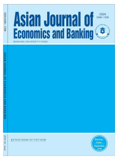
Asian Journal of Economics and Banking

ISSN 2588-1396

http://ajeb.buh.edu.vn/Home



Clarifying ASA's View on P-Values in Hypothesis Testing

William M. Briggs†, Hung T. Nguyen[1,2]

[1]Department of Mathematical Sciences, New Mexico State University, USA

[2]Faculty of Economics, Chiang Mai University, Thailand

Article Info

Received: 17/01/2019
Accepted: 17/06/2019
Available online: In Press

Keywords

Bayesian testing, Fisher's significance testing, Hypothesis testing, LASSO, Linear regression, Neyman-Pearson's hypothesis test, NHST, P-values

JEL classification

C1, C11, C12

MSC2010 classification

62F03, 62F15, 62J05

Abstract

This paper aims at clarifying both the ASA's Statements on P-values (2016) and the recent *The American Statistician* (TAS) special issue on "Statistical inference in the 21st century: Moving to a world beyond $p < 0.05$" (2019), as well as the US National Academy of Science's recent "*Reproducibility and Replicability in Science*" (2019). These documents, as a worldwide announcement, put a final end to the use of the notion of P-values in frequentist testing of statistical hypotheses.

Statisticians might get the impression that abandoning P-values only affects Fisher's significance testing, and not Neyman-Pearson's (N-P) hypothesis testing since these two "theories" of (frequentist) testing are different, although they are put in a combined testing theory called Null Hypothesis Significance Testing (NHST). Such an impression might be gained because the above documents were somewhat silent on N-P testing, whose main messages are "Don't say statistically significant" and "Abandon statistical significance". They do not specifically declare "The final collapse of the Neyman-Pearson decision theoretic framework" (as previously presented in Hurlbert and Lombard [14]). Such an impression is dangerous as it might be thought that N-P testing is still valid because P-values are not used *per se* in it.

[†] Corresponding author: William M. Briggs, Independent Researcher, New York, NY, USA. Email address: matt@wmbriggs.com

1 INTRODUCTION

Christensen [9] said "It is clear that p-values can have no role in N-P testing" and "N-P testing is not based on proof by contradiction as is Fisherian testing". Worse, the author had other misunderstandings about hypothesis testing which are dangerous for applied statisticians, exemplified by statements such as "One on the famous controversies in statistics is the dispute between Fisher and Neyman-Pearson about the *proper way to conduct a test*" (wrong, they conducted their test in the same way, using P-values, although their "frameworks" are different, noting that only Bayesians conduct their Bayesian tests differently!); "I am exposing a logical basis for testing that is distinct from N-P theory and that is related to Fisher's views" (It is clear that while Fisher's test and N-P's test are different in structure, they have the same testing philosophy, i.e., using the same (wrong) logic to conduct their tests). We will elaborate in details on these dangerous misunderstandings, for the good of applied statistics.

Thus, by "clarification" of ASA's announcements on P-values, we specifically spell out its "implicit implication", loud and clear, that "*N-P testing theory dies together with P-values".*

In view of the retirement of P-values from hypothesis testing which is the core of statistical inference, we will also address some "urgent" issues for applied statisticians in this 21st century (i.e., statistics without P-values) such as "How to test if you must?" (Answer: Use Bayesian testing, at least for the moment, because it is not wrong logically), and "How to do covariate selection in linear regression without P-values?" (Answer: Use LASSO).

In summary, we are talking about *statistics without P-values* for this 21st century. In fact, this revolution (or rather, this progress) in statistics, which is at least as significant as the one caused by the James-Stein estimator in 1961, has taken shape before the ASA's announcements, exemplified by publications such as "HCI Statistics without p-values" (Dragicevis [11]).

2 NEYMAN-PEARSON TESTING BASED ON P-VALUES

By now, statisticians should be, not only, aware of the "p-value crisis" (finally revealed through the serious problem of reproducibility and replicability of published results based on hypothesis testing, see e.g., Reproducibilty and Replicability in Science, 2019), but also understand of what to do next.

The message in ASA (2016) and (2019) is clear "Do not use p-values to conduct tests", see also Mcshan et al. [19]. Now, although we can formulate various kinds of testing problems, for each of them, we still need to specify, logically, how to carry out a test in it. A test is trusted if at least the "rule" to carry it out (i.e., jump to a conclusion) is logical, as, unlike statistical estimation and prediction, testing of hypotheses, an inference precedure, is not based on mathematical theorems, but only on logic (reasoning). Clearly, there are at least two kinds of testing frameworks: frequentist and Bayesian, as there are two such "schools of thought" in statistics! Bayesians do not need p-values to carry out their tests, they use Bayes factors instead.

Thus, only frequentist testing uses p-values to conduct frequentist testing problems.

The first frequentist testing framework is Fisher's "test of significance". Its structure is this.

Suppose a student asks "what kinds of tests do we use p-values to conduct?". Well, a teacher will immediately replies "tests of significance" because, not only the notion of p-value was born precisely to carry out such tests, but also this kind of tests is easy to explain why it needs p-values!

Roughly speaking, a statistical hypothesis is an assertion about the distribution of a random variable. As the distribution of a random variable plays the role of the law governing its dynamics, an analogy with physics is obvious. However, except quantum mechanics, natural science is deterministic, whereas in social sciences, we face uncertainty.

