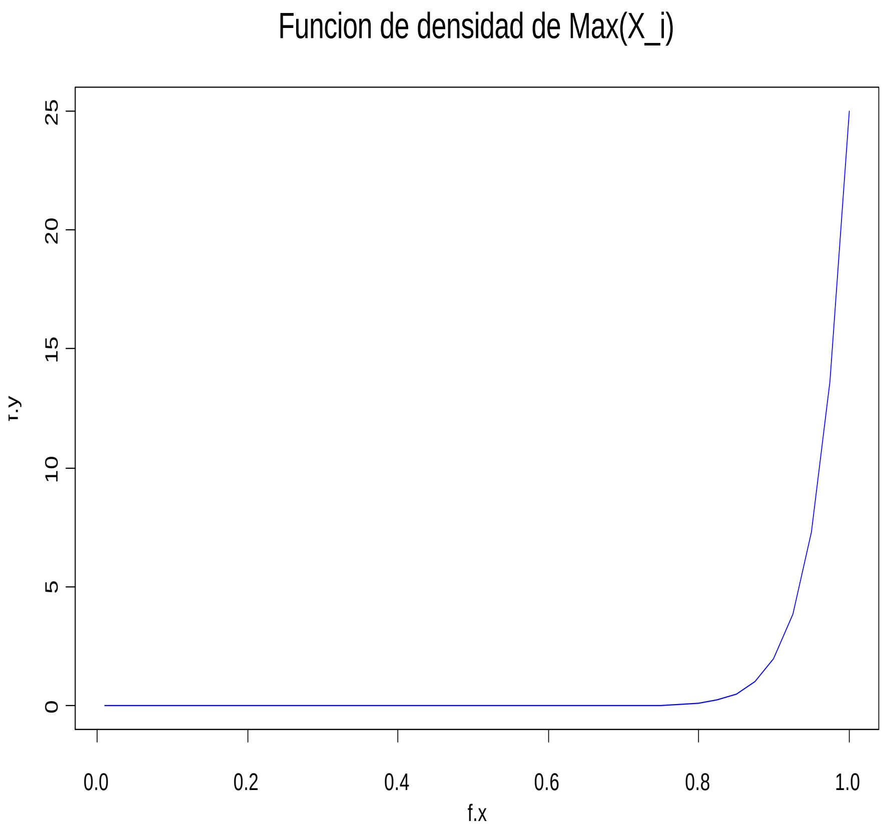


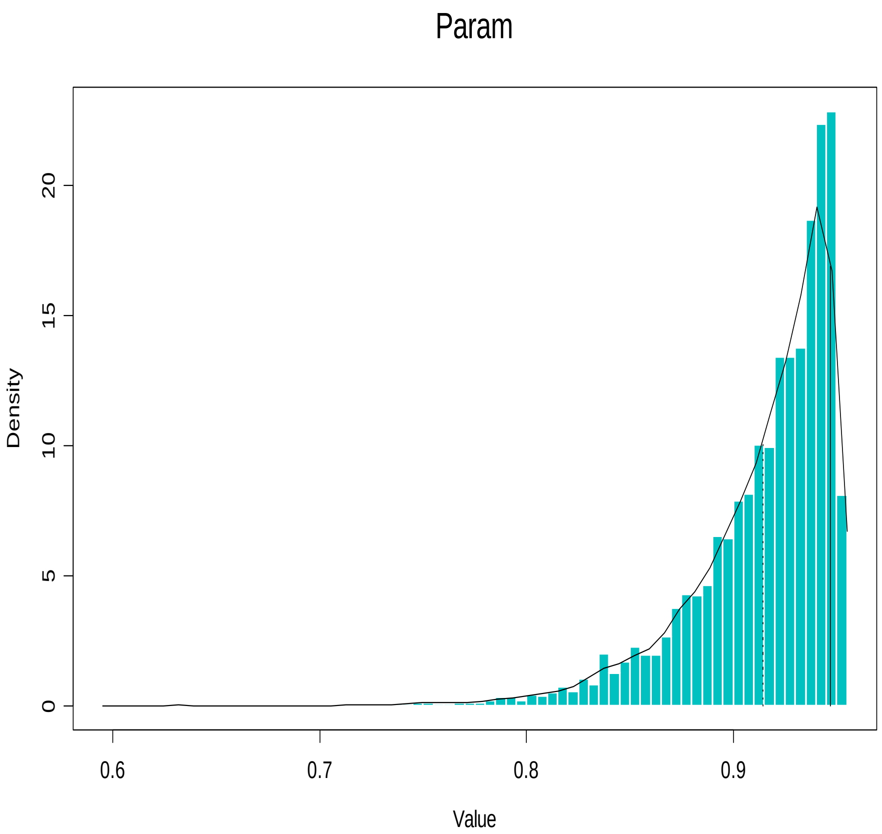


$$\Pr\{\widehat{\theta}^* = \widehat{\theta}\} = 0, \text{ but } \Pr^*\{\widehat{\theta}^* = \widehat{\theta}\} = 1 - (1 - 1/n)^n \to 1 - e^{-1}.$$

Bootstrap always?

One example where the bootstrap fails (cont.): A bootstrap solution is the *parametric bootstrap*:





Bootstrap in Linear Regression: Let $\big((Y_1, X_{1,1}, \ldots, X_{1,k}),$ $(Y_2, X_{2,1}, \ldots, X_{2,k}), \ldots, (Y_N, X_{N,1}, \ldots, X_{N,k})\big)$ be observations generated by the model:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k} + \varepsilon_i,$$

where the ε_i are i.i.d. with $\mathrm{E}[\varepsilon_i] = 0$, and $\mathrm{E}[\varepsilon_i^2] = \sigma^2$.

- Resampling of residuals (based on model):

1. We generate resamples $\varepsilon_1^*, \varepsilon_2^*, \ldots, \varepsilon_N^*$ i.i.d. from $F_{N,\widehat{\varepsilon}}$.
2. We construct bootstrap replicas by:

$$Y_i^* = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i,1} + \cdots + \widehat{\beta_k} X_{i,k} + \varepsilon_i^*,$$

   where the $\widehat{\beta}_i$ are estimates of the parameters β_i.
3. Finally, we compute the statistic of interest in the bootstrap replicas.

Bootstrap in Linear Regression:

- Resampling of "pairs" (not based on model):
  1. We generate bootstrap replicas $(Y_1^*, X_{1,1}^*, \ldots, X_{1,k}^*)$, $(Y_2^*, X_{2,1}^*, \ldots, X_{2,k}^*)$, $\ldots, (Y_N^*, X_{N,1}^*, \ldots, X_{N,k}^*))$ i.i.d. from $F_{N,(Y,X_1,\ldots,X_k)}$.
  2. We calculate the statistic of interest in the bootstrap replicas.

