A model-free approach to linear least squares regression with exact probabilities

Laurie Davies and Lutz Dümbgen
(University of Duisburg-Essen and University of Bern)

July 26, 2018

Abstract

In a regression setting with observation vector $y \in \mathbb{R}^n$ and given finite collection $(x_\nu)_{\nu \in N}$ of regressor vectors $x_\nu \in \mathbb{R}^n$, a typical question is whether a given subset of these regressors is sufficient to approximate $y$. A classical method for this question is the $F$ test, assuming that $y$ is a linear combination of the regressor vectors plus Gaussian white noise. In this note we show that the corresponding p-value has also a clear data-scientific interpretation without having to assume the data to be random. Then it is shown that such a dual interpretation is possible for a rather large family of tests, the underlying tool being normalized Haar measure on orthogonal groups.

1 Introduction

Let $y \in \mathbb{R}^n$ be an observation vector, and let $x_\nu \in \mathbb{R}^n$, $\nu \in N_*$, be a given finite collection of regressor vectors. The question is how well $y$ may be approximated by linear combinations of these regressors.

Specifically, suppose the raw data are given by a data matrix with $n$ rows

$$[y_i, z_1^\top] = [y_i, z_{i1}, \ldots, z_{id}], \quad 1 \leq i \leq n,$$

containing the values of a response and $d$ numerical covariables for each observation. Then the usual multiple linear regression model would consider the regressor vectors $x_0 := (1)_{i=1}^n$ and $x_j := (z_{ij})_{i=1}^n$, $1 \leq j \leq d$. More complex models would also include the $\binom{d}{2}$ interaction vectors $x_{j,k} := (x_{ij}x_{ik})_{i=1}^n$, $1 \leq j < k \leq d$. In general, with arbitrary types of covariables, one could think of $x_\nu = (f_\nu(z_i))_{i=1}^n$ with given basis functions $f_\nu, \nu \in N_*$.

For any subset $N$ of $N_*$ consider the linear space $\mathbb{V}_N := \text{span}(x_\nu : \nu \in N)$ and the orthogonal projection

$$\hat{y}_N := \arg \min_{\eta \in \mathbb{V}_N} \|y - \eta\|^2$$

of $y$ onto $\mathbb{V}_N$. This includes the case of $N = \emptyset$ with $\mathbb{V}_\emptyset := \{0\}$ and $\hat{y}_\emptyset := 0$. If $\hat{y}_N$ is viewed as an approximation of $y$, a raw measure of approximation error would be the sum of squared
residuals

$$SS_N := \|y - \hat{y}_N\|^2 = \|y\|^2 - \|\hat{y}_N\|^2.$$  

In case of $\forall_N \neq \forall_{N*}$, a common question in linear regression is whether the approximation of $y$ by $\hat{y}_N$, is “substantially better” than the one by $\hat{y}_{N*}$. Of course it follows from $\forall_N \subset \forall_{N*}$ that $SS_N \geq SS_{N*}$, so the question is whether the ratio $SS_{N*}/SS_N$ is “significantly small”. Let us first recall the classical answer as presented in standard textbooks, e.g. [4].

**Classical approach: Gaussian model and F test.** Suppose that the regressor vectors $x_{\nu}$, $\nu \in N_*$, are fixed and linearly independent with $0 \leq p := \#N < p_* := \#N_* < n$. (In case of random regressors, we consider conditional distributions given $(x_{\nu})_{\nu \in N_*}$.) Suppose that

$$y = \sum_{\nu \in N_*} \theta_{\nu} x_{\nu} + \epsilon$$

with unknown parameters $\theta_{\nu}$, $\nu \in N_*$, and a random vector

$$\epsilon \sim N_n(0, \sigma^2 I_n),$$

the standard deviation $\sigma > 0$ being unknown as well. Here $I_n$ denotes the identity matrix in $\mathbb{R}^{n \times n}$. Under the null hypothesis that

$$\theta_{\nu} = 0 \quad \text{for} \quad \nu \in N_* \setminus N,$$  

the random variables $SS_{N_*}$ and $SS_N - SS_{N_*}$ are stochastically independent with

$$\frac{SS_{N_*}}{\sigma^2} \sim \chi^2_{n-p_*} \quad \text{and} \quad \frac{SS_N - SS_{N_*}}{\sigma^2} \sim \chi^2_{p_* - p}.$$  

With the F test statistic

$$F := \frac{(SS_N - SS_{N_*})/(p_* - p)}{SS_{N_*}/(n - p_*)},$$

a corresponding p-value of the null hypothesis [1] is given by

$$1 - F_{p_* - p, n - p_*}(F),$$  

where $F_{k,\ell}$ denotes the distribution function of Fisher’s F distribution with $k$ and $\ell$ degrees of freedom.

