MONASH University

ETC3250

# Business Analytics

## 4. Cross-validation

17 August 2015

# Outline

# Choosing regression models

- When there are many predictors, how should we choose which ones to use?
- How do we choose the df for a spline?
- We need a way of comparing two competing models
- If there are a limited number of predictors, we can study all possible models.
- Otherwise we need a search strategy to explore some potential models.

# Choosing regression models

- When there are many predictors, how should we choose which ones to use?
- How do we choose the df for a spline?
- We need a way of comparing two competing models
- If there are a limited number of predictors, we can study all possible models.
- Otherwise we need a search strategy to explore some potential models.

# Choosing regression models

- When there are many predictors, how should we choose which ones to use?
- How do we choose the df for a spline?
- We need a way of comparing two competing models
- If there are a limited number of predictors, we can study all possible models.
- Otherwise we need a search strategy to explore some potential models.

# Choosing regression models

- When there are many predictors, how should we choose which ones to use?
- How do we choose the df for a spline?
- We need a way of comparing two competing models
- If there are a limited number of predictors, we can study all possible models.
- Otherwise we need a search strategy to explore some potential models.

# Choosing regression models

- When there are many predictors, how should we choose which ones to use?
- How do we choose the df for a spline?
- We need a way of comparing two competing models
- If there are a limited number of predictors, we can study all possible models.
- Otherwise we need a search strategy to explore some potential models.

# Choosing regression variables

## What not to do!

- Plot $Y$ against a particular predictor ($X_j$) and if it shows no noticeable relationship, drop it.
- Do a multiple linear regression on all the predictors and disregard all variables whose $p$ values are greater than 0.05.
- Maximize $R^2$
- Minimize SSE.

# Choosing regression variables

## What not to do!

- Plot $Y$ against a particular predictor ($X_j$) and if it shows no noticeable relationship, drop it.
- Do a multiple linear regression on all the predictors and disregard all variables whose $p$ values are greater than 0.05.
- Maximize $R^2$
- Minimize SSE.

# Choosing regression variables

## What not to do!

- Plot $Y$ against a particular predictor ($X_j$) and if it shows no noticeable relationship, drop it.
- Do a multiple linear regression on all the predictors and disregard all variables whose $p$ values are greater than 0.05.
- Maximize $R^2$
- Minimize SSE.

# Choosing regression variables

## What not to do!

- Plot $Y$ against a particular predictor ($X_j$) and if it shows no noticeable relationship, drop it.
- Do a multiple linear regression on all the predictors and disregard all variables whose $p$ values are greater than 0.05.
- Maximize $R^2$
- Minimize SSE.

# Comparing regression models

## Sum of squared errors

$$\text{SSE} = \sum_{i=1}^{n} e_i^2$$

Minimizing SSE will always choose the model with the most predictors.

## Estimated residual variance

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-k-1}$$

where $k =$ no. predictors.
Minimizing $\hat{\sigma}^2$ works quite well for choosing predictors (but better methods to follow).

# Comparing regression models

## Sum of squared errors

$$\text{SSE} = \sum_{i=1}^{n} e_i^2$$

Minimizing SSE will always choose the model with the most predictors.

## Estimated residual variance

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-k-1}$$

where $k =$ no. predictors.

Minimizing $\hat{\sigma}^2$ works quite well for choosing predictors (but better methods to follow).

# Comparing regression models

## Sum of squared errors

$$\text{SSE} = \sum_{i=1}^{n} e_i^2$$

Minimizing SSE will always choose the model with the most predictors.

## Estimated residual variance

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-k-1}$$

where $k =$ no. predictors.

Minimizing $\hat{\sigma}^2$ works quite well for choosing predictors (but better methods to follow).

# Comparing regression models

However . . .

- $R^2$ does not allow for "degrees of freedom".
- Adding *any* variable tends to increase the value of $R^2$, even if that variable is irrelevant.

To overcome this problem, we can use *adjusted* $R^2$:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

**Maximizing $\bar{R}^2$ is equivalent to minimizing $\hat{\sigma}^2$.**

# Comparing regression models

However ...

- $R^2$ does not allow for "degrees of freedom".
- Adding *any* variable tends to increase the value of $R^2$, even if that variable is irrelevant.

To overcome this problem, we can use *adjusted* $R^2$:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

**Maximizing $\bar{R}^2$ is equivalent to minimizing $\hat{\sigma}^2$.**

# Comparing regression models

However . . .

