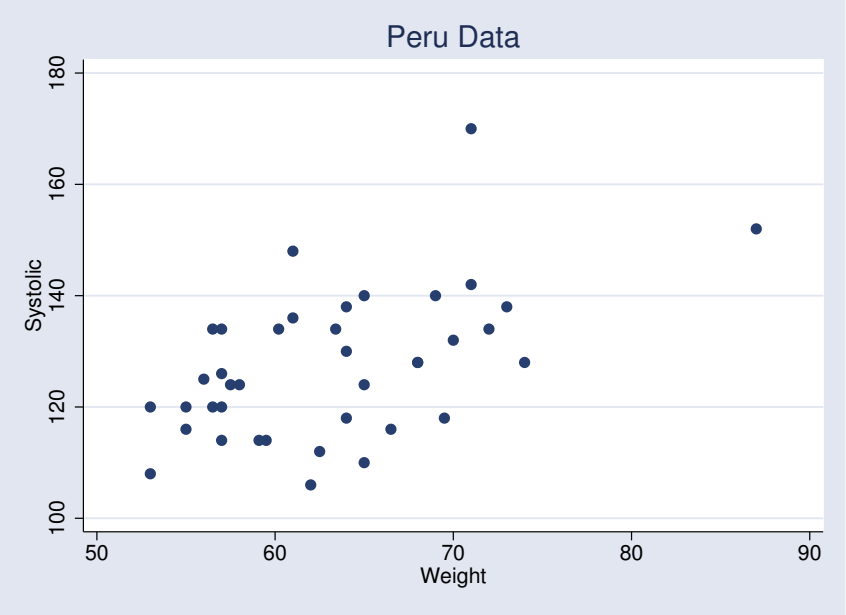
Section 2.2 - Correlation

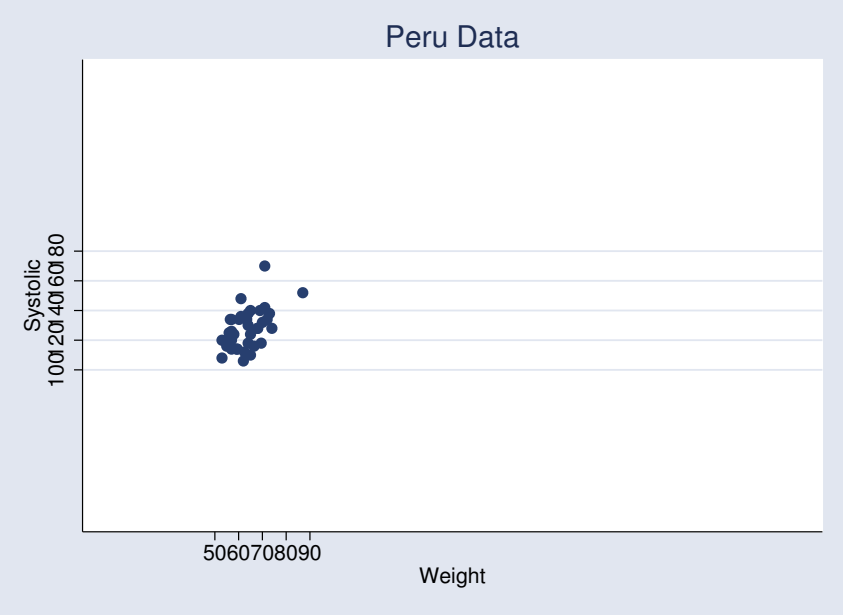
Statistics 104

Autumn 2004

Correlation

Would be nice to have a numerical measure of how strong a relationship between two variables is





The scale a relationship is plotted at can affect the level of perceived strength.

Correlation Coefficient r

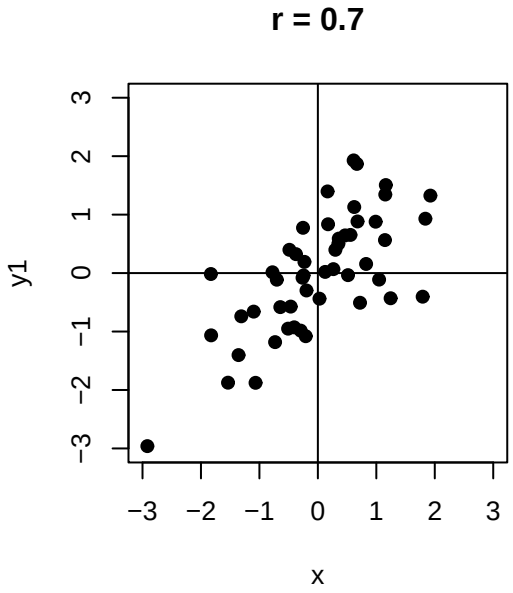
Measure the direction and strength of the **linear** relationship between two quantitative variables.

