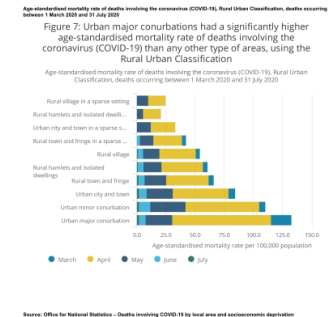


In-course assessment 2: background to take-home task

- Strategies for **reducing Covid mortality** require understanding of **risk factors**
- Age** known to be major risk factor; also **gender**, **social deprivation**, **pre-existing health conditions**, **ethnicity** etc.

- August 2020: UK Office for National Statistics (ONS) report gave **simple analysis of Covid death rates in England and Wales** between March and July 2020
- Analysis revealed **major differences between death rates in different areas** of England and Wales



Question

Why do death rates vary between areas?

In-course assessment 2: take-home component

The data

ID	Region	RUCode	Deaths	PopTot	PopM	PopF	PopComm	PopDens	HH	HH_1Pers	HH_1Fam	HH_Oth	HH_HealthPrb	HHNoCH	HHRooms	HHBedrooms	HHDepriv1	HHDepriv2	HHDepriv3
1	East Mid	C1	24	10992	5381	5611	270	23.3	5081	2001	2799	281	1403	118	5.2	2.5	1616	1025	272
2	East Mid	C1	1	6561	3174	3387	71	12.7	2778	867	1757	154	672	53	5.4	2.7	873	488	75
3	East Mid	C1	15	6047	3005	3042	143	34.1	2864	1315	1455	194	961	60	4.8	2.3	964	758	241
4	East Mid	D1	4	6647	3238	3409	33	4.5	2797	735	1970	92	645	26	6.0	3.0	884	414	48
5	East Mid	E1	1	8785	4310	4475	39	0.8	3551	764	2618	169	707	66	7.0	3.4	1016	384	44
6	East Mid	D1	7	12280	5992	6288	182	1.0	5113	1208	3700	205	1132	56	6.3	3.1	1532	658	80
7	East Mid	C1	9	8012	3922	4090	165	5.4	3241	759	2341	141	878	22	5.7	2.9	1118	735	194
8	East Mid	E1	0	7152	3482	3670	62	0.7	3152	767	2270	115	892	44	6.4	3.0	1132	568	76
9	East Mid	B1	2	7425	3443	3982	126	25.4	3297	1173	1918	206	1084	60	4.7	2.4	1097	1002	387
10	East Mid	C1	5	9132	4334	4798	85	32.5	4046	1246	2572	228	1008	50	5.5	2.8	1259	782	105
11	East Mid	D1	5	7073	3409	3664	89	6.2	3291	1128	2047	116	988	64	5.5	2.7	1120	718	146
12	East Mid	B1	15	7614	3750	3864	180	22.2	3124	743	2261	120	688	74	6.1	3.0	990	420	61

Your task

Given:

- numbers of Covid deaths** in some areas of England and Wales between March and July 2020;
- demographic information** for these areas (mostly from **2011 UK Census**) ;

Build a statistical model that will help you to:

- Understand** factors associated with variation in numbers of Covid deaths during the period;
- Estimate** numbers of deaths for other areas of England and Wales.

- Total **numbers of deaths** and **>80 social & demographic characteristics** of “Middle Layer Super Output Areas” (MSOAs i.e. statistical reporting areas) in England and Wales
 - Deaths are totals over period **March–July 2020**; most covariates are from **2011 UK Census**
- 7 201 MSOAs** in total (**death numbers missing for 1 800 of them** — but we know the missing values)

Detailed requirements

- You may use **either R or SAS**.
 - You may **work alone or in pairs**.
 - You **must sign up to a "group"** (yourself or your pair) on Moodle
- 1 **Read data** and carry out any **necessary recoding**
 - 2 **Exploratory analysis** to support subsequent model-building:
 - Identify appropriate set of **candidate variables** to consider;
 - Identify **important features of the data** that may have implications for modelling.
 - 3 **Develop a statistical model** that can be (a) used to **predict numbers of Covid deaths** in MSOAs where they are not known (b) interpreted to **understand why some areas have more deaths than others**.
 - Model must be either a **linear model**, a **generalized linear model** or a **generalized additive model**.
 - 4 Use model to **predict numbers of Covid deaths** for each of 1 800 cases with missing values, and give **standard deviations of prediction errors**.

