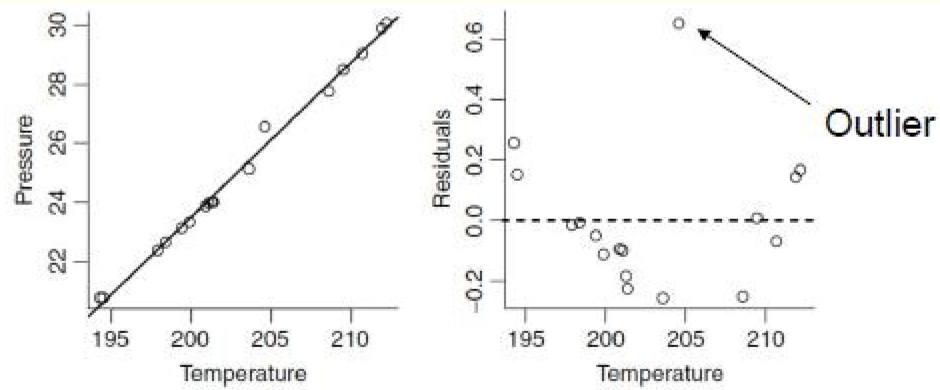
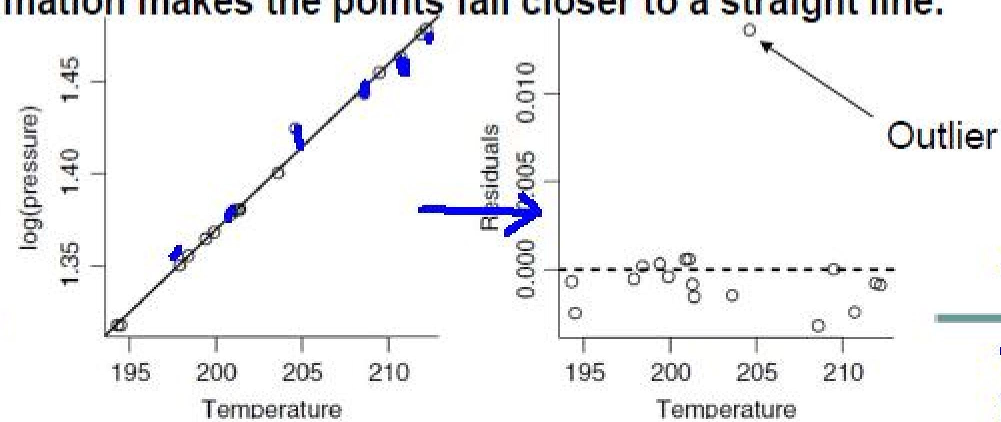
Examples 2 – Forbes Data
Measure pressure from boiling point of water

