Lecture 20

October 28, 2018

Criteria Based Variable Selection

- We have so far looked at the following criteria:

Criteria Based Variable Selection

- We have so far looked at the following criteria:
- Adjusted R^2

Criteria Based Variable Selection

- We have so far looked at the following criteria:
- Adjusted R^2
- AIC

Criteria Based Variable Selection

- We have so far looked at the following criteria:
- Adjusted R^2
- AIC
- BIC

Criteria Based Variable Selection

- We have so far looked at the following criteria:
- Adjusted R^2
- AIC
- BIC
- Mallows's C_p

Mallows's C_p

- Mallow's C_p is defined as:

Mallows's C_p

- Mallow's C_p is defined as:
- $$C_p(m) := \frac{RSS(m)}{\hat{\sigma}^2} - (n - 2 - 2p(m)).$$

Mallows's C_p

- Mallow's C_p is defined as:
- $$C_p(m) := \frac{RSS(m)}{\hat{\sigma}^2} - (n - 2 - 2p(m)).$$
- One picks the model m for which $C_p(m)$ is the smallest.

Mallows's C_p

- Mallow's C_p is defined as:
- $$C_p(m) := \frac{RSS(m)}{\hat{\sigma}^2} - (n - 2 - 2p(m)).$$
- One picks the model m for which $C_p(m)$ is the smallest.
- The idea here is: let us pick the submodel m whose fitted values give us the best possible estimate of $X\beta$.

Mallows's C_p

- Mallow's C_p is defined as:
- $$C_p(m) := \frac{RSS(m)}{\hat{\sigma}^2} - (n - 2 - 2p(m)).$$
- One picks the model m for which $C_p(m)$ is the smallest.
- The idea here is: let us pick the submodel m whose fitted values give us the best possible estimate of $X\beta$.
- The vector of fitted values in a submodel m is denoted by $H(m)Y$.

Mallows's C_p

- Mallow's C_p is defined as:
- $$C_p(m) := \frac{RSS(m)}{\hat{\sigma}^2} - (n - 2 - 2p(m)).$$
- One picks the model m for which $C_p(m)$ is the smallest.
- The idea here is: let us pick the submodel m whose fitted values give us the best possible estimate of $X\beta$.
- The vector of fitted values in a submodel m is denoted by $H(m)Y$. The risk of $H(m)Y$ was calculated in the last class to be:

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

Prediction error

- Let us say today we observe a data set Y generated as $Y \sim N(X\beta, \sigma^2 I)$.

Prediction error

- Let us say today we observe a data set Y generated as $Y \sim N(X\beta, \sigma^2 I)$.
- Let us say tomorrow we will collect data Z generated as $Z \sim N(X\beta, \sigma^2 I)$.

Prediction error

- Let us say today we observe a data set Y generated as $Y \sim N(X\beta, \sigma^2 I)$.
- Let us say tomorrow we will collect data Z generated as $Z \sim N(X\beta, \sigma^2 I)$. Then our guess, today, for Z is

$$H(m)Y,$$

  if our guess is based on the ols procedure with the submodel m.

Prediction error

- Let us say today we observe a data set Y generated as $Y \sim N(X\beta, \sigma^2 I)$.
- Let us say tomorrow we will collect data Z generated as $Z \sim N(X\beta, \sigma^2 I)$. Then our guess, today, for Z is

  $$H(m)Y,$$

  if our guess is based on the ols procedure with the submodel m.
- If we knew β our guess for Z would be $X\beta$, and our error would be $\mathbb{E}\left[\|X\beta - Z\|_2^2\right] \; = \; n\sigma^2$.

Prediction error

- Let us say today we observe a data set Y generated as $Y \sim N(X\beta, \sigma^2 I)$.
- Let us say tomorrow we will collect data Z generated as $Z \sim N(X\beta, \sigma^2 I)$. Then our guess, today, for Z is

  $$H(m)Y,$$

  if our guess is based on the ols procedure with the submodel m.
- If we knew β our guess for Z would be $X\beta$, and our error would be $\mathbb{E}\left[\|X\beta - Z\|_2^2\right] = n\sigma^2$.
- Using model m, the expected error we will make tomorrow satisfies

