Global Journal of Science Frontier Research: G
Bio-Tech & Genetics
Volume 21 Issue 1 Version 1.0 Year 2021
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

Null Distribution of the Test Statistic for Model Selection Via Marginal Screening: Implications for Multivariate Regression Analysis

By A. V. Rubanovich & V. A. Saenko
Russian Academy of Sciences

Summary- Marginal screening (MS) is the computationally simple and commonly used for the dimension reduction procedures. In it, a linear model is constructed for several top predictors, chosen according to the absolute value of marginal correlations with the dependent variable. Importantly, when k predictors out of *m* primary covariates are selected, the standard regression analysis may yield false-positive results if $m >> k$ (Freedman's paradox). In this work, we provide analytical expressions describing null distribution of the test statistics for model selection via MS. Using the theory of order statistics, we show that under MS, the common F-statistic is distributed as a mean of *k* top variables out of m independent random variables having a χ_1^2 distribution. Based on this finding, we estimated critical p-values for multiple regression models after MS, comparisons with which of those obtained in real studies will help researchers to avoid false-positive result. Analytical solutions obtained in the work are implemented in a free Excel spreadsheet program.

Keywords: linear models, multiple comparisons, freedman's paradox, extreme value theory, order statistics, genetic risk score.

GJSFR-G Classification: FOR Code: 060499

Strictly as per the compliance and regulations of:

Null Distribution of the Test Statistic for Model Selection Via Marginal Screening: Implications for Multivariate Regression Analysis

A. V. Rubanovich [α] & V. A. Saenko [σ]

Summary- Marginal screening (MS) is the computationally simple and commonly used for the dimension reduction procedures. In it, a linear model is constructed for several top predictors, chosen according to the absolute value of marginal correlations with the dependent variable. Importantly, when *k* predictors out of *m* primary covariates are selected, the standard regression analysis may yield false-positive results if $m >> k$ (Freedman's paradox). In this work, we provide analytical expressions describing null distribution of the test statistics for model selection via MS. Using the theory of order statistics, we show that under MS, the common *F*-statistic is distributed as a mean of *k* top variables out of *m* independent random variables having a χ_1^2 distribution. Based on this finding, we estimated critical *p*-values for multiple regression models after MS, comparisons with which of those obtained in real studies will help researchers to avoid false-positive result. Analytical solutions obtained in the work are implemented in a free Excel spreadsheet program.

Keywords: linear models, multiple comparisons, freedman's paradox, extreme value theory, order statistics, genetic risk score.

I. Introduction

Marginal screening (MS) is the simplest and most commonly used method of variable selection (Hastie, Tibshirani, 2003; Genovese et al, 2009, 2012; Leek, 2012). Its selection power is comparable to that of the Lasso, Stepwise, Compressed Sensing and other, but it is faster and easier than regularization approaches. In fact, MS is intuitively preferred by the researchers, when the number of objects (e.g. participants of a study, samples or outcomes) is much smaller than the number of explanatory variables (the so-called "n << p problem"). This problem is particularly acute in modern genetic studies such as GWAS, RNA-seq, microarray analysis, motif activity response analysis, etc. (Windle M., 2016). For all of them, a typical situation is when the number of objects is several orders of magnitude less than the number of covariates from which a statistically significant combination of predictors is derived. In fact, this is another side of an old problem of multiple comparisons, which is somewhat masked in regression analysis. Indeed, when a multiple regression model incorporating all investigated covariates is being established, no correction for multiple comparisons is required. For the standardized model

$$Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon, \qquad (1)$$

the mean of *p*-values (the *F*-test) is 0.5 under any *k* if all X_i are random and independent of *Y*. False-positives results are impossible even if *k* is large.

The situation changes dramatically under the model selection, e.g. via MS. If the predictors for a model (1) are pre-selected as top effects from a (very) large set of random variables, a false-positive result is practically inevitable. For example, if one selects top 5 predictors from 100 random variables and constructs a model (1) for 100 points, then, on average, the squared multiple correlation coefficient $R^2 = 0.23$, $p = 0.00015$, and a predictor with maximum effect with $p_1 = 0.01$ will be obtained. This phenomenon, known as Freedman's paradox (Freedman, 1983; Lukacs et al., 2010), is usually ignored in the textbooks and statistical software packages. In the literature, only analytical and empirical estimates of $E(R^2 \mid H_0)$ for MS are available (Alam, 1979; Rencher, Pun, 1980; Wray, 2013).

The purpose of this work was to explicitly address null-distribution of the *F*-statistic, which is used for testing the significance of a model (1), when model selection is performed with MS. In addition, we present related formulae for the case when the reduction of the number of independent variables is performed by calculating the so-called «risk score» (such as genetic risk score (GRS) or polygenic risk score (PGRS) in genetic studies). The derived analytical solutions are implemented in a free Excel program, which evaluates statistical significance of the results of a regression analysis after MS.

II. Notations and Abbreviations

MS - marginal (or correlation) screening;

RS - risk score; GRS – genetic risk score; PGRS – polygenic risk score;

Author α: Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia. e-mail: rubanovich@vigg.ru

Author σ: Department of Radiation Molecular Epidemiology, Atomic Bomb Disease Institute, Nagasaki University, Nagasaki, Japan.

CDF - cumulative distribution function;

PDF - probability density function;

iid - independently and identically distributed;

r.v. - random variable;

X – random variable X;

$E(\mathbf{X}), D(\mathbf{X})$ - the mean value and the variance of X;

$X \sim Y$ - X and Y are identically distributed;

$\mathbf{X} \mid H_0$ - X under null hypothesis;

$X \sim \mathbf{F}_{k,n}$ - X has F-distribution with parameters *k* and *n* (degrees of freedom);

$X \sim \boldsymbol{\chi}^2_k$ - X has chi-squared distribution with parameter *k* (degrees of freedom);

$X \sim \mathbf{N}(\mu, \sigma^2)$ - X has normal distribution with parameters μ and σ^2 (mean and variance, respectively);

$\Phi(x)$ - CDF of the standard normal distribution $\mathbf{N}(0,1)$;

$\Phi^{-1}(x)$ - the inverse CDF (quantile function) of the standard normal distribution;

n - sample size;

m - total number of independent variables (predictors);

k - the number of top predictors selected for a model construction.

