Overview

1. Overview
   - Theory: the organization of Observed and Latent variables
   - Data and scaling
   - Correlation, Regression, Multiple Regression
   - A latent variable approach to measurement
   - Types of validity; What are we measuring
   - Structural Equation Models
   - Individual differences

2. Correlation
   - Preliminaries
   - Getting the data and describing it
   - Transforming the data
   - Displaying the data
   - Alternative views of correlation

3. Multivariate Regression
   - Using the raw data
   - Using transformed data
Psychometric Theory: A conceptual Syllabus

Error  X  Latent X  Latent Y  Y  Error

\[ \delta_1 \rightarrow X_1 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_1 \rightarrow \epsilon_1 \]
\[ \delta_2 \rightarrow X_2 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_2 \rightarrow \epsilon_2 \]
\[ \delta_3 \rightarrow X_3 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_3 \rightarrow \epsilon_3 \]
\[ \delta_4 \rightarrow X_4 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_4 \rightarrow \epsilon_4 \]
\[ \delta_5 \rightarrow X_5 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_5 \rightarrow \epsilon_5 \]
\[ \delta_6 \rightarrow X_6 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_6 \rightarrow \epsilon_6 \]
\[ \delta_7 \rightarrow X_7 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_7 \rightarrow \epsilon_7 \]
\[ \delta_8 \rightarrow X_8 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_8 \rightarrow \epsilon_8 \]

\[ \delta_9 \rightarrow X_9 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow \xi_2 \rightarrow \xi_3 \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_3 \rightarrow Y_4 \rightarrow Y_5 \rightarrow Y_6 \rightarrow Y_7 \rightarrow Y_8 \rightarrow \epsilon_1 \rightarrow \epsilon_2 \rightarrow \epsilon_3 \rightarrow \epsilon_4 \rightarrow \epsilon_5 \rightarrow \epsilon_6 \rightarrow \epsilon_7 \rightarrow \epsilon_8 \]
Overview

Theory: the organization of Observed and Latent variables

Observed Variables

<table>
<thead>
<tr>
<th>Error</th>
<th>X</th>
<th>Latent X</th>
<th>Latent Y</th>
<th>Y</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>δ1</td>
<td>X1</td>
<td>χ1</td>
<td></td>
<td>Y1</td>
<td>δ1</td>
</tr>
<tr>
<td>δ2</td>
<td>X2</td>
<td></td>
<td>χ1</td>
<td>Y2</td>
<td>δ2</td>
</tr>
<tr>
<td>δ3</td>
<td>X3</td>
<td>χ2</td>
<td></td>
<td>Y3</td>
<td>δ3</td>
</tr>
<tr>
<td>δ4</td>
<td>X4</td>
<td></td>
<td>χ2</td>
<td>Y4</td>
<td>δ4</td>
</tr>
<tr>
<td>δ5</td>
<td>X5</td>
<td>χ1</td>
<td></td>
<td>Y5</td>
<td>δ5</td>
</tr>
<tr>
<td>δ6</td>
<td>X6</td>
<td></td>
<td>χ2</td>
<td>Y6</td>
<td>δ6</td>
</tr>
<tr>
<td>δ7</td>
<td>X7</td>
<td>χ3</td>
<td></td>
<td>Y7</td>
<td>δ7</td>
</tr>
<tr>
<td>δ8</td>
<td>X8</td>
<td></td>
<td>χ3</td>
<td>Y8</td>
<td>δ8</td>
</tr>
<tr>
<td>δ9</td>
<td>X9</td>
<td></td>
<td>χ3</td>
<td></td>
<td>δ9</td>
</tr>
</tbody>
</table>

Error terms: δ1, δ2, δ3, δ4, δ5, δ6, δ7, δ8, δ9

Latent variables: χ1, χ2, χ3

Observed variables: X1, X2, X3, X4, X5, X6, X7, X8, X9

Latent variables: χ1, χ2, χ3

Observed variables: Y1, Y2, Y3, Y4, Y5, Y6, Y7, Y8

Error terms: ϵ1, ϵ2, ϵ3, ϵ4, ϵ5, ϵ6, ϵ7, ϵ8

Error terms: δ1, δ2, δ3, δ4, δ5, δ6, δ7, δ8, δ9
Theory: the organization of Observed and Latent variables

**Latent Variables**

- **Observed Variables (X):**
  - \( X_1 \)
  - \( X_2 \)
  - \( X_3 \)
  - \( X_4 \)
  - \( X_5 \)
  - \( X_6 \)
  - \( X_7 \)
  - \( X_8 \)
  - \( X_9 \)

- **Error (δ):**
  - \( δ_1 \)
  - \( δ_2 \)
  - \( δ_3 \)
  - \( δ_4 \)
  - \( δ_5 \)
  - \( δ_6 \)
  - \( δ_7 \)
  - \( δ_8 \)
  - \( δ_9 \)

- **Latent Variables (χ):**
  - \( χ_1 \)
  - \( χ_2 \)
  - \( χ_3 \)

- **Latent Variables (ξ):**
  - \( ξ_1 \)
  - \( ξ_2 \)

- **Observed Variables (Y):**
  - \( Y_1 \)
  - \( Y_2 \)
  - \( Y_3 \)
  - \( Y_4 \)
  - \( Y_5 \)
  - \( Y_6 \)
  - \( Y_7 \)
  - \( Y_8 \)

- **Error (ϵ):**
  - \( ϵ_1 \)
  - \( ϵ_2 \)
  - \( ϵ_3 \)
  - \( ϵ_4 \)
  - \( ϵ_5 \)
  - \( ϵ_6 \)
  - \( ϵ_7 \)
  - \( ϵ_8 \)
Theory: the organization of Observed and Latent variables

Theory

Error X Latent X Latent Y Y Error

δ1 X1 χ1 ξ1 Y1 ε1
δ2 X2 ξ1 Y2 ε2
δ3 X3 χ2 Y3 ε3
δ4 X4 ξ2 Y4 ε4
δ5 X5 χ3 Y5 ε5
δ6 X6 ξ2 Y6 ε6
δ7 X7 χ3 Y7 ε7
δ8 X8 ξ2 Y8 ε8
δ9 X9

Overview Correlation Multivariate Regression Significance
A theory of data and fundamentals of scaling

Error  X  Latent X  Latent Y  Y  Error

δ₁  δ₂  δ₃  δ₄  δ₅  δ₆  δ₇  δ₈  δ₉

χ₁  ξ₁  ξ₂  ξ₃

X₁  X₂  X₃  X₄  X₅  X₆  X₇  X₈  X₉

Y₁  Y₂  Y₃  Y₄  Y₅  Y₆  Y₇  Y₈

ϵ₁  ϵ₂  ϵ₃  ϵ₄  ϵ₅  ϵ₆  ϵ₇  ϵ₈
Correlation, Regression, Partial Correlation, Multiple Regression

Error \[ \delta_1 \] \[ \delta_2 \] \[ \delta_3 \] \[ \delta_4 \] \[ \delta_5 \] \[ \delta_6 \] \[ \delta_7 \] \[ \delta_8 \] \[ \delta_9 \]

\[ X \]

\[ X_1 \] \[ X_2 \] \[ X_3 \] \[ X_4 \] \[ X_5 \] \[ X_6 \] \[ X_7 \] \[ X_8 \] \[ X_9 \]

\[ Y \]

\[ Y_1 \] \[ Y_2 \] \[ Y_3 \] \[ Y_4 \] \[ Y_5 \] \[ Y_6 \] \[ Y_7 \] \[ Y_8 \]

Error \[ \epsilon_1 \] \[ \epsilon_2 \] \[ \epsilon_3 \] \[ \epsilon_4 \] \[ \epsilon_5 \] \[ \epsilon_6 \] \[ \epsilon_7 \] \[ \epsilon_8 \]

\[ \beta_{y,x} \] \[ \beta_{x,y} \] \[ r_{xy} \] \[ r_{x_4 y_4 . x_5} \] \[ R_{y . x_6 x_7 x_8} \]
A latent variable approach to measurement

Overview
Correlation
Multivariate Regression
Significance

Measurement: A latent variable approach.

Error X Latent X Latent Y Y Error

δ1 X1 χ1 ξ1 Y1 ϵ1
δ2 X2 χ1 ξ1 Y2 ϵ2
δ3 X3 χ1 ξ1 Y3 ϵ3
δ4 X4 χ2 ξ2 Y4 ϵ4
δ5 X5 χ2 ξ2 Y5 ϵ5
δ6 X6 χ2 ξ2 Y6 ϵ6
δ7 X7 χ3 ξ2 Y7 ϵ7
δ8 X8 χ3 ξ2 Y8 ϵ8
δ9 X9
Reliability: How well does a test reflect one latent trait?

Error X Latent X

\[ \begin{align*}
\delta_1 & \rightarrow X_1 \\
\delta_2 & \rightarrow X_2 \\
\delta_3 & \rightarrow X_3
\end{align*} \]

Latent X

\[ \begin{align*}
\chi_1 & \rightarrow Y_1 \\
\chi_2 & \rightarrow Y_2 \\
\chi_3 & \rightarrow Y_3
\end{align*} \]

\[ \begin{align*}
\epsilon_1 & \leftarrow Y_1 \\
\epsilon_2 & \leftarrow Y_2 \\
\epsilon_3 & \leftarrow Y_3
\end{align*} \]

\[ \begin{align*}
\delta_4 & \rightarrow X_4 \\
\delta_5 & \rightarrow X_5 \\
\delta_6 & \rightarrow X_6 \quad \chi_1
\end{align*} \]

Latent Y

\[ \begin{align*}
\xi_1 & \rightarrow Y_4 \\
\xi_2 & \rightarrow Y_5 \\
\xi_3 & \rightarrow Y_6
\end{align*} \]

\[ \begin{align*}
\epsilon_4 & \leftarrow Y_4 \\
\epsilon_5 & \leftarrow Y_5 \\
\epsilon_6 & \leftarrow Y_6
\end{align*} \]

\[ \begin{align*}
\delta_7 & \rightarrow X_7 \\
\delta_8 & \rightarrow X_8 \\
\delta_9 & \rightarrow X_9
\end{align*} \]
Face, Concurrent, Predictive, Construct

Types of validity: What are we measuring

- Face (Faith)
- Concurrent
- Predictive
- Construct
Psychometric Theory: Data, Measurement, Theory

