MONASH University

ETC3250/5250: Introduction to Machine Learning

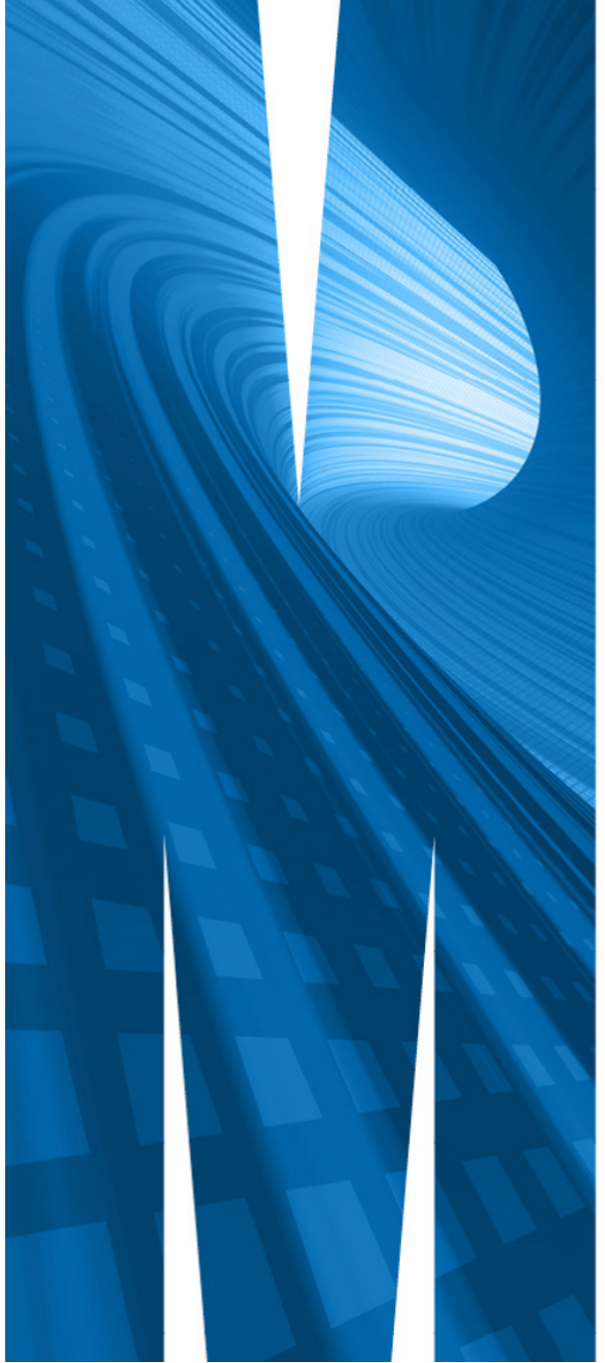
Model assessment

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 9a



```
library(statquotes)
search_quotes(search="Holdane", fuzzy=TRUE)
```

```
 ## In scientific thought we adopt the simplest theory which will explain
## all the facts under consideration and enable us to predict new facts of
     ## the same kind. The catch in this criterion lies in the world
 ## ``simplest.'' It is really an aesthetic canon such as we find implicit
## in our criticisms of poetry or painting. The layman finds such a law as
## $dx/dt = K(d^2x/dy^2)$ much less simple than "it oozes," of which it is
 ## the mathematical statement. The physicist reverses this judgment, and
   ## his statement is certainly the more fruitful of the two, so far as
 ## prediction is concerned. It is, however, a statement about something
 ## very unfamiliar to the plainman, namely, the rate of change of a rate
                                  ## of change.
 ## --- John Burdon Sanderson Haldane (1892--1964) Possible Worlds, 1927.
```

```
statquote(source="Box")
```

```
## It is the data that are real (they actually happened!) The model is a
  ## hypothetical conjecture that might or might not summarize and/or
              ## explain important features of the data
                    ## --- George E. P. Box
```

Know your data ✈️

Quantitative or qualitative response? Predictors all quantitative? Do you have independent observations?

Plot your data 🖼️

Is there a relationship between response and predictors? Is the relationship linear? Are boundaries linear? Is variability heterogeneous? Are groups distinct? Are there unusual observations?

Fit a versatile model 💻

Compute and plot model diagnostics. Where doesn't the model do well? How can it be refined?

