

# Simple Linear Regression

Applied Regression and Other Multivariable Methods  
Sections 5-1 - 5-6

4

## Regression with a Single Predictor

- Want to find a **curve** that best fits the data  
Have a sample of  $n$  individuals  
Observe pairs  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$
- 1 Does a linear, quadratic, or log-linear appear best?  
How do we quantify and search for the “best”
- 2 What are the parameters of curve that fits best?  
How do we estimate the model parameters

4-1

## Search for Best Type of Curve

- Scatterplot of  $Y$  vs  $X$  very helpful
- Allows quick assessment of several simple models
- Search approaches (discuss more in Chpt 16)
  - Forward: Start simple and build complexity
  - Backward: Start complex and reduce complexity
  - Use scientific theory and experience
- Will focus first on forward approach
- Simplest curve is a line
  - How do we find the “best” line?
  - How do we assess the fit?
  - How do we use line for inference?

4-2

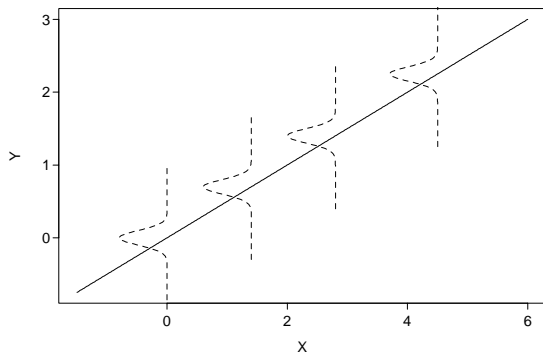
## Mathematical Properties of Line

- Straight line represented with equation
$$y = \beta_0 + \beta_1 x$$
- Can be defined based on two points
- Has two parameters,  $\beta_0$  and  $\beta_1$ 
  - $\beta_0$  is the  $y$ -intercept (where the line crosses the  $y$ -axis)
  - $\beta_1$  is the slope (change in  $y$  for a unit change in  $x$ )
- Positive slope  $\rightarrow$  if  $x$  increases,  $y$  increases
- Negative slope  $\rightarrow$  if  $x$  increases,  $y$  decreases

4-3

## The Straight-Line Model

- How do we determine the “best” line?
  - Very unlikely all obs pairs fall directly on line
  - Want to use sample to infer population line
  - What is meant by a population line?
- Graphical Representation of Linear Model



4-4

## Assumptions

- **Existence**

For any  $X$ ,  $Y$  is a rv with mean  $\mu_{Y|X}$  and var  $\sigma_{Y|X}^2$

- **Independence**

The  $Y$ 's are statistically independent of one another

Common violation - same indiv measured over time

Crucial assumption - can lead to spurious results

- **Linearity**

Mean of  $Y$  is a linear function of  $X$  ( $\mu_{Y|X} = \beta_0 + \beta_1 X$ )

Implies  $Y = \beta_0 + \beta_1 X + E$

$E$  is rv with mean 0 and var  $\sigma_{Y|X}^2$

$E_i = Y_i - (\beta_0 + \beta_1 X_i)$  : vertical dist between  $Y_i$  and  $\mu_{Y|X}$

Will use  $E$ 's to assess fit

4-5

## Assumptions

- **Homoscedasticity**
  - The variance of  $Y$  is constant for any  $X$  :  $\sigma_{Y|X}^2 \equiv \sigma^2$
- **Normal distribution**
  - For any  $X$ , the dist of  $Y \sim N(\beta_0 + \beta_1 X, \sigma)$
  - Will discuss procedures robust to normality assumption
  - Least critical assumption, especially if a large data set
- Assumptions pertain to  $Y|X$  (not  $Y$ )
- More convenient to assess assumptions using  $E$ 's
- Assumptions on  $Y|X \rightarrow E \sim N(0, \sigma)$

4-6

## Estimating the Best Line

- Wish to obtain estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$
- Will use  $\hat{\cdot}$  notation to denote any sample estimate
- Given  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , can also calculate  $\hat{E}_i$ ,  $\hat{\sigma}^2$ , etc
- The least-squares method

Determines line that minimizes  $\sum \hat{E}_i^2$

Minimizes sum of squared vertical deviations

Find  $\hat{\beta}_0$  and  $\hat{\beta}_1$  to minimize

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

$\sum (Y_i - \hat{Y}_i)^2$  known as sum of squares due to error

Least-squares minimizes sum of squares due to error (SSE)

LS parameters explain as much SS in  $Y$  as possible

4-7

## Least-Squares Solutions

- Other methods give same estimates
- Minimum variance
  - Unbiased  $E(\hat{\beta}_0) = \beta_0, E(\hat{\beta}_1) = \beta_1$
  - Variance  $\hat{\sigma}^2$  is minimized
  - Best linear unbiased estimators (BLUE)
- Maximum likelihood (if  $E$ 's normal)
- Formulas of estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\sigma}^2 = S_{Y|X}^2 = SSE / (n - 2)$$

