Chapter 2

Graphical Displays of Longitudinal Data

Displaying longitudinal data can present a greater challenge than the analysis of such data. This chapter focuses on the graphical display of repeated measures data - that is when an outcome (such as CD4 count) is measured repeatedly over time on an individual. The challenge is to highlight potentially meaningful patterns among messy and abundant data. Because the data is longitudinal, interest will often focus on trends in outcomes over time. However, other relationships are also of interest (such as changes in outcomes versus changes in explanatory variables). The optimal graph will be a function of the question being addressed, and thus there is no universally best plot to display longitudinal data. This chapter will provide some techniques that can help in specific circumstances and hopefully will provide at least a few general principles that can guide the reader's selection.

2.1 Plotting Raw Data (Outcomes) vs. Time

We will begin by examining CD4 measured longitudinally on HIV positive subjects from the time they begin highly active anti-retroviral therapy (HAART). The simplest plot is to compare the response through time of the different subjects. However, Figure 2.1 illustrates one of the challenges of displaying (abundant) longitudinal data - there is simply too much data to put on one plot.

One solution is to simply select a subset of the individuals to plot; these individuals can be selected randomly or systematically based on characteristics of the subject. First of all,

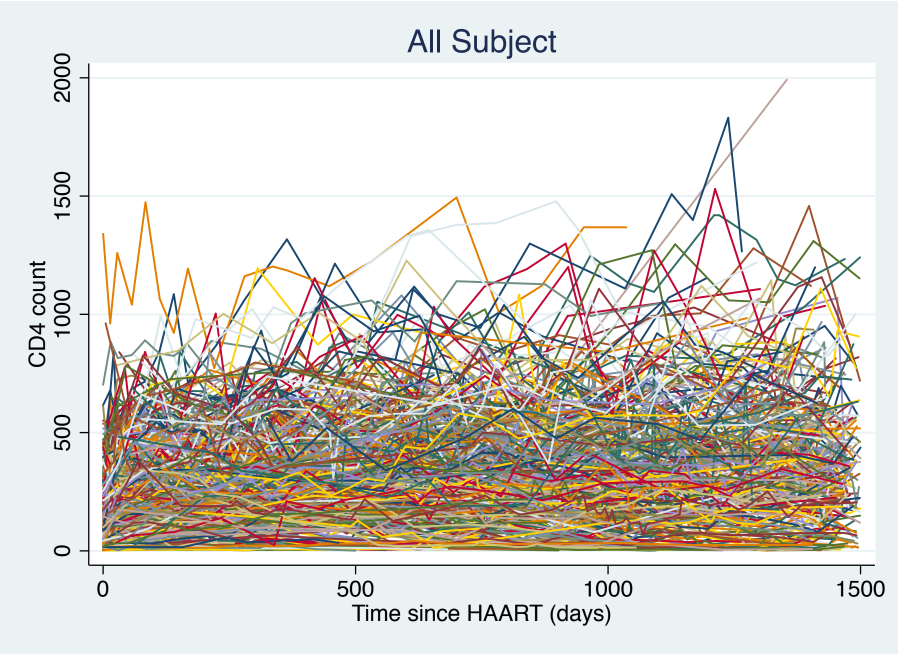


Figure 2.1: CD4 count versus time for all subjects

consider randomly selecting 5 percent of subjects and plotting their CD4 counts. There are various ways to algorithmically select subjects depending on the statistical package used; we provide an example using STATA. The variables used are the subject id, CD4 count (Y_{ij} for subject i, measurement j) and time (called `etime`, which we will refer to as T_{ij}). The data are in long format (see Table 1.1 for an excerpt of this data).

For all the plots, we provide STATA code on the website. The first such plot is for a random sample of 5 percent of the subjects and an overlay of their CD4 counts versus time (days). Figure 2.2 shows the resulting plot and some patterns now become more apparent; some subjects appear to hug the x-axis with consistently low CD4 counts throughout time, others appear to start low and respond positively to treatment, still others start high and stay there.

However, we might better characterize the distribution of responses by choosing among the subjects more systematically. Specifically, by randomly selecting subjects, we do not guarantee an 'even' distribution with respect to some characteristic of the individual. For instance, we might want subjects that represent that are evenly distributed with respect to average(CD4) count, averaged over the history of the subject. To do so requires us to 1) rank the subjects based on their average CD4 count and 2) choose evenly spaced subjects

