# Topic 1. Introduction to Longitudinal Studies and Exploring Longitudinal Data

In biomedical and epidemiological studies, such as clinical trials, disease progression follow-up studies, growth studies, etc., longitudinal data have been widely used and discussed.

**Characteristic of a longitudinal study:** Individuals are measured repeatedly through time.
**Characteristic of a cross-sectional study:** Single outcome is collected for each individual.

**General layout for longitudinal measurements-**

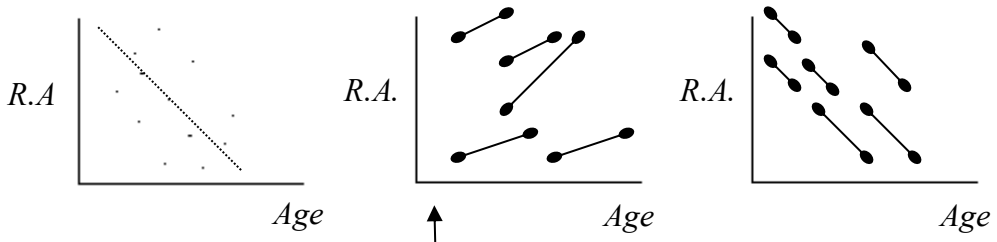
| subject (i) | time point (t) | response ($Y(t)$) | Covariates $X(t)=(X_1(t),\cdots,X_p(t))^T$ |
|---|---|---|---|
| 1 | $t_{11}$ | $Y_{11}$ | $X_{111}$ $X_{112}$ $\cdots$ $X_{11p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 1 | $t_{1m_1}$ | $Y_{1m_1}$ | $X_{1m_11}$ $X_{1m_12}$ $\cdots$ $X_{1m_1p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| i | $t_{i1}$ | $Y_{i1}$ | $X_{i11}$ $X_{i12}$ $\cdots$ $X_{i1p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| i | $t_{im_i}$ | $Y_{im_i}$ | $X_{im_i1}$ $X_{im_i2}$ $\cdots$ $X_{im_ip}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| n | $t_{n1}$ | $Y_{n1}$ | $X_{n11}$ $X_{n12}$ $\cdots$ $X_{n1p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| n | $t_{nm_n}$ | $Y_{nm_n}$ | $X_{nm_n1}$ $X_{nm_n2}$ $\cdots$ $X_{nm_np}$ |

## Advantages and disadvantages of longitudinal studies –

**Advantages:**

(a) It's possible to obtain information concerning individual patterns of change.

Hypothetical data on the relationship between reading ability and age.




Every one's reading ability is improved with time.

Note: the reading ability deteriorates with age.

(b) Subjects can serve as their own controls in that the outcome variable can be measured under both control and experimental conditions for each subject.

**Remark.** The variability between subjects can be excluded from the experimental error.

(c) Data can be collected more reliable in a study than in a cross-sectional study.

(Approximately the same result time after time)

**Disadvantages:**

(a) The analysis becomes complicated because the dependence among the repeated measurements within each individual

(b) Investigators often cannot control the circumstances for obtaining measurements so that data may be unbalanced or partially incomplete.

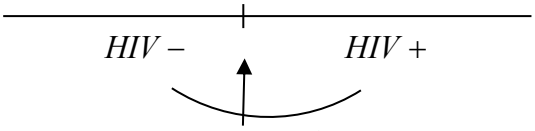
## Examples:

### 1. CD4+ cell numbers

Keywords:

HIV: the human immune deficiency virus which causes AIDS by reducing a person's ability to fight infection.

CD4+ cell: this cell orchestrates the body's immunoresponse to infection agents.




$CD4+ = 1100$ cells per milliitre of blood

Data: 369 seroconverters were collected from the Multicenter AIDS Cohort Study (MACS) with total 2376 repeated measurements during the study period.

Variables: time since seroconversion, CD4 counts, age, cesd, cigarettes smoked per day, drug use status, and number of sexual partners.

