ISSN 1973-2201 *STATISTICA, anno LXXXIV, n. 2, 2024*
DOI: https://doi.org/10.60923/issn.1973-2201/18816

CONNECTING MODEL-BASED AND MODEL-FREE APPROACHES TO LINEAR LEAST SQUARES REGRESSION

Lutz Dümbgen [1]
Department of Mathematics and Statistics, University of Bern, Bern, Switzerland
Laurie Davies
University of Duisburg-Essen, Essen, Germany

SUMMARY

In a regression setting with a response vector and given regressor vectors, a typical question is to what extent the response is related to these regressors, specifically, how well it can be approximated by a linear combination of the latter. Classical methods for this question are based on statistical models for the conditional distribution of the response, given the regressors. In the present paper it is shown that various p-values resulting from this model-based approach have also a purely data-analytic, model-free interpretation. This finding is derived in a rather general context. In addition, we introduce equivalence regions, a reinterpretation of confidence regions in the model-free context.

Keywords: Gaussian covariate; Haar measure; Projection.

1. INTRODUCTION

Statistical inference with general linear models is a well-established and indespensable tool for data analysis. The standard output of statistical software for linear models includes least squares estimators of parameters and their standard erros as well as p-values for various linear hypotheses. While the latter are based on certain model assumptions, linear models can also be viewed as tools for purely exploratory data analysis. In such a model-free context, one might wonder whether the p-values for, say, the relevance of certain covariates are still meaningful. The surprising answer is yes, these p-values do have a very precise and new interpretation.

To formulate a first result, suppose we observe a response vector $y \in \mathbb{R}^n$ and $p < n$ linearly independent regressor vectors (regressors) $x_1, \ldots, x_p \in \mathbb{R}^n$. For a given integer

[1] Corresponding Author. E-mail: lutz.duembgen@unibe.ch

$p_o \in \{0,\ldots,p-1\}$, we would like to know whether the least squares approximation of y by a linear combination $\hat{y}$ of $x_1,\ldots,x_p$, that is,

$$\|y-\hat{y}\|^2 = \min_{\beta\in\mathbb{R}^p} \Big\|y-\sum_{j=1}^{p}\beta_j x_j\Big\|^2$$

with the standard Euclidean norm $\|\cdot\|$, is substantially better than the restricted least squares approximation of y by a linear combination $\hat{y}_o$ of $x_1,\ldots,x_{p_o}$ only, where $\hat{y}_o := 0$ in case of $p_o = 0$. A classical, model-based answer is to assume that the x_j are fixed while

$$y \sim N_n\Big(\sum_{j=1}^{p}\beta_j x_j, \sigma^2 I\Big), \tag{1}$$

with an unknown parameter vector $\beta \in \mathbb{R}^p$ and an unknown standard deviation $\sigma > 0$. Then an exact p-value of the null hypothesis that

$$\beta_j = 0 \text{ for } p_o < j \le p$$

is given by

$$1 - \mathrm{F}_{p-p_o,n-p}\Big(\frac{\|\hat{y}-\hat{y}_o\|^2/(p-p_o)}{\|y-\hat{y}\|^2/(n-p)}\Big), \tag{2}$$

where $\mathrm{F}_{k,\ell}$ denotes the distribution function of Fisher's F distribution with k and ℓ degrees of freedom. Since $\|y-\hat{y}_o\|^2 = \|y-\hat{y}\|^2 + \|\hat{y}-\hat{y}_o\|^2$, one can deduce from well-known connections between chi-squared, gamma, F and beta distributions that the p-value (2) may be rewritten as

$$\mathrm{B}_{(n-p)/2,(p-p_o)/2}\Big(\frac{\|y-\hat{y}\|^2}{\|y-\hat{y}_o\|^2}\Big), \tag{3}$$

where $\mathrm{B}_{a,b}$ denotes the distribution function of the beta distribution $\mathrm{Beta}(a,b)$ with parameters $a, b > 0$.

Now let us view all observation vectors y and $x_1,\ldots,x_p$ as fixed. To measure to what extent $(x_j)_{p_o<j\le p}$ contributes substantially to the least squares fit $\hat{y}$, let $\hat{y}^*$ be the least squares fit of y after replacing $(x_j)_{p_o<j\le p}$ with a tuple $(x_j^*)_{p_o<j\le p}$ of independent random vectors $x_j^* \sim N_n(0, I)$. A precise measure of the relevance of $(x_j)_{p_o<j\le p}$ is given by the probability that $\|y-\hat{y}^*\|^2$ is smaller than $\|y-\hat{y}\|^2$. The smaller this probability, the higher is the relevance. Interestingly, it can be computed exactly and coincides with the p-value in Eq. (3).

LEMMA 1. *For arbitrary fixed, linearly independent vectors $y, x_1,\ldots,x_p \in \mathbb{R}^n$ and stochastically independent standard Gaussian random vectors $x_j^* \in \mathbb{R}^n$, $p_o < j \le p$,*

$$\frac{\|y-\hat{y}^*\|^2}{\|y-\hat{y}_o\|^2} \sim \mathrm{Beta}\big((n-p)/2,(p-p_o)/2\big).$$

In particular,

$$\mathbb{P}\big(||\boldsymbol{y}-\hat{\boldsymbol{y}}^*||^2 \le ||\boldsymbol{y}-\hat{\boldsymbol{y}}||^2\big) \;=\; \mathrm{B}_{(n-p)/2,(p-p_o)/2}\Big(\frac{||\boldsymbol{y}-\hat{\boldsymbol{y}}||^2}{||\boldsymbol{y}-\hat{\boldsymbol{y}}_o||^2}\Big).$$

This Lemma is essentially a variant of a classical result about the angle between a random linear subspace and a fixed vector, see for instance Theorem 1.1 of Frankl and Maehara (1990). A direct and self-contained proof will be given at the beginning of Section A. But Lemma 1 can be viewed as a special case of a more general connection between the model-based and model-free point of view which is elaborated in Section 2. In particular, the classical model-based p-values do not require a Gaussian distribution of $\boldsymbol{y}$, given $(\boldsymbol{x}_j)_{1\le j\le p}$, and for the model-free interpretation, the random tuple $(\boldsymbol{x}_j^*)_{p_o<j\le p}$ may have different distributions all of which lead to the p-value in Eq. (3). In Section 3 we discuss "equivalence" regions. In the model-based context, these are confidence regions for the unknown mean vector $\mu = \mathbb{E}(\boldsymbol{y})$. Under the model-free point of view, the interpretation of these regions is somewhat different. To illustrate the concept, we describe relatively simple equivalence regions for a sparse signal vector.

Some final comments and an outlook to future work are given in Section 4. In particular, we explain how the considerations and results in the present paper are related to previous work about permutation tests in regression settings, a key reference being Freedman and Lane (1983) and a review of Winkler *et al.* (2014).

Technical details and proofs are deferred to Section A. Throughout this paper we use standard results from multivariate statistics and linear models as presented in standard textbooks, e.g. Mardia *et al.* (1979), Eaton (1983) and Scheffé (1959), without further reference.

2. The F test and other methods revisited

We consider arbitrary vectors $\boldsymbol{y}$ and $\boldsymbol{x}_1,\dots,\boldsymbol{x}_p$ in $\mathbb{R}^n$. At first we discuss the question wether there is any association between $\boldsymbol{y}$ and $(\boldsymbol{x}_j)_{1\le j\le p}$. In the introduction, this corresponds to $p_o = 0$. In Section 2.3 we return to situations in which the contribution of $p_o \in \{1,\dots,p-1\}$ regressors $\boldsymbol{x}_1,\dots,\boldsymbol{x}_{p_o}$ is not questioned. This includes linear regression models with an intercept, accommodated by the trivial regressor $\boldsymbol{x}_1 = (1)_{i=1}^n$.

