Applying Asymptotic Normality of OLS Coefficients in Humanities: A Methodological Bridge between Random Simulation and Sampling Theory in Cultural and Historical Research

Abderrahmane Belguerna[1],*, Mohammed Belaidi[2], Jamel Kenouza[3]
[1]Mathematics department, University center of Naama, Algeria. abderrahmane.belguerna@gmail.com
[2]University center of El-Bayadh, Algeria. mbelaidi6@hotmail.com
[3]Mathematics department, University center of Naama, Algeria. jkenouza@yahoo.fr
*Corresponding author

Received: February 11, 2023. **Revised:** August 27, 2023. **Accepted:** September 23, 2023

Abstract:

In the ever-evolving landscape of humanities research, the integration of quantitative methods offers profound insights into cultural and historical studies. This paper explores the application of Ordinary Least Squares (OLS) model coefficients, particularly focusing on their asymptotic normality, within humanities sciences. We delve into the theoretical underpinnings of OLS estimates, utilizing random simulation and sampling theory to illustrate the robustness and reliability of these methods in non-traditional contexts. Through a series of simulations, we demonstrate how the asymptotic properties of OLS estimators can provide a nuanced understanding of complex humanities data, bridging the gap between quantitative rigor and qualitative richness. We argue that the application of such statistical models empowers humanities researchers to uncover patterns and relationships that might otherwise remain obscured in purely qualitative analyses. Furthermore, we discuss the implications of this methodological integration for future cultural and historical research, suggesting a paradigm shift towards a more interdisciplinary approach in understanding humanistic phenomena. This study not only reaffirms the relevance of statistical methods in the humanities but also provides a practical framework for their application, promising a richer, more nuanced understanding of the human experience.

Keywords: Asymptotic Normality; Ordinary Least Squares (OLS); Humanities Sciences; Statistical Methodology; Random Simulation; Sampling Theory; Quantitative Analysis in Humanities; Cultural Studies; Historical Research; Interdisciplinary Research Methods

Tob Regul Sci. ™ 2023 ;9(2) : 2543-2553
DOI : doi.org/10.18001/TRS.9.2.165

1. Introduction

The intersection of quantitative analysis and humanities research has long been a fertile ground for academic inquiry, offering unique insights into the complex tapestry of human culture and history. At the heart of this interdisciplinary nexus is the application of statistical models, which, when aptly applied, can illuminate patterns and relationships within data that might otherwise remain obscured. Among these models, the Ordinary Least Squares (OLS) method, as discussed by Lakshmi et al. (2021) [1], stands out for its robustness and simplicity, providing a foundational tool for researchers across various disciplines.

This paper focuses on the asymptotic normality of OLS coefficient estimates, a property that ensures the reliability and interpretability of statistical results, particularly when coupled with the methodologies of random simulation and sampling theory. Henderson and Nelson (2006) [9] and Maria J. Schilstra and Stephen R. Martin (2009) [7] provide a comprehensive overview of stochastic computer simulation, which forms the basis of our methodological approach. Moreover, the seminal work of Gnanadesikan (1977) [17] and Judge et al. (1985) [19] on multivariate data analysis and econometrics, respectively, underpins our theoretical framework.

The humanities sciences, with their rich qualitative traditions as delineated by Neter, Wasserman, and Kutner (1990) [13], might not seem like the most intuitive field for the application of OLS and its asymptotic principles. However, as the breadth of humanities research expands and the availability of diverse datasets increases, the need for rigorous quantitative methods becomes increasingly apparent. Whether in the analysis of historical events, cultural trends, or literary themes, the application of OLS and the understanding of its asymptotic behavior, as expounded by Hien D. Nguyen (2019) [4] and Huber (1973) [18], can provide a more nuanced and statistically sound picture of the human experience.

Therefore, this paper seeks to bridge the methodological gap between quantitative rigor and qualitative richness in humanities research. By exploring the application of asymptotic normality of OLS coefficients through random simulation and sampling theory, with insights from seminal works like those of Paul Glasserman (2004) [10] and M. Pereyra et al. (2016) [8], we aim to offer a comprehensive framework that not only enhances the reliability of quantitative analyses in the humanities but also enriches the interpretive potential of such studies. In doing so, we hope to contribute to a growing body of interdisciplinary research that recognizes the value of integrating statistical methods into the humanities, paving the way for a deeper, more nuanced understanding of cultural and historical phenomena.