Objectives of this study:
(a) Estimate the average time course of CD4+ depletion.
(b) Estimate the time course of CD4+ for each individual man.
(c) Characterize the degree of heterogeneity across men in the rate of progression.
(d) Identify factors which predict CD4+ cell changes.

**2. Indonesian children's health study**

Data: 275 Indonesian children were collected from the Indonesian Children's Health Study (ICHS) in the Aceh province of Indonesia to determine the causes and effects of vitamin A.

Variables: time (in months), gender, vitamin A status, age, and respiratory infection.

Keyword:
Vitamin A deficiency: one of the leading causes of morbidity and mortality in children from the developing world.

Objectives:
Estimate the increase in risk of respiratory infection for children who are vitamin A deficient while controlling for other demographics factors (age, weight, height, etc.), and to estimate the degree of heterogeneity in the risk of disease among children.

**3. Growth of Sitka spruce trees**

Data: 79 trees over two growing seasons were recorded by Dr. Peter Lucas of the Biological Sciences Division at Lancaster University

$$\begin{cases} \text{54 trees were growing with ozone exposure at 70 ppb} < \begin{matrix} 27 \\ 27 \end{matrix} \\ \text{25 trees were growing under control conditions} < \begin{matrix} 12 \\ 13 \end{matrix} \end{cases} \quad \text{four chambers}$$

Variables: log-size of trees, chamber, ozone, year (1988,1989).

Objective: Assess the effect of ozone pollution on the tree growth which is measured via $\log(\text{tree hight} \times (\text{diameter})^2)$

**4. Protein content of milk**

Data: Milks were collected weekly from 79 Australian cows and analyzed for their protein contents. The data were provided by Ms. Alison Frensham. The cows were maintained on one of the three diets - barley, mixture of barley and lupins, and lupins.

Objective: Determine how diet affects the protein in milk.

**Remark.**
1. The protein contents were measured in weeks since calving, and the experiment was terminated 19 weeks after the earliest calving.
2. Calving date may directly or indirectly associate with the physiological processes that also determine protein content.

**5. Crossover trial**

Data: The data were reported from Jones and Kenward (1987). 86 women joined the experiment of a three-period cross-over trial of analgesic drug for relieving pain from primary dysmenorrhoea (menstrual cramps).

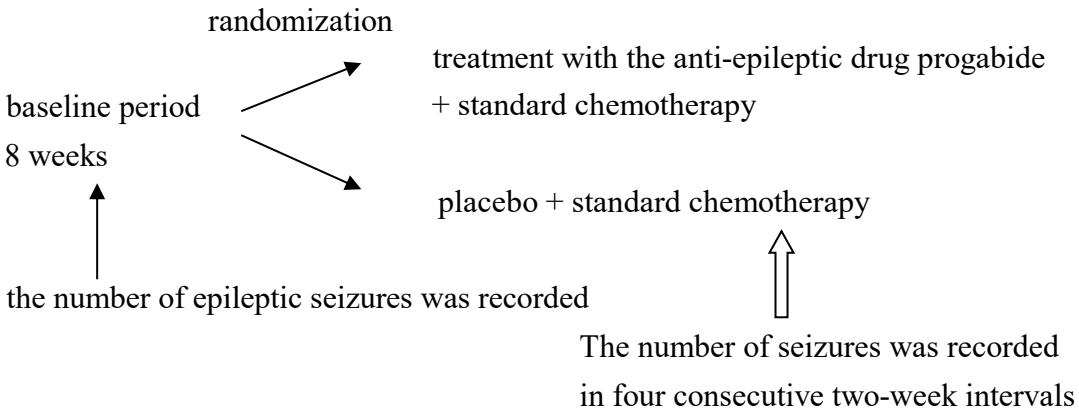
Note: Three levels of analgesic (control, low and high) were given to each woman and they were randomly allocated to one of the six possible treatment orders.

Objective: Study the effect of analgesic treatment for pain from the primary menstrual cramps on the pain relief.