Concerning the regressors $\boldsymbol{x}_j$, suppose the raw data are given by a data matrix with n rows

$$[y_i, \boldsymbol{w}_i^\top] = [y_i, w_{i,1},\dots,w_{i,d}], \quad 1\le i\le n,$$

containing the values of a response and d covariates for each observation. If the covariates are numerical or 0-1-valued, the usual multiple linear regression model would consider the regressors $\boldsymbol{x}_1 := (1)_{i=1}^n$ and $\boldsymbol{x}_j := (w_{i,j-1})_{i=1}^n$, $2 \le j \le d+1$. More complex models would also include the $\binom{d}{2}$ interaction vectors $\boldsymbol{x}_{j(a,b)} := (w_{i,a}w_{i,b})_{i=1}^n$, $1 \le a < b \le d$. In general, with arbitrary types of covariates, one could think of $\boldsymbol{x}_j = (f_j(\boldsymbol{w}_i))_{i=1}^n$ with given basis functions $f_1,\dots,f_p$.

Let us introduce some notation. The unit sphere of $\mathbb{R}^n$ is denoted by $\mathbb{S}_n$, and $\mathbb{O}_n$ stands for the set of orthogonal matrices in $\mathbb{R}^{n\times n}$.

2.1. *The model-based approach*

We consider the regressors $x_1,\ldots,x_p$ as fixed and y as a random vector. In settings with random regressors, the subsequent considerations concern the conditional distribution of y, given $x_1,\ldots,x_p$. For simplicity we assume throughout that the distribution of y is continuous, i.e. $Pr(y=x)=0$ for any $x\in R^n$.

The null hypothesis of no relationship between y and the regressors $x_1,\ldots,x_p$ can be specified by describing a distribution of y which does not depend on the latter vectors (or any other fixed regressors):

$$H_0:\ \text{The random vector } y \text{ has a spherically symmetric distribution on } R^n.$$

That means, its length $||y||$ and direction $||y||^{-1}y$ are stochastically independent, where $||\boldsymbol{y}||^{-1}\boldsymbol{y}\sim \mathrm{Unif}(\mathbb{S}_n)$, the uniform distribution on the unit sphere $\mathbb{S}_n$. It is well-known that this hypothesis H_0 encompasses the classical assumption that $\boldsymbol{y}\sim N_n(0,\sigma^2\boldsymbol{I})$ for some unknown $\sigma>0$.

In order to derive p-values for H_0, let $S(\boldsymbol{y})=S(\boldsymbol{y},\boldsymbol{x}_1,\ldots,\boldsymbol{y}_p)$ be a test statistic such that high values indicate a potential violation of H_0. Then, a p-value for H_0 is given by

$$\pi(\boldsymbol{y}) \;:=\; \mathbb{P}\big(S(||\boldsymbol{y}||\boldsymbol{u})\geq S(\boldsymbol{y})\,\big|\,\boldsymbol{y}\big), \tag{4}$$

where $\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{S}_n)$ is independent from $\boldsymbol{y}$. If $S(\boldsymbol{y})$ is scale-invariant in the sense that

$$S(c\boldsymbol{v}) \;=\; S(\boldsymbol{v}) \quad \text{for all } \boldsymbol{v}\in\mathbb{R}^n\setminus\{0\} \text{ and } c>0, \tag{5}$$

one can write

$$\pi(\boldsymbol{y}) \;=\; 1-\mathrm{F}(S(\boldsymbol{y})-), \tag{6}$$

with the distribution function F of $S(\boldsymbol{u})$,

$$\mathrm{F}(x) \;:=\; \mathbb{P}(S(\boldsymbol{u})\leq x).$$

Here one could also consider a random vector $\boldsymbol{z}\sim N_n(0,\boldsymbol{I})$ instead of $\boldsymbol{u}$.

EXAMPLE 2 (F TEST). *If $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p$ are linearly independent with $p<n$, and if $S(\boldsymbol{y})$ equals the F test statistic*

$$S(\boldsymbol{y}) \;:=\; \frac{||\hat{\boldsymbol{y}}||^2/p}{||\boldsymbol{y}-\hat{\boldsymbol{y}}||^2/(n-p)}, \tag{7}$$

then scale-invariance of the latter implies that the p-value $\pi(\boldsymbol{y})$ is given by the simplified formula in Eq. (6). Moreover, the distribution function F *in Eq. (6) equals* $\mathrm{F}_{p,n-p}$, *so $\pi(\boldsymbol{y})$ coincides with Eq. (2) in the special case of $p_o=0$. This follows from a standard*

argument for linear models: let $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n$ *be an orthonormal basis of* $\mathbb{R}^n$ *such that* $\operatorname{span}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = \operatorname{span}(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)$. *Then* $\boldsymbol{z} \sim N_n(\mathbb{0}, \boldsymbol{I})$ *has the same distribution as* $\tilde{\boldsymbol{z}} = \sum_{i=1}^n z_i \boldsymbol{b}_i$, *and*

$$S(\tilde{\boldsymbol{z}}) = \frac{\sum_{j=1}^p z_j^2/p}{\sum_{i=p+1}^n z_i^2/(n-p)}$$

has distribution function $\mathrm{F}_{p,n-p}$ *by definition of Fisher's F distributions.*

EXAMPLE 3 (MULTIPLE T TESTS). *Suppose that the linear span* $\mathbb{V}$ *of the regressors* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ *satisfies* $q := \dim(\mathbb{V}) < n$. *Further let* $\mathbb{A}$ *be a subset of* $\mathbb{V} \cap \mathbb{S}_n$. *With the orthogonal projection* $\hat{\boldsymbol{y}}$ *of* $\boldsymbol{y}$ *onto* $\mathbb{V}$, *a possible test statistic is given by*

$$S(\boldsymbol{y}) := \hat{\sigma}^{-1} \sup_{\boldsymbol{a} \in \mathbb{A}} |\boldsymbol{a}^\top \boldsymbol{y}| \quad \text{with} \quad \hat{\sigma} := (n-q)^{-1/2} ||\boldsymbol{y} - \hat{\boldsymbol{y}}||. \tag{8}$$

Note that under H_0, *each term* $\hat{\sigma}^{-1} \boldsymbol{a}^\top \boldsymbol{y}$ *follows student's t distribution with* $n-q$ *degrees of freedom. This Example of* $S(\cdot)$ *is motivated by Tukey's studentized maximum modulus or studentized range test statistics (Miller, 1981).*

EXAMPLE 4 (MULTIPLE F TESTS). *Let* p *and* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ *be arbitrary, and let* Λ *be a family of subsets* M *of* $\{1, \ldots, p\}$ *such that the vectors* $\boldsymbol{x}_j$, $j \in M$, *are linearly independent with* $\#M < n$. *With* Π_M *denoting the orthogonal projection from* $\mathbb{R}^n$ *onto* $\operatorname{span}(\boldsymbol{x}_j : j \in M)$, *a possible test statistic is given by*

$$S(\boldsymbol{y}) := \max_{M \in \Lambda} \frac{||\Pi_M \boldsymbol{y}||^2/\#M}{||\boldsymbol{y} - \Pi_M \boldsymbol{y}||^2/(n - \#M)}. \tag{9}$$

The idea behind this test statistic is that possibly $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ *with a random vector* $\boldsymbol{\varepsilon}$ *having spherically symmetric distribution and a fixed vector* $\boldsymbol{\mu} \in \mathbb{R}^n$ *such that*

$$||\Pi_M \boldsymbol{\mu}||^2 \gg ||\boldsymbol{\mu} - \Pi_M \boldsymbol{\mu}||^2$$

for some $M \in \Lambda$.

