The Statistician (1996)
45, *No.* 1, *pp.* 49–56

General Papers

The interpretation of Mallows's C_p-statistic

By STEVEN G. GILMOUR†

University of Reading, UK

[Received October 1994. Revised July 1995]

SUMMARY

When selecting variables in multiple-regression studies, the model with the lowest value of Mallows's C_p-statistic is often chosen. It is shown here that when the estimate of σ^2 comes from the full model an adjusted C_p, $\overline{C}_p$, has the property that $E(\overline{C}_p) = p$. It is suggested that a procedure be adopted which involves testing whether the model with minimum $\overline{C}_p$ is really better than a simpler model. Tables approximating the null distribution of the test statistics are given.

Keywords: Multiple regression; Multivariate F-distribution; Variable selection

1. The problem

Many procedures are available for selecting a subset of a set of k candidate regressors in multiple linear regression problems. One of the commonly used methods is to perform all possible regressions and to compare the results on the basis of Mallows's C_p-statistic. For a particular model with p parameters

$$C_p = \frac{\mathrm{SSE}_p}{\hat{\sigma}^2} - n + 2p, \tag{1}$$

where SSE_p is the error sum of squares from the model being considered, $\hat{\sigma}^2$ is an estimate of the error variance, σ^2, and n is the number of observations. The mean-square error (MSE) from the full model is often used as the estimate of σ^2. The standard texts, such as Draper and Smith (1981), Montgomery and Peck (1992) and Myers (1992), recommend plotting C_p against p for all possible regressions and choosing an equation with low C_p or with C_p close to p. If σ^2 is known, any model which provides unbiased estimates of the regression coefficients, i.e. which contains all important regressors, has $E(C_p) = p$.

Fig. 1 shows a plot of C_p against p for a set of 24 observations, given in Table 1, originally published by Narula and Wellington (1977), which is used to relate nine variables, $x_1, \ldots, x_9$, to the sale price, y, of houses. Only models with $C_p < 15$ are shown. From this plot the three-parameter model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{2}$$

would probably be chosen since it has the lowest C_p and looks clearly better than the two-parameter model

$$y = \beta_0 + \beta_1 x_1. \tag{3}$$

†*Address for correspondence*: Department of Applied Statistics, Harry Pitt Building, University of Reading, Whiteknights Road, PO Box 240, Reading, RG6 6FN, UK.
E-mail: S.G.Gilmour@reading.ac.uk

 0039–0526/96/45049

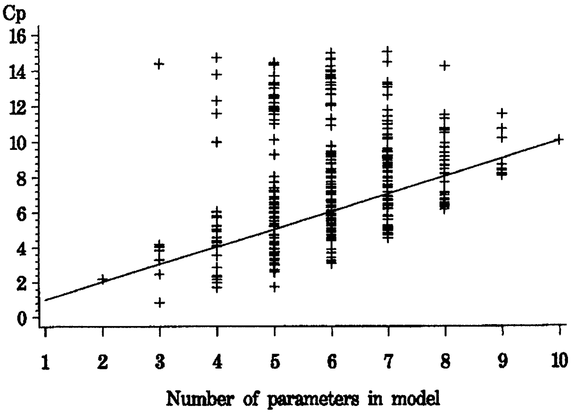

Fig. 1. C_p *versus* p for the house price data

One pattern in the above plot is the large number of models with $C_p < p$. This pattern can be observed for most data sets with a reasonably large number of regressors when many of them are unimportant. The comments in the above reference works are not very helpful. Draper and Smith (1981) say

> 'because of random variation, points representing well-fitting equations can also fall below the $C_p = p$ line'.

Myers (1992) says

> 'Since the residual mean square for the complete model need not be the smallest estimate of σ^2 among those for the candidate models, it is quite possible that the equation will yield a $C_p < p$ for a few of the candidate models'.

