# GDA of *DoctorAUS* (from **Ecdat**)

**Background** Data from an Australian household survey on health, a cross-sectional study from 1977-78.

**Aims** What kind of people were in the survey? How healthy did they think they were? How did they use the health service?

**Source** Cameron, A.C. and Trivedi P.K. (1998) Regression analysis of count data, Cambridge University Press, http://cameron.econ.ucdavis.edu/racd/racddata.html, chapter 3

**Structure** 5190 observations on 15 variables (2 discretised variables, 9 count variables, 1 ordinal variable, 3 factors)

The four demographic variables are shown in Figure 1. Some data manipulation has been carried out first to ease interpretation. Note that the count axes are all scaled individually and that the scales for the discretised age and income variables are not linear.

```
library(gridExtra)
data(DoctorAUS, package="Ecdat")
DoctorAUS <- DoctorAUS %>% mutate(sex=factor(sex,levels=c(0,1), labels=c("Male","Female")))
DoctorAUS <- DoctorAUS %>% mutate(age=factor(100*age))
DoctorAUS <- DoctorAUS %>% mutate(income=factor(10*income))
a1 <- ggplot(DoctorAUS, aes(sex)) + geom_bar() + ylab("")
a2 <- ggplot(DoctorAUS, aes(age)) + geom_bar() + ylab("")
a3 <- ggplot(DoctorAUS, aes(insurance)) + geom_bar() + ylab("")
a4 <- ggplot(DoctorAUS, aes(income)) + geom_bar() + ylab("")
grid.arrange(a1, a2, a3, a4, nrow=2, widths=c(2,3))
```

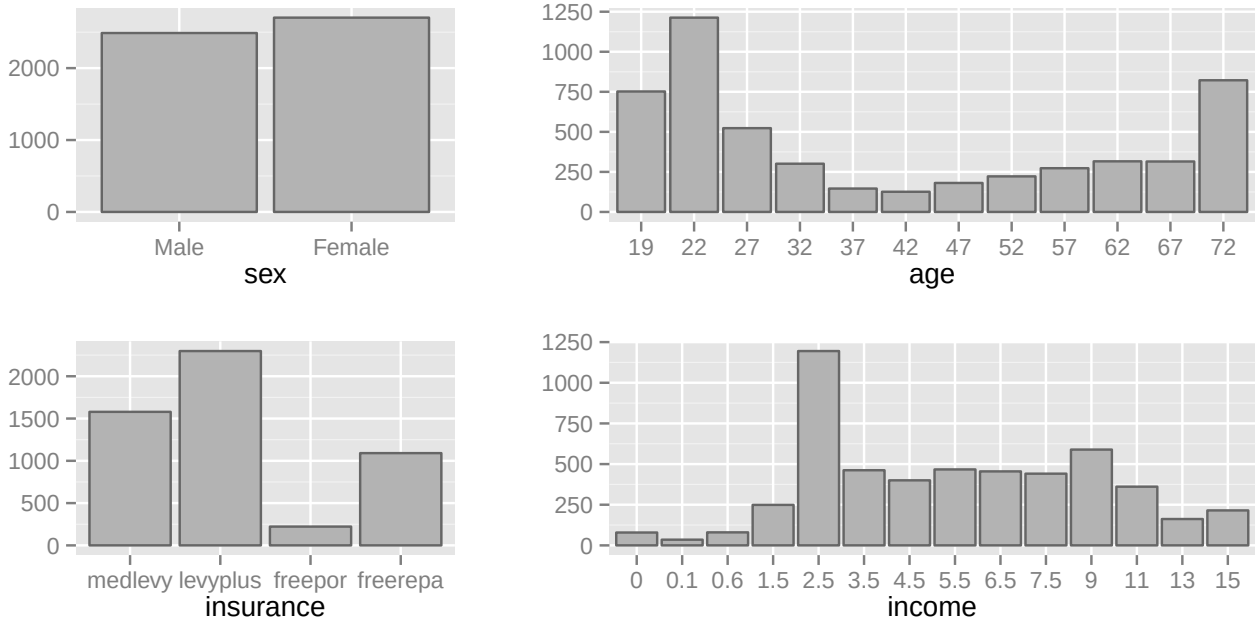


Fig 1: There are slightly more females than males. The age distribution is curiously U-shaped (see below). Almost half have private health insurance ('levyplus') and a few have government insurance due to low income ('freepor'). Income is measured in thousands of Australian dollars and shows a clear mode at 2.5 (for the interval 2001 to 3000). The final age and income categories represent 70+ in age and >14,000 in income respectively.