In "significance testing", we wish to find out whether a claim, called a hypothesis, can be confirmed. For that, we consider its negation, called a null hypothesis, denoted by H_o under which the distribution of the random variable of interest is known. Thus, we have one hypothesis with known distribution. We gather data from the variable and wish to find a way to "infer" that the data tell us that H_o could be "rejected" or not. If H_o is rejected, then we declare that our original claim is "significant", i.e., believable. This is a test about the "significance" of a claim.

The problem is "how to carry out such a test?". Fisher told us to do the following (such as in his "Lady tasting tea" story). Choose a statistic $T(X)$ to see whether its observed data is "consistent" or not with the known distribution of X under H_o. This "consistency" is measured by the probability $p(x) = P(T(X) \geq T(x)|H_o)$, where x is the observed data and the notation $(.|H_o)$ refers to "under H_o", i.e., when H_o is true (and not a conditional distribution!). This probability is called the *p-value* (of $T(X)$ when we observe x, where of course, p stands for probability). In general, the statistic $T(X)$ is chosen so that its large values reflect somehow the inconsistency of the data with respect to H_o.

Remark. Since

$$\begin{aligned} p(x) &= P(T(X) \geq T(x)|H_o) \\ &= P(-T(X) \leq -T(x)|H_o) \end{aligned}$$

where $P(-T(X) \leq -T(x)|H_o)$ is the value of the distribution function of the random variable $-T(X)$, under H_o, evaluated at $-T(x)$, i.e., $= F_{(-T(X)|H_0)}(-T(x))$, $p(X)$ is a statistic (taking values in $[0,1]$) equal to the statistic $F_{(-T(X)|H_0)}(-T(X))$ which is the probability integral transform of the random variable $-T(X)$, and hence stochastically dominates the uniform random variable on $[0,1]$, i.e., under H_o, we have $P(p(X) \leq \alpha|H_o) \leq \alpha$, for any $\alpha \in [0,1]$. See also Casella and Berger (2002), Rougier (2019).

Now, if the observed event is rare, i.e., has a very small chance to occur under H_o, and we got it, then it is not consistent with H_o, and that could "indicate" that H_o is not true. This type of reasoning can be rephrased as:

"If H_o is true, then the event is unlikely to occur, The event occured, then H_o is false".

which at first glance seems similar to a proof by contradiction in mathematics (or modus tollens in $0-1$ logic). Note right away that, it is well-known by now, among other reasons, the main one which destroys p-value as an inferential engine to conduct tests is that this "proof by contradiction" is not valid outside of binary logic. See also Nguyen [22].

To implement this (wrong) logic, Fisher first "defuzzified" the linguistic (fuzzy) term "unlikely" by putting a threshold $\alpha \in [0,1]$, some small (probability) number representing the chance of occurence for an event which can be considered as "rare". A threshold such as α is called a significance level, e.g., $\alpha = 0.05$.

The Fisher's testing procedure (i.e., jump to conclusion/make a decision) just consists simply of comparing the observed p-value (of the test statistics) with the given significance level, for example, if $p < \alpha$, reject H_o and declare that the test is (statistically) significant, so that the original "claim of interest" can be believed to be confirmed. Otherwise, the claim cannot be confirmed.

Fisher's testing is viewed as an "inference" since it leads to confirmation of a claim from data. Note however, while the focus is only on one hypothesis H_o, though *in practice but not in theory* there is a hidden hypothesis in the background, namely the negation H_o^c of H_o, but Fisher's program is not about choosing between these two hypotheses, a decision (or selection) problem (a behavior).

This point is crucial to understand. Under Fisher's tests of significance there is "only one hypothesis", as Christensen [9] emphasizes. This means something like the following. Suppose we know that under the model H_o the chance of seeing x is as small as you like, but not impossible. We see x. what can we conclude? Nothing, except the tautology, that since H_o is given, H_o is (locally) true.

If there *truly* is no alternative hypothesis, it is impossible to conclude anything except that H_o is true. One possible alternative hypothesis often considered is that "Something other than H_o is true" or its negation H_o^c. But we do not consider this alternative hypothesis under Fisher. Fisher says there are *no* alternative hypothesis, not even H_o^c. We start with H_o; H_o is all there is; we cannot move from H_o. Using a p-value is nothing but an act of will. This was Neyman's original critiscism, and which is formally proved in Briggs [4].

Obviously, people *do* consider alternative hypothesis , even informal ones like H_o^c. This is to say, nobody treats Fisherian tests in a logical manner. H_o^c is incredible vague; in cases with continuous parameterized probability models, it is infinitely vague. Suppose H_o insists a certain parameter in the model under consideration equals 0. This means, and here is a subtle point, that the vagueness is not-0 (say), but where the parameter is thought to be in definite range or value.

That means nobody really believes in a blanket H_o^c, but in a much more concrete alternative, even if this alternative is "the parameter is greater than 0". Once that is done (mentally), testing becomes of the Neyman-Pearson type, as shown on paper. Thus every use of Fisherian testing is by use or in practice a form of N-P testing. Again, this must be so. For if *all* we believe or know or are considering is H_o, then H_o is all we have. The moment we allow for hypotheses that are different from H_o, we chuck out p-values and test in a different way.

