An Insight of Linear Regression Analysis

Ng Set Foong[1], Chew Yee Ming[2], Chng Pei Eng[2], Ng Kok Shien[2]

[1]*Universiti Teknologi MARA, Kampus Pasir Gudang,*
81750 Masai, Johor

[2]*Universiti Teknologi MARA, Cawangan Pulau Pinang,*
13500 Permatang Pauh, Pulau Pinang

[1]*Email: setfoong@gmail.com*

ABSTRACT

Regression models are developed in various field of applications to help researchers to predict certain variables based on other predictor variables. The dependent variables in the regression model are estimated by a number of independent variables. Model utility test is a hypothesis testing procedure in regression to verify if there is a useful relationship between the dependent variable and the independent variable. The hypothesis testing procedure that involves p-value is commonly used in model utility test. A new technique that involves coefficient of determination R^2 in model utility test is developed in this paper. The effectiveness of the model utility test in testing the significance of regression model is evaluated using simple linear regression model with the significance level α = 0.01, 0.025 and 0.05. The study in this paper shows that a regression model that is declared to be a significant model by using model utility test, however it fails to guarantee a strong linear relationship between the independent variable and dependent variable. Based on the evaluation presented in this paper, it is shown that the p-value approach in model utility test is not a good technique in evaluating the significance of a regression model. The results of this study could serve as a reference for other researchers applying regression analysis in their studies.

Keywords: *regression, model utility test, hypothesis testing, p-value, coefficient of determination*

INTRODUCTION

Regression is a statistical technique widely applied by researchers in many fields to describe the nature of the relationship between variables. The relationship between the variables can be positive or negative, linear or non-linear [1]. In regression, the variables are categorised into independent variable and dependent variable. The dependent variable is a response variable that can be predicted by the independent variable. Hence, the independent variable is also called predictor variable.

In various field of applications, regression models are developed to help researchers to predict certain variables based on other predictor variables [2]. Regression models are developed using many techniques such as simple linear regression, multiple linear regression, non-linear regression, non-parametric regression and multivariate regression. Regression is widely used in all branches of engineering. In construction of capital projects, accuracy in estimation of the cost during the early stages is very important. Trost & Oberlender [3] applied factor analysis and multivariate regression in predicting the accuracy of early cost estimates in construction projects. Clark [4] used multivariate non-parametric regression to predict the future state of traffic on a motorway. Goyal & Goyal [5] applied artificial neural engineering and regression in forecasting shelf life of instant coffee drink. Moreover, regression is also applied in the field of management and economics. Regression is widely used to forecast variables such as prices of raw materials, sales, interest rates, inflation rates, product demand and labour cost [6-10].

A hypothesis testing procedure in the regression analysis called model utility test is conducted to verify if there is a useful relationship between the dependent variable and the predictor variable. The null hypothesis of the model utility test refers to no useful relationship between the dependent variable and the predictor variable. If null hypothesis is rejected, the predictor variable is considered to be useful in the regression model. Devore [11] stated that the regression model should not be used for further inferences such as predictions of future value or estimates of mean value unless the model utility test results in rejection of null hypothesis for a suitably small significance level of the test, α.

Therefore, model utility test is an important hypothesis testing procedure to verify if the regression model is significantly useful for further inferences. A rejection of null hypothesis in the test indicates that the regression model is significant.

The importance and usefulness of model utility test in regression analysis is well-known. Many studies related to regression included model utility test in their analysis. However, Woolston [12] reported that a prestigious international scientific journal has no longer accepted paper containing hypothesis significance testing procedure in their studies. The editorial of the journal emphasised that null hypothesis significance testing procedure is invalid. In hypothesis testing, p-value is a probability value that is used as a common index for strength of evidence. When the p-value is less than 0.05, the result is declared 'statistically significant'. When the p-value is less than 0.01, the result is declared 'very significant' [14]. The widely use of p-value eventually became the most powerful arbiter of credibility for many scientific researches such as testing the effect of genes on behaviour, the curative powers of new drugs or the health effects of pollutants [15]. Siegfried [15] emphasised that using a p-value from a study to make any conclusion has no logical basis. Siegfried [16] doubted if p-value can really help scientists to decide whether an experimental result is meaningful. However, some researchers commented that the decision of disapproving in the usage of p-value may have its own negative consequences and thus the proper use of inferential methods needs to be analysed [18].

