Data Visualization and Health Econometrics*

Andrew M Jones
Professor of Economics, Department of Economics and Related Studies, University of York, Heslington, York, YO10 5DD, United Kingdom

Abstract

This article reviews econometric methods for health outcomes and health care costs that are used for prediction and forecasting, risk adjustment, resource allocation, technology assessment and policy evaluation. It focuses on the principles and practical application of data visualization and statistical graphics and how these can enhance applied econometric analysis. Particular attention is devoted to methods for skewed and heavy-tailed distributions. Practical examples show how these methods can be applied to data on individual healthcare costs and health outcomes. Topics include: an introduction to data visualization; data description and regression; generalized linear models; flexible parametric models; semiparametric models; and an application to biomarkers.

---

* I have positions at the Centre for Health Economics, Monash University and the Department of Economics, University of Bergen. Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service. I am grateful to the Economic and Social Research Council for financial support via project "How can biomarkers and genetics improve our understanding of society and health?" (award no. ES/M008592/1). The funders, data creators and UK Data Service have no responsibility for the contents of this article. The article draws on joint work with Apostolos Davillas, James Lomas and Nigel Rice. It uses material presented in a keynote address at the European Health Economics Association (EuHEA) PhD Student-Supervisor Conference, Barcelona 2016 and short courses presented at the Australasian Workshop on Econometrics and Health Economics 2014, the Bavarian Graduate Programme in Economics 2014, the CINCH Academy Essen 2015, the Conference of the European Health Economics Association (EuHEA) Hamburg 2016.

Contents

1

Introduction

Econometric models for health outcomes and health care costs are used for prediction and forecasting in health care planning, risk adjustment by insurers and public providers of health care, geographic resource allocation, health technology assessment and health policy impact evaluations. Methods for risk adjustment focus on predicting the treatment costs for particular types of patient, often with very large survey or administrative datasets.

Microdata for individual medical expenditures and costs of treatment are typically non-normal. Survey data often feature a spike at zero, if there are non-users in the data. Both survey and administrative data, such as registers and discharge records, typically have a heavily skewed distribution and heavy tails. The spike at zero is often modelled by a two-part specification, with a binary choice model for the probability of any costs, and a conditional regression model for the positive costs (Jones, 2000). Due to the skewness and excess kurtosis of the data and the importance of influential observations, regression models applied directly to the raw data on the level of costs can perform poorly. Traditionally the positive observations have been transformed prior to fitting a regression model, most often by taking a logarithmic or, sometimes, a square root transformation. Once these models have been fitted then predictions have to be retransformed back to the original – raw cost – scale. This is not straightforward to do in a robust way, especially if there is heteroskedasticity in the data on the transformed scale (Manning, 1998; Manning and Mullahy, 2001; Mullahy,1998).

In the recent literature attention has shifted away from linear regression models to semiparametric and flexible parametric estimators. A popular semiparametric approach is to use generalized linear models or GLMs (e.g., Buntin and Zaslavsky, 2004; Manning and Mullahy, 2001; Manning et al., 2005; Manning, 2006). GLMs are built around a *link function* that specifies the relationship between the conditional mean and a linear function of the covariates and a *distributional family* that specifies the form of the conditional variance as a function of the conditional mean. GLM models are estimated using a quasi-likelihood approach derived from the quasi-score or "estimating equations".

In a conventional GLM the choice of link and distribution has to be specified a priori. In practice the most frequently used GLM specification for medical costs has been the log-link with a gamma variance (Blough et al., 1999; Manning and Mullahy, 2001; Manning et al., 2005). Basu and Rathouz (2005) have developed a flexible semiparametric approach to the problem of selecting the appropriate link and variance functions. Their extended estimating equations estimator (EEE) approach uses a Box-Cox transformation for the link function and either a power variance or quadratic variance function for the distribution. The particular form of the link and distribution are thereby estimated from the data at hand.

Other semiparametric methods that have appeared in the literature on modelling health care costs include the conditional density estimator and finite mixture models. The conditional density approach was advocated by Gilleskie and Mroz (2004) and divides the support of the distribution of the dependent variable into discrete intervals then applies discrete hazard models to these, implemented in practice as a series of sequential logit models. Finite mixture models use a discrete mixture of parametric models and, for example, have been applied to medical costs by Conway and Deb (2005). Combining simple distributions such as the gamma or log-normal in a mixture of relatively few components may approximate complex empirical distributions effectively, especially for distributions that are multi-modal.

In contrast to semiparametric methods, flexible parametric methods fully specify the distribution for health care costs. Building on standard distributions such as the log-normal and gamma distributions, they move to more flexible 3 and 4-parameter distributions such as the generalized gamma and the generalized

beta distribution of the second kind (GB2). This provides the additional flexibility to fit the high level of skewness and the heavy tails seen in cost data (Jones et al., 2014). The downside of this flexibility is a risk of over-fitting and, in practice, these approaches may be best used as a guide to selecting one of the special or limiting cases that are nested within the general models. In this respect the flexible parametric models can play a similar role to using the EEE approach to select the link and distribution functions to be used in a GLM.

Earlier literature reviews have synthesised and compared the wide range of approaches to modelling health care costs (e.g., Hill and Miller, 2010; Jones, 2000, 2011; Jones et al., 2013; Mullahy, 2009). In addition, studies using a quasi Monte Carlo design, based on English administrative data for patient level costs of hospital care, have provided an assessment of the relative performance of these approaches (Jones et al., 2014, 2015, 2016). To complement these earlier studies this article focuses on the principles and practice of data visualization and statistical graphics and how these can enhance empirical analysis of health care costs and outcomes, especially for skewed and heavy-tailed distributions. The scope of this review is limited to non-normal but continuous outcomes such as health care costs and biomarkers. Many health economics applications deal with categorical and ordered outcomes, count data, or duration data. Methods for these are reviewed in Jones (2000) and Jones et al. (2013). The methods and applications used here are limited to cross sectional data. For discussions of methods for panel data see Jones (2009) and for the use of cohort data (Von Hinke Kessler Scholder and Jones (2015).

Practical examples show how these graphical methods can be applied using the software package Stata, which is widely used in applied econometrics. Stata is not the obvious software of choice for specialist work in data visualization especially for users who wish to present their work online and to make use of animation or interactivity. Nevertheless, for many applied econometricians it is the workhorse for data management and econometric analysis. In this article Stata code, shown in the font `courier new`, is included to show how far it is possible to go within Stata so that graphical analysis can be integrated with statistical and econometric analysis within one piece of software and using one set of syntax.

The review of methods that have been developed for health care cost regressions is complemented by an empirical case study that

focuses on objectively measured health outcomes, whose distributions share many of the features of cost data. The case study applies the econometric and graphical methods to blood-based biomarkers as the dependent variables. The dataset is the UK Household Longitudinal Study (UKHLS), known as Understanding Society, which is a large nationally representative longitudinal study (Benzeval et al. 2016).

2

Data Visualization – a Primer

"Greatest number of ideas in the shortest time with the least ink in the smallest space"

Edward Tufte (1983)

Edward Tufte's ideas, expressed in his 1983 book *The Visual Display of Quantitative Information* and in subsequent publications, have been highly influential in the field of data visualization (Tufte 1983, 1990, 1997, 2001, 2006). Tufte has developed a set of principles of graphical excellence which are summarised here:

- Show the data.
- Induce thought about substance (not methods, visual style…).
- Do not distort the data.
- Present many numbers in a small space.
- Make large data coherent.
- Encourage the eye to compare.
- Reveal levels of details – from overview to fine structure.
- Serve a clear purpose (whether it be description, exploration, decoration, etc).
- Integrate the graphics within the analysis.

It is clear that this perspective favours functionalism, minimalism and clarity of design over embellishment and visual "fireworks". Tufte's influence is evident in the recent article by Jonathan Schwabish (2014) in the *Journal of Economic Literature,* "An economist's guide to visualizing data", which aims to bring Tufte's philosophy of graphical design to an economics audience by critically appraising and redesigning graphics that have appeared in articles published in the *American Economic Review*.

In addition to Edward Tufte's work other sources for the ideas and material presented in this brief primer include the books of Stephen Few (2009, 2012, 2013, 2015) which share the same focus on simplicity of design and clarity of purpose but with a more pragmatic approach aimed at those providing business information in tabular and graphical formats. Few's web page, *Perceptual Edge,* provides a useful source of case studies and critical appraisals of published visualizations (see the Appendix for information on this and other web links that are relevant sources of ideas for data visualization).

Nathan Yau's blog *Flowing Data* and his books *Visualize This* and *Data Points* (Yau, 2011, 2013) provide a visually elegant and contemporary guide to good practice that is rooted in statistical analysis with an emphasis on online resources and interactive graphics.

Alberto Cairo (2012, 2016) comes from a background of experience in data journalism and working with infographics. His books *The Functional Art* and *The Truthful Art* draw on the lessons of cognitive psychology and their implications for graphical design. Jorge Camoes's (2016) *Data at Work* shares a similar perspective. It provides an impressive glimpse of how far Microsoft Excel can be taken to produce effective and visually appealing graphics.

Lessons for statistical graphics from the psychology of visual perception are tested and put into practice in a classic article by Cleveland and McGill (1984) and are covered in a book by William Cleveland (1985) *The Elements of Graphing Data*. These lessons are put into practical form by Naomi Robbins (2005) in her book *Creating More Effective Graphs*.

A useful guide for anyone who wishes to explore and select graphical methods and see how they can be implemented within Stata is Michael Mitchell's (2012) *A Visual Guide to Stata Graphics*. The Guide allows readers to browse a catalogue of graphical styles, all produced in Stata, to find the one that best suits their own needs and to see the

corresponding code alongside. Mitchell's Guide is complemented by an encyclopaedic source of the various forms of statistical graphic that have been used in practice across a broad range of disciplines in Robert Harris's (1999) *Information Graphics: A Comprehensive Illustrated Reference*.

One feature of Mitchell (2012) is to stress the usefulness of graphics schemes in Stata. Schemes are text files that can be called upon to set options that affect the appearance of Stata graphics. This saves having to add numerous sub-commands when using individual graphics commands. The default colour graphics scheme used by Stata is called `s2color`, which can be set explicitly using:

```
set scheme s2color
```

Many other schemes are available. For example to adopt the visual style of the Economist magazine use:

```
set scheme economist
```

Alternatively a custom written scheme can be created and installed. The graphics illustrated in this article were mainly done using a very simple custom scheme that begins by including the default scheme and then modifies some of the colour options that are used by default to give a palette based on dark orange:

```
#include s2color
color background ltbluishgray
color histogram dkorange
color boxplot dkorgange
color p2 dkorange
color bar dkorange
color hbar dkorange
```

Figure 2.1 compares **histograms** and overlaid **kernel density plots** of the same set of raw data on the logarithm of annual medical costs produced using the custom scheme shown above and the Economist scheme. These schemes produce graphics that appear quite different despite displaying the same information.

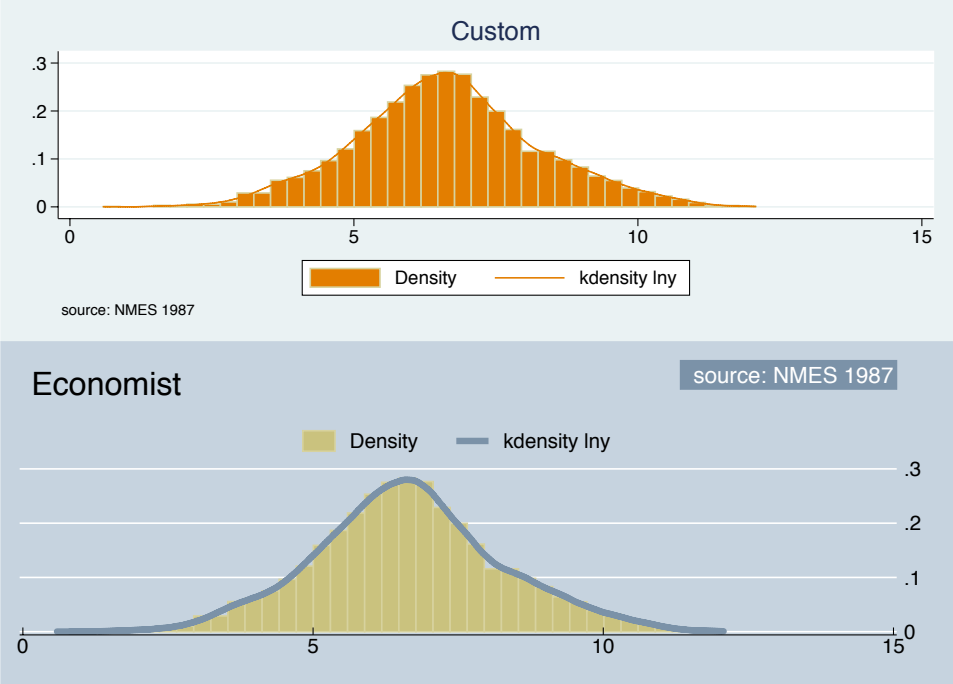


Fig 2.1 Histograms drawn with different schemes

Returning to Edward Tufte. Following his principles of graphical excellence, Tufte (1983) has identified many of the pitfalls that can arise when visualisation is done badly. These include:

Distortion: when a graphic creates a distorted picture of the underlying data. For example if the angles in a pie chart or the areas in a histogram do not match the variation in the data. A classic example is the way in which a non-zero baseline in a bar chart can be used to create a biased impression of the difference between the height of the bars.

Design variation: when the visual features of a graphic, such as shading or colouring do not match the variation in data that is being represented. The implication is that in good visual design the graphical variation - whether it is differences in the position of points, lengths of lines, size of areas or differences in shading or colour - will reflect and illuminate the underlying statistical variation in the data and hence will make the variation visible and "encourage the eye to compare".

"Chartjunk": which is essentially the use of uninformative embellishment in graphics, such as the vibrations and grid patterns

that are sometimes used to fill bars in a bar chart. A notorious example is the use of 3-D effects in graphics, such as bar charts, where the addition of 3-D typically reduces the clarity of the graphic.

Lack of content: where excessive space is devoted to graphing relatively little quantitative data. This is the antithesis of Tufte's call to present many numbers in a small space. In these cases tabulating may be more effective than graphing the data.

To put these ideas into practice and to help sift effective from ineffective graphics, Tufte (1983) introduced a couple of simple concepts that can be used to evaluate a graphic: the *data-ink ratio* and the *data density*. Data ink corresponds to marks on the page that represent data points, such as the dots in a scatter plot, while the remaining non-data ink represents all the other marks, which may be functional such as titles, labels, axes, gridlines and tick points or may be purely decorative. So:

$$\text{Data-ink ratio} = \text{data ink/total ink}$$

Implicit in this concept is the notion that the data-ink ratio equals one minus the proportion of ink that could be erased without loss of statistical information and that the design of a graphic should aim to minimize non-data ink. In practice there are limits on how far this can be taken as non-data ink often helps the eye to navigate and interpret a graphic and may help to make it eye-catching and memorable.

The data density takes the number of data points that are being graphed relative to the physical size of the graphic to capture the principle of aiming to present many numbers in a small space:

$$\text{Data density} = \frac{\text{number of entries in data matrix}}{\text{area of data graphic}}$$

For example this may come up with a value such as 3.8 numbers per square centimetre. This notion is rooted in graphics that appear on a printed page, whereas many now appear in digital format and are scalable, but the basic idea here is that, within reason, the data density should be maximized. This is an idea that is especially relevant for work with microdata and, in particular, 'big data' where graphics provide a way of managing the dimensionality of large volumes of data.

Tufte has suggested a related concept: the *Shrink principle* that "graphics can be shrunk way down" and that readers can cope with quite small graphics and with multiple graphics presented alongside each other. This principle is embodied in the use of **information dashboards** to combine multiple graphical and tabular views of data in a single dashboard format (Few, 2013).

Tufte is not alone in taking a prescriptive approach to visualization. Cairo (2016) also provides a couple of 'checklists' to guide practitioners. In the first he argues that "a good visualization is:

- reliable information;
- visually encoded so relevant patterns become noticeable;
- organised in a way that enables at least some exploration, when it's appropriate;
- and presented in an attractive manner, but always remembering that honesty, clarity and depth come first".

In the second, influenced by Enrico Bertini, Cairo (2016) argues that the qualities of a good chart are:

- It is truthful (based on thorough and honest research).
- It is functional (accurate depiction of the data).
- It is beautiful (pleasing for audience).
- It is insightful (reveals evidence).
- It is enlightening (changes minds for the better).

To illustrate how Edward Tufte's ideas might be put into practice start with a histogram of individual data on annual medical costs. These data are from the 1987 US National Medical Expenditure Survey (NMES). Data that have been used as part of the evidence surrounding tobacco litigation in the United States and in analyses of the health care costs attributable to smoking (Rubin, 2001; Johnson et al., 2003; Imai and Van Dyk, 2004). The dataset used here includes total annual health care costs measured in US dollars, measures of smoking that include indicators for current, ex- and never smokers and a variable for pack years, the product of years of smoking and packs per day, along with controls for age, sex, education, marital status, poverty status and region[1].
In Figure 2.2 the top left-hand panel shows the histogram for the logarithm of medical costs that is produced using the Stata default graphics settings. To modify these the top right-hand panel removes

[1] I am grateful to Elizabeth Johnson for supplying the dataset used in her paper.

the axis lines, which are non-data ink, and rotates the labels on the axis to make them more legible, while adding a note of the data source. The bottom-left panel takes the elimination of non-data ink a step further by removing the grid lines while keeping them as a reference point by superimposing white lines over the shaded area of the histogram and by removing the axis label. This process of elimination could have been taken further. For example, the shading of the bars is non-data ink and could be eliminated to leave hollow outlines, or simply horizontal lines to mark the height of each bar. In practice the elimination of non-data ink typically reaches a point where further simplification hampers the legibility of a graphic and here the process stops at shaded bars. In this case the purpose of the histogram is more qualitative, to give an impression of the overall shape of the distribution - is it symmetric, unimodal, heavy tailed? - rather than to read off precise quantitative information about the height of the density at different points. So the qualitative contrast between the shaded bars and the white background serves this purpose well. Finally, the bottom-right panel adds some further information by super-imposing a kernel density plot to smooth out the outline of the distribution.