Error | X | Latent X | Latent Y | Y | Error

\( \delta_1 \rightarrow X_1 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_1 \rightarrow \epsilon_1 \)
\( \delta_2 \rightarrow X_2 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_2 \rightarrow \epsilon_2 \)
\( \delta_3 \rightarrow X_3 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_3 \rightarrow \epsilon_3 \)
\( \delta_4 \rightarrow X_4 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_4 \rightarrow \epsilon_4 \)
\( \delta_5 \rightarrow X_5 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_5 \rightarrow \epsilon_5 \)
\( \delta_6 \rightarrow X_6 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_6 \rightarrow \epsilon_6 \)
\( \delta_7 \rightarrow X_7 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_7 \rightarrow \epsilon_7 \)
\( \delta_8 \rightarrow X_8 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_8 \rightarrow \epsilon_8 \)

\( \delta_9 \rightarrow X_9 \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_1 \rightarrow \epsilon_1 \)
\( \delta_{10} \rightarrow X_{10} \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_{10} \rightarrow \epsilon_{10} \)
\( \delta_{11} \rightarrow X_{11} \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_{11} \rightarrow \epsilon_{11} \)
\( \delta_{12} \rightarrow X_{12} \rightarrow \chi_1 \rightarrow \xi_1 \rightarrow Y_{12} \rightarrow \epsilon_{12} \)

\( \chi_2 \rightarrow \xi_2 \rightarrow Y_2 \rightarrow \epsilon_2 \)
\( \chi_3 \rightarrow \xi_2 \rightarrow Y_3 \rightarrow \epsilon_3 \)
\( \chi_4 \rightarrow \xi_2 \rightarrow Y_4 \rightarrow \epsilon_4 \)
\( \chi_5 \rightarrow \xi_2 \rightarrow Y_5 \rightarrow \epsilon_5 \)
\( \chi_6 \rightarrow \xi_2 \rightarrow Y_6 \rightarrow \epsilon_6 \)
\( \chi_7 \rightarrow \xi_2 \rightarrow Y_7 \rightarrow \epsilon_7 \)
\( \chi_8 \rightarrow \xi_2 \rightarrow Y_8 \rightarrow \epsilon_8 \)

\( \xi_1 \rightarrow Y_1 \rightarrow \epsilon_1 \)
\( \xi_2 \rightarrow Y_2 \rightarrow \epsilon_2 \)
\( \xi_3 \rightarrow Y_3 \rightarrow \epsilon_3 \)
\( \xi_4 \rightarrow Y_4 \rightarrow \epsilon_4 \)
\( \xi_5 \rightarrow Y_5 \rightarrow \epsilon_5 \)
\( \xi_6 \rightarrow Y_6 \rightarrow \epsilon_6 \)
\( \xi_7 \rightarrow Y_7 \rightarrow \epsilon_7 \)
\( \xi_8 \rightarrow Y_8 \rightarrow \epsilon_8 \)
Psychometric Theory: Data, Measurement, Theory

Error X Latent X Latent Y Y Error

Avoidance

Affect

Approach

Cog

Ability

δ₁ X₁ Y₁ €₁
δ₂ X₂ Y₂ €₂
δ₃ X₃ Y₃ €₃
δ₄ X₄ Y₄ €₄
δ₅ X₅ Y₅ €₅
δ₆ X₆ Y₆ €₆
δ₇ X₇ Y₇ €₇
δ₈ X₈ Y₈ €₈
δ₉ X₉
Correlation and Regression

1. Developed in 1886 by Francis Galton
   - Further developments by Karl Pearson and Charles Spearman

2. Correlation/regression are the root concept of psychometrics
   - Other statistics, including factor analysis are ways of partitioning correlation matrices
   - Reliability theory is merely an application of factor analysis
Francis Galton 1822-1911

Francis Galton (1822-1911) was among the most influential psychologists of the 19th century. He did pioneering work on the correlation coefficient, behavior genetics and the measurement of individual differences. He introspectively examined the question of free will and introduced the lexical hypothesis to the study of personality and character. In addition to psychology, he did pioneering work in meteorology and introduced the scientific use of fingerprints. Whenever he could, he counted.
Karl Pearson 1857-1936

Carl (Karl) Pearson was among the most influential statisticians of the early 20th century. Founder of the statistics department at University College London. He developed the Pearson Product Moment Correlation Coefficient, its special case the $\phi$ coefficient, and the tetrachoric correlation. Major behavior geneticist and eugenicist.
Charles Spearman 1863-1945

Charles Spearman (1863-1945) was the leading psychometrician of the early 20th century. His work on the classical test theory, factor analysis, and the g theory of intelligence continues to influence psychometrics, statistics, and the study of intelligence. More than 100 years after their publication, his most influential papers remain two of the most frequently cited articles in psychometrics and intelligence.
Galton’s height data

**Table**: The relationship between the average of both parents (mid parent) and the height of their children. The basic data table is from who used these data to introduce reversion to the mean (and thus, linear regression). The data are available as part of the **UsingR** or **psych** packages.

```r
> library(psych)
> data(galton)
> galton.tab <- table(galton)
> galton.tab[order(rank(rownames(galton.tab)),decreasing=TRUE),] #sort it by decreasing row values

<table>
<thead>
<tr>
<th>child</th>
<th>61.7</th>
<th>62.2</th>
<th>63.2</th>
<th>64.2</th>
<th>65.2</th>
<th>66.2</th>
<th>67.2</th>
<th>68.2</th>
<th>69.2</th>
<th>70.2</th>
<th>71.2</th>
<th>72.2</th>
<th>73.2</th>
<th>73.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>73</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>72.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>71.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>70.5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>12</td>
<td>18</td>
<td>14</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>69.5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>4</td>
<td>17</td>
<td>27</td>
<td>20</td>
<td>33</td>
<td>25</td>
<td>20</td>
<td>11</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>68.5</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>11</td>
<td>16</td>
<td>25</td>
<td>31</td>
<td>34</td>
<td>48</td>
<td>21</td>
<td>18</td>
<td>4</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>67.5</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>14</td>
<td>15</td>
<td>36</td>
<td>38</td>
<td>28</td>
<td>38</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>66.5</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>17</td>
<td>17</td>
<td>14</td>
<td>13</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>65.5</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>5</td>
<td>7</td>
<td>11</td>
<td>11</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>64.5</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>64</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```
Galton’s height data

Figure: Galton’s data can be plotted to show the relationships between mid parent and child heights. Because the original data are grouped, the data points have been jittered to emphasize the density of points along the median. The bars connect the first, 2nd (median) and third quartiles. The dashed line is the best fitting linear fit, the ellipses represent one and two standard deviations from the mean.
Bivariate Regression

\[ \hat{y} = \beta_{y,x} x + \epsilon \]

\[ \beta_{y,x} = \frac{\sigma_{xy}}{\sigma_x^2} \]
Bivariate Regression

\[ \hat{y} = \beta_{y,x} x + \epsilon \]

\[ \beta_{y,x} = \frac{\sigma_{xy}}{\sigma_x^2} \]

\[ \hat{x} = \beta_{x,y} y + \delta \]

\[ \beta_{y,x} = \frac{\sigma_{xy}}{\sigma_y^2} \]
Bivariate Correlation is the geometric average of the two regressions

\[
\hat{x} = \beta_{x.y} y + \delta
\]

\[
\hat{y} = \beta_{y.x} x + \epsilon
\]

\[
\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_y^2}
\]

\[
\beta_{x.y} = \frac{\sigma_{xy}}{\sigma_x^2}
\]

\[
r_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}
\]
The variance and the variance of a composite

1. If \( \mathbf{x}_1 \) and \( \mathbf{x}_2 \) are vectors of \( N \) observations centered around their mean (that is, deviation scores) their variances are
\[
V_{x_1} = \frac{\sum x^2_{i1}}{N - 1} \quad \text{and} \quad V_{x_2} = \frac{\sum x^2_{i2}}{N - 1},
\]
or, in matrix terms
\[
V_{x_1} = \mathbf{x}_1' \mathbf{x}_1 / (N - 1) \quad \text{and} \quad V_{x_2} = \mathbf{x}_2' \mathbf{x}_2 / (N - 1).
\]

2. The variance of the composite made up of the sum of the corresponding scores, \( \mathbf{x} + \mathbf{y} \) is just
\[
V_{(x_1+x_2)} = \frac{\sum(x_i + y_i)^2}{N - 1} = \frac{\sum x^2_i + \sum y^2_i + 2 \sum x_i y_i}{N - 1} = \frac{(\mathbf{x} + \mathbf{y})'(\mathbf{x} + \mathbf{y})}{N - 1}.
\]

Or, more generally,
\[
\mathbf{S} = \begin{pmatrix}
V_{x_1} & C_{x1x2} & \cdots & C_{x1xn} \\
C_{x1x2} & V_{x2} & \cdots & C_{x2xn} \\
\vdots & \vdots & \ddots & \vdots \\
C_{x1xn} & C_{x2xn} & \cdots & V_{xn}
\end{pmatrix}
\]
Sums as matrix products

\[ V_X = \sum \frac{X'X}{N - 1} = \frac{1'(X'X)1}{N - 1} \]

\[ V_Y = \sum \frac{Y'Y}{N - 1} = \frac{1'(Y'Y)1}{N - 1} \]

and

\[ C_{XY} = \sum \frac{X'Y}{N - 1} = \frac{1'(X'Y)1}{N - 1} \]
Use R
A nice feature of R is that you can read from remote data sets. The example dataset is on the personality-project.org server. Get it and describe it.

```r
> datafilename="http://personality-project.org/r/datasets/psychometrics.prob2.txt"
> mydata =read.table(datafilename,header=TRUE)  #read the data file
> describe(mydata,skew=FALSE)
```

```
   var  n  mean   sd median trimmed mad  min  max  range   se
  ID 1 1000 500.50 288.82  500.50  500.50 370.65 1.0 1000.0 999.00 9.13
GREV 2 1000 499.77 106.11  497.50  498.75 106.01 138.0 873.00 735.00 3.36
GREQ 3 1000 500.53 103.85  498.00  498.51 105.26  191.0 914.00 723.00 3.28
GREA 4 1000 498.13 100.45  495.00  498.67  99.33 207.0 848.00 641.00 3.18
  Ach 5 1000 49.93  9.84  50.00  49.88 10.38  16.0 79.00 63.00 0.31
  Anx 6 1000 50.32  9.91  50.00  50.43 10.38  14.0 78.00 64.00 0.31
 Prelim 7 1000 10.03  1.06  10.00  10.02  1.48  7.0 13.00  6.00 0.03
   GPA 8 1000  4.00  0.50  4.02  4.01  0.53  2.5  5.38  2.88 0.02
   MA 9 1000  3.00  0.49  3.00  3.00  0.44  1.4  4.50  3.10 0.02
```
Plot it using the `pairs.panels` function.

Use the `pairs.panels` function to show a splom plot (use `gap=0` and `pch='.'`).

