

# Variable selection for heteroscedastic data through variance estimation

Songjoon Baek<sup>1</sup>, Filiz Karaman<sup>2</sup> and Hongshik Ahn<sup>1\*</sup>

<sup>1</sup>Department of Applied Mathematics and Statistics  
Stony Brook University  
Stony Brook, NY 11794-3600

\*Corresponding author, *email*: hahn@ams.stonybrook.edu

<sup>2</sup>Department of Economics, Yeditepe University  
81120 Kadikoy-Istanbul, Turkey

## Abstract

In this paper, we extend some variable selection criteria in regression analysis to heteroscedastic models. First, a sequential test procedure is proposed to identify potential heteroscedasticity of the error variances. Next, we develop a variance estimation method to estimate the variance-covariance matrix for data with unequal variances. We improve Mallows's  $C_p$  and AIC using the proposed variance estimation method. This work is motivated by the poor behavior of  $C_p$  in highly heteroscedastic models and by the fact that  $C_p$  can be written as a linear function of an  $F$  statistic for testing the fit of a regression model. The proposed method performs well for both homoscedastic and heteroscedastic data. Simulation results show that our method is superior to  $C_p$  for data with significant heteroscedasticity and is comparable in accuracy for homoscedastic models. The new method is illustrated with real data.

KEY WORDS: Akaike information criterion, Experimental design, Homoscedasticity, Mallows'  $C_p$ , Regression, Variance estimation

## 1 Introduction

A number of methods have been developed for selecting the “best” or at least a “good” subset of variables in regression analysis. These methods include stepwise regression and all possible subsets regression. Stepwise regression is faster, but because the variables are deleted or added in a sequential fashion, the selected subset may not be the best possible. All possible subsets

regression does not have this weakness, but it requires a selection criterion such as Mallows' (1964)  $C_p$  statistic, Akaike's (1974) information criterion (AIC),  $R^2$ , adjusted  $R^2$ , or residual mean square.

Mallows'  $C_p$  can be expressed as a function of the residual mean square. Gorman and Toman (1966), Daniel and Wood (1971), Mallows (1973), Draper and Smith (1998) and Rawlings et al. (1998) discussed the statistic and its derivation, and studied several examples of its use. Kennard (1971) showed a one-to-one correspondence between  $C_p$  and adjusted  $R^2$ .  $C_p$  is also related to the  $R^2$  statistic. Hocking (1976) observed that  $C_p$  for a subset of the variables is linearly related to the  $F$ -ratio for testing significance of the coefficients of the rest of the variables in the full model. Akaike (1974) introduced AIC as an extension of the maximum likelihood principle for the problem of selecting the best model among models with different numbers of parameters. Because it is likelihood-based, AIC requires the user to make an assumption about the error distribution. It can be applied, however, to heteroscedastic models as well as homoscedastic models. When a heteroscedastic model is assumed, the performance of model selection partly depends on the estimation of the variance-covariance matrix.

On the other hand,  $C_p$  assumes a homoscedastic model. Therefore it is not expected to perform well on heteroscedastic data. When the error distribution is assumed to be normal, AIC is equivalent to  $C_p$  (Amemiya, 1980). In the present paper, a generalized  $C_p$  (denoted by  $GC_p$ ) is introduced for heteroscedastic data.  $GC_p$  is equivalent to AIC when a normal error distribution is assumed and the same variance estimates are used. We show by means of a simulation study that  $GC_p$  selects the correct model more often than  $C_p$  for data with significant heteroscedasticity. For homoscedastic models, however,  $C_p$  is still more accurate than  $GC_p$ . To overcome this problem, an adaptive procedure is developed that tests first for homoscedasticity. If the data are determined to be homoscedastic,  $C_p$  is used for subset selection; otherwise,  $GC_p$  is used. This adaptive procedure is denoted by  $AC_p$ . Because  $AC_p$  employs  $C_p$  on homoscedastic data and  $GC_p$  on heteroscedastic data, it tends to be more accurate than the application of  $C_p$  or  $GC_p$  alone. It can also be considered an improved AIC with a normal likelihood for heteroscedastic data.

The rest of the article is organized as follows. Section 2 reviews the selection criteria and the relationships among them. Section 3 develops the  $GC_p$  criterion for heteroscedastic data. A homoscedasticity testing procedure is proposed in Section 4, and a variance estimation method is proposed in Section 5. The proposed selection criterion for heteroscedastic data is compared with  $C_p$  in a simulation study in Section 6 and a real data set in Section 7. We conclude with some

closing remarks in Section 8.

## 2 Review of some selection criteria

### 2.1 Mallows' $C_p$

Consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad E(\boldsymbol{\epsilon}) = \mathbf{0}, \quad (1)$$

where  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$  is an  $n \times k$  matrix of rank  $k$ ,  $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$ , and  $\mathbf{I}_n$  is the  $n \times n$  identity matrix. Mallows (1964) introduced a criterion for model selection

$$\Gamma_p = \sigma^{-2} (E\hat{\mathbf{Y}}_p - \mathbf{X}\boldsymbol{\beta})' (E\hat{\mathbf{Y}}_p - \mathbf{X}\boldsymbol{\beta}) + p$$

where  $\mathbf{X}_p$  is the design matrix for the  $p$  selected variables,  $\hat{\mathbf{Y}}_p = \mathbf{X}_p \hat{\boldsymbol{\beta}}_p$ , and  $\hat{\boldsymbol{\beta}}_p$  is the least squares estimator of  $\boldsymbol{\beta}_p$ , the coefficient vector for the selected variables. Letting  $\text{SSB} = (E\hat{\mathbf{Y}}_p - \mathbf{X}\boldsymbol{\beta})' (E\hat{\mathbf{Y}}_p - \mathbf{X}\boldsymbol{\beta})$ , we see that  $\Gamma_p = \sigma^{-2} \text{SSB} + p$ , i.e.,  $\Gamma_p$  is a measure of the trade-off between the model bias (SSB) and model complexity ( $p$ ). Typically, SSB is large if  $p$  is small and vice versa. Therefore the best model is the one that minimizes  $\Gamma_p$ .

Since  $\Gamma_p$  is unknown, Mallows proposed that it be estimated by

$$C_p = \hat{\sigma}^{-2} \text{RSS}_p - (n - 2p) \quad (2)$$

where  $\text{RSS}_p = (\mathbf{Y} - \mathbf{X}_p \hat{\boldsymbol{\beta}}_p)' (\mathbf{Y} - \mathbf{X}_p \hat{\boldsymbol{\beta}}_p)$ ,  $\hat{\sigma}^2 = \text{RSS}/(n - k)$ ,  $\text{RSS} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ , and  $\hat{\boldsymbol{\beta}}$  is the least squares estimator of  $\boldsymbol{\beta}$ . The argument is that if a  $p$ -variable regression model adequately describes the data, the bias should be negligible, i.e.,  $\text{SSB} \simeq 0$ . In this case,  $\text{RSS}_p/(n - p)$  and  $\hat{\sigma}^2$  should be close since they both estimate  $\sigma^2$ . Therefore