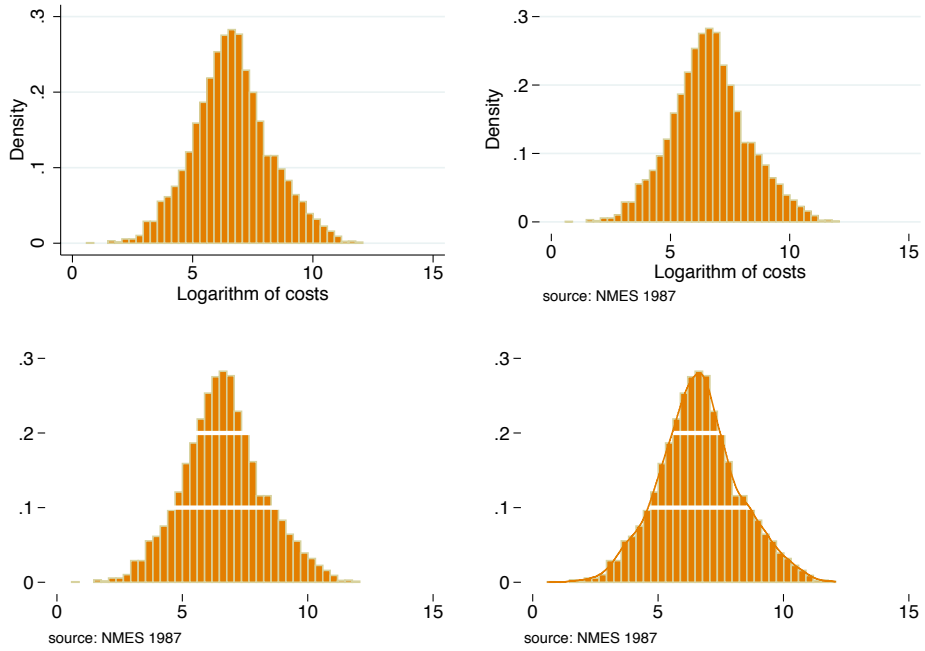


Fig 2.2 Refining the histogram

Now consider the default version of a **scatter plot** produced in Stata. Figure 2.3 shows the logarithm of annual health care costs plotted

against a measure of individual smoking history based on pack years and reveals their joint distribution through the density of data points. This was produced and saved as an encapsulated postscript (eps) file to be used for publication. This is a vector graphics format which, like pdf format, is scalable and should look sharp at any size. This is an alternative to raster graphics formats such as tif, jpeg and gif files. The following code is used:

```
tw scatter lny treat, title(Default scatter plot)
   name(scatter, replace)
graph export scatter.eps, replace
```

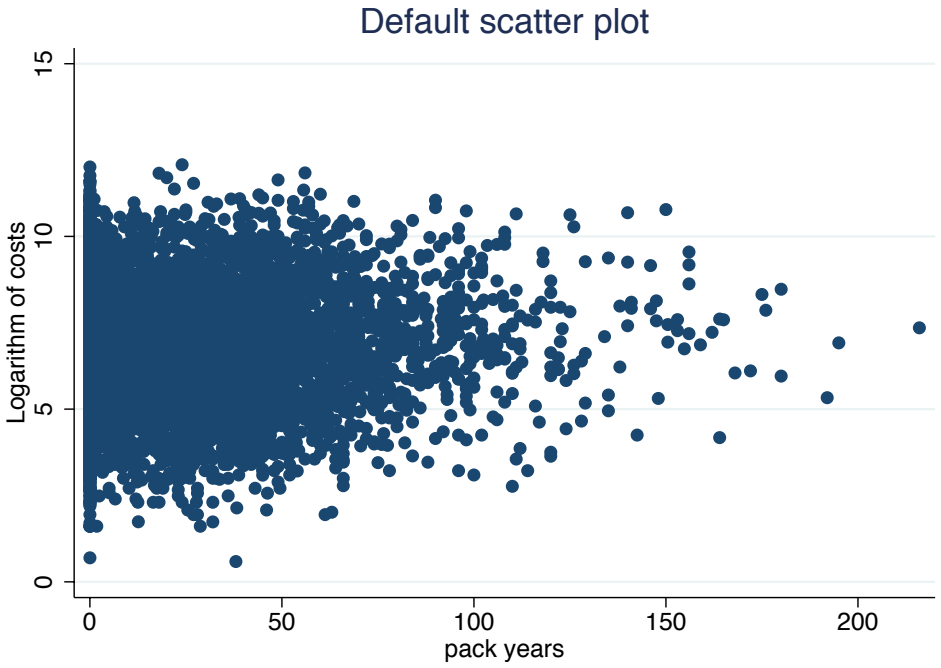


Fig 2.3 Scatter plot of log(costs) and pack years

One way to reduce the proportion of non-data ink used in this standard scatter plot is to replace the standard horizontal and vertical axes by a **range frame plot** so that the length of the lines that appear on the axes represent the range of the y and x variables and hence become data themselves. This is show in Figure 2.4. Note that the default scatter of 'dots' has been replaced by a cloud plot so that individual data points become visible and better represent the use of microdata and the extensive heterogeneity that is typical of such data. Also the tick points on the vertical axis have been rotated so

that they are easier to read. In this case using the range on the horizontal axis does not really add much but the vertical axis does reveal the limits on the range of the logarithm of costs.

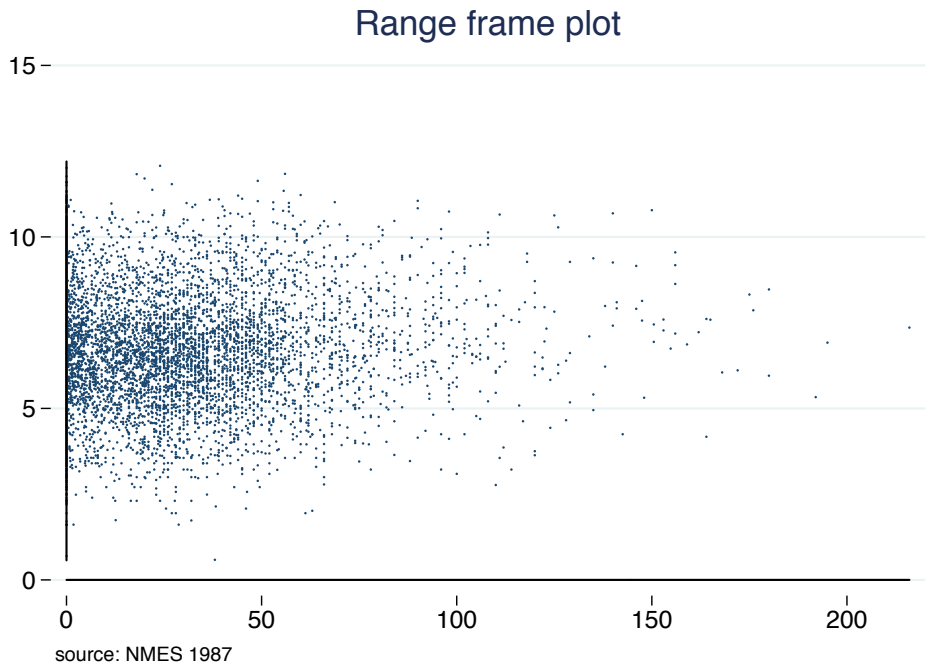


Fig 2.4 Range-frame plot of log(costs) and pack years

The range-frame plot was drawn with the following code where the axes are drawn as functions:

```
tw (scatter lny treat, msymbol(p) legend(off)
    ylabel( , angle(horizontal)) )
   (function y=0, range(0 216) lcolor(black))
   (function y=0, range(0.57 12.2) lcolor(black)
    horizontal), title(Range frame plot)
    ysca(noline) xsca(noline) name(rangeframe,
    replace) note($note)
```

A more informative alternative to the range frame plot is a **dot-dash plot** where the marginal distributions of y and x are shown using dashes for individual data points. Figure 2.5 reveals detail in the tails of these marginal distributions rather than just showing the overall range of the data so that the horizontal axis now becomes

informative as well as the vertical. In particular the skewness and sparsity of the observations in the right-hand tail of the distribution of pack years is revealed.

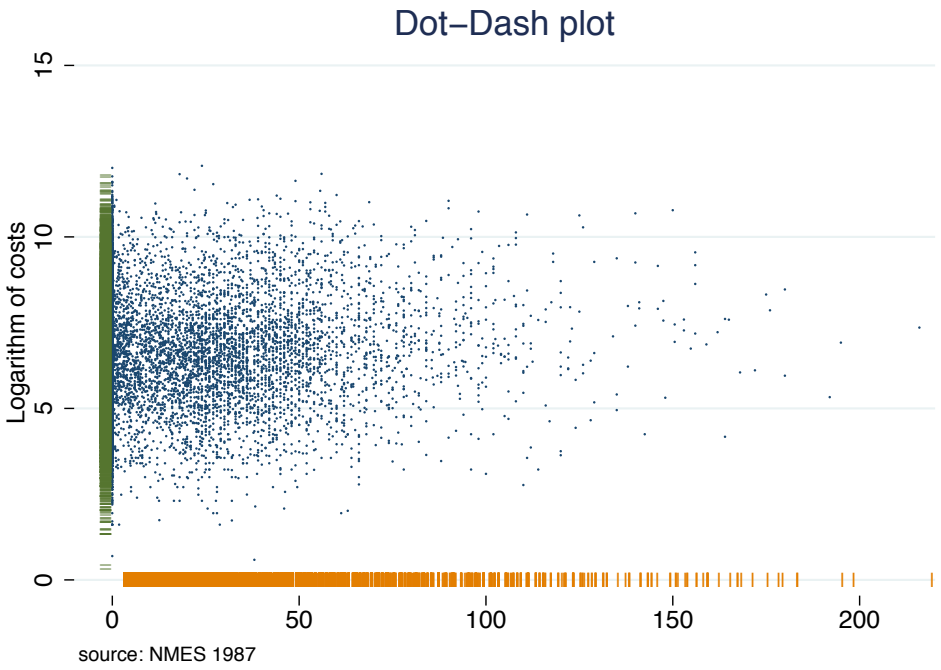


Fig 2.5 Dot-dash plot of log(costs) and pack years

The dot-dash plot was produced using the following code where the axes are drawn using scatter plots:

```
* Dot-Dash plot
gen where_x=0
gen where_y=-6
gen pipe = "|"
gen bar ="_"
tw (scatter lny treat, msymbol(p) xsca(noline)
    ysca(noline) legend(off)
   (scatter where_xtreat, ms(none) mlabel(pipe))
   (scatter lny where_y, ms(none) mlabel(bar)),
    note($note) title(Dot-Dash plot) name(dotdash,
    replace)
```

The idea of turning the axes of the figure into data can be taken a step further by showing the marginal distributions of y and x as histograms. This complements the scatter of points that show their joint distribution. At the same time both linear and quadratic fits of the data are added to the scatter plot to reflect the conditional relationship between y and x (see Figure 2.6).

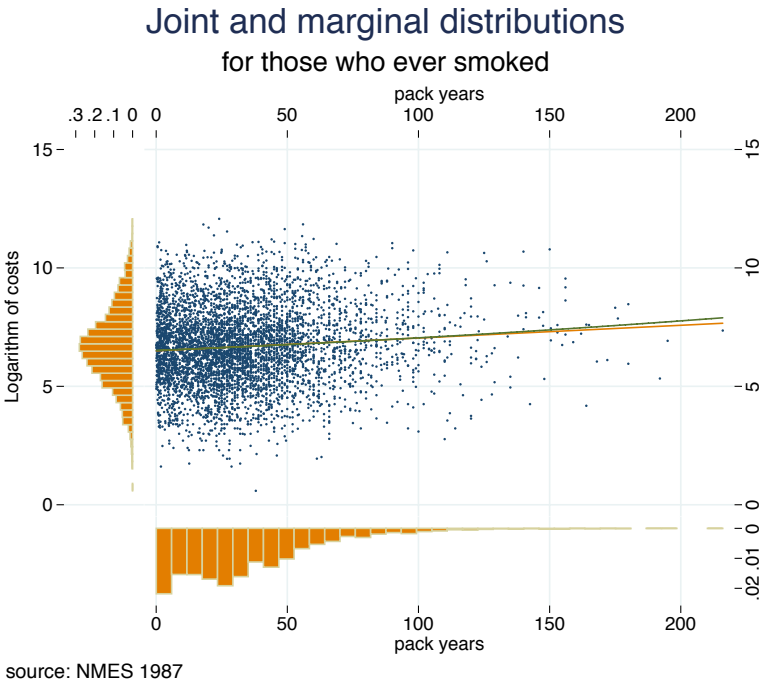


Fig 2.6 Scatter plot and histograms combined

The Stata code required to do this is rather trickier in order to correctly rotate and align the three graphs that are combined in the final image:

```
clonevar y_s=lny
clonevar x_s=treat
tw  (scatter y_s x_s, msymbol(p))
    (lfit y_s x_s)
    (qfit y_s x_s)
     if d==1,legend(off) yscale(alt noline)
     xscale(alt noline) xlabel(,grid gmax
     angle(horizontal))
     ylabel( , angle(vertical)) saving(yt, replace)
```

```
quietly tw histogram y_s if d==1,
     xsca(alt reverse noline) ysca(noline)
     fxsize(20) horiz ylabel( , angle(horizontal))
     xt(" ") saving(hy,replace)
quietly tw histogram x_s if d==1,
     yscale(alt reverse noline) xsca(noline)
     ylabel(0(.01).02, nogrid) yt("")
     xlabel(, grid gmax) fysize(20) saving(ht,
     replace)
graph combine hy.gph yt.gph ht.gph, hole(3)
     imargin(0 0 0 0) graphregion(margin(l=15
     r=15)) note($note) title("Joint and marginal
     distributions") subtitle("for those who ever
     smoked") saving(scatter&margs, replace)
```

Note in particular how the x and y scales for the component graphics have to be reversed (`reverse`) and moved to the alternative side (`alt`) of the graphics from the default settings and how the axis lines are removed (`noline`). The graphic is drawn for the sub-sample who have ever smoked, reflected in the condition `if d==1`. The command `graph combine` allows the three component graphics to be spliced together and note how the `hole(3)` subcommand is used to control their positioning by placing a space in the bottom left-hand corner.

Experimental research and statistical analysis by Cleveland and McGill (1984) has been highly influential and is cited by many recent authors (e.g., Cairo, 2012; Robbins, 2005; Yau, 2011). They demonstrated how there is a hierarchy in terms of the perception of visual cues and in how accurately people are able to read quantitative information from these cues. The hierarchy runs from the most accurate which is our ability to distinguish points along a common scale, or on different scales, then through differences in length, angles, direction, areas, volumes, shading/saturation, and finally colour/hue. The position of volume in this list is one reason that 3-D plots are inadvisable compared to simple 2-D plots.

If numerical precision is the primary goal of a graphic then the hierarchy of visual cues can be a helpful guide. For example, Cleveland and McGill (1984) show that a **dot plot** is likely to be more effective than a **pie chart**, that uses comparisons of angles. Similarly a **bar chart**, that also uses comparison of points along a common scale, will be more effective than a **tree map**, that relies on comparisons of areas.

Cairo (2013) and Nussbaumer Knaflic (2015) draw on further lessons from cognitive psychology and discuss the ways in which the gestalt principles of visual perception can be used to aid graphical design. These principles include:

Proximity: objects which appear close to each other are perceived as groups.

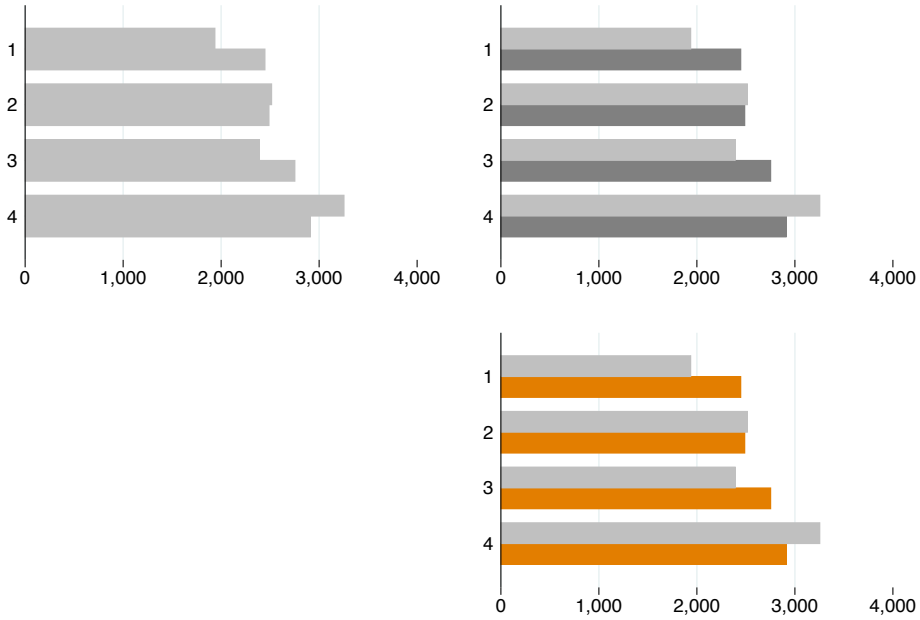
Similarity: visually identical objects are perceived as belonging to a group.

Connectedness: objects that are linked, for example by a line, are perceived as a group.

Closure: objects within crisp boundaries are perceived as belonging to a group.

Continuity: it is easier to perceive shapes when contours are smooth and rounded.

As an illustration of some of these ideas consider the horizontal bar chart in Figure 2.7.

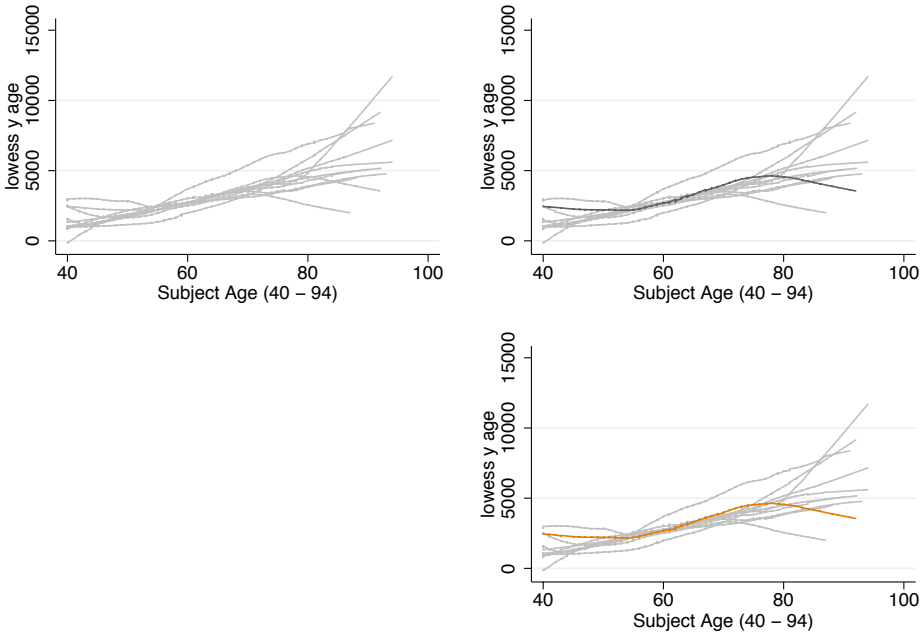


source: NMES 1987

Fig 2.7 Using saturation and hue to suggest groups

The length of the bars represents average levels of medical costs and these are split between those who had never smoked and those who had ever smoked. So, despite the fact that there are no labels on the graphic, due to their proximity and the use of space to separate them, the pairs of bars labelled 1, 2, 3 and 4 are likely to be perceived as a group. These groups do in fact correspond to four levels of educational attainment. In the top left panel it is hard to distinguish the bars for smokers and non-smokers as they have the same shading and colour. The distinction becomes clearer in the other two panels that use differences in shading or in hue. Again there are no labels to make this explicit but the dark shaded bars in the top panel or the dark orange bars in the bottom panel will be perceived as a group. In this case the darker/orange bars represent the smokers within each level of education.