2.2. *The model-free point of view*

To elaborate on the connection between model-based and model-free approach, note first that the null hypothesis H_0 is equivalent to an orthogonal invariance property. With $\stackrel{d}{=}$ denoting equality in distribution, the alternative formulation reads as follows:

$$H_0' : \boldsymbol{T}\boldsymbol{y} \stackrel{d}{=} \boldsymbol{y} \text{ for any fixed } \boldsymbol{T} \in \mathbb{O}_n.$$

Another equivalent formulation involves normalized Haar measure Haar_n on $\mathbb{O}_n$. This is the unique distribution of a random matrix $\boldsymbol{H} \in \mathbb{O}_n$ with left-invariant distribution in the sense that

$$\boldsymbol{T}\boldsymbol{H} \overset{d}{=} \boldsymbol{H} \quad \text{for any fixed } \boldsymbol{T} \in \mathbb{O}_n.$$

For a thorough account of Haar measure we refer to Eaton (1989); in Section A we mention two explicit constructions of $\boldsymbol{H}$ and resulting properties. For the moment it suffices to know that also

$$\boldsymbol{H}^\top \overset{d}{=} \boldsymbol{H} \overset{d}{=} \boldsymbol{H}\boldsymbol{T} \quad \text{for any fixed } \boldsymbol{T} \in \mathbb{O}_n.$$

Moreover, for any fixed unit vector $\boldsymbol{v} \in \mathbb{S}_n$, the random vector $\boldsymbol{H}\boldsymbol{v}$ is uniformly distributed on $\mathbb{S}_n$. Now the null hypothesis H_0' may be reformulated as follows:

$$H_0'' : \text{ If } \boldsymbol{H} \sim \mathrm{Haar}_n \text{ is independent from } \boldsymbol{y}, \text{ then } \boldsymbol{H}\boldsymbol{y} \overset{d}{=} \boldsymbol{y}.$$

The equivalence of the null hypotheses H_0, H_0' and H_0'' is explained in Section A.

From now on suppose that the test statistic $S(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is orthogonally invariant in the sense that

$$S(\boldsymbol{T}\boldsymbol{y}, \boldsymbol{T}\boldsymbol{x}_1, \ldots, \boldsymbol{T}\boldsymbol{x}_p) = S(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \quad \text{for all } \boldsymbol{y} \in \mathbb{R}^n \text{ and } \boldsymbol{T} \in \mathbb{O}_n. \tag{10}$$

Since $\boldsymbol{x} \mapsto \boldsymbol{T}\boldsymbol{x}$ preserves inner products, a sufficient condition for orthogonal invariance of the test statistic S is that $S(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ depends only on the inner products $\boldsymbol{y}^\top\boldsymbol{y}$, $\boldsymbol{y}^\top\boldsymbol{x}_j$ and $\boldsymbol{x}_j^\top\boldsymbol{x}_k$, $1 \le j \le k \le p$. Then, the p-value in Eq. (4) may be rewritten as follows:

$$\begin{aligned}
\pi(\boldsymbol{y}) &= \mathbb{P}\big(S(\|\boldsymbol{y}\|\boldsymbol{u}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \ge S(\boldsymbol{y}) \,\big|\, \boldsymbol{y}\big) \\
&= \mathbb{P}\big(S(\boldsymbol{H}\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \ge S(\boldsymbol{y}) \,\big|\, \boldsymbol{y}\big) \\
&= \mathbb{P}\big(S(\boldsymbol{y}, \boldsymbol{H}^\top\boldsymbol{x}_1, \ldots, \boldsymbol{H}^\top\boldsymbol{x}_p) \ge S(\boldsymbol{y}) \,\big|\, \boldsymbol{y}\big) \\
&= \mathbb{P}\big(S(\boldsymbol{y}, \boldsymbol{H}\boldsymbol{x}_1, \ldots, \boldsymbol{H}\boldsymbol{x}_p) \ge S(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \,\big|\, \boldsymbol{y}\big),
\end{aligned}$$

where $\boldsymbol{H} \sim \mathrm{Haar}_n$ is independent from $\boldsymbol{y}$. If we adopt the model-free point of view and consider all vectors $\boldsymbol{y}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ as fixed, we may write

$$\pi(\boldsymbol{y}) = \mathbb{P}\big(S(\boldsymbol{y}, \boldsymbol{H}\boldsymbol{x}_1, \ldots, \boldsymbol{H}\boldsymbol{x}_p) \ge S(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)\big).$$

Thus, $\pi(\boldsymbol{y})$ measures the strength of the apparent association between $\boldsymbol{y}$ and the regressor tuple $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, as quantified by the test statistic $S(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, by comparing the latter value with $S(\boldsymbol{y}, \boldsymbol{H}\boldsymbol{x}_1, \ldots, \boldsymbol{H}\boldsymbol{x}_p)$. That means, the regressor tuple $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ undergoes a random orthogonal transformation, and there is certainly no "true association" between $\boldsymbol{y}$ and $(\boldsymbol{H}\boldsymbol{x}_1, \ldots, \boldsymbol{H}\boldsymbol{x}_p)$. To make the latter point rigorous, note that if

$\boldsymbol{H},\boldsymbol{J} \sim \text{Haar}_n$ and $\boldsymbol{u} \sim \text{Unif}(\mathbb{S}_n)$ are independent (while $\boldsymbol{y}$ is fixed), then $\boldsymbol{J}^\top\boldsymbol{H} \sim \text{Haar}_n$ too, whence

$$\begin{aligned} S(\boldsymbol{y},\boldsymbol{H}\boldsymbol{x}_1,\ldots,\boldsymbol{H}\boldsymbol{x}_p) &\overset{d}{=} S(\boldsymbol{y},\boldsymbol{J}^\top\boldsymbol{H}\boldsymbol{x}_1,\ldots,\boldsymbol{J}^\top\boldsymbol{H}\boldsymbol{x}_p) \\ &= S(\boldsymbol{J}\boldsymbol{y},\boldsymbol{H}\boldsymbol{x}_1,\ldots,\boldsymbol{H}\boldsymbol{x}_p) \\ &\overset{d}{=} S(||\boldsymbol{y}||\boldsymbol{u},\boldsymbol{H}\boldsymbol{x}_1,\ldots,\boldsymbol{H}\boldsymbol{x}_p). \end{aligned}$$

Finally, recall that the method in the introduction with $p_o = 0$ amounts to replacing the fixed regressors $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p$ with independent random vectors $\boldsymbol{x}_1^*,\ldots,\boldsymbol{x}_p^* \sim N_n(0,\boldsymbol{I})$. But in connection with the F test statistic, this has the same effect as replacing the former with $(\boldsymbol{H}\boldsymbol{x}_1,\ldots,\boldsymbol{H}\boldsymbol{x}_p)$. Indeed, in case of linearly independent vectors $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p$, the value of $S(\boldsymbol{y})$ in Eq. (7) depends only on $\boldsymbol{y}$ and the p-dimensional linear space $\text{span}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p)$. Moreover, the distributions of $\text{span}(\boldsymbol{H}\boldsymbol{x}_1,\ldots,\boldsymbol{H}\boldsymbol{x}_p)$ and of $\text{span}(\boldsymbol{x}_1^*,\ldots,\boldsymbol{x}_p^*)$ coincide, see Section A.

2.3. *Composite null models*

Quite often, the potential influence of some regressors $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{p_o}$ with $1 \le p_o < p$ is out of question or not of primary interest, and the main question is whether the regressors $\boldsymbol{x}_{p_o+1},\ldots,\boldsymbol{x}_p$ are really relevant for the approximation of $\boldsymbol{y}$. One possibility to deal with that is to "residualize" the response $\boldsymbol{y}$ and the regressors $\boldsymbol{x}_{p_o+1},\ldots,\boldsymbol{x}_p$, that is, to project them onto the orthogonal complement of $\text{span}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{p_o})$. In case of $p_o = 1$ and $\boldsymbol{x}_1 = (1)_{i=1}^n$, this boils down to centering $\boldsymbol{y}$ and $\boldsymbol{x}_2,\ldots,\boldsymbol{x}_p$.

More formally, assuming without loss of generality that $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{p_o}$ are linearly independent, let $\boldsymbol{b}_1,\boldsymbol{b}_2,\ldots,\boldsymbol{b}_n$ be an orthonormal basis of $\mathbb{R}^n$ such that $\boldsymbol{b}_1,\ldots,\boldsymbol{b}_{p_o}$ is a basis of $\text{span}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{p_o})$. With $\boldsymbol{B} = [\boldsymbol{b}_{p_o+1},\ldots,\boldsymbol{b}_n] \in \mathbb{R}^{n\times(n-p_o)}$, the model Equation (1) implies that

$$\boldsymbol{B}^\top\boldsymbol{y} \sim N_{n-p_o}\Big(\sum_{j=p_o+1}^{p} \beta_j \boldsymbol{B}^\top\boldsymbol{x}_j, \sigma^2\boldsymbol{I}\Big).$$

Generally, the previous model-based and model-free considerations can be applied to $\boldsymbol{B}^\top\boldsymbol{y} \in \mathbb{R}^{n-p_o}$ in place of $\boldsymbol{y}$.

