Published in final edited form as:

Comput Stat Data Anal. 2010 April 1; 54(4): 1079–1088. doi:10.1016/j.csda.2009.10.018.

An Empirical Likelihood-Based Method for Comparison of Treatment Effects—Test of Equality of Coefficients in Linear Models

Haiyan Su and

Department of Mathematical Sciences, Montclair State University, Montclair, New Jersey 07043, USA

Hua Liang

Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642 USA

Summary

To compare two treatment effects, which can be described as the difference of the parameters in two linear models, we propose an empirical likelihood-based method to make inference for the difference. Our method is free of the assumptions of normally distributed and homogeneous errors, and equal sample sizes. The empirical likelihood ratio for the difference of the parameters of interest is shown to be asymptotically chi-squared. Simulation experiments illustrate that our method outperforms the published ones. Our method is used to analyze a data set from a drug study.

Keywords

Bartlett correction; Coverage probability; Projection; Profile empirical likelihood

1 Introduction

In epidemiologic and biomedical studies, a common concern that often arises is how to evaluate the difference between two treatments. Many methods have been proposed to evaluate the difference in special cases. One example is the Behrens-Fisher problem (Pfanzagl, 1974), which is a powerful and popular tool to study the difference between the means of two independent and normally distributed populations. However, in some situations, comparison of the response means from two populations ignores the fact that two populations may not be identical. Measuring a treatment effect may need to take into account the effect of other covariates. For instance, we may be interested in two HIV treatment effects, indicated by viral load RNA levels, on two groups of HIV patients. The patients from one group may be substantially younger and have higher CD4 counts than those from the other group. A difference in viral load RNA levels between the two groups may be due to the contribution of innate immunity rather than the treatment because younger patients tend to have stronger immunity systems. Bhuyan and Majumder (1996) give another simple example. This concern gives rise to a comparison of coefficients in linear regression models.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

A pioneering work on testing linear regression equality was done by Chow (1960), in which he proposed a statistical ratio to conduct the hypothesis test. The numerator of the ratio is found by the difference between the sum of residual squares, assuming equality, the sum of residual squares without assuming equality, which is also the denominator. The resulting ratio was shown to follow an F distribution under the null hypothesis. Toyada (1974) noticed that the Chow test is fairly robust to the inequality of the variances of the two error terms, provided that at least one of the two sample sizes is large. However, when both sample sizes are small, the significance level of the test is considerably affected by even moderate heteroscedasticity. Since then, various versions of the Chow test have been proposed. Jayatissa (1977) gave a small-sample exact test on the same problem, but the test is inefficient because confidence intervals based on the test statistic are large (Weerahandi, 1987). Schmidt and Sickles (1977) provided a formula to calculate the exact tail probability of the Chow test when the ratio of two variances is known. They further examined Toyoda's procedure and concluded that Toyoda's approximation is reasonably accurate only when the two variances and the sample sizes are of the same order of magnitude. Ali and Silver (1985) proposed two relatively robust tests on the basis of the Chow test and likelihood ratio statistics. However, the distributions of their tests are unknown and need to be approximated by the Pearson system of distributions using the moments of statistics under the null hypothesis. Conerly and Mansfield (1988) presented an approximate test. Their method modified the Chow statistic by replacing the pooled residual variance in the denominator with a weighted average of the residual variances from each group. This method provides an alternative for comparing heteroscedastic regression models. In comparing slopes of two linear regressions to calculate sample size and power function, Dupont and Plummer (1998) developed an intuitive test, which has been applied to clinical trials, and qualified its applicability when the two error terms have the same variance. In contrast to the above tests, Weerahandi (1987) gave a simple exact test, which involves only one-dimensional integration along with independent and normally distributed errors. Weerahandi showed that his test follows an F distribution under the null hypothesis, and claimed that his test is the conditionally uniformly most powerful invariant unbiased test. Yang and Zhao (2007) proposed a test of treatment effect via empirical likelihood using weighted log rank tests.

All of the above tests were derived from the likelihood principle and therefore need a distributional assumption. In addition, the results of these tests were obtained subject to additional information about the ratio of the variances, that is, either the ratio is known, or the magnitude of the variance is of the same order. These assumptions are not always satisfied or at least need to be diagnosed. In this paper, we develop a test that is free of these assumptions on the basis of the empirical likelihood principle, a data-driven nonparametric inference, which has been widely used to construct confidence regions. It was first proposed by Thomas and Grunkemeier (1975) and revived by Owen (1988). This nonparametric likelihood method makes no distribution assumption and therefore has the robustness and flexibility of nonparametric methods, with the efficiency of the likelihood principle. In particular, it does not involve asymptotic variance estimation, which may be complicated in non-parametric or semi-parametric models. The confidence interval or confidence region may not be symmetric or even concave (Owen, 2001). The confidence region has a data-determined shape, thereby better reflecting the true shape of the underlying distribution. These features make it more applicable in real studies, and it is more popular than its parametric counterparts such as normal approximation and bootstrap approximation (Hall and La Scala, 1990).

The contribution of this paper is to develop a more efficient approach to make inferences for the difference of coefficients in two linear regressions. We first derive an empirical likelihood-based statistic for the difference between the two regression coefficients and show that the statistic is asymptotically chi-squared, leading to tests and confidence regions for our concern. We also apply the Bartlett correction method to obtain an adjusted confidence interval as this

correction can provide a simple mechanism to improve the accuracy of the inference, using a similar theory to those developed by DiCiccio et al. (1991) for smooth functions of means, Chen and Hall (1993) for quantiles, and Chen and Cui (2006) for estimating equations with nuisance parameters.

This paper is organized as follows. In Section 2, we propose the empirical likelihood-based statistic and show that it has an asymptotic chi-squared distribution. Section 3 reports the results of the simulation experiments. Section 4 presents the results obtained for a drug study. The proof of the theoretical result is given in the Appendix.