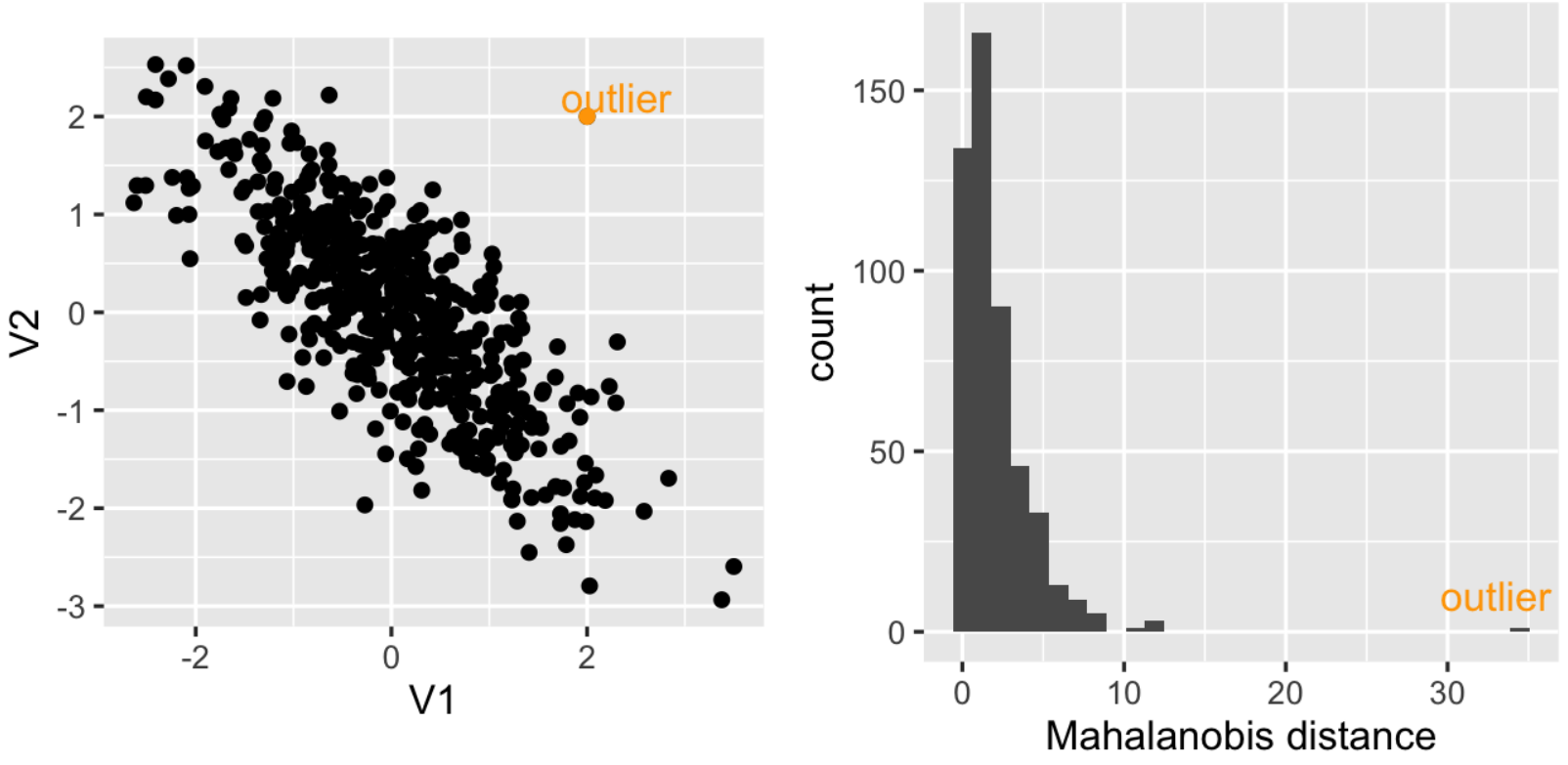
Check for missing values 🔥

Do some variables have too many missings to use them? Do some observations have too many missings to use them? What would be a useful imputation method to fix the sporadic missing value?

Data quality

Multivariate outliers

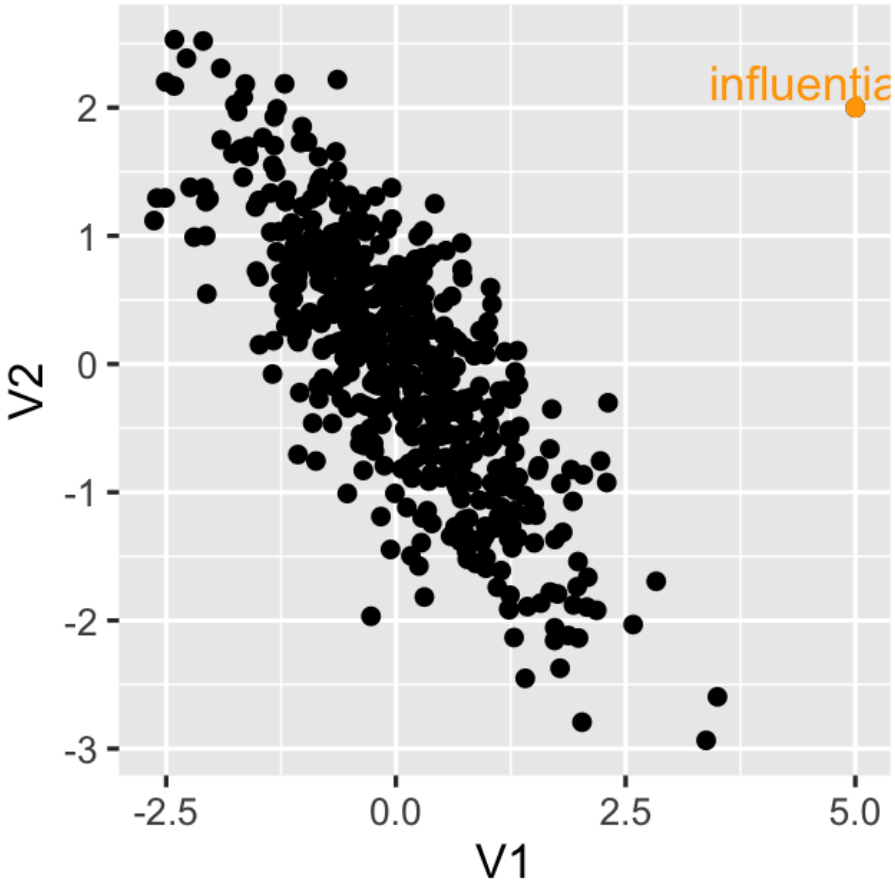
Mahalanobis distance measures the distance from the mean, relative to the variance-covariance matrix, and is useful for outlier detection: $D^2 = (X - \mu)' \Sigma^{-1} (X - \mu)$