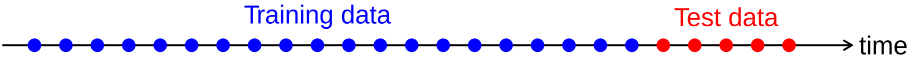
- $R^2$ does not allow for "degrees of freedom".
- Adding *any* variable tends to increase the value of $R^2$, even if that variable is irrelevant.

To overcome this problem, we can use *adjusted* $R^2$:

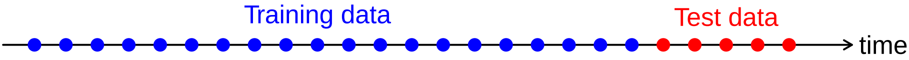
$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

**Maximizing $\bar{R}^2$ is equivalent to minimizing $\hat{\sigma}^2$.**

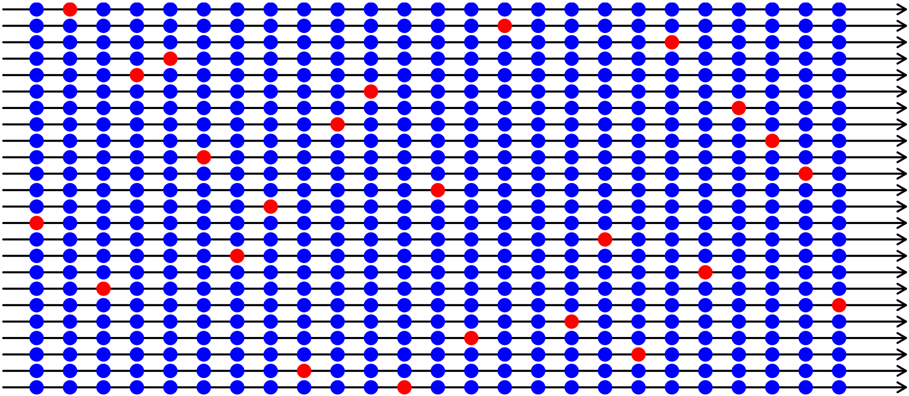
# Test sets



# Test sets




## Leave one-out cross-validation (LOOCV)



# LOO Cross-validation

Leave-one-out cross-validation (LOOCV) for regression can be carried out using the following steps.

1. Remove observation $i$ from the data set, and fit the model using the remaining data. Then compute the error ($e_i^* = y_i - \hat{y}_i$) for the omitted observation.
2. Repeat step 1 for $i = 1, \ldots, n$.
3. Compute the MSE from $\{e_1^*, \ldots, e_n^*\}$. We shall call this the CV.

The best model is the one with the smallest value of CV.

# LOO Cross-validation

Leave-one-out cross-validation (LOOCV) for regression can be carried out using the following steps.

1. Remove observation $i$ from the data set, and fit the model using the remaining data. Then compute the error ($e_i^* = y_i - \hat{y}_i$) for the omitted observation.
2. Repeat step 1 for $i = 1, \dots, n$.
3. Compute the MSE from $\{e_1^*, \dots, e_n^*\}$. We shall call this the CV.

The best model is the one with the smallest value of CV.

# LOO Cross-validation

Leave-one-out cross-validation (LOOCV) for regression can be carried out using the following steps.

1. Remove observation $i$ from the data set, and fit the model using the remaining data. Then compute the error ($e_i^* = y_i - \hat{y}_i$) for the omitted observation.
2. Repeat step 1 for $i = 1, \ldots, n$.
3. Compute the MSE from $\{e_1^*, \ldots, e_n^*\}$. We shall call this the CV.

The best model is the one with the smallest value of CV.

# LOO Cross-validation

Leave-one-out cross-validation (LOOCV) for regression can be carried out using the following steps.

1. Remove observation $i$ from the data set, and fit the model using the remaining data. Then compute the error ($e_i^* = y_i - \hat{y}_i$) for the omitted observation.
2. Repeat step 1 for $i = 1, \ldots, n$.
3. Compute the MSE from $\{e_1^*, \ldots, e_n^*\}$. We shall call this the CV.

The best model is the one with the smallest value of CV.

# LOO Cross-validation

Leave-one-out cross-validation (LOOCV) for regression can be carried out using the following steps.

