Correlation analysis

Tuan V. Nguyen
Garvan Institute of Medical Research
Sydney, Australia
Correlation analysis

- Some examples
- Theory of correlation
- R functions for correlation analysis
Relationship between whole body bone density and age
Percent body fat and body mass index
When correlation analysis is required?

- **t-test**
- **or**
- **ANOVA**

When the *independent variable* is *a categorical variable*
When correlation analysis is required?

• *Relationship* between 2 continuous variables
• Degree of *co-variation*
• Prediction
Theory and ... history
Sir Francis Galton (16/2/1822 – 17/1/1911)

- Half-cousin of Charles Darwin
- Geographer, meteorologist, tropical explorer, inventor of fingerprint identification, eugenicist
- Love to measure everything from the weather to female beauty
Correlation and the genetics of intelligence

Galton’s conclusions:

- **Nature dominates**: “families of reputation were much more likely than ordinary families to produce offspring of ability”
- **Recommended** “judicious marriages during several generations” to “produce a highly gifted race of men”
- His “genetic utopia”: “Bright, healthy individuals were treated and paid well, and encouraged to have plenty of children. Social undesirables were treated with reasonable kindness so long as they worked hard and stayed celibate.”

Didn’t have data on “intelligence” so instead studied HEIGHT

Research interest:

“Those qualifications of intellect and disposition which … lead to reputation”

Although a self-proclaimed genius, who wrote that he could read @2½, write/do arithmetic @4, and was comfortable with Latin texts @8, he couldn’t figure out how to model these data(!)
Karl Pearson (1857 – 1936)

• Realized Galton's idea of correlation

• Invented the "Pearson's correlation coefficient", Chi square test, and many others

• Author of "The Grammar of Science"

• With Galton, he founded the prestigious journal *Biometrika*
How to describe a relationship / association?

- X and Y are random variables from 2 observations
- To measure the variation of X or Y: **variance**

\[
\text{var}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
\text{var}(y) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2
\]

- We need a measure of covariation between X and Y
- Covariance = average of the product of X.Y

\[
\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})
\]
Variance and covariance: Geometry

The geometry of independence

\[ h^2 = x^2 + y^2 \]

\[ h^2 = x^2 + y^2 - 2xycos(H) \]

Covariance
Meaning of variance and covariance

• Variance is always +ve
• Covariance can be –ve or +ve
  – If covariance = 0, *X and Y are independent*
  – If covariance > 0, X and Y vary in the same direction
  – If covariance < 0, X and Y vary in opposite directions
• **Covariance = a measure of association/correlation**
Degrees of correlation

\[ r = 0.99 \]

\[ r = 0.90 \]

\[ r = 0.50 \]

\[ r = 0.25 \]

\[ r = 0.10 \]

\[ r = 0.01 \]
Purposes of correlation analysis

• Estimate the coefficient of correlation \( (r) \)

• Test the null hypothesis of \( r = 0 ? \)
Estimation of correlation coefficient

- The unit of measurement of covariance: X * Y
- **Coefficient of correlation** \((r)\) between \(X\) and \(Y\) is the **standardized covariance** (unitless)
- \(r\) is defined as:

\[
r = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \times \text{var}(y)}} = \frac{\text{cov}(x, y)}{SD_x \times SD_y}
\]
Diagram showing the range of correlation:

- +1: Perfect Positive Correlation
- 0: No Correlation
- -1: Perfect Negative Correlation
Test the hypothesis of $r = 0$

- Null hypothesis: $H_0: r = 0$
- Alternative hypothesis: $H_A: r \neq 0$
- Fisher’s z-transformation: $r \rightarrow z$

\[
z = \frac{1}{2} \ln \left( \frac{1 + r}{1 - r} \right)
\]

- Standard error of z

\[
SE(z) = \frac{1}{\sqrt{n - 3}}
\]

- T-test:

\[
t = \frac{z}{SE(z)}
\]
Assumptions

- X and Y follow the bivariate Normal distribution
- X and Y are linearly related
Using R

