Mallows C_p for Out-of-sample Prediction

Lawrence D. Brown
Statistics Department,
Wharton School, University of Pennsylvania
lbrown@wharton.upenn.edu

Purdue University, Nov. 11, 2016

Joint work with others in the **W**harton **L**inear **M**odels **R**esearch **G**roup:
including R. Berk, A. Buja, A. Kuchibotla and L. Zhao.

Linear Model

- "Conventional" Linear Model

(LM)
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e};\ \mathbf{Y}\ \&\ \mathbf{e} \sim n \times 1,\ \mathbf{X} \sim n \times r,\ \boldsymbol{\beta} \sim r \times 1$$
$$E(\mathbf{e}_i) = 0,\ \mathbf{e}_i \text{ independent},\ \mathrm{var}(\mathbf{e}_i) = \sigma^2 \text{ (unknown)}.$$

- Variable selection:

  Choose a subset of variables=columns of $\mathbf{X} = \mathbf{X}_{[p]}$.
  Reanalyze via Least Squares using $\mathbf{X}_{[p]}$ under (LM).
  Get $\hat{\boldsymbol{\beta}}_{[p]}$ and $\hat{\mathbf{Y}}_{[p]} = \mathbf{X}_{[p]}\hat{\boldsymbol{\beta}}_{[p]}$.

- C_p is a numerical measure often used to choose a good subset of variables.

Here **X** is a matrix of constants (not random). The first column of **X** is **1**. So, **ordinary regression is *p*=2**.

C_p

Mallows version for a sub-model of size p is

$$C_p = \left(SSE_{[p]} / \hat{\sigma}_r^2 \right) - n + 2p.$$

Note that $\hat{\sigma}_r^2$ comes from the full model.

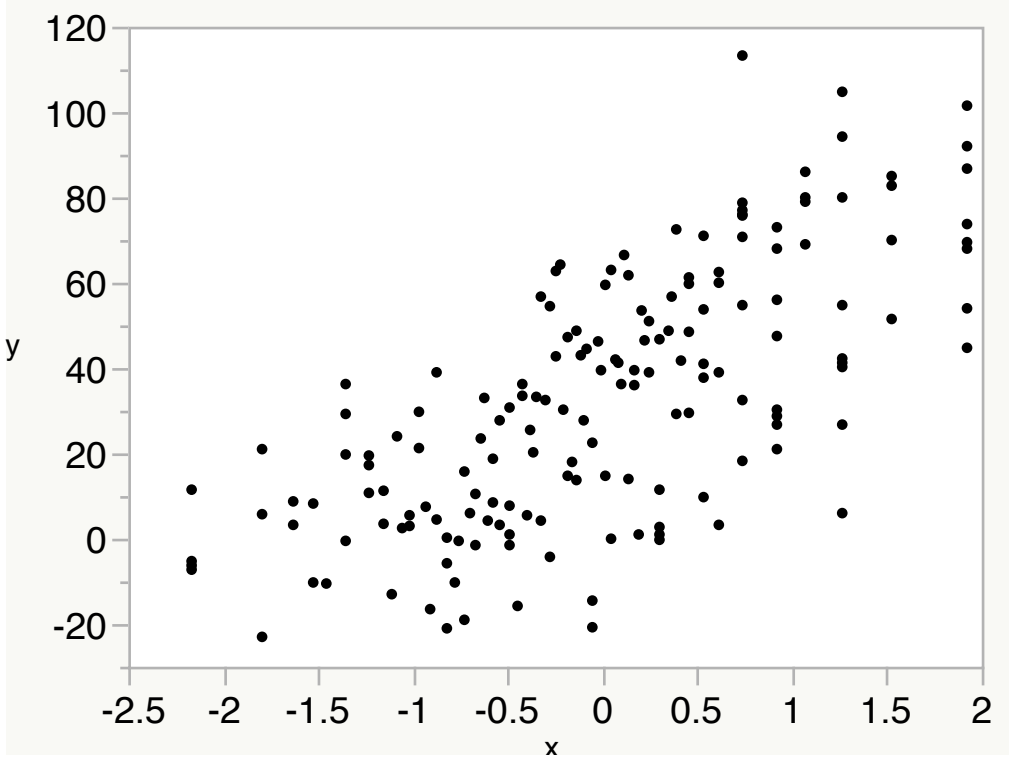
This is used to compare the suitability of various sub-models.

