PH7430 Statistical Methods for Correlated Data - Fall 2009

Exploratory Data Analysis

Kyle D. Rudser

September 10, 2009
Role of Statistics

- “Statistics is about science and science is about proving things to people”
- Advances in science based on scientific method
- Statistics formalizes (quantifies) evidence for distinguishing between (2) hypotheses
  - Differences between groups are observed
  - Quantify the precision (uncertainty) in estimates
  - Evaluate the strength of evidence for/against hypotheses of interest
General approach to data analysis (and manuscripts)
(1) Identify overall goal of the study
(2) Specification of specific aims (hypotheses)
(3) Materials and methods (describe and characterized)
(4) Results of analysis
(5) Interpretation of results (discussion of scientific impact)
General approach to data analysis: before looking at data

1. Identify the overall goal of the study
   - Identify scientific question of interest (may perhaps be vague)

2. Identify specific aims and how they relate to the overall goal
   (current state of scientific knowledge, competing hypotheses
   current study is designed to discriminate between)
   - Identify statistical hypotheses to address scientific ones
     (refinement from questions of interest)
     - Types of questions: cluster cases, cluster variables,
       quantification of distributions, detecting associations between
       groups (distributions), prediction. Which one pertains?
     - Chose relevant summary measure(s)
   - Often helpful to consider the design of an ideal experiment
     (ignoring limitations of resources, ethics, etc.) and then
     identify how close the actual is able to obtain
General approach to data analysis: before looking at data

(3) Materials and (Methods)

- Available data
  - Sampling scheme and impact on questions able to answer, features able to examine (e.g., retrospective vs. prospective, intervention vs. observational, inclusion/exclusion criteria)
  - Categorize variables in dataset according to meaning (demographic, physiology, risk factors/prognosis, treatment, etc.)
  - Categorize variables in dataset according to use in analysis (response, predictor of interest, precision, confounding, effect modification)
- Descriptive statistics...
“Every good statistical analysis begins with an ‘ocular test’, that is, a good look at the data”. Here are some recommendations for plotting correlated data:

1. Show as much of the raw data as possible; minimize summarizations.

2. Highlight aggregate patterns.

3. Identify patterns both within “clusters” and across “clusters.”

4. Identify both unusual observations and unusual clusters.

For ease of conversation on this, will try to be consistent in terminology: we’ll discuss these plotting techniques in terms of taking observations across time for each of a number of people. Many of these plots can be done for any type of correlated data.
Things to look for in exploratory data analysis (EDA) plots

- Do individuals in general show trajectories or trends in their response values? (e.g., increasing, decreasing, monotone, . . . )
- If there are trends, what form do they take? (e.g., linear, quadratic, exponential, . . . )
- If there are trends, is the form similar across individuals?
- If there are trends, do they differ by various covariate groups? (e.g., gender, age, . . . )
How does the variability in the response values change across time? Across clusters?

How does the variability in the response values change across levels of the response values? (e.g., do higher response values show higher variability)

How does the variability in the response values change across covariate values? (e.g., gender)
Raw Data Plots

(1) “Time plot” with lines connecting individuals’ observations
Disadvantages: can get very crowded for large m, and changes in variability can hide changes in level

(2) Time plot of within-person “residuals”
\[(y_{ij} - \bar{y}_i) / (\sqrt{s^2_i})\] with lines connecting individuals’ observations
Disadvantages: can get very crowded for large m, no residual for those with 1 observation

(3) Time plot of within-time point “residuals”
\[(y_{ij} - \bar{y}_j) / (\sqrt{s^2_j})\] with lines connecting individuals’ observations
Disadvantages: can get very crowded for large m
SIX CITIES STUDY

Forced expiratory volume in 1 second (liters)

Within-year residual of forced expiratory volume in 1 second (liters)

Within-person residual of forced expiratory volume in 1 second (liters)
Plot (1), (2), or (3) for a pre-specified representative subsample of individuals

**EX** calculate the median outcome for each person and rank the medians; make a time plot for those with rank 0, 25, 50, 75, and 100\(^{th}\) percentile.

**EX** calculate the sample variance (or standard deviation) of the residuals for each person and rank them; make a time plot for those with rank 0, 25, 50, 75, and 100\(^{th}\) percentile.

