# Graphical Methods for Exploratory Multivariate Longitudinal Data Analysis

Ozlem Ilk and Dianne Cook,
Department of Statistics, Iowa State University, Ames, IA 50011-1210
oilk@iastate.edu

**Key Words:** data mining, direct manipulation graphics, dynamic graphics, exploratory data analysis, linked brushing, statistical graphics, visualization.

## Abstract

This paper describes direct manipulation and dynamic graphics for analyzing multivariate longitudinal data. Longitudinal data measures individuals repeatedly in time, perhaps at irregular and unequal time points. There is an emphasis on studying the individual patterns as well as mean trends because we can. Static plots of individuals are messy and often unreadable because there are many overlapping lines. Direct manipulation graphics allow patterns amongst the individuals to be explored. Dynamic graphics enable patterns in multivariate responses to be viewed. Three very different data sets are used to illustrate the methods.

## 1 Introduction

In longitudinal (panel) data individuals are repeatedly measured through time which enables the direct study of change (Diggle, Heagerty, Liang & Zeger 2002). Each individual will have certain special characteristics, and measurements on several topics or variables may be taken each time an individual is measured. The reporting times can vary from individual to individual in number, dates and time between reporting. This deviation from equi-spaced, equal quantity time points, producing a ragged time indexing of the data, is common in longitudinal studies and it causes grief for many data analysts. It may be difficult to develop formal models to summarize trends and covariance, yet there may be rich information in the data. There is a need for methods to tease information out of this type of complex data (Singer & Willett 2003). Most documented analyses discuss equi-spaced, equal quantity longitudinal measurement, but ragged time indexed data is probably more common than the literature would have us believe. This paper discusses exploratory methods for difficult to model ragged time indexed longitudinal data.

The basic question addressed by longitudinal studies is how the responses vary through time, in relation to the covariates. Unique to longitudinal studies is the ability to study individual responses. This is different from repeated cross-sectional studies which take different samples at each measurement time, to measure the societal trends but not individual experiences. Longitudinal studies are similar to time series except that there are multiple time series, one for each individual. Software for time series can deal with one time series or even a couple, but the analysis of hundreds of them is not easily possible. The analysis of repeated measures could be considered to be a subset of longitudinal data analysis where the time points are equal in number and spacing (Crowder & Hand 1990).

Analysts want to explore many different aspects of longitudinal data - the distribution of values, temporal trends, anomalies, the relationship between multiple responses and covariates in relation to time. Exploration, which reveals the unexpected in data and is driven by rapidly changing questions, means it is imperative to have graphical software which is interactive and dynamic: software that responds in real time to an analyst's enquiries and changes displays dynamically, depending on the analyst's questions. Plots provide insight into multiple aspects of the data, overviews of the general behavior and tracking individuals. Analysts may also want to link recorded events, such as a graduation or job loss, to an individuals' behavior. Even with unequal time points the values for each individual can be plotted, for example, each variable against time, variable against variable with measurements for each individual connected with line segments. Linking between plots, using direct manipulation, enables the analyst to explore relationships between responses and covariates (Swayne & Klinke 1998, Unwin, Hofmann & Wilhelm 2002). Dynamic graphics such as tours (Asimov 1985) will enable the study of multivariate responses.

There is very little in the literature discussing graphical methods for longitudinal data. Both Diggle et al. (2002) and Singer & Willett (2003) state there is a need for graphics but have only brief chapters describing static graphics. Koschat & Swayne (1996) illustrated the use of direct manipulation for customer panel data. They applied tools such as case identification, linking multiple views and brushing on scatterplots, dot plots and clustering trees, and a plot they called the case-profile plot (time series plot of a specific subject). Case-profile plots are also known as parallel coordinates(Inselberg 1985, Wegman 1990), interaction plots, or profile plots in the repeated measures and ANOVA literature. Koschat and Swayne recommended looking at different views of the same data. Sutherland, Rossini, Lumley, Lewin-Koh, Dickerson, Cox & Cook (2000) demonstrate viewing multiple responses in relation to the time context using a tour. Faraway (1999) introduced what he called a graphical method for exploring the mean structure in longitudinal data. His approach fits a regression model and uses graphical displays of the coefficients as a function of time. Thus the method describes graphics for plotting model diagnostics but not the data: graphical method is an inaccurate title.

*** What are we doing that is different to Koschat and Swayne?

The purpose of this paper is to describe the use of direct manipulation and dynamic graphics to "slice-and-dice" multivariate longitudinal data in the spirit of exploratory data analysis. The next section describes longitudinal data, sets up a notation, and describes the types of questions that are typical for this kind of data. Section 3 describes approaches for studying mean trends and Section 4 describes approaches for exploring individual patterns. Finally we discuss what methods would ideally be available in a software for multivariate longitudinal data. To illustrate the graphical methods, three data sets are used. This is a lot of examples! But the methods are explained through examples, and the three data sets cover a representative range of different types of longitudinal data. The first two examples have been previously analyzed in the literature. In each case from the exploratory analysis we have discovered different and perhaps important structure than was not revealed by the previous analyses.

# 2 What is longitudinal data?

## 2.1 Notation

We denote the *response* variables to be $\mathbf{Y}_{ijt_i}$, and the time-dependent explanatory variables, or *covariates* to be $\mathbf{X}_{ikt_i}$, where $i = 1, \ldots, n$ indexes the number of individuals in the study, $j = 1, \ldots, q$ indexes the number of response variables, $k = 1, \ldots, p$ indexes the number of covariates, and $t_i = 1, \ldots, n_i$ indexes the number of times individual $i$ was measured. Note that $n$ is the number of subjects or individuals in the study, $n_i$ is the number of time points measured for individual $i$, $q$ is the number of response variables, and $p$ is the number of explanatory variables, measured for each individual and time. The explanatory variables may include indicator variables marking special events affecting an individual. There may also be time-independent explanatory variables or covariates, which we will denote as $\mathbf{Z}_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, r$. Simplifications to the notation can be made when the data is more constrained, such as equi-distant, equal number of time points.

Consider the wages data analyzed in Singer & Willett (2003). A subset of the data is shown Table 1.

For the wages data, there is only one response, `lnw`, and one time-dependent explanatory variable, `uerate`, thus $q = 1, p = 1$. The subset has $n = 2$ cases, with identities, $31, 36$, and the number of time measurements

| id | lnw | exper | black | hispanic | hgc | uerate |
|---|---|---|---|---|---|---|
| 31 | 1.491 | 0.015 | 0 | 1 | 8 | 3.215 |
| 31 | 1.433 | 0.715 | 0 | 1 | 8 | 3.215 |
| 31 | 1.469 | 1.734 | 0 | 1 | 8 | 3.215 |
| 31 | 1.749 | 2.773 | 0 | 1 | 8 | 3.295 |
| 31 | 1.931 | 3.927 | 0 | 1 | 8 | 2.895 |
| 31 | 1.709 | 4.946 | 0 | 1 | 8 | 2.495 |
| 31 | 2.086 | 5.965 | 0 | 1 | 8 | 2.595 |
| 31 | 2.129 | 6.984 | 0 | 1 | 8 | 4.795 |
| 36 | 1.982 | 0.315 | 0 | 0 | 9 | 4.895 |
| 36 | 1.798 | 0.983 | 0 | 0 | 9 | 7.400 |
| 36 | 2.256 | 2.040 | 0 | 0 | 9 | 7.400 |
| 36 | 2.573 | 3.021 | 0 | 0 | 9 | 5.295 |
| 36 | 1.819 | 4.021 | 0 | 0 | 9 | 4.495 |
| 36 | 2.928 | 5.521 | 0 | 0 | 9 | 2.895 |
| 36 | 2.443 | 6.733 | 0 | 0 | 9 | 2.595 |
| 36 | 2.825 | 7.906 | 0 | 0 | 9 | 2.595 |
| 36 | 2.303 | 8.848 | 0 | 0 | 9 | 5.795 |
| 36 | 2.329 | 9.598 | 0 | 0 | 9 | 7.600 |

Table 1: Long format: time-independent covariates repeated at each time point.

| id | black | hispanic | hgc |
|---|---|---|---|
| 31 | 0 | 1 | 8 |
| 36 | 0 | 0 | 9 |

| id | lnw | exper | uerate |
|---|---|---|---|
| 31 | 1.491 | 0.015 | 3.215 |
| 31 | 1.433 | 0.715 | 3.215 |
| 31 | 1.469 | 1.734 | 3.215 |
| 31 | 1.749 | 2.773 | 3.295 |
| 31 | 1.931 | 3.927 | 2.895 |
| 31 | 1.709 | 4.946 | 2.495 |
| 31 | 2.086 | 5.965 | 2.595 |
| 31 | 2.129 | 6.984 | 4.795 |
| 36 | 1.982 | 0.315 | 4.895 |
| 36 | 1.798 | 0.983 | 7.400 |
| 36 | 2.256 | 2.040 | 7.400 |
| 36 | 2.573 | 3.021 | 5.295 |
| 36 | 1.819 | 4.021 | 4.495 |
| 36 | 2.928 | 5.521 | 2.895 |
| 36 | 2.443 | 6.733 | 2.595 |
| 36 | 2.825 | 7.906 | 2.595 |
| 36 | 2.303 | 8.848 | 5.795 |
| 36 | 2.329 | 9.598 | 7.600 |

Table 2: Short format: Separate table for time-independent covariates.

for each subject are $n_1 = 8, n_2 = 10$. (The full data set has $n = 888$.) There are $r = 2$ time-independent covariates, `race`, coded into two dummy variables (`black`, `hispanic`), and highest grade completed, `hgc`. The data is written in long format in this table; the time-independent covariates repeated for each time value. A shorter format would be achieved by two tables linked by id, one for the time-independent covariates and one for time-dependent measurements, in Table 2. This latter format is the one that we use in the exploratory approach illustrated next. We set up the data in two tables, one containing the time-dependent measurements, another containing the time-independent covariates. The time-dependent measurements for each individual will be related in a plot using line segments.