**6. Epileptic seizures**

Data: The data from a clinical trial of 59 epileptics, analyzed by Thall and Vail (1990), and Breslow and Clayton (1993).

Variables: baseline epileptic seizures, treatment status, and age.




Objective: Whether the progabide reduces the rate of epileptic seizures.

**7. A clinical trial of drug therapies for schizophrenia**

Data source:

1. 523 patients of chronic schizophrenia were collected by Dr. Peter Ouyang, Janssen Research Foundation.

2. Patients were randomly allocated amongst the following six treatments:
placebo, haloperidol 20mg, and risperidone at dose levels 2, 6, 10, and 16mg.
Standard therapy.

Keyword:
PANSS: the positive and negative symptom rating scale, a measure of psychiatric disorder.

3. Measurements taken at weeks -1, 0, 1, 2, 4, 6, and 8
baseline
selection into the trial
⇑
Establish a stable regime of medication for each patient.
⇑
Egibility criteria: age 18-65 with good general health PANSS 60-120.

4. Within 523 patients, 253 patients complete the study and 270 patients drop out in the clinical trial for some reasons -abnormal lab result, adverse experience, inadequate response, inter-current illness, lost to follow up, uncooperative, withdraw consent, and other reasons.

Objective: Study the effects of treatments on the PANSS score.

## Approaches to longitudinal data analysis –

With one observation on each individual at each time, we are confined to modeling the population average of the response $Y(t)$, called the marginal mean response.

(a) Marginal models: Model the marginal mean as in a cross-sectional study.

Examples:

$$\begin{cases} 1\ \text{logit}(P(Y_{ij}=1)) = \beta_0 + \beta_1(\text{vitamin A deficient status})_{ij} \\ 2\ E[(\text{CD4+})_{ij}] = g(t_{ij}) \end{cases}$$

Since repeated values are not likely to be independent, the marginal model approach has the advantage of separately modeling the mean and covariance.

(b) Random effects model: Assume that correlation arises among repeated responses because the regression coefficients vary across individuals.

Example:

$E[Y_{ij} \mid \beta_i] = x_{ij}^T \beta_i$, where $x_{ij} = (1, x_{ij1}, \cdots, x_{ijp})^T$ and $\beta_i = (\beta_{i0}, \beta_{i1}, \cdots, \beta_{ip})^T$ with

$\beta_i = \beta + U_i$ and $U_i \overset{iid}{\sim} (0, \Sigma)$, a mean zero latent variable.

(c) Transition model: Focus on the conditional expectation of $Y_{ij}$ given past outcomes $(Y_{i1}, \cdots, Y_{ij-1})$ and the covariates $x_{ij}$.

Example:

$$\text{log}it\ P(Y_{ij}=1 \mid Y_{ij-1}, \cdots, Y_{i1}, x_{ij}) = x_{ij}^T \beta + \alpha Y_{ij-1}$$

↑ respiratory infection ↖ vitamin A deficiency

**Consequences of ignoring the correlation –**

1. Incorrect inferences about parameters.
2. Inefficient estimates of parameters.
3. Sub-optimal protection against biases causes by missing data.

**Two categories of longitudinal data analysis problems –**

(a) $n >> \max\{n_1, \cdots, n_m\}$ or the repeated times $t_{ij}$'$s$ occurring at the common set of times $\{t_1, \cdots, t_m\}$ : the robust variance estimates can be used to draw valid inferences about regression parameters.

(b) Correlation is of prime interest or $n$ is small: valid inferences can be obtained via approximately correctly specified models for mean and covariance.

**Exploratory data analysis (EDA)** – A detective work, which is the foundation stone of data analysis.

↕

**Confirmatory analysis** –

Guidelines:

1. Show as much of the relevant raw data as possible rather than only data summaries.
2. Highlight aggregate patterns of potential scientific interest.
3. Identify both cross-sectional and longitudinal patterns.
4. Make easy the identification of unusual individuals.