Figure 2.2: CD4 count versus time for randomly chosen subjects

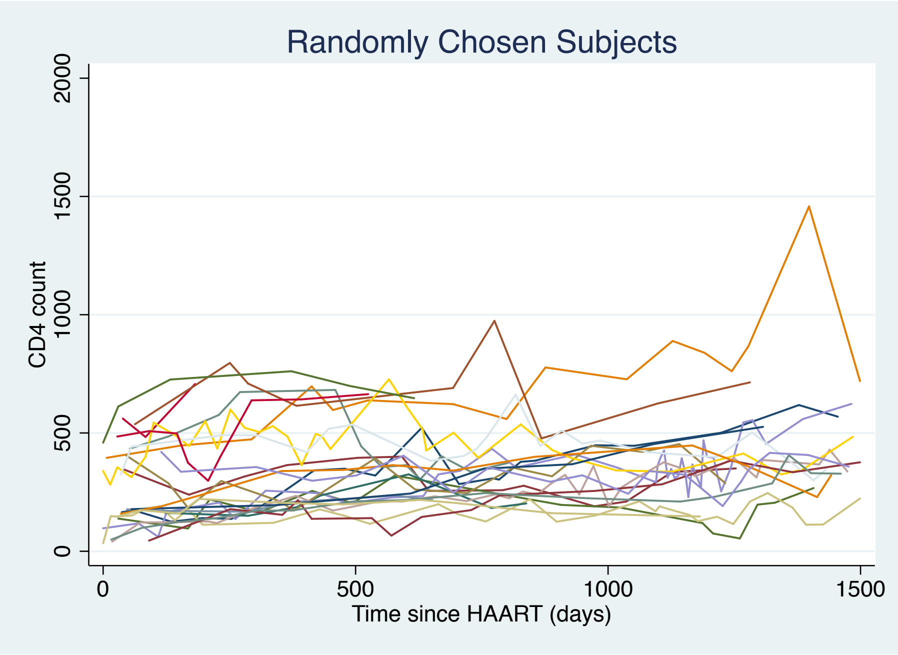
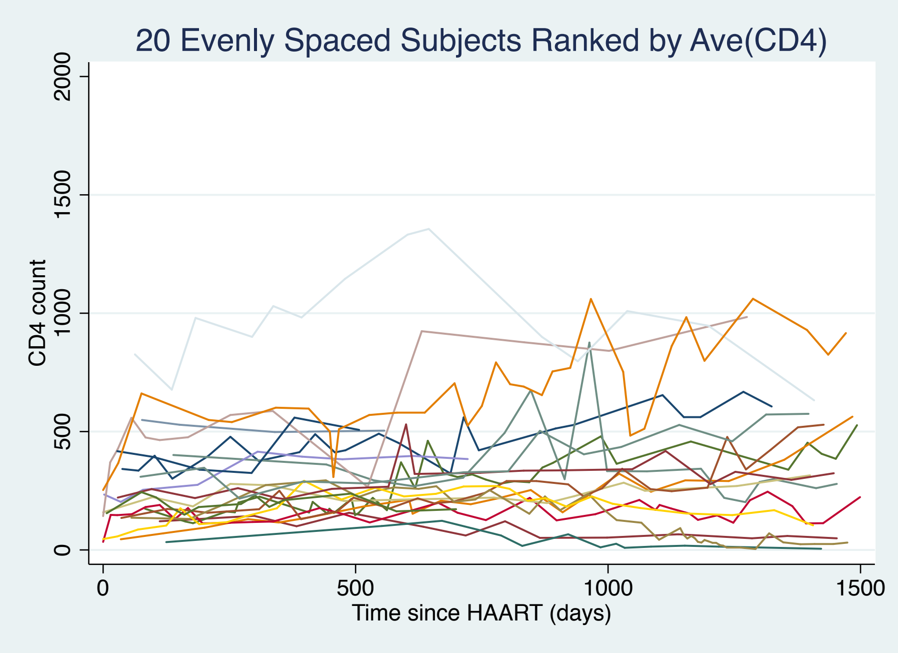


Figure 2.3: CD4 count versus time for subjects evenly distributed with respect to ave(CD4)



on this ranked list to plot their CD4 count (for instance, the 5th, 10th, 15th,..., id on ranked list). Below, we do so using STATA to plot curves ranked by average CD4 count. If you want to plot only k of the subjects and there are m subjects, then only take only every m/kth subject on list. For instance, if you want to plot 20 of the subjects and there are $m = 200$ subjects, take every 200/20 or 10th subject in ranked list and plot their CD4 count versus time. In the web supplementary code is an example for code that can be used for choosing subjects evenly distributed with respect to their average CD4 count.

The resulting plot in Figure 2.3 shows definite improvements over the random selection. One can now see more clearly a wider variety of CD4 patterns, from those with nearly no improvement (increase) in CD4 count after starting anti-retroviral therapy to those starting with relatively high CD4 counts and, on average, increasing over time. One can order and select the subjects to plot based on other summary statistics as well (for example, see the web supplementary code for subjects ordered by area under the curve).

2.2 Using Individual Level Models for Plot

Plots of raw, unprocessed data are always best for a first glance at data. However, unless the data is relatively noiseless and the underlying individual trajectories smooth, pre-processing can yield greater insight than the raw data plots. In addition, one may be willing to assume over-simplified statistical models to find patterns potentially obscured in raw data plots. For instance, assume that the mean trend for each subject is a simple linear trend (increase or decrease) in CD4 count with time - this kind of model underlies the random coefficient models that we will discuss later in Chapter 5. Specifically, for individual i, assume the mean of CD4 (Y_{ij}) at time, T_{ij} is,

$$E[Y_{ij} \mid T_{ij}] = b_{0i} + b_{1i}T_{ij}, \tag{2.1}$$

where (b_{0i}, b_{1i}) are the individuals i's intercept and slope. Although we will learn more sophisticated ways to estimate such individual-level coefficients in Chapter 5, one simple and legitimate method is to simply fit a separate linear regression of Y_{ij} on T_{ij} for every person, $i = 1, \ldots, m$; this results in m intercepts and slopes. In STATA, if one wishes to store the slopes and intercepts for each individual, as well as the predicted values for the linear models, then a simple program must be written (see appendix).

Finally, one can plot a sample of these trends, for instance subjects evenly distributed with respect to the magnitude of their slopes, much as we did for mean(CD4) above. Figure 2.4 shows a sample of two subjects with their respective linear trends plotted along with the raw data; Figure 2.5 has the estimated trends plotted $(\hat{b}_{0i} + \hat{b}_{1i}T_{ij})$ for 30 subjects evenly distributed with respect to their slope (equivalent to the example above using the mean(CD4) count). One can see that there appears to be a variety of subjects from those that start low and stay low, that is both $(\hat{b}_{0i}, \hat{b}_{1i})$ are close to 0. On the other extreme are subjects that have steep increases in CD4 count with time starting from relative high baseline CD4 counts. We can get a more global look at the distribution of slopes and intercepts by doing simple histograms of these across the subjects (Figure 2.6). We note that this anticipates random coefficient models, where the distribution of random coefficients within a population is made via mixed models, assuming certain distributions for these coefficients.