In view of the importance of hypothesis testing procedure in regression analysis, the model utility test in the regression analysis is further investigated in this paper. In this paper, a new technique that involves R^2 in model utility test is developed. The development of the new technique is presented in the Methodology. In Results and Discussion, model utility test was evaluated using p-value approach as well as the new technique that involves R^2.

3

METHODOLOGY

Regression Model

In this paper, a new technique that involves R^2 in model utility test was developed. The development of the new technique involves the use of the test statistic, p-value and coefficient of determination in hypothesis testing. Hence, some concepts related to regression model, model utility test and coefficient of determination are first presented in Definition 1-5.

Definition 1. A regression model that contains a dependent variable and k independent variable is defined as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$ where y is the dependent variable, xi is the independent variable, βi is the regression coefficient, $i = 1, 2, \ldots, k$ and ε is the error variable.

Definition 2. Model utility test is a hypothesis testing procedure to test if the regression model is significantly useful for further inferences. The null hypothesis, alternative hypothesis and the condition for the null hypothesis to be rejected are given as follows.

1. Null hypothesis,
   Ho: The regression model is not significant
   Alternative hypothesis,
   H1: The regression model is significant

2. The condition for the null hypothesis to be rejected is:
p-value $< \alpha$, where α = significance level.

3. If null hypothesis is rejected, we may conclude that the regression model is significantly useful for further inferences.

Definition 3. In model utility test, p-value is defined as
$$p\text{-value} = \text{Prob}(F > F_{test\ statistic}).$$

Definition 4. $F_{test\ statistic}$ is defined as
$$F_{test\ statistic}$$

$$= \frac{SS(Reg)/k}{SS(Error)/(n-k-1)}$$

$$= \frac{SS(Reg)/k}{[SS(Total)-SS(Reg)]/(n-k-1)}$$

$$= \frac{(n-k-1)SS(Reg)}{(k)[SS(Total)-SS(Reg)]} \; ;$$

$SS(Reg) = \sum(\hat{y}_i - \overline{y})$ is defined as the sum of squares of regression where it measures the explained variation. Note that $\hat{y}_i$ is the predicted value of an observed y_i value while $\overline{y}$ is the sample mean of observed y values.

$SS(Error) = \sum(y_i - \hat{y}_i)$ is defined as the sum of squares of error where it measures the unexplained variation.

$SS(Total) = \sum(y_i - \overline{y})$ is defined as the total sum of squares where it is the total variation.

Note that, $SS(Total) = SS(Reg) + SS(Error)$.

$F_{test\ statistic}$ has a F-distribution with k degrees of freedom in the numerator and $n - k - 1$ degrees of freedom in the denominator. Note that k is the number of independent variables in the regression model, n is the sample size representing number of pairs of observed values associated with the independent variable and dependent variable.

Definition 5. Coefficient of determination of the regression model is defined as

$$R^2 = \frac{SS(Reg)}{SS(Total)}$$

where R^2 measures the variation of the dependent variable that is explained by the regression model and independent variables.

The Development of the New Technique Involves R^2

In model utility test for a regression model, p-value is commonly used in the hypothesis testing procedure as described in Definition 2. In this paper, a new technique that involves R^2 in model utility test was developed and it is presented in Theorem 1. The proof of Theorem 1 is shown as follow.

Theorem 1. In model utility test, the condition for the null hypothesis to be rejected,

'p-value $< \alpha$' is equivalent to '$R^2 > \dfrac{((k)F((\alpha,k,n\text{-}k\text{-}1)))}{((n\text{-}k\text{-}1)+(k)F((\alpha,k,n\text{-}k\text{-}1))\,)}$'.

Proof.
The condition for the null hypothesis to be rejected,
p-value $< \alpha$ (1)

can be expressed as
Prob($F>F$(test statistic)) $< \alpha$.   (2)

The condition above (2) is equivalent to
F(test statistic)$>F$(critical value) ,   (3)

where F(critical value) $=F((\alpha,k,n\text{-}k\text{-}1))$.