```r
> pairs.panels(mydata, pch='.', gap=0)  # pch='.' makes for a cleaner plot
```
Plot a subset of the data using the c() function (concatenate).

Use the pairs.panels function to show a splom plot. Select a subset of variables using the c() function.

> pairs.panels(mydata[c(2:4,6:8)],pch=’.’)

![SPLOM plot with correlation coefficients]

- GREV: 0.73
- GREQ: 0.60
- GREA: -0.39
- Anx: -0.23
- Prelim: 0.42
- GPA: 0.60

**Correlation Matrix**

<table>
<thead>
<tr>
<th></th>
<th>GREV</th>
<th>GREQ</th>
<th>GREA</th>
<th>Anx</th>
<th>Prelim</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>GREV</td>
<td>0.73</td>
<td>0.60</td>
<td>-0.39</td>
<td>-0.23</td>
<td>0.42</td>
<td>0.60</td>
</tr>
<tr>
<td>GREQ</td>
<td>0.60</td>
<td>0.38</td>
<td>0.57</td>
<td>-0.22</td>
<td>0.42</td>
<td>0.37</td>
</tr>
<tr>
<td>GREA</td>
<td>-0.39</td>
<td>0.57</td>
<td>0.52</td>
<td>-0.22</td>
<td>0.42</td>
<td>-0.39</td>
</tr>
<tr>
<td>Anx</td>
<td>-0.23</td>
<td>-0.22</td>
<td>-0.22</td>
<td>0.42</td>
<td>-0.22</td>
<td>-0.22</td>
</tr>
<tr>
<td>Prelim</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>GPA</td>
<td>0.60</td>
<td>0.37</td>
<td>-0.39</td>
<td>-0.39</td>
<td>0.60</td>
<td>0.60</td>
</tr>
</tbody>
</table>
Do this for the first 200 subjects

```r
> pairs.panels(mydata[mydata$ID < 200, c(2:4,6:8)])
```

<table>
<thead>
<tr>
<th>GREV</th>
<th>0.78</th>
<th>0.67</th>
<th>-0.06</th>
<th>0.45</th>
<th>0.51</th>
</tr>
</thead>
<tbody>
<tr>
<td>GREQ</td>
<td>0.62</td>
<td>-0.05</td>
<td>0.44</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>GREA</td>
<td>-0.41</td>
<td>0.52</td>
<td>0.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anx</td>
<td>-0.19</td>
<td>-0.23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prelim</td>
<td></td>
<td></td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In order to do interaction terms in regressions, it is necessary to 0 center the data. We need to turn the result into a data.frame in order to use it in the regression function.

```r
> cent <- data.frame(scale(mydata,scale=FALSE))
> describe(cent,skew=FALSE)
```

<table>
<thead>
<tr>
<th>var</th>
<th>n</th>
<th>mean</th>
<th>sd</th>
<th>median</th>
<th>trimmed</th>
<th>mad</th>
<th>min</th>
<th>max</th>
<th>range</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>1000</td>
<td>0 288.82</td>
<td>0.00</td>
<td>0.00</td>
<td>370.65</td>
<td>-499.50</td>
<td>499.50</td>
<td>999.00</td>
<td>9.13</td>
<td></td>
</tr>
<tr>
<td>GREV</td>
<td>1000</td>
<td>0 106.11</td>
<td>-2.27</td>
<td>-1.02</td>
<td>106.01</td>
<td>-361.77</td>
<td>373.23</td>
<td>735.00</td>
<td>3.36</td>
<td></td>
</tr>
<tr>
<td>GREQ</td>
<td>1000</td>
<td>0 105.85</td>
<td>-2.53</td>
<td>-2.02</td>
<td>105.26</td>
<td>-309.53</td>
<td>413.47</td>
<td>723.00</td>
<td>3.28</td>
<td></td>
</tr>
<tr>
<td>GREA</td>
<td>1000</td>
<td>0 100.45</td>
<td>-3.13</td>
<td>0.54</td>
<td>99.33</td>
<td>-291.13</td>
<td>349.87</td>
<td>641.00</td>
<td>3.18</td>
<td></td>
</tr>
<tr>
<td>Ach</td>
<td>1000</td>
<td>0 9.84</td>
<td>0.07</td>
<td>-0.05</td>
<td>10.38</td>
<td>-33.93</td>
<td>29.07</td>
<td>63.00</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Anx</td>
<td>1000</td>
<td>0 9.91</td>
<td>-0.32</td>
<td>0.11</td>
<td>10.38</td>
<td>-36.32</td>
<td>27.68</td>
<td>64.00</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Prelim</td>
<td>1000</td>
<td>0 1.06</td>
<td>-0.03</td>
<td>0.00</td>
<td>1.48</td>
<td>-3.03</td>
<td>29.07</td>
<td>63.00</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>GPA</td>
<td>1000</td>
<td>0 0.50</td>
<td>0.02</td>
<td>0.00</td>
<td>0.53</td>
<td>-1.50</td>
<td>1.38</td>
<td>2.88</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>MA</td>
<td>1000</td>
<td>0 0.49</td>
<td>0.00</td>
<td>0.00</td>
<td>0.44</td>
<td>-1.60</td>
<td>1.50</td>
<td>3.10</td>
<td>0.02</td>
<td></td>
</tr>
</tbody>
</table>

The standard deviations and ranges have not changed. However, the means are all 0. We use the scale function with the scale=FALSE option.
Alternatively, we could standardize it.

```r
> z.data <- data.frame(scale(my.data))
> describe(z.data)

<table>
<thead>
<tr>
<th>var</th>
<th>n</th>
<th>mean</th>
<th>sd</th>
<th>median</th>
<th>trimmed</th>
<th>mad</th>
<th>min</th>
<th>max</th>
<th>range</th>
<th>skew</th>
<th>kurtosis</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>1000</td>
<td>0</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>1.28</td>
<td>-1.73</td>
<td>1.73</td>
<td>3.46</td>
<td>0.00</td>
<td>-1.20</td>
<td>0.03</td>
</tr>
<tr>
<td>GREV</td>
<td>1000</td>
<td>0</td>
<td>1</td>
<td>-0.02</td>
<td>-0.01</td>
<td>1.00</td>
<td>-3.41</td>
<td>3.52</td>
<td>6.93</td>
<td>0.09</td>
<td>-0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>GREQ</td>
<td>1000</td>
<td>0</td>
<td>1</td>
<td>-0.02</td>
<td>-0.02</td>
<td>1.01</td>
<td>-2.98</td>
<td>3.98</td>
<td>6.96</td>
<td>0.22</td>
<td>0.08</td>
<td>0.03</td>
</tr>
<tr>
<td>GREA</td>
<td>1000</td>
<td>0</td>
<td>1</td>
<td>-0.03</td>
<td>0.01</td>
<td>0.99</td>
<td>-2.90</td>
<td>3.48</td>
<td>6.38</td>
<td>-0.02</td>
<td>-0.06</td>
<td>0.03</td>
</tr>
<tr>
<td>Ach</td>
<td>1000</td>
<td>0</td>
<td>1</td>
<td>0.01</td>
<td>-0.01</td>
<td>1.05</td>
<td>-3.45</td>
<td>2.95</td>
<td>6.40</td>
<td>0.00</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>Anx</td>
<td>1000</td>
<td>0</td>
<td>1</td>
<td>-0.03</td>
<td>0.01</td>
<td>1.05</td>
<td>-3.67</td>
<td>2.79</td>
<td>6.46</td>
<td>-0.14</td>
<td>0.14</td>
<td>0.03</td>
</tr>
<tr>
<td>Prelim</td>
<td>1000</td>
<td>0</td>
<td>1</td>
<td>-0.02</td>
<td>0.00</td>
<td>1.40</td>
<td>-2.86</td>
<td>2.81</td>
<td>5.67</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>GPA</td>
<td>1000</td>
<td>0</td>
<td>1</td>
<td>0.03</td>
<td>0.01</td>
<td>1.06</td>
<td>-3.00</td>
<td>2.74</td>
<td>5.74</td>
<td>-0.07</td>
<td>-0.29</td>
<td>0.03</td>
</tr>
<tr>
<td>MA</td>
<td>1000</td>
<td>0</td>
<td>1</td>
<td>0.01</td>
<td>0.01</td>
<td>0.90</td>
<td>-3.23</td>
<td>3.04</td>
<td>6.27</td>
<td>-0.07</td>
<td>-0.09</td>
<td>0.03</td>
</tr>
</tbody>
</table>
```

Or, we can standardize it by dividing though by the standard deviation. We use the `scale` function to do this for us.
Displaying the data

Show how the correlations do not change with standardization

Find the correlations using the `lowerCor` function. This, by default, uses pairwise Pearson correlations and rounds to two decimals. Compare with the standard `cor` function.

```r
> lowerCor(my.data)
> lowerCor(z.data)
```

<table>
<thead>
<tr>
<th></th>
<th>ID</th>
<th>GREV</th>
<th>GREQ</th>
<th>GREA</th>
<th>Ach</th>
<th>Anx</th>
<th>Prelm</th>
<th>GPA</th>
<th>MA</th>
<th>ID</th>
<th>GREV</th>
<th>GREQ</th>
<th>GREA</th>
<th>Ach</th>
<th>Anx</th>
<th>Prelm</th>
<th>GPA</th>
<th>MA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GREV</td>
<td>-0.01</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GREV</td>
<td>-0.01</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GREQ</td>
<td>0.00</td>
<td>0.73</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GREQ</td>
<td>0.00</td>
<td>0.73</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GREA</td>
<td>-0.01</td>
<td>0.64</td>
<td>0.60</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GREA</td>
<td>-0.01</td>
<td>0.64</td>
<td>0.60</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ach</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.45</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ach</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.45</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anx</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.39</td>
<td>-0.56</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td>Anx</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.39</td>
<td>-0.56</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prelm</td>
<td>0.02</td>
<td>0.43</td>
<td>0.38</td>
<td>0.57</td>
<td>0.30</td>
<td>-0.23</td>
<td>1.00</td>
<td></td>
<td></td>
<td>Prelm</td>
<td>0.02</td>
<td>0.43</td>
<td>0.38</td>
<td>0.57</td>
<td>0.30</td>
<td>-0.23</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>GPA</td>
<td>0.00</td>
<td>0.42</td>
<td>0.37</td>
<td>0.52</td>
<td>0.28</td>
<td>-0.22</td>
<td>0.42</td>
<td>1.00</td>
<td></td>
<td>GPA</td>
<td>0.00</td>
<td>0.42</td>
<td>0.37</td>
<td>0.52</td>
<td>0.28</td>
<td>-0.22</td>
<td>0.42</td>
<td>1.00</td>
</tr>
<tr>
<td>MA</td>
<td>-0.01</td>
<td>0.32</td>
<td>0.29</td>
<td>0.45</td>
<td>0.26</td>
<td>-0.22</td>
<td>0.36</td>
<td>0.31</td>
<td>1.00</td>
<td>MA</td>
<td>-0.01</td>
<td>0.32</td>
<td>0.29</td>
<td>0.45</td>
<td>0.26</td>
<td>-0.22</td>
<td>0.36</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Show that the two matrices do not differ using the `lowerUpper` function