4-8

## Test Example Using SAS

```
options nocenter linesize=72;
goptions colors=(none);

title 'Data used in Chapter 5 Example';

data table5;
  infile 'C:\TeX\st502\DATA\example5-1.dat';
  input indiv sbp age;

proc print;

symbol1 v=circle i=none;
proc gplot; plot sbp*age/ frame;

proc sort; by age;
symbol1 v=circle i=sm55;
proc gplot; plot sbp*age/ frame;

symbol1 v=circle i=rl;
proc gplot; plot sbp*age/ frame;

proc reg simple;
  model sbp=age /cli clm;
  output out=fit r=res p=pred;
run;
```

4-9

## Log Window

```
options nocenter linesize=72;
goptions colors=(none);
title 'Data used in Chapter 5 Example';
data table5;
  infile 'I:\.www\datasets502\example5-1.dat';
  input indiv sbp age;

NOTE: The infile 'I:\.www\datasets502\example5-1.dat' is:
      File Name=I:\.www\datasets502\example5-1.dat,
      RECFM=V,LRECL=256

NOTE: 30 records were read from infile 'I:\.www\datasets502\example5-1.dat'.
      The minimum record length was 8.
      The maximum record length was 9.

NOTE: The data set WORK.TABLE5 has 30 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.21 seconds
      cpu time           0.04 seconds

proc print;

NOTE: There were 30 observations read from the data set WORK.TABLE5.
NOTE: PROCEDURE PRINT used:
      real time          0.03 seconds
      cpu time           0.00 seconds
.
.
.
symbol1 v=circle i=rl;
proc gplot; plot sbp*age/ frame;
run;

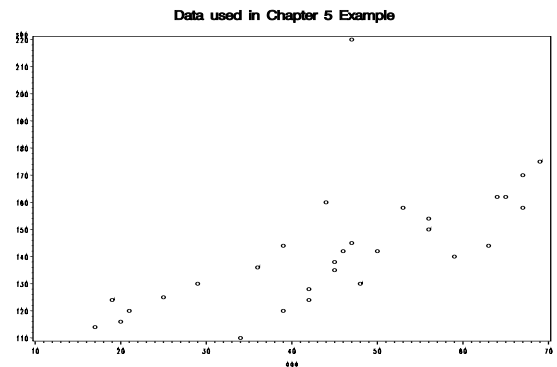
NOTE: Regression equation : sbp = 98.71472 + 0.97087*age.
```

4-10

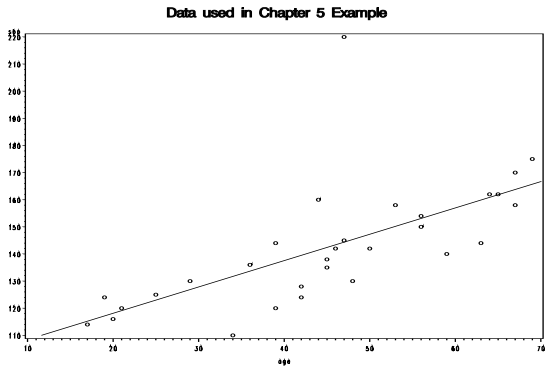
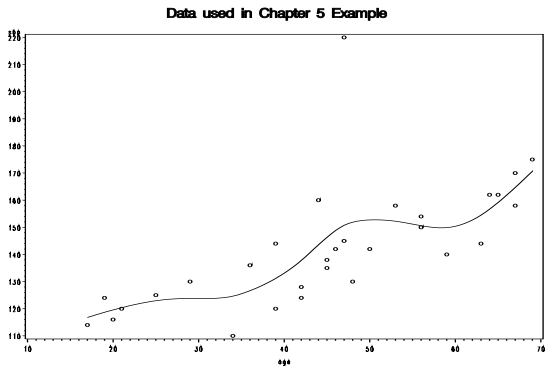
## Output

Data used in Chapter 5 Example 15:48 Thursday, January 4, 2001 7

Obs	indiv	sbp	age
1	1	144	39
2	2	220	47
3	3	138	45
4	4	145	47
5	5	162	65
6	6	142	46
7	7	170	67
8	8	124	42
9	9	158	67
10	10	154	56
11	11	162	64
.	.	.	.
.	.	.	.
.	.	.	.



4-11



4-12

## Output

The REG Procedure

Descriptive Statistics				
Variable	Sum	Mean	Uncorrected SS	Variance
Intercept	30.00000	1.00000	30.00000	0
age	1354.00000	45.13333	67894	233.91264
sbp	4276.00000	142.53333	624260	509.91264

Variable	Standard Deviation
Intercept	0
age	15.29420
sbp	22.58125

Dependent Variable: sbp

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6394.02269	6394.02269	21.33	<.0001
Error	28	8393.44398	299.76586		
Corrected Total	29	14787			

Root MSE	17.31375	R-Square	0.4324
Dependent Mean	142.53333	Adj R-Sq	0.4121
Coeff Var	12.14716		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	98.71472	10.00047	9.87	<.0001
age	1	0.97087	0.21022	4.62	<.0001

4-13