Proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/39/20240596

Application of hypothesis testing in estimating regression models

Jingxin Song

Department of Economics and Finance, Lancaster University, Lancaster, LA1 4YW, United Kingdom

songj14@lancaster.ac.uk

Abstract. Hypothesis testing is a typical technique used in data analysis to ascertain the accuracy of an estimated regression model obtained from the data. Therefore, this method is closely related to the accuracy of the conclusions drawn. There are many studies that have used hypothesis testing to draw conclusions, however, there is still a lack of generalizations about the accuracy of the estimated model. Therefore, this paper will shed light on how the accuracy of estimated regression models can be analyzed through F-tests as well as specification tests. The F-test confirms the significance of the overall fitness of the model, yields whether there are any omitted or irrelevant variables. Finally, the specification test confirms whether each variable should be included in the model. The accuracy of the estimated regression model can be interpreted after completing the specification test for each individual variable to obtain a final model that is statistically significant using the F-test. This work highlights the importance of hypothesis testing in estimating various regression models.

Keywords: hypothesis testing, t-test, F-test, specification test

1. Introduction

Recent studies on statistics education and learning have highlighted how important it is for research methods and scientific investigations that data interpretation can be understood as a multifaceted process involving both technical and cognitive components [1]. A hypothesis must be estimated and tested in order to draw a statistical conclusion. 'Testing of Hypothesis' is a critical debate to have while performing a statistical observation, and this is known as statistical inference [2]. Hypothesis testing is widely used to answer some questions by selecting some samples. For example, exploring the personality of a target person uses hypothesis testing to hypothesize the individual personality of the target person (which is the target person extroverted or are they introverted) and then uses the target person's social interaction data to test the hypothesis to draw conclusions [3]. Again, hypothesis testing can be a tedious and redundant process, sometimes the problem may be solved by simply repeating the study without any hypothesis testing, and sometimes the null hypothesis may be silly [4].

Most of the papers that have used the hypothesis testing methodology have done so by building a specific model and using sample data to answer the question, without a thorough analysis of the accuracy of the model. This study will then focus on the accuracy of the regression model estimated in refining the hypothesis testing to solve the problem to make a theoretical summary approach that applies to the estimated multiple regression model.

DOI: 10.54254/2753-8818/39/20240596

This research will be done by using hypothesis testing and testing the overall fitness of the estimated regression model. Model specification will then be used to ensure that the estimated regression model is free from irrelevant variables as well as omitted variables.

2. Methods and theory

2.1. Uses of hypothesis testing

Testing hypotheses is a useful tool for determining sample size and overall difference produced by sampling error or essential difference caused by statistical inference method [5]. In addition to being a popular technique for evaluating hypotheses, the significance test is regarded as a basic type of statistical inference. Its fundamental idea entails speculating on the general properties of the data, which includes the null hypothesis $H_0 : \beta \leq 0$ (unexpected value) and alternative hypothesis $H_A : \beta > 0$ (expected value)). Next, one may use statistical inference sampling study to determine whether to reject or accept the hypothesis in order to draw conclusions. The two categories of errors in hypothesis testing relate to the fact that, as a result of the sample information's constraints, errors will inevitably arise. The error is no more than two cases, and in statistics they are generally referred to as Type I and Type II errors. Specifically, Type I Error stands for Prob(Rejecting H_0 | H_0 is True) can be expressed as α, while Type II Error stands for Prob(Do not reject H_0 | H_0 is False) can be expressed as β.

One can test theories concerning specific slope coefficients using the t-test. This test is suitable for use when the sample meets several requirements, including independence, equal variance, and normality [6]. The t-test is now commonly used in econometrics for hypothesis testing because these are typically the case. The t-values can be calculated for each estimated coefficient of a common equation for multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \varepsilon_i, \tag{1}$$

and t-statistic for the k^{th} coefficient is

$$t_k = \frac{(\hat{\beta}_k - \beta_{H0})}{SE(\hat{\beta}_k)}. \tag{2}$$

Here, $\hat{\beta}_k$ stands for the k^{th} variable's estimated regression coefficient, β_{H0} for the variable's k^{th} border value coefficient, and $SE(\hat{\beta}_k)$ for the variable $\hat{\beta}_k$'s estimated standard error. A null hypothesis's rejection or acceptance is determined by contrasting the estimated t-value with the critical t-value. The value that distinguishes the acceptance area from the area of rejection is known as the critical t-value. The degrees of freedom, the type I error level, and the one- or two-sidedness of the test are taken into consideration when choosing the critical t-value, or t_c, from the corresponding table. Once a critical t-value (t_c) has been selected and calculated, the t-value (t_k) can be obtained. This leads to the decision rule: if $|t_k| > t_c$ and t_k bears the sign that H_A implies, reject H_0, and otherwise, do not reject H_0.

