JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Lecture 10

Model Checking and Key Extensions continued

Key Assumptions by Order of Importance

- 1 $E(Y|X) = X\beta$, i.e. the mean model is "correctly" specified
  - Misspecification of $X\beta$ can lead to biased β / misinterpretations
  - Omitted variable Bias
  - Correct functional form for continuous X $\Rightarrow$ Two plots $\hat{R}$ vs X_j, $\hat{R}$ vs $\hat{Y}$
- 2 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
- 3. Variance of residuals is constant
  - Often the variance is a function of some X
  - Show today: same impact and solution as violation of independence
- 4 Residuals are normally distributed
  - CLT, bootstrap procedure
- 5 There are not a small number of highly influencial observations
  - Sensitivity analyses

Independence Assumption

- Driven by the design of the study

- Longitudinal design
  - Sample/recruit i = 1, …., m subjects
  - Follow the subjects over time and measure the outcome of interest at n_i occasions
  - Data: (Y_ij, X_ij), where Y_ij is the jth value of the outcome for the ith subject
  - Level 1: Subject, Level 2: time

Y_{ij} i = level 1
j = level 2

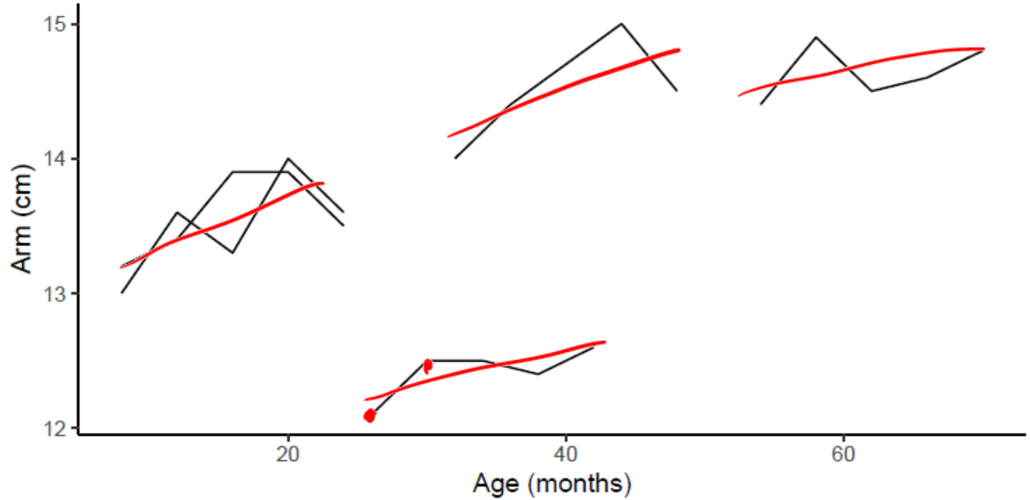
- Clustered design
  - Sample/recruit i = 1, …, m clusters (e.g. clinics, schools, villages)
  - Sample/recruit j = 1, …, n_i individuals within the ith cluster
  - Data (Y_ij, X_ij), where Y_ij is the outcome for the jth individual from cluster i
  - Level 1: Cluster, Level 2: subject

- Why do we care?
  - Observations Y_ij from the same level 1 unit tend to be correlated
  - Why? Context, genetics, sociodemographic factors, governmental policies, weather, etc.

Example: Nepali Anthropometry Data

- Design: i = 1, …, m = 200 children each measured at baseline (j = 1) and then every 4 months for 4 follow-up visits (j = 2, 3, 4, 5).

```
ggplot(d5,aes(x=age,y=arm,group = factor(id))) +
    geom_line() +
    labs(x='Age (months)', y ='Arm (cm)') +
    theme_classic()
```



