ACSC/STAT 3740
Predictive Analytics

WINTER 2025
Toby Kenney

In Class Examples

Story Time — Rumpelstiltzkin

Analysis Process

1. Identify statistical problem.
2. Determine and assess useable data.
3. Explore data for possible issues and approaches.
4. Research any relevant subject knowledge.
5. Fit initial models.
6. Validate models.
7. Fit better models
8. Report results.

Limitations

- Predictive modelling can only determine the relationship between variables. It cannot answer value judgements from the problem.

Criteria for success

- Explanation/Interpretation
- What prediction errors are acceptable?
- Are some errors worse than others?
- What is the relative importance of small vs large errors?
- How important is measuring the uncertainty?

Considerations

- Significance of problem.
- available data
- implementation challenges.

Example Problem

Should Dalhousie continue to require the use of masks in lectures?

Data Sources

- Could source be biased?
- Publication Bias.

Data Collection

- Survivorship Bias.
- Measurement error.
- Participation bias.

Processing

- Removed values.
- Binning.

Meaning of data

- Surrogate Variables

What is a Program?

- A sequence of commands that the computer follows.
- Plain text files.
- Case sensitive.

Why write a Program?

- More efficiently re-run the same analyses.
- Allow others to use same methods.
- Some analyses need to run similar steps repeatedly.

How to Write a Program

- Organise into functions.
- Include comments
- Structure the code clearly.

File Structure

- Break code into files.
- Use `source` to read files.
- Organise to help find code.
- `source`d files should contain only functions.
- Can also organise code into packages.

Programming Structures

- Comments
- Conditional Statements
- Loops
- Functions
- Exception handling

Comments

Not executed. Purpose is to explain how code works.

Conditional Statements

A block of code that will only run under certain conditions.

Loops

Block of code which is run repeatedly. Two types:

- Run for a fixed set of values
- Continue until condition

Functions

A block of code that can be reused at various times in the program.

Exception handling

When something goes wrong, the default behaviour can be overridden.

What is a function?

- A reusable block of code.
- May include parameters.
- May return values.
- Has enclosed scope.

Why write a function?

- Better organisation.
- Can reuse code.
- Easier to test.

A Good Function Should

- Be as general as possible.
- Have few parameters.
- Be easy to test.
- Include comments to explain the function, the inputs and the return values.

What Should be a Function?

- Reusable code
- Self-contained code

R Objects Have:

- Type — Data type
- Value — Data
- Attributes
- Class — magic.

Basic Data Types

- NULL
- character
- double (numeric)
- integer
- complex
- logical
- factor*
- date*
- closure (function)*

Vectors

- Other data types are all vectors.
- Lists — Vectors of mode "`list`" can contain any objects
- Matrices — vectors with "dimension".
- Objects (e.g. models) — structured lists with class.

Exploring Objects

- `typeof` — returns basic type.
- `class` — returns class
- `length` — length of a vector
- `attributes` — attributes

Variables

- A variable refers to a specific piece of data.
- Variables have types.

Types

- All variables have types
- Sometimes R will automatically convert between types. Sometimes it won't.
- Many errors happen because variables are the wrong type.

Scope

- The scope of a variable is the part of the code that can reference that variable.
- Global variables can be accessed everywhere
- Local variables can only be accessed from within the current block

A Complete List of Times when Global Variables Should be Used

What is Complexity?

- Complexity is about how the running time of a program grows with the size of the problem.
- If an algorithm is too inefficient, it may never finish.

How to Assess Complexity

- Complexity is approximately the number of basic operations.
- Be careful about operations in high-level languages.

Time sinks

- Loops
- Copying/concatenating vectors

R pitfalls

- Dynamic typing
- Vectorisation
- Default return values
- For loops
- The `:` operator

Dynamic Typing

- Type calculated at runtime.
- Operations and functions not well typed. Particularly subset operation `[]`.

Vectorisation

- Functions applied elementwise. May Be unexpected for matrices.
- Functions may need to do this.

Default return values

- Functions return last value computed.
- Trivial functions fractionally neater, but can induce bugs.

For loops

- Vector evaluated at start.

The `:` operator

- Has high precedence.
- Gives wrong value for `1:0`

Types of Problems

- Errors
- Non-termination
- Wrong Answers

Errors

Errors happen when the computer issues an error message, indicating that something is wrong.
Similar are warnings. These indicate that something is probably wrong, but it is possible to continue.

Non-termination

Program does not stop running within expected time-frame.
Possible reasons are:

- Expected time too short
- Program too slow
- Program will never finish.

Wrong Answers

Program runs and produces an answer. The answer does not match the expectations. Could be because answer is wrong, or because expectation is wrong.

For All Problems

- Are you sure it's wrong?
- When? Always? Cases?
- Check intermediate results.
- Test functions in isolation.
- Divide and conquer.
- External packages.

Errors & Warnings

- Error not when bug occurs.
- Messages cryptic — Google.
- Always read warnings.
- Error=no way to continue.
  Warning= probably wrong.

External Packages

- Check documentation.
- Produce minimal example.
- Contact the maintainer.
- Bad input — divide & conquer.

Program Doesn't Terminate

- Find expected running time.
- Test on simpler cases.
- Find where it gets stuck.
- Track termination condition.

Bad Results

- Check expectations.
- Try simple cases.
- Correctness checks.
- Track intermediate results.

What is Data Wrangling?