$$C_p = \hat{\sigma}^{-2} \text{RSS}_p - (n - 2p) \simeq p. \quad (3)$$

Now  $E(\text{RSS}_p) = n\sigma^2 + \boldsymbol{\beta}' \mathbf{X}' \mathbf{X} \boldsymbol{\beta} - E(\mathbf{Y}' \mathbf{A} \mathbf{Y})$ , where  $\mathbf{A} = \mathbf{X}_p (\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{X}'_p$  and  $E(\mathbf{Y}' \mathbf{A} \mathbf{Y}) = \boldsymbol{\beta}' \mathbf{X}' \mathbf{A} \mathbf{X} \boldsymbol{\beta} + \text{tr}[\mathbf{A} E(\boldsymbol{\epsilon} \boldsymbol{\epsilon}')] = \boldsymbol{\beta}' \mathbf{X}' \mathbf{A} \mathbf{X} \boldsymbol{\beta} + \sigma^2 p$ . Hence  $E(\text{RSS}_p) = (n - p)\sigma^2 + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I}_n - \mathbf{A}) \mathbf{X} \boldsymbol{\beta} = (n - p)\sigma^2 + \text{SSB}$  and  $C_p \simeq \sigma^{-2} E(\text{RSS}_p) - (n - 2p) = \sigma^{-2} \text{SSB} + p = \Gamma_p$ . In this sense, selection of the optimal set of variables involves identifying those sets which lead to the smallest  $C_p$  values

and, in the light of (3), choosing  $C_p$  values which are close to  $p$ . Note, however, that if  $p = k$ , then  $\text{RSS}_p = (n - k)\hat{\sigma}^2$  and  $C_p = k$ . Therefore, if we choose only the  $C_p$  value that is closest to  $p$ , the selected model is always the full model. Because of random variation, points representing well-fitting equations can fall below the  $C_p = p$  line (Draper and Smith, 1998). For this reason, we choose the model with the smallest  $C_p$  value in this paper. This is equivalent to AIC with the assumption of a normal error distribution.

## 2.2 Akaike information criterion (AIC)

Write  $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$  where  $\mathbf{X}_1$  has  $p$  columns and  $\mathbf{X}_2$  has  $r = k - p$  columns, so that  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$  with  $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$ . The  $p$ -variable model  $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$  is equivalent to the hypothesis  $Q\boldsymbol{\beta} = (O_{(r \times p)}, \mathbf{I}_r)\boldsymbol{\beta} = \mathbf{0}$ , that is, the last  $r$  elements of  $\boldsymbol{\beta}$  corresponding to the unselected variables are zero. Akaike (1973, 1974) proposed that model selection be based on minimization of the quantity  $\text{AIC} = -\ln L(\hat{\boldsymbol{\beta}}_1 | \mathbf{y}) + 2p$ , where  $L$  is the likelihood function and  $\hat{\boldsymbol{\beta}}_1$  is the maximum likelihood estimator of  $\boldsymbol{\beta}_1$  under the  $p$ -variable model. Suppose that  $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ , i.e., the errors are normally distributed with covariance matrix  $\mathbf{V}$ . If  $Q\boldsymbol{\beta} = \mathbf{0}$ , we have

$$\text{AIC} = n \ln(2\pi) + \ln |\hat{\mathbf{V}}| + (\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1)' \hat{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1) + 2p$$

where  $\hat{\mathbf{V}}$  is the maximum likelihood estimate of  $\mathbf{V}$ . Amemiya (1980) showed that if  $\mathbf{V} = \sigma^2 \mathbf{I}_n$ , then AIC selects the same model as  $C_p$ .

## 2.3 Relations among selection criteria

Kennard (1971) showed that  $(C_p + n - 2p)/(n - p) = \{1 - R_a^2(p)\}/\{1 - R_a^2(k)\}$ , where  $R_a^2(p)$  and  $R_a^2(k)$  are the adjusted  $R^2$  values for the  $p$ -variable model and the full model, respectively. Hence, there is a one-to-one correspondence between the adjusted  $R^2$  and  $C_p$ . Similarly, a correspondence exists between the unadjusted  $R^2$  and  $C_p$ . Furthermore, let

$$F = \frac{(\text{RSS}_p - \text{RSS})/r}{\text{RSS}/(n - k)} \quad (4)$$

be the  $F$ -statistic for testing the significance of the  $p$ -variable model. Hocking (1976) observed that  $C_p = rF + p - r$ .

### 3 Generalized $C_p$ ( $GC_p$ )

Although  $C_p$  is a popular criterion for model selection, it is designed for homoscedastic models. Hence it may not perform well on highly heteroscedastic data. But  $C_p$  does not make any assumptions about the form of the error distribution. On the other hand, being likelihood-based, AIC is applicable to nonnormal and heteroscedastic data. If the likelihood is modeled incorrectly, AIC can perform poorly.

We now generalize the  $C_p$  statistic in (2) to make it applicable to heteroscedastic models. Suppose that  $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ , with  $\mathbf{V}$  nonsingular. In this study, we restrict attention to the case of independent errors. Thus  $\mathbf{V}$  is a diagonal matrix with  $\sigma_i^2$  in the  $i$ th diagonal position. Let  $\boldsymbol{\beta}^* = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}$  be the weighted least squares estimator of  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)$  based on some estimate  $\hat{\mathbf{V}}$  of  $\mathbf{V}$ . Suppose the null hypothesis is

$$H_0 : \beta_{i_1} = \dots = \beta_{i_r} = 0, \quad 1 \leq i_1 \leq \dots \leq i_r \leq k.$$

Define

$$\begin{aligned} \boldsymbol{\beta}_p &= (\beta_1, \dots, \beta_{i_1-1}, \beta_{i_1+1}, \dots, \beta_{i_2-1}, \beta_{i_2+1}, \dots, \beta_{i_r-1}, \beta_{i_r+1}, \dots, \beta_k) \\ \mathbf{X}_p &= (\mathbf{x}_1, \dots, \mathbf{x}_{i_1-1}, \mathbf{x}_{i_1+1}, \dots, \mathbf{x}_{i_2-1}, \mathbf{x}_{i_2+1}, \dots, \mathbf{x}_{i_r-1}, \mathbf{x}_{i_r+1}, \dots, \mathbf{x}_k) \end{aligned}$$

and let  $\boldsymbol{\beta}_p^*$  be the weighted least squares estimator of  $\boldsymbol{\beta}_p$  under  $H_0$ . Further, define

$$\begin{aligned} \text{RSS}^* &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*)'\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*) \\ \text{RSS}_p^* &= (\mathbf{Y} - \mathbf{X}_p\boldsymbol{\beta}_p^*)'\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}_p\boldsymbol{\beta}_p^*) \\ F^* &= \frac{(\text{RSS}_p^* - \text{RSS}^*)/r}{\text{RSS}^*/(n-k)}. \end{aligned}$$

Since  $F^*$  is the counterpart to (4) for heteroscedastic models, we define the generalized  $C_p$  as

$$GC_p = rF^* + p - r.$$

Clearly,  $GC_p$  reduces to  $C_p$  if we assume  $\mathbf{V} = \sigma^2\mathbf{I}_n$ .