### 2.2 Types of questions common in longitudinal data

The questions that are common in cross sectional studies are also frequently asked for longitudinal data. For instance, we are usually interested in descriptive statistics such as the number of males in the study, or the number of CD4+ cells (a marker for HIV) for a group of patients on their first visit to the medical office.

However, because we have measurements recorded in a time context, the primary question is about temporal change in responses, with respect to covariates. The number of CD4+ cells at one time point is not very informative; the *change* in the number of cells over time is an indicator of disease. An healthy person has around 1100 cells per milliliter of blood, and this number decreases over time for an infected person. It is also interesting to know if the change has a different pattern depending on the patient's gender or ethnicity. We consider these to be questions about the mean trend.

Longitudinal data lends itself to the study of individuals, and groups of individuals such as families. For instance, we would like to explore the association among covariates, and the correlation structure among cases that belong to the same group. Any outstanding individual or groups or unexpected patterns are of peculiar interest. Finding these patterns might lead us to uncover errors in the process of collecting or typing data, or it might have a statistical implications such as interactions between covariates. Missing cases are common in longitudinal data, probably because data collection is a lengthy process. We might lose track of subjects because of death, or illness, or relocations, or subjects might simply be unwilling to cooperate after some time. Exploring the missing value mechanism is an important task in longitudinal data analysis.

These types of questions can be addressed by the methods described in this paper.

## 3 Mean Trends

The primary question is how responses vary with time. The immediate impulse is to plot $Y_{ij}$ against $t_i$, with values for each individual connected by line segments. These plots can be very messy, and practically useless (Diggle et al. 2002). To assess the trend with time requires some estimate of the trend to be plotted, along with enough information about the distribution of values to assess the strength of the trend. For ragged time indexed data we use a smoother to estimate the trend, otherwise we calculate the median or mean at each time. Plots and calculations are conditioned by time-independent categorical covariates. For non-ragged time indexed data we can also condition on time.

### 3.1 Example 1: Wages

This example is used to illustrate methods for ragged time data. This data, introduced in Section 2.1, is ragged time indexed data. There is only one response and one time-dependent covariate. Singer & Willett (2003) use this data to illustrate fitting mixed linear models to ragged time indexed data (Figure 1). The analysis reports that the average growth in wages is about 4.7% for each year of experience. There is no difference between whites and Hispanics, but a big difference from blacks. The model uses a linear trend (on the log wages) to follow these patterns. The within variance components of the model is significant. This says that the variability for each person is dramatically different. It is possible to estimate the individual trajectories to examine how people differ, and which people are different or similar to others. Although some

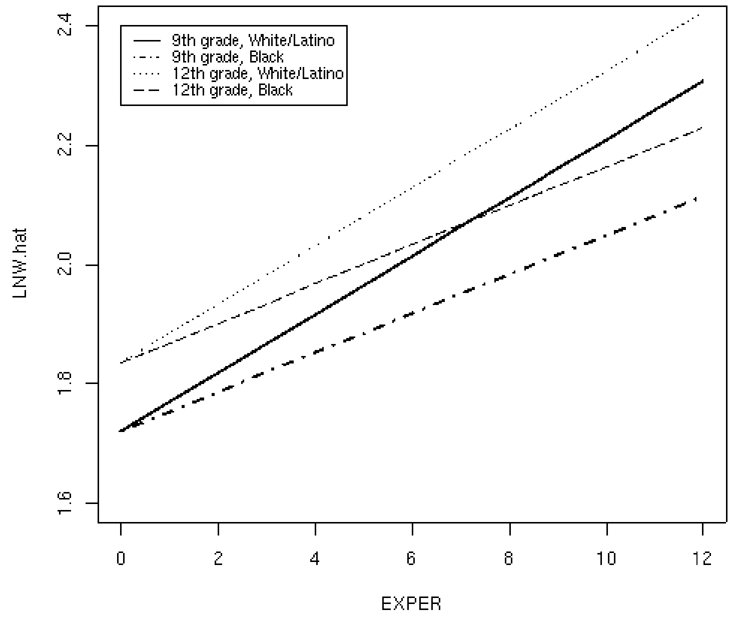

Figure 1: Linear models for wages by race and education.

software programs provide these estimates, it is usually impractical to calculate and save estimates for all subjects in moderate to large samples. They can be calculated for a randomly selected sample of subjects, but, it will be likely to miss interesting patterns on individual level. Here and in section 4 we show what we can learn about the data that is different from what Singer & Willett (2003) found.

Figure 2 (top left) displays the lowess smooth (Cleveland 1979) of the wages value in relation to experience, overlaid on a scatterplot of the data values $(Y_{ij}, t_i)$. The trend is that wages increase with increasing experience. This is not surprising. It is expected that with more experience wages will be higher.

The purpose of showing the scatterplot underneath the curve is to assess the variation around the mean trend. The variation in this data is huge. The mean trend is quite strong, but its also clear that the variation in wages is quite large across the full range of experience. We use very small glyphs, single pixels because there are a lot of points, and a lot of ink. If the individual profiles were drawn as line then there would be too much ink to see the smoothed line showing the mean trend (see Diggle et al. (2002) for more explanation). Estimates of variance could be added to this plot in a variety of ways, see Faraway (1999) for example and discussions in Diggle et al. (2002).

Figure 3 (middle and right plot) display the lowess smoothed lines of wages based on experience conditionally on race. There appears to be a difference in the wages for men with more workforce experience according to race: people who are black have somewhat lower wages on average than Hispanic and other races when they have similarly high levels of experience. For blacks the wages plateau out around 5-7 years of experience and then increase again. The scatterplots at right also show a dramatic difference. Whites and Hispanics have a clearly positive linear association between wages and experience, but the relationship is not positive linear for blacks. But there are also fewer blacks with 9 or more years of experience than whites and hispanics, which makes the results at the upper end of experience less reliable.

*How strong is the difference in trend?* Using ideas described in Buja (1999) which builds on considerable material on permutation tests described in Good (1993). The plots in Figure 4 are produced by permuting the race identity of each individual fifteen times. A lowess smoother is computed for each fake race group. The actual data is also plotted amongst these plots of permuted data to give sixteen plots. Now the way to

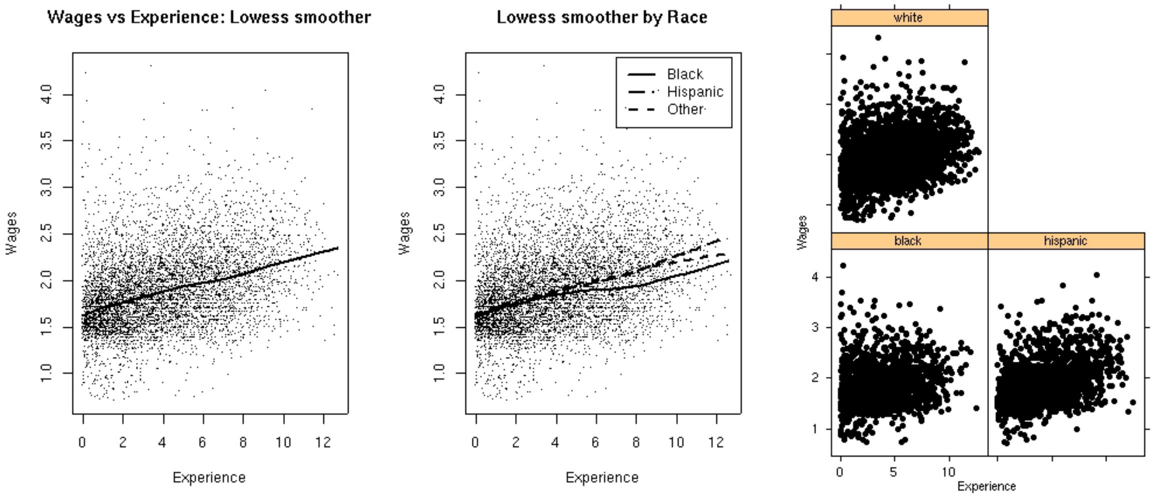

Figure 2: Mean trends using lowess smoother: (Left) Overall wages increase with experience. (Middle) Race makes a difference, as more experience is gained. (Right) The scatter plot of wages against experience conditioned on race. The pattern is different for the different races, in that whites and Hispanics appear to have a more positive linear dependence than blacks, and there are less blacks with the longest experiences. This latter fact could be a major reason for the trend difference.

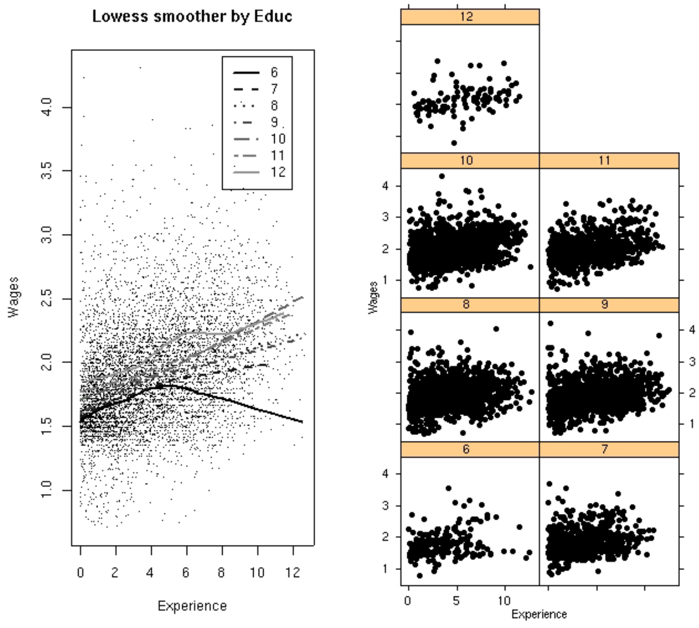

Figure 3: (Left) Mean trends using lowess smoother conditioned on last year of school. (Right) The scatter plot of wages against experience conditioned on last year of school.

