Mahidol University
Faculty of Medicine Ramathibodi Hospital
Section for Clinical Epidemiology and Biostatistics

# Multiple linear regression III

Sasivimol Rattanasiri, Ph.D

Section for Clinical Epidemiology and Biostatistics

Ramathibodi Hospital, Mahidol University

E-mail: sasivimol.rat@mahidol.ac.th

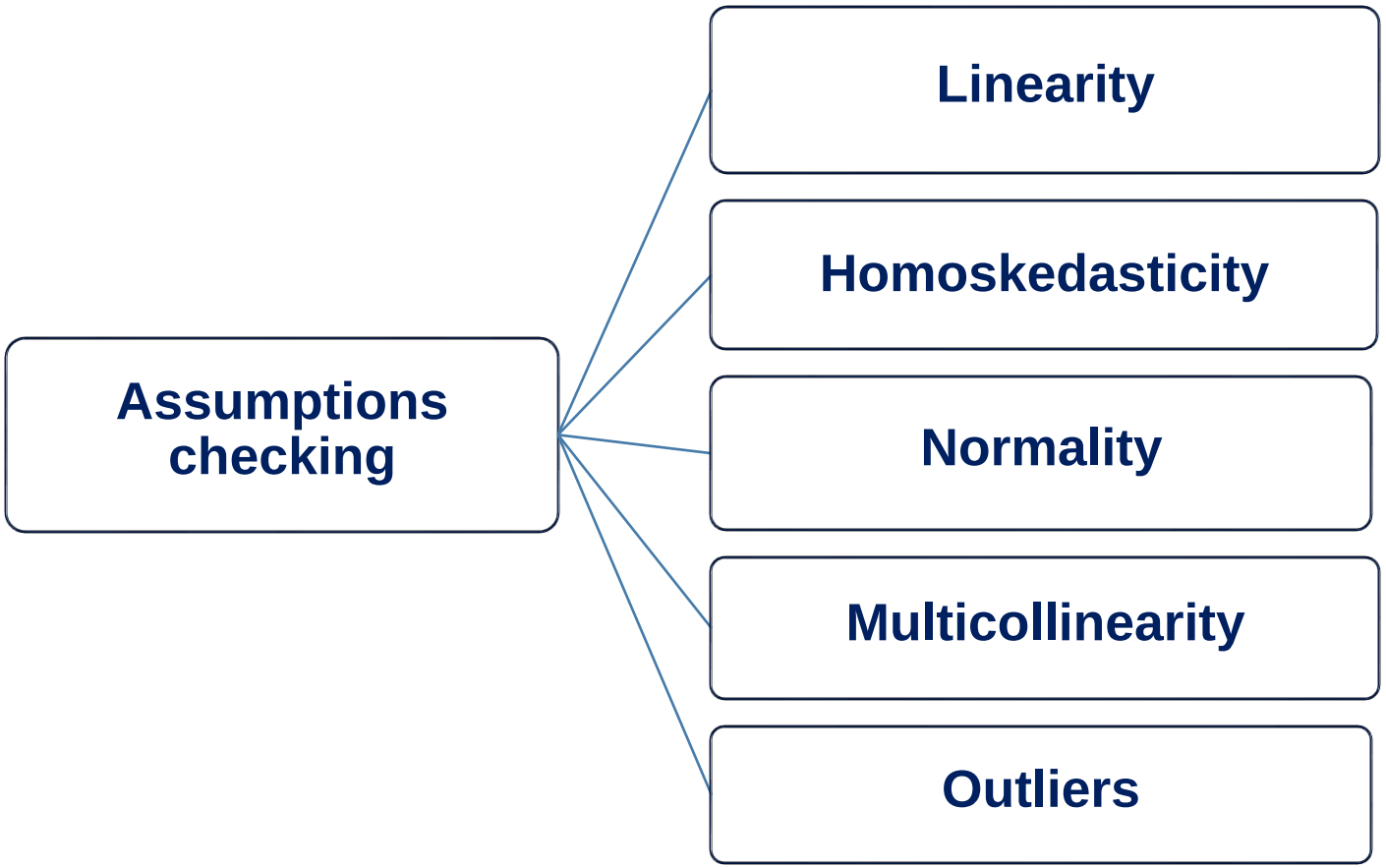



**Figure 1. Flow chart for assumptions checking for multiple linear regression**

# Final model for predicting SBP

```
. xi:regress sbp1 age bmi i.sex
i.sex             _Isex_1-2           (naturally coded; _Isex_1 omitted)

      Source |       SS           df       MS      Number of obs   =       294
-------------+----------------------------------   F(3, 290)       =     51.34
       Model |  45694.2664         3  15231.4221   Prob > F        =    0.0000
    Residual |   86038.347       290  296.683955   R-squared       =    0.3469
-------------+----------------------------------   Adj R-squared   =    0.3401
       Total |  131732.613       293  449.599363   Root MSE        =    17.225

------------------------------------------------------------------------------
        sbp1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .6305603   .0611661    10.31   0.000     .5101745    .7509461
         bmi |   1.340517    .235336     5.70   0.000     .8773341     1.8037
      _Isex_2 |  -7.709543   2.035356    -3.79   0.000    -11.71549    -3.7036
       _cons |     65.138   6.146222    10.60   0.000     53.04114    77.23486
------------------------------------------------------------------------------
```

# Assumption of linearity

- The multiple linear regression can only accurately estimate the relationship between outcome and predictors if the relationship are linear in nature.
- If the relationship between outcome and predictor is non-linear, the result of regression analysis will under-estimate the true relationship.
- This under-estimate can increase probability of type I and type II errors.

