Chapter 1

Scatterplot and Regression

Motivation example 1

- What is the value of gravity?
- Remember this?
  - $v=u+at$

Motivation example 1

- $v=u+gt$
- Experiment
  - Drop sth from the top of different buildings
  - Record the landing speed and travelling time

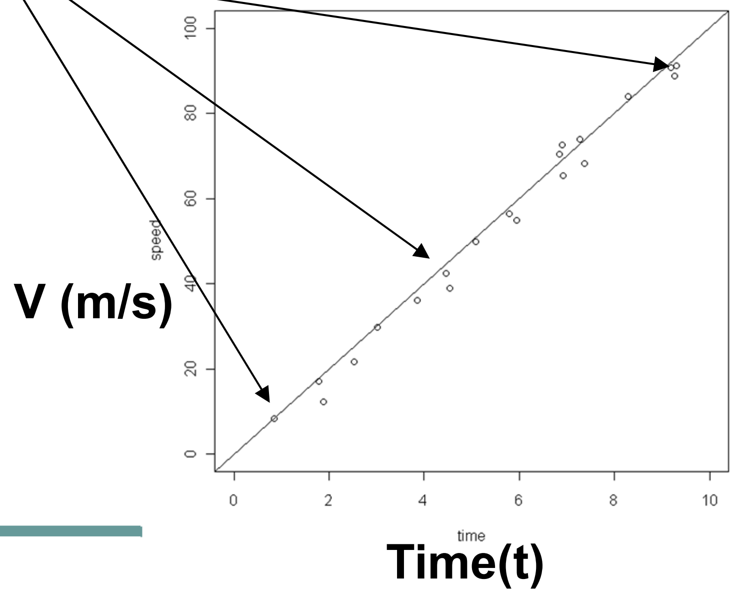
Building	V (final speed m/s)	t (time s)
LSB	11	1
MMW	50	5.2
IFC	93	9
…	…	…

Motivation example 1

- True relation: v=gt
- Interest: find g
- Estimated quantities

Building	V (final speed m/s)	t (time s)
LSB	11	1
MMW	50	5.2
IFC	93	9
…	…	…

  - The estimated quantities do not **exactly** follow v=gt
    - Measurement error
    - Air resistance
    - …
  - The intercept of the line ≈ 0
  - The slope of the line ≈ g.
- How to draw the line in a professional manner???



Motivation example 2

- Want: predict the grade point average (GPA) of all STAT3008 students.
- To do this:
  1. Select a random sample of past STAT3008 students.
  2. Record the GPA of each student
  3. Record some properties which may be useful for prediction, e.g. IQ, AL-score
  4. Use the information obtained in 2&3 to predict the GPA of this year's STAT3008 students.

Motivation example 2

- Suppose we decided to relate GPA(Y) to IQ(X)

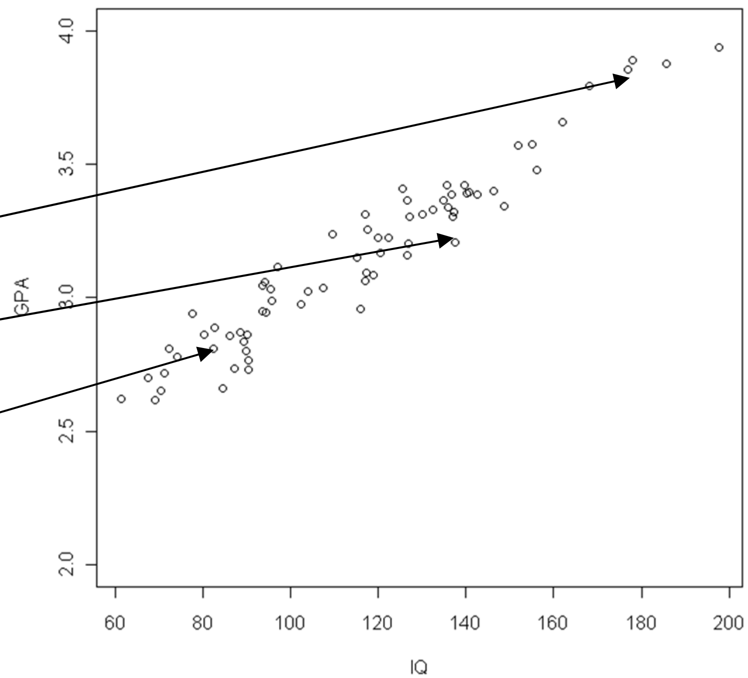
Student	GPA	IQ
Siu Chan	3.8	180
Siu Cheung	3.2	140
Siu Lee	2.7	90
Siu …	…	…

- How to understand the relationship between GPA and IQ?

