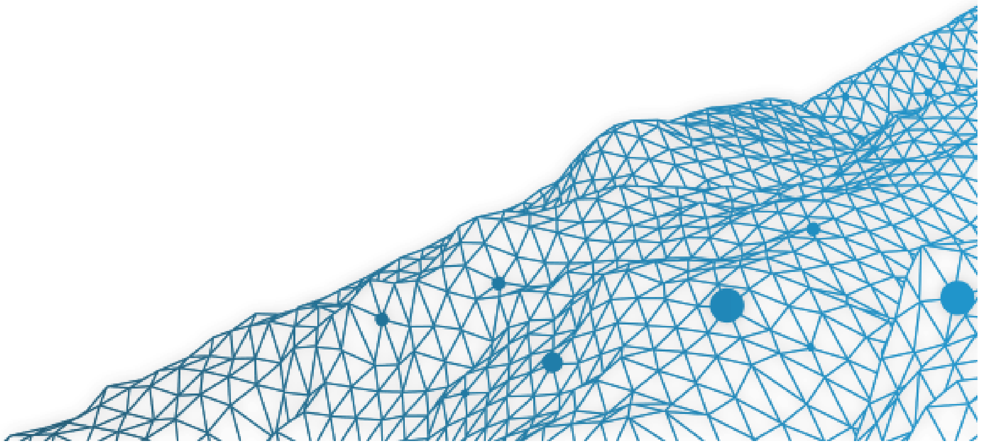
# Advanced Econometrics #2: Simulations & Bootstrap

A. Charpentier (Université de Rennes 1)

Université de Rennes 1,

Graduate Course, 2017.



## Motivation

Before computers, statistical analysis used probability theory to derive statistical expression for standard errors (or confidence intervals) and testing procedures, for some linear model

$$y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{j,i} + \varepsilon_i.$$

But most formulas are approximations, based on large samples ($n \to \infty$).

With computers, simulations and resampling methods can be used to produce (numerical) standard errors and testing procedure (without the use of formulas, but with a simple algorithm).

## Overview

Linear Regression Model: $y_i = \beta_0 + \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$

- Nonlinear Transformations : smoothing techniques
- Asymptotics vs. Finite Distance : boostrap techniques
- Penalization : Parcimony, Complexity and Overfit
- From least squares to other regressions : quantiles, expectiles, etc.

## Historical References

Permutation methods go back to Fisher (1935) The Design of Experiments and Pitman (1937) Significance tests which may be applied to samples from any population

(there are $n!$ distinct permutations)

Jackknife was introduced in Quenouille (1949) Approximate tests of correlation in time series, popularized by Tukey (1958) Bias and confidence in not quite large samples

Bootstrapping started with Monte Carlo algorithms in the 40's, see e.g. Simon & Burstein (1969) Basic Research Methods in Social Science

Efron (1979) Bootstrap methods: Another look at the jackknife defined a resampling procedure that was coined as "bootstrap".

(there are $n^n$ possible distinct ordered bootstrap samples)

# References

## Motivation

Bertrand, M., Duflo, E. & Mullainathan, 2004. Should we trust difference-in-difference estimators?. QJE.

## References

Davison, A.C. & Hinkley, D.V. 1997 Bootstrap Methods and Their Application. CUP.

Efron B. & Tibshirani, R.J. An Introduction to the Bootstrap. CRC Press.

Horowitz, J.L. 1998 The Bostrap, Handbook of Econometrics, North-Holland.

MacKinnon, J. 2007 Bootstrap Hypothesis Testing, Working Paper.

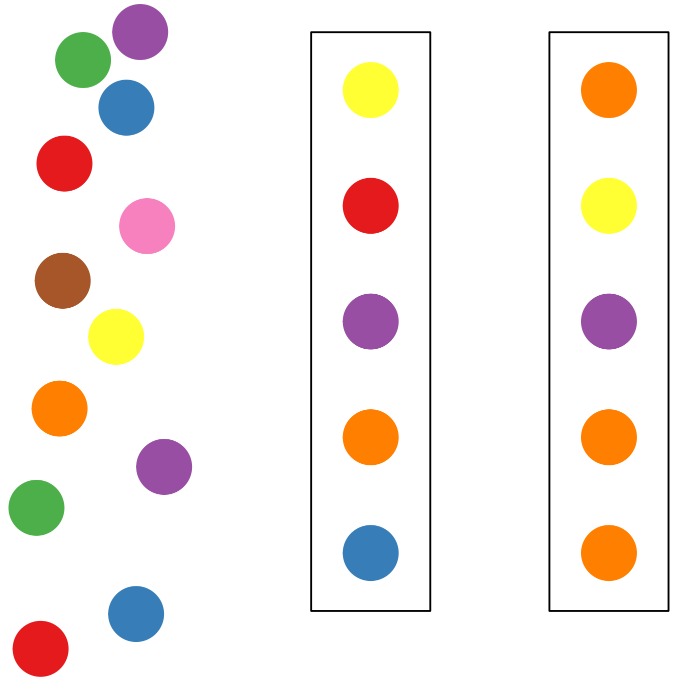
## Bootstrap Techniques (in one slide)

Bootstrapping is an asymptotic refinement based on computer based simulations.

Underlying properties: we know when it might work, or not

Idea : $\{(y_i, \boldsymbol{x}_i)\}$ is obtained from a stochastic model under $\mathbb{P}$

We want to generate other samples (not more observations) to reduce uncertainty.



## Heuristic Intuition for a Simple (Financial) Model

Consider a return stochastic model, $r_t = \mu + \sigma\varepsilon_t$, for $t = 1, 2, \cdots, T$, with $(\varepsilon_t)$ is i.id. $\mathcal{N}(0, 1)$ [Constant Expected Return Model, CER]

$$\widehat{\mu} = \frac{1}{T}\sum_{t=1}^{T} r_t \text{ and } \widehat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\big[r_t - \widehat{\mu}\big]^2$$

then (standard errors)

$$\widehat{se}[\widehat{\mu}] = \frac{\widehat{\sigma}}{\sqrt{T}} \text{ and } \widehat{se}[\widehat{\sigma}] = \frac{\widehat{\sigma}}{\sqrt{2T}}$$

then (confidence intervals)

$$\mu \in \Big[\widehat{\mu} \pm 2\widehat{se}[\widehat{\mu}]\Big] \text{ and } \sigma \in \Big[\widehat{\sigma} \pm 2\widehat{se}[\widehat{\sigma}]\Big]$$

What if the quantity of interest, $\theta$, is another quantity, e.g. a Value-at-Risk ?

## Heuristic Intuition for a Simple (Financial) Model

One can use nonparametric bootstrap

1. resampling: generate $B$ "bootstrap samples" by resampling with replacement in the original data,

$$\boldsymbol{r}^{(b)} = \{r_1^{(b)}, \cdots, r_T^{(b)}\}, \text{ with } r_t^{(b)} \in \{r_1, \cdots, r_T\}.$$

2. For each sample $\boldsymbol{r}^{(b)}$, compute $\widehat{\theta}^{(b)}$

3. Derive the empirical distribution of $\widehat{\theta}$ from $\{\widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(B)}\}$.

4. Compute any quantity of interest, standard error, quantiles, etc.

E.g. estimate the bias

$$\text{bias}[\widehat{\theta}] = \underbrace{\frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^{(b)}}_{\text{bootstrap mean}} - \underbrace{\frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}}_{\text{estimate}}$$

## Heuristic Intuition for a Simple (Financial) Model

E.g. estimate the standard error

$$\text{se}[\widehat{\theta}] = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta}^{(b)} - \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^{(b)}\right)^2}$$

E.g. estimate the confidence interval, if the bootstrap distribution looks Gaussian

$$\theta \in \left[\widehat{\theta} \pm 2\text{se}[\widehat{\theta}]\right]$$

and if the distribution does not look Gaussian

$$\theta \in \left[q_{\alpha/2}^{(B)}; q_{1-\alpha/2}^{(B)}\right]$$

where $q_{\alpha}^{(B)}$ denote a quantile from $\{\widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(B)}\}$.

## Monte Carlo Techniques in Statistics

Law of large numbers (---), if $\mathbb{E}[X]=0$ and $\text{Var}[X]=1$ : $\sqrt{n}\,\overline{X}_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$

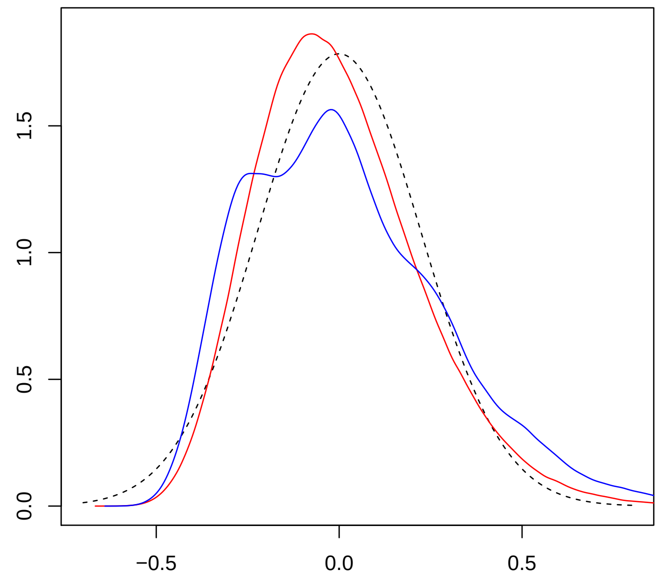
What if $n$ is small? What is the distribution of $\overline{X}_n$?

Example : $X$ such that $2^{-\frac{1}{2}}(X-1)\sim\chi^2(1)$

Use Monte Carlo Simulation to derive confidence interval for $\overline{X}_n$ (—).

Generate samples $\{x_1^{(m)},\cdots,x_n^{(m)}\}$ from $\chi^2(1)$, and compute $\overline{x}_n^{(m)}$

Then estimate the density of $\{\overline{x}_n^{(1)},\cdots,\overline{x}_n^{(m)}\}$, quantiles, etc.




Problem : need to know the true distribution of $X$.

What if we have only $\{x_1,\cdots,x_n\}$ ?

Generate samples $\{x_1^{(m)},\cdots,x_n^{(m)}\}$ from $\widehat{F}_n$, and compute $\overline{x}_n^{(m)}$ (—)

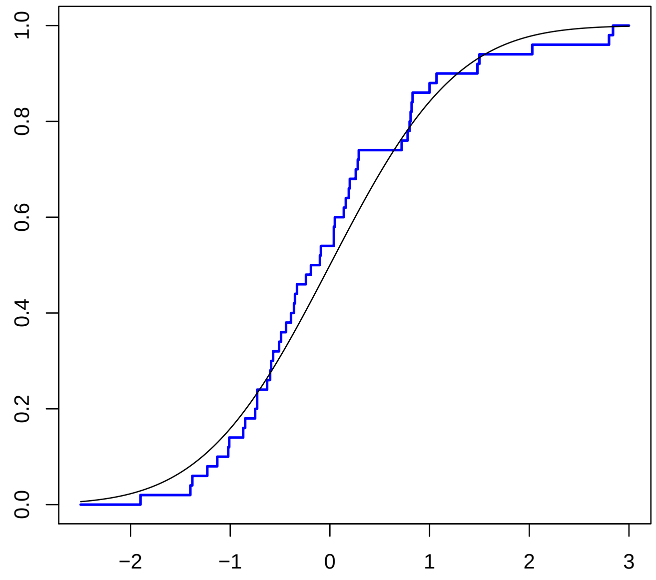
## Monte Carlo Techniques in Statistics

Consider empirical residuals from a linear regression, $\widehat{\varepsilon}_i = y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}$.

Let $\widehat{F}(z) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \mathbf{1}\left(\frac{\widehat{\varepsilon}_i}{\widehat{\sigma}} \leq z\right)$ denote the empirical distribution of Studentized residuals.

Could we test $H_0 : F = \mathcal{N}(0,1)$?

```
> X <- rnorm(50)
> cdf <- function(z) mean(X<=z)
```




Simulate samples from a $\mathcal{N}(0,1)$ (true distribution under $H_0$)

## Quantifying Bias

Consider $X$ with mean $\mu = \mathbb{E}(X)$. Let $\theta = \exp[\mu]$, then $\widehat{\theta} = \exp[\overline{x}]$ is a biased estimator of $\theta$, see Horowitz (1998) The Bostrap

**Idea 1** : Delta Method, i.e. if $\sqrt{n}[\widehat{\tau}_n - \tau] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$, then, if $g'(\tau)$ exists and is non-null,

$$\sqrt{n}[g(\widehat{\tau}_n) - g(\tau)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2[g'(\tau)]^2)$$

so $\widehat{\theta}_1 = \exp[\overline{x}]$ is asymptotically unbiased.

**Idea 2** : Delta Method based correction,

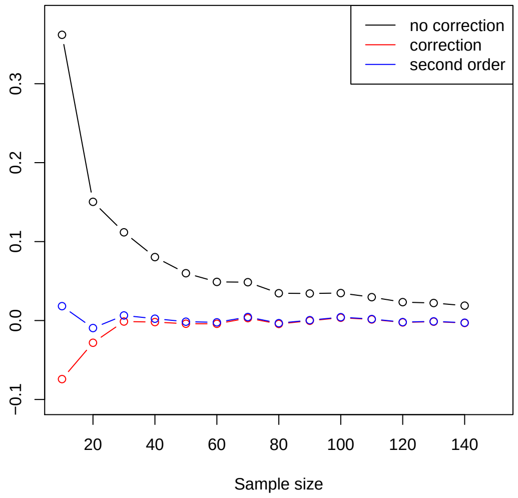
based on $\widehat{\theta}_2 = \exp\left[\overline{x} - \frac{s^2}{2n}\right]$ where $s^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}[x_i - \overline{x}]^2$.