# Methods for checking assumption of linearity

- Visual examination
  - Partial regression plot
  - Augmented component plus-residual
  - Residual plot

# Partial regression plot

- Partial regression plot is commonly used to identify the pattern of the relationship between outcome and a given predictor when other predictors are already in the model.
- Partial regression plots are also referred to as added variable plots, adjusted variable plots, and individual coefficient plots.

# Partial regression plot

- Partial regression plot is performed by:
  - Computing residuals of regressing an outcome against all predictors except given predictor.

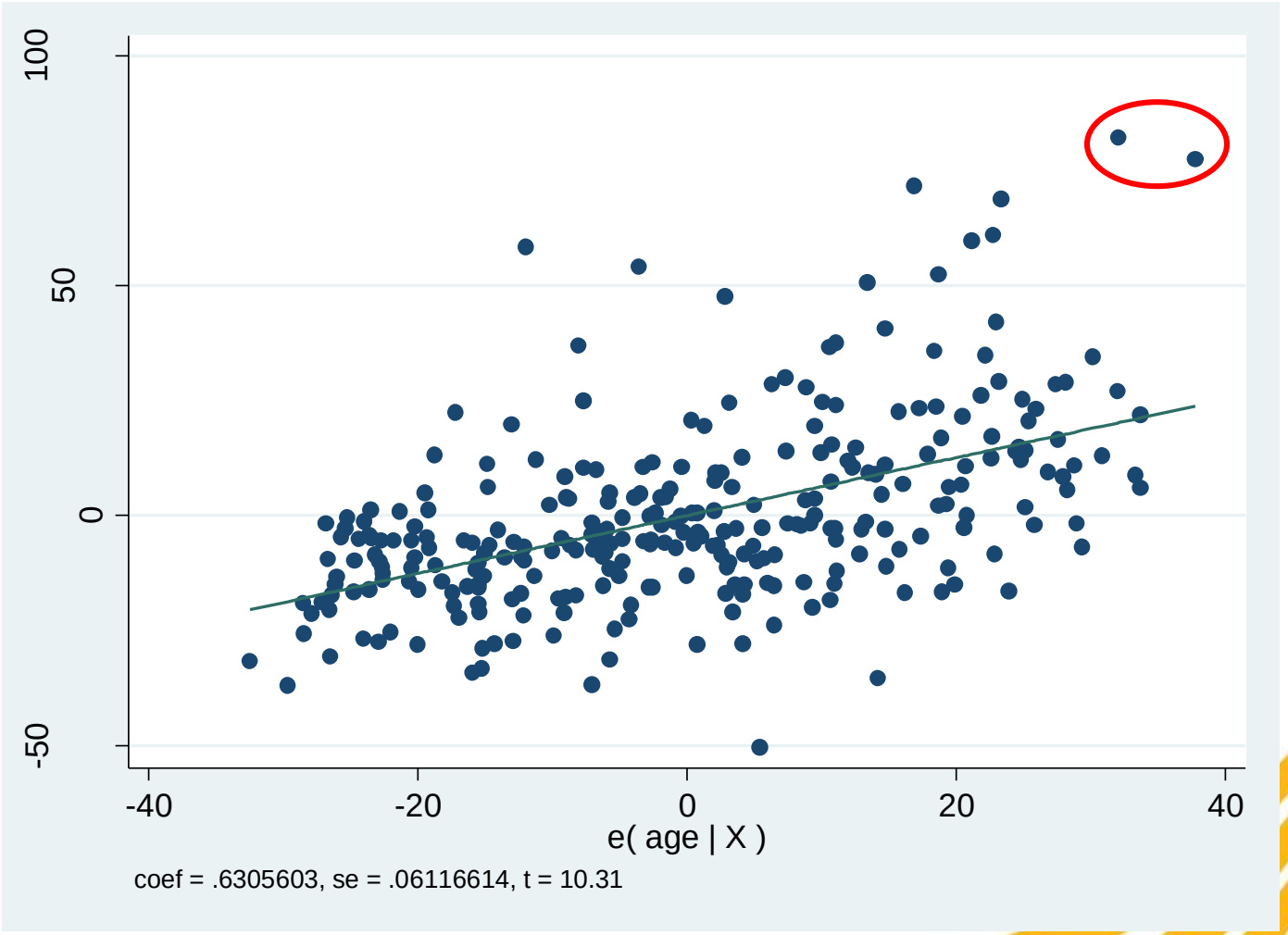
$$\mathbf{SBP}_i = \mathbf{b}_0 + \mathbf{b}_1(\mathbf{BMI}) + \mathbf{b}_2(\mathbf{Sex}) + \mathbf{e}_1$$