## 4 Homoscedasticity testing procedure

In the current and next sections, we propose a method for estimating  $\mathbf{V}$  by dividing heteroscedastic data into homoscedastic subgroups. These methods select the best subsets using the choice of either  $C_p$  or  $GC_p$  according to the homoscedasticity test results. As mentioned in Section 1, we denote this adaptive procedure as  $AC_p$ . In order to use the  $GC_p$  statistic, we need to estimate  $\mathbf{V}$ . Two solutions are proposed. The first is for data with replication (with qualitative predictors), and the second is for data without replication (with quantitative predictors).

### 4.1 For qualitative predictors

We propose a method for calculating  $AC_p$  for data with qualitative predictors where there are replicates. A sequence of homoscedasticity tests for variances is carried out for the choice of  $GC_p$  or  $C_p$ . Several tests for homoscedasticity have been proposed in the literature. Lim and Loh (1996) compared several of them, including the Levene (1960), Bartlett (1937), modified Bartlett (Boos and Brownie, 1989), and Box-Andersen (Box and Andersen, 1955) tests. They find that no test is uniformly best for all distributions and sample sizes. The Levene test, however, controls the probability of a type I error well for many distributions and it is robust against non-normality. Therefore, the Levene test is used for the sequential homoscedasticity tests in this paper.

First, fit the regression model given in (1). The Levene test is used for sequential homoscedasticity tests as follows:

1. For each predictor variable  $\mathbf{x}_j$  ( $j = 1, \dots, k$ ), divide the sample into two groups according to the values (levels) of  $\mathbf{x}_j$  such that every group has at least  $2k$  observations. Then there are at most  $2^{g_j-1} - 1$  possible choices of grouping into two if there are  $g_j$  distinct values in  $\mathbf{x}_j$ . Perform the Levene test of homoscedasticity on the corresponding residuals for all possible such pairs of the groups. Divide the sample into two groups with the smallest  $p$ -value among all the tests. Denote this  $p$ -value as  $p_j$ .
2. Let  $p_{j_0}$  be the smallest  $p$ -value among  $p_1, \dots, p_k$ . Define  $n_j$  as the number of Levene's tests conducted on  $\mathbf{x}_j$ , and let  $\alpha$  be a pre-specified level of significance. If  $p_{j_0} > \alpha / \sum_{j=1}^k n_j$  (a Bonferroni-type adjustment), stop and conclude that the data are homoscedastic for the current sample. Otherwise, divide the data ( $\mathbf{Y}$  and  $\mathbf{X}$ ) into two groups ( $G_{L_{j_0}}$  and  $G_{U_{j_0}}$ ) according to  $\mathbf{x}_{j_0}$  at the split point determined in the above step and proceed to the next step.

3. Repeat the above procedure recursively on each of the two groups  $G_{U_{j_0}}$  and  $G_{L_{j_0}}$ . In each of these groups, the significance levels are adjusted by the number of the tests conducted according to the number of subgroups as in Step 2.

## 4.2 For quantitative predictors

In this section, we develop a method for calculating  $AC_p$  in regression situations with quantitative predictors where there are insufficient replicates. First, observations that are near each other in  $\mathbf{X}$ -space are grouped together to form  $g$  groups, say. In order to form the groups, we fit the regression model (1) and perform the following procedure.

1. For each predictor variable  $\mathbf{x}_j$  ( $j = 1, \dots, k$ ), sort the observations according to the values of  $\mathbf{x}_j$  and partition them into  $h_j$  groups  $G_1, \dots, G_{h_j}$  of roughly equal size of at least  $2k$ . If the variable has observations with the same values, they are assigned to the same group. Perform the Levene (1960) test of homoscedasticity for the  $h_j - 1$  pairs of the adjacent groups  $(G_i, G_{i+1})$  of the corresponding residuals,  $i = 1, \dots, h_j - 1$ . Merge the two groups with the largest  $p$ -value from the  $h_j - 1$  tests. This procedure results in a partition of  $h_j - 1$  groups. In the same way, perform the Levene test for each pair of the adjacent groups of the corresponding residuals in the new partition and merge the pair with the largest  $p$ -value. Continue this process until two groups ( $G_{L_j}$  and  $G_{U_j}$ ) are left. Perform the Levene test one more time on these two groups and obtain its  $p$ -value. Denote this  $p$ -value as  $p_j$ .
2. Let  $p_{j_0}$  be the smallest  $p$ -value among  $p_1, \dots, p_k$ . Let  $\alpha$  be a pre-specified level of significance. If  $p_{j_0} > \alpha / \left[ \sum_{j=1}^k (h_j - 1) \right]$ , stop and conclude that the data are homoscedastic for the current sample. Otherwise, sort the data ( $\mathbf{Y}$  and  $\mathbf{X}$ ) according to  $\mathbf{x}_{j_0}$  and proceed to the next step.
3. Repeat the above procedure recursively on each of the two groups  $G_{U_{j_0}}$  and  $G_{L_{j_0}}$ . In each of these groups, the sample is partitioned into small groups with roughly equal sample sizes as described in Step 1. In each of the two groups  $G_{U_{j_0}}$  and  $G_{L_{j_0}}$ , the significance levels are adjusted by the number of subgroups as in Step 2.

## 5 Variance estimation

From the homoscedasticity testing procedure given in Section 4, if the data are determined to be homoscedastic, use  $C_p$  for variable selection. Otherwise, the above procedure yields the partitions

$\mathbf{Y}' = (\mathbf{Y}'_1, \dots, \mathbf{Y}'_g)'$  and  $\mathbf{X}' = (\mathbf{X}'_1, \dots, \mathbf{X}'_g)'$ , where  $\mathbf{Y}_i$  is an  $n_i$ -dimensional vector and  $\mathbf{X}_i$  is an  $n_i \times k$  matrix for  $i = 1, \dots, g$ . The error variance  $\sigma_i^2$  in the  $i$ th group is estimated as

$$\hat{\sigma}_i^2 = (n_i - k)^{-1} (\mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}})' (\mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}), \quad i = 1, \dots, g$$

where  $\hat{\boldsymbol{\beta}}$  is the ordinary least squares estimator. Let

$$\mathbf{W} = \begin{pmatrix} \hat{\sigma}_1^2 \mathbf{I}_{n_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\sigma}_g^2 \mathbf{I}_{n_g} \end{pmatrix}.$$

Next we update the estimate of  $\boldsymbol{\beta}$  with the weighted least squares estimator

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{-1} \mathbf{Y}.$$

This in turn yields the updated estimates

$$\tilde{\sigma}_i^2 = (n_i - p)^{-1} (\mathbf{Y}_i - \mathbf{X}_i \tilde{\boldsymbol{\beta}})' (\mathbf{Y}_i - \mathbf{X}_i \tilde{\boldsymbol{\beta}}), \quad i = 1, \dots, g$$

and

$$\mathbf{W}^* = \begin{pmatrix} \tilde{\sigma}_1^2 \mathbf{I}_{n_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \tilde{\sigma}_g^2 \mathbf{I}_{n_g} \end{pmatrix}.$$

The latter is used to produce the final updated estimates  $\boldsymbol{\beta}^* = (\mathbf{X}' \mathbf{W}^{*-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{*-1} \mathbf{Y}$  for the full model and  $\boldsymbol{\beta}_p^* = (\mathbf{X}'_p \mathbf{W}^{*-1} \mathbf{X}_p)^{-1} \mathbf{X}'_p \mathbf{W}^{*-1} \mathbf{Y}$  for a reduced model from which  $GC_p$  is calculated.

