This is to certify that the

dissertation entitled

Resampling Methods For Linear Models

presented by

Krzysztof Podgorski

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Statistics

Major professor

Date July 23, 1993

Resampling methods for linear models

By

Krzysztof Podgórski

A DISSERTATION

Submitted to
Michigan State University
in partial fulfilment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics and Probability

1993

Abstract

Resampling methods for linear models

By

Krzysztof Podgórski

In this dissertation we study resampling residuals by random permutation in linear regression with exchangeable errors. For given observations we consider the resampling distribution obtained by this method versus the distribution of errors *conditional on its order statistics.* We observe that *with high probability* both are approximately equal on *contrast vectors* after suitable normalization. The results remain valid if we consider the reduced regression model. This model comes from the original one by removing those data for which the corresponding errors take extreme values. We obtain general conditions under which confidence regions produced by such methods are accurate. No moment/distribution assumptions on the underlying errors other than exchangeability are required. We adopt the term *seamless resampling* to indicate this robustness. In the absence of finite second moments randomness of the limiting bootstrap distributions has been considered as a failure of the traditional bootstrap and no practical meaning has been given to the phenomena. For our method this type of randomness is still present in the limit but it has clear probabilistic interpretation as a conditional distribution which can be used to e.g. obtain confidence sets. We study this phenomena in detail for the case of independent errors with distribution from the domain of attraction of stable laws. We have developed computer software in which we implement these methods and provide convenient tools for the analysis of resampled data in general.

Acknowledgment

I wish to express my gratitude to my advisor Professor Raoul LePage for the introduction to the problem, fruitful discussions and constant encouragement. His exceptional support and excellent guidance during my doctoral studies at Michigan State University went far beyond what I could ever expect.

Contents

List of Figures

Part 1

Theoretical Results

1.1 Introduction

This paper considers a multiple regression model

$$Y = \mathbf{X}\beta + \epsilon, \tag{1}$$

where $\mathbf{X}$ is an $n \times d$ finite matrix of real numbers, β is a $d \times 1$ vector of real numbers, and ϵ is an $n \times 1$ vector of exchangeable random variables.

For reasons that will be discussed later we will also consider reduced regression models which come from (1) by deleting some data and the corresponding rows of the design matrix. To describe this submodel, let R_i be the rank of ϵ_i among $(\epsilon_i)_{i=1}^n$ and M be a subset of m integers from $\{1, \ldots, n\}$ with cardinality m. For any $n \times k$ matrix $\mathbf{A}$ by $\mathbf{A}_M$ we will denote an $(n-m) \times k$ matrix obtained from $\mathbf{A}$ after deleting those rows indexed by i whose R_i is in M. With this notation for any M we define a reduced regression model by the equality

$$Y_M = \mathbf{X}_M\beta + \epsilon_M. \tag{2}$$

Notice that for different vectors of errors ϵ with a fixed set M we can have different reduced models (2) and thus the design matrix $\mathbf{X}_M$ becomes random.

A $k \times 1$ vector v of real numbers is called a *contrast* if the entries of v sum to zero, equivalently if the Euclidean scalar product $v \cdot \mathbf{1} = 0$ where $\mathbf{1}$ denotes the vector of 1's in $\mathbb{R}^k$ (we will not indicate in notation dependence of $\mathbf{1}$ on a dimension k). Throughout the paper we will assume that $\mathbf{X}_M$ has rank d and also that the vector $\mathbf{1}$ is in the column space of $\mathbf{X}$ and thus of any $\mathbf{X}_M$

for a subset M of $\{1,\ldots,n\}$. Without losing generality we can assume that $\mathbf{1}$ is the first column of $\mathbf{X}$. Then contrast vectors come to (2) in a natural way. Namely, for $i > 1$ the least squares estimators $\hat{\beta}_i$ of β_i are of the form $\beta_i = V_i \cdot Y_M$ for a contrast V_i depending also on M. Indeed, V_i is the i-th row of the matrix $\mathbf{M} = (\mathbf{X}_M^T \mathbf{X}_M)^{-1} \mathbf{X}_M^T$ and $\mathbf{M}\mathbf{X}_M = \mathbf{I}$. This implies that any row V_i for $i > 1$ is orthogonal to the first column which is assumed to be equal to $\mathbf{1}$.

Ordinarily one is interested in approximating the joint distribution of $v \cdot \epsilon$ for several v since it is equal to the estimation error, i.e. $v \cdot \epsilon = (v \cdot Y) - (v \cdot \mathbf{X}\beta)$. For i.i.d. errors with finite variance, the central limit theorem is popularly used for this purpose. Bootstrap methods apply to this case as well (see: Efron (1979), Freedman (1981), Wu (1986)) but how are we to know if the underlying assumptions have practical application to the data at hand? See LePage (1990).

Instead approximating or estimating the unconditional distribution of $v \cdot \epsilon$, we propose, without moment assumptions, to estimate the sampling distribution of $v \cdot \epsilon$ *conditional* on the sigma field $\mathcal{F}$ generated by the order statistics of the coordinates of ϵ. For this purpose we will use randomly permuted residuals. To be more precise let us introduce the following notation. We will use $v \cdot \epsilon | \mathcal{F}$ to *denote the above conditional distribution.* Denote by $\epsilon^{\perp}$ the vector of residuals $Y - \hat{Y}$, where $\hat{Y} = Y/\mathbf{X}$ is the projection of Y to the column space of $\mathbf{X}$. Let π denote a uniformly distributed random permutation applied to the coordinates of n-space. Suppose that the distribution of π, conditional on ϵ, is also uniform over all $n!$ permutations, so that π is independent of ϵ. We will observe that, provided d/n is small, $v \cdot \pi\epsilon^{\perp} | Y$

provides *with high probability* a close approximation of $v \cdot \pi\epsilon|\mathcal{F}$. Notice that the latter distribution is equal to $v \cdot \epsilon|\mathcal{F}$ since ϵ has exchangeable coordinates.

Proposal 1 *For any finite set of contrast vectors* $\{v_k, k \leq l\}$, *estimate the joint* $\mathcal{F}$*–conditional sampling distributions of contrasts* $v_k \cdot \epsilon$ *by the distributions of* $v_k \cdot \pi\epsilon^{\perp}$ *conditional on* Y.

This proposal provides strong support for the approach taken by Freedman and Lane (1983) who develop *descriptive* tests of linear hypotheses. To quote from them, "our reference sets are derived by permuting residuals, and our significance level is a descriptive statistic rather than a probability." It will be seen below that $v \cdot \pi\epsilon^{\perp}|Y$ is an estimate of a *conditional* distribution insensitive to the moment assumptions. Thus it is possible to achieve robust estimation of the *conditional* sampling distributions of least squares estimators. Previously only M-estimators with bounded score functions were known to have this property, see Lahiri (1992).

As we will see Proposal 1 can be applied to a quite general class of exchangeable distributions. However sometimes the recovered distribution $v \cdot \epsilon|\mathcal{F}$ itself will posses certain unpleasant properties when some coordinates of ϵ will take relatively very large values. This may happen for example for long tailed distributions of ϵ. By our result also the distribution $v \cdot \pi\epsilon^{\perp}|Y$ will inherit this. It can happen, e.g. the example in Section 1.3, that confidence regions will not be in the form of single intervals but sums of widely separated intervals. In such a case to "improve" the distribution of interest it would be reasonable to exclude observations with outsized disturbance by errors and then apply our method of resampling by permutation to this reduced model.

Our results extends also to this situation.

Let us describe this in a more rigorous way. We shall fix a set M of those values of ranks $(R_i)_{i=1}^n$ for which we have decided to exclude corresponding observations. For example by taking $M = \{1, n\}$ we will exclude two observations: with the smallest and the largest error. Thus we will consider a reduced model as defined by (2) and by $\epsilon_M^\perp$ we will denote its residuals i.e. $\epsilon_M^\perp = Y_M - Y_M/\mathbf{X}_M$. We will use the same letter π for a random permutation in $\mathbb{R}^{n-m}$ (again disregarding dependence on a dimension). We will denote $R_{\bar{M}} = \{R_i : i \in M\}$. In the problem of estimation of coefficients β we will use contrast vectors being rows of $(\mathbf{X}_M^T\mathbf{X}_M)^{-1}\mathbf{X}_M^T$ and the matrix $\mathbf{X}_M$ is now dependent on values of $R_{\bar{M}}$. For that reason we will allow a contrast vector $V(R_{\bar{M}}) \in \mathbb{R}^{n-m}$ to be dependent on ϵ through its order statistics $R_{\bar{M}}$. Further let $\mathcal{F}_M$ denote the sigma field generated by the order statistics of ϵ_M and by $R_{\bar{M}}$ while $\mathcal{G}_M$ is the sigma field generated by Y_M and $R_{\bar{M}}$. With this notation and conventions we will find that if the ratio $d/(n-m)$ is small then, as before, the distribution $V(R_{\bar{M}}) \cdot \pi\epsilon_M^\perp | \mathcal{G}_M$ approximates the distribution $V(R_{\bar{M}}) \cdot \epsilon_M | \mathcal{F}_M$. This can be stated as the generalization of our proposal.

Proposal 2 *For any finite set of contrast vectors $\{V_k; k \leq l\}$, which may dependent upon $R_{\bar{M}}$, estimate the joint $\mathcal{F}_M$-conditional sampling distribution of contrasts $V_k \cdot \epsilon_M$ by the joint $\mathcal{G}_M$-conditional distribution of $V_k \cdot \pi\epsilon_M^\perp$.*

Notice that $R_{\bar{M}}$ provides information concerning which data might be excluded from the model and Y_M are observations remaining after this exclusion. Thus $\mathcal{G}_M$ represents information which should be given to us to consider a reduced model based on extreme values of errors.

1.2 Main Results

We assume the notation of the introduction. As before, ϵ is assumed to has exchangeable coordinates and to be independent of a random uniformly distributed permutation π. If u is an $k \times 1$ vector then by $\bar{u}, u^2$ we will denote the arithmetic average and the vector of squares of coordinates, respectively. By s_u^2 we will denote the $(k-1)$–divisor sample variance of u i.e. $s_u^2 = k\overline{(u-\bar{u})^2}/(k-1)$. Our goal is to show that for any contrast vector $V(R_{\bar{M}})$ dependent on ranks $R_{\bar{M}}$ the distribution $V(R_{\bar{M}}) \cdot \pi\epsilon_M^{\perp}|\mathcal{G}_M$ approximates the distribution $V(R_{\bar{M}}) \cdot \epsilon_M|\mathcal{F}_M$. Here as before $\mathcal{G}_M = \sigma(Y_M, R_{\bar{M}}) = \sigma(\epsilon_M, R_{\bar{M}})$. In all that follows for a random variable Z or a sigma field $\mathcal{H}$ by $E^Z, E^{\mathcal{H}}$ we will denote conditional expectation with respect to Z and $\mathcal{H}$, respectively. Also using $Z|\mathcal{H}$ we will mean both conditional distribution of a random variable with respect to a σ-field $\mathcal{H}$ and an example of a random variable with this distribution allowing context to distinguish between these two concepts. This convention will simplify notation later when we will consider weak convergence of conditional distributions.

