Quantifying the Health Impacts of Air Pollution

Day 1: The Health Impacts of Air Pollution

Prof. Gavin Shaddick
University of Exeter

7^{th} August 2017

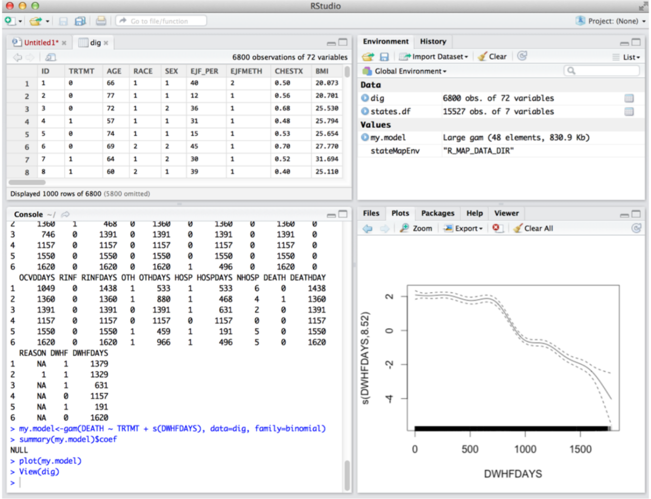
Outline

An Introduction to R

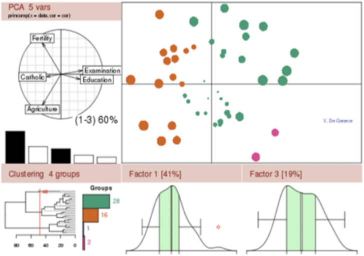
STATISTICAL SOFTWARE

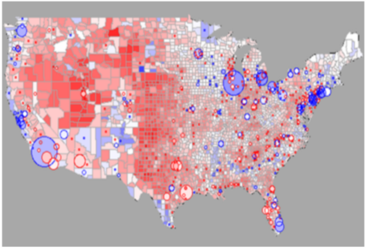
- Excel
  - simple descriptive statistics, plots, and regression can be done in the basic installation of Excel
  - the Analysis Toolpak allows many more methods to be used such as ANOVA and hypothesis tests.
- SPSS, SAS, Stata
  - general purpose statistical packages that can perform a very wide variety of analyses
  - cover everything from initial descriptive analyses to very complex methods.
  - GUI interfaces: functions found by menus.
- R
  - a language and environment for statistical computing and graphics
  - open source with many many user packages
  - it's free! (Open-Source)

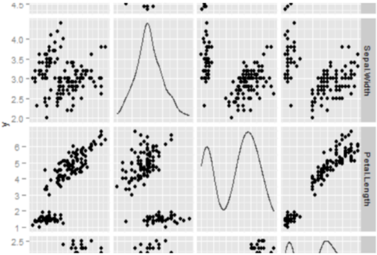
RSTUDIO

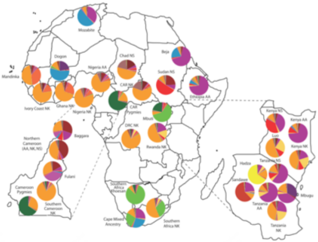

RStudio
Go to file/function
Project: (None)
Untitled1*
dig
6800 observations of 72 variables
ID TRTMT AGE RACE SEX EJF_PER EJFMETH CHESTX BMI
1 1 0 66 1 1 40 2 0.50 20.073
2 2 0 77 1 1 12 1 0.56 20.701
3 3 0 72 1 2 36 1 0.68 25.530
4 4 1 57 1 1 31 1 0.48 25.794
5 5 0 74 1 1 15 1 0.53 25.654
6 6 0 69 2 2 45 1 0.70 27.770
7 7 1 64 1 2 30 1 0.52 31.694
8 8 1 60 2 1 39 1 0.40 25.110
Displayed 1000 rows of 6800 (5800 omitted)
Environment History
Import Dataset
Clear
List
Global Environment
Data
dig 6800 obs. of 72 variables
states.df 15527 obs. of 7 variables
Values
my.model Large gam (48 elements, 830.9 Kb)
stateMapEnv "R_MAP_DATA_DIR"
Console ~/
3 746 0 1391 0 1391 0 1391 0 1391 0
4 1157 0 1157 0 1157 0 1157 0 1157 0
5 1550 0 1550 0 1550 0 1550 0 1550 0
6 1620 0 1620 0 1620 1 496 0 1620 0
OCVDDAYS RINF RINFDAYS OTH OTHDAYS HOSP HOSPDAYS NHOSP DEATH DEATHDAY
1 1049 0 1438 1 533 1 533 6 0 1438
2 1360 0 1360 1 880 1 468 4 1 1360
3 1391 0 1391 0 1391 1 631 2 0 1391
4 1157 0 1157 0 1157 0 1157 0 0 1157
5 1550 0 1550 1 459 1 191 5 0 1550
6 1620 0 1620 1 966 1 496 5 0 1620
REASON DWHF DWHFDAYS
1 NA 1 1379
2 1 1 1329
3 NA 1 631
4 NA 0 1157
5 NA 1 191
6 NA 0 1620
> my.model<-gam(DEATH ~ TRTMT + s(DWHFDAYS), data=dig, family=binomial)
> summary(my.model)$coef
NULL
> plot(my.model)
> View(dig)
>
Files Plots Packages Help Viewer
Zoom
Export
Clear All
s(DWHFDAYS,8.52)
2
0
-2
-4
0
500
1000
1500
DWHFDAYS