Checking the Independence Assumption

- Don't need to, we know the independence assumption is violated based on knowledge of the design
- We can explore covariance/correlation in the observed data
- Example: Consider the Nepali Anthropometry data where we have data for i = 1, …, m = 200 children each measured at baseline (j = 1) and then every 4 months for 4 follow-up visits (j = 2, 3, 4, 5).
  - Step 1: Regress Y on X assuming independence and estimate β and R
  - Step 2: Plot $\hat{R}_{ij}$ vs. $\hat{R}_{ik}$ for all j,k
  - Compute $Cov(\hat{R}_{ij}, \hat{R}_{ik}) = \sqrt{Var(\hat{R}_{ij})} \times \sqrt{Var(\hat{R}_{ik})} \times Corr(\hat{R}_{ij}, \hat{R}_{ik})$
  - Or standardize the residuals and plot $Corr(\hat{R}_{sij}, \hat{R}_{sik})$
- Compute the autocorrelation function

$$\underbrace{\rho(u)}_{\text{lag in measurement times}} = Corr(\hat{R}_{sij}, \hat{R}_{sij+u})$$

How do we rethink the model?

Our current assumption

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{bmatrix} \sim MVN\left(\begin{pmatrix} X_1\beta \\ X_2\beta \\ X_3\beta \\ X_4\beta \\ X_5\beta \\ X_6\beta \end{pmatrix}, \left[\begin{array}{ccc|ccc} \sigma^2 & & & & & \\ & \sigma^2 & & & 0 & \\ & & \sigma^2 & & & \\ \hline & & & \sigma^2 & & \\ & 0 & & & \sigma^2 & \\ & & & & & \sigma^2 \end{array}\right] \right)$$

$Corr(\varepsilon_i, \varepsilon_j) = 0$

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{bmatrix} \begin{matrix} \underset{\sim}{Y}_1 \\ \\ \underset{\sim}{Y}_2 \end{matrix} \sim MVN\left(\begin{pmatrix} X_{11}\beta \\ X_{12}\beta \\ X_{13}\beta \\ X_{21}\beta \\ X_{22}\beta \\ X_{23}\beta \end{pmatrix}, \left[\begin{array}{c|c} V_1 & 0 \\ \hline 0 & V_2 \end{array}\right] \right)$$

individuals are independent

How do we rethink the model?

$$Var(\underset{\sim}{Y}_1) = Var\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \end{pmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

$$Var(Y_{ij}) = \sigma_j^2$$

$$Cov(Y_{ij}, Y_{ik}) = \sigma_{jk}$$

$$Corr(Y_{ij}, Y_{ik}) = \frac{\sigma_{jk}}{\sqrt{\sigma_j^2 \sigma_k^2}}$$

$$\begin{pmatrix} \underset{\sim}{Y}_1 \\ \underset{\sim}{Y}_2 \\ \vdots \\ \underset{\sim}{Y}_m \end{pmatrix} \sim MVN\left[\begin{pmatrix} X_1\beta \\ X_2\beta \\ \vdots \\ X_m\beta \end{pmatrix}, \underbrace{\begin{bmatrix} V_1 & & \\ & V_2 & 0 \\ 0 & & \ddots \\ & & & V_m \end{bmatrix}}_{\Sigma}\right]$$

What if we apply least squares to correlated data?

True model: $\underbrace{Y}\sim MVN(\underbrace{X\beta}, \underbrace{\Sigma}) \sim \begin{bmatrix} v_1 & & 0 \\ & \ddots & \\ 0 & & v_m \end{bmatrix}$

$\Rightarrow$ We obtain $\hat{\beta}_{LS} = (X'X)^{-1}X'Y$

$$\underbrace{E(\hat{\beta}_{LS})} = E\left[(X'X)^{-1}X'Y\right] = (X'X)^{-1}X'E(Y)$$

$$= \underbrace{(X'X)^{-1}}\underbrace{X'X}\beta$$

$$= \beta$$

unbiased estimates for β

What if we apply least squares to correlated data?