Submission requirements

Deadline: Monday 26th April, 12:00 London time
Online submission via Moodle.

- **Report on analysis**, in three sections and not exceeding 2 500 words of text + two pages of graphics / tables (see detailed instructions on Moodle page).
- **R script or SAS program** generating analysis and predictions.
 - Should **run without user intervention**, providing data file is present in current working directory / current folder.
 - Should **produce any results mentioned in your report** including your graphs, together with **predictions and standard deviations**.
 - See detailed instructions for **file naming requirements**
- **Text file containing predictions** for the 1 800 records with missing numbers of deaths.
 - **Three columns** (Record ID, prediction, standard error), **separated by spaces** and with **no header**.
 - See detailed instructions for **file naming requirements**

Marking criteria

- **Report:** 40 marks — see instructions on Moodle for criteria.
- **Code:** 15 marks — ditto.
- **Predictions:** 20 marks — *you are competing against each other and against us!*

Assessment of predictions

- Prediction score is $S = \sum_{i=1}^{1800} \left[\log \sigma_i + \frac{(Y_i - \hat{\mu}_i)^2}{2\sigma_i^2} \right]$, where
 - Y_i is **actual number of deaths** for record i ;
 - $\hat{\mu}_i$ is **predicted number of deaths**;
 - σ_i is your **prediction error standard deviation**.
- **Accurate predictions with low uncertainty yield low values**
- **20 marks for best (lowest) scores**; worse scores, fewer marks.

Hints for tackling assessment

- 1 No single 'right' answer — exercise requires combination of **technical knowledge** (stats and computing) and **good judgement**:
- 2 Looking for **structured and critical** approach to analysis
- 3 Looking for **appropriate judgement** both in **approach to analysis** and in **choice of material to present**.
- 4 **Use what you know about the situation** e.g. web links / papers / reports in detailed instructions, your knowledge of the UK situation in early 2020, other commentaries on Covid, etc.

Hints: exploratory analysis

Dealing with many covariates — some ideas

- Some covariates may be **highly correlated** or **represent similar things** ⇒ **not all necessary**
 - E.g. 'Employment / occupation' and 'Social grade' variables are closely linked (see Appendix in detailed instructions)
- First tidy your room:** Use **background knowledge** and **exploratory analysis** to **simplify data** prior to modelling, e.g.:
 - Look at **other published commentaries** to see what is relevant (acknowledge your sources!)
 - Calculate **summary measures based on context** — e.g. aggregate age bands
 - Define **new variables based on statistical criteria**, and work with these (more later)
 - Use **preliminary 'automatic' procedures** to choose from similar subsets of variables — e.g. stepwise regression to identify most promising variables.
 - Plot, plot, plot ...**

Where to start? Preprocessing and exploratory analysis ...

Aims of an exploratory analysis (for the third time!)

- 1 To gain a **preliminary understanding of structure** in a dataset
- 2 To look for possible **outliers** or **data quality problems**
- 3 To **suggest some initial assumptions** (e.g. normality of residuals, constant variance) that may be reasonable as a starting point in subsequent modelling and analysis

Key questions for (3)

- How to deal with **many potential covariates** and **factors with many levels**?
- What kind of model** — linear, generalized linear, generalized additive?

New variables based on statistical criteria ...

