# Spring 2015 Statistics 151 (Linear Models) : Lecture Eighteen

Aditya Guntuboyina

02 April 2015

We are looking at criteria for variable selection. Last time we looked at adjusted $R^2$, AIC and BIC. Today we shall study Mallows' $C_p$. For this, we first need a simple formula for computing the expectation of quadratic forms.

## 1 Mallow's $C_p$

Next we look at Mallows's $C_p$. We need a formula for computing expectations of quadratic forms.

### 1.1 A Formula for computing expectations of quadratic forms

Suppose $Z$ is a random vector with mean $\mu$ and covariance matrix $\Sigma$. Then

$$\mathbb{E}(Z^T A Z) = tr(A\Sigma) + \mu^T A \mu. \tag{1}$$

To see how this is proved, write

$$\begin{aligned}\mathbb{E}(Z^T A Z) &= \mathbb{E}(Z-\mu+\mu)^T A (Z-\mu+\mu) \\ &= \mathbb{E}(Z-\mu)^T A (Z-\mu) + \mu^T A \mu + \mathbb{E}(Z-\mu)^T A \mu + \mathbb{E}\mu^T A (Z-\mu) \\ &= \mathbb{E}(Z-\mu)^T A (Z-\mu) + \mu^T A \mu.\end{aligned}$$

To evaluate $\mathbb{E}(Z-\mu)^T A(Z-\mu)$, we use the so called trace trick. Because $(Z-\mu)^T A(Z-\mu)$ is a scalar, it equals its trace.

$$\mathbb{E}(Z-\mu)^T A(Z-\mu) = \mathbb{E}tr\left((Z-\mu)^T A(Z-\mu)\right).$$

Since $tr(AB) = tr(BA)$, we get

$$\mathbb{E}(Z-\mu)^T A(Z-\mu) = \mathbb{E}tr\left(A(Z-\mu)(Z-\mu)^T\right).$$

Because $tr$ is a linear operator, we can bring it out of the expectation to obtain

$$\mathbb{E}(Z-\mu)^T A(Z-\mu) = tr\left(A\mathbb{E}\left((Z-\mu)(Z-\mu)^T\right)\right).$$

Now, the expectation of $(Z-\mu)(Z-\mu)^T$ is nothing but the covariance matrix of $Z$ which is $\Sigma$. We thus have

$$\mathbb{E}(Z-\mu)^T A(Z-\mu) = tr\left(A\Sigma\right).$$

This proves (1).

Before discussing Mallows's $C_p$, it makes sense to study a kind of biased estimation in the linear model.

## 1.2 Biased Estimation of $X\beta$

Consider the linear model $Y = X\beta + e$ with the usual assumptions on $e$: $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I$.

How does one estimate $X\beta$? The most natural estimator is of course $X\hat{\beta} = HY$. This is the least squares estimator. We have seen that this is BLUE (best linear unbiased estimator). Under the additional assumption of normality of the errors, this is the best among all unbiased estimators.

However, if biased estimators are allowed, then one might be able to do better than $X\hat{\beta}$.

Note that $X\hat{\beta} = HY$ is the vector of fitted values. One might try to obtain fitted values from a submodel of the original full model. Let us denote such a submodel by $m$ and let $X(m)$ denote its $X$-matrix. The hat matrix corresponding to $m$ submodel is $H(m) = X(m)(X(m)^T X(m))^{-1} X(m)^T$. Let us investigate the performance of $H(m)Y$ as an estimator of $X\beta$. A natural measure of the effectiveness of $H(m)Y$ as an estimator of $X\beta$ is

$$\mathbb{E}||X\beta - H(m)Y||^2 = \mathbb{E}(X\beta - H(m)Y)^T(X\beta - H(m)Y).$$

To calculate this, we use the formula (1) with $Z = X\beta - H(m)Y$ and $A = I$. The mean of $Z$ is

$$\mathbb{E}Z = X\beta - H(m)\mathbb{E}Y = X\beta - H(m)X\beta = (I - H(m))X\beta.$$

The covariance matrix of $Z$ is

$$Cov(Z) = Cov(X\beta - H(m)Y) = Cov(H(m)Y) = \sigma^2 H(m)H(m)^T = \sigma^2 H(m).$$

Therefore, from (1), we get

$$\mathbb{E}||H(m)Y - X\beta||^2 = \sigma^2 tr(H(m)) + \beta^T X^T (I - H(m))X\beta.$$

The trace of $H(m)$ equals the rank of $X(m)$ which equals the number of parameters in $X(m)$. If intercept is included, then $tr(H(m)) = 1 + p(m)$. This therefore gives

$$\mathbb{E}||H(m)Y - X\beta||^2 = \sigma^2 (1 + p(m)) + \beta^T X^T (I - H(m))X\beta. \tag{2}$$

When $m$ equals the full model $M$, we obtain

$$\mathbb{E}||HY - X\beta||^2 = \sigma^2(1 + p) + \beta^T X^T (I - H)X\beta = \sigma^2(1 + p) \tag{3}$$

because $HX = X$.

Comparing (2) with (3), we see that $H(m)Y$ is a better estimator of $X\beta$ than $HY$ provided

$$\beta^T X^T (I - H(m))X\beta < \sigma^2(p - p(m)).$$

## 1.3 Definition of Mallows's $C_p$

Mallows's idea for model selection is to choose the submodel $m$ for which

$$\mathbb{E}||H(m)Y - X\beta||^2$$

is the smallest. Note that estimating $X\beta$ is crucial for prediction.