First notice that for $\sigma \in \Sigma_{n-m}$ we have $(\epsilon_M, R_{\bar{M}}) \stackrel{d}{=} (\sigma\epsilon_M, R_{\bar{M}})$. Here Σ_k denotes the set of all permutations of coordinates of vectors in $\mathbb{R}^k$. To see that let consider an event $\{\epsilon_M \in A, R_{\bar{M}} = R_{\bar{M}}\}$. Define a permutation $\tilde{\sigma} \in \Sigma_n$ such that it permutes ϵ_i for $r_i \notin M$ the same way as σ and leaves ϵ_i unchanged for $r_i \in M$. Then $R_{\bar{M}}(\tilde{\sigma}\epsilon) = R_{\bar{M}}(\epsilon)$ on a set $\{R_{\bar{M}}(\epsilon) = R_{\bar{M}}\}$ thus

by exchangeability of ϵ we have

$$\begin{aligned} P(\epsilon_M(\epsilon) \in A, R_{\bar{M}}(\epsilon) = R_{\bar{M}}) &= P(\epsilon_M(\tilde{\sigma}\epsilon) \in A, R_{\bar{M}}(\tilde{\sigma}\epsilon) = R_{\bar{M}}) \\ &= P(\sigma\epsilon_M(\epsilon) \in A, R_{\bar{M}}(\epsilon) = R_{\bar{M}}). \end{aligned}$$

This implies that $V(R_{\bar{M}}) \cdot \epsilon_M | \mathcal{F}_M \stackrel{d}{=} V(R_{\bar{M}}) \cdot \pi\epsilon_M | \mathcal{G}_M$. Indeed, for each bounded and measurable function h and a random permutation of coordinates in $\mathbb{R}^{n-m}$ independent of ϵ we have

$$\begin{aligned} E^{\mathcal{F}_M} h(V(R_{\bar{M}}) \cdot \epsilon_M) &= \frac{1}{(n-m)!} E^{\mathcal{F}_M} \sum_{\sigma \in \Sigma_{n-m}} h(V(R_{\bar{M}}) \cdot \sigma\epsilon_M) \\ &= E^{\mathcal{F}_M} h(V(R_{\bar{M}}) \cdot \pi\epsilon_M) \\ &= E^{\mathcal{G}_M} h(V(R_{\bar{M}}) \cdot \pi\epsilon_M), \end{aligned}$$

where in the last equality we use that $\mathcal{F}_M$ is a sub-σ-field of $\mathcal{G}_M$ and the fact that the last conditional expectation is $\mathcal{F}_M$-measurable. Moreover both $V(R_{\bar{M}}) \cdot \pi\epsilon_M^{\perp} | \mathcal{G}_M$ and $V(R_{\bar{M}}) \cdot \epsilon_M | \mathcal{F}_M$ are centered at zero since, by independence of π and ϵ we have

$$\begin{aligned} E^{\mathcal{G}_M}(V(R_{\bar{M}}) \cdot \pi\epsilon_M^{\perp}) &= E^{\mathcal{G}_M}(\pi V(R_{\bar{M}}) \cdot \epsilon_M^{\perp}) \\ &= \frac{1}{(n-m)!} E^{\mathcal{G}_M} \sum_{\sigma \in \Sigma_{n-m}} (\sigma V(R_{\bar{M}}) \cdot \epsilon_M^{\perp}) \\ &= E^{\mathcal{G}_M} \left(E(\pi V(R_{\bar{M}}))|_{R_{\bar{M}} = R_{\bar{M}}} \cdot \epsilon_M^{\perp} \right) \\ &= 0, \end{aligned}$$

where in the last equality we use the fact that $E(\pi u) = \bar{u}\mathbf{1} = k^{-1}(u \cdot \mathbf{1})\mathbf{1}$ for any $k \times 1$ vector u and thus $E(\pi u) = 0$ if u is a contrast vector. Similarly

we have

$$\begin{aligned} E^{\mathcal{F}_M}(V(R_{\bar{M}}) \cdot \pi\epsilon_M) &= E^{\mathcal{F}_M}(\pi V(R_{\bar{M}}) \cdot \epsilon_M) \\ &= E^{\mathcal{F}_M}\left(E(\pi V(R_{\bar{M}}))|_{R_{\bar{M}}=R_{\bar{M}}} \cdot \epsilon_M\right) \\ &= 0. \end{aligned}$$

The main result can be stated as follows.

Proposition 1 *Let $M \subseteq \{1, \ldots, n\}$ have cardinality m. If* **1** *is in the column space of $\mathbf{X}_M$ which has rank d and π is a random permutation of coordinates in $\mathbb{R}^{n-m}$ then for each contrast vector $V(R_{\bar{M}})$ we have*

$$E^{\mathcal{F}} \frac{E^{\mathcal{G}_M}(V(R_{\bar{M}}) \cdot \pi\epsilon_M - V(R_{\bar{M}}) \cdot \pi\epsilon_M^{\perp})^2}{E^{\mathcal{G}_M}(V(R_{\bar{M}}) \cdot \pi\epsilon_M)^2} = \frac{d-1}{n-m-1}.$$

For the proofs of this and other results in this section see Section 1.4.

As a consequence of the above result we have that the distributions $V(R_{\bar{M}}) \cdot \epsilon_M | \mathcal{F}_M$ and $V(R_{\bar{M}}) \cdot \pi\epsilon_M^{\perp} | \mathcal{G}_M$ are approximately the same when scaled by $E^{\mathcal{F}_M}(V(R_{\bar{M}}) \cdot \epsilon_M)^2$ provided that $d/(n-m)$ is small. This follows directly by Markov's inequality applied to the second moment which stands in the denominator of the left side in Proposition 1 (see also the proof of Proposition 2; Section 1.4). One can use as well the Mallows metric d_2 as a measure of this closeness (see Bickel and Freedman (1981)). The result is interesting by itself. For example any picture illustration of a distribution is in general invariant on rescaling. But when no moment assumptions are involved the scaling $E^{\mathcal{F}_M}(V(R_{\bar{M}}) \cdot \epsilon_M)^2$ may diverge to infinity with n. In such a case we need more arguments to apply our methods of resampling to confidence regions.

The next result explains the practical meaning of our result when we do not use the scaling. For the sake of simplicity of notation we will state it for the case $M = \emptyset$ but it can be restated in the obvious way also in the general case. For a given contrast vector v, let

$$\gamma(\epsilon) = E^{\epsilon}(v \cdot \pi\epsilon - v \cdot \pi\epsilon^{\perp})^2 / E^{\epsilon}(v \cdot \pi\epsilon)^2,$$

$$C_1 = \sqrt{E^{\mathcal{F}}(v \cdot \epsilon)^2},$$

where in the latter we do not show explicitly dependence on the order statistics of ϵ and let

$$C_2(Y) = \sqrt{E^{\epsilon}(v \cdot \pi\epsilon^{\perp})^2}.$$

Denoting by F_1, F_2 the distribution functions of $v \cdot \epsilon | \mathcal{F}$, $v \cdot \pi\epsilon^{\perp} | Y^*$, respectively, we then have the following relations.

Proposition 2 *For any positive δ we have*

$$P(\gamma(\epsilon) > \sqrt{\frac{d-1}{n-1}}\delta) \leq \sqrt{\frac{d-1}{n-1}}\delta^{-1}, \tag{3}$$

$$F_1(x) \in [F_2(x - C_1\delta) - \gamma(\epsilon)\delta^{-2}, F_2(x + C_1\delta) + \gamma(\epsilon)\delta^{-2}] \tag{4}$$

for all $x \in \mathbb{R}$ and

$$E^{\mathcal{F}} C_2^2(Y) = \frac{n-d}{n-1} C_1^2.$$

We can use Proposition 2 to examine accuracy of using the distribution given by F_2 obtained by resampling residuals instead of the unknown conditional distribution given by F_1. For example (4) allows us to assess a value of the

confidence regions based on F_2 as approximations for corresponding confidence regions based on F_1. More precisely, we should adjust the first ones by taking their halo of thickness $C_1\delta$ and then their significance will differ at most by $\gamma(\epsilon)/\delta^2$. By (3) the random variable $\gamma(\epsilon)$ is small with high probability. Thus the accuracy of proposed approximation of F_1 by F_2 depends on how large $C_1\delta$ is. Suppose for example that we would like to use quantiles of F_2 for obtaining confidence intervals. Then we want $C_1\delta$ to be small relative to these quantiles and thus to lengths of confidence intervals. In the next paragraph we will explore this problem in more detail.

For any $p \in (0,1)$ let K_p^1, K_p^2 denote p-quantiles of F_1, F_2, respectively. By (3) for any δ_0 there exist Ω_0, a subset of underlying probability space, with $P(\Omega_0) > 1 - \delta_0$ and a constant $M > 0$ dependent neither on n nor on d such that on Ω_0 we have

$$\gamma(\epsilon) \leq Md/n.$$

By (4) on Ω_0 for each $\delta > 0$ we have

$$F_1(x) \in \left[F_2\left(x - \sqrt{\frac{Md}{\delta}}\frac{C_1}{\sqrt{n}}\right) - \delta, F_2\left(x + \sqrt{\frac{Md}{\delta}}\frac{C_1}{\sqrt{n}}\right) + \delta\right]$$

for all $x \in \mathbb{R}$. This implies that for any $p, \delta > 0$ such that $p - \delta, p + \delta$ still are in $(0,1)$ we have the following relations between corresponding quantiles

$$K_{p-\delta}^2 - \sqrt{\frac{Md}{\delta}}\frac{C_1}{\sqrt{n}} \leq K_p^1 \leq K_{p+\delta}^2 + \sqrt{\frac{Md}{\delta}}\frac{C_1}{\sqrt{n}}.$$

Thus it would be desirable to have $C_1/\sqrt{n}$, at least asymptotically, small relatively to K_p^1. We can say equivalently that we are interested in the cases

when quantiles of $\sqrt{n}\ W_n$ are convergent to infinity, where

$$W_n = \frac{(v \cdot \pi\epsilon)|\epsilon}{\sqrt{E^{\epsilon}(v \cdot \pi\epsilon)^2}}. \tag{5}$$

We will examine this question for errors sampled independently from a distribution belonging to the domain of attraction of a symmetric stable law.

Let $G(x) = x^{-\alpha}\mathbf{1}_{[1,\infty)}(x)$ and for any $n \in \mathbb{N}$ let $(U_{n,i})_{i=1}^n$ and $(\delta_i)_{i=1}^n$ be independent sequences of i.i.d. random variables such that $U_{n,i}$ has the uniform distribution on $(0,1)$ and $P(\delta_i = \pm 1) = 1/2$. Let us define a sequence $\epsilon_n = (\epsilon_{n,i})_{i=1}^n$ of errors by the equality

$$\epsilon_{n,i} = \delta_i G^{-1}(U_{n,i}) = \delta_i U_{n,i}^{-1/\alpha}. \tag{6}$$

It is well known that the distribution of $\epsilon_{n,i}$ belongs to the domain of attraction of a symmetric stable law with index of stability α. In fact for $x > 1$ we have $P(|\epsilon_{n,i}| > x) = x^{-\alpha}$ (see also LePage (1980)). Although we will assume here this special form of errors by application of invariance principle of the type obtained by Kinateder (1990) it is possible to extend our results on an arbitrary law from the domain of attraction of a stable distribution. We will also need a special form for contrast vectors. Let $v : [0,1] \to \mathbb{R}$ be a continuous function such that $\int_0^1 v(x)dx = 0$. Define a sequence $v_n = (v_{i,n})_{i=1}^n$ of contrast vectors by

$$v_{n,i} = v\left(\frac{i}{n}\right) - \frac{1}{n}\sum_{k=1}^{n} v\left(\frac{k}{n}\right). \tag{7}$$

To express the limit distribution of $W_n = (v_n \cdot \pi\epsilon_n)|\epsilon_n/(E^{\epsilon}(v_n \cdot \pi\epsilon)^2)^{1/2}$ we will need a notion and some basic properties of the stochastic integral

of a function v with respect to a Levy motion. For details see for example Schilder (1970) or Kuelbs (1973).

By a Levy motion we mean here a stochastic process $(\Lambda_t)_{t\in[0,1]}$ continuous in probability with independent and homogenous increments such that for any $t \in [0,1]$ a distribution of Λ_t is symmetric stable with characteristic function $\phi_t(u) = \exp(-t|u|^\alpha)$. Then the integral $\int_0^1 v d\Lambda$ is defined in a usual way by as a limit in topology of convergence in probability of simple functions. A random variable $\int_0^1 v d\Lambda$ has a stable distribution with the characteristic function $\phi_v(u) = \exp(-\int_0^1 |v(x)|^\alpha dx\ |u|^\alpha)$. There exists a version of Levy motion such that its trajectories with probability one are right continuous and left limits exist at any point $t \in [0,1]$ (cádlág) (cf. Doob (1953) p.422). Thus we can define a σ-field $\mathcal{J}$ generated by values of jumps of trajectories of this process. We can do it, for example, by ordering values of jumps of a given cádlág trajectory in a sequence since the number of jumps exceeding given positive value is finite. Then $\mathcal{J}$ is just a σ-field generated by such a sequence. By $\Gamma = (\Gamma_i)_{i\in\mathbf{N}}$ we will denote arrival times of a Poisson process with intensity one. In this situation we have the following limit behavior of W_n.

Theorem 1 *For each continuous function $v : [0,1] \to \mathbb{R}$ with probability one the conditional distributions*

$$W_n = \left. \frac{v_n \cdot \pi\epsilon_n}{\sqrt{E^{\epsilon_n}(v_n \cdot \pi\epsilon_n)^2}} \right| \epsilon_n$$

converge weakly to the distribution

$$\left. \frac{\int_0^1 v d\Lambda}{\sqrt{Z \int_0^1 v^2(x)dx}} \right| \mathcal{J},$$

where Z is $\mathcal{J}$ measurable and $Z \stackrel{d}{=} \sum_{i=1}^{\infty} \Gamma_i^{-2/\alpha}$.

Remark Proposition 1 implies that the joint distribution $(v_n \cdot \pi\epsilon_n^{\perp}, v_n \cdot \pi\epsilon_n)|\epsilon_n$ scaled by $\sqrt{E^{\epsilon_n}(v_n \cdot \pi\epsilon_n)^2}$ converge to $(\int_0^1 v d\Lambda, \int_0^1 v d\Lambda)|\mathcal{J}$ scaled by $\sqrt{Z \int_0^1 v^2(x)dx}$.