Montgomery and Peck (1992) say

> 'If the full model has several regressors that do not contribute significantly to the model, then MSE_{k+1} will often overestimate σ^2, and consequently the values of C_p will be small'.

Montgomery and Peck's comment is misleading. MSE_{k+1} will underestimate σ^2 slightly more often than it will overestimate it, as MSE_{k+1} is an unbiased estimator of σ^2, with a skewed distribution. The comments of Draper and Smith and Myers are true but miss the main point that, if there are several models of a particular size which all contain all the important regressors, then the model with the lowest C_p is very likely to have $C_p < p$.

Another point which is often ignored is that, if the MSE from the full model is used to estimate σ^2, the distribution of $\mathrm{SSE}_p/\hat{\sigma}^2$ can be obtained and gives an expected value of C_p which is not p. Both of these issues are addressed in this paper. In Section 2 the expected value of C_p is given and a modified statistic, $\overline{C}_p$, is defined. The correct interpretation of this statistic, allowing for the fact that there may be several models with p parameters, all of which allow unbiased estimation of parameters, is described in Section 3. The relationship of these results to other recent work on selection of variables is discussed in Section 4.

2. Modification of C_p

2.1. *Expected value of C_p*

The definition of the C_p-statistic was intended to ensure that, for a model including all important regressors, C_p had expected value p. Such a model will have

TABLE 1
House price data†

y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
25.9	4.9176	1.0	3.4720	0.9980	1.0	7	4	42	0
29.5	5.0208	1.0	3.5310	1.5000	2.0	7	4	62	0
27.9	4.5429	1.0	2.2750	1.1750	1.0	6	3	40	0
25.9	4.5573	1.0	4.0500	1.2320	1.0	6	3	54	0
29.9	5.0597	1.0	4.4550	1.1210	1.0	6	3	42	0
29.9	3.8910	1.0	4.4550	0.9880	1.0	6	3	56	0
30.9	5.8980	1.0	5.8500	1.2400	1.0	7	3	51	1
28.9	5.6039	1.0	9.5200	1.5010	0.0	6	3	32	0
35.9	5.8282	1.0	6.4350	1.2250	2.0	6	3	32	0
31.5	5.3003	1.0	4.9883	1.5520	1.0	6	3	30	0
31.0	6.2712	1.0	5.5200	0.9750	1.0	5	2	30	0
30.9	5.9592	1.0	6.6660	1.1210	2.0	6	3	32	0
30.9	5.0500	1.0	5.0000	1.0200	0.0	5	2	46	1
36.9	8.2464	1.5	5.1500	1.6640	2.0	8	4	50	0
41.9	6.6969	1.5	6.9020	1.4880	1.5	7	3	22	1
40.5	7.7841	1.5	7.1020	1.3760	1.0	6	3	17	0
43.9	9.0384	1.0	7.8000	1.5000	1.5	7	3	23	0
37.5	5.9894	1.0	5.5200	1.2560	2.0	6	3	40	1
37.9	7.5422	1.5	5.0000	1.6900	1.0	6	3	22	0
44.5	8.7951	1.5	9.8900	1.8200	2.0	8	4	50	1
37.9	6.0831	1.5	6.7265	1.6520	1.0	6	3	44	0
38.9	8.3607	1.5	9.1500	1.7770	2.0	8	4	48	1
36.9	8.1400	1.0	8.0000	1.5040	2.0	7	3	3	0
45.8	9.1416	1.5	7.3262	1.8310	1.5	8	4	31	0

†y, sale price (\$/1000); x_1, taxes (\$/1000); x_2, number of baths; x_3, lot size (ft^2/1000); x_4, living space (ft^2/1000); x_5, number of garage stalls; x_6, number of rooms; x_7, number of bedrooms; x_8, age (years); x_9, number of fireplaces.

$$E(\mathrm{SSE}_p) = (n-p)\sigma^2. \tag{4}$$

If $\hat{\sigma}^2$ is a 'good' estimate of σ^2 then, approximately,

$$E(C_p) = \frac{(n-p)\sigma^2}{\sigma^2} - n + 2p = p. \tag{5}$$

However, if $\hat{\sigma}^2$ is the MSE from the full model, it is possible to work out the distribution of C_p.