- Turn mess of files into useable data.
- File formatting and importing.
- Includes some aspects of data exploration — e.g. outlier detection.
- Many different types of data sources — SQL databases, html, text or csv files, proprietory formats.

Scope of this Course

- We will cover reformatting data in R, including:
  - Importing text or csv files.
  - Changing data types.
  - Reformatting tables.

Things not Covered in this Course (Computer Science)

- Handling various data formats.
- Processing text or csv files.

Text and CSV Files

`read.table` — Text file

`read.csv` — CSV file

File input

- `file` — filename, url, or "`stdin()`"

Row Processing

- `header` — row 1 = names
- `nrows` — no. of rows to read
- `skip` — no. of rows to skip
- `blank.lines.skip` —
- `comment.char` — comments

Field Processing

- `quote` — character for quoting.
- `row.names` — row names, or column no.
- `na.strings` — NA values
- `stringsAsFactors` — convert to factor.

Output

- A data frame.
- If row and column names not supplied, they are generated automatically.

Problems Reading Data

- Incorrect formatting
- Incorrectly specified options
- variables not converting
- non-standardised formatting

Converting Columns

- A number of `as.???` can convert data types.
- Using a logical subsetting operation, we can replace bad values. For example `X[X[,4]=="none",4]<-NA`.

Factor Variables

- Sometimes it is important to change the order of factors.
- Some levels may not appear in the dataset.
- Some levels might be labelled in multiple ways due to spelling errors or formatting differences. You will need to correct these first.
- Use `factor(x,levels=...)` to remove redundant levels or add additional levels.

Wide Table Format

- Rows are levels of one factor
- Columns are levels of another
- For each (or many) pairs of levels, there is a value.

Long Table Format

- One column for each factor variable, and one for the value.
- each factor variable class is repeated multiple times.
- If multiple values per level, can create a new factor variable to indicate which value is shown.

Converting Between Formats

- Use `tidyr::pivot_longer` to change wide to long.
- Use `tidyr::pivot_wider` to change long to wide.
- `reshape::melt` can change matrices to long format.

Merging Data Frames

- Use `dplyr::left_join` to merge data frames
- `tidyr::separate` splits columns. `tidyr::union` rejoins them.

Why Visualise Data?

- Patterns are often easier to identify from a figure than a table.
- Summary statistics can disguise important features, e.g. outliers.
- For complicated patterns, graphs can convey more information.
- Your eyes have fewer bugs than your `R` code.

What Data Should we Visualise?

- Sometimes omitting some data obscures the patterns.
- Conversely, putting too much in a single plot can make it difficult to see patterns.

Who is Looking at the Graph?

- Yourself — e.g. when first exploring data.
- Experts in your field.
- Non-experts willing to spend time examining the graph.
- Non-experts reading quickly.

What do you Want to Show Them?

- General trends.
- Specific patterns

Possible Graph Problems

- Bad data.
- Bad perception.
- Distracting aspects.

Channels for Conveying Continuous Data

- position
- length
- angle
- area
- depth
- brightness
- colour saturation
- shape

Channels for Conveying Categorical Data

- Hue (red, green, blue)
- Shape

Creating a Plot with ggplot

- Specify data and mappings.

```
ggplot(data=courses,
mapping=aes(x=students,y=average_grade,colour=
    subject,shape=term,linetype=as.factor(year)))+
```

- Specify plot type(s)

```
geom_point(colour="red",shape=2,aes(group=year))+
```

- Add labels captions, legends etc.

```
labs(x="This is the x axis (I used a log scale!)
    ",y="This is the y axis",title="An example
plot",subtitle="made with ggplot2")+
guides(fill=TRUE,shape=FALSE)+
```

Creating a Nice Plot with ggplot

- Specify axis transformations

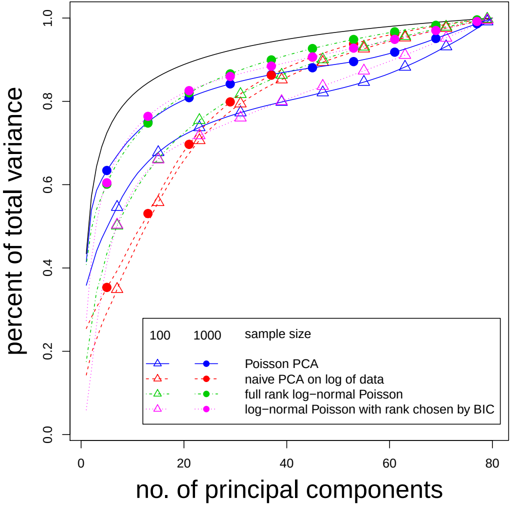
```
scale_x_log10()+
```

- Split into subplots

```
facet_wrap(~subject)+
```

- Make adjustments

```
theme(legend.position="left",
axis.text=element_text(size=12),
axis.title=element_text(size=14),
plot.title=element_text(size=25,hjust=0.5),
plot.subtitle=element_text(size=20,hjust=0.5))
```

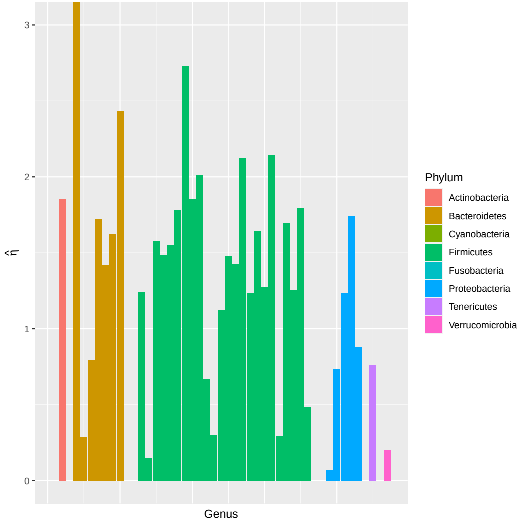


Comments

- Comparing 4 methods at two sample sizes.
- Black line shows theoretical maximum.
- Figure is in Black/White in printed article.