Motivation example 2

- Scatter plot

Student	GPA	IQ
Siu Chan	3.8	180
Siu Cheung	3.2	140
Siu Lee	2.7	90
Siu …	…	…

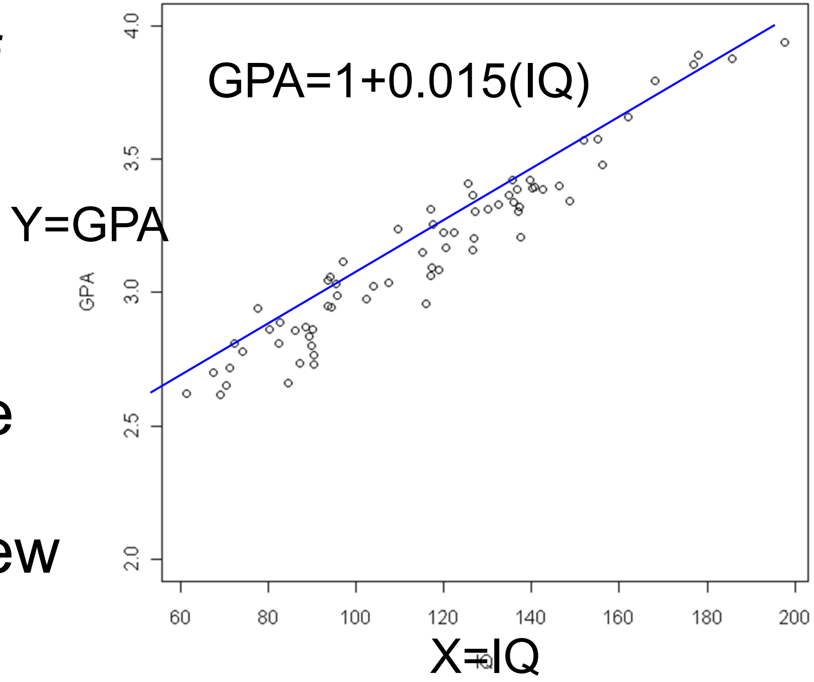


- How to use a mathematical model that relates Y(GPA) to X(IQ) and best fits the data?

Linear Regression in 1 page

Steps

1. select a random sample of 3008 students
2. record y(GPA) and x (IQ)
3. plot them on a scatterplot
4. Find the equation of a straight line that best fit the data points
5. predict the GPA using a new student's IQ.



Regression is the study of dependence between Predictors (X) and Responses (Y)