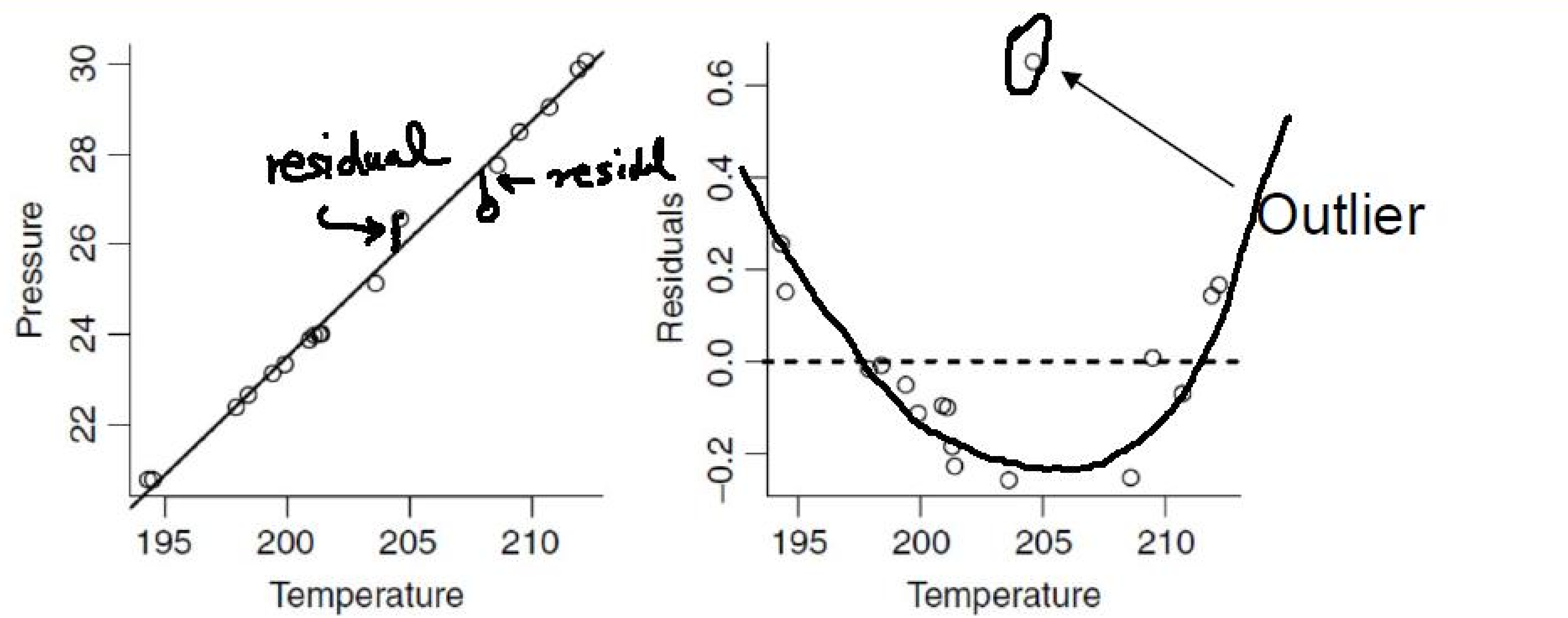


- There seems to be a systematic error (curve relationship) between y & x
- Transformation makes the points fall closer to a straight line.



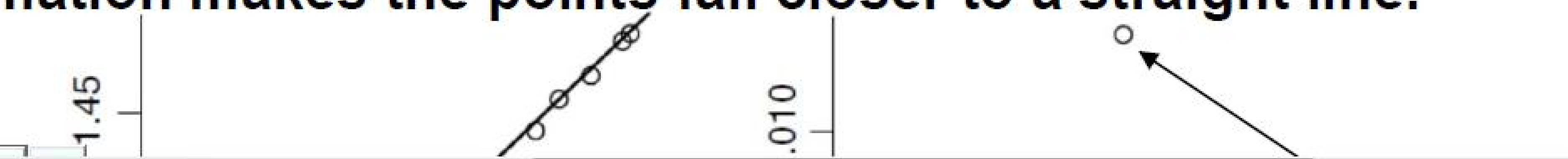
If residual plot is null plot then the regression is a good fit