One can use either an all subsets approach or a stepwise approach (just forward, or forward and backward, etc.).

Mallows (1964, 1966, oral presentations; 1973 *Technometrics*)
Gorman and Toman (1966, *Technometrics*)

Example: Polynomial Regression

Cholesterol data from Efron&Feldman (1991), Efron (2013, *JASA*)).



Y is improvement in cholesterol level over duration of study

X is a measure of compliance during study (normalized to approx. normal)

Stepwise Regression

Entered in order of degree of polynomial [Notation: X^k $= \left(X - \bar{X}\right)^k$]

Parameter	"Sig Prob"	RSquare	Cp	p	AICc	BIC
X	0.0000	0.4853	**3.137**	2	1477.3	1486.5
X^2	0.1988	0.4905	**3.467**	3	1477.7	1489.9
X^3	0.0816	0.5001	**2.427**	4	1476.7	1491.8
X^4	0.7572	0.5004	**4.331**	5	1478.7	1496.8
X^5	0.2654	0.5043	**5.089**	6	1479.7	1500.7
X^6	0.7656	0.5046	**7**	7	1481.8	1505.7

C_p suggests that the cubic model is best.

Note that all-subsets chooses linear and cubic; no quadratic term.

Alternate, equivalent form of C_p

Recall,

$$C_p = \left(SSE_{[p]} / \hat{\sigma}_r^2 \right) - n + 2p.$$

A more convenient form moving forward is

$$C_p^U = \frac{\hat{\sigma}_r^2}{n}\left(C_p + n \right) = \frac{SSE_{[p]}}{n} + 2p\frac{\hat{\sigma}_r^2}{n}.$$

Both select the same models.

C_p has a nice self-normalizing property: $C_{[r]} = r$.

But C_p^U has an unbiasedness property that resonates well later.

Continue to note that $\hat{\sigma}_r^2$ is derived from **full model** residuals.

Example: Variable Selection

PROSTATE data from Tibshirani (1996, JRSS-B),
Data adapted from Stamey, et.al (1989)

- Response is Y = "lcavol" (=log CancerVolume)

There are 8 potential explanatory variables. n = 97

- We used all-subsets with C_p^U. But I'll display as a step-wise table. In this example step-wise and all-subsets yielded the same models.

Note: I switched variables from the analysis in Tibshirani (1996); he used Y=lpsa and lcavol was one of the potential covariates. We use full data set; Tibshirani (1996) splits the data and uses only n = 71.

Step-wise Table

Parameter	"Sig Prob"	Seq SS	RSquare	Cp	C_p^U	AICc
constant	0.000	176.785		176.84	1.977	310.3
lpsa	0.000	71.938	0.539	32.20	0.643	237.2
lcp	0.000	14.143	0.646	5.37	0.508	214.0
age	0.151	1.040	0.653	5.25	0.507	214.1
lbph	0.098	1.359	0.664	**4.48**	**0.503**	213.4
pgg45	0.283	0.567	0.667	5.32	0.507	214.5
gleason	0.162	0.957	0.675	5.37	0.508	214.8
svi	0.549	0.175	0.676	7.01	0.516	216.8
lweight	0.903	0.007	0.676	9.00	0.526	219.3

- C_p chooses a 4-factor model (same via step-wise, as above, or via all subsets)
- I'll later return to this example.

[NOTE: I included the "constant" as a possible model. This corresponds to the model $Y = \beta_1$. This is not traditional in such a table.]

Estimation of Predictive Risk

- **The core concept within Mallows method**

Mallows noted:

- C_p is a [normalized] estimate of [excess] predictive risk

and also explicitly assumed

- The x's are taken to be from a **fixed design**, not from a random sample.