It is shown in Appendix A that, for a model which includes all important regressors,

$$C_p = (k-p+1)F + 2p - k - 1, \tag{6}$$

where $F \sim F_{k-p+1,n-k-1}$, the F-distribution with $k-p+1$ and $n-k-1$ degrees of freedom. This result was given by Mallows (1973) and Hocking (1976) who did not, however, note the following implications for the expected value of C_p and the interpretation of the C_p-plots. Since

$$E(F) = \frac{n-k-1}{n-k-3}, \tag{7}$$

$$E(C_p) = (k-p+1)\frac{n-k-1}{n-k-3} + 2p - k - 1 \tag{8}$$

$$= p + \frac{2(k-p+1)}{n-k-3} \tag{9}$$

$$= \frac{(n-k-5)p + 2(k+1)}{n-k-3}. \tag{10}$$

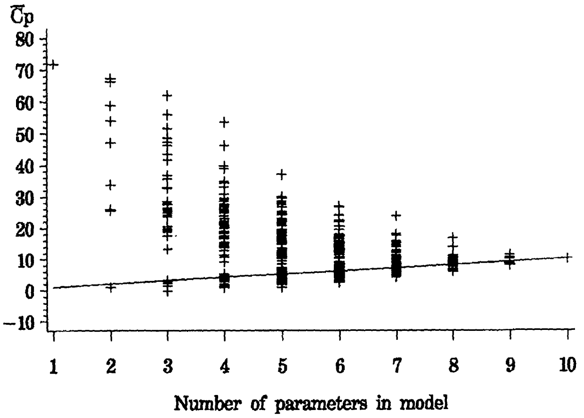

Fig. 2. $\overline{C}_p$ *versus* p for the house price data

This expectation can be considerably greater than p when $n - k$ is small, i.e. when the number of candidate regressors is not much fewer than the number of observations. For example, if there are 28 observations and 20 candidate regressors,

$$E(C_p) = \frac{3}{5}p + \frac{42}{5}.$$

In this case $E(C_1) = 9$, not 1, but $E(C_{21}) = 21$, as it should be. This is the general pattern. $E(C_p)$ is further from p for small values of p. Thus choosing the model with lowest C_p will tend to overfit, i.e. to suggest the inclusion of at least one unimportant regressor.

2.2. *$\overline{C}_p$-statistic*

Since the expected value of C_p for models which include all important regressors is not p, but $p + 2(k - p + 1)/(n - k - 3)$, if a plot is to be interpreted in the way described earlier, we should use not C_p but

$$\overline{C}_p = C_p - \frac{2(k - p + 1)}{n - k - 3}. \tag{11}$$

Clearly $E(\overline{C}_p) = p$, so plotting $\overline{C}_p$ against p should show models which include all important regressors falling near the line $\overline{C}_p = p$. Fig. 2 shows a plot of $\overline{C}_p$ against p for the introductory example on the house price data. It shows a large number of models with $\overline{C}_p$ close to p and others with $\overline{C}_p$ much greater than p.

3. Interpretation of $\overline{C}_p$

3.1. *Joint distribution of $\overline{C}_p$s*

In the previous section it was shown that using $\overline{C}_p$ instead of C_p ensured that the expected value of $\overline{C}_p$ for any particular model which includes all important regressors is p. It might then be thought that choosing the model with the minimum $\overline{C}_p$ would be a good selection criterion, as was originally envisaged with C_p. However, even this is likely to overfit.

