# Lecture 19

October 24, 2018

# Criteria Based Variable Selection

- If there are $p$ explanatory variables, then there are $2^p$ possible linear models.

# Criteria Based Variable Selection

- If there are $p$ explanatory variables, then there are $2^p$ possible linear models.
- In criteria-based variable selection, we fit all these models and choose the best one according to some criterion.

# Criteria Based Variable Selection

- If there are $p$ explanatory variables, then there are $2^p$ possible linear models.
- In criteria-based variable selection, we fit all these models and choose the best one according to some criterion.
- If $p$ is such that $2^p$ is prohibitively large, then one uses a stepwise procedure for generating candidate models and then compares them according to a criterion.

# Criteria Based Variable Selection

if two columns are collinear, var(beta) is large

- If there are $p$ explanatory variables, then there are $2^p$ possible linear models.
- In criteria-based variable selection, we fit all these models and choose the best one according to some criterion.
- If $p$ is such that $2^p$ is prohibitively large, then one uses a stepwise procedure for generating candidate models and then compares them according to a criterion.
- There are several criteria that one can use. Some of the common ones are given below.

# $R^2$

- For each candidate model $m$, recall that its $R^2(m)$ is defined as
$$R^2(m) := 1 - \frac{RSS(m)}{TSS}$$
where $RSS(m)$ is the residual sum of squares for the model $m$ and $TSS$ is the total sum of squares.

# $R^2$

- For each candidate model $m$, recall that its $R^2(m)$ is defined as
$$R^2(m) := 1 - \frac{RSS(m)}{TSS}$$
where $RSS(m)$ is the residual sum of squares for the model $m$ and $TSS$ is the total sum of squares.
- $R^2(m)$ **should NOT be used** as a criterion for variable selection because then we will always pick the full model $M$ which has the highest $R^2$ value among all the candidate models.

# Adjusted $R^2$

- Adjusted $R^2$ is defined as

$$(AdjR^2)(m) := 1 - \frac{RSS(m)/(n - p(m) - 1)}{TSS/(n-1)}$$

where $p(m)$ is the number of explanatory variables in model $m$.

# Adjusted $R^2$

- Adjusted $R^2$ is defined as

$$(AdjR^2)(m) := 1 - \frac{RSS(m)/(n - p(m) - 1)}{TSS/(n-1)}$$

  where $p(m)$ is the number of explanatory variables in model $m$.
- This is very similar to $R^2$ but has the desirable property that when an explanatory variable is removed from a model, the value of $AdjR^2$ does not necessarily decrease.

# Adjusted $R^2$

- Adjusted $R^2$ is defined as

$$(AdjR^2)(m) := 1 - \frac{RSS(m)/(n - p(m) - 1)}{TSS/(n-1)}$$

  where $p(m)$ is the number of explanatory variables in model $m$.
- This is very similar to $R^2$ but has the desirable property that when an explanatory variable is removed from a model, the value of $AdjR^2$ does not necessarily decrease.
- It might increase if the removed variable has no predictive power.

# Adjusted $R^2$

- Adjusted $R^2$ is defined as

$$(AdjR^2)(m) := 1 - \frac{RSS(m)/(n-p(m)-1)}{TSS/(n-1)}$$

  where $p(m)$ is the number of explanatory variables in model $m$.
- This is very similar to $R^2$ but has the desirable property that when an explanatory variable is removed from a model, the value of $AdjR^2$ does not necessarily decrease.
- It might increase if the removed variable has no predictive power.
- This can therefore be used as a criterion for variable selection.

# AIC

- AIC stands for Akaike Information Criterion and is one of the most popular model selection techniques not just in linear models but in other contexts as well.

# AIC

- AIC stands for Akaike Information Criterion and is one of the most popular model selection techniques not just in linear models but in other contexts as well.
- AIC for a model $m$ is defined as

$$\begin{aligned} AIC(m) := & -2\log(\text{maximum value of likelihood in } m) \\ & +2(\text{number of parameters in } m) \end{aligned} \tag{1}$$

# AIC

- AIC stands for Akaike Information Criterion and is one of the most popular model selection techniques not just in linear models but in other contexts as well.
- AIC for a model $m$ is defined as

$$\begin{aligned} AIC(m) := & -2\log(\text{maximum value of likelihood in } m) \\ & +2(\text{number of parameters in } m) \end{aligned} \tag{1}$$

- We pick models with small AIC.