A follow-up on Fisher's test of significance is *Neyman-Pearson's "test of hypotheses"* which is formulated in a decision framework. It is a problem of choosing between two hypotheses H_o and H_a, again using a data-based procedure $T(X)$, where H_a needs not be H_o^c. The new ingredient in the framework is two types of error, designed to control error in making decsions "in the long run". Note right away that such a decision-framework seems appropiate for situations such as in statistical quality control where a decision must be made which could be wrong, and some "guarantee" is needed.

Thus, consider two types of error when making decisions: the type-I error $\alpha = P(\text{Reject } H_o|H_o \text{ is true})$, and type-II error $\beta = P(\text{Accept } H_o|H_o \text{ is false})$, and find a way to conduct the test, i.e., a decision rule of rejecting or accepting H_o based on a statistic $T(X)$.

The N-P testing procedure is this. Specify in advance $\alpha \in [0,1]$, find a test statistics $T(X)$ so that $1 - \beta = P(\text{Accept } H_a|H_o \text{ is false})$ is as large as possible. This amounts to define a rejection region R_α determined by $P(T(X) \in R_\alpha|H_o) \leq \alpha$, so that the decision rule (i.e., the way to carry out the test) : If $T(X) \in R_\alpha$, reject H_o (hence, choose H_a); otherwise choose H_o.

What is the difference with Fisher's significance testing that is often referred to as the "incompatibility" among the two types of testing framework (an argument against putting these two frameworks together to form the *Null Hypothesis Significance Testing/ NHST* that text books even did not mention in their chapter on hypothesis testing)?

That difference is simply between Fisher's level of significance α, and N-P's type-I error α (N-P should not use the same notation α !). But what is the big deal about that? Suppose we use N-P framework with type-I error α. To conduct a N-P test means to determine the rejection region R_α. Once R_α is determined, the statistician looks at the value $T(x)$: If $T(x) \in R_\alpha$, she rejects H_o and takes H_a, protecting her from making the wrong decision with probability α (in a long run).

But, for example, for a rejection R_α of the form $R_\alpha = \{T(X) > t_\alpha\}$, i.e., $P(\{T(X) > t_\alpha\}|H_o) = \alpha$, it is determined simply by t_α which is the $\alpha-$ quantile of the distribution of $T(X)$ under H_o (i.e., the distribution of the statistic $T(X)$ when H_o is true), resulting in rejecting H_o when $T(x) > t_\alpha$, and this is strictly equivalent to *p-value* $= P(T(X) > T(x)|H_o) \leq \alpha$, regardless the meaning of α (it is just a number in $[0,1]$). α is just a threshold. See also Lehmann [18] and Kennedy-shaffer [16]. As as matter of fact, McShane et al. [19] stated "We propose to drop NHST paradigm-and the p-value threshold *in-*

trinsic to it".

In summary, the logic of N-P testing is based on P-value with threshold α, and hence it is based on a wrong "proof by contradiction", just like Fisher's significance testing.

In other words, while the frameworks and purposes are different, Fisher's test and N-P's test use the same *logic to conduct their tests, namely using p-values*.

3 HOW TO TEST WITHOUT P-VALUES IF YOU MUST?

One fact is do not test in the conventional sense and to cast problems in their predictive sense. If the statistician has two (or more) competing models for an observable y in mind, there are only two possibilities. The first is that uncertainty in not-yet-seen (usually future) values of y needs to be quantified. The second is guessing which process or cause was responsible for observed results. Both arfe predictions. See also Billheimer [1].

Suppose two models are under consideration, H_o and H_a. If there is no other prior information other than there are only these two possibilities, andonly these two possibilities, then by the statistical syllogism $P(H_o|B) = P(H_a|B) = 1/2$. Of course, the background information (B) could be different such that one model more receives more weight. Then

$$\begin{aligned} P(y \in s|B) = P(y \in s|H_oB)P(H_o|B) \\ + P(y \in s|H_aB)P(H_a|B) \quad (1) \end{aligned}$$

where s is a subset of interest of the observable y. If data D has been taken, then (1) becomes

$$\begin{aligned} &P(y \in s|DB) = \\ &P(y \in s|DH_oB)P(H_o|DB) \\ &+ P(y \in s|DH_aB)P(H_a|DB) \quad (2) \end{aligned}$$

Either (1) or (2) can be expanded in the obvious way for more than two models. In other words, the full uncertainty of the situation is considered and used to make predictions of the observable y. No choice need be made of any model; i.e.,no testing need be done.

The second idea is to calculate $P(H_o|DB)$ and $P(H_a|DB)$, which is extensible to more models in the obvious way. To decide between them is not solely a matter of picking which has the higher probability, for to make a decision requires considering cost and loss. If the cost-loss is symmetric, then picking the model with the highest posterior probability it the best bet.

For a handy, but potentially misleading, one number summary, the probability ration can also be calculated:

$$\frac{P(H_o|DB)}{P(H_a|DB)} \quad (3)$$

and this is equivalent to a Bayes factor (BF). See, e.g., Kock [17] for Bayesian Statistics, and Nguyen [23].