$$Var(\hat{\beta}_{LS}) = Var\left(\underbrace{(X'X)^{-1}X'}Y\right)$$

$$= [(X'X)^{-1}X']\,\underbrace{Var(Y)}\,[(X'X)^{-1}X']'$$

$$= [(X'X)^{-1}X']\,\Sigma\,[(X'X)^{-1}X']'$$

$$= (X'X)^{-1}X'\Sigma X(X'X)^{-1}$$

$$* \; Var(\hat{\beta}_{LS}) = \underbrace{\sigma^2(X'X)^{-1}} \text{ if } Var(Y) = \sigma^2 I$$

Solution: Weighted least squares

$Y \sim MVN(X\beta, \Sigma)$

Define a weight matrix $\Sigma^{-1/2}$ (symmetric) and a transformed Y and X $\Rightarrow$ in the transformed problem the residuals have var/cov $= I$

$Y^* = \Sigma^{-1/2} Y, \quad X^* = \Sigma^{-1/2} X$

$$\beta^*_{WLS} = (X^{*\prime} X^*)^{-1} X^{*\prime} Y^*$$

$$= \left[(\Sigma^{-1/2} X)' \Sigma^{-1/2} X\right]^{-1} (\Sigma^{-1/2} X)' \Sigma^{-1/2} Y$$

$$= \left[X' \Sigma^{-1/2} \Sigma^{-1/2} X\right]^{-1} X' \Sigma^{-1/2} \Sigma^{-1/2} Y$$

$$= (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} Y \Rightarrow$$ weighted least squares solution

Solution: Weighted least squares

$$\Sigma = \begin{bmatrix} v_1 & & & 0 \\ & v_2 & & \\ & & \ddots & \\ 0 & & & v_m \end{bmatrix} \qquad \Sigma^{-1} = \begin{bmatrix} v_1^{-1} & & & 0 \\ & v_2^{-1} & & \\ & & \ddots & \\ 0 & & & v_m^{-1} \end{bmatrix}$$

an estimate for $\underline{V_i} = (Y_i - X_i\hat{\beta})'(Y_i - X_i\hat{\beta})$

Weighted least squares

- The weighted least squares approach requires that we specify two models:
  - The model for the mean E(Y|X)
  - A model for the Var(Y) = Σ or V_i

$$V_i = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & & \\ \vdots & & \ddots & \\ \sigma_{1n} & & & \sigma_n^2 \end{bmatrix}$$

- Approaches:

1. Unstructured model: Estimate all the required σ_j^2 and σ_{jk} in V_i => this requires many parameters to be estimated. In a balanced design with n assessments: $n(n+1)/2$

Or simplify!

$$V_i = [\mathrm{diag}(V_i)^{-1/2}]' R_i [\mathrm{diag}(V_i)^{-1/2}]$$

$\mathrm{Corr}(Y_{ij}, Y_{ik}) = \rho_{jk}$

where R_i = correlation matrix = $\begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{12} & 1 & & \vdots \\ \vdots & \vdots & \ddots & \\ \rho_{1n} & \rho_{2n} & \cdots & 1 \end{bmatrix}$

* separately model R_i and $\mathrm{diag}(V_i)$

Weighted least squares

- Some parametric models to consider: Common models for R, see Lab 5

exchangeable aka compound symmetry

$$R = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

$$\underbrace{Corr(\varepsilon_{ij}, \varepsilon_{ik})}_{\text{Clustered designs}} = \rho$$

AR-1

$$R = \begin{bmatrix} 1 & \rho^1 & \rho^2 \\ \rho^1 & 1 & \rho^1 \\ \rho^2 & \rho^1 & 1 \end{bmatrix}$$

$$Corr(\varepsilon_{ij}, \varepsilon_{ik}) = \rho^{|j-k|}$$

best for discrete time follow-up studies

continuous time analog → exponential decay

Toeplitz

$$R = \begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{bmatrix}$$

$$Corr(\varepsilon_{ij}, \varepsilon_{ik}) = \rho_{|j-k|}$$

+ many others

Robust Variance Estimate

- Is there an alternative to using weighted least squares?
  - i.e. can we estimate β using least squares and then get an estimate for $Var(\hat{\beta})$ that accounts for the correlation in the data? ↳ unbiased estimate for β

when the truth is $Var(Y) = \Sigma$,

$$Var(\hat{\beta}_{LS}) = \underbrace{(X'X)^{-1}}_{*} \underbrace{X'\Sigma X}_{+} \underbrace{(X'X)^{-1}}_{*}$$