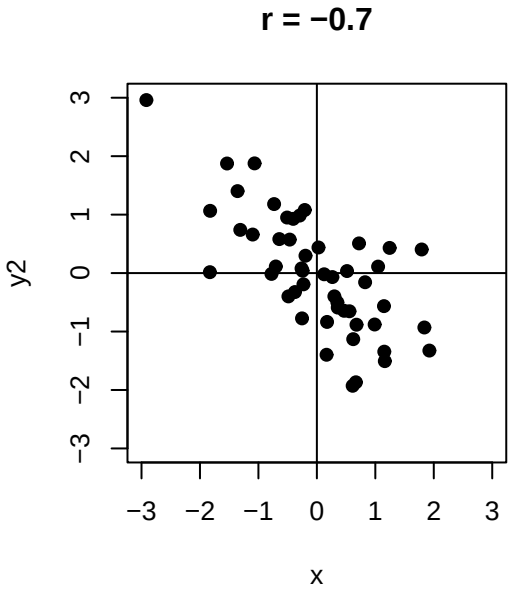
$$\begin{aligned} r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \\ &= \frac{1}{n-1}\frac{\sum x_i y_i - n\bar{x}\bar{y}}{s_x s_y} \qquad \{\text{Computational formula}\} \end{aligned}$$

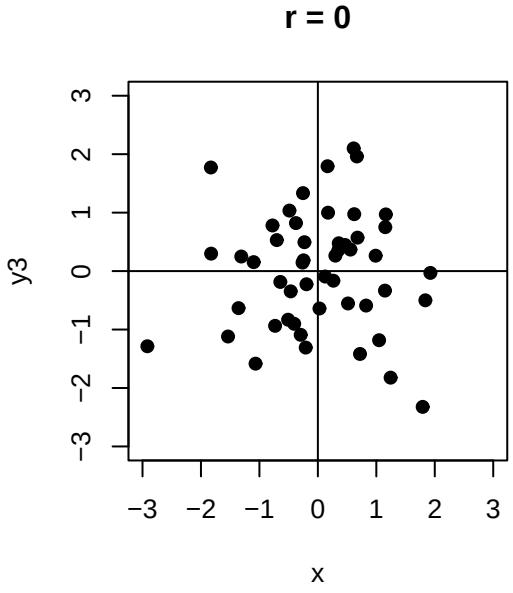
Note that the variables

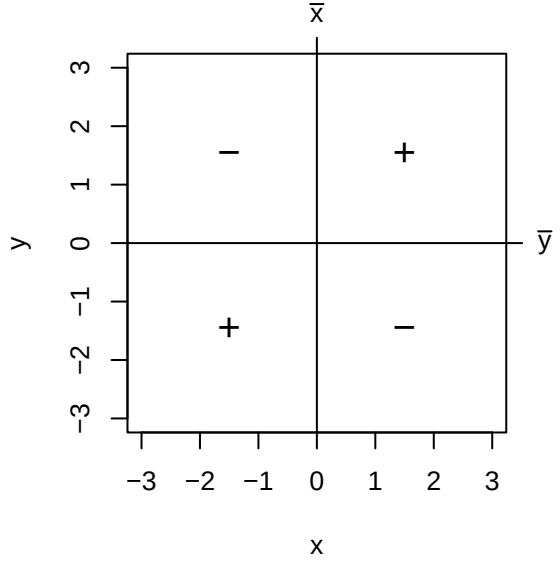
$$\frac{x_i-\bar{x}}{s_x} \quad \text{and} \quad \frac{y_i-\bar{y}}{s_y}$$

both have mean $= 0$ and standard deviation $= 1$.