## 6 Simulation results

We now report the results of a simulation to evaluate the proposed homoscedasticity testing procedure for use in  $AC_p$  (or AIC with normal likelihood based on our variance estimates) as well as to compare their performance with that of  $C_p$  (or AIC with a normal likelihood). For each error distribution, 10000 simulated data sets are generated from model (1) with  $k = 5$ . Level  $\alpha = 0.05$

is used throughout for all the tests.

## 6.1 Method for qualitative predictors

### 6.1.1 Homoscedasticity tests and variance estimation

The regression coefficients  $\beta = (1, 1, 0, 1, 1)$  are considered. Define  $\mathbf{a}_m$  as an  $m$  vector of  $a$ 's. Four different design matrices  $\mathbf{X} = (\mathbf{1}, \mathbf{X}_2, \dots, \mathbf{X}_5)$ , are considered as follows.

D1  $\mathbf{X}_2 = (-\mathbf{1}'_{72}, \mathbf{1}'_{72})'$ ,  $\mathbf{X}_3 = (\mathbf{0}'_{48}, \mathbf{1}'_{48}, \mathbf{2}'_{48})'$ ,  $\mathbf{X}_4$  is a vector of random numbers from  $\{-2, -1, 1, 2\}$  and  $\mathbf{X}_5$  is a vector of random numbers from  $\{-1, 1\}$ .

D2  $\mathbf{X}_2 = (\mathbf{0}'_{48}, \mathbf{1}'_{48}, \mathbf{2}'_{48})'$ ,  $\mathbf{X}_3 = (-\mathbf{1}'_{24}, \mathbf{1}'_{24}, -\mathbf{1}'_{24}, \mathbf{1}'_{24}, -\mathbf{1}'_{24}, \mathbf{1}'_{24})'$ ,  $\mathbf{X}_4 = (\mathbf{0}'_{24}, \mathbf{1}'_{48}, \mathbf{2}'_{48}, \mathbf{0}'_{24})'$  and  $\mathbf{X}_5$  is a vector of random numbers from  $\{-1, 1\}$ .

D3  $\mathbf{X}_2 = (-\mathbf{1}'_{144}, \mathbf{1}'_{144})'$ ,  $\mathbf{X}_3 = (\mathbf{0}'_{72}, \mathbf{1}'_{144}, \mathbf{2}'_{72})'$ ,  $\mathbf{X}_4$  is a vector of random numbers from  $\{0, 1, 2\}$  and  $\mathbf{X}_5$  is a vector of random numbers from  $\{-1, 1\}$ .

D4  $\mathbf{X}_2 = (-\mathbf{1}'_{72}, \mathbf{1}'_{72})'$ ,  $\mathbf{X}_3 = (-\mathbf{1}'_{36}, \mathbf{1}'_{36}, -\mathbf{1}'_{36}, \mathbf{1}'_{36})'$ ,  $\mathbf{X}_4 = (\mathbf{0}'_{48}, \mathbf{1}'_{48}, \mathbf{2}'_{48})'$  and  $\mathbf{X}_5$  is a vector of random numbers from  $\{0, 1, 2\}$ .

A diagram of the above design matrices is given in Figure 1.

The error distributions are chosen as follows.

M1  $\epsilon \sim N(\mathbf{0}, \mathbf{I}_n)$

M2  $\epsilon \sim N(\mathbf{0}, 4\mathbf{I}_n)$

M3  $\epsilon \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/2}, 4\mathbf{I}_{n/2})$

M4  $\epsilon \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/2}, 16\mathbf{I}_{n/2})$

M5  $\epsilon \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/2}, 36\mathbf{I}_{n/2})$

M6  $\epsilon \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/3}, 4\mathbf{I}_{n/3}, 16\mathbf{I}_{n/3})$

M7  $\epsilon \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/3}, 9\mathbf{I}_{n/3}, 81\mathbf{I}_{n/3})$

M8  $\epsilon \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/3}, 16\mathbf{I}_{n/3}, 256\mathbf{I}_{n/3})$

M9  $\epsilon \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/4}, 27\mathbf{I}_{n/4}, 9\mathbf{I}_{n/4}, 81\mathbf{I}_{n/4})$

