Holzmann, Min, Czado:

Validating linear restrictions in linear regression models with general error structure

Sonderforschungsbereich 386, Paper 478 (2006)

Online unter: http://epub.ub.uni-muenchen.de/

Validating linear restrictions in linear regression models with general error structure

Hajo Holzmann[1], Aleksey Min[2] and Claudia Czado[2]

[1]Institute for Mathematical Stochastics
Georg-August-University Göttingen
37073 Göttingen, Germany

and

[2]Center for Mathematical Sciences
Munich University of Technology
Boltzmannstr. 3
85747 Garching, Germany

Abstract

A new method for testing linear restrictions in linear regression models is suggested. It allows to validate the linear restriction, up to a specified approximation error and with a specified error probability. The test relies on asymptotic normality of the test statistic, and therefore normality of the errors in the regression model is not required. In a simulation study the performance of the suggested method for model selection purposes, as compared to standard model selection criteria and the t-test, is examined. As an illustration we analyze the US college spending data from 1994.

Keywords asymptotic normality, linear regression, model selection, model validation

Mathematics Subject Classification 62J05, 62J20

[1] corresponding author, Institut für Mathematische Stochastik, Universität Göttingen, Maschmühlenweg 8–10, 37073 Göttingen, Germany, Fon: +49/551/3913516, Fax: +49/551/39-13505
E-mail address: holzmann@math.uni-goettingen.de

1 Introduction

The choice of the relevant covariates in a linear regression model is an important and much studied problem. For this purpose, various methods have been suggested in the literature. One approach is via model selection criteria. Here one chooses the sub-model which minimizes a certain criterion function, e.g. the AIC (Akaike, 1974) or the BIC (Schwarz, 1978). Another approach is to specify the sub-model by testing the relevant linear restrictions. Toro-Vizcarrondo and Wallace (1968), see also Wallace (1972), observed that the sub-model may be superior to the complete model in terms of mean square error (MSE) even if the sub-model is incorrect. Therefore they suggested to test in which model the least squares estimator has smaller MSE. In this paper we suggest a related test which focuses on validating the sub-model. More precisely, the test allows to validate the sub-model up to a certain specified approximation error, and with a specified error probability. The test is based on asymptotic normality of the test statistic and therefore does not require normality of the errors in the regression model.

This paper is organized as follows. In Section 2 we introduce the model and the testing problem. Section 3 presents the test statistics and its asymptotic distribution. Further we discuss how to perform the test. In Section 4 we investigate the performance of our method, as compared to the *t*-test and some model selection criteria in a simulation study. Finally, in Section 5, we illustrate the practical usefulness of our method by analyzing the US college spending data from 1994.

2 Testing problem

Consider the homoscedastic linear regression model

$$Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $Y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n\times(p+q)}$ is the design matrix, which is assumed to be non-random, and $\boldsymbol{\beta} \in \mathbb{R}^{p+q}$ denotes the unknown regression parameter vector of interest. The errors $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$ are assumed to be independent identically distributed (i.i.d.) random variables with $E(\epsilon_1) = 0$ and $Var(\epsilon_1) = \sigma^2$.

Suppose that we want to check the validity of the sub-model

$$Y = X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}, \tag{2}$$

where $X = [X_1, X_2]$ and $X_1 \in \mathbb{R}^{n\times p}$, $X_2 \in \mathbb{R}^{n\times q}$, and $\boldsymbol{\beta}^t = [\boldsymbol{\beta}_1^t, \boldsymbol{\beta}_2^t]$, where $\boldsymbol{\beta}_1 \in \mathbb{R}^p$, $\boldsymbol{\beta}_2 \in \mathbb{R}^q$. Classically one verifies model (2) by testing the hypothesis

$$H_0 \ : \ \boldsymbol{\beta}_2 = 0.$$

Let $\hat{\boldsymbol{\beta}}$ denote the least squares (LS) estimator in the full model (1), and let $\hat{\boldsymbol{\beta}}_r$ be the restricted LS estimator in the submodel (2), which we also consider as a $(p+q)$-dimensional vector by filling the last q entries by 0. Suppose for the moment that in addition the errors are normally distributed, and let $SSE(b)$ denote the error sum of squares of an estimator b of $\boldsymbol{\beta}$. A popular statistic for testing H_0 is via the F-statistic

$$T = \frac{SSE(\hat{\boldsymbol{\beta}}_r) - SSE(\hat{\boldsymbol{\beta}})}{q\hat{\sigma}^2}, \tag{3}$$