2.2. Approaches to hypothesis testing

2.2.1. The t-test. T-tests are conducted in four phases. First, it will establish the alternative and null hypotheses. Second, it will decide on a critical t value and a significance level. Regression analysis is done in step three to determine the estimated t-value, also known as the t score. The fourth step uses the calculated t-value to decide whether to reject the H_0 or not. The chance of finding a calculated t-value higher than the critical value if H_0 were true is indicated by the degree of significance. It calculates the likelihood that a specific critical t-value indicates a Type I error. Most novice econometricians believe that importance levels should be as low as possible. Unfortunately, the likelihood of a Type II error rises sharply at low levels of significance. It is advised to use a 5-percent level of significance unless there is unique knowledge about the expenses associated with committing Type I or Type II errors [5]. If the null can be rejected at the 5% significance level, it is common to summarize and say: "The coefficient is statistically significant at the 5-percent level." For example, significance tests can be used to confirm if the variable's sign in the estimated regression model matches its theoretical sign.

DOI: 10.54254/2753-8818/39/20240596

It is found that these symbols satisfy the relation of

$$\hat{P}_t = 4.00 - 0.010PRP_t + 0.030PRB_t + 0.0035YD_t. \quad (3)$$

Here, the variables $\hat{P}_t$, PRP_t, PRB_t, and YD_t stand for the number of pounds of pork consumed per capita, the price of pork, the price of beef, and the amount of disposable income per capita over time period t, respectively. The standard deviations σ of the coefficients of PRP_t, PRB_t and YD_t in Eq. (3) are 0.004, 0.025, and 0.0005, respectively. The two hypotheses are given by $H_0: \beta_{PRP} \geq 0/\beta_{PRB} \leq 0/\beta_{YD} \leq 0$ and $H_1: \beta_{PRP} \leq 0/\beta_{PRB} \geq 0/\beta_{YD} \geq 0$, and the details are shown in Table 1.

Table 1. The coefficients are results of a selected hypothesis testing.

Coefficient	β_{PRP}	β_{PRB}	β_{YD}
Hypothesized sign	−	+	+
Calculated t-value	−2.5	+1.2	+7.0
t_c=1.708			
Result	reject H_0	do not reject H_0	reject H_0

2.2.2. The p-value. An alternative to the t-test is the marginal significance level, or p-Value. The p-value represents the study's likelihood of discovering the intended outcome if the null hypothesis is true. P-values are typically presented in quantitative studies. However, it is important to read them carefully because research consumers frequently misunderstand and interpret p-values incorrectly [7]. A t-score's p-value is the likelihood, expressed in absolute terms, of witnessing a t-value that size or higher if the H_0 were true. Software tools for standard regression automatically determine the p-values for each coefficient. Virtually every package report p-values for two-sided hypotheses. Apply the decision rule: If the p-value is less than the required level of significance and $\hat{\beta}_k$ has the sign that H_A implies, reject H_0, and otherwise, do not reject H_0.

2.2.3. The confidence intervals. A range of values including the real value of β at a specific significance level is called a confidence interval. The confidence interval formula:

$$\text{confidence interval} = \hat{\beta}_k \pm t_c * SE(\hat{\beta}_k), \quad (4)$$

where, for the selected significance level, t_c is the t-statistic's two-sided critical value. Apply the following decision rule: Reject H_0 at the X-percent level if β_{H0} is not in the confidence interval. Otherwise, do not reject H_0. For example, the range of values of a variable with a given probability can be calculated. Determine the Lancaster households' average income for the previous year: $\frac{1}{3000}\sum_{i=1}^{3000} I_i = \bar{I}$, where I represents the income of the household. The average income of every household in this county is the population parameter $E(I)$, and sample statistic $\bar{I}$ represents this parameter. Test the hypothesis that the average income of the population is £40,000 by using sample data. $\bar{I}$=40, n=3000, S=10 and for a 0.5% two-sided test with 2999, the critical t-value is 2.576. so $Prob(-2.576 \leq t \leq 2.576) = 0.99$. Substituting the data into the confidence interval formula Eq. (4) gives $40 - \frac{2.576*10}{\sqrt{3000}} \leq \mu_I \leq 40 + \frac{2.576*10}{\sqrt{3000}}$. Finally, one can simplify it to get $39.530 \leq \mu_I \leq 40.470$, which means this "random interval" (39.530, 40.470) has a 99% chance of containing the true μ_I.