2 Empirical Likelihood-Based Test

Assume that the data are sampled from the models

$$\begin{cases} y_1 = x_1^T \beta_1 + \varepsilon_1, \\ y_2 = x_2^T \beta_2 + \varepsilon_2, \end{cases} \quad (2.1)$$

where ε_1 and ε_2 are two independent random errors with a mean of zero and variances σ_1^2 and σ_2^2 respectively, and β_1 and β_2 are two unknown parameter vectors of length p , which indicate the treatment effects. We are interested in testing the hypothesis $H_0: \beta_1 = \beta_2$ against $H_1: \beta_1 \neq \beta_2$.

Let $\{(x_{11}, y_{11}), \dots, (x_{1n_1}, y_{1n_1})\}, \{(x_{21}, y_{21}), \dots, (x_{2n_2}, y_{2n_2})\}$ be the independent samples from model (2.1). Let $\mathbf{X}_1 = (x_{11}, \dots, x_{1n_1})^T$, $\mathbf{X}_2 = (x_{21}, \dots, x_{2n_2})^T$, and similarly for $\mathbf{Y}_1$ and $\mathbf{Y}_2$, ε_1 and ε_2 , then model (2.1) can be written in the form of

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}. \quad (2.2)$$

Let $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T)^T$, $\mathbf{X}$ be the block diagonal matrix $\text{diag}(\mathbf{X}_1, \mathbf{X}_2)$, $\beta = (\beta_1^T, \beta_2^T)^T$, and $\varepsilon = (\varepsilon_1^T, \varepsilon_2^T)^T$. Note that model (2.2) can further be expressed as

$$\mathbf{Y} = \mathbf{X}_1^* \beta_1^* + \mathbf{X}_2^* \beta_2 + \varepsilon, \quad (2.3)$$

where $\mathbf{X}_1^* = (\mathbf{X}_1^T, 0)^T$, $\mathbf{X}_2^* = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$, and $\beta_1^* = \beta_1 - \beta_2$. Therefore, H_0 is equivalent to $\beta_1^* = 0$. We now treat β_2 as a nuisance parameter. Our estimation method, motivated by partial regression plots, is to form partial residual vectors adjusting the influence of $\mathbf{X}_2^*$, i.e., $\mathbf{Y} - E(\mathbf{Y}|\mathbf{X}_2^*)$ and $\mathbf{X}_1^* - E(\mathbf{X}_1^*|\mathbf{X}_2^*)$. It follows from (2.3) that

$$\mathbf{Y} - E(\mathbf{Y}|\mathbf{X}_2^*) = \{\mathbf{X}_1^* - E(\mathbf{X}_1^*|\mathbf{X}_2^*)\} \beta_1^* + \varepsilon. \quad (2.4)$$

If $E(\mathbf{Y}|\mathbf{X}_2^*)$ and $E(\mathbf{X}_1^*|\mathbf{X}_2^*)$ are known, we can define estimates of β_1^* in a standard way. In what follows, we denote $\tilde{S} = S - E(S|\mathbf{X}_2^*)$ and $\widehat{S} = S - \widehat{E}(S|\mathbf{X}_2^*)$ for any vector S . For example,

$\tilde{Y} = Y - E(Y|X_2^*)$ and $\widehat{Y} = Y - \widehat{E}(Y|X_2^*)$. Let $\widehat{E}(Y|X_2^*)$ and $\widehat{E}(X_1^*|X_2^*)$ be the least square estimates of $E(Y|X_2^*)$ and $E(X_1^*|X_2^*)$ respectively. We may estimate β_1^* by solving the estimating equation

$$\frac{1}{n} \sum_{i=1}^n \widehat{X}_{1i}^{*T} (\widehat{Y}_i - \widehat{X}_{1i}^* \beta_1^*) = 0,$$

as $E[\widehat{X}_1^{*T} (\widehat{Y} - \widehat{X}_1^* \beta_1^*)] = 0$. This statement motivates us to define our empirical likelihood ratio as follows.

Let F be the distribution function which assigns probability p_i at points $(Y_i, X_{1i}^*, X_{2i}^*)$. The empirical likelihood ratio function for β_1^* is therefore defined as

$$\mathcal{R}_n(\beta_1^*) = \sup \left\{ \prod_{i=1}^n n p_i \mid \sum_{i=1}^n p_i \widehat{X}_{1i}^{*T} (\widehat{Y}_i - \widehat{X}_{1i}^* \beta_1^*) = 0, p_i \geq 0, \sum_{i=1}^n p_i = 1 \right\}. \quad (2.5)$$

Theorem 1

Assuming that $E(\|X_k\|^4) < \infty$, $E(\|X_k\|^2 Y_k^2) < \infty$ and $E(X_k^T X_k)$ are nonsingular for $k = 1, 2$, then $-2\log\{\mathcal{R}_n(\beta_1^*)\}$ converges in distribution to a chi-squared distribution with p degrees of freedom.

Based on Theorem 1, we can obtain an estimate of β_1^* and the associated $100(1 - \alpha)\%$ confidence region:

$$\{\beta_1^* : -2\log\{\mathcal{R}_n(\beta_1^*)\} \leq c_\alpha\}, \quad (2.6)$$

where c_α is the $1 - \alpha$ quantile of the χ_p^2 distribution satisfying $P(\chi_p^2 < c_\alpha) = 1 - \alpha$.

To further improve the accuracy of the inference, the empirical likelihood ratio may be Bartlett-corrected with a higher-order error than the usual error term of the order $O(n^{-1})$. The basic idea is to multiply the χ^2 threshold by a constant $(1 + B_c/n)$ instead of 1, where B_c is the Bartlett correction constant. That is, by applying the Bartlett-correction into formula (2.6), we have

$$\text{pr}\{-2\log\{\mathcal{R}_n(\beta_1^*)\} < c_\alpha(1 + n^{-1} B_c)\} = 1 - \alpha + O(n^{-b}), \text{ for some } b \in (1, 2].$$

Because it is very difficult to obtain an exact expression for B_c , we give an estimator of $(1 + n^{-1} B_c)$ by using the bootstrap procedure in the following steps, which have successfully been applied in a more complex setting by Chen and Cui (2006).

