AN EXPERIMENTAL PROBABILISTIC INTERPRETATION FOR CONFIDENCE INTERVALS IN REGRESSION

Lubomír Seif – Ondřej Vít – Miriam Helena Hudák – Lubomír Štěpánek

Abstract

This paper introduces a novel approach to hypothesis testing within the framework of classical regression analysis, aimed at enhancing the interpretability of regression coefficients. Traditionally, *t*-tests assess the significance of regression parameters against a single null hypothesis, typically $\beta = 0$. While effective, this method often provides a limited view of the underlying data structure. To address this, we propose Multi-Hypothesis tests method, which conducts a series of *t*-tests across a continuous range of potential parameter values, thereby generating a spectrum of *p*-values. These *p*-values are then plotted against the tested parameter values, offering a somewhat probabilistic interpretation akin to the Bayesian approach, but within a frequentist framework. We demonstrate this method using the `mtcars` dataset, revealing how it can uncover more nuanced insights into the behavior of regression coefficients. This approach bridges the gap between hypothesis testing and confidence intervals, potentially paving the way for more comprehensive statistical analysis. While further research is needed, our findings suggest that this method could significantly enhance the application of frequentist statistics in complex models.

Key words: *t*-tests, regression, simulation, Multi-Hypothesis

JEL Code: C12, C15, C20

Introduction

In recent years, Bayesian statistics have gained significant attention, offering powerful tools for data analysis and hypothesis testing. However, the classical frequentist approach remains foundational in statistical practice, particularly in regression analysis. This paper introduces a novel approach to hypothesis testing that aims to bridge the gap between confidence intervals and *p*-values, providing a more nuanced interpretation of parameter estimates. Specifically, we demonstrate this approach through individual *t*-tests of significance in regression analysis.

The paper begins with a review of the existing literature on significance testing in regression models, highlighting the strengths and limitations of the classical *t*-test. We then introduce our new method, which involves conducting multiple *t*-tests across a range of possible values for regression coefficients, thereby offering a probabilistic interpretation of these estimates. Finally, we apply this method to a sample dataset, illustrating its practical utility and potential to enhance the interpretability of regression results.

1 Literature Review for testing hypothesis in regression

In this section, we focus on the typical approaches for testing hypothesis in regression. We show what needs to be fulfilled to even start hypothesis testing. Furthermore, we overview methods that are used when certain assumptions of the classical regression model are violated.

1.1 The Classical *t*-Test in Regression Analysis

The *t*-test is one of the most fundamental tools in statistical analysis, particularly within the framework of regression analysis. Introduced by William Sealy Gosset under the pseudonym "Student" in 1908, the *t*-test allows researchers to determine whether a population coefficient (θ) is equal to a certain value (Student, 1908). The Gauss-Markov theorem provides the theoretical foundation for the Ordinary Least Squares (OLS) estimator, applied to the population coefficients' estimation, too, and asserting that, under certain assumptions, the OLS estimator is the Best Linear Unbiased Estimator (BLUE) (Greene, 2018).

In the classical linear regression model, the significance of an estimated coefficient is typically tested against the null hypothesis $H_0: \beta = 0$ using a *t*-test. This approach assumes that the errors are normally distributed and that the model satisfies the Gauss-Markov assumptions. If the calculated *t*-statistic exceeds a critical value from the *t*-distribution, the null hypothesis is rejected, indicating that the corresponding independent variable has a statistically significant effect on the dependent variable (Wooldridge, 2020).

1.2 Extensions and Modifications of the *t*-Test

Over time, the *t*-test has seen numerous extensions and modifications to account for various practical challenges in regression analysis. For instance, in the presence of heteroskedasticity – where the assumption of constant variance in the errors is violated – standard errors of the coefficients can be adjusted using robust estimates of standard

errors, a method introduced by White (1980). This adjustment allows the *t*-test to remain valid even when the Gauss-Markov assumptions are partially relaxed.