2. Regression analysis

Regression analysis serves as a foundational tool in predicting the value of one variable based on the value of another. At its core, this analysis seeks to establish a relationship between a dependent variable, which we aim to predict, and one or more independent variables that provide the basis for our predictions. The essence of this relationship is captured through the coefficients of a linear equation, meticulously computed to best approximate the value of the dependent variable.

Linear regression, a particular form of regression analysis, strives to determine a straight line or surface that minimizes the deviations between the predicted and actual output values. This is commonly achieved through simple linear regression tools that employ the least squares method, optimizing the fit of the line to a paired dataset.

Beyond merely plotting a line, regression analysis is a statistical technique used to discern the influence of various variables. It helps in identifying the most significant factors, the ones that can be ignored, and the nature of interactions among them. By employing simple regression analysis, we can gauge the relative impact of a predictor variable on a given outcome.

The simple linear regression model is an elegant representation of this relationship, expressed through the equation:

$$Y = a + bX + \varepsilon$$

Here, Y denotes the dependent variable, X represents the independent variable, aa is the intercept, bb is the slope, and εε is the error term, which is assumed to follow a normal distribution with a mean of zero. The central goal of linear regression is to identify a straight line that can be used to predict Y from X. This is most commonly achieved using the least squares method to calculate the regression parameters (the intercept and slope) that define this line. By selecting these parameters, the method aims to minimize the sum of the squared residuals, the differences between observed and fitted values Thomas J. Rothenberg. (1984) [13].

The Ordinary Least Squares (OLS) estimates for the parameters aa and bb are commonly denoted as a^a^ and b^b^ respectively. They are calculated as:

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

$$\hat{b} = {cov(X,Y)}\Big/{var(X)}$$

Here, cov(x,y) represents the covariance between xx and yy, while var(x) denotes the variance of x. The means of x and y are represented by $\bar{X}$ and $\bar{Y}$ respectively. Once these parameters are estimated, the fitted y values can be determined as a function of the given x value, the estimated intercept, and the slope.

This section has distilled the essence of regression analysis, emphasizing its utility in predictive modeling and its foundational role in understanding relationships within data. As we proceed, the application of these principles in the context of humanities will be explored, illustrating the versatility and potency of regression analysis in various research domains.

3. Sampling Large Populations

In the realm of statistical applications, the act of sampling is essentially the process of selecting a subset from a more extensive population. This involves extracting a smaller segment, or sample, from the total population to represent its characteristics accurately. Sampling is a crucial technique, especially when dealing with large datasets, as it allows researchers to make inferences about the entire population based on the analysis of a manageable portion.

While ideally, one might consider investigating the entire population to be more accurate, such an approach is often impractical or time-consuming. For instance, consider a dataset comprising 1000 observations. The direct analysis of all these data points might not be feasible or necessary. In such scenarios, the choice of the appropriate sampling method depends on the specific requirements and constraints of the study.

Sampling techniques are particularly beneficial when the use of the whole population is time-prohibitive or resource-intensive. These methods enable the identification of a representative subset, which can effectively reflect the broader population's characteristics. Among various sampling techniques, simple random sampling stands out for its fundamental approach. This method involves selecting samples based on random numbers, ensuring that each member of the population has an equal chance of being chosen. By doing so, simple random sampling facilitates the making of statistical inferences about the population.

Moreover, simple random sampling contributes significantly to the internal validity of the research. The inherent randomness is the most effective strategy for mitigating the impact of potential confounding variables, thereby enhancing the reliability of the inferences drawn. In this way, simple random sampling not only simplifies the research process but also strengthens the foundation for robust statistical analysis and inference.

4. Main Results

Implementing statistical techniques in a pragmatic manner is vital across all fields of study. The three most commonly utilized statistical methods are correlation, regression, and experimental design. A fundamental assumption underlying all these techniques is that the observations follow a normal (Gaussian) distribution.