Using Definition 4 and Definition 5, the followings are obtained.

$$F_{test\ statistic} > F_{critical\ value}$$

$$\frac{(n-k-1)SS(Reg)}{(k)[SS(Total)-SS(Reg)]} > F_{(\alpha,k,n-k-1)}$$

$$\frac{(n-k-1)\left[\frac{SS(Reg)}{SS(Total)}\right]}{(k)\left[1-\frac{SS(Reg)}{SS(Total)}\right]} > F_{(\alpha,k,n-k-1)}$$

$$\frac{(n-k-1)R^2}{(k)(1-R^2)} > F_{(\alpha,k,n-k-1)}$$

$$(n-k-1)R^2 > (k)(1-R^2)F_{(\alpha,k,n-k-1)}$$

$$R^2 > \frac{(k)F_{(\alpha,k,n-k-1)}}{(n-k-1)+(k)F_{(\alpha,k,n-k-1)}}.$$

RESULTS AND DISCUSSION

An Alternative Way to Conduct Model Utility Test

From Theorem 1 as shown in previous section, it was found that there is a relationship between p-value and R^2. Apart from the usual condition 'if p-value $< \alpha$, we reject null hypothesis', an equivalent condition involves R^2 is developed in Theorem 1, that is 'if $R^2 > \dfrac{(k)F((\alpha,k,n\text{-}k\text{-}1)))}{(n\text{-}k\text{-}1)+(k)F((\alpha,k,n\text{-}k\text{-}1))}$),

we reject null hypothesis'. With the establishment of Theorem 1, researchers are provided with an alternative way to conduct model utility test for the regression model.

Many researchers applied regression analysis as one of the statistical techniques in their data analysis. The regression model obtained from the regression analysis can be used to make prediction of the dependent variable from the independent variables. Before the regression model can be used for forecast purpose, a test of significance is needed to validate the significance of the model. As described in Definition 2, if null hypothesis is rejected, it is concluded that the regression model is significant. In order for the null hypothesis to be rejected, researchers often use the condition 'if p-value $< \alpha$, we reject null hypothesis'.

In regression analysis, coefficient of determination, usually denoted by R^2, is also a useful statistic obtained from the regression model. The value of coefficient of determination ranges from 0 to 1. A high value of R^2 that is close to 1 shows that the independent variables included in the regression model explained well on the variation of dependent variable. Hence, a regression model with high value of R^2 is always desired by the researchers.

The Effectiveness of the Model Utility Test with *p*-Value Approach

It is known that model utility test is used to evaluate the significance of the regression model. Apart from the usual p-value approach in model utility test, an alternative method by using R^2 in the hypothesis testing

procedure is developed in the previous section, i.e. Theorem 1. In this section, the effectiveness of the usual p-value approach in model utility test was evaluated. To perform the evaluation, Theorem 1 was applied in the case of simple linear regression model.

When a regression model contains one dependent variable and one independent variable, the model is called simple linear regression model. A simple linear regression model can be written as $y = \beta_0 + \beta_1 x_1 + \varepsilon$. In the case where only one independent variable involves, k is assigned to the value of 1 in Definition 1. When simple linear regression model is applied in Theorem 1, Corollary 1 is obtained.

Corollary 1. In model utility test of a simple linear regression model, the condition for the null hypothesis to be rejected is as follow:

'p-value $< \alpha$' is equivalent to '$R^2 > R^2_{min}$',

where $R^2_{min} = \dfrac{F((\alpha,1,n\text{-}2))}{((n\text{-}2)+F((\alpha,1,n\text{-}2)))}$.

A researcher who is working on a simple linear regression model may evaluate the significance of the model by using the model utility test. From Corollary 1, the researcher may use the usual condition 'if p-value $< \alpha$, we reject null hypothesis' or the condition 'if $R^2 > R^2_{min}$, we reject null hypothesis'. It is noticed that the condition related to p-value involves the value of α while the condition related to R^2 involves the values of α and n in R^2_{min}.

Case Study 1 was used to illustrate Corollary 1 when α is assigned to 0.01. The values of R^2_{min} with respect to different values of n were studied. In this Case Study, $n = 20$ to $n = 50$ were studied.

Case Study 1. Let $\alpha = 0.01$ and $n = 20$ to $n = 50$ in Corollary 1, the values of R^2_{min} were evaluated.

Results of Case Study 1. The results are presented in Table 1. Table 1 shows the equivalent conditions for the null hypothesis in model utility test to be rejected as stated in Corollary 1. When $\alpha = 0.01$, it is noticed that

a regression model is concluded to be significant when the model has a *p*-value that is smaller than 0.01. As stated in Corollary 1, when the condition inVolves R^2 is applied, a regression model with R^2 that is more than R^2_{min} can be concluded to be significant.