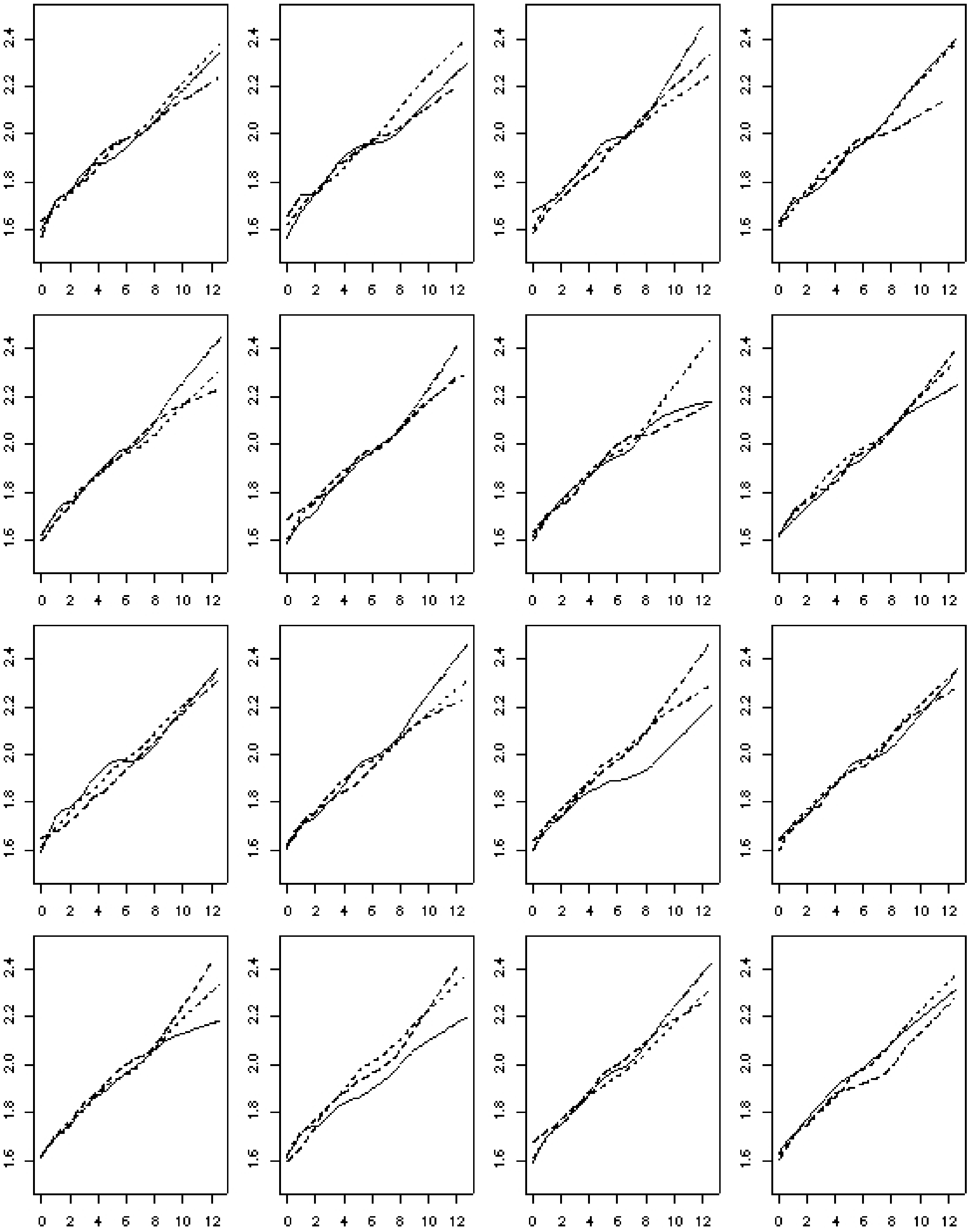

Figure 4: Assessing the significance of the difference in mean trend for race. The plots here were generated by permuting the race variable, and recalculating the lowess smoothed lines. One of the plots is generated using the correct labels. Which is the real data?

use this page of plots is to take it down the hallway to colleagues, or home to your family and neighbors, and ask them to identify the most unusual plot. You can't do this yourself because you've already seen the real plot. If your surveyees consistently point to the plot of the actual data then there is evidence to say that the difference in trend between the races is real. We think that they will, the real difference between wages and experience for blacks in comparison to whites and hispanics is the mid-experience plateau. This is much more extreme in the plot of the real data than in any of the permuted data plots. It is interesting to note that, at the higher end of experience a big difference in the three curves is observed in most of the plots. This says that the difference observed between races in wages for long-term workforce experience is probably not significant. The differences are likely due to the lack of data.

Figure 3 shows lowess smoothed lines of wages based on experience conditionally on education. For education, there is some difference in average wages when there is little experience and the gap widens with more experience. In general, more education means higher wages, especially with more experience. The interesting contradiction is for individuals with the least education (6 years) is that with more experience the wages drop dramatically. A slightly similar pattern can be seen with people with the most education (12 years). These trends are suspicious, and really can probably be explained by lack of data. The bottom right shows the wages vs experience plot conditioned on the education, and it can be seen that there are not too many people in the 6 and 12 years categories of education. An interesting observation is that with earlier dropout there are fewer men at the longer times of workforce experience.

### 3.2 Example 2: Panel Study of Income Dynamics(PSID)

Faraway (1999) suggests an exploratory method for examining the mean structure in longitudinal data. As the example he uses a random sample of 85 heads of household from a much larger study, the Panel Study of Income Dynamics(PSID). The response variable is annual household income recorded every year between 1968 and 1990. This is equi-spaced, equal number of time points data. There are several covariates, age at the onset of the study, number of years of education and gender of the head of household. Table 3 gives the median values for raw incomes for every fourth year. The overall median annual household income increases from \$4,530 to \$21,000 in twenty years. The ratio of household income for males compared to females is approximately 2 to 3. In the coming analysis income is logged (base $e$) to reduce skewness. Of the 85 heads of household, a little under half, 39, are female. The age of the household head ranges from 25 to 39 at the start of the study, and about half are younger than 33. Most subjects are either high-school (34) or college graduates (13). There is no information about occupation. However, in the full sample, the majority of subjects are reported to be craftsmen, sales workers, farm workers or farmers, professionals and managers, or not in the labor force in the previous year.

| | 1968 | 1972 | 1976 | 1980 | 1984 | 1988 |
|---|---|---|---|---|---|---|
| Overall | 4530 | 6525 | 7500 | 9375 | 15850 | 21000 |
| Male | 5995.5 | 8701 | 12000 | 14400 | 22300 | 28750 |
| Female | 2450 | 2950 | 3676 | 6240 | 9800 | 15000 |
| M/F | 2.45 | 2.95 | 3.26 | 2.31 | 2.28 | 1.92 |

Table 3: Median of raw incomes (\$) in PSID data over selected years.

Faraway's approach addresses only the mean trends in the data. He fits a model for each time point:

$$\log(inc\hat{o}me)_i(t) = \beta_0(t) + \beta_g(t)gender_i + \beta_e(t)education_i + \beta_a(t)age_i + \epsilon_i(t)$$

His "contribution" to graphics is that the estimated regression coefficients are plotted against time, interpolating through time. Lowess smooth curves (or true regression coefficients in simulation study) and $\mp$ 2 standard errors are plotted as well. The slope of the lines is examined to assess the change in the mean trend. Moreover, whether the pointwise confidence bands cover the zero line is examined to assess whether the covariate significantly affects the response. A major concern about this approach is that the data is not plotted. Only the model statistics are plotted which is dangerous. Studying only the model statistics is akin to testing only if the model is capturing something, not whether it's capturing the structure of the data or what the model is missing. Its important to plot the model in relation to the data to assess the fit.

He concluded that females earned substantially less than males but the difference decreased with time. An approximately linear (increasing over time) effect of education was observed. There was no important age effect, but some indication of quadratic age effect was reported. These are rather inadequate findings.

*What more could have been found in this data?* Using exploratory graphics we immediately find many more interesting observations about this data. Figures 5, 6 and 7 illustrate some aspects of the approach.

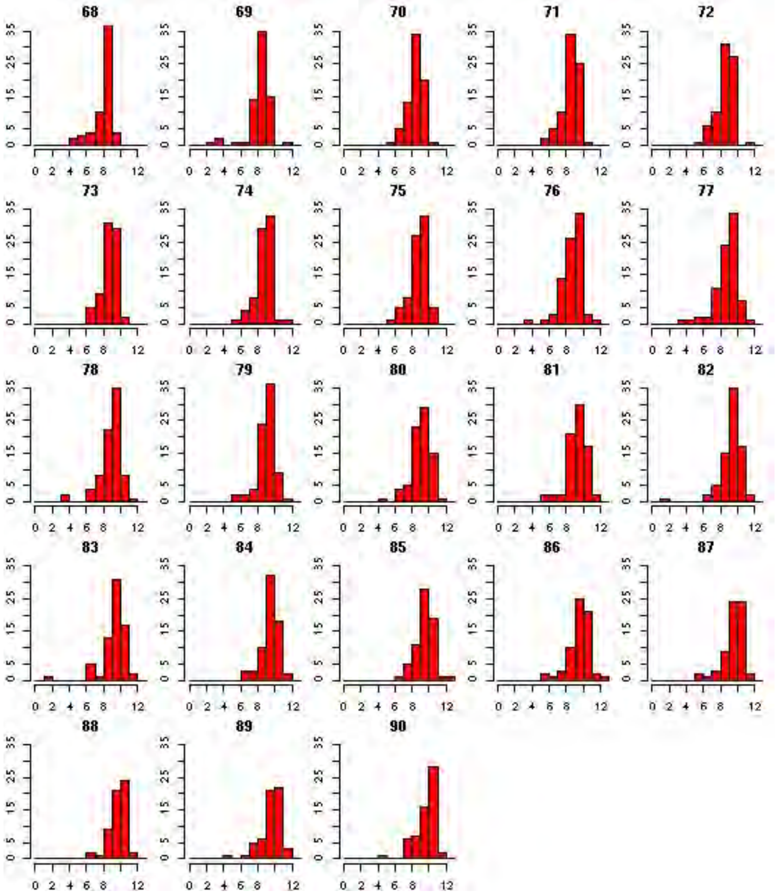

Figure 5: PSID Data: Yearly histograms of log(income). Incomes are increasing with time. The distributions in each year are roughly symmetric, and unimodal. There are several extreme points: low incomes in 69, 78, 82, 84.