Sandwich estimator

$*\ \sigma^2 (X'X)^{-1}$ based on LS

$+$ Truth

Estimate Σ using the residuals: $(Y - X\hat{\beta})'(Y - X\hat{\beta})$

Robust Variance Estimate

In general, suppose you assume the following model

$\Rightarrow Y \sim MVN(X\beta, \Sigma_m) \Rightarrow$ working model

But the truth is: $Y \sim MVN(X\beta, \Sigma_T)$

Then you use weighted least squares

$$\hat{\beta}_{WLS,m} = (X'\Sigma_m^{-1}X)^{-1} X'\Sigma_m Y \Rightarrow \text{unbiased for } \beta$$

$$Var(\hat{\beta}_{WLS,m}) = (X'\Sigma_m^{-1}X)^{-1} X'\Sigma_m^{-1} \Sigma_T \Sigma_m^{-1} X (X'\Sigma_m^{-1}X)^{-1}$$

$$\hat{Var}(\hat{\beta}_{robust}) = \underbrace{(X'\hat{\Sigma}_m^{-1}X)^{-1} X'\hat{\Sigma}_m^{-1}}_{*} \underbrace{\hat{\Sigma}_T}_{\uparrow} \underbrace{\hat{\Sigma}_m^{-1} X (X'\hat{\Sigma}_m^{-1}X)^{-1}}_{*}$$

* model based estimates

observed
estimated from residuals

Constant Variance Assumption

Now, let $Y_1, Y_2, \ldots, Y_n$ be independent observations from n units, but $Var(Y_i) = Var(\epsilon_i) = \sigma_i^2(X_i)$
Then the regression model is given by:

$$Y = X\beta + \epsilon, \epsilon \sim MVN(0, \Sigma), \Sigma = \text{ diag}(\sigma_j^2,\ \text{j} = 1, \ldots, \text{n})$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ 0 & \sigma_2^2 & \\ & & \ddots \\ & & & \sigma_n^2 \end{bmatrix}$$

Then apply the weighted least squares solution:

$$\hat{\beta}_{wls} = (X'\Sigma^{-1}X')^{-1}X'\Sigma^{-1}Y = \left[(\underbrace{\Sigma^{-1/2}X}_{X^*})'(\Sigma^{-1/2}X)\right]^{-1}(\Sigma^{-1/2}X)'(\underbrace{\Sigma^{-1/2}Y}_{Y^*})$$

where $\Sigma^{-1/2} = \text{ diag}(1/\sigma_j,\ \text{j} = 1, \ldots, \text{n})$.

$$\begin{bmatrix} 1/\sigma_1 & 0 & \ldots & 0 \\ 0 & 1/\sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/\sigma_n \end{bmatrix} \begin{bmatrix} 1 & X_{11} & \ldots & X_{p1} \\ 1 & X_{12} & \ldots & X_{p2} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & X_{1n} & \ldots & X_{pn} \end{bmatrix} = \begin{bmatrix} 1/\sigma_1 & X_{11}/\sigma_1 & \ldots & X_{p1}/\sigma_1 \\ 1/\sigma_2 & X_{12}/\sigma_2 & \ldots & X_{p2}/\sigma_2 \\ \vdots & \vdots & \ldots & \vdots \\ 1/\sigma_n & X_{1n}/\sigma_n & \ldots & X_{pn}/\sigma_n \end{bmatrix}$$

$$Y^* = \Sigma^{-1/2} Y \qquad X^* = \Sigma^{-1/2} X$$

Constant Variance Assumption

- What if we apply the least squares solution?

$\Rightarrow$ estimates $\hat{\beta}_{LS}$ are unbiased for β

$\Rightarrow$ $Var(\hat{\beta}_{LS})$ are biased $Var(\hat{\beta}_{LS}) \neq Var(\hat{\beta}_{WLS})$

- How do we estimate Σ?