As a corollary we obtain desired property for the quantiles of the distributions of $\sqrt{n}\, W_n$.

Corollary 1 *For a non-zero continuous function $v : [0,1] \to \mathbb{R}$ the quantiles of $\sqrt{n}\, W_n$, different from median, converge to plus or minus infinity with probability one.*

For the proofs we refer again to Section 1.4.

The special form of contrast vectors assumed above is in some sense essential for obtaining a result of this type. We will see that in some extreme cases i.e. for a very unusual choice of contrast vectors, the quantiles converge to zero. In such a case usage of quantiles of F_2 as approximations of corresponding ones for F_1 becomes worthless.

Example 1

Assume that our contrast vectors are of the form $\mathbf{v}_n = (v_1, \ldots, v_k, 0, \ldots, 0) \in \mathbb{R}^n$, where $k \leq n$ does not depend on n and $(v_1, \ldots, v_k)$ is a fixed contrast vector. Further let $\epsilon_n = (\epsilon_1, \ldots, \epsilon_n)$, where ϵ_i, $i \in \mathbb{N}$ are i.i.d. random variables such that $s_{\epsilon_n}/\sqrt{n}$ diverges in probability to infinity. In such a case $\mathbf{v}_n \cdot \pi\epsilon_n|\epsilon_n$ converges in distribution to the same limit as $\mathbf{v}_n \cdot \epsilon_n^*|\epsilon_n$, where ϵ_n^* means random variable obtained by traditional bootstrap i.e. by resampling with replacement form $(\epsilon_1, \ldots, \epsilon_n)$. It becomes more clear if we

notice that because of the special form of $\mathbf{v}_n$ permuting ϵ_n in $\mathbf{v}_n \cdot \pi\epsilon_n$ is equivalent to sampling without replacement k-element sample form a population of n-elements (we have here k fixed and n going to infinity). Thus the distribution of $\mathbf{v}_n \cdot \pi\epsilon_n | \epsilon_n$ converges weakly to the distribution of $\mathbf{v}_k \cdot \epsilon_k$. But assumption that $s_{\epsilon_k}/\sqrt{n}$ diverges in probability to infinity implies that $\sqrt{n}/\sqrt{E^{\epsilon}(\mathbf{v}_n \cdot \pi\epsilon_n)^2}$ will be convergent to zero so do the quantiles of $\sqrt{n}\, W_n$. See Section 1.4 for more rigorous arguments.

As a specified example of a random variable satisfying the required condition one can consider the errors given by (6) for $\alpha \in (0,1)$. Indeed, for such ϵ we have

$$
\begin{aligned}
\frac{s^2_{\epsilon_n}}{n} &= \frac{\sum_{i=1}^n \Gamma_i^{-2/\alpha}}{\Gamma_{n+1}^{-2/\alpha}} n^{-2} - \frac{\sum_{i=1}^n \delta_i \Gamma_i^{-1/\alpha}}{\Gamma_{n+1}^{-1/\alpha}} n^{-3} \\
&= n^{-2+2/\alpha} \left(\frac{\Gamma_{n+1}}{n}\right)^{2/\alpha} \sum_{i=1}^n \Gamma_i^{-2/\alpha} \qquad (8) \\
&\quad - n^{-3+1/\alpha} \left(\frac{\Gamma_{n+1}}{n}\right)^{1/\alpha} \sum_{i=1}^n \delta_i \Gamma_i^{-1/\alpha}.
\end{aligned}
$$

We are using here the well known fact that

$$(\Gamma_1/\Gamma_{n+1}, \ldots, \Gamma_n/\Gamma_{n+1}) \stackrel{d}{=} (U_{1,n}, \ldots, U_{n,n}).$$

By Law of Large Numbers Γ_n/n converges a.s. to 1. The series $\sum_{i=1}^{\infty} \delta_i \Gamma_i^{-1/\alpha}$ and $\sum_{i=1}^{\infty} \Gamma_i^{-2/\alpha}$ are convergent a.s. to the continuous distributions. Since $n^{-2+2/\alpha}$ for $\alpha \in (0,1)$ diverges thus in this case with probability one $s^2_{\epsilon_n}/n$ is convergent to infinity.

However in some cases our approximations will be justified regardless of the form of contrast vectors. Namely, this will happen if $s_{\epsilon_n}/\sqrt{n}$ is convergent

in probability to zero. Indeed, by simple application of Proposition 1 we can obtain the following corollary. In its formulation we do not indicate explicitly dependence of $s^2_{\epsilon_n}$ as well as F_1, F_2 on n and also it should be remembered that F_1, F_2 being conditional distributions are dependent on a random value of ϵ.

Theorem 2 *If s^2_ϵ/n converges in probability to zero then for any $\delta > 0$ and for sufficiently large n on a set of probability not less than $1 - \delta$ we have*

$$F_1(x) \in [F_2(x - \delta) - \delta, F_2(x + \delta) + \delta].$$

for all $x \in \mathbb{R}$.

Notice that the assumption required in this result is satisfied for errors in domain of attractions of stable laws when $\alpha \in (1, 2]$. In Section 1.4 there are given arguments for errors given (6) but by series representations of stable laws given by LePage (1980) they can be easily extended to the general case.

Theorem 1 and Theorem 2 validate approximation of the unknown distribution of errors of contrasts by permuting residuals under special assumptions on distributions of errors. The failure of this approximation in Example 1 is due to the asymptotically pathological choice of contrast. At this moment we do not know any example with an essentially different type of contrasts for which applicability of Proposition 1 for this approximation would not be valid.

We will end this section with the following result, uniform with respect to numbers of contrast vectors.

Proposition 3 *Under the assumptions of Proposition 1 and for contrast vectors* $V_k(R_{\bar{M}})$, $k = 1, \ldots, r$ *we have*

$$E^{\mathcal{F}} \frac{\sum_{k=1}^{r} E^{\mathcal{G}_M}[(V_k(R_{\bar{M}}) \cdot \pi\epsilon_M) - (V_k(R_{\bar{M}}) \cdot \pi\epsilon_{\bar{M}}^{\perp})]^2}{\sum_{k=1}^{r} E^{\mathcal{G}_M}(V_k(R_{\bar{M}}) \cdot \pi\epsilon_M)^2} = \frac{d-1}{n-m-1}.$$

1.3 Examples

Example 2

We now present a simple example which illustrates resampling by permutations and the advantage of using a reduced model. The vector of errors considered here has one value dominating the other ones and typifies what can happen when the distribution of errors has a long tail (in particular when our errors do not satisfy the usual assumptions about existence of moments). Consider an $n \times 2$ matrix $\mathbf{X}$ defined as follows

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{bmatrix}.$$

We assume that n is even and the columns $X_1 = \mathbf{1}$ and X_2 of $\mathbf{X}$ are orthog-

onal. Suppose that our actual errors consist of an n-vector ϵ given by

$$\epsilon = \begin{bmatrix} a \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

One can notice at once that

$$\begin{aligned} \epsilon/\mathbf{X} &= \frac{1 \cdot \epsilon}{||1||^2} \cdot 1 + \frac{X_2 \cdot \epsilon}{||X_2||^2} \cdot X_2 \\ &= \frac{a}{n} \cdot 1 + \frac{a}{n} \cdot X_2 = \begin{bmatrix} 2a/n \\ \vdots \\ 2a/n \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \end{aligned}$$

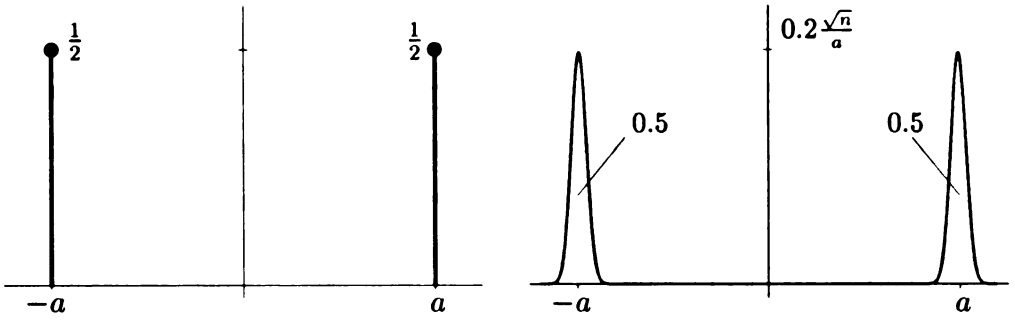
and thus

$$\epsilon^{\perp} = \begin{bmatrix} a - 2a/n \\ -2a/n \\ \vdots \\ -2a/n \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Let us take for our contrast vector v the second column X_2 of the matrix $\mathbf{X}$. One can easily observe that the distribution $v \cdot \epsilon | \mathcal{F}$ is concentrated on two

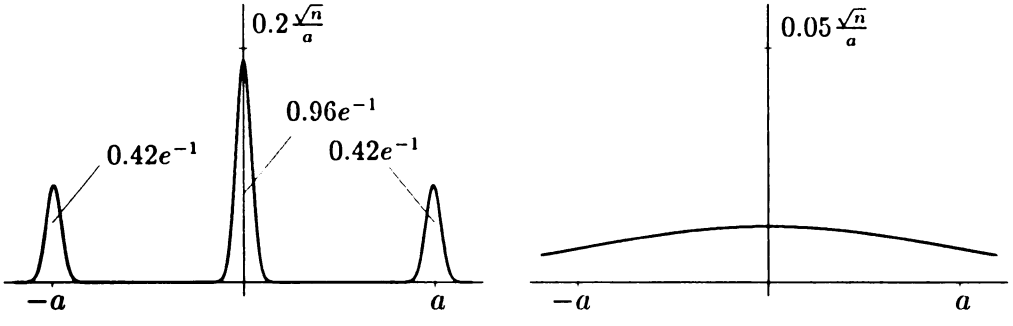
points a and $-a$ with the equal weights $1/2$. We will compare three different methods of estimating this distribution: the resampling by permutation of residuals, the traditional bootstrap with replacement and the normal approximation. On Figure 1 we present the distribution of interests itself versus first order approximations for $n = 625$ of distribution obtained by resampling residuals while on Figure 2 we present first order approximation of the distribution obtained by applying the traditional bootstrap with replacements and the normal approximation. See Appendix 2.5 for details.

If we consider the reduced model (2) for $M = \{n\}$ and if the residual $a - 2a/n$ is large compared to residuals $-2a/n$ we will delete the first row from our model and then we will obtain $\epsilon_M = \epsilon_M^{\perp} = 0$ and thus resampling of residuals gives us a perfect estimate of the distribution of errors.



Distribution of $v \cdot \epsilon | \mathcal{F}$ The resampling by permutation

Figure 1: The conditional distribution of errors and its recovery by permuting residuals

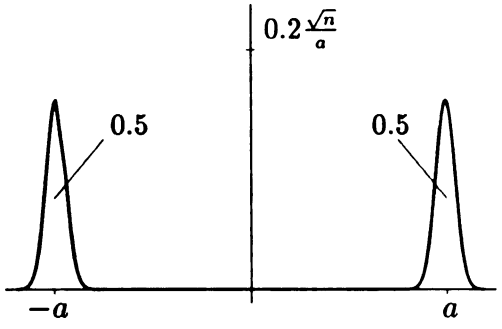


The traditional bootstrapping The normal approximation

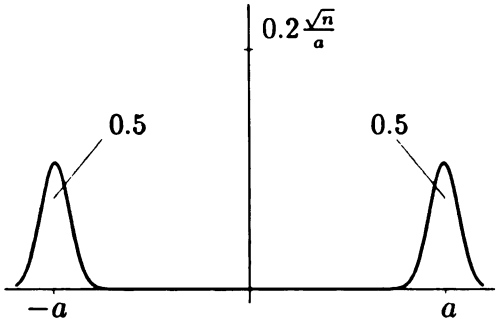
Figure 2: The distributions obtained by other methods

Example 3

As a second example, assume that errors are made from the errors in the first example by adding a sample vector from Gaussian white noise with a small deviation relative to the value of a. In our numerical simulations we will take it equal to $a/\sqrt{n}$. Let us take as before $M = \{n\}$. Figures 3 and 4 present the conditional distributions of error of the contrast in the reduced and unreduced model. It illustrates the advantage of the reduced model. Although approximations in both cases are accurate, in the case of the unreduced model the confidence region which should be used consists of two separated intervals around $-a$ and a while in the reduced model one can consider the natural confidence interval around the origin.