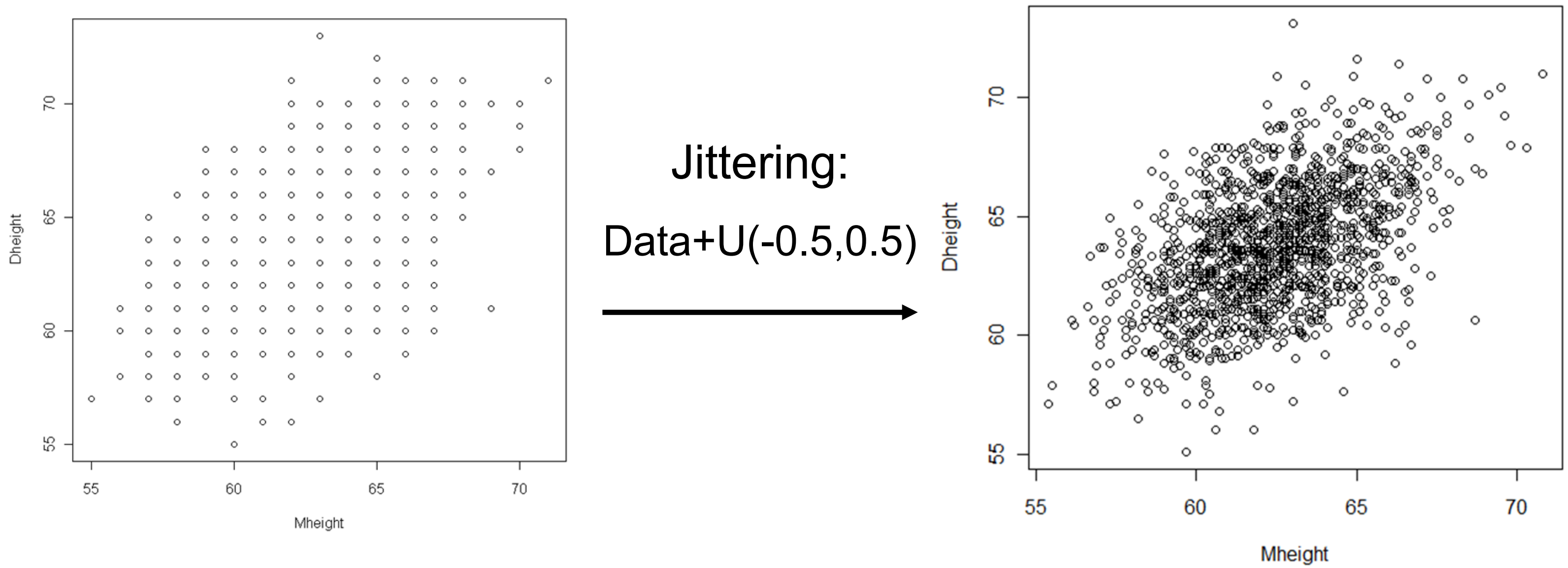
Linear Regression $Y=a+bX$

Regression is the study of dependence between Predictors (X) and Responses (Y)

Associated questions to consider

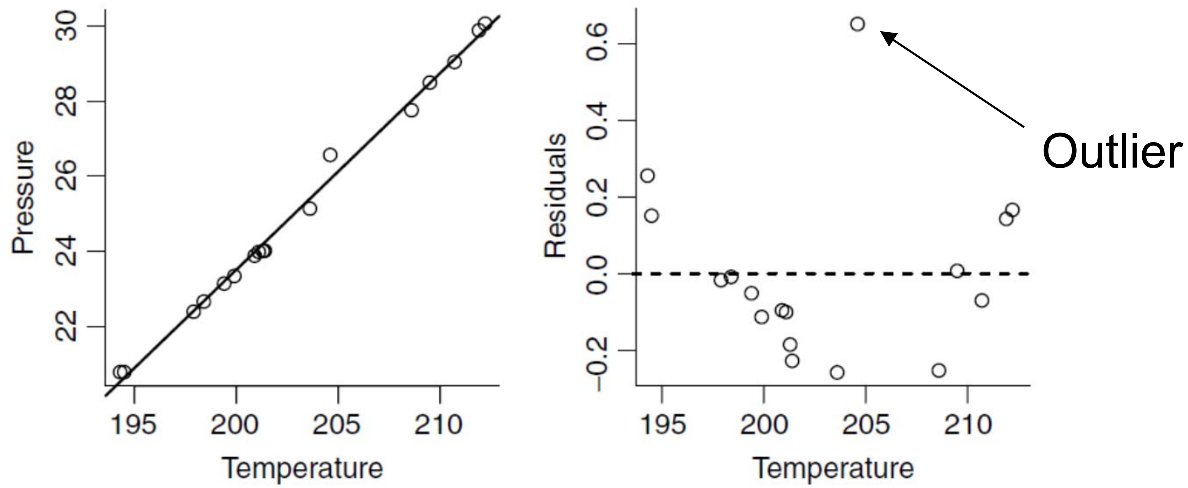
- Find the equation (intercept a and slope b) e.g. gravity
- Prediction of future values of a response (forecast unknown Y using observed X) e.g GPA vs IQ
- Discovering which predictors are important
- Does a straight line fits the data well?
- If the straight line doesn't fit well, how can we improve the fit?

Examples 1 – Heights data
"Do taller mothers have taller daughters?"