What about estimating Σ by:

$$\hat{\Sigma}^{-1/2} = \begin{bmatrix} 1/\mid y_1 - X_1\hat{\beta} \mid & 0 & \dots & 0 \\ 0 & 1/\mid y_2 - X_2\hat{\beta} \mid & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\mid y_n - X_n\hat{\beta} \mid \end{bmatrix}$$

This is not a great idea!

If $y_i \approx X_i\hat{\beta}$ then you are creating very large weights!

You want to smooth!

$\hat{\beta}$ are OLS estimates

Build a model for the variance

$\sigma^2(X_i) \Rightarrow$ variance has to be positive

$\log \sigma^2 \sim \underbrace{\text{Gamma}}(\text{mean}, \text{shape} = 1/a)$

<u>Generalized linear model</u> :> next term

$$\log \sigma_i^2 = \underbrace{\gamma_0 + \gamma_1 X_{1i} + \ldots + \gamma_p X_{pi}}$$

$$\hat{\sigma}_i^2 = \exp(\gamma_0 + \gamma_1 X_{1i} + \ldots + \gamma_p X_{pi})$$

$$\hat{\sigma}_i = \exp\left[\frac{\gamma_0 + \gamma_1 X_{1i} + \ldots + \gamma_p X_{pi}}{2}\right]$$

We observe $r_i = (y_i - X_i\beta)$ and $E(r_i^2) = \sigma_i^2$

$$\underbrace{\text{Log}}\, E(r_i^2) = \underbrace{\gamma_0 + \gamma_1 X_{1i} + \ldots + \gamma_p X_{pi}}$$

$\hookrightarrow$ Gamma model, log link to squared residuals

Two-step estimation approach

X_p

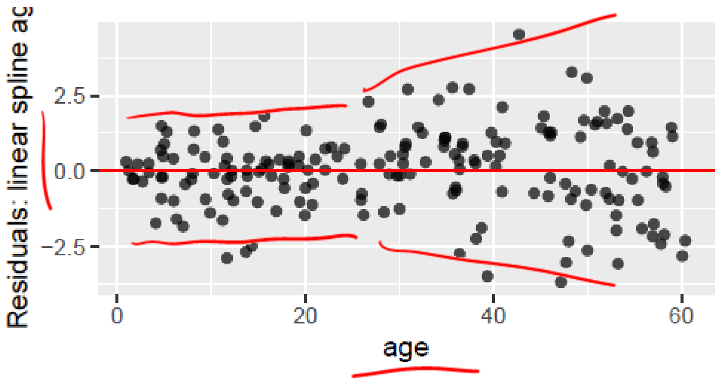
- Step 1: Fit the model for $E(Y|X) = X\beta$ using least squares
- Step 2: Obtain the residuals $\hat{r}_i = y_i - X_i\hat{\beta}$ and fit the regression of $\hat{r}_i^2$ on $ns(X_i\hat{\beta}, df = k)$ using a Gamma regression (log link) to estimate σ_i^2..
- Step 3: Compute $w_i = 1/\sqrt{\hat{\sigma}_i^2} = e^{-Z_i\hat{\gamma}/2}$ and fit the linear regression model using the weights.

$$\hat{\Sigma}^{-1/2} = \begin{bmatrix} 1/\sigma_1 & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_n \end{bmatrix}$$

Example: Nepali Anthropometry Data

```
load("C:\\Users\\Elizabeth\\Dropbox\\Biostat6532020\\Lecture34\\NepalAnthro.rdata")
d = nepal.anthro %>% select(names(.)[1:16]) %>% filter(.,num==1)
d = mutate(d,
agesp6=ifelse(age-6>0, age-6,0)
)
cc=complete.cases(select(d,age,wt))
d.cc=filter(d,cc)
d.cc = arrange(d.cc,age)
reg<-lm(data=d.cc, wt~age+agesp6)
d.cc$residuals = residuals(reg)
```

```
ggplot(d.cc,aes(x=age, y=residuals)) +
    geom_jitter(alpha = 0.7) +
    geom_hline(yintercept=0,color="red") +
    labs(y="Residuals: linear spline age")
```