3. Results and Application

3.1. F-test for overall fitness

The majority of frequently used statistical methods, such as multiple regression, multilevel modeling, t-tests for independent and dependent groups, analysis of variance and analysis of covariance, and structural equation modeling, can be described by linear regression models [8]. A linear regression model

DOI: 10.54254/2753-8818/39/20240596

is frequently used in data analysis, and hypothesis testing can be employed in linear regression models to evaluate the precision of parameter estimations [8]. The estimation of unknown values by regression analysis makes use of the correlation between known values and unknown variables. Here, the link between the independent and dependent variables is represented by a regression line, and the estimated value of the dependent variable is determined by using the independent variable's given value [9]. In order to determine how well the computed regression line matches the given data points, an F-test for overall significance is required.

The following are the primary presumptions of the regression analysis F-test of overall significance. The first is linearity, which implies that the independent and dependent variables in a linear regression must have a linear relationship. Secondly, it is the normality, every variable in a linear regression analysis needs to be a multivariate normal variable. The third is multicollinearity, which assumes that there is little or no multicollinearity in the data used in linear regression. The fourth is homoskedasticity, which indicates that if the data are homoscedastic. This means that the residuals on the regression line are equal, and a scatter plot is a useful tool for determining this. Regression mean square, or regression explained by the regression model, divided by residual mean square, or unexplained variance, is represented by the F statistic. As each predictor variable's significance was determined using the t-test only assesses the individual significance of each predictor variable. The F-test for overall significance determines if all predictor variables are jointly significant. Consequently, each predictor variable's combined significance is evaluated by the F-test. According to the F-test, every predictor factor taken is jointly significant, even though it is possible that none of the predictor variables are significant [9]. Formulas for testing the overall significance of regression models and testing subsets of coefficients

$$F = \frac{SSE \times df_{SSE}}{SSR \times df_{SSR}} = \frac{(n-k-1) \times R^2}{k(1-R^2)}. \tag{5}$$

Here, SSE denotes the regression model's explained sum of squares, and SSR stands for the residual sum of squares. The notations df_{SSE} and df_{SSR} represent the number of degrees of freedom with respect to the independent variable and the regression model, respectively. In addition,

$$F = \frac{(SSR_m - SSR) \div M}{SSR \div (n-k-1)} = \frac{(R^2 - R_M^2) \div M}{(1-R^2) \div (n-k-1)}. \tag{6}$$

Here, M denotes the number of restricted variables, $n-k-1$ is the number of degrees of freedom with respect to the unconstrained model. And SSR_m and SSR represent the residual sum of squares from the constrained and unconstrained equations, respectively.

Table 2. The coefficients, standard error, and confidence interval of the restaurant revenue model.

Y	Coeff.	Std. Err.	t	P>|t|	[95% Conf. Interval]
N	-9074.674	2052.674	-4.42	0.000	[-13272.86, -4876.485]
P	0.3546684	0.0726808	4.88	0.000	[0.2060195, 0.5033172]
I	1.287923	0.5432938	2.37	0.025	[0.1767628, 2.399084]
Cons	102192.4	12799.83	7.89	0.000	[76013.84, 128371.0]

Using hypothesis testing in a linear regression model with K independent variables, it is defined that $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ and H_1: At least one of these beta's value is not zero. A regression model's overall significance cannot be evaluated using the t-test. Specifically, it is not desirable to calculate the results of each test separately. This is because both coefficients are estimated using the same set of data. They are not self-contained entities. It is plausible that the common coefficient deviates from zero while the individual coefficients are all equal to zero. R^2 and adjusted R^2 $\left(\overline{R}^2\right)$ are not formal tests, but they

DOI: 10.54254/2753-8818/39/20240596

can quantify the equation's total deviance [5]. For example, one can test the overall significance of a restaurant revenue model

$$Y_i = \beta_0 + \beta_1 N_i + \beta_2 P_i + \beta_3 I_i + \varepsilon_i. \quad (7)$$

Here, the notations Y_i, N_i, P_i, and I_i stand for sales, competition, population, and income, respectively, and represent the gross sales volume of the i^{th} outlet, the number of direct market competitors within a three-mile radius of the i^{th} outlet, the population within a five-mile radius of the i^{th} outlet, and the average household income of the population as indicated by variable P. According to the data obtained from a sample with a sample size of 33 observations, it is found in Table 2 that $\hat{\beta}_0 = 102192, \hat{\beta}_1 = -9075, \hat{\beta}_2 = 0.355, \hat{\beta}_3 = 1.288$. They can be obtained from the data in the graph and then the estimated regression model is obtained as follows

$$\hat{Y}_i = 102192 - 9075N_i + 0.355P_i + 1.288I_i, R^2 = 0.618, n = 33. \quad (8)$$