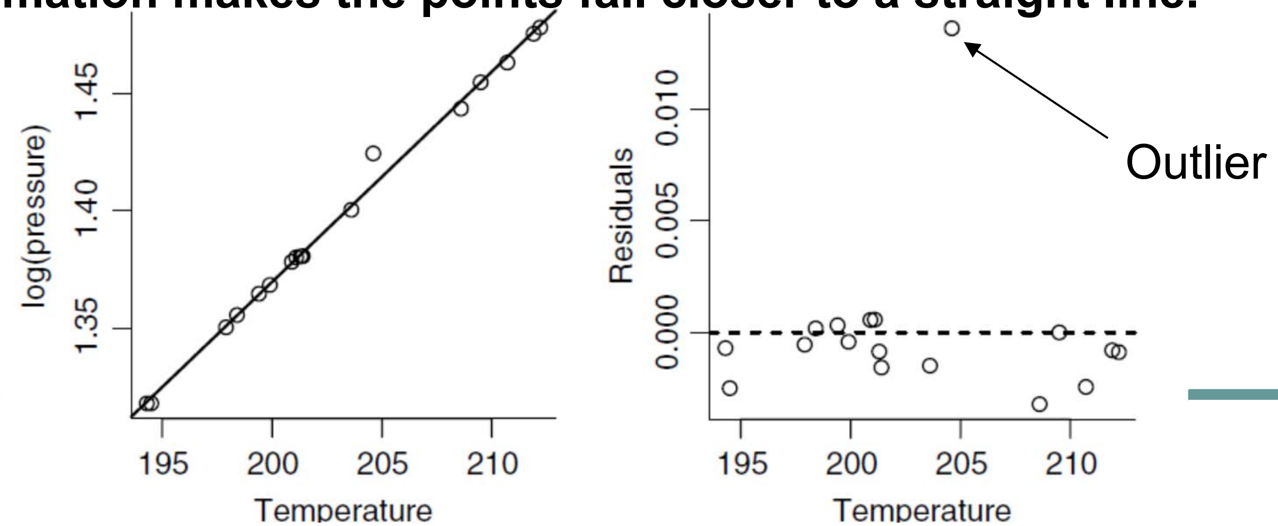


- **Axis are the same (55-70) → mother height ≈ daughter height**
- **Daughter height increases with mothers height**
- **Slope seems a little smaller than 45°. Daughter not as tall as mother**
- **The scatter of points appears elliptically shaped**

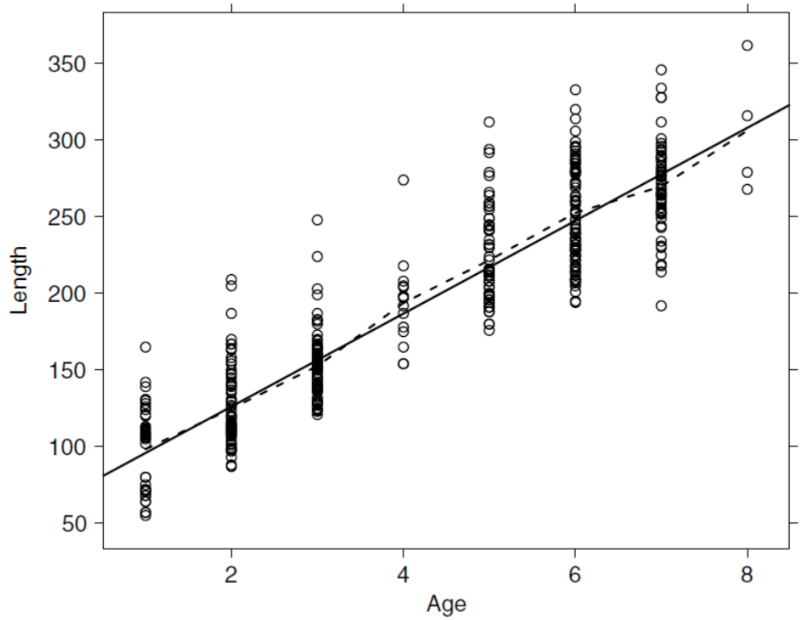
Examples 2 – Forbes Data
Measure pressure from boiling point of water



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

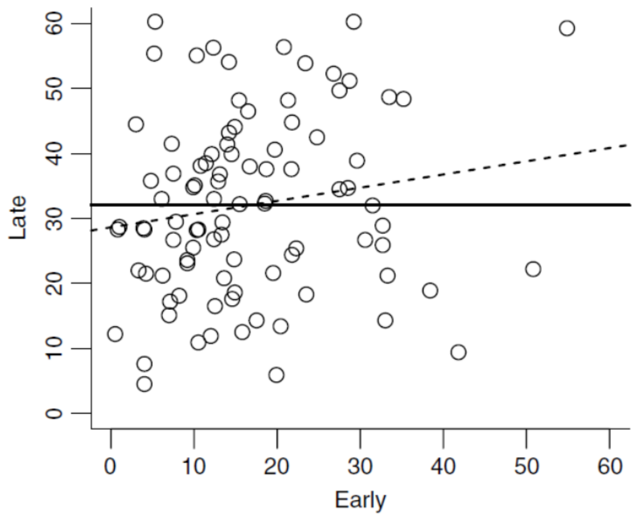


Examples 3 – Smallmouth Bass
Size vs Age of fish



- **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



- 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:

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

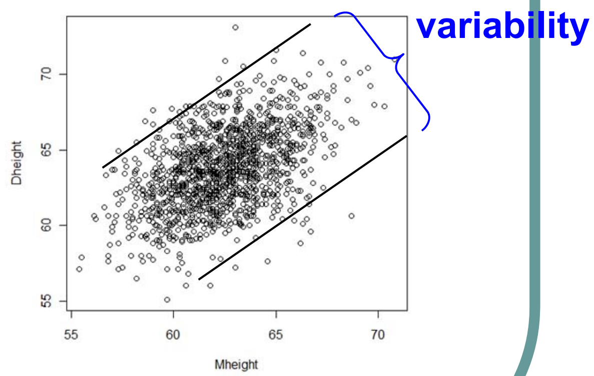
  - expected value of the response when the predictor is fixed as X=x
- e.g.,
  - Linear regression: $f(x)=a+bx$,
  - Polynomial regression: $f(x)=a + bx + cx^2$
  - Heights data
    - $E(\text{Dheight} \mid \text{Mheight} = x) = \beta_0 + \beta_1 x$
    - parameters: β_0 (intercept), β_1 (slope)
    - 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:

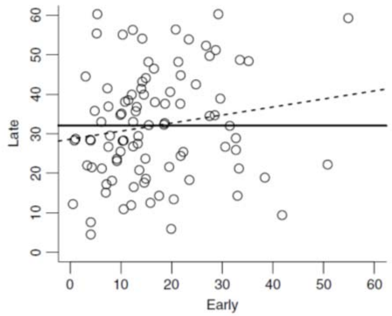
$$Var(Y \mid X = x) = \sigma^2$$

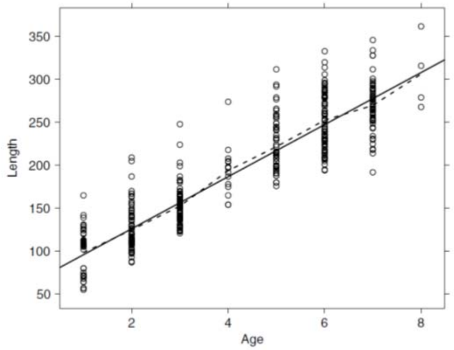
  - Variance of the response is the same for all value of predictor x
  - This is assumed for good statistical properties of the estimators
- e.g.,
  - Heights data
    - $Var(Dheight \mid Mheight = x) = \sigma^2$
    - from the scatterplot, the variance function for Dheight|Mheight is approximately the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

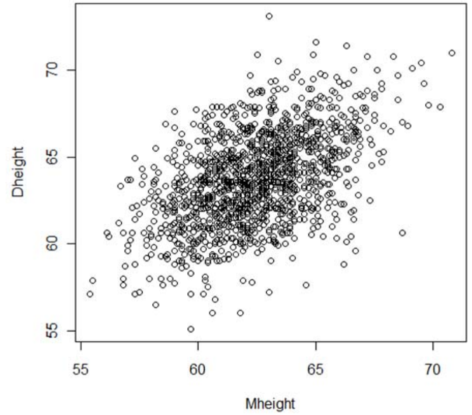


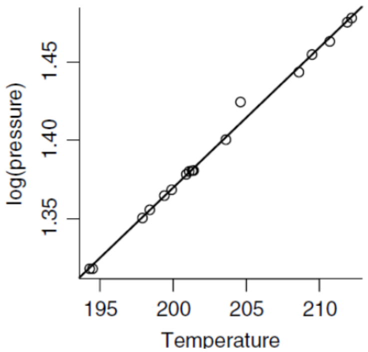
Variance functions

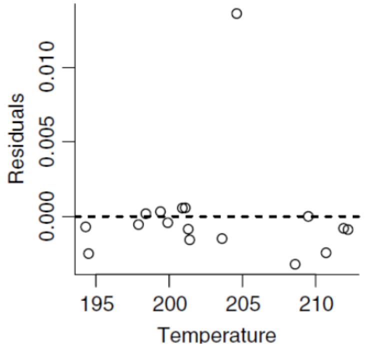
- Constant variance?