**Outline of this note.** In Section 2 we present a new interpretation of the p-value (2). Instead of viewing $y$ as a random vector with a rather specific distribution, we consider all vectors $y$ and $x_{\nu}$, $\nu \in N_*$, as fixed. Then we compare the decrease $SS_N - SS_{N_*}$ with the random decrease which would result if we replaced the $p_* - p$ vectors $x_{\nu}$, $\nu \in N_* \setminus N$, with independent random vectors
\[ z_\nu \sim \mathcal{N}_n(0, I_n), \nu \in N_\ast \setminus N. \] It turns out that the probability of this random decrease being greater than or equal to the actual decrease \( SS_N - SS_N^\ast \) is precisely the p-value (2). This provides a purely data-driven interpretation of this p-value.

The technical arguments for Section 2 are rather direct in the sense that we rely only on basic properties of standard Gaussian, beta and gamma distributions. There is, however, a more abstract approach based on Haar distributions on orthogonal groups. A good introduction to that topic can be found, for instance, in the monograph [3]. It is shown in Section 3 that both the classical and the new interpretation of the p-value (2) are a consequence of a basic invariance consideration. This viewpoint allows us to weaken our assumptions on the random noise vector \( \epsilon \) in the classical setup and on the artificial random regressor vectors \( z_\nu \) in the new approach. It also shows that several other tests, some of which are applicable for high-dimensional settings with \( p_\ast \geq n \), have a purely data-driven interpretation, too.

Technical details and proofs are deferred to Section 4.

Based on the findings of Section 2 there is a simple stepwise procedure for selecting covariates which outperforms the lasso and knockoff procedures introduced in [5] and [1], respectively. As an example, consider the well-known Boston housing data with \( n = 504 \) observations and \( d = 13 \) covariables. A linear model with interactions of order at most seven gives a data set with \( (n, p_\ast) = (504, 77520) \). This is much too large for knockoff which exits with the error message “cannot allocate a vector of the size 44.8 GB”. Ten repetitions of lasso gave between 4 and 116 selected regressors with a mean of 58. The time for each selection was about 100 seconds. The selection method based on Section 2 selects 10 regressors in less than two seconds. It also selects 10 regressors with interactions of order at most eight giving \( (n, p_\ast) = (504, 203490) \). The time required was five seconds. Interactions of order at least nine exceed the memory capacity of the laptop. A detailed description of the method and comparison with lasso and knockoff is given in [2].

2 A model-free interpretation of the F test

Rephrasing the p-value (2). In view of the subsequent considerations it is useful to rewrite the p-value (2) in terms of beta distribution functions. Let \( B_{a,b} \) be the distribution function of the beta distribution with parameters \( a, b > 0 \). Then

\[
F_{p_\ast-p,n-p_\ast}(F) = 1 - B_{(p_\ast-p)/2,(n-p_\ast)/2}\left(\frac{SS_N - SS_N^\ast}{SS_N}\right),
\]

(3)

\[
= B_{(n-p_\ast)/2,(p_\ast-p)/2}\left(\frac{SS_N^\ast}{SS_N}\right),
\]


see Section 4.

The new interpretation. As mentioned in the introduction, we now consider the data $y$ and $x_\nu$ as fixed vectors. To judge whether $\hat{y}_{N_*}$ is substantially better than $\hat{y}_N$, we compare the resulting reduction $SS_{N_*}/SS_N$ in the sum of squared residuals with the reduction one would obtain if $(x_\nu)_{\nu \in N_* \setminus N}$, would be replaced with pure white noise.

**Theorem 1.** Suppose that the $p + 1$ vectors $y$ and $x_\nu$, $\nu \in N$, are fixed and linearly independent, and let $p_* < n$. Suppose we replace the regressors $x_\nu$, $\nu \in N_* \setminus N$, with independent random vectors $z_\nu \sim N_n(0, I_n)$, $\nu \in N_* \setminus N$. Then $SS_{N_*}$ becomes a random variable such that

$$\frac{SS_{N_*}}{SS_N} \sim \text{Beta}\left(\frac{(n-p_*)/2}{(p_*-p)/2}\right).$$

Consequently, if $y$ and $x_\nu$, $\nu \in N_*$, are linearly independent and viewed as fixed vectors, then the p-value

$$B_{(n-p_*)/2,(p_*-p)/2}\left(\frac{SS_{N_*}}{SS_N}\right),$$

which is precisely the p-value (2), quantifies how extraordinary the reduction in the sum of squared residuals really is, without referring to a statistical model for $y$.