where $\hat{\sigma}^2$ is the LS estimator of σ^2 in the full model (1). Toro–Vizcarrondo and Wallace (1968) show that T has a F distribution with degrees of freedom q and $(n-(p+q))$ and non-centrality parameter (in the notation of Kotz and Johnsson, 1970),

$$\lambda = n\,\frac{d_n(\boldsymbol{\beta}_2)}{\sigma^2}, \quad d_n(\boldsymbol{\beta}_2) = \frac{1}{n}\boldsymbol{\beta}_2^t X_2^t M_{X_1} X_2 \boldsymbol{\beta}_2,$$

where $M_{X_1} = I_n - P_{X_1}$, $P_{X_1} = X_1(X_1^t X_1)^{-1}X_1^t$ and I_n is the identity matrix of dimension n. Thus, under H_0, T is central F distributed with q and $(n-(p+q))$ degrees of freedom.

For many purposes it is not adequate to base a decision for or against the sub-model (2) on testing the hypothesis H_0. For example, Toro–Vizcarrondo and Wallace (1968) pointed out that the estimator $\hat{\boldsymbol{\beta}}_r$ can have a smaller MSE (mean square error) than $\hat{\boldsymbol{\beta}}$, even if the model (2) is incorrect. Therefore they suggested to test the hypothesis

$$H_{MSE} \ : MSE(\hat{\boldsymbol{\beta}}_r) \leq MSE(\hat{\boldsymbol{\beta}}),$$

where $MSE(b) = E(b-\boldsymbol{\beta})(b-\boldsymbol{\beta})^t$, and $MSE(\hat{\boldsymbol{\beta}}_r) \leq MSE(\hat{\boldsymbol{\beta}})$ means that $MSE(\hat{\boldsymbol{\beta}}) - MSE(\hat{\boldsymbol{\beta}}_r)$ is positive semidefinite. Toro–Vizcarrondo and Wallace (1968) showed that the hypothesis H_{MSE} is equivalent to $\lambda \leq 1$, which they used to construct a uniformly most powerful test for H_{MSE} based on T. Hypotheses related to H_{MSE} were investigated by Wallace (1972) and by Yancey et al. (1973).

The hypothesis H_{MSE} still has some drawbacks. Instead of comparing models, it compares the performance of certain estimators. This is a somewhat arbitrary choice since there are other estimators (e.g. the ridge estimator, cf. Farbrother, 1975), which have smaller MSE than the LS estimator. Further, and more importantly, even if the hypothesis H_{MSE} (or H_0) cannot be rejected with a large p-value, this does not imply that the sub-model (2) is actually true. Therefore, we suggest to test a hypothesis which focuses on validating the sub-model (2). A related approach to validating parametric functional forms of regression models (against nonparametric alternatives) was suggested by Dette and Munk (1998).

To this end, note that $d_n(\boldsymbol{\beta}_2)$ is the normalized length (with factor n^{-1}) of the n vector $X_2\boldsymbol{\beta}_2$, when projected onto the orthogonal complement of the space spanned by the columns of X_1. Thus it provides a natural measure of distance between the restricted model (2) and the full model (1), and we propose to validate sub-model (2) by testing the hypothesis that

$$H_{\Delta,n} \ : \ d_n(\boldsymbol{\beta}_2) > \Delta \quad \text{against } K_{\Delta,n} \ : \ d_n(\boldsymbol{\beta}_2) \leq \Delta,$$

for some $\Delta > 0$. Under normality we have that $H_{\Delta,n}$ is equivalent to $H_{\lambda,n} : \lambda > n\Delta/\sigma^2$. Since σ^2 is unknown we cannot construct even under normality an exact test of $H_{\Delta,n}$. Therefore we give a condition under which $d_n(\boldsymbol{\beta}_2)$ converges as $n \to \infty$, say to $d(\boldsymbol{\beta}_2)$, and consider testing $H_\Delta \ : \ d(\boldsymbol{\beta}_2) > \Delta$ against $K_\Delta \ : \ d(\boldsymbol{\beta}_2) \leq \Delta$. For this testing problem we will construct an asymptotic test which does not require normality of the errors.

3 An asymptotic test

In order to formulate an asymptotic version of the hypotheses $H_{\Delta,n}$, we need the following assumption.

Assumption 1. The regressors X are non-random and we have $X^tX/n \to G$ as $n \to \infty$, where $G \in \mathbb{R}^{(p+q)\times(p+q)}$ is a symmetric positive definite matrix.