r = 0.7
y1
x
−3 −2 −1 0 1 2 3


r = −0.7
y2
x
−3 −2 −1 0 1 2 3


r = 0
y3
x
−3 −2 −1 0 1 2 3


x̄
ȳ
y
x
−
+
+
−
−3 −2 −1 0 1 2 3

Properties of r

1. $-1 \leq r \leq 1$

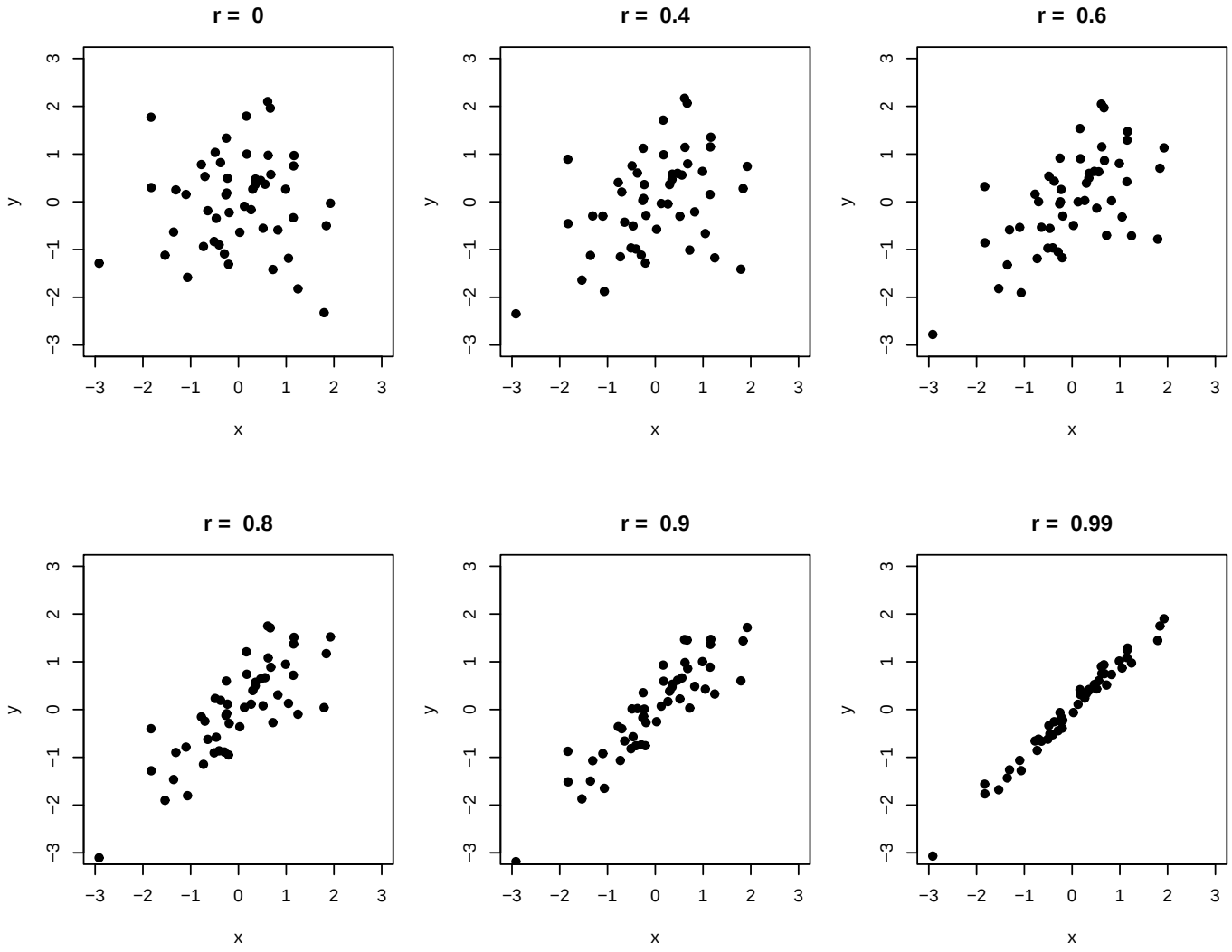
2. $r = 1$ or $r = -1$ only if all points lie exactly on a straight line

