Stat 145, Wed 15-Sep-2021 -- Wed 15-Sep-2021

Wednesday, September 15th 2021

Due::   PS03 due at 11 pm

Wednesday, September 15th 2021

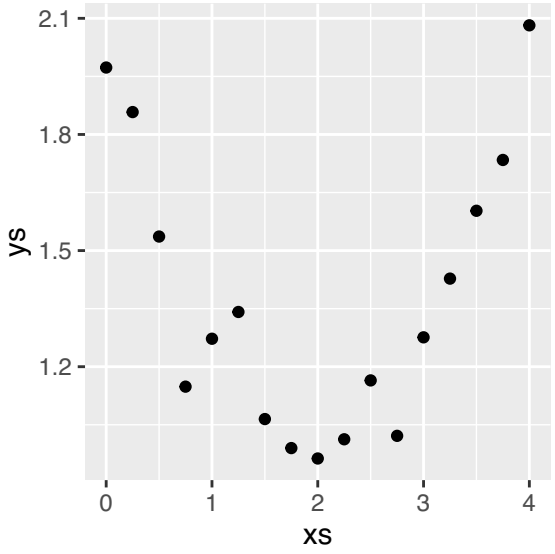
Wk 3, We
Topic:: Correlation
Read:: Lock5 2.5
Topic:: least-squares regression
Read:: Lock5 2.6

Warmup:
 - where to place mean and median on various density curves
 - need na.rm=TRUE switch with mean()

The Correlation Coefficient

- It is used for (near) linear relationships between *quantitative* variables. The data involved must be true *bivariate data*—i.e., two quantities measured on the same subjects/units.
  - These are the same kind of scenarios (variable-wise) as those in which a scatterplot is possible.
  - You could not talk about the correlation coefficient between these two variables: *model of car* and *price of car*.
- It measures direction and strength of a *linear* relationship.
  - distinction between variables *having an association* and variables being *correlated*. The authors use the phrase "two variables are correlated" as synonomous with say "the two variables have an association", which seems to add only to the confusion.
  - Be careful! Data that has a strong association, can have a correlation coefficient near zero. Look at your data to see if a correlation coefficient makes sense.

```
xs = seq(0,4,.25)
ys = (xs-2)^2 / 4 + 1 + rnorm(length(xs), 0, 0.1)
gf_point(ys ~ xs)
cor(ys ~ xs)
```

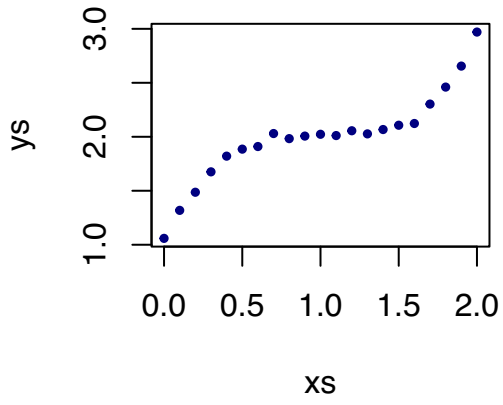



```
cor(ys ~ xs)
```

```
[1] 0.03695452
```

  - Similarly, data can produce a correlation coefficient close to (± 1), even though the relationship is not linear:

```
xs = seq(-1,1,.1) + 1
ys = (xs-1)^3 + rnorm(length(xs), 0, 0.05) + 2
plot(xs, ys, col="navy", pch=19, cex=.5)
```




```
cor(ys ~ xs)

[1] 0.9079114
```

- As with other quantities (the *mean*, for instance), there is a **population correlation** coefficient (denoted by ρ) and a **sample correlation** (denoted by r)
- Always a number between (-1) and 1.

  At the lower extreme (-1), a scatterplot of the two variables will exactly lie on a straight line with negative slope.

  At the upper extreme (1), a scatterplot of the two variables will exactly lie on a straight line with positive slope.

  Correlation coefficients near zero indicate a weak or nonexistent linear association.
- The sample correlation coefficient is calculated using some of the same kinds of squared deviations from the mean as "sum of squares" calculations for ANOVA, or standard deviations/variances:

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}} = \frac{1}{n-1}\sum_i \frac{(x_i - \overline{x})}{s_x}\frac{(y_i - \overline{y})}{s_y}.$$

  That makes it a fairly complicated number to calculate by hand. Once again, we will get the number using software. In R, you type `cor(y x)`, when `x` and `y` are vectors (with the same number of entries) whose correlation you seek.
- It is a dimensionless quantity—i.e., it has no units. It will not change if, say, your x-values are converted from inches to feet, or the like.
- It is fairly sensitive to outliers. See applet at

  `http://www.stat.sc.edu/~west/javahtml/Regression.html`

Q: What is wrong with this statement? "There is a strong correlation between length of stay in a job and whether you are married or not."

Play the correlation game.

Approximate stopping place

```
Q: True or False.  In the presence of two quantitative variables,
   is a 0 correlation the mark of no association?
   Follow up: What is?

More scatter plotting
 - spruce data:  Di.change ~ Ht.change
    add color for Fertilizer
    lm(Di.change ~ Ht.change, data = spruce)
 - draft data:  N69 ~ nday

Review features of a line  y = intercept + slope * x
 - intercept
 - slope
    meaning

least-squares regression:  \hat y = a + bx
 - identify slope as b, intercept as a
 - offers a "prediction" to value of y for given x
 - observed y vs. fitted/predicted \hat y-value
    residual = observed - predicted
      straight-line distance
      positive if data point is above line, negative if below
 - how data is used to choose a, b
    want to make overall measure of residuals as small as possible
    might add up residuals and try to make sum small
      sum r_i  does not prove to be effective
      two alternatives:
        sum |r_i|
        sum r_i^2    better setup for calculus to take over and produce
    b = r s_y / s_x
    a = ybar - b xbar
 - use app, have groups make guesses
```