1. Generate the bootstrap samples of size n by sampling with replacement from the samples $\{\widehat{X}_{1i}^*\}_{i=1}^n$ and $\{\widehat{Y}_i\}_{i=1}^n$ respectively after projection. Calculate $-2\log\{\mathcal{R}_n^*(\widehat{\beta}_1^*)\}$ based on the resampling, where $\widehat{\beta}_1^*$ is the global maximum empirical likelihood estimator of β_1^* based on the original samples $\{\widehat{X}_{1i}^*\}_{i=1}^n$ and $\{\widehat{Y}_i\}_{i=1}^n$.

2. Repeat Step 1 B times to obtain $-2\log\{\mathcal{R}_{n1}^*(\widehat{\beta}_1^*)\}, -2\log\{\mathcal{R}_{n2}^*(\widehat{\beta}_1^*)\}, \dots, -2\log\{\mathcal{R}_{nB}^*(\widehat{\beta}_1^*)\}$ and $-2B^{-1}\sum_{b=1}^B \log\{\mathcal{R}_{nb}^*(\widehat{\beta}_1^*)\}$, which is the bootstrap estimator of $E[-2\log\{\mathcal{R}_n(\widehat{\beta}_1^*)\}]$.

The bootstrap estimator of $\tau = 1 + n^{-1}B_c$ is $\widehat{\tau} = -2p^{-1}B^{-1}\sum_{b=1}^B \log\{\mathcal{R}_{nb}^*(\widehat{\beta}_1^*)\}$. Correspondingly, the Bartlett-corrected confidence region is given by

$$CI_\alpha = \{\beta_1^* | -2\log\{\mathcal{R}_n(\beta_1^*)\} \leq \widehat{\tau}c_\alpha\}. \quad (2.7)$$

Remark 1

These procedures can be generalized to the situations where x_1 and x_2 have different dimensions. We are interested in the effects of only partial elements of x_1 and x_2 , e.g., x_1^I and x_2^I . Let x_1^c and x_2^c be the remaining elements respectively. We regress y_1 and x_1^I on x_1^c , and y_2 and x_2^I on x_2^c , to remove the influence of $x_1^c(x_2^c)$ from y_1 and $x_1^I(y_2$ and $x_2^I)$. The produced partial residual models fit the setting given in (2.1). This partial projection procedure is particularly useful when we have multi-dimensional covariates. However, we are interested in the effects of only partial covariates. Although a profile empirical likelihood can be used in this situation, the computational burden is heavy. Implementing the proposed procedure will tremendously reduce the computational burden.

Remark 2

The proposed test can also be extended to do multi-treatment comparisons (Liu, Jamshidian, and Zhang, 2004). For instance, consider models

$$\begin{cases} y_1 = x_1^T \beta_1 + \varepsilon_1, \\ y_2 = x_2^T \beta_2 + \varepsilon_2, \\ y_3 = x_3^T \beta_3 + \varepsilon_3, \end{cases} \quad (2.8)$$

where $\varepsilon_1, \varepsilon_2$ and ε_3 are three independent random errors with a mean of zero and variances σ_1^2, σ_2^2 and σ_3^2 respectively, and β_1, β_2 and β_3 are unknown parameters of the treatment effects. We may be interested in the joint comparison $H_0: \beta_1 = \beta_2 = \beta_3$ against H_1 : at least one of them is different from others. In a similar way, we can express (2.8) as $Y = X^* \beta^* + \varepsilon$, where $Y = (y_1^T, y_2^T, y_3^T)^T$, $\varepsilon = (\varepsilon_1^T, \varepsilon_2^T, \varepsilon_3^T)^T$, and $\beta^* = (\beta_1 - \beta_2, \beta_2 - \beta_3, \beta_3)^T$. Testing H_0 is equivalent to testing $(\beta_1 - \beta_2, \beta_2 - \beta_3) = (0, 0)$.

3 Simulation Study

To evaluate the finite-sample performance of the proposed methods, we conduct extensive small-sample simulation experiments in this section to consider level and power studies. In Examples 1 and 2 we simulate data from model (2.1), while in Example 3 we simulate data from model (2.8).

We set the nominal level of the confidence interval to be 95% with type-I error $\alpha = 0.05$ and run each simulation 1,000 times, with the results being the average score of the 1,000 simulations. We then compare the performance of the proposed test with three other tests: the empirical t-test for comparing regression coefficients in two regression models (Sachs, 1982) (the ANOVA test for three regression models), the Weerahandi test, and the Dupont and

Plummer test. The first two tests require that the two error terms follow the normal distribution, while the third one also assumes that the two error terms have the same variance.

Example 1

X_1 and X_2 are generated as follows. Let $u_1 \sim \text{Unif}(1, 10)$ and $u_2 \sim \text{Unif}(1, 10)$, then $X_1 = \exp(u_1/4) + \varepsilon_1$ and $X_2 = \exp(u_2/4) + \varepsilon_2$, where $\varepsilon_j \sim N(0, 4)$. Let $\mathbf{Y}_j = 1 + \mathbf{X}_j\beta_j + \varepsilon_j$, where ε_j are the error terms with a mean of 0 and a variance of σ_j^2 . Sample sizes are $n_1 = 15$ and $n_2 = 15$. We conduct simulations, for eight different δ , the true difference between β_1 and β_2 : 0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35. We consider the following four cases for the error terms ε_j .

Case 1: ε_1 and ε_2 independently follow $N(0, \sigma_j^2)$ for $j = 1, 2$, and $\sigma_1^2 = \sigma_2^2 = 1$.

Case 2: Same as in Case 1, except σ_1 and σ_2 are different with $\sigma_1^2 = 1$ and $\sigma_2^2 = 5$. This design is to check the performance of the proposed test when the variances of two errors are unequal.