- Categorical variables** can perhaps be aggregated according to **similarity of relationships in different groups** ⇒ **clustering methods**
 - E.g. potential for **grouping rural / urban categories**, **occupation categories** or **social grades**?
- Several continuous variables:** hard to disentangle **effects of highly correlated variables**, may be better to **combine into single 'index'** e.g. **principal components analysis**:

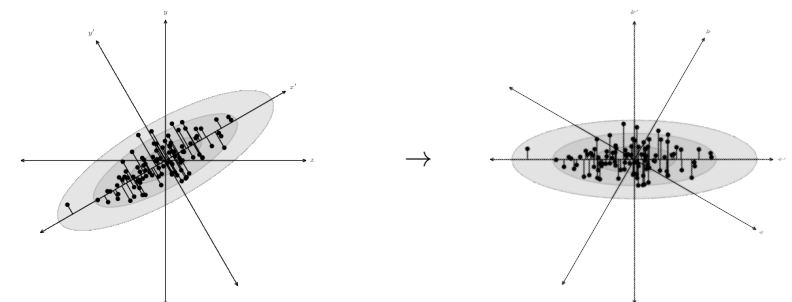


Figure taken from [AstroML](#)

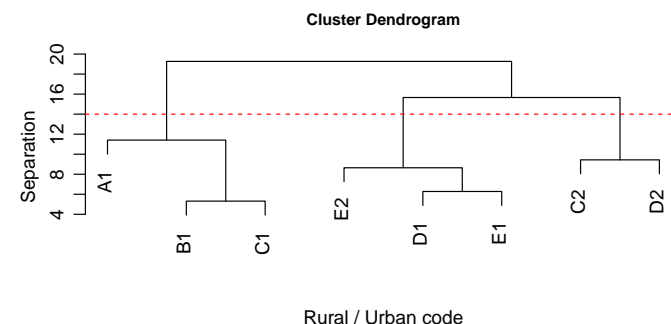
Data-driven grouping: hierarchical clustering

- **Idea:** group records based on **similar values of one or more variables**
- **Algorithm** for hierarchical clustering:
 - 1 Compute 'distance' between every pair of records e.g. Euclidean

$$\text{distance} = \sqrt{\sum_{k=1}^p (y_{ij} - y_{jk})^2}$$
 where p is total # of variables
 - May be useful to **standardise variables first**
 - **Other distance measures available** for non-continuous variables
 - 2 Combine **closest pair** into single group; calculate distance from this group to each other record
 - **Various options for determining distance** from group: 'single linkage' (closest group member), 'complete linkage' etc.
 - 3 Repeat **step (2)** until all records are in same group
- Results usually visualised in **cluster dendrogram**: 'cut the tree' to define groups ...

Hierarchical clustering in R

```
> NumVars <- (sapply(DeathData, is.numeric) & # Columns containing numeric
+ names(DeathData) != "Deaths") # covariates
> RUMeans <-
+ aggregate(DeathData[,NumVars], # Means of all numeric covariates
+ by=list(DeathData$RUCode), FUN=mean) # by rural / urban group
> rownames(RUMeans) <- RUMeans[,1]
> RUMeans <- scale(RUMeans[, -1]) # Standardise to mean 0 & SD 1
> Distances <- dist(RUMeans) # Pairwise distances
> ClusTree <- hclust(Distances, method="complete") # Do the clustering
> par(mar=c(3,3,3,1), mgp=c(2,0.75,0)) # Set plot margins
> plot(ClusTree, xlab="Rural / Urban code", ylab="Separation", cex.main=0.8)
> abline(h=14, col="red", lty=2)
```



Hierarchical clustering in R: defining new groups

```
> #
> # Cut tree to form three groups
> #
> NewGroups <- cutree(ClusTree, k=3)
> print(NewGroups)
A1 B1 C1 C2 D1 D2 E1 E2
1 1 1 2 3 2 3 3
> #
> # Code below shows how to add new groups to original data frame
> #
> DeathData <-
+ merge(data.frame(RUCode=names(NewGroups), NewGroup=NewGroups),
+ DeathData)
> table(DeathData[,c("NewGroup", "RUCode")],
+ dnn=c("New group", "Original code"))
      Original code
New group  A1  B1  C1  C2  D1  D2  E1  E2
1  2399  249 3206    0    0    0    0    0
2     0    0    0   21    0   29    0    0
3     0    0    0    0  645    0  566   86
```