Applying the model-based or model-free approach with the F test statistic in Eq. (7) to the transformed observations $\boldsymbol{B}^\top\boldsymbol{y}$ and $\boldsymbol{B}^\top\boldsymbol{x}_{p_o+1},\ldots,\boldsymbol{B}^\top\boldsymbol{x}_p$ yields the p-value of Eq. (3) in the introduction, see also the first part of the proof of Lemma 1.

3. Equivalence regions

3.1. General considerations

Model-based approach. Let $\mathbb{M} \subset \mathbb{R}^n$ be a given set. We assume that $\boldsymbol{y}$ is a random vector such that

$$\boldsymbol{y} \overset{d}{=} \mu + \boldsymbol{\varepsilon}, \tag{11}$$

with an unknown fixed parameter vector $\mu \in \mathbb{M}$ and a random vector $\boldsymbol{\varepsilon}$ with spherically invariant distribution on $\mathbb{R}^n$, where $\mathbb{P}(\boldsymbol{\varepsilon} = 0) = 0$. Now, let $S(\boldsymbol{y}) = S(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ be a test statistic which is scale-invariant in $\boldsymbol{y} \neq 0$, and let $S(0) = 0$. For a given (small) number $\alpha \in (0,1)$ we define the equivalence region

$$C_\alpha(\boldsymbol{y}) = C_\alpha(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) := \big\{\boldsymbol{m} \in \mathbb{M} : S(\boldsymbol{y} - \boldsymbol{m}) \le \varkappa_\alpha\big\},$$

where $\varkappa_\alpha$ is the $(1-\alpha)$-quantile of the distribution of $S(\boldsymbol{u}) \overset{d}{=} S(\boldsymbol{z})$ with random vectors $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}_n)$ and $\boldsymbol{z} \sim N_n(0, \boldsymbol{I})$. This defines an $(1-\alpha)$ confidence region for μ in the sense that in case of Eq. (11),

$$\mathbb{P}(C_\alpha(\boldsymbol{y}) \ni \mu) = \mathbb{P}(S(\boldsymbol{z}) \le \varkappa_\alpha) \ge 1 - \alpha.$$

Model-free interpretation. We consider $\boldsymbol{y}$ as fixed and assume in addition that $S(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is orthogonally invariant. Then, the equivalence region $C_\alpha(\boldsymbol{y}) = C_\alpha(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ consists of all vectors $\boldsymbol{m} \in \mathbb{M}$ such that the association between $\boldsymbol{y} - \boldsymbol{m}$ and the tuple $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is not substantially stronger than the association between $\boldsymbol{y} - \boldsymbol{m}$ and the randomly rotated tuple $(\boldsymbol{H}\boldsymbol{x}_1, \ldots, \boldsymbol{H}\boldsymbol{x}_p)$, where $\boldsymbol{H} \sim \mathrm{Haar}_n$. Precisely, the value $S(\boldsymbol{y} - \boldsymbol{m}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, our measure of association, is not larger than the $(1-\alpha)$-quantile of $S(\boldsymbol{y} - \boldsymbol{m}, \boldsymbol{H}\boldsymbol{x}_1, \ldots, \boldsymbol{H}\boldsymbol{x}_p)$.

Example 2 (continued) Let $\mathbb{M} = \mathrm{span}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = \boldsymbol{X}\mathbb{R}^p$ with the matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \in \mathbb{R}^{n \times p}$. Then, the equivalence region equals

$$C_\alpha(\boldsymbol{y}) = \big\{\boldsymbol{X}\eta : \eta \in \mathbb{R}^p, \|\hat{\boldsymbol{y}} - \boldsymbol{X}\eta\|^2 \le p F^{-1}_{p,n-p}(1-\alpha)\hat{\sigma}^2(\boldsymbol{y})\big\},$$

where $\hat{\sigma}^2(\boldsymbol{y}) = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2/(n-p)$. The corresponding set $\tilde{C}_\alpha(\boldsymbol{y}) = \big\{\eta \in \mathbb{R}^p : \boldsymbol{X}\eta \in C_\alpha(\boldsymbol{y})\big\}$ is Scheffé's well-known confidence ellipsoid for the unknown parameter $\beta \in \mathbb{R}^p$ such that $\mu = \boldsymbol{X}\beta$.

3.2. Inference on a sparse signal

Suppose that $p = n$, and that the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are linearly independent. In that case, $\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$ is nonsingular, and assuming Eq. (11), the least squares estimator of $\beta := \boldsymbol{X}^{-1}\mu$ is given by $\hat{\beta}(\boldsymbol{y}) := \boldsymbol{X}^{-1}\boldsymbol{y}$. Writing $\boldsymbol{X}^{-1} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]^\top$, the Gauss–Markov

estimator of β_i equals $\hat{\beta}_i(\boldsymbol{y}) = \boldsymbol{a}_i^\top \boldsymbol{y}$. In case of $\boldsymbol{y} \sim N_n(\mu, \sigma^2 \boldsymbol{I})$, it has distribution $N(\beta_i, \|\boldsymbol{a}_i\|^2\sigma^2) = N\big(\beta_i, ((\boldsymbol{X}^\top \boldsymbol{X})^{-1})_{ii}\sigma^2\big)$.

Suppose that $\mu = \boldsymbol{X}\beta$ is sparse in the sense that

$$\|\beta\|_0 := \#\{i \le n : \beta_i \ne 0\}$$

is relatively small compared with n. Then a possible test statistic for the null hypothesis "$\mu = 0$" is given by

$$S_\ell(\boldsymbol{y}) = S_\ell(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) := G_1(\boldsymbol{y})/G_\ell(\boldsymbol{y})$$

for some integer $\ell \in \{2, \ldots, n\}$, where $G_1(\boldsymbol{y}) \ge G_2(\boldsymbol{y}) \ge \cdots \ge G_n(\boldsymbol{y})$ are the values $\|\boldsymbol{a}_1\|^{-1}|\boldsymbol{a}_1^\top \boldsymbol{y}|$, $\|\boldsymbol{a}_2\|^{-1}|\boldsymbol{a}_2^\top \boldsymbol{y}|$, …, $\|\boldsymbol{a}_n\|^{-1}|\boldsymbol{a}_n^\top \boldsymbol{y}|$ in descending order. The idea behind this choice is that $G_\ell(\boldsymbol{y})$ depends mostly on the noise vector $\boldsymbol{\varepsilon}$ if $\|\beta\|_0 < \ell$, while $G_1(\boldsymbol{y})$ is mainly driven by μ in case of $\|\beta\| \gg 0$. Scale-invariance of the test statistic $S_\ell(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ in $\boldsymbol{y}$ is obvious. Since $\boldsymbol{X}^{-1} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, one may write $\|\boldsymbol{a}_i\|^{-2} = (\boldsymbol{X}^\top \boldsymbol{X})_{ii}$ and $\boldsymbol{a}_i^\top \boldsymbol{y} = \big((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}\big)_i$, whence $S_\ell(\cdots)$ is also orthogonally invariant.

Denoting the $(1-\alpha)$-quantile of $S_\ell(\boldsymbol{z})$, $\boldsymbol{z} \sim N_n(0, \boldsymbol{I})$, with $\varkappa_{\ell,\alpha}$ and setting $\mathbb{M} = \mathbb{R}^n$, we obtain the equivalence region

$$C_{\ell,\alpha}(\boldsymbol{y}) := \big\{\boldsymbol{m} \in \mathbb{R}^n : S_\ell(\boldsymbol{y} - \boldsymbol{m}) \le \varkappa_{\ell,\alpha}\big\}.$$

This region is rather useless per se. But if we restrict our attention to vectors $\boldsymbol{m} = \boldsymbol{X}\boldsymbol{b}$, where $\|\boldsymbol{b}\|_0$ is bounded by a given number, we end up with the equivalence regions

$$\tilde{C}_{k,\ell,\alpha}(\boldsymbol{y}) := \big\{\boldsymbol{b} \in \mathbb{R}^n : \|\boldsymbol{b}\|_0 \le k, S_\ell(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \le \varkappa_{\ell,\alpha}\big\}, \quad 1 \le k < n,$$

which are potentially useful in case of $k < \ell$. In this manuscript we only prove a first result about these equivalence regions.