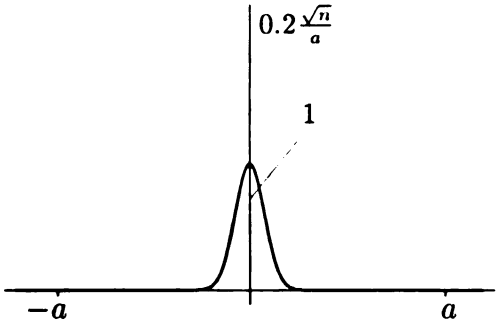


Distribution of $v \cdot \tilde{\epsilon}|\mathcal{F}$ with previous errors modified by adding white noise

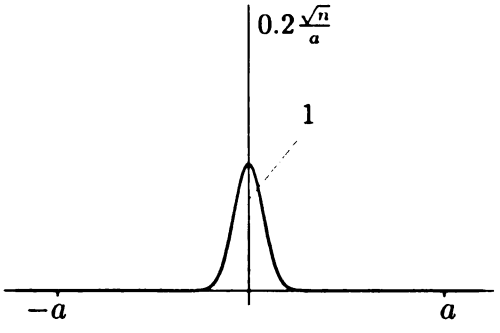


The resampling by permutation for modified errors

Figure 3: Conditional distributions in the unreduced model



Distribution of $v_M \cdot \tilde{\epsilon}_M|\mathcal{F}$ after deletion based on the maximal error



The resampling residuals by permutation after deletion

Figure 4: Conditional distributions in the reduced model

Example 4

As a third illustration we have made computer simulations which compare resampling permutations of errors vs resampling permutations of residuals in a linear regression model fitting a cubic polynomial. Consider a 101–dimensional symmetric α-stable error ϵ with independent coordinates and a 101×4 matrix $\mathbf{X}$ obtained by substitution of equally spaced x–values in the interval $[0,1]$ to the vector of monomials (x^0, x^1, x^2, x^3), so $x_{i,j} = x_i^{j-1}$ $j = 1,2,3,4$, $x_i = (i-1)/101$, $i = 1,\ldots,101$. We have used the Chambers et. al. formula (1983) to generate stable errors by pseudo–random numbers. Since our result does not require any assumption on moments of errors we have observed three notable cases: the normal distribution ($\alpha = 2$), Cauchy distribution ($\alpha = 1$) and $\alpha = 0.8$. We first compute three 101×1 row vectors v_i, $i = 1,2,3$ of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ as well as the matrix of projection on the column space of $\mathbf{X}$. Then, by generating pseudo–random permutations π we get a sample of two hundred errors $v_i \cdot \pi\epsilon$, residuals $v_i \cdot \pi\epsilon^{\perp}$ and projections of errors $v_i \cdot \pi\hat{\epsilon}$, $i = 1,2,3$. Visual comparison of the values of these samples is made on parallel plots. Since scale changes in ϵ apply equally to $v_i \cdot \pi\epsilon, v_i \cdot \pi\epsilon^{\perp}, v_i \cdot \pi\hat{\epsilon}$ it is the relative comparisons between these plots which are important. Four parallel lines represent the coordinate axes of four dimensional space and any point of this space corresponds to a polygonal line which joins values of the coordinates on each axis. For convenience all polygonal lines are started at the origin. For details concerning parallel plots see Chernoff (1973).

In these parallel plots we see a very close agreement between the multivariate sampling distribution of $\{v_i \cdot \pi\epsilon, i = 1,2,3\}$ and that of $\{v_i \cdot \pi\epsilon^{\perp}, i = 1,2,3\}$

conditional on the given ϵ. As might be expected the agreement is even better than suggested by the sampling distribution of $\{v_i \cdot \pi\hat{\epsilon},\ i = 1,2,3\}$. For this example$(d-1)/(n-1) = .03$.

In Part 2 we will give more detailed description of numerical methods and procedures we have used here.

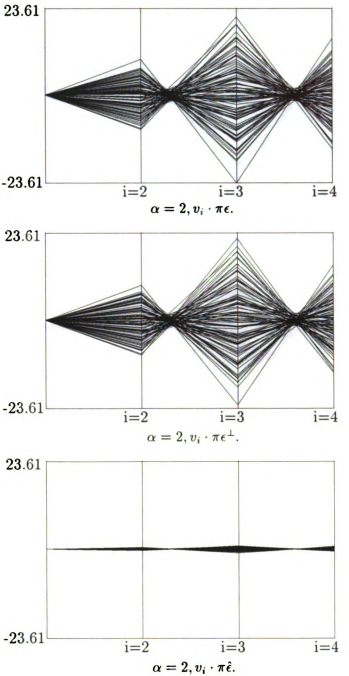


Figure 5: Results of resampling, Gaussian case

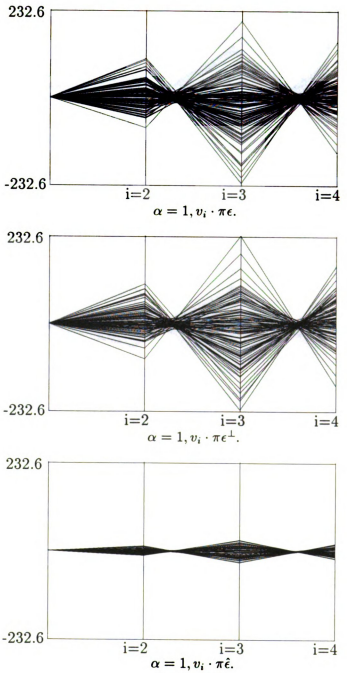
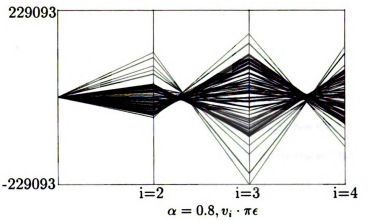
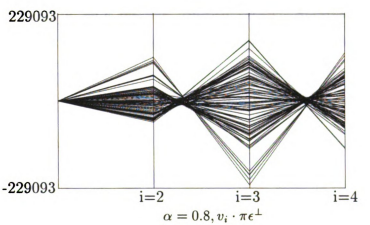


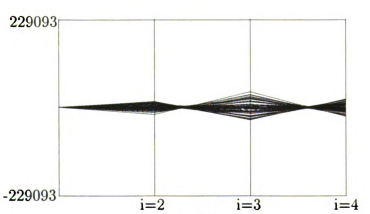
Figure 6: Results of resampling, Cauchy case



$\alpha = 0.8, v_i \cdot \pi\epsilon$



$\alpha = 0.8, v_i \cdot \pi\epsilon^{\perp}$



$\alpha = 0.8, v_i \cdot \pi\hat{\epsilon}$

Figure 7: Results of resampling, stable case with $\alpha = 0.8$

1.4 Proofs

Variants of the following lemma, in the special case when V is not random, are apparently well known Chernoff (1973) Section 4.1; Freedman and Lane (1983), pg 296, Lemma 1). Although formulated here in probabilistic language it illustrates rather a geometrical property of the group of permutations. We present the simple proof of this result.

Lemma 1 *Let V be a random contrast vector. Assume that V, a random $n \times 1$-vector ξ and a σ-field $\mathcal{G}$ all are independent of a random permutation π. Then*

$$E^{\mathcal{G}}(V \cdot \pi\xi)^2 = E^{\mathcal{G}}(||V||^2 s_\xi^2).$$

PROOF. We have

$$\begin{aligned} E^{\mathcal{G}}(V \cdot \xi)^2 &= E^{\mathcal{G}} \sum_{\sigma \in \Sigma_n} E^{\sigma(V,\mathcal{G})}(V \cdot \sigma\xi)^2/n! \\ &= E^{\mathcal{G}} \left(\sum_{a=1}^{n} V_a^2 \xi_{\sigma(a)}^2 + \sum_{a \neq b} V_a V_b \xi_{\sigma(a)} \xi_{\sigma(b)} \right) /n! \\ &= E^{\mathcal{G}} \left((\sum_{a=1}^{n} V_a^2)\overline{\xi^2} + \frac{1}{n(n-1)} (\sum_{a \neq b} V_a V_b)(\sum_{a \neq b} \xi_{\sigma(a)} \xi_{\sigma(b)}) \right) \\ &= E^{\mathcal{G}} \left\{ \|V\|^2 \overline{\xi^2} + \frac{1}{n-1} [\|V\|^2 - (V \cdot \mathbf{1})^2](\overline{\xi^2} - n\bar{\xi}^2) \right\}, \end{aligned}$$

where Σ_n denotes the set of all permutations of coordinates of vectors in $\mathbb{R}^n$. Since $(V \cdot \mathbf{1}) = 0$ we obtain

$$E^{\mathcal{G}}(V \cdot \pi\xi)^2 = E^{\mathcal{G}} \left(\|V\|^2 \frac{n}{n-1} (\overline{\xi^2} - \bar{\xi})^2 \right) = E^{\mathcal{G}}(\|V\|^2 (s_\xi^2)).$$

□

The above lemma and some elementary properties of conditional expectation are the only tools in the proof of our main result.

PROOF OF PROPOSITION 1. By Lemma 1 we have that

$$
\begin{aligned}
E^{\mathcal{G}_M}\left[(V(R_{\bar{M}})\cdot\pi\epsilon_M)-(V(R_{\bar{M}})\cdot\pi\epsilon_M^{\perp})\right]^2 &= E^{\mathcal{G}_M}(V(R_{\bar{M}})\cdot\pi\epsilon_M/\mathbf{X}_M)^2 \\
&= E^{\mathcal{G}_M}(||V(R_{\bar{M}})||^2\cdot s^2_{\epsilon_M/\mathbf{X}_M}) \\
&= s^2_{\epsilon_M/\mathbf{X}_M}E^{\mathcal{G}_M}||V(R_{\bar{M}})||^2 \qquad (9)
\end{aligned}
$$

where the last equality follows from measurability of $s^2_{\epsilon_M/\mathbf{X}_M}$ with respect to $\mathcal{G}_M$. By similar arguments we have also

$$
E^{\mathcal{G}_M}(V(R_{\bar{M}})\cdot\pi\epsilon_M)^2 = s^2_{\epsilon_M}E^{\mathcal{G}_M}||V(R_{\bar{M}})||^2. \qquad (10)
$$

Using the fact that $(\epsilon_M, R_{\bar{M}}) \stackrel{d}{=} (\sigma\epsilon_M, R_{\bar{M}})$ obtained in Section 1.2 and since both $s^2_{\epsilon_M}$ and $E^{\epsilon_M}s^2_{\pi\epsilon_M/\mathbf{X}_M}$ are $\mathcal{F}_M$ measurable we have

$$
\begin{aligned}
E^{\mathcal{F}}\frac{E^{\mathcal{G}_M}\left[(V(R_{\bar{M}})\cdot\pi\epsilon_M)-(V(R_{\bar{M}})\cdot\pi\epsilon_M^{\perp})\right]^2}{E^{\mathcal{G}_M}(V(R_{\bar{M}})\cdot\pi\epsilon_M)^2} &= E^{\mathcal{F}}\frac{s^2_{\epsilon_M}/\mathbf{X}_M}{s^2_{\epsilon_M}} \\
&= E^{\mathcal{F}}\frac{E^{\mathcal{F}_M}s^2_{\epsilon_M}/\mathbf{X}_M}{s^2_{\epsilon_M}} \qquad (11) \\
&= E^{\mathcal{F}_M}s^2_{\pi\epsilon_M/\mathbf{X}_M} \\
&= E^{\epsilon_M}s^2_{\pi\epsilon_M/\mathbf{X}_M}.
\end{aligned}
$$

The proof can be completed using an equality

$$
E^{\mathcal{F}}s^2_{\epsilon/\mathbf{X}} = ((d-1)/(n-1))s^2_{\epsilon}
$$

obtained by Box and Watson (1962); Section 4.1 in their study of F–distribution approximations for $s^2_{\epsilon/X}/s^2_\epsilon$ with non–normal errors ϵ and also mentioned by Freedman and Lane (1983); see (21), p. 267. In our case it may be obtained from Lemma 1 by expanding in an orthonormal basis $\{u_a\}$ of $\mathbf{X}_M$ with $u_1 = \mathbf{1}/(n-m)$ as follows

$$\begin{aligned} E^{\epsilon_M} s^2_{\pi\epsilon_M/\mathbf{X}_M} &= \frac{1}{n-m-1} E^{\epsilon_M}\left(\|\pi\epsilon_M/\mathbf{X}_M\|^2 - (\pi\epsilon_M/\mathbf{X}_M \cdot \mathbf{1})^2/(n-m)\right) \\ &= \frac{1}{n-m-1} E^{\epsilon_M}\left(\sum_{a=1}^{d}(u_a \cdot \pi\epsilon_M)^2 - (\epsilon_M \cdot \mathbf{1})^2/(n-m)\right) \\ &= \frac{1}{n-m-1}\sum_{a=2}^{d} E^{\epsilon_M}(u_a \cdot \pi\epsilon_M)^2 \\ &= \frac{1}{n-m-1}\sum_{a=2}^{d} \|u_a\|^2 s^2_{\epsilon_M} \\ &= \frac{d-1}{n-m-1} s^2_{\epsilon_M}, \end{aligned}$$

where Lemma 1 was used in the third equality.