Details:

- Consider **X,** fixed, with linear estimator $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and a new (vector) observation $\mathbf{Y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}^*$.
- The Total predictive Risk across the design is

$$\mathrm{TR}_{|\mathbf{X}} \triangleq E\left(\left\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2\right).$$

The average, per-coordinate, is $\mathrm{R}_{|\mathbf{X}} = n^{-1}\mathrm{TR}_{|\mathbf{X}}$.

- THEN, as implicit in Mallows,

(U) $$E\left(\mathrm{C}_p^U\right) = \mathrm{R}_{|\mathbf{X}}. \quad \text{[Exact unbiasedness]}$$

- Mallows has

(M) $$\mathrm{C}_p \approx \left(\mathrm{TR}_{|\mathbf{X}} - n\sigma^2\right)\Big/\sigma^2 = \left(\text{Excess total pred risk}\right)\Big/\sigma^2.$$

See Gilmour (1996, *JRSS-D*) for a slight modification.

- Because of (M)

  Many statisticians and statistical textbooks believe/claim that

"C_p is a suitable estimate of predictive risk." !!!

BUT

That's not true!

- **Here's why it's not true...** (one reason)
- The predictive goal is to predict (squared-error) risk for an individual who is **not** in the statistical sample.
- This is (with X^*, Y^* denoting the new measurements)

(predRisk) $$R_{X^*_{[p]}} \triangleq E\left\{\left(Y^* - X^{*\mathrm{T}}_{[p]}\hat{\beta}\right)^2\right\}.$$

- This is not the same as

(Mallows) $$\mathrm{R}_{|\mathbf{X}_{[p]}=\mathrm{X}} \triangleq n^{-1}E\left(\left\|\mathbf{Y} - \mathrm{X}\hat{\boldsymbol{\beta}}\right\|^2 \middle| \mathbf{X}_{[p]} = \mathrm{X}\right),$$

which is the target of C_p^U.

- Why are these two different? Synopsis follows. & see

"Models as Approximations -- A Conspiracy of Random Regressors and Model Deviations Against Classical Inference in Regression" WLMRG,

Estimating the Predictive Risk, *R*

Preamble:

- C_p assumes a well-specified linear model with fixed regressors, **X** and homoscedasticity of residuals.
- The following is for random regressors; it estimates predictive risk of linear estimators in such a setting.
- In addition, the following does not assume a well-specified linear model or homoscedasticity. INSTEAD
- It holds in the "assumption lean" framework – (see next slide).
- There is a small price relative to what one gets with C_p:
- The new estimator of risk is only an asymptotically unbiased estimator of its target.

Assumption Lean Linear Regression

- Observe iid Sample $\{X_i, Y_i\}$ with $X_i \in \Re^r$, $X, Y \sim F$.
- No assumptions about F, other than existence of low order moments.
- Contemplate **a** future observation $X^*, Y^* \sim F$ and
- Construct best bi-linear predictor to minimize predictive risk, R. [Predictor of the form $X^{*\mathrm{T}}\tilde{\beta}$ with $\tilde{\beta}$ linear in **Y**.]
- This is $X^{*\mathrm{T}}\hat{\beta}$ with $\hat{\beta}$ the usual LS estimator.
- Notation: The target/oracle predictor is $X^{*\mathrm{T}}\beta$

with β being the population LS parameter:

$$\beta = \arg\min_b E_F\left\{\left(Y - X^{\mathrm{T}}b\right)^2\right\} \Leftrightarrow$$

$$\beta = \left[E\left(XX^{\mathrm{T}}\right)\right]^{-1} E\left(XY\right).$$

Estimate of $R_{[\Gamma]}$; Definition

- For a submodel, Γ, compute the sample LS residuals

$$\hat{\rho}_{[\Gamma]i} = Y_i - X'_{[\Gamma]i}\hat{\beta}_{[\Gamma]},\ i = 1,..,n.$$

- Then compute the matrices

$$\hat{\mathbf{M}}_{[\Gamma]} = n^{-1}\mathbf{X}'_{[\Gamma]}\mathbf{X}_{[\Gamma]} \text{ and } \hat{\mathbf{W}}_{[\Gamma]} = n^{-1}\sum_{i=1}^{n} X_{[\Gamma]i}X'_{[\Gamma]i}\hat{\rho}^2_{[\Gamma]i}.$$

