May 24, 2005  Lecture Notes  (VIII. Correlation, regression)

Case Study 17

heights of fathers (f) & sons (s)

- We've talked about how to describe + do inference on 1 variable at a time; what about 2 at a time?

<table>
<thead>
<tr>
<th>father</th>
<th>son</th>
</tr>
</thead>
<tbody>
<tr>
<td>70 in</td>
<td>72 in</td>
</tr>
<tr>
<td>67 in</td>
<td>66 in</td>
</tr>
</tbody>
</table>

- 1 row for each family; 2 columns (for fathers & sons)

- height of f+s: both quantitative continuous variables

- mean: mean_y = 69 in mean_x = 66 in

- SD: s_y = 2.7 in s_x = 2.7 in

- Secular trend in height: sons taller than fathers by about 1 in. because of better nutrition

- both normal curve b/c CLT

- best fit line is best way to show the trend

- bivariate normal b/c both histograms follow normal curve

- keep plotting the points until you get same plot as on p. 262

- this graph is called a scatter plot here, called elliptical scatter plot (football shape)

- what you get when both variables follow the normal curve

- x: independent variable (predictor)

- y: dependent variable (outcome)

- Q: How strongly related (in a linear way) are x = y?

- r > 0

- increase x = increase y

- point in middle = point of averages

- slope positive: x+y positively associated

- r < 0

- increase x = decrease y

- weight of car

- slope negative: x+y negatively assoc.
flat trend line - no (linear) association

$r = 0$

$r = \text{correlation coefficient}$

$\text{divided into the 4 quadrants of scatter plot}$

$\sqrt{\text{most of data in plot}}$

\[ r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{S_x S_y} \]

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

$\text{What's the correlation between heights of fathers and sons?}$

$r = +$

\[ r \text{ always comes out between } [-1, +1] \]

$\text{see p.104 for chart of different } r \text{ values on scatter plots}$

$\text{facts about } r: 1)$

perfect negative linear assoc. $\rightarrow$ neg slope (not necessarily $-1$)

perfect positive linear association $\rightarrow$ pos. slope (not necessarily $+1$)
In this problem, \( r = +0.5 \) (moderate positive correlation) 
moderate strong determination

2) \( r \) can be fooled by 2 things:
   a) nonlinearity
   b) outliers

If correlation is near 1 or -1, the scatterplot can have only one 
possible appearance, but if \( r = 0 \), it could look like any of these 
3 plots:

- There is relationship, but nonlinear
  \( r = 0 \)

- Correlation \( r = +1 \), but overall, outlier creates 
  problem, esp. when \( n \) is small

- \( r = 0 \) (flat trend line)

3) Units of \( r \)?
   \[ r = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]
   ex: \( x = \text{income} (\text{dollars}) \) \( y = \text{height (in)} \) cancel out in formula
   \( r \) is a pure number without units

b) Switching \( x + y \) leaves \( r \) unchanged
   \( y = x \) plot folds over 45° line
c) Adding a positive or negative constant to $x$ also doesn't change $r$.

\[ y \]
\[ x \]
\[ 0 \]
\[ x \]

\[ y \]
\[ 0 \]
\[ y = 5 \]
\[ 0 \]
\[ x \]

\[ y \]
\[ 0 \]
\[ \text{multiply } x \text{ by 10} \]
\[ 0 \]
\[ x \]

(End of Case Study 17)

Case Study 18

- Major axis of ellipse
- SD line: slope $\frac{S_y}{S_x}$
- Not best line for predicting $y$ from $x$