**Idea 3** : Use Bootstrap, $\widehat{\theta}_3 = \dfrac{1}{B}\displaystyle\sum_{b=1}^{n}\exp[\overline{x}^{(b)}]$

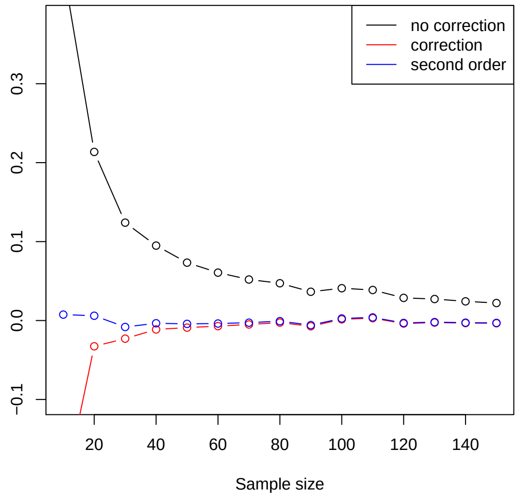
## Quantifying Bias

$X$ with mean $\mu = \mathbb{E}(X)$. Let $\theta = \exp[\mu]$. Consider three distributions

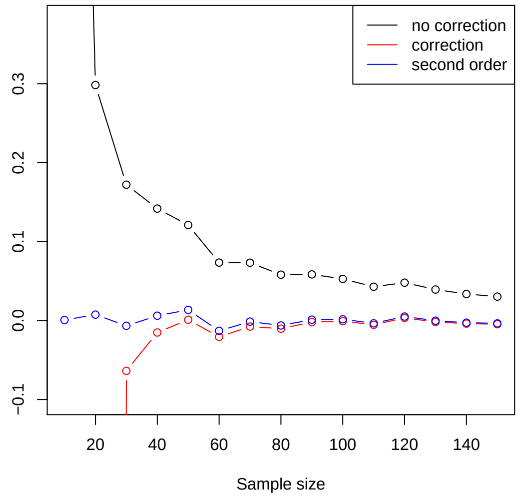
Log-normal




Student $t_{10}$

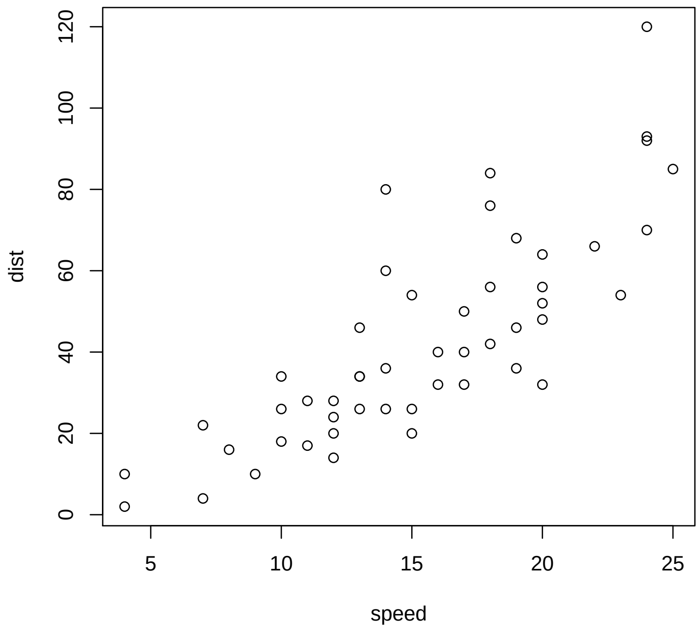



Student $t_5$



## Linear Regression & Bootstrap : Parametric

1. sample $\widetilde{\varepsilon}_1^{(s)},\cdots,\widetilde{\varepsilon}_n^{(s)}$ randomly from $\mathcal{N}(0,\widehat{\sigma})$
2. set $y_i^{(s)}=\widehat{\beta}_0+\widehat{\beta}_1x_i+\widetilde{\varepsilon}_i^{(s)}$
3. consider dataset $(x_i,y_i^{(b)})=(x_i,y_i^{(b)})$'s and fit a linear regression
4. let $\widehat{\beta}_0^{(s)},\widehat{\beta}_1^{(s)}$ and $\widehat{\sigma}^{2(s)}$ denote the estimated values



## Linear Regression & Bootstrap : Residuals

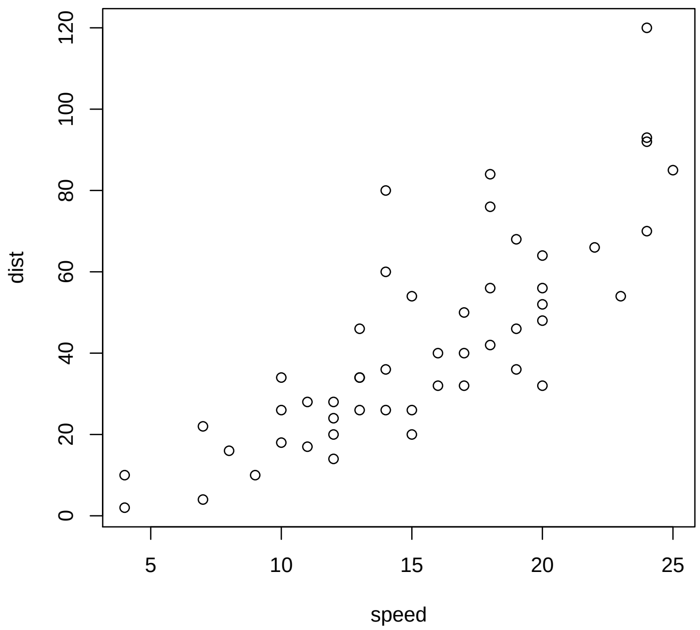
**Algorithm 6.1.** Davison & Hinkley (1997) Bootstrap Methods and Applications.

1. sample $\widehat{\varepsilon}_1^{(b)},\cdots,\widehat{\varepsilon}_n^{(b)}$ randomly with replacement in $\{\widehat{\varepsilon}_1,\widehat{\varepsilon}_2,\cdots,\widehat{\varepsilon}_n\}$
2. set $y_i^{(b)}=\widehat{\beta}_0+\widehat{\beta}_1x_i+\widehat{\varepsilon}_i^{(b)}$
3. consider dataset $(x_i,y_i^{(b)})=(x_i,y_i^{(b)})$'s and fit a linear regression
4. let $\widehat{\beta}_0^{(b)},\widehat{\beta}_1^{(b)}$ and $\widehat{\sigma}^{2(b)}$ denote estimated values

$$\widehat{\beta}_1^{(b)}=\frac{\sum[x_i-\overline{x}]\cdot y_i^{(b)}}{\sum[x_i-\overline{x}]^2}=\widehat{\beta}_1+\frac{\sum[x_i-\overline{x}]\cdot\widehat{\varepsilon}_i^{(b)}}{\sum[x_i-\overline{x}]^2}$$

hence $\mathbb{E}[\widehat{\beta}_1^{(b)}]=\widehat{\beta}_1$, while

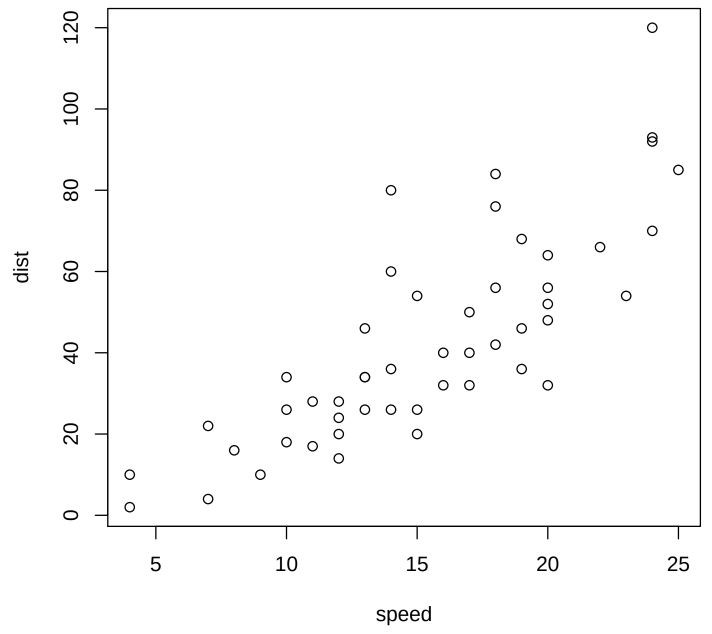
$$\text{Var}[\widehat{\beta}_1^{(b)}]=\frac{\sum[x_i-\overline{x}]^2\cdot\text{Var}[\widehat{\varepsilon}_i^{(b)}]}{\left(\sum[x_i-\overline{x}]^2\right)^2}\sim\frac{\sigma^2}{\sum[x_i-\overline{x}]^2}$$



## Linear Regression & Bootstrap : Pairs

**Algorithm 6.2.** Davison & Hinkley (1997) Bootstrap Methods and Applications.

1. sample $\{i_1^{(b)},\cdots,i_n^{(b)}\}$ randomly with replacement in $\{1,2,\cdots,n\}$
2. consider dataset $(x_i^{(b)},y_i^{(b)})=(x_{i_i^{(b)}},y_{i_i^{(b)}})$'s and fit a linear regression
3. let $\widehat{\beta}_0^{(b)},\widehat{\beta}_1^{(b)}$ and $\widehat{\sigma}^{2(b)}$ denote the estimated values




**Remark** $\mathbb{P}(i\notin\{i_1^{(b)},\cdots,i_n^{(b)}\})=\left(1-\dfrac{1}{n}\right)^n\sim e^{-1}$

Key issue : residuals have to be independent and identically distributed

## Linear Regression & Bootstrap

Difference between the two algorithms:

1) with the second method, we make no assumption about variance homogeneity

potentially more robust to heteroscedasticity

2) the simulated samples have different designs, because the $x$ values are randomly sampled

Key issue : residuals have to be independent and identically distributed

See discussion below on

- dynamic regression, $y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \varepsilon_t$
- heteroskedasticity, $y_i = \beta_0 + \beta_1 x_i + |x_i \cdot |\varepsilon_t$
- instrumental variables and two-stage least squares

## Monte Carlo Techniques to Compute Integrals

Monte Carlo is a very general technique, that can be used to compute any integral.

Let $X \sim \mathcal{C}auchy$ what is $\mathbb{P}[X > 2]$. Observe that

$$\mathbb{P}[X > 2] = \int_2^\infty \frac{dx}{\pi(1+x^2)} \qquad (\sim 0.15)$$

since $f(x) = \dfrac{1}{\pi(1+x^2)}$ and $Q(u) = F^{-1}(u) = \tan\left(\pi\left[u - \frac{1}{2}\right]\right)$.

Crude Monte Carlo: use the law of large numbers

$$\widehat{p}_1 = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(Q(u_i) > 2)$$

where $u_i$ are obtained from i.id. $\mathcal{U}([0,1])$ variables.

Observe that $\text{Var}[\widehat{p}_1] \sim \frac{0.127}{n}$.

Crude Monte Carlo (with symmetry): $\mathbb{P}[X>2]=\mathbb{P}[|X|>2]/2$ and use the law of large numbers

$$\widehat{p}_2=\frac{1}{2n}\sum_{i=1}^n \mathbf{1}(|Q(u_i)|>2)$$

where $u_i$ are obtained from i.id. $\mathcal{U}([0,1])$ variables.

Observe that $\text{Var}[\widehat{p}_2]\sim\frac{0.052}{n}$.

Using integral symmetries :

$$\int_2^\infty \frac{dx}{\pi(1+x^2)}=\frac{1}{2}-\int_0^2\frac{dx}{\pi(1+x^2)}$$

where the later integral is $\mathbb{E}[h(2U)]$ where $h(x)=\dfrac{2}{\pi(1+x^2)}$.

From the law of large numbers

$$\widehat{p}_3=\frac{1}{2}-\frac{1}{n}\sum_{i=1}^n h(2u_i)$$

where $u_i$ are obtained from i.id. $\mathcal{U}([0,1])$ variables.

Observe that $\text{Var}[\widehat{p}_3] \sim \frac{0.0285}{n}$.

