Published in final edited form as:

Commun Stat Theory Methods. 2015 ; 44(8): 1571–1579. doi:10.1080/03610926.2013.781633.

Inversion Theorem Based Kernel Density Estimation for the Ordinary Least Squares Estimator of a Regression Coefficient

Dongliang Wang¹ and Alan D. Hutson²

¹Department of Public Health and Preventive Medicine, State University of New York Upstate Medical University, Syracuse, New York, USA

²Department of Biostatistics, University at Buffalo, Buffalo, New York, USA

Abstract

The traditional confidence interval associated with the ordinary least squares estimator of linear regression coefficient is sensitive to non-normality of the underlying distribution. In this article, we develop a novel kernel density estimator for the ordinary least squares estimator via utilizing well-defined inversion based kernel smoothing techniques in order to estimate the conditional probability density distribution of the dependent random variable. Simulation results show that given a small sample size, our method significantly increases the power as compared with Wald-type CIs. The proposed approach is illustrated via an application to a classic small data set originally from Graybill (1961).

Keywords

Kernel density estimation; Characteristic function; Inversion theorem

1. Introduction

Consider a dependent variable Y and $p-1$ independent variables $X_1, X_2, \dots, X_{p-1}$. Classic regression analysis is to model the relationship between Y and $X_1, X_2, \dots, X_{p-1}$ via fitting the linear model

$$y = X\beta + e, \quad (1.1)$$

where y is the n -element column vector of values taken by Y , X is the concatenation of the $n \times 1$ column vector of 1's with the $n \times (p-1)$ matrix of values taken by the $p-1$ independent variables $X_1, X_2, \dots, X_{p-1}$ and e is a column of random error e 's. The major goal of regression analysis is to obtain the vector of point estimators for the unknown parameter vector β and corresponding confidence intervals. Under the assumptions that (1) X is fixed and has rank p and (2) e is a random vector with $Ee = \mathbf{0}$ and $Eee' = \sigma^2 \mathbf{I}$, it is well known that the Ordinary least squares (OLS) estimator $\hat{\beta} = (X'X)^{-1} X'y$ of β is the Best Linear Unbiased Estimator (BLUE), and has a multivariate normal distribution

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}), \quad (1.2)$$

if the random errors are normally distributed. The sampling distribution of $\hat{\beta}$ is still well approximated by (1.2) even if the errors don't follow a normal distribution. The wellness of the approximation depends on the extent to which the distribution of errors deviates from normal.

In this article we propose a new kernel estimator for the sampling distribution of OLS estimator $\hat{\beta}$ by utilizing the classic kernel density estimation in conjunction with the well-known inversion theorem. The theory and applications of kernel density estimation have been paid much attention over the past two decades, e.g., see Scott (2008) for a comprehensive overview. In the i.i.d. univariate continuous case, with support over the real line $\mathcal{R}$ and observed data $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$, a well-known form of kernel density estimator for the probability density function (pdf) $f(\cdot)$ is given by

$$\hat{f}_X(x) = \frac{1}{n} \sum_{k=1}^n \frac{1}{h} \phi\left(\frac{x - x_k}{h}\right), \quad (1.3)$$

where the kernel, $\phi(\cdot)$, is defined in this case as the standard normal pdf and $h(>0)$ is the bandwidth function. For a general unknown underlying population the choice of bandwidth still is an open problem in statistics; For more details see Scott (2008). In this article, we examine the performance of our proposed approach with the bandwidth h selected according to the Asymptotic Mean Integrated Squared Error (AMISE) criterion or the plug-in methods as described in Sheather and Jones (1991).

The rest of this article is organized as follows. In Sec. 2, we illustrate our novel approach of estimating the sampling distribution of OLS estimators in context of linear regression. Simulation results are presented in Sec. 3 to show the performance of our estimator in term of coverage probability and power. We apply our approach to a small data set in Sec. 4. In Sec. 5, we provide a summary of our approach and discuss possible improvement and related future works.