  - Computing residuals of regressing a given predictor against the remaining predictors.

$$\mathbf{Age}_i = \mathbf{b}_0 + \mathbf{b}_1(\mathbf{BMI}) + \mathbf{b}_2(\mathbf{Sex}) + \mathbf{e}_2$$

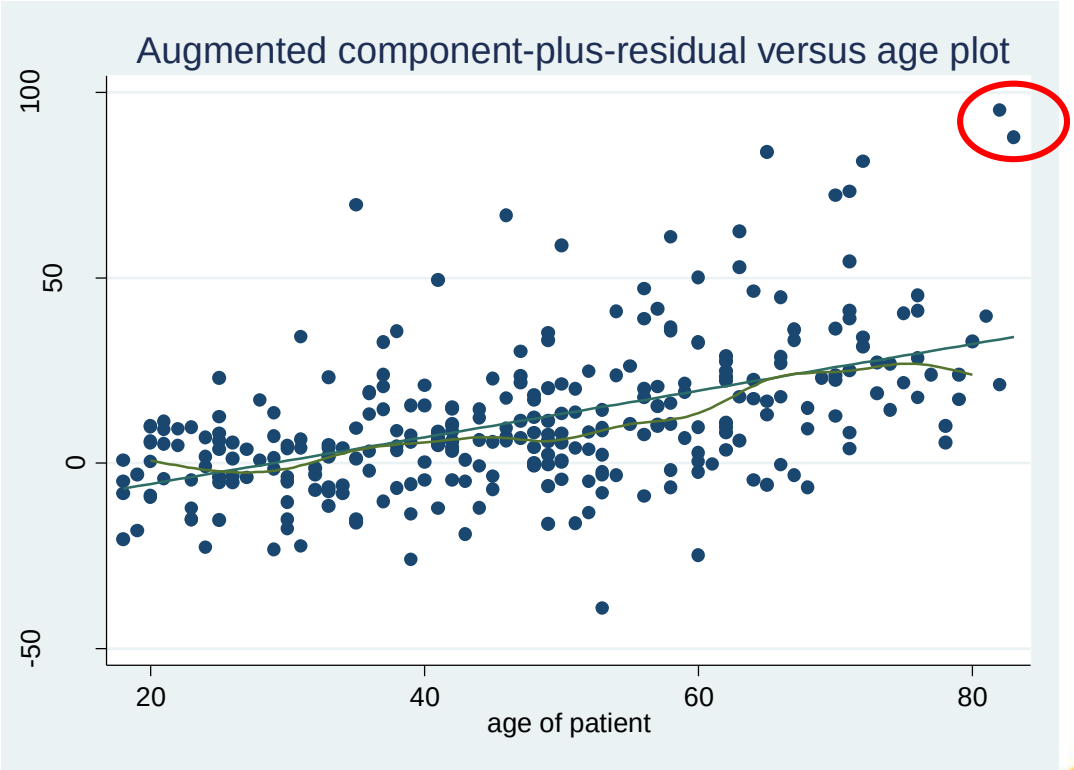
  - Plotting the residuals from ($e_1$) against the residuals from ($e_2$).