Using integral transformations :

$$\int_2^\infty \frac{dx}{\pi(1+x^2)} = \int_0^{1/2} \frac{y^{-2}dy}{\pi(1-y^{-2})}$$

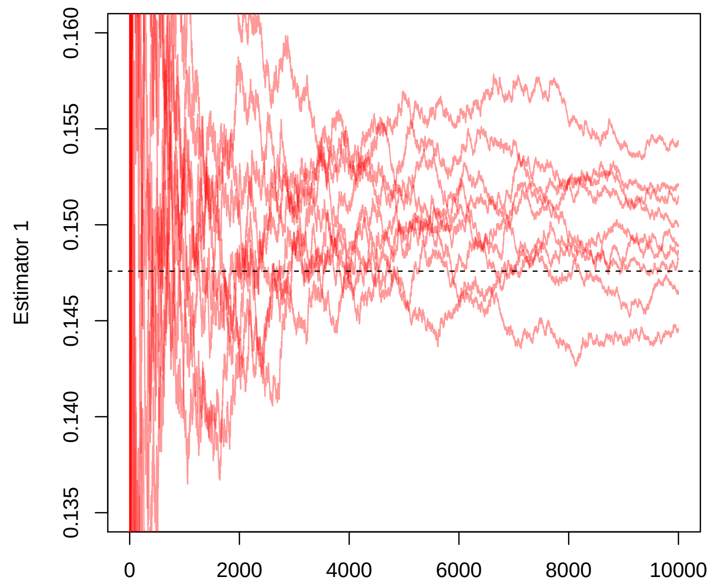
which is $\mathbb{E}[h(U/2)]$ where $h(x) = \dfrac{1}{2\pi(1+x^2)}$.

From the law of large numbers

$$\widehat{p}_4 = \frac{1}{4n}\sum_{i=1}^n h(u_i/2)$$

where $u_i$ are obtained from i.id. $\mathcal{U}([0,1])$ variables.

Observe that $\text{Var}[\widehat{p}_4] \sim \frac{0.0009}{n}$.



## Simulation in Econometric Models

(almost) all quantities of interest can be writen $T(\varepsilon)$ with $\varepsilon \sim F$.

E.g. $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{\varepsilon}$

We need $\mathbb{E}[T(\varepsilon)] = \displaystyle\int t(\epsilon)dF(\epsilon)$

Use simulations, i.e. draw $n$ values $\{\epsilon_1, \cdots, \epsilon_n\}$ since

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n T(\epsilon_i)\right] = \mathbb{E}[T(\varepsilon)] \text{ (unbiased)}$$

$$\frac{1}{n}\sum_{i=1}^n T(\epsilon_i) \xrightarrow{\mathcal{L}} \mathbb{E}[T(\varepsilon)] \text{ as } n \to \infty \text{ (consistent)}$$

## Generating (Parametric) Distributions

Inverse cdf Technique :

Let $U \sim \mathcal{U}([0,1])$, then $X = F^{-1}(U) \sim F$.

Proof 1:

$$\mathbb{P}[F^{-1}(U) \leq x] = \mathbb{P}[F \circ F^{-1}(U) \leq F(x)] = \mathbb{P}[U \leq F(x)] = F(x)$$

Proof 2: set $u = F(x)$ or $x = F^{-1}(u)$ (change of variable)

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x)dF^{\star}(x) = \int_0^1 h(F^{-1}(u))du = \mathbb{E}[h(F^{-1}(U))]$$

with $U \sim \mathcal{U}([0,1])$, i.e. $X \stackrel{\mathcal{L}}{=} F^{-1}(U)$.

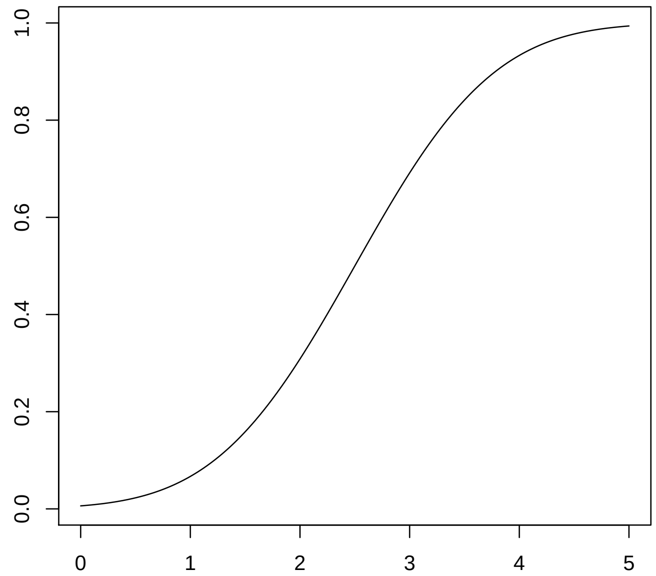
## Rejection Techniques

Problem : If $X \sim F$, how to draw from $X^\star$, i.e. $X$ conditional on $X \in [a,b]$ ?

Solution : draw $X$ and use accept-reject method

1. if $x \in [a,b]$, keep it (accept)
2. if $x \notin [a,b]$, draw another value (reject)

If we generate $n$ values, we accept - on average - $[F(b) - F(a)] \cdot n$ draws.



## Importance Sampling

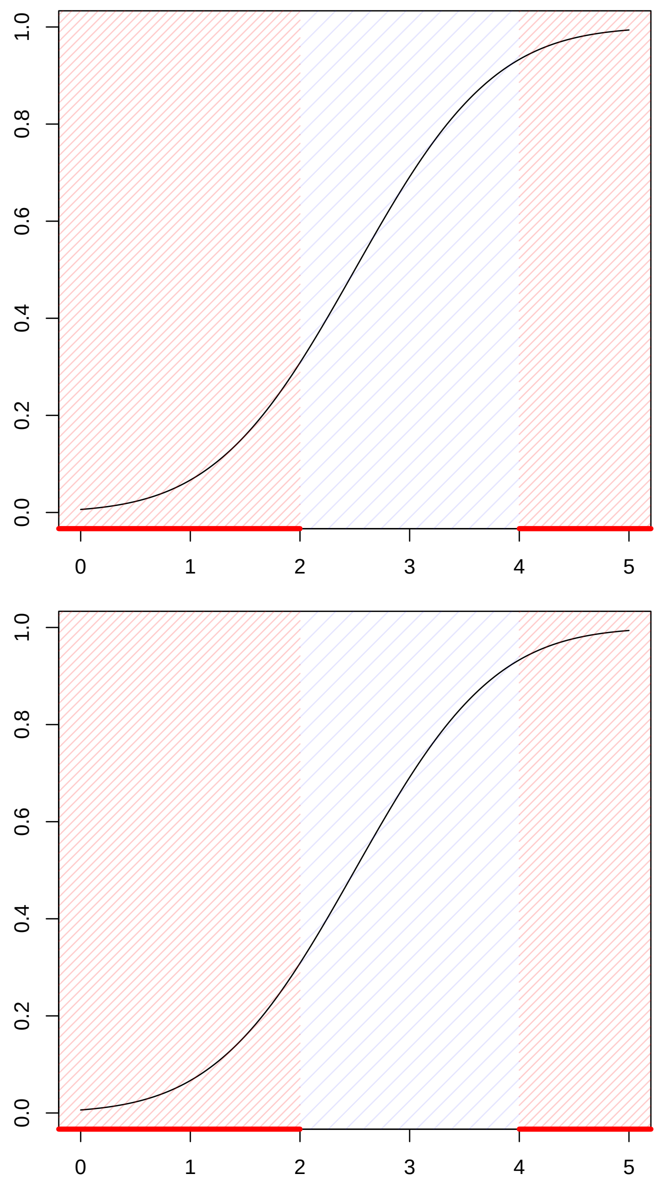
Problem : If $X \sim F$, how to draw from $X$ conditional on $X \in [a, b]$ ?

Solution : rewrite the integral and use importance sampling method

The conditional censored distribution $X^\star$ is

$$dF^\star(x) = \frac{dF(x)}{F(b) - F(a)} \mathbf{1}(x \in [a, b])$$

Alternative for truncated distributions : let $U \sim \mathcal{U}([0, 1])$ and set $\tilde{U} = [1 - U]F(a) + UF(b)$ and $Y = F^{-1}(\tilde{U})$



## Going Further : MCMC

Intuition : we want to use the Central Limit Theorem, but i.id. sample is a (too) strong assumtion: if $(X_i)$ is i.id. with distribution $F$,

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^n h(X_i) - \int h(x)dF(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\sigma^2), \text{ as } n\to\infty.$$

Use the ergodic theorem: if $(X_i)$ is a Markov Chain with invariant measure $\mu$,

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^n h(X_i) - \int h(x)d\mu(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\sigma^2), \text{ as } n\to\infty.$$

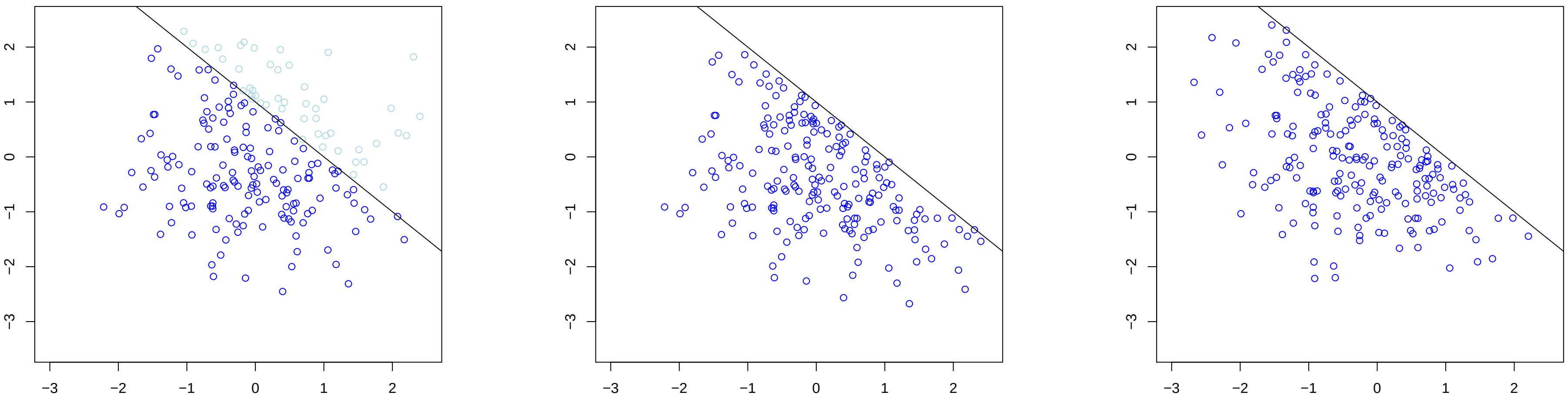
See Gibbs sampler

Example : complicated joint distribution, but simple conditional ones

## Going Further : MCMC

To generate $\boldsymbol{X}|\boldsymbol{X}^\mathsf{T}\mathbf{1} \leq m$ with $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ (in dimension 2)

1. draw $X_1$ from $\mathcal{N}(0,1)$
2. draw $U$ from $\mathcal{U}([0,1])$ and set $\tilde{U} = U\Phi(m - \epsilon_1)$
3. set $X_2 = \Phi^{-1}(\tilde{U})$




See Geweke (1991) Efficient Simulation from the Multivariate Normal and Distributions Subject to Linear Constraints

## Monte Carlo Techniques in Statistics

Let $\{y_1, \cdots, y_n\}$ denote a sample from a collection of $n$ i.id. random variables with true (unknown) distribution $F_0$. This distribution can be approximated by $\widehat{F}_n$.

parametric model : $F_0 \in \mathcal{F} = \{F_\theta; \theta \in \Theta\}$.

nonparametric model : $F_0 \in \mathcal{F} = \{F \text{ is a c.d.f.}\}$

The statistic of interest is $T_n = T_n(y_1, \cdots, y_n)$ (see e.g. $T_n = \widehat{\beta}_j$).

Let $G_n$ denote the statistics of $T_n$:

Exact distribution : $G_n(t, F_0) = \mathbb{P}_F(T_n \leq t)$ under $F_0$

We want to estimate $G_n(\cdot, F_0)$ to get confidence intervals, i.e. $\alpha$-quantiles

$$G_n^{-1}(\alpha, F_0) = \inf\left\{t; G_n(t, F_0) \geq \alpha\right\}$$

or $p$-values,

$$p = 1 - G_n(t_n, F_0)$$

## Approximation of $G_n(t_n, F_0)$

Two strategies to approximate $G_n(t_n, F_0)$ :

1. Use $G_\infty(\cdot, F_0)$, the asymptotic distribution as $n \to \infty$.
2. Use $G_\infty(\cdot, \widehat{F}_n)$

Here $\widehat{F}_n$ can be the empirical cdf (nonparametric bootstrap) or $F_{\widehat{\theta}}$ (parametric bootstrap).