□

Likewise we can obtain the result on the relative second moments of difference of $\pi\epsilon_M - \pi\epsilon^\perp_M$. For simplicity of the notation let us formulate it in the case $M = \emptyset$. If we denote by $\mathbf{X} \ominus \mathbf{1}$ the orthogonal complement of $\mathbf{1}$ in the column space of $\mathbf{X}$ we have then

$$E^{\mathcal{F}} \frac{E^\epsilon \|(\pi\epsilon - \pi\epsilon^\perp)/\mathbf{X} \ominus \mathbf{1}\|^2}{E^\epsilon \|(\pi\epsilon)/\mathbf{X} \ominus \mathbf{1}\|^2} = \frac{d-1}{n-1}.$$

Indeed, for any vector u we have by the definition of s^2_u that $s^2_u = s^2_{u/\mathbf{X}\ominus\mathbf{1}}$ and $\|u\|^2 = (u \cdot 1)^2/n + (n-1)s^2_u$. Thus

$$E^{\mathcal{F}} \|(\pi\epsilon - \pi\epsilon^\perp)/\mathbf{X} \ominus \mathbf{1}\|^2 = (n-1)E^{\mathcal{F}} s^2_{\pi(\epsilon/X)} = (n-1)E^{\mathcal{F}} s^2_{\epsilon/X}$$

and

$$E^{\mathcal{F}}\|(\pi\epsilon)/\mathbf{X} \ominus \mathbf{1}\|^2 = (n-1)E^{\mathcal{F}}s_\epsilon^2.$$

PROOF OF PROPOSITION 2. For any random variables Z_1, Z_2 with marginal distribution functions F_1, F_2 and for each $\delta > 0, C > 0$ we have that

$$F_2(x - C\delta) - P\left(\frac{Z_2 - Z_1}{C} \geq \delta\right) \leq F_1(x) \leq F_2(x + C\delta) + P\left(\frac{Z_2 - Z_1}{C} \geq \delta\right),$$

which applied to $Z_1 = v \cdot \pi\epsilon$, $Z_2 = v \cdot \pi\epsilon^\perp$, $P = P^\epsilon$ and $C = C_1$ (notice that $Z_1|\epsilon = v \cdot \epsilon|\mathcal{F}, Z_2|\epsilon = v \cdot \pi\epsilon^\perp|Y$ and $C_1 = \sqrt{E^\epsilon Z_1^2}$) together with the Markov inequality

$$P^\epsilon\left(\frac{|Z_2 - Z_1|}{C_1} \geq \delta\right) \leq \frac{E^\epsilon|Z_2 - Z_1|^2}{C_1^2}\delta^{-2}$$

gives us the second part of Proposition 2.

From the Markov inequality applied to non-negative $\gamma(\epsilon)$ we get

$$P(\gamma(\epsilon) > \delta) \leq E\ \gamma(\epsilon)/\delta.$$

Thus the first part of Proposition 2 follows from Proposition 1 since it implies that $E\ \gamma(\epsilon) = (d-1)/(n-1)$. The third part of Proposition 2 follows from Proposition 1 if we notice that it remains valid when we replace $\epsilon^\perp$ by $\epsilon - \epsilon^\perp$ and $d-1$ by $n-d$. In fact it follows from (11) that

$$E^{\mathcal{F}}\frac{E^\epsilon(v \cdot \pi\epsilon^\perp)^2}{C_1^2} = E^{\mathcal{F}}\frac{s^2_{\epsilon/\mathbf{X}^\perp}}{s_\epsilon^2},$$

where $\mathbf{X}^\perp$ denotes orthogonal complement of $\mathbf{X}$. So the expansion in orthonormal basis in $\mathbf{X}^\perp$ gives us, as before, that $E^{\mathcal{F}}s^2_{\epsilon/\mathbf{X}^\perp} = s_\epsilon^2(n-d)/(n-1)$ (notice that basis will consist of $n-d$ contrast vectors).

□

We will precede the proof of Theorem 1 by some considerations on constructions of series representations of stable laws and Lévy processes as introduced by LePage (1980). Let $u = (u_i)_{i=1}^{\infty}$ be any sequence of numbers belonging to $(0,1)$. We will define

$$I_1^n(u,t) = \mathbf{1}_{\{s\in[0,1]:u_1\leq[ns]/n\}}(t)$$

and for $1 < i \leq n$, by recursion,

$$I_i^n(u,t) = \mathbf{1}_{\{s\in[0,1]:u_i-\sum_{k=1}^{i-1} I_k^n(u,t)\leq[ns]/(n-i)\}}(t).$$

It is not difficult to verify the following list of properties of a sequence $(I_i^n(u,t))_{i=1}^n$ (see also LePage (1980) and Kinateder (1990)). First it is obvious that a sequence $(I_i^n(u,t))_{i=1}^n$ consists only of zeros and ones. Moreover for fixed t and n there are exactly $k = [nt]$ ones in this sequence. Such a sequence can be interpreted as a combination without replacement of k-elements form n-elements when positions of ones indicate elements of $\{1,\ldots,n\}$ which are chosen to this combination. On the other hand if we fix n, and u then the number of ones in the sequence will increase from 0 to n as t will increase form 0 to 1. A sequence of positions of subsequent appearance of ones or, equivalently, jumps in the sequence $(I_i^n(u,t))_{i=1}^n$ is a permutation of elements of $\{1,\ldots,n\}$. It is worth notice here that these jumps can occur only at $t = k/n$ and one at the time. The combination and permutation, which can be related in this way to $(I_i^n(u,t))_{i=1}^n$, depend in fact only on values of $u = (u_i)_{i=1}^{\infty}$. We will refer to this as a combination and a permutation chosen by u.

Now if $U = (U_i)_{i=1}^{\infty}$ is a sequence of i.i.d. random variables with uniform distribution on $(0,1)$ then for fixed $t \in [k/n, (k+1)/n)$, $k \in \{0, \ldots, n\}$, a random combination chosen by U has a uniform distribution over all possible $\binom{n}{k}$ combinations. Also the positions of jumps of $(I_i^n(U,\cdot))_{i=1}^n$ i.e. the permutation chosen by U will be uniformly distributed over all $n!$ possible permutations. Moreover $I_i^n(u,t)$ as a function of t, for fixed u, n and i, will change its value form zero to one at one and only one of the points $t = k/n$, and it is right continuous. It follows from that

$$I_i^n(u,t) = \mathbf{1}_{[0,t]}(J_i^n(u)),$$

where J_i^n denotes the jump of $I_i^n(u,\cdot)$. So for any sequence of random variables $\epsilon_n = (\epsilon_{n,i})_{i=1}^n$ independent of U a formula

$$L^n(t) = \sum_{i=1}^{n} \epsilon_{n,i} I_i^n(U,t), \tag{12}$$

defines cádlág process on $[0,1]$ constant on intervals $[(k-1)/n, k/n)$, where $k \in \{1, \ldots, n\}$ and with n-jumps at points $k/n, k = 1, \ldots, n$ equal to values of $\epsilon_{n,i}$, $i = 1, \ldots, n$ randomly permuted by U. Thus

$$(\Delta L_k^n)_{k=1}^n = \left(L^n\left(\frac{k}{n}\right) - L^n\left(\frac{k-1}{n}\right) \right)_{k=1}^n$$

is a random permutation of ϵ_n chosen by U. Consequently, for any vector $v \in \mathbb{R}^n$ we have

$$v \cdot \pi\epsilon_n | \epsilon_n = \sum_{i=1}^{n} v_k \Delta L_i^n | (\delta_i \Gamma_i)_{i=1}^{\infty}$$

From now on let $\Gamma = (\Gamma_i)_{i=1}^{\infty}$ will be arrival times of a Poisson process with intensity equal to one i.e. $\Gamma_i = Z_1 + \cdots + Z_i$ where Z_i are i.i.d. random

variables with exponential distribution with a parameter equal to one. Then $(\Gamma_i/\Gamma_{n+1})_{i=1}^n$ is a vector of i.i.d. random variables uniformly distributed over $(0,1)$. Let $\delta = (\delta_i)_{i=1}^\infty$ be random signs defined as in Section 1.2. Moreover let Γ and δ be mutually independent. Then

$$\epsilon_n = (\epsilon_{n,i})_{i=1}^n = \left(\delta_i \frac{\Gamma_i^{-1/\alpha}}{\Gamma_{n+1}^{-1/\alpha}}\right)_{i=1}^n$$

is a vector of i.i.d. random variables with a distribution form the domain of attraction of a symmetric α–stable law.

For a continuous function v on $[0,1]$ let Riemman type sums be defined by $\sum_{k=1}^N v(k/n)\Delta\Lambda_k^n$, where $\Delta\Lambda_k^n = \Lambda(k/n) - \Lambda((k-1)/n)$. They converge in probability to $\int_0^1 v d\Lambda$ (see Schilder (1970)). Kinateder (1990) proved that a sequences of cádlág processes $(L^n(t))_{t\in[0,1]}$ is weakly convergent under Skorohod metric to $(\Lambda^n(t))_{t\in[0,1]}$ (see also LePage (1980)). The main idea was to apply the series representation of LePage (1980) of a Lévy motion and integrals with respect to Lévy motion. This representation states that a symmetric α–stable Lévy motion with cádlág trajectories can be written as

$$\Lambda(t) = \sum_{i=1}^n \delta_i \Gamma_i^{-1/\alpha} \mathbf{1}_{[U_i,1]}(t) = \sum_{i=1}^n \delta_i \Gamma_i^{-1/\alpha} \mathbf{1}_{[0,t]}(U_i),$$

where U is independent of (δ, Γ). One can easily extend this formula for stochastic integral with respect to Lévy motion to obtain

$$\int_0^1 v d\Lambda = \sum_{i=1}^\infty \delta_i \Gamma_i^{-1/\alpha} v(U_i).$$

The next proof exploits this convenient representation once again and applies it in a similar manner to Kinateder (1990) to obtain that

$$\lim_{n\to\infty} \sum_{k=1}^n v(k/n)\Delta L_k^n | (\delta_i\Gamma_i)_{i=1}^\infty \overset{\text{a.s.}}{=} \int_0^1 v d\Lambda | \mathcal{J}.$$

PROOF OF THEOREM 1. By definition of v_n (see (7)) and Lemma 1 we have

$$\frac{v_n \cdot \pi\epsilon_n|\epsilon_n}{\sqrt{E^{\epsilon_n}(v_n \cdot \pi\epsilon_n)^2}} = \frac{||\epsilon_n||}{||v_n||s_{\epsilon_n}}\left(\frac{\sum_{i=1}^n v(i/n)\Delta L_i^n|(\delta_i\Gamma_i)_{i=1}^\infty}{||\epsilon_n||} - \frac{\sum_{i=1}^n v(i/n)}{n}\frac{\sum_{i=1}^n \Delta L_i^n}{||\epsilon_n||}\right).$$

Notice that

$$\frac{||\epsilon_n||}{||v_n||s_{\epsilon_n}} = \left(1-\frac{1}{n}\right)^{1/2}\left(\sum_{i=1}^n v^2\left(\frac{i}{n}\right)/n - \left(\sum_{i=1}^n v\left(\frac{i}{n}\right)/n\right)^2\right)^{-1/2} \left(\frac{\sum_{i=1}^n \Gamma_i^{-2/\alpha}}{\sum_{i=1}^n \Gamma_i^{-2/\alpha} - (\sum_{i=1}^n \delta_i\Gamma_i^{-1/\alpha})^2/n}\right)^{1/2},$$

and

$$\frac{\sum_{i=1}^n \Delta L_i^n}{||\epsilon_n||} = \frac{\sum_{i=1}^n \delta_i\Gamma_i^{-1/\alpha}}{\sqrt{\sum_{i=1}^n \Gamma_i^{-2/\alpha}}}.$$

By almost sure convergence of $\sum_{i=1}^\infty \delta_i\Gamma_i^{-1/\alpha}$ and $\sum_{i=1}^\infty \Gamma_i^{-2/\alpha}$ and the fact that $\lim_{n\to} \sum_{i=1}^n v\left(\frac{i}{n}\right)/n = \int v(t)dt = 0$ we have with probability one

$$\lim_{n\to\infty} \frac{||\epsilon_n||}{||v_n||s_{\epsilon_n}} = (\int v^2(t)dt)^{-1}, \tag{13}$$

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n v(i/n)}{n}\frac{\sum_{i=1}^n \Delta L_i^n}{||\epsilon_n||} = 0 \tag{14}$$

and

$$\lim_{n\to\infty} ||\epsilon_n|| = \sqrt{\sum_{i=1}^\infty \Gamma_i^{-2/\alpha}}. \tag{15}$$