# Partial regression plot

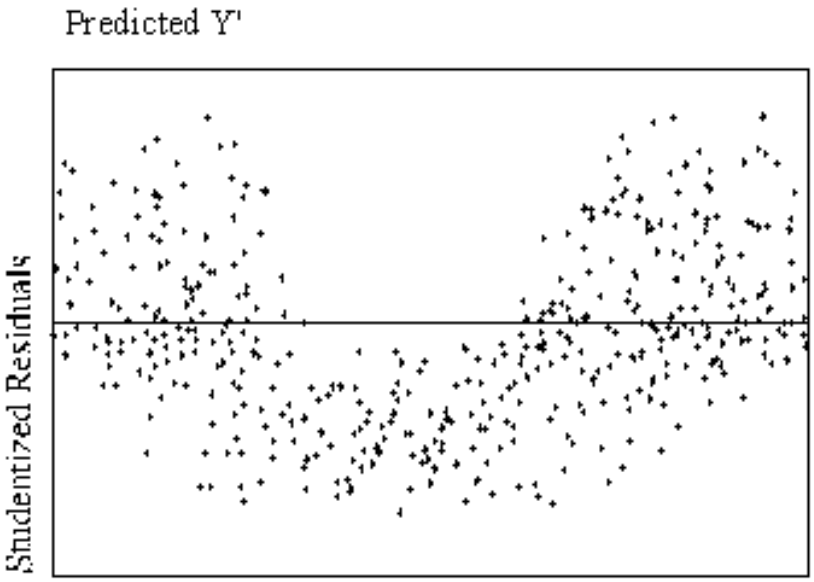


# Augmented component plus-residual

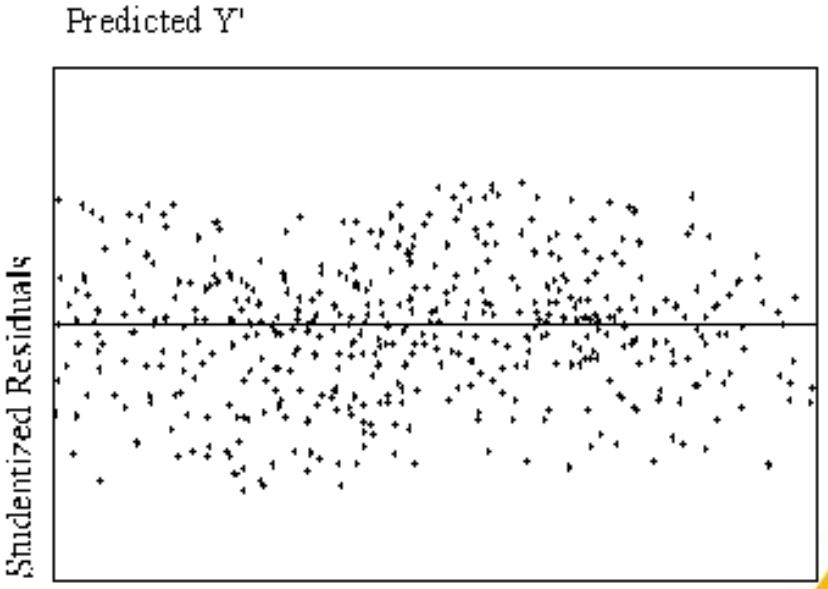
An Augmented component plus-residual is useful in detecting non-linearity.