Case 3: ε_1 and ε_2 independently follow $\text{Unif}(-1, 1)$ $\sqrt{3}\sigma_j$; $\sigma_1^2 = \sigma_2^2 = 1$. This is to check the proposed test when the normal distribution assumption is violated.

Case 4: Same as in Case 3, except $\sigma_1^2 = 1$ and $\sigma_2^2 = 5$. This is to check whether the proposed test works well when there is no distribution and homoscedasticity assumption.

Table 1 lists the type-I errors for the four tests: the proposed test, the Weerahandi test, the t-test, and the Dupont and Plummer test. These empirical levels are all close to the nominal level, except those of the Weerahandi test, which are slightly less than the nominal level. These levels are not very sensitive to the distribution or homoscedasticity of the error terms, as shown in Table 1.

The empirical power as a function of δ is shown in Figure 1, which increases when the true difference δ increases. It can be seen that both the proposed test and the Weerahandi test have better power performance, and the former is better than the latter when their power is less than 0.98. This superiority stands out in Case 4, when the normal distribution and homoscedasticity assumptions are violated. Both t test and Dupont and Plummer test give almost always identical power and type-I error. We didn't show the results for t test here for their indistinguishableness.

Example 2

We generate data from (2.1) with $n_1 = 25$ and $n_2 = 35$. X_1 and X_2 are 3-dimensional and 4-dimensional covariates, consisting of $\{\exp(u_1/4) + \varepsilon_{11}, u_1 + 5 + \varepsilon_{12}, 5 \log(u_1 + 1) + \varepsilon_{13}\}$, and $\{\exp(u_2/4) + \varepsilon_{21}, u_2 + 5 + \varepsilon_{22}, 5 \log(u_2 + 1) + \varepsilon_{23}, 4 \sqrt{u_2} + \varepsilon_{24}\}$ respectively, where $u_1 \sim \text{Unif}(1, 10)$ and $u_2 \sim \text{Unif}(1, 10)$ and ε_{jl} are independently generated from $N(0, 4)$ for $j = 1, 2, l = 1, \dots, 4$. We consider the same four cases for the structures of ε_j as in Example 1. In this example, we are only interested in δ , which is the difference of the effects in the first component from the two models.

As pointed out in Remark 1, we do a partial projection first, i.e., we project the first component and response variable on the remaining covariates, and then apply these four tests to the "residual" data. The type-I errors are reported in Table 2. The type-I errors of the proposed test are slightly higher than the nominal level when the errors are normally distributed, while the Weerahandi test has a much lower type-I error than the nominal level.

To examine the power performance of the proposed test, we generate data for six different δ values ranging in $[0, 0.25]$ with equal-interval for Cases 1 and 2 and 0, 0.05, 0.1, 0.15, 0.2,

0.25, 0.4, 0.6 for Cases 3 and 4, where δ is the true difference of β_1 and β_2 . The empirical powers of all four tests (Figure 2) have a similar pattern that increases as the true difference δ increases. Similarly we have dropped the t test power curve in Figure 2 since it is indistinguishable from Dupont and Plummer test. When the sample size grows (Table 3), the power decreases as the variance of the errors increases (Table 4). This may not be surprising because when the noise becomes more involved, detecting deviations caused by noise or by coefficients themselves is more difficult. We also see that the proposed test has a higher power to detect differences than the other three tests. It is also worthwhile mentioning that the Weerahandi test has lowest power, which seems to conflict the declaration of ‘to be hard to beat’. The reason may be because we do a projection first and the “residual” data are not independent.

Example 3

To test the equality of three group means, i.e., $X_j \equiv 1$, we generate data from (2.8), in which $n_1 = 35$, $n_2 = 45$, and $n_3 = 50$ respectively. The true differences, $(\beta_1 - \beta_2, \beta_2 - \beta_3)$, are (0, 0), (-0.3, 0.3), and (-0.4, -0.5) respectively. The four cases of ε_j are the same as those in Example 1, with variance values $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ being (1, 1, 1) in Cases 1 and 3, and (1, 2, 3) in Cases 2 and 4. We compare the simulation results for the proposed test, ANOVA, and the Weerahandi test, and list the results in Table 5. It can be seen that the proposed test has a comparable performance with other tests when the errors are normally distributed (Cases 1 and 2) and a superior performance to other tests when the normality assumption is violated (Cases 3 and 4).

4 Drug Study

In this section, we analyze a data set (Hocking, 2003) from a medical experiment involving four drugs to measure their effects on the response y to a particular stimulus. Because the individuals in this study may not be identical in their responses to the drugs, their response x to the stimulus prior to taking the drug was measured. For each drug, there are 9 observations (x, y) (Figure 3), where squares, octagons, triangles, and diamonds indicate the observed values with regard to drugs A, B, C, and D respectively. Linear model fitting of each drug is also shown in Figure 3, in which the solid, slashed, dotted, and broken-dotted lines correspond to drugs A, B, C, and D respectively.

By observing y values in the four treatments, one may see that the treatment effects in Groups A and C are stronger than those in Groups B and D. We are interested in whether these effects are significant and how we can provide a statistical justification for them.

We fit model $y_{ij} = \alpha_j + x_{ij}\beta_j$ and apply the four tests to this data set for a pair comparison. We obtain the empirical likelihood-based confidence intervals for the difference of the treatment effects for each pair as follows. A vs B: (0.048, 0.277); A vs C: (-0.161, 0.217); A vs D: (0.066, 0.291); B vs C: (-0.315, 0.002); B vs D: (-0.051, 0.113); and C vs D: (0.046, 0.358). These results indicate that the differences between A and B, A and D, and C and D are statistically significant. The results for the other three methods are presented in Table 6, and are inconsistent with those of the proposed test. That is, the results based on the Weerahandi test indicate that only the difference between A and D is significant, while the t-test and the Dupont and Plummer test fail to detect any differences in these paired comparisons. We prefer the results obtained by using the empirical likelihood-based test based on the simulation results and the patterns shown in Figure 3.