Data Visualisation

Examples From Papers



Comments

- This shows a single variable.
- However, the genus is arranged in a tree structure.
- In addition to the grouping by phylum, there are unshown subgroupings by class, order and family.

Purpose of Data Exploration

- Identify (and hopefully correct) data issues.
- Decide on suitable modelling frameworks for the data.
- Identify unforeseen hypotheses. These might lead to future studies, or indicate confounding variables that need to be addressed.

Data Exploration Tools

- Data Visualisation
- Dimension reduction

Missing Data

- Completely at random.
- At random.
- Not at random.

Outliers

- Large influence on results.
- May be data collection errors.
- Sometimes invalid values.

Duplicate Values

- Can influence the results.
- May be data collection errors.
- Can give misleading cross-validation/test results.

Missing Data

- Completely at random.
- At random.
- Not at random.

Outliers

- Large influence on results.
- May be data collection errors.
- Sometimes invalid values.

Duplicate Values

- Can influence the results.
- May be data collection errors.
- Can give misleading cross-validation/test results.

Handling Data Issues

- Find correct value.
- Remove.
- Impute.

Questions to Answer

- Linear or non-linear model?
- Outliers?
- Important variables?
- Are residuals normal?
- Additional features?
- High correlation between predictors?

Simple Visualisations

- Histograms or density plots
- Draw scatterplots.

Dimension Reduction

- PCA

Use Summary Statistics to Identify:

- Outliers.
- Rare values.
- Failure of assumptions.

Additional Features

- If most relations are linear, linear regression may be appropriate.
- Nonlinear functions can be fitted by adding transformations of original variables.
- Interaction terms can be added to model dependence between more than two variables.
- Very complicated models may be better modelled using random forest or other flexible methods.
- If predictors are strongly correlated and fairly high-dimensional, principal components may make good features.

Data Exploration

Research any relevant subject knowledge

Subject Knowledge

- What sort of relationship is expected?
- Which modelling assumptions are expected to be true?

Idea

- Transform data into smaller number of principal components.
- Principal components are linear combinations of variables.
- PCs are uncorrelated, and minimise squared error.
- Sometimes standardise (correlation matrix instead of covariance).

Assumptions

- If we assume the data are multivariate normal, then:
  - Squared error is supported by likelihood theory.
  - Principal components are independent.

Diagnostics

- Scree plot — used to choose number of principal components.

Limitations

- High dimensions
- Interpretability

Question 1

The data set `iris` contained in the `datasets` package in `R` contains measurements of several iris plants.
(a) Perform a principal component analysis to find the main directions of variation..
(b) Make a scree plot to assess how many principal components to analyse.
(c) Plot the loadings and show how this relates to species.
(d) Repeat this using correlation instead of covariance.

Idea

- Identify unknown groups within the data.
- Individuals from different groups follow different distributions.

Different clustering method

- K-means — normal groups with identity variance
- Mixture model — i.i.d. sample from a mixture of components.
- Hierarchical clustering — treat clusters as points and cluster them.

Choosing No. of Clusters

- Plot sum of squared distances vs. no. of clusters. Find elbow point.
- Gap statistic — compare with an expected graph (using simulations).

Limitations

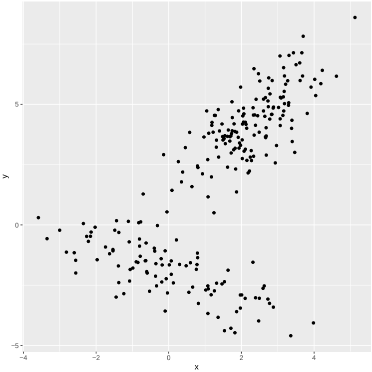
- Sometimes no true answer
- Can be sensitive to outliers.

Problem

- Fix no. K of clusters.
- Minimise sum of squared distance to cluster centres.

Algorithm

- Start with a random assignment to clusters.
- Calculate the cluster centres
- Reassign points to clusters.
- Repeat until convergence.

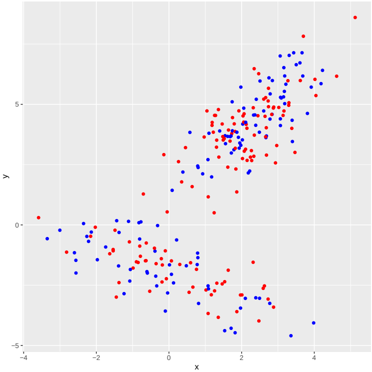


Problem

- Fix no. K of clusters.
- Minimise sum of squared distance to cluster centres.

Algorithm

- Start with a random assignment to clusters.
- Calculate the cluster centres
- Reassign points to clusters.
- Repeat until convergence.

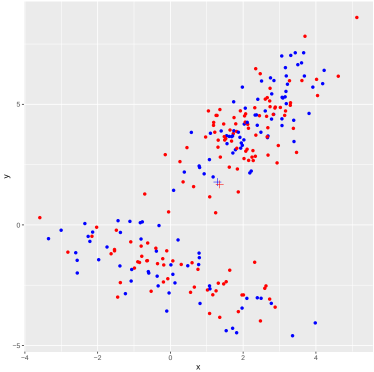


Problem

- Fix no. K of clusters.
- Minimise sum of squared distance to cluster centres.