# Residuals plots

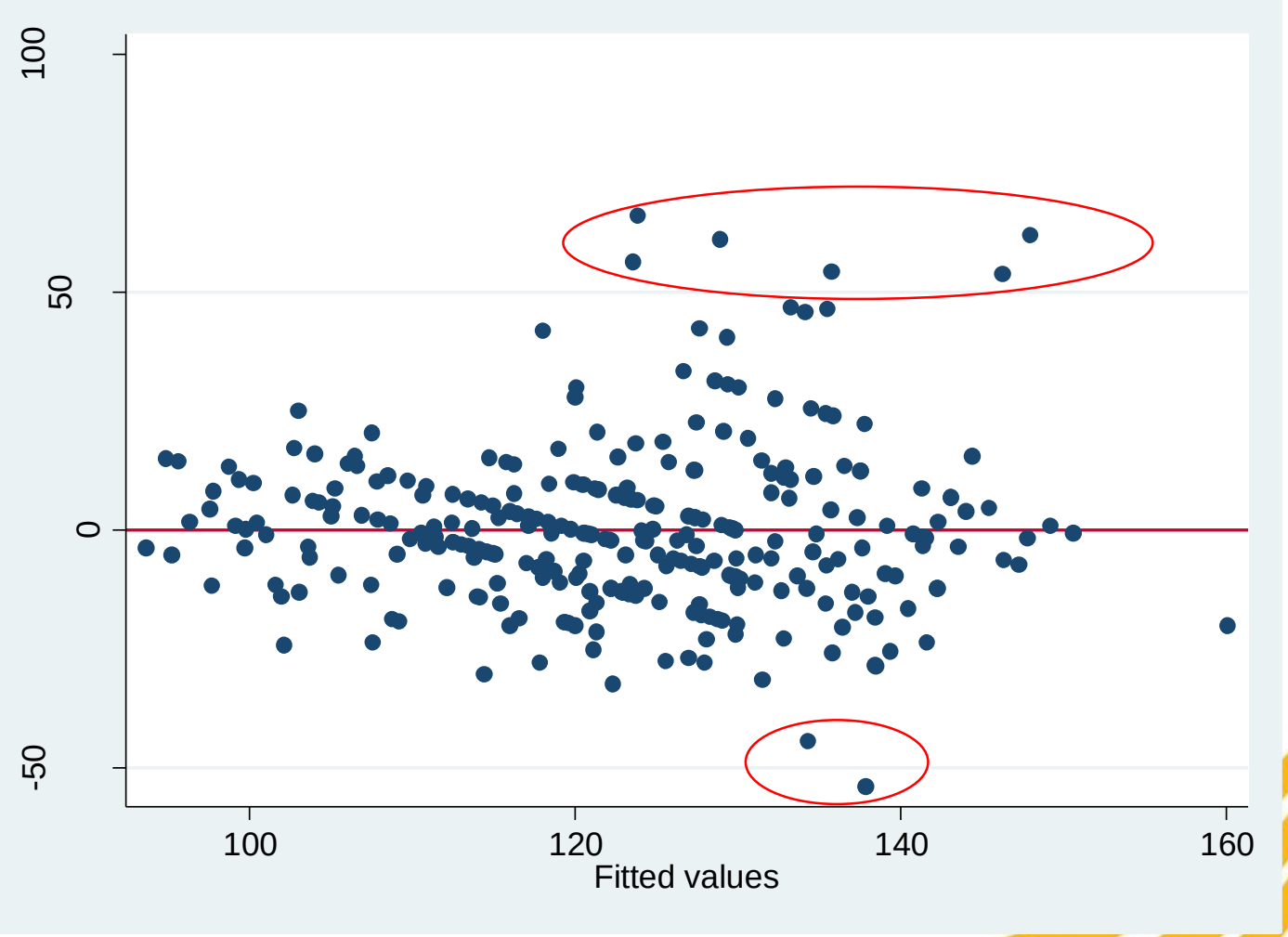



a) Curvilinear relationship




b) Linear relationship

# Residuals plots



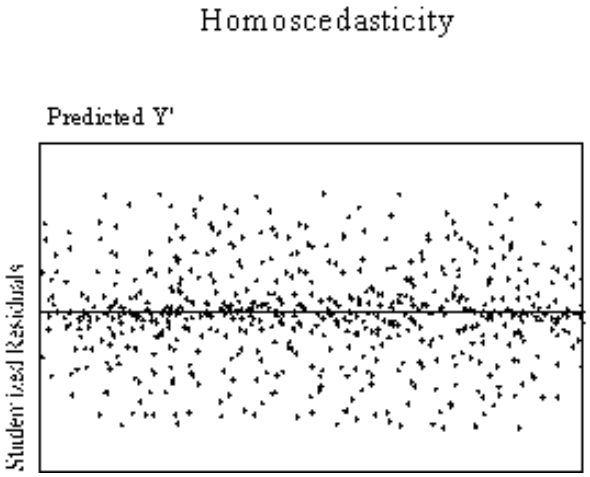
# Assumption of homoscedasticity

- Homoscedasticity means that the variance of errors is the same across all levels of predictor.
- When variance of errors differs at different values of predictor, heteroscedasticity is indicated.
- When heteroscedasticity is indicated, it can lead to serious distortion of findings, and thus increase the probability of type I error.

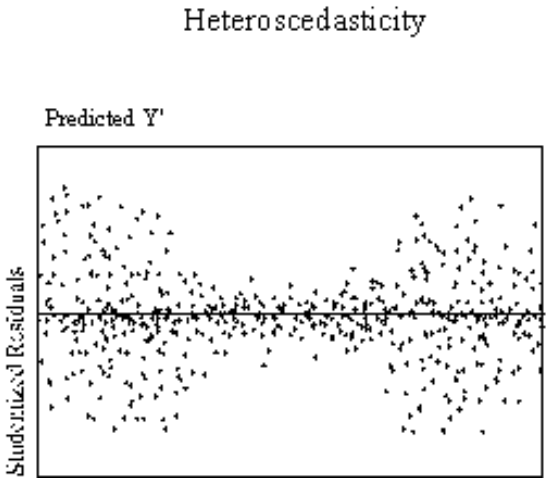
# Methods for checking assumption of homoscedasticity

- Visual examination
  - Residual plot
- Statistical test
  - Breusch-Pagan/Cook-Weisberg test