The BF is

$$\frac{P(D|H_oB)}{P(D|H_aB)} = \frac{P(H_o|DB)}{P(H_a|DB)} \times \frac{P(H_a|B)}{P(H_o|B)} (4)$$

If $P(H_o|B) = P(H_a|B) = 1/2$ then (3) is equivalent to (4). Now the model posterior for H_o is

$$P(D|H_oB) = \frac{P(D|H_oB)P(H_o|B)}{P(D|B)} \quad (5)$$

A similar calculation gives the posterior for H_a. Thus (3) is equivalent to

$$\frac{P(H_o|DB)}{P(H_a|DB)} = \frac{P(D|H_oB)P(H_o|B)}{P(D|H_aB)P(H_a|B)} \quad (6)$$

There is thus no logical difference in using the PR (probability ratio) or BF. The difference is emphasis, or in the ease of cinveying understanding. The PR is stated directly in terms of the probabilities of the models, which is after all what the decision is about: which is most likely true given the evidence? The BF is motivated by p-value like thinking. It asks for the probability of the observations, which while it is the same, puts the question the wrong way around because our goal is to make a decision about the model, not the data.

The warning about the real goal of the analysis cannot be understated. Often testing is done when what is really desired is quantifying uncertainty in the observable y. In that case, no testing is needed at all. The first method is applicable, and should be used. Too often scientists and statisticians think that they must *always* select between alternatives, even when the goal is not to pick the one best model. Picking the best model (in the sense of most likely, or by other decision analysis) is thus bound to led to over-certainty, even dramatic over-certainty when the number of models considered is greater than two. Which is most often the case in most problems.

Often what's really wanted is the ability, as in regression below, to make statements $P(y \in s|xDB)$ where $x = (x_1, x_2, ..., x_k)$ are covariates of y. How much does the probability of y change for a change in some x_i? That's almost always the science under question. The model doesn't appear in that statement unless there is only one model or hypothesis under consideration, in which case we write $P(y \in s|xHDB)$. If there is more than one model, then we have (1), or the version of that equation expanded for more than two models with the conditioning on x, i.e.,

$$P(y \in s|xB) = \sum_i P(y \in s|xH_iB) \times P(H_i|B) \quad (7)$$

In the best scientific sense, there is no sense in throwing out via testing any H_i that is implied by the background information B. This is discussed in more depth in Briggs [4]. See also Nuitj [24].

For more additional recent discussions on p-values and hypothesis testing, see e.g., Briggs ([3], [5], [6]), and Briggs, Nguyen and Trafimow [7].

4 LINEAR REGRESSION ANALYSIS WITHOUT P-VALUES

The ASA's documents (2019) mark the new statistics for this 21st century, a statistics without P-values. Let the past rest in peace. As already stated in recent literature, from now on we will not see publications involving statistics with hypothesis testing using P-values anymore. Let's move ahead to make the public trust scientific results based on statistics.

The lesson learned is simply this. Statistical methods need to be trusted. They should be founded upon logical reasoning, and empirical results coming out from them must be cleanly explained.

Having said that, we face an urgent task facing both education and research, namely how to "handle" linear regression analysis, the Bread-and- Butter (BB) tool of applied statistics, once P-values can be no longer "allowed" to use to conduct tests (for covariate selection)?

Clearly, testing in linear models is a typical situation where statisticians usually have to face. As we will see, it turns out that it seems that we are somewhat lucky to answer the question "How to test in linear regression?" simply as "Do not test, you don't have to". And that is because we have a modern method of estimation in linear models, called LASSO (Least Absolute Shrinkage and Selection Operator), due to Tibshirani [27]. Thus, in a sense, in the search for ways to do linear regression without p-values, we encounter modern estimation methods improving traditional Ordinary Least Squares (OLS) method of classical statistics.

In this section, we will be a bit tutorial on the road leading to LASSO, a type of supervised machine learning method to do parametric linear regression without p-values.

One popular situation in (statistical) model building is this. We have a response (scalar) variable Y of interest, for the sake of simplicity, and wish to describe, explain, predict and intervene (the four main goals of a scientific investigation, as spelled out in the US National Academy of Science's recent "*Reproducibility and Replicability in Science*", 2019). For that, we look for covariates (factors, not necessarily the causes) which, we "think" , could affect Y. Suppose the covariates that we can consider are $X_{.1}, X_{.2}, ..., X_{.k}$. Of course we are not sure either they are all "relevant", i.e., really contribute to Y or not, or there are other "relevant" covariates that we did not include in this set of covariates. The former issue is termed "covariate selection problem" (or subset selection), in the spirit of the principle of parsimony (Occam's razor), necessary especially for high-dimensional data (much more covariates than sample size); the latter is another effort to possibly improve a given model (in the context of nested models).

One thing at a time! Let's see first how we can come up with a "good" model for prediction purposes, even temporarily (to be improved later), when we have at our disposal, the set $\{X_{.1}, X_{.2}, ..., X_{.k}\}$ of covariates. Since we are going to predict Y based on $X_{.1}, X_{.2}, ..., X_{.k}$, we could consider the conditional mean $E(Y|X_{.1}, X_{.2}, ..., X_{.k})$, which is a function of the covariates, i.e., a statistic), if it exists of course! Suppose $E(Y|X_{.1}, X_{.2}, ..., X_{.k})$ exists and we take it as our predictor. Just like an estimator, we need to judge its performance which is its prediction error. Suppose, in addition, that all variables involved have finite second moments, so that the prediction error of $E(Y|X_{.1}, X_{.2}, ..., X_{.k})$ can be taken as its mean squared error (MSE). In this case, it is a mathematical theorem that

$E(Y|X_{.1}, X_{.2,...}, X_{.k})$ is the best predictor in the MSE sense.

