Regression Assumptions

- linear : $E(Y|X) = X\beta$
  - polynomial reg ch6
  - transformation ch7
- Y – continuous random variable

  e.g. $Y = 0, 1$

  $Y =$ pos. integers

  integer → $Y = \beta_0 + \underbrace{\beta_1 X_1 + \cdots + \beta_1 X_p + e}_{\text{range is unlimited never be integers}}$

  - logistic reg
  - Poisson reg ~~ch12~~ STAT 4006
- Variance is constant $e_i \sim N(0, \underbrace{\sigma^2}_{\text{constant}})$
  - Weighted Least Square ch5
- No outlier & influential points
  - ~~ch9~~

5.1 Weighted Least Square (WLS)

- Model

$$E(Y \mid X = x_i) = \beta' x_i$$

$$Var(Y \mid X = x_i) = \frac{\sigma^2}{w_i}$$

assumed known

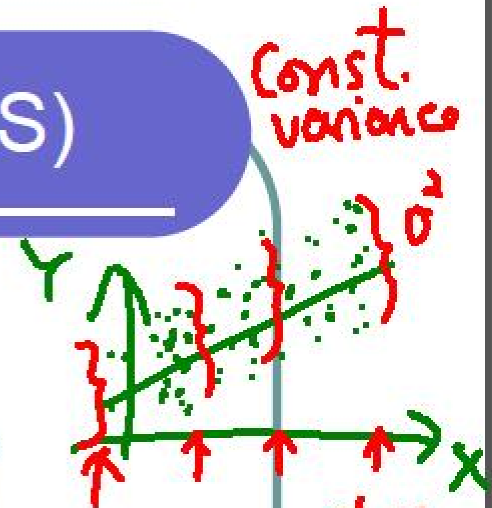
- Alternative representation

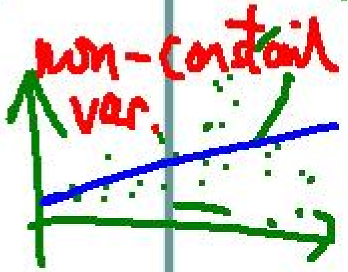
$e_i \sim^{id}$ $var(e_i) = \frac{\sigma^2}{w_i}$

$$Y = X\beta + e, \qquad Var(e) = \sigma^2 W^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{w_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{w_2} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{w_n} \end{pmatrix}$$

$n \times 1$

  - Errors are independent but not identically distributed





5.1 Weighted Least Square (WLS)

- Model

$$E(Y \mid X = x_i) = \beta' x_i$$

$$Var(Y \mid X = x_i) = \frac{\sigma^2}{w_i}$$

assumed known

- Alternative representation

$$Y = X\beta + e, \qquad Var(e) = \sigma^2 W^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{w_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{w_2} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{w_n} \end{pmatrix}$$

  - Errors are independent but not identically distributed

$$W = \begin{pmatrix} w_1 & & 0 \\ & w_2 & \\ 0 & & w_n \end{pmatrix}$$

assume that $Y_i \overset{iid}{\sim} \sigma^2$

	mean GPA G	mean IQ Q	class size
STAT 3003	3.3	130	40
STA 3004	3.4	140	50
STA 3005	3.5	150	60
STA 3007	3.7	170	30
STAT 2008	3.999	199	112

$$G_1 = \frac{\sum_1^{40} Y_i}{40} \qquad Var(G_1) = \frac{\sigma^2}{\underline{40}}$$

$$Var(Y|x_i) = \frac{\sigma^2}{W_i}$$

$$\Rightarrow W_1 = 40$$

$$G_5 = \frac{\sum_1^{120} Z_i}{120} \qquad Var(G) = \frac{\sigma^2}{\underline{120}}$$

$$\Rightarrow W_5 = 120$$

Page 5 of 13

Recording 00:00:00

	Y = monthly rainfall	temp
Jan	$Z_1 + \cdots + Z_{31}$	.
Feb	$Z_1 + \cdots + Z_{28}$	.
March	$Z_1 + \cdots + Z_{31}$	.
April	.	.
	⋮	.