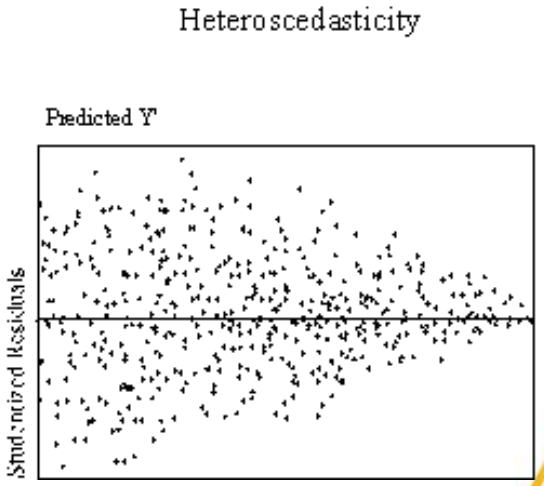
# Residuals plots




Horizontal band

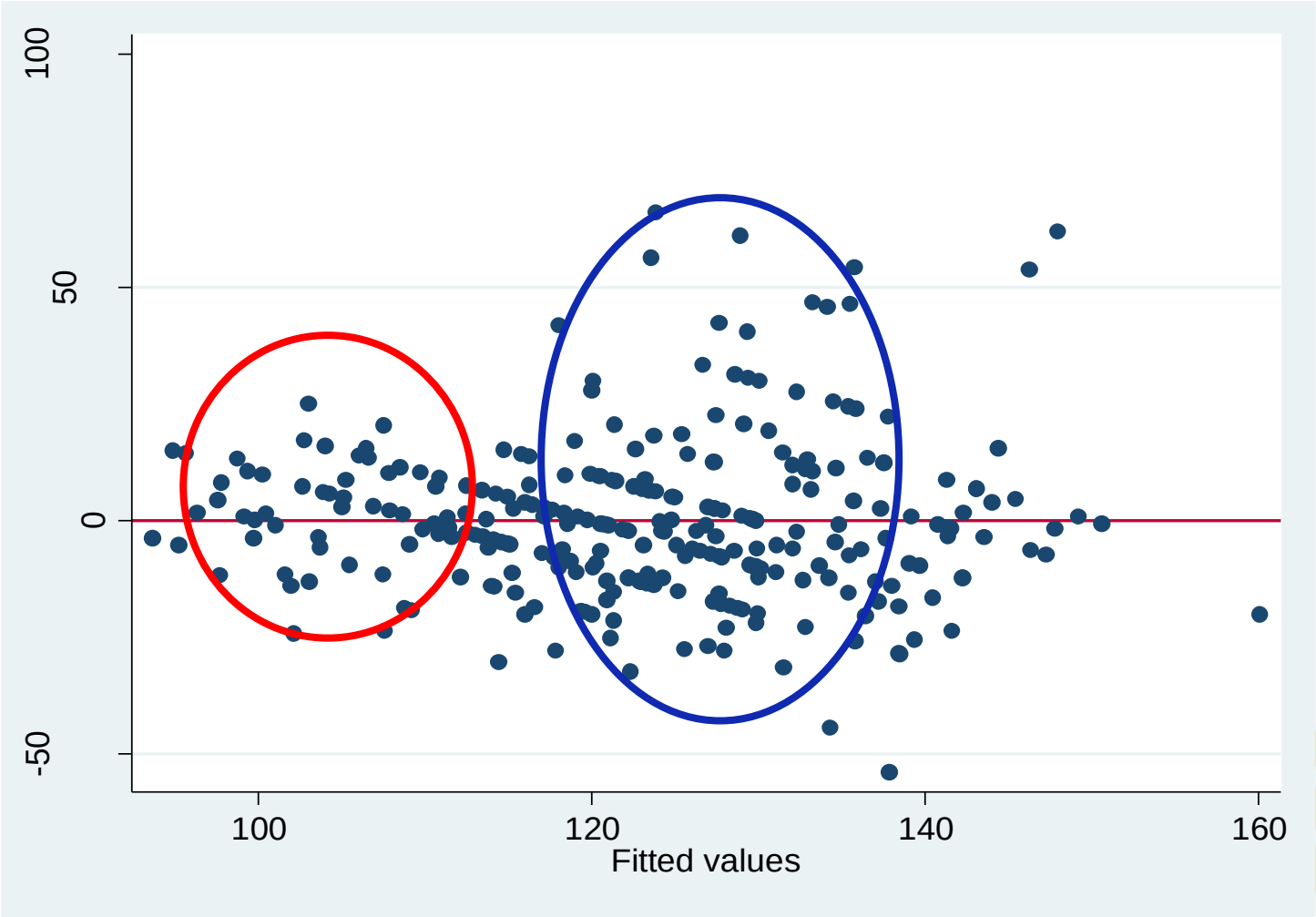



Bowtie shape




Fan shape

# Residual plot against predicted values



# Test for heteroscedasticity

```
. estat hettest res1

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: res1

         chi2(1)      =   123.46
         Prob > chi2  =   0.0000
```

# Assumption of normality

- Multiple regression assumes that residuals must have normal distribution.
- Violation of normality can distort relationships and significance tests.

# Methods for checking assumption of normality

- Visual examination
  - Histograms
  - Normal probability plot
- Descriptive statistics
  - Skewness and Kurtosis
- Statistical tests
  - Kolmogorov-Smirnov
  - Shapiro-Wilk test

# New consideration

- Adding more predictors to a multiple regression model creates more relationship among them.
- So not only are predictors related to the outcome variable, they are also potentially, related to each other.
- This called “Multicolinearity”.
- The ideal is for all of predictors to be correlated with the outcome, but NOT with each other.

# Assumption of multicollinearity

- This assumption does not matter for simple regression analysis, but it is important to check for the multiple regression model.
- Multicollinearity occurs when predictors in the multiple regression model are correlated among themselves.

# Effects of multicollinearity

- Large changes in the estimated regression coefficients when a predictor is added or deleted.
- Non-significant results in individual tests on the regression coefficients for significant predictors.
- Wide confidence intervals for the regression coefficients .

# Effects of multicollinearity

| | Coef. (95% CI) | |
|---|---|---|
| Predictors in model | Model 1 | Model 2 |
| Age | 0.63 (0.51, 0.75) | 0.62 (0.50, 0.75) |
| BMI | 1.34 (0.88, 1.80) | 1.52 (0.44, 2.61) |
| Sex | -7.71 (-11.72, -3.70) | -8.41 (-13.87, -2.95) |
| Weight | | -0.07 (-0.45, 0.31) |

# Multicollinearity diagnostics

There are two approaches to detect multicollinearity:

- Pairwise correlation analysis.
- Variance inflation factor (VIF).