e pressure from boiling point of wate

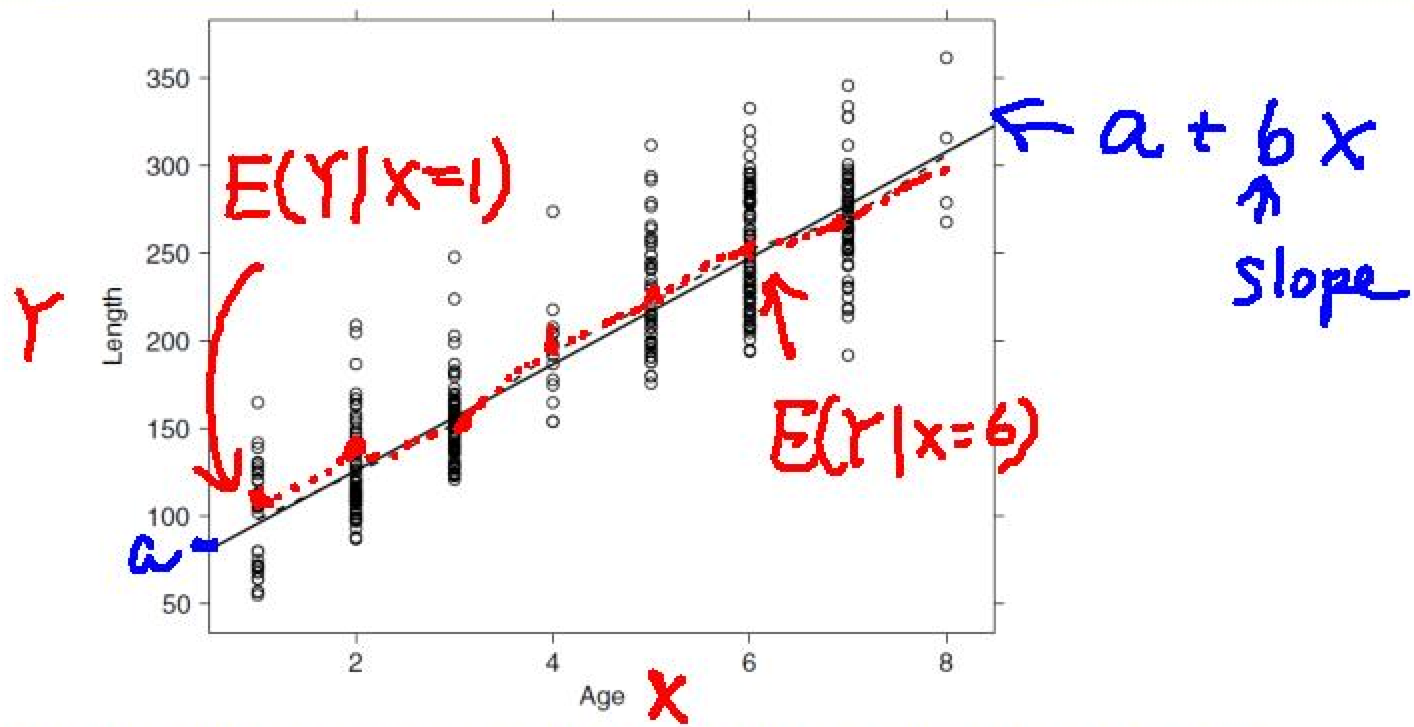


ems to be a systematic error (curve relationship) between

mation makes the points fall closer to a straight line.