LEMMA 5. *Let*

$$k_{\ell,\alpha}(\boldsymbol{y}) := \ell - 1 - \max\big\{j - i : 1 \le i \le j \le \ell, G_i(\boldsymbol{y})/G_j(\boldsymbol{y}) \le \varkappa_{\ell,\alpha}\big\}.$$

Then, $\tilde{C}_{\alpha,k,\ell}(\boldsymbol{y}) \ne \emptyset$ if and only if $k \ge k_{\ell,\alpha}(\boldsymbol{y})$.

In the model-based context, $k_{\ell,\alpha}$ is a lower $(1-\alpha)$-confidence bound for $\|\beta\|_0$, and $\tilde{C}_{k,\ell,\alpha}$ is a $(1-\alpha)$-confidence region for β under the additional assumption that $\|\beta\|_0 \le k$.

3.3. Special case: the Gaussian sequence model

Suppose that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is the standard basis of $\mathbb{R}^n$. Then $(G_i(\boldsymbol{z}))_{i=1}^n$ has the same distribution as $\big(\tilde{\Phi}^{-1}(U_{(n+1-i)})\big)_{i=1}^n$, where $U_{(1)} < \cdots < U_{(n)}$ are the order statistics of independent random variables $U_1, \ldots, U_n \sim \mathrm{Unif}[0,1]$, and $\tilde{\Phi}$ is the distribution function of $|z_1|$,

that is, $\tilde{\Phi}(r) = 2\Phi(r) - 1$ for $r \geq 0$. Since $U_{(n+1-\ell)} = 1 - \ell/n + O_p(n^{-1/2})$ as $n \to \infty$, a reasonable choice for ℓ seems to be $\ell \approx (1 - \tilde{\Phi}(1))n = 2\Phi(-1)n \approx 0.32\,n$. Then $\varkappa_{\ell,\alpha}$ is approximately the $(1-\alpha)$-quantile of $G_1(z) = \max_{1 \leq i \leq n} |z_i| = \sqrt{2 \log n}(1 + o_p(1))$.

Specifically, let $n = 100$ and $\ell = 32$. Numerical computations (outlined in Section A) yield $\varkappa_{\ell,0.01} = 4.083$.

Now we consider the model-based setting with $\boldsymbol{y} \sim N_n(\mu, \boldsymbol{I})$ and and two different choices for μ. The proof of Lemma 5 includes the construction of a particular vector $\hat{\mu}(\boldsymbol{y}) = \hat{\mu}_{\ell,\alpha}(\boldsymbol{y})$ in $C_{\ell,\alpha}(\boldsymbol{y})$ such that $||\hat{\mu}||_0 = k_{\ell,\alpha}(\boldsymbol{y})$. Now we investigated the distribution of the latter number, of the set $\{i \leq n : \hat{\mu}_i \neq 0\}$ and of the cosine of the angle between μ and $\hat{\mu}$.

In the first scenario, we considered the sparse vector $\mu = (10, -6, 3, 0, \ldots, 0)^\top$. In 100'000 Monte Carlo simulations we estimated the joint distribution of

$$\mathrm{TP}(\boldsymbol{y}) := \big\{i \in \{1,2,3\} : \hat{\mu}_i(\boldsymbol{y}) \neq 0\big\} \quad \text{and} \quad \mathrm{FP}(\boldsymbol{y}) := \#\{i > 3 : \hat{\mu}_i(\boldsymbol{y}) \neq 0\},$$

the *set* of nonzero components of μ which were detected correctly ("true positives") and the *number* of "false positives", respectively. Table 1 contains estimated probabilities rounded to four digits. It turned out that $\mathrm{FP}(\boldsymbol{y}) \geq 2$ with estimated probability less than $10^{-4}/2$, and $\mathrm{FP}(\boldsymbol{y}) \geq 1$ with estimated probability 0.0058. The first component of μ was always identified correctly as non-zero, whereas the second and third component stayed sometimes undetected.

In the second scenario, we considered the vector $\mu = \big((-1)^{i-1} 2^{-(i-1)/2}\big)_{i=1}^n$. Although $||\mu||_0 = n$, the first four to six components of μ contain the main signal, because $||\mu||^{-2} \sum_{i>4} \mu_i^2 = 0.0625$ and $||\mu||^{-2} \sum_{i>6} \mu_i^2 < 0.0157$. Figure 1 shows the (estimated) distribution of $k_{\ell,\alpha}(\boldsymbol{y})$. Recall that the latter is a lower $(1-\alpha)$-confidence bound for $||\beta||_0$.

Finally, Figure 2 show boxplots of the distribution of

$$\cos \angle(\mu, \boldsymbol{y}) \quad \text{and} \quad \cos \angle(\mu, \hat{\mu})$$

for both scenarios. These plots illustrate that the sparse estimator $\hat{\mu}$ captures μ better than the raw data $\boldsymbol{y}$.

TABLE 1
Joint distribution of $\mathrm{TP}(\boldsymbol{y})$ *and* $\mathrm{FP}(\boldsymbol{y})$ *in* 1^{st} *scenario (in percent).*

$\mathrm{TP}(\boldsymbol{y}) =$	$\{1,2,3\}$	$\{1,2\}$	$\{1,3\}$	$\{1\}$	Otherwise	Total
$\mathrm{FP}(\boldsymbol{y}) = 0$	11.13	82.70	0.35	5.24	0.00	99.42
$\mathrm{FP}(\boldsymbol{y}) = 1$	0.12	0.42	0.00	0.04	0.00	0.58
Total	11.25	83.12	0.35	5.28	0.00	100.00

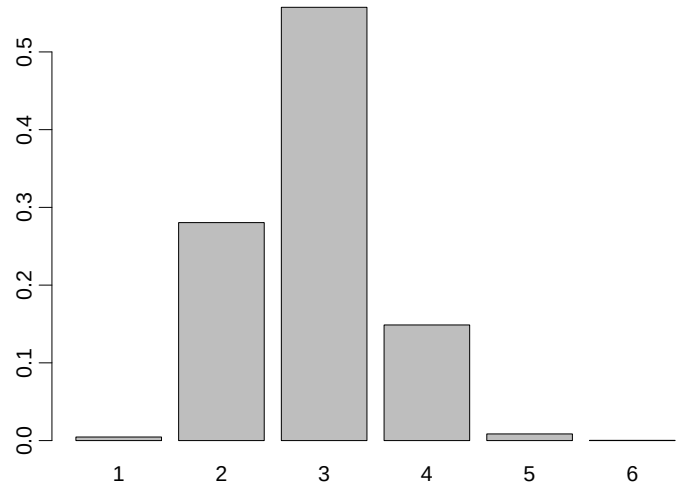

Figure 1 – Distribution of $k_{\ell,\alpha}(\boldsymbol{y})$ in 2nd scenario.

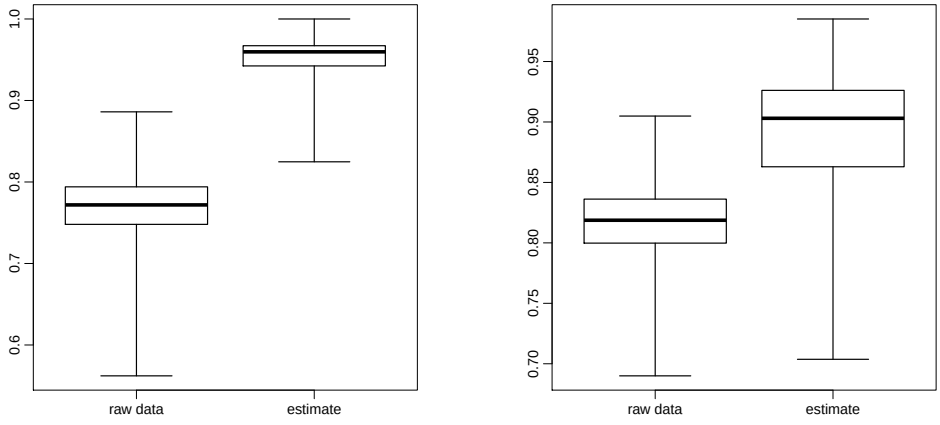

Figure 2 – Distribution of $\cos\angle(\mu, \boldsymbol{y})$ and $\cos\angle(\mu, \hat{\mu})$ in 1st scenario (left) and 2nd scenario (right).