To further justify these findings, we generate 1, 000 bootstrap samples and present the proportion of times that H_0 is rejected based on these resampling data for our proposed test, the t-test, and the Dupont and Plummer test (Table 7). This table does not include the bootstrap proportion of rejecting H_0 for the Weerahandi test as the Weerahandi test gives its own p-value.

The proportions of rejection of the equalities between A and B, A and D, and C and D using our proposed test (EL) are 66.1%, 63.1%, and 43.1% respectively, which are consistent with the test result given earlier. The percentages of rejection of the equalities between A and B, A and D, and C and D are the highest among the three tests in the table. The percentage of rejection of the equality between A and C, B and C, and B and D are the smallest compared to the other tests in the table (4.7%, 4.4%, and 3.3% respectively). That is also consistent with the results that the differences between A and C, B and C, and B and D are not significant. In summary, the bootstrap results shown in Table 7 support our conclusion obtained by the proposed test.

As mentioned in Remark 2, we may be also interested in test of equality of all slopes. In the same way as in Example 3, we implement the ANOVA, the Weerahandi test, and the proposed test. The first one fails to reject H_0 with p-value 0.305, while the Weerahandi test rejects H_0 with p-value 0.019. The proposed test also rejects the null hypothesis because the confidence region does not contain the vector $\mathbf{0}$.

In summary, based on the simulation results and the patterns shown in Figure 3, we prefer the results obtained by using the empirical likelihood-based test.

5 Discussion

To efficiently compare treatment effects under weaker assumptions on data, we develop an empirical likelihood-based approach to construct the confidence region for the difference between the coefficients of interest. The proposed approach is more applicable than its normal-distribution-based competitor. The finite-sample performance indicates the proposed test outperforms other tests in most situations for comparing two regression coefficients. It is reliable and powerful in detecting the difference of the parameters of interest.

The computation of the proposed test is simple, and the theoretical results do not need any distribution assumption nor the homoscedasticity assumption. This makes it applicable to real studies in detecting and comparing two treatment effects.

In this article, we only study general linear models. We may consider two partially linear models, both of which take the form

$$Y + X^T\beta + g(Z) + \varepsilon,$$

where $g(\cdot)$ is an unknown smoothing function and Z is a covariate, which nonlinearly contributes to Y . We may use the same idea to study the difference between two linear parameter β s.

Acknowledgments

The authors thank the editor, Associate editor, and one referee for their comments and suggestions, which improved the original manuscript. Liang's research was partially supported by grant AI059773 from NIH/NIAID and DMS 0806097 from NSF.

References

- Ali MM, Silver JL. Tests for equality between sets of coefficients in two linear regressions under heteroscedasticity. *J Amer Statist Assoc* 1985;80:730–735.
- Bhuyan KC, Majumder AK. Testing equality of regressions when error variances are heterogeneous to study the production differentials of rice in Bangladesh. *Biom J* 1996;38:819–828.

- Chen SX, Cui HJ. On bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika* 2006;93:215–220.
- Chen SX, Hall P. Smoothed empirical likelihood confidence intervals for quantiles. *Ann Statist* 1993;21:24–40.
- Chow GC. Tests of equality between sets of coefficients in two linear regressions. *Econometrica* 1960;28:591–605.
- Conerly MD, Mansfield ER. An approximate test for comparing heteroscedastic regression models. *J Amer Statist Assoc* 1988;83:811–817.
- DiCiccio T, Hall P, Romano J. Empirical likelihood is bartlett-correctable. *Ann Statist* 1991;19:1053–1061.
- Dupont WD, Plummer WD. Power and sample size calculations for studies involving linear regression. *Controlled Clinical Trials* 1998;19:589–601. [PubMed: 9875838]
- Hall P, La Scala B. Methodology and algorithms of empirical likelihood. *Intern Statist Rev* 1990;58:109–127.
- Hocking, RR. *Methods and Applications of Linear Models: Regression and Analysis of Variance*. 2. New Jersey: John Wiley & Sons; 2003.
- Jayatissa WA. Tests of equality between sets of coefficients in two linear regressions when disturbance variances are unequal. *Econometrica* 1977;45:1291–1292.
- Liang H, Wang S, Carroll RJ. Partially linear models with missing response variables and error-prone covariates. *Biometrika* 2007;94:185–198.
- Liu W, Jamshidian M, Zhang Y. Multiple comparison of several linear regression models. *J Amer Statist Assoc* 2004;99:395–403.
- Owen AB. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 1988;75:237–249.
- Owen, AB. *Empirical Likelihood*. London: Chapman and Hall/CRC; 2001.
- Pfanzagl J. On the Behrens-Fisher problem. *Biometrika* 1974;61:39–47.
- Sachs, L. *Applied Statistics: A Handbook of Techniques*. Springer-verlag; 1982.
- Schmidt P, Sickles R. Some further evidence on the use of the Chow test under heteroskedasticity. *Econometrica* 1977;45:1293–1298.
- Thomas DR, Grunkemeier GL. Confidence interval estimation of survival probability for censored data. *J Amer Statist Assoc* 1975;70:865–871.
- Toyoda T. Use of the Chow tests under heteroskedasticity. *Econometrica* 1974;42:601–608.
- Weerahandi S. Testing regression equality with unequal variances. *Econometrica* 1987;55:1211–1215.
- Yang S, Zhao Y. Testing treatment effect via empirical likelihood using weighted log rank tests. *Statist Probab Lett* 2007;77:1385–1393.

Appendix: Proof of Theorem 1

Suppose $\mathbf{X}_1^* = \gamma_0 + \mathbf{X}_2^* \gamma_1 + \varepsilon_1$, which along with (2.3) indicates that $\mathbf{Y} = \nu_0 + \mathbf{X}_2^* \nu_1 + \varepsilon_1 \beta_1^* + \varepsilon$. Let $\hat{\gamma}_k$ and $\hat{\nu}_k$ be the least squares estimators of γ_k and ν_k for $k = 0, 1$, respectively. Then $E(\mathbf{X}_1^* | \mathbf{X}_2^*) = \gamma_0 + \mathbf{X}_2^* \gamma_1$ and $E(\mathbf{Y} | \mathbf{X}_2^*) = \nu_0 + \mathbf{X}_2^* \nu_1$.