The age distribution is because only single people were included in the dataset. This explains why there are relatively many young people (not yet married) and old people (divorced or widowed in addition to never married). The next figure takes a closer look at age by looking at the sexes separately.

```
ggplot(DoctorAUS, aes(age)) + geom_bar() + facet_grid(sex~.) + ylab("")
```

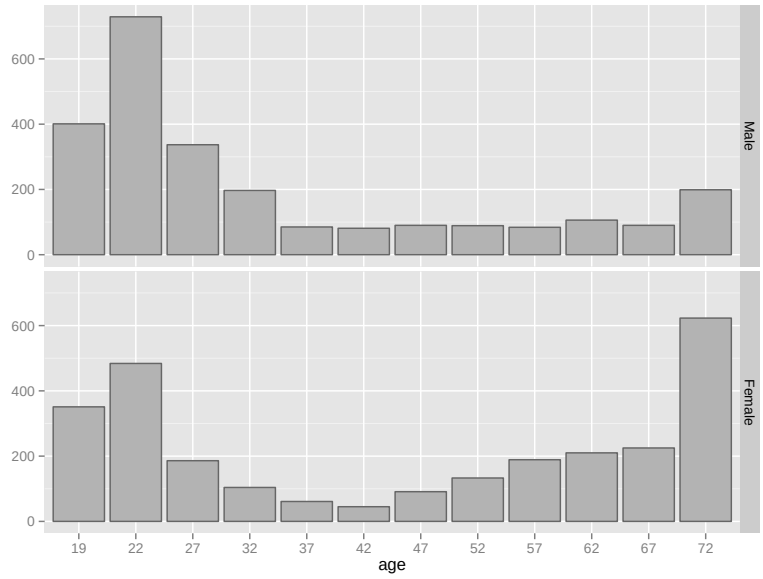


Fig 2: There are rather more young single men in the study than young single women and a lot more old single women than old single men.

The distribution of income may be influenced by the rules for determining how much insurance levy an individual has to pay. Cameron and Trivedi write that it starts for people with an income of $2604.

```
ggplot(DoctorAUS, aes(income)) + geom_bar() + facet_grid(sex~.) + ylab("")
```

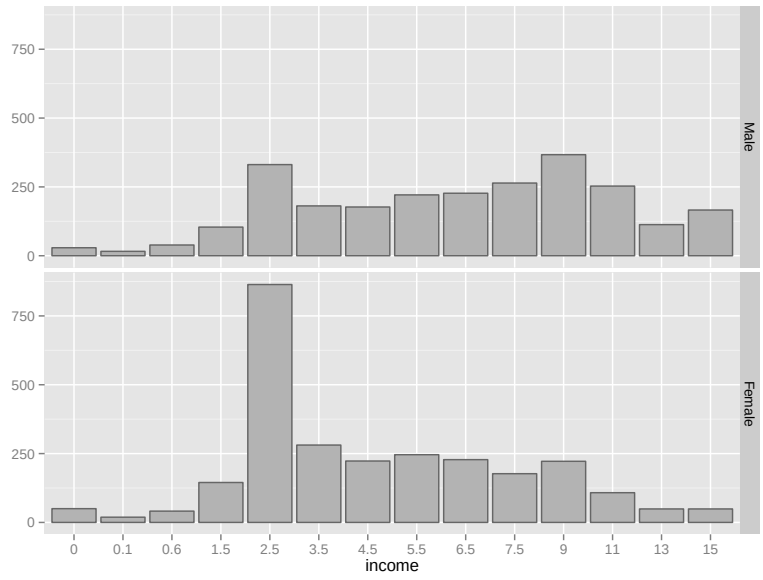


Fig 3: A large proportion of females have income of between two and three thousand dollars a year. The mode for the income distribution for all is predominantly due to females. More males than females have higher incomes.