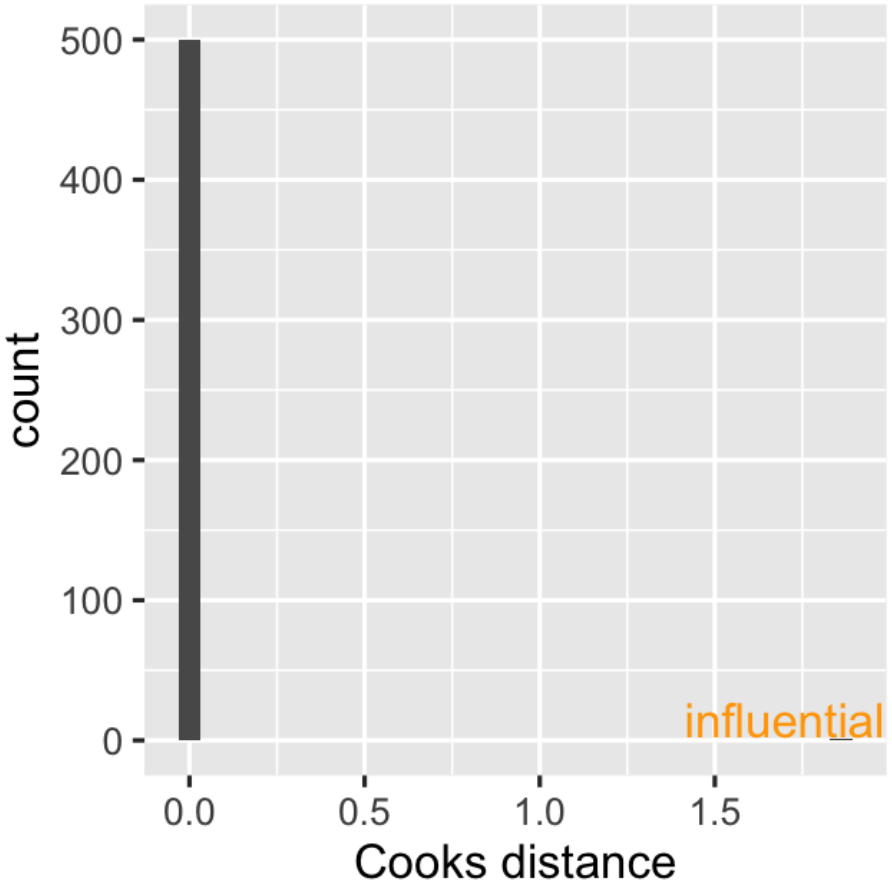


Related to "leverage" in regression diagnostics.

Influential observations

Cook's distance measures the change in the model estimates due to the observation: $D_i = \frac{e_i^2}{MSE \times p} \frac{h_i}{(1-h_i)^2}$ where h_i is the leverage of observation i





Model choice and comparison

Comparing statistical models using ANOVA

- Models are nested when one model is a particular case of the other model
  - Model 1: $\hat{y} = \beta_0 + \beta_1 X_1$
  - Model 2: $\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  - Model 3: $\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
- Nested models can be compared using ANOVA F test

$$F = \frac{(SSE_{reduced} - SSE_{full})/(p-k)}{SSE_{full}/(n-p-1)} \sim F_{p-k,n-p-1}$$

Material adapted from YaRrr! The Pirate's Guide to R by Nathaniel D. Phillips and Professor Cécile Ané Stat 572 slides.

Example

```
## Analysis of Variance Table
##
## Model 1: read ~ 1
## Model 2: read ~ country + year0 + television + math
## Model 3: read ~ country * year0 + country * television + country * math
## Model 4: read ~ country * year0 * television + country * math
## Model 5: read ~ country * year0 * television * math
##   Res.Df       RSS Df Sum of Sq          F    Pr(>F)
## 1  44837 510140630
## 2  44830 143570284  7 366570346 16754.0671 < 2.2e-16 ***
## 3  44818 140516614 12   3053670    81.4145 < 2.2e-16 ***
## 4  44813 140394227  5    122387     7.8312 2.226e-07 ***
## 5  44798 140022447 15    371780     7.9297 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remember the confusion table

		true	
		C1 (positive)	C2 (negative)
pred-	C1	*a*	*b*
icted	C2	*c*	*d*

- Sensitivity: *a/(a+c)* (true positive, recall)
- Specificity: *d/(b+d)* (true negative)
- False positive: *c/(a+c)*
- False negative: *b/(b+d)* (1-specificity)

From a quantitative prediction, a cutoff needs to be used to create a categorical prediction.

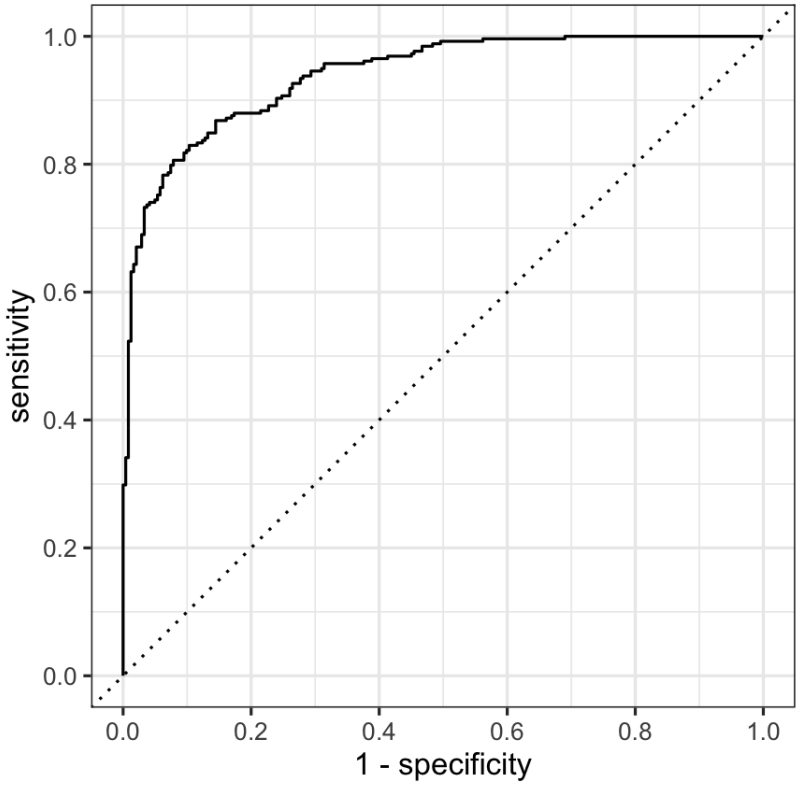
```
library(tidyverse)
library(yardstick)
options(digits=2)
glimpse(two_class_example)
```