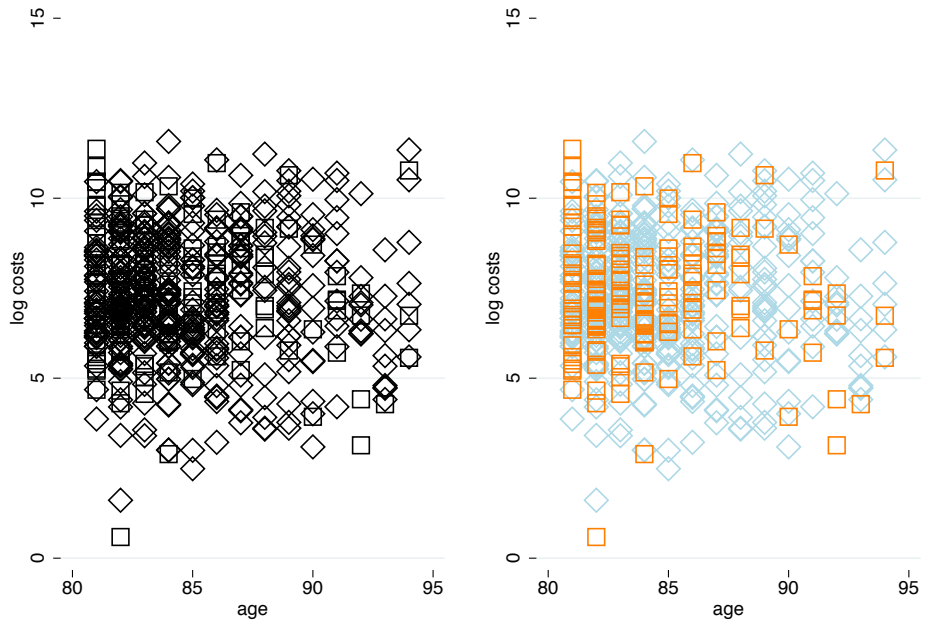
Similarity and difference can be used to draw attention to particular features of a graphic and to highlight certain results. Figure 2.8 shows lowess fits of the relationship between medical costs and age for those aged over 40 for different sub-samples. The right-hand panels show how contrasts in saturation or hue can be used to highlight one particular line, such as the results for the full sample or for a particular group of interest.



source: NMES 1987

Fig 2.8 Using saturation and hue to highlight

Figure 2.9 uses the idea that identical objects will be perceived as a group, here squares are used to denote smokers and diamonds are non-smokers in a scatter plot of the logarithm of medical costs against age for those aged 80 and over. The difference between the two groups is clearer to see when colour is added and the orange squares can be distinguished from the blue diamonds.



source: NMES 1987

Fig 2.9 Using shape and hue to suggest groups

The code for Figure 2.9, which sets the shape, size and colour used for the data points, is:

```
tw (scatter lny $xc if d==0 & age>80, msize(vlarge)
    msymbol(Dh) mcolor(black) ysca(noline)
    xsca(noline) legend(off) ytitle(log costs))
   (scatter lny $xc if d==1 & age>80, msize(vlarge)
    msymbol(Sh) mcolor(black) ysca(noline)
    xsca(noline) legend(off) ytitle(log costs)
    xtitle(age)), saving(sc1, replace)
tw (scatter lny $xc if d==0 & age>80, msize(vlarge)
    msymbol(Dh) mcolor(ltblue) ysca(noline)
    xsca(noline) legend(off) ytitle(log costs))
   (scatter lny $xc if d==1 & age>80, msize(vlarge)
    msymbol(Sh) mcolor(orange) ysca(noline)
    xsca(noline) legend(off) ytitle(log costs)
```

```
    xtitle(age)), saving(sc2, replace)
graph combine sc1.gph sc2.gph, note($note)
```

The findings of Cleveland and McGill (1984) on how visual cues are perceived may create the impression that points along a common scale should always be preferred over the use of differences in saturation or hue. In fact, extracting precise quantitative comparisons is only one purpose of an effective graphic and very often making broad qualitative comparisons is important as well. The lesson of the preceding examples is that the use colour or shading often works best to create an immediate and bold impression of groups and proximity.

Camoes (2016) summarises the uses of colour in graphics as being to:

- Categorise.
- Group.
- Emphasize.
- Sequence.
- Diverge.
- Alert.

A software package, such as Stata, will typically offer a palette of "built in" colours such as the shade of `dkorange` used in the custom scheme for this article. The available colours can be found under `colorstyle`. At a more fundamental level colour is typically defined and controlled using colour systems that can be used to mix your own colours. The *RGB model* defines colours in terms of their mix of red, green and blue. For example:

(0,0,0) - black
(255,0,0) - red
(0,255,0) - green
(0,0,255) - blue
(255,0,255) - magenta
(255,255,255) - white

Alternatively the *HSL model* defines colours in terms of their hue, saturation and luminance where each element varies from 0 to 255. Hues varies from red (0) to violet (255) and a luminance score of 128 corresponds to pure colour. Stata uses a variant of this known as the HSV (hue, saturation and value) where the first term is on a 360 degree scale and the other two are expressed as proportions.

Camoes (2016) notes that when colours are ordered by hue there is an analogy with qualitative (ordinal or categorical) data. While when they are ordered by luminance there is an analogy with continuous quantitative variables. So, for example, hue can be used to create a “diverging scale” that might be used to represent data from a Likert scale in a sequence such as red-orange-yellow-blue. This is typical of the **heat map** and **tree map** styles of graph. While “colour ramps”, created by selecting a hue and progressively changing the level of luminance, can capture the sense of continuous variation (if not the magnitudes). In Stata this can be done by taking a particular colorstyle and modifying its intensity, for example, `dkorange*.6`. Or if an RGB value is used, `0 255 255*.8`.

Figure 2.10 uses a diverging scale to represent self-assessed health (SAH) data. SAH is an ordinal variable with responses ranging from poor, fair, good, very good, to excellent, This is represented by a scale that borrows the idea of a heat map and runs from a cool blue for excellent health to a hot red for poor health, The **bar charts** show the distribution of SAH over vigiciles of household income ranging from 1 which is the poorest group to 20 which is the richest. The greater concentration of poor and fair health among poorer households is clear.

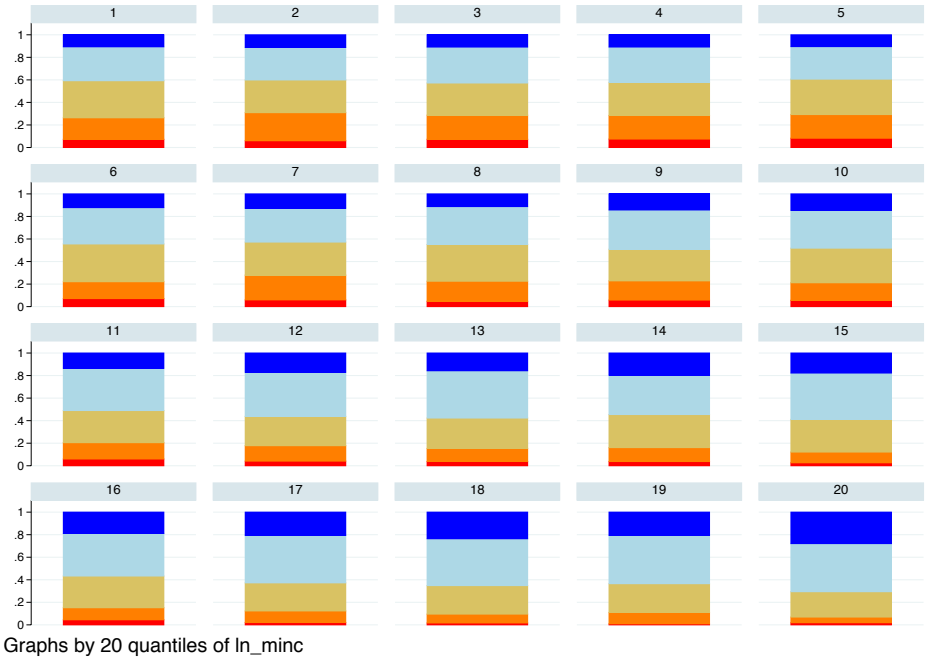


Fig 2.10 Income gradients in SAH in UKHLS

Figure 2.11 shows the relationship between cholesterol, age, gender and income. It presents a set of smoothed **lowess plots** for the ratio of total to HDL cholesterol against age in years. A diverging scale of blue (`ebblue`) and dark orange (`dkorange`) is used to contrast results for men, in blue, and women, in orange. Within these a colour ramp distinguishes separate plots by deciles of household income, with lighter hues corresponding to lower incomes. The graphic reveals higher levels of the cholesterol ratio and a more pronounced hump shape in middle age for men as well as evidence of income gradients for both men and women.

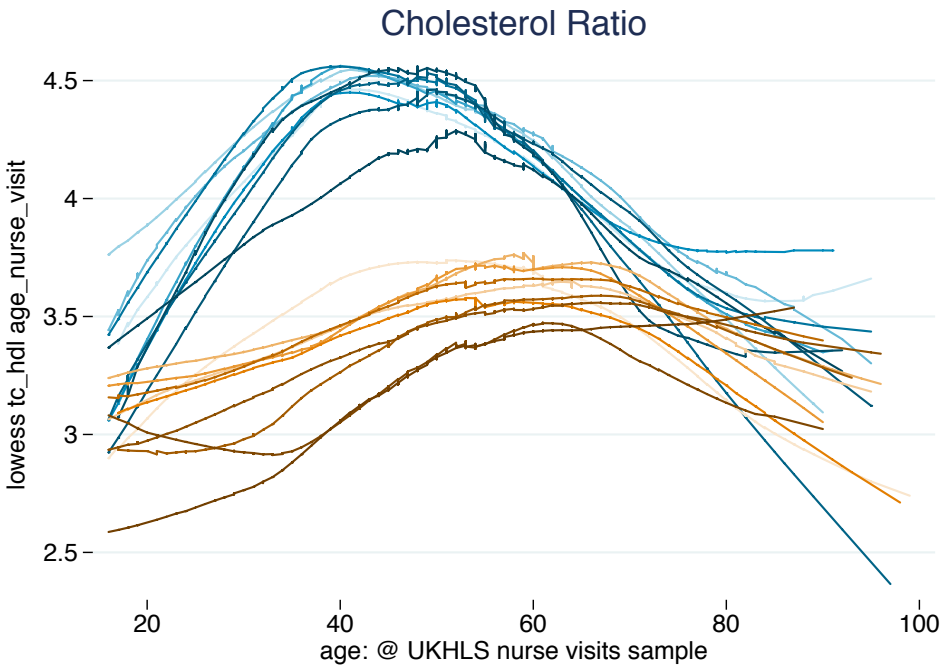


Fig 2.11 Cholesterol ratio by age, gender and income in UKHLS

A colour wheel can be used as a guide to select the particular palette of colours to use in a graphic in order to have a pleasing and harmonious appearance and hence capture and retain the reader's attention. When a two-colour scheme is used these can be complementary and appear at 180^{o} from each other on the colour wheel. Three-colour schemes include triadic harmony (120^{o} from each other), split complementary, analogous colours (close together on the wheel), or warm and cool combinations. A four-colour scheme might be based on the rectangle rule (90^{o} from each other).

It is worth bearing in mind that the best colour scheme to use will depend on the purpose of the graphic. For example, if the graphic is designed to appear on screen, in a presentation or web page, a dark background colour will often work best, while if it is to appear in a printed document a light or white background will work better. To achieve a white background `color background white` should be included in the graphics scheme or within a graphics command (the custom graphics scheme shown earlier was modified in this way to prepare the graphics for print publication).

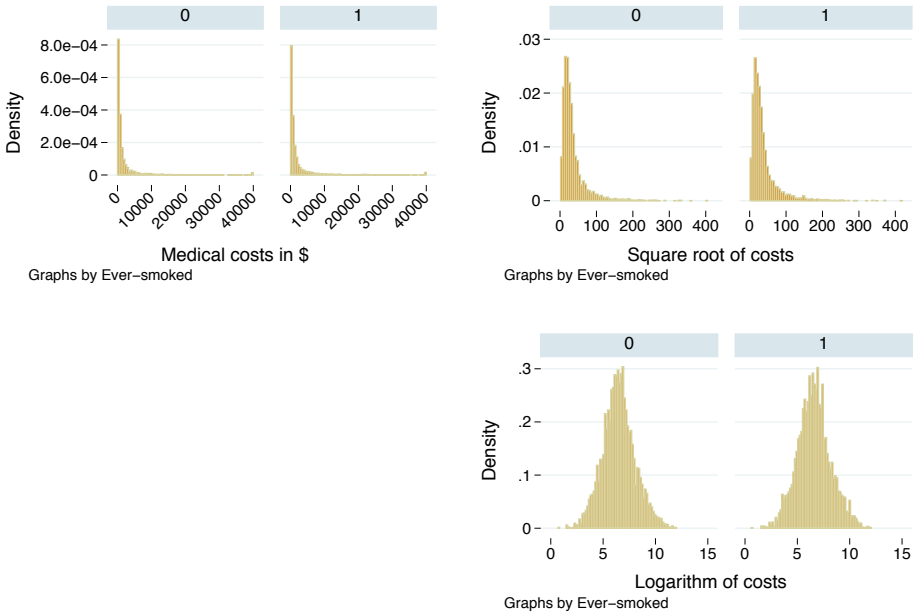
3

Methods

3.1 Data Description and Regression

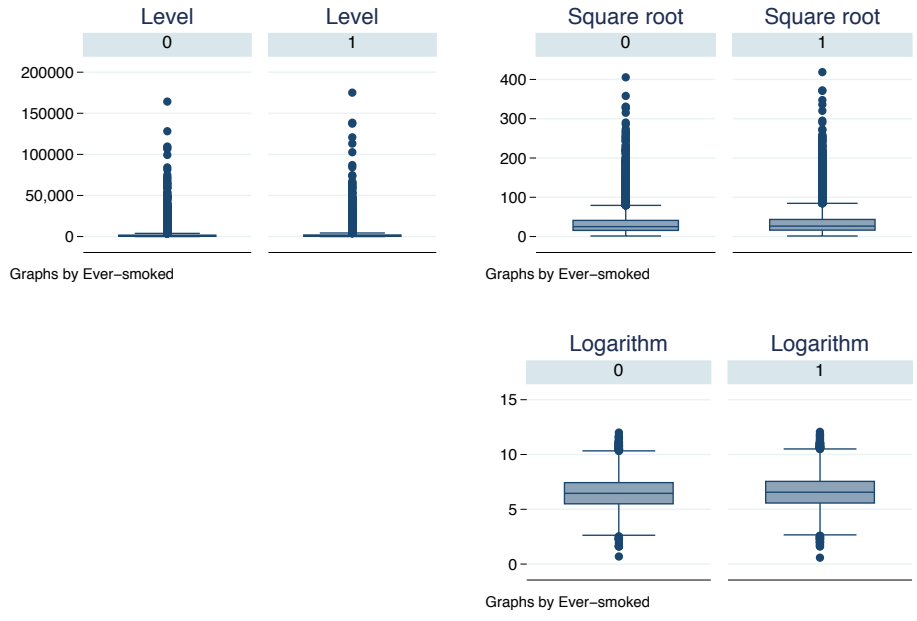
For a visual description of the marginal distribution of non-zero health care costs an obvious place to start is the histogram. Figure 3.1 shows histograms for annual medical costs in the 1987 NMES on the raw scale (truncated at $40,000) and after taking square root and logarithmic transformations. The distributions are shown separately for those who have never smoked (0) and those who have ever smoked (1). As in Jones (2011) these plots exclude the zero observations – which are relatively rare as these data are for totoal annual costs for an elderly population. Discussion of the use of two-part models and other limited dependent variable approaches to model the spike at zero can be found in Jones (2000) and Jones et al. (2013). The heavy degree of skewness and long right-hand tail makes the detail of the plot hard to read for the raw data, while the square root and logarithmic transformations make the distribution progressively more symmetric and closer to a Gaussian distribution.

The **histogram** gives a sense of the shape of the marginal distribution of the outcome and could be complemented by **kernel density plots** to give a smoother image of the outline shape of the distribution. **Box-and-whisker plots** provide more detail of the tails and outliers in the distribution, as shown in Figure 3.2.



source: NMES 1987

Fig 3.1 Histograms for medical costs



source: NMES 1987

Fig 3.2 Box plots for medical costs

The approximate log-normality of the marginal distribution of medical costs in the NMES data is reinforced by a ladder of powers of **normal quantile-quantile (QQ) plots**, where the empirical quantiles are plotted against those expected from a normal distribution with the same mean and variance. These plots are automated in Stata using the `qladder` command as shown in Figure 3.3:

```
qladder y, saving(qladder, replace) title(ladder of
     powers: normal QQ plots) ysca(noline)
     xsca(noline) ylabel( , angle(horizontal))
     note($note)
```

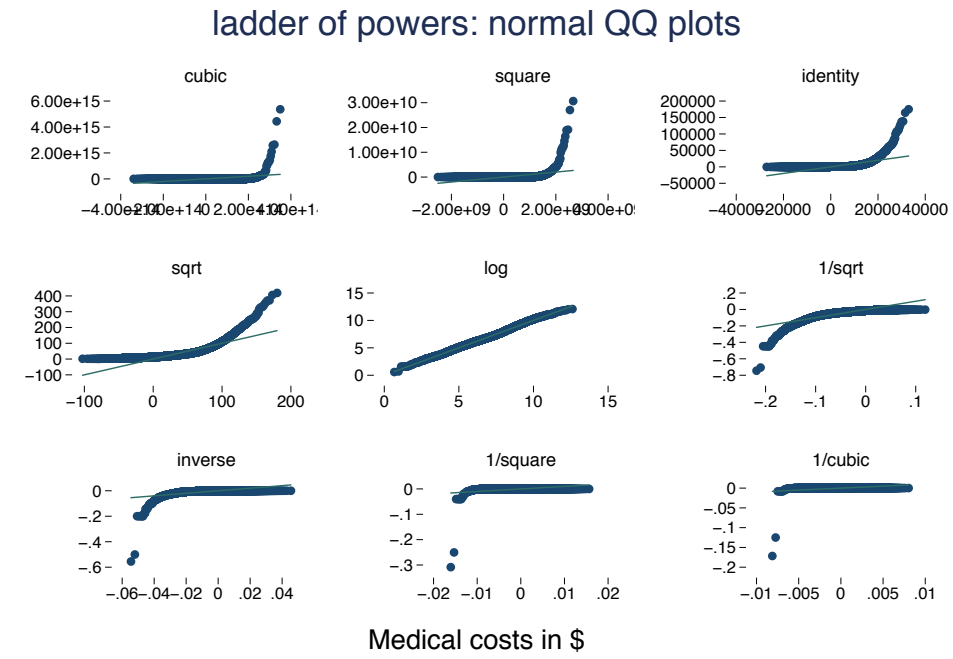


Fig 3.3 Normal QQ plots for medical costs

Kernel plots and histograms provide information about the marginal distribution of the outcome. Moving towards econometric models of health care costs shifts attention to the conditional distribution, given the observed covariates. Various options are available to describe and visualize the conditional distribution. The first borrows the notion of a **population pyramid**, a graphic that is widely used in demography to illustrate the size of the population stratified by sex and by age group. In Figure 3.4 this has been modified to show average levels of medical costs, given by the length of the bars, for

men and women separately and divided into vigiciles of age for those aged 40 and over.

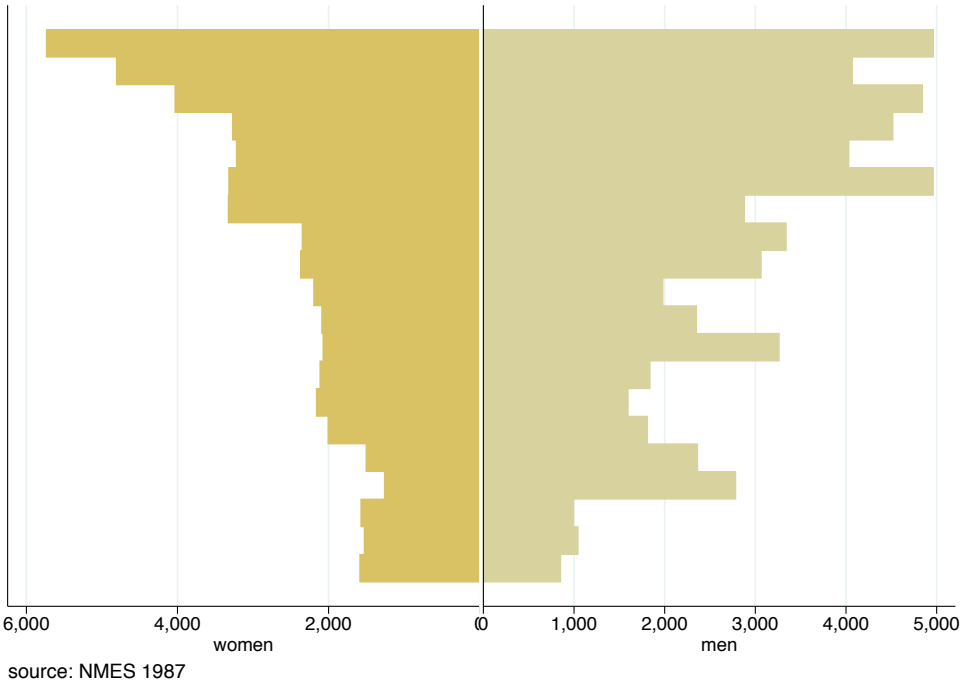


Fig 3.4 Pyramid plot of medical costs by age and gender

This modified pyramid plot is produced and saved using the following code which creates two horizontal bar charts one of which is reversed so that they can be combined together as an (inverted) pyramid:

```
xtile ageq=age, nq(20)
graph hbar y if male==0, yrev over(ageq, reverse
      label(nolabels) gap(0)) bar(1, color(sand))
      ytitle(women) saving(hb1, replace)
graph hbar y if male==1, over(ageq, reverse
      label(nolabels) gap(0)) bar(1, color(stone))
      ytitle(men) saving(hb2, replace)
graph combine hb1.gph hb2.gph, title(Population
      Pyramid) note($note) imargin(0 0 0 0)
graph export pyramid.tif, replace
```

A **strip plot**, as shown in Figure 3.5, provides another way of visualizing the conditional distribution and emphasizes the variation in the individual observations for the logarithm of costs at each level of age.

