Assumption for multiple linear regression.

การทดสอบสำหรับ linear regression ประกอบด้วย

1. Linearity
2. Normal distribution of error (residual)
3. Constant varience along predictors
4. Influential points

การทดสอบ "Normality with equal varience" เป็นสิ่งจำเพาะใน Linear regression

การทดสอบ Linearity และ Influential points ต้องการใน Logistic regression ด้วย

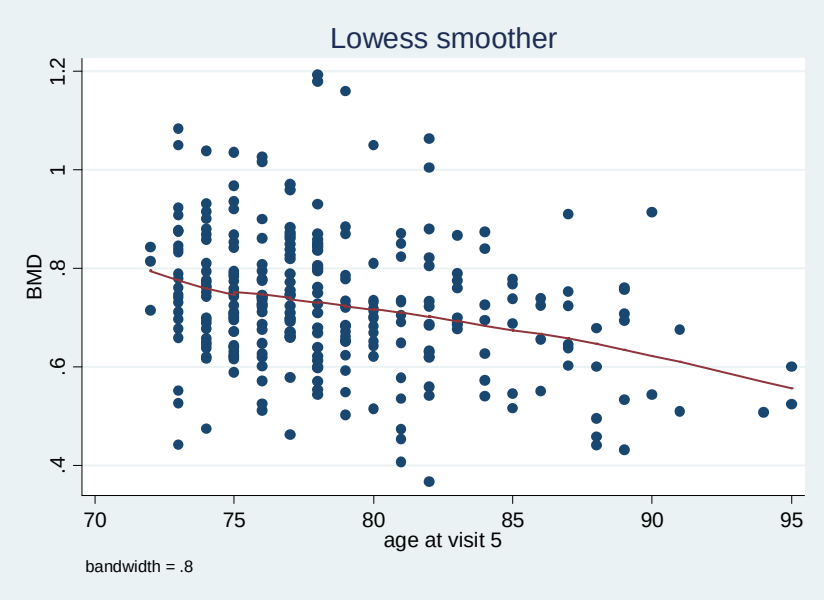
การทดสอบและลงวินิจฉัยเป็นเรื่องละเอียดซับซ้อน ใช้ทั้งศาสตร์และศิลป์ ดังนั้นเพียงรู้ในหลักการเบื้องต้นแล้วปรึกษานักสถิตินำจะเหมาะสม

****

Concep 1: Linearity -> CPR plot บอก slope

ใน simple linear regression (single predictor) เราสามารถดูความสัมพันธ์แบบ linear ได้ตรงไปตรงมา

```
. lowess bmd age, name(lowessbmd , replace)
```



แต่ใน multipredictors เราไม่สามารถทำกราฟแบบข้างต้นได้ การวิเคราะห์ต่างๆ จะเกี่ยวข้องกับ "Residual"

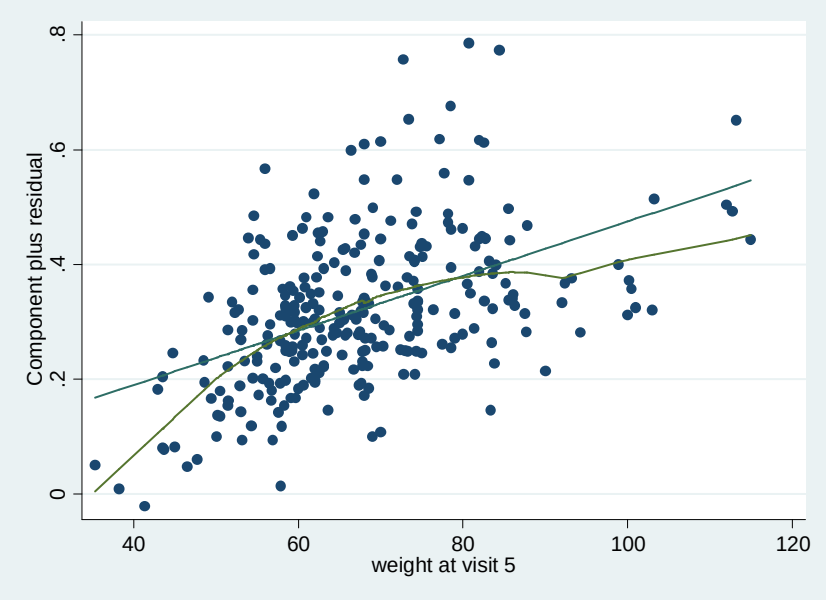
Note: Residual หรือ error ความต่างระหว่างข้อมูลจากการศึกษา กับค่าที่พยากรณ์ด้วย model