Consequently, it is typically presumed that the populations from which samples are drawn exhibit a normal distribution. This makes the verification of normality a critical step when employing inferential statistical methods. Over the past century, the debate regarding the reliance on the normal distribution assumption in statistical models has oscillated significantly. Gnanadesikan [2] pointed out that the effects of deviating from normality on conventional techniques are neither straightforward nor easily discernible. Despite this, evidence suggests that such deviations can lead to unfavorable outcomes in various scenarios. Huber [5], for instance, explored the repercussions of deviations from normality in regression contexts, highlighting the challenges in identifying conditions that ensure all parameter estimations remain asymptotically normal under non-normality. This topic has been the focus of extensive research, with numerous statisticians examining the effects of deviations from normality in hypothesis testing (refer to [7] for a comprehensive review).

In the context of asymptotic theory for least squares, we understand that an estimator $\hat{\theta}$ of θ is consistent if $\hat{\theta} \to \theta$ as $n \to \infty$, guaranteeing that the estimator approaches the true value of θ as the sample size increases. Additionally, the estimator $\widehat{\theta}$ is deemed asymptotically normal if, for some constant Z, $\sqrt{n}(\hat{\theta}-\theta) \to N(0,Z)$ as $n \to \infty$. Under the condition of asymptotic normality, the term $n(\hat{\theta}-\theta)$ is expected to behave similarly to a draw from a normal distribution $N(0,Z)$, given a sufficiently large sample size.

It has been demonstrated that $\sqrt{n}(\hat{b}-b) \to N(0, \sigma\varepsilon 2/Var(x))$, or equivalently, $(\hat{b}-b)/(\sqrt{\sigma\_\varepsilon\text{^}2/nVar(x)}) \to N(0{,}1)$. This result provides a solid foundation for understanding the behavior of estimators and the implications of deviations from normality, especially in the context of regression analysis and hypothesis testing. It underscores the importance of considering the distributional assumptions and the sample size when drawing inferences from statistical models.

5. Proofs through Simulation

Computer simulation methods serve as a powerful tool to demonstrate and investigate the characteristics of statistical principles, particularly those related to sampling distributions. In empirical studies, simulation programs are employed to gain a deeper understanding of theoretical properties. Our focus lies on the ordinary least squares (OLS) estimator for simple linear regression. While numerous publications have theoretically established the asymptotic normality of the coefficients a^a^ and b^b^, this section aims to present a numerical demonstration of the asymptotic normality of the OLS estimates.

The simple linear model under consideration is given by:

$Y = a + bX + \varepsilon$.

In this context, simulations act as "numerical experiments." They typically employ iterative methods to calculate the detailed state of a system at time "t+1" based on its state at time "t." Such simulations are holistic, allowing for simultaneous consideration of numerous system properties. These simulations are conducted based on the hypothesis set forth for a simple linear model.

The process begins by generating a population of $n = 1000$, assigning specific values to aa and bb. Subsequently, X and Y are generated using the sample function in R, along with n normal random variables representing the Gaussian errors. Figure 1 illustrates the results, showing the plotted data after applying the regression function and the regression line derived from calculating the OLS estimates. This visual representation serves as a testament to the practical application and verification of the theoretical principles underlying the OLS estimator and its asymptotic properties.