## Approximation of $G_n(t_n, F_0)$: Linear Model

Consider the test of $H_0 : \beta_j = 0$, $p$-value being $p = 1 - G_n(t_n, F_0)$

- Linear Model with Normal Errors $y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

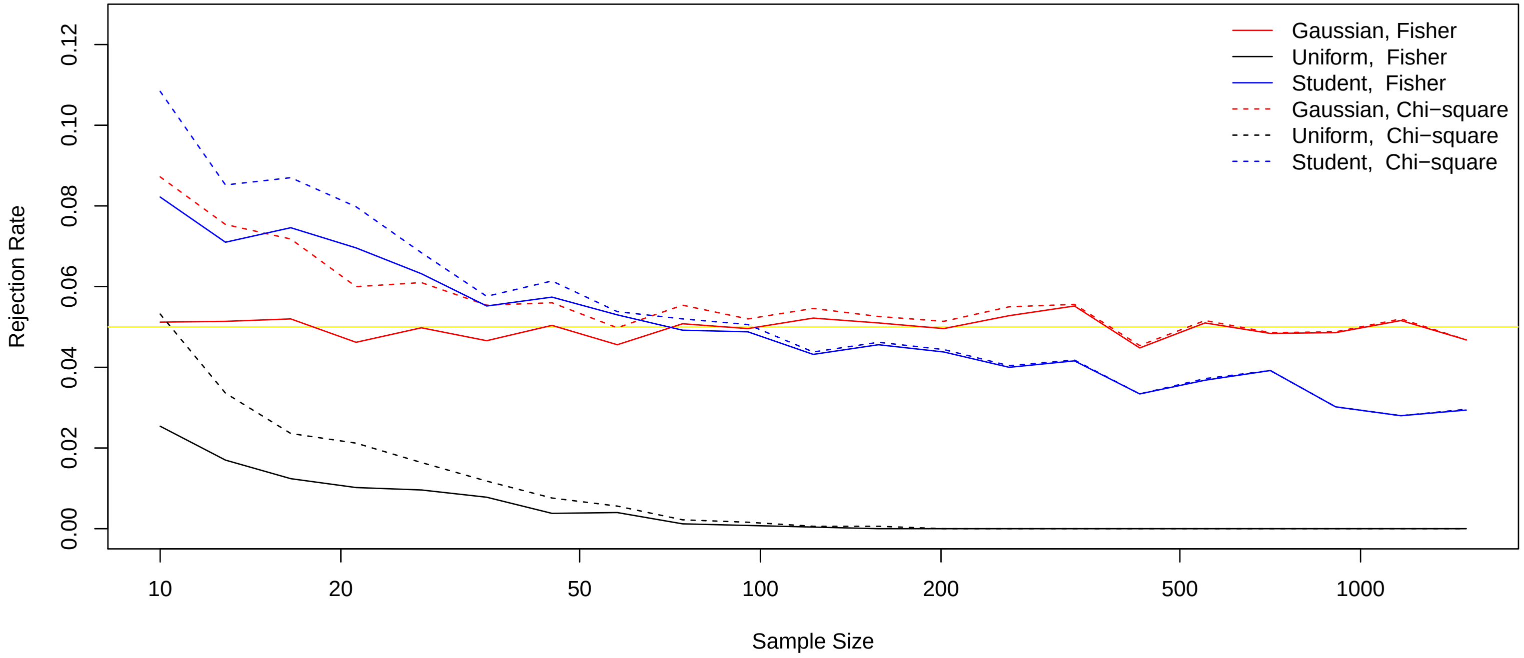
Then $\dfrac{(\widehat{\beta}_j - \beta_j)^2}{\widehat{\sigma}_j^2} \sim \mathcal{F}(1, n-k) = G_n(\cdot, F_0)$ where $F_0$ is $\mathcal{N}(0, \sigma^2)$

- Linear Model with Non-Normal Errors $y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i$, with $\mathbb{E}[\varepsilon_i] = 0$.

Then $\dfrac{(\widehat{\beta}_j - \beta_j)^2}{\widehat{\sigma}_j^2} \xrightarrow{\mathcal{L}} \xi^2(1) = G_\infty(\cdot, F_0)$ as $n \to \infty$.

## Approximation of $G_n(t_n, F_0)$: Linear Model

Application $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$, $\varepsilon \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{U}([-1,+1])$ or $\varepsilon \sim \mathcal{S}td(\nu = 2)$.




Here $F_0$ is $\mathcal{N}(0, \sigma^2)$

## Computation of $G_\infty(t,\widehat{F}_n)$

For $b\in\{1,\cdots,B\}$, generate boostrap samples of size $n$, $\{\widehat{\varepsilon}_1^{(b)},\cdots,\widehat{\varepsilon}_n^{(b)}\}$ by drawing from $\widehat{F}_n$.

Compute $T^{(b)}=T_n(\widehat{\varepsilon}_1^{(b)},\cdots,\widehat{\varepsilon}_n^{(b)})$, and use sample $\{T^{(1)},\cdots,T^{(B)}\}$ to compute $\widehat{G}$,

$$\widehat{G}(t)=\frac{1}{B}\sum_{b=1}^B \mathbf{1}(T^{(b)}\leq t)$$

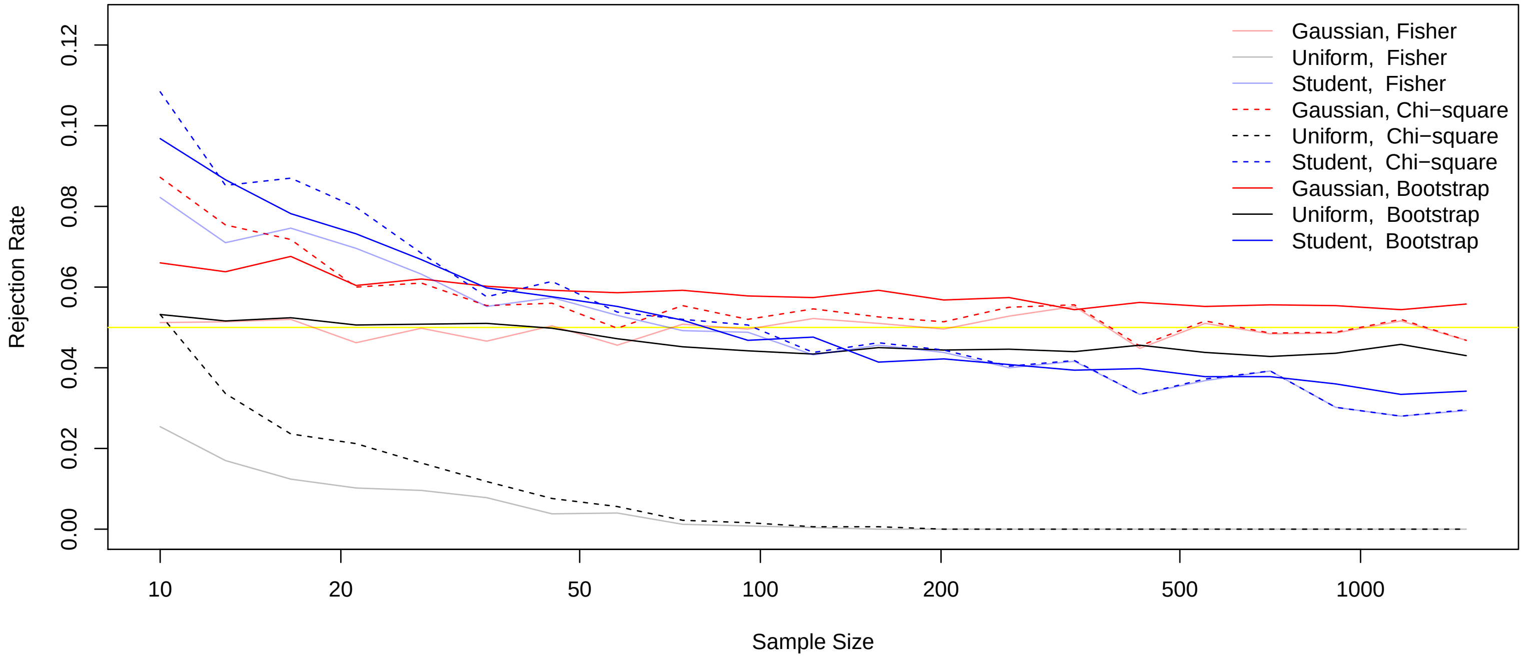
## Linear Model: computation of $G_\infty(t, \widehat{F}_n)$

Consider the test of $H_0 : \beta_j = 0$, $p$-value being $p = 1 - G_n(t_n, F_0)$

1. compute $t_n = \dfrac{(\widehat{\beta}_j - \beta_j)^2}{\widehat{\sigma}_j^2}$
2. generate $B$ boostrap samples, under the null assumption
3. for each boostrap sample, compute $t_n^{(b)} = \dfrac{(\widehat{\beta}_j^{(b)} - \widehat{\beta}_j)^2}{\widehat{\sigma}_j^{2(b)}}$
4. reject $H_0$ if $\displaystyle\frac{1}{B}\sum_{i=1}^{B} \mathbf{1}(t_n > t_n^{(b)}) < \alpha$.

## Linear Model: computation of $G_\infty(t, \widehat{F}_n)$

Application $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$, $\varepsilon \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{U}([-1,+1])$ or $\varepsilon \sim \mathcal{S}td(\nu = 2)$.



## Linear Regression

What does generate $B$ boostrap samples, under the null assumption means ?

Use residual boostrap technique:

Example : (standard) linear model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $H_0 : \beta_1 = 0$.

2.1. Estimate the model under $H_0$, i.e. $y_i = \beta_0 + \eta_i$, and save $\{\widehat{\eta}_1, \cdots, \widehat{\eta}_n\}$

2.2. Define $\widetilde{\boldsymbol{\eta}} = \{\widetilde{\eta}_1, \cdots, \widetilde{\eta}_n\}$ with $\widetilde{\eta} = \sqrt{\dfrac{n}{n-1}}\widehat{\eta}$

2.3. Draw (with replacement) residuals $\widetilde{\boldsymbol{\eta}}^{(b)} = \{\widetilde{\eta}_1^{(b)}, \cdots, \widetilde{\eta}_n^{(b)}\}$

2.4. Set $y_i^{(b)} = \widehat{\beta}_0 + \widetilde{\eta}_i^{(b)}$

2.5. Estimate the regression model $y_i^{(b)} = \beta_0^{(b)} + \beta_1^{(b)} x_i + \varepsilon_i^{(b)}$

## Going Further on Linear Regression

Recall that the OLS estimator satisfies

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i$$

while for the boostrap

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}^{(b)} - \widehat{\boldsymbol{\beta}}) = \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i^{(b)}$$

Thus, for i.id. data, the variance is

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i\right)^\top\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top\varepsilon_i^2\right]$$

## Going Further on Linear Regression

and similarly (for i.id. data)

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i^{(b)}\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i^{(b)}\right)^{\mathsf{T}}\middle|\boldsymbol{X},Y\right]=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}}\widehat{\varepsilon}_i^2$$

## Bootstrap with dynamic regression models

Example : linear model, $y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \varepsilon_t$ with $H_0 : \beta_1 = 0$.

2.1. Estimate the model under $H_0$, i.e. $y_t = \beta_0 + \beta_2 y_{t-1} + \eta_i$, and save $\{\widehat{\eta}_1, \cdots, \widehat{\eta}_n\}$ (estimated residuals from an AR(1))

2.2. Define $\widetilde{\boldsymbol{\eta}} = \{\widetilde{\eta}_1, \cdots, \widetilde{\eta}_n\}$ with $\widetilde{\eta} = \sqrt{\dfrac{n}{n-2}}\widehat{\eta}$

2.3. Draw (with replacement) residuals $\widetilde{\boldsymbol{\eta}}^{(b)} = \{\widetilde{\eta}_1^{(b)}, \cdots, \widetilde{\eta}_n^{(b)}\}$

2.4. Set (recursively) $y_t^{(b)} = \widehat{\beta}_0 + \widehat{\beta}_2 y_{t-1}^{(b)} + \widetilde{\eta}_t^{(b)}$

2.5. Estimate the regression model $y_t^{(b)} = \beta_0^{(b)} + \beta_1^{(b)} x_t + \beta_2^{(b)} y_{t-1}^{(b)} + \varepsilon_t^{(b)}$

**Remark** : start (usually) with $y_0^{(b)} = y_1$

## Bootstrap with heteroskedasticity

Example : linear model, $y_i = \beta_0 + \beta_1 x_i + |x_i| \cdot \varepsilon_t$ with $H_0 : \beta_1 = 0$.

2.1. Estimate the model under $H_0$, i.e. $y_i = \beta_0 + \eta_i$, and save $\{\widehat{\eta}_1, \cdots, \widehat{\eta}_n\}$

2.2. Compute $H_{i,i}$ with $\boldsymbol{H} = [H_{i,i}]$ from $\boldsymbol{H} = \boldsymbol{X}[\boldsymbol{X}^\top \boldsymbol{X}]^{-1}\boldsymbol{X}^\top$.

2.3.a. Define $\widetilde{\boldsymbol{\eta}} = \{\widetilde{\eta}_1, \cdots, \widetilde{\eta}_n\}$ with $\widetilde{\eta}_i = \pm \dfrac{\widehat{\eta}_i}{\sqrt{1 - H_{i,i}}}$

(here $\pm$ mean $\{-1, +1\}$ with probabilities $\{1/2, 1/2\}$)

2.4.a. Draw (with replacement) residuals $\widetilde{\boldsymbol{\eta}}^{(b)} = \{\widetilde{\eta}_1^{(b)}, \cdots, \widetilde{\eta}_n^{(b)}\}$

2.5.a. Set $y_i^{(b)} = \widehat{\beta}_0 + \widehat{\beta}_2 y_{i-1}^{(b)} + \widetilde{\eta}_i^{(b)}$

2.6.a. Estimate the regression model $y_i^{(b)} = \beta_0^{(b)} + \beta_1^{(b)} x_i + \varepsilon_i^{(b)}$

This was suggested in Liu (1988) Bootstrap procedures under some non - i.i.d. models

## Bootstrap with heteroskedasticity

Example : linear model, $y_i = \beta_0 + \beta_1 x_i + |x_i| \cdot \varepsilon_t$ with $H_0 : \beta_1 = 0$.

2.1. Estimate the model under $H_0$, i.e. $y_i = \beta_0 + \eta_i$, and save $\{\widehat{\eta}_1, \cdots, \widehat{\eta}_n\}$

2.2. Compute $H_{i,i}$ with $\boldsymbol{H} = [H_{i,i}]$ from $\boldsymbol{H} = \boldsymbol{X}[\boldsymbol{X}^\top \boldsymbol{X}]^{-1}\boldsymbol{X}^\top$.

