a column: a variable
a row: a unit — raw data matrix →

Binomial Data

- From each unit, observe data

$$(x_{j1}, x_{j2}, \ldots, x_{jm}, z_j),\ j = 1, 2, \ldots, K,$$

jth unit. In total, K units. fixed.
# of explanatory variables

➢ covariates (explanatory variables): $x=(x_1,\ldots,x_m)$ $\sim$ Bernoulli(p_{x_j})

➢ response: z_j's, *independent* random variables, z_j equals 1 with probability p_{x_j}, and 0 with probability $1-p_{x_j}$

sucess, 0 ← failure

assume p only depends on x. i.e., same x ⇒ same p

LM: $\mu_x = E(y_x) \leftrightarrow x$

⇒ objective: determine the relationship of $x=(x_1,\ldots,x_m)$ to p_x cf.

(cf., linear model)
① same objective (i.e., study the relationship btwn the parameters in the distribution of response & covariates).
② But, different in the concept of $Y = X\beta + \varepsilon$, i.e., response is the sum of the 2 component $X\beta$ & ε. 規 隨

- Covariate classes

convert (suff. stat.) categorical response (z's) into count response (y's)

➢ Sometimes, several units have same values of covariates — same x

$K = n_1 + \cdots + n_k$ ⇐ # of units in each group

➢ Split the total sample K into k groups of size $n_1, \ldots, n_k$, where each observation within a group has same values of covariates

➢ The groups are called *covariate classes*

Assume they are fixed values (i.e., not random) for now

- After grouping, data can be expressed as — ith covariate class

# of covariate classes

different i, distinct x_i

$$(x_{i1}, x_{i2}, \ldots, x_{im}, y_i),\ i = 1, 2, \ldots, k,$$

x_i — # (sum) of z_j's = 1 in ith group

∵ ✱ in LNp.3-1

➢ response: y_i's, independent r.v.'s with distribution $B(n_i, p_{x_i})$

- Advantage of grouping → $(z_1, \ldots, z_K)$ categorical $\xrightarrow{\text{suff. stat.}}$ $(y_1, y_2, \ldots, y_k)$ counts (numerical)

➢ Data is easier to view and store in the form of covariate classes (K units → k classes) — concept: replicates

➢ Grouped case and n_i's are large ⇒ can use Normal asymptotic theory or fit a linear model

① $y_x \sim B(n, p)$ $\approx N(np, np(1-p))$ as $n \to \infty$ (next slide)
② likelihood approach for estimation (MLE), test (LRT) & confidence interval (LNp.3-8~12)

n_i's → ∞ ⇓ ✗ K → ∞

(cf., un-grouped case, regarded as n_i=1 for all i — a unit, a group

⇒ different asymptotics) — K → ∞

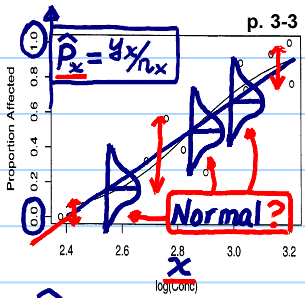
- Warning: z_j's $\xrightarrow[\text{transformation}]{\text{non-invertible}}$ y_i's ⇒ information lost — useful? important?

variable x range: 0~100 precision: 10 decimal places

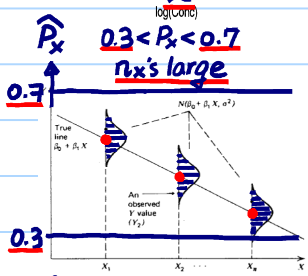
➢ $y_i \sim B(n_i, p_{x_i})$ when the corresponding z_j's are (1) independent, (2) identically distributed as (3) Bernoulli with same p_{x_i}. It should be checked whether the 3 conditions hold.

