Announcements:

- You can turn in homework until 6pm, slot on wall across from 2202 Bren. Make sure you use the correct slot! (Stats 8, closest to wall)
- We will cover Chs. 5 and 6 first, then 3 and 4.
- Mon, Oct 4 discussion is practice with R Commander. Discussion at 5pm, 6pm, go to **192 ICS** (Computer Lab). Discussion at 4pm (only), you have two options. Go to 192 ICS to work on lab computer, *or* watch presentation in 174 ICS and bring laptop if desired.

Homework (due **Fri**, Oct 8):

Ch. 5: # 5a, 17, 18, 76

For #76 use R Commander

Data on CD and website (dataset called **oldfaithful**)

TODAY: Chapter 5, Sections 5.1 and 5.2

Relationship between

Two Quantitative Variables

Algebra Review (Linear relationship)

Equation for a straight line:

$$y = b_0 + b_1x$$

b_0 = y-intercept, the value of y when $x = 0$

b_1 = slope, the increase in y when x goes up by 1 unit

Example: One pint of water weighs 1.04 pounds. ("A pint's a pound the world around.")

Suppose a bucket weighs 3 pounds. Fill it with x pints of water. Let y = weight of the filled bucket.

Example, continued:

b_0 = y-intercept, the value of y when x = 0

This is the weight of the empty bucket, so $b_0 = 3$

b_1 = slope, the increase in y when x goes up by 1 unit; this is the added weight for adding 1 pint of water, i.e. 1.04 pounds.

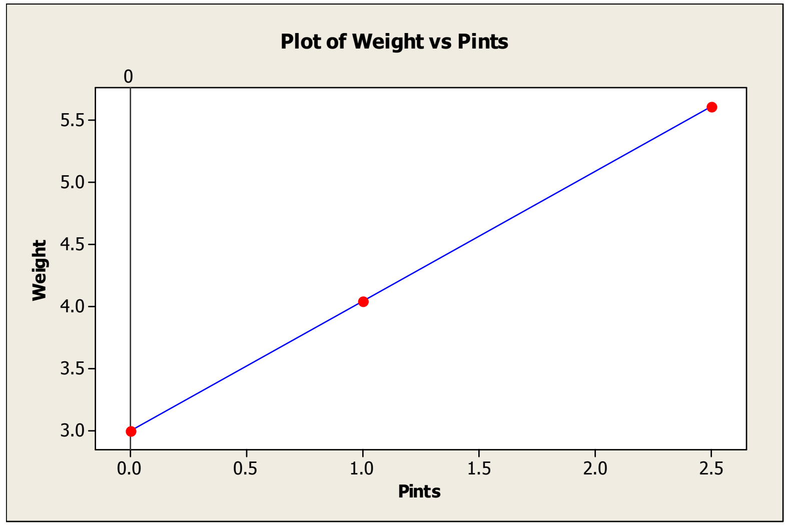
The equation for the line:

$$y = b_0 + b_1 x$$

$$y = 3 + 1.04\,x$$

x = 1 pint → $y = 3 + 1.04(1) = 4.04$ pounds

x = 2.5 pints → $y = 3 + 1.04(2.5) = 5.6$ pounds



You have just seen an example of a *deterministic relationship* – if you know x, you can calculate y.

Definition: In a statistical relationship there is *variation* in the possible values of y at each value of x.

If you know x, you can only find an *average* or *approximate* value for y.

We are interested in describing linear relationships between two quantitative variables. Usually we can identify one as the *explanatory variable* and one as the *response variable*. We always define:

x = explanatory variable

y = response variable

Examples:

		Example 5.12	Example 5.6
Explanatory Variable:	x = Average of parents' heights	x = Verbal SAT Score	x = Age
Response Variable:	y = Male's height	y = College GPA	y = Highway sign reading distance

Features we will look at for two quantitative variables:

1. Graph – "Scatter plot" – to *visually see* relationship
2. Regression equation – to ***describe*** *the "best" straight line* through the data, and ***predict*** y, given x in the future.
3. Correlation coefficient – to *describe the* ***strength and direction*** of the linear relationship

Example 1: Can height of male student be predicted by knowing the average of his parents' heights?