หลักการคือ ถ้า "ข้อมูลจากการศึกษา" Y (outcome) กับ X (predictor) มีความสัมพันธ์เป็นเส้นตรง Residual จะมีค่าเฉลี่ยคงที่

component plus residual (CPR) plot -> plot ที่บอก slope เพราะแกน Y คือ coef + residual มีข้อดีที่สามารถดูลักษณะ slope แล้วคาดได้ว่าการใช้ log transformation จะแก้ได้สำเร็จหรือไม่

residual versus predictor (rvp) plot -> plot ที่ไม่บอก slope เพราะแกน Y คือ residual เท่านั้น

```
. cprplot weight, lowess name(cprweight, replace)
```



CPR plot ดังภาพ เป็น downward curvature แสดงว่า outcome เปลี่ยนแปลงช้ากว่า rate ของ predictor การใช้ log predictor น่าจะได้ผล

กรณีที่ต้องใช้ RVP ประเมิน linearity คือกรณีต้องการตรวจสอบหลังจากใช้เทคนิค quadritic term ซึ่ง RVP การสร้าง LOWESS (อ่าน โลว์ อิส) ทำยาก แปลผลยาก ด้านล่างคือตัวอย่างการทำ quadirtic transformation ของ predictor weight

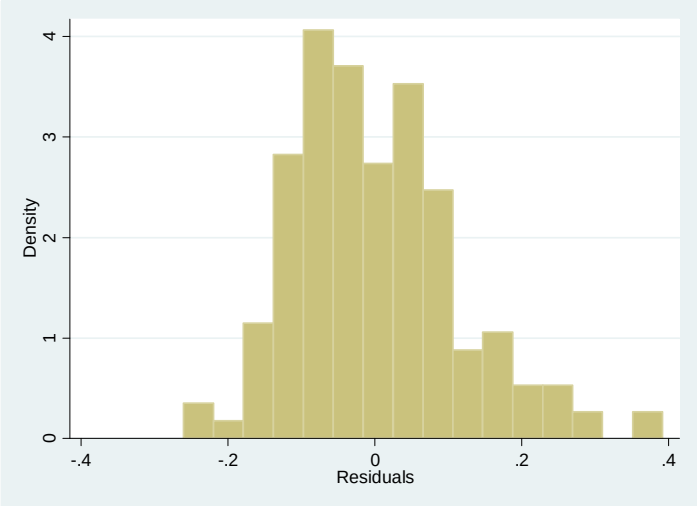
****

Concep 2: Normality of residual -> Histogram + Q-Q plots of residuals

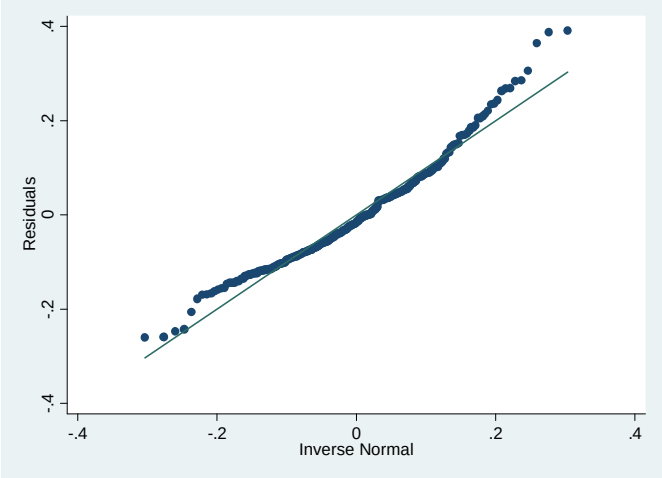
หากจินตนาการ ว่าแต่ละ X จะให้ Y กระจายตัวตามแกน linear เหมือนการนำเอาระฆังคว่ำมาเรียงต่อกันเป็นเส้นตรง

normal distribution ของ residual ก็หมายถึงระฆังไม่บุบเบี้ จึงมีขอบ "แต่ละด้าน" เท่ากัน

```
. reg bmd age weight weight2
. predict bmdresid, residual
. hist bmdredid, name(hisbmd, replace)
```




```
. qnorm bmdresid , name(qnormbmd, replace)
```



การแปลผล Q-Q norm ดูที่หัว (ขวามือ) ของกราฟ ถ้าส่วนหัวกระดก หมายถึง ข้อมูลด้านขวากระจาย extreme (Rt.skew ใน histogram) ในทางตรงข้ามถ้าหัวตก หมายถึง Left skewness

การพิจารณาว่า acceptable normality หรือไม่ ใช้หลักการของ Central limit theorem

กล่าวคือ ถ้า sample size ใหญ่ แล้ว skew เล็กน้อยก็ไม่เป็นปัญหา

แต่ถ้า sample size เล็กแล้ว skew มาก ควรใช้ trasformation ของ variable เข้ามาช่วย