R GRAPHICS


PCA 5 vars
Fertility
Examination
Education
Catholic
Agriculture
(1-3) 60%
Clustering 4 groups
Groups
Factor 1 [41%]
Factor 3 [19%]




Sepal.Width
Petal.Length



DOES R HAVE EPIDEMIOLOGICAL PACKAGES?

- The default installation of R does not have packages that specifically implement epidemiological applications.
- However, many of the statistical tools that epidemiologists use are readily available, including statistical models such as unconditional logistic regression, conditional logistic regression, Poisson regression, Cox proportional hazards regression, and much more.

DOES R HAVE EPIDEMIOLOGICAL PACKAGES?

- To meet the specific needs of public health epidemiologists and health data analysts, there are many epidemiology packages.
- There are also packages to perform sample size calculations, survival analysis, clustering, mapping, almost everything you can think of.
- "During the last decade, the momentum coming from both academia and industry has lifted the R programming language to become the single most important tool for computational statistics, visualisation and data science."

Base R comes with many statistical tools

Summary statistics

- `summary(),fivenum(),stem()` – examine the distribution of data
- `boxplots()`

Test statistics

- `t.test()` – returns a one-sample t-test. Can be used for a two-sample t-test by setting `paired=TRUE`
- `wilcox.test()` - returns a one-sample non-parametric Wilcoxon (Mann-Whitney) test. Similarly, can be used for a two-sample Wilcoxon test by setting `paired=TRUE`

Summary Statistics in R

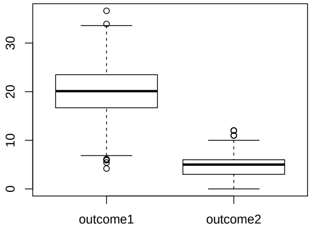
- If we have a dataset `mydat` then we can summarise as follows

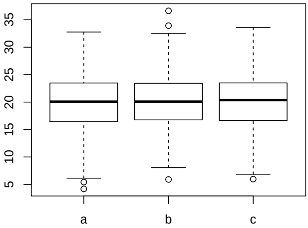
```
# Summary of data
summary(mydat$outcome1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.186  16.680  20.110  19.990  23.480  36.620

summary(mydat$outcome2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   3.000   5.000   4.975   6.000  12.000
```

Summary Statistics in R

- We can create boxplots as follows

```
boxplot(mydat)
boxplot(outcome1~grp, data=mydat)
```





Test Statistics in R

- We can perform test statistics such as t-test and Wilcoxon-Rank test as follows

```
t.test(mydat$outcome1,
          mydat$outcome2)
wilcox.test(mydat$outcome1,
               mydat$outcome2)
wilcox.test(mydat$outcome1,
               mydat$outcome2,
               paired=T)
```

Outputs

- Linear regression: `lm(formula, data)`
- Returns object of class `lm`
  - `summary(x)` comprehensive summary of results
  - `print(x)` precise version of the object
  - `deviance(x)` residuals
  - `plot(x)` returns plot: residuals, fitted values and some diagnostics
  - `coef(x)` extract regression coefficients
  - `predict(x, newdata=...)` second argument takes a vector or matrix of new data values you want predictions for
  - `step()` add or drop terms, model with smallest AICs returned.