III. Method

The distributions of top extreme (hereafter – top) values of r.v.s are examined in frame of the order statistics theory (Ahsanullah et al., 2013; Arnold et al., 2008).

Let r.v.s $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_m$ are independent and identically distributed, and $\mathbf{Z}_{1:m} \le \mathbf{Z}_{2:m} \le ... \le \mathbf{Z}_{m:m}$ are the same r.v.s in ascending order. Then r.v. $\mathbf{Z}_{i:m}$ is called the i^{th} order statistic and $\mathbf{Z}_{m:m} = \max\{\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_m\}$. The arithmetic mean of top *k* order statistics is called the «selection differential»:

$$D_{k,m}(\mathbf{Z}) \equiv k^{-1}\left(\mathbf{Z}_{m:m} + \mathbf{Z}_{m-1:m} + ... + \mathbf{Z}_{m-k+1:m}\right) = k^{-1} \sum_{i=m-k+1}^{m} \mathbf{Z}_{i:m}.$$

The properties of the r.v. $D_{k,m}(\mathbf{Z})$ has been studied in detail (Nagaraja, 1980, 1982, 2006; Arnold et al., 2008). Further reasoning involves the formalism published earlier (Stigler, 1973). Let $m \to \infty$ and $p = k/m = \text{const}$. Then r.v. $D_{k,m}(\mathbf{Z})$ is an asymptotically normally distributed r.v. with the following parameters:

$$D_{k,m}(\mathbf{Z}) \sim \mathbf{N}\left(\mu_p, \frac{1}{k}\left(\sigma_p^2 + (1-p)(\mu_p - z_p)^2\right)\right), \quad (2)$$

$$\mu_p = E(\mathbf{Z} \mid \mathbf{Z} > z_p) = \frac{1}{p}\int_{z_p}^{\infty} z f(z) dz,$$

$$\sigma_p^2 = D(\mathbf{Z} \mid \mathbf{Z} > z_p) = \frac{1}{p}\int_{z_p}^{\infty} (\mu_p - z)^2 f(z) dz,$$

where $F(z)$ and $f(z)$ are the CDF and PDF of the r.v. $\mathbf{z}$; $z_p = F^{-1}(1-p)$; μ_p and σ_p^2 are the mean and variance of the distribution obtained by *F* trimming to below z_p.

IV. Results

a) Additivity of Cohen's f^2 effect size for independent predictors

Our analytical results described below are essentially based on the Proposition 1. We believe this proposition may be of independent interest.

Proposition 1: Let correlations between independent variables X_i in a regression model (1) are

$$Cor(X_i, X_j) = r_i r_j, \; i \neq j,$$

where $r_i = Cor(Y, X_i)$ are marginal correlations between the dependent variable and predictors. Then the squared multiple correlation (R^2) for model (1) satisfies

$$\frac{R^2}{1-R^2} = \sum_{i=1}^{k} \frac{r_i^2}{1-r_i^2}. \quad (3)$$

Proof. It is known that $r = C\beta$ and $R^2 = r^T\beta$, where *r* is a column vector whose i^{th} entry is r_i; β is a column vector whose i^{th} entry is β_i; *C* is the correlation matrix for the predictor variables. By the condition of Proposition 1, $C = rr^T + D$, where *D* is a diagonal matrix with the elements $D_{ii} = 1 - r_i^2$ on the diagonal. Then, $r = C\beta = r(r^T\beta) + D\beta = R^2 r + D\beta$. Further, $\beta = (1-R^2)D^{-1}r$ and hence, $r^T\beta = (1-R^2)(r^T D^{-1} r)$. Therefore,

$$\frac{R^2}{1-R^2} = r^T D^{-1} r = \sum_{i=1}^{k} \frac{r_i^2}{1-r_i^2}.$$

Comments

1. Index $f^2 = R^2/(1-R^2)$ is known as Cohen's f^2 effect size (Cohen, 1988; 1992).
2. The condition $Cor(X_i, X_j) = r_{ij} = r_i r_j$ assumes that all predictors are statistically independent and their mutual correlations are fully due to the correlation with *Y*. In this case, the partial

correlations between the predictor variables (with fixed *Y*) are:

$$r_{ij.Y} = \frac{r_{ij} - r_i r_j}{\sqrt{(1-r_i^2)(1-r_j^2)}} = 0$$

Thus, according to the Proposition 1 proved above, Cohen's indices f^2 are additive when the predictors are independent.

b) Null distribution of the F-statistic for marginal screening

When testing the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ for the model (1), Fisher's *F*-statistic is calculated as

$$\mathbf{F} = \frac{n-k-1}{k}\frac{\mathbf{R}^2}{1-\mathbf{R}^2},$$

where $\mathbf{R}^2$ is the squared multiple correlation coefficient for the sample size *n*. If the model is initially constructed for variables $\{X_i\}$, $i = 1, \ldots, k$, then the null distribution of $\mathbf{F}$ is: $\mathbf{F} \mid H_0 \sim \mathbf{F}_{k,n-k-1}$. However, this is incorrect if the predictors were preliminarily selected as top from a large number of variables $\{X_i\}$, $i = 1, \ldots, m$, where $m >> k$.