Now we will show that $\sum_{i=1}^{n} v\left(\frac{i}{n}\right) \Delta L_i^n$ converges almost surely to $\int v d\Lambda$. Kinateder (1990) has proven that a sequence $L_n(t)$ is convergent with probability one to $\sum_{i=1}^{\infty} \Gamma_i^{-1/\alpha} \delta_i \mathbf{1}_{(0,t]}(U_i)$ (in fact convergence of the sequence of entire stochastic processes holds in Skorohod metric). Mostly the same arguments apply here when instead of $\mathbf{1}_{(0,t)}$ we will take a continuous function v. Namely with probability one

$$\lim_{n\to\infty} \sum_{i=1}^{n} \Gamma_i^{-1/\alpha} \delta_i v(R_i^n(U)) = \sum_{i=1}^{\infty} \Gamma_i^{-1/\alpha} \delta_i v(U_i),$$

since

$$\lim_{n\to\infty} R_i^n(U) \stackrel{\text{a.s.}}{=} U_i$$

(see Kinateder (1990); Proof of Proposition 3.1). In view of the series representation and definition of ΔL_i^n this is equivalent to

$$\lim_{n\to\infty} \sum_{i=1}^{n} v\left(\frac{i}{n}\right) \Delta L_i^n \stackrel{\text{a.s.}}{=} \sum_{i=1}^{\infty} \Gamma_i^{-1/\alpha} \delta_i v(U_i) \stackrel{d}{=} \int_0^1 v d\Lambda.$$

This implies that

$$\begin{aligned} \lim_{n\to\infty} v_n \cdot \pi\epsilon_n | \epsilon_n \quad &\stackrel{d}{=} \quad \lim_{n\to\infty} \sum_{i=1}^{n} v(i/n) \Delta L_i^n | (\delta_i \Gamma_i)_{i=1}^{\infty} \\ &\stackrel{\text{a.s.}}{=} \quad \sum_{i=1}^{\infty} \Gamma_i^{-1/\alpha} \delta_i v(U_i) | (\delta_i \Gamma_i)_{i=1}^{\infty} \\ &\stackrel{d}{=} \quad \int_0^1 v d\Lambda | \mathcal{J}. \end{aligned} \tag{16}$$

Combining (13)–(16) we obtain Theorem 1.

□

PROOF OF COROLLARY 1. By Theorem 1 it is enough to show that the limiting distribution of W_n does not have an atom at zero with probability one. Assume to the contrary that the event

$$\left\{P\left(\sum_{i=1}^{\infty}\Gamma_i^{-1/\alpha}\delta_i v(U_i)=0|(\Gamma_i\delta_i)_{i\in\mathbf{N}}\right)>0\right\}$$

has positive probability. Then

$$\begin{aligned} P\left(\sum_{i=1}^{\infty}\Gamma_i^{-1/\alpha}\delta_i v(U_i)=0\right) &= EP\left(\sum_{i=1}^{\infty}\Gamma_i^{-1/\alpha}\delta_i v(U_i)=0|(\Gamma_i\delta_i)_{i\in\mathbf{N}}\right) \\ &> 0. \end{aligned} \tag{17}$$

But $\sum_{i=1}^{\infty}\Gamma_i^{-1/\alpha}\delta_i v(U_i) = \int_0^1 v d\Lambda$ and since v is a non-zero function thus $\int_0^1 v d\Lambda$ has α–stable distribution which is continuous. This contradicts (17).

□

Let us discuss now the statements included in Example 1 in Section 1.2. For a fixed ω belonging to the underlying probability space by $\epsilon_n^*(\omega) = (\epsilon_{n,i}^*(\omega))_{i=1}^n$ we will denote a random variable obtained by traditional bootstrap i.e. by re-sampling with replacement from $(\epsilon_i(\omega))_{i=1}^n$. For ω from the set of probability one for $l = 1,\ldots,k$ we have

$$\lim_{n\to\infty} Ee^{itv_l\cdot\epsilon_{n,l}^*(\omega)} = \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} e^{itv_l\epsilon_j(\omega)} = Ee^{itv_l\epsilon_1}. \tag{18}$$

We will assume that for our fixed ω (18) holds. Denote $\epsilon_n(\omega) = (\epsilon_i(\omega))_{i=1}^n$. Then

$$\begin{aligned} |Ee^{itv_n\cdot\pi\epsilon_n(\omega)} - e^{itv_k\cdot\epsilon_k}| &\leq |Ee^{itv_n\cdot\pi\epsilon_n(\omega)} - e^{itv_n\cdot\epsilon_n^*(\omega)}| \\ &\quad + |Ee^{itv_n\cdot\epsilon_n^*(\omega)} - e^{itv_k\cdot\epsilon_k}| \\ &\leq \frac{n!}{(n-k)!}\left(\frac{(n-k)!}{n!} - \frac{1}{n^k}\right) \\ &\quad + \left|\prod_{l=1}^k Ee^{itv_l\epsilon_{n,k}^*(\omega)} - e^{itv_k\cdot\epsilon_k}\right|. \end{aligned}$$

Thus by (18) we get that $\mathbf{v}_n\cdot\pi\epsilon_n|\epsilon_n$ converge weakly to $\mathbf{v}_k\cdot\epsilon_k$ with probability one.

PROOF OF THEOREM 2. By Proposition 1 and Lemma 1 we have

$$E^{\mathcal{F}}(v\cdot\pi\epsilon^{\perp} - v\cdot\pi\epsilon)^2 = (d-1)||v||^2\frac{s_\epsilon^2}{n-1}.$$

Thus by our assumption $E^{\mathcal{F}}(v\cdot\pi\epsilon^{\perp} - v\cdot\pi\epsilon)^2$ converges to zero in probability and thus for a given $\delta > 0$ and for sufficient large n we have

$$P(E^{\mathcal{F}}(v\cdot\pi\epsilon^{\perp} - v\cdot\pi\epsilon)^2 > \delta) < \delta.$$

Now using Markov's inequality in the same manner as in the proof of Proposition 2 we immediately obtain the result.

□

The next proof is just a slight variation of the proof of Proposition 1.

PROOF OF PROPOSITION 3. By (9) and (10) we have

$$\begin{aligned}
&\frac{\sum_{k=1}^{r} E^{\mathcal{G}_M}[(V_k(R_{\bar{M}}) \cdot \pi\epsilon_M) - (V_k(R_{\bar{M}}) \cdot \pi\epsilon_M^{\perp})]^2}{\sum_{k=1}^{r} E^{\mathcal{G}_M}(V_k(R_{\bar{M}}) \cdot \pi\epsilon_M)^2} \\
&\quad = \frac{s^2_{\epsilon_M/\mathbf{X}_M} \sum_{k=1}^{r} E^{\mathcal{G}_M} \|(V_k(R_{\bar{M}})\|^2}{s^2_{\epsilon_M} \sum_{k=1}^{r} E^{\mathcal{G}_M} \|(V_k(R_{\bar{M}})\|^2} \\
&\quad = \frac{s^2_{\epsilon_M/\mathbf{X}_M}}{s^2_{\epsilon_M}}.
\end{aligned}$$

Now conclusion follows immediately from the proof of Proposition 1.

□

Part 2

Numerical Analysis

2.1 General Description of Software

This section contains the description of our software which was created to verify results on resampling methods for linear models. The software is intended to be developed to a package which, we believe, will become a very useful tool not only for verifying theoretical results and comparing different resampling methods but also for tracking new phenomena which can contribute to the development new effective statistical methods. The linear regression and time series are the two main areas in which these programs are assumed to be useful but we believe that there can be more applications for them. From the point of view of functionality our software has two complementary components. The first one includes all procedures for numerical analysis of data will be written in the C–language. To assure the maximal portability we use the standard C–language as defined in Kernighan and Ritchie (1978). The second component has as its goal a visualization of obtained sets of data. We would like to assure simplicity and functionality together with high quality of graphics and we found that the NeXTSTEP environment available on NeXT computers and IBM compatibles satisfy both requirements the best. The whole software in the portable C–language serving the numerical part of our project was written by the author while the front–end and graphical applications available in NeXTSTEP computer environment was programmed by Isaac Tsaiyi in cooperation with author and Dr. R.LePage.

2.2 The Numerical Procedures

As we have mentioned our software is designated to enable simulations and analysis of the resampling methods in linear regression models and time series. It will serve as an easy to use but powerful research tool for wide range of problems connected with resampling methods and linear models in statistics.

In its current stage the data analysis is made in the two steps. First all of our strictly numerical programs and procedures write their results into files with names usually chosen by a user. Then the graphical interface can be used for analysis of the data contained in these files. Thus the analysis and display of the data are separated processes here. In the future however we would like to control them under one joint application running on NeXTSTEP.

In its present form our software has subroutines which compute all linear algebra objects which are extensively used in studies of linear models such as inverse matrices, projections on given subspaces and more. We have also provided procedures for finding the least square estimates of unknown parameters of a given model. Many of them can be applied to time series as well as to linear regression without any change but, of course, the properties of these two classes of linear models are quite different. Figure 8 illustrates the difference in the character of data obtained when resampling by permutation is applied to two different models (linear regression and autoregression). In both cases the underlying distribution was normal.

We also have efficient procedures for generating samples of data which correspond to a given model when its parameters are specified. Once we

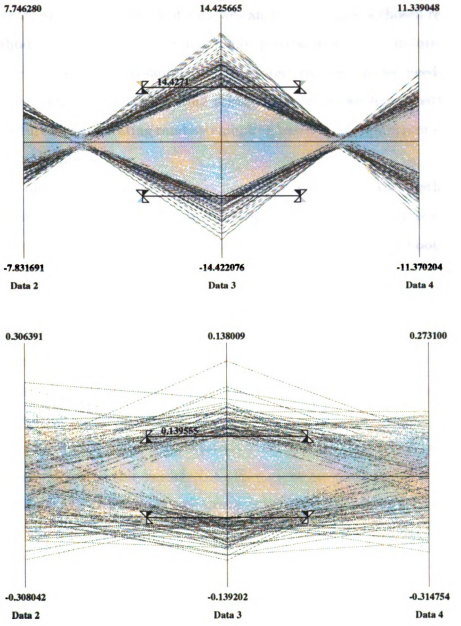


Figure 8: The sample of 1000 elements obtained by permuting residuals in a linear regression model versus an autoregression model

generate an appropriate set of data we can apply to them a chosen resampling method. We have incorporated many programs which compute different characteristics of the resampled data. In our simulations we needed a very efficient generator of resampling schemes therefore we have written quick procedures for generating random sampling based on the generator of pseudo-random numbers.

In our theoretical research we have concentrated on the method of resampling by permutation but in the proposed package a user has also access to other resampling methods like the traditional bootstrap or bootstrapping by signs (see LePage (1990)). Generating sets of data from as many classes of distributions as possible will be also assured. Currently because of our research interests we implemented methods of generating data from the domain of attraction of stable laws. In Figure 9 we can observe how the nature of underlying distributions can influence the obtained numerical results even for the same autoregression model.

