Exact Likelihood Ratio Test for the Parameters of the Linear Regression Model with Normal Errors

M. Chvosteková, V. Witkovský

Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia
Email: chvosta@gmail.com

Abstract. *We present an exact likelihood ratio (LR) based test for testing the simple null hypothesis on all parameters of the linear regression model with normally distributed errors. In particular, we consider simultaneous test for the regression parameters (beta) and the error standard deviation (sigma). The critical values of the LR test are presented for small sample sizes and small number of explanatory variables with standard significance level, alpha = 0.05.*

Keywords: Exact Likelihood Ratio Test, Linear Regression Model, Simultaneous Tolerance Intervals

1. Introduction

In the paper we present an exact likelihood ratio test (LRT) for testing the simple null hypothesis, $H_0 : (\beta, \sigma) = (\beta_0, \sigma_0)$ against the alternative $H_1 : (\beta, \sigma) \neq (\beta_0, \sigma_0)$, on the parameters β and σ of the linear regression model $Y = X\beta + \sigma Z$ with normally distributed errors, $Z \sim N(0, I_n)$. Although the derivation of the exact distribution of the likelihood-ratio based test statistic under the null hypothesis H_0 is straightforward, it seems that the result is not available in the standard statistical literature on linear regression models. The critical values of the LR test are presented for small sample sizes $n = k + 1, \ldots, 100$ with different number of explanatory variables, $k = 1, \ldots, 10$, and significance level 0.05.

2. Likelihood RatioTest of the Hypothesis $H_0 : (\beta, \sigma) = (\beta_0, \sigma_0)$

Consider the linear regression model $Y = X\beta + \sigma Z$ with normally distributed errors, where Y represents the n-dimensional random vector of response variables, X is the $n \times k$ matrix of non-stochastic explanatory variables (for simplicity, here we assume that X is a full-rank matrix), β is a k-dimensional vector of regression parameters, Z is an n-dimensional vector of standard normal errors, i.e. $Z \sim N(0, I_n)$, and σ is the error standard deviation, $\sigma > 0$.

Here we consider likelihood-ratio (LR) based test for testing the simple null hypothesis $H_0 : (\beta, \sigma) = (\beta_0, \sigma_0)$ against the alternative $H_1 : (\beta, \sigma) \neq (\beta_0, \sigma_0)$. Based on the above assumptions the log-likelihood function, denoted as $\ell(\beta, \sigma | Y = y)$, is given by

$$\ell(\beta, \sigma | y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta). \tag{1}$$

The (-2)-multiple of the likelihood ratio test (LRT) statistic, say $\lambda(y)$ for observed value y of Y, for testing the null hypothesis $H_0 : (\beta, \sigma) = (\beta_0, \sigma_0)$ is given by

$$\begin{aligned} \lambda(y) &= -2\left(\sup_{(\beta,\sigma)\in H_0} \ell(\beta, \sigma | y) - \sup_{(\beta,\sigma)} \ell(\beta, \sigma | y)\right) = -2\left(\ell(\beta_0, \sigma_0 | y) - \ell(\hat{\beta}_{ML}, \hat{\sigma}_{ML} | y)\right) \\ &= \frac{1}{\sigma_0^2}(y - X\beta_0)'(y - X\beta_0) - n\log\left(\frac{\hat{\sigma}^2_{ML}}{\sigma_0^2}\right) - n, \end{aligned} \tag{2}$$

where $\hat{\beta}_{ML} = \hat{\beta} = (X'X)^{-1}X'y$ is the standard least squares estimate (LSE) of β (which is also the MLE of β) and $\hat{\sigma}_{ML}$ is the maximum likelihood estimate (MLE) of the standard deviation σ, i.e. $\hat{\sigma}_{ML} = \sqrt{\frac{1}{n}(y - X\hat{\beta})'(y - X\hat{\beta})}$. Under given model assumptions, and under the null hypothesis H_0, it is straightforward to derive the distribution of the test statistic $\lambda(Y)$:

$$\begin{aligned}\lambda(Y) &\sim \frac{1}{\sigma_0^2}(Y - X\beta_0)'(Y - X\beta_0) - n\log\left(\frac{(Y - X\beta_0)'M_X(Y - X\beta_0)}{n\sigma_0^2}\right) - n \\ &\sim Z'Z - n\log\left(Z'M_X Z\right) + n\left(\log(n) - 1\right) \\ &\sim Z'(P_X + M_X)Z - n\log\left(Z'M_X Z\right) + n\left(\log(n) - 1\right) \\ &\sim Q_k + Q_{n-k} - n\log\left(Q_{n-k}\right) + n\left(\log(n) - 1\right),\end{aligned} \tag{3}$$