One can think of these linear fits as a form of smoothing. This form of smoothing, although certainly oversimplifying the true patterns with time, might benefit interpretation because patterns that are invisible given both the biological variability and measurement error start to become more apparent when some of the variance is reduced. As in all statistical modeling, there is a trade-off between bias and variance. We discuss now a method that assumes the pattern is ”smooth” without assuming it is linear.

Figure 2.4: The raw data and best fitting linear trends for two subjects

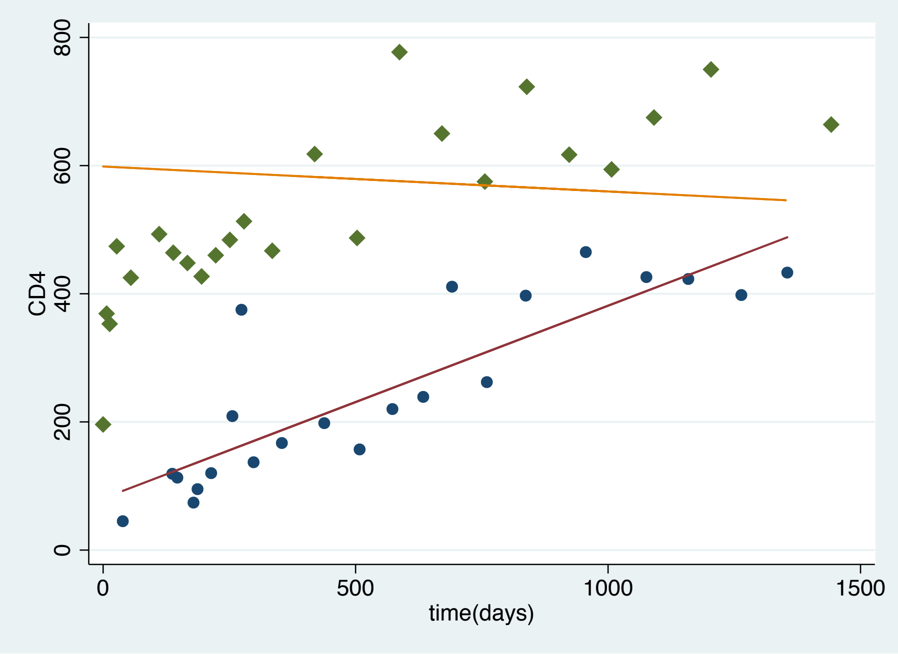


Figure 2.5: A random sample of subjects with their estimated linear trends plotted

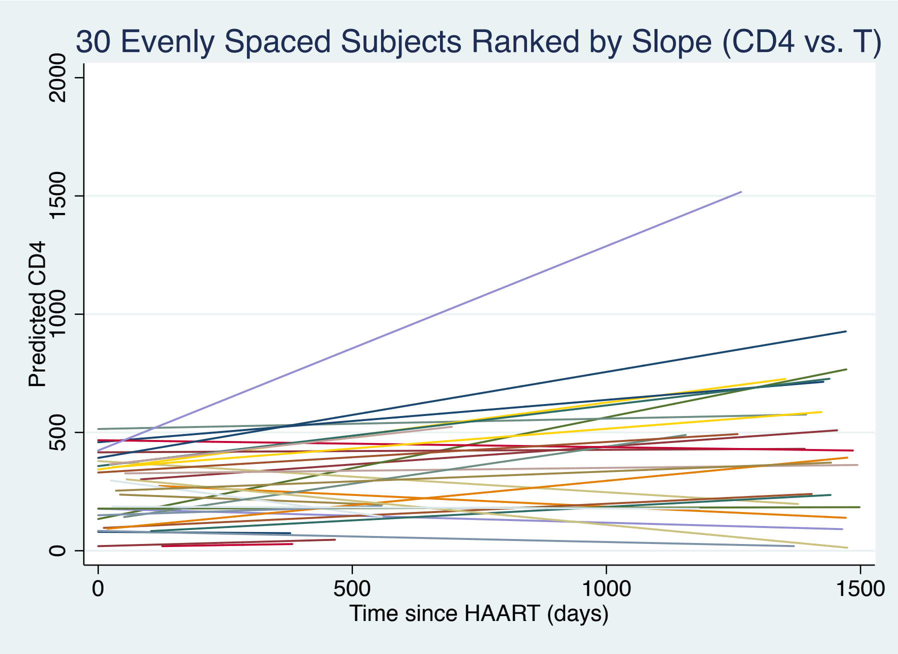
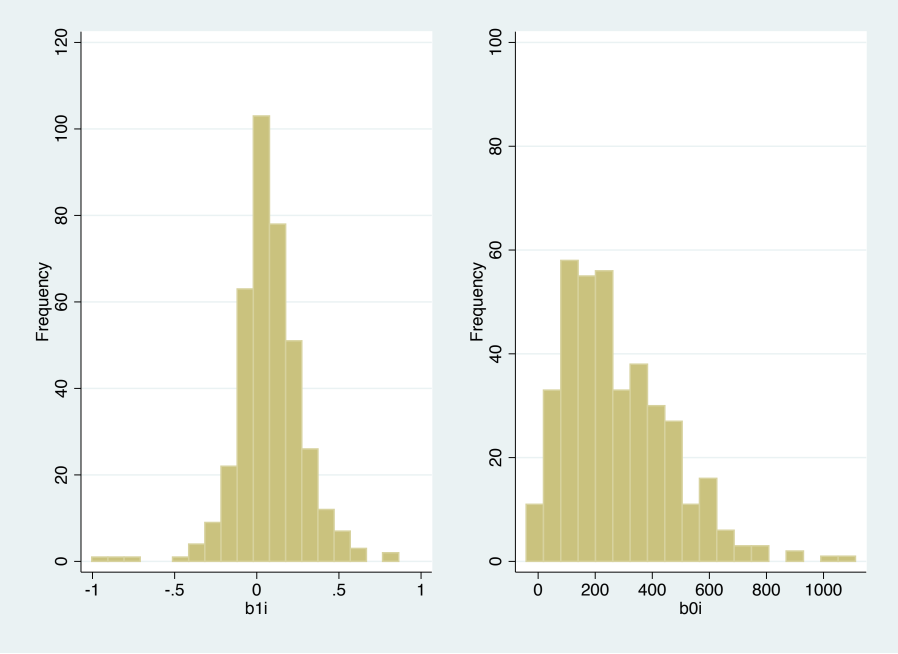