Consider an artificial example, where there are 20 candidate regressors, five of which have large effects and 15 of which have no effect. Then *any* model containing the five important regressors provides unbiased estimates of the regression coefficients, but the model containing *only* the important regressors provides the lowest variances of the regression coefficients. There are 15 models with six regressors which provide unbiased estimates, but only one such model with five regressors. The 'correct' model has $E(\overline{C}_6) = 6$ and each of the 15 six-variable models has $E(\overline{C}_7) =$

7. It would not be surprising if at least one of these models had $\overline{C}_7 < 6$. Thus a model with more regressors than the best model is likely to have minimum $\overline{C}_p$.

In this example the important question is how small the minimum $\overline{C}_7$ from the 15 six-variable models is compared with what would be expected. This suggests using the distribution of $\min(\overline{C}_7)$, but it is more convenient to use the distribution of $\max(\overline{C}_6 - \overline{C}_7)$ where both models include all important regressors. In general, assume that there are $q-1$ important regressors from the k candidates. Consider $\overline{C}_q$ for the 'correct' model and $\overline{C}_{q+1}^{(i)}$ for the ith model with one redundant regressor. Then

$$\overline{C}_q = \frac{(n-k-1)s^2 + \mathrm{SS}(\beta_q, \ldots, \beta_k | \beta_1, \ldots, \beta_{q-1})}{s^2} - n + 2q - \frac{2(k-q+1)}{n-k-3}, \tag{12}$$

where $\mathrm{SS}(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)$ is the extra sum of squares for $\boldsymbol{\beta}_1$ allowing for $\boldsymbol{\beta}_2$, and

$$\overline{C}_{q+1}^{(1)} = \frac{(n-k-1)s^2 + \mathrm{SS}(\beta_{q+1}, \ldots, \beta_k | \beta_1, \ldots, \beta_q)}{s^2} - n + 2(q+1) - \frac{2(k-q)}{n-k-3}. \tag{13}$$

Partitioning the sum of squares, as in a standard analysis of variance, we obtain

$$SS(\beta_q, \ldots, \beta_k | \beta_1, \ldots, \beta_{q-1}) = \mathrm{SS}(\beta_{q+1}, \ldots, \beta_k | \beta_1, \ldots, \beta_q) + \mathrm{SS}(\beta_q | \beta_1, \ldots, \beta_{q-1}) \tag{14}$$

and so

$$\overline{C}_{q+1}^{(1)} = \overline{C}_q - \frac{S_1}{s^2} + \frac{2(n-k-2)}{n-k-3}, \tag{15}$$

where $S_1 = \mathrm{SS}(\beta_q \mid \beta_1, \ldots, \beta_{q-1})$. Similarly

$$\overline{C}_{q+1}^{(i)} = \overline{C}_q - \frac{S_i}{s^2} + \frac{2(n-k-2)}{n-k-3}, \tag{16}$$

$i = 1, \ldots, k-q+1$, where $S_i = \mathrm{SS}(\beta_{q-1+i} | \beta_1, \ldots, \beta_{q-1})$. Under the null hypothesis, $S_i \sim \sigma^2\chi_1^2$ $(i = 1, \ldots, k-q+1)$, $(n-k-1)s^2 \sim \sigma^2\chi_{n-k-1}^2$ and $S_1, \ldots, S_{k-q+1}$ and s^2 are all mutually independent.

Define

$$F_i = \overline{C}_q - \overline{C}_{q+1}^{(i)} + \frac{2(n-k-2)}{n-k-3}.$$

The F_i jointly have a multivariate F-distribution with a common denominator, having $n-k-1$ degrees of freedom, and independent numerators, each having 1 degree of freedom. In the interpretation of $\overline{C}_q$, interest is in the minimum $\overline{C}_{q+1}$, and hence in the maximum F_i. If this is below the critical value in the distribution of the maximum of $k-q+1$ random variables with a multivariate F-distribution, then the null hypothesis that all important regressors have been included in the model would be rejected.