**EX** as above, but instead of sample variance, calculate the median absolute deviation for each person:

\[
\text{median}\{|r_{ij} - \bar{r}_i|, j = 1, \ldots, n\}
\]

which is less affected by outlying residuals.
(5) Plot (1), (2), or (3) but only for a random subsample of the individuals
Disadvantages: may get an unrepresentative sample (e.g., may hide outlying observations or individuals)

(6) Plot any of the above separately by covariate groups.
SIX CITIES STUDY

Lines drawn for participants with median FEV1 ranked 0, 10, 25, 50, 75, 90, or 100th percentile.
Raw Data Plots with Superimposed Summaries

(1) Scatterplot with a loess fit superimposed
Disadvantage: cannot distinguish cross-sectional from longitudinal effect since individuals are not tracked

(2) Scatterplot with averages superimposed
Disadvantage: (as above)

(3) Either of the above separately by covariate groups
SIX CITIES STUDY

Average trends in FEV1 over time by gender
Aside: Basic Scatterplot Smoothing

Sample means can be sensitive to “outlying” values. How can we get an idea of the trend in response values over time without taking the average response at each time point?

Fit a non-parametric curve to the data; these are often called smoothers. Smoothers typically place few or no constraints on the form of the curve (aside from being smooth - think differentiable).

EX loess (lowess)
To calculate a loess fit of $y$ at any point $x$, a “window” is placed around the $k$ “nearest neighbor” values of $x$. Those are the $k$ points to the left and right of $x$ which are closest to $x$. 
The square brackets show the nearest neighbor window. All the observed data points which fall within this window are called the nearest neighbor set N(x) of x. x is the point at which we want the fitted loess curve.
Denote $N(x)$ as the set of nearest neighbor points around $x$: $N(x) = \{x_i : x_i \ is \ a \ n.n. \ to \ x\}$.

1. Compute the maximum (horizontal) distance between $x$ and each $x_i$ in $N(x)$:

$$\Delta = \max_{x_i \in N(x)} |x - x_i|$$

2. Assign weights to each data point:

$$w(x_i) = \begin{cases} 
\left(1 - \left(\frac{|x-x_i|}{\Delta}\right)^3\right)^3 & x_i \in N(x) \\
0 & x_i \notin N(x) 
\end{cases}$$

Thus each data point in the nearest neighbor set gets a positive weight (which is closer to 1 as $x_i$ gets closer to $x$) and every other data point gets a zero weight.
3. Carry out a weighted least squares regression of $y_i$ on $x_i$ (or on $x_i$ and $x_i^2$ for more “smoothness”).

Repeat 1-3 for a very tight grid of $x$-values and plot.
$k$ is usually chosen by specifying the fraction of data points to be included in the nearest neighbor window. S-Plus and R call this the span and take a default span of $2/3$. SAS calls this the smoothing parameter.

- As span increases, the loess fit becomes smoother (less jagged).
- The WLS regressions within each nearest neighbor window are either simple linear or quadratic regressions.
- Kernal smoothers and splines are similar and could be used instead.
The missing Kernel smoothing slide: Kernel smoothing

- another example of a nonparametric approach
- for a given ‘test point’, we have \( y_o = \theta(x_o) \), which is obtained by weighting all of the points in the sample through a weighting function (kernel): \( K_\lambda(x_o, \vec{x}) \)
- the weight for each observation \( x_j \) is based on some metric to quantify its distance from \( x_o \) (weight could be zero, could be different for different \( x_o \))
- generally nearby observations to \( x_o \) will be given more weight, hence making a larger contribution to the estimate \( y_o \)
- the parameter \( \lambda \), called bandwidth will determine the speed at which the influence of surrounding points die away
- formally we have \( K_\lambda(x_o, x_j) = g \left( \frac{d(x_o, x_j)}{\lambda(x_o, x_j)} \right) \) where \( d(\cdot) \) denotes the distance metric and \( g(\cdot) \) denotes the weighting function
Kernel smoothing