Age will have some influence on income, so Figure 4 looks at income distribution by sex and age.

```
ggplot(DoctorAUS, aes(income)) + geom_bar() + facet_grid(age~sex) + ylab("") +
  scale_y_continuous(breaks=c(0,200,400))
```

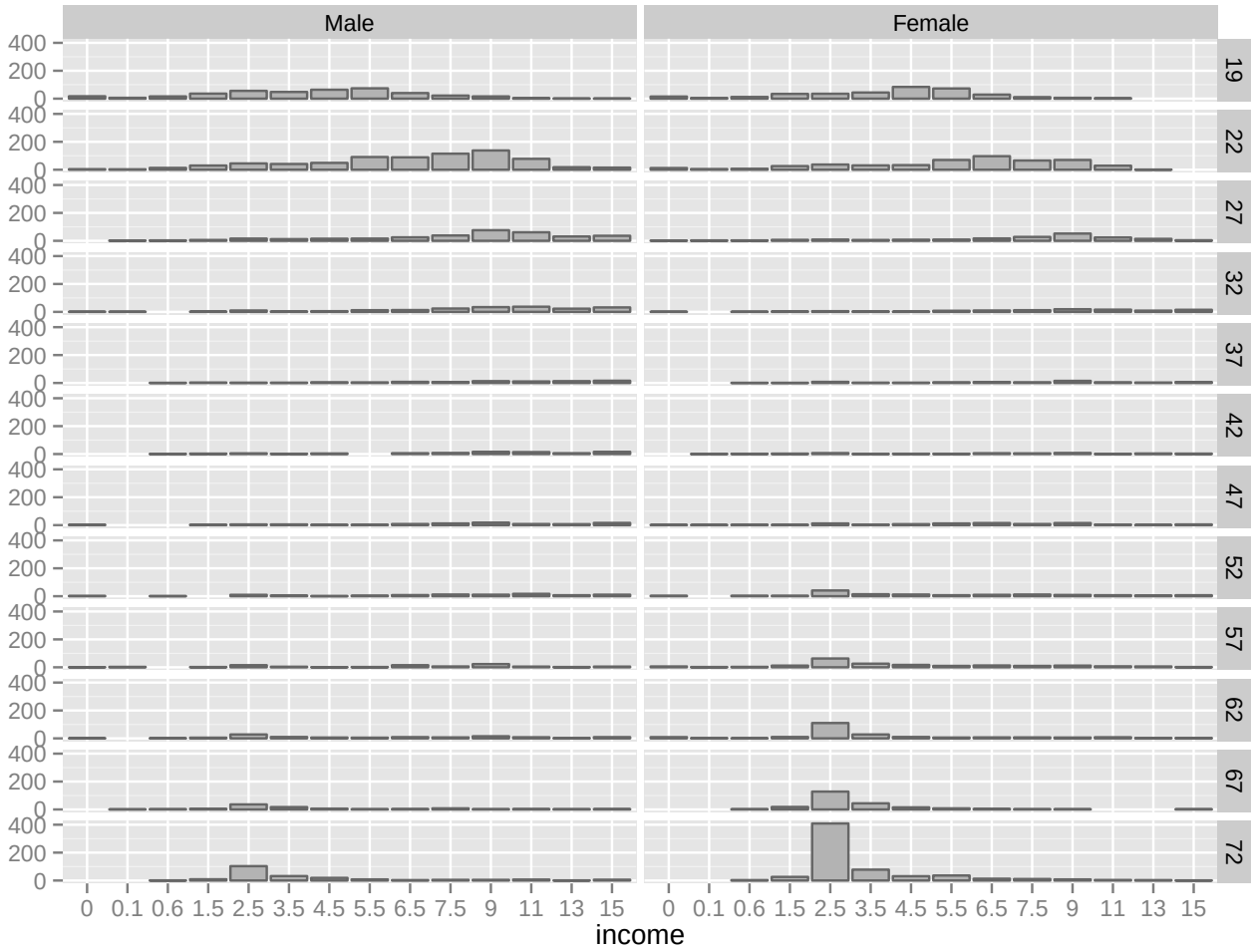


Fig 4: It is now clear that the large group of females in the income class 2.5 are all over 70. There are relatively more males in the class as well, so maybe that is the amount of a standard pension.