Algorithm

- Start with a random assignment to clusters.
- Calculate the cluster centres
- Reassign points to clusters.
- Repeat until convergence.

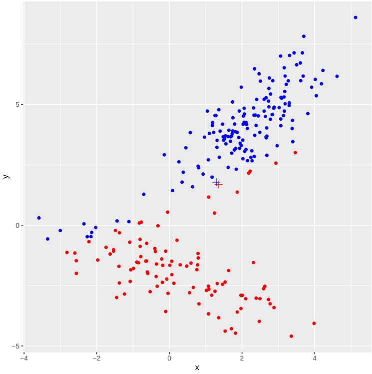


Problem

- Fix no. K of clusters.
- Minimise sum of squared distance to cluster centres.

Algorithm

- Start with a random assignment to clusters.
- Calculate the cluster centres
- Reassign points to clusters.
- Repeat until convergence.

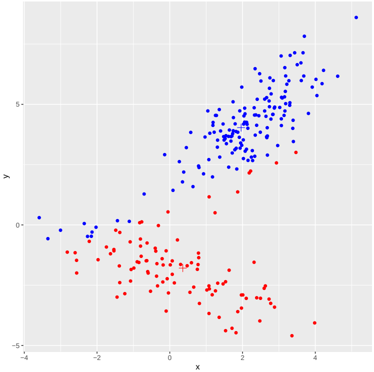


Problem

- Fix no. K of clusters.
- Minimise sum of squared distance to cluster centres.

Algorithm

- Start with a random assignment to clusters.
- Calculate the cluster centres
- Reassign points to clusters.
- Repeat until convergence.

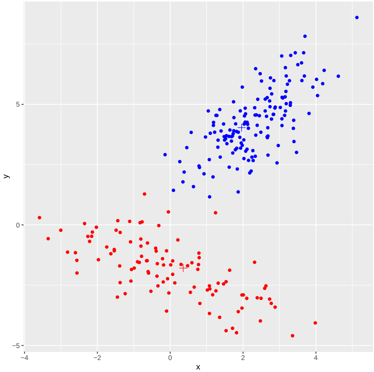


Problem

- Fix no. K of clusters.
- Minimise sum of squared distance to cluster centres.

Algorithm

- Start with a random assignment to clusters.
- Calculate the cluster centres
- Reassign points to clusters.
- Repeat until convergence.



Question 2

For the iris data set:

(a) use *K*-means clustering on the measurements (not using the species) to cluster the plants.

(b) Choose the appropriate number of clusters.

(c) How do the clusters compare with the species of the plants?

Linear Regression

- Fits models of the form $y = X\beta + \beta_0 + E$ for some vector β.
- Can add functions of existing predictors as new predictors.
- Fitted by least squares. This is MLE for normal residuals.

Assumptions

- Error is normal.
- Errors are independent.
- Homoskedasticity

Diagnostics

- Residuals vs. fitted values
- Q-Q plot of residuals.

Limitations

- High dimensions.
- Correlation between predictors.

Question 1

The data set `Boston` contained in the `MASS` package in `R` describes house prices.
(a) Perform a linear regression of median value on the other variables.
(b) Perform diagnostics to assess whether the linear regression model is suitable.
(c) Use a transformation of median value to improve the regression.
(d) Add additional predictors to improve the regression.

Idea

- Specified conditional distribution for response (e.g. Bernoulli)
- Transformed conditional expectation $f(\mu_i)$ is fitted via regression.
- Coefficients fitted via MLE.

Assumptions

- Conditional distribution of response follows specified distribution.
- Homoskedasticity or fixed value of other parameters

Diagnostics

- Raw residuals $y_i - \mu_i$ don't have good properties.
- Several alternative residuals.
- Deviance residuals — root of log-likelihood difference.

Limitations

- High dimensions.
- Correlated predictors.

Question 1

The dataset `iris` in `R` contains measurements of three different species of iris plants.
(a) Use logistic regression to classify the samples from the `versicolor` and `verginica` species.
(b) Plot the deviance residuals.
(c) The predictor `Sepal.Width` is not strongly correlated with the other predictors. Fit a model with this predictor removed. Why do the coefficients change?

Idea

- Often not obvious what transformations to create.
- Flexible family called splines approximate any smooth function.
- Can use the same link functions as for GLM.
- Can use splines for some predictors and linear functions (or selected transformations) for others.

(Cubic) Splines

- Piecewise cubic functions.
- Boundaries between pieces are called knots.
- Pieces aligned so that function twice continuously differentiable.
- Often use natural splines where the outermost pieces are linear.
- Each knot adds one parameter.
- Choose a basis with good properties (not too much dependance).
- Can choose large number of knots and penalise coefficients

Assumptions

- Conditional distribution of response follows specified distribution.
- Homoskedasticity or fixed value of other parameters.
- Conditional mean of response is an additive function of predictors. (No interactions.)

Diagnostics and Interpretation

- As for GLM.
- Can plot the fitted additive functions for each predictor.

Limitations

- High dimensions.
- Correlated predictors.
- Dependant predictors.

Idea

- Assumption of independent errors is not valid.
- Add previous values as predictors.
- Also need to add time (and possibly functions of it) as a predictor.

