3.3 Least-Squares Regression

Regression line (line of best fit):

- straight line
- predicts y if given x
- requires explanatory & response variables
- describes relationship between these variables

http://bcs.whfreeman.com/tps3e/pages/bcs-main.asp?v=category&s=00020&n=99000&i=99020.01&o=|00510|00520|00530|00540|00590|00010|00020|00030|00040|00050|00100|00060|00070|00080|00090|00110|00120|00300|0P000|01000|02000|03000|04000|05000|06000|07000|08000|09000|10000|11000|12000|13000|14000|15000|99000|

3.3 McDonaldsData.ftm

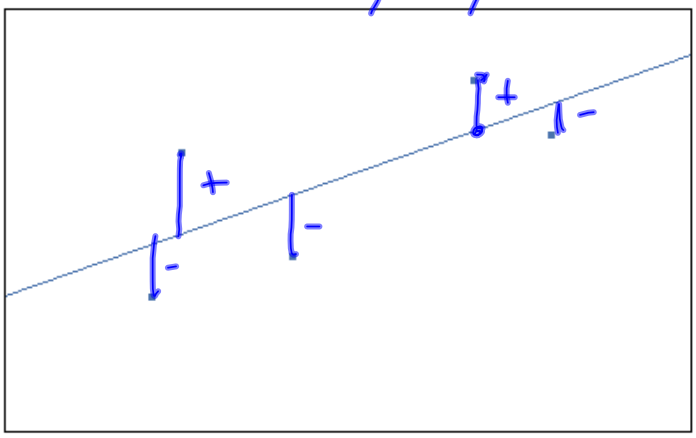
Least-squares regression line (LSRL)

- minimizes sum of squares of vertical distances between each data point and line
- through $(\bar{x},\bar{y})$
- $\hat{y} = a+bx$ $\hat{y} = b_0+b_1x$
- slope: $b=(r)(s_y)\div(s_x)$

$$b_1 = r\,\frac{s_y}{s_x}$$

obs - predicted

$y - \hat{y}$



I. Descriptive Statistics

$$\bar{x} = \frac{\sum x_i}{n}$$

$$s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}}$$

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$r = \frac{1}{n-1}\sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

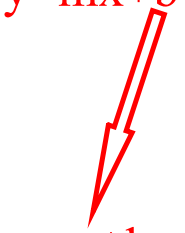
$$b_1 = r\frac{s_y}{s_x}$$

$$s_{b_1} = \frac{\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$y = mx + b$, $m = b_1$, $(\bar{x}, \bar{y})$

y intercept

y=mx+b



y=a+bx

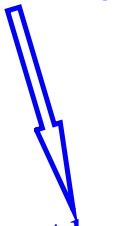
algebra form
y=mx+b

TI-83/84 form
y=a+bx

college board $\hat{y} = b_0+b_1x$

slope

y=mx+b



y=a+bx

$y = a_0 + a_1x^1 + a_2x^2 + a_3x^3 + \ldots$

example: Given: $(\bar{x},\bar{y})$ = (3,7)
r = .89
s_x = 1.5 and
s_y = 2.5
what is the equation of the LSRL?

slope = $b_1 = r\frac{s_y}{s_x}$ =(.89)(2.5)/(1.5) = 1.483.

The LSRL passes through $(\bar{x},\bar{y})$, so
7 = a + 1.483(3)
7 = a + 4.45
2.55 = a

LSRL equation: $\hat{y}$ = 2.55 + 1.483x

slope = $(.89)\left(\frac{2.5}{1.5}\right)$ = 1.483

$y = mx + b$
$7 = 1.483(3) + b$
$7 = 4.45 + b$
$2.55 = b$
$\hat{y} = 1.483x + 2.55$

example: suppose that $(\bar{x}, \bar{y}) = (14,10)$

$r = .6$

$s_x = 2$ and

$s_y = 3$

what is the equation of the LSRL?

Start on exercise 41 on page 157.

slope = $b_1 = r\frac{s_y}{s_x}$ $= (.6)(3)/(2) = .9.$

slope $= .6\left(\frac{3}{2}\right) = .9$

The LSRL passes through $(\bar{x}, \bar{y})$, so

$10 = a + .9(14)$

$10 = a + 12.6$

$-2.6 = a$

$y = mx + b$

$10 = .9(14) + b$

$10 = 12.6 + b$

$-2.6 = b$

$\hat{y} = .9x - 2.6$

LSRL equation: $\hat{y} = -2.6 + .9x$

or $\hat{y} = .9x - 2.6$

Tough Love vs. Spanking

Some Americans think it very improper to spank children, so some try other methods to control a child during one of "those moments". Some find it very effective to just take the child for a car ride and talk. The child usually calms down and stops misbehaving after a little car ride together.