From Table 1, $R^2_{min} = 0.32$ for *n* = 20 is obtained. It is noticed that R^2_{min} decreases from $R^2_{min} = 0.32$ to $R^2_{min} = 0.13$ when n increases from *n* = 20 to *n* = 50. The $R^2_{min} = 0.13$ corresponding to *n* = 50 implies that a regression model that consists of 50 pairs of observed values of dependent variable and independent variable can be concluded to be a significant model if its R^2 is more than 0.13.

Table 1: The Equivalent Conditions for the Null Hypothesis in Model Utility Test to be Rejected when *α* = 0.01

n	Condition involves *p*-value, 'if *p*-value < *α*, we reject null hypothesis'	Condition involves R^2, 'if $R^2 > R^2_{min}$, we reject null hypothesis'
20	*p*-value < 0.01	$R^2 > 0.32$
21	*p*-value < 0.01	$R^2 > 0.30$
22	*p*-value < 0.01	$R^2 > 0.29$
23	*p*-value < 0.01	$R^2 > 0.28$
24	*p*-value < 0.01	$R^2 > 0.27$
25	*p*-value < 0.01	$R^2 > 0.26$
26	*p*-value < 0.01	$R^2 > 0.25$
27	*p*-value < 0.01	$R^2 > 0.24$
28	*p*-value < 0.01	$R^2 > 0.23$
29	*p*-value < 0.01	$R^2 > 0.22$
30	*p*-value < 0.01	$R^2 > 0.21$
31	*p*-value < 0.01	$R^2 > 0.21$
32	*p*-value < 0.01	$R^2 > 0.20$
33	*p*-value < 0.01	$R^2 > 0.20$
34	*p*-value < 0.01	$R^2 > 0.19$
35	*p*-value < 0.01	$R^2 > 0.18$
36	*p*-value < 0.01	$R^2 > 0.18$
37	*p*-value < 0.01	$R^2 > 0.17$
38	*p*-value < 0.01	$R^2 > 0.17$

39	p-value < 0.01	$R^2 > 0.17$
40	p-value < 0.01	$R^2 > 0.16$
41	p-value < 0.01	$R^2 > 0.16$
42	p-value < 0.01	$R^2 > 0.15$
43	p-value < 0.01	$R^2 > 0.15$
44	p-value<0.01	$R^2 > 0.15$
45	p-value < 0.01	$R^2 > 0.14$
46	p-value < 0.01	$R^2 > 0.14$
47	p-value < 0.01	$R^2 > 0.14$
48	p-value < 0.01	$R^2 > 0.14$
49	p-value < 0.01	$R^2 > 0.13$
50	p-value < 0.01	$R^2 > 0.13$

Through this Case Study, it is found that, assuming $\alpha = 0.01$ and considering $n = 20$ to $n = 50$, the values of R^2_{min} are in the range of 0.13 to 0.32.

Case Study 2. Considering $n = 20$ to $n = 50$ in Corollary 1, the values of R^2_{min} versus $n = 20$ to $n = 50$ for $\alpha = 0.05$, $\alpha = 0.025$ and $\alpha = 0.01$ were evaluated.

Results of Case Study 2. Case Study 2 was used to illustrate how the values of R^2_{min} is affected by different values of α, that are $\alpha = 0.05$ and $\alpha = 0.025$. It is known that a p-value that is less than 0.05 is conventionally deemed 'statistically significant' while a p-value that is less than 0.01 is considered 'very significant' [15]. Therefore, the results for $\alpha = 0.05$, 0.025, 0.01 are summarised in Figure 1.

From Figure 1, it is noticed that R_min^2 decreases with respect to n regardless the values of α. Through this Case Study, it is found that, by considering $n = 20$ to $n = 50$ in Corollary 1, the values of R^2_{min} are in the range of 0.08 to 0.20 when $\alpha = 0.05$; the values of R^2_{min} are in the range of 0.10 to 0.25 when $\alpha = 0.025$; the values of R^2_{min} are in the range of 0.13 to 0.32 when $\alpha = 0.01$. The values of R^2_{min} are smaller for a larger α.