## Graphical presentation of longitudinal data –

(a) Scatter plot of the response against time.

(b) Lines connect the repeated observations for each individual.

↑

{ raw data: $y_{ij} \leftrightarrow t_{ij}$

{ standardized data: $y_{ij}^* = \dfrac{(y_{ij} - \bar{y}_i)}{S_i} \leftrightarrow t_{ij}$

↑

This plot is able to highlight the degree of 'tracking', whereby subjects tend to maintain their relative size over time.

Display changes through time for individuals.

Drawback: the plot may be extremely busy, reducing its usefulness.

Solution: Individual curves are ordered with some characteristic that is relevant to the model of interest.

(c) Fitting smooth curves to longitudinal data when the number of different times is large enough.

Data: $\{((t_{i1}, Y_{i1}), \cdots, (t_{im_i}, Y_{im_i})) : 1 \le i \le n\}$

$Y_{ij} = g(t_{ij}) + \varepsilon_{ij}$, where $\varepsilon_{ij}$'$s$ are mean zero errors, and $g(\cdot)$ is a non-parametric smooth function. The estimator $\hat{g}(t)$ for $g(t)$ can be obtained by the local averaging procedure, which is defined by

$$\hat{g}(t) = \arg\min N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \omega(t, t_{ij})(Y_{ij} - g(t))^2 = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \omega(t, t_{ij}) Y_{ij},$$

where $N = \sum_{i=1}^{n} m_i$ and $\omega(t, \cdot)$ is a weight function $\frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \omega(t, t_{ij}) = 1$.

(c1) kernel estimator: $\omega(u, v) \triangleq \frac{1}{h} k(\frac{u - v}{h})$ with $h > 0$.

| kernel function | k(s) |
|---|---|
| Rectangular | $\frac{1}{2} 1_{(\lvert s \rvert \le 1)}$ |
| Gaussian | $(2\pi)^{\frac{1}{2}} e^{\frac{-1}{2} s^2}$ |
| Triangular | $(1 - \lvert s \rvert) 1_{(\lvert s \rvert \le 1)}$ |
| Biweight | $\frac{15}{16}(1 - s^2)^2 1_{(\lvert s \rvert \le 1)}$ |
| Epanechnikov | $\frac{3}{4}(1 - s^2) 1_{(\lvert s \rvert \le 1)}$ |

**Remark.** When $h$ increases, $\hat{g}_h(t)$ becomes smoother since $\lvert Bias(\hat{g}_h(t)) \rvert$ increases and $Var(\hat{g}_h(t))$ decreases. When $h$ decreases, $\hat{g}_h(t)$ oscillates seriously since $\lvert Bias(\hat{g}_h(t)) \rvert$ decreases and $Var(\hat{g}_h(t))$ increases.

(c2) smoothing spline estimator: (special case: natural cubic spline)

$$\hat{g}_\lambda(t) = \arg\min_g\left(\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Y_{ij}-g(t_{ij}))^2 + \lambda\int_{T_L}^{T_U}(g^{(k)}(v))^2 dv\right) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i} S_\lambda(t,t_{ij})Y_{ij},$$

where $\lambda$ is a non-negative smoothing parameter and $S_\lambda(t,u)$ is a spline function.

**Remark**. $\lambda$ controls the trade-off between smoothness and goodness of fit to the data.

(c3) local polynomial estimator:

Assume that $g(\cdot)$ is $(k+1)$ times differentiable at $t$. Then, $g(s) \approx \sum_{j=0}^{k}\frac{g^{(j)}(t)}{j!}(s-t)^j$ for $s$ in a neighborhood of $t$. Let $\beta_j \triangleq \frac{g^{(j)}(t)}{j!}$, the local polynomial estimator can be derived as $(\tilde{\beta}_0,\cdots,\tilde{\beta}_k) = \underset{(\beta_0,\beta_1,\cdots,\beta_k)}{\arg\min}\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(Y_{ij}-\sum_{j=0}^{k}\frac{g^{(j)}(t)}{j!}(t_{ij}-t)^j)^2 k_h(t-t_{ij})$, where $k_h(v) = \frac{1}{h}k(\frac{v}{h})$. It implies that $\tilde{g}_h(t) = \tilde{\beta}_0$.