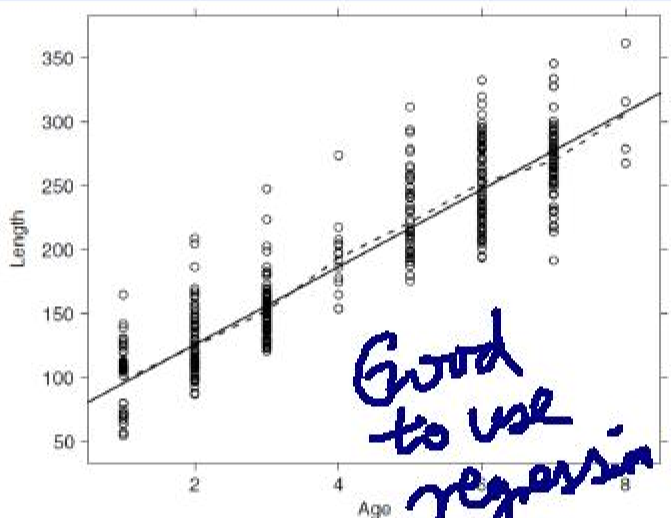


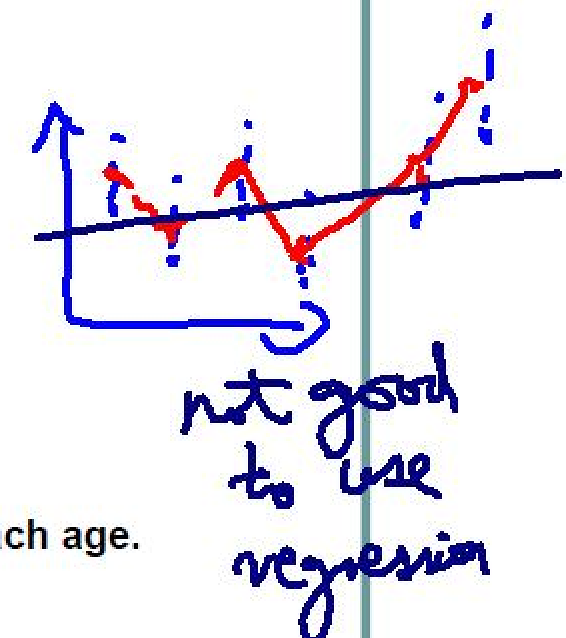
mples 3 – Smallmouth Bass
vs Age of fish



- ne dash line joins the average observed length at each age.
- . mean of length at age i, i=1,2,..,8.
  - This summary of data needs 8 numbers.
- ne solid line is the regression line. Y=a+bX.

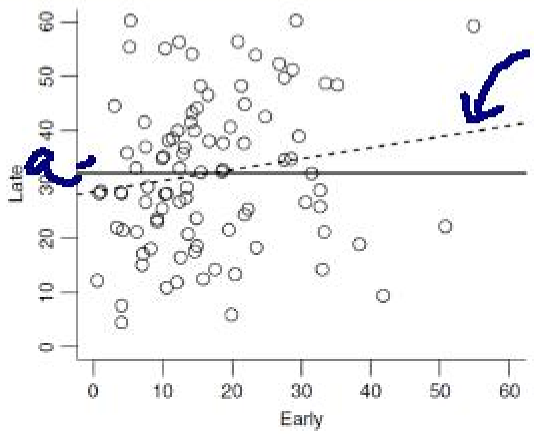
Examples 3 – Smallmouth Bass Size vs Age of fish





- **The dash line joins the average observed length at each age. i.e. mean of length at age i, i=1,2,..,8.**
  - This summary of data needs 8 numbers.
- **The solid line is the regression line, Y=a+bX.**
  - This summary of data needs 2 numbers (slope and intercept).
  - Regression gives a good summary for this dataset

Examples 4 – Predicting the weather
Predict late season snowfall from early snowfall



$y = a + 0.1x$

$y = a + 0x$

- Early (predictor): early winter snowfall from Sep 1 until Dec 31 (inches)
- Late (response): late winter snowfall from Jan 1 to Jun 30 (inches)
- Dash line = regression line
- Solid line = average Late snowfall (slope=0)
- Can Early predict Late? (Is the slope significantly different from 0?)

Mean functions

- Two characteristics of the distribution of the Y given X = x:
  - 1. mean functions
  - 2. variance functions
- define mean function:

$$E(Y \mid X = x) = f(x)$$

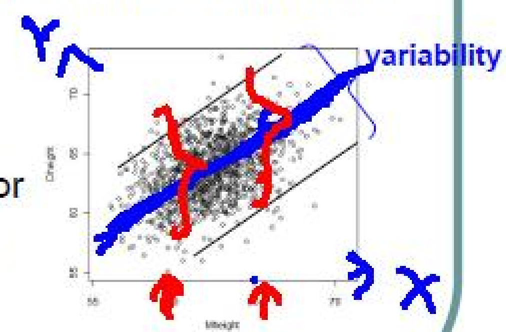
  - expected value of the response when the predictor is fixed as X=x
- e.g.,
  - Linear regression: $f(x)=a+bx$
  - Polynomial regression: $f(x)=a + bx + cx^2$
  - Heights data
    - $E(\text{Dheight} \mid \text{Mheight} = x) = \beta_0 + \beta_1 x$
    - parameters: β_0 (intercept), β_1 (slope)
    - β_0 ,β_1 need to be estimated from data
      - It is found that β_1's estimate <1. e.g. Mheight=70inch →E(Dheight)=68
      - Regression – extreme values regress towards the mean