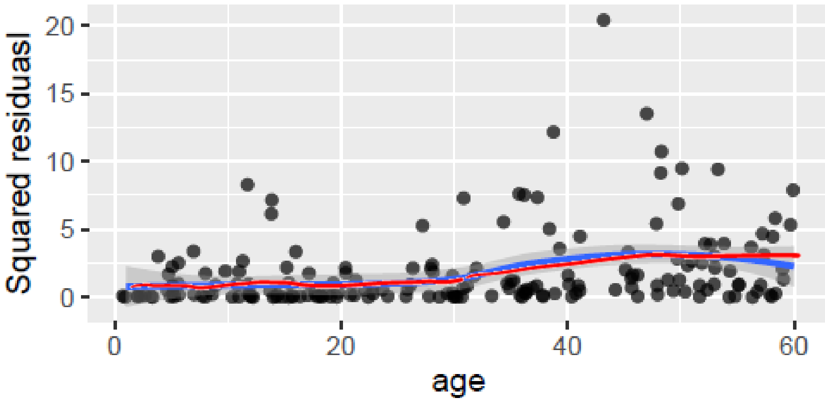


Example: Nepali Anthropometry Data

```
d.cc = mutate(d.cc,r2 = residuals^2)
ggplot(d.cc,aes(x=age, y=r2)) +
    geom_jitter(alpha = 0.7) +
    geom_smooth() +
    labs(y="Squared residuasl")
```

$\Rightarrow E(r^2) \Rightarrow \sigma^2(x_i)$

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Example: Nepali Anthropometry Data

```
v=predict.glm(glm(r2 ~ ns(age,3),data=d.cc, family=Gamma(link="log")),type="response")
regw = lm(wt~age+agesp6,data=d.cc,weights=1/sqrt(v))
summary(reg)
```

$\hat{\sigma}_i^2$

```
##
## Call:
## lm(formula = wt ~ age + agesp6, data = d.cc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.6768  -0.7575   0.0366   0.8998   4.5174
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.2112     0.7168   4.480 1.32e-05 ***
## age            0.5793     0.1303   4.445 1.53e-05 ***
## agesp6        -0.4307     0.1328  -3.243  0.00141 **
## ---
```

OLS

```
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -2.9081  -0.6539   0.0860   0.7535   3.4893
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.2750     0.5904   5.547 1.01e-07 ***
## age            0.5554     0.1080   5.142 6.99e-07 ***
## agesp6        -0.4042     0.1104  -3.662 0.000328 ***
## ---
```

mean 0
var 1

WLS procedure

Iteratively Reweighted Least Squares

The general estimation procedure requires an interative process.

The algorithm is:

1. Regress Y on X to obtain $\hat{\beta}_{ls} = \hat{\beta}^{(0)}$.
2. Regress $\hat{r}_i^2 = (y_i - X_i\hat{\beta}^{(k)})^2$ on Z_i using a Gamma regression with log link. Obtain $\hat{\gamma}^{(k)}$.
3. Regress Y_i on X_i with weights $1/\hat{\sigma}_j^{(k)} = e^{-Z_i\hat{\gamma}^{(k)}/2}$.
4. Repeat steps 2. and 3. until convergence:

$$\frac{(\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)})'(\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)})}{\hat{\beta}^{(k)\prime}\hat{\beta}^{(k)}} << \delta$$

Normality Assumption of the Residuals

- The normality assumption is used to derive the formulas for the variance and standard error of the estimated regression coefficients.
- If the data are far from normally distributed and sample size is not sufficiently large for the central limit theorem to protect us, then we can get misleading inferences
- Look at: histogram $\hat{r}_i = (y_i - X_i\hat{\beta})$

  quantile-quantile plot for $\hat{r}_i$
- Solution:

  Sample size is large enough $\to$ CLT $\hat{\beta} \sim \text{approx } N$

  Bootstrap procedure

MLR setting

$\hat{\beta} \sim MVN(\beta, \sigma^2(X'X)^{-1})$

$\hat{Y} \sim MVN$

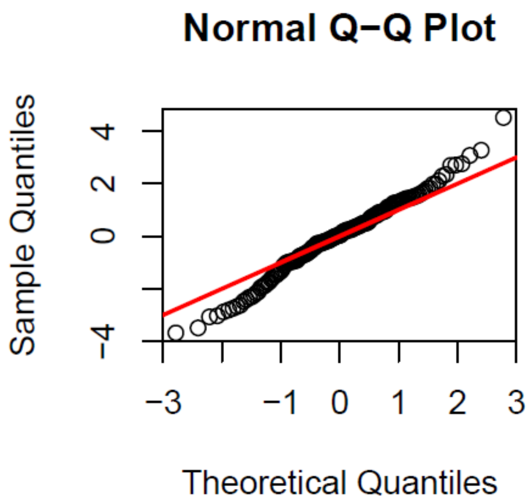
$\hat{R} \sim MVN$

$\varepsilon_i \sim N(0, \sigma^2)$

Example: Nepali Anthropometry Data

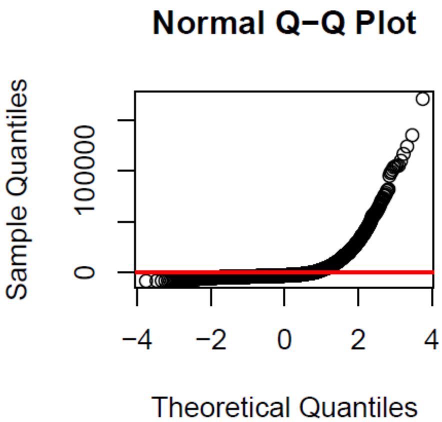
Back to the analysis of child weight vs. age.

```
qqnorm(d.cc$residuals)
abline(0,1,col="red",lwd=2)
```



Example: NMES

```
load("C:\\Users\\Elizabeth\\Dropbox\\Biostat6532020\\Problem Set 2\\nmes.rdata")
d = nmes %>% select(names(.)[c(1,2,3,15)]) %>% filter(.,lastage>=65)
d = mutate(d,
agec=lastage-65,
agesp1 = ifelse(lastage-75>0, lastage-75,0),
agesp2 = ifelse(lastage-85>0, lastage-85,0)
)
reg = lm(totalexp~(agec+agesp1+agesp2)*male,data=d)
qqnorm(reg$residuals);abline(0,1,col="red",lwd=2)
```

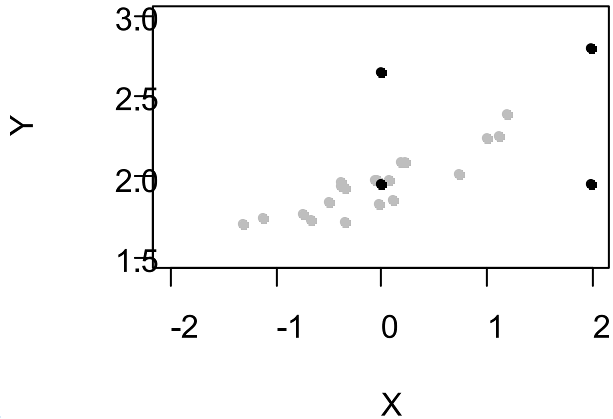


→ Bootstrap

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 $\hat{Y}$.