where $P_X = X(X'X)^{-1}X'$, $M_X = I_n - P_X$, $Z \sim N(0, I_n)$, $Q_k \sim \chi_k^2$ and $Q_{n-k} \sim \chi_{n-k}^2$ are two independent random variables with chi-square distributions, with k and $n - k$ degrees of freedom, respectively. This LRT rejects the null hypothesis $H_0 : (\beta, \sigma) = (\beta_0, \sigma_0)$ for large values of the observed test statistic $\lambda(y)$, i.e. for the given significance level $\alpha \in (0, 1)$ the test rejects the null hypothesis if

$$\lambda(y) > \lambda_{1-\alpha}, \tag{4}$$

where $\lambda_{1-\alpha}$ is the $(1 - \alpha)$-quantile of the distribution of the random variable $\lambda(Y)$, given by Eq. (3). The quantiles $\lambda_{1-\alpha}$ could be evaluated numerically, by inverting the cumulative distribution function, of the random variable $\lambda(Y)$, denoted by $\mathcal{F}_{LR}(\cdot)$:

$$\begin{aligned}\mathcal{F}_{LR}(x) &= \Pr(\lambda(Y) \leq x) \\ &= \Pr(Q_k \leq x - Q_{n-k} + n\log(Q_{n-k}) - n\left(\log(n) - 1\right)) \\ &= \int_0^\infty \mathcal{F}_{\chi_k^2}\left(x - q_{n-k} + n\log(q_{n-k}) - n(\log(n) - 1)\right) f_{\chi_{n-k}^2}(q_{n-k})\,\mathrm{d}q_{n-k},\end{aligned} \tag{5}$$

where $\mathcal{F}_{\chi_k^2}(\cdot)$ denotes the cumulative distribution function of the chi-square distribution with k degrees of freedom, and $f_{\chi_{n-k}^2}(\cdot)$ denotes the probability density function of the chi-square distribution with $n - k$ degrees of freedom. For illustration, the critical values of the LR test are presented in Table 1 for different number of explanatory variables, $k = 1, \ldots, 10$, selected small sample sizes, $n = k + 1, \ldots, 100$, and the significance level $\alpha = 0.05$. Notice that since the family of normal distributions meets regularity conditions, from standard asymptotic result about the distribution of the LRT we get $\lambda_{1-\alpha} \to \chi_{k+1,1-\alpha}^2$as $n \to \infty$, where by $\chi_{k+1,1-\alpha}^2$ we denote the $(1 - \alpha)$-quantile of chi-square distribution with $k + 1$ degrees of freedom.

The LRT could be equivalently based on the test statistic $F^\star$ defined as $F^\star = \lambda(Y)/(kS^2/\sigma_0^2)$, where $S^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(n - k)$ and $\hat{\beta} = (X'X)^{-1}X'Y$:

$$F^\star = \frac{1}{k}\frac{(\hat{\beta} - \beta_0)'X'X(\hat{\beta} - \beta_0)}{S^2} + \frac{n-k}{k} - \frac{n}{k}\frac{\log\left((n-k)S^2/n\sigma_0^2\right) + 1}{S^2/\sigma_0^2}. \tag{6}$$

Note, that the leading term in $F^\star$ is the standard F-statistic for testing the hypothesis on regression parameters $H_0 : \beta = \beta_0$ against the alternative $H_1 : \beta \neq \beta_0$. Under null hypothesis $H_0 : (\beta, \sigma) = (\beta_0, \sigma_0)$ we directly get

$$F^\star \sim \frac{Q_k/k}{Q_{n-k}/n-k} + \frac{n-k}{k} - \frac{n}{k}\frac{\log(Q_{n-k}/n) + 1}{Q_{n-k}/n-k}. \tag{7}$$