Figure 2.6: Distribution of the estimated slopes (b_{1i}) and intercepts (b_{0i}) across the subjects.



2.2.1 Smoothing

One way to examine mean trends over time is to assume no particular form (linear, quadratic, etc.), but "smooth" the data (perform smooth regression). Let the value of the smooth at time t be $m(t)$. Although there are many types of smoothing (moving average, kernel density, smoothing splines, local linear, etc.), most of the techniques used to calculate $m(t)$ can be understood as "local" (weighted) averages of the Y's in a neighbor around t (Hardle,1990). How local these averages are determines how smooth $m(t)$ is. First, think of the simple moving average, which just places a window along the X-axis, with t at the center of the window, and determines $m(t)$ by a simple average of the Y's in the window around t.

The problem with such an approach is that the contribution of a Y changes drastically as it moves just outside of the window, from contributing a full observation to contributing nothing to $m(t)$. Thus, we expect simply moving averages will not be very smooth, particularly if sample sizes are small. An elegant solution is to not have a fixed window, but instead provide a weight for each observation, Y, which is related to the distance from the point, t, at which the smoothed is being calculated. Thus, the contribution of an observation changes gradually as the point, t, moves either closer or farther from the observation, the weight becoming either larger or smaller, respectively. The weight function is often referred to as a kernel. Assume for now we only have one subject and n observations of (Y_j, T_j), $j = 1, \ldots, n$; when calculating $m(t)$, the weighted average is:

$$m(t) = \frac{\sum_{j=1}^{n} w(\frac{|T_j - t|}{h}) Y_j}{\sum_{j=1}^{n} w(\frac{|T_j - t|}{h})} \tag{2.2}$$

where w is the weight function (kernel) that decreases from $w(0)$ on both sides and h is referred to as the bandwidth. The bandwidth determines how quickly wt decreases away from a point t; as h gets smaller then the weight decreases quickly and the resulting $m(t)$ becomes rougher. The form of the kernel, w, varies but one commonly used kernel is the normal density. In this case the bandwidth, h, is the same as the standard deviation of the normal distribution. Figure 2.7 shows an example of raw data (Y vs. t) and two possible kernel weight functions (normal) at $t = 50$ with different bandwidths superimposed on raw data; the larger h results in a wider weight function and thus greater smoothness relative to the smaller bandwidth. There are many types of smoothers and other technical issues regarding issues like smoothing near the limits of data or variable bandwidths that are covered in other books (see Hardle, 1990).

STATA has a locally weighted polynomial regression or *lowess* (Cleveland, 1979) smoother