Figure 5 displays the distributions of incomes in each year. The distributions are roughly symmetric and unimodal. There are some years where a few individuals have extremely low incomes. There's nothing surprising to be found from these plots.

Figure 6 displays the median incomes each year plotted using connected lines, and the medians calculated conditionally on several of the covariates. There is a strong increasing trend for household income by year. Despite the substantial variation around this trend there's no doubt that the increase is significant. There is very little difference in income for older vs younger head of households (top right plot). We do observe substantial differences in income for both gender and education (bottom row plots). Households headed by females have lower incomes for all years. Household headed by college educated people have higher incomes

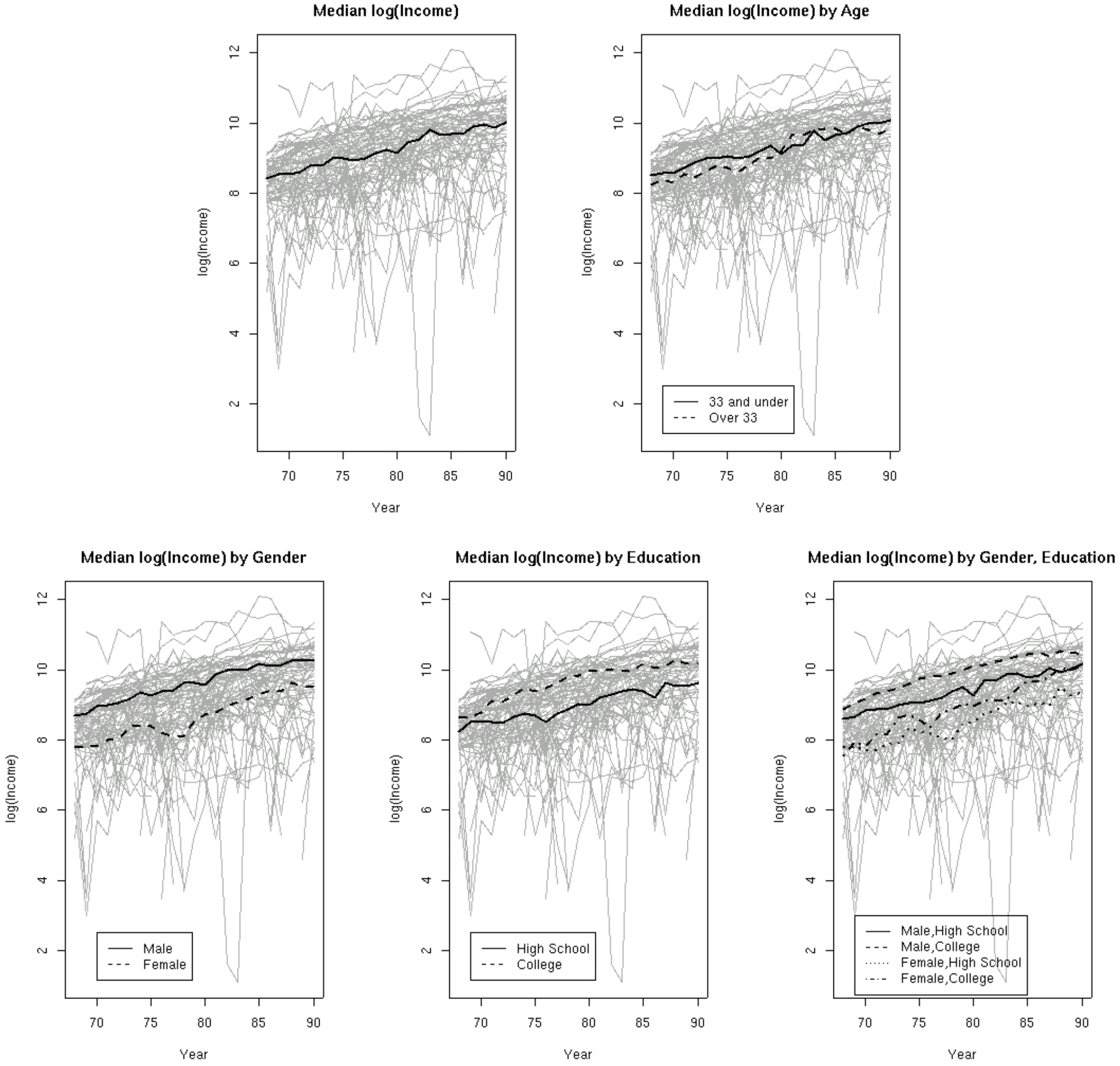

Figure 6: PSID Data: Medians over time plotted over individual profiles, and conditioned by several covariates. There is a strong upward trend in household income by year, although there is substantial variability around this trend. There is little difference in trend for age, substantial difference in household income by gender and education, and a clear interaction effect in gender and education.

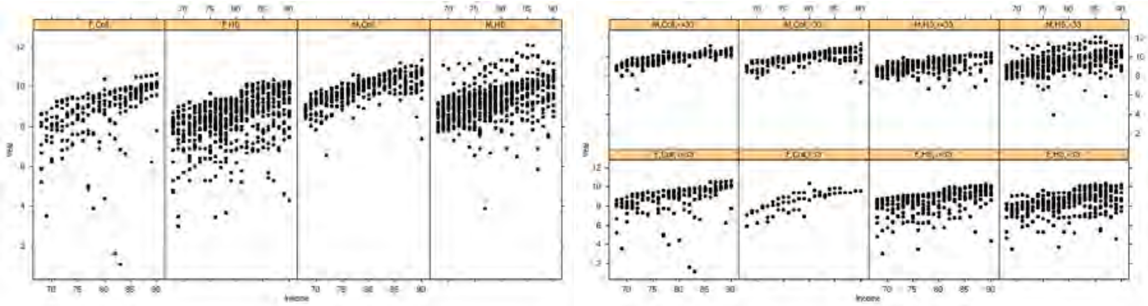

Figure 7: PSID Data: Conditional scatterplots for gender and education (left) and gender, education and age (right).

than those headed by high school educated people, although the difference was almost non-existent at the start of the study period. There is an interesting interaction between gender and education (bottom right plot). The trend lines for males, college and high school educated, are roughly parallel to each other for most of the years, with households headed by a college educated male enjoying higher income. For females the trend is different. Households headed by college educated females see great strides in income equity over the years to be at least equal in relation to households headed by high schooled educated males, and gaining on households headed by college educated males. Unfortunately a similar pattern is not seen for household headed by high school educated women.

The scatterplots supporting the trend plots are shown in Figure 7. At left are the four plots of income by year for gender and education. Households headed by college educated males show very little variation in income. The variation is greatest in the incomes of female headed households, and generally in households headed by high school educated people. On the later years there does seem to be a reduction in income variability for female college educated headed households. The plots at right show the full breakdown of plots by the three covariates, gender, education and age. Households headed by college educated males have consistently high incomes regardless of age. Households headed by high school educated males show more variability in incomes for older males. There are very few households headed by older college educated females. More of the lower incomes arise is female headed households.

From these plots we've learned these things about the mean trend that Faraway didn't: *** this needs more explanation ***

1. On average income is increasing over time. In Faraway's method, constant function $\hat{\beta}_0(t)$ does not correspond to mean income.

2. There is an interaction between gender and education.

3. The decomposition of variation in the data with respect to covariates. *** what?

## 3.3 Example 3: Iowa Youth and Families Project (IYFP)

To illustrate the approach to examining multivariate dependencies we use data from the Iowa Youth and Families Project (IYFP), a longitudinal study of 451 rural families that began in 1989 and continues at regular intervals. The IYFP began in the wake of the financial "farm crisis" that shocked rural America in the mid-to-late 1980s. The purpose of the study was to examine the effects of sudden economic hardship on families and to detect covariates that would help predict why some families survived economic hardship without major disruption while others experienced disturbing outcomes such as divorce and adolescent behavior problems. Data in this paper is part of 11-year follow up of 451 families from the eight counties of north central Iowa. Measurements are taken at 8 time points: 1989,1990, 1991, 1992, 1994, 1995, 1997, 1999. Targets were selected to be seventh graders, with an average age of 12.7 years at the start of the project, with two married biological parents and with a sibling within four years of age. There are 215 male and 236 female subjects in the study. Graduation from high school coincides with 1994. The response variables, anxiety, hostility and depression, were measured by using a symptom check list. Original responses were ordinal, ranging from 1 to 5, with 1 corresponding to no problem. Due to skeweness, responses are transformed into logarithm scale with base 10. Some of the symptoms for distress include nervousness or shakiness, an urge to break things, or feeling low in energy. The project is documented in Elder & Conger (2000).

*How do we understand the relationship between the three responses in relation to the temporal context?* The approach is to start with univariate plots and progress to multivariate plots. Plot each response separately to examine the mean trend in relation to time. Plot each pair of responses with time represented in the plot using line segments. The three responses are examined using 3D rotation with time represented with line segments. Figure 8 illustrates the approach. The univariate plots for depression are shown at top. On average depression is fairly flat during the high school years then dips after graduation. The middle row of plots display bivariate plots of depression and hostility. At left are the means for each year, with sequential years being connected by line segments. The plots to the right show the scatterplots of the two responses conditioned by year. Depression and hostility both decrease with time, but both stay relatively flat during high school. The distribution of points for each year shows that depression and hostility reporting are moderately positively associated, a high depression score is usually associate with a high hostility score.