- In the case of linear models, we can show that

$$AIC(m) = n\log\left(\frac{RSS(m)}{n}\right) + n\log(2\pi e) + 2(1 + p(m)) \quad (2)$$

- In the case of linear models, we can show that

$$AIC(m) = n \log\left(\frac{RSS(m)}{n}\right) + n\log(2\pi e) + 2(1 + p(m))) \quad (2)$$

- To see this observe that the log-likelihood function in the linear model $Y \sim N(X\beta, \sigma^2 I)$ equals

$$\frac{-n}{2}\left(\log(2\pi) + \log\sigma^2\right) - \frac{||Y - X\beta||^2}{2\sigma^2}.$$

- In the case of linear models, we can show that

$$AIC(m) = n \log\left(\frac{RSS(m)}{n}\right) + n\log(2\pi e) + 2(1 + p(m))) \quad (2)$$

- To see this observe that the log-likelihood function in the linear model $Y \sim N(X\beta, \sigma^2 I)$ equals

$$\frac{-n}{2}\left(\log(2\pi) + \log\sigma^2\right) - \frac{||Y - X\beta||^2}{2\sigma^2}.$$

- It is easy to see that this is maximized when

$$\hat{\beta} = (X^T X)^{-1} X^T Y \text{ and } \hat{\sigma}^2_{mle} := \frac{RSS}{n}.$$

- Plugging these values in the log-likelihood function and simplifying, we see that the maximized log-likelihood for the model is

- Plugging these values in the log-likelihood function and simplifying, we see that the maximized log-likelihood for the model is

$$\frac{-n}{2}\left(\log(2\pi) + \log \hat{\sigma}^2_{mle}\right) - \frac{||Y - X\hat{\beta}||^2}{2\hat{\sigma}^2_{mle}}.$$

$$=$$

- Plugging these values in the log-likelihood function and simplifying, we see that the maximized log-likelihood for the model is

$$
\begin{aligned}
&\frac{-n}{2}\left(\log(2\pi) + \log \hat{\sigma}^2_{mle}\right) - \frac{||Y - X\hat{\beta}||^2}{2\hat{\sigma}^2_{mle}}. \\
&= \frac{-n}{2}\left(\log(2\pi) + \log\left(\frac{RSS}{n}\right)\right) - \frac{RSS}{2\frac{RSS}{n}}. \\
&=
\end{aligned}
$$

- Plugging these values in the log-likelihood function and simplifying, we see that the maximized log-likelihood for the model is

$$
\begin{aligned}
&\frac{-n}{2}\left(\log(2\pi) + \log \hat{\sigma}^2_{mle}\right) - \frac{||Y - X\hat{\beta}||^2}{2\hat{\sigma}^2_{mle}}. \\
&= \frac{-n}{2}\left(\log(2\pi) + \log\left(\frac{RSS}{n}\right)\right) - \frac{RSS}{2\frac{RSS}{n}}. \\
&= \frac{-n}{2}\left(\log(2\pi) + \log\left(\frac{RSS}{n}\right)\right) - \frac{n}{2}. \\
&=
\end{aligned}
$$

- Plugging these values in the log-likelihood function and simplifying, we see that the maximized log-likelihood for the model is

$$
\begin{aligned}
&\frac{-n}{2}\left(\log(2\pi) + \log\hat{\sigma}^2_{mle}\right) - \frac{||Y - X\hat{\beta}||^2}{2\hat{\sigma}^2_{mle}}. \\
&= \frac{-n}{2}\left(\log(2\pi) + \log\left(\frac{RSS}{n}\right)\right) - \frac{RSS}{2\frac{RSS}{n}}. \\
&= \frac{-n}{2}\left(\log(2\pi) + \log\left(\frac{RSS}{n}\right)\right) - \frac{n}{2}. \\
&= -\frac{n}{2}\log(2\pi e) - \frac{n}{2}\log\left(\frac{RSS}{n}\right).
\end{aligned}
$$

- AIC is used as a criteria to compare various models. The term $n \log(2\pi e)$ clearly is the same for all models and therefore, one often simply drops it and defines the AIC for linear models as

$$AIC(m) := n \log \left( \frac{RSS(m)}{n} \right) + 2(1 + p(m)))$$

- AIC is used as a criteria to compare various models. The term $n \log(2\pi e)$ clearly is the same for all models and therefore, one often simply drops it and defines the AIC for linear models as

$$AIC(m) := n \log\left(\frac{RSS(m)}{n}\right) + 2(1 + p(m)))$$

- Note that if $m_1$ is a sub-model of $m_2$, then $RSS(m_1) \geq RSS(m_2)$ while $p(m_1) \leq p(m_2)$

