STOR 455
STATISTICAL METHODS I

Jan Hannig

Regression

- Read Chapter 2
- We will now begin with *straight line regression* (Chapter 3)

SAS Example (Task 2.3.1)

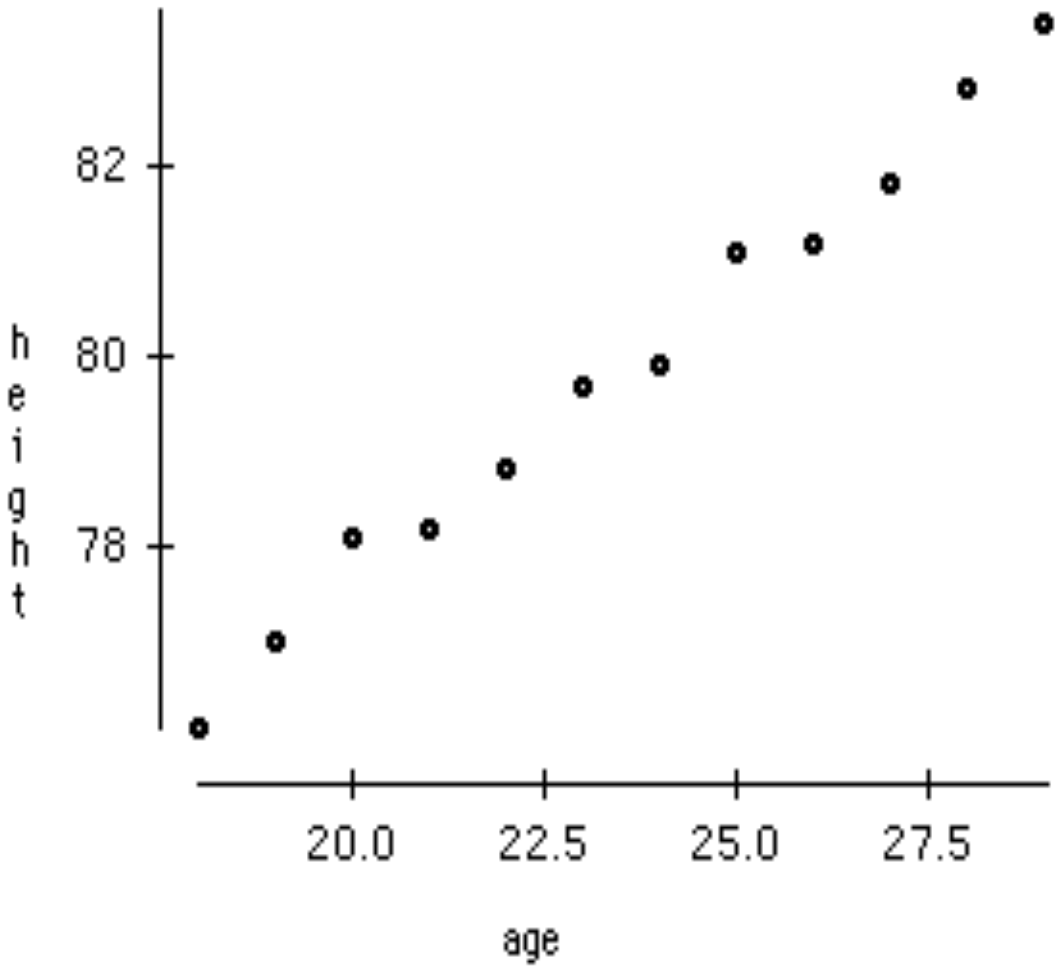
- ```
  data car;
  infile 'T:\...\CAR.DAT';
  input carno mtcost price miles;
  run;
  ```
- ```
  proc contents data=car;
  run;
  ```
- ```
  data subpop;
  set car;
  if miles=14000;
  proc print data=subpop;
  run;
  ```
- ```
  proc chart data=car;
  hbar mtcost;
  run;
  ```
- ```
  proc plot data=car;
  plot mtcost*miles='*'/hpos=50 vpos=15;
  run;
  ```
  ```

Example: Age and Height (Section 3.2)

- The height of a baby is not stable but increases over time.
- Pattern of growth varies from child to child.
- To understand the general growth pattern, use the ave. of several children's heights

Age	Height
18	76.1
19	77
…	
29	83.5

Scatterplot of Height vs. Age



Example: Age and Height

- Response variable: Height (Y)
- Explanatory variable: Age (X)
- Can we use a straight line to summarize the relationship? (Description)
- Can we predict a new baby's height if we know its age? (Prediction)

Simple Linear Regression Models

- Describe statistical relationship between variables (vs. functional relationship)
- Linear relationship (vs. non-linear)
- Simple and easy to study
- May be good approximation within a region even if the true relationship is nonlinear
- A base for nonlinear models

Where can we use simple linear regression?

- Give me some example where simple linear regression model might be appropriate:

- Give me some example where simple linear regression model might not be appropriate:

Data for Simple Linear Regression

- Y_i the response variable
- X_i the explanatory variable
- i = 1 to n, n is the sample size

Simple Linear Regression Model (Section 3.3)

- $Y_i = \beta_0 + \beta_1 X_i + \xi_i$
  - Y_i is the value of the response variable for the i^{th} case
  - X_i is the value of the explanatory variable for the i^{th} case
  - ξ_i is a (normally distributed) random error with mean 0 and variance σ^2

Parameters of SLRM

- $Y_i = \beta_0 + \beta_1 X_i + \xi_i$
  - β_0 the intercept
  - β_1 the slope
  - $\sigma^2 = var(\xi_i)$ the variance of the error term
- Alternative form of SLR requires (X,Y) to be bivariate normal.

Features of SLRM

- Y_i random, X_i fixed
- $E(Y_i) = \beta_0 + \beta_1 X_i$
- $Var(Y_i) = var(\xi_i) = \sigma^2$
- $\rho(Y_i, Y_j) = 0$

Likelihood

$$Y_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma\right)$$

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2}$$

$$L = f_1 \cdot f_2 \cdots f_n - \text{ likelihood function}$$

Fitted Reg. Equation and Residuals

- $\hat{Y}_i = b_0 + b_1 X_i$
  - b_0 is the estimated intercept
  - b_1 is the estimated slope
- e_i residual for case i
- $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$

Least Squares

- minimize $\Sigma(e_i)^2 = \Sigma(Y_i - (b_0 + b_1 X_i))^2$
- use calculus
- take derivative with respect to b_0 and with respect to b_1
- set the two resulting equations equal to zero and solve for b_0 and b_1

Least Squares Solution

$$b_1 = \frac{SXY}{SXX}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$SXY = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$SSX = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- $(\bar{X}, \bar{Y})$ on fitted regression equation.

Using Regression Equation

- Estimate mean response $E(Y_h)$ by $\hat{Y}_h = b_0 + b_1 X_h$
- Predicting new observation Y_h by $\hat{Y}_h$
- Interpretation of b_1
- Properties of fitted value $\hat{Y}_i$ and residuals e_i
- See age and height example

Using Regression Equation

- Estimate mean response $\mu_{Yh}=E(Y_h)$ by $\hat{Y}_h = b_0 + b_1 X_h$
- Predicting new observation Y_h by $\hat{Y}_h$
- Interpretation of b_1
- Properties of fitted value $\hat{Y}_i$ and residuals e_i
- See age and height example

SAS

- Import data: Data step
- Print data: proc print
- Plot data: proc gplot
- Regression: proc reg

SAS