(a) $\alpha = 0.05$

(b) $\alpha = 0.025$
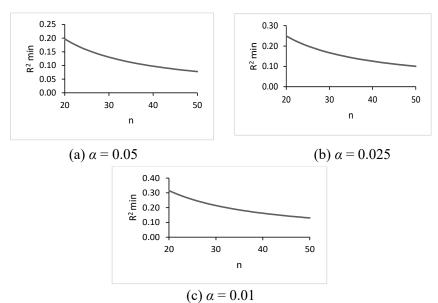

(c) $\alpha = 0.01$

Figure 1: Graph of R^2_{min} versus n for (a) $\alpha = 0.05$, (b) $\alpha = 0.025$, (c) $\alpha = 0.01$

DISCUSSION

When null hypothesis in the model utility test is rejected, it is concluded that the regression model is significant. Thus, there is a useful relationship between the dependent variable and the predictor variable. Most of the time, when the condition involves p-value is used, it may be decided to reject the null hypothesis when 'p-value $< \alpha$'. Thus, the regression model is significant.

In Theorem 1, a condition involves R^2 for the null hypothesis to be rejected was developed. The condition involves R^2 is equivalent to the condition involves p-value.

Corollary 1 was obtained when Theorem 1 was applied in the case of simple linear regression model. Corollary 1 stated that for the null hypothesis to be rejected in model utility test, the condition 'p-value $< \alpha$' is equivalent to the condition '$R^2 > R^2_{min}$'. This implies that a regression model that is considered to be significant should have R^2 that is more than R^2_{min}. Thus, R^2_{min} is the minimum value of R^2 for a significant model.

11

Through Case Study 1 and 2, the values of R^2_{min} for various sample sizes and significance levels were obtained. The values of R^2_{min} are small as the range of R^2_{min} is 0.08-0.20, 0.10-0.25, 0.13-0.32 when $\alpha = 0.05$, $\alpha = 0.025$ and $\alpha = 0.01$, respectively. The values of R^2_{min} in the range from 0.08 to 0.32 implies that a regression model with a minimum of 0.08 to 0.32 in R^2 is considered a significant model.

It is known that R^2 measures the variation of the dependent variable that is explained by the regression model and the independent variable. For instant, a regression model with a value of 0.1 in R^2 shows that 10% of the variation in the dependent variable was explained by the regression model and the independent variable. The rest of the 90% of the variation in the dependent variable is unexplained. When the regression model and the independent variable explain only a small percentage of the variation in the dependent variable, the significance of the regression model is questionable. A small value of R^2 in the regression model shows that the existing model need to be improved. For example, an addition of other independent variables in the model or a nonlinear model could be considered.

A useful and significant regression model should be able to explain a large proportion of the variation in the dependent variable. A high value of R^2 in the regression model shows that the model successfully explain a large portion of the variation of the dependent variable. Thus, a significant regression model should have a high value of R^2.

In simple linear regression model, the strength of a linear relationship between the dependent variable and the independent variable is measured by correlation coefficient, r. A reasonable rule of thumb to say a correlation is weak if the magnitude of correlation coefficient, $|r|$, is the range of 0-0.5, the correlation is strong if the magnitude of r is the range of 0.8-1, and moderate otherwise [11]. The square of the correlation coefficient gives the value of R^2 in simple linear regression model. Thus, for a given value of R^2_{min}, the corresponding $|r|$ can be obtained. Table 2 summarises the range of R^2_{min} with the corresponding range of $|r|$.

Table 2: The Range of R^2_{min} and $|r|$

α	Range of R^2_{min}	Range of $|r|$	Strength of a linear relationship
0.05	0.08-0.20	0.28-0.45	Weak
0.025	0.10-0.25	0.32-0.50	Weak
0.01	0.13-0.32	0.36-0.57	Weak to moderate

The significance of the regression model is commonly evaluated by model utility test. The hypothesis testing procedure in model utility test usually involves p-value technique. An equivalently technique involves R^2 in the hypothesis testing procedure was developed in this paper. A regression model with R^2 that is more than R^2_{min} is concluded to be a significant model in model utility test. However, it was found that when a regression model that is concluded to be a significant model from the model utility test, the linear relationship between the dependent variable and the independent variable is weak (for $\alpha = 0.05$ and $\alpha = 0.025$) or weak to moderate (for $\alpha = 0.01$) (Table 2). This shows that even if the regression model is significant at $\alpha = 0.05$, $\alpha = 0.025$ or $\alpha = 0.01$, the variables might not have a strong linear relationship.