- Estimate of $R_{[\Gamma]}$, the predictive risk for sub-model Γ is

$$C_p^{\oplus} = n^{-1}\left\|\mathbf{Y} - \mathbf{X}_{[\Gamma]}\hat{\beta}_{[\Gamma]}\right\|^2 + 2n^{-1}\mathrm{tr}\left(\hat{\mathbf{M}}^{-1}_{[\Gamma]}\hat{\mathbf{W}}_{[\Gamma]}\right)$$

$$= n^{-1}SSE_{[\Gamma]} + 2n^{-1}\hat{\varsigma}^2_{[\Gamma]}$$ (See derivation on next pages)

- Compare this to $C_p^U = n^{-1}SSE_{[\Gamma]} + 2n^{-1}\left(p\hat{\sigma}^2_r\right)$ --

The difference is the last term, which estimates variance.

Derivation

- Follows the pattern for derivation of C_p.
- Uses Δ method and weak LLN (or CLT),
- And the Sandwich estimator.
- Let $\approx$ denote suitable asymptotic approximation.
- Let $\hat{\Psi}_{[\Gamma]}$ denote the usual sandwich estimator for the covariance matrix of $\hat{\beta}_{[\Gamma]}$ (assumption-lean setting):

$$\hat{\boldsymbol{\Psi}}_{[\Gamma]} = \hat{\mathbf{M}}_{[\Gamma]}^{-1} \hat{\mathbf{W}}_{[\Gamma]} \hat{\mathbf{M}}_{[\Gamma]}^{-1}.$$

- Then (suppressing the subscript $_{[\Gamma]}$ for convenience)

Derivation, continued:

$$R \triangleq E\left(Y^* - X^{*\mathrm{T}}\hat{\boldsymbol{\beta}}\right)^2 = E\left(Y^* - X^{*\mathrm{T}}\boldsymbol{\beta}\right)^2 + E\left(X^{*\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)^2$$

$$\approx E\left(Y^* - X^{*\mathrm{T}}\boldsymbol{\beta}\right)^2 + E\left(X^{*\mathrm{T}}\hat{\Psi}X^*/n\right) \qquad \text{(Sandwich)}$$

$$\approx n^{-1}\left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 + n^{-1}tr\left(\hat{\Psi}\mathbf{X}^{\mathrm{T}}\mathbf{X}/n\right) \qquad \text{(Empirical moment)}$$

$$\approx n^{-1}\left\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2 + n^{-1}tr\left(\hat{\Psi}\mathbf{X}^{\mathrm{T}}\mathbf{X}/n\right) + n^{-1}tr\left(\hat{\Psi}\mathbf{X}^{\mathrm{T}}\mathbf{X}/n\right)$$

$$= n^{-1}\left\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2 + 2n^{-1}tr\left(\hat{\mathbf{M}}\hat{\mathbf{W}}\right) \triangleq C_p^{\oplus}$$

end of derivation.

Asymptotic Results

Tracking the error terms in the preceding derivation (and being precise about [benign] assumptions on existence of moments) yields –

T1: $C_p^{\oplus}$ is asymptotically unbiased in the sense that

$$E\left(C_p^{\oplus}\right) = R + O\left(1/n\right).$$

T2: $C_p^{\oplus}$ has considerable asymptotic variability about its mean. To be precise,

$$C_p^{\oplus} = R + O_P\left(1/\sqrt{n}\right).$$

- T2 is disappointing. Here's why

A Disappointment

T2 Says $$C_p^{\oplus} = R + O_P\left(1/\sqrt{n}\right).$$

- $C_p^{\oplus}$ can be written: $C_p^{\oplus} = n^{-1}SSE_{[\Gamma]} + 2n^{-1}\hat{\varsigma}_{[\Gamma]}^2$.
- 1st term is an underestimate of R. The 2nd term is a correction meant to fix the optimism in 1st term.
- According to T2 the second term is asymptotically negligible in comparison to natural randomness in $C_p^{\oplus}$.
- Hence **if the goal** of $C_p^{\oplus}$ **were** to estimate R, then the 2nd term is asymptotically useless, and one might just as well use $n^{-1}SSE_{[\Gamma]}$.
- Analogous comments are equally true about C_p^U. Hence (asymptotic) unbiasedness is irrelevant for estimation of R.

