PS2 → grading
Quiz2 → tomorrow
PS3 → post / Due late next week

Lab 6

Lab 7 and 8

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

# Lecture 11

## Finish Model Checking

## Implementation of WLS and robust variance estimation in R

# Key Assumptions by Order of Importance

- E(Y|X) = Xβ, i.e. the mean model is "correctly" specified
  - Misspecification of Xβ can lead to biased β / misinterpretations
  - Omitted variable Bias
  - Correct functional form for continuous X

- Residuals are independent
  - This assumption is violated due to the design of the study
  - Longitudinal study
  - Clustered design
  - Show today: ignoring the correlation will impact $Var(\hat{\beta})$ and derive weighted least squares

- Variance of residuals is constant
  - Often the variance is a function of some X
  - Show today: same impact and solution as violation of independence

WLS

- Residuals are normally distributed
  - CLT, bootstrap procedure

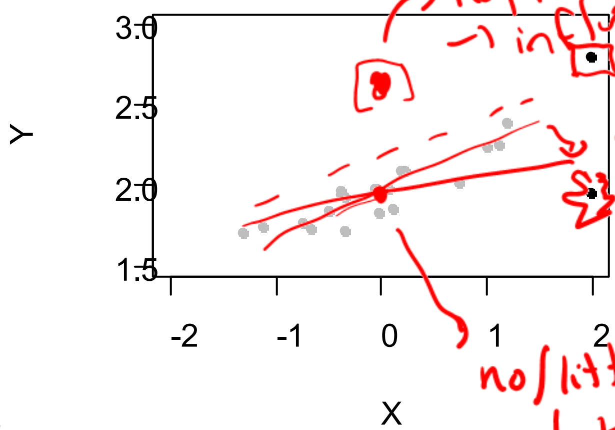
↓ Bootstrap is also a solution
→ fit OLS → $\hat{\beta}$
→ resample people
fit OLS → $\hat{\beta}_k$

- There are not a small number of highly influencial observations
  - Sensitivity analyses

# Leverage and Influence

- Leverage: A measure of how far an individual's predictors (X_i) are from the mean X_i
  - Hat matrix: $h_{ii} = \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}$

- Influence: An observation $(Y_i, X_i)$ such that including this value would greatly change the fitted values: $\hat{\beta}$ and $\hat{Y}$.




no/little leverage
influence on estimate of $\beta_0$

high leverage
no/little influence

high leverage
high influence

no/little leverage
no/little influence

# Influence statistics

There are several influence statistics that are used in practice:

DFBETA$_{ij}$

- $DBETA_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}$
- $DBETAS_{ij} = \dfrac{DBETA_{ij}}{\hat{se}(\hat{\beta}_{j(-i)})}$ *

DFBETAS$_{ij}$

- $DFIT_i = \hat{Y}_i - \hat{Y}_{i(-i)}$
- $DFITS_i = \dfrac{DFIT_i}{\hat{se}(\hat{Y}_{i(-i)})}$ *

i = observation, 1, ..., n
j = coefficient, 0, 1, 2, ..., p

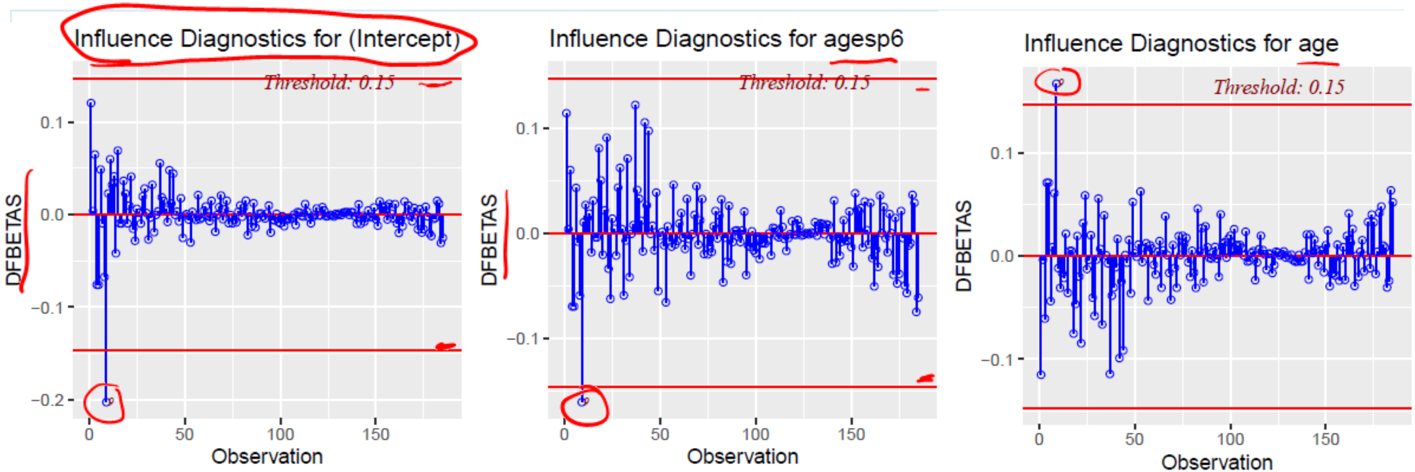
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots$

→ use all n → $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$

→ remove obs i → $\hat{\beta}_{0(-i)}, \hat{\beta}_{1(-i)}, \dots$

→ $\hat{Y}_{i(-i)} = \hat{\beta}_{0(-i)} + \hat{\beta}_{1(-i)} X_{1i} + \dots$ $\hat{\beta}_{p(-i)}$