OUTPUTS

- Assign the function to an object to extract additional output:

```
my.reg <- lm(outcome1~outcome2, mydat)
summary(my.reg)

Call:
lm(formula = outcome1 ~ outcome2, data = mydat)
Residuals:
    Min      1Q  Median      3Q     Max
-15.806  -3.423   0.260   3.491  16.834
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.48190    0.39883  48.847   <2e-16 ***
outcome2     0.10208    0.07351   1.389    0.165
```

Packages are additional collections of functions

- `rmeta`
- `ggplot2`
- `epicalc`
- `epi`
- `epitools`
- `epiR`
- `pwr`
- `MASS`
- `lattice`
- `lme4`
- `mgcv`
- `Survival`
- ...

Health Impact Assessment associated with outdoor air pollution

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each $10\mu gm{-}3$ $PM_{2.5}$ increase
- Counterfactual: $5\mu gm{-}3$

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each 10μgm−3 $PM_{2.5}$ increase
- Counterfactual: 5μgm−3

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5					

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each 10μgm-3 $PM_{2.5}$ increase
- Counterfactual: 5μgm-3

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00				

EXAMPLE: HIA OF AIR POLLUTION

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each $10\mu gm-3$ $PM_{2.5}$ increase
- Counterfactual: $5\mu gm-3$

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0			

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each $10\mu gm{-}3$ $PM_{2.5}$ increase
- Counterfactual: $5\mu gm{-}3$

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15					

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each 10μgm-3 $PM_{2.5}$ increase
- Counterfactual: 5μgm-3

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15	1.06				

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each 10μgm$-$3 $PM_{2.5}$ increase
- Counterfactual: 5μgm$-$3

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15	1.06	84.8			

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each 10μgm$-$3 $PM_{2.5}$ increase
- Counterfactual: 5μgm$-$3

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15	1.06	84.8	8,480		

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each 10μgm-3 $PM_{2.5}$ increase
- Counterfactual: 5μgm-3

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15	1.06	84.8	8,480	480.0	480.0
25					

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each $10\mu gm-3$ $PM_{2.5}$ increase
- Counterfactual: $5\mu gm-3$

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15	1.06	84.8	8,480	480.0	480.0
25	1.124				

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each $10\mu gm{-}3$ $PM_{2.5}$ increase
- Counterfactual: $5\mu gm{-}3$

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15	1.06	84.8	8,480	480.0	480.0
25	1.124	89.9			

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each $10\mu gm{-}3$ $PM_{2.5}$ increase
- Counterfactual: $5\mu gm{-}3$

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15	1.06	84.8	8,480	480.0	480.0
25	1.124	89.9	8,989		

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each $10\mu gm{-}3$ $PM_{2.5}$ increase
- Counterfactual: $5\mu gm{-}3$

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15	1.06	84.8	8,480	480.0	480.0
25	1.124	89.9	8,989	988.8	508.8
35					

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each $10\mu gm{-3}$ $PM_{2.5}$ increase
- Counterfactual: $5\mu gm{-3}$

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15	1.06	84.8	8,480	480.0	480.0
25	1.124	89.9	8,989	988.8	508.8
35	1.19				

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each $10\mu gm{-}3$ $PM_{2.5}$ increase
- Counterfactual: $5\mu gm{-}3$

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15	1.06	84.8	8,480	480.0	480.0
25	1.124	89.9	8,989	988.8	508.8
35	1.19	95.3			

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each 10μgm-3 $PM_{2.5}$ increase
- Counterfactual: 5μgm-3

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15	1.06	84.8	8,480	480.0	480.0
25	1.124	89.9	8,989	988.8	508.8
35	1.19	95.3	9,528		

Example: HIA of Air Pollution

- City XYZ, Population: 1,000,000
- Overall Mortality rate: 80/10000
- 6% increase in mortality per each $10\mu gm{-}3$ $PM_{2.5}$ increase
- Counterfactual: $5\mu gm{-}3$

$PM_{2.5}$	RR	Rate	Deaths	Extra	Gain
5	1.00	80.0	8,000		
15	1.06	84.8	8,480	480.0	480.0
25	1.124	89.9	8,989	988.8	508.8
35	1.19	95.3	9,528	1528.1	539.3
45	1.26	101.0	10,100	2099.8	571.7
55	1.34	107.1	10,706	2705.8	606.0
65	1.42	113.5	11,348	3348.2	642.3
75	1.5	120.3	12,029	4029.0	680.9
85	1.59	127.5	12,751	4750.8	721.7
95	1.69	135.2	13,516	5515.8	765.0
105	1.79	143.3	14,327	6326.8	810.9