Comparison of Sub-Models

- In practice, the use of Mallows' type measures is to compare two (or more) sub-models.
- Let Γ_1, Γ_2 denote two designated sub-models whose predictive power is to be compared in an analysis. Then

T3: $$R_{[\Gamma_2]} - R_{[\Gamma_1]} = C^{\oplus}_{[\Gamma_2]} - C^{\oplus}_{[\Gamma_1]} + O_P\left(1/n\right).$$

- The stochastic error in the comparison of two sub-models is of the same order as the second terms in the definitions of $C^{\oplus}_p$. Thus those terms may improve the estimate of comparative predictive risks.
- But they do not guarantee success.
- Unfortunately, nothing can do so! (**T4**.)

$C_p^{\oplus}$ Does Not Guarantee Successful Comparisons

not even asymptotically

- Suppose $Y_i = a + \varepsilon_i$, $\varepsilon_i \sim N(0,1)$, indep,
- Compare two models: $\Gamma_1 : Y = \beta_1$ & $\Gamma_2 : Y = \beta_1 + \beta_2 X$.
- Then Γ_1 is "correct" and yields the better predictions.
- BUT, as $n \to \infty$

$$P\left(C_{\Gamma_2}^{\oplus} < C_{\Gamma_1}^{\oplus}\right) \to P\left(\chi_1^2 > 2\right) = 0.157$$

$$= P(\text{wrong choice})$$

- This is how far from perfectly $C_p^{\oplus}$ does. C_p will do the same since this is the conventional setup.

Can C_p (or C_p^U) and $C_p^{\oplus}$ give different model choices?

- They can! But in well-behaved examples they often do not.
- Here are two moderately well-behaved real data examples to illustrate this;
- followed by a third example, from a simulation, that demonstrates a more extreme case.

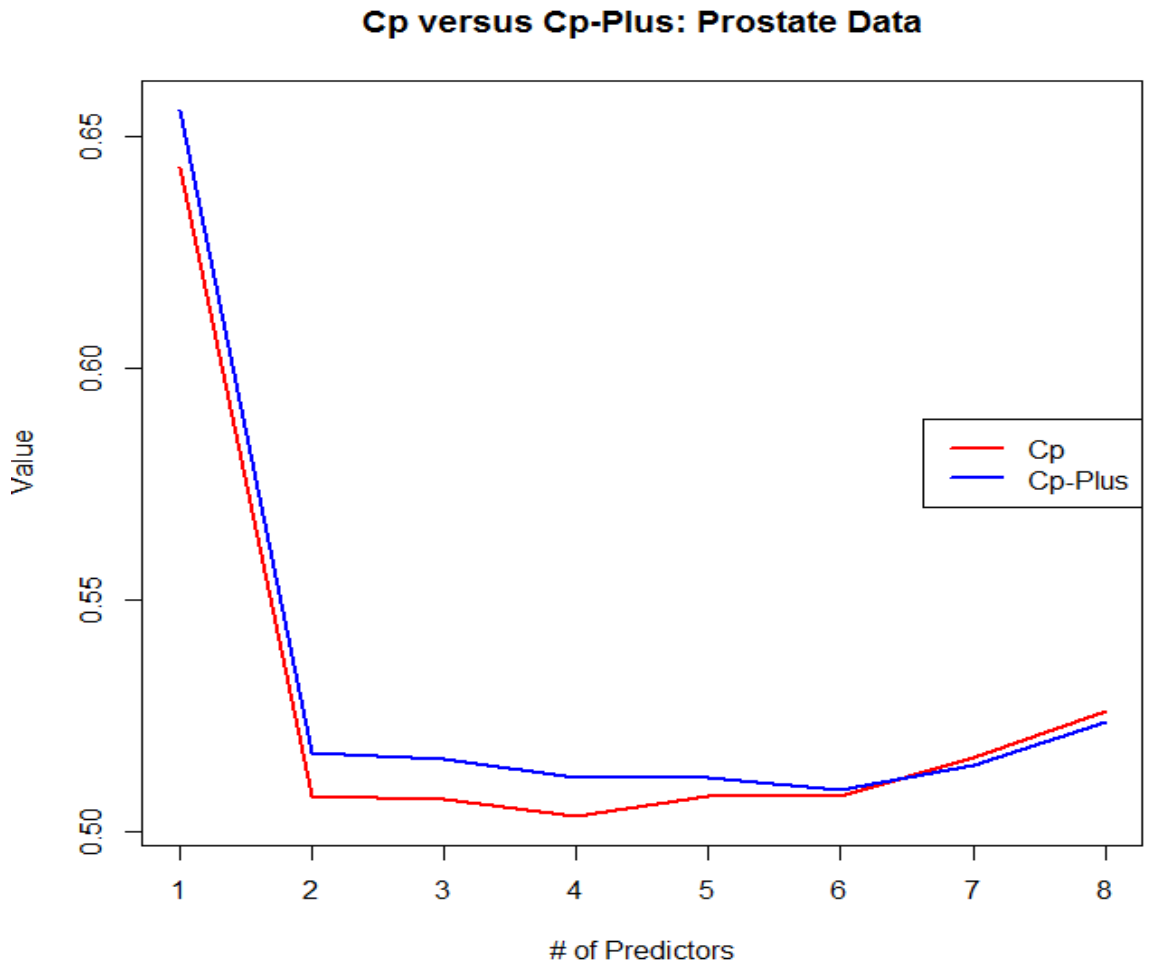
Example: Prostate Data from Tibshirani (1996)