Example 2: Can college GPA be predicted from Verbal SAT?

Example 3: Can the distance at which a driver can see a road sign be predicted from the driver's age?

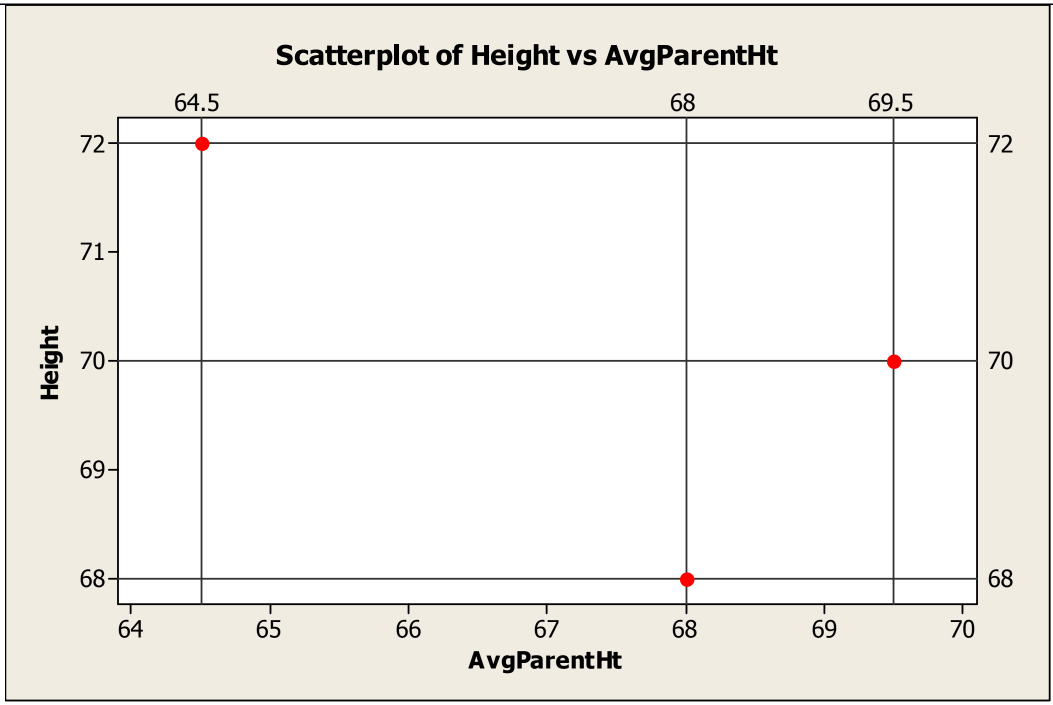
Creating a scatter plot:

- Create axes with the appropriate ranges for x (horizontal axis) and y (vertical axis)
- Put in one “dot” for each (x,y) pair in the data set.