- AIC is used as a criteria to compare various models. The term $n \log(2\pi e)$ clearly is the same for all models and therefore, one often simply drops it and defines the AIC for linear models as

$$AIC(m) := n \log\left(\frac{RSS(m)}{n}\right) + 2(1 + p(m)))$$

- Note that if $m_1$ is a sub-model of $m_2$, then $RSS(m_1) \geq RSS(m_2)$ while $p(m_1) \leq p(m_2)$ so $AIC(m_1)$ may or may not be smaller than $AIC(m_2)$.

- AIC is used as a criteria to compare various models. The term $n\log(2\pi e)$ clearly is the same for all models and therefore, one often simply drops it and defines the AIC for linear models as

$$AIC(m) := n\log\left(\frac{RSS(m)}{n}\right) + 2(1 + p(m)))$$

- Note that if $m_1$ is a sub-model of $m_2$, then $RSS(m_1) \geq RSS(m_2)$ while $p(m_1) \leq p(m_2)$ so $AIC(m_1)$ may or may not be smaller than $AIC(m_2)$. If it is smaller, we would prefer $m_1$; otherwise, we prefer $m_2$.

# BIC

- BIC stands for Bayesian Information Criterion. BIC for a model $m$ is defined as

$$\begin{aligned} BIC(m) := -2\log(&\text{maximum value of likelihood in } m)+ \\ &(\log n)(\text{number of parameters in } m). \end{aligned} \tag{3}$$

# BIC

- BIC stands for Bayesian Information Criterion. BIC for a model $m$ is defined as

$$
\begin{aligned}
BIC(m) := -2 \log(&\text{maximum value of likelihood in } m)+ \\
&(\log n)(\text{number of parameters in } m).
\end{aligned} \tag{3}
$$

- In model selection via the BIC, one selects models with small BIC.

# BIC

- BIC stands for Bayesian Information Criterion. BIC for a model $m$ is defined as

$$\begin{aligned} BIC(m) := -2\log(&\text{maximum value of likelihood in } m) + \\ &(\log n)(\text{number of parameters in } m). \end{aligned} \tag{3}$$

- In model selection via the BIC, one selects models with small BIC.
- Note that the only difference between the formulae for AIC and BIC is the factor of the number of parameters term which is 2 for AIC and $\log n$ for BIC.

# BIC

- BIC stands for Bayesian Information Criterion. BIC for a model $m$ is defined as

$$BIC(m) := -2\log(\text{maximum value of likelihood in } m) + (\log n)(\text{number of parameters in } m). \tag{3}$$

- In model selection via the BIC, one selects models with small BIC.
- Note that the only difference between the formulae for AIC and BIC is the factor of the number of parameters term which is 2 for AIC and $\log n$ for BIC.
- Because $\log n$ is typically larger than 2, the size of models selected by BIC is smaller than those selected by AIC.

In the case of linear models, one has

$$BIC(m) := n \log\left(\frac{RSS(m)}{n}\right) + (\log n)(1 + p(m)).$$

# Mallow's $C_p$

- For a submodel $m$, the Mallows's $C_p$ criterion is defined as

$$C_p(m) := \frac{RSS(m)}{\hat{\sigma}^2} - (n - 2(1 + p(m)))$$

where $\hat{\sigma}^2 := RSS(M)/(n - p(M) - 1)$ is the estimate of $\sigma^2$ in the full model.

# Mallow's $C_p$

- For a submodel $m$, the Mallows's $C_p$ criterion is defined as

$$C_p(m) := \frac{RSS(m)}{\hat{\sigma}^2} - (n - 2(1 + p(m)))$$

where $\hat{\sigma}^2 := RSS(M)/(n - p(M) - 1)$ is the estimate of $\sigma^2$ in the full model. The Mallows's method picks the model $m$ for which $C_p(m)$ is the smallest.

- The justification for the $C_p$ criterion comes from unbiased risk estimation as explained below.