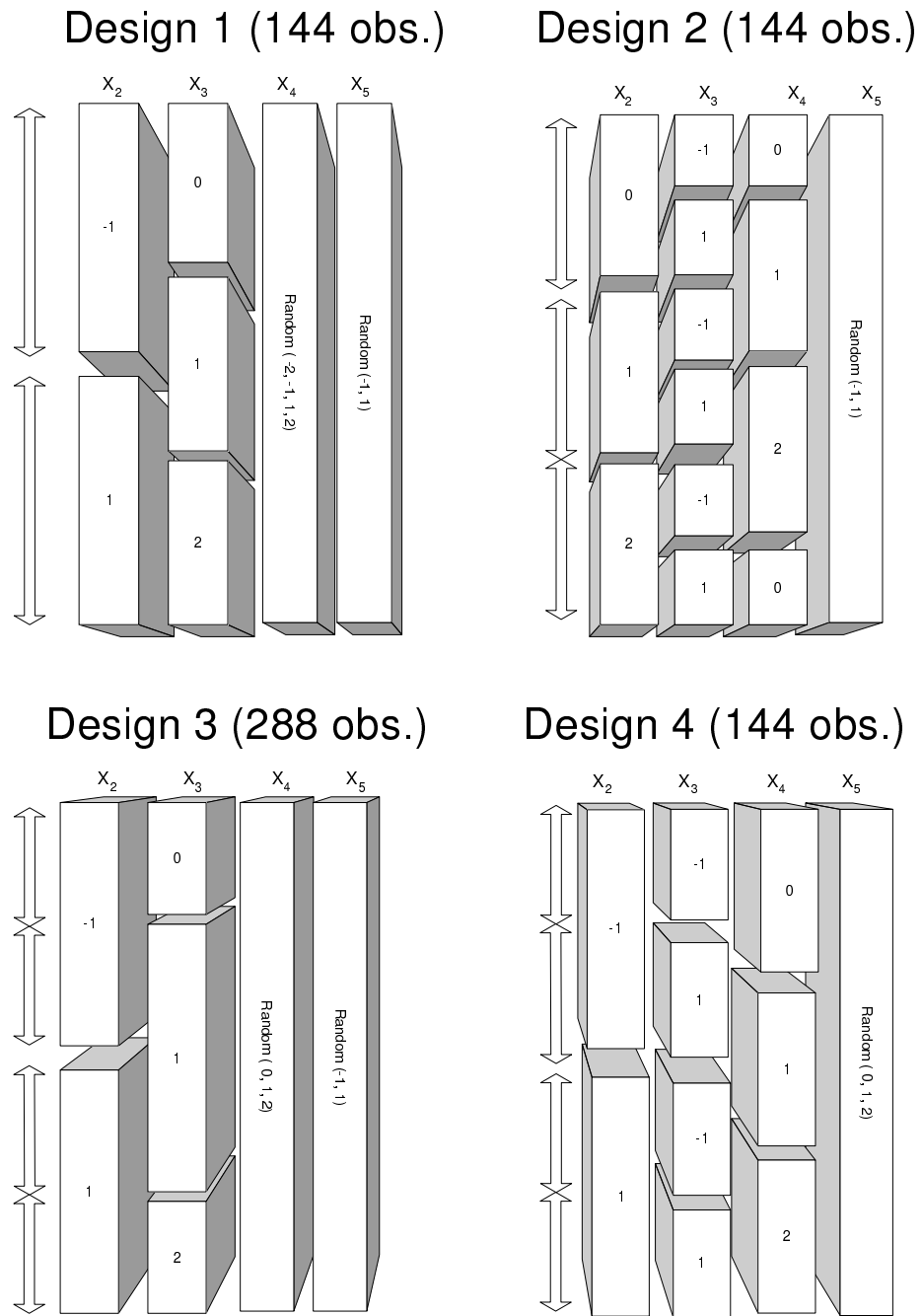


Figure 1: Diagram of the four design matrices considered in the simulation study.

M10  $\epsilon \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/4}, 64\mathbf{I}_{n/4}, 9\mathbf{I}_{n/4}, 81\mathbf{I}_{n/4})$

M11  $\epsilon \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/4}, 4\mathbf{I}_{n/4}, \mathbf{I}_{n/4}, 4\mathbf{I}_{n/4})$

M12  $\epsilon \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/4}, 16\mathbf{I}_{n/4}, \mathbf{I}_{n/4}, 16\mathbf{I}_{n/4})$

M13  $\epsilon \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/4}, 64\mathbf{I}_{n/4}, \mathbf{I}_{n/4}, 64\mathbf{I}_{n/4})$

M14  $\epsilon_i \sim t_3$  ( $t$  distribution with 3 degrees of freedom)

M15  $\epsilon \sim$  multivariate  $t$  such that  $\epsilon_i = 3t_{20}/\sqrt{10}$  with  $\text{Var}(\epsilon_i) = 1$  for  $i = 1, \dots, n/2$  and  $\epsilon_i = 2\sqrt{3}t_3$  with  $\text{Var}(\epsilon_i) = 36$  for  $i = (n/2) + 1, \dots, n$

M16  $\epsilon \sim$  multivariate  $t$  such that  $\epsilon_i = 3t_{20}/\sqrt{10}$  with  $\text{Var}(\epsilon_i) = 1$  for  $i = 1, \dots, n/3$ ,  $\epsilon_i = 3\sqrt{3}t_8/2$  with  $\text{Var}(\epsilon_i) = 9$  for  $i = (n/3) + 1, \dots, 2n/3$  and  $\epsilon_i = 3\sqrt{3}t_3$  with  $\text{Var}(\epsilon_i) = 81$  for  $i = (2n/3) + 1, \dots, n$

M17  $\epsilon \sim$  multivariate  $t$  such that  $\epsilon_i = 3t_{20}/\sqrt{10}$  with  $\text{Var}(\epsilon_i) = 1$  for  $i = 1, \dots, n/4$ ,  $\epsilon_i = 6\sqrt{6}t_{14}/\sqrt{7}$  with  $\text{Var}(\epsilon_i) = 36$  for  $i = (n/4) + 1, \dots, n/2$ ,  $\epsilon_i = 3\sqrt{3}t_8/2$  with  $\text{Var}(\epsilon_i) = 9$  for  $i = (n/2) + 1, \dots, 3n/4$  and  $\epsilon_i = 3\sqrt{3}t_3$  with  $\text{Var}(\epsilon_i) = 81$  for  $i = (3n/4) + 1, \dots, n$

M18  $\epsilon \sim$  multivariate  $t$  such that  $\epsilon_i = 3t_{20}/\sqrt{10}$  with  $\text{Var}(\epsilon_i) = 1$  for  $i = 1, \dots, n/4$  and  $i = (n/2) + 1, \dots, 3n/4$ ,  $\epsilon_i = 2\sqrt{3}t_3$  with  $\text{Var}(\epsilon_i) = 36$  for  $i = (n/4) + 1, \dots, n/2$  and  $i = (3n/4) + 1, \dots, n$

M1, M2 and M14 are for evaluating the Type I error rate, and M3 – M13 and M15 – M18 are for evaluating the power.

Frequency of the correct partition is reported for each configuration. Table 1 exhibits Type I error rates (rates of not rejecting the assumption of homoscedasticity) of the proposed homoscedasticity testing procedure. The Type I error rates are close to the nominal significance level. It shows an excellent control of the Type I error rates for all the design matrices and error distributions.

Table 2 displays the frequency of the correct and wrong partitions using the proposed test procedure for the data from each model. The correct partition is indicated using boldface. The frequency of correct classification is quite high for the models chosen in this study. The power can be obtained by adding the frequencies of selecting more than one group.

Table 1: Simulated Type I error rate (%) for the proposed homoscedasticity test procedure for quantitative predictors. Frequencies are computed from 10000 replicates. Significance level  $\alpha = 0.05$  is used.

Design matrix	Error distribution		
	M1	M2	M14
D1	4.57	4.43	4.41
D2	4.85	5.37	4.63
D3	4.75	4.48	4.51
D4	4.86	4.57	4.57

Table 2: Frequency (%) of classification of groups using the proposed homoscedasticity test procedure. Frequencies are computed from 10000 replicates for each model. Correct partition is indicated using boldface. Significance level  $\alpha = .05$  is used.

Design matrix	Partition formed	Error distribution			
		M3	M4	M5	M15
D1	One group	0.97	0.00	0.00	0.58
	$(y_1, \dots, y_{72}), (y_{73}, \dots, y_{144})$	<b>84.89</b>	<b>90.44</b>	<b>90.00</b>	<b>86.05</b>
	Others	14.14	9.56	10.00	13.37
D2	One group	0.00	0.00	0.00	0.09
	$(y_1, \dots, y_{48}), (y_{49}, \dots, y_{96}), (y_{97}, \dots, y_{144})$	<b>74.39</b>	<b>88.47</b>	<b>87.61</b>	<b>84.76</b>
	Others	25.61	11.53	12.39	15.15
D3	One group	0.00	0.00	0.00	0.00
	$(y_1, \dots, y_{72}), (y_{73}, \dots, y_{144}), (y_{145}, \dots, y_{216}), (y_{217}, \dots, y_{288})$	<b>75.84</b>	<b>80.06</b>	<b>68.16</b>	
	Others	24.16	19.94	31.84	
D4	One group	0.93	0.00	0.00	0.54
	$(y_1, \dots, y_{36}, y_{73}, \dots, y_{108}), (y_{37}, \dots, y_{72}, y_{109}, \dots, y_{144})$	<b>90.70</b>	<b>91.29</b>	<b>82.97</b>	<b>87.77</b>
	Others	8.37	8.71	17.03	11.69

Table 3: Frequency (%) of selecting the correct model for homoscedastic data. Frequencies are computed from 10000 replicates.