Remark: The bootstrap observations in the resampling of "pairs" could be interpreted as a sampling with replacement from the N vectors $p \times 1$ in the original data.

Remark: What happens if ε_i is not an i.i.d. sequence? For example, if the ε_i is an AR(1) process? or if it is an heteroscedastic sequence?

Example # 4: The data are taken from an operation of a plant for the oxidation of ammonia to nitric acid, measured on 21 consecutive days (Chapter 6 of Draper and Smith (1966)):

loss	percent of ammonia lost
Air.Flow	air flow to the plant
Water.Temp	cooling water inlet temperature
Acid.Conc.	acid concentration as a percentage

Bootstrap results:

```
bootstrap(data = stack, statistic = coef(lm(stack.loss ~ Air.Flow
    + Water.Temp + Acid.Conc., stack)), B = 1000, seed = 0, trace = F)

Summary Statistics:
            Observed      Bias     Mean     SE
(Intercept) -39.9197  0.829215 -39.0905 8.8239
   Air.Flow   0.7156  0.004886   0.7205 0.1749
 Water.Temp   1.2953 -0.031415   1.2639 0.4753
 Acid.Conc.  -0.1521 -0.005164  -0.1573 0.1180
```

Example # 4 (cont):

```
Summary Statistics:
              Observed      Bias     Mean     SE
(Intercept) -39.9197  0.829215 -39.0905 8.8239
   Air.Flow   0.7156  0.004886   0.7205 0.1749
 Water.Temp   1.2953 -0.031415   1.2639 0.4753
 Acid.Conc.  -0.1521 -0.005164  -0.1573 0.1180

Empirical Percentiles:
                2.5%        5%       95%     97.5%
(Intercept) -55.4846 -52.7583 -23.4913 -17.84522
   Air.Flow   0.3844   0.4454   1.0136   1.05255
 Water.Temp   0.3913   0.4768   2.0544   2.23920
 Acid.Conc.  -0.4181  -0.3604   0.0209   0.06103
```

Example # 4 (cont):