**Remark.**

$k=0$: local constant fit, i.e. kernel estimator.

$k=1$: local linear fit. $\tilde{g}_h(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\omega(t,t_{ij})Y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\omega(t,t_{ij})}$,

where $\omega(x,x_{ij}) = k_h(t-t_{ij})\{S_{n,2}+(t-t_{ij})S_{n,1}\}$ with $S_{n,j} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}k_h(t-t_{ij})(t_{ij}-t)^j$.

(c4) locally weighted scatter plot smoothing (LOWESS):

Let $k(t) = \frac{70}{81}(1-|t|^3)^3\ 1_{(|t|\le 1)}$ and $k^*(t_{i_1j_1},t_{i_2j_2}) = k(\frac{1}{h(t_{i_2j_2})}(t_{i_1j_1}-t_{i_2j_2}))$.

Step1. Compute

$$(\tilde{\beta}_0(t_{ij}),\tilde{\beta}_1(t_{ij}),\cdots,\tilde{\beta}_k(t_{ij})) = \underset{(\beta_0,\beta_1,\cdots,\beta_k)}{\arg\min}\frac{1}{N}\sum_{i_1=1}^{n}\sum_{j_1=1}^{n_{i_1}}(Y_{ij}-\sum_{l=0}^{k}\beta_l(t_{i_1j_1}-t_{ij})^l)\frac{k^*(t_{i_1j_1},t_{ij})}{h(t_{ij})}$$

and obtain $\hat{Y}_{ij} = \tilde{\beta}_0(t_{ij})$.

Step 2. Let $Y_{ij} = Y_{ij}-\hat{Y}_{ij}$, $M = median\{|Y_{ij}|: 1\le i\le n\ ;\ 1\le j\le n_i\}$, and $\delta_{ij} = B(\frac{Y_{ij}}{6M})$,

where $B(t) = (1-t^2)^2\ 1_{(|t|\le 1)}$.

Step 3. Compute

$$(\tilde{\beta}_0^*(t_{ij}), \tilde{\beta}_1^*(t_{ij}), \cdots, \tilde{\beta}_k^*(t_{ij})) = \underset{(\beta_0, \beta_1, \cdots, \beta_k)}{\arg\min} \frac{1}{N} \sum_{i_1=1}^{n} \sum_{j_1=1}^{n_{i_1}} (Y_{ij} - \sum_{l=0}^{k} \beta_l (t_{i_1 j_1} - t_{ij})^l) \frac{\delta_{ij} k^*(t_{i_1 j_1}, t_{ij})}{h(t_{ij})}$$

and obtain $\hat{Y}_{ij}^* = \tilde{\beta}_0^*(t_{ij})$.

Step 4. Repeat Steps 2 and 3 until $\hat{Y}_{ij}^*$'$s$ converge.

**Bandwidth or Smoothing parameter selection:**

Find $h$ (or $\lambda$) to minimize the average predictive squared error (PSE) defined by $PSE(h) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{m_i} E[(Y_{ij}^* - \hat{g}_h(t_{ij}))^2]$, where $Y_{ij}^*$ is a new observation at $t_{ij}$.

**Remark.** $PSE(h)$ can be estimated by the cross-validation criterion:

$$CV(h) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{m_i} E[(Y_{ij} - \hat{g}_{h(-i)}(t_{ij}))^2],$$

where $\hat{g}_{h(-i)}(t)$ is an estimator of $g(t)$ without including the $(i, j)$th observation.

## Exploring Correlation Structure –

**Quantitative responses:**

(a) $t_{ij}$'$s$ occur at the regularly spaced time point $t_1, \cdots, t_m$.