Barcharts for chronic condition (chcond) and general health score (hscore) variables are shown in Figure 5.

```
DoctorAUS <- DoctorAUS %>% mutate(chc=factor(chcond, levels=c("np","nla","la")))
b1 <- ggplot(DoctorAUS, aes(chc)) + geom_bar() + ylab("") +
  xlab("chronic condition")
b2 <- ggplot(DoctorAUS, aes(factor(hscore))) + geom_bar() + ylab("") +
  xlab("General health score")
grid.arrange(b1, b2, nrow=1, widths=c(2,3))
```

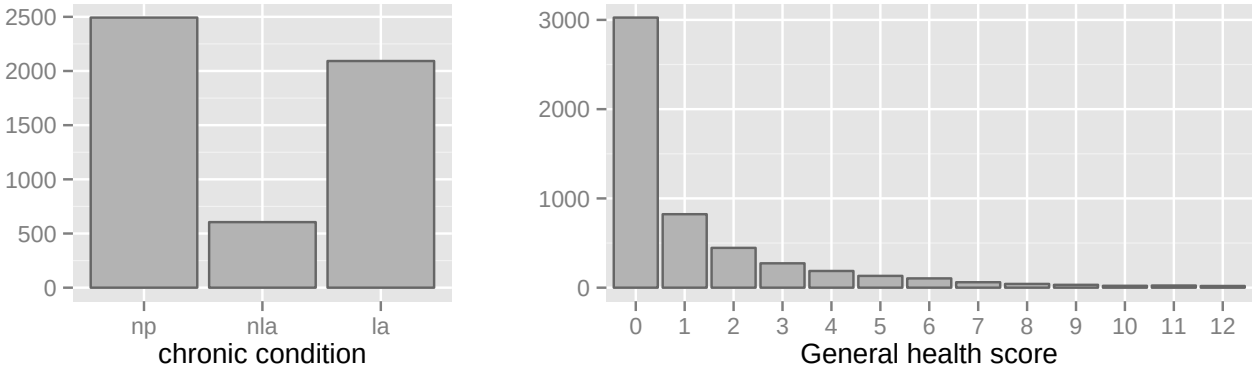


Fig 5: Most people in the survey have a chronic condition which limits their activity or no chronic problem. Around 10% have a chronic condition which does not limit their activity. Some 60% of the survey participants reported good health according to the general health score (apparently this part of the questionnaire was concerned with assessing general well being rather than health).

Two of the remaining variables are plotted in Figure 6. The other seven variables have skew distributions with the majority of cases on any variable having the value 0, reflecting no related health problem.

```
c1 <- ggplot(DoctorAUS, aes(factor(illness))) + geom_bar() + ylab("") +
  xlab("Number of illnesses")
c2 <- ggplot(DoctorAUS, aes(factor(actdays))) + geom_bar() + ylab("") +
  xlab("Number of days of reduced activity")
grid.arrange(c1, c2, nrow=1, widths=c(2,3))
```

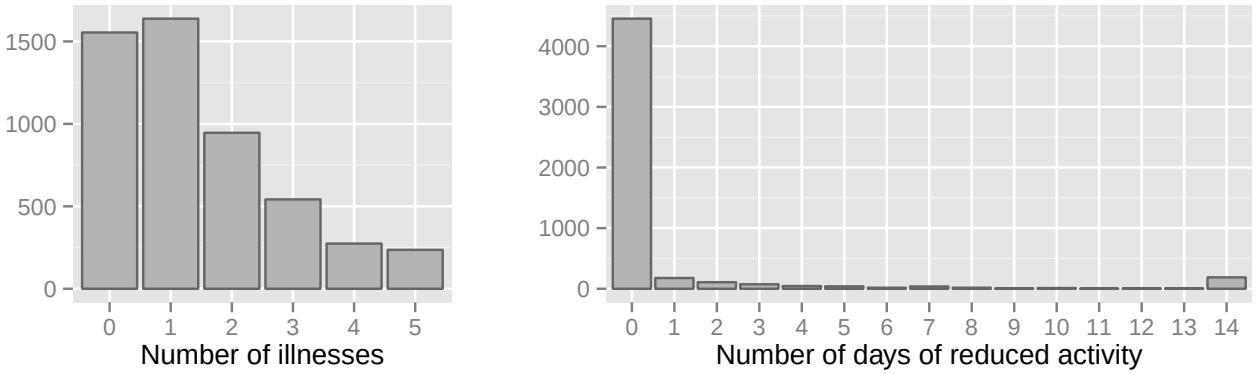


Fig 6: More than two thirds of participants reported having one or more illnesses in the previous two weeks. The vast majority of participants had no days of reduced activity due to illness or injury in the previous two weeks, but a few had reduced activity for the whole two weeks.