Therefore, the hypotheses are $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ and H_1: At least one of these beta's is not zero. Thus, $F = \frac{29\times0.618}{3\times(1-0.618)} = 15.64$. For 3 and 29 degrees of freedom, F_c = 2.9340 if a significance level of 5% is chosen. Applying the decision rule: Reject H_0 if 15.64 > 2.9340. H_0 is rejected, which states that either all slope coefficients are simultaneous equal to zero or the model is overall insignificant, as 15.64 > 2.9340 in fact.

Table 3. The coefficients, standard error, and confidence interval of the restaurant revenue model.

Y	Coeff.	Std. Err.	t	P>|t|	[95% Conf. Interval]
N	-1683.542	2074.623	-0.81	0.423	[-5914.763, 2547.679]
Cons	133032.0	9923.290	13.41	0.000	[112793.3, 153270.7]

Then, suspecting that population and income may have no effect upon sales revenue, a second model can be specified as

$$Y_i = \beta_0 + \beta_1 N_i + \varepsilon_i \quad (9)$$

New data obtained from the same sample with two variables removed. It is found in Table 3 that $\hat{\beta}_0 = 133032$ and $\hat{\beta}_1 = -1683.542$. The new estimated regression model obtained with the two variables removed is then as follows

$$\hat{Y}_i = 133032 - 1683.542N_i, R_M^2 = 0.02, n = 33. \quad (10)$$

Therefore, the hypotheses are $H_0: \beta_2 = \beta_3 = 0$ and H_1: At least one of these beta's is not zero. Thus, $F_k = \frac{(0.62-0.02)\div2}{(1-0.62)\div29} = 22.89$. For 2 and 29 degrees of freedom, F_c = 3.3277 if a significance level of 5% is chosen. Applying the decision rule: Reject H_0 if 22.67 > 3.3277. Since 22.69>3.32, the null hypothesis that both β_2 and β_3 are simultaneously equal to zero is rejected and conclude that at least one of them is not zero. The research indicates that sales volumes are significantly impacted by both income and population and the Eq. (7) is more accurate than Eq. (8).

3.2. Model specification

After doing the overall test and subsets of coefficients test, it is knowing whether the estimated regression model is significant overall. It is then possible to identify some certain omitted independent variables as well as irrelevant independent variables, and specify them individually. Model specification is the process of selecting the independent variables to include or exclude from a regression equation [10]. Theoretical factors should have a greater influence on regression model specification than empirical or methodological ones. Regression analysis can be seen to involve three separate stages: model specification, parameter estimation, and parameter interpretation. The first and most important step in the process is model specification, as accurate model specification is necessary for both the estimation and interpretation of a model's parameters. As a result, specifying a model can lead to issues. There are

DOI: 10.54254/2753-8818/39/20240596

two main categories of specification errors: (1) Add a theoretically nonsensical independent variable to the regression equation, misspecification the model in the process, hence specifying the model. (2) If an independent variable that is theoretically significant is omitted from the regression equation, the model is specified [10].

For each addition or subtraction of a variable, there are four specification criteria to consider every time. People first use theory to determine if the placement of the variables in the equation is unambiguous and theoretically sound. Secondly, the significance of the variable's coefficient in the expected direction was verified using a t-test. The third use the $\overline{R}^2$, i.e., whether adding the variable results in a better overall fit of the equation. The fourthly use the bias, i.e., whether the coefficients of the other variables have changed significantly after the addition of the variable. For example, add an independent variable to the estimate regression model: Demand for Brazilian coffee is determined by the actual price of Brazilian coffee $(P_{bc}, -)$, actual price of tea $(P_t, +)$ and actual disposable income in U.S. $(Y_d, +)$. Gather data and get the estimated regression model

$$\widehat{COFFEE}_i = 9.1 + 7.8P_{bc} + 2.4P_t + 0.0035Y_d. \quad (11)$$

The standard deviations σ of the coefficients of P_{bc}, P_t and Y_d in Eq. (11) are 15.6, 1.2, and 0.0010, respectively, then the t-score can be determined by applying Eq. (2) to obtain 0.5, 2.0, and 3.5, respectively. The $\overline{R}^2$ of this estimated regression model is 0.60 and the sample size is 25. The sign of the actual price of Brazilian coffee (P_{bc}) is unexpected and insignificant. However, in theory the actual price of Brazilian coffee (P_{bc}) should have been included in the regression model, so it is possible to find that there is an omitted variable that has a positive sign and is negatively correlated with the actual price of Brazilian coffee (P_{bc}), Alternatively, that is positively correlated with the actual price of Brazilian coffee (P_{bc}) but has a negative sign. Adding the actual price of Colombian coffee (P_{cc}) to the original Eq. (9), one can get [11]

$$\widehat{COFFEE}_i = 10.0 + 8.0P_{cc} - 5.6P_{bc} + 2.6P_t + 0.0030Y_d. \quad (12)$$