The bottom rows of plots in Figure 8 are snapshots from 3D rotations of the three responses. Motion is one of the basic visual tools that can be used to examine multivariate distributions. Tours (Asimov 1985, Cook, Buja, Cabrera & Hurley 1995, Cook & Buja 1997, Wegman 1991, Tierney 1991, Wegman & Carr 1993, Wegman 2003, Wegman & Solka 2002) are created by generating a sequence of low-dimensional projections of a high-dimensional space. In each of the bottom two rows 2D projections of the 3D response space are shown. Each row shows the same 2D projection of different parts of the data: the mean trend at left, and the full scatterplot of data with each year highlighted in the subsequent eight plots. The data is a 3-column matrix, $\mathbf{Y}_{451\times 8,3}$ where column 1 is anxiety, column 2 is hostility and column 3 is depression. This data matrix is projected into 2D using the projection matrix:

$$\begin{bmatrix} -0.024 & 0.711 \\ 1.000 & 0.017 \\ 0.000 & 0.703 \end{bmatrix}$$

to give the bottom row of plots. The plot of the means is generated by multiplying the matrix of yearly means, $\bar{\mathbf{Y}}_{8,3}$, by this projection matrix. Sequential time points are joined by line segments. From viewing the rotations we learn that the dominant relationship is that the means are positively associated in time: on average if depression is high, then anxiety and hostility are high. Generally the three mean responses are high during school and then drop after graduation. From the conditional scatterplots it is also noticeable that the variance, or spread, of response values changes in time, the spread of values shrinks dramatically when adolescents leave school.

If depression is high, anxiety is high and hostility is high. The temporal trend is from high to low.

There is nothing surprising in this data, when looking at the 3D. For the most part I'd guess we wouldn't expect to see anything unusual here, because we have correlated measurements. If we were to see something odd, it may appear as a non-linear dependency, wouldn't expect to see clustering. Now there is a lot of variation in this data which is not captured by the mean trends, and it suggest that we may find some interesting profiles of individuals.

Mean trends by covariates....

Ragged time points, brushing on time still can be done....

One of the main questions to answer with this data is "Are there gender differences in reporting emotions?" Some analysis related to this question is reported in Ge, Conger & Elder (2001). Figures 9, 10 contain plots for exploring the mean trends conditional on gender. Distress levels for both genders decrease over time. Average anxiety and hostility scores for both gender are close to each other, slightly higher for females in some years. However, females clearly report higher depression scores compared to males. There is less difference in the baseline year. The univariate response scores suggest that females in this study are more likely to report distress than males. The bivariate relationships are interesting. Male and females have similar mean trends. However, there are some distinctive features. Anxiety, hostility and depression jointly decrease in time. There is little difference between males and females in anxiety and hostility. The big differences can be seen when depression is plotted against the other two responses: females tend to have an increase in depression while anxiety stays constant during high school, whereas males tend to report a decrease in both anxiety and depression early in high school, with a jump in depression as graduation approaches.

*** I included this: The difference between males and females does seem to be more in the negative diagonal which says bivariate relationship is important, a female is more likely to report higher depression but lower hostility and anxiety. This is not true with the anxiety and hostility, males are more likely to report marginally higher hostility than anxiety. (*** maybe skip this??? On average, for females, depression scores are likely to be higher than their hostility scores, but, for males, especially before graduation, exactly the opposite is observed. ) This information cannot be extracted from the previous marginal trends. *** does this need a separate plot?

Mean scores for males have larger variability in both responses compared to females. Nothing interesting to report was seen in the 3D plots of mean trend for gender.

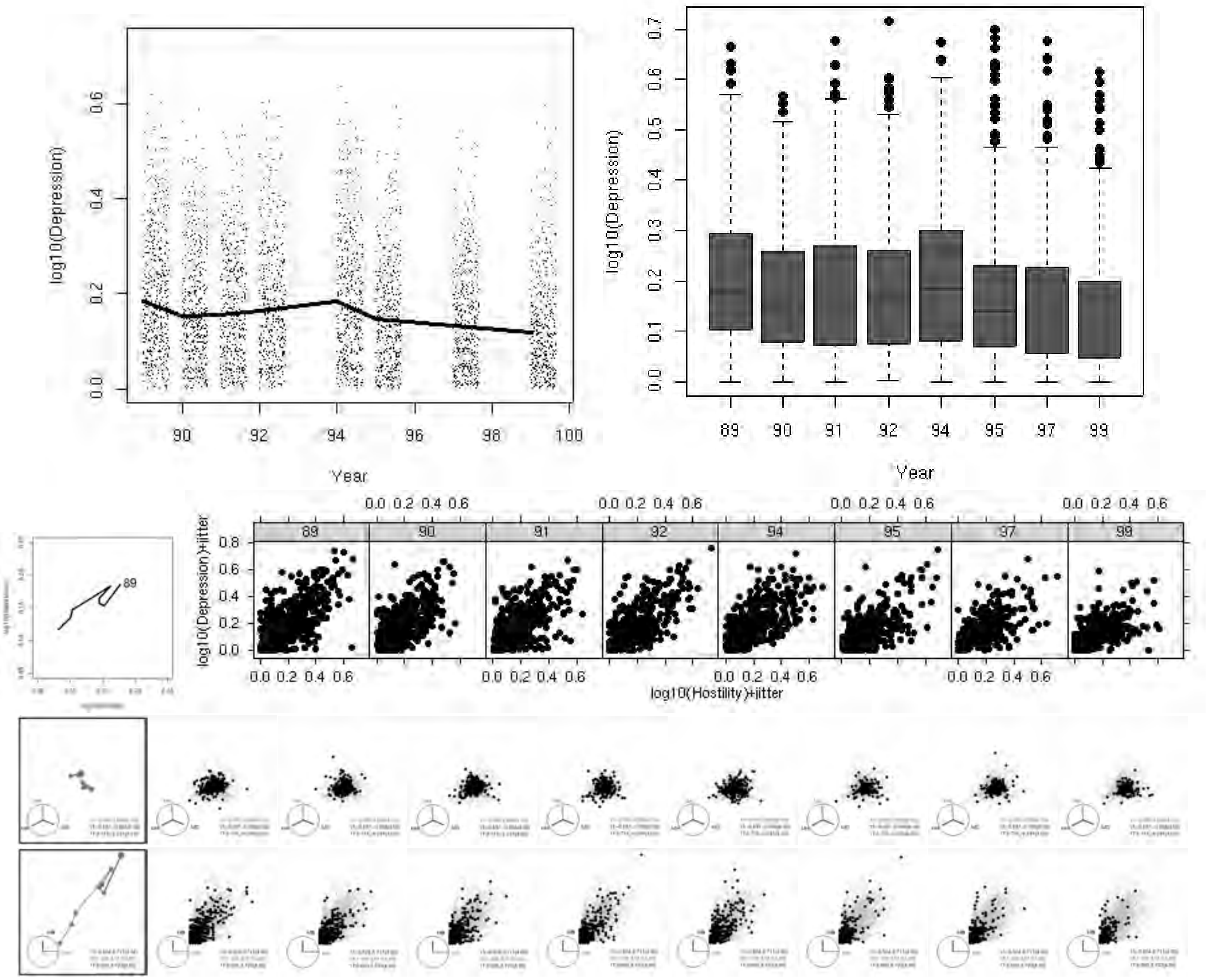

Figure 8: IYFP Data: Mean trends. (Top row) Mean trend for depression: stays relatively flat during high school then dips after graduation. (Middle row) Mean trend for depression and hostility: depression and hostility both decrease with time, but both stay relatively flat during high school. The distribution of points for each year show that depression and hostility reporting are moderately positively associated, a high depression score is usually associate with a high hostility score. (Bottom two rows) These are selected views from studying mean trend for anxiety, hostility and depression. The strongest dependency amongst the three responses, is that on average they are moderately positively associated in time, as the student moves through school and beyond the scores for all three decrease. During the high school years the responses are similar, and drop after school. Its also noticeable that the variance, or spread, of response values changes in time, the spread of values shrinks dramatically when adolescents leave school. With the exception of a few, most adolescents report feeling less anxious, less depressed and less hostile.

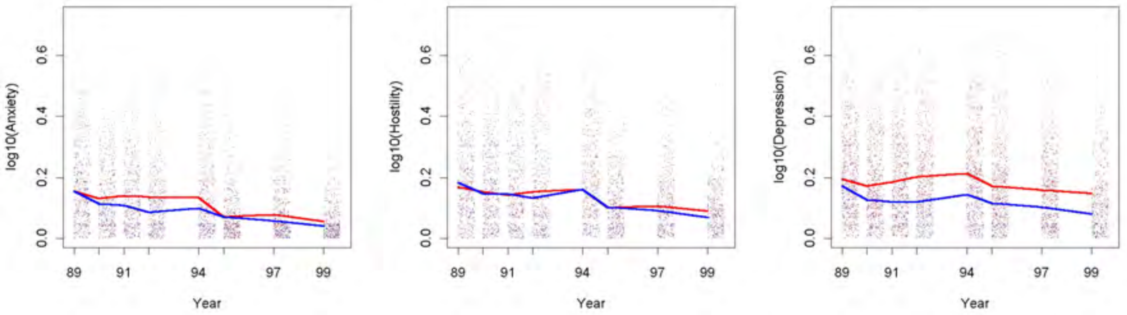

Figure 9: IYFP Data: Scatterplots of anxiety, hostility, and depression by time for two gender groups. Females are brushed into red, and males are into blue. Females are more likely to report distress, especially depression.

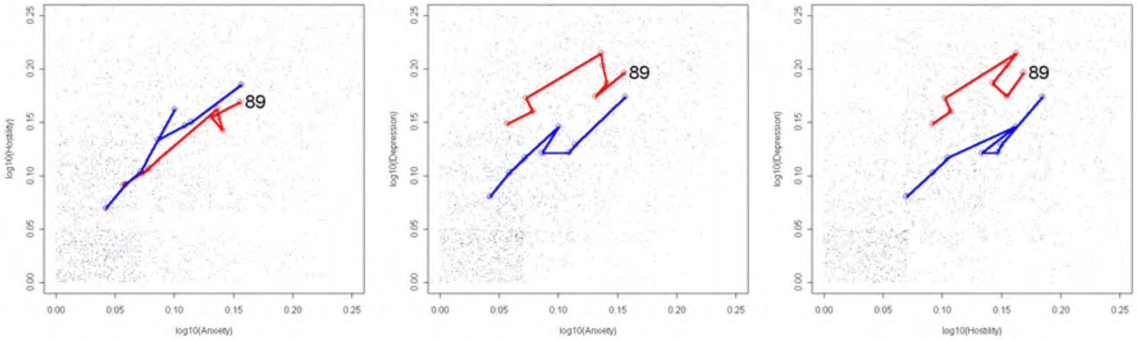

Figure 10: IYFP Data: Scatterplots of hostility by anxiety; depression by anxiety; and depression by hostility for two gender groups. Mean at the first time point is marked as ’89’. Females are brushed into red, and males are into blue. There is little difference between males and females in anxiety and hostility. Females tend to have an increase in depression while anxiety stays constant during high school, whereas males tend to report a decrease in both anxiety and depression early in high school, with a jump in depression as graduation approaches. Mean scores for males have larger variability in both responses compared to females.