Design matrix	Model	Selection criterion	
		$C_p$	$AC_p$
D1	M1	.8405	.8373
	M2	.8360	.8335
	M14	.8385	.8378
D2	M1	.8421	.8411
	M2	.8434	.8424
	M14	.8333	.8328
D3	M1	.8452	.8439
	M2	.8403	.8399
	M14	.8382	.8377
D4	M1	.8362	.8343
	M2	.6884	.6879
	M14	.7686	.7669

### 6.1.2 Variable selection

We now report the results of a simulation study for evaluating the performance of the  $AC_p$  based on the proposed homoscedasticity test procedure. If the data are determined to be homoscedastic from the proposed test procedure, use  $C_p$  for variable selection. Recall from Section 1 that AIC reduces to  $C_p$  for a normal likelihood with homoscedasticity assumption. If our test procedure rejects the null hypothesis of homoscedasticity,  $GC_p$  (which is equivalent to AIC for heteroscedastic model with a normal likelihood and the variance estimates using the proposed method) is used according to the variance matrix estimated under a heteroscedastic model based on the procedure given in Section 5.

Table 3 shows the results for homoscedastic data. The correct model is one without the third covariate. In this simulation, the correct subset of the variables is chosen more than 80% of the time except for some error distributions with design matrix D4. Since the heteroscedasticity test does not reject the null hypothesis most of the time for these models, the frequencies are almost the same for the two criteria. This is because  $AC_p$  (or AIC with a normal likelihood) is equivalent to  $C_p$  for homoscedastic models.

Table 4 shows the results of the variable selection using our test procedure for heteroscedastic data. In general,  $AC_p$  (AIC with a normal likelihood utilizing our heteroscedasticity test) selects

Table 4: Frequency (%) of selecting the correct model for heteroscedastic data. Frequencies are computed from 10000 replicates.

Design matrix	Model	Selection criterion	
		$C_p$	$AC_p$
D1	M3	.8358	.8294
	M4	.7979	.8235
	M5	.6086	.7482
	M15	.6402	.7647
D2	M6	.7897	.8107
	M7	.2002	.4940
	M16	.2750	.5390
D3	M9	.6485	.8126
	M10	.4968	.7901
	M17	.6332	.8136
D4	M11	.7799	.8153
	M12	.3365	.7708
	M13	.0137	.7274
	M18	.1206	.7494

the correct model more often than  $C_p$  alone. The frequency of selecting the correct model does not vary much for different design matrices and error distributions for  $AC_p$ , but it substantially decreases as the degree of heteroscedasticity increases when  $C_p$  is used. These results suggest that a criterion based on heteroscedasticity in conjunction with the proposed heteroscedasticity test is necessary.

## 6.2 Method for quantitative predictors

The simulated data are generated from model (1) with  $n = 100$  and  $\boldsymbol{\beta} = (4, 0, 4, 4, 4)$ . The  $\mathbf{X}$  matrix consists of the intercept and independent random numbers  $x_{ij}$  from Uniform(0, 1),  $i = 2, \dots, 5$ ,  $j = 1, \dots, n$ . Therefore the correct model is one without the second covariate. We consider some of the error distributions given in Section 6.1 and the following additional models.

M19  $\epsilon_i \sim$  double exponential with density function  $f(x) = \exp(-|x|/2)/4$ ,  $i = 1, \dots, n$

M20  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/2}, 64\mathbf{I}_{n/2})$

M21  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/2}, 100\mathbf{I}_{n/2})$

Table 5: Frequency (%) of selecting the correct model for homoscedastic data. Frequencies are computed from 10000 replicates for each model. Significance level  $\alpha = .05$  is used.

Distribution	Selected model	$C_p$	$AC_p$
M1	Full model	.1547	.1577
	<b>Correct model</b>	<b>.8453</b>	<b>.8423</b>
	Other models	.0000	.0000
M14	Full model	.1628	.1639
	<b>Correct model</b>	<b>.8335</b>	<b>.8350</b>
	Other models	.0037	.0011
M19	Full model	.1603	.1637
	<b>Correct model</b>	<b>.8181</b>	<b>.8157</b>
	Other models	.0216	.0206

M22  $\epsilon \sim N(\mathbf{0}, \mathbf{V})$ , where  $\mathbf{V} = \text{diag}(\mathbf{I}_{n/4}, 4\mathbf{I}_{n/4}, 16\mathbf{I}_{n/4}, 64\mathbf{I}_{n/4})$

M23  $\epsilon \sim$  multivariate  $t$  such that  $\epsilon_i = .2t_{20}$  with  $\text{Var}(\epsilon_i) = .044$  for  $i = 1, \dots, n/2$  and  $\epsilon_i = 5t_3$  with  $\text{Var}(\epsilon_i) = 75$  for  $i = (n/2) + 1, \dots, n$ .

The following table displays frequency of the correct partition for each configuration, which is the Type I error rate. The Type I error rate is close to the nominal significance level for M1 and M19, but it is quite high for M14.

Error distribution	M1	M14	M19
Type I error rate (%)	3.61	10.51	5.7

Table 5 compares the frequencies of the selected models by the two criteria for the homoscedastic data.  $AC_p$  selects the correct model as frequent as  $C_p$ . The correct model is selected more frequently than 80%. In most of the cases,  $C_p$  is used in  $AC_p$  procedure because our variance estimation method assumed the data as homoscedastic according to the homoscedasticity test procedure.

Table 6 compares the results of variable selection using our test procedure for heteroscedastic data.  $AC_p$  outperforms  $C_p$  in selecting the correct variables for all the heteroscedastic models considered in this section. The accuracy of selecting the correct model by  $AC_p$  is near or above 80% for most of the distributions.  $C_p$  performs poorly for all the models and the frequency of selecting the correct model decreases as heteroscedasticity becomes more substantial.

Table 6: Frequency (%) of selecting the correct model for heteroscedastic data. Frequencies are computed from 10000 replicates.

Distribution	Selected model	$C_p$	$AC_p$
M5	Full model	.1181	.1827
	<b>Correct model</b>	<b>.5892</b>	<b>.8173</b>
	Other models	.2927	.0000
M20	Full model	.0655	.1780
	<b>Correct model</b>	<b>.3153</b>	<b>.8220</b>
	Other models	.6192	.0000
M21	Full model	.0323	.1755
	<b>Correct model</b>	<b>.1681</b>	<b>.8245</b>
	Other models	.7996	.0000
M22	Full model	.0217	.1973
	<b>Correct model</b>	<b>.0670</b>	<b>.6729</b>
	Other models	.9113	.1428
M23	Full model	.0689	.1315
	<b>Correct model</b>	<b>.3504</b>	<b>.8512</b>
	Other models	.5807	.0173

## 7 Example: Swedish motor insurance data

The problem concerning which factors would influence the number of claims in car insurance is an important problem that has been dealt with in insurance companies. In this study, we investigate this problem by analyzing the Third Party Motor Insurance data for Sweden in 1977 described by Andrews and Herzberg (1985). The data can be accessed from URL [www.statsci.org/data/general/motorins.html](http://www.statsci.org/data/general/motorins.html). The original data contain over 3000 cases including observations in which the number of claims is zero. In this analysis, we consider the data from geographical zone 1, which consists of the three largest cities; Stockholm, Göteborg, Malmö and surroundings. We exclude 20 cases with no claim in 1977. This leaves 295 observations.