```
## Rows: 500
## Columns: 4
## $ truth     <fct> Class2, Class1, Clas
## $ Class1    <dbl> 0.00359, 0.67862, 0.
## $ Class2    <dbl> 1.0e+00, 3.2e-01, 8.
## $ predicted <fct> Class2, Class1, Clas
```

Set threshold to 0.5

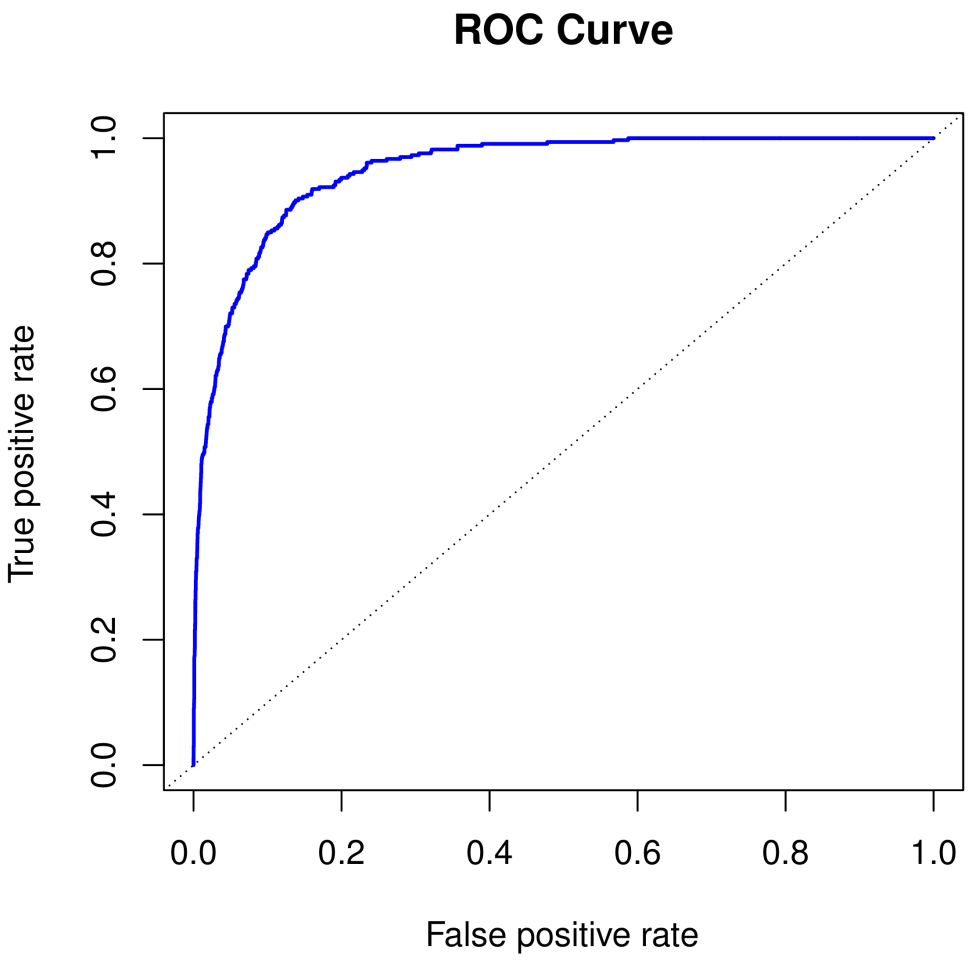
```
##          Truth
## Prediction Class1 Class2
##     Class1    227     31
##     Class2     50    192
```

sensitivity = 0.82, 1-specificity = 0.14



Your turn: Set the threshold to be 0.75, re-compute the confusion matrix, and sensitivity, specificity.

ROC



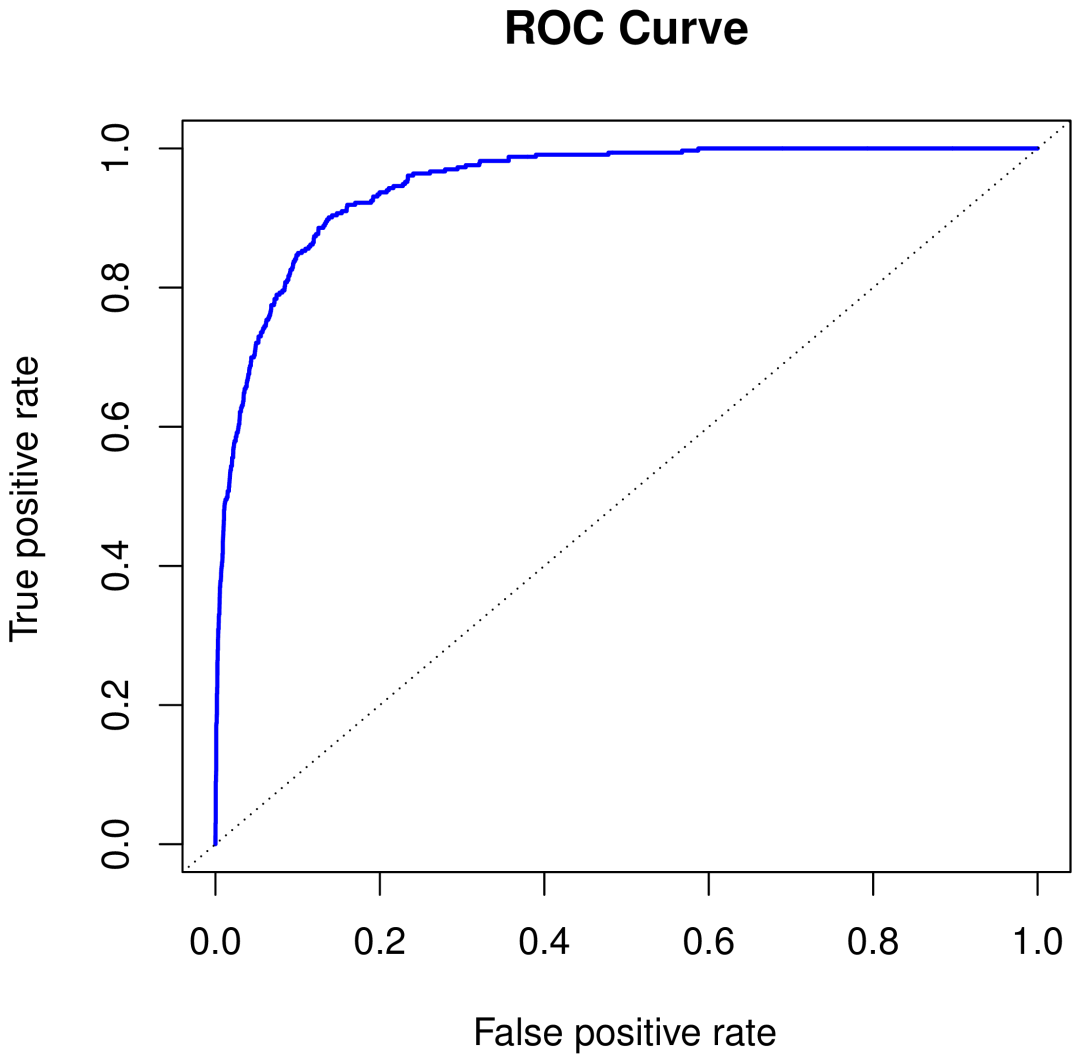
The true positive rate is the sensitivity: the fraction of defaulters that are correctly identified, using a given threshold value.

The false positive rate is 1-specificity: the fraction of non-defaulters that we classify incorrectly as defaulters, using that same threshold value.