Proposition 2: Let the model (1) is constructed for *k* predictors which were chosen as the top $|Cor(Y, X_i)|$ out of *m* variables. Then, when $n >> k$, the *F*-statistic is approximately distributed as

$$\mathbf{F} \mid H_0 \sim \frac{1}{k}(\mathbf{Z}_{m:m} + \mathbf{Z}_{m-1:m} + ... + \mathbf{Z}_{m-k+1:m}) \equiv D_{k,m}(\boldsymbol{\chi}_1^2),$$

$$\mathbf{Z} \sim \boldsymbol{\chi}_1^2.$$

Proof. Let in a regression model all *Y*, X_1, X_2 …, X_m are independent and identical r.v.s having standard normal distribution *N*(0,1). Consider r.v.s $\mathbf{f}_i^2 = \mathbf{r}_i^2/(1-\mathbf{r}_i^2)$, where $\mathbf{r}_i^2$ are sample correlations. If the number of points in the model is *n*, then $(n-2)\mathbf{f}_i^2 \sim \mathbf{F}_{1,n-2} \to \boldsymbol{\chi}_1^2$ under $n \to \infty$. Consequently, the "top" predictors correspond to *k* greatest values among the *m* values of independent r.v.s $\mathbf{Z} \sim \chi_1^2$. Then, when $n >> k$, according to Proposition 1:

$$\mathbf{F} = \frac{n-k-1}{k}\frac{\mathbf{R}^2}{1-\mathbf{R}^2} = \frac{n-k-1}{k}\sum_{i=1}^{k}\mathbf{f}_i^2 \sim \frac{n-k-1}{k(n-2)}(\mathbf{Z}_{m:m} + \mathbf{Z}_{m-1:m} + ... + \mathbf{Z}_{m-k+1:m}) \approx D_{k,m}(\boldsymbol{\chi}_1^2).$$

Corollaries

1. Model selection leads to the substantial inflation of $E(\mathbf{F} \mid H_0)$. Without model selection, $k = m$ and, consequently, $E(\mathbf{F} \mid H_0) = E(D_{m,m}(\boldsymbol{\chi}_1^2)) = E(\boldsymbol{\chi}_1^2) = 1$. In case $k << m$ (model selection by MS), $E(\mathbf{F} \mid H_0) = E(D_{k,m}(\boldsymbol{\chi}_1^2)) \to \infty$ when $m \to \infty$. In any case, the distribution of $\mathbf{F}|\ H_0$ is weakly dependent on the sample size *n*.
2. As $n \to \infty$, $m \to \infty$ and $p = k/m = \text{const}$, the null distribution of the *F*-statistic is approximately $\mathbf{F} \mid H_0 \sim \mathbf{N}(\mu_\mathbf{F}, \sigma_\mathbf{F}^2)$, where

$$\mu_\mathbf{F} = 1 + \frac{1}{p}\sqrt{\frac{2z_p}{\pi}}e^{-\frac{z_p}{2}},$$

$$\sigma_\mathbf{F}^2 = \frac{1}{k}\left(3 + \frac{1}{p}\sqrt{\frac{2z_p}{\pi}}(3+z_p)e^{-\frac{z_p}{2}} - \mu_p^2 + (1-p)(\mu_p - z_p)^2\right) \quad (4)$$

and $z_p = F^{-1}(1-p) = \Phi^{-1}(1-p/2)^2$. Here F^{-1} and Φ^{-1} are the quantile functions of $\boldsymbol{\chi}_1^2$ and normally distributed r.v., respectively. Expressions (4) are obtained through the PDF of $\boldsymbol{\chi}_1^2$ substitution into formula (2). Series expansion of the expression (4) under $z_p \to \infty$ leads to an approximation for the distribution of r.v. $\mathbf{F}$ for the case $k << m$:

$$\mathbf{F} \mid H_0 \sim \mathbf{N}\left(2 + z_p, \frac{2}{k}\left(1 + \sqrt{\frac{2z_p}{\pi}}\right)\right). \quad (5)$$

3. Let $\mathbf{p} = 1 - F_{k,n-k-1}(\mathbf{F})$ is a r.v. corresponding to the spurious *p*-value for the model (1) under MS. Equation (4) allows to estimate the lower 5% quantile of its null-distribution: $Q_{5\%}(\mathbf{p} \mid H_0) \approx 1 - F_{k,n-k-1}(\mu_\mathbf{F} + 1.65\sigma_\mathbf{F})$
4. When $m >> k$, expectation of the squared multiple correlation coefficient under H_0 approximately equals to

$$E(\mathbf{R}^2 \mid H_0) = \frac{k\mu_\mathbf{F}}{n-k-1+k\mu_\mathbf{F}}. \quad (6)$$

c) Significance of the regression coefficient β_1 for a predictor based on the maximum effect

Usually the researcher desires to control not only significance of the whole model, but also of the predictors in the model (1). Here, we focus on the statistical significance of a predictor with the maximum

effect, denoted as X_1: $|\beta_1| \equiv \max_i |\beta_i|$. The standard procedure for testing H_0: $\beta_1 = 0$ for the model (1) involves the calculation of a statistic

$$\mathbf{F}_1 = (n-k-1)\frac{\mathbf{R}^2 - \mathbf{R}^2_{2,\ldots,k}}{1-\mathbf{R}^2},$$

where $\mathbf{R}^2_{2,\ldots,k}$ are the squared multiple correlations for predictors X_2, X_3 …, X_m. It is known that without model selection, $\mathbf{F}_1 \sim \mathbf{F}_{1,n-k-1} \to \boldsymbol{\chi}^2_1$ under $n \to \infty$. It turns out that with MS, r.v. $\mathbf{F}_1$ approximately equals to the maximum value of m independent realizations of r.v. $\mathbf{Z} \sim \boldsymbol{\chi}^2_1$. Thus, we arrive to

Proposition 3. Let the model (1) is constructed for k predictors, which were selected as top of $|\mathbf{r}_i| = |Cor(\mathbf{Y}, \mathbf{X}_i)|$ from m variables. Then $\mathbf{F}_1 \mid H_0 \sim \mathbf{Z}_{m:m}$, where $\mathbf{Z} \sim \boldsymbol{\chi}^2_1$.