The data contain both qualitative and quantitative variables. Thus we use a combination of the homoscedasticity testing procedures given in Sections 4.1 and 4.2. The data contain three explanatory variables as follows.

1. Traveled (T) – The distance an insured car traveled in kilometers and the ranges of the values are given by: less than 1000 km per year (1), 1000-15000 km per year (2), 15000-20000 km per year (3), 20000-25000 km per year (4), and more than 25000 km per year (5).
2. Bonus (B) – A measure of individual claim history. The insured starts with the class B=1. Every year he/she is moved up one class if there is no claim. It can reach up to B=7.
3. Make (M) – Classified into 9 (labeled 1–9) models. There is no further information on this variable how the car model is classified.

These data also include number of insured (INS), number of claims, and total claim payment (PMT). It is evident that PMT tends to increase as the number of insured increases. Thus, we used PPI (average payment per insured =  $PMT/INS$ ) to determine how causal variables affect average payment of each group of insured.

Since higher Bonus (B) implies that the insured customer is less likely to have an accident, we assume this variable as quantitative. Variable T is also considered a quantitative variable. Since M is categorical, eight 0-1 variables are created for this variable. The dummy variables are named M1 through M8. Here,  $M_i = 1$  if  $M = i$  and  $M_i = 0$  otherwise for  $i = 1, \dots, 8$ . It is evident that the count of car accidents is related to how much cars are exposed to the roads (T) and driver's history of car accidents (B). Therefore, this study focuses on investigating how the count of accidents responses with driver's car model.

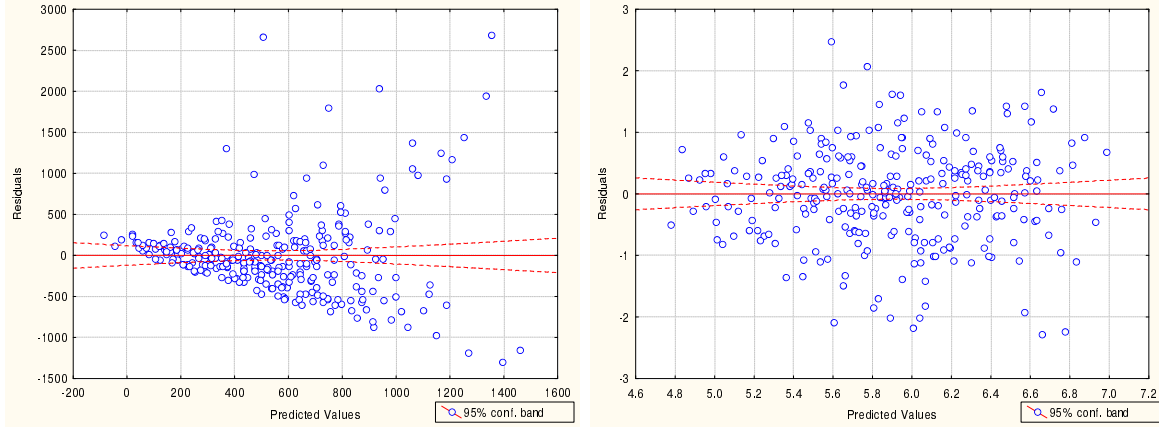


Figure 2: Residual plots of PPI (left) and  $\log(\text{PPI})$  (right)

The first plot of Figure 2 displays the residual plot of the regression by the ordinary least squares of PPI fitted by T, B, and M. In this figure, we notice that magnitude of the residuals is increasing as the corresponding predicted values are increasing. It suggests a transformation on PPI to eliminate this tendency. Hence, we use  $\log(\text{PPI})$  as our dependent variable to alleviate this problem. The second plot of Figure 2 displays the residual plot after a log-transformation. The transformed data appear to be heteroscedastic.

Our homoscedasticity test procedure splits the data into 5 groups of different variances. As shown in Figure 3, data are first split on M8 (Make = 8 or not). A low  $p$ -value ( $6 \times 10^{-9}$ ) of the Levene test indicates that these data are highly heteroscedastic. The data are further split on T, M7 and B.

The first group contains 112 observations and its sample variance is estimated to be .278. The second group is of size 76 and the variance estimate is .842. The third group has 60 observations with variance estimate .389. The fourth and the fifth groups have 18 and 29 observations with variance estimates 3.832 and 2.494, respectively. The last split occurs on Bonus between 4 and 5. This split indicates that among groups of higher mileage (more than 15,000 km per year), there is a significant difference in variances between the high Bonus group and the low Bonus group. Figure 3 also suggests that the data are highly heteroscedastic. The covariance matrix  $\mathbf{V} = \text{diag}(.278\mathbf{I}_{112}, .842\mathbf{I}_{76}, .389\mathbf{I}_{60}, 3.832\mathbf{I}_{18}, 2.494\mathbf{I}_{29})$  is used in the weighted least squares regression on  $\log(\text{PPI})$  when  $AC_p$  is used for selection procedure.

For the data,  $C_p$  selects the model consisting of T, B, M2, M4, M6 and M7, while  $AC_p$  selects

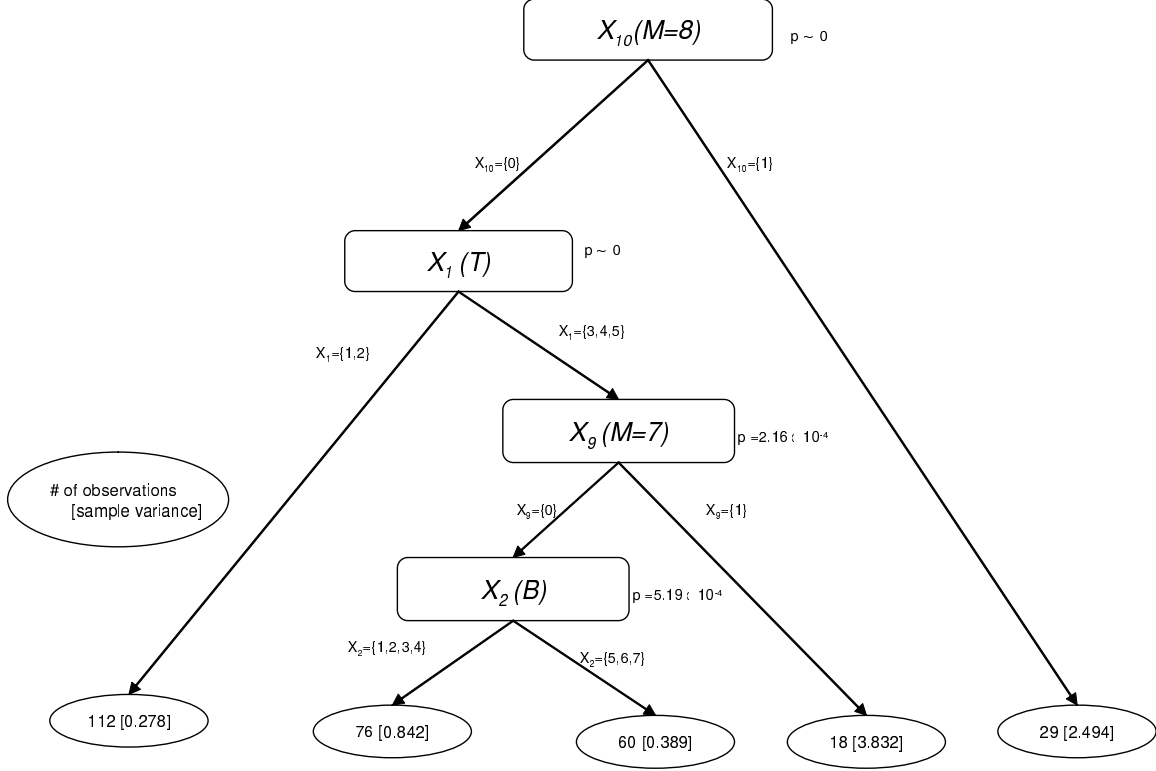


Figure 3: Hierarchy diagram of the homoscedasticity test for the Swedish auto insurance data.