It is perhaps unfortunate that the age information has been discretised, especially for the oldest group, where the demand for health services is possibly greatest. A new age variable was created to assess this, grouping the participants into young (18-24), midage (25-64), old (65 and over). As an example, medecine (sic) was used as the dependent variable and type of insurance was included as an additional explanatory variable.

```
DoctorAUS <- within(DoctorAUS, {
  ageZ=factor(age)
  levels(ageZ)=c("young", "young", "midage", "midage", "midage", "midage",
                 "midage", "midage", "midage", "midage", "old", "old")
 })
ggplot(DoctorAUS, aes(factor(medecine))) + geom_bar() + facet_grid(insurance~ageZ) +
  xlab("") + ylab("")
```

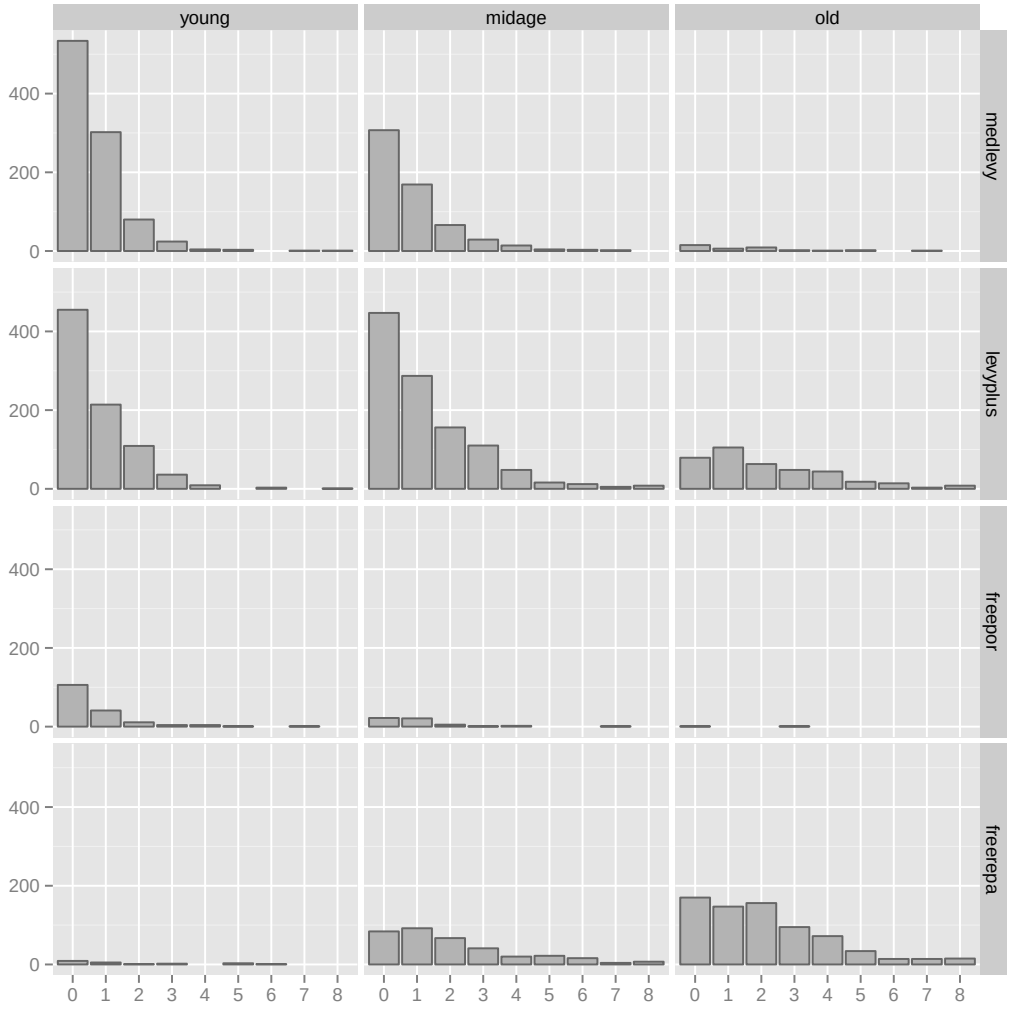


Fig 7: Medlevy and freepor insurances barely exist for old people. Freerepa is specifically for old age disability and veterans. There is some evidence of demand increasing with age in each insurance group, but it is not graphically striking (although probably statistically significant).

The dataset is used in the book by Cameron and Trivedi and in several of their article publications. They modelled count variables as functions of the others. Detailed explanations of the variables can be found in Cameron, A.C. and P.K. Trivedi (1988) "A Microeconometric Model of the Demand for Health Care and Health Insurance in Australia", The Review of Economic Studies, Vol. 55, No. 1, 85-106