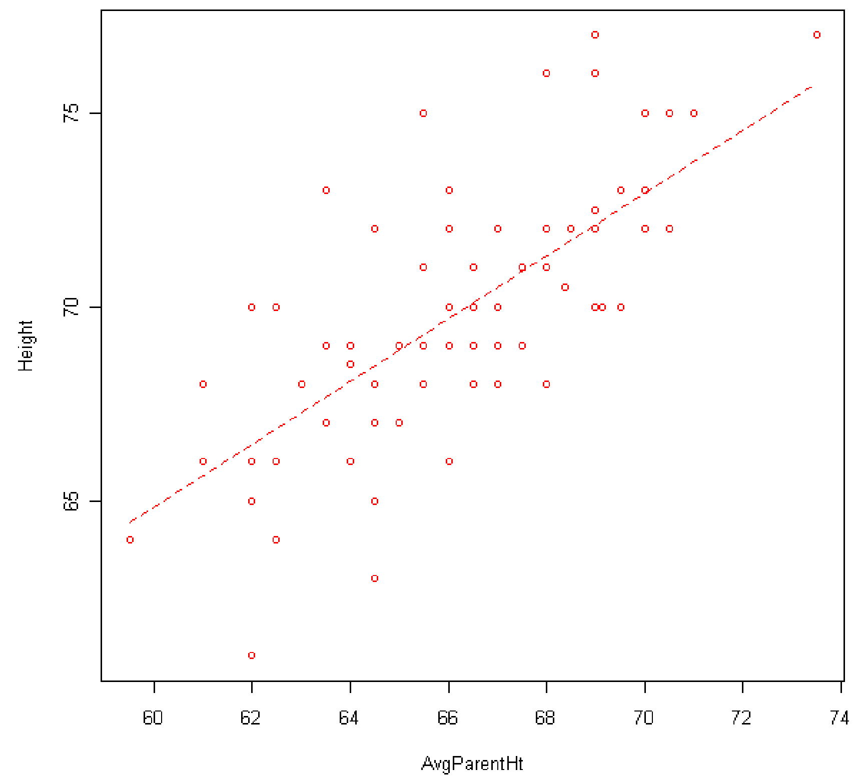
Example 1: Scatterplot of 3 points, x = avg parent ht, y = height

First 3 points in the data:

x	y
64.5	72
68	68
69.5	70



Scatterplot of all 73 individuals, with a line through them



What to notice in a scatterplot:

1. If the *average* pattern is linear, curved, random, etc.
2. If the trend is a *positive association* or a *negative association*
3. How *spread out* the **y-values** are **at each value of x** (*strength of relationship)*
4. Are there any *outliers* – unusual *combination* of (x,y)?

1. Average pattern looks *linear*
2. It's a *positive association* (as x goes up, y goes up, on average)
3. Student heights are quite spread out at each average parents' height
4. There are no obvious outliers in the combination of (x,y)

REGRESSION LINE (REGRESSION EQUATION)

Basic idea: Find the "best" line to

1. *Estimate* the *average value of y* at a given value of x
2. *Predict* y in the future, when x is *known* but y is not

Definition: A regression line or least squares line is a straight line that best* describes how values of a quantitative response variable (y) are related to a quantitative explanatory variable (x).

*"Best" will be defined later.

Notation for the regression line is:

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

"y-hat = b-zero + b-one times x"

Example 1: $\hat{y} = 16.3 + 0.809x$

For instance, if parents' average height = 68 inches,

$$\hat{y} = 16.3 + 0.809x$$

16.3 + 0.809(68) = 71.3 inches

Interpretation – the value 71.3 can be interpreted in two ways:

1. An *estimate* of the *average* height of all males whose parents' average height is 68 inches
2. A *prediction* for the height of a *single* male whose parents' average height is 68 inches

NOTE: It makes sense that we predict a male to be *taller* than the average of his parents. Presumably, a female would be predicted to be *shorter* than the average of her parents.

Example 1, continued

Interpreting the y-intercept and the slope:

Intercept = 16.3 is the estimated male height when parents' average height is 0. This makes no sense in this example!

Slope = +0.809 is the difference in estimated height for two males whose parents' average heights differ by 1 inch.

For instance, if parents' average height is 65 inches,

$$\hat{y} = 16.3 + 0.809(65) = 68.9 \text{ inches}$$

One inch higher parents' average height is 66 inches, and

$$\hat{y} = 16.3 + 0.809(66) = 69.7 \text{ inches}$$

(difference of .809 rounded to .8)

Prediction Errors and Residuals

Individual y values can be written as:

y = predicted value + prediction error

or

y = predicted value + residual

or

$$y = \hat{y} + residual$$

For each individual, *residual* = $y - \hat{y}$

Example: x = 66 inches, y = 69 inches.

Then $\hat{y}$ = 69.7 inches, so residual = 69 – 69.7 = –0.7 inches

The person is just 0.7 inches *shorter* than predicted.