2. A Kernel Density Estimator for OLS Estimators in Linear Model

The development of our kernel density estimator relies on the classic Inversion Theorem. For cumulative distribution function (cdf), F continuous everywhere, and $dF = f(x)dx$, the following relationships are well known:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx) \psi_X(t) dt, \quad (2.1)$$

$$F(x) = \frac{1}{2} + \frac{1}{2\pi} \int_0^{\infty} \frac{\exp(itx) \psi_X(-t) - \exp(-itx) \psi_X(t)}{it} dt, \quad (2.2)$$

where $\psi_X(t)$ denotes the characteristic generating function (cgf) as

$$\psi_X(t) = \int_{-\infty}^{\infty} \exp(itx) dF, \quad (2.3)$$

and $i = \sqrt{-1}$; see more details in Gil-Pelaez (1951), Shephard (1991), and Waller et al. (1995).

Letting $V = (X'X)^{-1}X'$, the OLS estimator $\hat{\beta}$ can be expressed as

$$\hat{\beta} = Vy = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & & \vdots \\ v_{p1} & v_{p2} & \cdots & v_{pn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad (2.4)$$

which allow us to write the individual parameter estimator $\hat{\beta}_i$ as a weighted average of y_i 's,

$$\hat{\beta}_i = \sum_{j=1}^n v_{ij} y_j. \quad (2.5)$$

Thus the distribution of $\hat{\beta}$ depends on the conditional density distribution of Y given $X = x$, which is equivalent to the distribution of random error e given the independence of X and e . Towards this end, we estimate the error terms by

$$\hat{e} = y - X(X'X)^{-1}X'y = HY, \quad (2.6)$$

where $H = I - X(X'X)^{-1}X'$. Theil (1965) has shown that the error estimates at (2.6) are best linear unbiased estimates in the sense of minimizing the expected sum of squares of the estimation error $(H - I)e$. Based on the n error estimates $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n$ at (2.6), the conditional distribution $f_{Y|X=x}(y)$ of $Y|X=x$ can be directly estimated through the traditional methodology of kernel density estimation at (1.3), that is

$$\hat{f}_{Y|X=x}(y) = \frac{1}{n} \sum_{i=1}^n \phi\left(\frac{\hat{e}_i - (y - x\beta)}{h}\right). \quad (2.7)$$

The characteristic function of $\hat{\beta}$ can be readily derived from (2.5) as

$$\psi_{\hat{\beta}_i}(t) = \prod_{j=1}^n \psi_{Y|X=x_j}(tv_{ij}), \quad (2.8)$$

where $\psi_{Y|X=x_j}(t)$ is the characteristic function of Y conditional on $X = x_j$. In reality, $\psi_{Y|X=x_j}$ can be estimated based on the conditional kernel density estimator at (2.7). Thus, an empirical estimator of cgf of $\hat{\beta}$ can be expressed in the form

$$\hat{\psi}_{\hat{\beta}_i}(t) = \prod_{j=1}^n \frac{\exp(itv_{ij}x_j\hat{\beta})}{n} \sum_{l=1}^n \exp(itv_{ij}e_l - \frac{1}{2}t^2v_{ij}^2h^2) \quad (2.9)$$

by noting that

$$\hat{\psi}_{Y|X=x}(t) = E e^{ity} | f_{Y|X=x}(y) = \frac{\exp(it\mathbf{x}\boldsymbol{\beta})}{n} \sum_{l=1}^n \exp\left(it e_l - \frac{1}{2}(th)^2\right) \quad (2.10)$$

and replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$.