Variance functions

- Two characteristics of the distribution of the Y given X = x:
  - 1. mean functions
  - 2. variance functions
- Define variance function:

$$\mathrm{Var}(Y \mid X = x) = \sigma^2$$

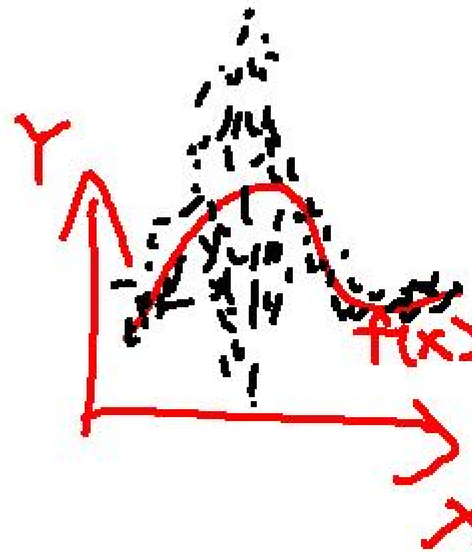
  - Variance of the response is the same for all value of predictor x
  - This is assumed for good statistical properties of the estimators
- e.g.,
  - Heights data
    - $\mathrm{Var}(\text{Dheight} \mid \text{Mheight} = x) = \sigma^2$
    - from the scatterplot, the variance function for Dheight|Mheight is approximately the same across Mheight