Assumptions

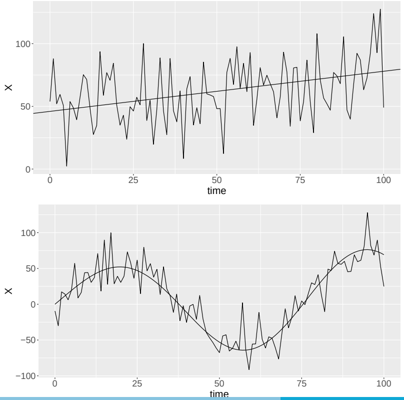
- Error is normal.
- Homoskedasticity
- Stationarity

Diagnostics

- Residuals vs. fitted values
- Q-Q plot of residuals.
- Dickey-Fuller Test

Idea

- Model $y_t = f(t) + \epsilon$ a time trend plus random error.
- Time trend is usually long-term trend plus (or multiplied by) seasonal trend.



Note

- Error is i.i.d..
- Top figure shows linear trend
- Bottom figure shows linear trend multiplied by seasonal trend.

Question 1

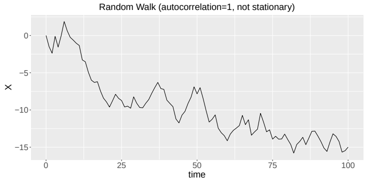
The data set `EuStockMarkets` in `R` contains daily stock market data for four european markets between 1991 and 1998.
[The data are every working day, so are not evenly spaced. However, for the purpose of this analysis, we will assume they are evenly spaced.]

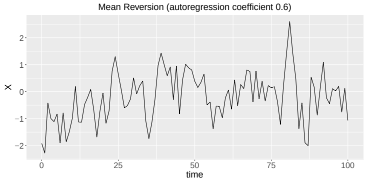
(a) Fit a quadratic trend to the log-DAX value.

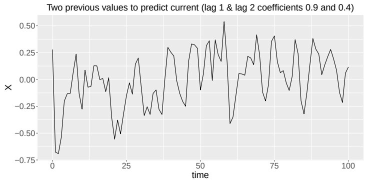
(b) Plot the residuals over time and other diagnostic plots.

Idea

- Even after removing trend, the values at different time points are not independent, with adjacent time points much nearer.
- Deal with this by including previous time points as predictors.







Question 2

Using the detrended data from the previous question:

(a) Fit an autogressive model on the DAX.

(b) Plot the residuals over time and other diagnostic plots.

Idea

- Linear moving average $s_t = \frac{x_t + \cdots + x_{t-k+1}}{k}$
- Exponential moving average $s_t = (1 - w) \sum_{i=1}^{t} w^{t-i} y_i$

Notes

- Exponential moving average of i.i.d. variables is AR 1
- Still assuming Homoskedasticity
- Weighted linear moving average of AR process is i.i.d.

ARMA model

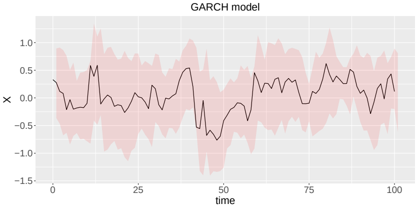
- moving averages follow AR process
- arises naturally as sum of AR processes

Idea

- Sometimes the variance of a time series follows a time series of its own.
- ARCH(p) — Conditional variance $\sigma_t{}^2 = w + \gamma_1 \epsilon_{t-1}{}^2 + \cdots + \gamma_p \epsilon_{t-p}{}^2$
- GARCH(p,q) — $\sigma_t{}^2 - \delta_1 \sigma_{t-1}{}^2 + \cdots + \delta_q \sigma_{t-q}{}^2 = w + \gamma_1 \epsilon_{t-1}{}^2 + \cdots + \gamma_p \epsilon_{t-p}{}^2$