In order to proceed we note that $\hat{\psi}_{\hat{\beta}_i}(t)$ satisfies *exactly* the necessary and sufficient conditions to be considered a characteristic function, namely:

1. $\hat{\psi}_{\hat{\beta}_i}(t)$ is continuous in t ;
2. $\hat{\psi}_{\hat{\beta}_i}(t)$ is defined in every finite t -interval;
3. $\hat{\psi}_{\hat{\beta}_i}(0) = 1$;
4. $\hat{\psi}_{\hat{\beta}_i}(t)$ and $\hat{\psi}_{\hat{\beta}_i}(-t)$ are conjugate quantities; and
5. $|\hat{\psi}_{\hat{\beta}_i}(t)| \leq 1$.

Proof

The conditions (1)–(4) above are fairly obvious. Condition (5) is satisfied since

$$|\hat{\psi}_{\hat{\beta}_i}(t)| = \left| \prod_{j=1}^n \hat{\psi}_{Y|X=x_j}(tv_{ij}) \right| \leq \prod_{j=1}^n |\hat{\psi}_{Y|X=x_j}(tv_{ij})| = 1$$

Combining (2.1) and (2.2) with (2.9), and noting that the necessary and sufficient conditions for $\hat{\psi}_{\hat{\beta}_i}(t)$ to be characteristic function given above are satisfied, we have the following.

Definition 2.1

The inversion theorem based kernel estimators for the pdf and cdf for the OLS estimator $\hat{\beta}_i$ at (2.5) given the observed data $\mathbf{y}$ and design matrix $\mathbf{X}$ are defined as

$$\hat{f}_{\hat{\beta}_i}(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itz) \prod_{j=1}^n \frac{\exp(itv_{ij}\mathbf{x}_j\boldsymbol{\beta})}{n} \sum_{l=1}^n \exp(itv_{ij}e_l - \frac{1}{2}t^2v_{ij}^2h^2) dt, \quad (2.11)$$

$$\begin{aligned} \hat{F}_{\hat{\beta}_i}(z) = & \frac{1}{2} + \frac{1}{2\pi} \int_0^{\infty} \frac{1}{it} \left(\exp(itz) \prod_{j=1}^n \frac{\exp(itv_{ij}\mathbf{x}_j\boldsymbol{\beta})}{n} \sum_{l=1}^n \exp(itv_{ij}e_l - \frac{1}{2}t^2v_{ij}^2h^2) \right. \\ & \left. - \exp(-itz) \prod_{j=1}^n \frac{\exp(itv_{ij}\mathbf{x}_j\boldsymbol{\beta})}{n} \sum_{l=1}^n \exp(itv_{ij}e_l - \frac{1}{2}t^2v_{ij}^2h^2) \right) dt. \end{aligned} \quad (2.12)$$

The expressions for the pdf and cdf estimators of $\hat{\beta}_i$ at (2.11) and (2.12) follow from substituting $\hat{\psi}_{\hat{\beta}_i}(t)$ at (2.9) into (2.1) and (2.2), respectively. Direct calculation of these expressions is quite straightforward given observed data values and can be readily programmed in commonly used symbolic software packages, such as *Mathematica*.

Percentile Intervals

The $(1 - \alpha) \times 100$ percentile-type confidence interval for the parameter β_i based on our new approach is defined as

$$\left(\hat{F}_{\hat{\beta}_i}^{-1}(\alpha/2), \hat{F}_{\hat{\beta}_i}^{-1}(1 - \alpha/2) \right), \quad (2.13)$$

where $\hat{F}_{\hat{\beta}_i}^{-1}(\cdot)$ is the inverse of $\hat{F}_{\hat{\beta}_i}(\cdot)$ defined at (2.12) above. Even though the calculation of the percentile interval requires numerical inversion of the cdf it is quite straightforward using common statistical software packages. The derivation of $(1 - \alpha) \times 100$ percentile-type confidence interval can be readily employed for hypothesis testing. Consider a set of

hypotheses $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$. We reject H_0 if $0 \notin \left(\hat{F}_{\hat{\beta}_i}^{-1}(\alpha/2), \hat{F}_{\hat{\beta}_i}^{-1}(1 - \alpha/2) \right)$. The coverage probability and power of the CIs will be examined in Sec. 3.