Hierarchical clustering in SAS

```
PROC MEANS DATA=STAT0023.Covid NOPRINT NONOBS;
  CLASS RUCode;
  VAR PopTot PopM PopF PopComm PopDens
      HH HH_1Pers HH_1Fam [snip snip] QualOther Stud18plus;
  OUTPUT OUT=RUMeans (WHERE=( _type_=1)) MEAN= /AUTONAME;

DATA RUMeans; /* Drop unwanted variables created by SAS */
  SET RUMeans;
  DROP _TYPE_ _FREQ_;

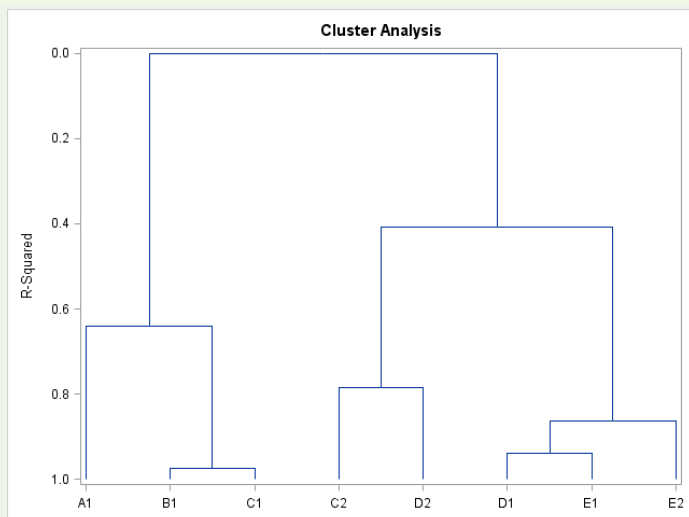
PROC CLUSTER DATA=RUMeans OUTTREE=RUTree METHOD=COMPLETE
  STANDARD PLOTS=DENDROGRAM(VERTICAL HEIGHT=RSQ) PRINT=0;
  ID RUCode;
RUN;

/* Use PROC TREE to create new dataset, including groups */
/* defined by clusters explaining 50% of overall variance */

PROC TREE DATA=RUTree OUT=CovidGroups HEIGHT=RSQ LEVEL=0.5;
RUN;
```

Hierarchical clustering in SAS: plot from PROC CLUSTER

Clustering of rural / urban categories



Dimension reduction: principal components analysis (PCA)

- Most popular way to produce 'indices' from multiple variables: transform variables $X_1 \dots X_k$ (usually with zero means) into

$$\begin{aligned} X_1^* &= a_{11}X_1 + a_{12}X_2 + \dots + a_{1k}X_k \\ X_2^* &= a_{21}X_1 + a_{22}X_2 + \dots + a_{2k}X_k \\ &\vdots \\ X_k^* &= a_{k1}X_1 + a_{k2}X_2 + \dots + a_{kk}X_k \end{aligned}$$

such that

- $\text{Var}(X_1^*) \geq \text{Var}(X_2^*) \geq \dots \geq \text{Var}(X_k^*)$ (see diagram on earlier slide)
 - X_1^*, \dots, X_k^* are mutually uncorrelated.
 - $\sum_{j=1}^k a_{ij}^2 = 1 \forall i = 1, \dots, k$.
- Can show that:
 - $\mathbf{a}_i = (a_{i1} \ a_{i2} \ \dots \ a_{ik})'$ is i th eigenvector of covariance matrix of X_1, \dots, X_k ;
 - $\text{Var}(X_i^*)$ is corresponding eigenvalue.