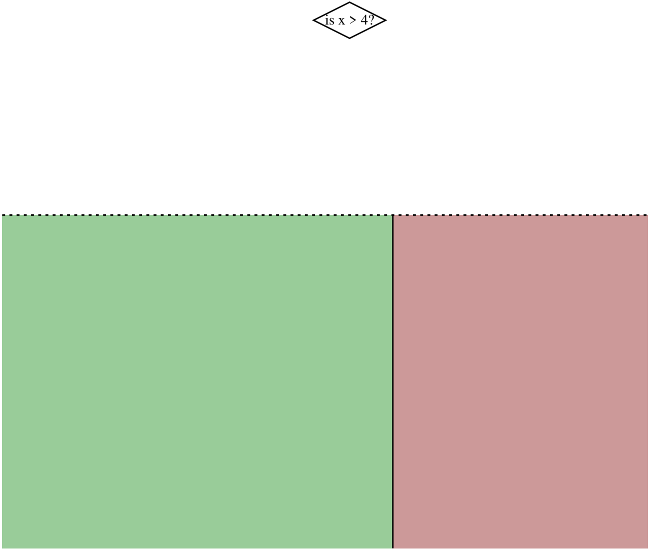
Notes

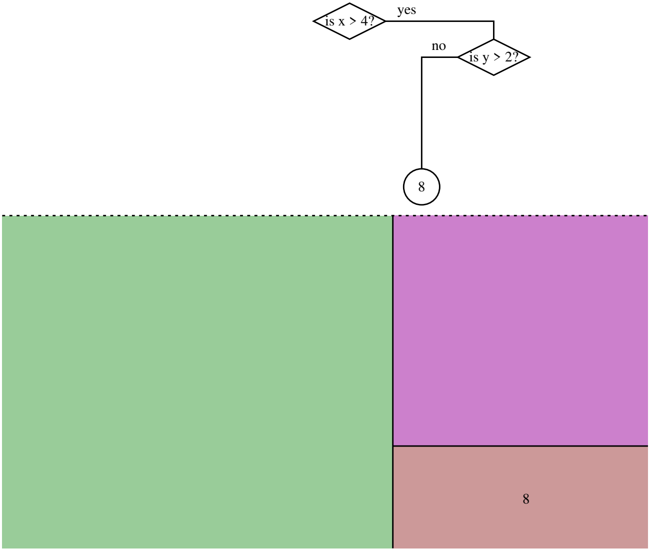
- Assume error is normal.
- Usually fitted by MLE
- Variance often important for time series, particularly financial.



Question 3

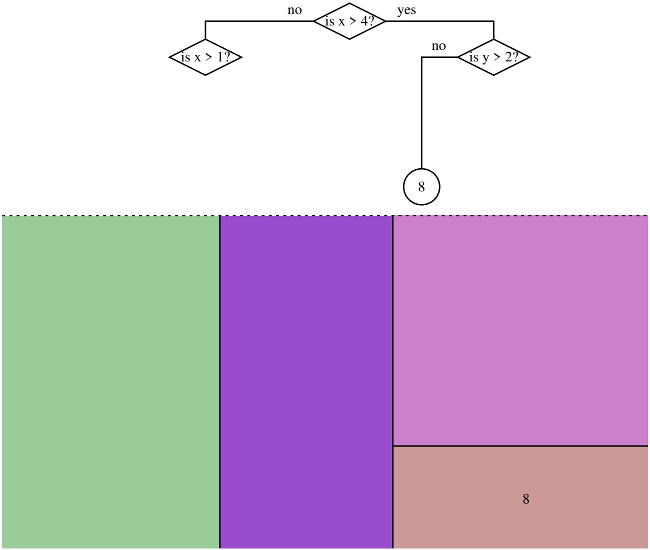
Fit a GARCH model to the DAX data studied in the previous questions.


is x > 4?