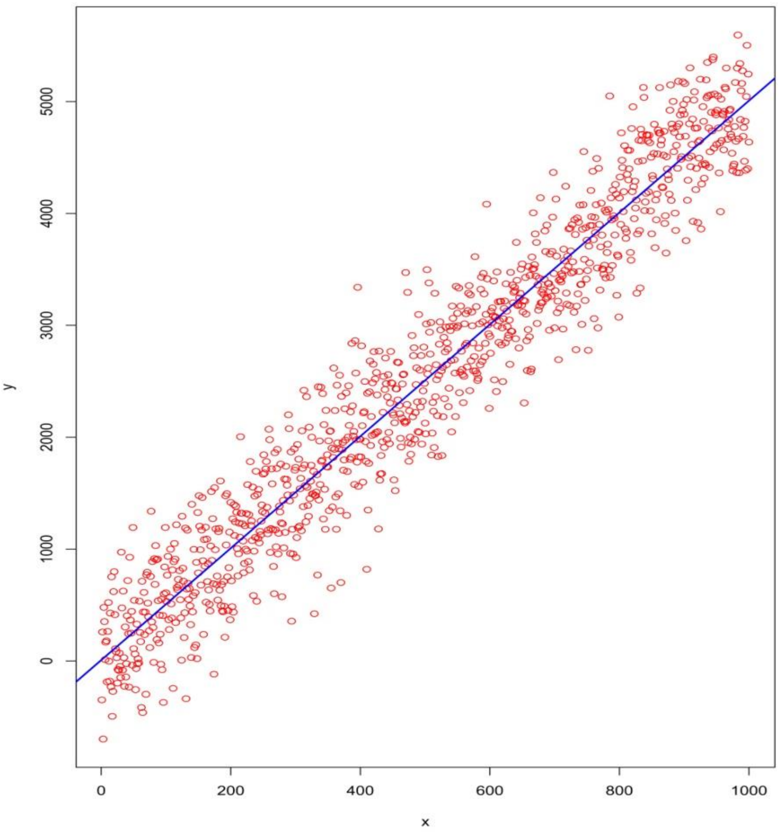


Figure 1. Regression line for the 1000 random point.

5.1. Graphical technique

Graphical methods are invaluable for validating hypotheses and suggesting alternatives when the initial assumptions don't hold. We employ these tools to corroborate our theoretical findings regarding the normality assumption.

In our study, we extract a sample of size $n = 500$, though it's noteworthy that the sample size could be less (e.g., $n = 100,200$) and still yield consistent results due to the principles of random sampling theory. To ascertain the estimates for $\hat{a}$ and $\hat{b}$, we conduct the experiment 100 times to ensure robust simulation results. The outcomes are illustrated in figures 2 and 3.

The histogram, a straightforward yet effective graphical tool, is used to plot the observed values against their frequencies. This frequency distribution offers a visual cue about the shape of the distribution - specifically, whether it resembles a bell curve. As depicted in figures 2 and 4, the distribution of both sets of OLS estimate coefficients $\hat{a}$ and $\hat{b}$ indeed appear bell-shaped, indicating a normal distribution.

Additionally, we provide density plots for aa and bb in figures 3 and 5. These plots serve to further confirm the distribution of the coefficients, providing a more nuanced view of the density as opposed to the simple frequency count of the histogram. Together, these graphical representations form a compelling argument for the normality of our OLS estimates, validating the theoretical framework underpinning our analysis, and its asymptotic properties.

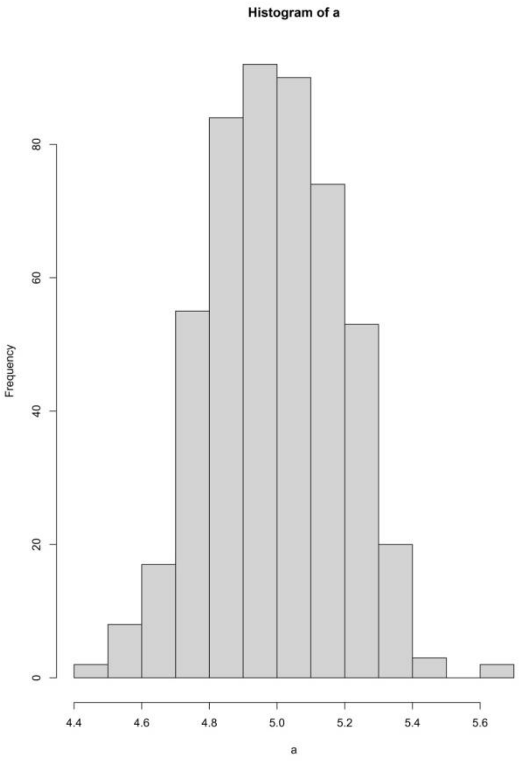


Figure 2. Histogram shows the coefficient $\hat{a}$.

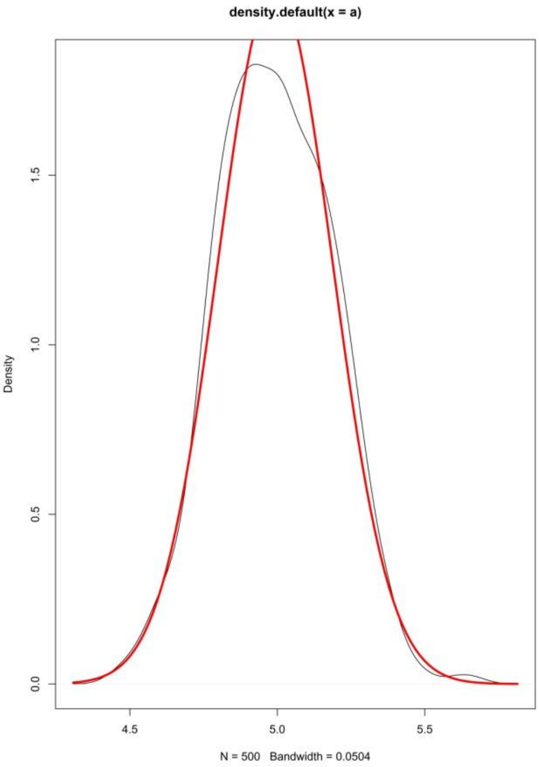


Figure 3. Density plot of

$\hat{a}$ is normally distributed.