### 3.4 General principles

Several general principles emerge from initial exploration of longitudinal data:

- The primary intention is to understand the relationship between responses and the temporal context. Thus the basic plot is response(s) against time.

- Plotting all the individual traces on the one plot can produce an unreadable plot. The purpose is to digest the mean trend, so that in general, it may be more useful to plot the points only.

- Along with a representation of the mean trend, a representation of the variation is important. A scatter plot of the points overlaid by the trend representation is the simplest approach to assessing the variation around the trend that works for all types of longitudinal data. In some constrained types of longitudinal data it is possible to use boxplots to display the distribution or display confidence intervals at common time points.

- Use conditional plots for assessing the trend in relation to categorical covariates, or common time points.

- Use grand tours for assessing the multivariate time trend in the presence of multiple responses. Plots of joint distributions usually reveal patterns that plots of marginal distributions cannot.

- Plots of model estimates are no substitute for plots of data.

## 4 Individuals

The ability to study the individual is a defining characteristic of longitudinal data analysis. With a tangle of overplotted profiles this can be a daunting task. There are two approaches in common use: (1) sample the individuals to reduce the number of lines plotted (Diggle et al. 2002), and (2) show one individual at a time, animating over all individuals. Neither of these provides satisfying insights into individual patterns. With sampling there can be too much missing to find the interesting individuals. To make a successful animation there needs to be continuity from frame to frame, and the order in which individuals appear in a data set is unlikely to be ordered in a way that will produce continuity. Animations over individuals invariably produce quick flashes of radically differing profiles from frame to frame, allowing little chance for digesting any patterns. This section describes some alternative approaches to studying individuals.

### 4.1 Example 1: Wages

The purpose of studying the individual profiles is that we want to get a sense for the individual wage vs experience pattern as it differs from a common trend. On average more experience means more wages but is it usual that as an individual gets more experience that their wages will go up. How common is this? Or is it more typical that a persons wages will bounce around regardless of experience? Figure 11 displays profiles of individuals who are at the extremes in the data to some extent. We take a brush and select an observation on the extremes of the wages and experience plot, and their record is highlighted. We can observe quite a range in the individual differences. Two people (top two at left) with extremely high wages with long experience both received substantial late jumps in wages. The first person had a quick rise in wages in their early experience, and then a sharp drop, oscillated around an average wage for several years and then jumped substantially at 12 years experience. Another person (third plot) with high wage, at 7-8 years of experience, took a dramatic drop in wages in the later years of experience. The people with low wages later in experience also had quite dramatically different patterns. One person (fifth plot) began their career with relatively high wages early on, and their wages have continued to drop. Another person (sixth plot) has consistently earned low wages despite more experience. The individual wage vs experience pattern is really quite varied!

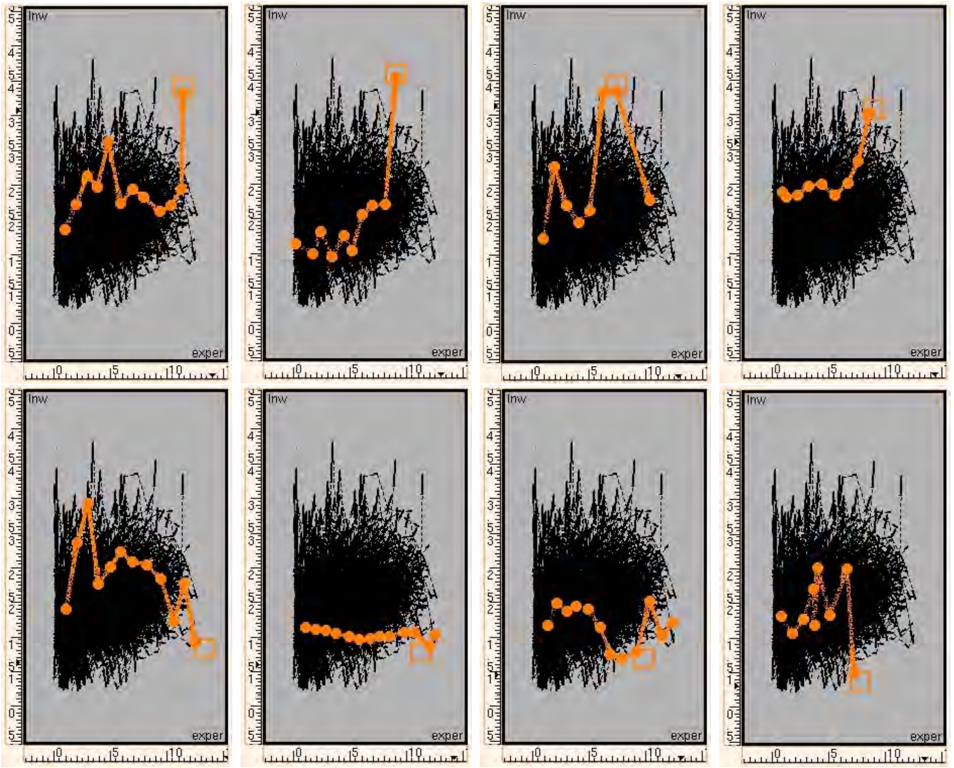

Figure 11: Extreme values in wages and experience are highlighted revealing several interesting individual profiles: large jumps and dips late in experience, early peaks and then drops, constant wages.

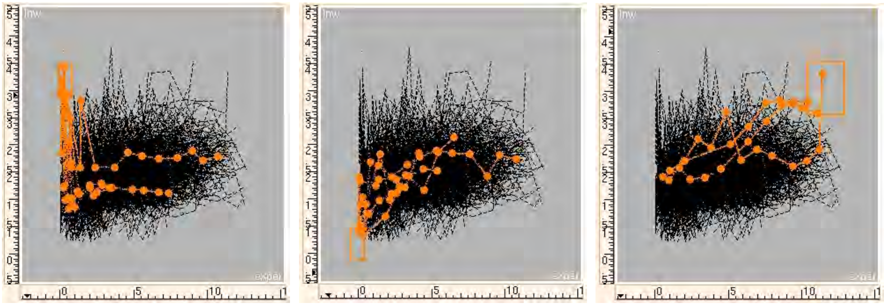

Figure 12: Early high/low earners, late high earners with experience.

Figure 12 shows several typical cases we might wish to explore. *How do people with high wages early fare with more experience? How do people with lots of experience and high wages get there?* If a person starts off with a high wage how do they fare with more experience. The top early wage earners are highlighted in the left plot, and only two of these people are retained for the full study, and their wages end up just at moderate levels. The middle plot highlights individuals whose wages started off very low. Again only two of these people were retained for the full study and their incomes did increase to be moderately high with more experience. The third plot highlights several individuals with high wages and more experience. It's interesting to see how they got there. All three started off with moderate wages. Two steadily increased their wages and the third person had a quite volatile wage history.

*Searching for Special Trends:* With what we've seen about the individual variability, a next stage is to search for particular types of patterns: the individuals that have the most volatility in their wages, the individuals who's wages steadily increase or decrease. That is, we want to construct numerical summaries for each individual that will identify particular types of patterns. As an example, we've created two new variables to measure volatility and smoothness of trend. The first measure (`SDWages`) is that standard deviation of values for each individual, measuring overall variability in the person's wages. The second measure (`UpWages`) is about smoothness of trend. It examines the difference between pairs of measurements, calculates the variance and changes the sign to be negative if the differences are mostly negative. When these two variables are incorporated into the analysis, they help identify individuals with particular patterns in wage history: volatile changes, steady increases, or declines. Figure 13 shows a few of these profiles. The first plot shows a person who has had an extremely volatile wage history. The second plot shows two high earning people that have had dramatic increases in wages as they have gained experience. The third plot shows three people with more steady increases in wages as they have become more experienced. The fourth plot shows many individuals who have had very little change in their wages with increasing experience. The fifth plot shows a person who has had their wage steadily decline despite increasing experience. The sixth plot shows an individual who's had some dramatic changes in wages with increasing experience, some volatility early in their career, and then a period of high earning with moderate experience to be declining in wages with more experience.

*So what have we learned about wages and experience?* A brief summary of what we have learned about the general patterns and individual variation from this data is:

- On average wages increase with experience, but there is a lot of variation in wages depending on experience.
- The amount of increase differs according to race and educational experience in the later years of experience.
- The individual patterns are dramatically different. We found several individuals have extremely volatile wages in relation to experience, several who have very constant wages despite more experience, and several people who saw a decline in their wages as they gained more experience in the workforce.

### 4.2 Example 2: PSID

The PSID data has some interesting household trends which differ substantially from the mean trend. Figure 14 shows the profiles for two interesting households. The highest earning household is headed by an older high school educated male. The household with an extremely low income is headed by a college educated young female. In general this household has a more average income.

Investigating the individuals illustrates substantial deviation from the mean trend. Many of the highest earning household are headed by high school educated people. Most of the lowest earning households are headed by females. Some female-headed households earn better than average. Some households have very steady incomes, some households have substantial year to year variability in income.