An obvious approximation to $E(Y|X_{.1}, X_{.2,...}, X_{.k})$ is the linear model $X'\beta = \sum_{j=1}^{k} \beta_k X_{.j}$, where $\beta = (\beta_1, \beta_2, ..., \beta_k)' \in \mathbb{R}^k$ (where $(.)'$ denotes transpose), $X = (X_{.1}, X_{.2,...}, X_{.k})'$, with by abuse of language, or refering to history (F. Galton's early work on heredity), we call this linear model a linear regression model. To accomodate for possible deviations from the true relationship, we add a random component e to obtain our statistical linear regression model $Y = X'\beta + \mathbf{e}$ with the assumption $E(X|e) = 0$, so that we do have $E(Y|X) = X'\beta$.

Of course, we need to validate such a linear model before using it! Suppose we observe data on the covariates as (Y_i, X_{ij}), $j = 1, 2, ..., k$; $i = 1, 2, ..., n$, so that

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + e_i$$

For $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)' \in \mathbb{R}^n$, $\mathbf{e} = (e_1, e_2, ..., e_n)' \in \mathbb{R}^n, \beta = (\beta_1, \beta_2, ..., \beta_k)'$, and the $(n \times k)$ data matrix

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & . & . & X_{1k} \\ X_{21} & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ X_{n1} & X_{n2} & . & . & X_{nk} \end{bmatrix}$$

The matrix form of the above is

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

Having a model in place, we proceed now to "specify" it for applications, i.e., to estimate the model parameter β from the data matrix $\mathbf{X}$.

Traditionally, for a linear model, we estimate its parameter by OLS which is the same as that of Maximum Likelihood (MLE) when the random error is assumed to be normally distributed, and that consists of minimizing the convex objective function

$$\beta \to \varphi(\beta) = ||\mathbf{Y} - \mathbf{X}\beta||_2^2$$

over $\beta \in \mathbb{R}^k$, where $||.||_2$ denotes the L_2-norm of $\mathbb{R}^n$.

Just like MLE where only for regular models that their MLE are "trusted" (since at least, they are consistent estimators), OLS is not applicable universally, i.e., there cases where OLS estimators do not exist. Indeed, the "normal" equation of OLS method is

$$(\mathbf{X'X})\beta = \mathbf{X'Y}$$

There are two cases:

(i) Only if $\mathbf{X}$ is of full column rank then $(\mathbf{X'X})^{-1}$ exists, and the OLS estimator $\hat{\beta}$ of β exists and is unique, given, in closed form, by

$$\hat{\beta} = (\mathbf{X'X})^{-1}(\mathbf{X'Y})$$

(ii) If not, we do not have OLS estimator! i.e., we cannot use OLS method to estimate parameters in our linear model! The "practical consequence" is : the expression $(\mathbf{X'X})^{-1}(\mathbf{X'Y})$ cannot be evaluated numerically (in software)! For example, in high dimensional data $(k > n)$, model parameters cannot be estimated by OLS.

What should we do then? Well, if $(\mathbf{X'X})$ is not inversible, you can obtain a "pseudo-solution" (not unique) by using a "pseudo-inverse" M of $(\mathbf{X'X})$ (e.g., Moore-Penrose), at the place of

$(\mathbf{X}'\mathbf{X})^{-1}$, i.e., a matrix M such that $\mathbf{X}'\mathbf{X}M\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}$. Specifically, the solution of the normal equation is only determined up to an element of a non trivial space V , i.e., $M(\mathbf{X}'\mathbf{Y}) + v$, for any $v \in V$. Thus, there is no unique estimator of β by OLS. But when solutions are not unique, we run into the serious problem of "model identifiability".

Roughly speaking, among all vector $\beta \in \mathbb{R}^k$ which minimize $||\mathbf{Y} = \mathbf{X}\beta||_2^2$ (a convex function in β), the one with shortest norm $||\beta||_2$ is $\beta = \mathbf{X}^*\mathbf{Y}$ (viewing as "a solution for the least squares problem") where $\mathbf{X}^*$ is the pseudo-inverse of $\mathbf{X}$. Using the *singular value decomposition (*SVD) of $\mathbf{X}$, this pseudo-inverse is easily computed.

Remark. In the past (where by the "past", we mean before 1970, the year where Ridge Regression was discoverd by Hoerl and Kennard [13], precisely to handle this "non existence of OLS solution", but, as "usually", awareness of new progress in science, in general, is slow; exemplified right now with the "ban" of using P-values in hypothesis testing!), statisticans and mathematicians tried to "save" the OLS (as a "golden culture" of statistics since Gauss) by proposing the SVD of matrices as a way to produce the pseudo-inverse of the data matrix $\mathbf{X}$, so that you still can use OLS, even its solutions so obtained are not unique. But, non - uniqueness is a "big" problem in statistics as it cretates the non-identifiability problem!

Note also that there is another alternative to OLS, called "Partial Least Squares" (PLS), generalizing principal component analysis, which seems somewhat "popular" in applied research, especially with high-dimensional data. However, like OLS and Ridge Regression, the analysis using PLS involves hypothesis testing using P-values.