Proof. Let r.v.'s Y, X_1, X_2 …, X_m are independent r.v.s, which have normal distribution $N(0,1)$. Let $\mathbf{f}_i^2 = \mathbf{r}_i^2/(1-\mathbf{r}_i^2)$ и $\mathbf{f}^2 = \mathbf{R}^2/(1-\mathbf{R}^2)$. Then according to Proposition 1,

$$\frac{\mathbf{R}^2_{2,\ldots,k}}{1-\mathbf{R}^2_{2,\ldots,k}} = \sum_{i=2}^{k} \mathbf{f}_i^2 = \mathbf{f}^2 - \mathbf{f}_1^2$$

Hence, under $n \to \infty$

$$\mathbf{F}_1 \mid H_0 = (n-k-1)\left(\frac{\mathbf{f}^2}{1+\mathbf{f}^2} - \frac{\mathbf{f}^2 - \mathbf{f}_1^2}{1+\mathbf{f}^2-\mathbf{f}_1^2}\right)(1+\mathbf{f}^2) = \frac{(n-k-1)\mathbf{f}_1^2}{1+\mathbf{f}^2-\mathbf{f}_1^2} \sim \frac{(n-k-1)\mathbf{Z}_{m:m}}{n-2+\sum_{i=m-k+1}^{m-1}\mathbf{Z}_{i:m}} \approx \mathbf{Z}_{m:m},\ \mathbf{Z} \sim \boldsymbol{\chi}^2_1.$$

Corollaries.

1. Marginal screening does not substantially change the $\mathbf{F}_1$ statistic, which only slightly depends on k and n provided $n > m$ (see Supplement 1, Fig. 2).
2. Let $\mathbf{p}_1$ is a r.v. related to the spurious p-value for the predictor with the maximum effect under H_0. Then, approximately, $\mathbf{p}_1 \mid H_0 \sim 1 - \mathbf{U}_{m:m}$ under $n >> 1$, where U has the uniform distribution on [0,1]. Indeed, under $n >> 1$

$$\mathbf{p}_1 \mid H_0 \sim 1 - F_{1,n-k-1}(\mathbf{F}_1) \approx 1 - F_{\chi^2_1}(\mathbf{Z}_{m:m}) = 1 - F_{\chi^2_1}(F^{-1}_{\chi^2_1}(\mathbf{U}_{m:m})) = 1 - \mathbf{U}_{m:m}$$

Here, the relation: $\mathbf{X}_{i:m} = F_X^{-1}(\mathbf{U}_{i:m})$ is used, where F_X^{-1} is the quantile function of a r.v. X.

Hence, $E(\mathbf{p}_1 \mid H_0) \approx 1 - E(\mathbf{U}_{m:m}) = 1/(m+1)$ (Ahsanullah et al., 2013), and the lower α-quantile for $\mathbf{p}_1$ equals to $Q_\alpha(\mathbf{p}_1 \mid H_0) \approx 1 - (1-\alpha)^{1/m} \approx \alpha / m$.

d) Risk summation

A special method for reducing the number of predictors is widely used in modern genetic studies, termed "risk summation" (RS). For example, in «case - control» studies, the so-called genetic risk scores (GRS) are often calculated as the total number of "risk alleles", which are more frequent in patients or affected participants. Then, a comparison of GRS between samples of healthy controls and affected individuals is performed using e.g. Student's t-test. In the case of quantitative trait, its correlation with the number of alleles is sometimes calculated, resulting in the increase of the dependent variable.

Let us calculate the null distribution of the t-statistic for RS comparison. For the model (1) we define the new variable «risk score» X_{RS} as

$$X_{RS} = \mathrm{Sign}(\beta_1)X_1 + \mathrm{Sign}(\beta_2)X_2 + \ldots + \mathrm{Sign}(\beta_k)X_k,$$

and consider a simple regression $Y = \beta X_{RS} + \varepsilon$. Let $R_{RS} \equiv Cor(Y, X_{RS}) = \beta$. Corresponding F-statistic for this model (F_{RS}) is calculated in case of a continuous dependent variable, and Student's t-test is applied if the dependent variable is binary. The t-statistic is $\mathbf{T}_{RS} = \sqrt{\mathbf{F}_{RS}}$. Then, if H_0 is true, the following proposition takes place:

Proposition 4.

$$\mathbf{T}_{RS} \mid H_0 \sim \frac{1}{\sqrt{k}}(\mathbf{Z}_{m:m} + \mathbf{Z}_{m-1:m} + \ldots + \mathbf{Z}_{m-k+1:m}) = \sqrt{k} D_{k,m}(\boldsymbol{\chi}_1),\ \mathbf{Z} \sim \boldsymbol{\chi}_1 .$$

Proof. Indeed, for a standardized model $\sigma_{X_i} = \sigma_Y = 1$, $i = 1,\ldots,k$. Then, according to Proposition 1 under H_0:

$$\mathbf{R}_{RS} = Cor(\mathbf{Y}, \mathbf{X}_{RS}) = \left(\left(\sum_{i=1}^{k} |\mathbf{r}_i| \right)^2 + \sum_{i=1}^{k} (1 - \mathbf{r}_i^2) \right)^{-1/2} \sum_{i=1}^{k} |\mathbf{r}_i|$$

from which under $n \to \infty$

$$\mathbf{F}_{RS} = (n-2)\frac{\mathbf{R}_{RS}^2}{1-\mathbf{R}_{RS}^2} = (n-2)\frac{\left(\sum_{i=1}^{k} |\mathbf{r}_i| \right)^2}{\sum_{i=1}^{k} (1-\mathbf{r}_i^2)} \approx (n-2)\left(\sum_{i=1}^{k} \frac{\mathbf{f}_i}{\sqrt{k}} \right)^2,$$

and $\mathbf{f}_i^2 = \mathbf{r}_i^2 / (1-\mathbf{r}_i^2)$. Taking into account that $\mathbf{T}_{RS} = \sqrt{\mathbf{F}_{RS}}$, and $\sqrt{(n-2)}\mathbf{f}_i \sim \boldsymbol{\chi}_1$, we arrive to Proposition 4.

Corollaries.