3. Simulation Results

We performed simulation studies to evaluate and compare the efficiency of our methods and the classic central limit theorem (CLT) based methods in term of coverage probability and power. We considered a simple linear model

$$Y_i = \beta_0 + \beta_1 X_i + e_i. \quad (3.1)$$

The value of independent variable X was either 0 or 1, divided evenly by sample size, i.e., half of the total values have a covariate level of 0 and half of the values have a covariate level of 1. The noise terms were simulated under the following situations:

- i. normal: $e_i \sim \text{i.i.d normal}(0, 1)$;
- ii. Laplace: $e_i \sim \text{i.i.d Laplace}(0, 1)$;
- iii. t_2 : $e_i \sim \text{i.i.d student } t_2$;
- iv. exp: $e_i + 4 \sim \text{i.i.d exponential}(1)$;
- v. lnorm1: $e_i + \exp(0.5) \sim \text{i.i.d log-normal}(0, 1)$;
- vi. lnorm2: $e_i + \exp(0.72) \sim \text{i.i.d log-normal}(0, 1.2)$; and
- vii. lnorm3: $e_i + \exp(0.98) \sim \text{i.i.d log-normal}(0, 1.4)$.

The shapes of the underlying distributions are shown in Fig. 1. We generated 5000 Monte Carlo simulations for sample size $n = 10, 20, 50$ and compared the standard t-interval approximate 95% confidence interval for $\beta_1 = 2$ and the conditional kernel based estimator of the 95% confidence interval given at (2.13). The calculation of pdf and cdf estimators at (2.11) and (2.12) for slope estimates β_1 was performed in *Mathematica*. The empirical coverage probabilities of the 95% CIs are given in Table 1. The kernel based CIs are referred to as KE1 if the bandwidths were selected by the AMISE criterion and as KE2 if the bandwidths were selected by the plug-in methods as described in Sheather and Jones (1991). For small sample sizes ($n = 10$ or 20), the CIs derived from our kernel density estimator for β_1 have equivalent or slightly superior coverage probabilities as compared to the asymptotic

theory based CIs in terms of the difference to the nominal coverage probability of 95%. Our method tends to yield a lower coverage probability than the asymptotic method. As the sample sizes go up to $n = 50$, the performance of the two methods is comparable. Note that the empirical probability of a 95% confidence interval under 5000 simulations ranges from 0.944 and 0.956.

The power comparison for testing the hypothesis $H_0 : \beta_1 = 0$ vs. the simple two-sided alternative $H_a : \beta_1 \neq 0$ are carried out under the simple alternative that $\beta_1 = 2$ at level $\alpha = 0.05$. The results are shown in Table 2. Note that the coverage probabilities presented in Table 1 can be regarded as empirical Type I errors after a simple location shift. At $n = 10$, the power for our method is 6–10% higher as compared with the asymptotic method. At $n = 20$, our method still statistically significantly elevates the power as confirmed by a two-proportion comparison test. What is interesting to note is that the skewness of exponential(1), log-normal(0,1), log-normal(0,1.2) and log-normal(0.1.4) is 2, 6.18, 11.16, and 22.47, respectively. Thus, our simulation results suggest the advantage of our method over the normal method may become larger as the data become more skewed. As the sample size increases to $n = 50$, the power of the two methods are comparable and no statistically significant difference is observed.