The standard deviations σ of the coefficients of P_{cc}, P_{bc}, P_t and Y_d in Eq. (12) are 4.0, 2.0, 1.3 and 0.0010, respectively, then the t-score can be determined by applying Eq. (2) to obtain 2.0, -2.8, 2.0 and 3.0, respectively. The $\overline{R}^2$ of this estimated regression model is 0.65 instead and the sample size is still 25. Finally, the author can apply four specification criteria. Initially, it was always appropriate to include both prices. The second, the new variable P_{cc} has a t-score of 2.0, which is significant at most levels. The third, $\overline{R}^2$ rises when P_{cc} is introduced suggests that the variable was left out. Fourthly, the coefficient on P_{bc} changes dramatically, suggesting bias in the first result, even while two of the coefficients essentially stay the same. This should indicate the correlation between them and P_{cc} is low. This leads to the conclusion that P_{cc} is an omitted variable.

4. Conclusion

This study explains in detail how to analyze the estimated regression model's accuracy after obtaining the estimated regression model from the data. A F-test is performed to determine the overall fitness of the model, and if the results are not significant at a certain level (usually five percentage points), it is possible that there are omitted or irrelevant variables in the estimated regression model. In addition, two different regression models can be obtained from the same data (one including the variables to be specified and the other not). In cases where all variables are identical except for the variable to be specified, specification tests are performed to confirm whether a variable belongs to the calculated regression model. After individual specification tests are performed on all variables, a final estimated regression model can be obtained that is appropriate and accurate for the given data. This paper mainly uses hypothesis testing to improve the estimated regression model, which is conducive to a more accurate use of data and makes the conclusions drawn more relevant to the data, avoiding the separation of data and conclusions resulting in biased or misleading results. The methodology of this study did not

DOI: 10.54254/2753-8818/39/20240596

allow for specific omitted variables. Namely, if the data did not contain a variable that the estimated regression model ought to have contained, it was only possible to find that the estimated regression had omitted a variable. However, it was not possible to conclude what variables had been omitted or to confirm how many variables had been omitted. Future research could try to find a way to get a specific omitted variable or a possible direction to get the omitted variable.

References

[1] Queiroz, T., Monteiro, C., Carvalho, L., & François, K. (2017). Interpretation of statistical data: The importance of affective expressions. Statistics Education Research Journal, 16(1), 163-180.

[2] Dash, B., & Ali, A. (2019). Importance of Hypothesis Testing, Type I, and Type II Errors–A Study of Statistical Power. Type I, and Type II Errors–A Study of Statistical Power (December 5, 2019).

[3] Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. Journal of personality and social psychology, 36(11), 1202.

[4] Eberhardt, L. L. (2003). What should we do about hypothesis testing?. The Journal of wildlife management, 27, 241-247.

[5] Travers, J. C., Cook, B. G., & Cook, L. (2017). Null hypothesis significance testing and p values. Learning Disabilities Research & Practice, 32(4), 208-215.

[6] Wooldridge, J. M. (2014). Introduction to econometrics: Europe, middle east and africa edition. Cengage Learning.

[7] Kim, T. K. (2015). T test as a parametric statistic. Korean journal of anesthesiology, 68(6), 540.

[8] Travers, J. C., Cook, B. G., & Cook, L. (2017). Null hypothesis significance testing and p values. Learning Disabilities Research & Practice, 32(4), 208-215.

[9] Hayes, A. F., Glynn, C. J., & Huge, M. E. (2012). Cautions regarding the interpretation of regression coefficients and hypothesis tests in linear models with interactions. Communication Methods and Measures, 6(1), 1-11.

[10] Sureiman, O., & Mangera, C. M. (2020). F-test of overall significance in regression analysis simplified. Journal of the Practice of Cardiovascular Sciences, 6(2), 116-122.

[11] Allen, M. P. (1997). Model specification in regression analysis. Understanding regression analysis, 166-170.