Stat 13, Thur 5/24/12.
1. Scatterplot.
2. Correlation, r.
3. Residuals
4. Def. of least squares regression line.
5. Example.
6. Extrapolation.
7. Interpreting b.

Final exam is Thur, 6/7, in class.
Hw7 is due Tue, 6/5, and is from the handout, which is from “An Introduction to the Practice of Statistics” 3rd ed. by Moore and McCabe.
Problems 2.20, 2.30, 2.38, 2.44, 2.46, and 2.102.

TWO VARIABLES NOW!

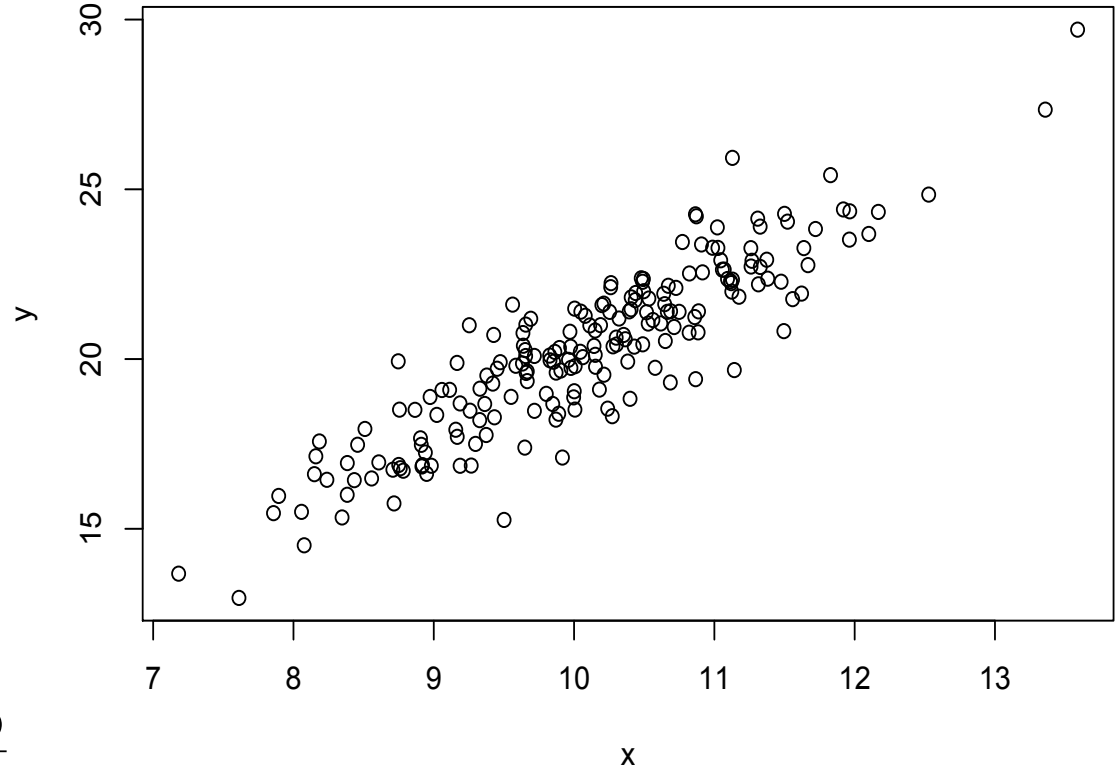
1. Scatterplot.

A plot of X versus Y. Each point represents one observational unit (1 person, typically).

Convention: explanatory variable on X axis, response on Y axis.

e.g. X = cups of coffee per week,

Y = body mass index (BMI).



2. Correlation, r.

Measures strength of linear relationship: how closely the points follow a line.

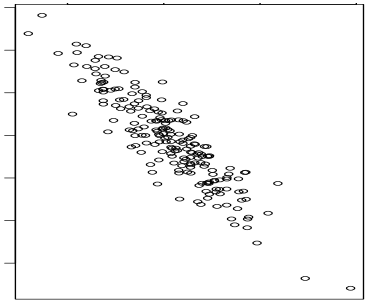
Two sample means: $\bar{x}$ and $\bar{y}$.