We also provides the power curves given β_1 is equal to 0–3. When the underlying distribution is normal as shown in Fig. 2a, the performance of our method is almost equivalent to the asymptotic method. The increase of power at $n = 10$ is partially caused by the inflation of Type I error. Our method is uniformly better than the asymptotic method across the whole range of β_1 values given other non-normal alternatives, particularly for t_2 , log-normal(0,1), log-normal(0,1.2), and log-normal(0,1.4). For the purpose of illustrating the pattern, the power curve given noise term from log-normal(0,1) is displayed in Fig. 2b. The elevation of power is caused by the construction of a more precise confidence interval, which results in a Type I error closer to 0.05. There is no significant difference between our kernel-based method and the asymptotic CLT based method at $n = 50$. In addition, the comparison between KE1 and KE2 suggests that the performance of our approach does not heavily depend on bandwidth selection.

4. Examples

For our example, we consider a classic small data set given in Table 3 originally from Graybill (1961), which is also consider by Adichie (1967) and Sen (1968). The ordinary least squares estimate of β_1 is 4.02 with the 95% large sample confidence interval calculated as (3.802, 4.241), while the kernel density based 95% confidence interval for β_1 is (3.843, 4.197) if the bandwidth is selected based on the normal reference rule and (3.850, 4.190) if the bandwidth is selected based on the plug-in method. The confidence intervals derived from our proposed approach are narrower and reflect the left skewness of the original data. The kernel density estimator for $\hat{\beta}_1$ is given in Fig. 3.

Sen (1968) provided a robust point estimate 4.00 for β_1 associated with a 93% confidence interval (3.71, 4.18) through utilizing Kendall's Tau. The 93% confidence interval based on the large sample theory is (3.826, 4.217). Although Sen's CI is a bit wider than the large

sample theory based CI, it illustrate the left skewness of the original data. The 93% confidence intervals from our proposed approach are (3.856, 4.185) and (3.862, 4.179). Again, both are much narrower and slightly show the left skewness.

5. Conclusions and Discussions

In this article, we developed a novel approach for estimating the kernel density distribution of an ordinary least squares estimate of linear regression coefficient via well-defined kernel smoothing techniques in conjunction with classic inversion theory. Although the empirical characteristic function estimators could be a reasonable alternative, we prefer to utilize the form of kernel density estimators in order to take advantage of the normal form and smoothness with respect to the numerical integration. Our simulation study shows that the $(1 - \alpha) \times 100$ percentile-type confidence intervals based on our proposed method significantly increase the power as compared with the CIs based on the large sample theory. We illustrated our approach through an application to a small dataset from Graybill (1961) and demonstrated the relative merits to our approach in terms of tighter confidence intervals.

It is fairly obvious that the estimation of the error terms may impact the performance of our kernel density estimator and that the resulting confidence intervals by altering the conditional density estimation of the dependent variable. Given a set of nonnormal data, one might estimate the residuals via nonparametric alternatives, such as local smoothing techniques or other robust estimator of regression coefficient. Even though we focused on developing a kernel density estimator for ordinary least squares estimators, what we have provided is a general nonparametric approach that can be utilized to develop kernel density estimator for any estimate in a form of linear combination of dependent variables in context of linear regression.

It is also worthy noting that essentially our approach is derived following a “bootstraplike” principle via plugging in the conditional probability density distribution estimator of the dependent random variable. However, it is distinct from the bootstrap methodology in that it can be executed through a direct calculation and, hence, does not suffer from additional Monte Carlo error arising from routine resampling approaches.

References

1. Adichie JN. Estimates of regression parameters based on rank tests. *Ann. Mathemat. Statist.* 1967; 38:894–904.
2. Gil-Pelaez J. Note on the inversion theorem. *Biometrika.* 1951; 38:481–482.
3. Graybill, FA. *Introduction to Linear Statistical Models*. Vol. 1. New York: McGraw-Hill Book Company; 1961.
4. Sen PK. Estimates of the regression coefficients based on Kendall’s Tau. *J. Amer. Statist. Assoc.* 1968; 63:1379–1389.
5. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B.* 1991; 53:683–690.
6. Shephard NG. From characteristic function to distribution function: a simple framework for the theory. *Econometric Theory.* 1991; 4:519–529.
7. Scott, DW. *Multivariate Density Estimation*. New York: Wiley-Interscience Publication; 2008.