PCA and units of measurement

- Note that:**
 - If $\text{Var}(X_i) \gg \text{Var}(X_j)$ for any $j \neq i$ then X_i will dominate PC1 by definition
 - $\text{Var}(X_i)$ depends on measurement units e.g. change from metres to centimetres increases variance by factor of 10 000
- Hence results of PCA depend on measurement units — unsatisfactory
- Solution:** standardise each variable to have mean zero and variance 1 before carrying out PCA — equivalent to using correlation matrix instead of covariance
- General guideline:** standardise if original variables have different measurement units.
 - PCs are linear combinations of standardised variables in this case

PCA in R: a final trip to the Galapagos

- Command is `prcomp()`
- Use `prcomp(..., scale.=TRUE)` to standardise variables

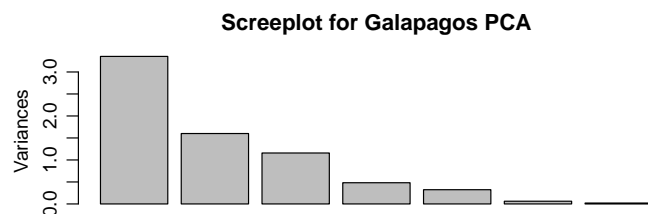
```
> Galapagos.PC <- prcomp(species.data, scale.=TRUE)
> print(Galapagos.PC, digits=2)
Standard deviations (1, .., p=7):
[1] 1.83 1.27 1.08 0.69 0.57 0.25 0.14
```

```
Rotation (n x k) = (7 x 7):
      PC1    PC2    PC3    PC4    PC5    PC6    PC7
Species  0.494 -0.0301  0.309 -0.2651  0.211 -0.480 -0.5607
Endemics 0.505 -0.0524  0.255 -0.3029  0.183  0.070  0.7399
Area     0.445 -0.0064 -0.016  0.7888 -0.305 -0.266  0.1247
Elevation 0.508 -0.1235 -0.250 -0.0049 -0.040  0.743 -0.3318
Nearest -0.061 -0.7021  0.183 -0.2329 -0.643 -0.051 -0.0033
Scruz    -0.103 -0.6972 -0.117  0.2809  0.639 -0.038  0.0224
Adjacent 0.174 -0.0449 -0.854 -0.2876 -0.077 -0.371  0.1097
```

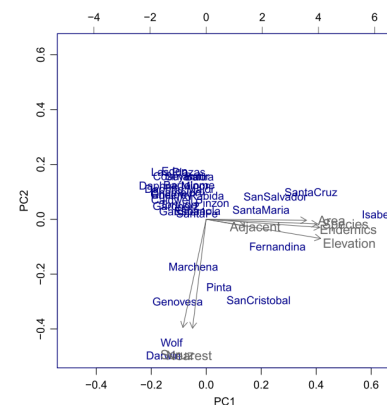
Galapagos PCA: comments on initial results

- PC1 has variance 1.83², PC2 has variance 1.27² etc.
- 'Rotation matrix' columns give loadings i.e. coefficients $\{a_{ij}\}$
- PC1 has large loadings for Species, Area, Endemics & Elevation — measure of **island size / capacity** (more later)
- PC2 has large loadings for Nearest & Scruz — measure of **isolation**
- How many PCs to retain? Some considerations:
 - Proportion of total variance (screeplot / output of summary())
 - Interpretability of components e.g. 'capacity' / 'isolation'

```
> par(mar=c(3,3,3,1), mgp=c(2,0.75,0)) # Set plot margins
> plot(Galapagos.PC, main="Screeplot for Galapagos PCA")
```



PCA visualisation: the biplot



- Points show **scores on first two PCs** for each observation (island) — shows clearly islands with **high capacity** (PC1) and **high isolation** (PC2)
- Arrows show **relation between original variables and PCs** — roughly, arrow for variable X shows scores for **hypothetical island** with **above-average value for X** but **average values for everything else**

PCA in SAS: the same again

```
PROC PRINCOMP DATA=Galapagos OUT=GalapagosWithPCs;
  VAR Species Endemics Area Elevation Nearest Scruz Adjacent;
RUN;
```

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.35378815	1.75330246	0.4791	0.4791
2	1.60048569	0.44234073	0.2286	0.7078
3	1.15814495	0.67614388	0.1654	0.8732
4	0.48200107	0.15849484	0.0689	0.9421
5	0.32350624	0.26159785	0.0462	0.9883
6	0.06190839	0.04174288	0.0088	0.9971
7	0.02016551		0.0029	1.0000