2.3.b. Define $\widetilde{\boldsymbol{\eta}} = \{\widetilde{\eta}_1, \cdots, \widetilde{\eta}_n\}$ with $\widetilde{\eta}_i = \xi_i \dfrac{\widehat{\eta}_i}{\sqrt{1 - H_{i,i}}}$

(here $\xi_i$ takes values $\left\{\dfrac{1-\sqrt{5}}{2}, \dfrac{1+\sqrt{5}}{2}\right\}$ with probabilities $\left\{\dfrac{\sqrt{5}+1}{2\sqrt{5}}, \dfrac{\sqrt{5}-1}{2\sqrt{5}}\right\}$)

2.4.b. Draw (with replacement) residuals $\widetilde{\boldsymbol{\eta}}^{(b)} = \{\widetilde{\eta}_1^{(b)}, \cdots, \widetilde{\eta}_n^{(b)}\}$

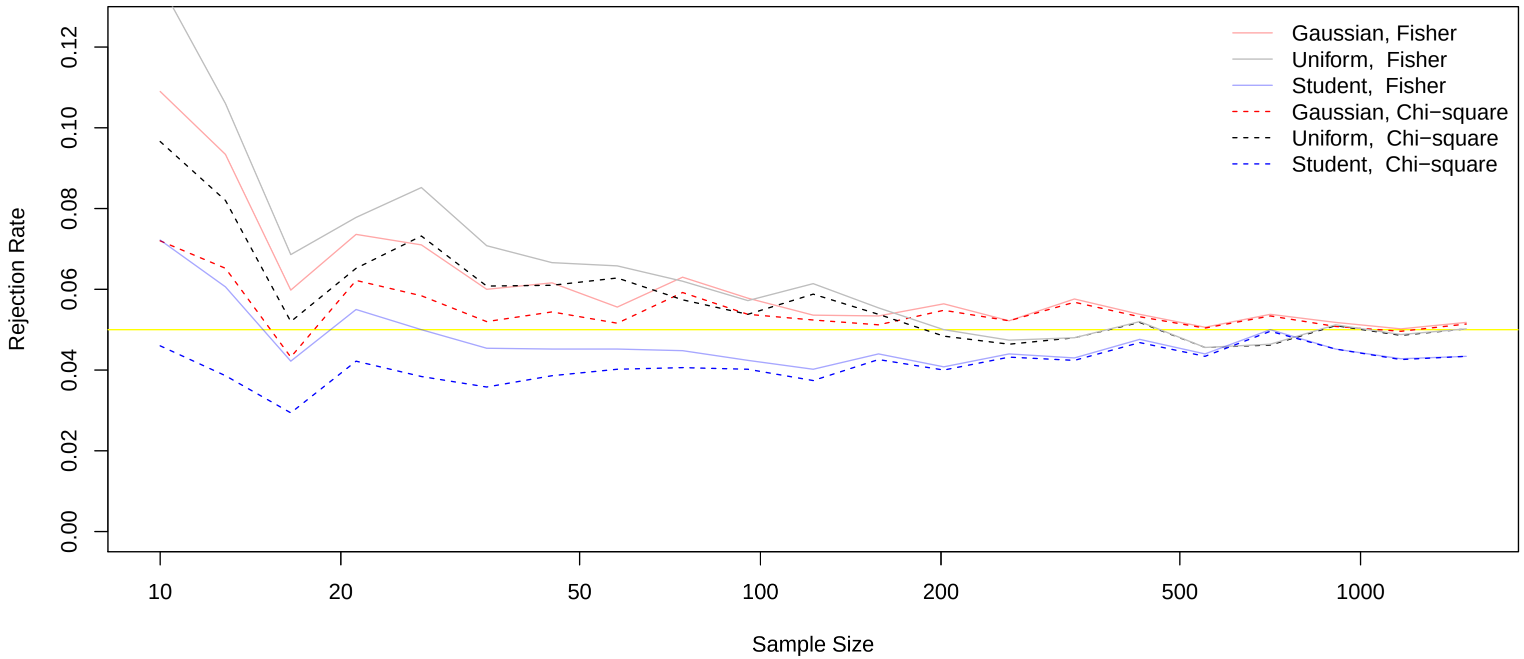
2.5.b. Set $y_i^{(b)} = \widehat{\beta}_0 + \widehat{\beta}_2 y_{i-1}^{(b)} + \widetilde{\eta}_i^{(b)}$

2.6.b. Estimate the regression model $y_i^{(b)} = \beta_0^{(b)} + \beta_1^{(b)} x_i + \varepsilon_i^{(b)}$

This was suggested in Mammen (1993) Bootstrap and wild bootstrap for high dimensional linear models, $\xi_i$'s satisfy here $\mathbb{E}[\xi_i^3] = 1$

## Bootstrap with heteroskedasticity

Application $y_i = \beta_0 + \beta_1 x_i + |x_i| \cdot \varepsilon_i$, $\varepsilon \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{U}([-1,+1])$ or $\varepsilon \sim \mathcal{S}td(\nu = 2)$.



## Bootstrap with 2SLS: Wild Bootstrap

Consider a linear model, $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ where $\boldsymbol{x}_i = \boldsymbol{z}_i^\top \boldsymbol{\gamma} + \boldsymbol{u}_i$.

Two-stage least squares:

1. regress each column of $\boldsymbol{x}$ on $\boldsymbol{z}$, $\widehat{\gamma} = [\boldsymbol{Z}^\top \boldsymbol{Z}]\boldsymbol{Z}^\top \boldsymbol{X}$ and consider the predicted value

$$\widehat{\boldsymbol{X}} = \boldsymbol{Z}\widehat{\gamma} = \underbrace{\boldsymbol{Z}[\boldsymbol{Z}^\top \boldsymbol{Z}]\boldsymbol{Z}^\top}_{\Pi_{\boldsymbol{Z}}} \boldsymbol{X}$$

2. regress $y$ on predicted covariates $\widehat{\boldsymbol{X}}$, $y_i = \widehat{\boldsymbol{x}}_i^\top \boldsymbol{\beta} + \varepsilon_i$

## Bootstrap with 2SLS: Wild Bootstrap

Example : linear model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_t$ where $x_i = \boldsymbol{z}_i^\top \boldsymbol{\gamma} + u_i$ and $\text{Cov}[\varepsilon, u] = \rho$, with $H_0 : \beta_1 = 0$.

So called Wild Boostrap, see Davidson & Mackinnon (2009) Wild bootstrap tests for IV regression

2.1. Estimate the model under $H_0$, i.e. $y_i = \beta_0 + \eta_i$, by 2SLS and save $\widehat{\boldsymbol{u}} = \{\widehat{\eta}_1, \cdots, \widehat{\eta}_n\}$

2.2. Estimate $\boldsymbol{\gamma}$ from $x_i = \boldsymbol{z}_i^\top \boldsymbol{\gamma} + \delta \widehat{\eta}_i + u_i$

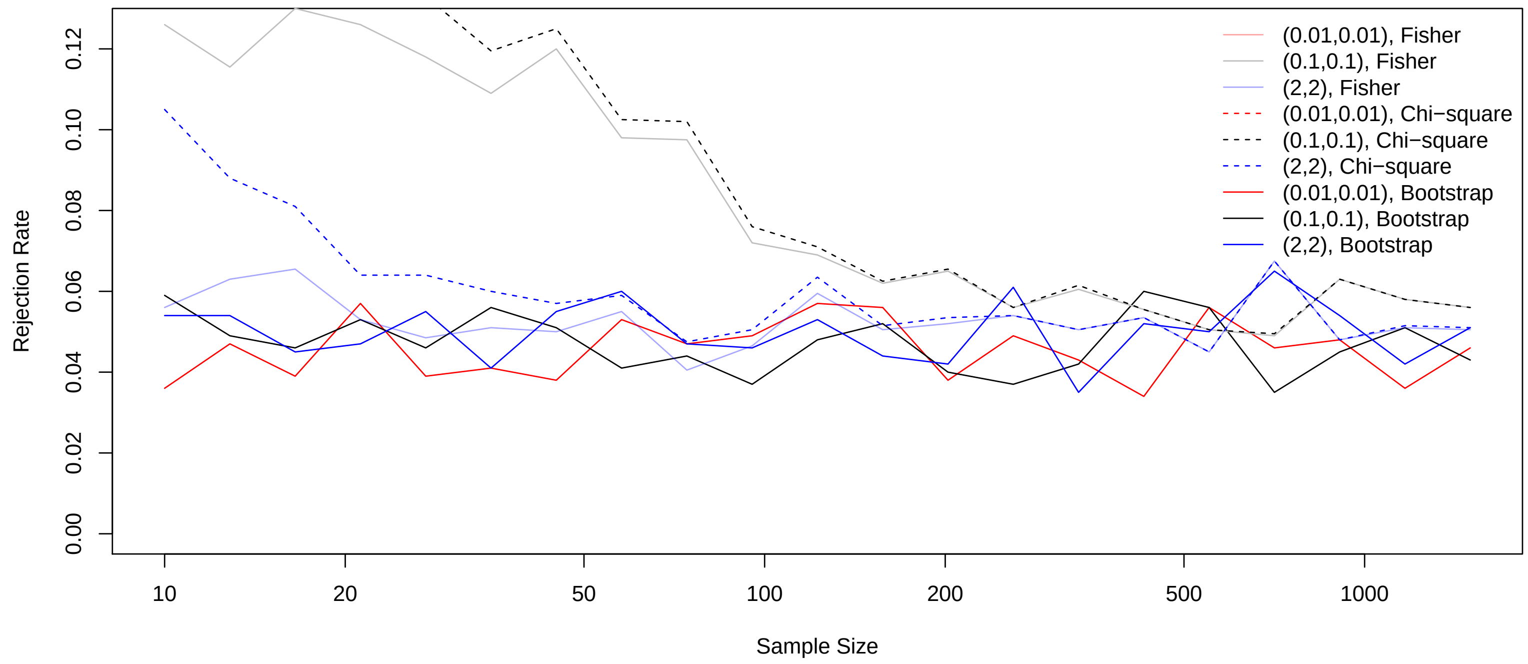
2.3. Define $\widetilde{\boldsymbol{u}} = \{\widetilde{u}_1, \cdots, \widetilde{u}_n\}$ with $\widetilde{u}_i = X_i - \boldsymbol{z}_i^\top \widehat{\boldsymbol{\gamma}}$

2.4. Draw (with replacement) pairs of residuals $(\widehat{\boldsymbol{\eta}}^{(b)}, \widetilde{\boldsymbol{u}}^{(b)})$ of $(\widehat{\eta}_i^{(b)}, \widetilde{u}_i^{(b)})$'s

2.5. Set $x_i^{(b)} = \boldsymbol{z}_i^\top \widehat{\boldsymbol{\gamma}} + \widetilde{u}_i^{(b)}$ and $y_i^{(b)} = \widehat{\beta}_0 + \widehat{\eta}_i^{(b)}$

2.6. Estimate (using 2SLS) the regression model $y_i^{(b)} = \beta_0^{(b)} + \beta_1^{(b)} x_i^{(b)} + \varepsilon_i^{(b)}$, where $x_i^{(b)} = \boldsymbol{z}_i^\top \boldsymbol{\gamma} + u_i$

# Bootstrap with 2SLS: Wild Bootstrap




See example Section 5.2 in Horowitz (1998) The Bostrap.

## Estimation of Various Quantifies of Interest

Consider a quadratic model,

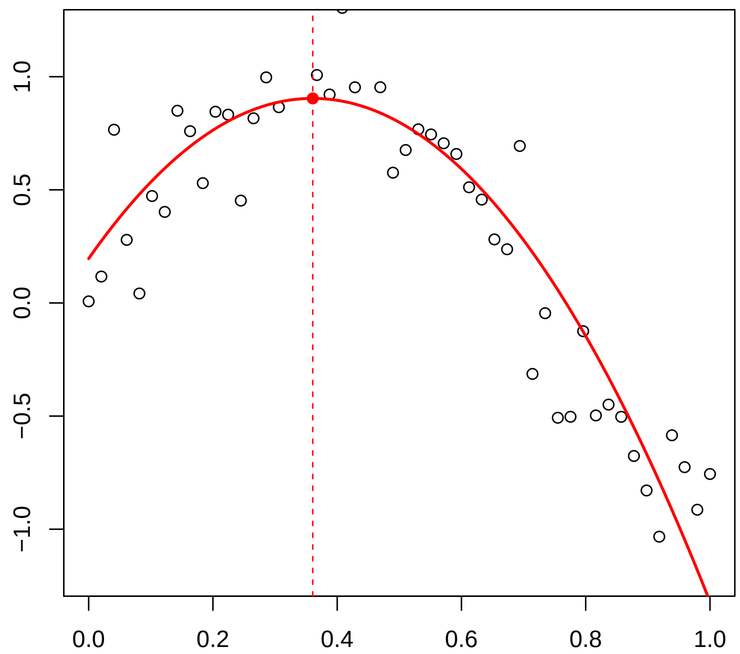
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

The minimum is obtained in $\theta = -\beta_1/2\beta_2$.

What could be the standard error for $\theta$ ?