We have also made simulations for an autoregression time series. Let us briefly describe this model. Let $(X_n)_{n\in\mathcal{Z}}$ be an autoregressive process of order d or AR(d) for shortness i.e. a stochastic process such that for each $n \in \mathcal{Z}$ the following equation is satisfied

$$X_n = \sum_{k=1}^{d} \beta_k X_{n-k} + \epsilon_n,$$

where $(\epsilon_n)_{n\in\mathcal{Z}}$ is a sequence of i.i.d. random variables and $(\beta_k)_{k=1}^{d}$ is a finite sequence of real numbers. Let $n \in \mathcal{Z}$ and $m \geq d$ will be fixed integers. Denote

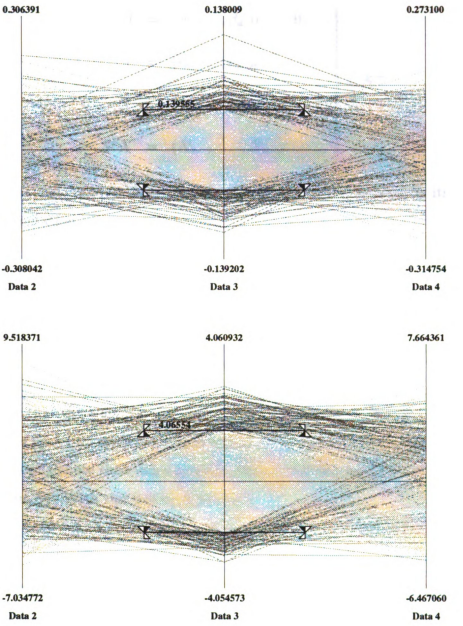


Figure 9: Resampling residuals in an autoregression model with Gaussian errors versus Cauchy errors.

$$\left.\begin{aligned} \beta &= (\beta_1, \ldots, \beta_d, 0, \ldots, 0) \\ \epsilon &= (\epsilon_n, \ldots, \epsilon_{n-m+1}) \\ X &= (X_n, \ldots, X_{n-m+1}) \\ S^k X &= (X_{n-k}, \ldots, X_{n-k-m+1}) \end{aligned}\right\} \in \mathcal{R}^m$$

A symmetric random operator $M : \mathcal{R}^k \to \mathcal{R}^k$ will be defined by its matrix representation

$$M = \begin{bmatrix} X_{n-1} & X_{n-2} & \cdots & X_{n-m} \\ X_{n-2} & X_{n-3} & \cdots & X_{n-m-1} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n-m} & X_{n-m-1} & \cdots & X_{n-2m+1} \end{bmatrix}$$

Notice that the columns of this matrix constitute the vectors $SX, \ldots, S^m X$. Denote by $\mathcal{M}$ a vector space spanned by them and for any vectors $x, y \in \mathcal{R}^m$ by $x \cdot y$ we denote their inner product while $x/\mathcal{M}$ will stand for the projection of x on $\mathcal{M}$. With this notation we have

$$X = M\beta + \epsilon.$$

Moreover,

$$M\beta \in \mathcal{M}.$$

Indeed,

$$\begin{aligned} M\beta &= (SX \cdot \beta, \ldots, S^m X \cdot \beta) \\ &= (\sum_{k=1}^{d} \beta_k X_{n-k}, \ldots, \sum_{k=1}^{d} \beta_k X_{n-k-m+1}) \\ &= \sum_{k=1}^{d} \beta_k S^k X \end{aligned}$$

By *ordinary residuals* for this model we will mean the coefficients of a random vector

$$\epsilon^{\perp} = X - X/\mathcal{M}$$

Since $M\beta \in \mathcal{M}$ we have

$$\epsilon^{\perp} = \epsilon - \epsilon/\mathcal{M}$$

The term *adaptive residuals* will be used when referring to a random vector

$$\begin{aligned} \epsilon^a &= \begin{pmatrix} (X & - & X/\mathcal{M})_1 \\ & \vdots & \\ (S^{m-1}X & - & S^{m-1}X/\mathcal{M})_1 \end{pmatrix} \\ &= \begin{pmatrix} X_n & - & (X/\mathcal{M})_1 \\ & \vdots & \\ X_{n-m+1} & - & (S^{m-1}X/\mathcal{M})_1 \end{pmatrix} \end{aligned}$$

where $(x)_k$ stands the $k-$th coordinate of a vector $x \in \mathcal{R}^m$. Of course, we have

$$\epsilon^a = \begin{pmatrix} (\epsilon & - & \epsilon/\mathcal{M})_1 \\ & \vdots & \\ (S^{m-1}\epsilon & - & S^{m-1}\epsilon/S^{m-1}\mathcal{M})_1 \end{pmatrix}$$

The estimate $\hat{\beta}$ of β which is defined by the equation

$$X/S^{m-1}\mathcal{M} = S^{m-1}M\hat{\beta}$$

was considered even in the case when the second moment of ϵ_n does not exist see Kanter and Steiger (1974), Hannan and Kanter (1977).

Then we have

$$X = M\hat{\beta} + \epsilon^{\perp}.$$

It seems thus to be reasonable to consider the adaptive estimate $\hat{\beta}^a$ of β by the equation

$$X = M\hat{\beta}^a + \epsilon^a.$$

For such a model we have carried on analogous resampling methods. An autoregression model of the order four was considered with the same coefficients in the both cases but in the first one we deal with the Gaussian distribution of underlying stationary random sequence while in the second one a sample from Cauchy distribution was generated. Although we do not have any specefic theoretical results verifying applicability of our method of resampling to this model, results of simulations strongly sugest that results of this type can be true also in this case.

2.3 The Graphical Interface

In our studies of resampling by permutations we have used the *LaPlot* application developed by R. LePage, K. Podgórski and I. Tsaiyi. It exploits the idea of parallel plots. The main window of this application is presented

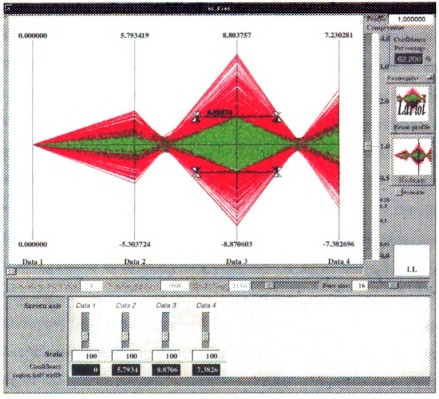


Figure 10: The main window of the *LaPlot* application for analysis of multivariate data.

on Figure 10. On the upper right side of the main window there is the "Create profile" icon which makes it possible to set a profile on the parallel plot symmetric with respect to the horizontal axis. This profile can be compressed (rescaled) by using the "Profile Compression" slide. Those points on the parallel plot which lie outside of the profile are automatically painted in a different color than the others. A user can select the colors used. In Figure 10 a profile was set up on the second and third vertical axes. The profile is indicated by two horizontal lines which show its actual position. Using the "Modify profile" icon one can easily change it. The current percentage of points which are not pruned by our profile is shown in the "Confidence Percentage" display. Confidence interval half–widths appear in the smaller dark windows at the bottom. A change of the horizontal gap between axes can be also regulated by a horizontal slide on the right bottom side of the display window.

The slide just below the graph enable to track the data in the cases when the number of axes is such big that dimensions of our graph exceed horizontal dimension of the graph window. This can be also regulated by a change of the horizontal gap between axes by the horizontal slide on the right bottom side of the main window. The rest of the slides corresponds to the axes showed on the plot and can be used for changing the scale of each of them. Numbers on the top and the bottom of each axis indicates the actual value of the corresponding coordinate of the first point which is not pruned by a given profile.

The data for this graphical interface should be prepared in a file where coordinates of each n dimensional point are written in a one separate row.

In other words a file should have the form of a matrix where rows represents points from the n dimensional Euclidian space which we want to view on the parallel plot. Many additional features which come usually with any NeXTSTEP application are also added as for example possibility of saving a picture as a Post Script file.

2.4 Numerical Experiments

In this section we will discuss some simulation data obtained for few different specifications of the model (2). In our simulations we check the accuracy of taking I_Y as conditional confidence intervals based on resampling residuals by comparing them with random intervals I_ϵ obtained by resampling real errors. We also compare our method with the methods based on the traditional bootstrap. The conditional distributions of both Z and Z' will be examined too. Here we will include only a small variety possible numerical studies. Some others requires additional procedures and supporting programs which are not yet available. The descriptions of existing software, as used here, are given in Appendix B.

The proposed method of resampling is based on rearrangements of residuals by random permutations which corresponds to π in our notation. Although hypothetically we can calculate the exact values of quantiles of a distribution obtained by resampling by considering all permutations of n–dimensional vectors, in practice it would not be possible even for not very large n (52! estimates the number of subatomic particles in the visible universe). We can omit this by sampling independently from the distribution

of π and then using the empirical distribution $G_N(x) = \sum_{k=1}^{N} 1_{(Z'_k \leq x)}$, where $Z'_k = v_j \cdot \pi^k(Y - \hat{Y})$ and π^k, $k = 1, \ldots, N$ are independent permutations of vectors in $\mathcal{R}^n$. Then by the law of large numbers or, better, by the Glivienko–Cantelli Theorem the $\alpha/2$ and $1 - \alpha/2$ sample quantiles a^N and b^N of G_N will be approximations of corresponding quantiles for actual errors a' and b'. In our computer programs we have used this method for finding conditional confidence intervals for a parameter β in a linear regression model.

Example 5

Here we will consider the model based on fitting of the cubic polynomial as described in Section 1.3. The vectors of errors will be an independent sample from stable distributions with the parameter of stability equal to 2, 1 or, in other words, with Gaussian and Cauchy distributions. More precisely, we will consider the matrix X with the coefficients given by $x_{i,j} = x_i^{j-1}$ $j = 1, 2, 3, 4$, $x_i = (i-1)/101$, $i = 1, \ldots, 101$ and as the contrast vectors the rows of the matrix of projection on the column space of X.

Generating independently permutations we have obtained the empirical distributions of random variables denoted earlier by Z and Z'. We can observe that the computable distribution of Z', independently of the underlying distribution, is quite close to the unknown distribution of Z. On Figures 11 and 12 vertical lines on the third axis indicate approximately the same .85 confidence interval for β_3 whether using resampled errors or residuals.

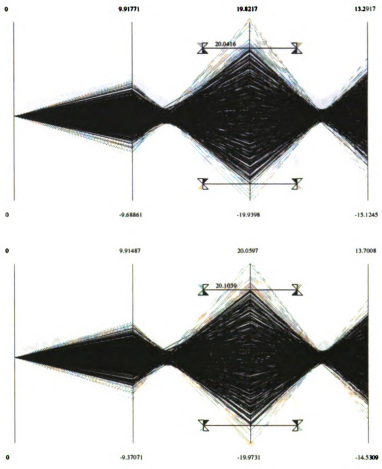


Figure 11: The distribution of resampled errors and residuals, Gaussian case.

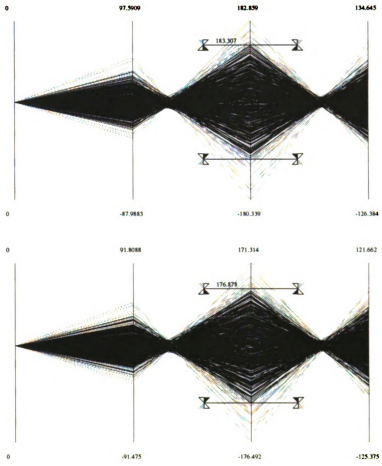


Figure 12: The distribution of resampled errors and residuals, Cauchy case.

2.5 Conclusion

We propose resampling from the permutations of residuals as a method of estimating the conditional sampling distributions of contrasts; without moment assumptions and assuming only exchangeable errors. The idea of resampling permutations of residuals occurs in Box and Watson (1962), Freedman and Lane (1983) and possibly elsewhere. Both papers place considerable emphasis on establishing conditions under which F–distributions associated with standard tests in regression are recovered by resampling permutations. In fact Box and Watson (1962) do not otherwise recommend the method while Freedman and Lane (1983) recommend it, but only as a *descriptive* method. Neither paper makes the crucial observation that the relations of Propositions 1, 2 imply that with probability tending to one the distribution $v \cdot \pi\epsilon^{\perp}|Y$ approximates $v \cdot \epsilon|\mathcal{F}$, when rescaled by $\sqrt{E^{\mathcal{F}}(v \cdot \epsilon)^2}$, as $(d-1)/(n-1) \to 0$. In fact we have independently discovered Propositions 1,2 in the course of extending the results of LePage (1990) which exploit resampling the signs δ of residuals in much the same way: $v \cdot \delta\epsilon^{\perp}|Y \approx v \cdot \epsilon|\,|\epsilon|$ for symmetric errors. We believe that methods which approximately recover conditional distributions by resampling are more widely applicable, e.g. to time series, and we have adopted the term *seamless resampling* to indicate their robust character.

Appendices

Appendix A

We present some facts about exact distributions and their approximations which were described and illustrated in Section 1.3.

We propose $v \cdot \pi\epsilon^{\perp}|\epsilon$ as a good approximation of $v \cdot \pi\epsilon|\mathcal{F}$ except in unusual circumstances. Notice first that this random variable has the same distribution as $\pi v \cdot \epsilon^{\perp}|\epsilon$ and thus takes the values

$$\begin{aligned} y_k^{\pm} &= \pm(a - 2a/n - 2ak/n + 2a(l-k-1)/n) \\ &= \pm 2a(1 - 2(k+1)/n) \end{aligned}$$

with corresponding probabilities

$$\frac{\binom{l-1}{k}\binom{l}{l-1-k}}{2\binom{n-1}{l-1}},$$

where $k = 0, \ldots, l-1$ and $l+1 = n/2$. This follows from the fact that it corresponds to the distribution which samples one's or minus one's. The

above probability is assigned to the event when the choice consists of k one's (minus one's) and $l-1-k$ minus one's (one's). The weight $1/2$ comes from the two symmetric events in which the first value of πv is equal either to one or to minus one.