1. Remove observation $i$ from the data set, and fit the model using the remaining data. Then compute the error ($e_i^* = y_i - \hat{y}_i$) for the omitted observation.
2. Repeat step 1 for $i = 1, \ldots, n$.
3. Compute the MSE from $\{e_1^*, \ldots, e_n^*\}$. We shall call this the CV.

The best model is the one with the smallest value of CV.

# LOOCV vs test sets

- CV has less bias
  - We repeatedly fit the statistical learning method using training data that contains $n - 1$ obs., i.e. almost all the data set is used
- LOOCV produces a less variable MSE
  - The validation approach produces different MSE when applied repeatedly due to randomness in the splitting process, while performing LOOCV multiple times will always yield the same results, because we split based on 1 obs. each time
- LOOCV is (usually) computationally intensive
  - We fit each model $n$ times!

# LOOCV for linear models

## Fitted values

$$\hat{\boldsymbol{Y}} = \mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the "hat matrix".

## Leave-one-out residuals

Let $h_1, \dots, h_n$ be the diagonal values of $\boldsymbol{H}$, then the cross-validation statistic is

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}[e_i/(1-h_i)]^2,$$

where $e_i$ is the residual obtained from fitting the model to all $n$ observations.

# LOOCV for linear models

## Fitted values

$$\hat{\boldsymbol{Y}} = \mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the “hat matrix”.

## Leave-one-out residuals

Let $h_1, \dots, h_n$ be the diagonal values of $\boldsymbol{H}$, then the cross-validation statistic is

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}[e_i/(1-h_i)]^2,$$

where $e_i$ is the residual obtained from fitting the model to all $n$ observations.

# Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2(k+1)$$

where $L$ is the likelihood and $k$ is the number of predictors in the model.

- This is a *penalized likelihood* approach.
- *Minimizing* the AIC gives the best model for prediction.
- AIC penalizes terms more heavily than $\bar{R}^2$.
- Minimizing the AIC is asymptotically equivalent to minimizing MSE via leave-one-out cross-validation.

# Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2(k+1)$$

where $L$ is the likelihood and $k$ is the number of predictors in the model.

- This is a *penalized likelihood* approach.
- *Minimizing* the AIC gives the best model for prediction.
- AIC penalizes terms more heavily than $\bar{R}^2$.
- Minimizing the AIC is asymptotically equivalent to minimizing MSE via leave-one-out cross-validation.

# Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2(k+1)$$

where $L$ is the likelihood and $k$ is the number of predictors in the model.

- This is a *penalized likelihood* approach.
- *Minimizing* the AIC gives the best model for prediction.
- AIC penalizes terms more heavily than $\bar{R}^2$.
- Minimizing the AIC is asymptotically equivalent to minimizing MSE via leave-one-out cross-validation.

# Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2(k+1)$$

where $L$ is the likelihood and $k$ is the number of predictors in the model.

- This is a *penalized likelihood* approach.
- *Minimizing* the AIC gives the best model for prediction.
- AIC penalizes terms more heavily than $\bar{R}^2$.
- Minimizing the AIC is asymptotically equivalent to minimizing MSE via leave-one-out cross-validation.

# Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2(k+1)$$

where $L$ is the likelihood and $k$ is the number of predictors in the model.

- This is a *penalized likelihood* approach.
- *Minimizing* the AIC gives the best model for prediction.
- AIC penalizes terms more heavily than $\bar{R}^2$.
- Minimizing the AIC is asymptotically equivalent to minimizing MSE via leave-one-out cross-validation.

# Corrected AIC

For small values of $n$, the AIC tends to select too many predictors, and so a bias-corrected version of the AIC has been developed.

$$\text{AIC}_\text{C} = \text{AIC} + \frac{2(k+2)(k+3)}{n-k-1}$$

As with the AIC, the $\text{AIC}_\text{C}$ should be minimized.

# Schwartz Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (k+1)\log(n)$$

where $L$ is the likelihood and $k$ is the number of predictors in the model.

- *Minimizing* the BIC gives the best model for prediction.
- BIC penalizes terms more heavily than AIC
- Also called SBIC and SC.
- Minimizing BIC is asymptotically equivalent to leave-$v$-out cross-validation when $v = n[1 - 1/(\log(n) - 1)]$.

# Schwartz Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (k+1)\log(n)$$

where $L$ is the likelihood and $k$ is the number of predictors in the model.