8. Theil H. The analysis of disturbances in regression analysis. *J. Amer. Statist. Assoc.* 1965; 60:1067–1079.
9. Waller LA, Turnbull BW, Hardin JM. Obtaining distribution functions by numerical inversion of characteristic functions with applications. *Amer. Statistician.* 1995; 49:346–350.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.
The distributions for the error term used in simulations.

(a) $e \sim \text{Normal}(0, 1)$ (b) $e + \exp(0.5) \sim \text{log-normal}(0, 1)$ **Figure 2.**

Size/power curve given normally and non-normally distributed noise term.

Figure 3. Kernel density and distribution estimation of the ordinary least squares estimator $\hat{\beta}_1$ given a small data set from Graybill (1961).

Table 1

Z = CLT based, KE = kernel estimator based. The bandwidths are selected based on normal reference rule for KE1 and plug-in method for KE2.

	n = 10			n = 20			n = 50		
	Z	KE1	KE2	Z	KE1	KE2	Z	KE1	KE2
normal	0.951	0.926	0.924	0.957	0.959	0.959	0.945	0.958	0.956
Laplace	0.962	0.934	0.930	0.956	0.953	0.953	0.949	0.955	0.954
t_2	0.960	0.937	0.935	0.966	0.961	0.960	0.958	0.959	0.959
exp	0.956	0.926	0.923	0.951	0.942	0.940	0.954	0.959	0.954
lnorm1	0.970	0.937	0.933	0.965	0.960	0.957	0.959	0.960	0.958
lnorm2	0.973	0.940	0.936	0.972	0.963	0.961	0.964	0.963	0.960
lnorm3	0.977	0.942	0.938	0.969	0.959	0.956	0.968	0.967	0.965

Table 2

The empirical power for testing $H_0 : \beta_1 = 0$ vs. the two-sided alternative $\beta_1 \neq 0$ given $\beta_1 = 2$. $Z = \text{CLT based}$, $\text{KE} = \text{kernel estimator based}$. The bandwidths are selected based on normal reference rule for KE1 and plug-in method for KE2.

	n = 10			n = 20			n = 50		
	Z	KE1	KE2	Z	KE1	KE2	Z	KE1	KE2
normal	0.785	0.846 ^{\$\$\$}	0.844 ^{\$\$\$}	0.989	0.989	0.989	1.000	1.000	1.000
Laplace	0.570	0.653 ^{\$\$\$}	0.657 ^{\$\$\$}	0.849	0.859	0.863 [*]	0.997	0.996	0.997
t_2	0.402	0.494 ^{\$\$\$}	0.495 ^{\$\$\$}	0.601	0.628 ^{**}	0.632 ^{**}	0.808	0.811	0.812
exp	0.810	0.860 ^{\$\$\$}	0.868 ^{\$\$\$}	0.967	0.972	0.973	1.000	1.000	1.000
lnorm1	0.524	0.604 ^{\$\$\$}	0.615 ^{\$\$\$}	0.698	0.720 [*]	0.730 [@]	0.893	0.894	0.896
lnorm2	0.374	0.461 ^{\$\$\$}	0.472 ^{\$\$\$}	0.485	0.518 ^{**}	0.530 ^{\$\$\$}	0.677	0.680	0.687
lnorm3	0.271	0.345 ^{\$\$\$}	0.361 ^{\$\$\$}	0.336	0.363 ^{**}	0.374 ^{\$\$\$}	0.452	0.458	0.465

The differences between Z and KE are tested with significance codes defined as

^{\$\$\$}, < 0.0001

[@], < 0.001

^{**}, < 0.01

^{*}, < 0.05.

Table 3

A small data set from Graybill (1961) with $n = 7$

i_x	1	2	3	4	10	12	18
i_y	9	15	19	20	45	55	78