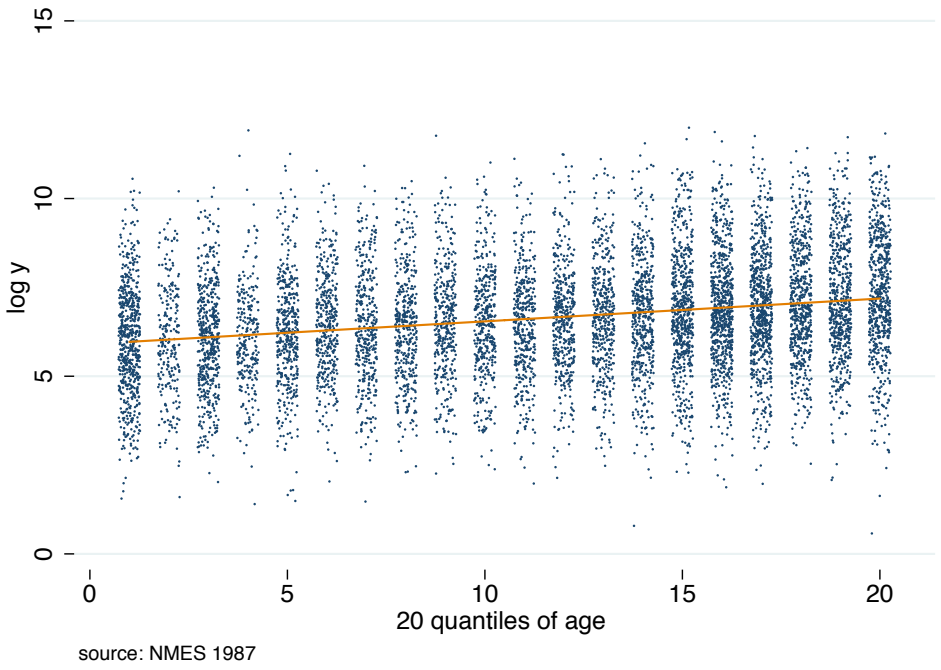


Fig 3.5 Strip plot for medical costs and age

The strip plot is produced and saved using the following code. Jittering is used here to add some artificial variation along the horizontal scale and produce the cloud of data points at each quantile of age. Without this jittering the vertical bars would become solid lines and would not be legible. Note that a seed is set so that the figure can be replicated each time the random jittering is carried out:

```
tw (scatter lny ageq, msymbol(p) jitter(3)
    jitterseed(12345))
    (lfit lny ageq), ytitle(log y) ysca(noline)
     xsca(noline r(0 21)) legend(off) note($note)
     saving(strip, replace)
```

The strip plot provides an impressionistic view of the distribution of y at different levels of x based on the intensity of the cloud of data points. Another perspective on this is given by stacking together histograms for y at each level of x to show the conditional distributions. This can be done for an observed x variable such as by deciles of age, as shown in Figure 3.6. Alternatively, in Figure 3.7, a composite measure of the covariates as a whole is used by first running a linear regression, then computing deciles of the fitted values and plotting histograms and kernel density plots for the logarithm of health care costs for the sub-sample of observations within each decile. Both of these plots show that there are not only changes in the conditional mean and median of the logarithm of costs moving across the deciles of the covariates but also that, although the conditional distributions all seem roughly log-normal, there are differences in the range, dispersion and tail probabilities across the conditional distributions.

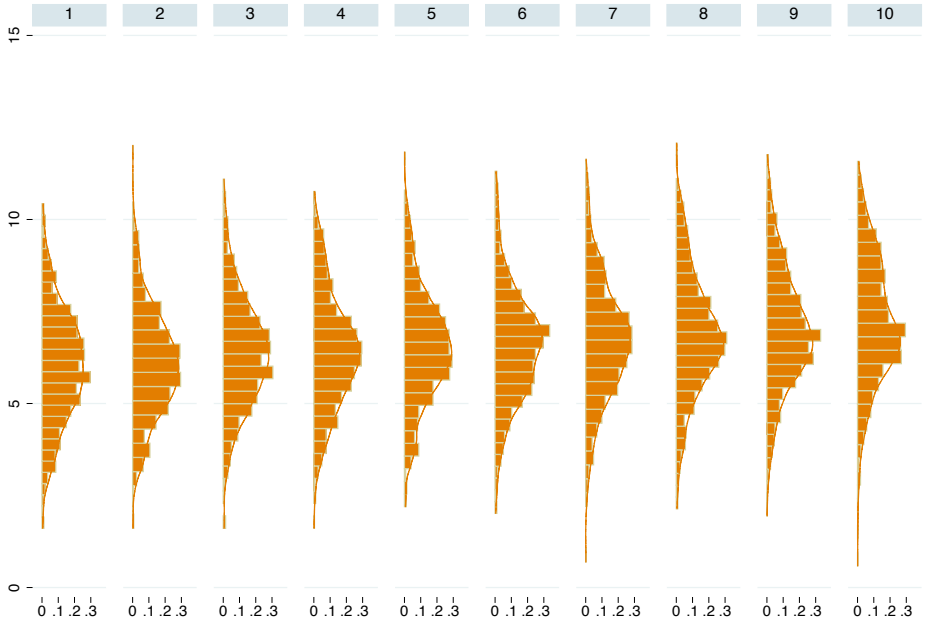


Fig 3.6 Histograms for medical costs and age

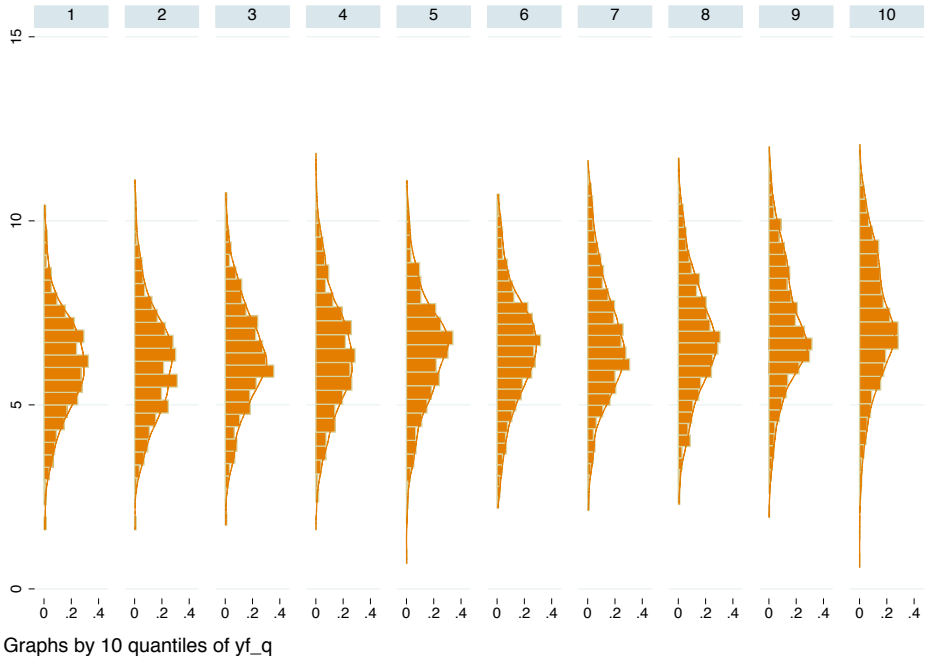


Fig 3.7 Histograms for medical costs and fitted values

The plot based on the fitted values is computed as follows. Note the use of the global `$xs` to save having to provide the full list of regressors each time the regression model is run:

```
* Drop any zero expenditures and run a regression
drop if y==0
quietly regress y $xs
predict yf_q

* Plot of distribution by deciles of the fitted
values
xtile yq10=yf_q, nq(10)
tw   (histogram lny, horiz)
     (kdensity lny, horiz),
     ysca(noline) xsca(noline) by(yq10, row(1))
     saving(hist_deciles_yf, replace)
```

The descriptive plots shown so far illustrate some of the issues that arise for econometric modelling of individual health care costs. The distribution of the raw cost data is typically non-normal with high levels of skewness and excess kurtosis associated with long right-

hand and heavy tails. This reflects the underlying data-generating process (Jones, 2011). Patients with severe health conditions may attract substantial services and relatively rare events and medical procedures might be very expensive. This means that a relative minority of patients are responsible for a high proportion of health care costs, reflected in the fact that mean costs are often much greater than the median. The variability of costs over different risk adjusters means that the data tend to be inherently heteroskedastic. Survey data might feature a mass point at zero, as costs are truncated at zero. Finally, the relationship between costs and covariates may not be linear for example the impact of risk adjusters may be multiplicative rather than additive.

A simple diagnostic of the performance of the linear regression model with the NMES data is provided by plotting means of the actual values of y against the fitted values and comparing these to a 45^{o} line that shows a perfect fit as shown in Figure 3.8. These means are computed within sub-samples defined by the deciles of the fitted values.

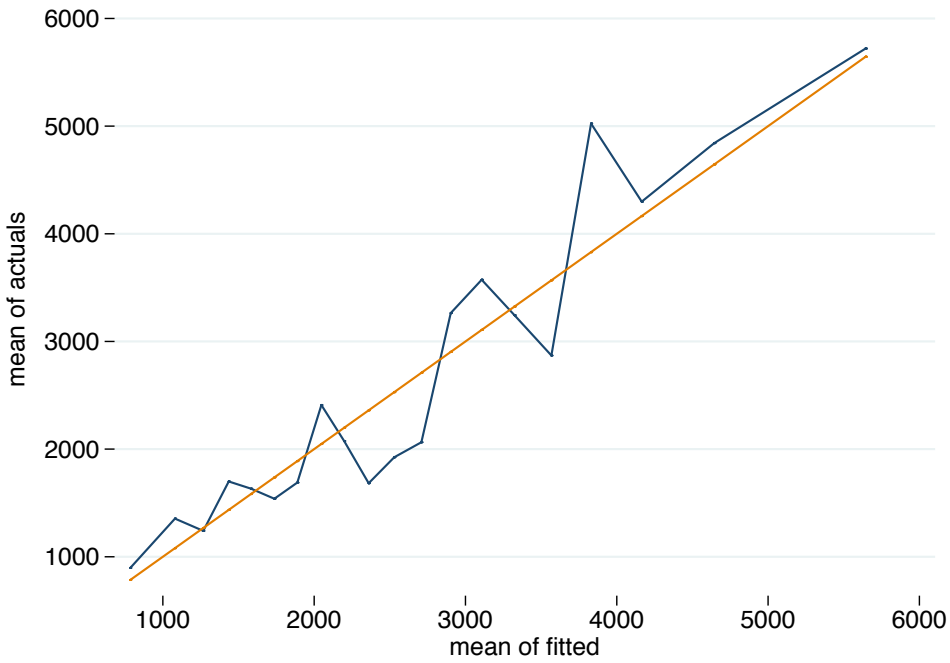


Fig 3.8 Observed versus fitted medical costs

This plot is created by first running the regression model and then saving the fitted values using `predict`:

```
regress y $xs
predict yf if e(sample)
xtile yfd=yf, nq(10)
bysort yfd:egen yfbar=mean(yf)
bysort yfd: egen ybar=mean(y)

tw line ybar yfbar yfbar, title(OLS on y)
     ytitle(mean of actuals) xtitle(mean of fitted)
     ylabel( , angle(horizontal)) xsca(noline)
     ysca(noline) legend(off) saving(dec_y_olsy,
     replace)
```

The non-normality of medical costs often leads researchers to consider transforming the dependent variable to produce a more symmetric distribution prior to running a regression. Most often this involves taking a log transformation or a square-root transformation. A consequence is that the analysis is no longer working on the raw cost scale and retransformation becomes an issue when computing predictions of costs on the original scale (see Jones (2000, 2011) for further discussion).

3.2 Generalized Linear Models

Unlike regressions on transformed costs, Generalized Linear Models (GLMs) specify the conditional mean directly:

$$E(y|x) = \mu = f(x'\beta) \tag{3.1}$$

For example, with a “log link”:

$$E(y|x) = \exp(x'\beta) \tag{3.2}$$

An advantage of this approach is that predictions are made on the original cost scale and that no retransformation is required. By specifying a distribution as well they allow for heteroskedasticity through the choice of distributional family (albeit limited to functions of the mean).

So, the full GLM framework requires a link function and a distribution. In general, the link function, *g(.)*, relates the conditional mean to a linear index of covariates:

$$g(\mu) = x'\beta \Longrightarrow \mu = E(y|x) = g^{-1}(x'\beta) = f(x'\beta) \qquad (3.3)$$

A distribution, that belongs to the natural exponential family (NEF), is used to specify the relationship between the conditional variance and the conditional mean:

$$Var(y|x) = \nu(\mu) \qquad (3.4)$$

The power function form of the variance:

$$Var(y|x) \propto (E(y|x))^{\theta} \qquad (3.5)$$

gives a menu of well-known distributions:

- Gaussian: constant variance; θ=0.
- Poisson: variance proportional to the mean; θ=1.
- Gamma: variance proportional to the square of the mean; θ=2.
- Inverse Gaussian: variance proportional to cube of the mean; θ=3.

These distributions and link functions can be used in any combination, although there are canonical links such as the log link used with Poisson variance and identity link with Gaussian variance.

As described by Holly (2009), in GLMs "all the moments of the distribution are functions of the mean, and those functions depend of the particular specification of the members of the NEF". In particular the skewness and kurtosis are "completely and uniquely determined once the member of the exponential family has been specified". The implication of this is that the higher moments that are likely to be of interest in our modelling work can be tied back to the conditional mean. As shown above, for most distributions included in the NEF the variance function is a polynomial in μ. Also following Holly (2009), within this family of distributions, skewness (S) and kurtosis (K) take the form:

$$S = \alpha CV \qquad (3.6)$$

$$K = \beta + \gamma CV \qquad (3.7)$$

where CV is the coefficient of variation. The implication of these results is that prior to adopting the GLM approach and to choosing a particular GLM specification it may be helpful to describe the relationship between the conditional mean and conditional variance and to assess whether there is a linear relationship between conditional skewness and the coefficient of variation and a quadratic relationship between conditional kurtosis and the coefficient of variation.

A simple way to produce plots that describe the relationships between the conditional moments of the sample data is shown here. This takes the predicted values from a simple linear regression of y on x to condition on the covariates. The sample is then divided into equal sized groups according to these predicted values, here 20 groups, or vigiciles, are used. Within each of these groups the mean, variance, skewness, kurtosis and coefficient of variation are computed and saved:

```
xtile yq=yf_q, nq(20)
gen yqmean=0
gen yqvar=0
gen yqskew=0
gen yqkurt=0
gen yqCV=0
forvalues i=1/$qy {
      quietly summ y if yq==`i', detail
      replace yqmean=r(mean) if yq==`i'
      replace yqvar=r(Var) if yq==`i'
      replace yqskew=r(skewness) if yq==`i'
      replace yqkurt=r(kurtosis) if yq==`i'
      replace yqCV=r(sd)/r(mean) if yq==`i'
      }
```

Armed with these statistics various plots can be produced. The first, Figure 3.9, is intended to show the relationship between conditional mean and conditional variance and hence, give an indication which distribution is relevant within the GLM framework. For the NMES data it is clear that there is a positive relationship between mean and variance suggesting that a Gaussian constant variance would not be appropriate but that linear (Poisson) and quadratic (gamma or negative binomial) fits might work well.

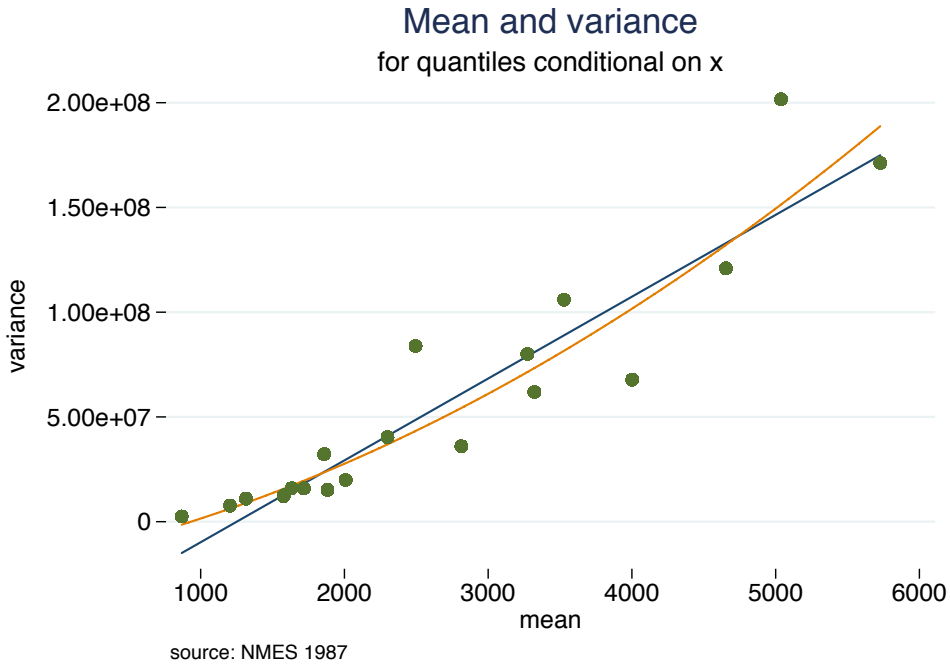


Fig 3.9 Conditional variance versus mean

The code used to produce Figure 3.9 combines a scatter plot with both linear and quadratic fits:

```
tw    (lfit yqvar yqmean)
      (qfit yqvar yqmean)
      (scatter yqvar yqmean, msize(medium)),
       title("Mean and variance") subtitle(for quantiles
       conditional on x) ytitle(variance) xtitle(mean)
       ylabel( , angle(horizontal)) ysca(noline)
       xsca(noline) note($note) legend(off) saving(mean&var,
       replace)
```

Then, to assess whether a linear relationship does appear to hold a scatter plot of skewness against the coefficient of variation from each of the sub-samples can be drawn (Figure 3.10). Similarly a scatter plot of kurtosis against the coefficient of variation can be used to check for a quadratic relationship (Figure 3.11). Note that both of these plots are heavily influenced by the sub-sample that contains those with the highest predicted medical costs that appears as an outlier in the graphs.