- The justification for the $C_p$ criterion comes from unbiased risk estimation as explained below.
- Model $Y$ as $N(X\beta, \sigma^2 I_n)$ and take $X$ to be deterministic as usual. Set $\mu := X\beta$.

- The justification for the $C_p$ criterion comes from unbiased risk estimation as explained below.
- Model $Y$ as $N(X\beta, \sigma^2 I_n)$ and take $X$ to be deterministic as usual. Set $\mu := X\beta$.
- Consider the problem of estimating $\mu$ based on $Y$ and $X$.

- The justification for the $C_p$ criterion comes from unbiased risk estimation as explained below.
- Model $Y$ as $N(X\beta, \sigma^2 I_n)$ and take $X$ to be deterministic as usual. Set $\mu := X\beta$.
- Consider the problem of estimating $\mu$ based on $Y$ and $X$.
- Here is a reasonable candidate estimator. Select a submodel $m$ and estimate $\mu$ by the vector of fitted values in the linear regression for $Y$ based on the explanatory variables in $m$.

- The justification for the $C_p$ criterion comes from unbiased risk estimation as explained below.
- Model $Y$ as $N(X\beta, \sigma^2 I_n)$ and take $X$ to be deterministic as usual. Set $\mu := X\beta$.
- Consider the problem of estimating $\mu$ based on $Y$ and $X$.
- Here is a reasonable candidate estimator. Select a submodel $m$ and estimate $\mu$ by the vector of fitted values in the linear regression for $Y$ based on the explanatory variables in $m$.
- Let us denote this estimator by $H(m)Y$ where $H(m)$ is the hat matrix in the submodel $m$.

- The justification for the $C_p$ criterion comes from unbiased risk estimation as explained below.
- Model $Y$ as $N(X\beta, \sigma^2 I_n)$ and take $X$ to be deterministic as usual. Set $\mu := X\beta$.
- Consider the problem of estimating $\mu$ based on $Y$ and $X$.
- Here is a reasonable candidate estimator. Select a submodel $m$ and estimate $\mu$ by the vector of fitted values in the linear regression for $Y$ based on the explanatory variables in $m$.
- Let us denote this estimator by $H(m)Y$ where $H(m)$ is the hat matrix in the submodel $m$. The performance of this estimator is evaluated by the term:

$$R(m) := \mathbb{E}\|H(m)Y - \mu\|^2.$$

- This quantity $R(m)$ is called the risk of the estimator $H(m)Y$. And as we show below:

$$R(m) = \|H(m)\mu - \mu\|^2 + \sigma^2(1 + p(m)). \tag{4}$$

where, again, $1 + p(m)$ is the number of columns of $X(m)$ (which is the $X$-matrix in the submodel $m$).

- This quantity $R(m)$ is called the risk of the estimator $H(m)Y$. And as we show below:

$$R(m) = \|H(m)\mu - \mu\|^2 + \sigma^2(1 + p(m)). \tag{4}$$

  where, again, $1 + p(m)$ is the number of columns of $X(m)$ (which is the $X$-matrix in the submodel $m$).
- Note the tradeoff between complicated and simple models in the right hand side of (4). If $m$ is a complicated model (i.e., if it has many explanatory variables), then $p(m)$ will be large while $\|H(m)\mu - \mu\|^2$ will be small.

- This quantity $R(m)$ is called the risk of the estimator $H(m)Y$. And as we show below:

$$R(m) = \|H(m)\mu - \mu\|^2 + \sigma^2(1 + p(m)). \tag{4}$$

  where, again, $1 + p(m)$ is the number of columns of $X(m)$ (which is the $X$-matrix in the submodel $m$).
- Note the tradeoff between complicated and simple models in the right hand side of (4). If $m$ is a complicated model (i.e., if it has many explanatory variables), then $p(m)$ will be large while $\|H(m)\mu - \mu\|^2$ will be small.
- On the other hand, if $m$ is a simple model, then $p(m)$ will be small but $\|H(m)\mu - \mu\|^2$ might be large. It may be helpful to note here that $\|H(m)\mu - \mu\|^2$ equals the squared distance from $\mu$ to the column space generated by the columns of $X(m)$.