From (2), Mallows's idea is equivalent to choosing the model $m$ for which

$$\sigma^2(1 + p(m)) + \beta^T X^T (I - H(m))X\beta \tag{4}$$

is the smallest.

The problem here is that the above quantity depends on $\beta$ and $\sigma^2$ which are unknown. One therefore estimates them from data and then minimizes the estimate instead.

To estimate (4), let us first compute the expectation of $Y^T(I - H(m))Y$. Using the formula (1) with $Z = Y$ and $A = (I - H(m))$, we obtain

$$\begin{aligned}\mathbb{E}Y^T(I - H(m))Y &= \sigma^2 tr(I - H(m)) + \beta^T X^T(I - H(m))X\beta \\ &= \sigma^2(n - p(m) - 1) + \beta^T X^T(I - H(m))X\beta.\end{aligned}$$

Therefore an unbiased estimator of $\beta^T X^T(I - H(m))X\beta$ is

$$Y^T(I - H(m))Y - \sigma^2(n - p(m) - 1) = RSS(m) - \sigma^2(n - p(m) - 1).$$

As a result, an unbiased estimator of (4) is given by

$$RSS(m) - \sigma^2(n - 2 - 2p(m)).$$

This cannot exactly be used as a model selection criterion either because $\sigma^2$ is unknown. But it is natural to estimate it via $\hat{\sigma}^2 := RSS(M)/(n - p(M) - 1)$.

Therefore an unbiased estimator of (4) is

$$RSS(m) - \hat{\sigma}^2(n - 2 - 2p(m)).$$

Mallows's criterion is just a rescaling of this. Mallows's $C_p$ is defined as

$$C_p(m) := \frac{RSS(m)}{\hat{\sigma}^2} - (n - 2 - 2p(m)).$$

Pick the model $m$ for which $C_p(m)$ is the smallest. Note that $C_p(M) = p + 1$.

# 2 Cross-Validation

This is probably the most natural method for variable selection. Among a collection of models, we need to pick the model which has the best predictive performance. If we had access to future data, we can evaluate our models based on their predictive performance on that future data. How can one do this based on the existing data alone?

The most natural idea is the following: Split the data into $K$ roughly equal-sized parts. For the $k$th part, fit each model to the other $K - 1$ parts of the data and calculate the prediction error of each fitted model on this $k$th part of the data. Do this for each $k = 1, \ldots, K$ and combine the $K$ estimates of prediction error.

This is called $K$-fold Cross-validation. The case $K = n$ corresponds to $n$-fold cross-validation or Leave One Out Cross Validation.

## 2.1 Leave One Out Cross-Validation

For each $i = 1, \ldots, n$, fit the model $m$ to the $(n-1)$ observations obtained by excluding the $i$th observation. Predict the response for the $i$th observation using this model $m$ and the values of the explanatory variables for the $i$th observation. Record the prediction error. Do this for each $i = 1, \ldots, n$ and then add the squares of the prediction errors. This gives the Leave One Out Cross Validation score for the model $m$. Pick the model $m$ for which this score is the smallest.

Observe that the Leave One Out Cross Validation score for $m$ is nothing but the sum of the squares of the predicted residuals of $m$. Therfore, the Leave One Out Cross Validation Score is also called PRESS (Predicted REsidual Sum of Squares).

Recall that the $i$th predicted residual is defined as

$$\hat{e}_{[i]}(m) := y_i - x_i^T \hat{\beta}_{[i]}(m)$$

where $\hat{\beta}_{[i]}$ is the estimate of $\beta$ in the model $m$ fitted to the data excluding the $i$th observation. We showed (in Lecture 13) that

$$\hat{e}_{[i]}(m) = \frac{\hat{e}_i(m)}{1 - h_{ii}(m)}$$

where $h_{ii}(m)$ is the leverage of the $i$th observation in the model $m$.

Therefore

$$PRESS(m) := \sum_{i=1}^{n} \frac{\hat{e}_i^2(m)}{(1 - h_{ii}(m))^2}.$$

# 3 Generalized Cross-Validation

The problem with leave-one-out Cross-Validation or the PRESS statistic is that one has to calculate the leverages $h_{ii}(m)$ for each model $m$. In Generalized Cross Validation (GCV), one changes the PRESS statistic by replacing the individual leverages $h_{ii}(m)$ by their average $(1 + p(m))/n$.

This results in

$$GCV(m) = \sum_{i=1}^{n} \frac{\hat{e}_i^2(m)}{(1 - (1 + p(m))/n)^2} = \left(1 - \frac{1 + p(m)}{n}\right)^{-2} RSS(m).$$

$GCV(m)$ is very closely connected to Mallow's $C_p$. Indeed, if $n$ is much larger than $p$, then the approximation

$$\left(1 - \frac{1 + p(m)}{n}\right)^{-2} \approx 1 + 2\frac{1 + p(m)}{n}$$

leads to

$$GCV(m) \approx RSS(m) + 2(1 + p(m))\frac{RSS(m)}{n}. \tag{5}$$

You may recall that Mallow's $C_p$ is equivalent to minimizing the criterion

$$RSS(m) + 2(1 + p(m))\hat{\sigma}^2. \tag{6}$$

The only difference between (5) and (6) is that $\hat{\sigma}^2$ in (6) is replaced by $RSS(m)/n$ in (5). Note that $RSS(m)/n$ is the MLE of $\sigma^2$ in the model $m$. Thus the only difference between Mallow's $C_p$ and GCV is in the estimate of $\sigma^2$ used.