Moreover, the development of alternative hypothesis testing approaches, such as Bayesian methods, has offered new perspectives on evaluating regression coefficients. Unlike the frequentist *t*-test, which assesses the probability of observing data given a null hypothesis, Bayesian methods compute the probability of a hypothesis given the observed data, incorporating prior information into the analysis (Gelman et al., 2013). These approaches have provided more flexibility in hypothesis testing, particularly in complex models where traditional *t*-tests may be less effective.

Another noteworthy extension is the introduction of permutation tests and bootstrap methods, which have been used to assess the significance of regression coefficients without relying on specific distributional assumptions (Efron & Tibshirani, 1993). These non-parametric approaches allow for greater robustness in hypothesis testing, particularly in small samples or when model assumptions might be violated.

1.3 Gaps in the Current Literature

While the *t*-test remains a cornerstone of regression analysis, its application is often limited to a single hypothesis – typically $H_0: \beta = 0$. This narrow focus can potentially obscure important information about the true nature of the relationship between variables, particularly in the presence of uncertainty or model misspecification. Recent studies have suggested that more granular approaches, such as performing multiple hypothesis tests across a range of β values could provide deeper insights into the behavior of regression coefficients (Leeb & Pötscher, 2005).

Furthermore, while robust and alternative hypothesis testing methods address some of the limitations of the classical *t*-test, they do not fully exploit the potential insights available from exploring a spectrum of null hypotheses. There is a growing recognition that such multi-hypothesis approaches could offer a more nuanced understanding of the underlying data-generating process, particularly in complex or high-dimensional models where traditional tests might fail to capture subtleties in the data (Genton, 2001).

In summary, while the *t*-test has been extensively studied and modified, there is a gap in the literature regarding the systematic exploration of multiple hypotheses around a central value. This paper seeks to address this gap by proposing a method that performs a series of *t*-tests across a range of β values, thereby offering a more comprehensive picture of the significance and behavior of regression coefficients.

2 Methodology

Let $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, be a theoretical linear regression model represented in a matrix form where:

- $\boldsymbol{y}$ – a vector of values of dependent variable
- $\boldsymbol{X}$ – a model matrix with independent variables and vector of ones as its first column
- $\boldsymbol{\beta}$ – a vector of population parameters
- $\boldsymbol{\epsilon}$ – a vector of errors (nonsystematic variable)

Then let $\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e}$ denote a sample linear regression function in a matrix form with vectors and matrix being the sample counterparts to the theoretical ones. The "$\boldsymbol{e}$" denotes the residuals.

The population parameters $\boldsymbol{\beta}$ are being estimated using the ordinary least squares method (OLS) as $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$. We assume Gauss-Markov assumptions hold (linearity, strict exogeneity, no multicollinearity, spherical errors – homoskedasticity and uncorrelated error terms, normality) and that this estimate is BLUE under the Gauss-Markov theorem.

With all these stated we can discuss the significance of $\boldsymbol{\beta}$ parameters using partial *t*-tests. First, we choose a significance level (like $\alpha = 0.05$) on which we do any hypothesis testing. In general, partial *t*-tests are comprised of hypotheses and the test statistic,

$$\mathrm{H}_0: \beta_\mathrm{i} = 0; \mathrm{H}_1: \beta_\mathrm{i} \neq 0, \quad (1)$$

$$T_i = \frac{\widehat{\beta_\mathrm{i}} - \beta_\mathrm{i}}{se\left(\widehat{\beta_\mathrm{i}}\right)}, \quad (2)$$

where $se\left(\widehat{\beta_\mathrm{i}}\right)$ is the standard error of coefficient estimate $\widehat{\beta_\mathrm{i}}$, maintaining the previous notation. The test statistic T_i from formula (2) is typically written without the subtraction of β_i part because we test for the value of 0. But for the next part of this article, we keep it in the numerator. After calculating test statistic T_i, we would either find the critical region or calculate *p*-value to determine whether we reject the null hypothesis or not. For the purpose of this article, we focus on the latter, i.e., the *p*-values.