### 4.3 Example 3: IYFP

This section illustrates observing the individual patterns in 3D. Figure 15 contains several plots where interesting individuals are highlighted. Each of these individuals differs from the trend: they each respond

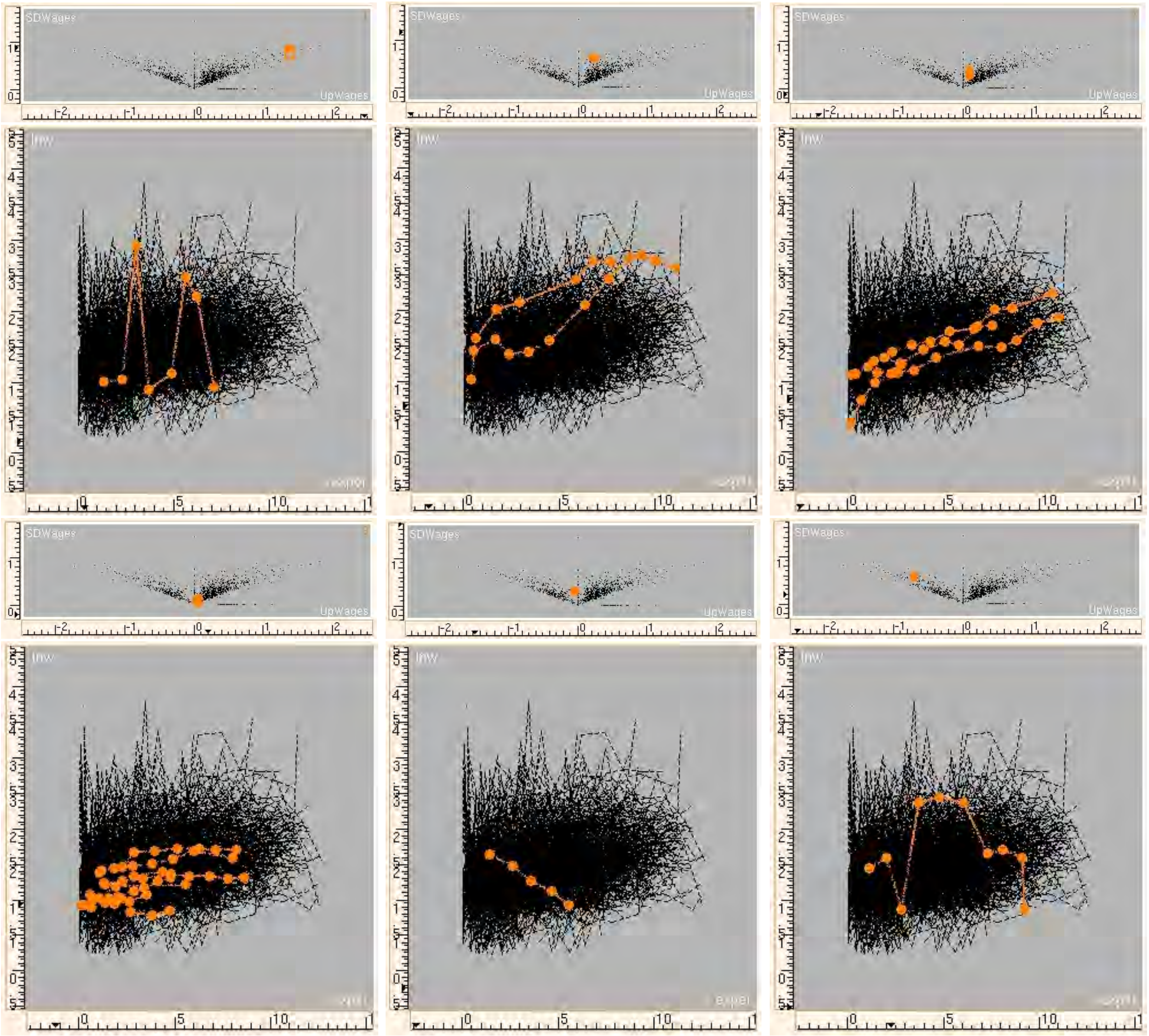

Figure 13: Special patterns: with some quick calculations to create indicators for particular types of structure we can find individuals with volatile wage histories and those with steady increases or declines in wages.

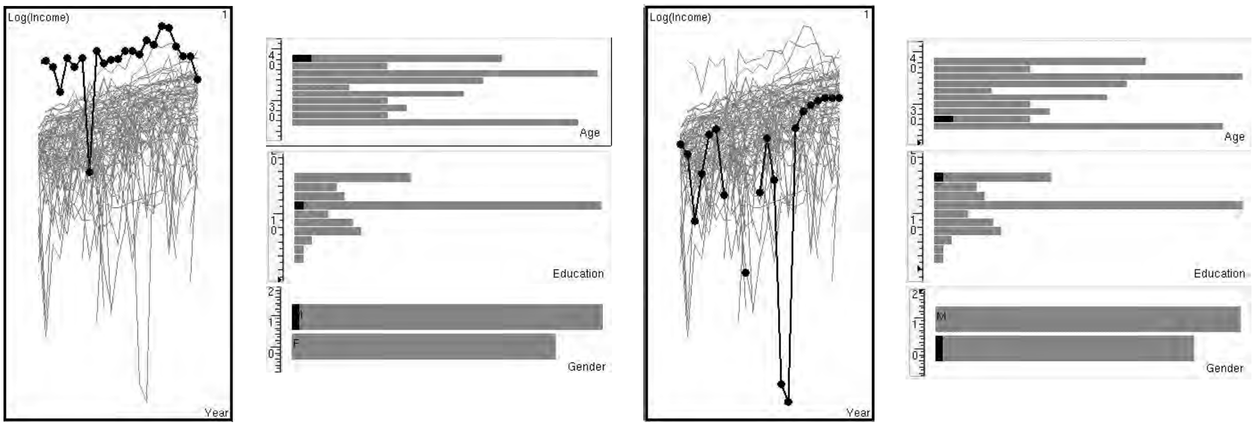

Figure 14: PSID Data: Looking for individuals. (Left) The highest earning household is headed by a male, age near 40 at the start of the study, with only a high school education, and in one year the household income took a serious dive. (Right) The household with an extremely low income in one year is headed by a college-educated female who was under 30 at the start of the study. For the most part, though, this household has closer to average income.

quite differently on the three responses at each interview, and there reported distress is dramatically different from year to year. The individual highlighted in the top row of plots has scores for the three variables running the gamut of high to low. The middle top plot is simple, anxiety plotted against depression, and the top right plot is effectively anxiety plotted against hostility. In responses on anxiety and depression, this individual answers similarly on all but two of the surveys, seen because the values are lying close to the center line along the point scatter. In responses on anxiety and hostility they answer quite differently, seen because these scores zig-zag around the plot space.

*So what have we learned about distress levels of Iowa youth?* From this data, we have seen:

- On average all three distress scores are high during high school and drop after graduation.
- There is gender differences observed. On average females report higher depression than males; and females are more likely to report higher depression scores compared to their hostility scores.
- On average subjects respond very similar on the three distress measurements, but there are individual trends quite different.

## 4.4 General principles

What have we learned about the data sets, and how does this generalize?

- Often the overall trend explains very little of the variation in the data. The trends in relation to covariates also can explain very little of the variation in the data. The variation from individual to individual can be immense, and this is of great interest to explore.
- Individual temporal trends can be dramatically different from the mean trend, and interactive graphics allows us to identify these individuals.
- There is a need for simple diagnostic statistics to explore and detect specific types of patterns in individuals.
- Dynamic plots of multiple responses enable individuals with extreme patterns to be detected, and then digested using profile plots of each response against time.
- Studying the individual patterns is similar to cluster analysis. We would like to build a catalog of particular types of profiles, which could be viewed as clusters.