Now, even in case where OLS estimator exists, are you really satisfied with it? You might say "what a question!" since by Gauss-Markov theorem, OLS estimator is a BLUE! Well, we all know that the notion of unbiased estimators was invented to have a "theory" of estimation in which we can claim there is a best estimator, in MSE sense, and not to rule out "bad" estimators, since "unbiasedness" does not mean "good". This is so since, afer all, the performance of an estimator is judged by its MSE only.

It took a research work like that of James and Stein [15] for statisticians to change their mind that biased estimators could be even better than unbiased ones. But that is a good sign! Statisticians should behave nicely, and correctly like physicists! There should be no "in defense of p-values"!

Now, since an OLS estimator is a MLE estimator, it can be improved by the shrinkage technique of James and Stein. Thus, there is a hope to improve unbiased OLS estimators by biased shrinkage estimators. Although originally considered to solve the uniqueness of solution of OLS, namely, replacing, in an ad-hoc manner, the possible OLS solution $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$ by $(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})^{-1}(\mathbf{X}'\mathbf{Y})$, where $\lambda > 0$, and $\mathbf{I}$ denotes the identity matrix of $\mathbb{R}^k$, since the matrix $\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}$ is always invertible (adding the positive definite matrix $\lambda\mathbf{I}_k$, for some $\lambda > 0$, to the semi-positive definite matrix $\mathbf{X'X}$ will make the matrix $\mathbf{X'X}+\lambda\mathbf{I}_k$ positive definite and hence invertible), this now classic ridge regression method of estimation improves OLS since it is based on shrinkage "technology". Indeed, the ridge estimator $\hat{\beta}_r(\lambda) = (\mathbf{X'X}+\lambda\mathbf{I})^{-1}(\mathbf{X'Y})$, while being the unique solution of the minimization of the strictly convex objective function $||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_2^2$ over $\beta \in \mathbb{R}^k$, is in fact, equivalently, derived from a minimization of $||\mathbf{Y} - \mathbf{X}\beta||_2^2$ under the constraint $||\beta||_2^2 \leq c$, exhibiting the shrinkage effect for its estimator. As such, ridge estimator, while being a biased estimator, has smaller MSE than OLS estimator, and that is important for prediction which is based on estimation. However, like OLS, ridge regression does not do covariate selection by itself.

How covariable selection is done in OLS regression?

Even until quite recently, you still see text books, lecture notes, and research papers with the headline *"Hypothesis testing in multiple linear regression"*. And then you wander "what happens to all these "stuff" once it is revealed that using P-values to carry out tests cannot be trusted?". Well, they are *a thing of the past.* We cannot blame them (I mean lots of them!). It is not easy to find ways to accept or reject hypotheses if we just have statistical data. And, it is not easy to see why using P-values to test is *not OK either! But now, it's done:* We will not ever use P-values to test hypotheses.

Here, we ask "Why there are tests in regression analysis, in the first place?", or, more directly "what for?"

Well, all tests of the form $H_o : \beta_j = 0$ vs $H_a : \beta_j \neq 0$, or simultaneous test $H_o : \beta_1 = \beta_2 = ... = \beta_k = 0$ vs $H_a : \beta_j \neq 0$ for some j, are designed to do covariable selection, i.e., to exclude some covariates from consideration in the model building.

Indeed, you read (and learn!) statements (from text books) like "*Tests like the above play an important role in model building. Model building is the task of selecting a subset of relevant predictors from a larger set of available predictors to build a good regression model. This kind of tests is well suited for this task, because it tests whether additional predictors contribute significantly to the quality of the model, given the predictors that are already included",* and "*P-values and coefficients in regression analysis work together to tell you which relationships in your model are "statistically significant", the p-values for the coefficients indicate whether these relationships are "statistical significant".*

Well, as spelled, loud and clear, in Wasserstein et al. [29], "*statistically significant: don't say it and don't use it",* we could just ignore the above "recommendations"! and instead, ask "If testing (following OLS estimation) is not trusted any more, what else could we do to replace it, for the sake of the important task of performing subset selection?".

Anyway, it is clear that, after using OLS to estimate the preliminary model's parameters, "*statisticians of the past"* carried out tests to do subset selection. Two things to note: OLS estimation method does not do (by itself) subset selection; the subset selection is a follow-up different procedure based on "statistical inference" (i.e., testing), although this statistical inference (i.e., the way to jump to decisions/ reject or accept hypotheses) is based on P-values which can be computed from statistical properties of OLS estimates. Using tests to exclude irrelevant covariates, however, is not a "reliable" (or not correct!) procedure, as "they" admitted "when there is multicollinearity in the data, the power of tests are very low, resulting in failing to reject a null hypothesis and hence exclude (wrongly) an important covariate".

There are probability-based methods that can be used to select covariates in regression. *None* of these should be used since uncertainty in the observable y is the main interest. As said above, in many cases covariate selection is carried out when there is no need to do so. There simply is no good reason to reject a covariate that might be informative just because a statistical threshold has been passed.

It is sometimes that covariate choice is important. Suppose a model for some medical observable y is conditioned on a covariate which is an expensive test. It would be useful to know whether adding that covariate to the model conveys useful information, conditional on the other information already in the model. If not, then some procedure to "reject" it would be of great use. If the researcher is merely unsure whether an easy-to-measure covariate should be in the model or not, then it turns out the problem is the same, as demonstrated next.