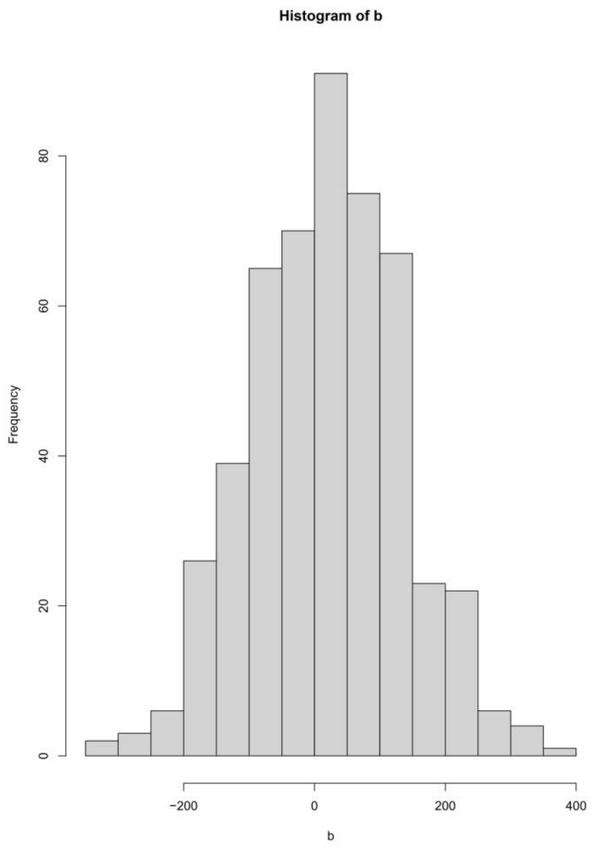


Figure 2. Histogram shows the coefficient $\hat{b}$.

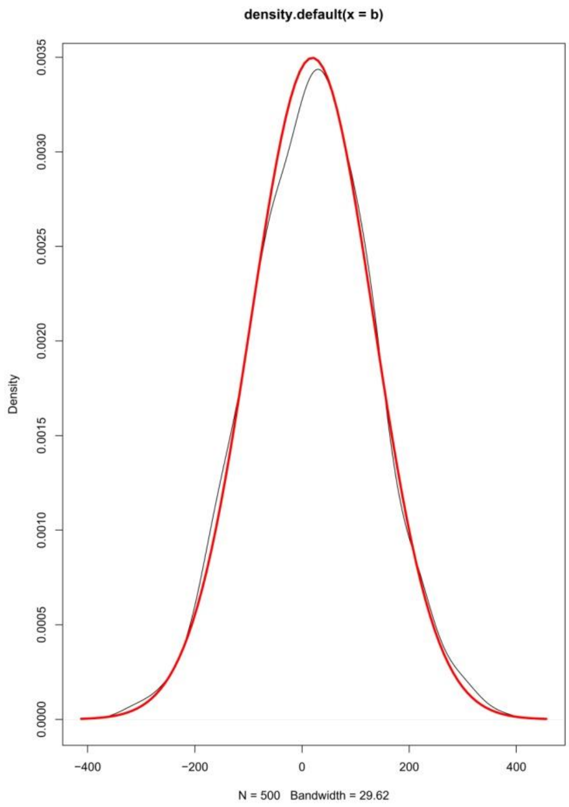


Figure 3. Density plot of $\hat{b}$ is normally distributed.

