ETC3250: Resampling

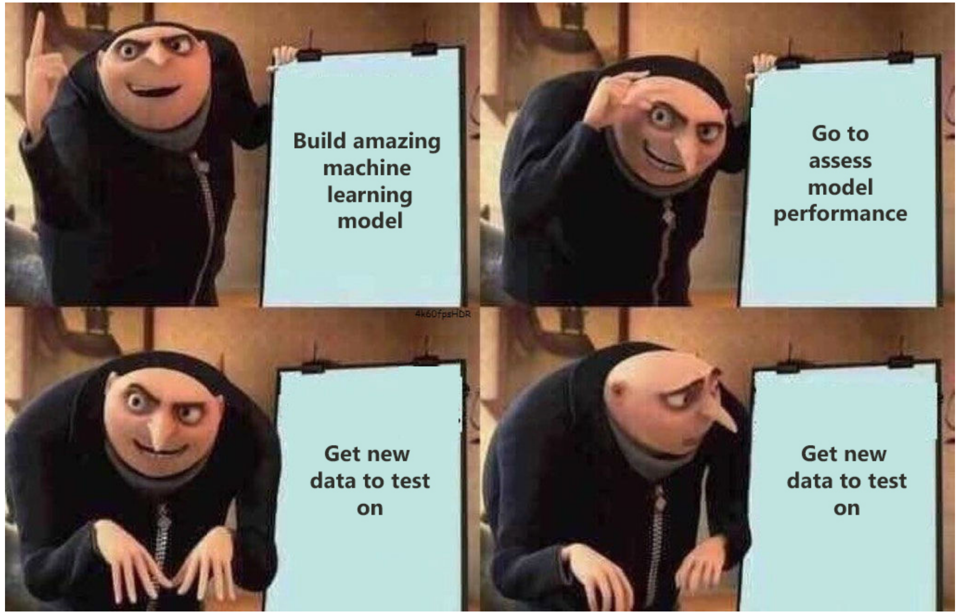
Semester 1, 2020

Professor Di Cook

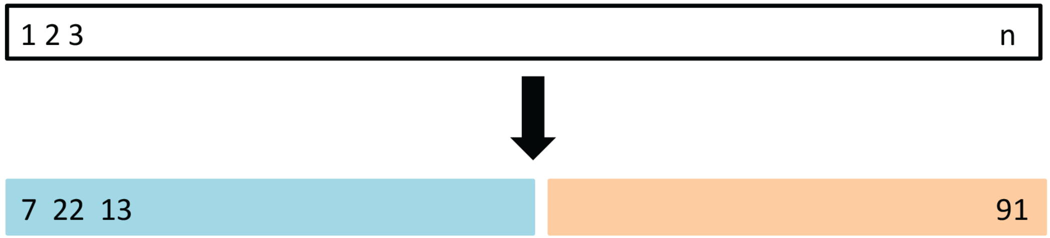
Econometrics and Business Statistics
Monash University

Week 3 (b)

Model assessment



Validating models



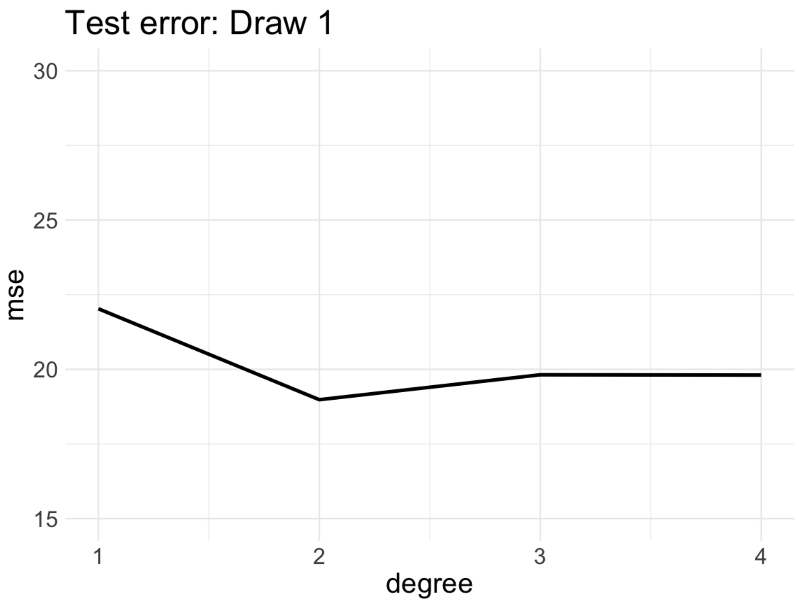
A set of n observations are randomly split into a training set (blue, containing observations 7, 22, 13, ...) and a validation set (yellow, all other observations not in training set).