3. $|r|$ is unchanged under linear scale changes

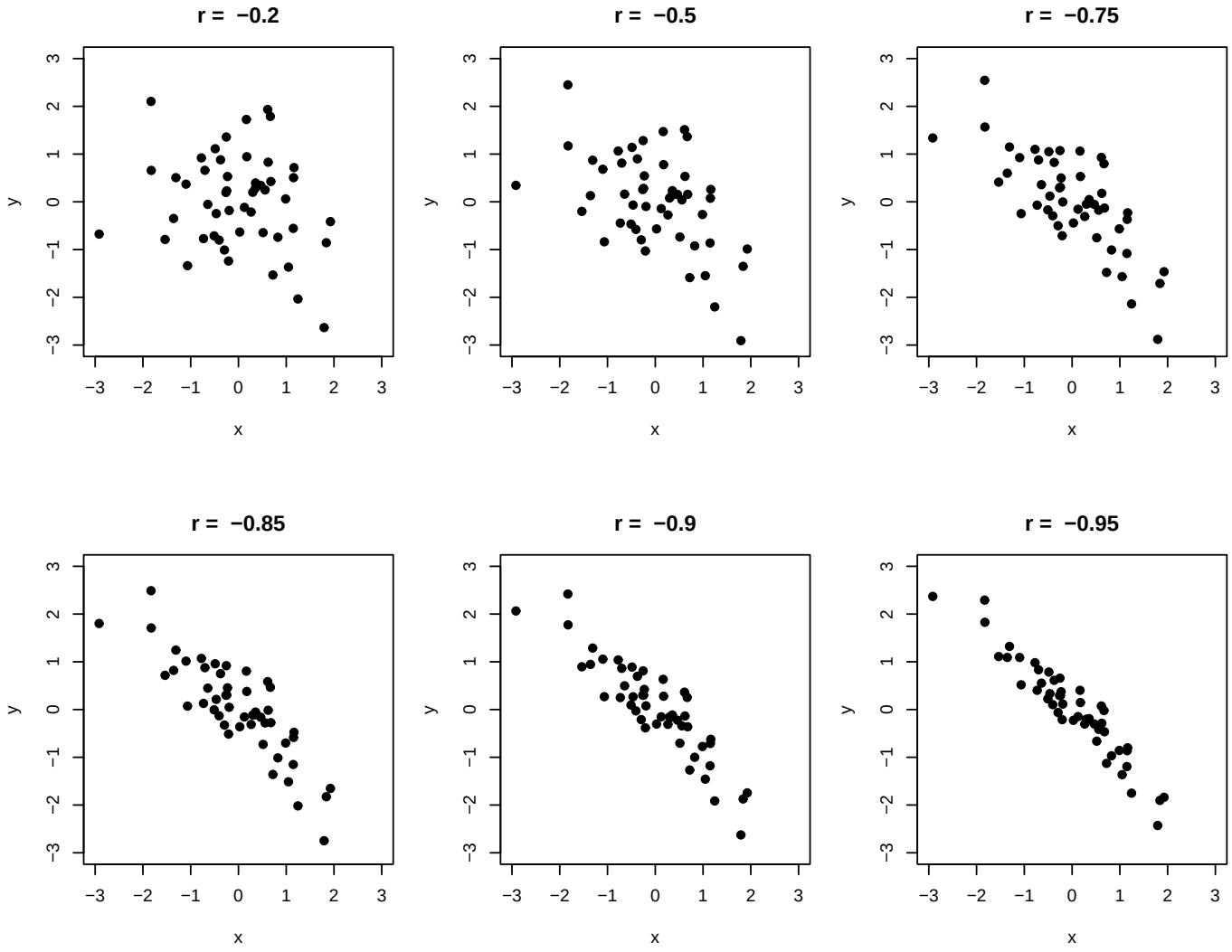
   e.g. r for MPG and Weight(lbs) is the same as for r for km/l and Weight (kg).