1. Use of the Delta-Method

$$\theta = g(\beta_1, \beta_2) = \frac{-\beta_1}{2\beta_2}$$



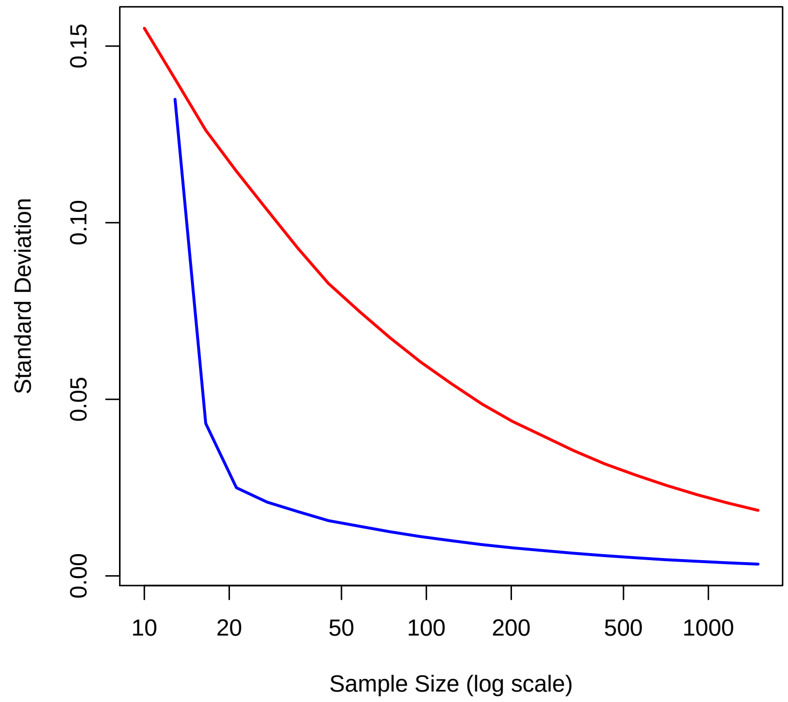

Since $\dfrac{\partial\theta}{\partial\beta_1} = \dfrac{-1}{2\beta_2}$ and $\dfrac{\partial\theta}{\partial\beta_2} = \dfrac{\beta_1}{2\beta_2^2}$, the variance is

$$\frac{1}{4}\begin{bmatrix}\dfrac{-1}{2\beta_2} & \dfrac{\beta_1}{2\beta_2^2}\end{bmatrix}\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}\begin{bmatrix}\dfrac{-1}{2\beta_2} & \dfrac{\beta_1}{2\beta_2^2}\end{bmatrix}^{\mathsf{T}} = \frac{\sigma_1^2\beta_2^2 - 2\beta_1\beta_2\sigma_{12} + \beta_1^2\sigma_2^2}{4\beta_2^2}$$

## Estimation of Various Quantifies of Interest

2. Use of Bootstrap

standard deviation of $\widehat{\theta}$,
— delta method vs.
— bootstrap.



## Box-Cox Transform

$$y_\lambda = \beta_0 + \beta_1 x + \varepsilon, \text{ with } y_\lambda = \frac{y^{\lambda - 1}}{\lambda}$$

with the limiting case $y_0 = \log[y]$.

We assume that for some (unkown) $\lambda_0$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

As in Horowitz (1998) The Bostrap, use residual bootstrap:

$$y_i^{(b)} = \big(\lambda[\widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\varepsilon}^{(b)}]\big)^{1/\lambda}$$

## Kernel based Regression

Consider some kernel based regression of estimate $m(x) = \mathbb{E}[Y|X = x]$,

$$\widehat{m}_h(x) = \frac{1}{nh\widehat{f}_n(x)} \sum_{i=1}^{n} y_i k\left(\frac{x - x_i}{h}\right) \text{ where } \widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right)$$

We have seen that the bias was

$$b_h(x) = \mathbb{E}[\tilde{m}(x)] - m(x) \propto h^2 \left(\frac{1}{2} m''(x) + m'(x) \frac{f'(x)}{f(x)}\right)$$

and the variance

$$v_h(x) \propto \frac{\text{Var}[Y|X = x]}{nhf(x)}$$

Further

$$Z_n(x) = \frac{\widehat{m}_{h_n}(x) - m(x) - b_{h_n}(x)}{\sqrt{v_h h_n(x)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \text{ as } n \to \infty.$$

## Kernel based Regression

Idea: convert $Z_n(x)$ into an asymptotically pivotal statistic

Observe that

$$\widehat{m}_h(x) - m(x) \sim \frac{1}{nhf(x)} \sum_{i=1}^{n} [y_i - m(x)] k\left(\frac{x - x_i}{h}\right)$$

so that $v_n(x)$ can be estimated by

$$\widehat{v}_n(x) = \frac{1}{(nh\widehat{f}_n(x))^2} \sum_{i=1}^{n} [y_i - \widehat{m}_h(x)]^2 k\left(\frac{x - x_i}{h}\right)^2$$

then set

$$\widehat{\theta} = \frac{\widehat{m}_h(x) - m(x)}{\sqrt{\widehat{v}_n(x)}}$$

$\widehat{\theta}$ is asymptotically $\mathcal{N}(0,1)$ and it is an asymptotically pivotal statistic

## Poisson Regression

| Diagnosis period | | Reporting-delay interval (quarters): | | | | | | | | | Total reports to end |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | Quarter | $0^\dagger$ | 1 | 2 | 3 | 4 | 5 | 6 | ··· | $\geq 14$ | of 1992 |
| 1988 | 1 | 31 | 80 | 16 | 9 | 3 | 2 | 8 | ··· | 6 | 174 |
| | 2 | 26 | 99 | 27 | 9 | 8 | 11 | 3 | ··· | 3 | 211 |
| | 3 | 31 | 95 | 35 | 13 | 18 | 4 | 6 | ··· | 3 | 224 |
| | 4 | 36 | 77 | 20 | 26 | 11 | 3 | 8 | ··· | 2 | 205 |
| 1989 | 1 | 32 | 92 | 32 | 10 | 12 | 19 | 12 | ··· | 2 | 224 |
| | 2 | 15 | 92 | 14 | 27 | 22 | 21 | 12 | ··· | 1 | 219 |
| | 3 | 34 | 104 | 29 | 31 | 18 | 8 | 6 | ··· | | 253 |
| | 4 | 38 | 101 | 34 | 18 | 9 | 15 | 6 | ··· | | 233 |
| 1990 | 1 | 31 | 124 | 47 | 24 | 11 | 15 | 8 | ··· | | 281 |
| | 2 | 32 | 132 | 36 | 10 | 9 | 7 | 6 | ··· | | 245 |
| | 3 | 49 | 107 | 51 | 17 | 15 | 8 | 9 | ··· | | 260 |
| | 4 | 44 | 153 | 41 | 16 | 11 | 6 | 5 | ··· | | 285 |
| 1991 | 1 | 41 | 137 | 29 | 33 | 7 | 11 | 6 | ··· | | 271 |
| | 2 | 56 | 124 | 39 | 14 | 12 | 7 | 10 | | | 263 |
| | 3 | 53 | 175 | 35 | 17 | 13 | 11 | | | | 306 |
| | 4 | 63 | 135 | 24 | 23 | 12 | | | | | 258 |
| 1992 | 1 | 71 | 161 | 48 | 25 | | | | | | 310 |
| | 2 | 95 | 178 | 39 | | | | | | | 318 |
| | 3 | 76 | 181 | | | | | | | | 273 |
| | 4 | 67 | | | | | | | | | 133 |

Example : see Davison & Hinkley (1997) Bootstrap Methods and Applications, UK AIDS diagnoses, 1988-1992.

Reporting delay can be important

Let $j$ denote year and $k$ denote delay. Assumption

$$N_{j,k} \sim \mathcal{P}(\lambda_{j,k}) \text{ with } \lambda_{j,k} = \exp[\alpha_j + \beta_k]$$

Unreported diagnoses for period $j$ : $\displaystyle\sum_{k \text{ unobserved}} \lambda_{j,k}$

Prediction : $\displaystyle\sum_{k \text{ unobserved}} \widehat{\lambda}_{j,k} = \exp[\widehat{\alpha}_j] \sum_{k \text{ unobserved}} \exp[\widehat{\beta}_k]$

Poisson regression is a GLM : confidence intervals on coefficients are asymptotic.

Let $V$ denote the variance function, then Pearson residuals are

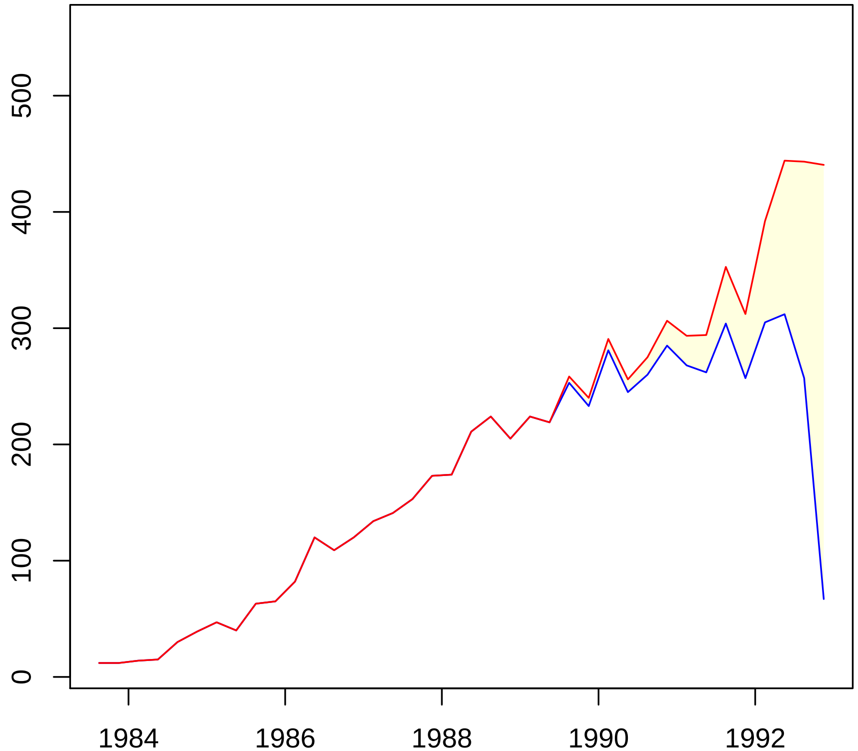
$$\widehat{\epsilon}_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{V[\widehat{\mu}_i]}}$$

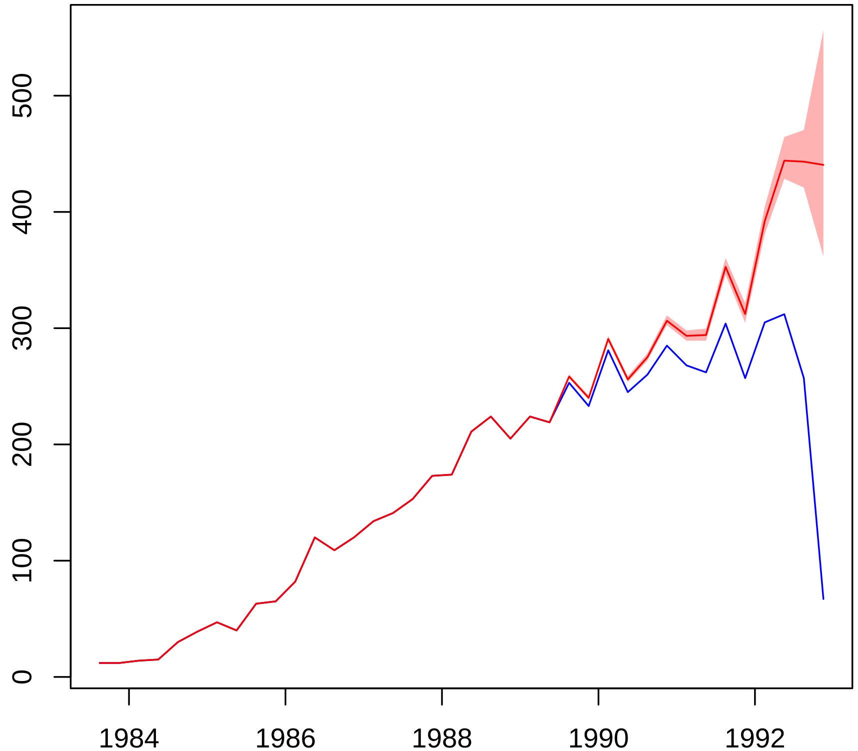
so here

$$\widehat{\epsilon}_{j,k} = \frac{n_{j,k} - \widehat{\lambda}_{j,k}}{\sqrt{\widehat{\lambda}_{j,k}}}$$

## Poisson Regression

So bootstrapped responses are $n^\star_{j,k} = \widehat{\lambda}_{j,k} + \sqrt{\widehat{\lambda}_{j,k}} \cdot \widehat{\epsilon}^\star_{j,k}$



## Pivotal Case (or not)

In some cases, $G(\cdot, F)$ does not depend on $F$, $\forall F \in \mathcal{F}$.

Then $T_n$ is said to be pivotal, relative to $\mathcal{F}$.