is x > 4?
yes
no
is y > 2?
8
8

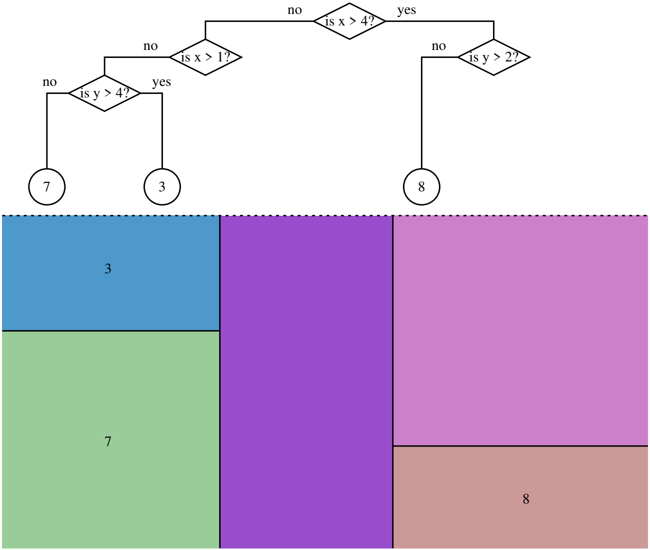
Tree-based Methods

Revision



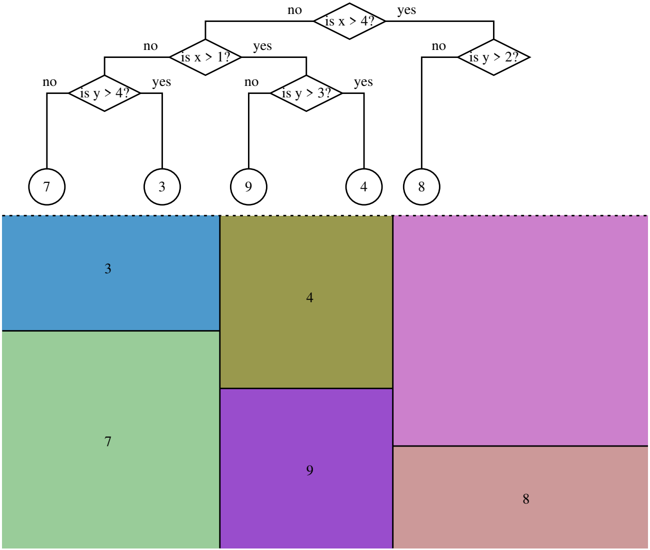
Tree-based Methods

Revision



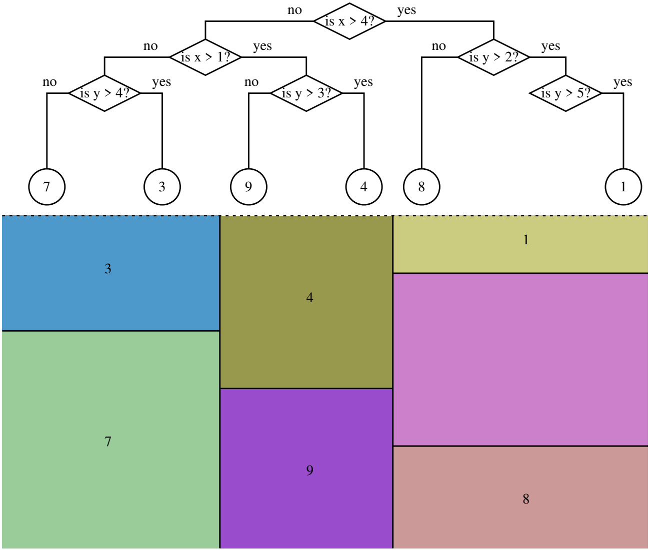
Tree-based Methods

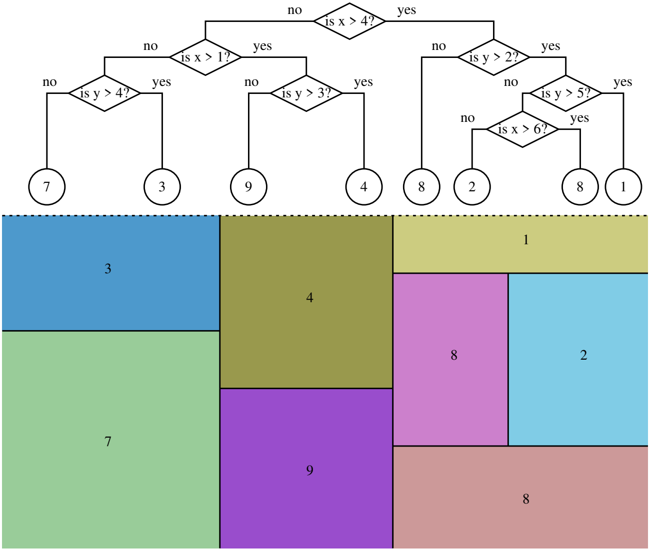
Revision



Tree-based Methods

Revision




is x > 4?
no
yes
is x > 1?
is y > 2?
is y > 4?
is y > 3?
is y > 5?
is x > 6?
7
3
9
4
8
2
8
1
3
4
1
8
2
7
9
8

Idea

- Divide region into rectangular blocks, assign value to each block.
- Equivalent to a decision tree.
- Various methods to avoid overfitting.
- One tree not flexible enough, so average many trees.

Decision Trees

- Cut each leaf node to best improve results.
- Limit complexity either with maximum depth or minimum node size or a complexity measure.

Random Forest

- Fits many decision trees.
- Subsets data and variables to make trees different.

Boosted trees

- Fits trees using residuals from current model.

Question 1

The data file `pollution.txt` contains pollution data from McDonald, G.C. and Schwing, R.C. (1973) 'Instabilities of regression estimates relating air pollution to mortality', *Technometrics*, vol.15, 463– 482. (Downloaded from `http://lib.stat.cmu.edu/datasets/pollution`)
(a) Fit a decision tree to estimate Nitrous Oxide pollution (`NOX`). Use a number of tuning parameter values on a training sample, and compare the accuracy on test data.
(b) Fit a random forest model to estimate `NOX`, using several different tuning parameters, and compare the accuracy.
(c) Use boosted trees to estimate `NOX`.

Variable Selection and Regularisation

Revision

Idea

- Too many predictors result in bad models or even no model.
- Select only the most important predictors, get better results.

Variable Selection Methods

- Search based on goodness of fit.
- Penalty based — LASSO, ridge regression.

Search methods

- Forward Selection
- Backward Selection

Goodness of Fit

- Information Criteria.
- (Generalised) Cross validation
- Hypothesis testing

Penalties

- L^0 – $|\{i|\beta_i \neq 0\}|$
- LASSO – $\sum |\beta_i|$
- Ridge Regression – $\sum {\beta_i}^2$

Variable Selection and Regularisation

Revision

Question 1

The data set `longley` in `R` contains a number of economic data points.
(a) Fit a linear model to predict the variable `Employed` from the other variables.
(b) Use forward selection to select only the important variables in this model.
(c) Use backward selection to select the important variables.
(d) Use LASSO to select the important variables.
(e) Use ridge regression to fit a model.

Normal Errors

- Q-Q plots

Independent Errors

- Difficult to detect unless good reason to suspect particular failures.
- Time series models make specific assumptions.

Homoskedasticity

- Conditional variance of response variable does not depend on predictor variables.
- Plot residuals against predicted values.

Question 1

The dataset `UKDriverDeaths` gives the monthly number of drivers killed or seriously injured in Great Britain.
A statistician uses the commands

```
library(forecast)
UK_driver_deaths<-
    data.frame(month=seq_len(192),
               deaths=as.vector(UKDriverDeaths))
UK_driver_arma<-auto.arima(UK_driver_deaths$deaths,d=0)
```

to fit an ARMA model to this data set.
Test the assumptions in this model.

Validate models

Measuring Performance

Information Criteria

- Training accuracy with correction for model complexity.
- Several versions — AIC, AICc, BIC, . . .

Test Error

- Training error has overfit.
- Test data results more accurately assess model performance on new data.
- Results in smaller training data set.

Cross-Validation

- Multiple training-test splits, average test error over the splits.
- Provides more test data results.

Question 2

The dataset `UKDriverDeaths` gives the monthly number of drivers killed or seriously injured in Great Britain.
A statistician uses the commands

```
library(forecast)
UK_driver_deaths<-
    data.frame(month=seq_len(192),
               deaths=as.vector(UKDriverDeaths))
UK_driver_arma<-auto.arima(UK_driver_deaths$deaths,d=0)
```

to fit an ARMA model to this data set.
Assess the performance of the model on this dataset.