Two sample sds: s_x and s_y.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y}$$

The numbers ($\bar{x}$, $\bar{y}$, s_x , s_y , r) are sometimes called the 5 number summary.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y}$$

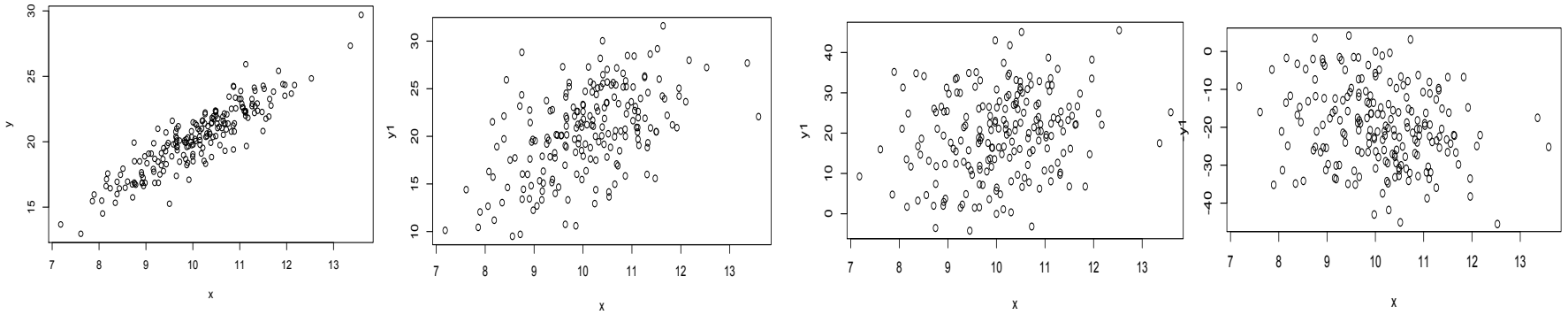


r = -0.9.

Facts about r.

- $-1 \le r \le 1$.
- positive r means positive association, negative means negative.
- r near 0 means weak linear relationship. r near 1 means the points fall near a line sloping up. r near -1 means the points fall near a line sloping down.
- r doesn't depend on the units of x or y.
- In computing r, it doesn't matter which is x and which is y. However, for regression, it does matter which is x and which is y.
- r only measures strength of the linear relationship between x and y. If there is curvature, the relationship between x and y can be clear and strong and yet r could be 0.
- r is sensitive to outliers. It is not resistant.

r = 0.9. r = 0.5. r = 0.2. r = -0.2.



3. Residuals.

Imagine fitting a line to the data. One wants the best-fitting line, esp. for prediction of Y based on X.

$$\hat{y}_i = a + bX_i.$$

a = intercept, b = slope.

Observed minus predicted values.

$e_i = y_i - \hat{y}_i$

Want residuals to be as small as possible. How can we quantify this?

For the best fitting line, the residuals will always average zero.

We can find the line so that the average *size* if the residuals is as small as possible, or similarly, the line where the average *square* of the residuals is as small as possible.

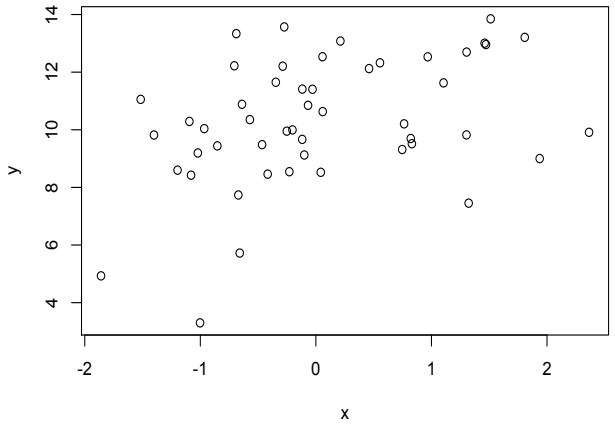
4. Definition of the least squares regression line.

THE REGRESSION LINE IS THE LINE THAT MINIMIZES THE MEAN SQUARE OF THE RESIDUALS.

How can we find this line?

We want to find a and b.

One way would be to try out all different a's and b's, and for each pair (a,b), find the residuals e_i, calculate the mean of e_i^2, and choose the (a,b) that minimize this.