   You could have a sign change. Let $x^* = -x$. Then $r_{x^*y} = -r_{xy}$

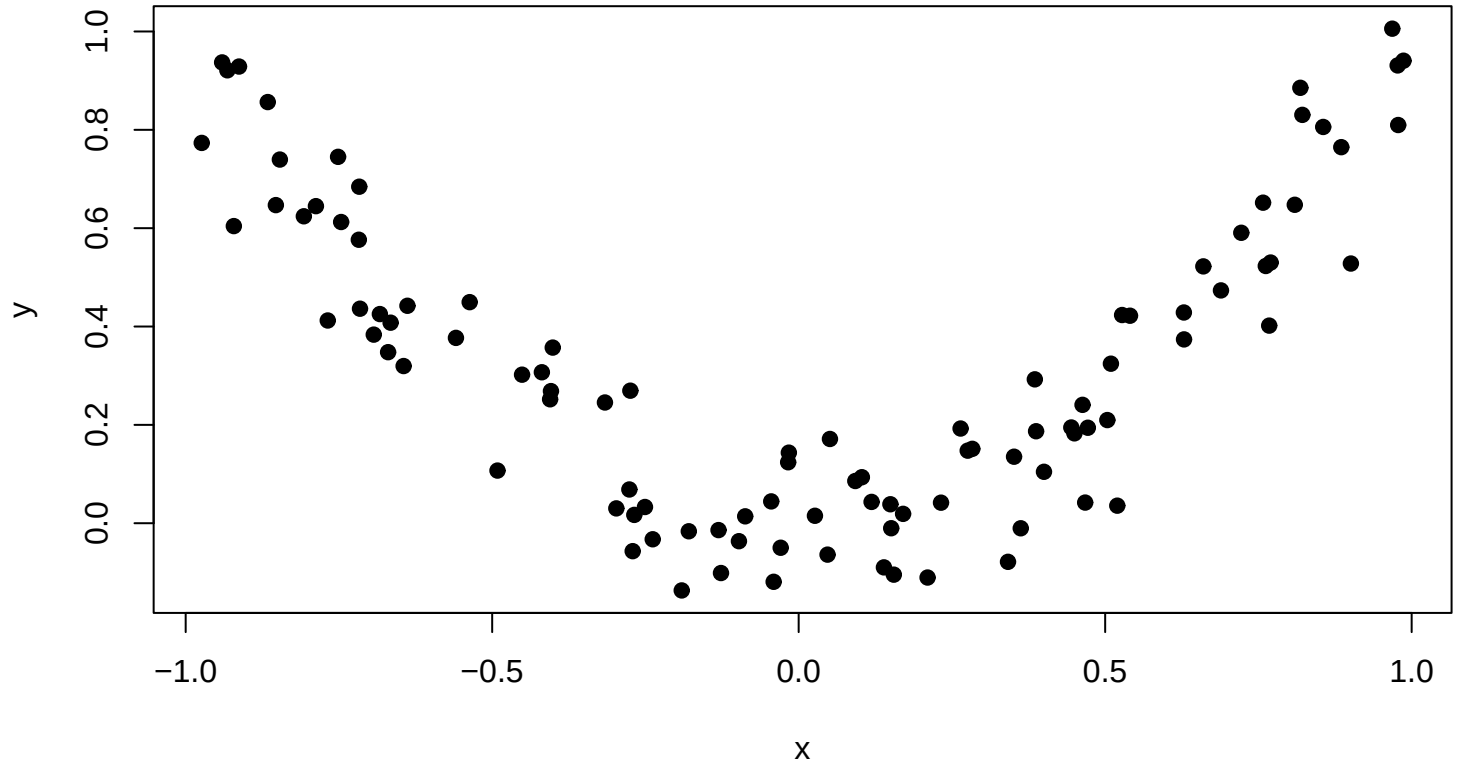
4. $r > 0$: positive association



$r < 0$: negative association



5. r only measures the strength of a linear association



Pearson correlation of x and y is $r = -0.02$.