The dotted line is "no information" classifier; class and predictor are not associated.

The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate.

Credit example in textbook figure 4.8



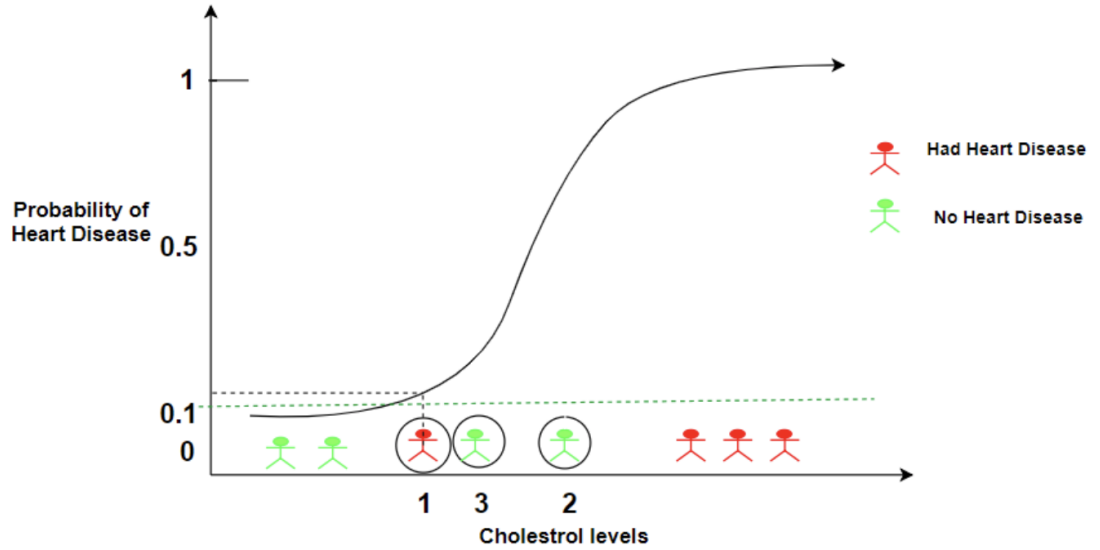
If the classifier returns a prediction between 0 and 1, interpret as the probability of a positive, then threshold (split the data) at different values, e.g. 0.1, 0.2, 0.3, 0.4, 0.5, ...

Compute the confusion table for each split, record the sensitivity and specificity and plot the resulting numbers.

Really nice explanation by Parul Pandey here and video by Josh Starmer here.

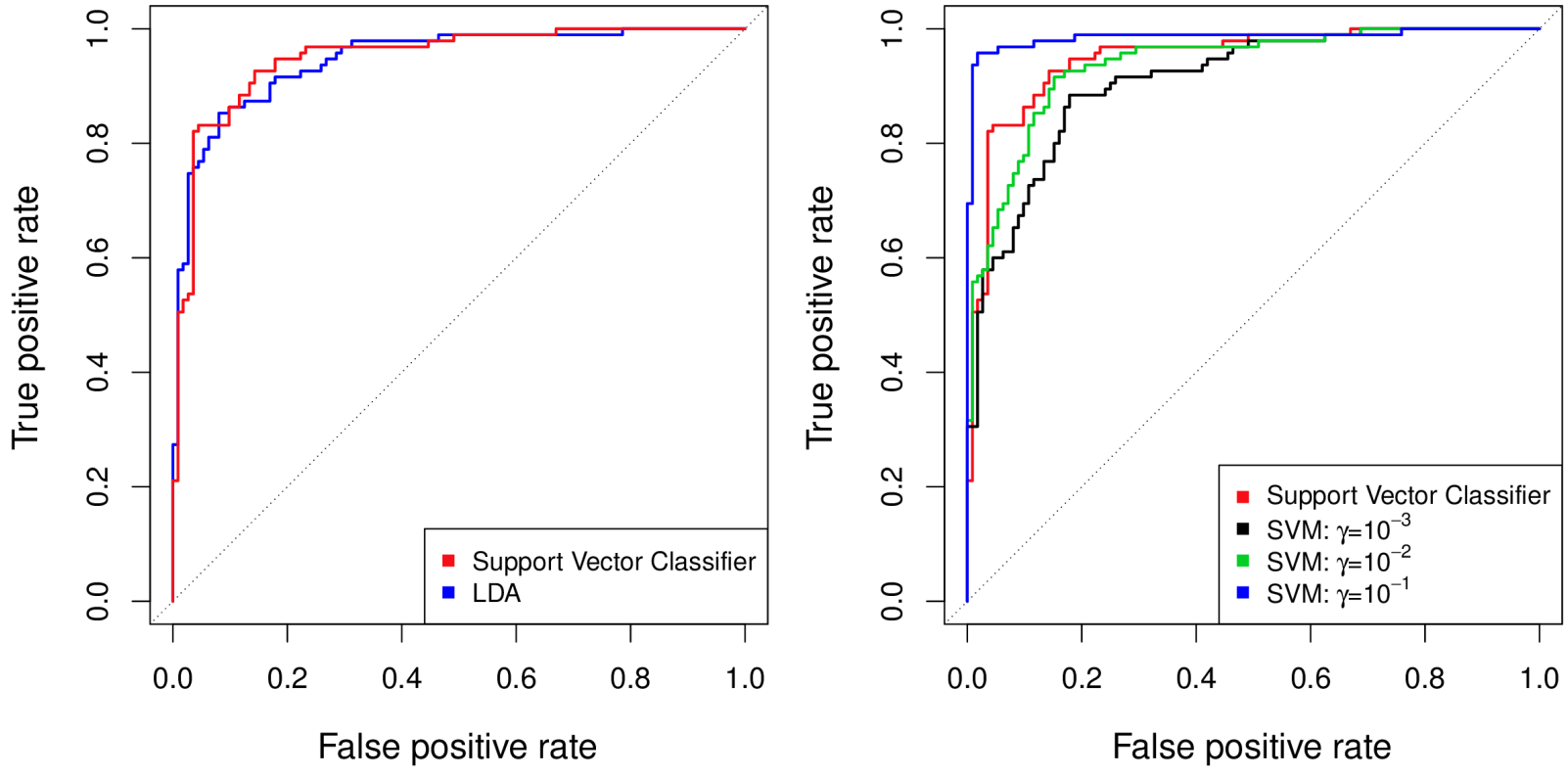
- **Setting the Threshold to 0.1**

This would correctly identify all people who have heart disease. The person labeled 1 is also correctly classified to be a heart patient.