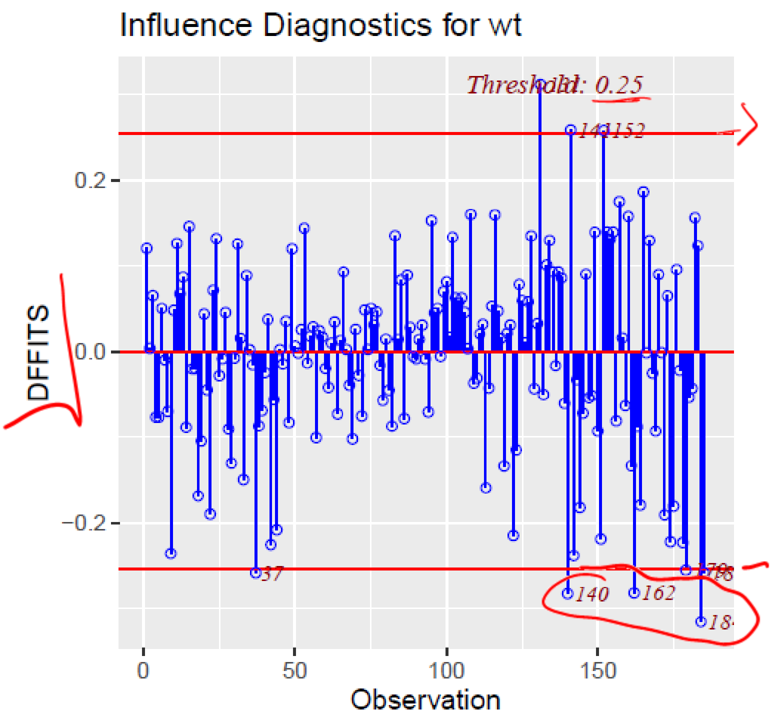
# Example: Nepali Anthropometry Data




1st visit data $n=185$ obs

$$wt = \beta_0 + \beta_1 age + \beta_1 (age-6)^+ + \varepsilon$$

# Example: Nepali Anthropometry Data

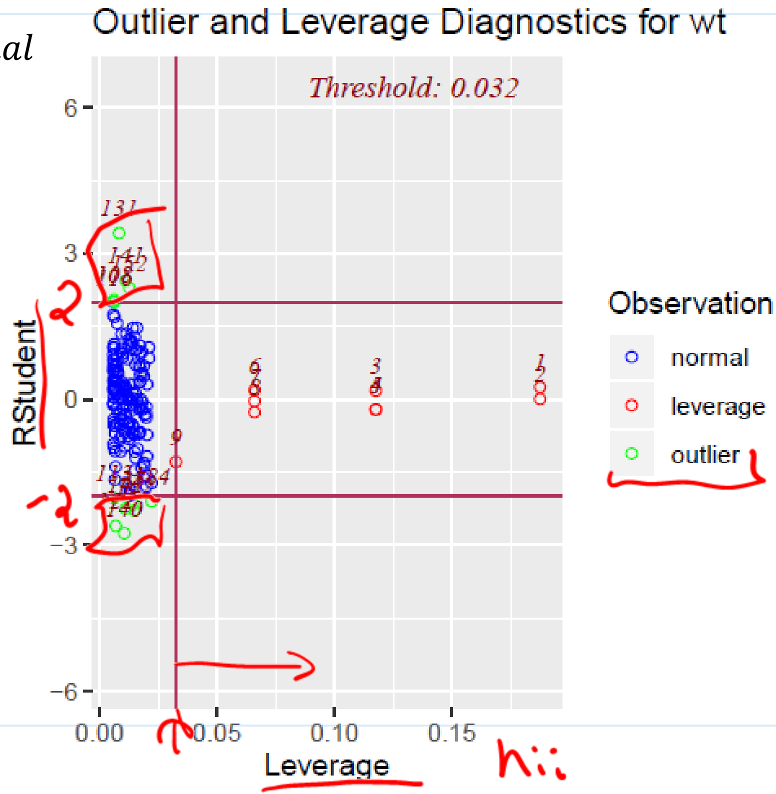



$$\frac{\hat{y}_i - \hat{y}_{i(-i)}}{se(\hat{y}_{i(-i)})}$$

# Example: Nepali Anthropometry Data

- $RStudent: studentized\ residual$
- $$\frac{y_i - \hat{y}_{(i)}}{se(y_i - \hat{y}_{(i)})} = \frac{\boxed{y_i - \hat{y}_i}}{\sqrt{\underbrace{MSE_{(i)}}_{(-i)}(1 - h_{ii})}}$$

$$\frac{Y_i - \hat{Y}_{i(-i)}}{se(Y_i - \hat{Y}_{i(-i)})}$$



# Implementation of WLS in R

- For the remainder of the lecture, we will work through some analyses to demonstrate how to fit WLS in R

$\Sigma_{a \times a} \begin{bmatrix} V_1 & & \\ & V_2 & \\ & & V_3 \end{bmatrix} \quad V_{i_{3\times 3}}$

- In addition, I will show one approach for obtained robust variance estimates for different working correlation models. Here we will use the *gee* package in R.

- See Handout 11.

$$Y = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix} \qquad \Sigma = Var(Y) = \begin{bmatrix} \overbrace{\begin{bmatrix} Var(Y_{11}) & Cov(Y_{11}, Y_{12}) \\ * & Var(Y_{12}) \end{bmatrix}}^{Var(\underset{\sim}{Y_1})} & \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix} \\ \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} \begin{bmatrix} Var(Y_{21}) & \overset{*}{C}(Y_{21}, Y_{22}) \\ * & Var(Y_{22}) \end{bmatrix} & \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} \\ \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix} & \begin{bmatrix} Var(Y_{31}) & \\ & Var(Y_{32}) \end{bmatrix} \end{bmatrix}_{6\times 6} \qquad \sigma^2 I$$