Write $\Omega_i = \{X_{1i}^* - \widehat{E}(X_{1i}^* | X_{2i}^*)\}^T [Y_i - \widehat{E}(Y_i | X_{2i}^*) - \{X_{1i}^* - \widehat{E}(X_{1i}^* | X_{2i}^*)\} \beta_1^*]$ and $\tilde{\Omega}_i = \{X_{1i}^* - E(X_{1i}^* | X_{2i}^*)\}^T [Y_i - E(Y_i | X_{2i}^*) - \{X_{1i}^* - E(X_{1i}^* | X_{2i}^*)\} \beta_1^*]$ for $i = 1, \dots, n$. A standard simplification as in Owen (2001) yields

$$p_i = \frac{1}{n(1 + a^T \Omega_i)}, i = 1, \dots, n,$$

where a is the solution of the equation

$$n^{-1} \sum_{i=1}^n \frac{\Omega_i}{1+a^T \Omega_i} = 0.$$

A direct calculation yields that

$$\begin{aligned} \Omega_i &= (\widehat{X}_{1i} - \widetilde{X}_{1i})^T [(\widehat{Y}_i - \widetilde{Y}_i) - (\widehat{X}_{1i} - \widetilde{X}_{1i})\beta_1^* + (\widetilde{Y}_i - \widetilde{X}_{1i})\beta_1^*] \\ &\quad + \widetilde{X}_{1i}^T [(\widehat{Y}_i - \widetilde{Y}_i) - (\widehat{X}_{1i} - \widetilde{X}_{1i})\beta_1^* + (\widetilde{Y}_i - \widetilde{X}_{1i})\beta_1^*] \\ &= \widetilde{\Omega}_i + o_p(1). \end{aligned} \quad (\text{A.1})$$

where $o_p(1)$ is independent of the index i .

Using the arguments similar to the proof of Theorem 3.2 of Owen (2001), we have $\|a\| = O_p(n^{-1/2})$ and $\max_{1 \leq i \leq n} \|\widetilde{\Omega}_i\| = o_p(n^{1/2})$. By (A.1), we have $\max_{1 \leq i \leq n} \|\Omega_i\| \leq \max_{1 \leq i \leq n} \|\widetilde{\Omega}_i\| + o_p(1) = o_p(n^{1/2})$.

Using the same argument as the proof of Theorem 4 in Liang et al. (2007), we have

$$\begin{aligned} 2\log\{\mathcal{R}_n(\beta_1^*)\} &= \sum_{i=1}^n a^T \Omega_i \Omega_i^T + o_p(1) \\ &= (n^{-1/2} \sum_{i=1}^n \Omega_i)^T (n^{-1} \sum_{i=1}^n \Omega_i \Omega_i^T)^{-1} (n^{-1/2} \sum_{i=1}^n \Omega_i) + o_p(1). \end{aligned}$$

To show Theorem 1, we first show that $n_1^{-1/2} \sum_{i=1}^{n_1} \Omega_i$ and $n_1^{-1/2} \sum_{i=1}^{n_1} \widetilde{\Omega}_i$ have the same limiting normal distribution, $n_2^{-1/2} \sum_{i=n_1+1}^n \Omega_i$ and $n_2^{-1/2} \sum_{i=n_1+1}^n \widetilde{\Omega}_i$ have the same limiting normal distribution, $n_1^{-1} \sum_{i=1}^{n_1} \Omega_i \Omega_i^T$ and $n_1^{-1} \sum_{i=1}^{n_1} \widetilde{\Omega}_i \widetilde{\Omega}_i^T$ have the same limiting value, and $n_2^{-1} \sum_{i=n_1+1}^n \Omega_i \Omega_i^T$ and $n_2^{-1} \sum_{i=n_1+1}^n \widetilde{\Omega}_i \widetilde{\Omega}_i^T$ have the same limiting value. Without loss of generality, we will prove it for the first group, that is, for $i = 1, \dots, n_1$. The second group, $i = n_1 + 1, \dots, n$, can be proved similarly.

Note that $\widetilde{\Omega}_i - \Omega_i$ can be decomposed as

$$\begin{aligned} &\widetilde{X}_{1i}^* [\widehat{E}(Y_i|X_{2i}^*) - E(Y_i|X_{2i}^*) - \{ \widehat{E}(X_{1i}^*|X_{2i}^*) - E(X_{1i}^*|X_{2i}^*) \} \beta_1^*] \\ &- \{ \widehat{E}(X_{1i}^*|X_{2i}^*) - E(X_{1i}^*|X_{2i}^*) \}^T [\widehat{E}(Y_i|X_{2i}^*) - E(Y_i|X_{2i}^*) - \{ \widehat{E}(X_{1i}^*|X_{2i}^*) - E(X_{1i}^*|X_{2i}^*) \} \beta_1^*] \\ &+ \{ \widehat{E}(X_{1i}^*|X_{2i}^*) - E(X_{1i}^*|X_{2i}^*) \}^T (\widetilde{Y}_i - \widetilde{X}_{1i} \beta_1^*) \end{aligned}$$

Note that

$$\frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \widetilde{X}_{1i}^* [\widehat{E}(Y_i|X_{2i}^*) - E(Y_i|X_{2i}^*)] = \sqrt{n_1}(\widehat{\nu}_1 - \nu_1) \frac{1}{n_1} \sum_{i=1}^{n_1} \varepsilon_{1i} X_{2i}^* + \sqrt{n_1}(\widehat{\nu}_0 - \nu_0) \frac{1}{n_1} \sum_{i=1}^{n_1} \varepsilon_{1i}.$$

The right-hand side is of order $o_p(1)$ since $\sqrt{n_1}(\widehat{\nu}_k - \nu_k) = O_p(1)$ for $k = 0, 1$, $1/n_1 \sum_{i=1}^{n_1} \varepsilon_{1i} X_{2i}^* = o_p(1)$, and $1/n_1 \sum_{i=1}^{n_1} \varepsilon_{1i} = o_p(1)$. It follows that $n_1^{-1/2} \sum_{i=1}^{n_1} \widetilde{X}_{1i}^T \{\widehat{E}(Y_i | X_{2i}^*) - E(Y_i | X_{2i}^*)\} = o_p(1)$.