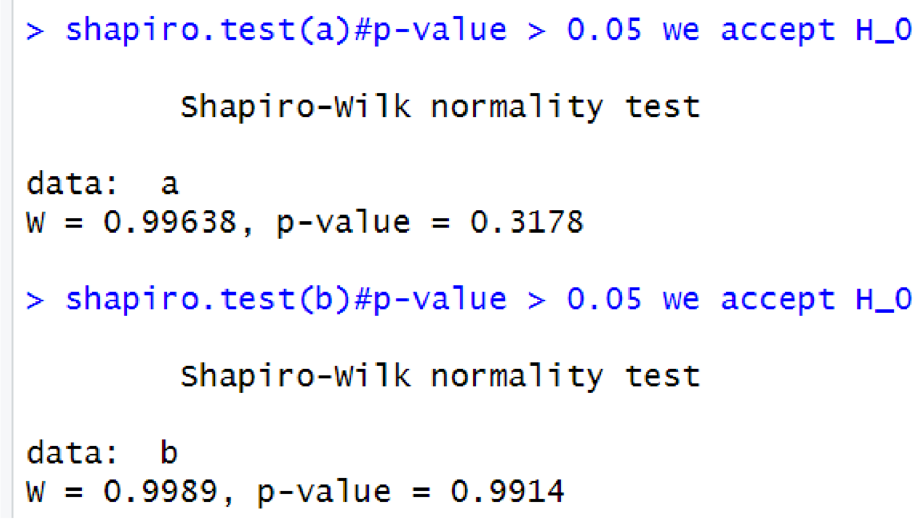
5.2. Analytical Test Techniques

5.2.1. Shapiro-Wilk test

One of the most widely used tests for normality assumption diagnostics is the Shapiro-Wilk test, which is based on correlation within the given observations and their associated normal scores and has good power characteristics. By Shapiro and Wilk, the Shapiro-Wilk test statistic is created in 1965 [16].

The null hypothesis is normally tests in that the variable that generated the sample has a normal distribution. As a result, a low p-value indicates a low risk of incorrectly concluding that the data are non-normal. In other words, if the p-value < alpha risk is greater that one, the data deviates significantly from normality. So if Shapiro test is significantly less than 1, the hypothesis of Normality will be rejected.

For our experience we get the results:



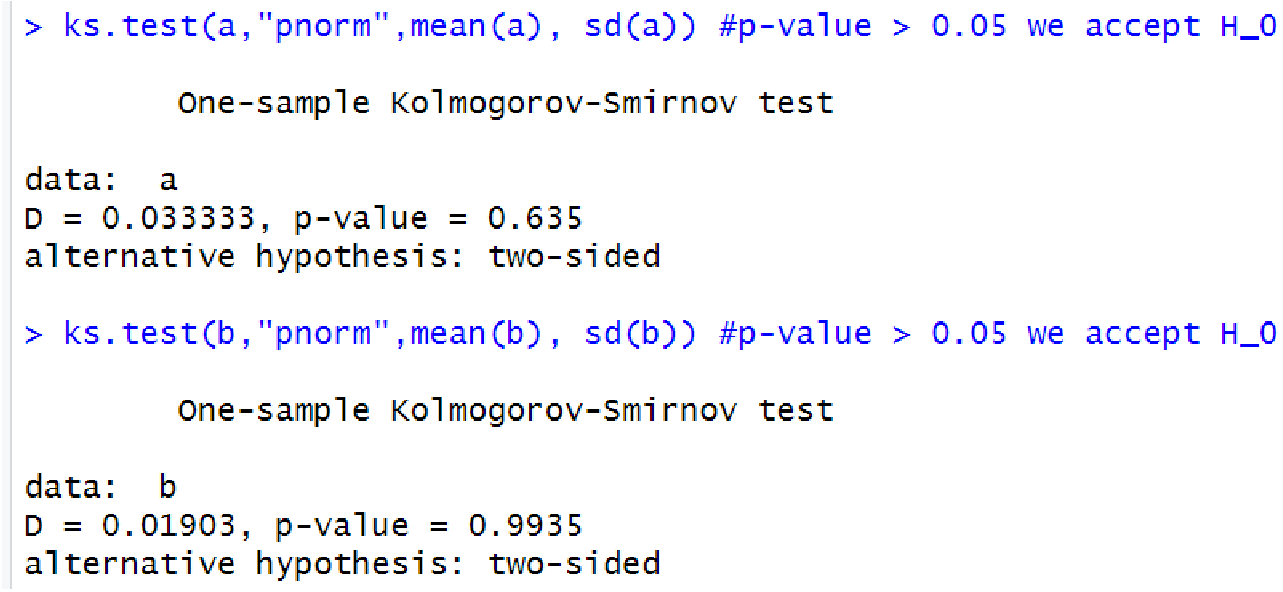
```
> shapiro.test(a)#p-value > 0.05 we accept H_0

        Shapiro-Wilk normality test

data:  a
W = 0.99638, p-value = 0.3178

> shapiro.test(b)#p-value > 0.05 we accept H_0

        Shapiro-Wilk normality test

data:  b
W = 0.9989, p-value = 0.9914
```