DEFINING THE "BEST" LINE

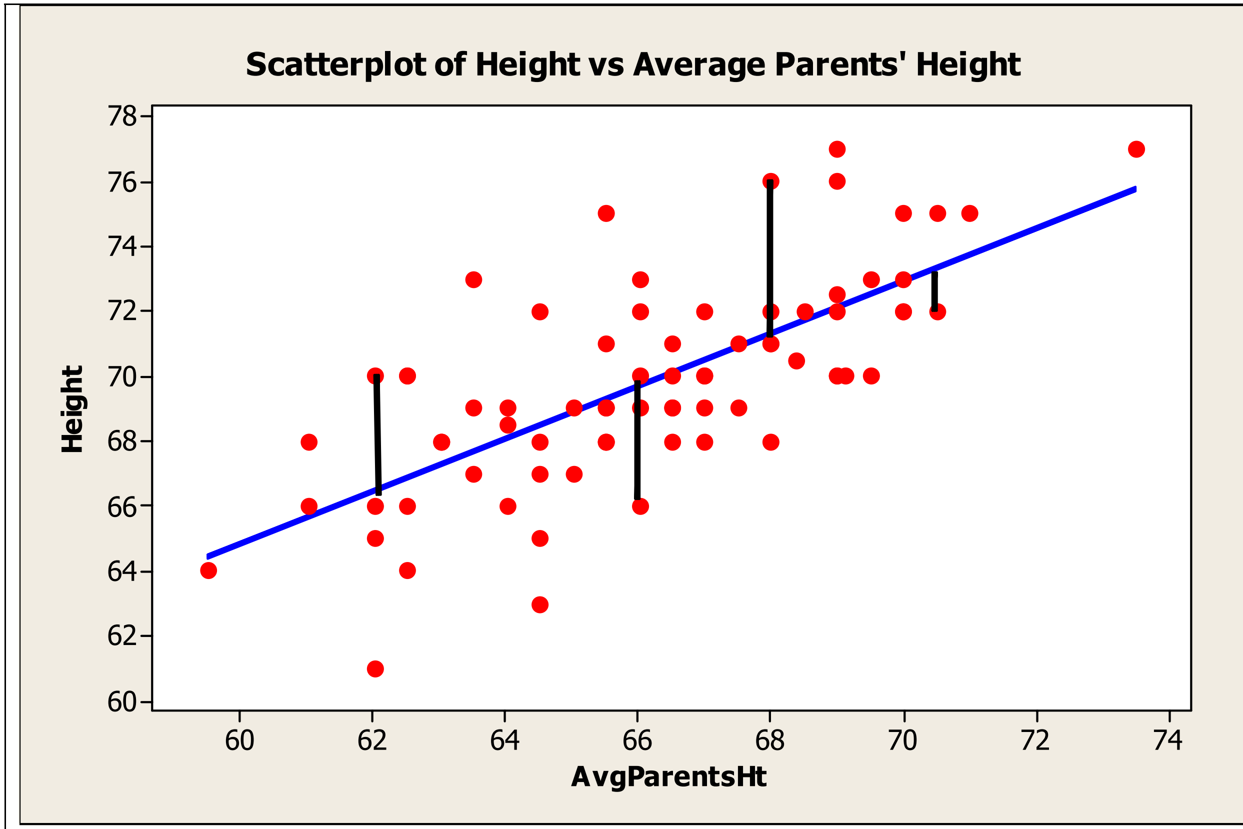
Basic idea: Minimize how far off we are when we use the line to predict y, based on x, by comparing to actual y.

For each individual in the data
"error" = "residual" = $y - \hat{y}$ = observed y – predicted y

Definition: The *least squares regression line* is the line that minimizes the sum of the squared residuals for all points in the dataset. The *sum of squared errors* = SSE is that minimum sum.

See picture on next page.