**Remark.** The artificial regressor tuple $(z_\nu)_{\nu \in N_* \setminus N}$ in Theorem 1 need not be pure white noise. The proof of Theorem 1 and the considerations in the next section reveal that the following property is sufficient: With the orthogonal projection $\Pi$ from $\mathbb{R}^n$ onto $V_{N_*}^\perp$, consider the random linear space

$$M := \text{span}(\Pi z_\nu : \nu \in N_* \setminus N).$$

Then $\dim(M) = p_* - p$ almost surely, and

$$\mathcal{L}(SM) = \mathcal{L}(M)$$

for any fixed orthogonal matrix $S \in \mathbb{R}^{n \times n}$ such that $Sv = v$ for all $v \in V_N$.

### 3 Further considerations in terms of orthogonal invariance

Let us first introduce some notation and recall some concepts from measure theory and algebra: With $O_n$ we denote the set of all orthogonal matrices $S \in \mathbb{R}^{n \times n}$. For a linear subspace $V$ of $\mathbb{R}^n$ with $q := \dim(V) < n$ let

$$O_n(V) := \{ S \in O_n : Sv = v \text{ for all } v \in V \}.$$
If \( b_1, \ldots, b_n \) is an orthonormal basis of \( \mathbb{R}^n \) such that \( \mathbb{V} = \operatorname{span}(b_i : 1 \leq i \leq q) \), then \( S \in O_n(\mathbb{V}) \) may be represented as

\[
S = B \begin{bmatrix} I_q & 0_{q \times (n-q)} \\ 0_{(n-q) \times q} & S_o \end{bmatrix} B^T
\]

with \( B := [b_1, \ldots, b_n] \) and an orthogonal matrix \( S_o \in \mathbb{R}^{(n-q) \times (n-q)} \).

Normalized Haar measure on \( O_n(\mathbb{V}) \), denoted by \( \text{Haar}_{n, \mathbb{V}} \), is the unique probability distribution on \( O_n(\mathbb{V}) \) such that a random variable \( T \sim \text{Haar}_{n, \mathbb{V}} \) satisfies

\[
\mathcal{L}(ST) = \mathcal{L}(T) \quad \text{for any fixed } S \in O_n(\mathbb{V}).
\]

The latter property also implies that

\[
\mathcal{L}(T^\top) = \mathcal{L}(T) \quad \text{and} \quad \mathcal{L}(TS) = \mathcal{L}(T) \quad \text{for any fixed } S \in O_n(\mathbb{V}).
\]

Moreover, for any fixed vector \( x = v + w \) with \( v \in \mathbb{V} \) and \( w \in \mathbb{V}^\perp \), the random vector \( Tx \) has the same distribution as

\[
v + \|w\|u
\]

where \( u \) is uniformly distributed on the unit sphere of \( \mathbb{V}_N^\perp \). Specifically, if \( b_{q+1}, \ldots, b_n \) is an orthonormal basis of \( \mathbb{V}_N^\perp \), then \( u \) is distributed as

\[
\left( \sum_{i=q+1}^{n} Z_i^2 \right)^{-1} \sum_{i=q+1}^{n} Z_i b_i
\]

with independent random variables \( Z_i \sim \mathcal{N}(0, 1) \), \( q < i \leq n \).

### 3.1 A generalization of the classical setting

Throughout this and the next subsection we consider a fixed subset \( N \) of \( N_* \) with \( p < n - 1 \) elements such that the vectors \( x_\nu, \nu \in N \), are linearly independent. We write \( \hat{y} := \hat{y}_N \) and consider the residual vector

\[
\hat{\varepsilon} := y - \hat{y},
\]

i.e. the orthogonal projection of \( y \) onto \( \mathbb{V}_N^\perp \). In the classical setting, \( y \) is viewed as a random vector. The next lemma specifies a null hypothesis which is appropriate for the F test as well as several other tests.