However, it would also increase the number of False Positives since now person 2 and 3 will be wrongly classified as having heart disease.

ROC for classification



- LDA and SVM similar (example on left).
- SVM radial basis with $\gamma = 10^{-1}$ is the best (example on the right).

Fig 9.10

Quantifying uncertainty

Utilising bagging

Remember the vote matrix available from random forests:

$$V = (V_1 V_2 \dots V_K)$$
$$= \begin{bmatrix} p_{11} p_{12} & \dots & p_{1K} \\ p_{21} p_{22} & \dots & p_{2K} \\ \dots\dots & & \dots \\ p_{n1} p_{n2} & \dots & p_{nK} \end{bmatrix}$$

With bagging, multiple out of bag predictions produces uncertainty measure for each observation. It's possible that observations with higher uncertainty are outliers.

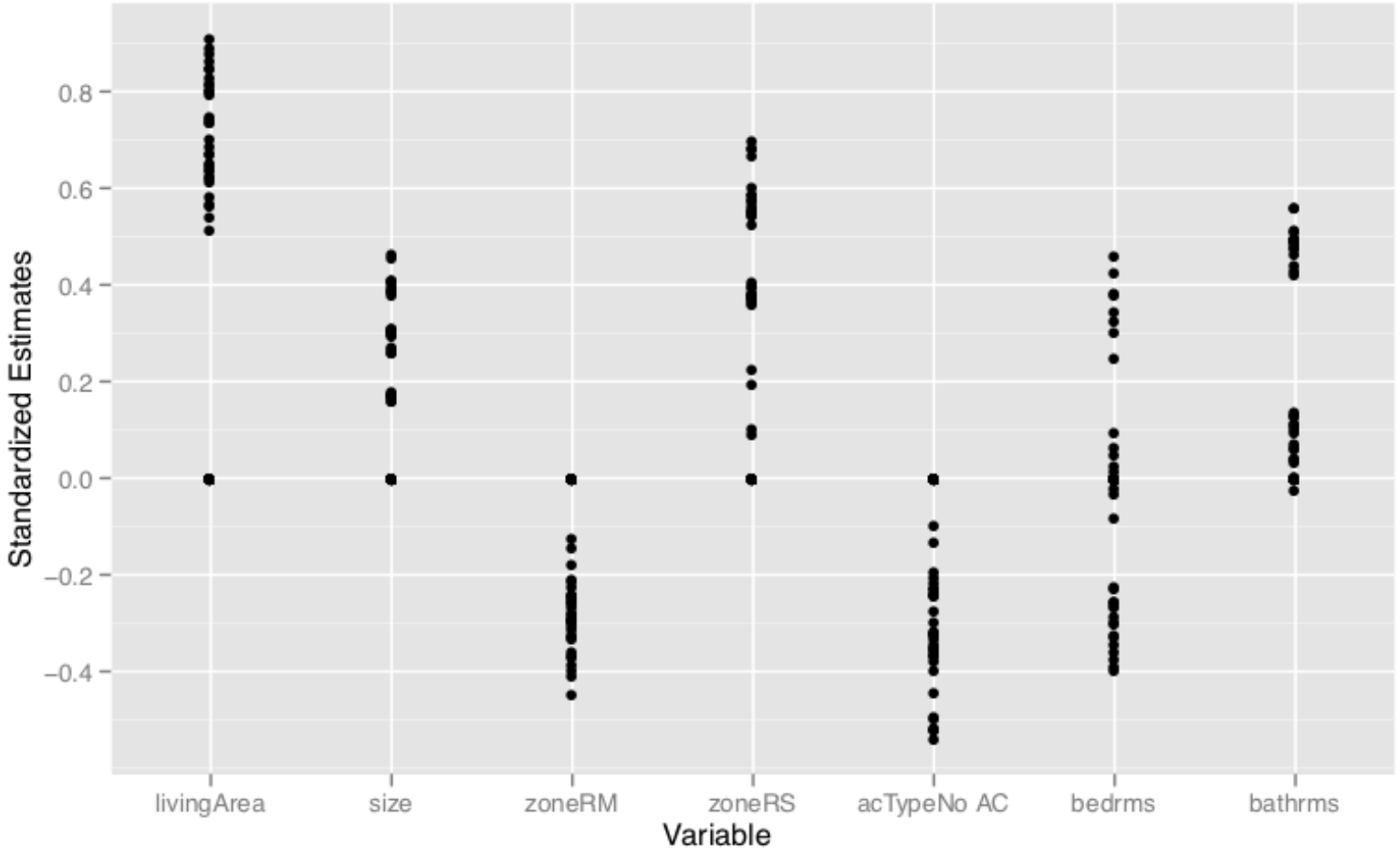
Variable importance

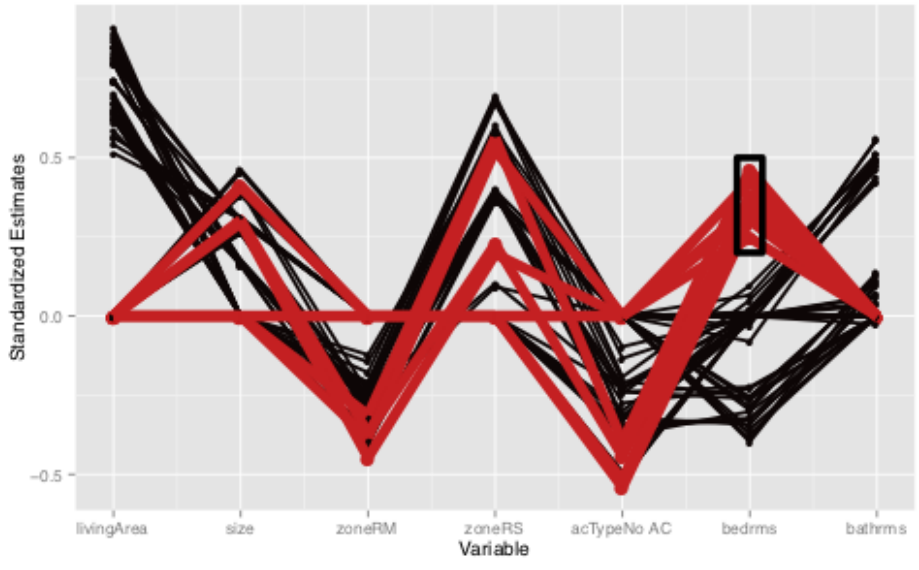