4. Final comments and outlook

Let us relate the considerations in the present paper to research about permutation tests in general regression models. We start from the model-based approach with $y = X\beta + \varepsilon$ with fixed $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$ and a random error $\varepsilon \in \mathbb{R}^n$. When testing the null hypothesis that $(\beta_j)_{p_o < j \le p} = 0$, the goal is to relax the assumption of a spherically symmetric distribution of ε to assuming exchangeability only. That is, for any fixed permutation τ of $\{1, \ldots, n\}$, the distribution of $\tau(\varepsilon)$ coincides with the distribution of ε, where $\tau(v) := (v_{\tau(i)})_{i=1}^n$ for $v \in \mathbb{R}^n$. Indeed, if $p_o = 0$ or $p_o = 1$ and $x_1 = (1)_{i=1}^n$, the null hypothesis can be tested exactly with a permutation test in which the original data (y, X) are compared with $(\tau(y), X)$ for all (or many randomly chosen) permutations τ. In other cases, however, it is not obvious how to perform a valid permutation test. Kennedy (1995) describes potential pitfalls, and Winkler *et al.* (2014) provide a good overview of proposed permutation schemes. Apart from the aforementioned simple setting, all these proposals are justified by asymptotic considerations only. A particularly interesting reference is Freedman and Lane (1983), because they motivated their permutation test also by the wish to interpret the resulting p-values in a model-free way. Their approach works as follows: Let $\hat{\varepsilon}_o := y - \hat{y}_o$ be the residual vector under the null model. Then they propose to compare the F test statistic for the original observations (y, X) with the F test statistic applied to $(\hat{y}_o + \tau(\hat{\varepsilon}_o), X)$. The collection of data $(\hat{y}_o + \tau(\hat{\varepsilon}_o), X)$, where τ is an arbitrary permutation of $\{1, \ldots, n\}$, serves as a data-generated reference set of data to be compared with the original (y, X). If the latter sticks out in terms of the F test statistic, this is viewed as model-free evidence that the regressors x_j, $p_o < j \le p$, are relevant. Freedman and Lane (1983) show that under certain regularity assumptions, the resulting p-value is asymptotically equivalent to the classical p-value (2). Thus our Lemma 1 may be viewed as a computationally simple and exact alternative to the paradigm of Freedman and Lane (1983).

Sections 3.2 and 3.3 describe a model-free approach to variable selection. Since it is computationally intensive, at least for non-orthogonal regressors $x_1, \ldots, x_n$, Davies and Dümbgen (2024) develop an alternative simpler procedure for model-free variable selection. The price to be paid for the simplification is that the selection procedure is possibly conservative in the sense of selecting too few variables. The idea of Gaussian covariate vectors might also be applicable for order selection in autoregressive time series models.

Acknowledgements

The authors are grateful to a reviewer and the editor for constructive comments. Part of this work was supported by Swiss National Science Foundation.

APPENDIX

A. TECHNICAL DETAILS AND PROOFS

A.1. Proof of Lemma 1

Note that $\boldsymbol{y}$ is the sum of the three orthogonal vectors $\hat{\boldsymbol{y}}_o$, $\hat{\boldsymbol{y}}^* - \hat{\boldsymbol{y}}_o$ and $\boldsymbol{y} - \hat{\boldsymbol{y}}^*$. Thus,

$$\frac{||\boldsymbol{y} - \hat{\boldsymbol{y}}^*||^2}{||\boldsymbol{y} - \hat{\boldsymbol{y}}_o||^2} = \frac{||\boldsymbol{y} - \hat{\boldsymbol{y}}_o||^2 - ||\boldsymbol{y} - \hat{\boldsymbol{y}}^*||^2}{||\boldsymbol{y} - \hat{\boldsymbol{y}}_o||^2} = 1 - \frac{||\hat{\boldsymbol{y}}^* - \hat{\boldsymbol{y}}_o||^2}{||\boldsymbol{y} - \hat{\boldsymbol{y}}_o||^2},$$

whence the claim is equivalent to

$$\frac{||\hat{\boldsymbol{y}}^* - \hat{\boldsymbol{y}}_o||^2}{||\boldsymbol{y} - \hat{\boldsymbol{y}}^*||^2} \sim \text{Beta}\big((p - p_o)/2, (n - p)/2\big).$$

Let us first recall two well-known facts about a standard Gaussian random vector $\boldsymbol{z} = (z_i)_{i=1}^n$ in $\mathbb{R}^n$:
(F1) For any fixed matrix $\boldsymbol{B} \in \mathbb{O}_n$, the random vector $\boldsymbol{Bz}$ is standard Gaussian too. Equivalently, for any orthonormal basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ of $\mathbb{R}^n$, the random vector $\sum_{i=1}^n z_i \boldsymbol{b}_i$ is standard Gaussian.
(F2) For any $k \in \{1, \ldots, n-1\}$, the random variable $\sum_{j=1}^k z_j^2 \big/ \sum_{i=1}^n z_i^2$ follows the beta distribution with parameters $k/2$ and $(n-k)/2$.

Step 1: Reduction to the case of $p_o = 0$. Suppose that $p_o \geq 1$. Let Π_o be the orthogonal projection from $\mathbb{R}^n$ onto $\text{span}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p_o})$, so $\hat{\boldsymbol{y}}_o = \Pi_o \boldsymbol{y}$. The vector $\hat{\boldsymbol{y}}^* - \hat{\boldsymbol{y}}_o$ is the orthogonal projection of $\boldsymbol{y} - \Pi_o \boldsymbol{y}$ onto the linear span of the vectors $\boldsymbol{x}_j^* - \Pi_o \boldsymbol{x}_j^*$, $p_o < j \leq p$. All these vectors lie in the $(n - p_o)$-dimensional linear space $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p_o}\}^\perp$, and the independent random vectors $\boldsymbol{x}_j^* - \Pi_o \boldsymbol{x}_j^*$, $p_o < j \leq p$, follow a standard Gaussian distribution on that space. The latter claim follows from (F1) when choosing an orthonormal basis of $\mathbb{R}^n$ such that $n - p_o$ of these basis vectors span $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p_o}\}^\perp$. Hence, we may assume without loss of generality that $p_o = 0$ and $\hat{\boldsymbol{y}}_o = 0$.

Step 2: The case of $p_o = 0$. We argue similarly as Frankl and Maehara (1990). With the random matrix $\boldsymbol{X} := [\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_p^*] \in \mathbb{R}^{n \times p}$, the orthogonal projection $\hat{\boldsymbol{y}}^*$ is given by $\hat{\boldsymbol{y}}^* = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$, and

$$||\hat{\boldsymbol{y}}^*||^2 = \boldsymbol{y}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

The distribution of this random variable depends only on $||\boldsymbol{y}||^2$. Indeed, let $\boldsymbol{B} \in \mathbb{O}_n$ such that $\boldsymbol{By} = ||\boldsymbol{y}|| \boldsymbol{e}$ with $\boldsymbol{e} = (1, 0, \ldots, 0)^\top$. Since $\tilde{\boldsymbol{X}} := \boldsymbol{BX}$ has the same distribution as $\boldsymbol{X}$, we may conclude that

$$||\hat{\boldsymbol{y}}^*||^2 = ||\boldsymbol{y}||^2 \boldsymbol{e}^\top \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{e} \overset{d}{=} ||\boldsymbol{y}||^2 \boldsymbol{e}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{e}.$$