Drawback: Only one split of data made, may not adequately estimate test error.

Validating models

Want to choose best degree of polynomial, for $\text{mpg} = \beta_0 + \beta_1 f(\text{horsepower}) + \varepsilon$

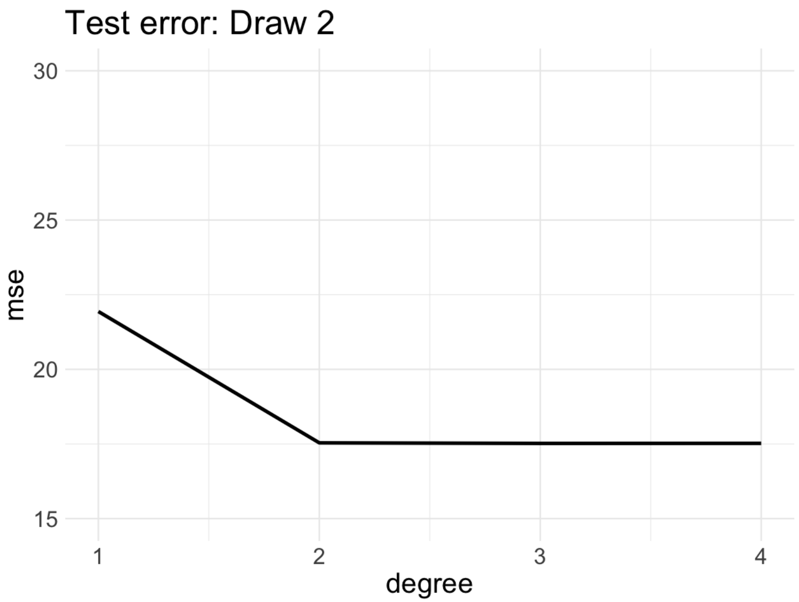
```
##  [1]  2  4  7  9 10 11 12 14 15 16 18 21 23
```



Validating models

Want to choose best degree of polynomial, for
$\text{mpg} = \beta_0 + \beta_1 f(\text{horsepower}) + \varepsilon$

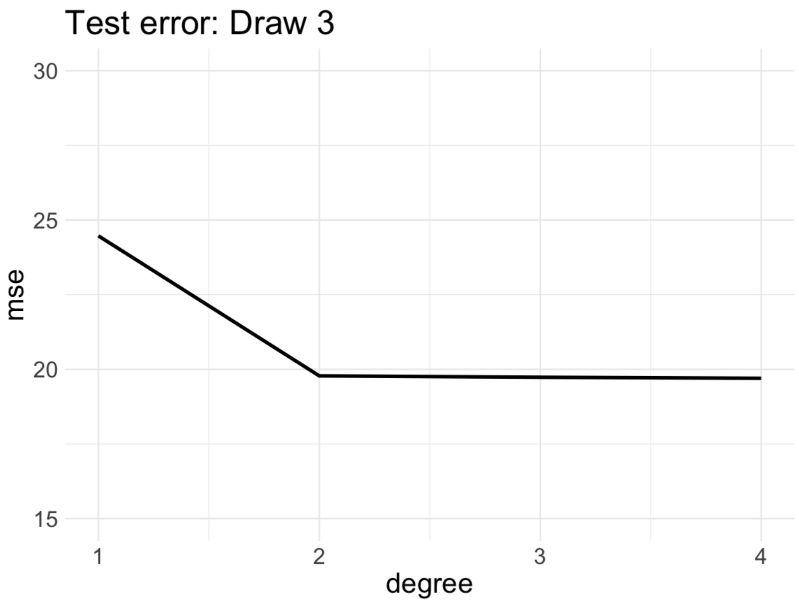
```
##  [1]  3  4  5  6  9 10 13 14 17 20 23 24 25
```



Validating models

Want to choose best degree of polynomial, for $\text{mpg} = \beta_0 + \beta_1 f(\text{horsepower}) + \varepsilon$