1. When $n \to \infty$, $m \to \infty$ и $p = k/m = \text{const}$, null-distribution of the *t*-statistic is:
$\mathbf{T} \mid H_0 \sim \mathbf{N}\left(\mu_{\mathbf{T}}, \sigma_{\mathbf{T}}^2\right)$, where

$$\mu_{\mathbf{T}} = \frac{1}{p}\sqrt{\frac{2k}{\pi}} e^{-\frac{z_p}{2}},$$

$$\sigma_{\mathbf{T}}^2 = 1 - \frac{2}{\pi p} e^{-z_p} - \sqrt{\frac{2z_p}{\pi}} \frac{(2p-1)}{p} e^{-\frac{z_p}{2}} + (1-p) z_p \quad (7)$$

and $z_p = \Phi^{-1}(1-p/2)^2$. Expressions (7) are obtained through the PDF of $\boldsymbol{\chi}_1$, which is $f(z) = \sqrt{2/\pi} e^{-z^2/2}$, substitution into the formula (2). For $p < 0.1$, the following approximation can be used:

$$\mathbf{T} \mid H_0 \sim \mathbf{N}\left(\sqrt{k}\Phi^{-1}\left(1 - \frac{k}{2m}\right),\ 2\Phi^{-1}\left(1 - \frac{k}{2m}\right)^{-1} \right),$$

where Φ^{-1} is the quantile of the normal distribution. When $k > 0.2m$,

$$\mathbf{T} \mid H_0 \sim \mathbf{N}\left(\sqrt{2k/\pi},\ 1 - 2/\pi\right). \quad (8)$$

2. $\mathbf{R}_{RS}^2 \le \mathbf{R}^2$. This inequality follows from Proposition 1 and the convexity of the $f(x) = x(1-x)^{-1}$ function. When $k << m$, an approximate equality takes place: $\mathbf{R}_{RS}^2 \approx \mathbf{R}^2$, and $\mathbf{F}_{RS} \approx k\mathbf{F}$, where F and $\mathbf{R}^2$ are defined in Proposition 2.

V. Related Works

Analytical approximation for the distribution $F \mid H_0$ under MS, is described here for the first time to the best of the authors' knowledge. Previously, the analytical and empirical expressions for $E(\mathbf{R}^2 \mid H_0)$ (Rencher, Pun, 1980; Wray, 2013) and the top quintile $Q_{1-\alpha}(\mathbf{R}^2 \mid H_0)$ (Diehr, Hoflin, 1974; Salt, 2007) were proposed. The empirical formula for $Q_\alpha(\mathbf{p} \mid H_0)$ was obtained by Salt (2007). A possible link between the null-distribution of $\mathbf{R}^2$ and order statistics for χ_1^2 under MS was first noted by Alam and Wallenius (Alam, Wallenius, 1979).

Our results are closest to the recent work of J. Fan that addressed "spurious correlations" (Fan et. al., 2017). Applicably to our work, those results are as follows. Assume that the *Y* and X_i in (1) are independent and identical r.v.s having the standard normal distribution *N*(0,1). The maximum spurious correlation is defined as $\hat{\mathbf{R}}_n^2(k,m) = \max_{\beta} Cor(Y, \sum_{i=1}^{m} \beta_i X_i)^2$ subject to $\| \beta \|_0 = k$ (the number of nonzero β_i's equals *k*). If $n >> (k \ln m)^7$, then the asymptotic distributions of the maximum spurious correlation is:

$$n\hat{\mathbf{R}}_n^2(k,m) \sim \mathbf{Z}_{m:m} + \mathbf{Z}_{m-1:m} + \ldots + \mathbf{Z}_{m-k+1:m},\ \mathbf{Z} \sim \chi_1^2 . \quad (9)$$

The condition $n >> (k \ln m)^7$ means that the practical use of this formula is possible only with very large *n*. Even at $m = 100$ the required sample size is $n >> 44000$. Otherwise, the formula (9) leads to inadequate estimates: $E(\hat{\mathbf{R}}_n^2(k,m)) > 1$. Our result (Proposition 2) is meaningfully similar to (9) yet, in contrast, requires only $n >> k$ and therefore it is devoid of such a disadvantage.

Recently, J. Taylor and colleagues developed an innovative approach to the problem of statistical inference after model selection (Lee, Taylor, 2014; Tibshirani, Taylor, 2016). For a regression model with Gaussian errors, they presented a method for exact inference, conditional on a polyhedral constraint on the observations *Y*. This approach is applicable to different model selection procedures including marginal screening, forward stepwise regression, lasso, least angle regression and other sequential regression techniques being an exciting new advance. Note,

however, that it is a conditional test in which any estimator used for inference is influenced by correlations with predictors that are not selected. As an example, consider the simplest case $k = 1$ (one top predictor). Than the "polyhedral lemma" (Lee, Taylor, 2014) would yield the following distribution of the regression coefficient β_1 under H_0:

$$\beta_1 \mid H_0 \sim TN[|r_2|; \infty].$$

Here r_2 is the second top marginal correlation between the dependent variable and predictors; $TN[a; b]$ is the normal distribution truncated to the interval $[a; b]$. Naturally, the p-value obtained will depend heavily on the unselected predictor X_2. In contrast, our unconditional test (Proposition 2) depends only on the selected first top predictor X_1:

$$(n-2)\frac{r_1^2}{1-r_1^2} \mid H_0 \sim Z_{m:m},\ Z \sim \chi_1^2.$$

We believe that the interpretation of the conditional "exact" p-value is still unclear since the expected value is also affected by the unselected predictors. It is also clear that a conditional post-selection test has lower power compared to the unconditional test, since for any X

$$\alpha > P(\text{Type 1 error} \mid X) > P(\text{Type 1 error, X}).$$

Post-selection inference is a very promising approach to inference that appears to work very well in case $n > m$. How well it performs when $m > n$ needs to be demonstrated in the future, since the method remains very new at this point.

VI. Practical Relevance

Our theoretical considerations have immediate application since the derived formulas enable quick and efficient evaluation of statistical significance of a multiple regression model, predictors in which are selected via MS. The accuracy of approximations was tested using computer simulations for a wide range of parameters (Supplement 1). More accurate estimates can be obtained by applying the $F_{1,n-2}$ instead of χ_1^2.