Split G into blocks as follows

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}$$

Then the asymptotic version of the distance $d_n(\boldsymbol{\beta})$ is defined as

$$d(\boldsymbol{\beta}_2) = \boldsymbol{\beta}_2^t(G_{22} - G_{21}G_{11}^{-1}G_{12})\boldsymbol{\beta}_2,$$

and the corresponding version of $H_{\Delta,n}$ as

$$H_\Delta \ : \ d(\boldsymbol{\beta}_2) > \Delta \quad \text{against } K_\Delta \ : \ d(\boldsymbol{\beta}_2) \leq \Delta.$$

In fact, under assumption 1 one can show that $d_n(\boldsymbol{\beta}_2) \to d(\boldsymbol{\beta}_2)$ as $n \to \infty$. Note that the matrix $G_{22} - G_{21}G_{11}^{-1}G_{12}$, used in the definition of $d(\boldsymbol{\beta}_2)$, is the Schur complement of the block matrix G_{11} and is positive definite since G is assumed to be positive definite.
Let $P_X = X(X^tX)^{-1}X^t$ and consider the test statistic

$$R_n = \frac{1}{n}\big(SSE(\hat{\boldsymbol{\beta}}_r) - SSE(\hat{\boldsymbol{\beta}})\big) = \frac{1}{n}Y^t\big(P_X - P_{X_1}\big)Y,$$

which estimates $d_n(\boldsymbol{\beta}_2)$.

Theorem 1. *Under assumptions 1 - 4 (cf. the appendix), if $d(\boldsymbol{\beta}_2) > 0$ we have that*

$$\sqrt{n}\big(R_n - d(\boldsymbol{\beta}_2)\big) \xrightarrow{\mathcal{L}} N\big(0, 4\sigma^2 d(\boldsymbol{\beta}_2)\big) \quad \text{as} \quad n \to \infty.$$

The proof of theorem 1 is given in the appendix. Using theorem 1, we construct an asymptotic test for H_Δ as follows. Given $\Delta > 0$, reject H_Δ with level $\alpha > 0$ if

$$\sqrt{n}\,\frac{R_n - \Delta}{2\hat{\sigma}\sqrt{\Delta}} \leq u_\alpha, \tag{4}$$

where u_α denotes the α-quantile of the standard normal distribution. Thus, the choice of Δ is evidently critical for the test decision. Note that for a given level α (e.g. $\alpha = 0.05$), one can determine the threshold $\Delta_{\text{crit},\alpha}$ for which $H_{\Delta_{\text{crit},\alpha}}$ can be rejected at level α, while H_Δ cannot be rejected for $\Delta < \Delta_{\text{crit},\alpha}$:

$$\Delta_{\text{crit},\alpha} = \Big(\big(R_n + \hat{\sigma}^2u_\alpha^2/n\big)^{1/2} - \hat{\sigma}u_\alpha/\sqrt{n}\Big)^2.$$

Now Δ is a threshold for $d(\boldsymbol{\beta}_2)$, the limit of the distance $d_n(\boldsymbol{\beta}_2)$, which as mentioned above measures the normalized (with factor n^{-1}) distance of the projected vector $X_2\boldsymbol{\beta}_2$. Therefore, we suggest to normalize $\Delta_{\text{crit},\alpha}$ by an estimate of the total normalized length $\boldsymbol{\beta}^t X^t X\boldsymbol{\beta}/n$:

$$D_{\alpha,n} = \frac{\Delta_{\text{crit},\alpha}}{\hat{\boldsymbol{\beta}}^t X^t X\hat{\boldsymbol{\beta}}/n}.$$

The quantity $D_{\alpha,n}$ can be nicely interpreted as the estimated maximal relative error one makes (with level α) if one uses sub-model (2) instead of the full model (1). In fact, one has $D_{\alpha,n} \to d(\boldsymbol{\beta}_2)/(\boldsymbol{\beta}^t G\boldsymbol{\beta})$ in probability as $n \to \infty$. Model validation now proceeds in terms of $D_{\alpha,n}$: If $D_{\alpha,n}$ is less than some fixed value which we allow as maximal relative error (say 0.1 or 0.05), then we use the smaller sub-model.

4 Simulation study

In this section we conduct a small simulation study in which we investigate the performance of our method for model selection as compared to the AIC, the BIC and the t-test. Here, for the computation of the AIC and the BIC we use the residual sum of squares (with appropriate penalty term), in spite of the fact that for non-normally distributed errors, it is not the maximized log-likelihood function. This is because we do not want to assume a specific distributional structure of the errors to be known in advance.