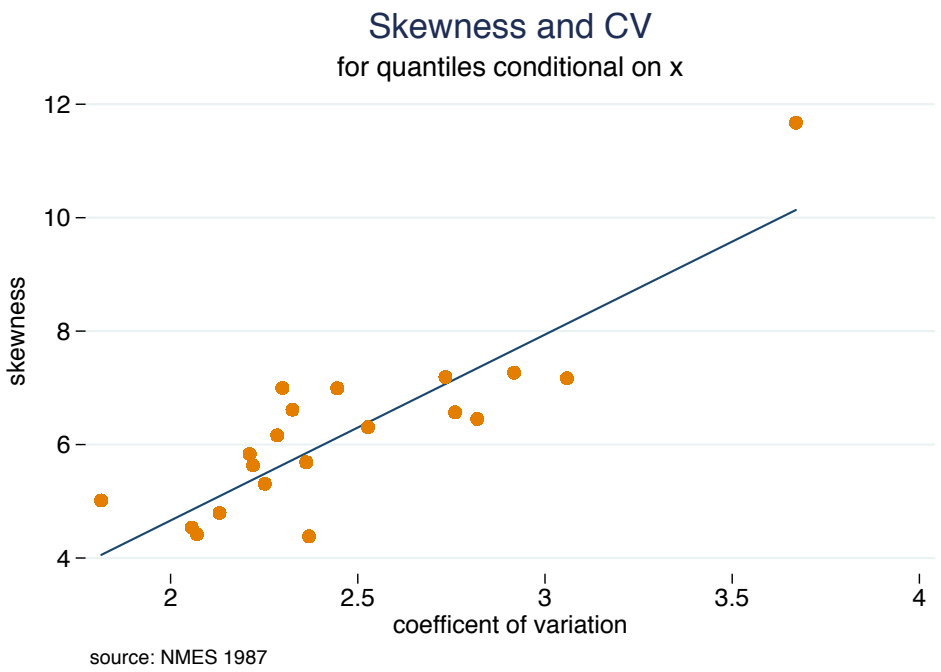


Fig 3.10 Conditional skewness versus coefficient of variation

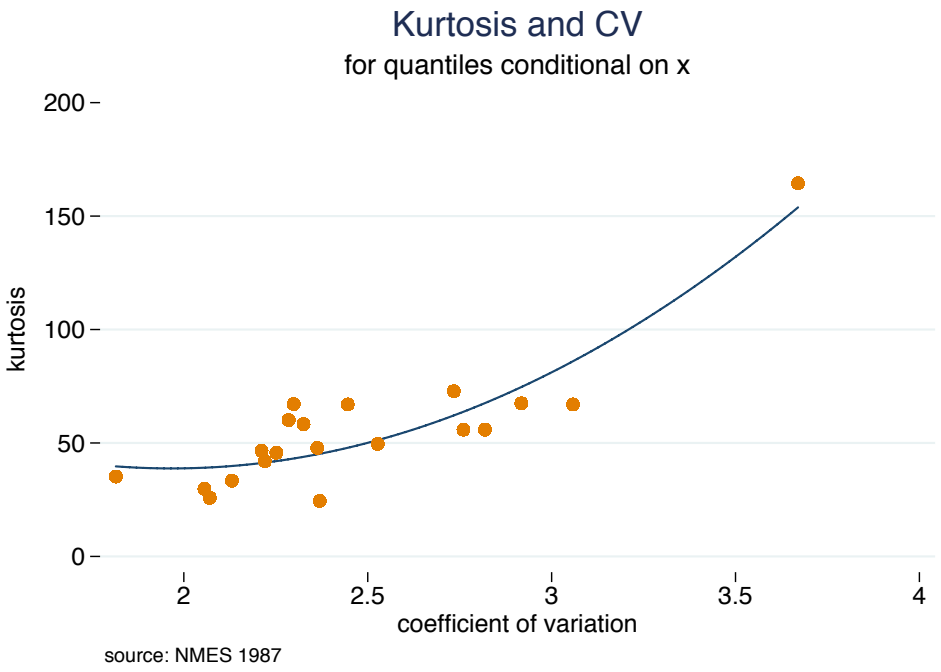


Fig 3.11 Conditional kurtosis versus coefficient of variation

As emphasised by Holly (2009), using a GLM and hence a distribution drawn from the natural exponential family limits the combinations of skewness and kurtosis values that are allowed. This is most obvious if a Gaussian distribution is fitted to the data as skewness is then constrained to equal 0 and kurtosis to equal 3. McDonald et al. (2011) develop a graphical approach to show how distributional choices limit the values of higher moments that can be estimated. For example, with a gamma distribution the values of skewness and kurtosis are constrained to lie on a particular locus of points. This is illustrated by the curve in Figure 3.12 which can be compared to the sample values of skewness and kurtosis. Note that this curve is a theoretical property of the gamma distribution and is not a fitted curve of the type shown in the two previous Figures. In this case the NMES data appear to be compatible with the curve implied by a gamma distribution. The implications of this issue are explored further in the next section that looks at flexible parametric models.

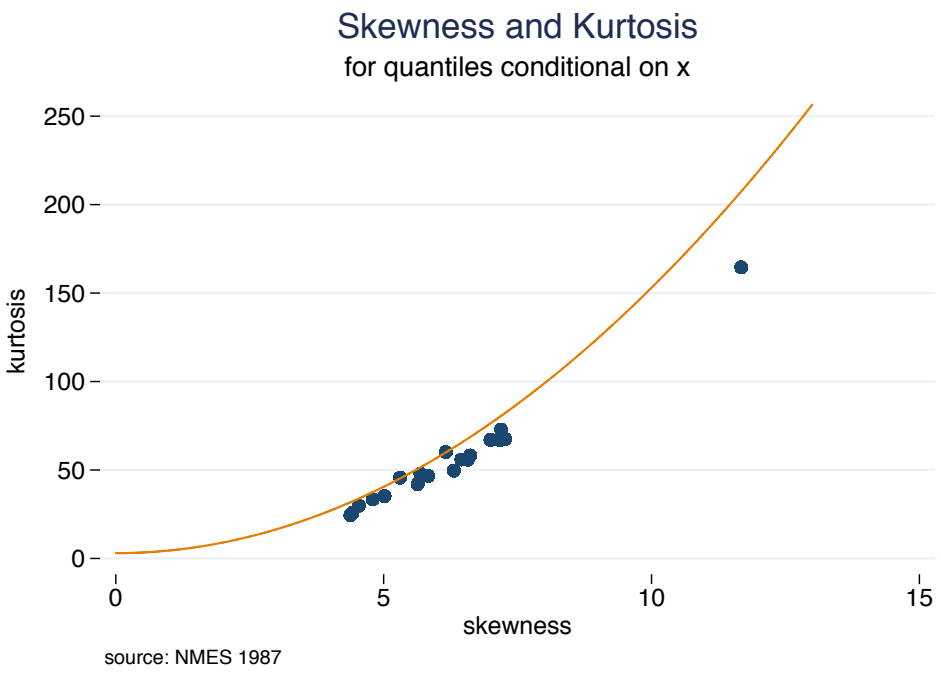


Fig 3.12 Limits on skewness and kurtosis

The code to produce Figure 3.12 is:

```
tw (scatter yqkurt yqskew, msize(medium))
```

```
(function gamma_locus=3+(3/2)*x^2, range(0 13)),
title(Skewness and Kurtosis)
subtitle(for quantiles conditional on x)
ytitle(kurtosis) xtitle(skewness) saving(skew&kurt,
replace) ylabel( , angle(horizontal)) ysca(noline)
xsca(noline) legend(off) note($note)
```

The descriptive graphs shown above provide guidance on whether a GLM is appropriate and, if so, which link and distribution function to use. This can be handled more formally. In particular, Basu and Rathouz (2005) suggest a flexible semiparametric approach to the problem of selecting the appropriate link and variance functions for the GLM model. Their extended estimating equations (EEE) approach, uses a Box-Cox transformation for the link function:

$$x'b = \frac{\mu^{\lambda}-1}{\lambda} \quad where \ \mu = E(y|x) \tag{3.8}$$

This includes the log, square root and identity links as special cases along with other power functions of *y*.

This is combined with a general power function for the variance:

$$Var(y|x) = \theta_1 \mu^{\theta_2} \tag{3.9}$$

which gives a flexible specification that nests the common GLM distributions and allows these nested models to be tested. The additional parameters are estimated, along with the regression coefficients, by quasi-maximum likelihood estimation.

3.3 Flexible Parametric Models

The widely used log-link version of the GLM is connected to a broader range of what can be called Exponential Conditional Mean (ECM) models (Jones, 2011). The ECM directly assumes a nonlinear relationship:

$$E(y|x) = \mu = \exp(x'\beta) \tag{3.10}$$

Or, more generally:

$$E(y|x) \propto \exp(x'\beta) \ or \ E(y|x) = \phi \exp(x'\beta) \tag{3.11}$$

Note that this implies that the effect of the covariates is "proportional" rather than additive, as in a proportional hazard model.

The ECM, and extensions, can be estimated in a variety of ways including nonlinear least squares (NLS); the Poisson quasi-ML estimator (QML); and by using ML estimation for parametric hazard models such as the exponential, Weibull, or generalized gamma. In particular, Manning et al. (2005) propose that the generalized gamma, which is typically used as a flexible "3-parameter" distribution for survival models, is well suited for use with medical cost data. The generalized gamma has a density function and conditional expectation that take the form (using notation taken from the Stata manual):

$$f(y;\kappa,\mu,\sigma) = \frac{\gamma^{\gamma}}{\sigma t\sqrt{\gamma}\,\Gamma(\gamma)}\exp(z\sqrt{\gamma} - u)$$

$$where \quad \gamma = |\kappa|^{-2}, z = sign(\kappa)\{\ln(y) - \mu\}, u = \gamma\exp(|\kappa|z), \mu = x'\beta$$

$$E(y \mid x) = \exp(x'\beta)\left[\kappa^{2\sigma/\kappa}\frac{\Gamma\left(\frac{1}{\kappa^2} + \frac{\sigma}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa^2}\right)}\right] = \exp(x'\beta)\phi \tag{3.12}$$

Special or limiting cases of the generalized gamma are the Weibull (κ=1), exponential (κ=1,σ=1), and lognormal (κ=0) distributions.

Jones et al. (2014) suggest adding further flexibility by modelling medical cost data based on the generalized beta of the second kind (GB2) distribution and its nested and limiting cases. The GB2 provides a useful tool for choosing between competing distributions. The GB2 is a 4-parameter distribution that is often used to model the size distribution of earnings and in analyses of income inequality (e.g., Jenkins, 2009). The mean of the GB2 distribution is:

$$E(y) = b\left[\frac{\Gamma\left(p+\frac{1}{a}\right)\Gamma\left(q-\frac{1}{a}\right)}{\Gamma(p)\Gamma(q)}\right] \tag{3.13}$$

Using $b = \exp(x'\beta)$ and treating the other parameters as scalars puts this in the ECM class of models. The Burr-Singh-Maddala

distribution is a special case when p=1, the Dagum is a special case when q=1 and p=q=1 gives the log-logistic. Also, the generalized gamma, and hence the gamma and Weibull, are limiting cases of the GB2. Figure 3.13 shows the nesting of distributions within the GB2 family.

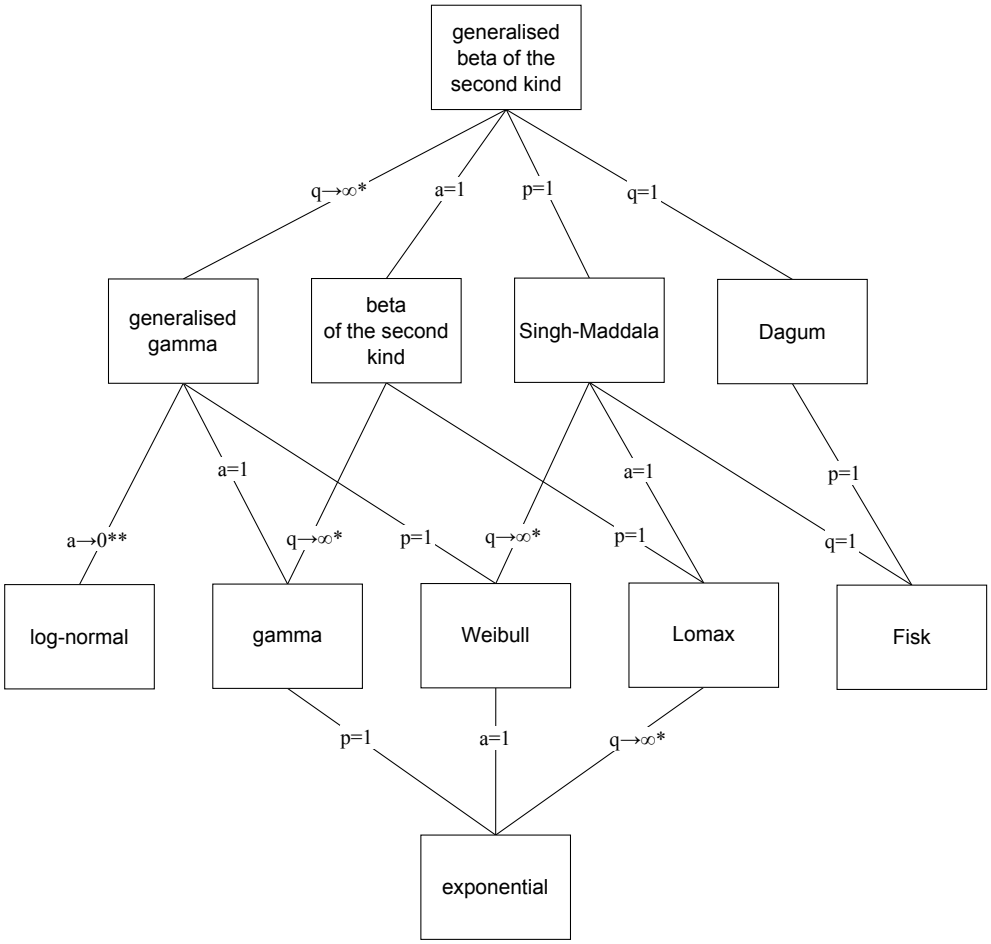


Fig 3.13 The GB2 and nested distributions
(source: Jones et al., 2014)

Jones et al. (2014) apply the GB2 distribution to English administrative data from the Hospital Episode Statistics (HES) dataset. They use a quasi-experimental design that estimates the models on a subset of the data and holds out another subset to evaluate their predictive performance. This design is used to compare the GB2 distribution with its nested and limiting cases. They

find that the GB2 can be a useful starting point for selecting which distribution to apply, with the beta-2 distribution and generalized gamma distribution performing the best within their dataset.
Figure 3.14 extends the analysis introduced above in Figure 3.12. It shows how the values of skewness and kurtosis measured in the HES data compare to the values that are compatible with the upper and lower limits implied by the GB2 distribution ($GB2_U$ and $GB2_L$) and the distributions nested within it. The dots relate to deciles of the fitted values and the 'target' symbol shows the overall sample values. Note that these all lie within the range implied by the generalized gamma distribution (between GG_U and GG_L).

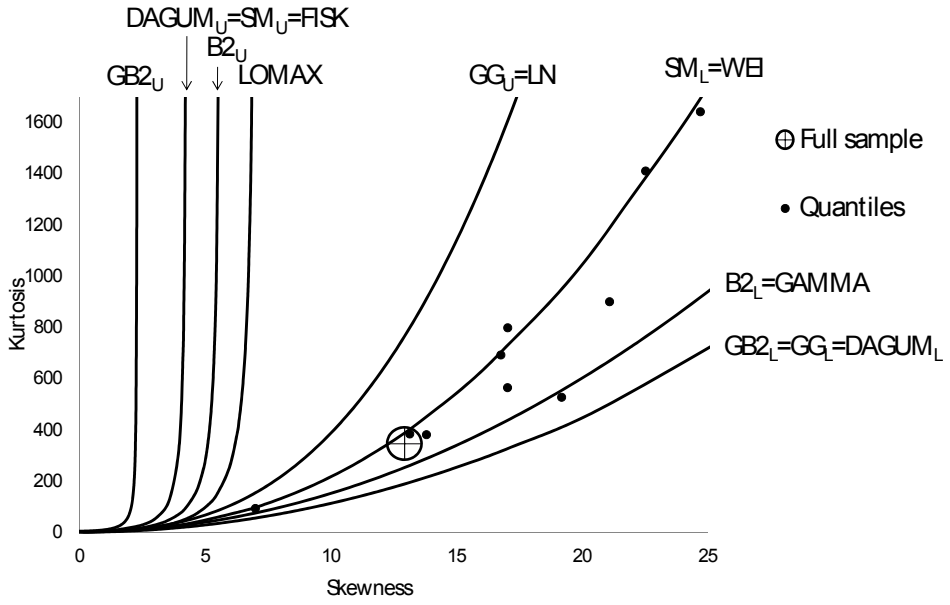


Fig 3.14 Skewness and kurtosis in HES data
(source: Jones et al., 2014)

3.4 Semiparametric Models

Deb and Trivedi (1997) proposed the use of finite mixture models (FMM) as an alternative to the hurdle model in the empirical modelling of count data measures of health care utilisation such as the number of visits to a doctor over a fixed period such as the previous month. In a FMM the population is assumed to be divided

in C distinct components in proportions $\pi_1, \ldots, \pi_C \; where \; \sum_{j=1}^{C} \pi_j = 1$, $0 \leq \pi_j \leq 1$. The C-point finite mixture model is given by:

$$f(y|.) = \sum_{j=1}^{C} \pi_j f_j(y|.) \qquad (3.14)$$

where $f_j(.)$ is the specified parametric distribution for each component and the mixing probabilities π_j are estimated along with all the other parameters of the model. Finite mixture models have been applied to cost data by, for example, Deb and Holmes (2000) and Conway and Deb (2005).

The use of a discrete conditional density estimator (CDE) for the moments of the distribution of medical costs was proposed by Gilleskie & Mroz (2004). The method divides the support of y into fixed number (K) of discrete intervals and then uses the approximation:

$$E(y|x) = \int yf(y|x)dy \approx \sum_k \bar{y}\, p(y_{k-1} \leq Y < y_k) \qquad (3.15)$$

Gilleskie and Mroz (2004) suggest a discrete hazard specification to estimate the probabilities $p(y_{k-1} \leq Y < y_k)$.

Jones et al. (2016) provide a comprehensive assessment of the relative performance of many of the approaches that are used in the literature again using a quasi-Monte Carlo design. The range of methods compared includes GLMs, flexible parametric models based on the GB2 family as well as the semiparametric finite density and finite mixture models. The models are compared in terms of their ability to predict the conditional mean of costs in the forecast sample. For their HES dataset the best performing model in terms of bias is linear regression with a square root transformed dependent variable, and a GLM with square root link function and Poisson distribution performs best in terms of goodness-of-fit.

3.5 Distributional Methods

Jones et al. (2015) compare methods that are designed to estimate the full conditional distribution of healthcare costs. This is motivated by the idea that it can be important to go 'beyond the mean'. Much of the work done on modelling health care costs has focused on prediction of the conditional mean of costs. But this can neglect other features of the full conditional distribution such as quantiles and tail

probabilities. For example, the varying size and shape of the tail probability associated with annual costs greater than £10,000 in the HES data is shown in Figure 3.15. Here the conditional distributions of log(costs) are split into quintiles ranging from those patients whose risk adjusters are associated with the lowest level of costs to those in the top quintile who are associated with the highest levels.

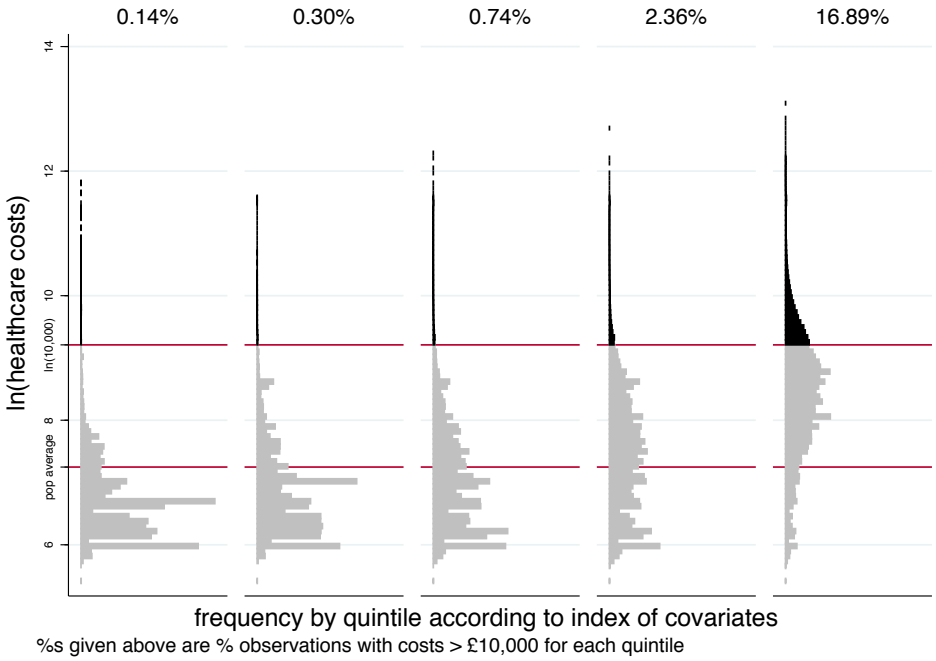


Fig 3.15 Tail probabilities in HES data
(source: Jones et al., 2015)

Jones et al. (2015) again use the English NHS inpatient data in a quasi-Monte Carlo design that compares fourteen different methods for modelling the distribution of healthcare costs. Nine of the methods use flexible parametric models, based on distributions that have been used in the literature to fit healthcare costs. These include the generalized gamma and the generalized beta of the second kind (GB2) models as well as the use of finite mixture models. Linear regressions and GLMs are not included in the comparison as they do not offer a ready way to estimate tail probabilities.