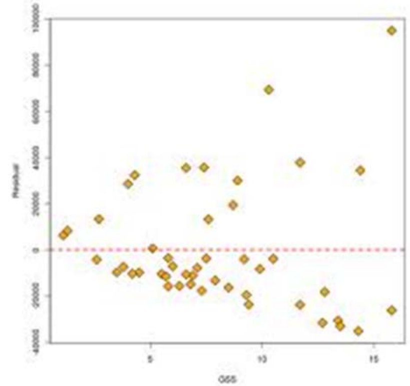






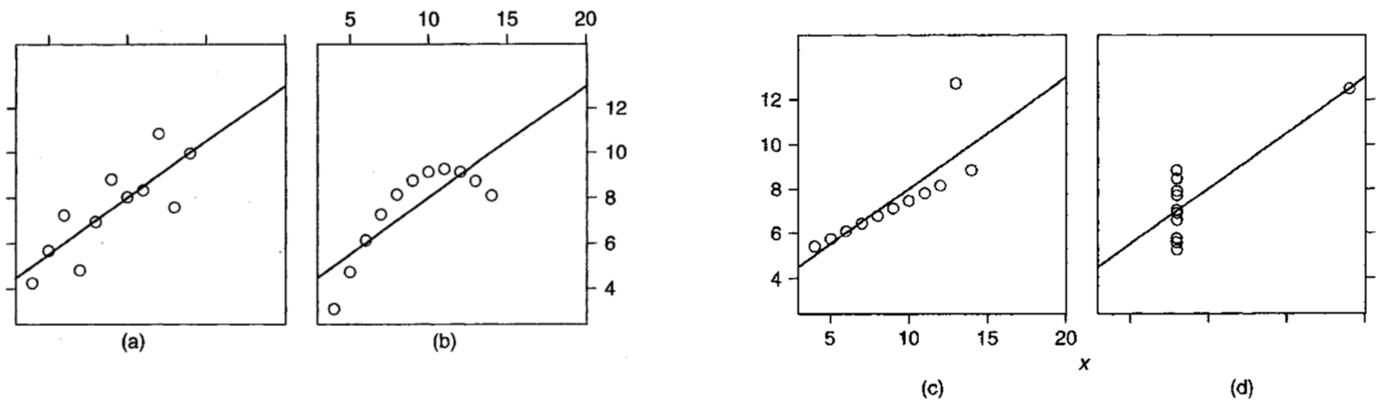






Four hypothetical data sets

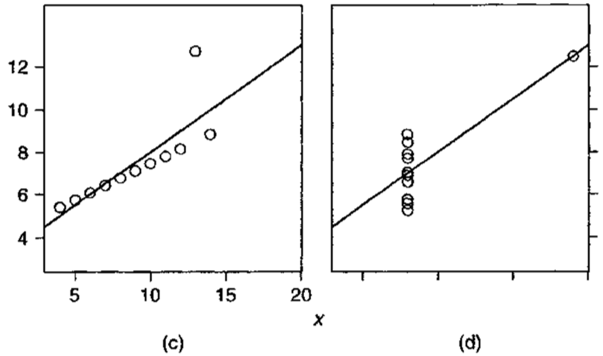
- See Textbook Table 1.1 for exact values of 4 data sets
- each data set leads to the same results (estimated intercept and slope, other summary statistics)



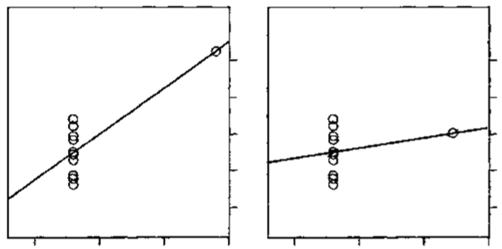
- Conclusions
  - Dependence is not limited to E(Y|X)=a+bX. (may be a curve)
  - Summary statistics may not give a good summary of dependence
  - Need to examine summary graph (scatterplot) first

Separated points

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



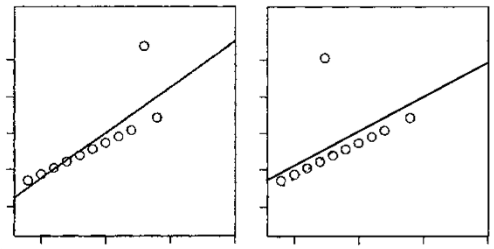
- Horizontal : leverage point (leverage effect to the line)



leverage = affect the regression line

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

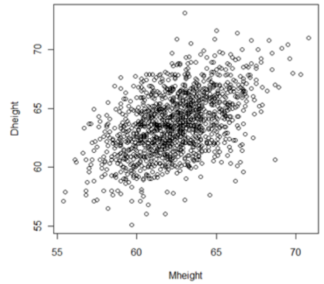
- Vertical : Outlier (lie outside the line)



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

Tools for looking at Scatterplots

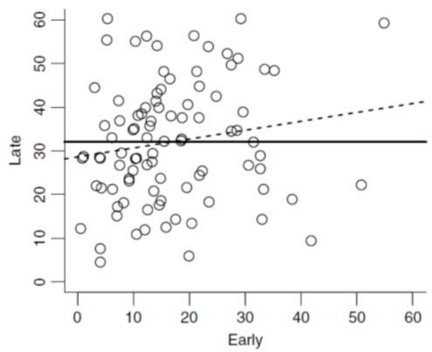
- Scatter plot shows
  - mean function
  - variance function
  - any separated points