**Lemma 2** (A null hypothesis \( H_N \)). The following three statements about the distribution of \( y \) are equivalent:

(i) For any fixed \( S \in O_n(\mathbb{V}_N) \),

\[
\mathcal{L}(Sy) = \mathcal{L}(y).
\]
(ii) Let $T$ be a random matrix with distribution $\text{Haar}_{n, V_N}$ such that $y$ and $T$ are stochastically independent. Then
\[
\mathcal{L}(Ty) = \mathcal{L}(y).
\]

(iii) Let $u$ be a random vector with uniform distribution on the unit sphere of $V_N^\perp$ such that $y$ and $u$ are stochastically independent. Then
\[
\mathcal{L}(y) = \mathcal{L}(\hat{y} + \|\hat{\epsilon}\|u).
\]

The null hypothesis $H_N$ described in Lemma 2 is satisfied, for instance, if
\[
y = \mu + \epsilon
\]
with fixed vector $\mu \in V_N$ and a random vector $\epsilon$ with orthogonally invariant distribution in the sense that
\[
\mathcal{L}(Su) = \mathcal{L}(u) \quad \text{for any fixed } S \in O_n.
\]

A general test. Let $\tau : \mathbb{R}^n \to \mathbb{R}$ be an arbitrary test statistic. Then a p-value for the null hypothesis $H_N$ specified in Lemma 2 is given by
\[
\pi(y) := \mathbb{P}(\tau(Ty) \geq \tau(y) \mid y) = \mathbb{P}(\tau(\hat{y} + \|\hat{\epsilon}\|u) \geq \tau(y) \mid y)
\]
with $T$ and $u$ as in Lemma 2.

Example: F test. Suppose that the vectors $x_\nu, \nu \in N_*$, are linearly independent with $p < p_* = \#N_* < n$. Let $b_1, \ldots, b_n$ be an orthonormal basis of $\mathbb{R}^n$ such that
\[
\text{span}(b_1, \ldots, b_p) = V_N \quad \text{and} \quad \text{span}(b_1, \ldots, b_{p_*}) = V_{N_*}.
\]
Then the F test statistic $F$ may be written as $F = \tau(y)$ with
\[
\tau(y) = \frac{\sum_{i=p+1}^{p_*} (b_i^\top y)^2 / (p_* - p)}{\sum_{i=p_*+1}^{n} (b_i^\top y)^2 / (n - p_*)}
\]
and the convention $0/0 := 0$. Now let’s replace $y$ with $\hat{y} + \|\hat{\epsilon}\|u$, where
\[
u = \mathcal{L} \left( \sum_{i=p+1}^{n} Z_i^2 \right)^{-1/2} \sum_{i=p+1}^{n} Z_i b_i
\]
with independent random variables $Z_i \sim \mathcal{N}(0, 1)$, independent from $y$. Then we obtain
\[
\tau(\hat{y} + \|\hat{\epsilon}\|u) = 1_{\hat{\epsilon} \neq 0} \frac{\sum_{i=p+1}^{p_*} Z_i^2 / (p_* - p)}{\sum_{j=p_*+1}^{n} Z_j^2 / (n - p_*)},
\]
and the latter fraction follows $F_{p_*-p,n-p_*}$. Hence the p-value (4) coincides with (2).
**Example: Multiple T test.** Suppose that $\mathbb{V}_N \neq \mathbb{R}^n$. Further suppose that the vectors $x_\nu$, $\nu \in N_s \setminus N$, have been standardized to be orthogonal to $\mathbb{V}_N$ and have unit length. Then a possible test statistic which is similar in spirit to Tukey’s studentized range statistic is given by

$$
\tau(y) := \max_{\nu \in N_s \setminus N} \frac{|x_\nu^\top y|}{\text{SS}_N^{1/2}}
$$

Note that the vectors $x_\nu$, $\nu \in N_s \setminus N$, need not be linearly independent.

**Example: Multiple F test.** If $\mathbb{V}_N = \mathbb{R}^n$ or if $n - \dim(\mathbb{V}_N)$ is rather small, one could think about a finite collection $(M_\lambda)_{\lambda \in \Lambda}$ of subsets of $N_s \setminus N$ all of which satisfy $\dim(\mathbb{V}_N \cup M_\lambda) = p + \#M_\lambda \ll n$. Then one could consider the test statistic

$$
\max_{\lambda \in \Lambda} F_\lambda
$$

where $F_\lambda$ is defined as $F$ with $N \cup M_\lambda$ in place of $N_s$. The idea behind this test statistic is that possibly $y = \sum_{\nu \in N_s} \theta_\nu x_\nu + \epsilon$ with a random vector $\epsilon$ having orthogonally invariant distribution and fixed real parameters $\theta_\nu$, $\nu \in N_s$, such that

$$
\sum_{\nu \in M_\lambda} \theta_\nu^2 \gg \sum_{\nu \in N_s \setminus (N \cup M_\lambda)} \theta_\nu^2
$$

for some $\lambda \in \Lambda$.