(cor, cor.test)
Determinant of bone density study

• Cross-sectional study

• Bone mineral density (BMD) measured at the femoral neck

• Variables of interest: age, weight, femoral neck BMD

• Research questions:
  – is there a correlation between age and femoral neck BMD?
  – is the correlation statistically significant?
Measurement of femoral neck BMD
Femoral neck bone density and age
"cor" function in R

- If there are no missing values
  \[ \text{cor}(x, y) \]

- If there are missing values
  \[
  \text{cor}(x, y, \\
  \hspace{1cm} \text{use}="\text{pairwise.complete.obs}"
  )
  \]
  \[
  \text{cor}(x, y, \hspace{1cm} \text{use}="\text{complete.obs}"
  )
  \]
Example

```r
> cor(age, wt, use="complete.obs")
[1] -0.236508
```
Guide to interpretation of r

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>Interpretation (tentative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05 – 0.19</td>
<td>Very weak</td>
</tr>
<tr>
<td>0.20 – 0.39</td>
<td>Weak</td>
</tr>
<tr>
<td>0.40 – 0.59</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.60 – 0.79</td>
<td>Strong</td>
</tr>
<tr>
<td>0.80+</td>
<td>Very strong</td>
</tr>
</tbody>
</table>
> cor.test(fnbmd, age)

Pearson's product-moment correlation

data: fnbmd and age
t = -14.4162, df = 556, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:
-0.5795310 -0.4584638

sample estimates:
cor
-0.5216183
• $r$ is the coefficient of correlation

• $R^2$ is the **coefficient of determination**

  the proportion of variation in $Y$ can be explained by $X$

• $r$(weight, BMD) = 0.33 means $R^2 = (0.33)^2 = 0.11$.

  11% variation in BMD is attributable to weight
Genetic application
How to estimate the index of heritability

- Heritability = proportion of variance of a trait due to genetic factors
- Study design
  - Twin study
  - Family study
The contributions of genetic and environmental factors to the association among bone mineral density (BMD), lean mass, and fat mass were assessed in the Sydney Twin Study of Osteoporosis (Australia), 1995–1996, in 57 monozygotic and 55 dizygotic female twin pairs of Caucasian background, aged 52.8 (standard deviation, 13) years. In multiple regression analysis, lean mass was a significant determinant of areal BMD; however, fat mass was a principal determinant of volumetric BMD. Univariate model-fitting analyses indicated that 80% and 65% of variance of lean mass and fat mass, respectively, were attributable to genetic factors. The estimated heritability of BMD for lumbar spine, femoral neck, and total body BMD was 78%, 76%, and 79%, respectively. Multivariate analyses suggested that, while the association between lean mass and fat mass was attributable mainly to environmental factors ($r_e = 0.53$, $p < 0.01$), the association among the three BMD sites was attributable to both genetic and environmental factors ($r_g = 0.64–0.75$, $p < 0.001$; $r_e = 0.57–0.70$, $p < 0.001$). Furthermore, genetic factors that affect lean mass or fat mass have minor effects on BMD. It is concluded that lean mass and fat mass, as well as bone density, are under strong genetic regulation. However, the associations between BMD and fat mass or between lean mass and fat mass appear to be mediated mainly via environmental influences. Am J Epidemiol 1998;147:3–16.
Theory of twin study

FIGURE 1. Classic twin model. Path diagrams illustrate the univariate twin model. Latent variables are in circles; observed variables are depicted by squares. A, additive genetic factors; D, dominant genetic factors; C, shared environmental factors; and E, nonshared environmental factors (including measurement error). The numbers associated with each factor denote twin 1 and twin 2. The correlation between A1 and A2 is 1 for monozygotic (MZ) pairs and 0.5 for dizygotic (DZ) pairs; between D1 and D2: 1 for MZ and 0.25 for DZ pairs. The correlation between shared environmental factors (C1 and C2) is assumed to be unity for both zygosities.
Intraclass correlation between twin 1 and twin 2 for (A) lean mass and (B) fat mass.
Multivariate correlation
cor.test (psych)

- Can estimate multiple coefficients of correlation by using package “psych”
- Want to estimate the correlation between age, weight, height, lumbar spine BMD, femoral neck BMD
library(psych)
vars = cbind(age, wt, ht, lsbmd, fnbmd)
cor.test(vars)
pairs.panels(vars)
> corr.test(vars)