```r
r <- lowerCor(my.data)  # find the original correlations
z <- lowerCor(z.data)  # find the z transformed correlations
lu <- lowerUpper(r,z,diff=TRUE)  # combine into one matrix and take the difference

round(lu,2)
```

<table>
<thead>
<tr>
<th></th>
<th>ID</th>
<th>GREV</th>
<th>GREQ</th>
<th>GREA</th>
<th>Ach</th>
<th>Anx</th>
<th>Prelim</th>
<th>GPA</th>
<th>MA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>NA</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>GREV</td>
<td>-0.01</td>
<td>NA</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>GREQ</td>
<td>0.00</td>
<td>0.73</td>
<td>NA</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>GREA</td>
<td>-0.01</td>
<td>0.64</td>
<td>0.60</td>
<td>NA</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>Ach</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.45</td>
<td>NA</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>Anx</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.39</td>
<td>-0.56</td>
<td>NA</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>Prelim</td>
<td>0.02</td>
<td>0.43</td>
<td>0.38</td>
<td>0.57</td>
<td>0.30</td>
<td>-0.23</td>
<td>NA</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>GPA</td>
<td>0.00</td>
<td>0.42</td>
<td>0.37</td>
<td>0.52</td>
<td>0.28</td>
<td>-0.22</td>
<td>0.42</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>MA</td>
<td>-0.01</td>
<td>0.32</td>
<td>0.29</td>
<td>0.45</td>
<td>0.26</td>
<td>-0.22</td>
<td>0.36</td>
<td>0.31</td>
<td>NA</td>
</tr>
</tbody>
</table>
Scatter Plot Matrix showing correlation and LOESS regression

UgradGPA

GRE.Q

GRE.V

Correlation

Multivariate Regression

Significance
The effect of selection on the correlation

- Consider what happens if we select a subset
  - The “Oregon” model
  - \((GPA + (V+Q)/200) > 11.6\)
- The range is truncated, but even more important, by using a compensatory selection model, we have changed the sign of the correlations.
Although the correlation is very sensitive, regression slopes are relatively insensitive to restriction of range.
R code for regression figures

gradq <- subset(gradf, gradf[2] > 700) # choose the subset
with(gradq, lm(GRE.V ~ GRE.Q)) # do the regression
Call:
  lm(formula = GRE.V ~ GRE.Q)
Coefficients:
     (Intercept)      GRE.Q
  258.1549     0.4977

# show the graphic
op <- par(mfrow=c(1,2)) # two panel graph
with(gradf, {
  plot(GRE.V ~ GRE.Q, xlim=c(200,800), main='Original data', pch=16)
  abline(lm(GRE.V ~ GRE.Q))
})
text(300,500, 'r = .46    b = .56')
with(gradq, {
  plot(GRE.V ~ GRE.Q, xlim=c(200,800), main='GRE Q > 700', pch=16)
  abline(lm(GRE.V ~ GRE.Q))
})
text(300,500, 'r = .18    b = .50')
op <- par(mfrow=c(1,1)) # switch back to one panel
Show many correlations with a heat map using `cor.plot`.

Big 5 Inventory Items from SAPA
Table: A number of correlations are Pearson $r$ in different forms, or with particular assumptions. If $r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$, then depending upon the type of data being analyzed, a variety of correlations are found.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>symbol</th>
<th>X</th>
<th>Y</th>
<th>Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>$r$</td>
<td>continuous</td>
<td>continuous</td>
<td></td>
</tr>
<tr>
<td>Spearman</td>
<td>rho ($\rho$)</td>
<td>ranks</td>
<td>ranks</td>
<td></td>
</tr>
<tr>
<td>Point bi-serial</td>
<td>$r_{pb}$</td>
<td>dichotomous</td>
<td>continuous</td>
<td></td>
</tr>
<tr>
<td>Phi</td>
<td>$\phi$</td>
<td>dichotomous</td>
<td>dichotomous</td>
<td></td>
</tr>
<tr>
<td>Bi-serial</td>
<td>$r_{bis}$</td>
<td>dichotomous</td>
<td>continuous</td>
<td>normality</td>
</tr>
<tr>
<td>Tetrachoric</td>
<td>$r_{tet}$</td>
<td>dichotomous</td>
<td>dichotomous</td>
<td>bivariate normality</td>
</tr>
<tr>
<td>Polychoric</td>
<td>$r_{pc}$</td>
<td>categorical</td>
<td>categorical</td>
<td>bivariate normality</td>
</tr>
</tbody>
</table>
The $\phi$ coefficient is just a Pearson r on dichotomous data

**Table**: The basic table for a phi, $\phi$ coefficient, expressed in raw frequencies in a four fold table is taken from ?

<table>
<thead>
<tr>
<th></th>
<th>Success</th>
<th>Failure</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept</td>
<td>A</td>
<td>B</td>
<td>$R_1 = A + B$</td>
</tr>
<tr>
<td>Reject</td>
<td>C</td>
<td>D</td>
<td>$R_2 = C + D$</td>
</tr>
<tr>
<td>Total</td>
<td>$C_1 = A + C$</td>
<td>$C_2 = B + D$</td>
<td>$n = A + B + C + D$</td>
</tr>
</tbody>
</table>

In terms of the raw data coded 0 or 1, the *phi coefficient* can be derived directly by direct substitution, recognizing that the only non zero product is found in the A cell

$$n \sum X_iY_i - \sum X_i \sum Y_i = nA - R_1 C_1$$

$$\phi = \frac{AD - BC}{\sqrt{(A + B)(C + D)(A + C)(B + D)}}.$$  (2)
Correlation size ≠ causal importance

Table: The relationship between sex and pregnancy (hypothetical data)

<table>
<thead>
<tr>
<th></th>
<th>Pregnant</th>
<th>Not Pregnant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercourse</td>
<td>2</td>
<td>1,041</td>
<td>1,043</td>
</tr>
<tr>
<td>No intercourse</td>
<td>0</td>
<td>6,257</td>
<td>6,257</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>7,298</td>
<td>7,300</td>
</tr>
</tbody>
</table>

> sex <- c(2, 1041, 0, 6257)
> phi(sex)

[1] 0.04
The biserial correlation estimates the latent correlation

- $r = 0.9$, $r_{pb} = 0.71$, $r_{bis} = 0.89$
- $r = 0.6$, $r_{pb} = 0.48$, $r_{bis} = 0.6$
- $r = 0.3$, $r_{pb} = 0.23$, $r_{bis} = 0.28$
- $r = 0$, $r_{pb} = 0.02$, $r_{bis} = 0.02$
The tetrachoric correlation estimates the latent correlation.
Correlation size ≠ causal importance – tetrachoric correlation

Table: The relationship between sex and pregnancy (hypothetical data)

<table>
<thead>
<tr>
<th></th>
<th>Pregnant</th>
<th>Not Pregnant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercourse</td>
<td>2</td>
<td>1,041</td>
<td>1,043</td>
</tr>
<tr>
<td>No intercourse</td>
<td>0</td>
<td>6,257</td>
<td>6,257</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>7,298</td>
<td>7,300</td>
</tr>
</tbody>
</table>

Phi = 0.04
\( \rho_{tet} = 0.95 \)

```r
> sex <- c(2, 1041, 0, 6257)
> phi(sex)
[1] 0.04
> tetrachoric(sex, correct=FALSE)
Call: tetrachoric(x = sex, correct = FALSE)
tetrachoric correlation
[1] 0.95

with tau of
[1] -3.5 -1.1```
Displaying the data

Pearson r versus tetrachoric correlation on dichotomous ability data

> tet <- tetrachoric(ability)
Loading required package: mvtnorm
Loading required package: parallel
> per <- lowerCor(ability)
> per.tet <- lowerUpper(tet$rho,per)
> per.tet.diff <- lowerUpper(tet$rho,per,diff=TRUE)
> round(per.tet[1:8,1:8],2)

<table>
<thead>
<tr>
<th>reason.4</th>
<th>reason.16</th>
<th>reason.17</th>
<th>reason.19</th>
<th>letter.7</th>
<th>letter.33</th>
<th>letter.34</th>
<th>letter.58</th>
</tr>
</thead>
<tbody>
<tr>
<td>reason.4</td>
<td>NA</td>
<td>0.28</td>
<td>0.40</td>
<td>0.30</td>
<td>0.28</td>
<td>0.23</td>
<td>0.29</td>
</tr>
<tr>
<td>reason.16</td>
<td>0.45</td>
<td>NA</td>
<td>0.32</td>
<td>0.25</td>
<td>0.27</td>
<td>0.20</td>
<td>0.26</td>
</tr>
<tr>
<td>reason.17</td>
<td>0.61</td>
<td>0.51</td>
<td>NA</td>
<td>0.34</td>
<td>0.29</td>
<td>0.26</td>
<td>0.29</td>
</tr>
<tr>
<td>reason.19</td>
<td>0.46</td>
<td>0.40</td>
<td>0.53</td>
<td>NA</td>
<td>0.25</td>
<td>0.25</td>
<td>0.27</td>
</tr>
<tr>
<td>letter.7</td>
<td>0.45</td>
<td>0.43</td>
<td>0.47</td>
<td>0.40</td>
<td>NA</td>
<td>0.34</td>
<td>0.40</td>
</tr>
<tr>
<td>letter.33</td>
<td>0.37</td>
<td>0.32</td>
<td>0.42</td>
<td>0.39</td>
<td>0.52</td>
<td>NA</td>
<td>0.37</td>
</tr>
<tr>
<td>letter.34</td>
<td>0.46</td>
<td>0.41</td>
<td>0.47</td>
<td>0.43</td>
<td>0.60</td>
<td>0.56</td>
<td>NA</td>
</tr>
<tr>
<td>letter.58</td>
<td>0.47</td>
<td>0.35</td>
<td>0.48</td>
<td>0.40</td>
<td>0.51</td>
<td>0.43</td>
<td>0.50</td>
</tr>
</tbody>
</table>