The remaining five methods are drawn from recent developments on the literature on regression-based decomposition analysis and methods that are designed to estimate counterfactual distributions. These methods involve distributional regressions and are estimated

using two broad approaches. The first set aim to estimate the distribution function directly and are essentially based on logit models for step functions (Han and Hausman, 1990; Foresi and Peracchi, 1995; Chernozhukov et al., 2013). The second set estimate quantile functions, either directly or through the use recentred influence function (RIF) regression, which are then inverted to give the distribution function and corresponding tail probabilities (Machado and Mata, 2005; Melly, 2005; Firpo et al., 2009).

Once again the design of the study split the data between an estimation set, used to draw random samples of varying sizes to fit the models, and a validation set, used to check the predictions of tail probabilities. The quasi-Monte Carlo comparisons focus on the estimation of specific tail probabilities rather than the distribution as a whole. They show that no single method is dominant and indicate that there is a trade-off between bias and precision in the forecasts. It is clear that the distributional regression methods demonstrate significant potential for estimating conditional tail probabilities, particularly with larger sample sizes when the variability of predictions is reduced. The parametric models including the log-normal, generalized gamma and GB2 estimate tail probabilities with high precision, but with varying bias depending upon the threshold for the tail probability that is used.

4

An Application to Biomarkers

This section puts the econometric and graphical methods described above into practice with a new application to health outcomes. This uses objective blood-based biomarkers, rather than medical costs, as the dependent variables. Biomarkers are biological or physiological measures used to indicate the presence of a disease or the likelihood of developing a disease. They are used to identify risk factors and as objective measures of health that avoid contamination by reporting bias (Benzeval et al. 2016). The distributions of these biomarkers vary and the descriptive and analytical methods, including graphical approaches, outlined above can be used to select appropriate models for each of them.

The dataset is the new UK national panel; the UK Household Longitudinal Study (UKHLS), known as Understanding Society, is a large nationally representative longitudinal study. The study began at wave 1 with a new sample of the members of about 32,000 households, known as the general population sample or GPS (Knies, 2015). Then at wave 2 (2010-2011), the existing sample of around 8,000 households from the previous national panel, the British Household Panel Study (BHPS) was incorporated into the UKHLS. The BHPS has been used heavily by economists and has 18 waves collected annually between 1991 and 2009.

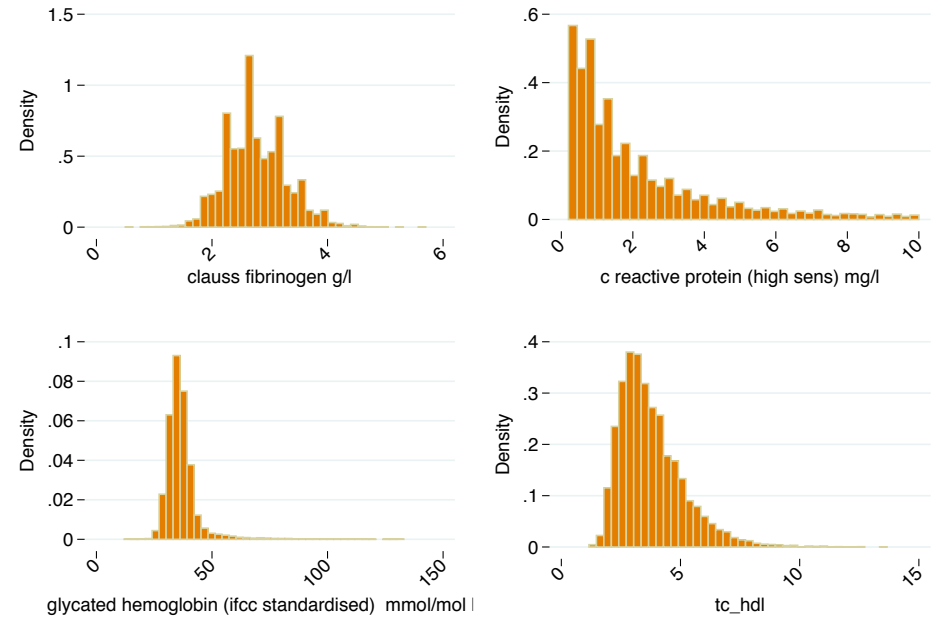
The UKHLS, and the BHPS before it, involves annual interviews for each member of the household aged 16 and over. In the UKHLS the fieldwork for each wave takes two calendar years. The questionnaire covers a broad set of questions about family composition, employment, education, housing, neighbourhoods, consumer durables, savings, wealth and income, health, health behaviours, well-being, cognition and personality, social support and engagement, transport, leisure, environmental and political behaviours.

As part of the UKHLS data collection a set of objective health measures (biomeasures), such as height, weight and blood pressure, as well as a non-fasted blood sample (for biomarkers) were collected by trained nurses. This was done at wave 2 for the general population sample and as part of wave 3 for the members of the BHPS sample (Benzeval et al., 2014; McFall et al., 2014). These nurse visits have not been repeated as yet so the health and biomarkers data are cross-sectional and the analysis presented here pools the available observations from waves 2 and 3. To be eligible for the nurse visit and collection of the blood samples respondents had to take part in the main survey, be aged 16 and over, be resident in Great Britain (thereby excluding Northern Ireland), and not be pregnant (McFall et al., 2014). In addition blood samples were restricted to those without clotting or bleeding disorders and who had not had a fit (Benzeval et al., 2014).

Following Davillas et al. (2017), four biomarkers are compared here. These are selected because of their differing distributions and because they are linked to major chronic health problems including coronary heart disease and diabetes. Two biomarkers of inflammation are used: c-reactive protein (CRP) and fibrinogen. CRP (in mg/ L) indicates general chronic or systemic inflammation. It has been shown that the risk of ischaemic vascular disease, metabolic syndrome and mortality are gradually increasing in CRP. Fibrinogen (in g/L) is a glycoprotein that stops bleeding by helping blood clots to form. As such, fibrinogen is directly related to coronary artery thrombosis; however, it is also regarded as an inflammatory biomarker. Glycated haemoglobin (HbA1c) measures ‘sugar in the blood’ and is used as an indicator for diabetes. Cholesterol measures “fat in the blood”. The cholesterol ratio is calculated as the ratio of total cholesterol over the high-density lipoprotein (HDL) cholesterol concentration in the blood. This is a predictor of cardiovascular morbidity and mortality risks.

Following Davillas et al. (2017), the biomarkers are modelled as a function of household income and other socioeconomic variables. The UKHLS includes current household income as a derived variable. In the regression models the logarithm of income is used in order to allow for the concavity of the biomarker-income relationships. The other covariates include fourteen age dummies for each gender to allow for a flexible association between health, age and sex. Ethnicity dummies are also included along with marital status, education attainment and dummies to capture regional variations. For the application presented a complete case analysis is used using all non-missing observations for the four biomarkers and associated regressors and excluding those with CRP values above 10 mg/L who are most likely experiencing an acute infection at the time of the blood sample. This sample includes 10,683 individuals.

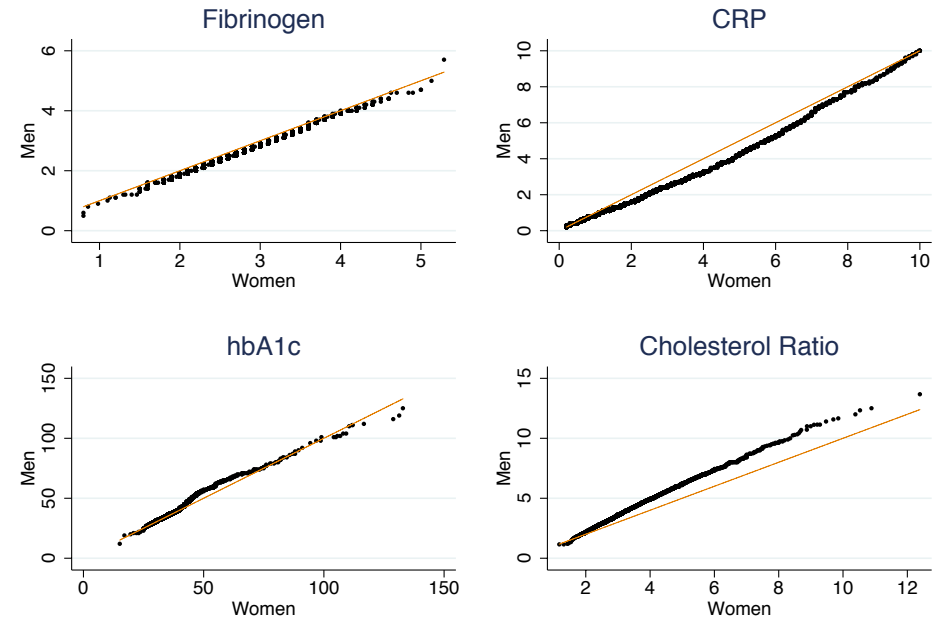
Now consider the shape of the marginal distributions of each of the four biomarkers, as represented by histograms in Figure 4.1. Each exhibits distinctive and 'non-normal' features. Fibrinogen appears symmetric but with heaping at particular values. CRP is highly skewed, even though the plot has been truncated at a value of 10 mg/L. HbA1c and the cholesterol ratio are less skewed than CRP but both exhibit long right-hand tails.



Source: UKHLS

Fig 4.1 Histograms of the biomarkers

Figure 4.1 pools the data across men and women. To assess whether the shape of the distributions differ conditional on gender Figure 4.2 shows **quantile-quantile (QQ) plots**. These indicate little difference between men and women in the distributions of fibrinogen CRP and HbA1c but show a difference, indicated by deviations from the 45^{o} line, especially in the right-hand tail, for the cholesterol ratio. This would need to be controlled for either by splitting the sample or by the way gender is included in the regression models.



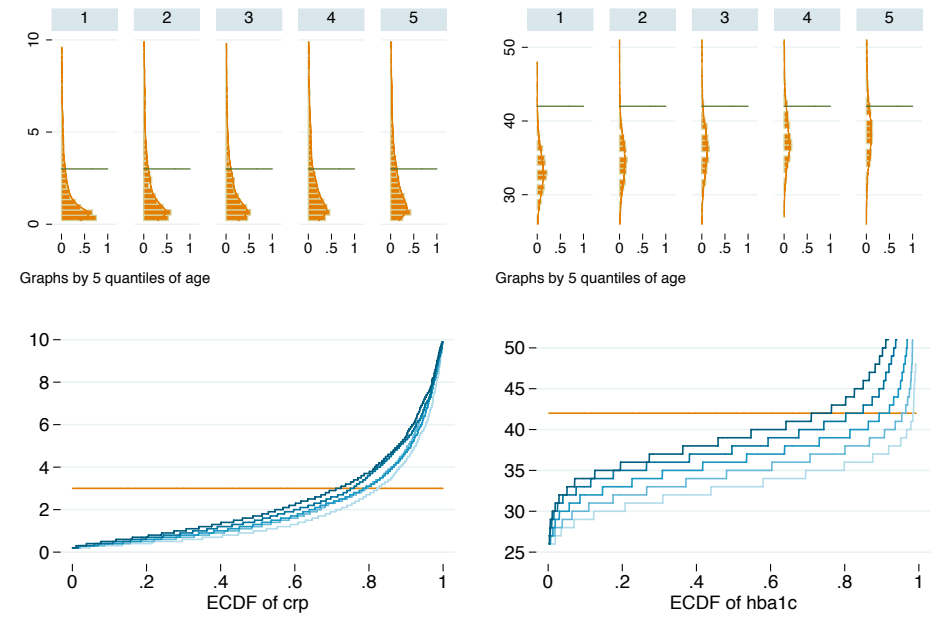
Source: UKHLS

Fig 4.2 QQ plots of the biomarkers by gender

The code used for the QQ plots is illustrated here for the case of fibrinogen:

```
gen cfib_m=cfib if male==1
gen cfib_f=cfib if male==0
qqplot cfib_m cfib_f, xtitle(Women) ytitle(Men)
    title(Fibrinogen) graphregion(color(white))
    msize(vsmall) mc(black) rlopt(lw(thin)
    lc(dkorange)) nodraw saving(qqcfib, replace)
```

The way in which the distributions of the biomarkers vary by age is illustrated for CRP and HbA1c, both of which have commonly applied clinical thresholds (see e.g., Davillas et al., 2017). Values of CRP over 3 mg/L are used to indicate elevated risk for cardiovascular diseases. Values over 10mg/L are typically seen as a sign of a current acute infection and these cases are excluded here. Levels of HbA1c between 42 and 48 mmol/mol are used to indicate pre-diabetes risk, with values above 48 indicating a diagnosis of diabetes. Figure 4.3 shows how the likelihood of crossing these thresholds varies with age. It plots both density functions and **Pareto charts** of the inverted empirical distribution functions for both of the biomarkers split by quintiles of the individual's age. The clinical thresholds are included as horizontal bars to show how the tail probabilities increase with age. This is most visible in the distribution plots, especially for HbA1c, showing the growing burden of pre-diabetes risk and diabetes with ageing.



Source: UKHLS

Fig 4.3 Distributions of CRP and HbA1c by age

The code for the empirical distribution functions illustrates the use of a colour ramp which takes one colour, `ebblue` from the *Economist*

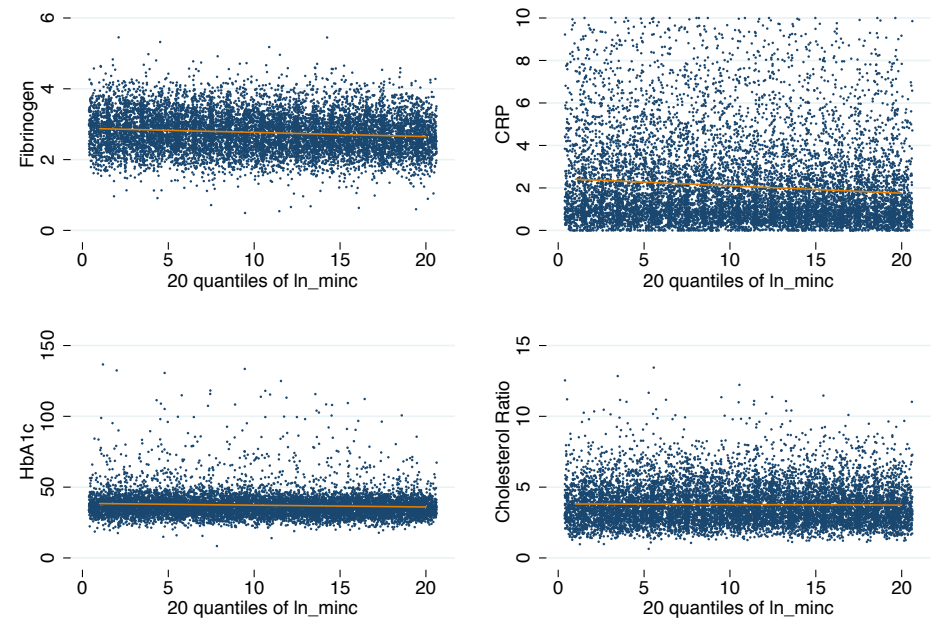
palette, and changes its intensity for each of the curves using the `lcolor` subcommand:

```
* Create indicator of quintiles of age
xtile yqage=age, nq(5)
* Create clinical thresholds
gen cutcrp=3
gen cuthba1c=42

* Pareto charts (edfs) using cumulated variables
bysort yqage: cumul crp, gen(cmcrp)
bysort yqage: cumul hba1c, gen(cmhba1c)

tw (line cutcrp cmcrp, sort lwidth(medium)
   lcolor(dkorange), ysca(noline) xsca(noline)
   ylabel( , angle(horizontal)))
   (line crp cmcrp if yqage==1, sort lwidth(medium)
   lcolor(ebblue*.3), ysca(noline) xsca(noline)
   ylabel( , angle(horizontal)))
   (line crp cmcrp if yqage==2, sort lwidth(medium)
   lcolor(ebblue*.6), ysca(noline) xsca(noline)
   ylabel( , angle(horizontal)))
   (line crp cmcrp if yqage==3, sort lwidth(medium)
   lcolor(ebblue*.9), ysca(noline) xsca(noline)
   ylabel( , angle(horizontal)))
   (line crp cmcrp if yqage==4, sort lwidth(medium)
   lcolor(ebblue*1.2), ysca(noline) xsca(noline)
   ylabel( , angle(horizontal)))
   (line crp cmcrp if yqage==5, sort lwidth(medium)
    lcolor(ebblue*1.5), ysca(noline) xsca(noline)
    ylabel( , angle(horizontal))) if crp<10,
    legend(off) saving(cmcrp, replace)
```

Similar code for HbA1c and for plotting the histograms has been omitted.

Figure 4.4 shows the bivariate relationships between the biomarkers and the logarithm of household income using strip plots. Note that the biomarkers are increasing in poor health so the bivariate regression lines for the conditional mean of the biomarkers as a function of income, while shallow, are downward sloping and, on average, the biomarker scores improve (get smaller) with higher income. However the strip plots show considerable heterogeneity around the regression lines and they show the extent of the right-

hand tails of the distributions, especially for CRP and the cholesterol ratio. HbA1c values are the most concentrated around the regression line.

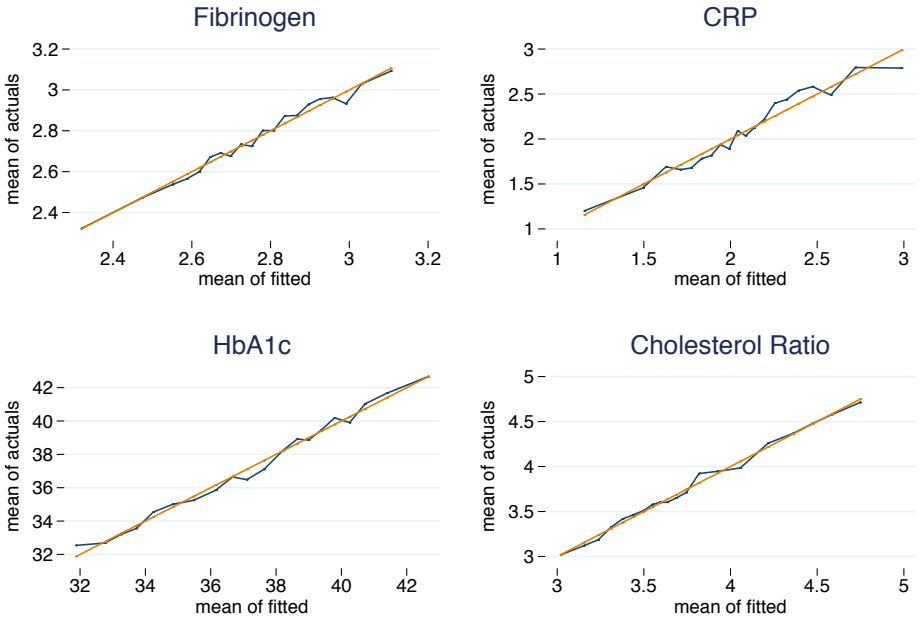


Source: UKHLS

Fig 4.4 Strip plots of the biomarkers by ln(income)

Linear regressions are fitted for each of the biomarkers, on their raw scale, using the list of regressors described earlier. The fit of these models is summarised in Figure 4.5, which plots the average of actual and fitted values for twenty intervals of the fitted values. As expected, based on Figure 4.4, HbA1c and the cholesterol ratio show the tightest fit with CRP showing deviations in the right-hand tail.

The scatter of observed data around the fitted regression lines is displayed in Figure 4.6. This is drawn using a variant of the observed versus fitted values plot with the command `ovfplot`. This is taken from Nicholas Cox's 'Modeldiag' package that extends the range of Stata graphical commands for regression diagnostics.