ILLUSTRATING THE LEAST SQUARES LINE



SSE = 376.9 (average of about 5.16 per person)

Example 1:

This picture shows the residuals for 4 of the individuals. The blue line comes closer to the points than any other line, where "close" is defined by SSE =

$$\sum_{all\ values} residual^2$$

R Commander does the work for you!

Statistics -> Fit models -> Linear regression

Then highlight the variables you want (response = y and explanatory = x) in the popup box. The results look like this:

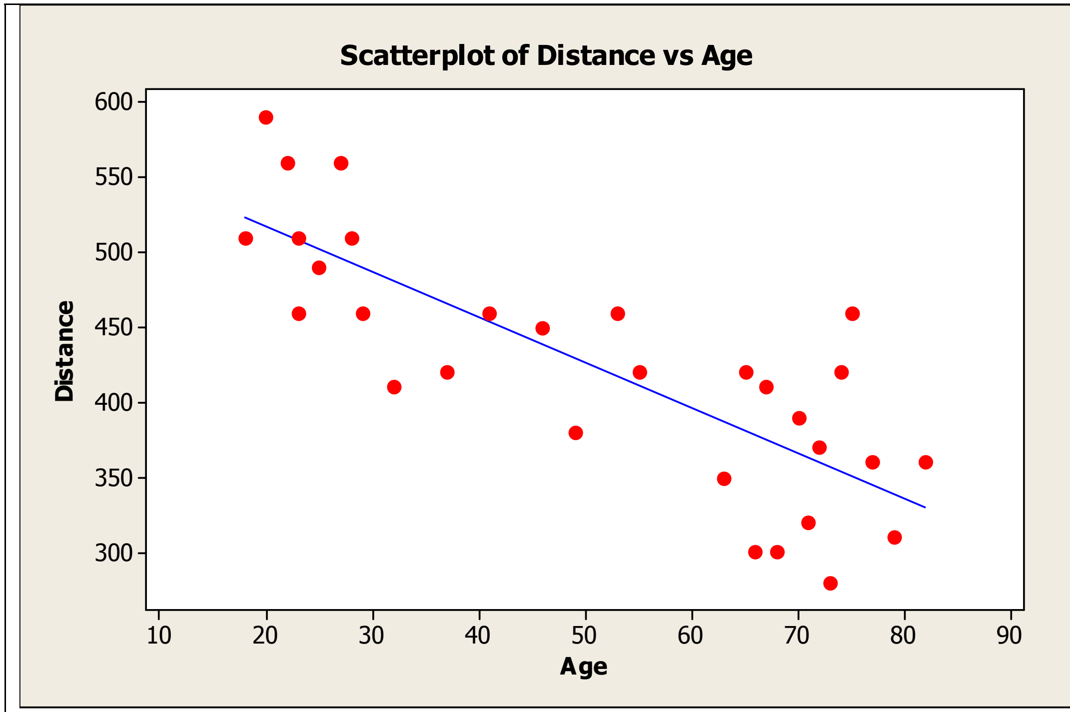
```
Call:
lm(formula = Height ~ AvgHt, data = UCDavisMLecture4)

Residuals:
    Min       1Q   Median       3Q      Max
-5.4768 -1.3305 -0.2858  1.2427  5.7142

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.3001     6.3188   2.580   0.0120 *
AvgHt         0.8089     0.0954   8.479 2.16e-12 ***
```

EXAMPLE OF A NEGATIVE ASSOCIATION

- A study was done to see if the distance at which drivers could read a highway sign at night changes with age.
- Data consist of n = 30 (x,y) pairs where x = Age and y = distance at which the sign could first be read (in feet).