> round(per.tet.diff[1:8,1:8],2)

<table>
<thead>
<tr>
<th>reason.4</th>
<th>reason.16</th>
<th>reason.17</th>
<th>reason.19</th>
<th>letter.7</th>
<th>letter.33</th>
<th>letter.34</th>
<th>letter.58</th>
</tr>
</thead>
<tbody>
<tr>
<td>reason.4</td>
<td>NA</td>
<td>0.17</td>
<td>0.21</td>
<td>0.17</td>
<td>0.16</td>
<td>0.14</td>
<td>0.17</td>
</tr>
<tr>
<td>reason.16</td>
<td>0.45</td>
<td>NA</td>
<td>0.19</td>
<td>0.15</td>
<td>0.16</td>
<td>0.13</td>
<td>0.16</td>
</tr>
<tr>
<td>reason.17</td>
<td>0.61</td>
<td>0.51</td>
<td>NA</td>
<td>0.19</td>
<td>0.18</td>
<td>0.16</td>
<td>0.18</td>
</tr>
<tr>
<td>reason.19</td>
<td>0.46</td>
<td>0.40</td>
<td>0.53</td>
<td>NA</td>
<td>0.14</td>
<td>0.14</td>
<td>0.15</td>
</tr>
<tr>
<td>letter.7</td>
<td>0.45</td>
<td>0.43</td>
<td>0.47</td>
<td>0.40</td>
<td>NA</td>
<td>0.18</td>
<td>0.20</td>
</tr>
<tr>
<td>letter.33</td>
<td>0.37</td>
<td>0.32</td>
<td>0.42</td>
<td>0.39</td>
<td>0.52</td>
<td>NA</td>
<td>0.19</td>
</tr>
<tr>
<td>letter.34</td>
<td>0.46</td>
<td>0.41</td>
<td>0.47</td>
<td>0.43</td>
<td>0.60</td>
<td>0.56</td>
<td>NA</td>
</tr>
<tr>
<td>letter.58</td>
<td>0.47</td>
<td>0.35</td>
<td>0.48</td>
<td>0.40</td>
<td>0.51</td>
<td>0.43</td>
<td>0.50</td>
</tr>
</tbody>
</table>
 Pearson r versus polychoric correlation on 6 alternative BFI data

```r
> poly <- polychoric(bfi[1:10])
> pearson <- cor(bfi[1:10], use="pairwise")
> poly.pear <- lowerUpper(poly$rho, pearson)
> poly.pear.diff <- lowerUpper(poly$rho, pearson, diff=TRUE)
> poly.pear

   A1   A2  A3  A4  A5  C1  C2  C3  C4  C5
A1 NA -0.34 -0.27 -0.15 -0.18 0.03 0.02 -0.02 0.13 0.05
A2 -0.41 NA  0.49  0.34  0.39 0.09 0.14 0.19 -0.15 -0.12
A3 -0.32  0.56 NA  0.36  0.50 0.10 0.14 0.13 -0.12 -0.16
A4 -0.18  0.39  0.41 NA  0.31 0.09 0.23 0.13 -0.15 -0.24
A5 -0.23  0.45  0.57  0.36 NA 0.12 0.11 0.13 -0.13 -0.17
C1  0.00  0.12  0.12  0.11  0.16 NA 0.43 0.31 -0.34 -0.25
C2  0.01  0.16  0.16  0.27  0.14 NA 0.36 -0.38 -0.30
C3 -0.02  0.23  0.16  0.17  0.15  0.34  0.40 NA -0.34 -0.34
C4  0.15 -0.19 -0.16 -0.20 -0.17 -0.40 -0.43 -0.38 NA  0.48
C5  0.06 -0.16 -0.19 -0.28 -0.20 -0.29 -0.33 -0.38  0.53 NA
> round(poly.pear.diff,2)

   A1   A2  A3  A4  A5  C1  C2  C3  C4  C5
A1 NA -0.07 -0.06 -0.03 -0.05 -0.02 -0.01  0.00  0.02  0.01
A2 -0.41 NA  0.07  0.05  0.06  0.02  0.02  0.03 -0.05 -0.03
A3 -0.32  0.56 NA  0.05  0.07  0.03  0.02  0.03 -0.04 -0.03
A4 -0.18  0.39  0.41 NA  0.05  0.02  0.04  0.04 -0.04 -0.04
A5 -0.23  0.45  0.57  0.36 NA  0.04  0.03  0.02 -0.04 -0.03
C1  0.00  0.12  0.12  0.11  0.16 NA  0.06  0.04 -0.06 -0.04
C2  0.01  0.16  0.16  0.27  0.14  0.48 NA  0.04 -0.05 -0.03
C3 -0.02  0.23  0.16  0.17  0.15  0.34  0.40 NA -0.04 -0.04
C4  0.15 -0.19 -0.16 -0.20 -0.17 -0.40 -0.43 -0.38 NA  0.05
C5  0.06 -0.16 -0.19 -0.28 -0.20 -0.29 -0.33 -0.38  0.53 NA
```
Spearman vs. Pearson on BFI data

```r
> spear <- cor(bfi[1:10], use="pairwise", method="spearman")
> spear.pear <- lowerUpper(spear, pearson, diff=TRUE)
> round(spear.pear, 2)
```

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>NA</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.01</td>
<td>-0.04</td>
<td>-0.05</td>
<td>-0.03</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>A2</td>
<td>-0.37</td>
<td>NA</td>
<td>0.02</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>-0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td>A3</td>
<td>-0.30</td>
<td>0.50</td>
<td>NA</td>
<td>0.00</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>-0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td>A4</td>
<td>-0.16</td>
<td>0.34</td>
<td>0.36</td>
<td>NA</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>-0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td>A5</td>
<td>-0.22</td>
<td>0.40</td>
<td>0.53</td>
<td>0.31</td>
<td>NA</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>-0.03</td>
<td>-0.02</td>
</tr>
<tr>
<td>C1</td>
<td>-0.02</td>
<td>0.11</td>
<td>0.12</td>
<td>0.10</td>
<td>0.15</td>
<td>NA</td>
<td>0.02</td>
<td>0.01</td>
<td>-0.04</td>
<td>-0.01</td>
</tr>
<tr>
<td>C2</td>
<td>-0.01</td>
<td>0.14</td>
<td>0.15</td>
<td>0.25</td>
<td>0.13</td>
<td>0.45</td>
<td>NA</td>
<td>0.01</td>
<td>-0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>C3</td>
<td>-0.04</td>
<td>0.21</td>
<td>0.16</td>
<td>0.15</td>
<td>0.14</td>
<td>0.32</td>
<td>0.37</td>
<td>NA</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>C4</td>
<td>0.15</td>
<td>-0.18</td>
<td>-0.16</td>
<td>-0.18</td>
<td>-0.16</td>
<td>-0.38</td>
<td>-0.40</td>
<td>-0.35</td>
<td>NA</td>
<td>0.01</td>
</tr>
<tr>
<td>C5</td>
<td>0.06</td>
<td>-0.15</td>
<td>-0.18</td>
<td>-0.26</td>
<td>-0.19</td>
<td>-0.26</td>
<td>-0.30</td>
<td>-0.35</td>
<td>0.49</td>
<td>NA</td>
</tr>
</tbody>
</table>
Comments on these alternative correlations

1. The assumption is that there was an underlying bivariate, normal distribution that was somehow artificially dichotomized.

2. But some things are in fact dichotomous, not normally distributed
   - Alive/Dead
   - Vaccinated/Not vaccinated

3. Polychromic and tetrachoric correlations are found by iteratively fitting bivariate normal distributions with varying correlations until the best fit for a n x n table is found.

4. This is done using the tetrachoric or polychoric functions. They are not fast! (In comparison to Pearson r).
### Cautions about correlations—The Anscombe data set

Consider the following 8 variables:

<table>
<thead>
<tr>
<th>var</th>
<th>n</th>
<th>mean</th>
<th>sd</th>
<th>median</th>
<th>trimmed</th>
<th>mad</th>
<th>min</th>
<th>max</th>
<th>range</th>
<th>skew</th>
<th>kurtosis</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1</td>
<td>9.0</td>
<td>3.32</td>
<td>9.00</td>
<td>9.00</td>
<td>4.45</td>
<td>4.00</td>
<td>14.00</td>
<td>10.00</td>
<td>0.00</td>
<td>-1.20</td>
<td>1.00</td>
</tr>
<tr>
<td>x2</td>
<td>2</td>
<td>9.0</td>
<td>3.32</td>
<td>9.00</td>
<td>9.00</td>
<td>4.45</td>
<td>4.00</td>
<td>14.00</td>
<td>10.00</td>
<td>0.00</td>
<td>-1.20</td>
<td>1.00</td>
</tr>
<tr>
<td>x3</td>
<td>3</td>
<td>9.0</td>
<td>3.32</td>
<td>9.00</td>
<td>9.00</td>
<td>4.45</td>
<td>4.00</td>
<td>14.00</td>
<td>10.00</td>
<td>0.00</td>
<td>-1.20</td>
<td>1.00</td>
</tr>
<tr>
<td>x4</td>
<td>4</td>
<td>9.0</td>
<td>3.32</td>
<td>8.00</td>
<td>8.00</td>
<td>0.00</td>
<td>8.00</td>
<td>19.00</td>
<td>11.00</td>
<td>2.47</td>
<td>11.00</td>
<td>1.00</td>
</tr>
<tr>
<td>y1</td>
<td>5</td>
<td>7.5</td>
<td>2.03</td>
<td>7.58</td>
<td>7.49</td>
<td>1.82</td>
<td>4.26</td>
<td>10.84</td>
<td>6.58</td>
<td>-0.05</td>
<td>-0.53</td>
<td>0.61</td>
</tr>
<tr>
<td>y2</td>
<td>6</td>
<td>7.5</td>
<td>2.03</td>
<td>8.14</td>
<td>7.79</td>
<td>1.47</td>
<td>3.10</td>
<td>9.26</td>
<td>6.16</td>
<td>-0.98</td>
<td>0.85</td>
<td>0.61</td>
</tr>
<tr>
<td>y3</td>
<td>7</td>
<td>7.5</td>
<td>2.03</td>
<td>7.11</td>
<td>7.15</td>
<td>1.53</td>
<td>5.39</td>
<td>12.74</td>
<td>7.35</td>
<td>1.38</td>
<td>4.38</td>
<td>0.61</td>
</tr>
<tr>
<td>y4</td>
<td>8</td>
<td>7.5</td>
<td>2.03</td>
<td>7.04</td>
<td>7.20</td>
<td>1.90</td>
<td>5.25</td>
<td>12.50</td>
<td>7.25</td>
<td>1.12</td>
<td>3.15</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Cautions, Anscombe continued