Source: UKHLS

Fig 4.5 Actual versus fitted plots for linear regression models

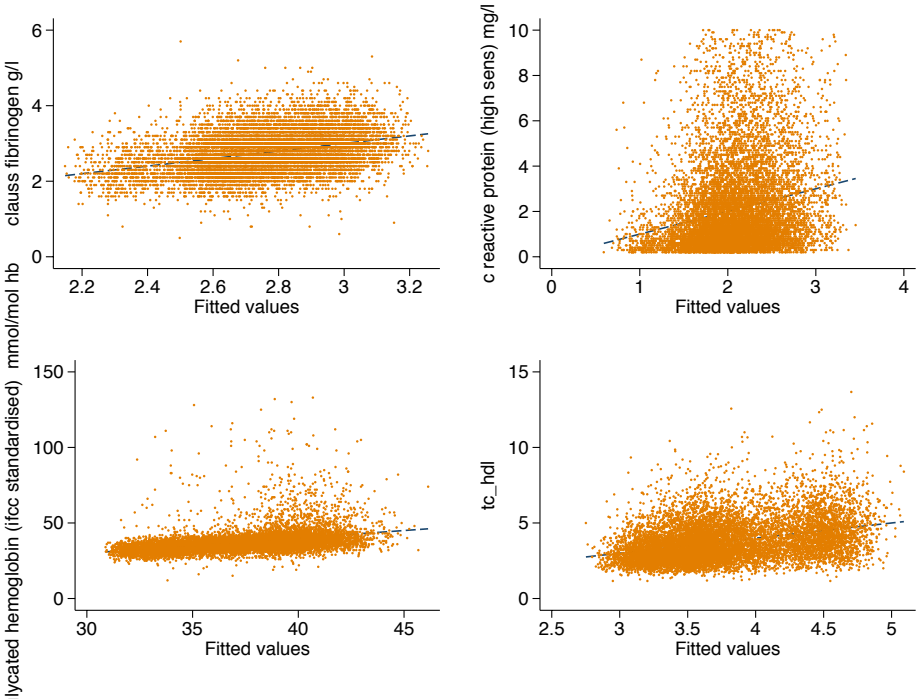
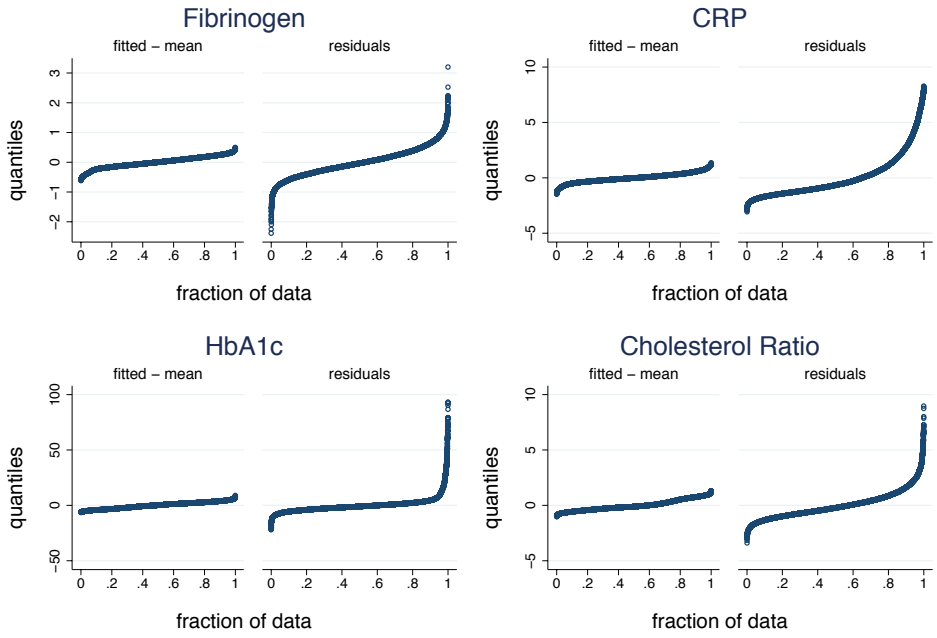


Fig 4.6 `Ovfplot` for linear regression models

Further insight is provided by another plot from the Modeldiag package. The `qfrplot` shown in Figure 4.7 shows the distribution of the quantiles of the fitted values and the residuals. This emphasizes the narrow range of variation in the fitted values of the conditional mean function. For most of the distributions the range of variation in the residuals is also quite narrow but the small fractions of very large positive residuals in the right-hand tail reflect the skewness of the distributions of CRP, HbA1c and the cholesterol ratio.



Source: UKHLS

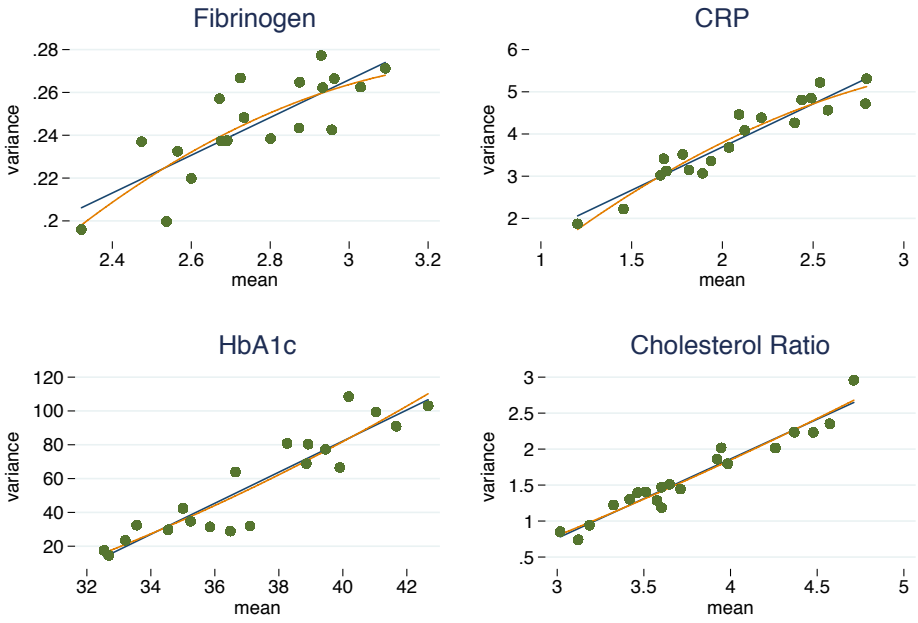
Fig 4.7 `Qfrplot` for linear regression models

The syntax for the `ovfplot` and `qfrplot`, for the case of fibrinogen, is:

```
ovfplot, legend(off) msymbol(p)
         ylabel( , nogrid angle(horizontal))
qfrplot, title(Fibrinogen)
```

The clouds of data points in Figure 4.6 shows that, on the raw scale, the conditional variance increases as the fitted values increase and, most likely, so do the higher conditional moments.

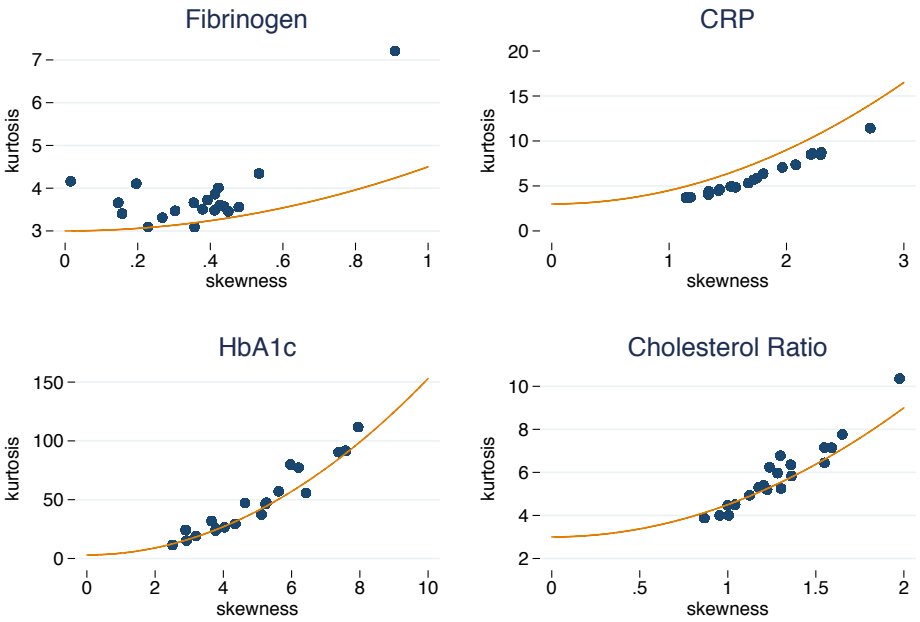
As described above, GLM models specify the conditional mean (link) and conditional variance (distribution) as (potentially) nonlinear functions of a linear index of the regressors. Figure 4.8 graphs the relationship between the conditional mean and variance for each of the biomarkers, using vigiciles of fitted values. These suggest that the variance is an increasing function of the mean in all cases.



Source: UKHLS

Fig 4.8 Conditional mean and variance of biomarkers

Figure 4.9 shows the relationship between conditional kurtosis and skewness. These are reasonably close to the locus implied by a gamma distribution (shown by the orange curves).



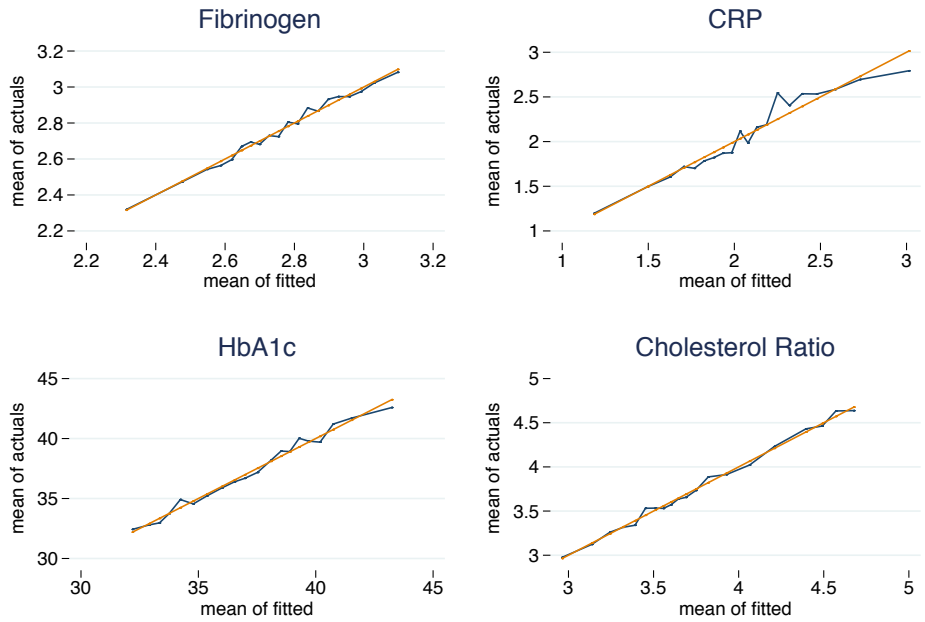
Source: UKHLS

Fig 4.9 Conditional kurtosis and skewness of biomarkers

The GLM framework is implemented based on the extended estimating equations (EEE) approach of Basu and Rathouz (2005) and the associated user-written program `pglm`. There can be computational problems in fitting these models and here the outcome variable is scaled relative to its mean prior to estimation (shown here for fibrinogen):

```
clonevar y=cfib
quietly summ y, meanonly
gen scy = y/r(mean)
global sc = r(mean)
pglm scy $xs
pglmpredict yfc if e(sample), mu scale($sc)
```

The fit of models estimated using the extended estimating equations (EEE) approach is shown in Figure 4.10. These look rather similar to those for the linear regressions shown above with the best fits for HbA1c and the cholesterol ratio and the poorest for CRP.



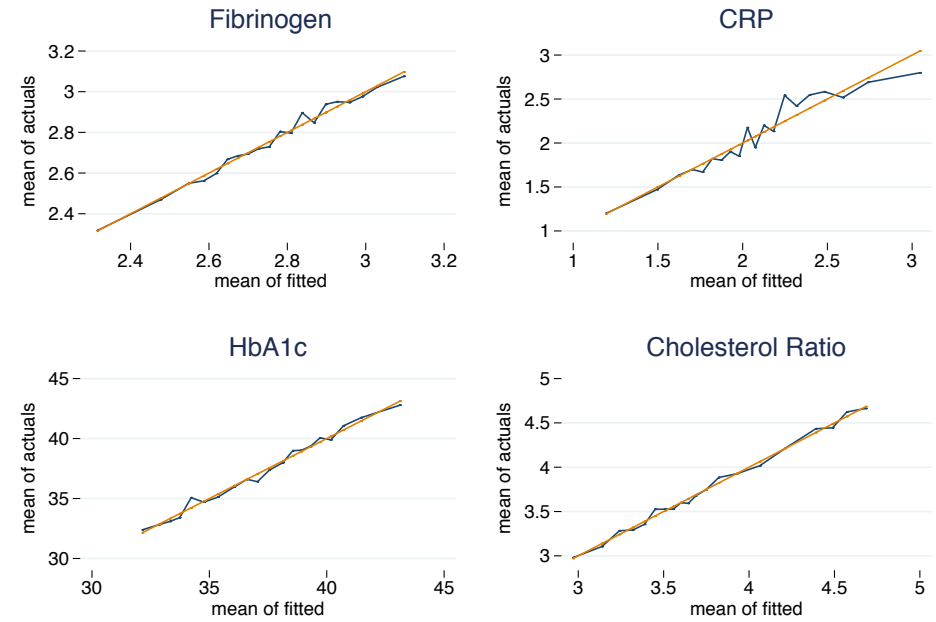
Source: UKHLS

Fig 4.10 Actual versus fitted plot for EEE models

The EEE approach shown here uses distributions based on power functions of the mean, for example the gamma distribution where the variance is proportional to the square of the mean. Another option, that might be consistent with Figure 4.8, is the negative binomial where the variance is a quadratic function with linear and squared terms. This can be accommodated in EEE using the option `vf(q)` but, to illustrate here, GLM models are estimated directly with the negative binomial family and using link functions close to those implied by the EEE estimates:

```
glm cfib $xs, link(power 3) family(nb)
      vce(robust) eform nolog
glm crp $xs, link(power 0.5) family(nb)
      vce(robust) eform nolog
glm hba1c $xs, link(power -2) family(nb)
      vce(robust) eform nolog
glm tc_hdl $xs, link(power 3) family(nb)
      vce(robust) eform nolog
```

Figure 4.11 compares the fit of these GLM models, with CRP once again appearing problematic.



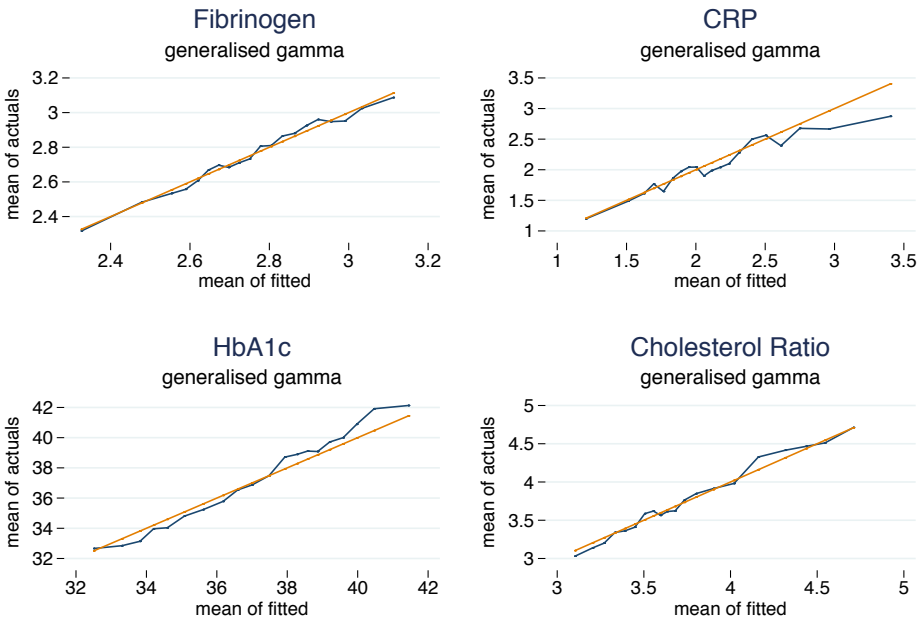
Source: UKHLS

Fig 4.11 Actual versus fitted plot for GLM models

Turn now to flexible parametric models; the generalized gamma model is estimated for each biomarker. The code to estimate the model and create predictions of the conditional means is:

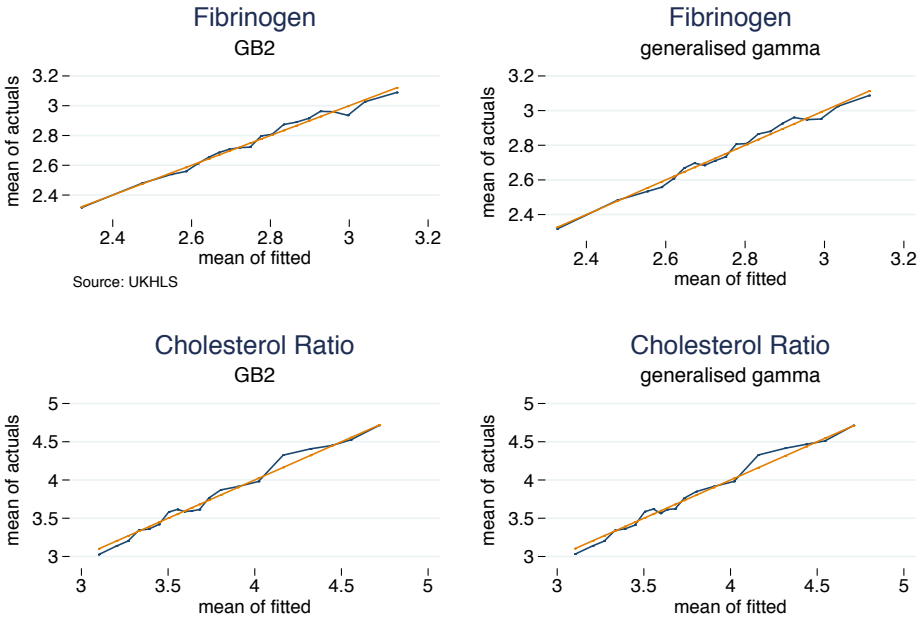
```
stset cfib
streg $xs, dist(ggamma) vce(robust) skip
      nonrtolerance nolog time
predict yfs if e(sample), xb
replace yfs=exp(yfs+(e(sigma)/e(kappa))
         *log(e(kappa)^2)+lngamma((1/e(kappa)^2)
         +e(sigma)/e(kappa))-lngamma(1/e(kappa)^2))
```

Note that the dataset has to be `stset` prior to estimation in order to use the survival regression, `streg`, command. Figure 4.12 presents the actual versus fitted values. These show a good fit for fibrinogen and a reasonable fit for the cholesterol ratio but suggest that there is likely to be an issue of misspecification for CRP, where there is a poor fit in the right-hand tail, and HbA1c, where there is systematic over-prediction at the bottom end of the distribution and under-prediction at the top end. Figure 4.13 compares the generalized gamma with the GB2 distribution for both fibrinogen and the cholesterol ratio. These are very similar in appearance.