Compute the residuals $r_{ij}$ based on the saturated model $h(x_{ij}; \beta)$ and draw the scatter plot matrices of $r_{ij}$ against $r_{ik}$ for $j < k = 2, \cdots, m$. The correlation of errors at $t_j$ and $t_k$ is estimated by

$$\hat{Corr}(\varepsilon_j, \varepsilon_k) = \left( \frac{\sum_{i=1}^{n} \sum_{j_1 \neq j_2} r_{ij_1} r_{ij_2} \delta_{ij_1 j_2}^{(j,k)}}{\sum_{i=1}^{n} \sum_{j_1 \neq j_2} \delta_{ij_1 j_2}^{(j,k)}} \right) \Bigg/ \left( \frac{\sum_{i=1}^{n} \sum_{j_1 \neq j_2} r_{j_1}^2 \delta_{ij_1}^{(j)}}{\sum_{i=1}^{n} \sum_{j_1 \neq j_2} \delta_{ij_1}^{(j)}} \right)^{1/2} \left( \frac{\sum_{i=1}^{n} \sum_{j_1 \neq j_2} r_{j_2}^2 \delta_{ij_2}^{(k)}}{\sum_{i=1}^{n} \sum_{j_1 \neq j_2} \delta_{ij_2}^{(k)}} \right)^{1/2},$$

where $\delta_{ij_1 j_2}^{(j,k)} = 1(t_{ij_1} = t_j, t_{ij_2} = t_k)$, $\delta_{ij_1}^{(j)} = 1(t_{ij_1} = t_j)$, and $\delta_{ij_2}^{(k)} = 1(t_{ij_2} = t_k)$.

**Remark.** If the errors $\varepsilon_j = Y_j - h(X_j; \beta)$'$s$ have a constant variance and the correlation of errors at $t_j$ and $t_k$ is a function of $|t_j - t_k|$, say, $\rho(|t_j - t_k|)$, the error process is said to

be weakly stationary, where $\rho(\cdot)$ is called the autocorrelation function.

Keyword:
Strictly stationary: The joint probability function is associated with the time only.
Let $u$ be the time lag between any two time point in $\{t_1,\cdots,t_m\}$. The correlation $\rho(\cdot)$ can be estimated by

$$\hat{\rho}(u)=\frac{(\sum_{i=1}^{n}\sum_{j_1\neq j_2} r_{ij_1}r_{ij_2}\delta_{ij_1j_2}^{*(u)} \Big/ \sum_{i=1}^{n}\sum_{j_1\neq j_2}\delta_{ij_1j_2}^{*(u)})}{(\sum_{i=1}^{n}\sum_{j=1}^{n_i} r_{ij}^2 \Big/ N)}, \text{ where } \delta_{ij_1j_2}^{*(u)}=1\,[\,|\,t_{ij_1}-t_{ij_2}\,|=u\,].$$

(b) Irregularly spaced time points $t_{ij}$: (consider $Y(t)=g(t)+\varepsilon(t)$)

(b1) Let $\sigma^2(t)=Var(\varepsilon(t))$, $Cov(t,s)=Cov(\varepsilon(t),\varepsilon(s))$, and $\rho(t,s)=\dfrac{Cov(t,s)}{\sigma(t)\sigma(s)}$.

$\sigma^2(t)$ and $Cov(t,s)$ can be estimated separately via $\hat{\sigma}_h^2(t)=\dfrac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i} r_{ij}^2 k_h(t-t_{ij})$

and $\hat{Cov}_{h_1h_2}(t,s)=\dfrac{1}{N}\sum_{i=1}^{n}\sum_{j_1\neq j_2} r_{ij_1}r_{ij_2}k_{h_1}(t-t_{ij_1})k_{h_2}(s-t_{ij_2})$, where $r_{ij}=Y_{ij}-\tilde{g}_h(t_{ij})$.

(b2) $\varepsilon(t)$ is weakly stationary:

Define variogram $r(u)=\dfrac{1}{2}E[(\varepsilon(t)-\varepsilon(t-u))^2]$, $u>0$.