CONCLUSION

The study in this paper shows that a regression model that is declared to be a significant model by using model utility test could not guarantee a strong linear relationship between the independent and dependent variable. This could inevitably affect the accuracy of prediction using such regression model. Therefore, it is shown that model utility test in regression is not a good technique in evaluating the significance of a regression model.

In view of the weakness in model utility test, other measures should be explored as alternatives to model utility test. With the reference to the strength of correlation in [11], an alternative measure is recommended as follow:

•If the R^2 of a regression model is in the range of 0-0.25, the regression model is not significant. The regression model should not be used for further inferences.

•If the R^2 of a regression model is in the range of 0.64-1, the regression model is strongly significant.

•If the R^2 of a regression model is in between 0.25 and 0.64, the inferences such as predictions made from the regression model should be interpreted with caution.

In this study, a measure is recommended as an alternative to the model utility test. In future, other measures could also be explored in evaluating a regression model.

REFERENCES

[1]   A. G. Bluman, 2014. *Elementary Statistics: A Step by Step Approach*, McGraw-Hill, New York.

[2]   W. Navidi, 2011. *Statistics for Engineers and Scientists*, McGraw-Hill, New York.

[3]   S. M. Trost, and G. D. Oberlender, 2003. Predicting accuracy of early cost estimates using factor analysis and multivariate regression, *Journal of Construction Engineering and Management, Vol. 129*(2), pp. 198-204. Doi: https://doi.org/10.1061/(ASCE)0733-9364(2003)129:2(198).

[4]   S. Clark, 2003. Traffic prediction using multivariate nonparametric regression, *Journal of Transportation Engineering, Vol. 129*(2), pp. 161-168. Doi: https://doi.org/10.1061/(ASCE)0733-947X(2003)129:2(161).

[5]   S. Goyal, and G. K. Goyal, 2011. Application of artificial neural engineering and regression models for forecasting shelf life of instant coffee drink, *International Journal of Computer Science, Vol. 8*(4), pp. 320-324.

[6]   G. Keller, and B. Warrack, 2003. *Statistics for Management and Economics*, Brooks Cole, USA.

[7]   J. W. Taylor, 2007. Forecasting daily supermarket sales using exponentially weighted quantile regression, *European Journal of Operational Research, Vol. 178*(1), pp. 154-167. Doi: 10.1016/j.ejor.2006.02.006

[8]   S. Nardini, 2009. Electricity consumption forecasting in Italy using linear regression models, *Energy, Vol. 34*(9), pp. 1413-1421. Doi: 10.1016/j.energy.2009.06.034.

[9]   C. J. Lu, and Y. W. Wang, 2010. Combining independent component analysis and growing hierarchical self-organizing maps with support vector regression in product demand forecasting, *International Journal of Production Economics, Vol. 128*(2), pp. 603-613. Doi: 10.1016/j.ijpe.2010.07.004.

[10] W. I. Lee, C. W. Chen, K. H. Chen, T. H. Chen, and C. C. Liu, 2012. Comparative study on the forecast of fresh food sales using logistic regression, moving average and BPNN methods, *Journal of Marine Science and Technology, Vol. 20*(2), pp. 142-152.

[11] J. L. Devore, 2012. *Probability and Statistics for Engineering and the Sciences*, Brooks Cole, USA.

[12] C. Woolston, 2015. Psychology journal bans P values, *Nature, Vol. 519*(7541), pp. 9. Doi: 10.1038/519009f.

[13] D. Trafimow, and M. Marks, 2015. Editorial, *Journal Basic and Applied Social Psychology, Vol. 37*, pp. 1-2. Doi: https://doi.org/10.1080/01973533.2015.1012991.

[14] R. Nuzzo, 2014. Scientific method: Statistical errors, *Nature, Vol. 506*(7487), pp. 150-152. Doi: 10.1038/506150a.

[15] T. Siegfried. Odds are, its wrong, Science News, March 12, 2010. https://www.sciencenews.org/article/odds-are-its-wrong.

[16] T. Siegfried. P value ban: small step for a journal, giant leap for science, Science News, March 17, 2015. https://www.sciencenews.org/blog/context/p-value-ban-small-step-journal-giant-leap-science

[17] R. L. Wasserstein, 2015. ASA comment on a journals ban on null hypothesis statistical testing, American Statistical Association.