We use a linear regression model with 7 covariates and the intercept, where the covariates are drawn uniformly from $[-1, 1]$. The vector of true regressions coefficients is chosen as

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)^t = (2, 2, 0.1, 0.1, 0.1, 2, 0.1, 2)^t.$$

Evidently, the relevant covariates that we want to identify are the 1st, 5th and 7th covariate and the intercept (which corresponds to $\boldsymbol{\beta}_0$ and in the following is assumed to be contained in all submodels).

The distinct methods are applied in a backward selection procedure. More specifically, consider the method suggested in section 3. In the first step, we compute $D_{\alpha,n}$ with $\alpha = 0.05$ for all submodels of the full model with 6 covariates and the intercept. Let M_1 be the submodel with minimal $D_{\alpha,n}$, denoted $D^1_{\alpha,n}$. If $D^1_{\alpha,n}$ is smaller than some threshold, which we take as 0.05, then we continue with model M_1, otherwise we select the full model. In the next step, consider all submodels of M_1 with 5 covariates and the intercept, and compute $D_{\alpha,n}$ for all these models, relative to M_1 (i.e. the denominator is computed in model M_1). Let M_2 denote the submodel with minimal $D_{\alpha,n}$, denoted $D^2_{\alpha,n}$. If $D^2_{\alpha,n}$ is smaller than 0.05, then we continue with model M_2, otherwise we select M_1. We proceed in this way until a model is selected or all covariates are discarded (and only the intercept remains). The other methods are applied in a similar fashion. For the information criteria, we iteratively discard covariates as long as the AIC and the BIC decreases in a submodel, and continue with the submodel with the smallest AIC or BIC. Finally, for the t-test, in the first step for

Table 1: Results of a single backward selection procedure for $n = 100$

step i	submodel	discarded cov.	$D^i_{\alpha,n}$	BIC	AIC	p_i of t-test
1	$x_0, x_1, x_2, x_4, x_5, x_6, x_7$	x_3	0.027	363.02	395.05	0.736
2	$x_0, x_1, x_2, x_5, x_6, x_7$	x_4	0.030	359.26	338.98	0.377
3	x_0, x_1, x_5, x_6, x_7	x_2	0.030	355.57	339.94	0.354
4	x_0, x_1, x_5, x_7	x_6	0.031	352.01	341.02	0.320
5	x_0, x_1, x_7	x_5	0.264	405.47	342.18	0.000

each submodel with 6 covariates and the intercept we compute the p-value for the t-test that the coefficient β_i of the missing covariate is zero. Let M_1 be the submodel for which the corresponding t-test has maximal p-value p_1. If $p_1 > 0.05$, we continue with model M_1, otherwise we choose the full model. In the next step for each submodel of M_1 with 5 covariates and the intercept we compute the p-value for the t-test that the coefficient β_j of the covariate missing from M_1 is zero. If M_2 denotes the submodel for which the corresponding t-test has maximal p-value p_2, we continue with M_2 if $p_2 > 0.05$, otherwise we choose M_1. We refer to Miller (2002) for other selection methods than backward selection. For example, one may modify our method in order to construct a forward selection procedure by considering K_Δ as the null hypothesis and H_Δ as the alternative.

The simulation is conducted as follows. After drawing the covariates once, these remain fixed subsequently, and we generate responses on model (1) for 1000 iterations, and for sample sizes $n = 100$ and $n = 200$. In each case, we apply the backward selection procedures described above.

Further, we use two kinds of error distributions, namely a t distribution with 6 degrees of freedom and an exponential distribution. For each distribution we consider two distinct scaling parameters. For the t distribution, we use scaling factors of $\tau = 1$ and of $\tau = \sqrt{1.33}$, which gives for the error variance 1.5 for $\tau^2 = 1$, and 1.995 for $\tau^2 = 1.33$, respectively. For the exponential distribution, we use $\lambda = 1$ and $\lambda = 1/\sqrt{2}$, giving variances of 1 ($\lambda = 1$) and 2 ($\lambda = 1/\sqrt{2}$). Further, we center the errors by their expectation. For the scaled t distribution with $\tau = 1$ ($\tau = \sqrt{1.33}$) we observe that 50% of the regression data have a signal to noise ratio (mean divided by standard error) larger than 1.65 (1.44) . For exponentially distributed errors, the signal to noise ratio for 50% of the regression data with $\lambda = 1$ ($\lambda = 1/\sqrt{2}$) is larger than 2.17 (1.54).