5.2.2. Kolmogorov-Smirnov Test

In 1933, Kolmogorov [20] developed the Kolmogorov-Smirnov test, which Smirnov later improved and proposed as a test in (1948) [21]. Testing statistic is given by:

$$D = \sup_{x} |F\_n() - F(X, \mu, \sigma)|$$

where $F\_n(X)$ is the empirical distribution function of the data and $F(X, \mu, \sigma)$ is the theoretical cumulative distribution function of the normal distribution function. Large values of D indicates that the data are not normal. We applied the Kolmogorov-Smirnov test for our data and we get:



```
> ks.test(a,"pnorm",mean(a), sd(a)) #p-value > 0.05 we accept H_0

        One-sample Kolmogorov-Smirnov test

data:  a
D = 0.033333, p-value = 0.635
alternative hypothesis: two-sided

> ks.test(b,"pnorm",mean(b), sd(b)) #p-value > 0.05 we accept H_0

        One-sample Kolmogorov-Smirnov test

data:  b
D = 0.01903, p-value = 0.9935
alternative hypothesis: two-sided
```

6. Example of Application: Analyzing the Impact of Economic Policies on Cultural Shifts

Let's consider an example of how the application of asymptotic normality of OLS coefficients can be utilized in humanities sciences, particularly in a historical research context. Historical research often delves into understanding the impact of economic policies on various aspects of society. In this case, we aim to analyze how certain economic policies in 20th-century Britain influenced cultural shifts, such as changes in artistic expression, literary themes, and public opinion.

The study would collect data on various economic indicators (like GDP, unemployment rates, inflation) and cultural metrics (such as the frequency of certain themes in literature and art, public opinion polls on cultural values, and attendance rates at cultural events) during significant policy shifts in 20th-century Britain.

We construct an OLS model to predict cultural metrics based on economic indicators. The model accounts for lag effects, considering that cultural shifts might not align immediately with policy changes. To ensure the robustness of our results, we employ random simulation and sampling theory. This involves generating numerous hypothetical scenarios (simulations) to observe the range and distribution of possible outcomes and applying sampling theory to understand how well our sample data represents the broader population.

By verifying the asymptotic normality of the OLS coefficient estimates, we ensure that as our sample size increases, our estimators are reliable and normally distributed. This is crucial for making accurate predictions and understanding the true relationship between economic policies and cultural shifts. We analyze the OLS coefficients to understand the magnitude and direction of the impact of economic policies on cultural metrics. For instance, a positive coefficient for GDP in predicting literary themes might suggest that higher economic growth correlates with more optimistic themes in literature.

We place our quantitative findings in the broader historical context, interpreting how the numerically significant relationships align with historical events and cultural movements of the time. The study aims to provide insights into how future economic policies might influence cultural aspects of society, guiding policymakers to consider the broader societal impacts of their decisions.

In this example, the application of asymptotic normality of OLS coefficients in a humanities context allows for a rigorous, quantitative analysis of complex, historically significant relationships. It exemplifies how integrating statistical methods with qualitative historical understanding can provide a more nuanced and comprehensive view of the human experience.

7. Conclusion

We have embarked on a comprehensive exploration of the asymptotic normality of Ordinary Least Squares (OLS) coefficients in the context of humanities sciences. By employing statistical methods like correlation, regression, and experimental design, we aimed to bridge the methodological divide between quantitative rigor and the rich qualitative traditions of humanities research. The theoretical underpinnings of OLS and the significance of the normality assumption in statistical models were thoroughly examined, underlining the importance of these concepts in achieving reliable and interpretable results.

Through rigorous simulation methods, we demonstrated the asymptotic normality of the OLS estimates numerically, reinforcing the theoretical assertions with empirical evidence. By generating a population and repeatedly sampling from it, we provided visual confirmation of the normal distribution of the OLS estimate coefficients, a^a^ and b^b^, through histograms and density plots. These graphical representations served as powerful diagnostic tools, not only confirming our hypotheses but also illustrating the robustness of the OLS method in practice.