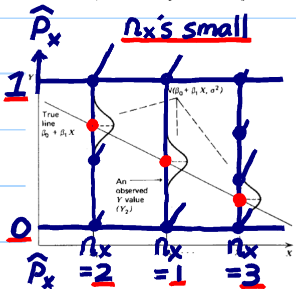
actually, only (1) & (2)

- Be careful about reducing raw data to covariate classes, e.g., the reduction throws away the time order of z_j's which may be relevant
  - x_i → 000111100111110··· (time): might violate (1), 0, 1 clustered together
  - x_i → 010010 ··· 110111 (time): might violate (2), less 1's ... more 1's

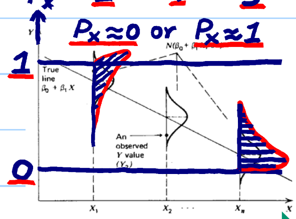


- A linear model approach for binomial data: $\hat{P}_x = y_x/n_x = X\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2 I)$
  - MLE at uni-variate (1 group)
  - $E(y_x/n_x) = P_x$
  - or $t(\hat{P}_x)$, t: transformation, e.g., logit
  - cf. a problem: $Var(y_x/n_x) = \frac{P_x(1-P_x)}{n_x}$ ⇒ not a constant, change with mean.
  - Q: when is the approach appropriate? (i.e., approximation OK?)
    - $\{0, \frac{1}{n_x}, \frac{2}{n_x}, \dots, \frac{n_x-1}{n_x}, 1\}$ ⇒ The pmf shape of the distribution of y_x/n_x is similar to the pdf shape of Normal (When x is fixed at each x_i)
      - $B \approx N$ when n_x large (fast $p\approx 0.5$, slow $p \approx 0$ or 1)
    - p_x's do not change too much (say, $0.3 < p_x < 0.7$) (When x varies)
      - also, $logit(p) \approx a + bp$ within the range (LN p.3-5)





- Some problems with this approach
  - Predicted probability may > 1 or < 0 ($-\infty < X\beta < \infty$, especially in extrapolation)
  - Variance of binomial is not constant → can use WLS or variance-stabling transformation



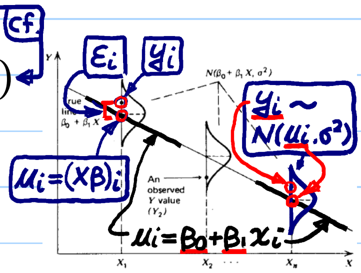
- Normal approximation might be too much a stretch when n_i's not large enough or $p_x \approx 1$ or 0 → Recall. LM data with response having upper/lower bound, & many observations of the response are close/equal to the bounds.
- Some of these problems could be corrected by using transformation and weighting
  - e.g. $\log(\hat{P}_x/(1-\hat{P}_x)) = X\beta + \varepsilon$, $0 \le$ predicted prob. ≤ 1
  - for {non-constant variance, non-linearity}
  - for non-constant variance

A suit fitting better for binomial response

2/26

Generalized Linear Model for Binomial Data

- Recall: linear model
  - binomial response $\sim B(n_x, P_x)$: $\{0, 1, \dots, n_x\} \ni y_x = X\beta + \varepsilon$ (cf.)
  - model description 1: $Y = X\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2 I)$ (equivalent in LM)
  - model description 2: $Y \sim N(X\beta, \sigma^2 I)$



  - response 隨機: $Y \sim N(\mu_x, \Sigma_x)$ — random; parameters (unknown but fixed values)
  - covariates 規律: $X\beta = \sum_{i=1}^{p} \beta_i \cdot h_i(X_1, \dots, X_m) \equiv \eta_x(\beta)$ — coefficients (parameters); known functions (base functions); matrix form; functional form; expressed as a function or vector
    - a function of x with a linear structure btwn h_i's & β_i's (of main interest in regression)
  - build the link btwn parameters in Y & $X\beta$: $\mu_x = \eta_x,\ \Sigma_x = \sigma^2 I$ — can be assigned a function of x (with parameters) if necessary
- Q: which description can be generalized to binomial data?