- the notion invoked is “borrowing information”
- want accurate estimates for each point $x_o$
- to the extent that examining the outcome from other points ‘similar’ to $x_o$ will give information as to what would be the expected outcome for $x_o$, then we can perhaps get a better estimate
- two ways to view a kernel smoother
  - surrounding observations of $x_o$ contribute to the estimate at that point
  - the value at $x_o$ is smeared out across the surrounding points in a specific fashion
- Library in R: *KernSmooth*
- Reference: Hastie, Tibshirani, Friedman: “The Elements of Statistical Learning”
**CODE: R and S-Plus**

```r
lowess(x, y, f= )  f = span, degree = 1
scatter.smooth(x, y, span = , degree= )
loess.smooth(x, y, span = , degree= )
loess(y ~ x, span = , degree = )
supsmu(x, y, span="cv", periodic=FALSE, bass=0)
```

`scatter.smooth()` creates the scatter plot and puts the fitted smooth on it. The other three only create the fitted smooth, which then must be put on an already-created scatterplot using the function `lines()`. The last three require the package `modreg` to be used in R.

**NOTE:** *missing values are not allowed!* Remove them from the data vectors beforehand... except for `supsmu`, that can handle missing values.
CODE: SAS

ODS OUTPUT OutputStatistics = outdata;
PROC LOESS DATA = indata;
MODEL y = x / SMOOTH = DEGREE = ;
RUN;

indata should contain x and y. outdata will contain x, y, and the fitted smooth values, which can then be used in PROC GPLOT or exported to another program for graphing.
SIX CITIES STUDY

FEV1 by baseline weight with loess smooth
SIX CITIES STUDY

FEV1 by baseline weight with loess smooth

Women
Men

6c.fev.blwtgender.ps
CPCRA 010: WOMEN'S FUNGAL STUDY

Candidiasis culture: loess by clinical unit
Summarized Data Plots

(1) Boxplot at each time point or for each cluster
Disadvantages: gives a very crude view of changes in level and changes in variability. Can’t be done for binary outcomes.

(2) Plot of average or median at each time point or for each cluster
Disadvantage: also crude.

(3) Scatterplot of within time (or within cluster) average vs. standard deviation

(4) Any of the above separately by covariate groups
SIX CITIES STUDY

FEV1 by year with width proportional to sqrt(sample size)
CPCRA 010: WOMEN’S FUNGAL STUDY

Standard error of $P(\text{positive culture})$ by average $P(\text{positive culture})$

![Diagram showing the relationship between standard error and average positive culture within months.]

- A: month 0
- B: month 3
- C: month 6
- D: month 9
- E: month 12

The diagram illustrates the standard error of $P(\text{positive culture})$ plotted against the average $P(\text{positive culture})$ for different months.
Plots with Time-Constant (Cluster-Level) Baseline Covariates

We will use the terminology

“time-constant covariate” (or time-independent covariate)

and

“time-dependent covariate” (or time-varying covariate)
(1) Plot $y_{i0}$ vs. $x_{i0}$
Disadvantage: cannot see how the baseline covariate may be associated with subsequent $y_{ij}$ values

(2) Plot $y_{ij}$ vs. $x_{i0}$ for each $j$ separately
Disadvantage: cannot see correlation across time (within cluster)
SIX CITIES STUDY

Year 1 FEV1

Year 4 FEV1

Year 7 FEV1
Plots with Time-Varying (Within-Cluster) Covariates

(1) Plot \( y_{ij} \) vs. \( x_{ij} \) for all \( j \) values together. Use different symbols for the first time point, the last time point, and then all points in between. Draw lines connecting individuals’ observations. Disadvantage: Can be difficult to detect a pattern in the \((x_{ij}, y_{ij})\) relation, especially if there are many time points.