Figure 2.7: Y vs. t with a kernel weight function at t=50

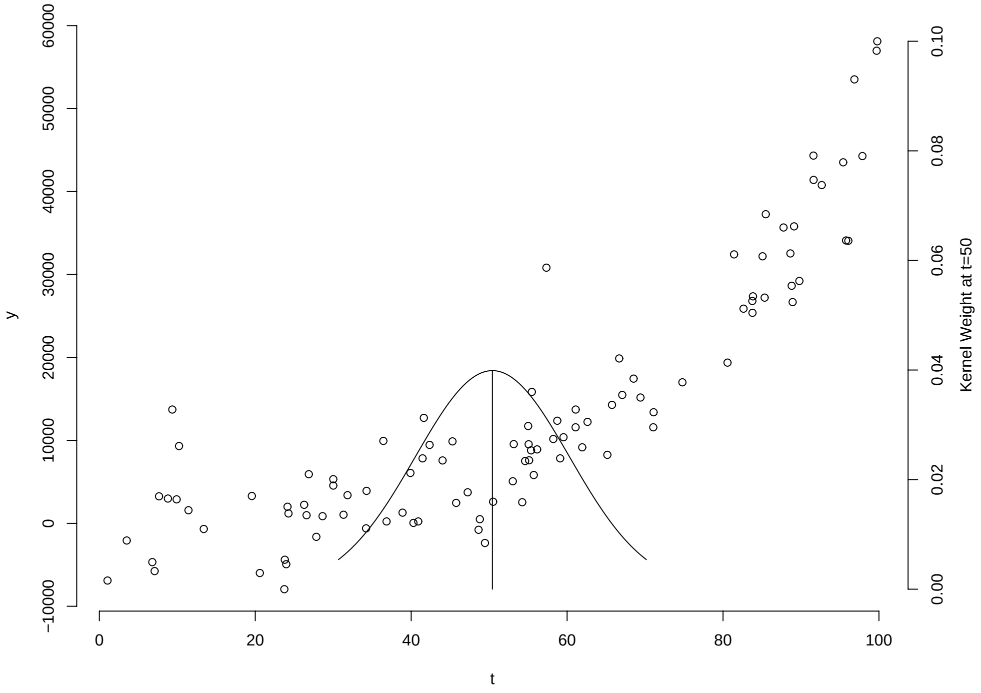
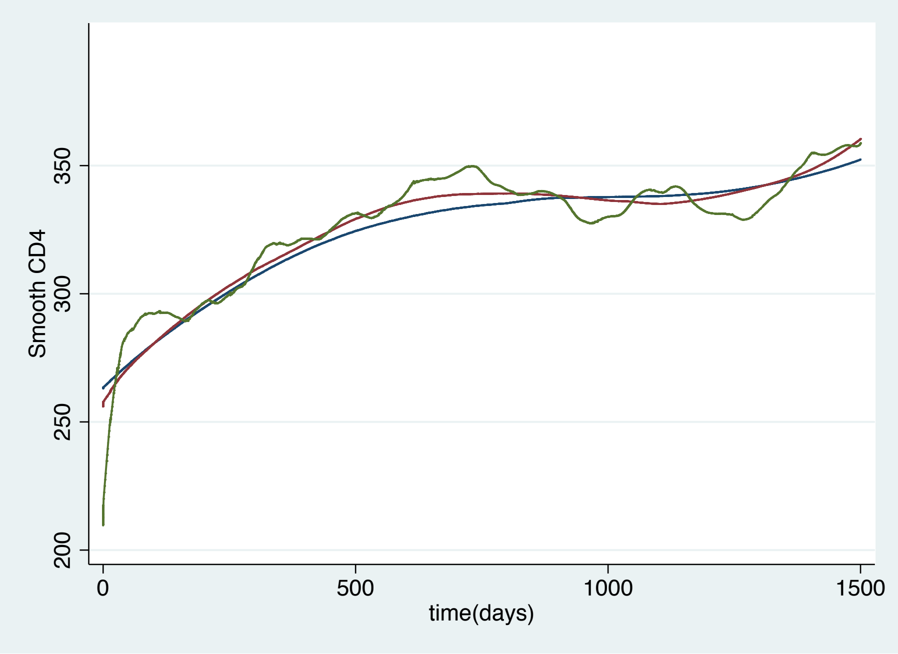


Figure 2.8: Smoothed CD4 at 3 different bandwidths



that we use on the CD4 data. As opposed to specifying a bandwidth, one chooses the proportion of data one wishes to use use to calculate $m(t)$ at each t and then the program determines the bandwidth that corresponds to that proportion. The following code is used to produce three smooths at progressively smaller bandwidths and, as expected (see Figure 2.8), the plot becomes rougher as a smaller proportion of the data is used to determine the smooth.

Interpreting the smooth of CD4 versus time since HAART of all subjects, on average, there is an early increase in CD4 count, a plateau around two years after initiation of therapy at a CD4 count of about 325, and finally a possible increase in CD4 near 3.5 years after HAART. Although all curves provide a similar picture, it is doubtful the roughest curve is picking up real variation in the underlying trend in average CD4 count. There are more formal ways of choosing the "optimal" bandwidth (such as cross-validation) than simple visualization of the plots and we leave that discussion to more specialized texts.

Besides looking at the average over all subjects, one can also smooth by subject and look at the distribution of smooth curves to examine the variation in smoothed individual trends, similar to what was done for linear trends by subject above. The following code is used to estimate smooths by subject.

Figure 2.9 has the lowess fits for two subjects; Figure 2.10 has smoothed trajectories for a random sample of 10 percent of the subjects. First of all, for most subjects the linear assumption appears reasonable; most of the smooths, which assume no particular functional form, look linear. However, there is variation in the trends, from almost flat CD4 with time to concave down, etc. This information can be used to fit a model that is flexible enough to accommodate the distribution of patterns and also incorporate the obvious variability among subjects (see Chapter 8).

2.3 Exploring the relationship of outcomes and explanatory variables

Besides the relationship of Y_{ij} and T_{ij}, one is often interested in the impact of explanatory variables (other than time) on the mean of the outcome of interest. First, we will consider a singe baseline covariate X_i and its impact on the future trend of Y_{ij} versus T_{ij}. Using the same data, consider categorical baseline HIV viral load (vl): $X_i = 1$ ($vl < 70000$), $X_i = 2$ ($70000 \leq vl < 220000$) and $X_i = 3$ ($vl \geq 22000$). One question is how baseline viral load predicts the future success of treatment as measured by CD4 count, which can be translated as the trends in CD4 count versus time stratified by baseline viral load (X_i).