The first method to select covariates, if covariates must be selected is the following.

Ordinary regression for an observable y the uncertainty of which is characterized by a normal distribution is written like this

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{1}$$

where the x_i are the covariates under consideration. Usually one of the x_j is under special view, the other x_i thought to be a necessary part of the model. Without loss of generality, we consider that problem: there may be, for instance, no other x_i, or other x_k that are also being considered, but this framework is applicable to all these scenarios.

As above, let D be the data, $x = (x_1, x_2, \ldots, x_p)$, and let M_j be the model with x_j in it, and M_{-j} the model without x_j. Calculate the posterior predictive probabilities

$$P(y \in s | xDM_j) \tag{2}$$

and

$$P(y \in s | x_{-j}DM_{-j}) \tag{3}$$

If (2) equals (3), then x_j is (conditionally) irrelevant, and it can be excluded from the model. If the difference of (2) and (3) is "small", where *small* is defined from researcher to researcher depending on their cost and loss, then x_j is said to be unuseful, and again it can be excluded from the model. If covariate selection must be done, then the consequences of having or removing x_j from the model are thus fully, completely, and probabilistically given. Probabilities can be entered into the decision analysis, which might differ from researcher to researcher. There is not, and should not be, a probability difference that is universally "significant", as with p-values.

It should be clear this procedure works beyond just regression, but for any probability model.

If covariate selection is not crucial, and there is (as there will always be) prior knowledge on whether x_j should be in the model, then we use the *full uncertainty* of the situation. Calculate:

$$\begin{aligned} &P(y \in s|xDB) = \\ &P(y \in s|xDM_jB) \times P(M_j|xDB) \\ &+ P(y \in s|x_{-j}DM_{-j}) \times P(M_{-j}|xDB) \end{aligned}$$

where the posteriors on the model $P(M_j|xDB)$ and $P(M_{-j}|xDB)$ are calculated as in (3) etc. above.

This approach will help stem the rising tide of over-certainty, which has led to the so-calle replicability crisis. Having covariate selection where none is needed, or in failing to state the full uncertainty of covariates always causes over-certainty.

Now, there is another shrinkage method for estimating parameters in linear regression models, due to Tibshirani [27], call Least Absolute Shrinkage and Selection Operator (LASSO), similar to ridge regression, but having the additional advantage of being able to do covariate selection by itself (i.e., the covariate selection is obtained simultaneously with the estimation process, and not as a follow-up one based on testing), see Hastie et al. [12]. As Boelaret and Ollion [2] declared, it is a *Great Regression: Parametric models without p-values.* Roughly speaking, it is so, since instead of just finding the parameters that minimize the sum of squared errors, the LASSO also seeks to limit the complexity of the fitted model, by forcing some parameter estimates to be equal exactly to zero, correponding to irrelevant covariates (to be exclude from the final model building).

In the same "spirit" of ridge regression, i.e., shrinkage estimation, LASSO is an estimation method for estimating parameters in linear regression models, but by shrinking the parameters with respect to another norm, namely L^1-norm, rather than L^2-norm.

Specifically, LASSO provides a solution to the minimization under constraint problem

$$\min_{\beta\in\mathbb{R}^k}[||\mathbf{Y}-\mathbf{X}\beta||_2^2] \quad \text{subject to } ||\beta||_1 \leq t$$

Note that the objective function is the same, but the constraint is different than that of a ridge regression.

It is the change from L^2-norm to L^1- norm which provides the automatic covariate selection. Some elaborations are as follows.

Similar to ridge regression, an equivalent formulation of this optimization under constraint is

$$\min_{\beta\in\mathbb{R}^k}\{[||\mathbf{Y}-\mathbf{X}\beta||_2^2] \ + \lambda||\beta||_1\}$$

for some tunning parameter $\lambda > 0$.

However, since $||\beta||_1 = \sum_{j=1}^k |\beta_j|$, the objective function $\beta \in \mathbb{R}^k \to E[||\mathbf{Y}-\mathbf{X}\beta||_2^2] \ + \lambda||\beta||_1$, while convex (but not stricly convex, so that are possibly more than one solution), is not differentiable. And as such, there is no "close form" solution to LASSO, hence its solution should be carried out numerically.

Now, the objective function in the LASSO estimation is convex but not strictly convex, so that the LASSO estimate (of β) is not unique. However, as solutions of a convex minimization problem, the set of LASSO solutions forms a convex set in $\mathbb{R}^k$. However, as far as prediction is concerned, just as in Machine Learning (viewing LASSO as a surpervised learning algorithm), this is not a problem since the linear predictor based on LASSO is unique. Note that this situation reminds us of an analogous situation in estimation by MLE: For regular models, when the log-likelihood function has several maximizers, any one of them can be used as a MLE, since any one of them is consistent.

But, say, in Econometrics, where we are also concerned with explaining the variable of interest from its covariates, for various reasons, the non-uniqueness of LASSO's solutions should be investigated with great care. Appropriate theoretical results (see e.g., Hastie et al. [12]) are somewhat available for justifying the use of LASSO in applications, including "covariate selection consistency" issue which could be investigated in the setting of (finite) *Random Set Theory*, e.g., Nguyen [21], Das and Resnick [10], and estimation consistency, recalling that the popular neural netwoks, as also a supervised machine learning algorith, is justified by its universal approximation (Stone-Weierstrass Theorem), see e.g., Nguyen et al. [20].