Eigenvectors							
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
Species	0.494253	0.030095	-0.308734	0.265074	-0.211093	0.479835	-0.560737
Endemics	0.504517	0.052446	-0.255452	0.302906	-0.182622	-0.070393	0.739864
Area	0.445310	0.006423	0.015504	-0.788758	0.304618	0.266340	0.124692
Elevation	0.508201	0.123522	0.250143	0.004910	0.039991	-0.743104	-0.331775
Nearest	-0.060722	0.702147	-0.182681	0.232856	0.642704	0.051273	-0.003254
Scruz	-0.103464	0.697153	0.117114	-0.280862	-0.639301	0.038287	0.022400
Adjacent	0.173826	0.044856	0.854124	0.287601	0.076959	0.370900	0.109728

More notes on PCA

- Variables are **centred automatically** by most software packages
- Variables are **standardised by default in SAS**, but not in R (need `.scale=TRUE`)
- **Sign of components** is arbitrary (X_j^* & $-X_j^*$ have same variance) — note difference between R and SAS results
- For interpretation, ask '**how to get large positive / large negative score?**' Examples:
 - **Galapagos PC1**: loadings for standardised Species, Area, Endemics & Elevation are large & positive, so islands with **above-average values of all these variables** give high positive score, islands with below-average values give high negative score.
 - **Galapagos PC3**, SAS version:¹ large positive loading for Adjacent, moderate negative loadings for Species & Endemics \Rightarrow high positive score corresponds to island with **fewer species than average** & where **adjacent island is larger than average**.

¹Opposite signs in R.

Which type of model?

Linear, generalized linear or additive? Main questions:

- ① Conditional on covariates, can response variable be assumed to follow **normal distribution** with **constant variance**?
 - ② Are covariate effects best represented **parametrically** or **nonparametrically**?
- **Normality**: response variable is **non-negative**, hence can't be exactly normal — but perhaps residuals from linear regression model will be **approximately normal**?
 - **Constant variance**: common for variability of non-negative quantities to increase with mean — look at **residual plots**
 - **Does it matter?** Depends on (a) **how serious are departures from normality / constant variance** (b) **whether you think it's worth the effort** of moving away from linear model.
 - **If variance varies substantially** between MSOAs, could possibly **improve prediction score by accounting for it** using (e.g.) Poisson / quasipoisson / binomial (etc.) generalized linear / additive model.
 - **Parametric / nonparametric?** **Plot, plot, plot** ...

Hints: model-building

Model-building: some recommendations

- Don't start till you've done a **really thorough exploratory analysis**
- Take **structured approach**. Example (just one possibility):
 - ① Fit **initial model** containing **all terms suggested by exploratory analysis**
 - ② Check model to ensure **no gross violations of assumptions**
 - ③ If not, **remove terms that seem insignificant**, **one at a time** (recall sheep energy example from Week 9) — use **nested model comparisons, AIC etc.** to check at each stage
 - ④ When finished, **check model again**
 - ⑤ Do some **more exploratory analysis with residuals**: are there any **relationships with additional candidate covariates** that you didn't consider at first? What about **interactions**?
 - ⑥ If you find anything, **expand the model** and go through a similar sequence of steps
 - ⑦ **Stop** when happy / bored / defeated (**hopefully happy!**)
- Use both your **statistical knowledge** and your **understanding of the context**

Other hints

- **Interpret p -values with care**: data set is large so '**statistical significance**' may not be the same as **practical relevance** (null hypothesis is $H_0 : \beta_j = 0$, but who cares if $\beta_j = 0.000001$?)
- **Consider interactions** e.g. **are relationships the same in rural and urban areas?**
- **Calculation of prediction error standard deviations**:
 - Use $SD(Y_i - \hat{\mu}_i) = \sqrt{\hat{V}ar(Y_i) + Var(\hat{\mu}_i)}$
 - $\hat{V}ar(Y_i)$ from **chosen distribution** e.g. if linear model gamma then $\hat{V}ar(Y_i) = \hat{\sigma}^2$, if Poisson then $\hat{V}ar(Y_i) = \hat{\mu}_i$.
 - $Var(\hat{\mu}_i)$ directly from R / SAS output.