A photo is attached showing such a session, in case you would like to use this highly effective technique.

Jul 22 - 2:43 PM

LSRL

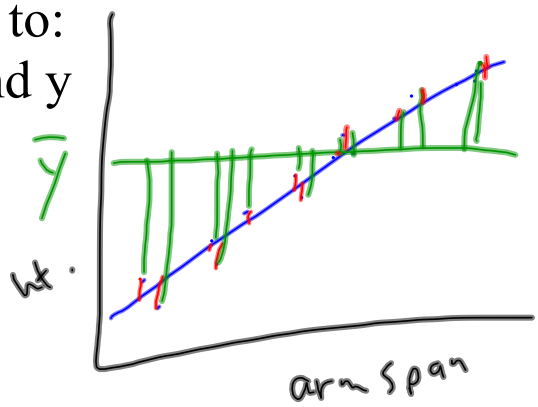
- model
- near data points

$y - \hat{y}$

residuals = observed − predicted

total variation around the line due to:

- linear relationship between x and y
- measurement error
- natural variablility
- lurking variables
- etc.



Coefficient of Determination

- r^2
- $0 \leq r^2 \leq 1$
- =1 when all points on line
- tells strength of linear association
- "% of variation in y explained by linear relationship between x & y"
- "% of variation in y attributed to the linear regression model for x & y"

Slope in Statistics

- Average change in y for every unit of increase in x.

Sep 24-7:15 AM

coasters through july 2010.ftm

Discuss r^2 and slope.

represents the percent of the data that is the closest to the line of best fit.

r squared example:
x = height & y = weight.
If $r = 0.922$, $r^2 = 0.850$.

85% of the total variation in *weight,* can be explained by the linear relationship between *height,* and *weight*. The other 15% of the total variation in *weight,* is due to some other variable(s).

Slope example:
x = height & y = weight, and slope = 6 lbs/inch
For every additional inch of height, we expect an average of 6 additional pounds of weight.

If every point is on LSRL, it explains all the variation. The further the line is from the points, the less it explains.

Another example:
x = child's verbal IQ & y = child's reading test score
If r = 0.8, r^2 = 0.64.
64% of total variation in reading score can be explained by the linear relationship between verbal IQ & reading score. The other 36% of the total variation in the reading score remains unexplained.

Slope example:
x = child's verbal IQ & y = child's reading test score and slope is 1.2, then

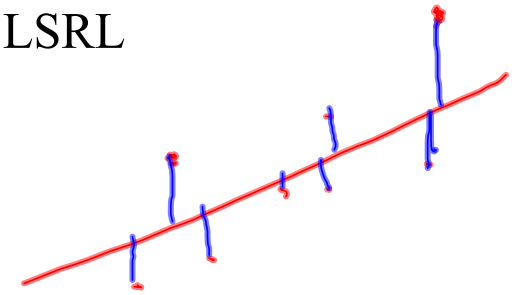
- For every 1 point increase in verbal IQ, we would expect the reading score test to increase by 1.2 points, on average.

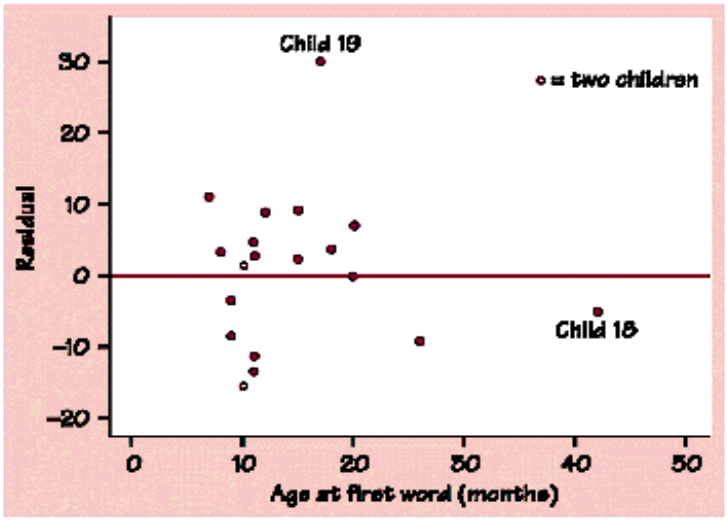
Residual:

- difference between observed y and predicted y.
- observed y - predicted y = $y - \hat{y}$
- mean = 0.

Residual plot:

- scatterplot of x and residuals
- use to assess the fit of LSRL





Now try exercises 43 and 45 on page 165.