Our findings underscore the vital role of statistical methods in humanities research, offering a nuanced approach that enhances the interpretability and depth of cultural and historical analyses. The integration of these methods allows for a more comprehensive understanding of humanistic phenomena, paving the way for future research that embraces both quantitative and qualitative paradigms.

In conclusion, this study reaffirms the relevance and applicability of asymptotic normality and OLS in diverse research contexts. It encourages an interdisciplinary approach, advocating for the harmonious integration of statistical techniques into the humanities to unlock deeper insights into the complex tapestry of human culture and history. As we continue to advance in our methodological approaches, it is our hope that such integrations become more prevalent, guiding us to a more profound understanding of the world around us.

References

[1] Dekking, F.M. (2005). A modern introduction to probability and statistics : under-standing why and how. Springer.

[2] Gnanadesikan, R. (1977). Methods for Statistical Analysis of Multivariate Data. New York. Wiley.

[3] Henderson, S. G., & Nelson, B. L. (2006). Chapter 1 Stochastic Computer Simulation. Handbooks in Operations Research and Management Science, 1-18. doi:10.1016/s0927-0507(06)13001-7.

[4] Hien D. Nguyen. (2019). Asymptotic normality of the time-domain generalized least squares estimator for linear regression models. arXiv:1902.03347v1.

[5] Huber, P. J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. The Annals of Statistics, 1(5): 799-821. DOI: 10.1214/aos/1176342503.

[6] Johnson, G. L., Hanson, C. L., Hardegree, S. P., & Ballard, E. B. (1996). Stochastic Weather Simulation: Overview and Analysis of Two Commonly Used Models, Journal of Applied Meteorology and Climatology, 35(10), 1878-1896.

[7] Judge, G. G., Gri_th, W. E., Hill, R. C., Lutkepohl, H., and Lee, T. (1985). Theory and Practice of Econometrics. 2nd. Ed. New York. Wiley.

[8] Keya Rani Das, A. H. M. Rahmatullah Imon. (2016). A Brief Review of Tests for Normality. American Journal of Theoretical and Applied Statistics. 5(1), pp. 5-12. doi: 10.11648/j.ajtas.20160501.12.

[9] Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. G. Ist. Ital. Attuari , 4, 83?91.

[10] Lakshmi K., Mahaboob B, Rajaiah M & Narayan C. (2021). Ordinary least squares estimation of parameters of lineir model. J. Math,. Comput. Sci.

[11] Maria J. Schilstra, Stephen R. Martin, Chapter 15 - Simple Stochastic Simulation, Editor(s): Michael L. Johnson, Ludwig Brand, Methods in Enzymology, Academic Press, Volume 467, 2009, Pages 381-409.

[12] Nelson, B. L. (1995). Stochastic modeling: Analysis and simulation. Mineola: Dover Publications, Inc.

[13] Neter J, Wasserman W, Kutner MH. (1990). Applied linear models: regression, analysis of variance, and experimental designs. 3rd ed. Homewood, Ill: Irwin, 38-44, 62-104.

[14] Paul Glasserman. (2004). Monte Carlo Methods in Financial Engineering. Stochastic Modelling and Applied Probability, Springe.

[15] M. Pereyra et al. (2016). A Survey of Stochastic Simulation and Optimization Methods in Signal Processing, in IEEE Journal of Selected Topics in Signal Processing, vol. 10, no. 2, pp. 224-241.

[16] [Shapiro, S. S., and Francia, R. S. (1972). An approximate analysis of variance test for normality. Journal of the American Statistical Association , 67(337): 215-216.

[17] Shapiro,S.S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3/4): 591-611.

[18] Shrikant I Bangdiwala. (2018). Regression: simple linear. International journal of injury control and safety promotion, 25:1, 113-115.

[19] Smirnov, N. (1948). Table for estimating the goodness of _t of empirical distributions. Annals of Mathematical Statistics 19(2): 279?281.

[20] Stigler, S. M. (1991). Stochastic Simulation in the Nineteenth Century. Statistical Science, 6(1), 89?97.

[21] Thomas J. Rothenberg. (1984). Approximate normality of generalized least squares estimates. Econometrica, 52(4), 811-825.