Question 3

The dataset `urine` from the package `boot` contains chemical analysis of urine samples. The objective is to predict the presence or absence of calcium oxalate crystals from the other predictors.
A statistician uses the commands

```
library(boot)
urine_logistic<-glm(r~.,data=urine,
                    family=binomial(link=logit))
```

to perform logistic regression to predict the outcome.
Assess the performance of the model on this dataset, and check the assumptions in the model.

Audience

- Account for audience's level of technical and subject knowledge.
- The report may need to be targeted to multiple audiences.

Logical Structure

- Organise the report in a consistent way.
- Start with more general ideas, and develop into more details.

Executive Summary/Abstract

- A short concise summary of the conclusions in the report.
- Should inform the reader of the main conclusions of your analysis.
- Often written last.

Introduction

- A short clear definition of the problem and its context.
- This should be precise enough to be answered from the data.
- Include a literature review where appropriate.
- Describe source and nature of data.
- May be appropriate to end with outline of remainder of report.

Data Characteristics

- A summary of the main observations in data exploration.

Communication

Parts of Report

Model Selection and Interpretation

- Start by clearly stating the recommended model.
- Interpret the model.
- Justify the model in comparison to alternative models.
- The statement and interpretation of the recommended model are the main conclusions for non-technical readers.
- Model justification should be streamlined.

Summary and Conclusions

- Repeat the main conclusions.
- Might be more technical than the abstract/executive summary. Usually more detailed.
- May also include suggestions for future studies.

Tables and Graphs

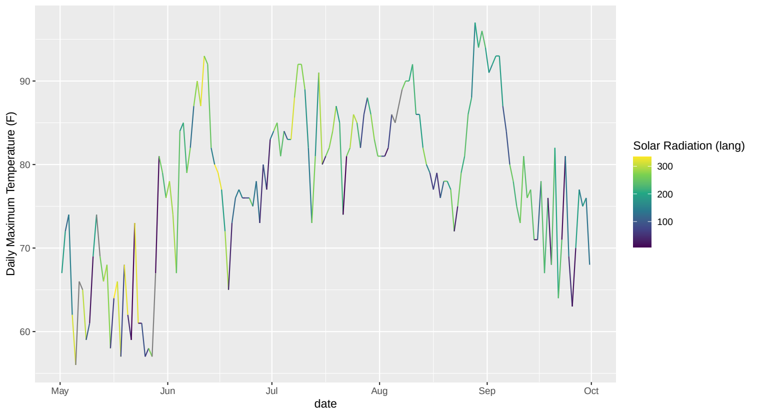
- Tables and Graphs in the main document should all make some point.
- Tables and graphs should be easy to read.
- Should be self-contained.
- Only include necessary information.

Summarising Tables and Graphs

- Focus on interpretation.
- Identify your conclusions. These should require as little context as possible. They should be related to the problem statement.
- Identify the aspects of the data that support the conclusion.

Question 1

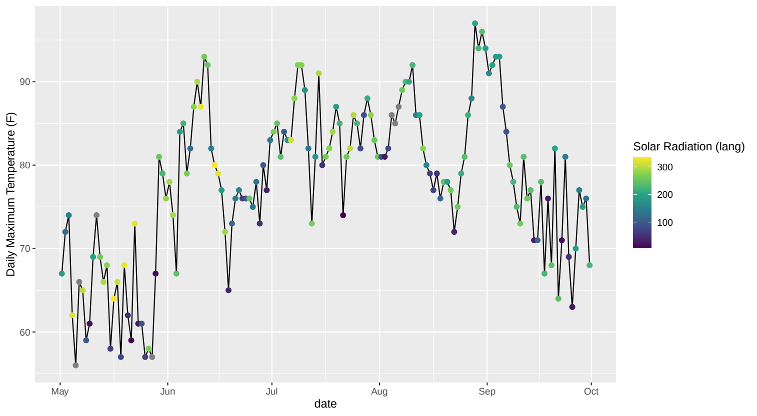
(a) How should this graph be edited to better show the conclusions?



(b) Summarise the main features of the graph.

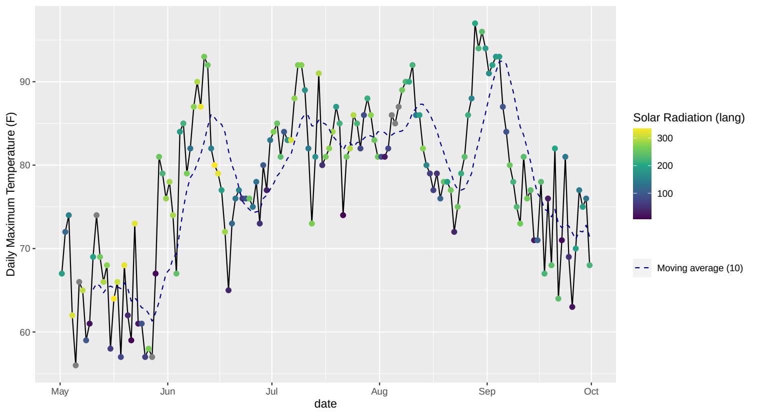
Answer to Question 1

(a)



Answer to Question 1

(a)



Answer to Question 1

(b)

- Very low solar radiation corresponds to low temperatures for the time of year.
- High solar radiation is more common in spring and early summer.
- Low solar radiation is common throughout the year.
- Temperature is highest in summer.
- Daily temperature fluctuation was highest in late September
- Fluctuations in the 10-day moving average were higher in summer.