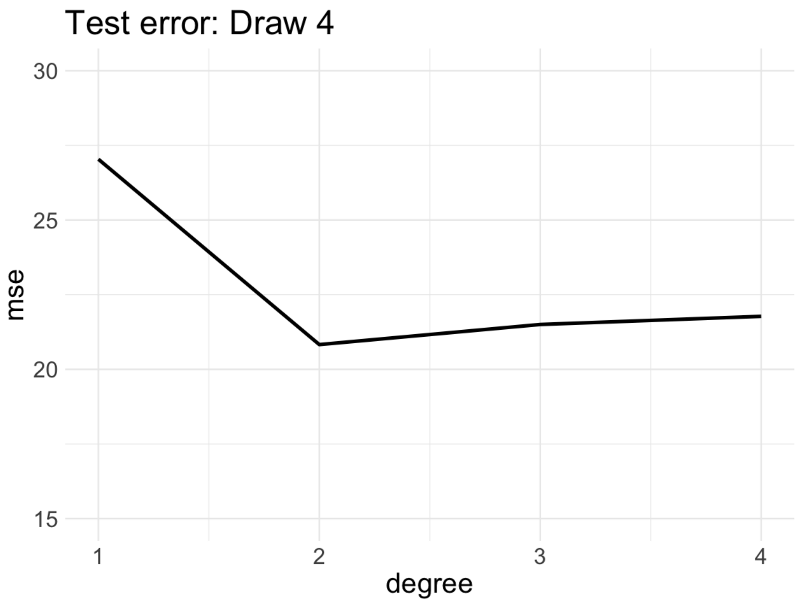
```
##  [1]  1  3  4  5  8  9 10 12 13 14 15 16 20
```



Validating models

Want to choose best degree of polynomial, for
$\text{mpg} = \beta_0 + \beta_1 f(\text{horsepower}) + \varepsilon$

```
##  [1]  1  2  3  4  5 11 13 15 18 19 21 23 24
```



The variability between different draws of test sets can be large. This can provide poor choice of model, or misleading estimate of error.

LOOCV

Leave-one-out (LOOCV) cross-validation: n validation sets, each with ONE observation left out.

LOOCV

Leave-one-out (LOOCV) cross-validation: n validation sets, each with ONE observation left out. For each set, $i = 1, \ldots, n$, compute the MSE_i.

The LOOCV estimate for the test MSE is the average of these n test error estimates:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} MSE_i$$

LOOCV

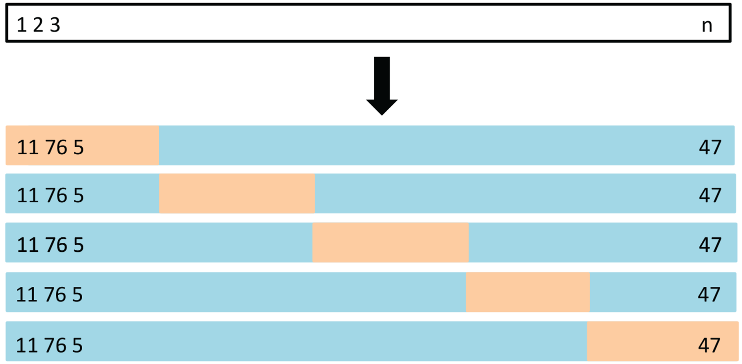
There is a computational shortcut, for linear or polynomial models, where not all n models need to be fitted.

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y})^2}{1 - h_i}$$

where $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'}^{n} (x_{i'} - \bar{x})^2}$ (known as *leverage* from regression diagnostics).