Then, the test rejects the null hypothesis if

$$F^{\star}_{obs} > F^{\star}_{1-\alpha}, \tag{8}$$

where $F^{\star}_{obs}$ denotes the observed value of the statistic $F^{\star}$ and $F^{\star}_{1-\alpha}$ is the $(1-\alpha)$-quantile of the distribution of the random variable $F^{\star}$. The quantiles $F^{\star}_{1-\alpha}$ could be evaluated by inverting the cumulative distribution function of the random variable $F^{\star}$, denoted by $\mathcal{F}_{F^{\star}}(x)$:

$$\begin{aligned}\mathcal{F}_{F^{\star}}(x) &= \Pr(F^{\star} \leq x)\\ &= \Pr\left(Q_k \leq \frac{xkQ_{n-k}}{n-k} - Q_{n-k} + n\left(\log\left(\frac{Q_{n-k}}{n}\right)+1\right)\right)\\ &= \int_0^{\infty} \mathcal{F}_{\chi^2_k}\left(\frac{xkq_{n-k}}{n-k} - q_{n-k} + n\left(\log\left(\frac{q_{n-k}}{n}\right)+1\right)\right) f_{\chi^2_{n-k}}(q_{n-k})\,\mathrm{d}q_{n-k}.\end{aligned} \tag{9}$$

The MATLAB function for computing the quantiles $\lambda_{1-\alpha}$ and $F^{\star}_{1-\alpha}$ is available upon request from the authors. More details on the numerical algorithm, as well as on its possible application for construction of the simultaneous tolerance intervals, could be found in the extended version of the paper, in Chvosteková and Witkovský (2009).

3. Discussion

The exact LR test for testing the simple null hypothesis $H_0 : (\beta, \sigma) = (\beta_0, \sigma_0)$ could be directly used to construct the exact confidence region for the parameters of the linear regression model. In particular, the exact $(1-\alpha)$-confidence region for the parameters β and σ is given as $\mathcal{C}_{1-\alpha}(Y) = \{(\beta, \sigma) : \lambda(Y) \leq \lambda_{1-\alpha}\}$. Moreover, this could be directly used for constructing the simultaneous tolerance intervals in linear regression model with normal errors, as suggested in Witkovský and Chvosteková (2009). These intervals are constructed such that, with confidence coefficient $1-\alpha$, we can claim that at least a specified proportion, say $1-\gamma$ of the population is contained in the tolerance interval, for all possible values of the predictor variates, see e.g. Lieberman and Miller (1963), Limam and Thomas (1988), De Gryze et al (2007), and Krishnamoorthy and Mathew (2009). For further details see also Chvosteková and Witkovský (2009).

Acknowledgements

The research was supported by the grants VEGA 1/0077/09, VEGA 2/7087/27, APVV-SK-AT-0003-09, and APVV-RPEU-0008-06.

References

[1] Chvosteková M. and Witkovský V. Exact likelihood ratio test for the parameters of the linear regression model with normal errors. *Measurement Science Review*, 9(1): 1–8, 2009.

[2] De Gryze S., Langhans I. and Vandebroek M. Using the correct intervals for prediction: A tutorial on tolerance intervals for ordinary least-squares regression. *Chemometrics and Intelligent Laboratory Systems*, 87: 147–154, 2007.

[3] Krishnamoorthy K. and Mathew T. *Statistical Tolerance Regions: Theory, Applications, and Computation*. Wiley, ISBN: 978-0-470-38026-0, 512 pages, 2009.

[4] Lieberman G. J. and Miller R. G. Jr. "Simultaneous Tolerance Intervals in Regression. *Biometrika,* 50: 155–168, 1963.

[5] Limam M.M.T. and Thomas D.R. Simultaneous Tolerance Intervals for the Linear Regression Model. *Journal of the American Statistical Association*, 83(403): 801–804, 1988.

[6] Witkovský V. and Chvosteková M. Simultaneous Tolerance Intervals for the Linear Regression Model. In *MEASUREMENT 2009. Proceedings of the International Conference on Measurement*, Smolenice, May 20–23, 2009.