General

$$E(Y \mid X = x) = f(x)$$

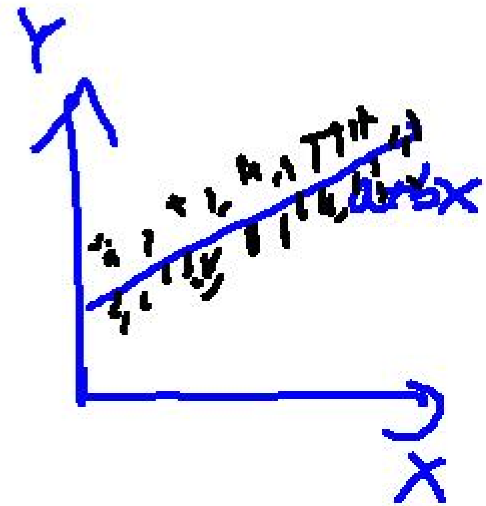
$$Var(Y \mid X = x) = g(x)$$



Linear regression

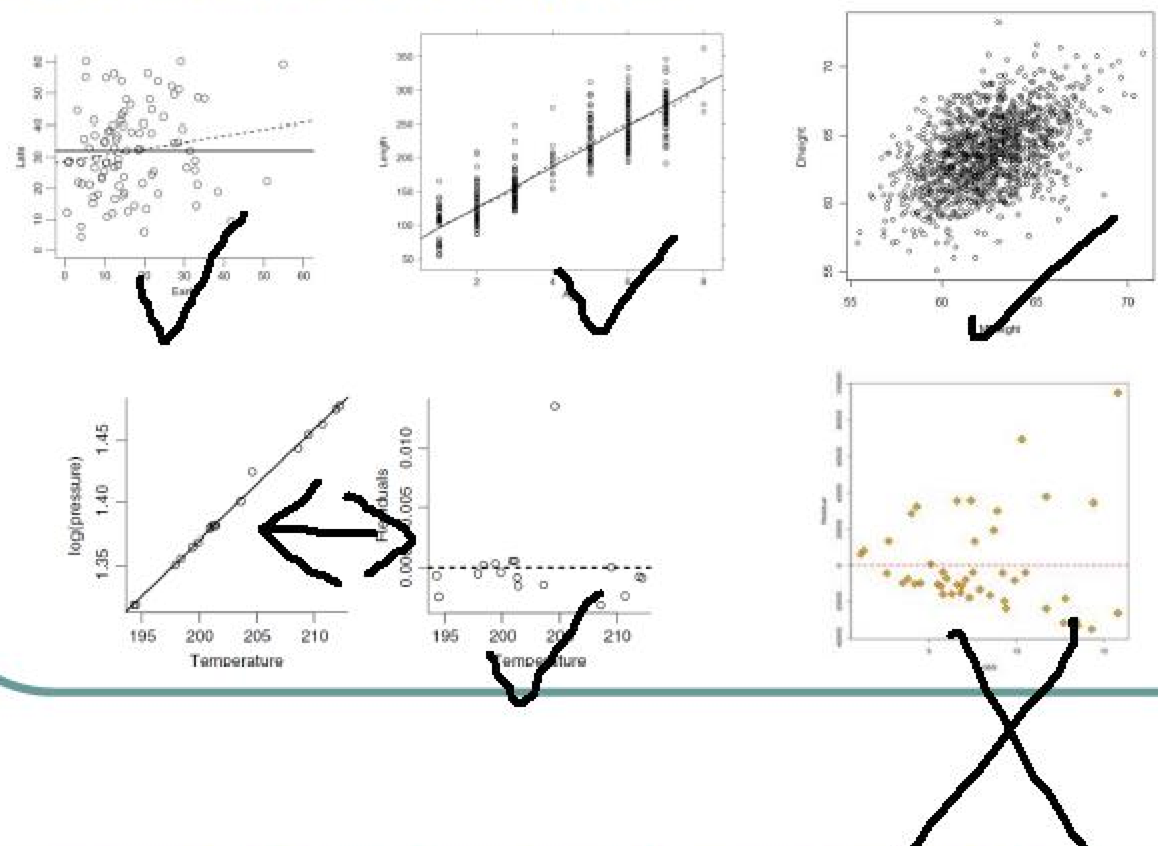
$$E(Y \mid X = x) = a + bX$$

$$Var(Y \mid X = x) = \sigma^2$$



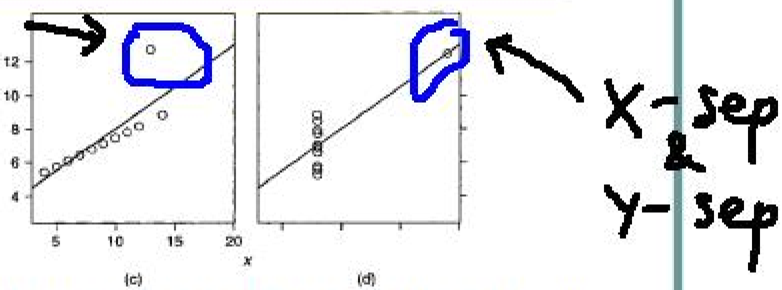
Variance functions

- Constant variance?

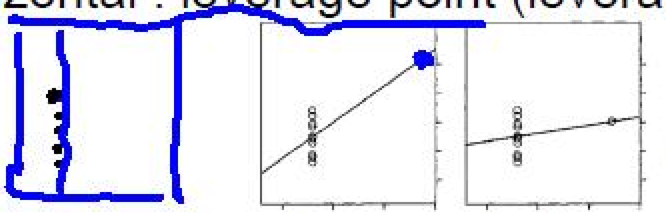


Separated points

- separated points:
  - points are well separated from the other points (horizontal or vertical)