- Working with standardised variables helps, because magnitude of coefficients is then directly interpreted as importance
- Permutation approach in random forests is useful more broadly. Compare magnitude of coefficients between models built on original and permuted variable.
- Effect of one predictor with the response can depend on their relationship with one another. Called multicollinearity in regression.

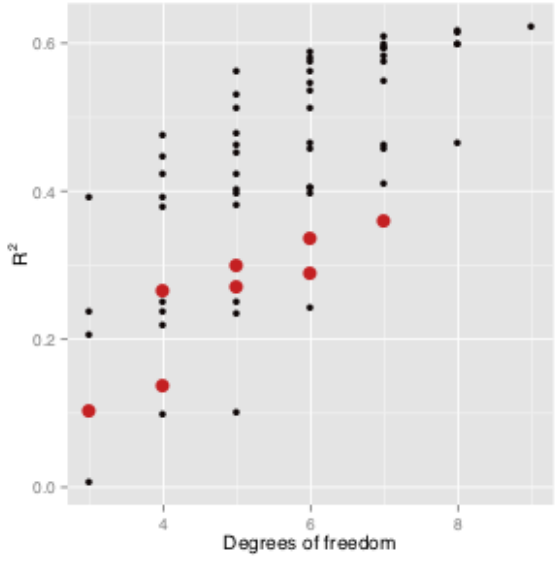
Beyond the optimal

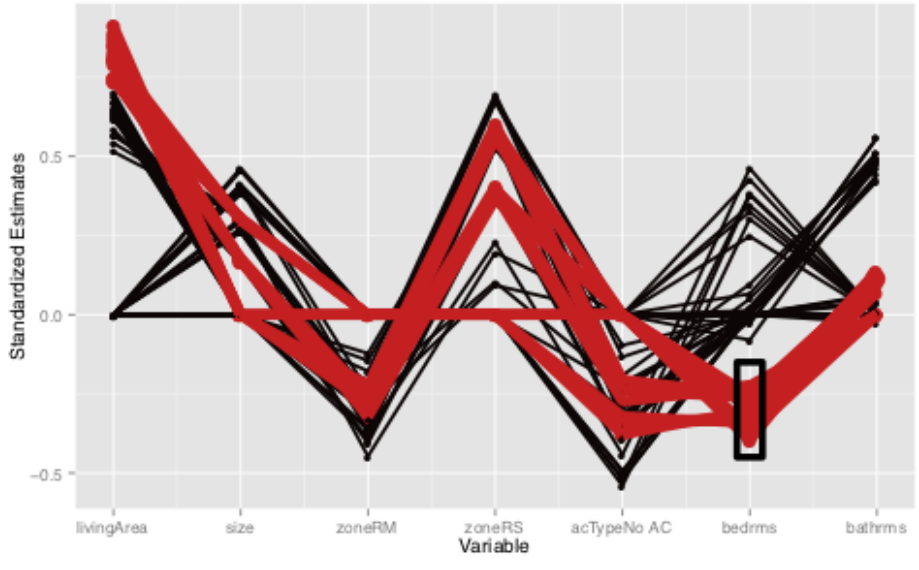
Bigger picture

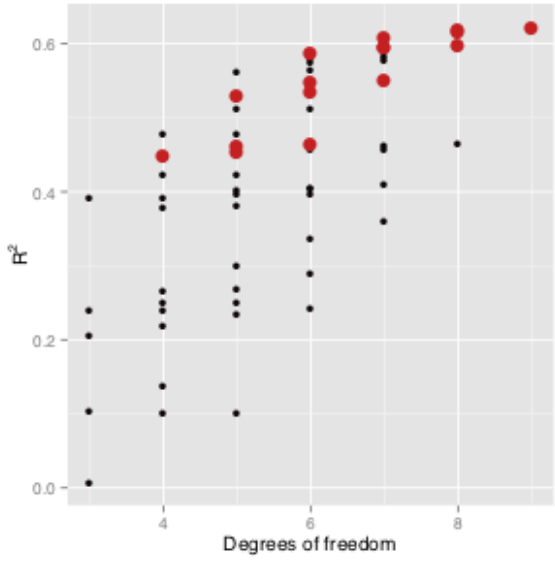
All possible model fits to housing data with 7 variables, from Wickham et al (2015) Removing the Blindfold











Three typical estimates for bedrooms: big positive, close to 0, big negative.

Models with big positive coefficients for bedrooms tend to have weaker fits. They tend to occur with models that have no livingArea contribution, and more negative coefficients for zoneRM, and no air con.