Let γ denote the hypergeometric distribution with parameters $n-1, l-1, l-1$. Then around points $\pm a$ our conditional distribution corresponds to the distribution of the random variable $\pm\xi = \pm 2a(1-2(\gamma+1)/n)$. Since

$$E(\gamma) = (l-1)^2/(n-1) \approx l/2 = n/4\,,$$

and

$$Var(\gamma) = \frac{(l-1)^2 l^2}{(n-1)^2(n-2)} \approx n/16\,,$$

for large n we have approximately

$$E(\pm\xi) \approx \pm(a - 4a/n) \approx \pm a\,,$$

and

$$\begin{aligned} Var(\xi) &\approx E(-4a/n(\gamma - n/4) - 4a/n))^2 = \\ &= 16a^2/n^2 E(\gamma - n/4)^2 + 16a^2/n^2 \\ &\approx a^2/n(1+16/n) \approx a^2/n\,. \end{aligned}$$

This confirms the fact that our distribution is concentrated quite close to points a and $-a$ and thus is reasonable approximation of the distribution of interest. One can also argue through the normal approximation of γ for large n. Namely, it follows from Theorem 3 of Bickel and Freedman (1984) that $(4\gamma - n)/\sqrt{n}$ converges in law to the standard normal distribution $\mathcal{N}(0,1)$.

So for large n distribution of γ is approximately equal to $\mathcal{N}(n/4, n/16)$ and consequently the distribution of ξ is equal to $\mathcal{N}(a, a^2/n)$.

Now consider the traditional bootstrap where the distribution of interest is approximated by the distribution of $v \cdot (\epsilon^{\perp})^* | \epsilon$, where $(\epsilon^{\perp})^*$ indicates random variable obtained by sampling n-times with replacement from values of the vector $\epsilon^{\perp}$. This random variable is distributed over points of the form

$$(k_1 - k_2)(a - 2a/n) + 2a(l_2 - l_1)/n$$

with probabilities

$$T_{l,\frac{1}{n},\frac{1}{2}-\frac{1}{n}}(k_1, l_1)\, T_{l,\frac{1}{n},\frac{1}{2}-\frac{1}{n}}(k_2, l_2),$$

where $k_1, k_2 = 0, \ldots, n$, $l_1 = 0, \ldots, l - k_1$, $l_2 = 0, \ldots, l - k_2$. Here and henceforth T_{l,p_1,p_2} stands for the trinomial distribution given by

$$T_{l,p_1,p_2}(x, y) = \begin{pmatrix} l \\ x \end{pmatrix} \begin{pmatrix} l - x \\ y \end{pmatrix} (p_1)^x (p_2)^y (p_3)^{l-x-y},$$

where $x = 0, \ldots, l$, $y = 0, \ldots, l - x$ and $p_3 = 1 - p_1 - p_2$. Similarly, $B_{n,p}$ will denote the binomial distribution. Now one can notice that the distribution of $v \cdot (\epsilon^{\perp})^* | \epsilon$ has asymptotically nonvanishing mass around zero.

Indeed, from the following equality

$$T_{l,p_1,p_2}(x, y) = B_{l,p_1}(x)\, B_{l-x,\frac{p_2}{1-p_1}}(y),$$

and taking $k_1 = k_2 = 0$ we obtain that the points $2a(l_2 - l_1)/n$, $l_1, l_2 = 0, \ldots, l$ are distributed with probabilities

$$(1 - 1/n)^n B_{l,\frac{n-2}{2n-2}}(l_1)\, B_{l,\frac{n-2}{2n-2}}(l_2).$$

and using the Poisson approximation for the binomial distribution and Central Limit Theorem we will obtain for large n the approximate distribution equal to $e^{-1}\mathcal{N}(0, a^2/n)$, which gives asymptotically the mass equal to e^{-1}. In fact the mass will be even bigger since we should take into account not only $k_1 = k_2 = 0$ but also all cases when $k_1 - k_2 = 0$. This implies that in this case the traditional bootstrap fails to recover the distribution $v \cdot \epsilon | \mathcal{F}$.

It is possible to explore the distribution of $v \cdot (\epsilon^{\perp})^* | \epsilon$ even more carefully since it is equal to the distribution of $\bar{\gamma}^s \cdot \begin{bmatrix} a - 2a/n \\ 2a/n \end{bmatrix}$, where a two dimensional vector γ has a trinomial distribution with parameters $l, 1/n, 1/2 - 1/n$ and $\bar{X}^s$ denotes symmetrization of a random variable X. So we can find that the distribution will have the positive mass around all points of the form $\pm ka$, $k = 0, 1, 2, \ldots$.

As the last method considered here we will consider the normal approximation $N(0, \hat{\sigma}^2 ||v||^2)$ of our original distribution, where

$$\hat{\sigma}^2 = s^2_{Y^{\perp}} = \frac{(a - 2a/n)^2 + (l-1)(2a/n)^2}{n-1}.$$

We have then $\hat{\sigma}^2 ||v||^2 \approx a^2(1 - 2/n)$ and this method fails completely in recovering the distribution of $v \cdot \epsilon | \mathcal{F}$, and of course the unconditional distribution may be far from normal.

Appendix B

Now we will give short descriptions of the computer programs we have used to obtain the numerical simulations presented in the previous section. In the

near future we intend to develop with Dr. LePage this software to a package which will serve as a useful tool for exploring of the resampling phenomena of the linear models in statistics.

From the point of view of functionality we have divide our tools into two independent although complementary parts, numerical and graphical.

Numerical subroutines and programs. All strictly numerical programs were written in the standard C–language as described in Kernighan and Ritchie (1978) so can be run on any computer with a C–language compiler. All numerical results of these programs are held in files with names specified by a user. The main routines for resampling linear models are called **blr, plr, slr**. They provide quick and convinient tool for resampling in these models. Below we put more detailed description of their usage.

The programs **plr**, **blr**, **slr** create sets of bootstrap replicate betahat vectors using three different methods of resampling: by permutations, by traditional bootstrapping and by sign change, respectively. For example

```
plr 1000 observ design output1  output2 + <ENTER/RETURN>
```

will produce a file **output1** with the least square fit of **observ** to the model **design** and a file **output2** containing 1000 bootstrap betahat vectors obtained by applying ordinary least squares to 1000 independent uniform random permutations of the residuals (i.e. residuals from the original fit).

The programs can also be used with only four parameters if a user wants to have a design matrix and a vector of observations in one file. Then one should just type

```
plr 1000 input output1  output2 + <ENTER/RETURN>
```

The observation vector y is assumed to be the first column in the input file.

The programs **blr** and **slr** work in an analogous way but instead of permuting residuals **blr** uses with-replacement sampling of the residuals while **slr** uses independent sign changes.

Example 6

Consider input files with the following data

```
observations:  design:

    -3         1   1
     2         1   0
     1         1  -1
```

Notice then that the vector of residuals has the form

```
-1
 2
-1
```

Here are examples of output files obtained by applications of our programs to these data. Thus

```
plr 1000 observ design output1  output2 + <ENTER/RETURN>
```

will produce a file **output2** with data

```
 betahat1*    betahat2*

  0.000000   -1.500000
  0.000000   -1.500000
  0.000000   -1.500000
  0.000000    0.000000
  0.000000    1.500000
  0.000000    1.500000
  0.000000    1.500000
  0.000000   -1.500000
  0.000000    1.500000
  0.000000    1.500000
```

Similarly for two others programs.

```
blr 1000 observ design output1  output2 + <ENTER/RETURN>

 betahat1*    betahat2*

-1.000000     0.000000
 1.000000    -1.500000
 2.000000     0.000000
 0.000000     1.500000
-1.000000     0.000000
 0.000000    -1.500000
-1.000000     0.000000
```

```
 0.000000     0.000000
-1.000000     0.000000
 1.000000    -1.500000
```

```
slr 1000 observ design output1  output2 + <ENTER/RETURN>
```

```
 betahat1*    betahat2*

 1.333333     0.000000
-0.666667     1.000000
-1.333333     0.000000
 1.333333     0.000000
 0.666667     1.000000
 0.666667     1.000000
 0.666667    -1.000000
 0.000000     0.000000
 0.000000     0.000000
-0.666667    -1.000000
```

The file **output 1** in all three cases will include the fit:

```
 0.000000
-2.000000
```

Here is the list of other subroutines with their short descriptions.

vanderm0.c This programs creates matrices which are used to define a linear regression models when fitting coefficients of polynomial is consid-

ered. In linear algebra matrices of this type are named Vandermonde matrices and they are of the form $[x_{ij}] = [x_i^j], i = 1, ..., n, j = 1, ..., m$.

stab_err.c We are using the generator of pseudorandom numbers and formula given in Chambers et. al. (1983) to generate a sample from symmetric stable distribution. The values of a sample are written into a file specified by a user.

pow_uni.c Here we are generating a sample from a distribution which is power of the uniform distribution on a given interval.

lin_regr.c For a given regression model we compute all matrices and vectors which are important for statistical analysis (including the projection matrix, the least square fit of unknown parameters and the vector of residuals).

perm_res.c This program generates a sample obtained by resampling a given vector by a random permutation and then taking the inner product with another given vector.

bootstrp.c For a defined regression model and a given vector of errors this program produces three samples connected with our technique of resampling. Namely, we will get samples from exact and approximated conditional distributions as described above as well as a sample which is difference between them. These program was used to verify the accuracy of our approximation.

trad_bts.c The same function as the above program but when the tradi-

tional bootstrapping is considered instead of resampling by permutations.

res_err.c This program computes the distribution of the ends of random intervals defined in Section 1.2 for a given regression model.

conf_res.c A sample from the conditional distribution of the ends of confidence intervals proposed in Section 1.2 is computed.

clt_appr.c A sample of the ends of confidence interval obtained when the normal approximation is used to our distributions. This program was used to compare two methods.

Graphical support. To visualize obtained numerical data we have used two systems Turbo–C, Borland implementation of the C–language available on IBM–compatibles and utilities provided by the NeXTSTEP. The last one we have found particularly useful for creation of the high level graphical utilities and therefore finally we would like to have our package available in the compact form on this computer.

Turbo–C graphics We have created several programs using Turbo Graphic which deliver some pictures illustrating properties of our numerical data. The one which was used here to draw histograms of obtained samples enables to produce a histogram of any set of data (not necessary bell shaped) and illustrates how heavy are the "tails" of our sample which is important when distributions without finite moments are considered.

LaPlot The new application in the NeXTSTEP computer environment was developed to simplify analysis of our data. It enables to track multidimensional data exploiting the idea of parallel plot (see Chernoff (1973)). Confidence regions can be observed in great detail. High quality graphics produce pictures. This application will be developed to the full software serving resampling techniques in linear models. See also Section 2.3.

Bibliography

[1] Bickel, P.J. and Freedman, D.A. (1981). Some assymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196–1217.

[2] Bickel, P.J. and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist.* **12**, 470-482.

[3] Box, C.E.P. and Watson, G.S. (1962). Robustness to non–normality of regression tests. *Biometrica*, **49**, 93–106.

[4] Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983). *Graphical methods for data analysis.* Belmont:Wadsworth.

[5] Chernoff, H. (1973), Using faces to represent points in k–dimensional space. *J. Am. Statist. Assoc.*, **68**, 361-368.

[6] Doob, J.L. (1953), *Stochastic Processes.* Wiley: New York *J. Am. Statist. Assoc.*, **68**, 361-368.

[7] Efron, B. (1979), Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, **7**, 1-26.

[8] Freedman, D.A. (1981). Bootstrapping regression models. *Ann. Statist.*, **9**, 1218-1228.

[9] Freedman, D.A. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. J. Business Econ. Statist. **1**, 292-298.

[10] Hannan, E.J. and Kanter, M. (1977) Autoregressive processes with infinite variance. *J. Appl. Prob.* **14** 411-415.

[11] Kanter, M. and Steiger, W.L. (1974) Regression and autoregression with infinite variance. *Adv. Appl. Prob.* **6** 768-783.

[12] Kernighan, B.W. and Ritchie, D.M. (1978) *The C Programming Language.* Prentice–Hall Inc., Englewood Cliffs, New Jersey

[13] Kinateder, J. (1990), An invariance principle applicable to the bootstrap. *PhD Thesis*, Michigan State University.

[14] Kuelbs, J. (1973), A representation theorem for symmetric stable processes and stable measures. *Z. Wahrsch. Verw. Gebiete* **26**, 259-271.

[15] Lahiri, S.N. (1992). Bootstrapping M–estimators of a multiple linear regression parameter. *Ann. Statist.*, **20**, 1548-1593.

[16] LePage, R. (1980), Multidimensional infinitely divisible variables and processes. Part I: Stable case. *Technical Report # 292*, Statistics Department, Stanford.

[17] LePage, R. (1990), Bootstrapping signs. in *Exploring the limits of bootstrap*, R. LePage and L. Billard Eds., John Wiley and Sons Publ., (1992).

[18] Schilder, M. (1970), Some structure theorems for the symmetric stable laws. *Ann. Math. Statist.*, **41**, 522-536.

[19] Wu, C.F.J. (1986), Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, **14**, 1261-1295.