In summary, the LASSO is a modern estimation method for linear regression models which the unique distinction, among all other alternatives, that it performs variable selection, togther with improved estimation, without using testing, and hence without using P-values.

In a "modern statistics world" where P-values should never be used, LASSO is the obvious tool to linear regression analysis.

Final Remark. It is "interesting" to note that there is such thing as "A significance test for the LASSO" in the lierature (but prior of 2015)! It was about testing for the "significance" of an additional covariate to a linear regression model after runing LASSO for that model (for covariate selection). We suspect that, in view of the actual ASA's documents (2019), such test will disappear from the literature? For a problem such as this, why not run again a LASSO with the new covariate to find out whether it does contribute to the response variable?

References

[1] Billheimer, D. (2019), Predictive inference and scientific reproducubility, *The American Statistician 73(51),* 291-295

[2] Boelaret, J. and Ollion, E. (2018), The Great regression. Machine learning, econometrics, and the future of quantitative social science, *hal-01841413*

[3] Briggs, W. (2015), The crisis of evidence: Why probability and statistics

cannot discover cause, *arXiv:1507.07244*

[4] Briggs, W. (2016), *Uncertainty: The Soul of Modeling, Probability, and Statistics,* Springer

[5] Briggs, W. (2017), The substitute of p-values, *Journal Amer. Statist.Assoc. 112(519),* 897-898

[6] Briggs, W. (2019), Everything wrong with p-values under one roof, in *Beyond Traditional Probabilistic Methods in Economics, Studies in Computational Intelligence 809, Springer,* 22-44

[7] Briggs, W. , Nguyen, H. T., and Trafimow, D. (2019), The replacement for hypothesis testing , in *Structural Changes and Their Econometric Modeling, Studies in Computational Intelligence 808, Springer,* 3-17

[8] Casella, G. and Berger, R. L. (2002), *Statistical Inference,* Duxbury

[9] Christensen, R. (2005), Testing Fisher, Neyman, Pearson, and Bayes, *The American Statistician 59(2),* 121-126

[10] Das, B. and Resnick, S. I. (2008), QQ plots, random sets and data from a heavy tailed distribution, *Stochastic Models 24(1),* 103-132

[11] Dragicevic, P. (2015), HIC Statistics without p-values, *Research Report # 8738, Research Centre, Saclay, France*

[12] Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations,* Chapman and Hall/ CRC Press

[13] Hoerl, A.E., and Kennard, R.W. (1970), Ridge regression: biased estimation for nonorthogonal problems, *Technometrics 12(1),* 55-67

[14] Hurlbert, S. H. and Lombardi, C. M. (2009), Final collapse of the Neyman-Pearson decision theoretic framework and the rise of the neoFisherian, *Ann. Zool. Fennici. (46),* 311-349

[15] James, W. and Stein, C. (1961) Estimation with quadratic loss, *Proceedings of the Fourth Berkeley symposium on Mathematical Statistics and Probability,* 361-379

[16] Kennedy-Shaffer, L. (2019), Before $p < 0.05$ to beyond $p < 0.05$: Using history to contextualize p-values and significance testing, *The American Statistician 73(51), 82-90*

[17] Kock, K. R. (2007). *Introduction to Bayesian Statistics,* Springer

[18] Lehmann, E.L. (1993), The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two?, *J. Amer. Statist. Assoc. (88),* 1242-1249

[19] Mcshan, B. B., Gal, D., Gelman, A.,Robert, C. and Tackett, J. L. (2019), Abandon statistical significance, *The American Statistician 73(51), 235-245*

[20] Nguyen, H. T., Prasad, N. P., Walker, C. L., and Walker, E. A. (2003), *A First Course in Fuzzy and Neural Control,* Chapman and Hall/ CRC Press

[21] Nguyen, H.T. (2006), *An Introduction to Random Sets*, Chapman and Hall/CRC Press

[22] Nguyen, H. T. (2016), On evidential measures of support for reasoning with integrated uncertainty: A lesson from the ban of p-values in statistical inference. In *Integrated Uncertainty in Knowledge Modeling and Decision Making*, LNAI 9978, Springer, 3-15

[23] Nguyen, H. T. (2019), How to test without P-values?, *Thailand Statistician,* to appear July 2019

[24] Nuitj, h. (2019), The limitations of p-values: An appeal for alternatives (Google)

[25] *Reproducibility and Replicability in Science* (2019), the National Academies Press

[26] Rougier, J. (2019), P-values, Bayes factors, and sufficiency, *The American Statistician 73(51),* 148-151

[27] Tibshirani, R. (1996), Regression shrinkage and selection via the LASSO, *J. Royal. Statist. Soc. 58(1),* 267-288

[28] Wasserstein, R. L. and Lazar, N. A. (2016), the ASA's Statement on P-values: Context, Process, and Purpose, *The American Statistician (70)*, 129-133

[29] Wasserstein, R. L., Schirm, A. L. and Lazar, N. A. (2019), Editorial: Moving to a world beyond $p < 0.05$", *The American Statistician 73(51),* 1-19