Figure 2.9: CD4 VERSUS TIME AND LOWESS SMOOTHS FOR TWO SUBJECTS

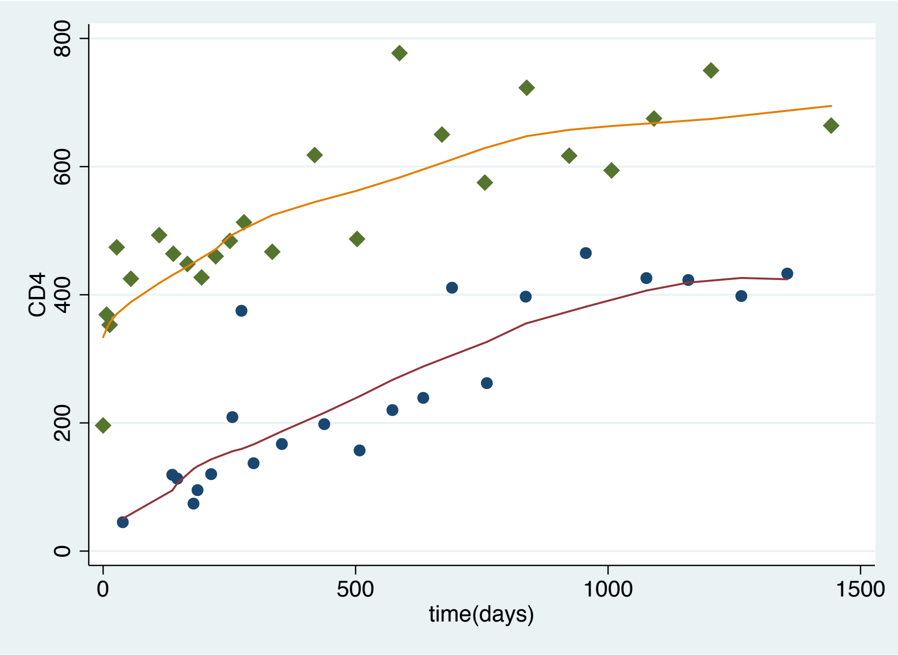


Figure 2.10: Smoothed CD4 versus time for a random sample of subjects

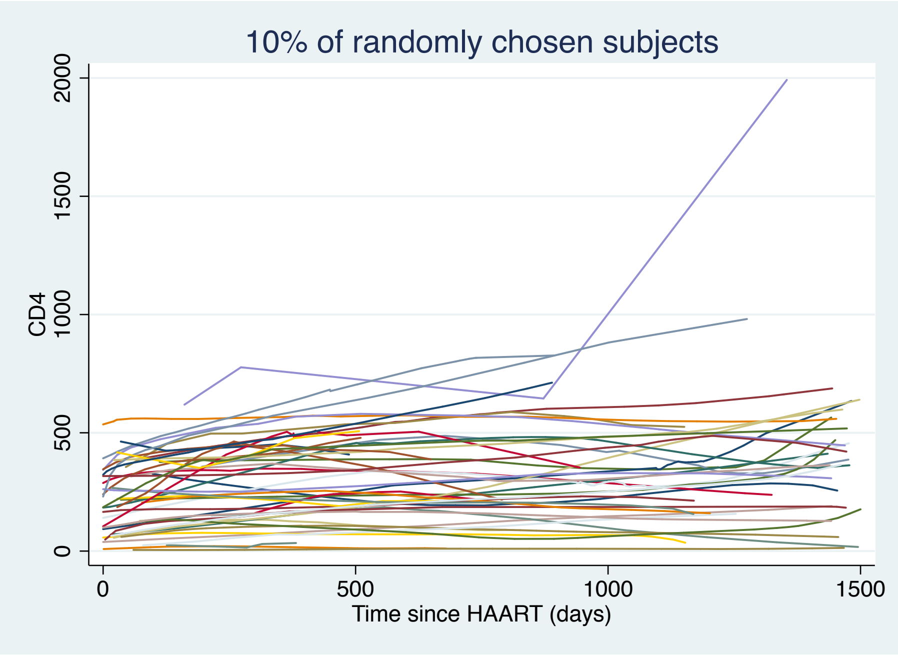
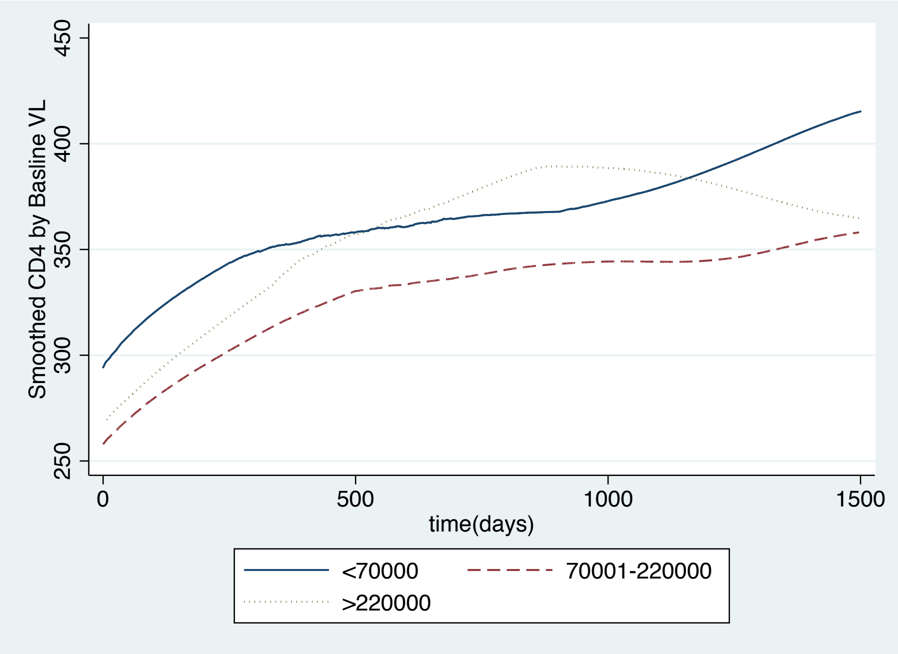


Figure 2.11: Smoothed CD4 versus time stratified by baseline viral load



In this case, we wish to estimate $E[Y_{ij} \mid T_{ij} = t, X_i = x]$ for a set of t's from 0 to 1500 and for each of the three strata; assuming no particular model, we simply smooth the data within strata of X_i.