### 3.2 A model-free interpretation of the p-value

Again we consider the data $y$ and $(x_\nu)_{\nu \in N_s}$ as fixed. All examples of the test statistic $\tau$ may be written as

$$
\tau(y) = \tau(y, (x_\nu)_{\nu \in N_s}),
$$

and one can verify in each case that the latter value depends only on the inner products

$$
y^\top y, \quad x_\nu^\top y \quad \text{and} \quad x_\nu^\top x_\omega
$$

for certain $\nu, \omega \in N_s$. For instance, if $X \in \mathbb{R}^{n \times q}$ contains linearly independent regressors $x_\nu$, $\nu \in M$, then

$$
\text{SS}_M = y^\top y - y^\top X(X^\top X)^{-1}X^\top y.
$$

Now let $T$ be a random matrix with distribution Haar$_{n, N_s}$. Since

$$
(Ty)^\top (Ty) = y^\top y, \quad (Ty)^\top x_\nu = y^\top (T^\top x_\nu) \quad \text{and} \quad x_\nu^\top x_\omega = (T^\top x_\nu)^\top (T^\top x_\omega)
$$
for arbitrary $\nu, \omega \in N$, and since $L(T^\top) = L(T)$, we may rewrite the p-value (4) as

$$\pi(y) = \mathbb{P}(\tau(y,(T^\top x_\nu)_{\nu \in N_*}) \geq \tau(y,(x_\nu)_{\nu \in N_*})).$$

In other words, the p-value (4) results from comparing the relation between $y$ and $(x_\nu)_{\nu \in N_*}$ with the relation between $y$ and the randomized regressor tuple $(T^\top x_\nu)_{\nu \in N_*}$. Note that

$$T x_\nu = x_\nu \text{ for } \nu \in N$$

and

$$(T x_\nu)^\top (T x_\omega) = x_\nu^\top x_\omega \text{ for } \nu, \omega \in N_*.$$

So the randomized tuple $(T x_\nu)_{\nu \in N_*}$ has the same geometry as the original $(x_\nu)_{\nu \in N_*}$, and the linear space $V_N$ remains unchanged.

### 3.3 Confidence and plausibility regions

Consider the classical setting with observation vector

$$y = \mu + \epsilon,$$

where $\mu$ is an unknown fixed vector in $\mathbb{R}^n$ and $\epsilon$ is a random vector with orthogonally invariant distribution on $\mathbb{R}^n$. Let $\beta$ be the orthogonal projection of $\mu$ onto $V_{N_*} \cap V_N^\perp$. The p-value (4) gives rise to a $(1 - \alpha)$-confidence region for $\beta$:

$$C_\alpha(y) := \{ \beta \in V_{N_*} \cap V_N^\perp : \pi(y - \beta) \geq \alpha \}.$$

In case of the usual F-test, this yields Scheffé’s confidence ellipsoid for $\beta$. The coverage probability of $C_\alpha(y)$ equals

$$\mathbb{P}(\beta \in C_\alpha(y)) \begin{cases} = 1 - \alpha & \text{if } \mathbb{P}(\epsilon = 0) = 0 \text{ and } \mu \in V_{N_*}, \\ > 1 - \alpha & \text{otherwise.} \end{cases}$$

If we view all data as fixed, and if the test statistic $\tau$ in (4) depends only on inner products of the data vectors, we may interpret $C_\alpha(y)$ as a $(1 - \alpha)$-plausibility region. It consists of all vectors $\beta \in V_{N_*} \cap V_N^\perp$ such that the association between $y - \beta$ and $(x_\nu)_{\nu \in N_*}$, as measured by $\tau$, is not significantly stronger than the association between $y - \beta$ and the randomized regressor tuple $(T^\top x_\nu)_{\nu \in N_*}$, where $T \sim \text{Haar}_{n,V_{N_*}}$. 