Source: UKHLS

Fig 4.12 Actual versus fitted plot for generalized gamma models



Source: UKHLS

Fig 4.13 Actual versus fitted plot for generalized gamma and GB2 models

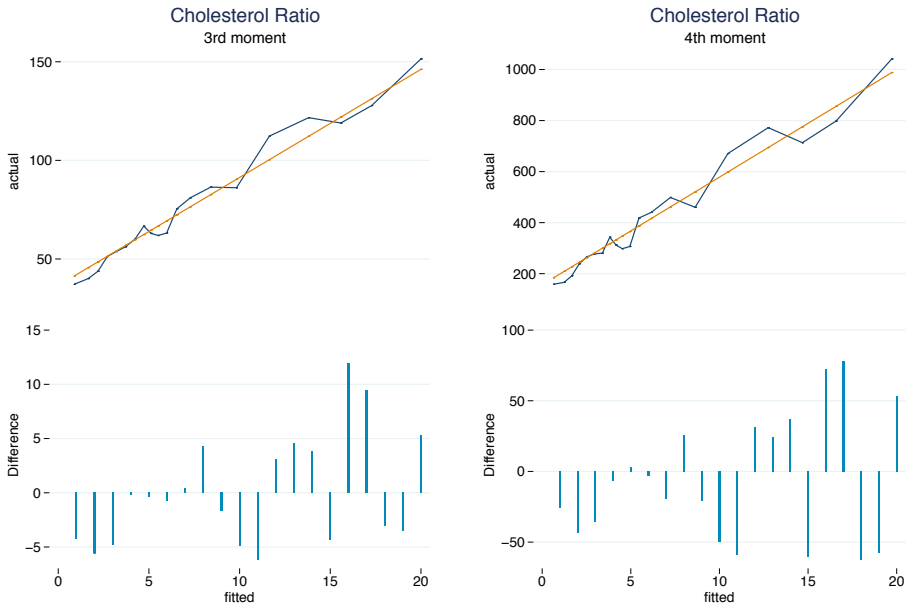
Estimation of the GB2 model is by maximum likelihood and uses a program from Jones et al. (2014) along with commands to compute and graph the fitted values of the conditional mean (shown here for fibrinogen):

```
program gb2log
      args lnf lnb a p q
      local y1 "$ML_y1"
      quietly {
      replace `lnf' = ln(abs(`a')) +
      (abs(`a')*abs(`p') - 1)*ln(`y1')
       - (abs(`p') + abs(`q'))
      *ln(1+(`y1'/exp(`lnb'))^abs(`a'))
       - abs(`a')*abs(`p')*`lnb' - lngamma(abs(`p'))
       - lngamma(abs(`q'))+lngamma(abs(`p')+abs(`q'))
            }
      end

* Estimation
ml model lf gb2log (lnb:  cfib = $xs) /a /p /q
ml search
ml max, noclear
ml display
* Predictions
predict double a_gb2, eq(a)
predict double p_gb2, eq(p)
predict double q_gb2, eq(q)
gen double gb2_lnbest = [lnb]_b[_cons]
foreach var of varlist $xs {
 replace gb2_lnbest = gb2_lnbest+[lnb]_b[`var']*`var'
}
gen double b_gb2 = exp(gb2_lnbest)
gen double yf=b_gb2*((exp(lngamma(p_gb2+(1/a_gb2)))
    *exp(lngamma(q_gb2-(1/a_gb2)))))
    /(exp(lngamma(p_gb2))*exp(lngamma(q_gb2)))

* Plot of averages of actual and fitted
xtile yfd=yf, nq(20)
bysort yfd: egen yfbar=mean(yf)
bysort yfd: egen ybar=mean(cfib)
tw line ybar yfbar yfbar, title(Fibrinogen)
  subtitle(GB2) ytitle(mean of actuals)
  xtitle(mean of fitted) ylabel( , angle(horizontal))
  xsca(noline) ysca(noline) legend(off)
  saving(deccfib, replace) note($note)
```

Figures 4.12 and 4.13 evaluate the generalized gamma and GB2 models in terms of their ability to fit the conditional mean of the data. However the attraction of these flexible 3 and 4-parameter distributions is more about their ability to fit the higher moments of the distribution, especially skewness and kurtosis, as well as conditional tail probabilities. The moments of the GB2 distribution are defined by (Jones et al., 2014):

$$E(y^r) = b^r \left[\frac{\Gamma\left(p+\frac{r}{a}\right)\Gamma\left(q-\frac{r}{a}\right)}{\Gamma(p)\Gamma(q)} \right] \tag{4.1}$$

So, to give a sense of the performance of the GB2 model in this respect, Figure 4.14 plots the actual and fitted values of the third and fourth moments of the cholesterol ratio. These are complemented by **spike plots** to show the absolute differences between the actual and fitted values of the third and fourth moments, shown at each of the 20 intervals of the fitted values of the mean that are used to split the sample and hence condition on different levels of the regressors.



Source: UKHLS

Fig 4.14 Actual versus fitted higher moments for GB2

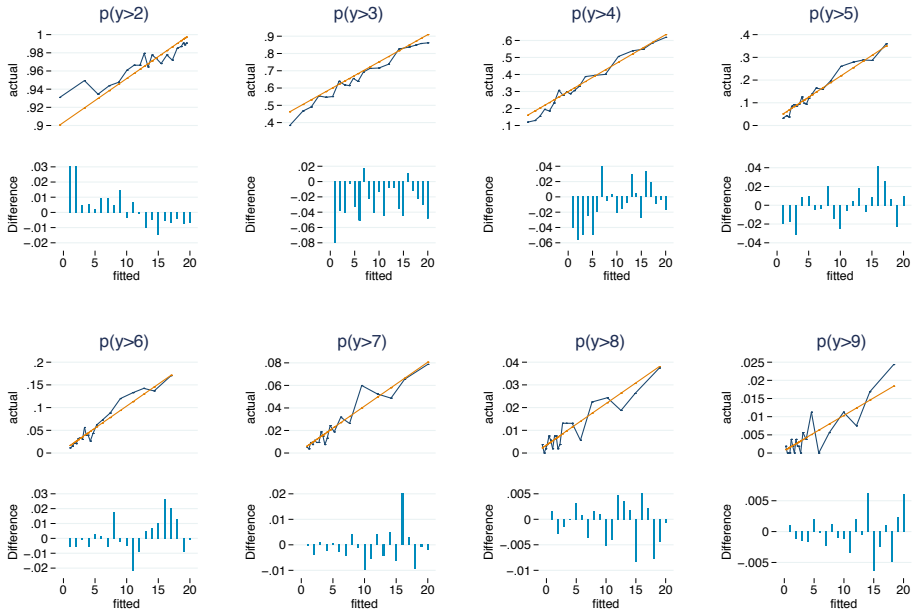
The code used for the plots of the third moments is shown here:

```
gen double yf_3=((b_gb2)^3)
    *((exp(lngamma(p_gb2+(3/a_gb2)))
    *exp(lngamma(q_gb2-(3/a_gb2)))))
    /(exp(lngamma(p_gb2))*exp(lngamma(q_gb2)))
gen y_3=tc_hdl^3
bysort yfd: egen yf3bar=mean(yf_3)
bysort yfd: egen y3bar=mean(y_3)
tw line y3bar yf3bar yf3bar, title(Cholesterol Ratio)
   subtitle(3rd moment) ytitle(actual)
   xtitle(fitted) ylabel( , angle(horizontal))
   xsca(off) ysca(noline) legend(off)
   saving(sk3, replace)

gen diff3=y3bar-yf3bar
tw spike diff3 yfd, nodraw xsca(noline)
   ysca(noline) lcolor(ebblue) legend(off)
   ylabel( , angle(horizontal)) ytitle(Difference)
   xtitle(fitted) saving(spk3, replace)
graph combine sk3.gph spk3.gph, cols(1)
   saving(skspk3, replace)
```

Finally, Figure 4.15 shows how the GB2 model fits the conditional tail probabilities of the distribution. The tail probabilities for the GB2 model involve the incomplete beta function `ibeta` (see Jones et al., 2014). The tail probabilities are computed at different levels of the cholesterol ratio that span the distribution, ranging from 2 to 9. Once again line plots are combined with spike plots of the absolute difference between the average of the actual and fitted values (the bias) at different levels of the fitted mean:

```
foreach j of numlist 2 3 4 5 6 7 8 9 10 {
      gen tp_`j'=0
      replace tp_`j'=1 if tc_hdl>`j'
      gen tpf_`j' =
           1 - ibeta(p_gb2,q_gb2,((`j'/b_gb2)^a_gb2)
               /(1+(`j'/b_gb2)^a_gb2))
      bysort yfd: egen tpfbar_`j'=mean(tpf_`j')
      bysort yfd: egen tpbar_`j'=mean(tp_`j')
      gen diff_`j'=tpbar_`j'-tpfbar_`j'
```

```
        tw line tpbar_`j' tpfbar_`j' tpfbar_`j',
           title(p(y>`j')) ytitle(actual)
           xtitle(fitted) ylabel( , angle(horizontal))
            xsca(off) ysca(noline) legend(off)
            nodraw saving(tptc_hdl_`j', replace)
         tw spike diff_`j' yfd, nodraw xsca(noline)
           ysca(noline) lcolor(ebblue) legend(off)
            ylabel( , angle(horizontal))
           ytitle(Difference) xtitle(fitted)
           saving(spk_`j', replace)
          graph combine tptc_hdl_`j'.gph spk_`j'.gph,
            cols(1) nodraw saving(tpspk_`j', replace)
}
graph combine tpspk_2.gph tpspk_3.gph tpspk_4.gph
      tpspk_5.gph tpspk_6.gph tpspk_7.gph tpspk_8.gph
      tpspk_9.gph,  cols(4) note($note)
```



Source: UKHLS

Fig 4.15 Actual versus fitted tail probabilities

Figure 4.15 shows that the ability of the model to predict the conditional tail probabilities is reasonable up to a cholesterol ratio of

6, corresponding to probabilities of 0.05 to 0.15, but the performance deteriorates beyond that.

5

Conclusion

This article introduces the principles and practice of data visualization and aims to show how these can enhance empirical analysis of health care costs and outcomes, especially for skewed and heavy-tailed distributions. The survey of regression methods for health care costs is complemented by an application of the econometric and graphical methods to blood-based biomarkers from the UK Household Longitudinal Study (UKHLS), known as Understanding Society (Benzeval et al. 2016; Davillas et al., 2017).

The article has shown how graphical methods can be applied using Stata so that the graphics can be integrated with statistical and econometric analysis within one piece of software and using one set of syntax. This means that the graphics presented have been designed for publication in print rather than online and are static and non-interactive. Of course other software packages are available and are well suited for data visualization. Many visualization practitioners use the open source programming language and software environment R; often by installing the package ggplot, which is a purpose built plotting system for R (Wickham, 2011). Specialist commercial packages such as Tableau provide interactive data visualization products and specialist tools are available for those who wish to prepare dynamic and interactive visualizations to be used in web browsers, such as D3.js (Data-Driven Documents), which is a JavaScript library. A source of inspiration for the use of interactivity and animation in health-related visualizations is the Gapminder project, initiated by the late Hans Rosling. This and a selection of other web resources are listed in the Appendix.

Appendix

A.1 Web Resources

Useful web pages and resources for data visualisation methods can be found at:

Excel Charts (Jorge Camoes):
https://excelcharts.com/author/jorge-camoes/

Flowing Data (Nathan Yau):
http://flowingdata.com

Perceptual Edge (Stephen Few):
http://www.perceptualedge.com

PolicyViz (Jonathan Schwabish):
https://policyviz.com

The Functional Art (Alberto Cairo):
http://www.thefunctionalart.com

Pages that focus on health issues and take a visual approach include:

Gapminder:
http://www.gapminder.org

Institute for Health Metrics and Evaluation (IHME):
http://www.healthdata.org/results/data-visualizations

Programs and data for health care cost regressions can be found at:

Health, Econometrics and Data Group (HEDG)
https://www.york.ac.uk/economics/postgrad/herc/hedg/

References

Basu, A. and Rathouz, P.J. (2005), 'Estimating marginal and incremental effects on health outcomes using flexible link and variance function models'. *Biostatistics* **6**, 93-109.

Benzeval, M., Davillas, A., Kumari, M., Lynn, P. (2014), *Understanding Society - UK Household Longitudinal Study: Biomarker User Guide and Glossary*. Colchester: University of Essex,

Benzeval, M., Kumari, M. and Jones, A.M. (2016), 'How do biomarkers and genetics contribute to Understanding Society?'. *Health Economics* **25**, 1219-1222.

Blough, D.K., Madden, C.W. and Hornbrook, M.C. (1999), 'Modeling risk using generalized linear models', *Journal of Health Economics* **18**, 153-71.

Buntin, M.B. and Zaslavsky, A.M. (2004), 'Too much ado about two-part models and transformation?: comparing methods of modeling Medicare expenditures', *Journal of Health Economics* **23**, 525-42.

Cairo, A. (2012), *The Functional Art.* New Riders.

Cairo, A. (2016), *The Truthful Art.* New Riders.

Camoes, J. (2016), *Data at Work, New Riders.*

Chernozhukov, V., Fernandez-Val, I. and Melly, B. (2013), 'Inference on counterfactual distributions'. *Econometrica* **81**, 2205-2268.

Cleveland, W.S. and McGill, R. (1984), 'Graphical perception: theory, experimentation, and application to the development of graphical methods'. *Journal of the American Statistical Association* **79**, 531-554.

Cleveland, W.S. (1985), *The Elements of Graphing Data*. Wadsworth Advanced Books and Software.

Conway, K.S. and Deb, P. (2005), 'Is prenatal care really ineffective? Or, is the 'devil' in the distribution?'. *Journal of Health Economics* **24**, 489-513.

Davillas, A., Jones, A.M. and Benzeval, M. (2017), 'The income-health gradient: evidence from self-reported health and biomarkers using longitudinal data on income', *HEDG Working Paper WP 17/*04.

Deb, P. and Holmes, A.M. (2000), 'Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models'. *Health Economics* **9**, 475-89.

Deb, P. and Trivedi, P.K. (1997), 'Demand for medical care by the elderly: a finite mixture approach'. *Journal of Applied Econometrics,* **12**, 313-36.

Few, S. (2009), *Now you see it. Simple visualization techniques for quantitative analysis*, Analytics Press.

Few, S. (2012), *Show me the Numbers*, 2nd ed., Analytics Press.

Few, S. (2013), *Information Dashboard Design, 2nd ed.*. Analytics Press.

Few, S. (2015), *Signal. Understanding what matters in a world of noise*, Analytics Press.

Firpo S, Fortin, N.M., Lemieux, T. (2009), 'Unconditional quantile regressions'. *Econometrica* **77**, 953–973.

Foresi, S. and Peracchi, F. (1995), 'The conditional distribution of excess returns: an empirical analysis'. *Journal of the American Statistical Association* **90**, 451-466.

Gilleskie, D.B. and Mroz, T.A. (2004), 'A flexible approach for estimating the effects of covariates on health expenditures'. *Journal of Health Economics* **23**, 391-418.

Han, A. and Hausman, J. (1990), 'Flexible parametric estimation of duration and competing risk models'. *Journal of Applied Econometrics* **5**, 1-28.

Harris, R.L. (1999), *Information Graphics. A Comprehensive Illustrated Reference*. Oxford University Press.

Hill, S.C. and Miller, G.E. (2010), 'Health expenditure estimation and functional form: application of the generalized gamma and extended estimating equation models'. *Health Economics* **19**, 608-627.

Holly, A. (2009), 'Modelling risk using fourth order pseudo maxiumum likelihood methods'. Institute of Health Economics and Management (IEMS), University of Lausanne: Switzerland.

Imai, K and Van Dyk, DA. (2004), 'Causal inference with general treatment regimes: generalizing the propensity score'. *Journal of the American Statistical Association* **99**, 854-866.

Jenkins, S.P. (2009), 'Distributionally-sensitive inequality indices and the GB2 income distribution'. *The Review of Income and Wealth* **55**, 392-398.

Johnson, E., Dominici, F., Griswold, M. and Zeger, S.L. (2003), 'Disease cases and their medical costs attributable to smoking: an analysis of the National Medical Expenditure Survey'. *Journal of Econometrics* **112**, 135-151.

Jones, A.M. (2000), 'Health Econometrics'. In Culyer, A. J. and J. P. Newhouse (eds), *Handbook of Health Economics.* Elsevier.

Jones, A.M. (2009), 'Panel data methods and applications to health economics', in Mills, T.C. and Patterson, K. (eds), *Palgrave Handbook of Econometrics. Volume 2*. Palgrave MacMillan.

Jones, A.M. (2011), 'Models for health care', in *Oxford Handbook of Economic Forecasting*, Hendry, D. and Clements, M. (eds.), Oxford University Press.

Jones, A.M., Rice, N., Bago d'Uva, T. and Balia, S. (2013), *Applied Health Economics, 2nd Edition* Routledge, 2013.

Jones, A.M., Lomas, J. and Rice, N. (2014), 'Applying beta-type size distributions to healthcare cost regressions'. *Journal of Applied Econometrics* **29**, 649-670.

Jones, A.M., Lomas, J., and Rice, N. (2015), 'Healthcare cost regressions: going beyond the mean to estimate the full distribution'. *Health Economics* **24**, 1192-1212.

Jones, A.M., Lomas, J., Moore, P. and Rice, N. (2016), 'A quasi-Monte Carlo comparison of recent developments in parametric and

semi-parametric regression methods for heavy tailed and non-normal data: with an application to health care costs', *Journal of the Royal Statistical Society Series A* **179**, 951-974.

Knies, G. (ed.) (2015), *Understanding Society – UK Household Longitudinal Study: Wave 1-5, 2009-2014, User Manual*. Colchester: University of Essex.

McDonald, J.B., Sorensen, J. and Turley P.A. (2011), 'Skewness and kurtosis properties of income distribution models'. *Review of Income and Wealth* **59**, 360-374.

McFall, S.L., Petersen, J., Kaminska, O., Lynn, P. (2014), *Understanding Society – UK Household Longitudinal Study: Waves 2 and 3 Nurse Health Assessment, 2010-2012, Guide to Nurse Health Assessment.* Colchester: University of Essex.

Machado, J.A.F. and Mata, J. (2005), 'Counterfactual decomposition of changes in wage distributions using quantile regression'. *Journal of Applied Econometrics* 20, 445-465.

Manning, W. (1998), 'The logged dependent variable, heteroscedasticity, and the retransformation problem'. *Journal of Health Economics* **17**, 283-95.

Manning, W. (2006), 'Dealing with skewed data on costs and expenditure.' In Jones, A.M. (ed) *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar.

Manning, W.G., Basu, A. and Mullahy, J. (2005), 'Generalized modeling approaches to risk adjustment of skewed outcomes data'. *Journal of Health Economics* **24**, 465-88.

Manning, W.G. and Mullahy, J. (2001), 'Estimating log models: to transform or not to transform?'. *Journal of Health Economics* **20**, 461-94.

Melly, B. (2005), 'Decomposition of differences in distribution using quantile regression'. *Labour Economics* **12**, 577-590.

Mitchell, M.N. (2012), *A Visual Guide to Stata Graphics.* Stata Press.

Mullahy, J. (1998), 'Much ado about two: reconsidering retransformation and the two-part model in health econometrics'. *Journal of Health Economics* **17**, 247-81.

Mullahy, J. (2009), 'Econometric modeling of health care costs and expenditures. A survey of analytical issues and related policy considerations'. *Medical Care* **47**, S104-S108.

Nussbaumer Knaflic, C. (2015), *Storytelling with Data. A Data Visualization Guide for Business Professionals,* Wiley.

Robbins, N.B. (2005), *Creating More Effective Graphs.* Wiley.

Rubin, D.B. (2001), 'Using propensity scores to help design observational studies: application to the tobacco litigation'. *Health Services & Outcomes Research* **2**, 169-188.

Schwabish, J.A. (2014), 'An economist's guide to visualizing data'. *Journal of Economic Perspectives* **28**, 209-234.

Tufte, E.R. (1990). *Envisioning Information*. Graphics Press.

Tufte, E.R. (1997), *Visual Explanations*. Graphics Press.

Tufte, E.R. (2001), *The Visual Display of Quantitative Information*, 2nd ed., Graphics Press.

Tufte, E.R. (2006), *Beautiful Evidence*. Graphics Press.

Von Hinke Kessler Scholder, S. and Jones, A.M. (2015) 'Cohort data in health economics', in Baltagi, B. (ed.), *Oxford Handbook of Panel Data*. Oxford University Press.

Wickham, H. (2011), *ggplot2: Elegant Graphics for Data Analysis,* Springer.

Yau, N. (2011), *Visualize This.* Wiley.

Yau, N. (2013), *Data Points*. Wiley.