The critical points of the distribution of the maximum F_i were tabulated for a few special cases by Finney (1941) and an approximation was recommended for the general problem. This approximation, however, does not work well when the numerators are based on 1 degree of freedom, the case which must be considered here. A much better approximation to the p-value can be obtained by simulating from the null distribution of the F_i. This is easily done by simulating from the null distributions of $S_1, \ldots, S_{k-q+1}$ and $(n-k-1)s^2$ which are all independent. Table 2 gives the 10%, 5% and 1% points of this distribution.

3.2. *Example*

To illustrate the procedure, we shall analyse the house price data. Fig. 2 shows a plot of $\overline{C}_p$ against p and Fig. 3 shows the same for models with $p \leqslant 5$ and $\overline{C}_p < 10$. The model with the lowest $\overline{C}_p$ is the three-parameter model with x_1 and x_2. The model with only x_1 has a slightly higher

TABLE 2
Critical points in distribution of the maximum of r random variables, each having an F-distribution with a common denominator having t degrees of freedom and independent numerators each having 1 degree of freedom

$t = n - k - 1$	*% points for the following values of $r = k - q + 1$:*									
	1	*2*	*3*	*4*	*6*	*8*	*10*	*15*	*20*	*30*
10% point										
4	4.53	7.08	8.82	10.2	12.3	14.0	15.2	17.5	19.2	21.2
5	4.05	6.19	7.65	8.79	10.5	11.7	12.7	14.5	16.1	18.0
6	3.78	5.66	6.96	7.97	9.44	10.6	11.5	13.4	14.4	16.5
8	3.47	5.11	6.22	7.03	8.32	9.23	9.89	11.3	12.2	13.6
10	3.27	4.79	5.82	6.53	7.72	8.52	9.17	10.2	11.4	12.6
12	3.18	4.62	5.56	6.29	7.34	8.06	8.69	9.85	10.6	11.9
15	3.10	4.40	5.32	5.96	6.94	7.66	8.18	9.19	9.97	11.0
20	2.97	4.24	5.08	5.74	6.57	7.25	7.75	8.68	9.44	10.4
30	2.86	4.11	4.87	5.43	6.24	6.85	7.30	8.20	8.88	9.80
5% point										
4	7.70	11.4	14.0	16.01	19.0	21.5	23.3	26.6	29.0	32.0
5	6.61	9.55	11.5	13.01	15.4	17.2	18.7	21.1	23.1	25.8
6	5.99	8.43	10.2	11.51	13.5	15.0	16.2	18.7	20.2	23.3
8	5.32	7.40	8.78	9.77	11.4	12.4	13.2	15.0	16.1	17.7
10	4.93	6.80	8.05	8.87	10.2	11.2	12.0	13.1	14.6	15.9
12	4.72	6.45	7.58	8.41	9.64	10.4	11.2	12.4	13.5	15.1
15	4.58	6.07	7.16	7.89	9.01	9.77	10.4	11.5	12.4	13.6
20	4.33	5.75	6.74	7.44	8.42	9.16	9.65	10.7	11.6	12.8
30	4.15	5.53	6.39	6.95	7.89	8.46	8.93	9.91	10.7	11.6
1% point										
4	21.1	29.6	35.7	40.6	47.0	52.8	57.5	65.3	72.4	75.0
5	16.4	22.3	26.0	29.0	33.8	37.5	40.1	44.2	48.9	53.6
6	13.6	18.0	21.4	23.5	27.0	29.5	32.6	37.4	38.8	44.7
8	11.1	14.5	16.7	18.3	20.9	22.2	24.0	25.4	29.2	30.6
10	9.91	12.6	14.4	15.8	17.7	19.1	20.1	21.9	23.8	26.2
12	9.37	11.5	13.1	14.5	15.9	17.1	18.0	20.1	21.3	23.8
15	8.72	10.7	12.2	13.1	14.4	15.2	16.2	17.4	18.6	21.1
20	8.04	10.0	11.0	11.1	13.2	14.1	14.8	15.6	17.0	18.5
30	7.55	9.27	10.1	10.8	11.9	12.6	13.1	14.1	15.0	16.0

$\overline{C}_p$, so here the hypothesis that $\beta_2 = 0$, i.e. that x_2 has no explanatory power, can be tested.