as discussed previously

Stepwise and All-subsets yield same model choices. Here is a stepwise table comparing results from C_p^U and $C_p^{\oplus}$.

p	C_p^U	$C_p^{\oplus}$
[2] lpsa	0.6433	0.6557
[3] lcp	0.5076	0.5169
[4] age	0.5070	0.5155
[5] lbph	**0.5031**	0.5117
[6] pgg45	0.5074	0.5114
[7] gleason	0.5076	**0.5090**
[8] svi	0.5159	0.5143
[9] lwt	0.5259	0.5236

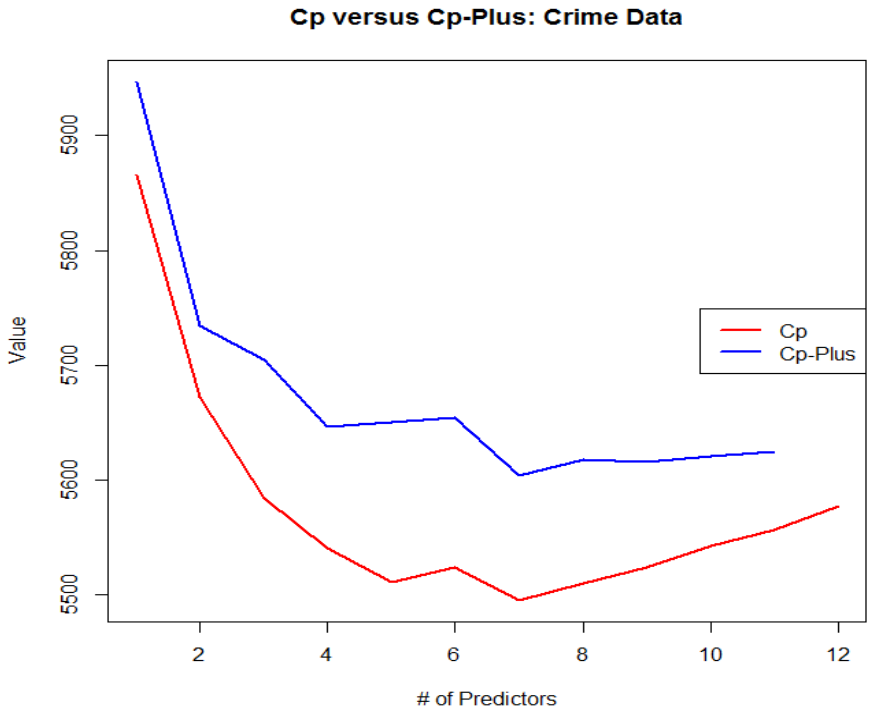
Scree plot ---


Cp versus Cp-Plus: Prostate Data
Value
0.50
0.55
0.60
0.65
1
2
3
4
5
6
7
8
# of Predictors
Cp
Cp-Plus