- *Minimizing* the BIC gives the best model for prediction.
- BIC penalizes terms more heavily than AIC
- Also called SBIC and SC.
- Minimizing BIC is asymptotically equivalent to leave-$v$-out cross-validation when $v = n[1 - 1/(log(n) - 1)]$.

## Schwartz Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (k+1)\log(n)$$

where $L$ is the likelihood and $k$ is the number of predictors in the model.

- *Minimizing* the BIC gives the best model for prediction.
- BIC penalizes terms more heavily than AIC
- Also called SBIC and SC.
- Minimizing BIC is asymptotically equivalent to leave-$v$-out cross-validation when $v = n[1 - 1/(log(n) - 1)]$.

# Schwartz Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (k+1)\log(n)$$

where $L$ is the likelihood and $k$ is the number of predictors in the model.

- *Minimizing* the BIC gives the best model for prediction.
- BIC penalizes terms more heavily than AIC
- Also called SBIC and SC.
- Minimizing BIC is asymptotically equivalent to leave-$v$-out cross-validation when $v = n[1 - 1/(log(n) - 1)]$.

# Schwartz Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (k+1)\log(n)$$

where $L$ is the likelihood and $k$ is the number of predictors in the model.

- *Minimizing* the BIC gives the best model for prediction.
- BIC penalizes terms more heavily than AIC
- Also called SBIC and SC.
- Minimizing BIC is asymptotically equivalent to leave-$v$-out cross-validation when $v = n[1 - 1/(log(n) - 1)]$.

## Best subsets regression

- Fit all possible regression models using one or more of the predictors.
- Choose the best model based on CV (or an asymptotic equivalent: AIC, AICc).

## Warning!

- If there are a large number of predictors, this is not possible.
  For example, 44 predictors leads to 18 trillion possible models!

# Choosing regression variables

## Best subsets regression

- Fit all possible regression models using one or more of the predictors.
- Choose the best model based on CV (or an asymptotic equivalent: AIC, AICc).

## Warning!

- If there are a large number of predictors, this is not possible.
  For example, 44 predictors leads to 18 trillion possible models!

# Choosing regression variables

## Best subsets regression

- Fit all possible regression models using one or more of the predictors.
- Choose the best model based on CV (or an asymptotic equivalent: AIC, AICc).

## Warning!

- If there are a large number of predictors, this is not possible.
  For example, 44 predictors leads to 18 trillion possible models!

# Choosing regression variables

## Best subsets regression

- Fit all possible regression models using one or more of the predictors.
- Choose the best model based on CV (or an asymptotic equivalent: AIC, AICc).

## Warning!

- If there are a large number of predictors, this is not possible.
  For example, 44 predictors leads to 18 trillion possible models!

# Choosing regression variables

## Backwards stepwise regression

- Start with a model containing all variables.
- Try subtracting one variable at a time. Keep the model if it has lower CV, AICc or AIC.
- Iterate until no further improvement.

Notes:

# Choosing regression variables

## Backwards stepwise regression

- Start with a model containing all variables.
- Try subtracting one variable at a time. Keep the model if it has lower CV, AICc or AIC.
- Iterate until no further improvement.

**Notes:**

- Stepwise regression is not guaranteed to lead to the best possible model.

# Choosing regression variables

## Backwards stepwise regression

- Start with a model containing all variables.
- Try subtracting one variable at a time. Keep the model if it has lower CV, AICc or AIC.
- Iterate until no further improvement.

## Notes:

- Stepwise regression is not guaranteed to lead to the best possible model.
- If you are trying several different models, use the CV, AICc or AIC value to select between them.

# Choosing regression variables

## Backwards stepwise regression

- Start with a model containing all variables.
- Try subtracting one variable at a time. Keep the model if it has lower CV, AICc or AIC.
- Iterate until no further improvement.

## Notes:

- Stepwise regression is not guaranteed to lead to the best possible model.
- If you are trying several different models, use the CV, AICc or AIC value to select between them.

# Choosing regression variables

## Backwards stepwise regression

- Start with a model containing all variables.
- Try subtracting one variable at a time. Keep the model if it has lower CV, AICc or AIC.
- Iterate until no further improvement.

**Notes:**

- Stepwise regression is not guaranteed to lead to the best possible model.
- If you are trying several different models, use the CV, AICc or AIC value to select between them.

# Choosing regression variables

## Backwards stepwise regression

- Start with a model containing all variables.
- Try subtracting one variable at a time. Keep the model if it has lower CV, AICc or AIC.
- Iterate until no further improvement.