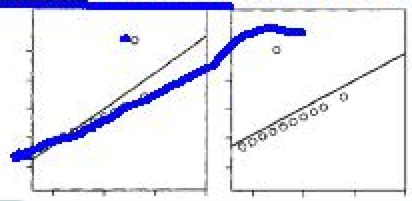
- Horizontal : leverage point (leverage effect to the line)



leverage = affect the regression line

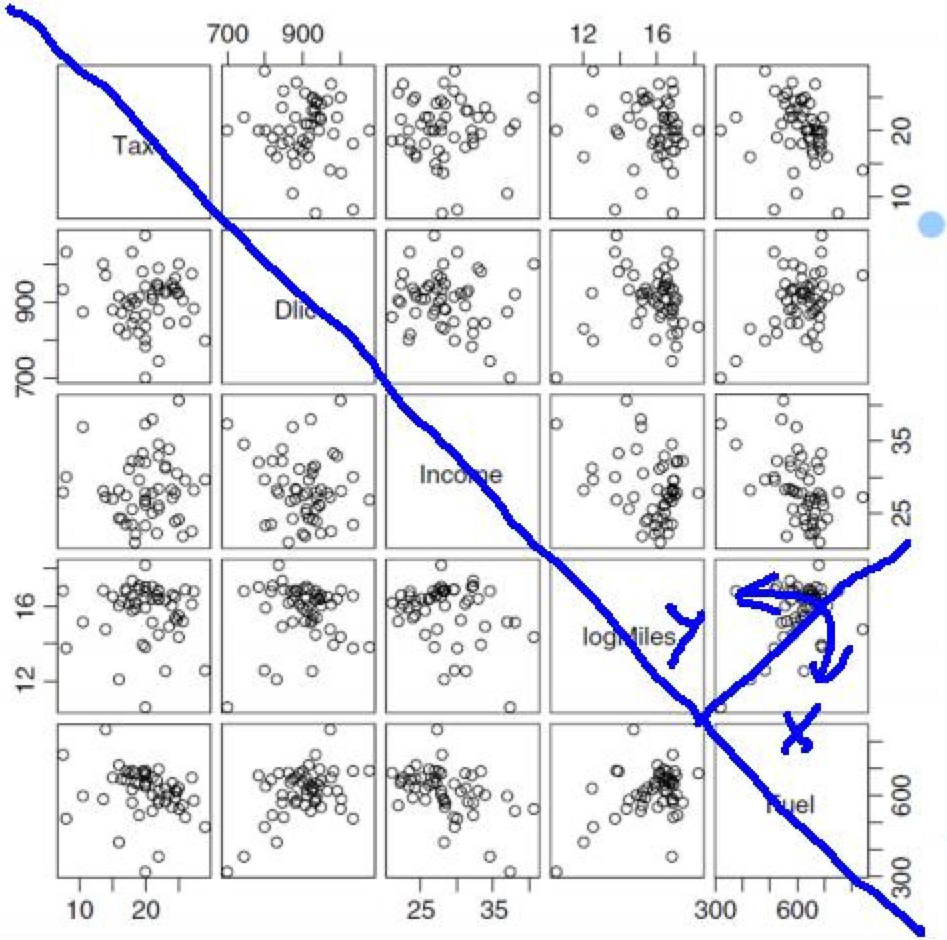
i.e., the regression lines with and without the point are very different.

- Vertical : Outlier (lie outside the line)



Usually separate points in the y co-ord but not x co-ord does not affect the line much

- Scatterplot for every combination
- Caution!!
  - Only marginal relationship between two variables is
  - Joint relationship (3 or more variables' interaction) c



- **In this example,**
  - Fuel is response
  - Relationship b/w predictors are rathe (null plots)
  - So marginal plots informative already about higher order
  - Note that the ma

Computer program: R

- Example of loess (textbook p.14)

```
x=read.table("C://heights.txt",header=T)
library(alr3);data(heights); x=heights
plot(x$Mheight,x$Dheight)
    with(x,plot(Mheight,Dheight))
fit<-lm(x$Dheight~x$Mheight)
    abline(a=fit$coef[1],b=fit$coef[2])
with(x,lines(lowess(Dheight~Mheight,f=0.2),lty=2,
col=4))
```

linear model

lm(y~ x1 + x2 +x3)

$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3$