Criminal Sentencing

This data is analyzed in WLMRG (2015, "Conspiracy") and McCarthy, Zhang & WLMRG (2016, Double bootstrap). Used here is a random sample of size 500 from a much larger data-set collected and analyzed by R. Berk.

The Y variable is the length of criminal sentence. The covariates are demographic descriptors of the sentenced individuals & their prior criminal justice experience and the type and severity of the crime for which they are sentenced; $r = 14$. Here is the scree plot for all subsets regression, showing Mallows' type values for the best subset for each value of p.



Note the best subset size was 7 under both measures. The subsets were the same. The curves are each fairly horizontal. (Note the vertical scale.). If operating stepwise one *might* decide to use $p = 4$ under $C_p^{\oplus}$ and $p = 5$ under C_p . $C_p^{\oplus}$ is above C_p, as is typical but not necessary.

Simulation

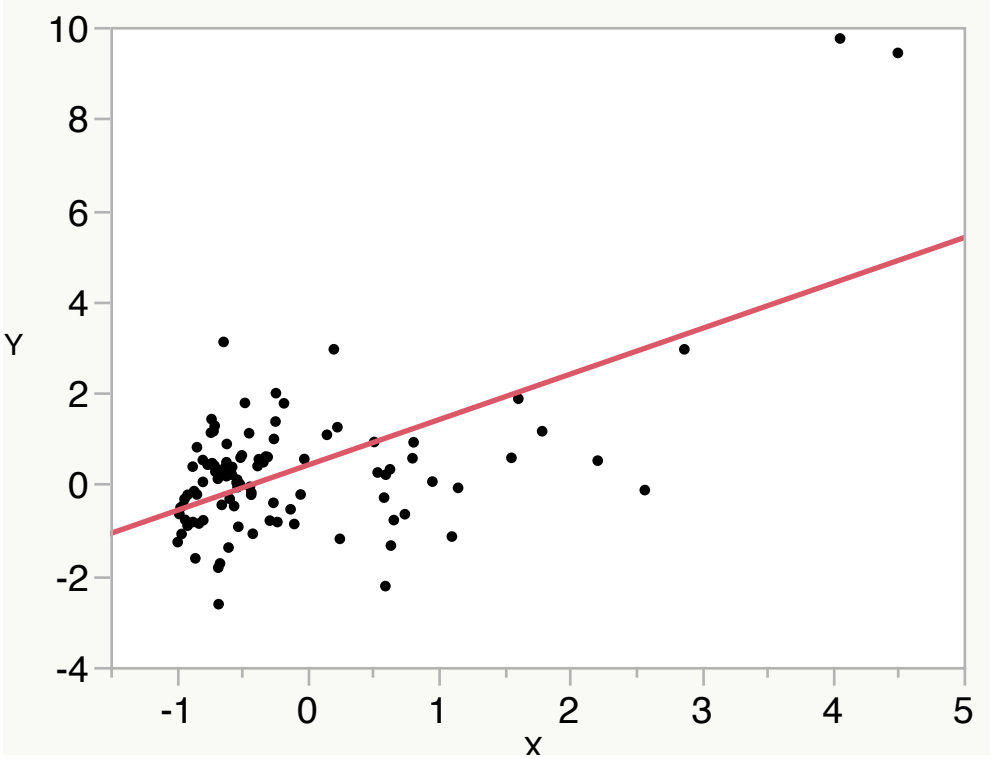
- In order to show what can happen in less well-behaved settings we simulated data from a particular joint distribution, *F*.
- It was chosen by trial and error to yield interesting results. See next page for a sample data set.
- The analytical models to compare were

$$\Gamma_1 : Y = b_1; \quad \Gamma_2 : Y = b_1 + b_2 X.$$

- [Neither model is a correct, perfect description of *F.*]
- Following is a scatterplot for a typical data set, followed by other info. You can see both the constant and linear models are mis-specified and there is strong leverage.