Example : consider the case of Gaussian residuals, $\mathcal{F} = \mathcal{F}_{\text{gaussian}}$. Then

$$T = \frac{\overline{y} - \mathbb{E}[Y]}{\widehat{\sigma}} \sim \mathcal{S}td(n-1)$$

which does not depend on $F$ (but it does depend on $\mathcal{F}$)

If $T_n$ is not pivotal, it is still possible to look for bounds on $G_n(t, F)$,

$$B_n(t) = \left[\inf_{F \in \mathcal{F}_\star} \{G_n(t, F)\}; \sup_{F \in \mathcal{F}_\star} \{G_n(t, F)\}\right]$$

for instance, when a set of *reasonable values* for $\mathcal{F}_\star$ is provided, by an expert.

## Pivotal Case (or not)

$$B_n(t) = \left[ \inf_{F \in \mathcal{F}_\star} \{G_n(t,F)\}; \sup_{F \in \mathcal{F}_\star} \{G_n(t,F)\} \right]$$

In the parametric case, set

$$\mathcal{F}_\star = \{F_\theta, \theta \in IC\}$$

where $IC$ is some confidence interval.

In the nonparametric case, use Kolomogorov-Smirnov statistics to get bounds, using quantiles of

$$\sqrt{n} \sup\{|\widehat{F}_n(t) - F_0(t)|\}$$

## Pivotal Function and Studentized Statistics

It is interesting to studentize any statistics.

Let $v$ denote the variance of $\widehat{\theta}$ (computed using $\{y_1,\cdots,y_n\}$. Then set

$$Z=\frac{\widehat{\theta}-\theta}{\sqrt{v}}$$

If quantiles of $Z$ are known (and denoted $z_\alpha$), then

$$\mathbb{P}\Big(\widehat{\theta}+\sqrt{v}z_{\alpha/2}\leq\theta\leq\widehat{\theta}+\sqrt{v}z_{1-\alpha/2}\Big)=1-\alpha$$

Idea : use a (double) boostrap procedure

## Pivotal Function and Double Bootstrap Procedure

1. Generate a bootstrap sample $\boldsymbol{y}^{(b)} = \{y_1^{(b)}, \cdots, y_n^{(b)}\}$
2. Compute $\widehat{\theta}^{(b)}$
3. From $\boldsymbol{y}^{(b)}$ generate $\beta$ bootstrap sample, and compute $\{\widehat{\theta}_1^{(b)}, \cdots, \widehat{\theta}_\beta^{(b)}\}$
4. Compute $\widehat{v}^{(b)} = \dfrac{1}{\beta}\displaystyle\sum_{j=1}^{\beta}\big(\widehat{\theta}_j^{(b)} - \overline{\theta}^{(b)}\big)^2$
5. Set $z^{(b)} = \dfrac{\widehat{\theta}^{(b)} - \widehat{\theta}}{\sqrt{\widehat{v}^{(b)}}}$

Then use $\{z^{(1)}, \cdots, z^{(B)}\}$ to estimate the distribution of $z$'s (and some quantiles).

$$\mathbb{P}\left(\widehat{\theta} + \sqrt{v} z_{\alpha/2}^{(B)} \leq \theta \leq \widehat{\theta} + \sqrt{v} z_{1-\alpha/2}^{(B)}\right) = 1 - \alpha$$

## Why should we studentize ?

Here $Z \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ as $n \to \infty$ (CLT). Using Edgeworth series,

$$\mathbb{P}[Z \leq z|F] = \Phi(z) + n^{-1/2}p(z)\varphi(z) + O(n^{-1})$$

for some quadratic polynomial $p(\cdot)$. For $Z^{(b)}$

$$\mathbb{P}[Z^{(b)} \leq z|\widehat{F}] = \Phi(z) + n^{-1/2}\widehat{p}(z)\varphi(z) + O(n^{-1})$$

where $\widehat{p}(z) = p(z) + O(n^{-1/2})$, so

$$\mathbb{P}[Z \leq z|F] - \mathbb{P}[Z^{(b)} \leq z|\widehat{F}] = O(n^{-1})$$

But if we do not studentize, $Z = (\widehat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\nu)$ as $n \to \infty$ (CLT). Using Edgeworth series,

$$\mathbb{P}[Z \leq z|F] = \Phi\left(\frac{z}{\sqrt{\nu}}\right) + n^{-1/2}p'\left(\frac{z}{\sqrt{\nu}}\right)\varphi\left(\frac{z}{\sqrt{\nu}}\right) + O(n^{-1})$$

for some quadratic polynomial $p(\cdot)$. For $Z^{(b)}$

$$\mathbb{P}[Z^{(b)} \leq z|\widehat{F}] = \Phi\left(\frac{z}{\sqrt{\widehat{\nu}}}\right) + n^{-1/2}\widehat{p}'\left(\frac{z}{\sqrt{\widehat{\nu}}}\right)\varphi\left(\frac{z}{\sqrt{\widehat{\nu}}}\right) + O(n^{-1})$$

recall that $\widehat{\nu} = \nu + 0(n^{-1/2})$, and thus

$$\mathbb{P}[Z \leq z|F] - \mathbb{P}[Z^{(b)} \leq z|\widehat{F}] = O(n^{-1/2})$$

Hence, studentization reduces error, from $O(n^{-1/2})$ to $O(n^{-1})$

## Variance estimation

The estimation of $\text{Var}[\widehat{\theta}]$ is necessary for studentized boostrap.

- double bootstrap (used here)
- delta method
- jackknife (leave-one-out)

### Double Bootstrap

Requieres $B \times \beta$ resamples, e.g. $B \sim 1,000$ while $\beta \sim 100$

### Delta Method

Let $\widehat{\tau} = g(\widehat{\theta})$, with $g'(\theta) \neq 0$.

$$\mathbb{E}[\widehat{\tau}] = g(\theta) + O(n^{-1})$$

$$\text{Var}[\widehat{\tau}] = \text{Var}[\widehat{\theta}]g'(\theta)^2 + O(n^{-3/2})$$

## Variance estimation

Idea: find a transformation such that $\text{Var}[\widehat{\tau}]$ is constant. Then

$$\text{Var}[\widehat{\theta}] \sim \frac{\text{Var}[\widehat{\tau}]}{g'(\widehat{\theta})^2}$$

There is also a nonparametric delta method, based on the influence function.

## Influence Function and Taylor Expansion

Taylor expansion

$$t(y) = t(x) + \int_x^y f'(z)cdz\ t(x) + (y-x)f'(x)$$

$$t(G) = t(F) + \int_{\mathbb{R}} L_t(z,F)dG(z)$$

where $L_t$ is the Fréchet derivative,

$$L_t(z,F) = \left.\frac{\partial[(1-\epsilon)F + \epsilon\Delta_z]}{\partial\epsilon}\right|_{\epsilon=0}$$

where $\Delta_z(t) = \mathbf{1}(t > z)$ denote the cdf of the Dirac measure in $z$.

For instance, observe that

$$t(\widehat{F}_n) = t(F) + \frac{1}{n}\sum_{i=1}^n L_t(y_i, F)$$

## Influence Function and Taylor Expansion

This can be used to estimate the variance. Set

$$V_L = \frac{1}{n^2}\sum_{i=1}^n L(y_i, F)^2$$

where $L(y,F)$ is the influence function for $\theta = t(F)$ for observation at $y$ when distribution is $F$.

The empirical version is $\ell_i = L(y_i, \widehat{F})$ and set

$$\widehat{V}_L = \frac{1}{n^2}\sum_{i=1}^n \ell_i^2$$

Example : let $\theta = \mathbb{E}[X]$ with $X \sim F$, then

$$\widehat{\theta} = \overline{y}_n = \sum_{i=1}^n \frac{1}{n} y_i = \sum_{i=1}^n \omega_i y_i \text{ where } \omega_i = \frac{1}{n}$$

## Influence Function and Taylor Expansion

Change $\omega$'s in direction $j$ :

$$\omega_j = \epsilon + \frac{1-\epsilon}{n}, \text{ while } \forall i \neq j, \omega_i = \frac{1-\epsilon}{n},$$

then $\widehat{\theta}$ changes in

$$\underbrace{[y_h - \widehat{\theta}]}_{\ell_j}\epsilon + \widehat{\theta}$$

Hence, $\ell_j$ is the standardized chance in $\widehat{\theta}$ with an increase in direction $j$, and

$$\widehat{V}_L = \frac{n-1}{n}\frac{\text{Var}[X]}{n}.$$

Example : consider a ratio, $\theta = \dfrac{\mathbb{E}[X]}{\mathbb{E}[Y]}$, then

$$\widehat{\theta} = \frac{\overline{x}_n}{\overline{y}_n} \text{ and } \ell_j = \frac{x_j - \widehat{\theta}y_j}{\overline{y}_n}$$

## Influence Function and Taylor Expansion

so that

$$\widehat{V}_L = \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{x_j - \widehat{\theta} y_j}{\overline{y}_n}\right)^2$$

Example : consider a correlation coefficient,

$$\theta = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\cdot\mathbb{E}[Y]}{\sqrt{\left(\mathbb{E}[X^2] - \mathbb{E}[X]^2\right)\cdot\left(\mathbb{E}[Y^2] - \mathbb{E}[Y]^2\right)}}$$

Let $\overline{xy} = n^{-1}\sum x_i y_i$, so that

$$\widehat{\theta} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sqrt{\left(\overline{x^2} - \overline{x}^2\right)\cdot\left(\overline{y^2} - \overline{y}^2\right)}}$$

## Jackknife

An approximation of $\ell_i$ is $\ell_i^\star = (n-1)\big(\widehat{\theta} - \widehat{\theta}_{(-j)}\big)$ where $\widehat{\theta}_{(-j)}$ is the statistics computed from sample $\{y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n\}$

One can define Jackknife bias and Jackknife variance

$$b^\star = \frac{-1}{n}\sum_{i=1}^n \ell_i^\star \text{ and } v^\star = \frac{1}{n(n-1)}\left(\sum_{i=1}^n \ell_i^{\star 2} - nb^{\star 2}\right)$$

cf numerical differentiation when $\epsilon = -\dfrac{1}{(n-1)}$.

## Convergence

Given a sample $\{y_1, \cdots, y_n\}$, i.id. with distribution $F$, set

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(y_i \leq t)$$

Then

$$\sup\left\{|\widehat{F}_n(t) - F_0(t)|\right\} \xrightarrow{\mathbb{P}} 0, \text{ as } n \to \infty.$$

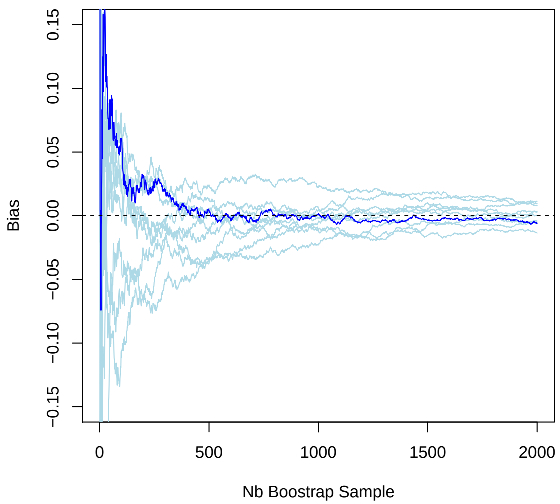
## How many Boostrap Samples?

Easy to take $B \geq 5000$

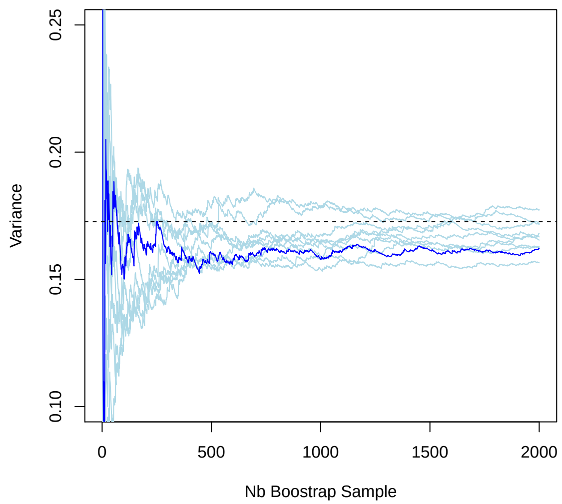
$R > 100$ to estimate bias or variance

$R > 1000$ to estimate quantiles

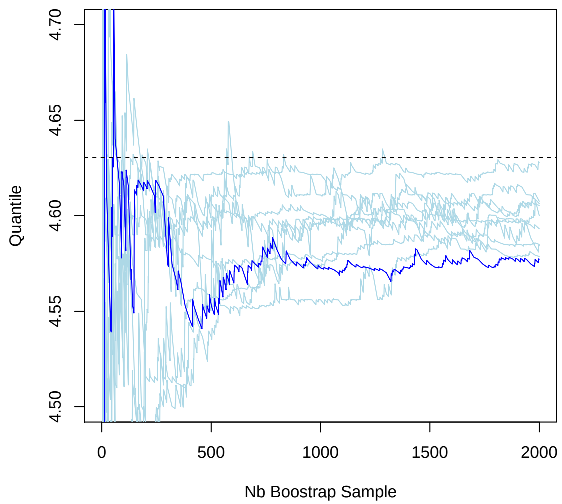
Bias




Variance




Quantile



## Consistency

We expect something like

$$G_n(t,\widehat{F}_n) \sim G_\infty(t,\widehat{F}_n) \sim G_\infty(t,F_0) \sim G_n(t,F_0)$$

$G_n(t,\widehat{F}_n)$ is said to be consistent if under each $F_0 \in \mathcal{F}$,

$$\sup_t \in \mathbb{R}\big\{|G_n(t,\widehat{F}_n) - G_\infty(t,F_0)|\big\} \overset{\mathbb{P}}{\to} 0$$

Example: let $\theta = \mathbb{E}_{F_0}(X)$ and consider $T_n = \sqrt{n}(\overline{X} - \theta)$. Here

$$G_n(t,F_0) = \mathbb{P}_{F_0}(T_n \leq t)$$

Based on boostrap samples, a bootstrap version of $T_n$ is

$$T_n^{(b)} = \sqrt{n}\big(\overline{X}^{(b)} - \overline{X}\big) \text{ since } \overline{X} = \mathbb{E}_{\widehat{F}_n}(X)$$

and $G_n(t,\widehat{F}_n) = \mathbb{P}_{\widehat{F}_n}(T_n \leq t)$

## Consistency

Consider a regression model $y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i$

The natural assumption is $\mathbb{E}[\varepsilon_i|\boldsymbol{X}] = 0$ with $\varepsilon_i$'s i.id.$\sim F$.