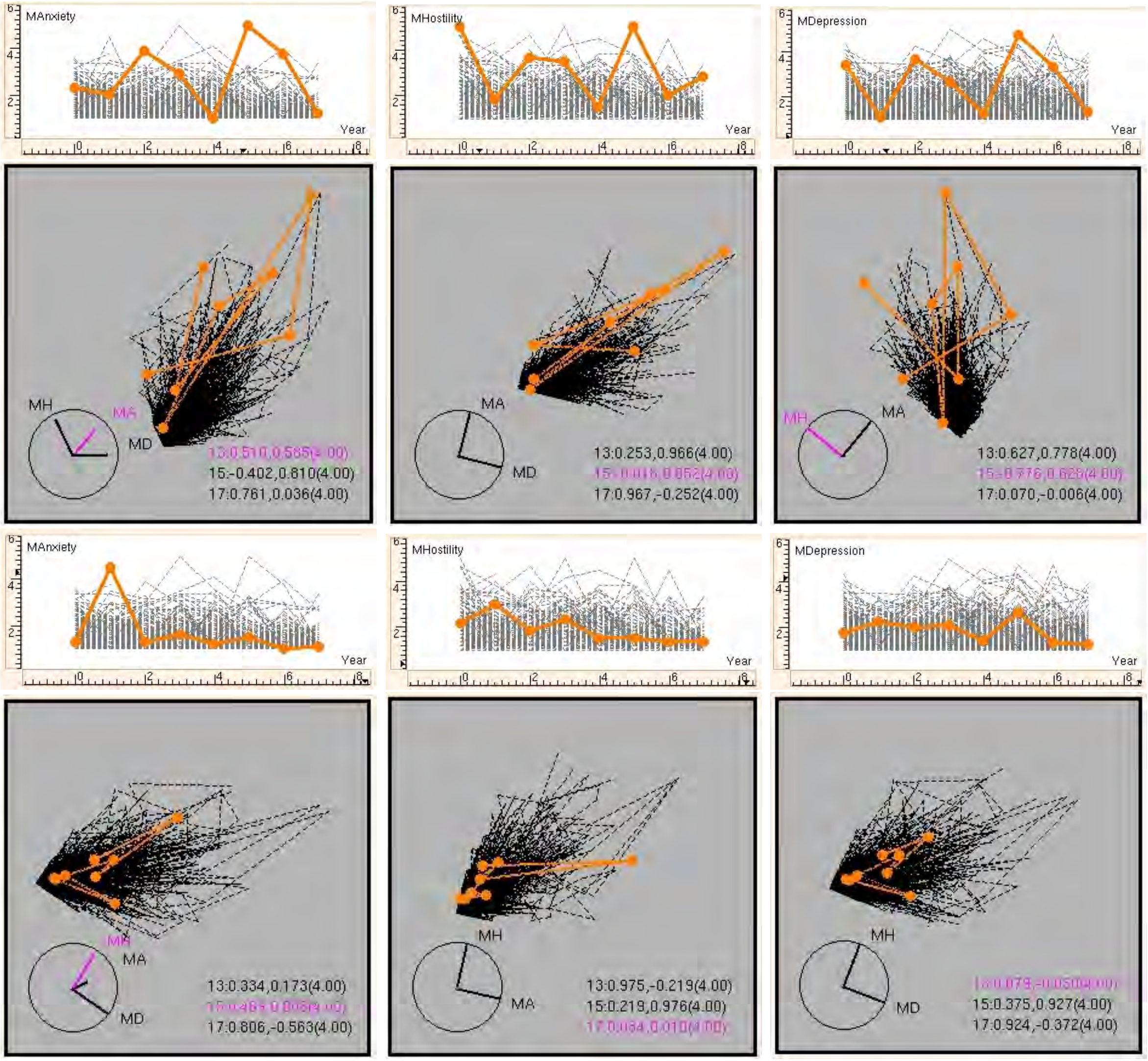

Figure 15: This is a part of the Iowa Youth and Family Project data: the three response variables measured at each survey, anxiety, hostility and depression, shown in a tour. Three tour projections of the data are shown: the axes are shown at lower left, and projection coefficients are given lower right. One particularly strongly responding individual is highlighted in the top row of plots. This individual has responded as being very anxious, hostile and depressed in some surveys and not at all in other surveys. The individual almost always answers similarly on anxiety and depression (top middle plot) but often reports quite different in anxiety and hostility (top third plot). For the most part it seems that individuals respond similarly on the three responses: that if they are feeling depressed they are also, anxious and hostile. There are a few that report differently about the three responses. One such individual is highlighted in the bottom row of plots. This individual reports being highly anxious once early in high school and once being highly depressed later in the study period.

# 5 Summary and Discussion

This paper has described truly graphical approaches providing methods that can be used for even the most difficult longitudinal data. New methods for analyzing multivariate longitudinal data are an important contribution to the study of our society and the environment we live in. Modeling longitudinal data remains a problem. The models in Diggle et al. (2002) and Faraway (1999) are useful for only a limited set of longitudinal data, when individuals are all measured at the same time. Singer & Willett (2003) describe mixed linear models for more complex, ragged time longitudinal data, but because these models are linear they may miss subtle but important nuances in the data. More broadly, modeling may ride rough-shod over a wealth of information about intricate individual variability.

*** Broader impacts... microarray data...

The initial explorations of some of the examples were conducted using XGobi, an older version of GGobi. Brushing covariates in XGobi automatically resulted in smooth curves for each group. However, the XGobi codes to calculate smooth curves are replicated in R/S+, thus this redundancy was not continued into GGobi code. Rather the new approach is to use the RGGobi package to calculate smooth curves using R and communicate this to GGobi. Given this new framework for computing, users can experiment and explore different smoothing methods to study the mean trend in longitudinal data. The special packages of R combined with the interactive tools of GGobi empowers the user to analyse even such complex data structures. With these tools, it is possible to reveal the unexpected, to explore the interaction between responses and covariates, to observe the individual variations, and understand structure in multiple dimensions.

## Acknowledgements

The authors would like to thank Dr. Frederick O. Lorenz, and Dr. Heike Hofmann for their valuable discussions and suggestions. We would also like to thank Dr. Rand Conger for making the IYFP data available, and Dr. Becky Burzette for her help in data collection.

GGobi is freely available from `http://www.ggobi.org`. Deborah F. Swayne is the primary developer of GGobi and Duncan Temple Lang has developed RGGobi and other advanced ways to communicate to and from GGobi.

## Appendix

Setting up data for an exploratory analysis...

This is a sample of the way the wages data is described using xml.

```
<?xml version="1.0"?>
<!DOCTYPE ggobidata SYSTEM "ggobi.dtd">

<ggobidata count="3">
<data name="wages">
<description>
This is XML created by GGobi
</description>
<variables count="15">
  <categoricalvariable name="id" nickname="id">
    <levels count="888">
      <level value="0"> 31  </level>
      <level value="1"> 36  </level>
....
      <level value="887"> 12543  </level>
    </levels>
  </categoricalvariable>
  <realvariable name="lnw" nickname="ln"/>
  <realvariable name="exper" nickname="ex"/>
....
</variables>
<records count="6402" glyph=". 1" color="8">
<record id="1" label="31">
0 1.491 0.015 1 0.015 0 1 8 -1 3.215 -3.785 0 3.215 0 3.215
</record>
<record id="2" label="31">
0 1.433 0.715 1 0.715 0 1 8 -1 3.215 -3.785 0 3.215 0 3.215
</record>
....
</records>
</data>
<data name="edges">
<description>
Profiles for each individual.
</description>
<variables count="5">
  <categoricalvariable name="id" nickname="id">
    <levels count="888">
      <level value="0"> 31  </level>
      <level value="1"> 36  </level>
....
     <level value="887"> 12543  </level>
    </levels>
  </categoricalvariable>
....
</variables>
<records count="5514" glyph=". 1" color="8">
<record source="1" destination="2" label="31">
0 1 0 1 8
</record>
<record source="2" destination="3" label="31">
0 1 0 1 8
</record>
....
</records>
</data>
<data name="demog">
<description>
This the demographic information associated with each individual.
</description>
<variables count="5">
  <categoricalvariable name="id" nickname="id">
    <levels count="888">
      <level value="0"> 31  </level>
      <level value="1"> 36  </level>
....
      <level value="887"> 12543  </level>
    </levels>
  </categoricalvariable>
....
```

```
</variables>
<records count="888" glyph=". 1" color="8">
<record label="31">
0 1 0 1 8
</record>
<record label="36">
1 1 0 0 9
</record>
</records>
</data>
</ggobidata>
```

# References

Asimov, D. (1985), ‘The Grand Tour: A Tool for Viewing Multidimensional Data’, *SIAM Journal of Scientific and Statistical Computing* **6**(1), 128–143.

Buja, A. (1999), ‘Inference for Data Visualization’, Invited Talk, JSM 1999, slides available at `http://www-stat.wharton.upenn.edu/∼buja/PAPERS/jsm99.ps.gz`.

Cleveland, W. S. (1979), ‘Robust Localy Weighted Regression and Smoothing Scatterplots’, *Journal of American Statistics Association* **74**, 829–836.

Cook, D. & Buja, A. (1997), ‘Manual Controls For High-Dimensional Data Projections’, *Journal of Computational and Graphical Statistics* **6**(4), 464–480. Also see `www.public.iastate.edu/∼dicook/research/papers/manip.html`.

Cook, D., Buja, A., Cabrera, J. & Hurley, C. (1995), ‘Grand Tour and Projection Pursuit’, *Journal of Computational and Graphical Statistics* **4**(3), 155–172.

Crowder, M. J. & Hand, D. J. (1990), *Analysis of Repeated Measures*, Chapman and Hall, London.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y. & Zeger, S. L. (2002), *Analysis of Longitudinal Data*, Oxford University Press, Oxford, UK.

Elder, J. J. & Conger, R. D. (2000), *Children of the Land*, The University of Chicago Press, Chicago.

Faraway, J. J. (1999), ‘A Graphical Method of Exploring the Mean Structure in Longitudinal Data Analysis’, *Journal of Computational and Graphical Statistics* **8**(1), 60–68.

Ge, X., Conger, R. D. & Elder, G. H. (2001), ‘Pubertal Transition, Stressful Life Events, and the Emergence of Gender Differences in Adolescent Depressive Symptoms’, *Developmental Psychology* **37**(3), 404–417.

Good, P. (1993), *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer-Verlag, New York.

Inselberg, A. (1985), ‘The Plane with Parallel Coordinates’, *The Visual Computer* **1**, 69–91.

Koschat, M. A. & Swayne, D. F. (1996), ‘Interactive graphical methods in the analysis of customer panel data (with discussion)’, *Journal of Business and Economic Statistics* **14**(1), 113–132.

Singer, J. D. & Willett, J. B. (2003), *Applied Longitudinal Data Analysis*, Oxford University Press, Oxford, UK.

Sutherland, P., Rossini, A., Lumley, T., Lewin-Koh, N., Dickerson, J., Cox, Z. & Cook, D. (2000), ‘Orca: A Visualization Toolkit for High-Dimensional Data’, *Journal of Computational and Graphical Statistics* **9**(3), 509–529.

Swayne, D. F. & Klinke, S. (1998), ‘Editorial commentary’, *Computational Statistics: Special Issue on The Use of Interactive Graphics.*

Tierney, L. (1991), *LispStat: An Object-Orientated Environment for Statistical Computing and Dynamic Graphics*, Wiley, New York, NY.

Unwin, A., Hofmann, H. & Wilhelm, A. (2002), ‘Direct Manipulation Graphics for Data Mining’, *Journal of Image and Graphics* **2**(1), 49–65.

Wegman, E. (1990), ‘Hyperdimensional Data Analysis Using Parallel Coordinates’, *Journal of American Statistics Association* **85**, 664–675.

Wegman, E. J. (1991), ‘The grand tour in k-dimensions’, *Computing Science and Statistics: Proceedings of the 22nd Symposium on the Interface* pp. 127–136.

Wegman, E. J. (2003), ‘Visual data mining’, *Statistics in Medicine* **22**, 1383–1397+10 color plates.

Wegman, E. J. & Carr, D. B. (1993), Statistical Graphics and Visualization, *in* C. R. Rao, ed., ‘Handbook of Statistics, Vol. 9’, Elsevier Science Publishers, Amsterdam, pp. 857–958.

Wegman, E. J. & Solka, J. L. (2002), ‘On some mathematics for visualizing high dimensional data’, *Sanhkya (A)* **64**, 429–452.