Table 1. Critical values of the likelihood ratio test (LRT) for testing the null hypothesis on parameters of the normal linear regression model with $k = 1, \ldots, 10$ explanatory variables, $H_0 : (\beta, \sigma) = (\beta_0, \sigma_0)$ against the alternative $H_1 : (\beta, \sigma) \neq (\beta_0, \sigma_0)$, for selected small sample sizes $n = k + 1, \ldots, 100$ and the significance level $\alpha = 0.05$.

n/k	1	2	3	4	5	6	7	8	9	10
2	11.8545	-	-	-	-	-	-	-	-	-
3	8.8706	19.3470	-	-	-	-	-	-	-	-
4	7.8893	13.4989	27.1540	-	-	-	-	-	-	-
5	7.4046	11.5844	18.3296	35.2052	-	-	-	-	-	-
6	7.1164	10.6358	15.4138	23.3410	43.4538	-	-	-	-	-
7	6.9257	10.0694	13.9556	19.3806	28.5094	51.8679	-	-	-	-
8	6.7901	9.6927	13.0778	17.3814	23.4748	33.8153	60.4240	-	-	-
9	6.6888	9.4241	12.4904	16.1687	20.9120	27.6847	39.2429	69.1047	-	-
10	6.6103	9.2228	12.0691	15.3518	19.3464	24.5412	31.9998	44.7793	77.8962	-
11	6.5477	9.0663	11.7521	14.7630	18.2858	22.6089	28.2622	36.4110	50.4142	86.7873
12	6.4966	8.9412	11.5047	14.3179	17.5176	21.2931	25.9522	32.0684	40.9101	56.1388
13	6.4541	8.8388	11.3063	13.9693	16.9345	20.3360	24.3719	29.3717	35.9540	45.4906
14	6.4182	8.7536	11.1435	13.6888	16.4763	19.6068	23.2179	27.5192	32.8629	39.9134
15	6.3874	8.6813	11.0075	13.4581	16.1065	19.0319	22.3357	26.1616	30.7317	36.4216
16	6.3608	8.6195	10.8923	13.2649	15.8015	18.5667	21.6383	25.1207	29.1648	34.0061
17	6.3375	8.5659	10.7933	13.1008	15.5456	18.1821	21.0724	24.2955	27.9601	32.2250
18	6.3170	8.5190	10.7074	12.9596	15.3278	17.8587	20.6035	23.6244	27.0028	30.8522
19	6.2988	8.4776	10.6321	12.8369	15.1400	17.5829	20.2085	23.0673	26.2225	29.7589
20	6.2825	8.4408	10.5656	12.7292	14.9765	17.3448	19.8711	22.5971	25.5736	28.8659
21	6.2679	8.4079	10.5064	12.6339	14.8329	17.1371	19.5793	22.1946	25.0249	28.1220
22	6.2546	8.3783	10.4533	12.5490	14.7056	16.9544	19.3244	21.8462	24.5546	27.4919
23	6.2426	8.3515	10.4056	12.4728	14.5920	16.7923	19.0998	21.5414	24.1468	26.9512
24	6.2316	8.3271	10.3623	12.4041	14.4901	16.6475	18.9004	21.2725	23.7897	26.4816
25	6.2216	8.3048	10.3230	12.3419	14.3981	16.5175	18.7221	21.0335	23.4743	26.0700
26	6.2123	8.2844	10.2870	12.2852	14.3146	16.3999	18.5618	20.8196	23.1936	25.7060
27	6.2038	8.2657	10.2540	12.2334	14.2386	16.2932	18.4167	20.6270	22.9421	25.3817
28	6.1959	8.2483	10.2236	12.1858	14.1689	16.1959	18.2849	20.4526	22.7155	25.0910
29	6.1885	8.2323	10.1955	12.1420	14.1050	16.1067	18.1646	20.2941	22.5102	24.8288
30	6.1817	8.2174	10.1695	12.1014	14.0460	16.0247	18.0543	20.1492	22.3234	24.5911
40	6.1328	8.1115	9.9864	11.8187	13.6384	15.4640	17.3081	19.1807	21.0900	23.0435
50	6.1038	8.0497	9.8809	11.6577	13.4094	15.1533	16.9006	18.6599	20.4377	22.2394
60	6.0847	8.0092	9.8122	11.5537	13.2627	14.9557	16.6437	18.3343	20.0335	21.7459
70	6.0712	7.9806	9.7640	11.4811	13.1606	14.8190	16.4668	18.1115	19.7584	21.4120
80	6.0611	7.9594	9.7282	11.4274	13.0855	14.7187	16.3376	17.9493	19.5590	21.1709
90	6.0533	7.9429	9.7006	11.3861	13.0279	14.6421	16.2391	17.8259	19.4078	20.9886
100	6.0470	7.9299	9.6788	11.3534	12.9823	14.5816	16.1615	17.7290	19.2892	20.8460
∞	5.9915	7.8147	9.4877	11.0705	12.5916	14.0671	15.5073	16.9190	18.3070	19.6751