Prediction error

- Let us say today we observe a data set Y generated as $Y \sim N(X\beta, \sigma^2 I)$.
- Let us say tomorrow we will collect data Z generated as $Z \sim N(X\beta, \sigma^2 I)$. Then our guess, today, for Z is

$$H(m)Y,$$

  if our guess is based on the ols procedure with the submodel m.
- If we knew β our guess for Z would be $X\beta$, and our error would be $\mathbb{E}\left[\|X\beta - Z\|_2^2\right] \; = \; n\sigma^2$.
- Using model m, the expected error we will make tomorrow satisfies

$$\mathbb{E}\left[\|Z - H(m)Y\|_2^2\right] \quad =$$

Prediction error

- Let us say today we observe a data set Y generated as $Y \sim N(X\beta, \sigma^2 I)$.
- Let us say tomorrow we will collect data Z generated as $Z \sim N(X\beta, \sigma^2 I)$. Then our guess, today, for Z is

$$H(m)Y,$$

  if our guess is based on the ols procedure with the submodel m.
- If we knew β our guess for Z would be $X\beta$, and our error would be $\mathbb{E}\left[\|X\beta - Z\|_2^2\right] \;=\; n\sigma^2$.
- Using model m, the expected error we will make tomorrow satisfies

$$\mathbb{E}\left[\|Z - H(m)Y\|_2^2\right] \;=\; \mathbb{E}\left[\|(Z - X\beta) + (X\beta - H(m)Y)\|_2^2\right]$$

Prediction error

- Let us say today we observe a data set Y generated as $Y \sim N(X\beta, \sigma^2 I)$.
- Let us say tomorrow we will collect data Z generated as $Z \sim N(X\beta, \sigma^2 I)$. Then our guess, today, for Z is

  $$H(m)Y,$$

  if our guess is based on the ols procedure with the submodel m.
- If we knew β our guess for Z would be $X\beta$, and our error would be $\mathbb{E}\left[\|X\beta - Z\|_2^2\right] \;=\; n\,\sigma^2$.
- Using model m, the expected error we will make tomorrow satisfies

$$\begin{aligned}
\mathbb{E}\left[\|Z - H(m)Y\|_2^2\right] \;&=\; \mathbb{E}\left[\|(Z - X\beta) + (X\beta - H(m)Y)\|_2^2\right] \\
&=\; \mathbb{E}\left[\|X\beta - Z\|_2^2\right] \;+\; \mathbb{E}\left[\|X\beta - H(m)Y\|_2^2\right]
\end{aligned}$$

Prediction error

- Let us say today we observe a data set Y generated as $Y \sim N(X\beta, \sigma^2 I)$.
- Let us say tomorrow we will collect data Z generated as $Z \sim N(X\beta, \sigma^2 I)$. Then our guess, today, for Z is

$$H(m)Y,$$

  if our guess is based on the ols procedure with the submodel m.
- If we knew β our guess for Z would be $X\beta$, and our error would be $\mathbb{E}\left[\|X\beta - Z\|_2^2\right] \;=\; n\sigma^2$.
- Using model m, the expected error we will make tomorrow satisfies

$$\begin{aligned}
\mathbb{E}\left[\|Z - H(m)Y\|_2^2\right] &= \mathbb{E}\left[\|(Z - X\beta) + (X\beta - H(m)Y)\|_2^2\right] \\
&= \mathbb{E}\left[\|X\beta - Z\|_2^2\right] \;+\; \mathbb{E}\left[\|X\beta - H(m)Y\|_2^2\right] \\
&= n\sigma^2 \;+\; \mathbb{E}\left[\|X\beta - H(m)Y\|_2^2\right].
\end{aligned}$$

- But we cannot minimize $\mathbb{E}||H(m)Y - X\beta||^2$ over m because this depends on β.

- But we cannot minimize $\mathbb{E}||H(m)Y - X\beta||^2$ over m because this depends on β. Mallows's realized that

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

- But we cannot minimize $\mathbb{E}||H(m)Y - X\beta||^2$ over m because this depends on β. Mallows's realized that

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

- Thus one can use

$$RSS(m) - \hat{\sigma}^2(n - 2 - 2p(m)) \tag{1}$$

as a proxy for $\mathbb{E}||H(m)Y - X\beta||^2$.

- But we cannot minimize $\mathbb{E}||H(m)Y - X\beta||^2$ over m because this depends on β. Mallows's realized that

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

- Thus one can use

$$RSS(m) - \hat{\sigma}^2(n - 2 - 2p(m)) \tag{1}$$

as a proxy for $\mathbb{E}||H(m)Y - X\beta||^2$.
- Minimizing (1) over submodels m is equivalent to minimizing $C_p(m)$.

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.

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.

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?