****

Concep 3 : Constant variance -> RVP

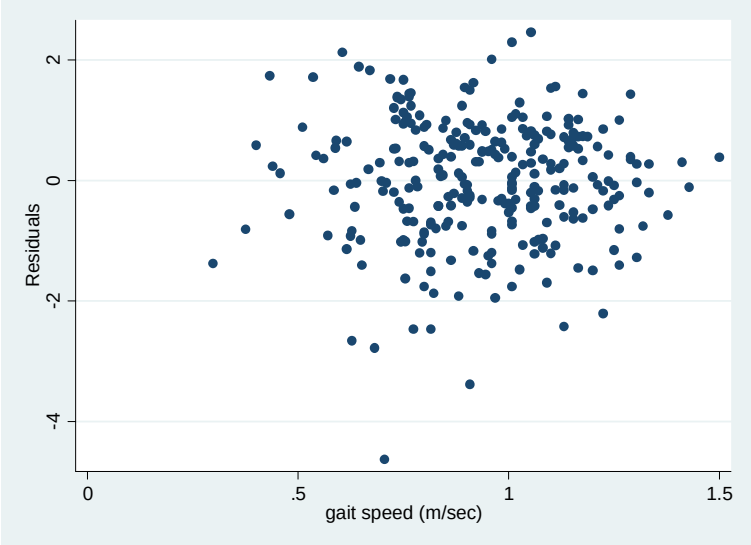
หากจินตนาการ ว่าแต่ละ X จะให้ Y ที่มีการกระจายตัวแบบ normal distribution เหมือนการนำเอาระฆังคว่ำมาเรียงต่อกัน constant of varience ก็หมายถึงขอบระฆัง "แต่ละอัน" กว้างพอๆ กันนั่นเอง

Constant varience ไม่สามารถใช้ Central limit theorem เข้ามาช่วย จึงจำเป็นต้องทดสอบด้วย RVP:
Residuals vs predictor
heteroskedasticity = non-constant variance

```
. reg leeu age poorhlth gaitspd

. rvfplot

. rvpplot gaitspd
```



ดูว่าลักษณะการกระจายของ residual แผ่ออกเป็น funnel shape หรือไม่ แต่ในตัวอย่างที่แสดงไม่มีลักษณะเป็น funnel จึงสรุปได้ว่า มี constant varience

****

Concept 4 Influential points -> BFBETA + box plot + list

Influential point คือ outlier ที่ส่งผลกระทบต่อการแปลผลโดยรวม

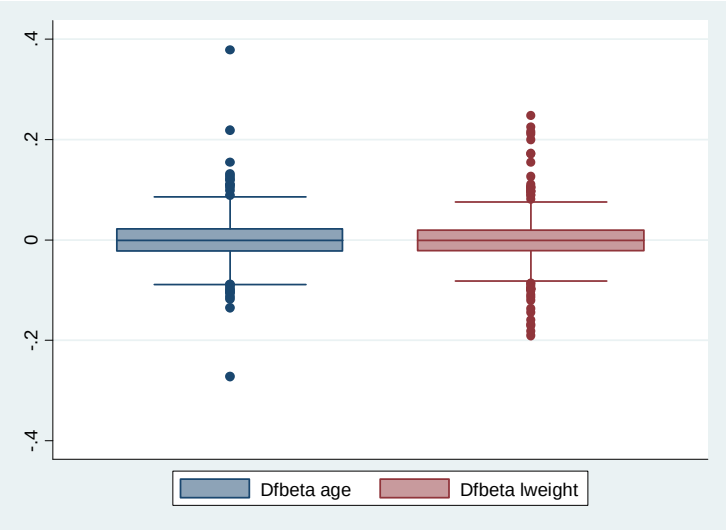
การประเมินว่า outlier ตัวไหนมีพฤติกรรมเป็น influential point จะใช้เทคนิคทางสถิติเรียกว่า DFBETA เป็นตัว "grade" ระดับความเป็น

ผู้ร้าย (influential) ของแต่ละ observation ของ outcome

หลังจากนั้น สามารถใช้ graph box ดูการแจกแจงของ DFBETA นั้น

ตามด้วย list รายชื่อผู้ร้าย ว่ามีใครสมควรถูกตัดออกไป

```
. dfbeta
. graph box _dfbeta_1 _dfbeta_2 , name(gboxdfbeta, replace)
```




```
. list bmd age lweight _dfbeta_1 if abs(_dfbeta_1) > 0.2 & _dfbeta_1
~= .

     +-----------------------------------+
     |  bmd   age    lweight   _dfbeta_1 |
     |-----------------------------------|
140. | .458    88   4.234107   -.2716317 |
151. | .914    90   4.023564    .3781947 |
200. |  .91    87   4.219508    .2188265 |
     +-----------------------------------+
```

```
. list bmd age lweight _dfbeta_2 if abs(_dfbeta_2) > 0.2 & _dfbeta_2
~= .
```

	bmd	age	lweight	_dfbet~2
94.	1.18	78	4.435567	.2476206
108.	.366	82	3.720862	.2242683
231.	.441	73	3.642836	.2110856
262.	1.193	78	4.390738	.2162125