8
4 Technical details and proofs

**Gamma, beta and chi-squared distributions.** Recall that the gamma distribution with shape parameter \(a > 0\) and scale parameter \(c > 0\), denoted by \(\text{Gamma}(a, c)\), is the distribution on \((0, \infty)\) with density

\[
\gamma_{a,c}(y) := \Gamma(a) c^{-1} e^{-y/c} y^{a-1}, \quad y > 0,
\]

where \(\Gamma(a) := \int_0^{\infty} x^{a-1} e^{-x} \, dx\). We also write \(\text{Gamma}(a, c) = \text{Gamma}(a, 1)\). Furthermore, the beta distribution with parameters \(a, b > 0\), denoted by \(\text{Beta}(a, b)\) is the distribution on \((0, 1)\) with density

\[
\beta_{a,b}(u) := B(a, b) u^{a-1} (1-u)^{b-1}, \quad 0 < u < 1,
\]

where \(B(a, b) := \int_0^1 u^{a-1} (1-u)^{b-1} \, du\). The following two results are well-known:

**Lemma 3.** For arbitrary integers \(\ell \geq 1\),

\[
\chi^2_\ell = \text{Gamma}(\ell/2, 2).
\]

**Lemma 4.** For \(a, b, c > 0\) let \(Y_a\) and \(Y_b\) be independent random variables with distribution \(\text{Gamma}(a, c)\) and \(\text{Gamma}(b, c)\), respectively. Then the random variables \(Y_a + Y_b\) and \(U := Y_a/(Y_a + Y_b)\) are stochastically independent, where

\[
Y_a + Y_b \sim \text{Gamma}(a + b, c) \quad \text{and} \quad U \sim \text{Beta}(a, b).
\]

**Proof of (3).** By definition of Fisher’s F distribution and Lemma 3 \(F_{k,\ell}\) describes the distribution of

\[
\tilde{F} := \frac{Y_{k/2}/k}{Y_{\ell/2}/\ell}
\]

with independent random variables \(Y_{k/2} \sim \Gamma(k/2, 2)\) and \(Y_{\ell/2} \sim \Gamma(\ell/2, 2)\). But then Lemma 4 implies that

\[
\frac{Y_{k/2}}{Y_{k/2} + Y_{\ell/2}} = \frac{(k/\ell)\tilde{F}}{(k/\ell)\tilde{F} + 1} \sim \text{Beta}(k/2, \ell/2),
\]

and the latter random variable is a strictly increasing function of \(\tilde{F}\). Hence

\[
F_{k,\ell}(x) = B_{k/2,\ell/2}(\frac{(k/\ell)x}{(k/\ell)x + 1}) \quad \text{for} \quad x \geq 0.
\]

With \(k := p_* - p, \ell := n - p_*\) and \(x := F\), these considerations show that the p-value \(2\) is equal to

\[
1 - F_{p_*-p,n-p_*}(F) = 1 - B_{(p_*-p)/2,(n-p_*)/2}(\frac{SS_N - SS_{N*}}{SS_N})
\]

\[
= B_{(n-p_*)/2,(p_*-p)/2}(\frac{SS_{N*}}{SS_N}).
\]
The latter equation follows from the elementary fact that $U \sim \text{Beta}(a, b)$ if, and only if, $1 - U \sim \text{Beta}(b, a)$.

In the proof of Theorem 1 we utilize another well-known result about beta distributions which is an easy consequence of Lemma 4.

Lemma 5. For $a, b, c > 0$ let $U \sim \text{Beta}(a, b)$ and $V \sim \text{Beta}(a + b, c)$ be independent random variables. Then $UV \sim \text{Beta}(a, b + c)$.

Corollary 6. For $a, \delta > 0$ and an integer $k \geq 2$ let $U_1, \ldots, U_k$ be independent with $U_j \sim \text{Beta}(a + (j - 1)\delta, \delta)$. Then $\prod_{j=1}^k U_j \sim \text{Beta}(a, k\delta)$.

Proof of Lemma 5. Our starting point are independent random variables $G_a \sim \text{Gamma}(a)$, $G_b \sim \text{Gamma}(b)$ and $G_c \sim \text{Gamma}(c)$. Now we apply Lemma 4 three times: We first conclude that $U := G_a/(G_a + G_b)$, $G_a + G_b$ and $G_c$ are independent, where $U \sim \text{Beta}(a, b)$ and $G_a + G_b \sim \text{Gamma}(a + b)$. Then we may conclude that $U$ and $V := (G_a + G_b)/(G_a + G_b + G_c)$ are independent with $V \sim \text{Beta}(a + b, c)$. Finally, $UV = G_a/(G_a + G_b + G_c)$ has distribution $\text{Beta}(a, b + c)$.