For calculations according to the formulas from Propositions 2-4, we suggest using the H0_Model_Selection.xlsx spreadsheet (Supplement 2). One needs to enter three numbers $\{n, m, k\}$, where n - sample size, m - the initial number of covariates (independent variables), and k - the number of top predictors selected to construct the model (1). The program instantly estimates the mean value and 95% quantiles under H_0 for the F-statistic, squared multiple correlation coefficient, and critical p-values for the model and for the top predictor. The program also includes an option of adjustment of the spurious p-values for the whole model and for a top predictor obtained in the experiment taking into account MS. In case of RS comparison, the program calculates mean values of the Student's t-statistic and of corresponding p-value under H_0 and recalculates the actual p-value based on the distribution of $\mathbf{T} \mid H_0$.

The major limitation of our theoretical formalism and, consequently, of the program stems from the $n \to \infty$ assumption. If sample size is very small ($n < 30$), formula 4 yields the F-statistic underestimated for 30-40%, depending on k, according to our computer simulations (data not shown). The problem of the F underestimation is negligible, however, if sample size exceeds 50.

VII. Discussion

Our work demonstrates that when MS is performed, the F-statistic under H_0 is distributed as a mean of k top variables selected out of m independent random variables having a χ_1^2 distribution. It is an intuitively expected and a clear finding. If MS is not performed (k=m), $E(F \mid H_0) = E(\chi_1^2) = 1$, precluding false-positive result. However, if the $m >> k$ condition takes place, $E(F \mid H_0)$ may become arbitrarily large that brings about a misleading result.

Besides of theoretical interest, our work resulted in a practical implementation in a form of software solution, which may be useful for a wide range of scientific investigations, including but not limited to genetic association studies and those employing regression analysis with stepwise selection (common in e.g. psychology, ecology and economics). For example, a regression model for 100 points with 3 predictors selected from the initial set of 15 covariates returns a p-value of 0.005 for the model. Having entered three values {100, 15, 3} to the program, one would find that under the null hypothesis the expected mean p-value for such a model is 0.025 and the 5%-quantile is 0.0019. This means that in fact the obtained model cannot be considered statistically significant. Entering the p-value of 0.005 to the program will result in the adjusted p-value of 0.155 for the model. Note that, among 10 independent covariates, there will be, on average, 2 top predictors for which the model (1) is formally "significant": $p \approx 0.04$ if $n = 100$.

Chance of obtaining a false-positive result is also high in the studies addressing risk scores. Equation (8) shows that when two samples are compared by RS, five factors are sufficient to obtain a "significant" difference using Student's t-test. Indeed, when $k = m = 5$,

$$\mathbf{T} \mid H_0 \sim \mathbf{N}\left(\sqrt{2 \cdot 5/\pi},\ 1 - 2/\pi\right) \approx N(1.78, 0.36),$$

which corresponds to the mean p-value of 0.037. Numerous examples of works with strongly

overestimated significance of the effects based on GRS are given in our earlier work (Rubanovich, Khromov-Borisov, 2016).

It should be acknowledged that the formulas (4) only partly solve the problem of null distribution of the test statistic under model selection. More sophisticated algorithms may lead to the higher values of $R^2 \mid H_0$ than MS. This is especially noticeable for the methods based on sparsity solutions such as Lasso or Compressed Sensing (CS). Our computer simulations for $k > 3$ and $m > 200$ demonstrate that usually $E(\mathbf{R}^2_{MS} \mid H_0) < E(\mathbf{R}^2_{CS} \mid H_0)$. This hints that if model selection is preformed using the advanced methods, actual critical *p*-values can be even lower than those obtained with MS. However, theoretical and analytical basis for these situations is not available so far.

In conclusion, our work provides formal explanation of the distribution of the test statistic for multiple regression models which utilize a subset of top predictors derived from a large set of explanatory variables through the preliminary marginal screening. Although some approximations are used, the approach has sufficient accuracy. Software implementation of the analytical expressions is available and can be conveniently used even by an unexperienced researcher willing to evaluate the validity of own finding. We believe this implementation is a useful means of avoiding false-positive results in the studies, which might have been flawed by no fault of the investigators but rather by insufficiency of theoretical and practical solutions for the problem of "multiple testing" in regression analysis.

Supplementary Materials

Supplement 1. Computer simulations.

Supplement 2. H0_Model_Selection.xlsx spreadsheet

Acknowledgments

Conflict of Interest: None declared.

Funding

This work was supported in part by KAKENHI Grant Number 15K09438 from the Japan Society for the Promotion of Science (JSPS).

References Références Referencias

1. Ahsanullah, M., Nevzorov, V.N., and Shakil M. (2013). An Introduction to Order Statistics. Amsterdam – Paris – Beijing, Atlantis Press.
2. Alam, K., and Wallenius, K. (1979). Distribution of a sum of order statistics. *Scandinavian Journal of Statistics* 6, 123-126.
3. Arnold, B.C., Balakrishnan, N., and Nagaraja, H.N. (2008). A First Course in Order Statistics. SIAM-Society for Industrial and Applied Mathematics.
4. Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. 2nd ed, Routledge. Cohen, J. (1992). A Power Primer. *Psychological Bulletin* 112, 155-159.
5. Diehr, G., and Hoflin, D.R. (1974). Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics* 16, 317-320.
6. Fan, J., Shao, Q., and Zhou, W. (2017). Are Discoveries Spurious? Distributions of Maximum Spurious Correlations and Their Applications. arXiv:1502.04237 [math.ST]
7. Foster, D.P., and Stine, R.A. (2006). Honest confidence intervals for the error variance in stepwise regression. *Journal of Economic and Social Measurement* 31, 89-102.
8. Freedman, D.A. (1983). A Note on Screening Regression Equations. *The American Statistician*, 37, 152-155.
9. Genovese, C.R., Jin, J., and Wasserman, L. (2009). Revisiting marginal regression. *arXiv*:0911.4080v1 [math.ST].
10. Genovese, C.R., Jin, J., Wasserman, L., and Yao, Z.A. (2012). Comparison of the lasso and marginal regression. *Journal of Machine Learning Research* 13, 2107-2143.
11. Hastie, T., and Tibshirani, R. (2003). Expression arrays and the $p >> n$ problem. Available at https://web.stanford.edu/~hastie/Papers/pgtn.pdf
12. Lee, J.D., Taylor, J.D. (2014). Exact Post Model Selection Inference for Marginal Screening. arXiv:1402.5596 [stat.ME].
13. Leek, J. (2016). Prediction: the lasso vs just using the top 10 predictors. *Blog "Simply Statistics"*. Available at http://simplystatistics.tumblr.com/post/18132467723/prediction-the-lasso-vs-just-using-the-top-10.
14. Lukacs, P.M., Burnham, K.P., and Anderson, D.R. (2010). Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics* 62, 117–125.
15. Nagaraja, H.N. (1980). Contributions to the theory of the selection differential and to order Statistics. Dissertation. Digital Repository @ Iowa State University. Available at http://lib.dr.iastate.edu/rtd/6746/
16. Nagaraja, H.N. (1982). Some Nondegenerate Limit Laws for the Selection Differential," *Annals of Statistics* 10, 1306-1310.
17. Nagaraja, H.N. (2006). Order Statistics from Independent Exponential Random Variables and the Sum of the Top Order Statistics. In "Advances in Distribution Theory, Order Statistics, and Inference", eds. N. Balakrishnan, J.M. Sarabia, and E. Castillo.
18. Rencher, A.C., and Pun, F.C. (1980). Inflation of R^2 in best subset regression. *Technometrics* 22, 49-53.
19. Rubanovich, A.V., and Khromov-Borisov N.N. (2016). Genetic risk assessment of the joint effect