1. Mean function – linear
2. Variance function – constant
3. No separate point

- null plot
  - constant mean function (=0)
  - constant variance function
  - no separated points

Snowfall data



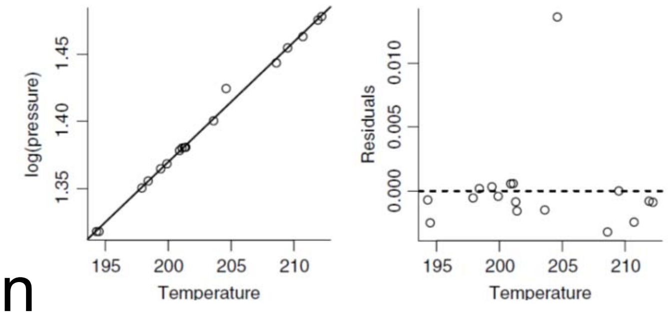
- 3 tools enhancing scatterplot
  - Size
  - Transformation
  - Smoothers

Tools for looking at Scatterplots

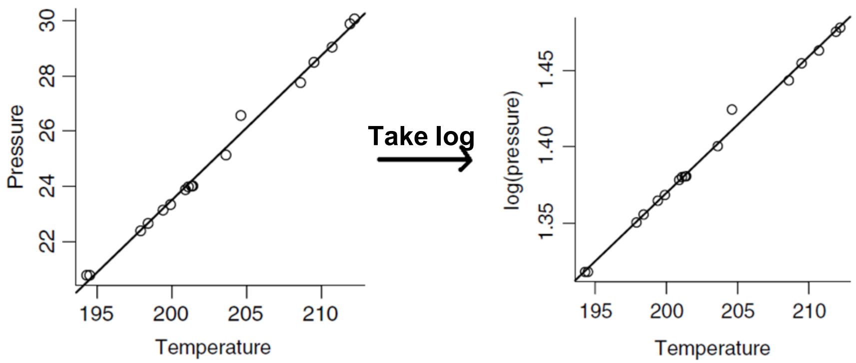
- **Three tools for scatterplot**
  - 1. Size   2. Transformation   3. Smoothers

1. Size

- Changing scales, removing linear trend



2. Transformation

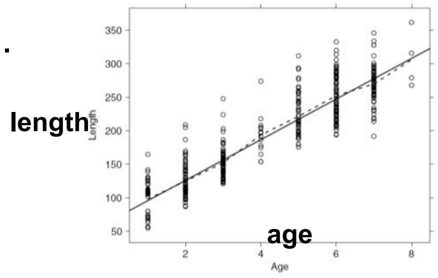


Tools for looking at Scatterplots

- **Three tools for scatterplot**
  - 1. Size   2. Transformation   3. Smoothers

3. Smoothers (e.g. loess: locally weighted scatterplot smoothing)

- A scatterplot smoother
  - estimates the mean function E(Y |X = x) as x varies
- No parametric assumptions about the mean function.
  - e.g. E(length|Age) is estimated for each age in Fish data.

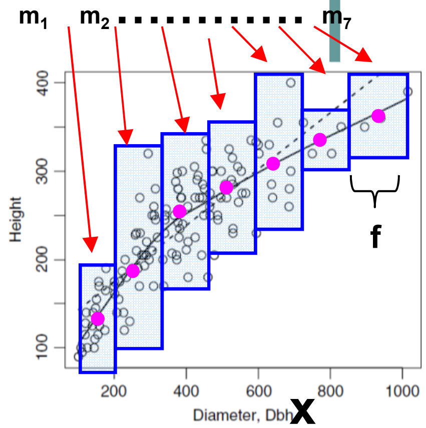


- loess smoother (locally weighted scatterplot smoother):
  - Idea: Use the “local data” (observation near x) to find E(Y|X=x), for various x.