In the same way, we obtain the following statements.

$$\begin{aligned} & \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \widetilde{X}_{1i}^T \{\widehat{E}(X_{1i}^* | X_{2i}^*) - E(X_{1i}^* | X_{2i}^*)\} \beta_1^* = o_p(1). \\ & \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \{\widehat{E}(X_{1i}^* | X_{2i}^*) - E(X_{1i}^* | X_{2i}^*)\}^T \{\widehat{E}(Y_i | X_{2i}^*) - E(Y_i | X_{2i}^*)\} = o_p(1). \\ & \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \{\widehat{E}(X_{1i}^* | X_{2i}^*) - E(X_{1i}^* | X_{2i}^*)\}^T \{\widehat{E}(X_{1i}^* | X_{2i}^*) - E(X_{1i}^* | X_{2i}^*)\} \beta_1^* = o_p(1). \\ & \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \{\widehat{E}(X_{1i}^* | X_{2i}^*) - E(X_{1i}^* | X_{2i}^*)\}^T \widetilde{Y}_i = o_p(1). \\ & \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \{\widehat{E}(X_{1i}^* | X_{2i}^*) - E(X_{1i}^* | X_{2i}^*)\}^T \widetilde{X}_{1i} \beta_1^* = o_p(1). \end{aligned}$$

These arguments imply that $n_1^{-1/2} \sum_{i=1}^{n_1} \Omega_i$ and $n_1^{-1/2} \sum_{i=1}^{n_1} \widetilde{\Omega}_i$ have the same limiting normal distribution, and $n_1^{-1} \sum_{i=1}^{n_1} \Omega_i \Omega_i^T$ and $n_1^{-1} \sum_{i=1}^{n_1} \widetilde{\Omega}_i \widetilde{\Omega}_i^T$ have the same limiting value. In addition, we know that $(n_1^{-1/2} \sum_{i=1}^{n_1} \widetilde{\Omega}_i)^T (n_1^{-1} \sum_{i=1}^{n_1} \widetilde{\Omega}_i \widetilde{\Omega}_i^T) (n_1^{-1/2} \sum_{i=1}^{n_1} \widetilde{\Omega}_i)$ converges to χ_p^2 in distribution. It follows that $(n_1^{-1/2} \sum_{i=1}^{n_1} \Omega_i)^T (n_1^{-1} \sum_{i=1}^{n_1} \Omega_i \Omega_i^T) (n_1^{-1/2} \sum_{i=1}^{n_1} \Omega_i)$ converges to χ_p^2 in distribution.

Furthermore, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{\Omega}_i = \frac{\sqrt{n_1}}{\sqrt{n}} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \widetilde{\Omega}_i + \frac{\sqrt{n_2}}{\sqrt{n}} \frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^n \widetilde{\Omega}_i,$$

which converges to a summand of two independent normal random variables with two weights, that is, $\sqrt{\xi} Z_1 + \sqrt{1-\xi} Z_2$, where $\xi = \lim_{n \rightarrow \infty} n_1/n$. So we know that $\frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{\Omega}_i$ converges to a normal distribution, $N(0, \Sigma)$, where $\Sigma = \text{cov}(\Omega_1)$. On the other hand,

$$\frac{1}{n} \sum_{i=1}^n \widetilde{\Omega}_i \widetilde{\Omega}_i^T = \frac{n_1}{n} \frac{1}{n_1} \sum_{i=1}^{n_1} \widetilde{\Omega}_i \widetilde{\Omega}_i^T + \frac{n_2}{n} \frac{1}{n_2} \sum_{i=n_1+1}^n \widetilde{\Omega}_i \widetilde{\Omega}_i^T \rightarrow \xi \Sigma + (1-\xi) \Sigma = \Sigma.$$

As a consequence, we conclude that $(n^{-1/2} \sum_{i=1}^n \widetilde{\Omega}_i)^T (n^{-1} \sum_{i=1}^n \widetilde{\Omega}_i \widetilde{\Omega}_i^T) (n^{-1/2} \sum_{i=1}^n \widetilde{\Omega}_i) \sim \chi_p^2$. Since $n^{-1/2} \sum_{i=1}^n \Omega_i$ and $n^{-1/2} \sum_{i=1}^n \widetilde{\Omega}_i$ have the same limiting normal distribution, and $n^{-1} \sum_{i=1}^n \Omega_i \Omega_i^T$ and $n^{-1} \sum_{i=1}^n \widetilde{\Omega}_i \widetilde{\Omega}_i^T$ have the same limiting value, thereby $-2 \log\{\mathcal{R}_n(\beta_1^*)\}$ converges to χ_p^2 . We complete the proof.

Figure 1.

Example 1: The power performances of the proposed test (solid line), Weerahandi test (dashed line) and the Dupont and Plummer test (dotted line) for the normal error (left panel), and for the non-normal error (right panel).

Figure 2.

Example 2: The power performances of the proposed test (solid line), the Weerahandi test (dashed line) and the Dupont and Plummer test (dotted line) for the normal error (left panel), and for the non-normal error (right panel).

Figure 3. Data from a drug study. Responses of treatment effect against dose level for four treatment agents: A (square), B(circle), C(triangle), D(diamond), and the associated four linear model fittings.

Table 1

Example 1: Type-I error of comparing two regression coefficients.