Table 1 shows the results for one simulation with t distributed errors (with $\tau = 1$) and $n = 100$. Since all methods depend monotonically on the statistic $SSE(\hat{\boldsymbol{\beta}}_r) - SSE(\hat{\boldsymbol{\beta}})$, they proceed in the same steps. The desired model appears in step 4, which is selected by all methods except for the AIC (which includes too many covariates). Tables 2 and 3 show, for scaled t distributed and exponentially distributed errors, respectively, how often among 1000 iterations the desired model was selected. Here different rows correspond to different random covariates, whereas within the rows these covariates are fixed.

For $n = 100$ and $\tau = 1$ or $\lambda = 1$ (yielding higher signal to noise ratios), the $D_{\alpha,n}$ method selects the desired model in more than 90% of the simulations, and for $n = 200$

Table 2: Number of iterations in which the desired submodel consisting of x_0, x_1, x_5, x_7 is selected; errors are scaled t distributed with 6 df. For $D_{\alpha,n}$, we choose $\alpha = 0.05$ and the threshold value also equal to 0.05.

sample size	scenario	$D_{\alpha,n}$	BIC	AIC	t test	τ^2
$n = 100$	1	981	793	382	739	1
		910	808	406	760	1.33
	2	952	769	380	713	1
		856	791	405	731	1.33
	3	917	778	381	722	1
		823	767	353	693	1.33
	4	968	731	326	664	1
		838	797	389	747	1.33
	5	962	774	363	716	1
		836	789	396	744	1.33
$n = 200$	1	1000	810	320	649	1
		1000	835	349	691	1.33
	2	1000	781	310	635	1
		1000	807	346	675	1.33
	3	1000	819	320	674	1
		1000	846	362	703	1.33
	4	999	801	309	638	1
		994	826	340	676	1.33
	5	1000	810	308	658	1
		998	833	342	695	1.33

Table 3: Number of iterations in which the desired submodel consisting of x_0, x_1, x_5, x_7 is selected; errors are centered exponentially distributed with $\lambda = 1$ and $\lambda = 1/\sqrt{2}$. For $D_{\alpha,n}$, we choose $\alpha = 0.05$ and the threshold value also equal to 0.05.

sample size	scenario	$D_{\alpha,n}$	BIC	AIC	t test	λ
$n = 100$	1	998	769	362	698	1
		850	831	423	753	$1/\sqrt{2}$
	2	994	756	337	696	1
		773	807	383	754	$1/\sqrt{2}$
	3	996	764	352	716	1
		803	808	413	753	$1/\sqrt{2}$
	4	996	712	313	647	1
		820	780	373	724	$1/\sqrt{2}$
	5	999	777	351	714	1
		916	825	410	769	$1/\sqrt{2}$
$n = 200$	1	1000	817	330	658	1
		999	864	394	741	$1/\sqrt{2}$
	2	1000	754	259	607	1
		1000	834	364	704	$1/\sqrt{2}$
	3	1000	798	303	645	1
		999	861	390	738	$1/\sqrt{2}$
	4	1000	770	263	608	1
		1000	853	357	706	$1/\sqrt{2}$
	5	1000	747	246	574	1
		1000	843	340	704	$1/\sqrt{2}$