Proof of Theorem 1. Let us first consider the case $p = p + 1$, so $\{x_\nu : \nu \in N \setminus N\} = \{z\}$ with only one random vector $z \sim N_n(0, I_n)$. Note that $y - \hat{y}_N$ is a nonzero vector in the linear space

$$\mathbb{V}_N^\perp = \{x \in \mathbb{R}^n : x^T \eta = 0 \text{ for all } \eta \in \mathbb{V}_N\}.$$ 

Let $b_1, b_2, \ldots, b_n$ be an orthonormal basis of $\mathbb{R}^n$ such that

$$\mathbb{V}_N = \text{span}(b_1, \ldots, b_p) \text{ and } y - \hat{y}_N = S_{SS}^{1/2} b_{p+1}.$$ 

By rotational symmetry of the standard Gaussian distribution on $\mathbb{R}^n$, $Z_j := b_j^T z$ defines stochastically independent, standard Gaussian random variables $Z_1, Z_2, \ldots, Z_n$, and the orthogonal projection of $z$ onto $\mathbb{V}_N^\perp$ is given by

$$\tilde{z} := \sum_{j=p+1}^n Z_j b_j.$$ 

In particular,

$$\mathbb{V}_N^* = \text{span}(1, \ldots, b_p, z) = \text{span}(1, \ldots, b_p, \tilde{z})$$ 

and

$$y = \hat{y}_N + S_{SS}^{1/2} b_{p+1},$$

$$\hat{y}_N^* = \hat{y}_N + \frac{z^T y}{\|z\|^2} \tilde{z} = \hat{y}_N + S_{SS}^{1/2} \frac{z^T b_{p+1}}{\|z\|^2} \tilde{z}.$$
Consequently,

\[
\frac{SS_{N*}}{SS_N} = \left\| b_{p+1} - \frac{z^\top b_{p+1} z}{\|z\|^2} \right\|^2 = 1 - \frac{(z^\top b_{p+1})^2}{\|z\|^2} = \frac{\sum_{j=p+2}^n Z_j^2}{\sum_{j=p+1}^n Z_j^2}
\]

\[
\sim \text{Beta}\left((n - p - 1)/2, 1/2\right) = \text{Beta}\left((n - p_*)_2, (p_* - p)/2\right)
\]

by Lemmas 3 and 4.

In case of \( k := p_* - p > 1 \), one may apply the previous argument inductively to show that

\[
\frac{SS_{N*}}{SS_N} = \prod_{\ell=1}^k U_\ell
\]

in distribution, where \( U_1, \ldots, U_k \) are stochastically independent with

\[
U_\ell \sim \text{Beta}\left((n - p - \ell)/2, 1/2\right).
\]

In other words, for \( j = 1, \ldots, k \),

\[
U_{k+1-j} \sim \text{Beta}\left((n - p_*)/2 + (j - 1)/2, 1/2\right).
\]

Applying Corollary 5 with \( a = (n - p_*)/2 \) and \( \delta = 1/2 \) yields the assertion that \( SS_{N*}/SS_N \) follows \( \text{Beta}\left((n - p_*)/2, k/2\right) \).

\[
\square
\]

**Haar measure on \( \mathcal{O}_n \).** For the reader’s convenience we collect some standard arguments to provide a self-contained account of that topic. We start with two specific constructions of a random matrix \( T \in \mathcal{O}_n \) such that

\[
\mathcal{L}(ST) = \mathcal{L}(T) \quad \text{for any fixed } S \in \mathcal{O}_n.
\]

(5)

In both cases the starting point is a random matrix \( Z = [z_1, \ldots, z_n] \) with \( d^2 \) independent components with standard Gaussian distribution. With probability one, the columns \( z_1, \ldots, z_n \) are linearly independent. Hence

\[
T := Z(Z^\top Z)^{-1/2}
\]

is well-defined almost surely and easily seen to belong to \( \mathcal{O}_n \). For fixed \( S \in \mathcal{O}_n \) it follows from \( \mathcal{L}(SZ) = \mathcal{L}(Z) \) that the distribution of \( T \) coincides with the distribution of

\[
(SZ)((SZ)^\top (SZ))^{-1/2} = SZ(Z^\top Z)^{-1/2} = ST.
\]

Thus \( T \) satisfies (5).
The same conclusion holds true if we construct \( T = [t_1, \ldots, t_n] \) by means of the Gram-Schmidt orthogonalization: We start with \( t_1 := \frac{1}{\|z_1\|} z_1 \) and then set

\[
t_k := \left( \frac{\|z_k\|^2}{k-1} - \sum_{j=1}^{k-1} (t_j^\top z_k)^2 \right)^{-1/2} \left( z_k - \sum_{j=1}^{k-1} (t_j^\top z_k) t_j \right)
\]

for \( k = 2, \ldots, n \). This representation shows that the first column of \( T \) has the same distribution as a standard Gaussian random vector normalized to have length one.