But then we may replace $\boldsymbol{y}$ with $||\boldsymbol{y}||||\boldsymbol{z}||^{-1}\boldsymbol{z}$ with a standard Gaussian random vector $\boldsymbol{z}$ which is independent from $\boldsymbol{X}$. Conditional on $\boldsymbol{X}$, this vector $\boldsymbol{z}$ has the same distribution as $\sum_{i=1}^n z_i \boldsymbol{b}_i$ with an orthonormal basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ of $\mathbb{R}^n$ such that $\mathrm{span}(\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_p^*) = \mathrm{span}(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)$, and the orthogonal projection of $\boldsymbol{z}$ onto the latter space equals $\sum_{j=1}^p z_j \boldsymbol{b}_j$. Consequently, conditional on $\boldsymbol{X}$,

$$\frac{||\hat{\boldsymbol{y}}^*||^2}{||\boldsymbol{y}||^2} \stackrel{d}{=} \Big\|\sum_{i=1}^n z_i \boldsymbol{b}_i\Big\|^{-2} \Big\|\sum_{j=1}^p z_j \boldsymbol{b}_j\Big\|^2 = \sum_{j=1}^p z_j^2 \Big/ \sum_{i=1}^n z_i^2$$

follows a beta distribution with parameters $p/2$ and $(n-p)/2$. Since this does not depend on $\boldsymbol{X}$, we may conclude that the ratio $||\hat{\boldsymbol{y}}^*||^2/||\boldsymbol{y}||^2$ has distribution $\mathrm{Beta}(p/2, (n-p)/2)$. □

A.2. Haar measure on $\mathbb{O}_n$ in a nutshell

Recall that Haar_n is the unique probability measure on $\mathbb{O}_n$ such that a random matrix $\boldsymbol{H} \sim \mathrm{Haar}_n$ satisfies $\boldsymbol{T}\boldsymbol{H} \stackrel{d}{=} \boldsymbol{H}$ for any fixed $\boldsymbol{T} \in \mathbb{O}_n$ (left-invariance). For the reader's convenience, we explain a few well-known basic facts here.

Existence. To construct such a random matrix explicitly, let $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$ be a random matrix such that $\boldsymbol{T}\boldsymbol{Z} \stackrel{d}{=} \boldsymbol{Z}$ for any fixed $\boldsymbol{T} \in \mathbb{O}_n$, and $\mathrm{rank}(\boldsymbol{Z}) = n$ almost surely. This is true, for instance, if the n^2 entries of $\boldsymbol{Z}$ are independent with distribution $N(0,1)$. Then one can easily verify that $\boldsymbol{H} := \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1/2}$ is a random orthogonal matrix with left-invariant distribution. Another possible construction would be to apply Gram-Schmidt orthogonalization to the columns $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n$ of $\boldsymbol{Z}$. This leads to a random matrix $\boldsymbol{H} \in \mathbb{O}_n$ with left-invariant distribution, because the coefficients for the Gram–Schmidt procedure involve only the inner products $\boldsymbol{z}_i^\top \boldsymbol{z}_j$, and these remain unchanged if $\boldsymbol{Z}$ is replaced with $\boldsymbol{T}\boldsymbol{Z}$ for some $\boldsymbol{T} \in \mathbb{O}_n$.

Inversion-invariance and uniqueness. Let $\boldsymbol{J}, \boldsymbol{H}$ be independent random matrices with left-invariant distribution on $\mathbb{O}_n$. By conditioning on $\boldsymbol{J}$ one can show that $\boldsymbol{J}^\top\boldsymbol{H} \stackrel{d}{=} \boldsymbol{H}$, and conditioning on $\boldsymbol{H}$ reveals that $\boldsymbol{J}^\top\boldsymbol{H} = (\boldsymbol{H}^\top\boldsymbol{J})^\top \stackrel{d}{=} \boldsymbol{J}^\top$. Hence, $\boldsymbol{J}^\top \stackrel{d}{=} \boldsymbol{H}$. In the special case that $\boldsymbol{J}$ is an independent copy of $\boldsymbol{H}$, this shows that

$$\boldsymbol{H}^\top \stackrel{d}{=} \boldsymbol{H}.$$

And then, with the original $\boldsymbol{J}$, we see that $\boldsymbol{J} = (\boldsymbol{J}^\top)^\top \stackrel{d}{=} \boldsymbol{H}^\top \stackrel{d}{=} \boldsymbol{H}$, i.e.

$$\boldsymbol{J} \stackrel{d}{=} \boldsymbol{H}.$$

Right-invariance. If $\boldsymbol{H} \sim \text{Haar}_n$, then its distribution is right-invariant in the sense that $\boldsymbol{HT} \stackrel{d}{=} \boldsymbol{H}$ for any fixed $\boldsymbol{T} \in \mathbb{O}_n$. This follows easily from left-invariance and inversion-invariance.

Random linear subspaces of $\mathbb{R}^n$. The second construction of $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n]$ via Gram–Schmidt may be applied to a random matrix $\boldsymbol{Z}$ with independent columns $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n \sim N_n(\mathbf{0}, \boldsymbol{I})$. This shows that the first column of $\boldsymbol{H}$ has distribution $\text{Unif}(\mathbb{S}_n)$. Combined with right-invariance, applied to permutation matrices, this shows that any column of $\boldsymbol{H}$ has distribution $\text{Unif}(\mathbb{S}_n)$. Furthermore, for $k \in \{1, 2, \ldots, n\}$ and arbitrary linearly independent vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^n$,

$$\text{span}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k) \stackrel{d}{=} \text{span}(\boldsymbol{H}_1, \ldots, \boldsymbol{H}_k) \stackrel{d}{=} \text{span}(\boldsymbol{Hx}_1, \ldots, \boldsymbol{Hx}_k).$$

The first equality follows from the construction of $\boldsymbol{H}$. For the second one, let $\boldsymbol{B} \in \mathbb{R}^{k \times k}$ be a nonsingular matrix such that the columns of $[\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k] = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k]\boldsymbol{B}$ are orthonormal. Extending them to a orthonomal basis $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n$ of $\mathbb{R}^n$,

$$\text{span}(\boldsymbol{Hx}_1, \ldots, \boldsymbol{Hx}_k) = \boldsymbol{H}[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k]\mathbb{R}^k = \boldsymbol{H}[\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k]\mathbb{R}^k,$$

and this is the linear span of the first k columns of $\boldsymbol{H}[\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n]$. But by right-invariance, the latter matrix has the same distribution as $\boldsymbol{H}$, because $[\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n] \in \mathbb{O}_n$.

A.3. Equivalence of the null hypotheses H_0, H_0' and H_0''

Recall that we consider a random vector $\boldsymbol{y}$ with continuous distribution on $\mathbb{R}^d$. Suppose first that H_0 is true. Then $\boldsymbol{y}$ has the same distribution as $||\boldsymbol{y}||||\boldsymbol{z}||^{-1}\boldsymbol{z}$, where $\boldsymbol{y}$ and $\boldsymbol{z} \sim N_n(\mathbf{0}, \boldsymbol{I})$ are independent. For any fixed $\boldsymbol{T} \in \mathbb{O}_n$, orthogonal invariance of the standard Gaussian distribution implies that $\boldsymbol{Tz} \stackrel{d}{=} \boldsymbol{z}$, so

$$\boldsymbol{Ty} \stackrel{d}{=} ||\boldsymbol{y}||||\boldsymbol{z}||^{-1}\boldsymbol{Tz} = ||\boldsymbol{y}||||\boldsymbol{Tz}||^{-1}\boldsymbol{Tz} \stackrel{d}{=} ||\boldsymbol{y}||||\boldsymbol{z}||^{-1}\boldsymbol{z} \stackrel{d}{=} \boldsymbol{y}.$$

Consequently, H_0' is true as well.

Now suppose that H_0' is true. If $\boldsymbol{H} \sim \text{Haar}_n$ and $\boldsymbol{y}$ are independent, then for any measurable set $B \subset \mathbb{R}^n$,

$$\mathbb{P}(\boldsymbol{Hy} \in B) = \mathbb{E}\,\mathbb{P}(\boldsymbol{Hy} \in B \,|\, \boldsymbol{H}) = \mathbb{E}\,\mathbb{P}(\boldsymbol{y} \in B \,|\, \boldsymbol{H}) = \mathbb{P}(\boldsymbol{y} \in B),$$

where $\mathbb{P}(\boldsymbol{Hy} \in B \,|\, \boldsymbol{H}) = \mathbb{P}(\boldsymbol{Y} \in B \,|\, \boldsymbol{H})$ by H_0'.