P-values for partial *t*-tests in regression can be calculated like this, $p\text{-value} = 2 * \big(1 - F(T)\big)$, where $F(T)$ is the cumulative distribution function of Student's *t*-distribution with $(n - p)$ degrees of freedom; n represents the sample size (number of rows of analyzed data), and p represents the number of estimated parameters. For a model with intercept, p is equal to the number of independent variables plus one. After calculating *p*-value, we compare it to the significance level we have chosen prior to the calculation. If and only if $p\text{-value} \leq \alpha$, we reject the null hypothesis H_0 in favor to the alternative hypothesis H_1. The *p*-value is the probability of obtaining a test result as extreme as the result actually observed, or even more extreme. This means that lower *p*-values indicate that it is unlikely that we would obtain this or a more extreme value of statistic under the assumption that the null hypothesis is valid. We use this probability and interpretation for our Multi-Hypothesis *t*-tests.

2.1 Multi-Hypothesis *t*-tests

Conducting a single *t*-test to evaluate whether the analyzed parameter is significant or not seems to be weak. This is one of the reasons to be also conducting interval estimates and building confidence intervals (CIs). Through them, we can decide not only if the parameter is significant, but we can also look for other values with one calculation. But even this approach isn't flawless. Confidence intervals, unlike credible intervals in the Bayesian approach to hypothesis testing, don't have a probabilistic interpretation. Usually, interpretations of CIs go like this "let the estimation process to be repeated over and over with random samples from the same population, then 95% of the calculated intervals would be expected to contain the true value" (Hazra, 2017). Instead of this, we would like something simple and like the Bayesian approach where we could tell: "With probability *P* the parameter is in this interval".

Something like this can be achieved using Multi-Hypothesis testing. We propose that instead of performing only one test; we can perform (in theory) multiple (thousands) tests, calculate *p*-values for each of these tests, and evaluate them against the values of tested parameter. This approach would aim to find which values are more probable to be the *true* value. In regression for partial *t*-test we can run the *t*-tests for each $\widehat{\beta_i}$ with differing null hypothesis like: $H_0: \beta_i = a_j$, where $a_j \in [-1; 1]$. In practice, we could run for each parameter 2 001 or 20 001 individual tests in total with 0.001 or 0.0001 incremental step, respectively, going through the entire sequence from -1 to 1. This would also contain the value 0, which we use for evaluation of the significance of the tested parameter. In each step, we would calculate *p*-value and save them. Afterward, we plot these *p*-values against the hypothesized values of the parameter and analyze

the behavior around the value 0. This approach aims to assess how distributed these *p*-values are around the 0 value and if they have any tendencies to show which tested values are more likely to be seen. By using *p*-values, we keep the "cumulative" interpretation meaning for each tested value and calculated test statistic; we can talk about the probability of obtaining the given test result with a specific tested value.

With this approach, we can cross the bridge between confidence intervals (CI) and hypothesis testing. In this quick thought process, we show that if the value of one of the boundaries of CI were to be the real value of the population parameter, we would obtain *p*-value equal to the significance level α.

Thought process.

- The following hypothesis testing procedure is not valid because we must test the parameter against a constant. In this case, we are "testing" parameter β_i against a random variable, which is a nonsense.
- A confidence interval for β parameters is as follows,

$$CI_i = \left(\widehat{\beta_i} - t_{1-\alpha/2}(n-p) * se\left(\widehat{\beta_i}\right); \widehat{\beta_i} + t_{1-\alpha/2}(n-p) * se\left(\widehat{\beta_i}\right)\right)$$

- Let's "test" whether the population parameter of some β_i could be equal to one of the CIs boundaries with two-sided t-test.

$$H_0: \beta_i = \widehat{\beta_i} \pm t_{1-\alpha/2}(n-p) * se\left(\widehat{\beta_i}\right)$$

$$H_1: nonH_0$$

- T statistic: $T_i = \frac{\widehat{\beta_i} - \left(\widehat{\beta_i} \pm t_{1-\alpha/2}(n-p) * se\left(\widehat{\beta_i}\right)\right)}{se\left(\widehat{\beta_i}\right)} = \pm t_{1-\alpha/2}(n-p)$
- P-value: $p\text{-value} = 2 * Pr(T_i \geq |t|) = 2 * Pr\left(T_i \geq \left|\pm t_{1-\alpha/2}(n-p)\right|\right) = 2 * \frac{\alpha}{2} = \alpha$

3 Example on data

The dataset used to showcase our new approach is `mtcars`, data about cars extracted from 1974 Motor Trend magazine (Henderson, 1981). This dataset is available in R by default (R Core Team, 2024).