Case	β_1	CI	Type-I error				T	DUP
			EL	W				
1	-0.002	(-0.108, 0.104)	0.055	0.036			0.051	0.051
2	-0.005	(-0.190, 0.182)	0.057	0.034			0.058	0.058
3	-0.003	(-0.110, 0.104)	0.060	0.037			0.050	0.050
4	-0.003	(-0.200, 0.194)	0.058	0.043			0.056	0.056

note. EL: the proposed test (EL); W: the Weerahandi test; T: the t-test; DUP: the Dupont and Plummer test; CI: the Bartlett-corrected empirical likelihood confidence interval for β_1^* .

Table 2

Example 2: Type-I errors of comparing two regression coefficients.

Case	β_1	CI	Type-I error				DUP
			EL	W	T		
1	0.000	(-0.144, 0.148)	0.059	0.019	0.058		0.058
2	0.001	(-0.258, 0.259)	0.056	0.017	0.041		0.041
3	-0.002	(-0.150, 0.143)	0.073	0.021	0.062		0.062
4	-0.003	(-0.257, 0.256)	0.055	0.019	0.041		0.040

Table 3

Example 2: The empirical power of 0.05 level for alternative hypothesis ($\beta_1^* = \beta_1 - \beta_2 = 0.15$) with different sample sizes under four cases of the error structures.

Case	n_1	n_2	β_1^*	CI	power			
					EL	W	T	DUP
1	50	70	-0.148	(-0.272, -0.026)	0.623	0.312	0.498	0.500
	70	90	-0.146	(-0.252, -0.044)	0.755	0.550	0.745	0.745
	100	120	-0.151	(-0.249, -0.054)	0.840	0.624	0.794	0.794
	120	140	-0.148	(-0.238, -0.062)	0.897	0.749	0.866	0.867
2	50	70	-0.148	(-0.367, 0.067)	0.254	0.086	0.188	0.188
	70	90	-0.143	(-0.337, 0.036)	0.338	0.169	0.322	0.322
	100	120	-0.151	(-0.326, 0.017)	0.397	0.189	0.336	0.336
	120	140	-0.149	(-0.298, 0.001)	0.467	0.258	0.408	0.408
3	50	70	-0.152	(-0.280, -0.029)	0.661	0.321	0.519	0.520
	70	90	-0.149	(-0.253, -0.045)	0.809	0.578	0.759	0.759
	100	120	-0.150	(-0.247, -0.056)	0.854	0.618	0.771	0.770
	120	140	-0.152	(-0.240, -0.063)	0.926	0.758	0.886	0.886
4	50	70	-0.152	(-0.379, 0.066)	0.270	0.099	0.204	0.204
	70	90	-0.146	(-0.332, 0.040)	0.318	0.155	0.299	0.299
	100	120	-0.149	(-0.321, 0.017)	0.397	0.185	0.341	0.341
	120	140	-0.154	(-0.303, -0.001)	0.494	0.278	0.430	0.430

Table 4

Example 2: The empirical power of 0.05 level for alternative hypothesis ($\beta_1^* = \beta_1 - \beta_2 = 0.40$) when $n_1 = 25$, $n_2 = 35$ with various error variances.

Case	σ_1^2	σ_2^2	β_1^*	CI	power			
					EL	W	T	DUP
1	1	1	-0.400	(-0.549, -0.256)	0.997	0.983	0.995	0.995
		5	-0.405	(-0.676, -0.142)	0.781	0.597	0.746	0.746
2	5	5	-0.396	(-0.728, -0.073)	0.613	0.371	0.572	0.572
		10	-0.389	(-0.815, 0.001)	0.436	0.219	0.395	0.394
3	1	1	-0.401	(-0.548, -0.253)	0.999	0.991	0.997	0.997
		5	-0.404	(-0.684, -0.140)	0.820	0.584	0.761	0.761
4	5	5	-0.392	(-0.727, -0.069)	0.612	0.347	0.555	0.555
		10	-0.386	(-0.801, 0.010)	0.436	0.203	0.366	0.366

Table 5

Example 3: The proportion of rejecting H_0 in 1000 simulated samples using Bartlett-corrected empirical likelihood-based test (EL), normal approximation method (Norm), the ANOVA and Weerahandi (W) test for comparing **three group means**.

Case	$\mathcal{J}_1 - \mathcal{J}_2$	$\mathcal{J}_2 - \mathcal{J}_3$	Joint comparison			
			Norm	EL	ANOVA	W
1	0	0	0.051	0.049	0.050	0.049
	-0.3	0.3	0.288	0.285	0.281	0.285
	-0.4	-0.5	0.960	0.956	0.960	0.965
2	0	0	0.044	0.043	0.043	0.047
	-0.3	0.3	0.136	0.146	0.134	0.150
	-0.4	-0.5	0.707	0.764	0.703	0.774
3	0	0	0.059	0.054	0.058	0.057
	-0.3	0.3	0.277	0.272	0.273	0.263
	-0.4	-0.5	0.971	0.976	0.968	0.968
4	0	0	0.046	0.056	0.044	0.058
	-0.3	0.3	0.131	0.153	0.125	0.150
	-0.4	-0.5	0.726	0.803	0.722	0.786

Table 6

Results of comparing responses of treatment effect against dose level for four treatment agents using the proposed test (EL) and the other three tests: the t-test (T) and the Dupont and Plummer test (DUP), the Weerahandi test (W).

Rejection of H_0						
Test	A vs B	A vs C	A vs D	B vs C	B vs D	C vs D
EL	yes	no	yes	no	no	yes
DUP	no	no	no	no	no	no
T	no	no	no	no	no	no
W	no	no	yes	no	no	no

Table 7

The proportion of rejecting H_0 in 1000 bootstrap samples using Bartlett-corrected empirical likelihood-based test (EL), the t-test (T) and the Dupont and Plummer test (DUP) for the paired comparison between the four treatment agents.

Test	Comparison Group					
	A vs B	A vs C	A vs D	B vs C	B vs D	C vs D
EL	0.661	0.047	0.631	0.042	0.033	0.431
DUP	0.113	0.049	0.265	0.202	0.156	0.281
T	0.114	0.065	0.207	0.249	0.126	0.306