```
/* Import data in file "ageheight.dat" */
data ageheight;
input age height;
cards;
18 76.1
19 77.0
20 78.1
21 78.2
22 78.8
23 79.7
24 79.9
25 81.1
26 81.2
27 81.8
28 82.8
29 83.5
;
run;
/* Look at the data */
proc print data=ageheight;
run;
```

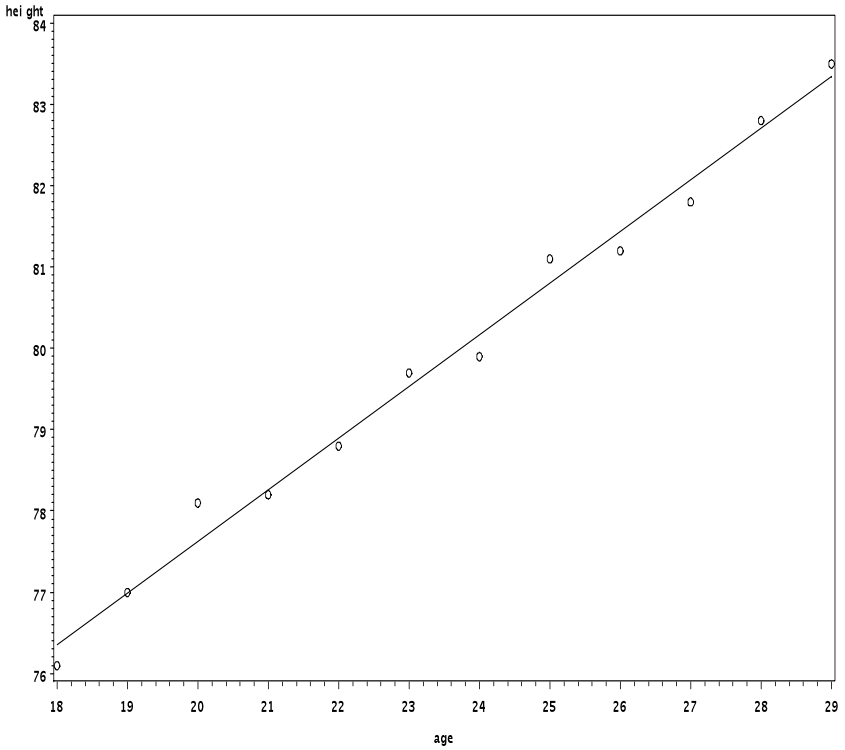
SAS

Obs	age	height
1	18	76.1
2	19	77.0
3	20	78.1
4	21	78.2
5	22	78.8
6	23	79.7
7	24	79.9
8	25	81.1
9	26	81.2
10	27	81.8
11	28	82.8
12	29	83.5

SAS

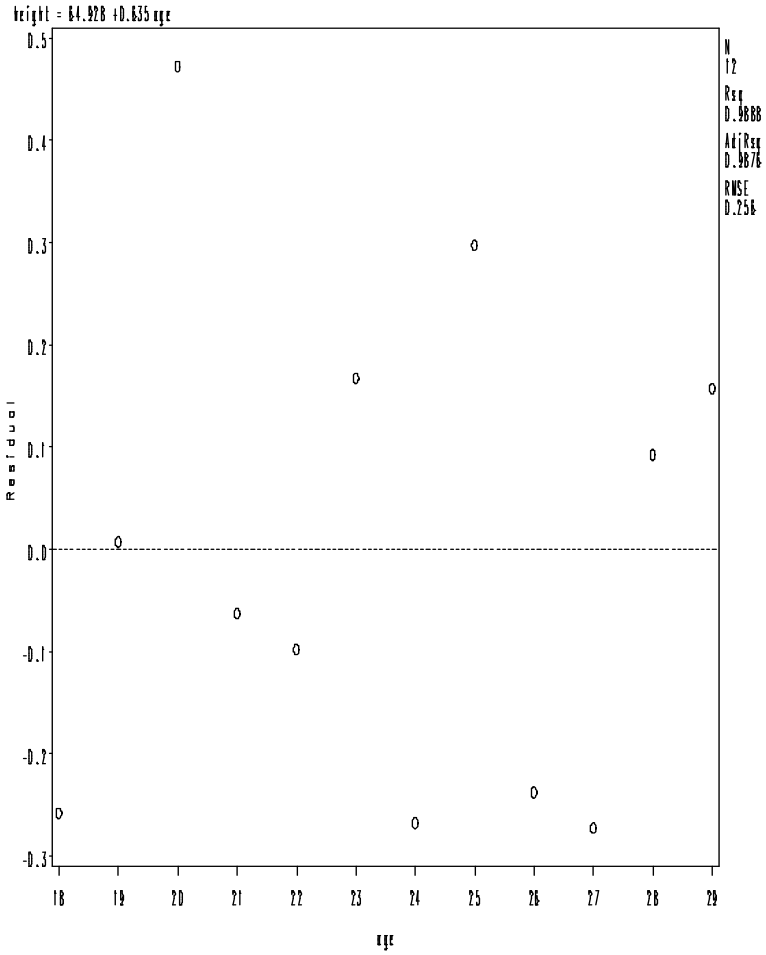
```
/* Plot the data */
symbol1 v=circle;
proc gplot data=ageheight;
    plot height*age;
run;