The dataset contains 11 cars specific variables with 32 observations (objects).

We will model cars consumption (`mpg` – miles per US gallon) on their engine power (`hp` – horsepower) and engine displacement (`disp`). See Tab. 1 for descriptive statistics of those variables.

Tab. 1: Variables of the `mtcars` dataset used in a model

name	vars	n	mean	sd	median	min	max	range	skew	kurtosis
mpg	1	32	20.09	6.03	19.2	10.4	33.9	23.5	0.61	-0.37
hp	2	32	146.69	68.56	123.0	52.0	335.0	283.0	0.73	-0.14
disp	3	32	230.72	123.94	196.3	71.1	472.0	400.9	0.38	-1.21

Source: compiled by the authors

3.1 Regression model

Let's assume a multiple linear regression model $mpg = \beta_0 + \beta_1 hp + \beta_2 disp + \epsilon$.

We estimate its parameters using OLS. Let's assume that all Gauss-Markov assumptions[1] are valid, and the estimates are BLUE. We use a 5% significance level ($\alpha = 0.05$) for hypothesis testing and 95% confidence intervals.

Tab. 2: Summary of the estimated regression model

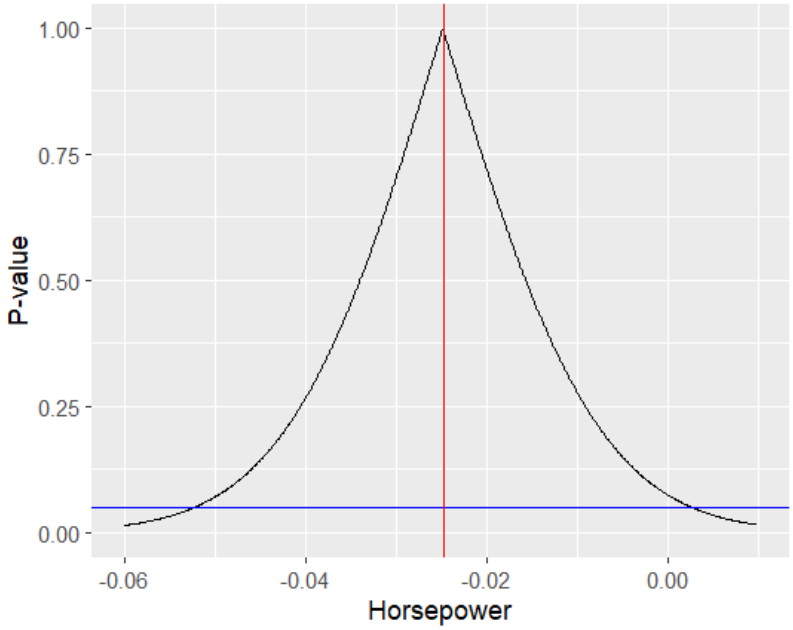
term	estimate	std.error	statistic	p.value
(Intercept)	30.736	1.332	23.083	< 0.001
hp	-0.025	0.013	-1.856	0.074
disp	-0.030	0.007	-4.098	< 0.001

Source: compiled by the authors

Tab. 2 shows OLS estimate. *P*-values are calculated for $H_0: \beta_i = 0$. From the table we can see that *p*-value for the variable `hp` is greater than our significance level, therefore we would assume that the population parameter is equal to 0. In other words, we are assuming that horsepower has no effect on consumption (`mpg`) in this specific model. In Fig. 1, we demonstrate how to use our proposed approach to identify the behavior of this estimate around zero.

[1] We are aware that these assumptions should be at least tested, but we use this model just to show on how to use Multi-Hypothesis *t*-tests. There is a slightly higher multicollinearity, and the normality assumption might be violated. This would lead to worse estimates and evaluation, but let's assume that it is not here because we want to demonstrate our Multi-Hypothesis approach.