# regress sbp1 age bmi i.sex

```
. estat vif

    Variable |       VIF        1/VIF
-------------+----------------------
         bmi |      1.02     0.981659
      _Isex_2 |      1.02     0.982055
         age |      1.01     0.987186
-------------+----------------------
    Mean VIF |      1.02
```

# regress sbp1 age bmi i.sex wt

```
. estat vif

    Variable |       VIF        1/VIF
-------------+----------------------
          wt |      5.99    0.166865
         bmi |      5.55    0.180145
      _Isex_2 |      1.89    0.530055
         age |      1.12    0.896557
-------------+----------------------
    Mean VIF |      3.64
```

# Variance Inflation Factor (VIF)

- This parameter measures how the variances of regression coefficients are inflated compared to when the predictors are not linearly related.
- The VIF > 10 is suggested as collinearity and the value close to 1 is no evidence of collinearity.

# Outliers

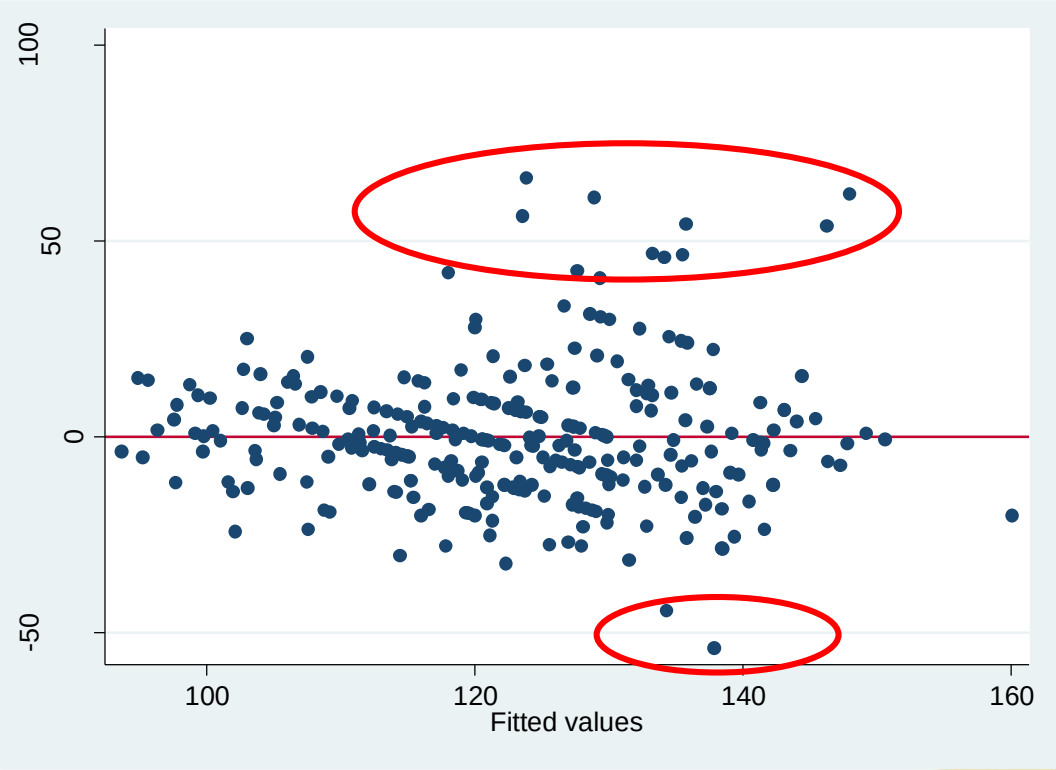
- Outliers are observations with large residuals.
- If the outliers are present, data should be checked to make sure that there is no error during data entry, or no error due to measurement.
- If the error came from measurement, that observation must be omitted.
- Analyses by Osborne (2001) show that removal of outliers can reduce type I and type II errors, and improve accuracy of estimate.



27

# Identify outlying cases

The outlying cases involve large residuals and often have influential on the fitted least squares regression function.



# Identify outlying cases

- It is therefore important to identify outlying cases and decide whether they should be retained or eliminated in the fitting process of the regression model.
- For regression with one or two predictors, identification of outlying cases can be done by simple graphics.
- When more than two predictors are included in the regression model, we need the special tools to detect the outliers and influential cases

# Tools for identify outliers and influential cases

## I. Identify outlying X observations

- Leverage values

## II. Identify outlying Y observations

- Studentized deleted residuals

## III. Identify influential cases

- Cook's distance
- DFFITS
- DFBETAS

# I. Identify outlying X observations

$h_{ij}$ (hat or leverage) is used to identify X outliers, which can be defined as:

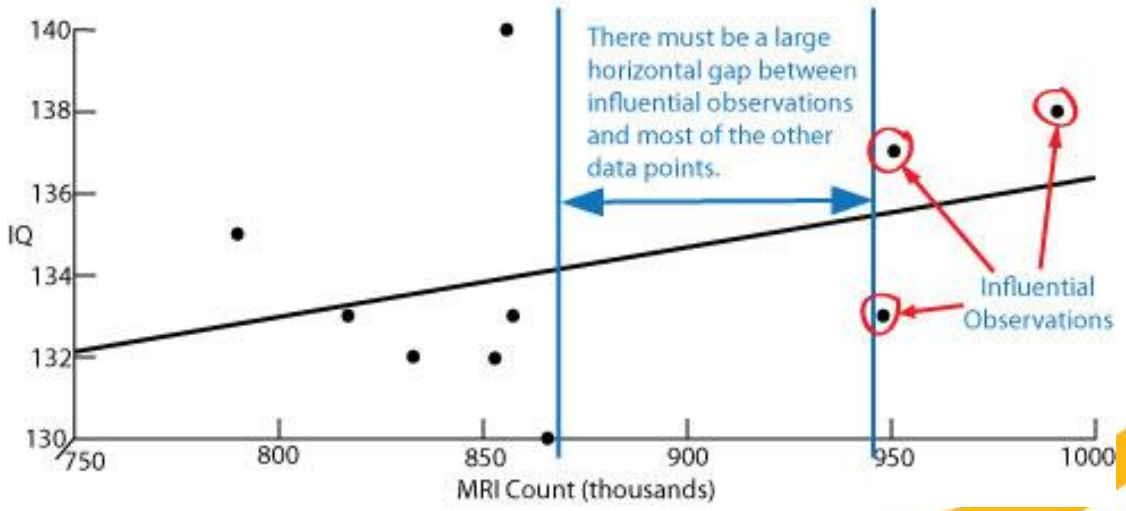
$$h_{ii} = X_i^{'}(X'X)^{-1}X_i$$

$$0 \le h_{ii} \le 1, \sum_{i=1}^{n} h_{ii} = p$$

where p = number of regression parameters including constant term.

# I. Identify outlying X observations

- The leverage is a measure of the distance between the X values for the *i* th case and the means of the X values for all n cases.
- A large leverage value indicates that the *i* th case is distant from the center of all X observations.



# I. Identify outlying X observations

- A leverage value is usually considered to large if it is more than twice as large as the mean leverage value.
- Hence, leverage values greater than 2p/n are considered as outlying cases with regard to their X values.

# II. Identify outlying Y observations

This can be done using studentized deleted residuals, which is calculated by

$$d_i^* = e_i \left( \frac{n-p-1}{SSE(1-h_{ii})-e_i^2} \right)^{1/2}$$

- Studentized deleted residual has t-distribution with n-p-1 degree of freedom.
- It is considered as the outlier, if absolute values of studentized deleted residual is higher than *t* value at tail areas of 0.05.

# III. Identify influential cases

- After identifying cases that are outlying with respect to their X values (values of predictors) and/or their Y values (values of outcome),
- Next step is to ascertain whether or not these outlying cases are influential on the fitted regression function.

# III. Identify influential cases

- We shall consider a case to be influential if its exclusion causes major changes in the fitted regression function.
- We consider three measures of influence that are widely used in practice.
  - Influence on predicted values
  - Influence on partial regression coefficients
  - Influence on all regression coefficients

# Influential on predicted value

DFFITS is used to indentify influential cases on predicted value. It can be defined as:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

The absolute value of DFFITS exceed 1 is considered as influential case for small-medium sample size and $2\sqrt{p/n}$ for a large sample size.

# Influential on all regression coefficients

Cook's distance ($D_i$) measures the impact of the $i^{th}$ case on all of regression coefficients. It is defined as:

$$D_i = \frac{(\mathbf{b} - \mathbf{b}_{(i)})' \mathbf{X}' \mathbf{X} (\mathbf{b} - \mathbf{b}_{(i)})}{p\text{MSE}}$$

if $D_i > 4/n$ or $D_i > F(p, n-p, 1-\alpha)$, the $i^{th}$ case has substantially influenced on estimate coefficients

# Influential on partial regression coefficients

**DFBETAS** measures the influence of the $i^{th}$ case on estimation of each regression coefficient. It can be defined as:

$$DFBETAS_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}} \text{ where}$$

$c_{kk}$ = the k - th diagonal element of $(X'X)^{-1}$

$b_{k(i)}$ = regression coefficient if the i - th case is omitted

# Influential on partial regression coefficients

- The DFBETAS determines the standardized difference of coefficients that estimated with and without the $i^{th}$ cases.
- The absolute DFBETAS > 1 and > $2/\sqrt{n}$ are supposed to be influential cases for small-medium and large sample sizes.

# Summary

- An outlying influential case should not be automatically removed, because it may entirely correct and simply represents an unlikely events.
- If outlying influential case can definitely be shown to be the result of measurement error, it would be appropriate to remove this case.

# Summary

- If the outlying influential case is accurate, it may not represent an unlikely event but rather than a failure of the model.
- The failure may be:
  - Omission of some important predictor variables.
  - Omission of an important interaction term.
  - Incorrect functional form, such as omission of a curvature effect for some predictor variables.