The results (Figure 2.11) do not imply a simple relationship of future CD4 count and baseline viral load. Though the lowest viral load group (< 70000) does start at a higher average CD4 count, the next highest baseline CD4 group appears to be those with highest baseline viral load (> 220000). All groups on average go up relatively steeply in CD4 through time until about 500 days where they begin to plateau, with the exception of the group with highest baseline viral load, which keeps gaining. Interestingly, near the end of the data (days > 1200), the two lowest baseline viral load groups ($X_i = 0, 1$) continue to increase in average CD4 count whereas the highest baseline viral load group ($X_i = 2$) starts to declining, suggesting a higher probability of treatment failure. Of course, this is not formal statistical inference; future chapters will turn these informal graphical explorations into formal estimates and inference regarding the behavior of CD4, viral load and time.

2.3.1 Change in Y_{ij} versus change in X_{ij}

Finally, we end this section with suggestions for plots to explore how changes in the outcome Y_{ij} are related to changes in a time-dependent explanatory variable, X_{ij}. Again we will use viral load, but will log_{10} transform for viral load as is typically done because of the great range of values in viral load and the small subset of subjects with enormous values. What we are doing is relatively simple - graph the change in CD4 versus change in log_{10}(viral load) at two different measurements in time. Note, that there are several ways one could draw such observations, for instance, a single random pair of observations from each subject (one $Y_{ij}, X_{ij}, Y_{ij^*}, X_{ij^*}$, $j \neq j^*$), all possible pairs of observations, (all unique pairs $Y_{ij}, X_{ij}, Y_{ij^*}, X_{ij^*}$, $j \neq j^*$) and adjacent pairs in time (all pairs $Y_{ij}, X_{ij}, Y_{i(j+1)}, X_{i(j+1)}$, $j = 1, \ldots, n_i - 1$). This will depend on several factors including the amount of data, how observations are clustered in time, convenience, etc. In our case, there is so much variation in the time between measurements, that we use the latter approach. Now, define the new outcome and explanatory variable as

$$Y^*_{ij} = Y_{ij} - Y_{i(j-1)} \tag{2.3}$$

$$X^*_{ij} = X_{ij} - X_{i(j-1)} \tag{2.4}$$

where the j's are ordered in time. Note, we remove observations for which viral load remains at the detection limit ($vl \leq 500$) two times in a row so the change is 0 only because there is no further down a viral load can go. We thus make the assumption that no change from one time to the next at a measurable viral load is a much different phenomena than no change at the lowest viral load possible. Next, we simply throw all these observations in one pool and smooth Y^*_{ij} versus X^*_{ij} (see web supplement for code used to create Figure 2.12).

The results imply an average decrease in CD4 of about 60 for an increase in 2 orders of magnitude (e.g., 1000 to 100000) for viral load, whereas an equivalent decrease in viral load is associated with an increase in CD4 of about 80. This is a graphical estimate of the longitudinal effect of log_{10} viral load on CD4 count; we can also compare this to the cross-sectional effect by simply smoothing all the raw CD4 counts versus the corresponding viral loads.

Similar to Figure 2.12, Figure 2.13 shows a decline in CD4 count with increasing log_{10}(viral load). However, the cross-sectional effect appears much steeper, suggesting a 200 unit decline in CD4 for every 2 orders of magnitude increase in viral load. As mentioned in Chapter 1, this shows one of the virtues of longitudinal data. The more interesting effect of interest is how, within a subject, the change in viral load is associated with a change in CD4 count. In this case, the cross-sectional estimate, which is equivalent to the association of CD4 and log_{10}(viral load) in a random sample of individuals who are measured just once,

Figure 2.12: Smoothed change in CD4 versus change in log_{10}(viral load) - Longitudinal Effect