Finally, suppose that H_0'' is true. Then for any measurable set $B \subset \mathbb{R}^n$,

$$\begin{aligned}\mathbb{P}(\boldsymbol{y} \in B) = \mathbb{P}(\boldsymbol{Hy} \in B) &= \mathbb{E}\,\mathbb{P}(\boldsymbol{Hy} \in B \,|\, \boldsymbol{y}) \\ &= \mathbb{E}\,\mathbb{P}\big(||\boldsymbol{y}||\boldsymbol{u} \in B \,\big|\, \boldsymbol{y}\big) = \mathbb{P}\big(||\boldsymbol{y}||\boldsymbol{u} \in B\big),\end{aligned}$$

where $\boldsymbol{u} \sim \text{Unif}(\mathbb{S}_n)$ and $\boldsymbol{y}$ are independent. Thus H_0 is satisfied as well.

A.4. Proof of Lemma 5

We first construct a particular point $\hat\beta = \hat\beta_{\ell,\alpha}(y) \in \mathbb{R}^n$ such that $S_\ell(y - X\hat\beta) \le \varkappa_{\ell,\alpha}$ and $\|\hat\beta\|_0 = k_{\ell,\alpha}(y)$. To this end let $(\sigma(1), \sigma(2), \ldots, \sigma(n))$ be a permutation of $(1, 2, \ldots, n)$ such that $G_i(y) = \|a_{\sigma(i)}\|^{-1} |a_{\sigma(i)}^\top y|$ for $1 \le i \le n$. Let $k_{\ell,\alpha}(y) = \ell - 1 - j_o + i_o$ with indices $1 \le i_o \le j_o \le \ell$ such that $|G_{i_o}(y)|/|G_{j_o}(y)| \le \varkappa_{\ell,\alpha}$. Now we set

$$\hat\beta_{\sigma(i)} := \begin{cases} 0 & \text{if } i_o \le i \le j_o \text{ or } i > \ell, \\ \operatorname{sign}(a_{\sigma(i)}^\top y)\big(|a_{\sigma(i)}^\top y| - \|a_{\sigma(i)}\| G_{i_o}(y)\big) & \text{if } 1 \le i < i_o, \\ \operatorname{sign}(a_{\sigma(i)}^\top y)\big(|a_{\sigma(i)}^\top y| - \|a_{\sigma(i)}\| G_{j_o}(y)\big) & \text{if } j_o < i \le \ell. \end{cases}$$

Obviously, this defines a vector $\hat\beta \in \mathbb{R}^n$ such that $\|\hat\beta\|_0 = k_{\ell,\alpha}(y)$. Note also that the numbers $\|a_{\sigma(i)}\|^{-1} |a_{\sigma(i)}^\top (y - X\hat\beta)|$ are nonincreasing in $i \in \{1, \ldots, n\}$ with

$$\|a_{\sigma(1)}\|^{-1} |a_{\sigma(1)}^\top (y - X\hat\beta)| \;=\; G_{i_o}(y), \quad \|a_{\sigma(\ell)}\|^{-1} |a_{\sigma(\ell)}^\top (y - X\hat\beta)| \;=\; G_{j_o}(y).$$

Consequently, $\hat\beta \in \tilde C_{k,\ell,\alpha}(y)$ for $k \ge k_{\ell,\alpha}(y)$.

On the other hand, let $k_{\ell,\alpha}(y) > 0$. We have to show that $S_\ell(y - b) > \varkappa_{\ell,\alpha}(y)$ for any $b \in \mathbb{R}^n$ with $k := \|b\|_0 < k_{\ell,\alpha}(y)$. Then there exist indices $1 \le i(1) < i(2) < \cdots < i(n-k) \le n$ and $1 \le j(1) < j(2) < \cdots < j(n-k) \le n$ such that $G_{i(s)}(y) = G_{j(s)}(y - Xb)$ for $1 \le s \le n - k$. But at least $n - k - (n - \ell) = \ell - k$ of the indices $j(1), \ldots, j(n-k)$ are contained in $\{1, \ldots, \ell\}$. Consequently, if $S_\ell(y - Xb) \le \varkappa_{\ell,\alpha}$, then

$$\begin{aligned} \varkappa_{\ell,\alpha} &\ge G_1(y - Xb)/G_\ell(y - Xb) \\ &\ge G_{j(1)}(y - Xb)/G_{j(\ell-k)}(y - Xb) \\ &= G_{i(1)}(y)/G_{i(\ell-k)}(y), \end{aligned}$$

whence $k_{\ell,\alpha}(y) \le \ell - 1 - (i(\ell - k) - i(1)) \le \ell - 1 - (\ell - k - 1) = k$, a contradiction to $k < k_{\ell,\alpha}(y)$. □

A.5. Technical details for Section 3.3

For any $\varkappa \ge 1$,

$$\begin{aligned} \mathbb{P}\big(S_\ell(z) \le \varkappa\big) &= \mathbb{P}\big(\tilde\Phi^{-1}(U_{(n)}) \le \varkappa \tilde\Phi^{-1}(U_{(n+1-\ell)})\big) \\ &= \mathbb{E}\,\mathbb{P}\big(U_{(n)} \le \tilde\Phi[\varkappa \tilde\Phi^{-1}(U_{(n+1-\ell)})] \,\big|\, U_{(n+1-\ell)}\big) \\ &= \mathbb{E}\Big(\Big(\frac{\tilde\Phi[\varkappa \tilde\Phi^{-1}(U_{(n+1-\ell)})] - U_{(n+1-\ell)}}{1 - U_{(n+1-\ell)}}\Big)^{\ell-1}\Big), \end{aligned}$$

where the last step follows from the fact that conditionally on $U_{(n+1-\ell)}$, the random variable $U_{(n)}$ has the same distribution as the maximum of $\ell-1$ independent random variables with uniform distirbution on $[U_{(n+1-\ell)},1]$. Since $U_{(n+1-\ell)}$ has distribution Beta$(n+1-\ell,\ell)$, the latter expectation may be expressed as

$$\int_0^1 \Big(\frac{\tilde{\Phi}[\varkappa\tilde{\Phi}^{-1}(B^{-1}(u))]-B^{-1}(u)}{1-B^{-1}(u)}\Big)^{\ell-1} du,$$

where B^{-1} is the quantile function of Beta$(n+1-\ell,\ell)$. This integral can be computed numerically, and the quantile $\varkappa_{\ell,\alpha}$ can be found via bisection.

References

L. Davies, L. Dümbgen (2024). *Gaussian covariates and linear regression*. Preprint in preparation.

M. L. Eaton (1983). *Multivariate Statistics: a Vector Space Approach*. Wiley Series in Probability and Statistics, Wiley and Sons, Chichester.

M. L. Eaton (1989). *Group invariance applications in statistics*. NSF-CBMS Regional Conference Series in Probability and Statistics, 1. Institute of Mathematical Statistics, Hayward, CA; American Statistical Association, Alexandria, VA.

P. Frankl, H. Maehara (1990). *Some geometric applications of the beta distribution*. Annals of the Institute of Statistical Mathematics, 42, no. 3, pp. 463–474.

D. Freedman, D. Lane (1983). *A nonstochastic interpretation of reported significance levels*. Journal of Business & Economic Statistics, 1, no. 4, pp. 292–298.

F. E. Kennedy (1995). *Randomization tests in econometrics*. Journal of Business & Economic Statistics, 13, no. 1, pp. 85–94.

K. Mardia, J. Kent, J. Bibby (1979). *Multivariate Analysis*. Academic Press, London.

R. G. Miller, Jr. (1981). *Simultaneous statistical inference, Second ed.* Springer-Verlag, New York.

H. Scheffé (1959). *Analysis of Variance*. John Wiley and Sons, New York.

A. M. Winkler, G. R. Ridgway, M. A. Webster, S. M. Smith, T. E. Nichols (2014). *Permutation inference for the general linear model*. NeuroImage, 92, pp. 381–397.