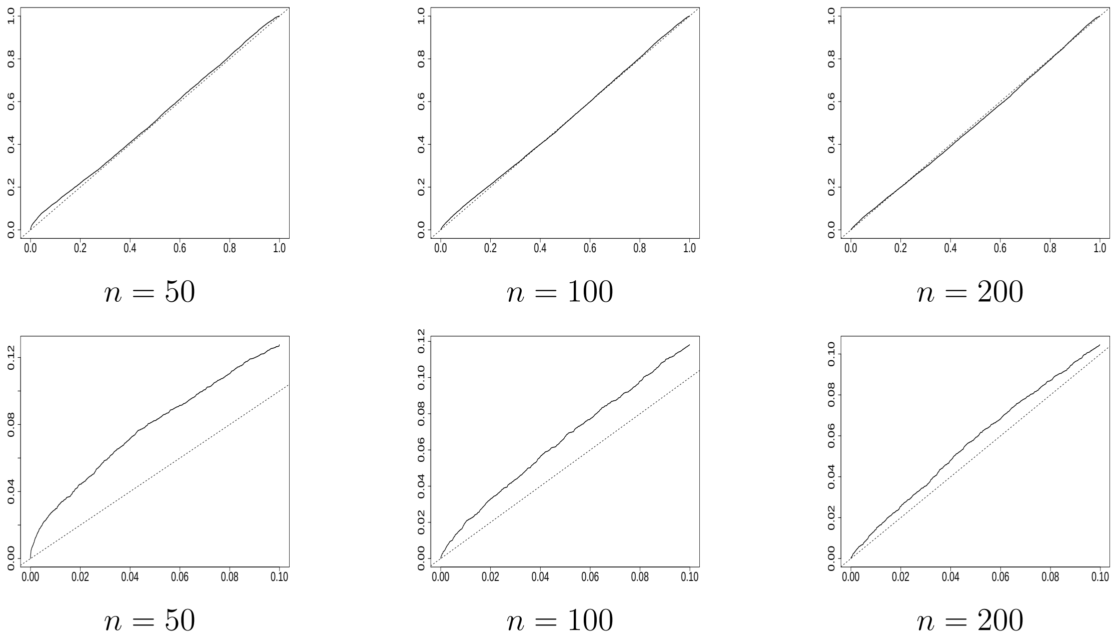
it does so almost always. In contrast, the BIC, the AIC and the t-test more often select larger models. This is mainly due to the thresholding used for the $D_{\alpha,n}$. Only if the $D_{\alpha,n}$ increases significantly (namely becomes larger than 0.05) we stop the model selection procedure. Observing Table 1, for the first four covariates the values of AIC and BIC change little, although they might increase slightly, which leads to the choice of a larger model. A huge increase only occurs if the 5th covariate is discarded. Therefore, if one used a threshold (say 350 for the BIC), one would get similarly precise results as for the $D_{\alpha,n}$ method. However, for the $D_{\alpha,n}$ method such a threshold has a natural interpretation as maximal relative error, whereas there is no such interpretation for the values of the BIC and the AIC. The t-test also uses a threshold, i.e. for the p value. If we chose it much smaller (e.g. 0.005) we would also recover the relevant model almost always. However, such a high precision is unnatural for a sample size $n = 100$ or $n = 200$. Furthermore, if we do not reject with a p-value of 0.04, this does not say anything about how good the smaller model still is.
Finally, we investigate the quality of the normal approximation in Theorem 1. We have to consider a testing situation where the hypothesis $H_{\Delta=0.05}$ is true, and where the complete model is included under this hypothesis. Therefore, we test the complete model against the model where the covariate x_7 is excluded, and simulate the statistic R_n 10000 times for sample sizes $n = 30, 50$ and 100 and centered exponentially distributed errors. For visualization in Figure 1 we use P-P plots, which show for each $\alpha \in [0, 1]$ the empirical probability of the event $\{\sqrt{n}[R_n - d_n(\boldsymbol{\beta}_2)] \leq Q_\alpha\}$, where Q_α is the α-quantile of the asymptotic normal distribution with consistently estimated variance. From the top row of Figure 1 we see that the asymptotic approximation is quite good already for rather small sample sizes. Note that for the test decision (4), the approximation for small α's is relevant, which can be assessed using the bottom row.

Summarizing the results of the simulation study we see that the performance of our method depends to some extend on the signal to noise ratio, especially for small sample sizes. In such cases ($n = 100$), it performs well for signal to noise ratios larger than 1.5. For large n, the dependence on the signal to noise ratio becomes weaker.

5 College spending data

To illustrate our method in a practical application we analyze the college spending data from U.S. News and World Report 1994 College Guide. The complete data can be found in Dielman (1996) and its short description is given in Table 4. The variable of interest is educational spending per full-time equivalent (SPEND) given for 147 US colleges. A simple explorative data analysis shows that there is a presence of variance heterogeneity and a log transformation of the response SPEND is needed. Further, for numerical stability, all variables including the response SPEND are centered and normalized by their sample mean and sample standard deviation. In Table 5, the results of a backward selection procedure for the $D_{\alpha,n}$ method, the BIC, AIC and the t-test, applied to the college spending data, are given. Here, we always keep the intercept in the submodels. Further, for the $D_{\alpha,n}$ we

Figure 1: P-P plots for $\sqrt{n}[R_n - d_n(\boldsymbol{\beta}_2)]$ based on 10000 replications (top row $\alpha \in (0,1)$, bottom row $\alpha \in (0, 0.1)$)



$n = 50$ $n = 100$ $n = 200$

$n = 50$ $n = 100$ $n = 200$

use a level of $\alpha = 0.05$ and a threshold of 0.05.