Models with big negative coefficients on bedrooms tend to have stronger fits. All contrast with livingArea (high positive coefficients).

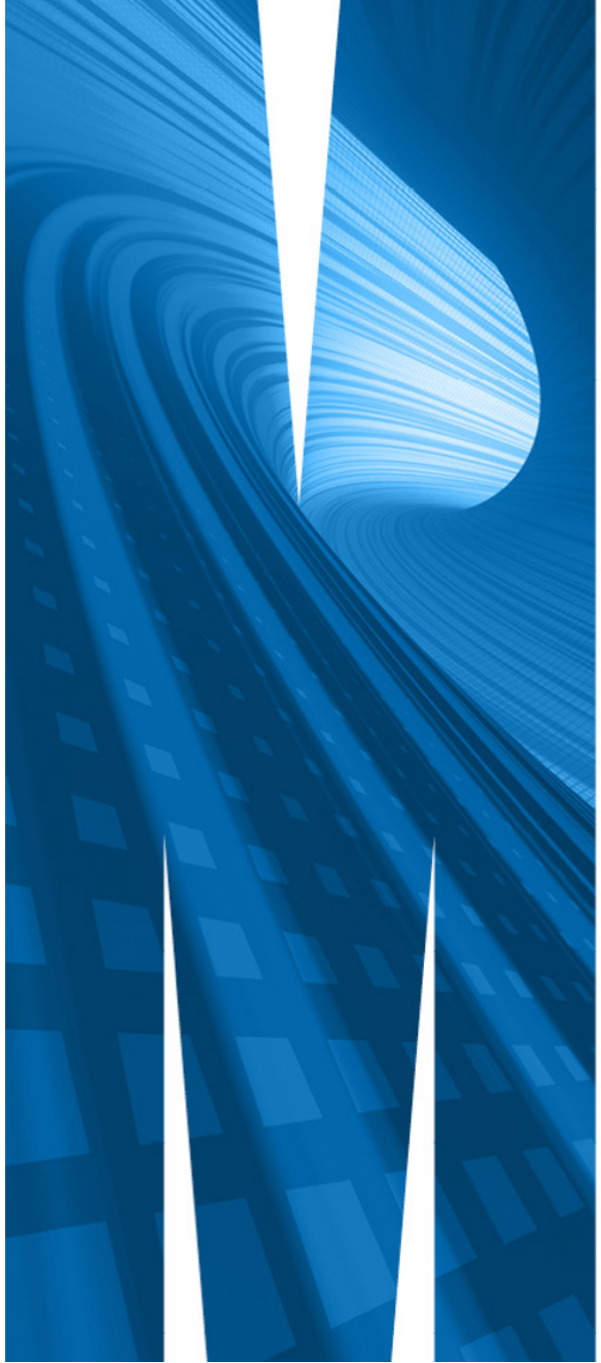
If bedrooms contribute to the model, bathrooms do not.

Model choice - robustness of conclusions

Whatever way you model the data, the interpretations should be consistent.

- Bias can explain difference in predictions between models, flexible vs inflexible can provide a spectrum on what the data predicts.
- Broad changes in a model when some cases or some variables are not used, should evoke suspicions (your "spidey sense").
- Model fit statistics are a measure of predictive power. A weak model can still be useful if there is a large cost involved.

MONASH University



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 9a