/* Add regression line */
symbol1 v=circle i=rl;
proc gplot data=ageheight;
plot height*age;
run;
```



SAS

```
/* Linear
   regression, with
   residual plot */
proc reg
   data=ageheight;
   model
   height=age;
   plot r. *age;
run;
```



SAS

The REG Procedure
Model: MODEL1
Dependent Variable: height

Number of Observations Read 12
Number of Observations Used 12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	57.65483	57.65483	879.99	<.0001
Error	10	0.65517	0.06552		
Corrected Total	11	58.31000			

Root MSE	0.25596	R-Square	0.9888
Dependent Mean	79.85000	Adj R-Sq	0.9876
Coeff Var	0.32056		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	64.92832	0.50841	127.71	<.0001
age	1	0.63497	0.02140	29.66	<.0001

Why Least Square Estimators?

- They are unbiased. In fact, they are Best Linear Unbiased Estimators (BLUE)
- Other approach possible
  - maximum likelihood estimators if error has normal distribution
  - Bayesian estimators

Gauss-Markov Theorem

- LS estimators are unbiased:

  $E(b_0) = \beta_0$

  $E(b_1) = \beta_1$

- LS estimators have minimum variance among all unbiased estimators.

Estimation of σ^2

$$s^2 = \frac{\sum\left(\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i}\right)^2}{n-2} = \frac{\sum \hat{e}_\mathrm{i}^{\,2}}{n-2} = \frac{SSE}{df_E} = MSE$$

$$s = \sqrt{s^2} = \text{Root } MSE$$

Alternative expression for SSE

$$SSE = SSY - \frac{(SXY)^2}{SSX}$$