(2) Plot response residuals vs. covariate residuals (use symbols as above)

\[
\text{EX} \quad \frac{y_{ij} - \bar{y}_j}{\sqrt{s^2_j}} \quad \text{vs.} \quad \frac{x_{ij} - \bar{x}_j}{\sqrt{s^2_{x,j}}} \quad \text{within-time “residuals”}
\]

\[
\text{EX} \quad \frac{y_{ij} - \bar{y}_i}{\sqrt{s^2_i}} \quad \text{vs.} \quad \frac{x_{ij} - \bar{x}_i}{\sqrt{s^2_{x,i}}} \quad \text{within-person “residuals”}
\]
SIX CITIES STUDY

FEV1 by weight across years for women

- o year 1
- . year 4
- x year 7
SIX CITIES STUDY

FEV1 by weight across years for men

weight (kg)
forced expiratory volume in 1 sec (liters)
40 60 80 100 120
1 2 3 4 5
year 1
year 4
year 7
C. Exploring Correlated Data – Correlation Structure

Recall that our models will look like:

\[ y_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + \varepsilon_{ij}. \]

Which part of the model will contain information on variances and correlations among the \( y_{ij} \)?

How do some people estimate this part?

⇒ Fit a model that assumes no correlation (e.g., independence model, ordinary least squares multiple regression) and that accounts for as many covariate effects as possible (as “saturated” as possible) including all reasonable interactions. Compute the residuals \( r_{ij} = y_{ij} - \hat{y}_{ij} \).
When the data were collected at equally spaced time points, make a scatterplot “matrix” of residuals ($r_{ij}$ vs. $r_{ij'}$ for each $j \neq j'$).

Next, compute the correlation within each plot if the plots look roughly like a bivariate normal distribution (an oval, not cone nor crescent shaped, cloud of points).

- Do the plots have roughly equivalent spread vertically and horizontally?
- Are the correlations constant within a “lag”? 
  
<table>
<thead>
<tr>
<th>Lag</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 and 4</td>
<td>3</td>
</tr>
<tr>
<td>1 and 7</td>
<td>6</td>
</tr>
<tr>
<td>4 and 7</td>
<td>3</td>
</tr>
</tbody>
</table>

- Are the correlations constant across lags?
Suppose we consider the Six Cities Study which had data collected at three equally spaced time points: years 1, 4, and 7. We fit a multiple linear regression using as many covariates as possible. We calculate the residuals:

\[ r_{ij} = y_{ij} - \hat{y}_{ij}, \quad i = 1, \ldots, m, \quad j = 1, 4, 7 \]

and group them according to year: \( \{r_{i1}\}, \{r_{i4}\}, \{r_{i7}\} \). Then we plot \( \{r_{i1}\} \) vs. \( \{r_{i4}\} \), \( \{r_{i1}\} \) vs. \( \{r_{i7}\} \), and \( \{r_{i4}\} \) vs. \( \{r_{i7}\} \).
Row variable plotted on vertical axis. Column variable plotted on horizontal axis.
What if the data were collected across time but not at equally spaced time points? The scatterplot matrix of residuals can still be made, but beware of how you interpret changes as the lag increases. (More on this later.)

However the time points are spaced, compute the average of the residuals within time point and the variance of the residuals within time point. Are they constant? Do they increase or decrease?

What if the data are correlated within a cluster and not collected across time? Again fit a saturated regression model and compute the residuals. Now compute the average residual and variance of residuals within cluster.

Later in the course, we’ll also learn how to compute sample auto correlation functions and sample variograms.
NOTE: When creating a scatterplot matrix of residuals, be very careful if your data have missing values!! We are plotting $r_{ij}$ vs. $r_{ij'}$ and consider $j = 1$ and $j' = 3$:

$\begin{bmatrix}
3.2 \\
2.8 \\
3.7 \\
\vdots \\
4.1 \\
\end{bmatrix}$ vs. $\begin{bmatrix}
3.0 \\
\cdot \\
3.6 \\
\vdots \\
3.6 \\
\end{bmatrix}$
Correct with missing data:

\[
\begin{bmatrix}
3.2 \\
3.7 \\
\vdots \\
4.1
\end{bmatrix}
\text{ vs. }
\begin{bmatrix}
3.0 \\
3.6 \\
\vdots \\
3.6
\end{bmatrix}
\]

Incorrect with missing data:

\[
\begin{bmatrix}
3.2 \\
2.8 \\
3.7 \\
\vdots \\
4.1
\end{bmatrix}
\text{ vs. }
\begin{bmatrix}
3.0 \\
3.6 \\
2.2 \\
\vdots \\
\vdots
\end{bmatrix}
\]