Make a Residual Plot:

1 data
STAT
1: Edit

L1	L2	L3 3
4	2	
10	6.4	
12	7.2	
14	9.2	
20	12	
26	16.1	
30	18.5	

L3(1)=

2 LSRL
STAT
CALC
4: LinReg (ax+b)

EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7↓QuartReg

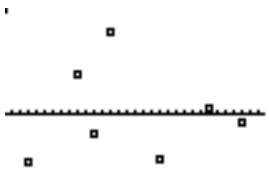
LinReg(ax+b) L1, L2

3 plot
2nd Y=

Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Ylist:RESID
Mark:

for RESID hit
2nd STAT

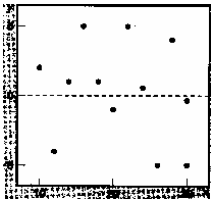
4 graph
ZOOM 9



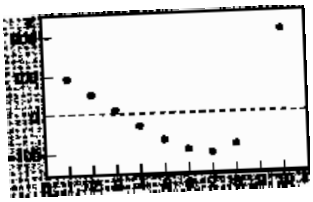
analyze the fit of your model

Interpreting residual plots

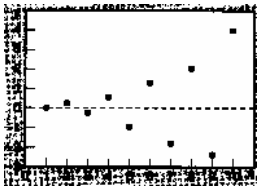
uniform, even, random scatter
LSRL fits data well
good model



curved pattern
relationship is not linear
poor model



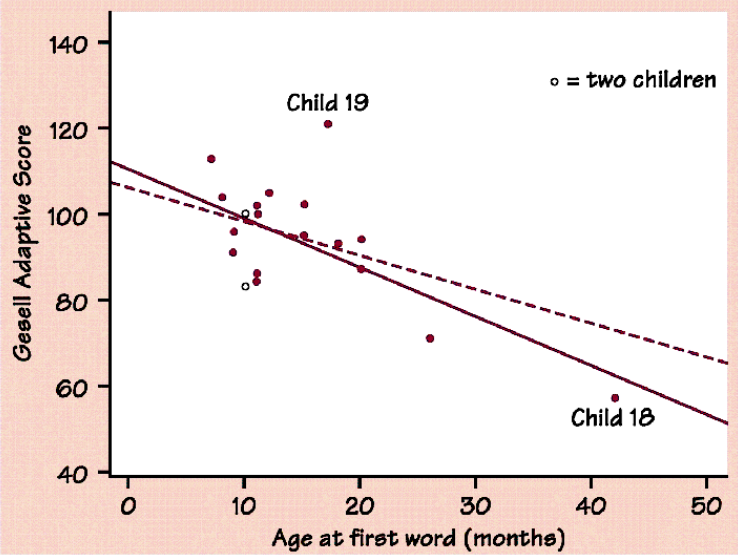
changing spread
prediction of y less accurate where spread out



Points with large residuals are outliers (graph on p. 172)

Points far to right or left can affect LSRL a lot

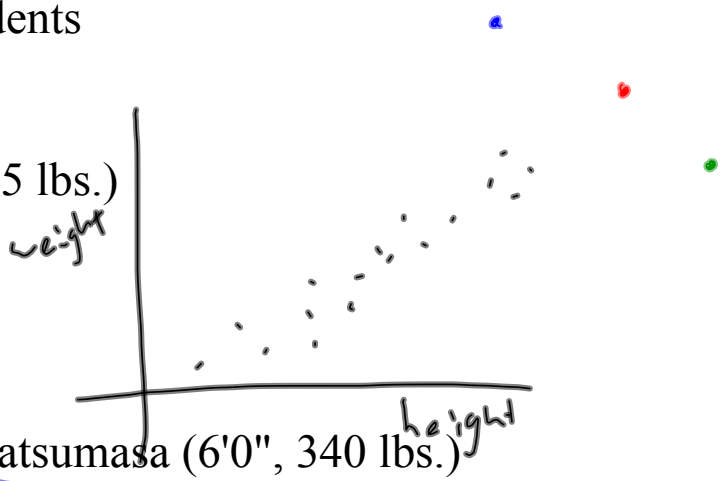
Points with strong effects called influential points



_ with child 18
-- without child 18

- child 18 is an influential point
- child 19 is an outlier

Imagine a scatterplot
heights & weights of H. S. students
Form, Direction, Strength?



Add Shaquille O'Neal (7'1", 325 lbs.)
Outlier in height or weight?
Still fit the overall pattern?
Influential?

Add sumo wrestler Kasugao Katsumasa (6'0", 340 lbs.)
Outlier in height or weight?
Still fit the overall pattern?
Influential?

Add Manute Bol, (7'6" 200 lbs)
Outlier in height or weight?
Still fit the overall pattern?
Influential?

Now try exercises 47 and 49 on page 173.

Sep 12-7:18 PM

Attachments

 coasters through july 2010.ftm