k-fold cross validation

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



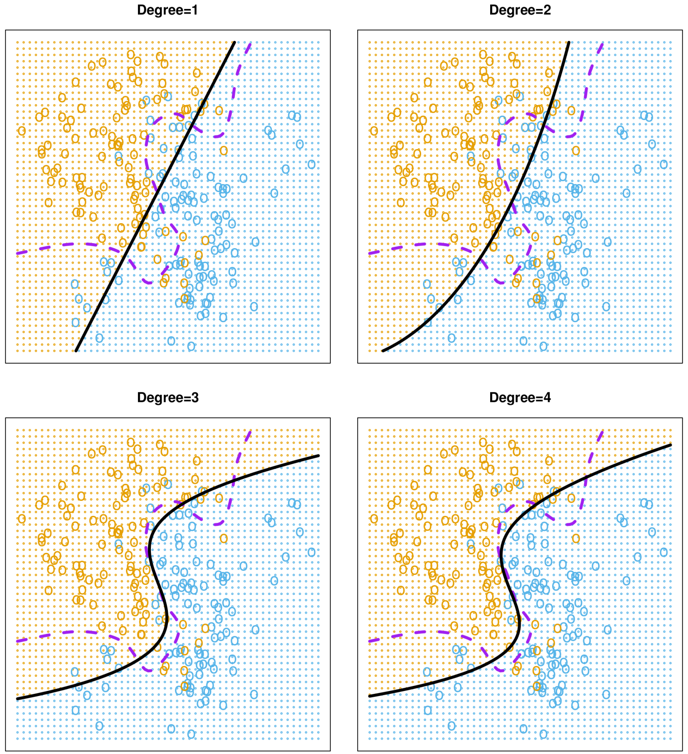
k-fold cross validation

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

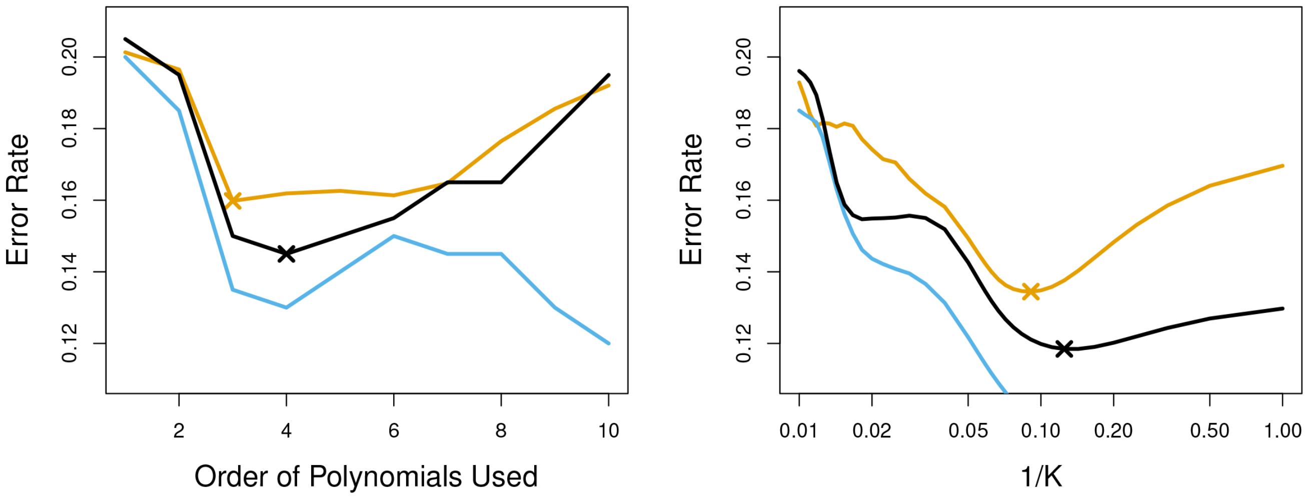
$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{n}\frac{(y_i-\hat{y})^2}{1-h_i}$$

- LOOCV is a special case of k-fold cross-validation.
- Bias-variance trade-off:
  - one vaidation set overestimates test error, LOOCV approximately unbiased estimates, k-fold CV intermediate
  - LOOCV has higher variance than does k-fold CV
  - choice of $k = 5$ or 10 is a compromise

Classification



Classification



Black line is 10-fold CV; training and TRUE test error for different choices of polynomial (left) and KNN classifier (right).

Bootstrap procedure

- Draw B independent bootstrap samples $X^{*(1)}, \ldots, X^{*(B)}$ from $\hat{P}$:

$$X_1^{*(b)}, \ldots, X_n^{*(b)} \sim \hat{P} \quad b = 1, \ldots, B.$$

- Evaluate the bootstrap replications:

$$\hat{\theta}^{*(b)} = s(X^{*(b)}) \quad b = 1, \ldots, B.$$

- Estimate the quantity of interest from the distribution of the $\hat{\theta}^{*(b)}$

Example - boostrap model

- Fit the model on a set of bootstrap samples, and then keep track of how well it predicts the original dataset

$$\mathrm{Err}_{\mathrm{boot}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{f}^{*b}(x_i))$$

- Each of these bootstrap data sets is created by sampling with replacement, and is the same size as our original dataset. As a result some observations may appear more than once in a given bootstrap data set and some not at all.

Out of bag (OOB) error

For estimating error, only use predictions from bootstrap samples not containing that observation. The leave-one-out bootstrap estimate of prediction error can be defined as

$$\mathrm{Err}_{\text{loo-boot}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

where C^{-i} is the set of indices of the bootstrap samples b that do not contain observation i.

Uses and variants of the boostrap

- Common uses:
  - Computing standard errors for complex statistics
  - Prediction error estimation
  - Bagging (Bootstrap aggregating) ML models
- Types of bootstrap based on different assumptions:
  - block bootstrap
  - sieve bootstrap
  - smooth bootstrap
  - residual bootstrap
  - wild bootstrap

👨‍💻 Made by a human with a computer

Slides at https://iml.numbat.space.

Code and data at https://github.com/dicook/Business_Analytics.

Created using R Markdown with flair by **xaringan**, and **kunoichi** (female ninja) style.