Now let \( T_1, T_2 \) be stochastically independent random matrices in \( O_n \) satisfying (5). Then for any Borel set \( B \subset O_n \),

\[
\Pr(T_1^\top T_2 \in B) = \mathbb{E} \Pr(T_1^\top T_2 \in B \mid T_1) = \Pr(T_2 \in B),
\]

and

\[
\Pr(T_1^\top T_2 \in B) = \mathbb{E} \Pr((T_2^\top T_1)^\top \in B \mid T_2) = \Pr(T_1^\top \in B).
\]

Hence \( T_1^\top \) and \( T_2 \) have the same distribution. From this one can easily deduce that there is only one distribution \( \text{Haar}_n \) on \( O_n \) such that a random matrix \( T \) with that distribution satisfies (5).

The previous considerations show that a random matrix \( T \sim \text{Haar}_n \) satisfies also

\[
\mathcal{L}(T^\top) = \mathcal{L}(T) = \mathcal{L}(TS) \quad \text{for any fixed } S \in O_n.
\]

(6)

Moreover, if \( s_1, \ldots, s_q \) are fixed orthonormal vectors in \( \mathbb{R}^n \), \( 1 \leq q < n \), then

\[
\mathcal{L}(\{Ts_1, \ldots, Ts_q\}) = \mathcal{L}(\{t_1, \ldots, t_q\}),
\]

where \( t_1, \ldots, t_n \) are the columns of \( T \). This follows by extending \( s_1, \ldots, s_q \) to an orthonormal basis \( s_1, \ldots, s_n \) of \( \mathbb{R}^n \) and applying (6) with \( S = [s_1, \ldots, s_n] \in O_n \). In particular, for any fixed unit vector \( s \in \mathbb{R}^n \), the random vector \( Ts \) has the same distribution as \( u := \|z\|^{-1}z \) with a standard Gaussian random vector \( z \in \mathbb{R}^n \).

**Proof of Lemma 2.** Suppose that \( y \) satisfies (i). Then for any Borel set \( B \subset \mathbb{R}^n \),

\[
\Pr(Ty \in B) = \mathbb{E} \Pr(Ty \in B \mid T) = \Pr(y \in B),
\]

Hence (ii) is satisfied as well.

Now suppose that \( y \) satisfies (ii). That means \( y \) has the same distribution as \( Ty \), where \( T \sim \text{Haar}_{n, V_N} \) is stochastically independent from \( y \). By conditioning on \( y = \tilde{y} + \hat{e} \) one sees that
The vector $Ty$ has the same distribution as $\hat{y} + \|\hat{\epsilon}\|u$, where $u$ is uniformly distributed on the unit sphere of $V_N^\perp$. Thus $y$ satisfies (iii) as well.

Finally, suppose that $y$ satisfies (iii), that means, $y$ has the same distribution as $\hat{y} + \|\hat{\epsilon}\|u$, where $u$ is uniformly distributed on the unit sphere of $V_N^\perp$ and stochastically independent from $y$. But for any fixed $S \in O_n$ the distributions of $Su$ and $u$ are identical, so for any Borel set $B \subset \mathbb{R}^p$,

$$
\mathbb{P}(Sy \in B) = \mathbb{P}(S(\hat{y} + \|\hat{\epsilon}\|u) \in B) = \mathbb{P}(\hat{y} + \|\hat{\epsilon}\|Su \in B) = \mathbb{E}\mathbb{P}(\hat{y} + \|\hat{\epsilon}\|Su \in B | y) = \mathbb{E}\mathbb{P}(\hat{y} + \|\hat{\epsilon}\|u \in B | y) = \mathbb{P}(\hat{y} + \|\hat{\epsilon}\|u \in B) = \mathbb{P}(y \in B).
$$

Consequently, $y$ satisfies (i) as well.

\[\square\]

References

[1] Candès, E., Fan, Y., Janson, L. and Lv, J. (2017). Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. Preprint, [arXiv:1610.02351](https://arxiv.org/abs/1610.02351).

[2] Davies, L. (2018). Lasso, knockoff and Gaussian covariates: a comparison. Preprint, [arXiv:1805.01862](https://arxiv.org/abs/1805.01862).

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

[4] Scheffé, H. (1959). Analysis of Variance. Wiley and Sons.

[5] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B.* 58(1), 267–288.