- To show the formula:

$$R(m) = \|H(m)\mu - \mu\|^2 + \sigma^2(1 + p(m)).$$

we can use a well known fact: Suppose $Z$ is a random vector with mean $\mu$ and covariance matrix $\Sigma$. Then

$$\mathbb{E}(Z^T A Z) = tr(A\Sigma) + \mu^T A \mu. \tag{5}$$

- To show the formula:

$$R(m) = \|H(m)\mu - \mu\|^2 + \sigma^2(1 + p(m)).$$

we can use a well known fact: Suppose $Z$ is a random vector with mean $\mu$ and covariance matrix $\Sigma$. Then

$$\mathbb{E}(Z^T A Z) = tr(A\Sigma) + \mu^T A \mu. \tag{5}$$

- Simply take $Z = X\beta - H(m)Y$ and $A = I$. The mean of $Z$ is

- To show the formula:

$$R(m) = \|H(m)\mu - \mu\|^2 + \sigma^2(1 + p(m)).$$

we can use a well known fact: Suppose $Z$ is a random vector with mean $\mu$ and covariance matrix $\Sigma$. Then

$$\mathbb{E}(Z^T AZ) = tr(A\Sigma) + \mu^T A\mu. \quad (5)$$

- Simply take $Z = X\beta - H(m)Y$ and $A = I$. The mean of $Z$ is

- To show the formula:

$$R(m) = \|H(m)\mu - \mu\|^2 + \sigma^2(1 + p(m)).$$

we can use a well known fact: Suppose $Z$ is a random vector with mean $\mu$ and covariance matrix $\Sigma$. Then

$$\mathbb{E}(Z^T A Z) = tr(A\Sigma) + \mu^T A\mu. \tag{5}$$

- Simply take $Z = X\beta - H(m)Y$ and $A = I$. The mean of $Z$ is

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

- To show the formula:

$$R(m) = \|H(m)\mu - \mu\|^2 + \sigma^2(1 + p(m)).$$

  we can use a well known fact: Suppose $Z$ is a random vector with mean $\mu$ and covariance matrix $\Sigma$. Then

$$\mathbb{E}(Z^T A Z) = tr(A\Sigma) + \mu^T A \mu. \tag{5}$$

- Simply take $Z = X\beta - H(m)Y$ and $A = I$. The mean of $Z$ is

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

- To show the formula:

$$R(m) = \|H(m)\mu - \mu\|^2 + \sigma^2(1 + p(m)).$$

we can use a well known fact: Suppose $Z$ is a random vector with mean $\mu$ and covariance matrix $\Sigma$. Then

$$\mathbb{E}(Z^T A Z) = tr(A\Sigma) + \mu^T A\mu. \quad (5)$$

- Simply take $Z = X\beta - H(m)Y$ 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

$$Cov(Z) = Cov(X\beta - H(m)Y)$$

- The covariance matrix of $Z$ is

$$\begin{aligned} Cov(Z) &= Cov(X\beta - H(m)Y) \\ &= Cov(H(m)Y) \end{aligned}$$

- The covariance matrix of $Z$ is

$$
\begin{aligned}
Cov(Z) &= Cov(X\beta - H(m)Y) \\
&= Cov(H(m)Y)
\end{aligned}
$$

- The covariance matrix of $Z$ is

$$\begin{aligned} Cov(Z) &= Cov(X\beta - H(m)Y) \\ &= Cov(H(m)Y) \\ &= \sigma^2 H(m)H(m)^T \end{aligned}$$

- The covariance matrix of $Z$ is

$$\begin{aligned} Cov(Z) &= Cov(X\beta - H(m)Y) \\ &= Cov(H(m)Y) \\ &= \sigma^2 H(m)H(m)^T \\ &= \sigma^2 H(m). \end{aligned}$$

- The covariance matrix of $Z$ is

$$\begin{aligned} Cov(Z) &= Cov(X\beta - H(m)Y) \\ &= Cov(H(m)Y) \\ &= \sigma^2 H(m)H(m)^T \\ &= \sigma^2 H(m). \end{aligned}$$

- Therefore, from (5), 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 covariance matrix of $Z$ is

$$\begin{aligned} Cov(Z) &= Cov(X\beta - H(m)Y) \\ &= Cov(H(m)Y) \\ &= \sigma^2 H(m)H(m)^T \\ &= \sigma^2 H(m). \end{aligned}$$

- Therefore, from (5), 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

$$\mathbb{E}||H(m)Y - X\beta||^2 = \sigma^2 (1 + p(m)) + \beta^T X^T (I - H(m)) X\beta. \tag{6}$$

- This therefore gives

$$\mathbb{E}||H(m)Y - X\beta||^2 = \sigma^2 (1 + p(m)) + \beta^T X^T (I - H(m))X\beta. \tag{6}$$

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

$$\mathbb{E}||HY - X\beta||^2 = \sigma^2(1+p) + \beta^T X^T (I - H)X\beta = \sigma^2(1+p) \tag{7}$$

because $HX = X$.