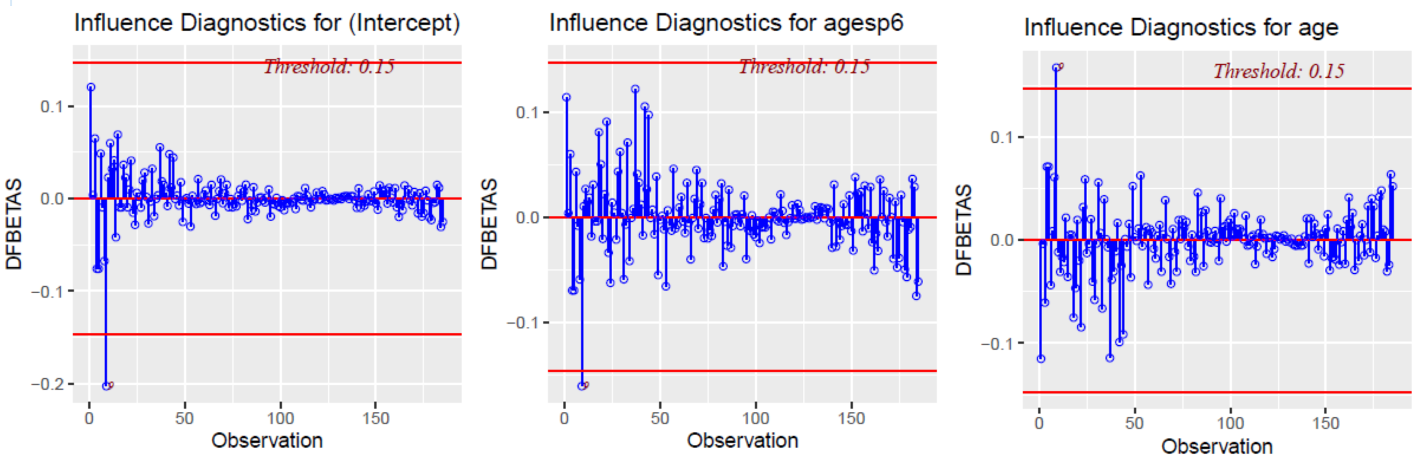


Influence statistics

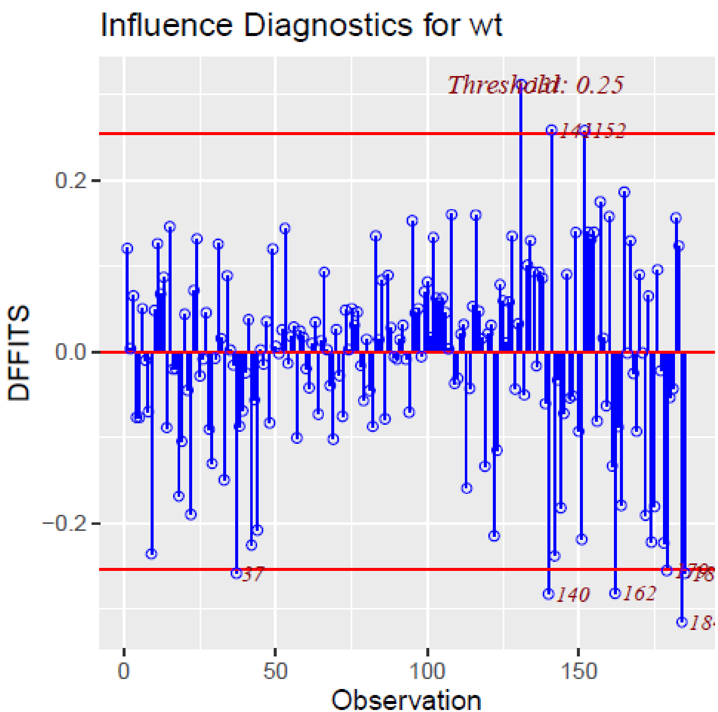
There are several influence statistics that are used in practice:

- $DBETA_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}$
- $DBETAS_{ij} = \frac{DBETA_{ij}}{\hat{se}(\hat{\beta}_{j(-i)})}$
- $DFIT_i = \hat{Y}_i - \hat{Y}_{i(-i)}$
- $DFITS_i = \frac{DFIT_i}{\hat{se}(\hat{Y}_{i(-i)})}$

Example: Nepali Anthropometry Data



Example: Nepali Anthropometry Data



Example: Nepali Anthropometry Data