With regressions of

|     | Estimate | Std. Error | t value | Pr(>|t|) |
|-----|----------|------------|---------|----------|
| (Intercept) | 3.00009   | 1.1247468  | 2.667348 | 0.025734051 |
| x1    | 0.50009   | 0.1179055  | 4.241455 | 0.002169629 |

[[2]]

|     | Estimate | Std. Error | t value | Pr(>|t|) |
|-----|----------|------------|---------|----------|
| (Intercept) | 3.0009    | 1.1253024  | 2.666758 | 0.025758941 |
| x2    | 0.50000   | 0.1179637  | 4.238590 | 0.002178816 |

[[3]]

|     | Estimate | Std. Error | t value | Pr(>|t|) |
|-----|----------|------------|---------|----------|
| (Intercept) | 3.00245   | 1.1244812  | 2.670080 | 0.025619109 |
| x3    | 0.49973   | 0.1178777  | 4.239372 | 0.002176305 |

[[4]]

|     | Estimate | Std. Error | t value | Pr(>|t|) |
|-----|----------|------------|---------|----------|
| (Intercept) | 3.00172   | 1.1239211  | 2.670763 | 0.025590425 |
| x4    | 0.49991   | 0.1178189  | 4.243028 | 0.002164602 |
Cautions about correlations: Anscombe data set
Further cautions about correlations—the problem of levels

1. Correlations taken at one level of analysis can be unrelated to those at another level

2. \[ r_{xy} = \eta_{xwg} \times \eta_{ywg} \times r_{xywg} + \eta_{xbg} \times \eta_{ybg} \times r_{xybg} \]

3. Where \( \eta \) is the correlation of the data with the within group values, or the group means.

4. The within group and between group correlations can even be of different sign!

5. The withinBetween data set is an example of this problem.

6. The statsBy function will find the within and between group correlations for this kind of multi-level design.
Cautions about correlations: Within versus between groups
Bias, or just Simpson’s Paradox?

Table: Hypothetical Admissions data showing sex discrimination

<table>
<thead>
<tr>
<th></th>
<th>Admit</th>
<th>Reject</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>40</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>Female</td>
<td>10</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

\[ \Phi = \frac{(VP - HR \times SR)}{\sqrt{HR \times (1-HR) \times SR \times (1-SR)}} = 0.60 \]

Polychoric rho = 0.81
Displaying the data

Calculate the $\phi$ and tetrachoric correlations

```r
> admit <- c(40,10,10,40)
> phi(admit)

[1] 0.6

> phi2poly(.6,.5,.5)

[1] 0.8090178

> tetrachoric(admit)

Call: tetrachoric(x = admit)
tetrachoric correlation

[1] 0.81

with tau of

[1] 0 0
```

1. Input the four cell counts
2. Find the $\phi$ coefficient
3. Convert this to a tetrachoric correlation by specifying the marginals
4. Or, just call tetrachoric with these cell entries
Sex discrimination by department shows opposite effect

Table: Hypothetical Admissions data showing sex discrimination

<table>
<thead>
<tr>
<th></th>
<th>Admit</th>
<th>Reject</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>40</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>Female</td>
<td>10</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

Table: Males: unselective

<table>
<thead>
<tr>
<th></th>
<th>Admit</th>
<th>Reject</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>40</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>Female</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>5</td>
<td>50</td>
</tr>
</tbody>
</table>

Table: Females: selective

<table>
<thead>
<tr>
<th></th>
<th>Admit</th>
<th>Reject</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Female</td>
<td>5</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>45</td>
<td>50</td>
</tr>
</tbody>
</table>

$\phi = -0.11, \rho = -0.95$
### Displaying the data

The ubiquitous correlation coefficient

#### Table: Alternative Estimates of effect size. Using the correlation as a scale free estimate of effect size allows for combining experimental and correlational data in a metric that is directly interpretable as the effect of a standardized unit change in x leads to r change in standardized y.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Estimate</th>
<th>$r$ equivalent</th>
<th>as a function of $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson correlation</td>
<td>$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}$</td>
<td>$r_{xy}$</td>
<td>$r_{xy}$</td>
</tr>
<tr>
<td>Regression</td>
<td>$b_{y,x} = \frac{C_{xy}}{\sigma_y^2}$</td>
<td>$r = b_{y,x} \frac{\sigma_y}{\sigma_x}$</td>
<td>$b_{y,x} = r \frac{\sigma_x}{\sigma_y}$</td>
</tr>
<tr>
<td>Cohen’s $d$</td>
<td>$d = \frac{X_1 - X_2}{\sigma_x}$</td>
<td>$r = \frac{d}{\sqrt{d^2 + 4}}$</td>
<td>$d = \frac{2r}{\sqrt{1 - r^2}}$</td>
</tr>
<tr>
<td>Hedge’s $g$</td>
<td>$g = \frac{X_1 - X_2}{s_x}$</td>
<td>$r = \frac{g}{\sqrt{g^2 + 4(df/N)}}$</td>
<td>$g = \frac{2r\sqrt{df/N}}{\sqrt{1 - r^2}}$</td>
</tr>
<tr>
<td>t-test</td>
<td>$t = \frac{d \sqrt{df}}{2}$</td>
<td>$r = \frac{t^2}{(t^2 + df)}$</td>
<td>$t = \frac{r \sqrt{2df}}{1 - r^2}$</td>
</tr>
<tr>
<td>F-test</td>
<td>$F = \frac{d^2 df}{4}$</td>
<td>$r = \sqrt{F/(F + df)}$</td>
<td>$F = \frac{r^2 df}{1 - r^2}$</td>
</tr>
<tr>
<td>Chi Square</td>
<td>$\chi^2 = r^2 n$</td>
<td>$r = \sqrt{\chi^2/n}$</td>
<td>$\chi^2 = r^2 n$</td>
</tr>
<tr>
<td>Odds ratio</td>
<td>$d = \frac{\ln(OR)}{1.81}$</td>
<td>$r = \frac{\ln(OR)}{1.81 \sqrt{(\ln(OR)/1.81)^2 + 4}}$</td>
<td>$\ln(OR) = \frac{3.62r}{\sqrt{1 - r^2}}$</td>
</tr>
<tr>
<td>$r_{equivalent}$</td>
<td>$r$ with probability $p$</td>
<td>$r = r_{equivalent}$</td>
<td>$r_{equivalent}$</td>
</tr>
</tbody>
</table>
Correlation as the average of regressions

Galton’s insight was that if both $x$ and $y$ were on the same scale with equal variability, then the slope of the line was the same for both predictors and was measure of the strength of their relationship. ? converted all deviations to the same metric by dividing through by half the interquartile range, and ? modified this by converting the numbers to standard scores (i.e., dividing the deviations by the standard deviation). Alternatively, the geometric mean of the two slopes ($b_{xy}$ and $b_{yx}$) leads to the same outcome:

$$r_{xy} = \sqrt{b_{xy}b_{yx}} = \sqrt{\frac{\text{Cov}_{xy}\text{Cov}_{yx}}{\sigma^2_x\sigma^2_y}} = \frac{\text{Cov}_{xy}}{\sqrt{\sigma^2_x\sigma^2_y}} = \frac{\text{Cov}_{xy}}{\sigma_x\sigma_y}$$

(3)

which is the same as the covariance of the standardized scores of $X$ and $Y$.

$$r_{xy} = \text{Cov}_{z_xz_y} = \text{Cov}_{\frac{x}{\sigma_x} \frac{y}{\sigma_y}} = \frac{\text{Cov}_{xy}}{\sigma_x\sigma_y}$$

(4)
The slope $b_{y,x}$ was found so that it minimizes the sum of the squared residual, but what is it? That is, how big is the variance of the residual?

$$V_r = \frac{\sum_{i=1}^{n} (y - \hat{y})^2}{n} = \frac{\sum_{i=1}^{n} (y - b_{y,x}x)^2}{n}$$

$$V_r = \frac{\sum_{i=1}^{n} (y^2 + b_{y,x}^2 x^2 - 2b_{y,x}xy)}{n}$$

$$V_r = V_y + \frac{\text{Cov}_{xy}^2}{V_x} - 2 \frac{\text{Cov}_{xy}^2}{V_x} = V_y - \frac{\text{Cov}_{xy}^2}{V_x}$$

$$V_r = V_y - r_{xy}^2 V_y = V_y (1 - r_{xy}^2) \quad (5)$$

That is, the variance of the residual in Y or the variance of the error of prediction of Y is the product of the original variance of Y and one minus the squared correlation between X and Y. The squared correlation between x and y is thus an index of the amount of variance in Y that is linearly predicted by X. This squared correlation is known as the index of determination.
The various relationships between correlations, predicted scores, the variance of the predicted scores, and the variances of the residuals may be seen in the following table (11).

**Table**: The basic relationships between Variance, Covariance, Correlation and Residuals

<table>
<thead>
<tr>
<th></th>
<th>Variance</th>
<th>Covariance with $X$</th>
<th>Covariance with $Y$</th>
<th>Correlation with $X$</th>
<th>Correlation with $Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>$V_X$</td>
<td>$V_X$</td>
<td>$C_{XY}$</td>
<td>1</td>
<td>$r_{xy}$</td>
</tr>
<tr>
<td>$Y$</td>
<td>$V_Y$</td>
<td>$C_{XY}$</td>
<td>$V_Y$</td>
<td>$r_{xy}$</td>
<td>1</td>
</tr>
<tr>
<td>$\hat{Y}$</td>
<td>$r_{xy}^2 V_Y$</td>
<td>$C_{XY} = r_{xy} \sigma_X \sigma_Y$</td>
<td>$r_{xy} V_Y$</td>
<td>1</td>
<td>$r_{xy}$</td>
</tr>
<tr>
<td>$Y_r = Y - \hat{Y}$</td>
<td>$(1 - r_{xy}^2) V_Y$</td>
<td>0</td>
<td>$(1 - r_{xy}^2) V_Y$</td>
<td>0</td>
<td>$\sqrt{1 - r^2}$</td>
</tr>
</tbody>
</table>
Set theoretic approach: Partitioning the variance in $Y$

$X_1 \rightarrow \beta_{y.x} \rightarrow Y$ 

$\hat{y} = \beta_{y.x} x$

$r_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$ 

$V_r = V_y + \frac{\text{Cov}_{xy}^2}{V_x} - 2 \frac{\text{Cov}_{xy}^2}{V_x}$

$V_r = V_y - \frac{\text{Cov}_{xy}^2}{V_x}$

$V_r = V_y - r_{xy}^2 V_y$

$V_r = V_y(1 - r_{xy}^2)$

Variance in $Y$ predicted by $X = r_{xy}^2 \sigma_y^2$
Distance in the observational space