The parameter of interest is $\theta = \beta_j$, and let $\widehat{\beta}_j = \theta(\widehat{F}_n)$.

1. The statistics of interest is $T_n = \sqrt{n}\big[\widehat{\beta}_j - \beta_j\big]$.

We want to know $G_n(t, F_0) = \mathbb{P}_{F_0}(T_n \leq t)$.

Let $\boldsymbol{x}^{(b)}$ denote a bootsrap sample.

Compute $T_n^{(b)} = \sqrt{n}\big(\widehat{\beta}_j^{(b)} - \widehat{\beta}_j\big)$, and then

$$G_n(t, F_n) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}(T_n^{(b)} \leq t)$$

## Consistency

2. The statistics of interest is $T_n = \sqrt{n}\dfrac{[\widehat{\beta}_j - \beta_j]}{\sqrt{\text{Var}[\widehat{\beta}_j]}}$.

We want to know $G_n(t, F_0) = \mathbb{P}_{F_0}(T_n \leq t)$.

Let $\boldsymbol{x}^{(b)}$ denote a bootsrap sample.

Compute $T_n^{(b)} = \sqrt{n}\dfrac{[\widehat{\beta}_j^{(b)} - \widehat{\beta}_j]}{\sqrt{\text{Var}^{(b)}[\widehat{\beta}_j]}}$, and then

$$G_n(t, F_n) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}(T_n^{(b)} \leq t)$$

This second option is more accurate than the first one :

## Consistency

The approximation error of bootstrap applied to asymptotically pivotal statistic is smaller than the approximation error of bootstrap applied on asymptotically non-pivotal statistic, see Horowitz (1998) The Bostrap.

Here, asymptotically pivotal means that

$$G_\infty(t,F)=G_\infty(t),\ \forall F\in\mathcal{F}.$$

Assume now that the quantity of interest is $\theta=\text{Var}[\widehat{\boldsymbol{\beta}}]$.

Consider a bootstrap procedure, then one can prove that

$$\begin{aligned}&\underset{B,n\to\infty}{\text{plim}}\left\{\frac{1}{B}\sum_{b=1}^{B}\sqrt{n}\big(\widehat{\boldsymbol{\beta}}^{(b)}-\widehat{\boldsymbol{\beta}}\big)\sqrt{n}\big(\widehat{\boldsymbol{\beta}}^{(b)}-\widehat{\boldsymbol{\beta}}\big)^{\mathsf{T}}\right\}\\ &=\underset{n\to\infty}{\text{plim}}\left\{n\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\big)\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\big)^{\mathsf{T}}\right\}\end{aligned}$$

## More on Testing Procedures

Consider a sample $\{y_1, \cdots, y_n\}$. We want to test some hypothesis $H_0$. Consider some test statistic $t(\boldsymbol{y})$

Idea: $t$ takes large values when $H_0$ is not satisfied.

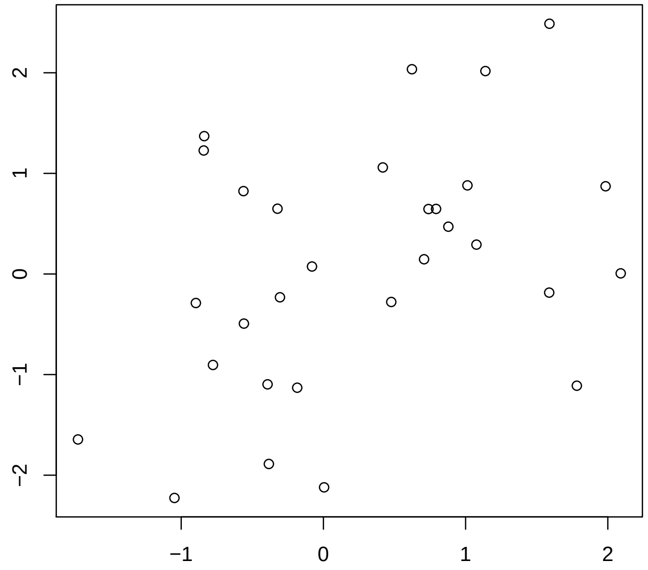
The $p$-value is $p = \mathbb{P}[T > t_{\text{obs}}|H_0]$.

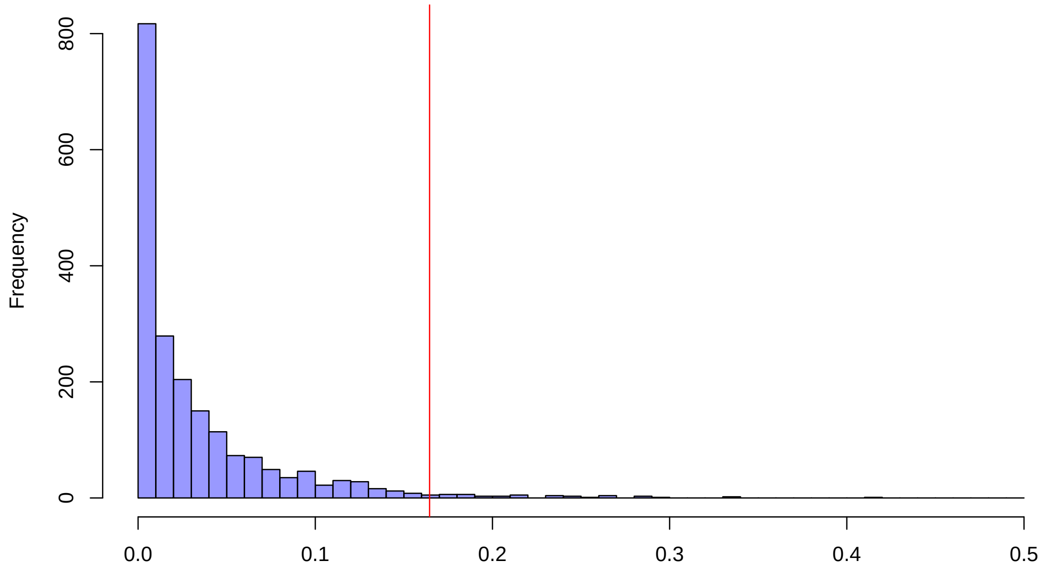
Boostrap/simulations can be used to estimate $p$, by simulation from $H_0$.

1. Generate $\boldsymbol{y}^{(s)} = \{y_1^{(s)}, \cdots, y_n^{(s)}\}$ generated from $H_0$.
2. Compute $t^{(s)} = t(\boldsymbol{y}^{(s)})$
3. Set $\widehat{p} = \dfrac{1}{1+S}\left(1 + \sum_{s=1}^{S} \mathbf{1}(t^{(s)} \geq t_{\text{obs}})\right)$

Example : testing independence, let $t$ denote the square of the correlation coefficient.

Under $H_0$ variables are independent, so we can bootstrap independently $x$'s and $y$'s.




With this bootstap procedure, we estimate

$$\widehat{p} = \mathbb{P}\big(T \geq t_{\text{obs}} | \widehat{H}_0\big)$$

which is not the same as

$$p = \mathbb{P}\big(T \geq t_{\text{obs}} | H_0\big)$$

## More on Testing Procedures

In a parametric model, it can be interesting to use a sufficient statistic $W$. One can prove that

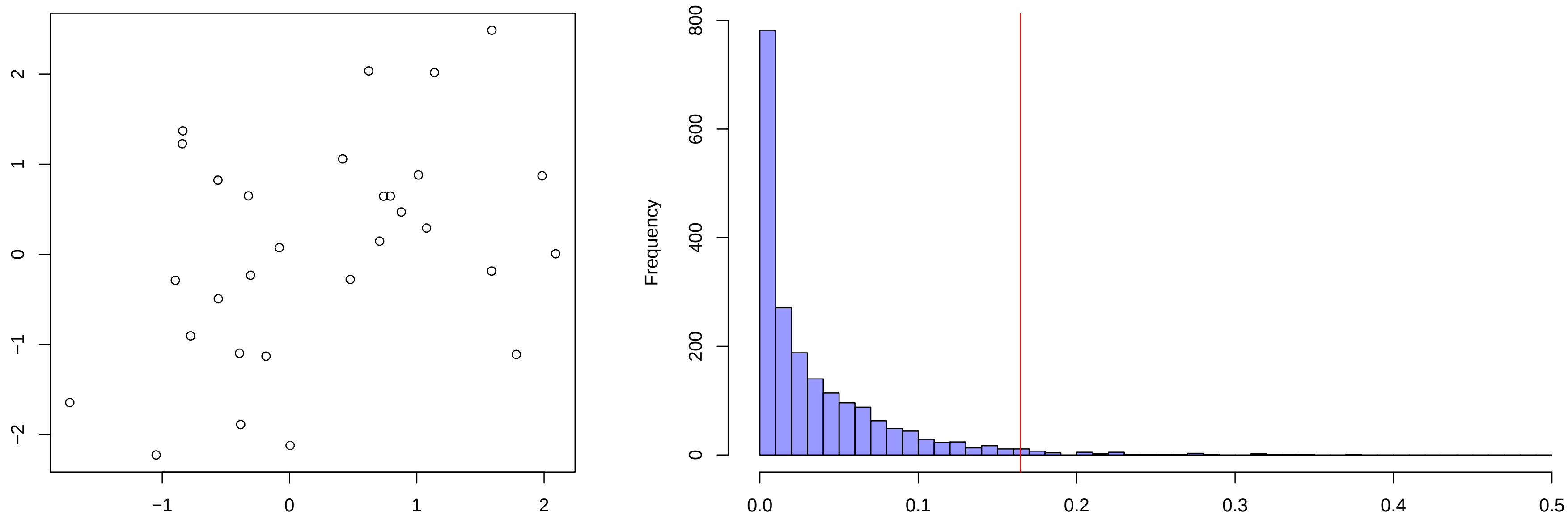
$$p = \mathbb{P}\big(T \geq t_{\text{obs}}|\widehat{H}_0, W\big)$$

The problem is to generate from this conditional distribution...

Example : for the independence test, we should sample from $\widehat{F}_x$ and $\widehat{F}_y$ with fixed margins.

Bootstrap should be here without replacement.

## More on Testing Procedures



## More on Testing Procedures

But this nonparametric bootstrap fails when

Gaussian Central Limit Theorem does not apply

Mammen's theorem

Example $X \sim$ Cauchy : limit distribution $G_\infty t, F$ is not continuous, in $F$

Example : distribution of the maximum of the support (see Bickel and Freedman (1981) ): $X \sim \mathcal{U}([0,\theta_0])$

$T_n = n(\theta_n - \theta_0)$ with $\theta_n = \max\{X_1, \cdots, X_n\}$

Set $T_n^{(b)} = n(\theta_n^{(b)} - \theta_n)$, and $\theta_n^{(b)} = \max\{X_1^{(b)}, \cdots, X_n^{(b)}\}$

Here $T_n \xrightarrow{\mathcal{L}} \mathcal{E}(1)$, exponential distribution, but not $T_n^{(b)}$, since $T_n^{(b)} \geq 0$ (we juste resample), and

$$\mathbb{P}[T_n^{(b)} = 0] = 1 - \mathbb{P}[T_n^{(b)} > 0] = 1 - \left(1 - \frac{1}{n}\right)^n \sim 1 - e^{-1}.$$

## Resampling or Subsampling ?

Why not draw subsamples of size $m < n$?

- with replacement, see $m$ out of $n$ boostrap
- without replacement, see subsampling boostrap

Less accurate than bootstrap when bootstrap works... but might work when bootstrap does not work

Exemple : maximum of the support, $Y_i \sim \mathcal{U}([0,\theta])$,

$$\mathbb{P}_{\widehat{F}_n}[T_m^{(b)} = 0] = 1 - \left(1 - \frac{1}{n}\right)^m \sim 1 - e^{-m/n} \sim 0$$

if $m = o(n)$.

## From Bootstrap to Bagging

Bagging was introduced in Breiman (1996) Bagging predictors

1. sample a boostrap sample $(y_i^{(b)}, \boldsymbol{x}_i^{(b)})$ by resampling pairs
2. estimate a model $\widehat{m}^{(b)}(\cdot)$

The bagged estimate for $m$ is then $m_{\text{bag}}(\boldsymbol{x}) = \dfrac{1}{B}\sum_{b=1}^{B}\widehat{m}^{(b)}(\boldsymbol{x})$

## From Bagging to Random Forests