Longitudianl Data Analysis

2. Graphical representation of longitudinal data

1. Y **against** $Time$

   - Check the time trend for every individual
   - Check the variability change over time (esp. at beginning vs end)

   Instead of using Y, use the standardized residuals:

   $$y_{iJ}^* = \frac{y_{iJ} - \bar{y}_{\cdot J}}{s_J}$$

   $$\bar{y}_{\cdot J} = \frac{1}{n}\sum_{i=1}^{n} y_{iJ}, s_J^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_{iJ} - \bar{y}_{\cdot J})^2$$

   - **Spaghetti plot**
   - **ZAP plot**: using residuals

     (a) regression y_{ij} on t_{ij} and get the residuals r_{ij}

     (b) choose on dimensional summary of the residuals, for example g_i=median$(r_{i1}, \cdots, r_{in_i})$

     (c) plot r_{ij} versus t_{ij} using points (quantiles of g_i)

     (d) order units by g_i (min., 10th, 25th, 50th, 75th, 90th, max.)

     (e) add lines for selected quantiles of g_i

2. Y **against** X

   - AV plot
   - Graphic methods to separate CS information from LS information

     The model

     $$y_{ij} = \beta_C x_{i1} + \beta_L(x_{ij} - x_{i1}) + \varepsilon_{ij}, i = 1, \cdot, m; j = 1, \cdots, n$$

     suggests making two scatterplots:

     (a) y_{i1} against x_{i1} for $i = 1, \cdots, m$

     (b) $y_{ij} - y_{i1}$ against $x_{ij} - x_{i1}$ for $i = 1, \cdots, m; j = 2, \cdots, n$

     From the plots, we can then ask:

     (a) are the differences across subjects? (CS effect)

     (b) are there changes across time within subject? (LS effect)

Fitting smotth curves

$$Y_i = \mu(t_i) + \varepsilon_i, i = 1, \cdots, m$$

We want to estimate an unknown mean response curve $\mu(t)$ in the model. Non-parametric regression models that can be used to estimate the mean response profile as a function of time.
Alway, the smaller of the window width the less smothness we get.

Kernel estimation

- selection of window centered at time t;
- $\hat{\mu}(t)$ is the average of Y of all points that are visible in that window
- slide a window from the extreme left to extreme right, calculating $\hat{\mu}(t_i)$ for each.
- a "better" method is to use a weight function that changes smoothly with time and gives weights to the observations closer to t. E.g. Gaussian kernel $K(t_i) = \exp(-0.5t_i^2)$ $(i.e. N(0,1))$
- at window with $t = t_1$,

  $$\hat{\mu}(t_1) = \sum_{i=1}^{m} \omega(t_1, t_i, h) y_i / \sum_{i=1}^{m} \omega(t_1, t_i, h)$$

  where h controls the smoothness.

Smoothing Spline

- **Smoothing spline** (cubic): the function $s(t)$ that minimizes the criterion

  $$J(\lambda) = \sum_{i=1}^{m} \{y_i - s(t_i)\}^2 + \lambda \int s''(t)^2 dt$$

- $s(t)$ stisfies the criterion if and only if it is a piecewise cubic polynomial
- λ smaller $\Rightarrow$ curve less smooth
- $\int s''(t)^2 dt$: roughness penalty

Loess

1. Center a window at time t_i
2. fit weighted least squares
3. calculate the residuals

4. down weight the outliers and repeat 1, 2, 3 many times
5. the result is a fitted line that is insensitive to the observations with outlying Y values

Exploring correlation structure

- Let $y_{ij} = \beta_0 + x_{ij}\beta + \varepsilon_{ij}$, we should be clear that

  $$cor(y_{ij}, y_{ik}) \neq cor(\varepsilon_{ij}), \varepsilon_{ik} = cor(y_{ij}, y_{ik}|x_{ij}, x_{ik})$$

  similarly, $var(y) \neq var(\varepsilon)$.

- We explore the correlation structure based on the **residuals** of the model. And it is used for **equally spanced** data, not for irregular data.

- **Weakly stationary**: if residuals have constant mean and variance and if $corr(y_{ij}, y_{ik})$ depends only on $|t_{ij} - tik|$, then the process Y_{ij} is said to be weakly stationary.

- **Autocorrelation function (ACF)**:

  $$\rho(u) = corr(Y_{ij}, Y_{ij-u})$$

  for all i, which is pooling observation pairs along the diagonals of the scatterplot matrix.

  $$se(\rho(u)) \approx 1/\sqrt{N}$$

  where N is the number of independent pairs of observations in the calculation.

- ACF is most effective for studying equally spaced data that are roughly stationary.

- For irreguarly-sapced data, we can use **Variogram**:

  $$\gamma(u) = \frac{1}{2}E\left[\{Y(t) - Y(t-u)\}^2\right], u \geq 0$$

- If $Y(t)$ is *stationary*, the Variogram is directly related to the ACF $\rho(u)$ by

  $$\gamma(u) = \sigma^2\{1 - \rho(u)\}$$

  where σ^2 is the variance of Y:

  $$\begin{aligned} E(X-Y)^2 &= [E(X-Y)]^2 + Var(X-Y) \\ &= 0 + Var(X) + Var(Y) - 2Cov(X,Y) \\ &= 2\sigma^2 - 2\rho\sigma^2 \\ &= \sigma^2(1-\rho) \end{aligned}$$

- Calculating $\gamma(u)$, the Vriagram:

  1. Starting with the *residuals* r_{ij} and the time t_{ij}, compute all possible

     $$v_{ijk} = \frac{1}{2}(r_{ij} - r_{ik})^2$$

     and

     $$u_{ijk} = t_{ij} - t_{ik} \text{ for } j < k$$

  2. Smooth v_{ijk} against u_{ijk} (using lowess)
  3. Estimate the total variance as

     $$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{ij}(r_{ij} - \hat{r})$$

  4. If the time t_{ij} are not total irregular, i.e. there will be more than one observation at each value of u. Then let

     $$\hat{\gamma}(u) = \frac{\sum_{i=1}^{n_i} v_{ijk}}{n_i}$$