The regression equation is

$$\hat{y} = 577 - 3x$$

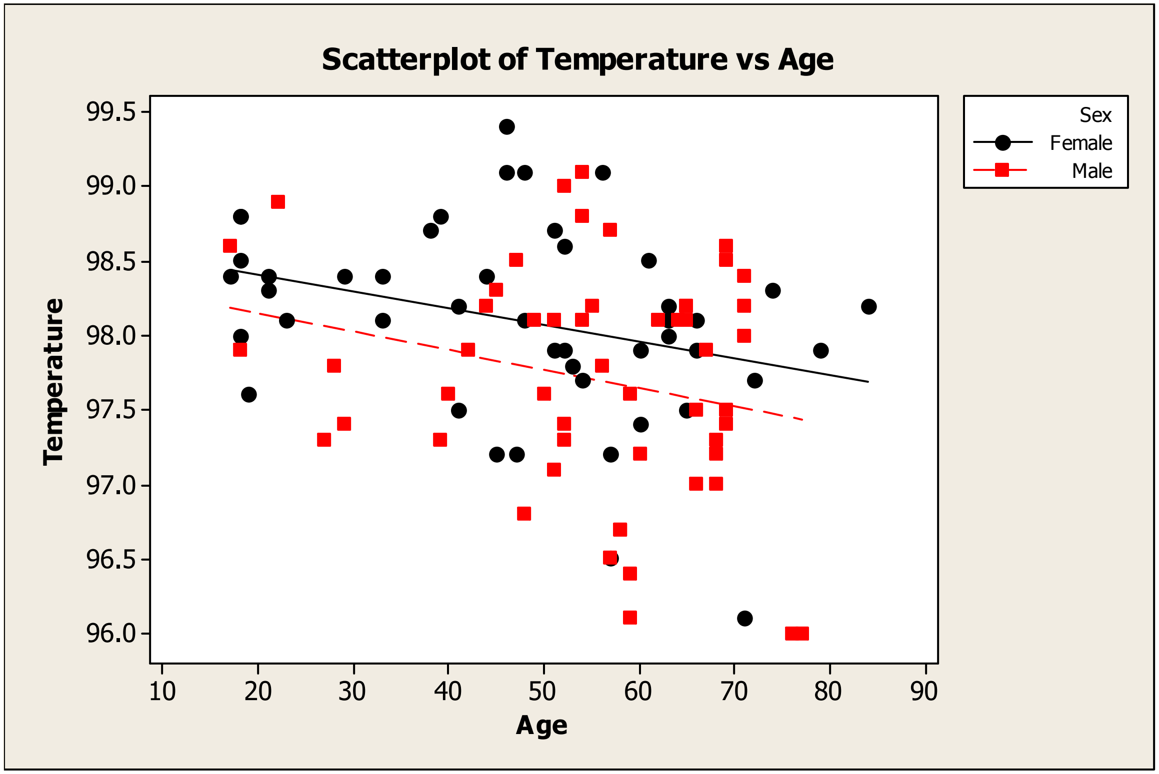
Notice *negative* slope

Ex: 577 – 3(20) = 577 – 60 = 517

Age	Pred. distance
20 years	517 feet
50 years	427 feet
80 years	337 feet

Separating Groups in Regression and Correlation

Example: Body temperature for 100 adults aged 17 to 84



Note females slightly higher at all ages. Regression equations:

Males: $\hat{y} = 98.4 - .0126(age)$

Females: $\hat{y} = 98.6 - .0112(age)$

Not easy to find the best line by eye!

Applets:

http://onlinestatbook.com/stat_sim/reg_by_eye/index.html

http://www.rossmanchance.com/applets/Reg/index.html

SUMMARY OF WHAT YOU SHOULD KNOW

1. How to read a scatterplot to look for
   a. Linear trend or not (curved, etc.)
   b. positive or negative association (or neither)
   c. strength of relationship (how close points are to line)
   d. outliers

2. Given a regression equation,
   a. Use it to *predict* y and *estimate* y for *given* x (useful when using the equation in the future, x known, y not)
   b. Interpret slope and intercept
   c. Find residual for a given individual, when given x and y for that individual.

Homework (due **Fri**, Oct 8):

Ch. 5: # 5a, 17, 18, 76

For #76 use R Commander

Data on CD and website (dataset called **oldfaithful**)