Z_i = daily rainfall
iid σ^2

$Y_1 = Z_1 + \cdots + Z_{31}$

$Var(Y_1) = 31\sigma^2$

$Y_2 = Z_1 + \cdots + Z_{28}$

$Var(Y_2) = 28\sigma^2$

$Var(Y_i) = \frac{\sigma^2}{W_i}$

$\Rightarrow W_1 = \frac{1}{31}$

$W_2 = \frac{1}{28}$

$\vdots$

Page
8 of 13

Recording
00:00:00

5.1 Weighted Least Square (WLS)

- Model $E(Y \mid X = x_i) = \beta' x_i,$ $Var(Y \mid X = x_i) = \dfrac{\sigma^2}{w_i}$ assumed known

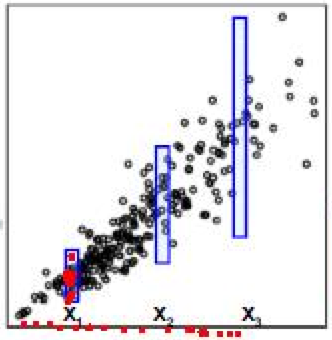
- Examples of known w_i
  - The i-th Observations is an average of n_i variables

    $$Y_i = \frac{z_{i1} + z_{i2} + \ldots + z_{in_i}}{n_i}, \quad Var(Y_i \mid X = x_i) = \frac{Var(z_{i1} \mid X = x_i)}{n_i} = \frac{\sigma^2}{n_i} \Rightarrow w_i = n_i$$

  - The i-th Observations is a sum of n_i variables

    $$Y_i = z_{i1} + z_{i2} + \ldots + z_{in_i}, \quad Var(Y_i \mid X = x_i) = n_i Var(z_{i1} \mid X = x_i) = n_i \sigma^2 \Rightarrow w_i = 1/n_i$$

  - When sample size is large, estimate $Var(Y \mid X = x_i)$ by computing sample variance of the Y with X close to x_i.
  - Subject knowledge…
  - Guess from the scatterplot…



X_1	$Var(Y_1)$
X_2	$V(Y_2)$
X_3	$V(Y_3)$
X_4	$V(Y_4)$
X_5	$\vdots$

ordinary → OLS : $RSS = \sum (y_i - \underbrace{\hat{y}_i}_{X\beta})^2$

5.1 Estimators for the parameters

$Var(y_i) = \dfrac{\sigma^2}{w_i}$

- Residual sum of square
  - Standardized by the variance of each observation

$$\mathrm{RSS}(\beta) = \sum \frac{(y_i - \hat{y}_i)^2}{Var(y_i)} = \frac{1}{\sigma^2} \sum w_i (y_i - \hat{y}_i)^2$$

$$\propto \sum w_i (y_i - \hat{y}_i)^2$$

$$= \begin{pmatrix} y_1 - \hat{y}_1 & \cdots & y_n - \hat{y}_n \end{pmatrix} \begin{pmatrix} w_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_n \end{pmatrix} \begin{pmatrix} y_1 - \hat{y}_1 \\ \vdots \\ y_n - \hat{y}_n \end{pmatrix}$$

$$= (Y - X\beta)' W (Y - X\beta)$$

5.1 Estimators for the parameters

- Residual sum of square

$$\sqrt{W} = \begin{pmatrix} \sqrt{w_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sqrt{w_n} \end{pmatrix}, \qquad \sqrt{W}\times\sqrt{W} = W$$

$\sqrt{W}$ $\sqrt{W}$

$$\text{RSS}(\beta) = (Y - X\beta)'W(Y - X\beta) = (\sqrt{W}Y - \sqrt{W}X\beta)'(\overbrace{\sqrt{W}Y}^{Z} - \overbrace{\sqrt{W}}^{M}X\beta)$$

Recall: In multiple linear regression,

$\hat{\beta} = (X'X)^{-1}X'Y$ minimizes $(Y - X\beta)'(Y - X\beta)$

$\hat{\beta}_w = (M'M)^{-1}M'Z$ $(Z-M\beta)'(Z-M\beta)$

- Therefore, the WLS estimator is

$$\boxed{\hat{\beta}_w = ((\sqrt{W}X)'\sqrt{W}X)^{-1}(\sqrt{W}X)'\sqrt{W}Y = (X'WX)^{-1}X'WY}$$

$M = \sqrt{W}X$

$Z = \sqrt{W}Y$