of several genes: critical appraisal. *Russian Journal of Genetics* 52, 757–769.
20. Salt, D.W., Ajmani, S., Crichton, R., and Livingstone, D.J. (2007). An improved approximation to the estimation of the critical F values in best subset regression. *Journal of Chemical Information and Modelingl* 47, 143-149.
21. Stigler, S.M. (1973). The asymptotic distribution of trimmed mean. *Annals of Statistics* 1, 472-477.
22. Tibshirani, R.J., Taylor, J., Richard Lockhart, R., Tibshirani, R. (2016). Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association*, 111:514, 600-620.
23. Windle, M. (ed.) (2016). Statistical Approaches to Gene x Environment Interactions for Complex Phenotypes. MIT Press.
24. Wray, N.R., Yang, J., Hayes, B.J., Wray, N.R., Yang, J., Hayes, B.J., Price. A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* 14, 507-515.

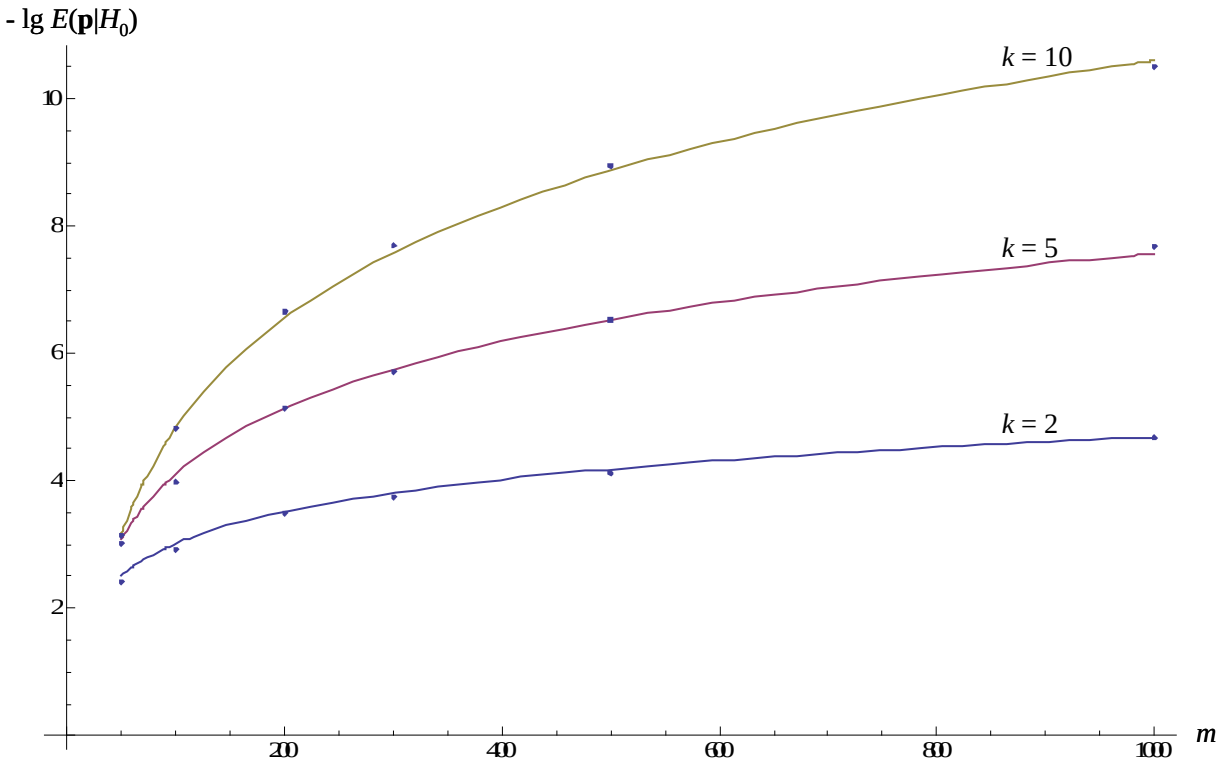

Fig. 1: Expected spurious *p*-value for the regression model (1) with regard to the total number of independent variables (*m*) and the number of top of them (*k*) used in the model under H_0 (n = 200). Plotted points are the result of numerical simulations (averaged for 1000 random repeats)