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:

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.

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 Kth part, fit each model to the other $K - 1$ parts of the data and calculate the prediction error of each fitted model on this kth part of the data.

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 Kth part, fit each model to the other $K - 1$ parts of the data and calculate the prediction error of each fitted model on this kth part of the data.
- Do this for each $k = 1, \dots, K$ and combine the K estimates of prediction error.

- This is called K-fold Cross-validation.

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

- This is called K-fold Cross-validation. The case $K = n$ corresponds to n-fold cross-validation or Leave One Out Cross Validation.
- **Leave One Out Cross-Validation** For each $i = 1, \ldots, n$, fit the model m to the $(n-1)$ observations obtained by excluding the ith observation.

- This is called K-fold Cross-validation. The case $K = n$ corresponds to n-fold cross-validation or Leave One Out Cross Validation.
- **Leave One Out Cross-Validation** For each $i = 1, \ldots, n$, fit the model m to the $(n - 1)$ observations obtained by excluding the ith observation.
- Predict the response for the ith observation using this model m and the values of the explanatory variables for the ith observation.

- This is called K-fold Cross-validation. The case $K = n$ corresponds to n-fold cross-validation or Leave One Out Cross Validation.
- **Leave One Out Cross-Validation** For each $i = 1, \ldots, n$, fit the model m to the $(n-1)$ observations obtained by excluding the ith observation.
- Predict the response for the ith observation using this model m and the values of the explanatory variables for the ith observation.
- Record the 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.
- **Leave One Out Cross-Validation** For each $i = 1, \dots, n$, fit the model m to the $(n - 1)$ observations obtained by excluding the ith observation.
- Predict the response for the ith observation using this model m and the values of the explanatory variables for the ith observation.
- Record the prediction error.
- Do this for each $i = 1, \dots, n$ and then add the squares of the prediction errors.

- This is called K-fold Cross-validation. The case $K = n$ corresponds to n-fold cross-validation or Leave One Out Cross Validation.
- **Leave One Out Cross-Validation** For each $i = 1, \dots, n$, fit the model m to the $(n - 1)$ observations obtained by excluding the ith observation.
- Predict the response for the ith observation using this model m and the values of the explanatory variables for the ith observation.
- Record the prediction error.
- Do this for each $i = 1, \dots, n$ and then add the squares of the prediction errors.
- This gives the Leave One Out Cross Validation score for the model m.

- This is called K-fold Cross-validation. The case $K = n$ corresponds to n-fold cross-validation or Leave One Out Cross Validation.
- **Leave One Out Cross-Validation** For each $i = 1, \dots, n$, fit the model m to the $(n-1)$ observations obtained by excluding the ith observation.
- Predict the response for the ith observation using this model m and the values of the explanatory variables for the ith observation.
- Record the prediction error.
- Do this for each $i = 1, \dots, n$ 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.

- 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.
- Therefore, the Leave One Out Cross Validation Score is also called PRESS (Predicted REsidual Sum of Squares).

- 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*.
- Therefore, 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 β in the model *m* fitted to the data excluding the *i*th observation.

- 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.
- Therefore, the Leave One Out Cross Validation Score is also called PRESS (Predicted REsidual Sum of Squares).
- Recall that the ith 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 β in the model m fitted to the data excluding the ith observation.
- We showed that

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

  where $h_i(m)$ is the leverage of the ith observation in the model m.

- Therefore

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

Generalized Cross-Validation

- The problem with leave-one-out Cross-Validation or the PRESS statistic is that one has to calculate the leverages $h_i(m)$ for each model m.

Generalized Cross-Validation

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

Generalized Cross-Validation

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

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

- $GCV(m)$ 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

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

- $GCV(m)$ 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

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

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

$$RSS(m) + 2(1 + p(m))\hat{\sigma}^2. \qquad (3)$$

- The only difference between (2) and (3) is that $\hat{\sigma}^2$ in (3) is replaced by $RSS(m)/n$ in (2).

- The only difference between (2) and (3) is that $\hat{\sigma}^2$ in (3) is replaced by $RSS(m)/n$ in (2).
- Note that $RSS(m)/n$ is the MLE of σ^2 in the model m.

- The only difference between (2) and (3) is that $\hat{\sigma}^2$ in (3) is replaced by $RSS(m)/n$ in (2).
- Note that $RSS(m)/n$ is the MLE of σ^2 in the model m.
- Thus the only difference between Mallow's C_p and GCV is in the estimate of σ^2 used.