The BIC, the t-test and the $D_{\alpha,n}$ method choose a submodel consisting of the three covariables SAT, TOP10 and RATIO, and only the AIC prefers a model with 4 covariates. This is in agreement with the simulation results in Section 4. Let us stress that in contrast to the BIC and the t-test (with a p-value of 0.89 in the final step), the $D_{\alpha,n}$-method allows for a clear interpretation of the quality of the resulting submodel, namely that the maximal relative error we make when using this smaller submodel is less than 0.05, with probability 0.95.

6 Conclusion

In this paper we introduced a new method for testing linear restrictions in linear regression models. It allows to test the validity of the linear restriction, up to a specified approximation error and with a specified error probability. The method can also be used to estimate a quantity $D_{\alpha,n}$, which can be interpreted as the estimated maximal relative error (with level α) that one makes when using the smaller submodel. This quantity $D_{\alpha,n}$ can be conveniently used for model-selection purposes. In contrast to classical model selection criteria such as the AIC and the BIC, the value $D_{\alpha,n}$ has a clear interpretation (as maximal relative error), and therefore allows for model selection strategies based on a threshold value for $D_{\alpha,n}$. As illustrated in a simulation study as well as a real data example, this might lead to good results in the model selection process.

Table 4: Variables of college spending data in USA from 1994

Notation	**Short description**
SAT	average SAT score
TOP10	freshmen in the top 10% of their high school class (in percentage)
ACCRATE	acceptance rate (in percentage)
PHD	faculty with PhD (in percentage)
RATIO	student faculty ratio
SPEND	educational spending per full-time equivalent student (in dollars)
GRADRATE	graduation rate (in percentage)
ALUMNI	alumni giving rate (in percentage)

Table 5: Results of a backward selection procedure for college spending data

step i	submodel	discarded cov.	$D^i_{\alpha,n}$	BIC	AIC	p_i of t-test
1	SAT, TOP10, ACCRATE, PHD, RATIO, GRADRATE	ALUMNI	0.025	253.7	229.8	0.789
2	SAT, TOP10, ACCRATE, PHD, RATIO	GRADRATE	0.026	248.8	227.9	0.808
3	SAT, TOP10, PHD,RATIO	ACCRATE	0.037	245.5	227.5	0.211
4	SAT, TOP10, RATIO	PHD	0.042	243.5	228.5	0.089
5	TOP10, RATIO	SAT	0.110	247.0	235.1	0.004

7 References

Akaike, H. (1974), A new look at the statistical model identification. System identification and time-series analysis. *IEEE Trans. Automatic Control* **19**, 716–723.

Dielman T. E. (1996), *Applied regression analysis for business and economics.* 2nd edition. Duxbury Press, Belmont.

Dette, H. and Munk, A. (1998), Validation of linear regression models. *Ann. Statist.* **26**, 778–800.

Farebrother, R. W. (1975), The minimum mean square error linear estimator and ridge regression. *Technometrics* **17**, 127–128.

Johnson, N. L. and Kotz, S. (1970), *Distributions in statistics. Continuous univariate distributions - 2.* Houghton Mifflin Co., Boston, Mass.

Miller, A. (2002), *Subset selection in regression.* 2nd edn. Chapman & Hall/CRC, Florida.

Schwarz, G. (1978), Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Seber, G. A. F. and Lee, A. J. (2003), *Linear Regression Analysis.* 2nd edt., John Wiley & Sons, Hoboken, New Jersey.

Theil, H. (1971), *Principles of Econometrics.* John Wiley & Sons, New York.

Toro-Vizcarrondo, C. and Wallace, T. D. (1968), A test of the mean square error criterion for restrictions in linear regression. *J. Amer. Statist. Assoc.* **63**, 558–572.

Wallace, T. D. (1972), Weaker criteria and tests for linear restrictions in regression. *Econometrica* **40**, 689–698.

Yancey, T. A., Judge, G. G. and Bock, M. E. (1973) Wallace's weak mean square error criterion for testing linear restrictions in regression: a tighter bound. *Econometrica* **41**, 1203–1206.

Appendix

Assumption 2. The errors $\epsilon_1, \dots \epsilon_n$ are i.i.d. with $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ and $E(\epsilon_1^4) < \infty$.

Assumption 3. We have that

$$\sqrt{n}\big(n^{-1}X^tX - G\big) \to 0. \tag{5}$$

Assumption 4. The entries of the covariate matrix X_2 lie in a compact set $K \subset \mathbb{R}$ for all n.

Note that from Assumptions 3 and 4 it follows that

$$\sqrt{n}\Big[\big(\frac{1}{n}X^tX\big)^{-1} - G^{-1}\Big] \to 0 \tag{6}$$

since taking the inverse of a matrix is a Lipschitz continuous mapping on compact sets.