6. Correlation makes no distinction between explanatory and response variables

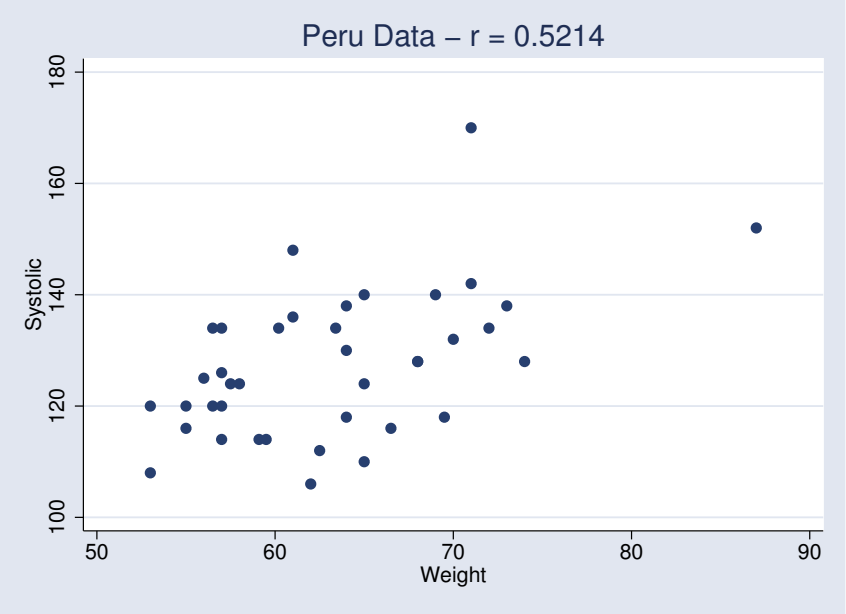
$$r_{xy} = r_{yx}$$

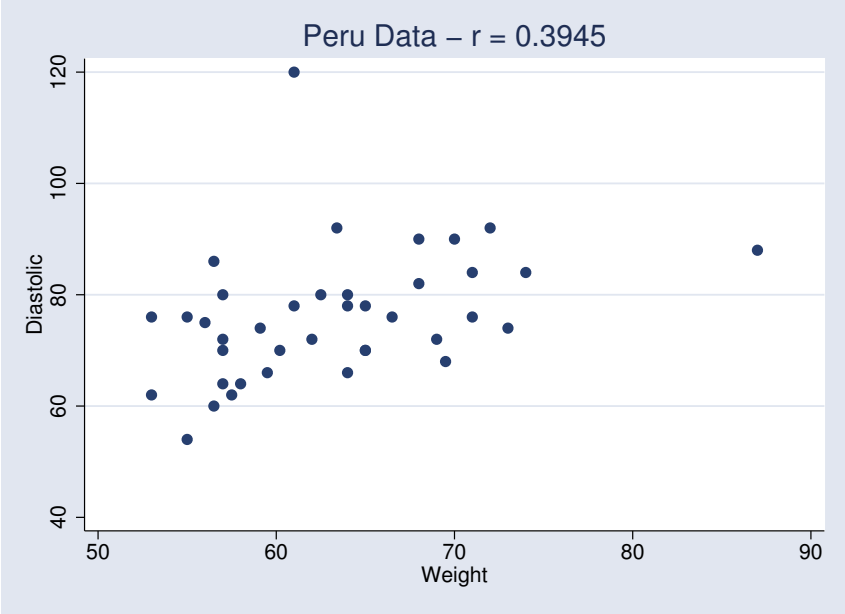
7. Correlation only makes sense for two quantitative variables.

Examples:

Peru Study:

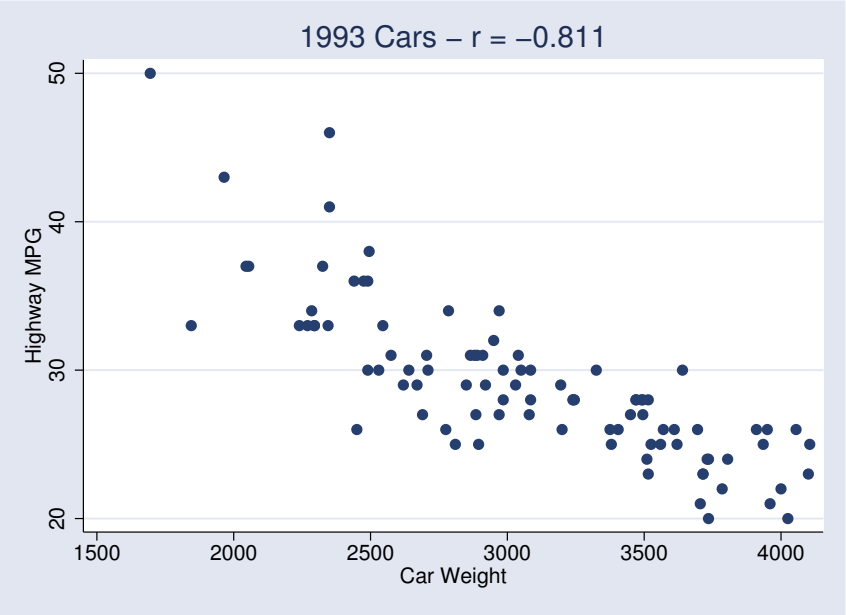
They studied the blood pressure of 39 Indians who migrated from a very primitive environment, high in the Andes mountains, into the mainstream of Peruvian society, at a much low altitude.