Call: corr.test(x = vars)
Correlation matrix

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>wt</th>
<th>ht</th>
<th>lsbmd</th>
<th>fnbmd</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>1.00</td>
<td>-0.24</td>
<td>-0.21</td>
<td>-0.11</td>
<td>-0.34</td>
</tr>
<tr>
<td>wt</td>
<td>-0.24</td>
<td>1.00</td>
<td>0.58</td>
<td>0.51</td>
<td>0.58</td>
</tr>
<tr>
<td>ht</td>
<td>-0.21</td>
<td>0.58</td>
<td>1.00</td>
<td>0.45</td>
<td>0.49</td>
</tr>
<tr>
<td>lsbmd</td>
<td>-0.11</td>
<td>0.51</td>
<td>0.45</td>
<td>1.00</td>
<td>0.66</td>
</tr>
<tr>
<td>fnbmd</td>
<td>-0.34</td>
<td>0.58</td>
<td>0.49</td>
<td>0.66</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Spearman's correlation
Charles Spearman (1863 – 1945)

- Psychologist, interested in human intelligence
- Pioneer in factor analysis
- Invented the Spearman's rank correlation

Found that schoolchildren's grades across seemingly unrelated subjects were positively correlated, and proposed that these correlations reflected the influence of a dominant factor, which he termed $\gamma$ for "general" intelligence.
Spearman's rank correlation

- The $n$ raw scores $X_i$, $Y_i$ are converted to ranks $x_i$, $y_i$
- The differences $d_i = x_i - y_i$ of each observation on the two variables are calculated
- If there are no tied ranks, then $\rho$ is given by this formula:

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$
Spearman's rank correlation correlation rho

data:  age and fnbmd
S = 2124863812, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
  rho
-0.3248905
• Coefficient of correlation: a measure of association between 2 continuous variables

• R functions
  - `cor(x, y, use="complete.obs")`
  - `cor.test(x, y)`

Association between many variables: `psych`
  - `pairs.panels(vars)`
  - `corr.test(vars)`
Warning: Correlation is not causation

- A correlation between X and Y does not necessarily imply X causes Y
- The best we can say is "association"
Shocking: Chocolate consumption and Nobel prizes

Be careful!

- Anscombe's 4 datasets
- All have $r = 0.816$

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th></th>
<th>II</th>
<th></th>
<th>III</th>
<th></th>
<th>IV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
</tr>
<tr>
<td>---</td>
<td>------</td>
<td>---</td>
<td>------</td>
<td>---</td>
<td>------</td>
<td>---</td>
<td>------</td>
<td>---</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
<td>7.46</td>
<td>8.0</td>
<td>6.58</td>
<td></td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
<td>6.77</td>
<td>8.0</td>
<td>5.76</td>
<td></td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
<td>12.74</td>
<td>8.0</td>
<td>7.71</td>
<td></td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
<td>7.11</td>
<td>8.0</td>
<td>8.84</td>
<td></td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
<td>7.81</td>
<td>8.0</td>
<td>8.47</td>
<td></td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
<td>8.84</td>
<td>8.0</td>
<td>7.04</td>
<td></td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
<td>6.08</td>
<td>8.0</td>
<td>5.25</td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
<td>5.39</td>
<td>19.0</td>
<td>12.50</td>
<td></td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
<td>8.15</td>
<td>8.0</td>
<td>5.56</td>
<td></td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
<td>6.42</td>
<td>8.0</td>
<td>7.91</td>
<td></td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
<td>5.73</td>
<td>8.0</td>
<td>6.89</td>
<td></td>
</tr>
</tbody>
</table>
Be careful!
# read in data file

dat = read.csv("~/Google Drive/Garvan Lectures 2014/Data/does_vn07.csv", header=T)

# select only women data

women = subset(dat, gender == "Female")

# plot the relationship between fnbmd and age

plot(fnbmd ~ age, pch=16)

abline(lm(fnbmd ~ age), col="red", lwd=2)