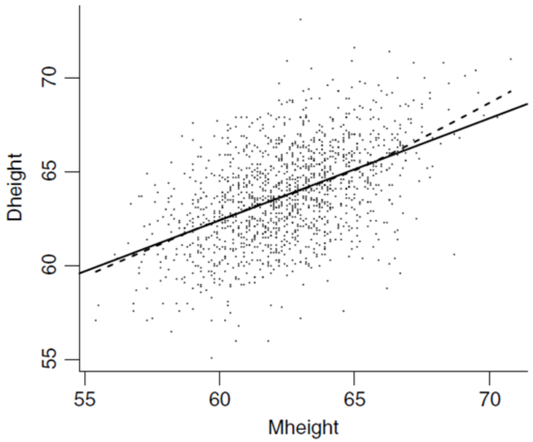
Tools for looking at Scatterplots

- Loess (locally weighted scatterplot smoother smoother):
  - estimates E(Y |X = x) by fitting a straight line to a fraction (f) of point closest to x
  - Giving more weight to points close to x than to points distant from x
  - Procedure:

1. Specify f and some x_i . 2. Find $E(Y|X=x_i)$ for each x_i. 3. Join $(x_i, E(Y|X=x_i))$

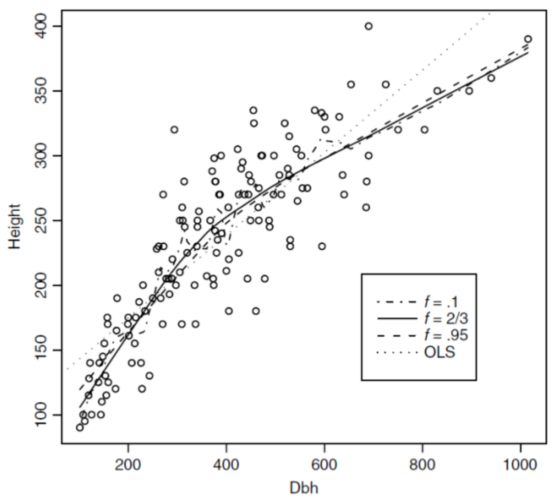


- Example : Heights data
  - dash line: smoother
  - solid line: Linear Regression



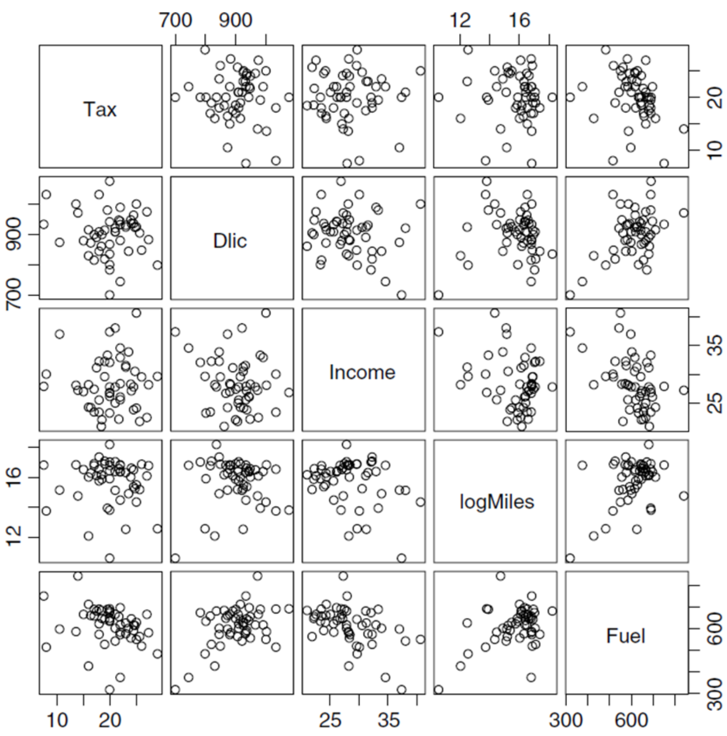
- Example : See p.277
  - Compare to the Heights data,
  - Linear Regression is bad. Need transformation

ESTIMATING E(Y|X) USING A SMOOTHER



Scatterplots Matrix

- What to do if there are more than 2 variables?
  - Scatterplot for every combination
  - Caution!!
    - Only marginal relationship between two variables is observed.
    - Joint relationship (3 or more variables' interaction) can't be seen



- **In this example,**
  - Fuel is response(y)
  - Relationship b/w pairs of predictors are rather weak (null plots)
  - So marginal plots are quite informative already, don't worry about higher order interaction
  - Note that the matrix is symmetric

Computer program: R

- Data can be obtained from
  - http://www.stat.umn.edu/alr/data/
  - Or the download the R package alr3
- Example (see textbook p.15)

```
x=read.table("C://fuel2001.txt",header=T) #(read .txt file)
library(alr3);data(fuel2001); x=fuel2001
    x$Fuel=x$FuelC/x$Pop
    x$Dlic=x$Drivers/x$Pop
    x$LogM=log(x$Miles,2)
pairs(x[,c(7,9,3,10,8)])
    with(x,pairs(cbind(Tax,Dlic,Income,LogM,Fuel)))
plot(x[,8],x[,10])
```

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))
```