- This therefore gives

$$\mathbb{E}||H(m)Y - X\beta||^2 = \sigma^2 (1 + p(m)) + \beta^T X^T (I - H(m)) X\beta. \tag{6}$$

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

$$\mathbb{E}||HY - X\beta||^2 = \sigma^2(1+p) + \beta^T X^T (I-H) X\beta = \sigma^2(1+p) \tag{7}$$

  because $HX = X$.
- Comparing (6) with (7), we see that $H(m)Y$ is a better estimator of $X\beta$ than $HY$ provided

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

- If we knew $R(m)$, we would pick the model $m$ for which $R(m)$ is the smallest.

- If we knew $R(m)$, we would pick the model $m$ for which $R(m)$ is the smallest.
- The trouble though is that $R(m)$ is unknown because $\mu$ is unknown (and also $\sigma^2$).

- If we knew $R(m)$, we would pick the model $m$ for which $R(m)$ is the smallest.
- The trouble though is that $R(m)$ is unknown because $\mu$ is unknown (and also $\sigma^2$).
- Mallows's had this clever idea that $R(m)$ can be estimated unbiasedly from the data. In fact, it is easy to check that

$$\mathbb{E}\left(RSS(m) - \sigma^2(n - 2r(m))\right) = R(m).$$

- If we knew $R(m)$, we would pick the model $m$ for which $R(m)$ is the smallest.
- The trouble though is that $R(m)$ is unknown because $\mu$ is unknown (and also $\sigma^2$).
- Mallows's had this clever idea that $R(m)$ can be estimated unbiasedly from the data. In fact, it is easy to check that

$$\mathbb{E}\left(RSS(m) - \sigma^2(n - 2r(m))\right) = R(m).$$

- This can be verified using (5).

- If we knew $R(m)$, we would pick the model $m$ for which $R(m)$ is the smallest.
- The trouble though is that $R(m)$ is unknown because $\mu$ is unknown (and also $\sigma^2$).
- Mallows's had this clever idea that $R(m)$ can be estimated unbiasedly from the data. In fact, it is easy to check that

$$\mathbb{E}\left(RSS(m) - \sigma^2(n - 2r(m))\right) = R(m).$$

- This can be verified using (5). Because of this, a natural idea for selecting $m$ is to minimize $RSS(m) - \sigma^2(n - 2(1 + p(m)))$ over $m$. But this still involves $\sigma^2$.

- If we knew $R(m)$, we would pick the model $m$ for which $R(m)$ is the smallest.
- The trouble though is that $R(m)$ is unknown because $\mu$ is unknown (and also $\sigma^2$).
- Mallows's had this clever idea that $R(m)$ can be estimated unbiasedly from the data. In fact, it is easy to check that

$$\mathbb{E}\left(RSS(m) - \sigma^2(n - 2r(m))\right) = R(m).$$

- This can be verified using (5). Because of this, a natural idea for selecting $m$ is to minimize $RSS(m) - \sigma^2(n - 2(1 + p(m)))$ over $m$. But this still involves $\sigma^2$.
- One simply replaces $\sigma^2$ by the usual estimate from the full model i.e., $\hat{\sigma}^2 := RSS(M)/(n - p(M) - 1)$. This leads to the minimization of

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

- Mallows's criterion is just a rescaling of this where we divide through by $\hat{\sigma}^2$.

- Mallows's criterion is just a rescaling of this where we divide through by $\hat{\sigma}^2$. Pick the model $m$ for which $C_p(m)$ is the smallest.

- Mallows's criterion is just a rescaling of this where we divide through by $\hat{\sigma}^2$. Pick the model $m$ for which $C_p(m)$ is the smallest. Note that $C_p(M) = p + 1$.

- Mallows's criterion is just a rescaling of this where we divide through by $\hat{\sigma}^2$. Pick the model $m$ for which $C_p(m)$ is the smallest. Note that $C_p(M) = p + 1$.
- For models of a given size, all the methods above will select the model with the smallest RSS.