Fig. 1: Plot of *p*-values for the `hp` estimate



Source: compiled by the authors

In Fig. 1, the red line represents our point estimate of the effect of horsepower on consumption. For this estimate, the *p*-value is logically the largest since the numerator is equal to zero and Student's *t*-distribution is symmetric around zero. Blue line represents significance level $\alpha = 0.05$. From the plot we can clearly see that we would not reject the null hypothesis $\beta_{hp} = 0$. But from the interpretation of the *p*-value for this null hypothesis, we obtain that there is a probability of 0.074 to obtain the same or more extreme test result if this given null hypothesis would be valid. Clearly from this plot, we can see that β_{hp} values lower than 0 are more probable to be seen, meaning the test result for example with a null hypothesis $H_0: \beta_{hp} = \alpha_j, j \in (-0.04, 0)$ has higher *p*-value signaling that there is higher probability for these test results to occur than a test result for a null hypothesis testing zero. In this case, we shouldn't disregard the effect of horsepower of a car on its consumption.

As we have shown in the thought process, *p*-values, for the null hypothesis that β_i is one of the boundaries of a 95% CI, are equal to 0.05 and in the plot, we can see it when the significance level line crosses the curve of *p*-values. The 95% two-sided CI is (-0.0522, 0.0025).

For the null hypothesis $H_0: \beta_{hp} = -0.0157$ we obtain *p*-value 0.5 meaning there is almost 50 % chance of obtaining same or more extreme test result. Compare this with different null hypothesis like $H_0: \beta_{hp} = -0.02$.

If point estimate and standard error remain the same, then only the numerator in test statistic changes. We obtain a smaller *p*-value indicating that the first tested value ($\beta_{hp} = -0.0157$) is more consistent with the observed data than the latter one ($\beta_{hp} = -0.02$).

Conclusion

While we have not achieved a full probabilistic interpretation of confidence intervals, this paper presents a new approach to hypothesis testing that offers valuable insights into the likely values of regression parameters. By conducting Multi-Hypothesis *t*-tests and analyzing the resulting *p*-values, researchers can gain a deeper understanding of which parameter values are more consistent with the observed data. This approach provides a richer framework for evaluating regression coefficients, complementing traditional methods.

We believe that this method has the potential for further exploration and development, potentially leading to a generalization that could enhance the frequentist approach to hypothesis testing. Although the debate between hypothesis testing and confidence intervals is ongoing within the statistical community, we hope that our work will inspire further research and foster a more integrated understanding of these two important concepts.

Acknowledgment

This research was supported by the grant no. F4/50/2023 which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

References

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. CRC press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC press.

Genton, M.G. (2001). Robustness Problems in the Analysis of Spatial Data. In: Moore, M. (eds) Spatial Statistics: Methodological Aspects and Applications. Lecture Notes in Statistics, vol 159. Springer, New York

Greene, W. H. (2018). *Econometric analysis* (8th ed.). Pearson.

Leeb, H., & Pötscher, B. M. (2005). *Model selection and inference: Facts and fiction*. Econometric Theory, 21(1), 22-59.

Student. (1908). *The probable error of a mean*. Biometrika, 6(1), 1-25.

White, H. (1980). *A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity*. Econometrica: Journal of the Econometric Society, 48(4), 817-838.

Wooldridge, J. M. (2020). *Introductory econometrics: A modern approach* (7th ed.). Cengage Learning.

Hazra A. (2017). Using the confidence interval confidently. *Journal of thoracic disease*, *9*(10), 4125–4130.

Henderson, H. V., & Velleman, P. F. (1981). Building Multiple Regression Models Interactively. *Biometrics*, *37*(2), 391–411.

R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

Contact

Bc. Lubomír Seif
Ing. Ondřej Vít
Bc. Miriam Helena Hudák
Ing. MUDr. Lubomír Štěpánek, Ph.D.

Prague University of Economics and Business
W. Churchill Sq. 1938/4, 130 67 Prague 3, Czech Republic

{seil02, vito02, hudm07, lubomir.stepanek}@vse.cz