(Simulation had 4,000 replications.)

Typical scatterplot from simulation



Note the possible non-linearity and/or high leverage outlier.

(What do you think is the true distribution that generated this data?)

p	R_p	$\bar{C}_p$	$\bar{C}_p^{\oplus}$
1	**2.84**	2.84	2.85
2	3.26	2.28	2.72

Table 1: Actual value of the predictive squared risk and the averages of its estimates.

- Neither Mallows' type of estimate does a good average job of estimating R_2. (PS: Simulation error was nearly negligible.) But $C_2^{\oplus}$ does better on average than C_2.
- The better model as between the two is $p = 1$. And/But

  $C_1^{\oplus} < C_2^{\oplus}$ (as hoped) **52%** of the time, **while**

  $C_1 < C_2$ only **25%** of the time.
- SO, as theory predicts, $C_p^{\oplus}$ does better than C_p at comparing, though neither does sparklingly well.

Take aways:

- For observational data Mallows' C_p is an unjustified estimate of predictive risk. $C_p^{\oplus}$ is a valid estimate.
- Both have high noise level $\Rightarrow$ may not be very useful.
- Even for estimating Δ between sub-models $C_p^{\oplus}$ **has a high noise to signal ratio (as does C_p)**, and so may not give right answer in challenging situations.
- BUT at least $\hat{C}_p^{\oplus}$ aims at the correct target, even in assumption-lean settings. Also, it's almost as easy to compute as C_p.

Here's what Mallows wrote/said about C_p, and is equally true about $\hat{C}_p^{\oplus}$:

Mallows view of how C_p should – and should not – be used (1973):

“The discussion above does not lend any support to the practice of taking the lowest point on a Cp-plot as defining a "best" subset of terms. The present author feels that the greatest value of the device is that it helps the statistician to examine some aspects of the structure of his data and helps him to recognize the ambiguities that confront him. The device cannot be expected to provide a single "best" equation when the data are intrinsically inadequate to support such a strong inference.”...

“ ...the ambiguous cases where the "minimum Cp" rule will give bad results are exactly those where a large number of subsets are close competitors for the honor. With such data no selection rule can be expected to perform reliably.”

Further research in progress reveals non-trivial limits on the possibility of correctly identifying better and/or best models. It turns out that C_p, when applicable, and $C_p^{\oplus}$ more generally provide a (nearly) optimal statistic for such purposes if used properly.

Take-aways (cont.)

- As Andreas Buja likes to express this idea:

"Look at the Scree plot. It will typically fall steeply, sort of level out, and then may gradually rise. Keep the variables for which the plot is falling steeply, throw out all those after the plot has noticeably risen, ***and do what you want*** *with the variables in between. For a parsimonious and usually satisfactory model throw out all the variables after the steep fall."*

P.S.: It is possible to pursue further distribution theory about C_p and $C_p^{\oplus}$ to provide a statistical quantification as to what "*steeply*" and "*noticeably*" mean. We know how to do this for C_p. We're still working on this for $C_p^{\oplus}$, and to improve results for C_p.

end

Parenthetical note: General "conspiracy" theory suggests that if the covariates are random but the linear model is first and second order well-specified then fixed X analysis is correct. However that's not necessarily true with respect to the C_p goal of estimating predictive risk when using sub-models. Even for such a case one should use $C_p^{\oplus}$ or analogous measure. The reason is that the sub-model need not be well-specified. Thus, even though $E\left(Y\middle|X\right)$ is linear in the vector X it need not be the case that $E\left(Y\middle|X_{[p]}\right)$ is linear in the reduced model vector $X_{[p]}$.

PS in the final simulation

- The data are from

$$X \sim \exp(1) - 1,\ Y = b_1 + b_2 X + b_3 X^3 + Z,\ n=100.$$

$b_1 = b_2 = b_3 = 0.1$ (This choice gave interesting results.)