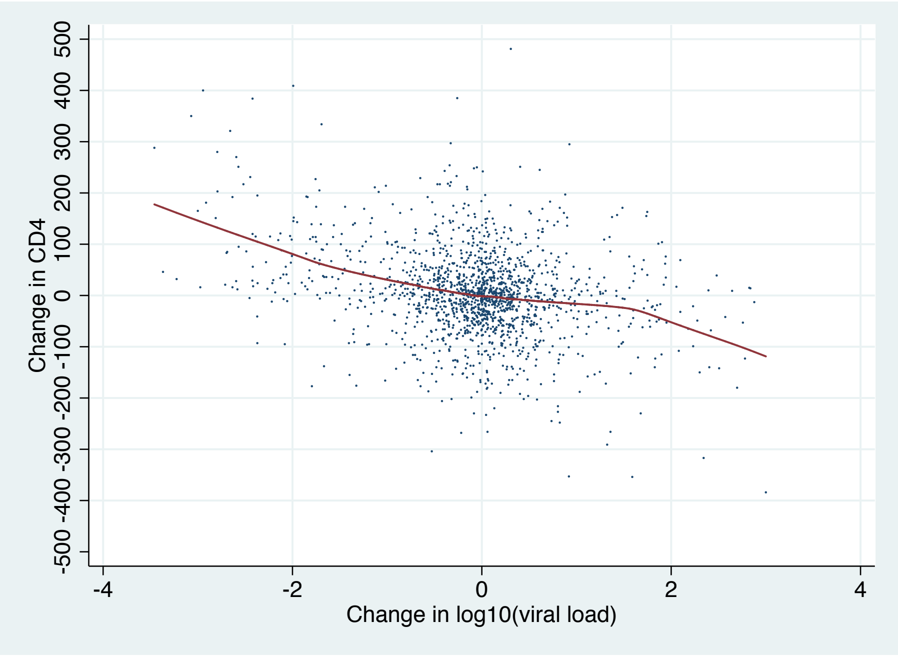
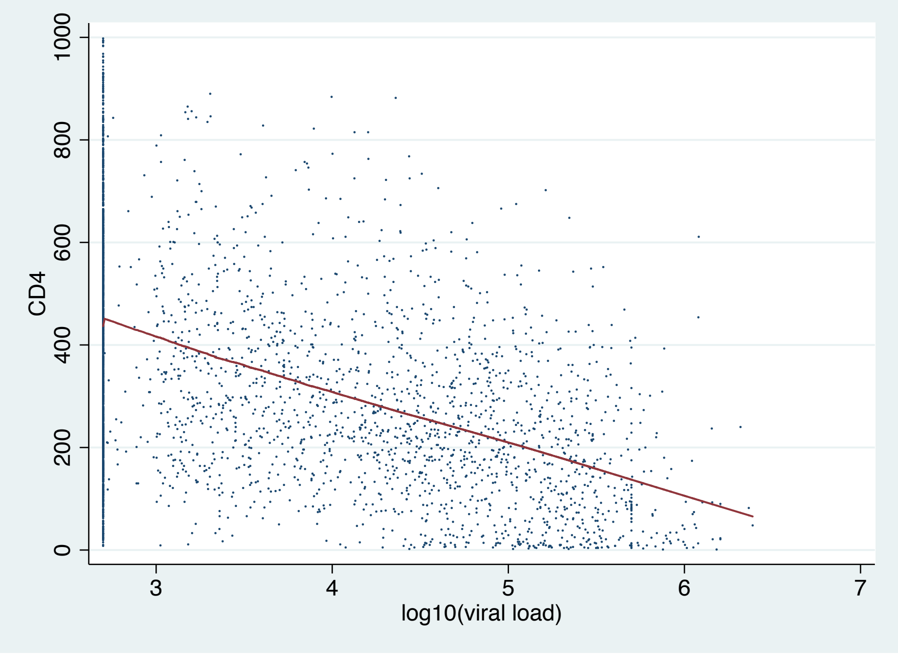


Figure 2.13: Smoothed CD4 versus log_{10}(viral load) - Cross-sectional Effect



overestimates the effect of changing viral load on CD4 count.

2.4 Problems

Question 2.1 Download the file `strength.dta` and then complete the following tasks and answer the questions.

Strength Data This is clinical trial data consisting of 3 treatment groups:

- No training (`tx = 1`)
- Weight training with light weights and high repetition (`tx = 2`)
- Weight training with heavy weights and low repetition (`tx = 3`)

Subjects were followed for 7 weeks and a measure of muscle strength was recorded each week. Answer the following questions using graphical methods (and a formal statistical approach if specified).

(a) For each treatment, make individual line plots of all subjects' trajectories of strength versus week. As a bonus, make the 3 plots appear on the same graph panel.

(b) Fit a linear regression by **subject**, and insert the regression lines on the plot from part (a).

(c) Fit a linear regression in each **treatment** group, and create a line plot of strength versus week that contains these (3) lines.

(d) Make a box plot of the distribution of strength by both week and tx (should have 3 panels, and 7 box plots per panel).

(e) *Optional.* Recreate the plots from parts (b) and (c), but instead of fitting a linear regression model, fit a smoothed average (using the `lowess` function in `Stata`). Compare these smoothed trajectories with the results from the parametric regression models.

(f) Answer the following questions based on the plots created in (a) - (d):

  - Does weight training have any impact on strength?

  - Is there a difference between treatments 2 and 3?
  - Which training program works quickest to increase strength?

(g) *Comprehensive Learning.* Reduce the data for each subject to 1 number: the slope of the change in strength estimated in each person separately (Stata code to do this is included in the web supplement).

  (i) Test the treatment effect of `tx=3` versus `tx=2` on this outcome (ignoring `tx=1`) using a standard two-sample t-test.

  (ii) Repeat using the sampling distribution generated by a permutation test to get a p-value.

  (iii) What do the results suggest? Turn in your results and a few sentences explaining your conclusions along with your code.

Question 2.2 Download the file `schitzophrenia.dta`. Answer the following questions using a few sentences and an appropriate graphical method (ideally one presented in this chapter).

(a) *One dimensional summaries.* Graphical supplements to your answers are not required here.

  - How many unique individuals are there in this dataset?
  - Of those, how many are being treated and how many are not?
  - How many females versus males are there?
  - What is the distribution of observations over the weeks (i.e. is it a balanced design, do most people miss visits)?

(b) *Two dimensional summaries.* Please include at least one graphic with each answer.

  - Are trends in severity of schitzophrenia attacks different in people who were treated versus people who were not treated? Describe each trend separately, and then qualitatively compare the two trends.
  - *Optional.* Is there a difference in average severity of episodes between the two treatment groups? In other words, is the drug effective in reducing the severity of schitzophrenia attacks?
  - Are trends in severity of schitzophrenia attacks different in women versus men? Describe each trend separately, and then qualitatively compare the two trends.

References

Cleveland, W.S. (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association, 74:829836, 1979.

Diggle, P., Heagerty, P., Liang, K-Y. & Zeger, S. (2002). *Analysis of Longitudinal Data*. Second Edition. Oxford: Oxford University Press.

Hardle, W. (1990). Applied nonparametric regression. Cambridge University Press, Econometric Society monographs.

StataCorp. 2005. Stata Statistical Software: Release 9. College Station, TX: StataCorp LP.