3. Residuals.

Imagine fitting a line to the data. One wants the best-fitting line, esp. for prediction of Y based on X.

$$\hat{y}_i = a + bX_i.$$

a = intercept, b = slope.

Observed minus predicted values.

$e_i = y_i - \hat{y}_i$

Want residuals to be as small as possible. How can we quantify this?

For the best fitting line, the residuals will always average zero.

We can find the line so that the average *size* if the residuals is as small as possible, or similarly, the line where the average *square* of the residuals is as small as possible.

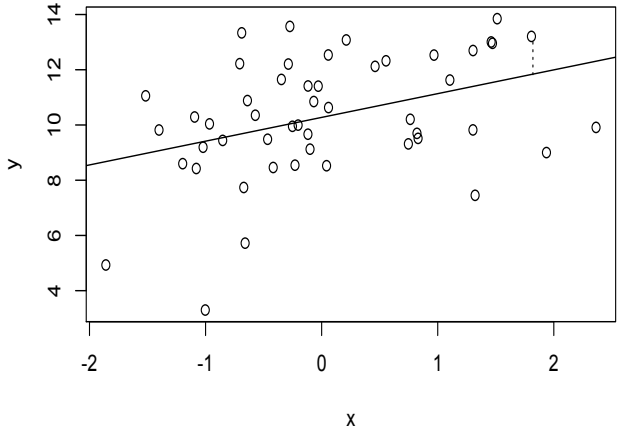
4. Definition of the least squares regression line.

THE REGRESSION LINE IS THE LINE THAT MINIMIZES THE MEAN SQUARE OF THE RESIDUALS.

How can we find this line?

We want to find a and b.

One way would be to try out all different a's and b's, and for each pair (a,b), find the residuals e_i, calculate the mean of e_i^2, and choose the (a,b) that minimize this.



It turns out that there is a much easier way to find a and b!

$b = rs_y/s_x$.

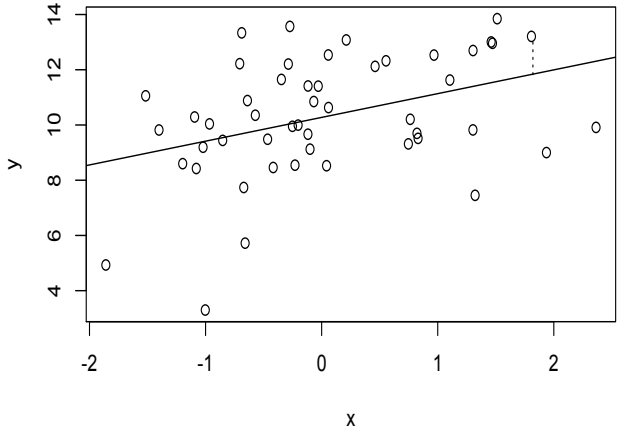
$a = \bar{y} - b\bar{x}$.

$\hat{y}_i = a + bX_i$.

5. Example. X = coffee cups/day, Y = BMI.
Suppose $\bar{x} = 2$, $\bar{y} = 27$, $s_x = 1.2$, $s_y = 5$, and $r = 0.3$. Find the regression line.
Use it to predict the BMI of a person drinking 3 cups of coffee per day.

$b = rs_y/s_x = 0.3(5)/1.2 = 1.25$.
$a = \bar{y} - b\bar{x} = 27 - (1.25)(2) = 24.5$.

So, the regression line is $\hat{y}_i = 24.5 + 1.25X_i$.
If $X_i = 3$, then $\hat{y}_i = a + bX_i = 24.5 + 1.25(3) = 28.25$.



6. Extrapolation.
Beware of making predictions far outside the observed range of the data, because even if the data fall near a line in the observed range, they might not fall near the line outside of this range.
Rats and saccharin: extrapolation from very high doses in rats to very low doses in humans.

7. Confounding factors and concluding causation from b.
Usually the slope b is the thing to interpret. b indicates the amount your prediction of Y increases, on average, per one unit increase in X. However, the relationship is not necessarily causal. There could be confounding factors.
For example, increasing your coffee consumption might not cause an increase in your longevity.