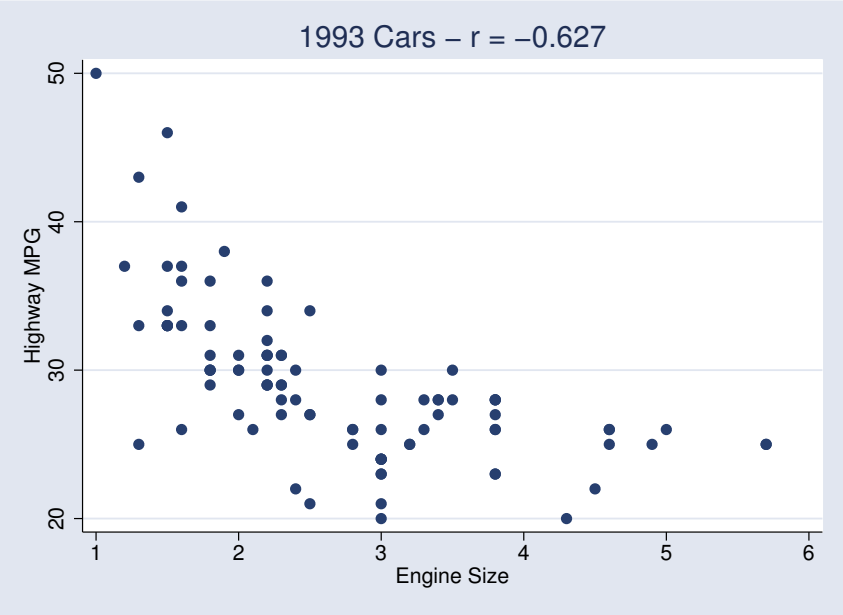


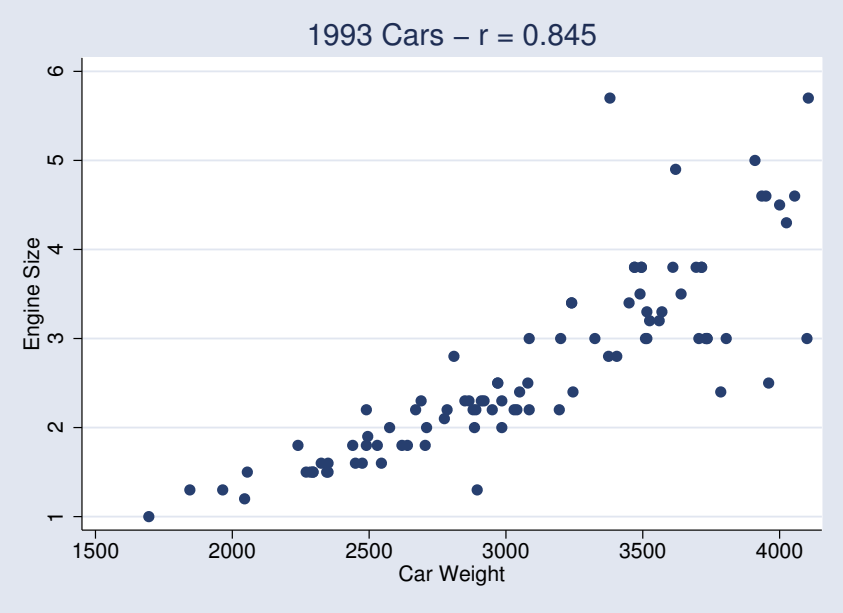


```
. correlate Weight  Height Systolic Diastolic
(obs=39)

             |   Weight   Height Systolic Diasto~c
-------------+------------------------------------
      Weight |   1.0000
      Height |   0.4503   1.0000
    Systolic |   0.5214   0.2191   1.0000
   Diastolic |   0.3945   0.2530   0.4752   1.0000
```


1993 Cars – r = –0.811
Highway MPG
20
30
40
50
1500
2000
2500
3000
3500
4000
Car Weight


1993 Cars – r = –0.627
Highway MPG
20
30
40
50
1
2
3
4
5
6
Engine Size


1993 Cars – r = 0.845
Engine Size
1
2
3
4
5
6
1500
2000
2500
3000
3500
4000
Car Weight

```
. correlate HighMPG CityMPG EngSize Weight (obs=93)

             |  HighMPG  CityMPG  EngSize   Weight
-------------+------------------------------------
     HighMPG |   1.0000
     CityMPG |   0.9439   1.0000
     EngSize |  -0.6268  -0.7100   1.0000
      Weight |  -0.8107  -0.8431   0.8451   1.0000
```