Because $X$ and $Y$ are vectors in the space defined by the observations, the covariance between them may be thought of in terms of the average squared distance between the two vectors in that same space. That is, following Pythagoras, the distance, $d$, is simply the square root of the sum of the squared distances in each dimension (for each pair of observations), or, if we find the average distance, we can find the square root of the sum of the squared distances divided by $n$:

$$d_{xy}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2. \tag{6}$$

which is the same as

$$d_{xy}^2 = V_x + V_y - 2C_{xy}$$

$$d_{xy} = \sqrt{2 \times (1 - r_{xy})}.$$
Distance, correlations, and the law of cosines

Compare this to the trigonometric law of cosines,

\[ c^2 = a^2 + b^2 - 2ab \cdot \cos(ab), \]

and we see that the distance between two vectors is the sum of their variances minus twice the product of their standard deviations times the cosine of the angle between them. That is, the correlation is the cosine of the angle between the two vectors. The next figure shows these relationships for two Y vectors. The correlation, \( r_1 \), of X with Y1 is the cosine of \( \theta_1 = \) the ratio of the projection of Y1 onto X. From the Pythagorean Theorem, the length of the residual Y with X removed (Y_x) is \( \sigma_Y \sqrt{1 - r^2} \).
A geometric version of correlation

Correlations as cosines

\[
\begin{align*}
\sqrt{1-r_2^2} & \quad y_2 \\
\sqrt{1-r_1^2} & \quad y_1
\end{align*}
\]

\[
\begin{align*}
\theta_2 & \quad \theta_1
\end{align*}
\]

\[
\begin{align*}
-r_2 & \quad r_1
\end{align*}
\]
The Ideal model of predicting $Y$ from $X_1$ and $X_2$

Variance in $Y$ predicted by $X_1$ and $X_2$ if $X_1$ and $X_2$ are independent. $\hat{V}_y = V_{y} r_{x_1y}^2 + V_{y} r_{x_2y}^2$
The usual case of predicting $Y$ from $X_1$ and $X_2$

Variance in $Y$ predicted by $X_1$ and $X_2$ if $X_1$ and $X_2$ - overlapping predictions

\[ \hat{V}_y = V_y r_{x_1,y}^2 + V_y r_{x_2,y}^2 - \text{overlap} \]

But what is the overlap?
Multiple correlations

\[ X_1 \quad r_{x_1y} \quad Y \]

\[ X_2 \quad r_{x_1x_2} \quad X_1 \]

\[ r_{x_2y} \]
Multiple Regression

\[ \beta_{y.x_1} \]

\[ \beta_{y.x_2} \]

\[ r_{x_1 x_2} \]
Multiple Regression: decomposing correlations

\[ Y = \beta_{y,x_1} X_1 + \beta_{y,x_2} X_2 + \epsilon \]

\[ r_{xy} = \rho_{X_1Y} \rho_{X_2Y} \]

\[ X \rightarrow Y \]

\[ X_1 \rightarrow Y \]

\[ X_2 \rightarrow Y \]

\[ X_1 \rightarrow X_2 \]

\[ X_2 \rightarrow X_1 \]

\[ \epsilon \rightarrow Y \]
Multiple Regression: decomposing correlations

\[ r_{x_1y} = \beta_{y.x_1} + r_{x_1x_2} \beta_{y.x_2} \]

\[ r_{x_2y} = \beta_{y.x_2} + r_{x_1x_2} \beta_{y.x_1} \]
Multiple Regression: decomposing correlations

\[ r_{x_1y} = \beta_{y.x_1} + r_{x_1x_2}\beta_{y.x_2} \]

\[ r_{x_2y} = \beta_{y.x_2} + r_{x_1x_2}\beta_{y.x_1} \]

\[ \beta_{y.x_1} = \frac{r_{x_1y} - r_{x_1x_2}r_{x_2y}}{1 - r_{x_1x_2}^2} \]

\[ \beta_{y.x_2} = \frac{r_{x_2y} - r_{x_1x_2}r_{x_1y}}{1 - r_{x_1x_2}^2} \]
Multiple Regression: decomposing correlations

\[ r_{x_1y} = \beta_{y.x_1} + r_{x_1x_2}\beta_{y.x_2} \]
\[ r_{x_2y} = \beta_{y.x_2} + r_{x_1x_2}\beta_{y.x_1} \]

\[ \beta_{y.x_1} = \frac{r_{x_1y} - r_{x_1x_2}r_{x_2y}}{1 - r_{x_1x_2}^2} \]
\[ \beta_{y.x_2} = \frac{r_{x_2y} - r_{x_1x_2}r_{x_1y}}{1 - r_{x_1x_2}^2} \]

\[ R^2 = r_{x_1y}\beta_{y.x_1} + r_{x_2y}\beta_{y.x_2} \]
What happens with 3 predictors? The correlations

\[ r_{x_1y} \]

\[ r_{x_1x_2} \]

\[ r_{x_1x_3} \]

\[ r_{x_2y} \]

\[ r_{x_2x_3} \]

\[ r_{x_3y} \]
What happens with 3 predictors? $\beta$ weights

$X_1$ $\beta_{y.x_1}$ $X_2$ $\beta_{y.x_2}$ $X_3$ $\beta_{y.x_3}$

$r_{x1x2}$ $r_{x1x3}$ $r_{x2x3}$

$X$ $Y$ $\epsilon$

$X_1$ $X_2$ $X_3$
What happens with 3 predictors?

\[
\begin{align*}
    r_{x_1,y} &= \beta_{y,x_1} + r_{x_1x_2}\beta_{y,x_1} + r_{x_1x_3}\beta_{y,x_3} \\
    r_{x_2,y} &= \ldots \\
    r_{x_3,y} &= \ldots \\
\end{align*}
\]

The math gets tedious.
Multiple regression and linear algebra

- Multiple regression requires solving multiple, simultaneous equations to estimate the direct and indirect effects.
  - Each equation is expressed as a $r_{xi,y}$ in terms of direct and indirect effects.
  - Direct effect is $\beta_{y,xi}$
  - Indirect effect is $\sum_{j \neq i} beta_{y,xj} r_{xj,y}$
- How to solve these equations?
- Tediously, or just use linear algebra.
3 special cases of regression

Orthogonal predictors

Correlated predictors

Suppressive predictors
Three basic cases

- **Independent**: $X_1$, $X_2$, $Y$
- **Correlated**: $X_1$, $X_2$, $Y$
- **Suppressor**: $X_2$, $X_1$, $Y$
3 special cases of regression

Orthogonal predictors

Correlated predictors

Suppressive predictors

\[
\beta_{y.x_1} = \frac{r_{x_1y} - r_{x_1x_2}r_{x_2y}}{1 - r^2_{x_1x_2}}
\]

\[
\beta_{y.x_2} = \frac{r_{x_2y} - r_{x_1x_2}r_{x_1y}}{1 - r^2_{x_1x_2}}
\]

\[
R^2 = r_{x_1y}\beta_{y.x_1} + r_{x_2y}\beta_{y.x_2}
\]
Three basic cases: Theoretical examples

Independent

Correlated

Suppressor
Find the regression of rated Prelim score on GREV

```r
> mod1 <- lm(GPA~GREV, data=mydata)
> summary(mod1)

Call:
  lm(formula = GPA ~ GREV, data = mydata)

Residuals:
     Min      1Q  Median      3Q     Max
-1.45807 -0.32322  0.00107  0.32811  1.44850

Coefficients:
                         Estimate Std. Error  t value Pr(>|t|)
(Intercept)            3.011729   0.069434   43.38  <2e-16 ***
GREV                   0.001984   0.000136   14.60  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4558 on 998 degrees of freedom
Multiple R-squared: 0.176, Adjusted R-squared: 0.1751
F-statistic: 213.1 on 1 and 998 DF,  p-value: < 2.2e-16
```
Regression on z transformed data

> mod2 <- lm(GPA~GREV, data=z.data)
> summary(mod2)

Call:
  lm(formula = GPA ~ GREV, data = z.data)

Residuals:
    Min     1Q Median     3Q    Max
-2.90526 -0.64404  0.00213  0.65377  2.88619

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.888e-17  2.872e-02   0.00       1
      GREV  4.195e-01  2.873e-02  14.60  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9082 on 998 degrees of freedom
Multiple R-squared: 0.176,   Adjusted R-squared: 0.1751
F-statistic: 213.1 on 1 and 998 DF,  p-value: < 2.2e-16

Note that the slope is the same as the correlation.
> mod3 <- lm(GPA~GREV, data=cent)
> summary(mod3)

Call:
lm(formula = GPA ~ GREV, data = cent)

Residuals:
               Min         1Q       Median         3Q        Max
-1.45807  -0.32322  0.00107  0.32811  1.44850

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.332e-17  1.441e-02   0.00     1
GREV        1.984e-03  1.359e-04  14.60  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4558 on 998 degrees of freedom
Multiple R-squared: 0.176, Adjusted R-squared: 0.1751
F-statistic: 213.1 on 1 and 998 DF,  p-value: < 2.2e-16

Note that the slope of the centered data is in the same units as the raw data, just the intercept has changed.
Multiple Regression: decomposing correlations

\[ r_{X_1Y} = \beta_{Y.X_1} + r_{X_1X_2}\beta_{Y.X_2} \]

\[ r_{X_2Y} = \beta_{Y.X_2} + r_{X_1X_2}\beta_{Y.X_1} \]

\[ \beta_{Y.X_1} = \frac{r_{X_1Y} - r_{X_1X_2}r_{X_2Y}}{1 - r_{X_1X_2}^2} \]

\[ \beta_{Y.X_2} = \frac{r_{X_2Y} - r_{X_1X_2}r_{X_1Y}}{1 - r_{X_1X_2}^2} \]

\[ R^2 = r_{X_1Y}\beta_{Y.X_1} + r_{X_2Y}\beta_{Y.X_2} \]
> summary(lm(GPA ~ GREV + GREQ , data= cent))

Call:
lm(formula = GPA ~ GREV + GREQ, data = cent)

Residuals:

            Min      1Q  Median      3Q     Max
-1.42442 -0.33228  0.00616  0.32465  1.43765

Coefficients:

                 Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.651e-17  1.435e-02  0.000 1.00000
GREV          1.534e-03  1.976e-04  7.760 2.10e-14 ***
GREQ          6.314e-04  2.019e-04  3.127 0.00182 **