For these two models, $\overline{C}_2 = 0.8474$ and $\min(\overline{C}_3) = -0.3410$, so that

$$\max(F_i) = 0.8474 + 0.3410 + \frac{2 \times 13}{12} = 3.355.$$

Comparing this with the tables for $r = 8$ and $t = 14$, it can be seen that there is little evidence that x_2 contributes anything to the regression.

The model with only the intercept has $\overline{C}_1 = 71.83$ and so

$$\max(F_i) = 71.83 - 0.8474 + \frac{13}{6} = 73.15.$$

Comparing this with the tables for $r = 9$ and $t = 14$, there is very strong evidence that x_1 is an important explanatory variable. Therefore the most appropriate model appears to be the model containing only the intercept and x_1.

Fig. 3 illustrates very clearly the problem with simply choosing the model with lowest $\overline{C}_p$. The eight models including x_1 and one other regressor have the following values of $\overline{C}_p$:

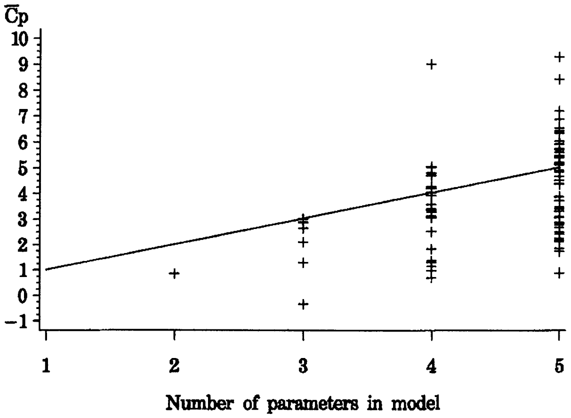

Fig. 3. $\overline{C}_p$ *versus p* for the house price data

$$-0.341 \quad 1.268 \quad 2.087 \quad 2.620 \quad 2.636 \quad 2.849 \quad 2.866 \quad 2.985.$$

These are scattered randomly about their expected value, given $\overline{C}_2$ from the model with only x_1, which is 1.8474 and with a skewed (F-) distribution. So, although the lowest $\overline{C}_p$ is quite low, the distribution of the eight points is no different from what would be expected if none of the variables other than x_1 has an effect.

4. Discussion

The C_p-statistic was first proposed by Mallows (1964) and first published by Gorman and Toman (1966). They noted that a 'good' estimate of σ^2 had to be obtained and suggested that the MSE from the full model be used. Mallows (1973) first noted the distributional implications of using the MSE from the full model and Hocking (1976) discussed the relationship between these and distributional results for other statistics. None of these researchers noted the implications for the interpretation of C_p discussed in this paper.

The results in this paper show that an improved interpretation of the C_p-statistic can be made by adjusting it and then interpreting the plots more conservatively than is usual. This avoids the overfitting which is almost inevitable when there is a large number of candidate regressors, many of which do not have any explanatory power. An interactive approach to variable selection is often useful; for example some of the variable-by-variable interactions can be added to the model to see what difference they make. The final choice of model will depend on subject-matter knowledge, as well as conclusions drawn from the $\overline{C}_p$-plot. However, the results in this paper show that overfitting is always likely if a model is chosen mainly because it has a low $\overline{C}_p$. It is often worthwhile to perform a hypothesis test like that described in Section 3 to decide whether there is really much evidence that a candidate model is better than a more parsimonious model.

Ronchetti and Staudte (1994) defined a robust version of C_p, which also has $\hat{\sigma}^2$ as a divisor and so may be expected to behave similarly to C_p. However, the numerator in their statistic is a weighted residual sum of squares, the weights being calculated from the data, so it is not possible to work out distributional results.