Proof of Theorem 1. First note that from (5) and (6) it follows that

$$\sqrt{n}\big(d_n(\boldsymbol{\beta}_2) - d(\boldsymbol{\beta}_2)\big) \to 0.$$

Since by assumption, $d(\boldsymbol{\beta}_2) > 0$, $d_n(\boldsymbol{\beta}_2)$ will be bounded away from 0 and we get

$$\sqrt{n}\,\frac{R_n - d(\boldsymbol{\beta}_2)}{2\sigma\sqrt{d_n(\boldsymbol{\beta}_2)}} = \sqrt{n}\,\frac{\frac{1}{n}Y^t(P_X - P_{X_1})Y - d_n(\boldsymbol{\beta}_2)}{2\sigma\sqrt{d_n(\boldsymbol{\beta}_2)}} + o(1) \tag{7}$$

¿From Theil (1973), p. 146,

$$P_X - P_{X_1} = M_{X_1}X_2(X_2^t M_{X_1} X_2)^{-1} X_2 M_{X_1} =: Q,$$

where $M_{X_1} = I_n - P_{X_1}$. The matrix Q is symmetric and idempotent and satisfies $QX_1 = 0$. Therefore

$$\begin{aligned} \frac{1}{n}Y^t(P_X - P_{X_1})Y &= \frac{1}{n}\epsilon^t Q\epsilon + \frac{2}{n}\boldsymbol{\beta}_2^t X_2^t Q\epsilon + \frac{1}{n}\boldsymbol{\beta}_2^t X_2^t Q X_2 \boldsymbol{\beta}_2 \\ &= S_1 + S_2 + d_n(\boldsymbol{\beta}_2). \end{aligned} \tag{8}$$

Now $ES_1 = tr\ Q/n \leq q/n$, and from Seber and Lee (2003, Theorem 1.6),

$$Var\,(S_1) = \frac{1}{n^2}\left[(\mu_4 - 3\sigma^4)h^t h + 2\sigma^4 tr(Q)\right],$$

where $\mu_4 = E(\epsilon_1^4)$ and h is the vector of diagonal elements of the matrix Q, for which $h^t h \leq q^2$. Thus

$$S_1 = O_P(|ES_1| + |S_1 - ES_1|) = O_P(n^{-1}).$$

Furthermore, $ES_2 = 0$ and

$$Var\,(S_2) = \frac{4}{n}\cdot\sigma^2 d_n(\boldsymbol{\beta}_2) \sim \frac{4}{n}\cdot\sigma^2 d(\boldsymbol{\beta}_2),$$

and therefore the term S_2 dominates the asymptotics in (8). It remains to show asymptotic normality of S_2. To this end we check the Lyapounov condition

$$\frac{1}{n^{3/2}}\sum_{i=1}^n E\,|b_i\varepsilon_i|^3 = \frac{E|\varepsilon_1|^3}{n^{3/2}}\sum_{i=1}^n |b_i|^3 \to 0 \quad \text{as} \quad n \to \infty,$$

where $b := 2\boldsymbol{\beta}_2^t X_2^t Q = (b_1, \ldots, b_n)$. It will be enough to show that the entries b_i are uniformly bounded. To this end, from assumption 4,

$$\begin{aligned} \max_{i=1,\ldots,n}|b_i| &= \max_{i=1,\ldots,n}|[QX_2\boldsymbol{\beta}_2]_i| \\ &\leq \max_{i=1,\ldots,n}\left\{\sum_{k=1}^n |[Q]_{ik}|\cdot|[X_2\boldsymbol{\beta}_2]_k|\right\} \\ &\leq C\max_{i=1,\ldots,n}\left\{\sum_{k=1}^n |[Q]_{ik}|\right\}, \end{aligned}$$

where $C > 0$ and $[\ \cdot\ ]_{ik}$ denotes the (i,k)-th entry of the corresponding matrix. Since Q is symmetric and positive semi-definite, $|[Q]_{ik}| \leq (Q_{ii} + Q_{kk})/2$, and thus

$$\begin{aligned} \max_{i=1,\ldots,n} |b_i| &\leq C \max_{i=1,\ldots,n} \left\{ \frac{1}{2} \sum_{k=1}^{n} [Q]_{ii} + [Q]_{kk} \right\} \\ &= C \max_{i=1,\ldots,n} \left\{ \frac{1}{2}[Q]_{ii} + \frac{1}{2} tr(Q) \right\} \\ &\leq C \cdot q. \end{aligned}$$

This finishes the proof of theorem 1. □