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4538 on 997 degrees of freedom
Multiple R-squared: 0.184, Adjusted R-squared: 0.1823
F-statistic: 112.4 on 2 and 997 DF,  p-value: < 2.2e-16
Do the same regression, but on the z transformed data. The units are now in correlation units.

```
> z.data <- data.frame(scale(my.data))
> summary(lm(GPA ~ GREV + GREQ, data= z.data))
```

Call:
```
lm(formula = GPA ~ GREV + GREQ, data = z.data)
```

Residuals:
```
   Min     1Q  Median     3Q    Max
-2.8382 -0.6621  0.0123  0.6469  2.8646
```

Coefficients:
```
                Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 3.205e-17   2.860e-02   0.000    1.000
GREV         3.242e-01   4.179e-02   7.760   2.10e-14 ***
GREQ         1.306e-01   4.179e-02   3.127    0.00182 **
```

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9043 on 997 degrees of freedom
Multiple R-squared: 0.184, Adjusted R-squared: 0.1823
F-statistic: 112.4 on 2 and 997 DF,  p-value: < 2.2e-16
The 3 correlations produce the beta weights

```r
> R.small <- cor(my.data[c(2,3,8)])
> round(R.small,2)

<table>
<thead>
<tr>
<th></th>
<th>GREV</th>
<th>GREQ</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>GREV</td>
<td>1.00</td>
<td>0.73</td>
<td>0.42</td>
</tr>
<tr>
<td>GREQ</td>
<td>0.73</td>
<td>1.00</td>
<td>0.37</td>
</tr>
<tr>
<td>GPA</td>
<td>0.42</td>
<td>0.37</td>
<td>1.00</td>
</tr>
</tbody>
</table>

> solve(R.small[1:2,1:2])

<table>
<thead>
<tr>
<th></th>
<th>GREV</th>
<th>GREQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>GREV</td>
<td>2.133</td>
<td>-1.555</td>
</tr>
<tr>
<td>GREQ</td>
<td>-1.555</td>
<td>2.133</td>
</tr>
</tbody>
</table>

> beta <- solve(R.small[1:2,1:2], R.small[3,1:2])

<table>
<thead>
<tr>
<th></th>
<th>GREV</th>
<th>GREQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>GREV</td>
<td>0.324</td>
<td>0.131</td>
</tr>
</tbody>
</table>

> beta.1 <- (.42 - .73*.37)/(1-.73^2)
> beta.1

[1] 0.321

> beta.2 <- (.37 - .73 *.42)/(1-.73^2)
> beta.2

[1] 0.136
```

- Find the correlation matrix
- Display it to two decimals
- Find the inverse of GREV and GREQ correlations
- Show them
- Find the beta weights by solving the matrix equation
  - show them
- Find the beta weights by using the formula
  - Show them
3 predictors, no interactions

Use three predictors, but print it with only 2 decimals

```r
> print(summary(lm(GPA ~ GREV + GREQ + GREA, data= cent)), digits=3)
```

Call:

```
lm(formula = GPA ~ GREV + GREQ + GREA, data = cent)
```

Residuals:

```
  Min       1Q   Median       3Q      Max
-1.2668  -0.3038   0.0073   0.3051   1.3022
```

Coefficients:

```
                              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)                 -6.89e-17 1.35e-02     0.00        1.00000
GREV                        6.66e-04 2.00e-04    3.32        0.00092 ***
GREQ                        7.75e-05 1.96e-04    0.40        0.69233
GREA                        2.08e-03 1.81e-04   11.52 < 2e-16 ***
```

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.427 on 996 degrees of freedom
Multiple R-squared: 0.28, Adjusted R-squared: 0.278
F-statistic: 129 on 3 and 996 DF,  p-value: <2e-16
3 predictors, no interactions

Use three predictors, but just the middle 200 subjects

```r
> mod4 <- lm(GPA ~ GREV + GREQ + GREA , data= cent[400:600,])
> summary(mod4)
```

Call:

```r
lm(formula = GPA ~ GREV + GREQ + GREA, data = cent[400:600, ])
```

Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.03553</td>
<td>-0.30799</td>
<td>-0.00889</td>
<td>0.29320</td>
<td>1.20228</td>
</tr>
</tbody>
</table>

Coefficients:

|                  | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------|----------|------------|---------|---------|
| (Intercept)      | 0.0397399| 0.0310412  | 1.280   | 0.202   |
| GREV             | 0.0004706| 0.0004530  | 1.039   | 0.300   |
| GREQ             | 0.0005236| 0.0004515  | 1.160   | 0.248   |
| GREA             | 0.0017904| 0.0004360  | 4.107   | 5.88e-05*** |

```

Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1  1

Residual standard error: 0.4394 on 197 degrees of freedom
Multiple R-squared: 0.2259, Adjusted R-squared: 0.2141
F-statistic: 19.16 on 3 and 197 DF,  p-value: 6.051e-11
Interaction terms are just products in regression

- To interpret all effects, the data need to be 0 centered.
  - This makes the main effects orthogonal to the interaction term.
  - Otherwise, need to compare model with and without interactions

- Graph the results in non-standardized form

- Consider a real data set of SAT V, SAT Q and Gender

```r
> data(sat.act)
> colors=c("black","red") #choose some nice colors
> symb=c(19,25)
> colors=c("black","red") #choose some nice colors
> with(sat.act,plot(SATQ~SATV,pch=symb[gender], col=colors[gender],
                      bg=colors[gender],cex=.6,main="SATQ varies by SATV and gender"))
> by(sat.act,sat.act$gender,function(x)
    abline(lm(SATQ~SATV,data=x)))
```
Plotting interactions and regressions

An example of an interaction plot

satQ varies by satV and gender

> data(sat.act)
> c.sat <- data.frame(scale(sat.act,scale=FALSE))
> summary(lm(satQ~satV * gender,data=c.sat))

Call:
  lm(formula = satQ ~ satV * gender, data = c.sat)

Residuals:
    Min      1Q  Median      3Q     Max
-294.423  -49.876   5.577  53.210  291.100

Coefficients:
                      Estimate Std. Error   t value     Pr(>|t|)
(Intercept)          -0.26696   3.31211  -0.0810    0.93596
satV                 0.65398    0.02926  22.3503  < 2.2e-16 ***
gender               -36.71820   6.91495  -5.3102  1.48e-07 ***
satV:gender          -0.05835    0.06086  -0.9591    0.33810

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 0.1 1

Residual standard error: 86.79 on 683 degrees of freedom
(13 observations deleted due to missingness)
Multiple R-squared: 0.4391, Adjusted R-squared: 0.4367
F-statistic: 178.3 on 3 and 683 DF,  p-value: < 2.2e-16
Interaction of Anxiety with Verbal

```r
> mod5 <- lm(GPA ~ GREV * Anx, data=cent)
> summary(mod5)
```

Call:
`lm(formula = GPA ~ GREV * Anx, data = cent)`

Residuals:
```
               Min        1Q      Median        3Q        Max
-1.49677 -0.31527 -0.00054  0.31223  1.32156
```

Coefficients:
```
                Estimate  Std. Error  t value  Pr(>|t|)  
(Intercept) -2.375e-04   1.395e-02  -0.017  0.986
  GREV       1.996e-03   1.316e-04  15.167 < 2e-16 ***
       Anx    -1.131e-02   1.414e-03  -7.997 3.51e-15 ***
GREV:Anx     2.219e-05   1.377e-05   1.612 0.107
```

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.4412 on 996 degrees of freedom
Multiple R-squared: 0.2294, Adjusted R-squared: 0.227
F-statistic: 98.81 on 3 and 996 DF,  p-value: < 2.2e-16
## Testing for the significance of correlations

```r
> corr.test(sat.act)
```

Call: `corr.test(x = sat.act)`

Correlation matrix

<table>
<thead>
<tr>
<th></th>
<th>gender</th>
<th>education</th>
<th>age</th>
<th>ACT</th>
<th>SATV</th>
<th>SATQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>1.00</td>
<td>0.09</td>
<td>-0.02</td>
<td>-0.04</td>
<td>-0.02</td>
<td>-0.17</td>
</tr>
<tr>
<td>education</td>
<td>0.09</td>
<td>1.00</td>
<td>0.55</td>
<td>0.15</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>age</td>
<td>-0.02</td>
<td>0.55</td>
<td>1.00</td>
<td>0.11</td>
<td>-0.04</td>
<td>-0.03</td>
</tr>
<tr>
<td>ACT</td>
<td>-0.04</td>
<td>0.15</td>
<td>0.11</td>
<td>1.00</td>
<td>0.56</td>
<td>0.59</td>
</tr>
<tr>
<td>SATV</td>
<td>-0.02</td>
<td>0.05</td>
<td>-0.04</td>
<td>0.56</td>
<td>1.00</td>
<td>0.64</td>
</tr>
<tr>
<td>SATQ</td>
<td>-0.17</td>
<td>0.03</td>
<td>-0.03</td>
<td>0.59</td>
<td>0.64</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Sample Size

<table>
<thead>
<tr>
<th></th>
<th>gender</th>
<th>education</th>
<th>age</th>
<th>ACT</th>
<th>SATV</th>
<th>SATQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>687</td>
</tr>
<tr>
<td>education</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>687</td>
</tr>
<tr>
<td>age</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>687</td>
</tr>
<tr>
<td>ACT</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>687</td>
</tr>
<tr>
<td>SATV</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>700</td>
<td>687</td>
</tr>
<tr>
<td>SATQ</td>
<td>687</td>
<td>687</td>
<td>687</td>
<td>687</td>
<td>687</td>
<td>687</td>
</tr>
</tbody>
</table>

Probability values (Entries above the diagonal are adjusted for multiple tests.)

<table>
<thead>
<tr>
<th></th>
<th>gender</th>
<th>education</th>
<th>age</th>
<th>ACT</th>
<th>SATV</th>
<th>SATQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>0.00</td>
<td>0.17</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0</td>
</tr>
<tr>
<td>education</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1</td>
</tr>
</tbody>
</table>
Various tests of significance

1. Is the correlation different from 0? cor.test, corr.test (for more than two variables)
2. Does a correlation differ from another correlation, r.test with or without a third variable.
3. Does a correlation matrix differ from an Identity matrix? cortest
4. Bootstrapping confidence intervals for correlations cor.ci
The correlation coefficient

1. Perhaps the most powerful and useful statistic ever developed
2. Special cases of the correlation are used throughout statistics.
3. The basic concepts of correlation are very straightforward
4. Many ways to be misled with correlations.