Much other recent work on selection of variables procedures has involved studying the bias induced in the estimated parameters by the selection procedure. This work was described by Miller (1990). Because $\overline{C}_p$ selects different subsets from C_p the selection bias will not be the same. However, selection bias still exists and can be studied in the same way for $\overline{C}_p$ as for C_p, for example by bootstrapping the residuals from a fitted model.

In conclusion, $\overline{C}_p$ should be plotted instead of C_p when the estimate of error is the MSE from the full model. Then each model containing all important regressors has expected value p. However, a model should not be chosen simply because it has the lowest $\overline{C}_p$. Instead a test should be performed to check that this model is indeed better than a simpler model with fewer regressors.

Acknowledgements

The author would like to thank Robert Curnow and Roger Mead for helpful discussions on this work and two referees whose comments led to a substantial improvement in the paper.

Appendix A: distribution of C_p

When the residual mean square from the full model is used as the estimate of σ^2,

$$C_p = \frac{\mathrm{SSE}_p}{\mathrm{SSE}_{k+1}/(n-k-1)} + 2p - n. \tag{17}$$

To obtain the distribution of C_p in the case where all important regressors are included in the model, assume, without loss of generality, that $\beta_p = \ldots = \beta_k = 0$, i.e. that $x_p, \ldots, x_k$ are unimportant regressors. Then

$$C_p = (n-k-1)\frac{\mathrm{SSE}_{k+1} + \mathrm{SS}(\beta_p \ldots \beta_k | \beta_0 \ldots \beta_{p-1})}{\mathrm{SSE}_{k+1}} + 2p - n \tag{18}$$

$$= (n-k-1)\left\{1 + \frac{\mathrm{SS}(\beta_p \ldots \beta_k | \beta_0 \ldots \beta_{p-1})}{\mathrm{SSE}_{k+1}}\right\} + 2p - n \tag{19}$$

$$= (k-p+1)\frac{\mathrm{SS}(\beta_p \ldots \beta_k | \beta_0 \ldots \beta_{p-1})/(k-p+1)}{\mathrm{SSE}_{k+1}/(n-k-1)} + 2p - k - 1 \tag{20}$$

$$= (k-p+1)\frac{U/(k-p+1)}{V/(n-k-1)} + 2p - k - 1 \tag{21}$$

where $U \sim \chi^2_{k-p+1}$, $V \sim \chi^2_{n-k-1}$ and U and V are independent. Hence

$$C_p = (k-p+1)F + 2p - k - 1$$

where $F \sim F_{k-p+1, n-k-1}$.

References

Draper, N. R. and Smith, H. (1981) *Applied Regression Analysis*, 2nd edn. New York: Wiley.

Finney, D. J. (1941) The joint distribution of variance ratios based on a common error mean square. *Ann. Eugen.*, **11**, 136–140.

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.

Mallows, C. L. (1964) Choosing variables in a linear regression: a graphical aid. *Cent. Regl Meet. Institute of Mathematical Statistics, Manhattan, May 7th–9th.*

——— (1973) Some comments on C_p. *Technometrics*, **15**, 661–675.

Miller, A. J. (1990) *Subset Selection in Regression*. London: Chapman and Hall.

Montgomery, D. C. and Peck, E. A. (1992) *Introduction to Linear Regression Analysis*, 2nd edn. New York: Wiley.

Myers, R. L. (1992) *Classical and Modern Regression Analysis*, 2nd edn. New York: Wiley.

Narula, S. C. and Wellington, J. F. (1977) Prediction, linear regression and minimum sum of relative errors. *Technometrics*, **19**, 185–190.

Ronchetti, E. and Staudte, R. G. (1994) A robust version of Mallows's C_P. *J. Am. Statist. Ass.*, **89**, 550–559.