a simpler model consisting of T, B, M2, M4 and M6. The regression model selected by  $AC_p$  is

$$\log(\text{PPI}) = 6.5113 + 0.0791T - 0.1840B + 0.3572M2 - 0.6129M4 - 0.4892M6$$

and the model selected by  $C_p$  is

$$\log(\text{PPI}) = 6.5698 + 0.0584T - 0.1779B + 0.3037M2 - 0.6075M4 - 0.5190M6 - 0.3812M7.$$

Table 7 shows that all the regression coefficients of the model selected by  $AC_p$  are highly significant. However, Table 8 shows that not all the variables selected by  $C_p$  are significant at level .05. The  $p$ -value for T from the model selected by  $C_p$  (based on the ordinary least squares fit) is .0735 (Table 8), while the  $p$ -value of T from the model selected by  $AC_p$  (based on the weighted least squares fit) is .003 (Table 7). Moreover, the  $p$ -value of M2 from the model selected by  $C_p$  is

Table 7: Regression coefficients and standard errors of a weighted least squares regression with the model selected by  $AC_p$ . The  $p$ -value is based on the two-sided test.

	Estimate	Standard Error	$t$ -value	$p$ -value
Intercept	6.51142	.10081	64.594	$< 2 \times 10^{-16}$
Traveled	.07907	.02632	3.004	0.002897
Bonus	-.18400	.01761	-10.447	$< 2 \times 10^{-16}$
M2	.35717	.10607	3.367	0.000862
M4	-.61292	.11378	-5.387	$1.49 \times 10^{-7}$
M6	-.48921	.10607	-4.612	$5.99 \times 10^{-6}$

Table 8: Regression coefficients and standard errors of ordinary least squares regression with the model selected by  $C_p$ . The  $p$ -value is based on the two-sided test.

	Estimate	Standard Error	$t$ -value	$p$ -value
Intercept	6.56981	.14417	45.571	$< 2 \times 10^{-16}$
Traveled	.05841	.03252	1.796	.073568
Bonus	-.17792	.02246	-7.922	$5.14 \times 10^{-14}$
M2	.30369	.14490	2.096	.036963
M4	-.60753	.16223	-3.745	.000218
M6	-.51899	.14490	-3.582	.000400
M7	-.38123	.15046	-2.534	.011814

.037, while that from the model selected by  $AC_p$  is .0009. These results suggest that  $AC_p$  is more appropriate for heteroscedastic data.

## 8 Conclusion

Because Mallows'  $C_p$  assumes the homoscedastic error variance, it is not expected to show good results in variable selection for the data with high heteroscedasticity. We proposed a generalized  $C_p$  for heteroscedastic data, which is based on the weighted least squares fit using the positive definite variance-covariance matrix  $\mathbf{V}$  of the errors. There is no generic algorithm to obtain the estimate of  $\mathbf{V}$  which is applicable to various kinds of data.

In this study, we proposed methods to obtain  $\hat{\mathbf{V}}$ . The methods consist of a sequential homoscedasticity test procedure and a variance estimation.  $C_p$  is used if the data are determined to be homoscedastic, and  $GC_p$  is used otherwise. This adaptive procedure can be considered an improved AIC under the assumption of normal error distribution. The likelihood function for AIC can be obtained using our estimate of the variance-covariance matrix.

Because  $GC_p$  reduces to  $C_p$ , it tends to select the same model as  $C_p$  if homoscedasticity is assumed in the model. On the other hand, if heteroscedasticity is assumed,  $GC_p$  is usually more accurate than  $C_p$  when there is substantial heteroscedasticity. In our simulation study,  $AC_p$  outperformed  $C_p$  for highly heteroscedastic data with replication. AIC and  $GC_p$  select the same model under the normal homoscedastic model assumption, if we use the same variance estimates.  $AC_p$  performed better than  $C_p$  for unequal variance situations with moderate to large variance differences.

From the analysis of Swedish auto insurance data, we observed that the proposed testing procedure adequately formed homoscedastic subgroups. In this analysis,  $AC_p$  selected the model containing less predictor variables than that selected by  $C_p$ . All the variables selected by  $AC_p$  were highly significant. Based on this observation, the selection procedure seems to be benefited by the consideration of heteroscedasticity.

## Acknowledgement

The authors would like to thank Kenny Ye for helpful discussions.

## References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium of Information Theory*, B. N. Petrov and F. Csaki, eds. Akademia Kiado, Budapest, 267–281.
- Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transaction on Auto. Control.* **19**, 716–723.
- Amemiya, T. (1980). Selection of regressors. *International Economic Review.* **21**, 331–354.
- Andrews, D. F. and Herzberg, A. M. (1985). *Data*. Springer-Verlag, New York. pp. 413–421.
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society Series A.* **160**, 268–282.
- Boos, D. and Brownie, C. (1989). Bootstrap methods for testing homogeneity of variances. *Technometrics.* **31**, 69–82.

- Box, G. E. P. and Andersen, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society Series B*. **17**, 1–26.
- Daniel, C. and Wood, F. (1971). *Fitting Equations to Data*. New York: Wiley.
- Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. 3rd edition, New York: Wiley. p. 300.
- Gorman, J. W. and Toman, R. J. (1966). Selection of variables for fitting equations to data. *Technometrics*. **8**, 27–51.
- Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*. **32**, 1–49.
- Kennard, R. W. (1971). A note on the  $C_p$  statistic. *Technometrics*. **13**, 899–900.
- Levene, H. (1960). Robust tests for equality of variances. in I. Olkin, Ghurye, S. G., Hoffding, W., Madow, W. G. and Mann, H. B. eds., *Contributions to Probability and Statistics*. Stanford: Stanford University Press, 278–292.
- Lim, T.-S. and Loh, W.-Y. (1996). A comparison of tests of equality of variances. *Computational Statistics and Data Analysis*. **22**, 287–301.
- Mallows, C. L. (1964). Choosing variables in a linear regression: A graphical aid. *Presented at the Central Regional meeting of the Institute of Mathematical Statistics*, Manhattan, Kansas
- Mallows, C. L. (1973). Some comments on  $C_p$ . *Technometrics*. **15**, 661–675.
- Rawlings, J. O., Pantula, S. G. and Dickey, D. A. (1998). *Applied Regression Analysis*. New York: Springer-Verlag.