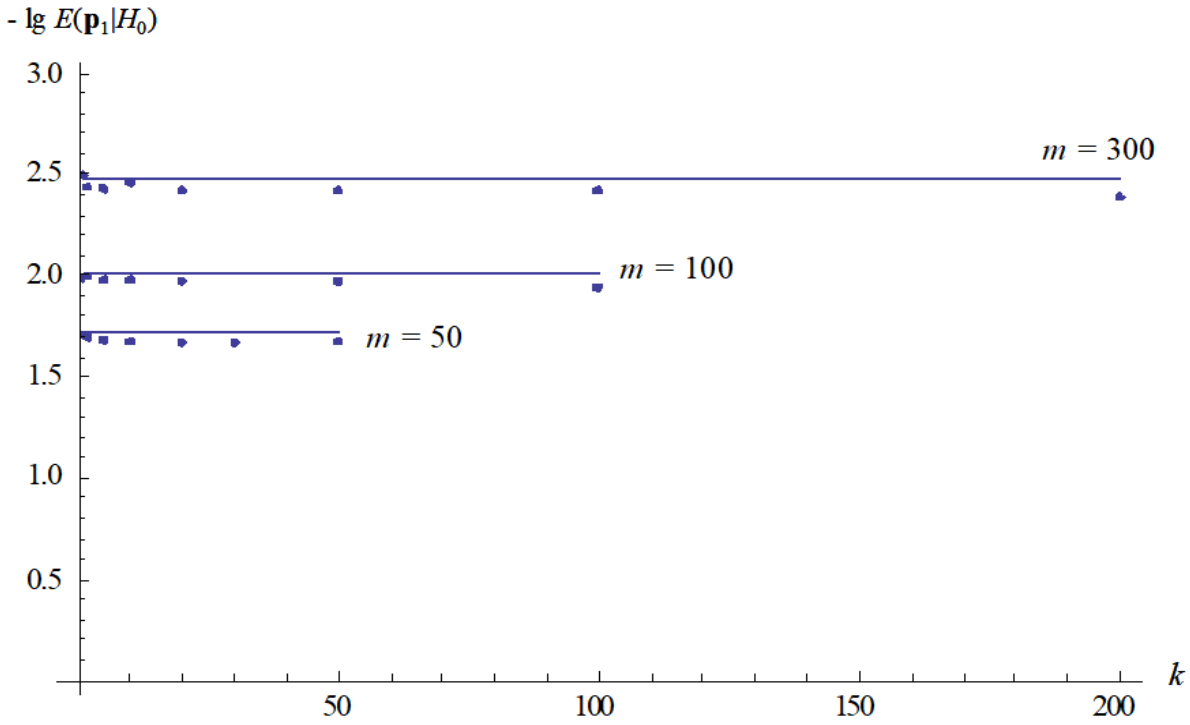

Fig. 2: Relationship between the spurious p-value for the top predictor (p_1) and the number of top predictors (k) used in the model (1) under H_0 ($n = 500$). Plotted points are the result of numerical simulations (averaged for 1000 random repeats). Horizontal lines represent the result described in Corollary 3.2: $E(p_1|H_0) \approx 1/(m + 1)$. When $n > m$, the mean $E(p_1|H_0)$ only weakly depends on k

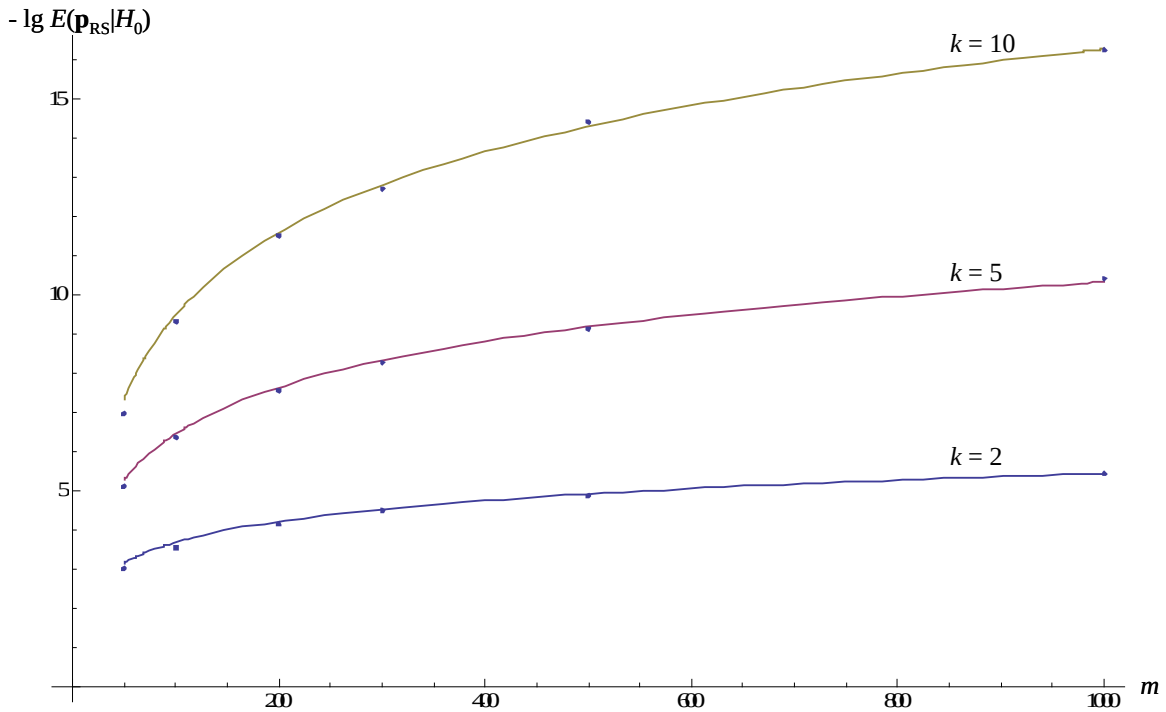

Fig. 3: Expected spurious p-value for the risk score (X_{RS}) comparison using Student's t-test with regard to the total number of independent variables (m) and the number of top of them (k) used to calculate the X_{RS} under H_0 ($n = 200$). Plotted points are the result of numerical simulations (averaged for 1000 random repeats)

Supplement 2
File H0_Model_Selection.xlsx