**Notes:**

- Stepwise regression is not guaranteed to lead to the best possible model.
- If you are trying several different models, use the CV, AICc or AIC value to select between them.

# Outline

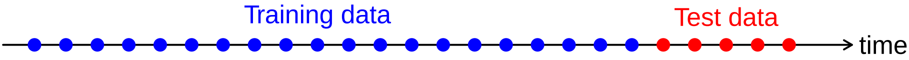
1 Choosing regression variables

2 Cross-validation

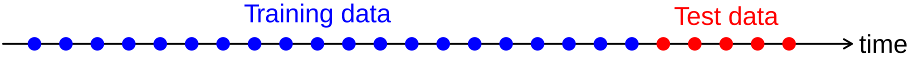
# Cross-validation

- The computational trick for computing LOOCV for linear models doesn't work in other contexts.
- In general, LOOCV is too computationally intensive. So we use $k$-fold CV instead (where $k = 5$ and $k = 10$ are common choices).

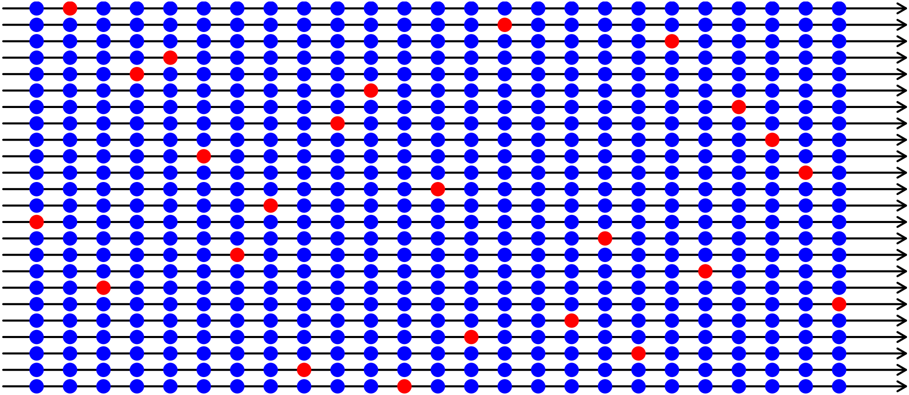
# Cross-validation



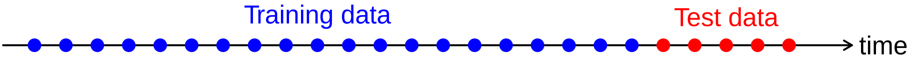
# Cross-validation



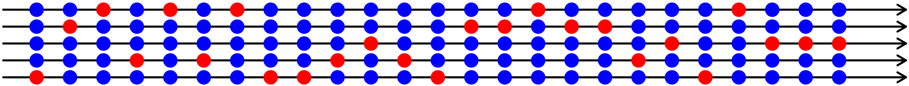

## Leave one-out cross-validation (LOOCV)



# Cross-validation




## 5-fold cross-validation



# $k$-fold Cross-validation

- Divide the data set into $k$ different parts.
- Remove one part, fit the model on the remaining $k-1$ parts, and compute the MSE on the omitted part.
- Repeat $k$ times taking out a different part each time

By averaging the $k$ MSEs we get an estimated validation (test) error rate for new observations.

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i$$

LOOCV is a special case where $k = n$.

# $k$-fold Cross-validation

- Divide the data set into $k$ different parts.
- Remove one part, fit the model on the remaining $k-1$ parts, and compute the MSE on the omitted part.
- Repeat $k$ times taking out a different part each time

By averaging the $k$ MSEs we get an estimated validation (test) error rate for new observations.

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

LOOCV is a special case where $k = n$.

# $k$-fold Cross-validation

- Divide the data set into $k$ different parts.
- Remove one part, fit the model on the remaining $k-1$ parts, and compute the MSE on the omitted part.
- Repeat $k$ times taking out a different part each time

By averaging the $k$ MSEs we get an estimated validation (test) error rate for new observations.

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

LOOCV is a special case where $k = n$.

# $k$-fold Cross-validation

- Each training set is only $(k-1)/k$ as big as the original data set. So the estimates of prediction error will be biased upwards.
- Bias minimized when $k = n$ (LOOCV).
- But variance increases with $k$ (as there are overlapping observations in each part).
- $k = 5$ or $k = 10$ provide a good compromise for this bias-variance tradeoff.