**Remark.** When $\varepsilon(t)$ is weakly stationary, $r(u)=\sigma^2\{1-\rho(u)\}$, where $\sigma^2=Var(\varepsilon(t))$.

Sample variogram: $v_{ijk}=\dfrac{1}{2}(r_{ij}-r_{ik})^2 \underset{corresponding}{\leftrightarrow}$ time difference $u_{ijk}=t_{ik}-t_{ij}$.

$\hat{r}(u)$ can be estimated via smoothing the data $(u_{ijk},v_{ijk})$ $j\neq k$ with

$$\hat{\sigma}^2=(0.5\sum_{i_1>i_2}\sum_{j_1,j_2}(r_{i_1j_1}-r_{i_2j_2})^2)\Big/(\sum_{i_1>i_2}\sum_{j_1,j_2}1).\ \text{Thus, } \hat{\rho}(u)=1-\frac{\hat{r}(u)}{\hat{\sigma}^2}.$$

**Categorical Responses – Binary response**

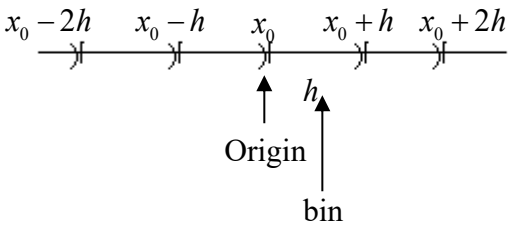
Define the log-odds-ratio, called lorelogram, as $LOR=\ln(r(Y_j,Y_k))$, where

$$r(Y_i,Y_k)=\frac{P(Y_i=1,Y_k=1)P(Y_i=0,Y_k=0)}{P(Y_i=1,Y_k=0)P(Y_i=0,Y_k=1)}.$$

## Appendix

$X_1, \cdots, X_n \overset{iid}{\sim} f(x)$

**Histograms:**




The histogram is defined by:

$$\hat{f}(x) = \frac{1}{h}\{\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in [x_0 + mh, x_0 + (m+1)h)\}, \quad x \in [x_0 + mh, x_0 + (m+1)h)$$

$$= \frac{1}{nh}\sum_{i=1}^{n} 1(X_i \in [x_0 + mh, x_0 + (m+1)h)$$

**Naive estimator:**

From the definition of the p.d.f. $f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$, one has

$$\hat{f}_h(x) = \frac{1}{2h}\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in (x - h, x + h)) = \frac{1}{2nh}\sum_{i=1}^{n} 1(X_i \in (x - h, x + h))$$

$$= \frac{1}{nh}\sum_{i=1}^{n} \omega(\frac{x - X_i}{h}) \text{ with } \omega(t) = \frac{1}{2} 1_{(|t| \le 1)}$$

It's not a continuous function. (jumps occur at the points $X_i \pm h$)

**Kernel estimator:**

Substituting a kernel function $k(\cdot)$ for the weight function $\omega(\cdot)$ which satisfies $\int k(x)dx = 1$, the kernel estimator is defined to be $\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x - X_i}{h})$, where $h$ is the window width and is called the smoothing parameter or bandwidth.

Remark: Generally, $k(\cdot)$ is assigned to be a symmetric probability density function.

**The nearest neighbor method:**

For a sample of size $n$, we would expect about $2rnf(t)$ observations to fall within the

interval $[t-r, t+r], \forall r > 0$.

Let $d_k(t)$ be the $k$th nearest distance of $t$ and $\{X_1, \cdots, X_n\}$. It implies that there are exact $k$ observations falling in the interval $[t-d_k(t), t+d_k(t)]$.

Thus, the nearest neighbor estimator can be obtained as $\hat{f}_{nk}(t) = \dfrac{k}{2nd_k(t)}$.

**Remark.** The naive estimator is based on the number of observations falling in a box of fixed width, while the nearest neighbor estimator is inversely proportional to the size of the box need to contain a given number of observations.