Math 1710 Class 20

Association, Correlation, Regression
Dr. Back

Oct. 14, 2009
Son’s Heights from Their Fathers

Galton’s Original 1886 Data

Son’s Heights vs Father Heights
Son’s Heights from Their Fathers

If you know a father’s height, what can you say about his son’s?
If you know a father’s height, what can you say about his son’s?

<table>
<thead>
<tr>
<th></th>
<th>hts</th>
<th>V2</th>
<th>V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>65.04851</td>
<td>59.77827</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>63.25094</td>
<td>63.21404</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>64.95532</td>
<td>63.34242</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>65.75250</td>
<td>62.79238</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>61.13723</td>
<td>64.28113</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>63.02254</td>
<td>64.24221</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>65.37053</td>
<td>64.08231</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>64.72398</td>
<td>63.99574</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>66.06509</td>
<td>64.61338</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>66.96738</td>
<td>63.97944</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Son’s Heights from Their Fathers

If you know a father’s height, what can you say about his son’s?

> summary(hts$V3)  

(Sons)

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1stQu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rdQu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sons</td>
<td>58.51</td>
<td>66.93</td>
<td>68.62</td>
<td>68.68</td>
<td>70.47</td>
<td>78.36</td>
</tr>
</tbody>
</table>

> summary(hts$V2)  

(Fathers)

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1stQu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rdQu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fathers</td>
<td>59.01</td>
<td>65.79</td>
<td>67.77</td>
<td>67.69</td>
<td>69.60</td>
<td>75.43</td>
</tr>
</tbody>
</table>
Son’s Heights from Their Fathers

If you know a father’s height, what can you say about his son’s?

> summary(hts$V3)  (Sons)

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1stQu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rdQu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sons</td>
<td>58.51</td>
<td>66.93</td>
<td>68.62</td>
<td>68.68</td>
<td>70.47</td>
<td>78.36</td>
</tr>
</tbody>
</table>

> summary(hts$V2)  (Fathers)

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1stQu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rdQu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fathers</td>
<td>59.01</td>
<td>65.79</td>
<td>67.77</td>
<td>67.69</td>
<td>69.60</td>
<td>75.43</td>
</tr>
</tbody>
</table>

> sd(hts$V3)
[1] 2.814702

> sd(hts$V2)
[1] 2.744868
Son’s Heights from Their Fathers

If you know a father’s height, what can you say about his son’s?

```r
> summary(hts$V3)  # (Sons)

  Min.  1stQu.  Median    Mean  3rdQu.   Max.  
  58.51  66.93  68.62  68.68    70.47  78.36

> summary(hts$V2)  # (Fathers)

  Min.  1stQu.  Median    Mean  3rdQu.   Max.  
  59.01  65.79  67.77  67.69    69.60  75.43

> cor(hts$V3, hts$V2)
[1] 0.5013383
```
Son’s Heights from Their Fathers

Galton Data with Line of Regression

Diagram showing a scatter plot with a line of regression for Son's Heights and Father Heights.
Son’s Heights from Their Fathers

Suppose $y = \text{height}_{\text{son}}$ were linearly related to $x = \text{height}_{\text{father}}$ by $y = \beta_1 x + \beta_0$. 
Son’s Heights from Their Fathers

Suppose $y = \text{height}_{\text{son}}$ were linearly related to $x = \text{height}_{\text{father}}$ by $y = \beta_1 x + \beta_0$. We’d then have $\bar{y} = \beta_1 \bar{x} + \beta_0$ as well and

$$(y - \bar{y}) = \beta_1 (x - \bar{x}).$$
Suppose \( y = \text{height}_{\text{son}} \) were linearly related to \( x = \text{height}_{\text{father}} \) by \( y = \beta_1 x + \beta_0 \).

We’d then have \( \bar{y} = \beta_1 \bar{x} + \beta_0 \) as well and

\[
(y - \bar{y}) = \beta_1 (x - \bar{x}).
\]

Furthermore, the slope \( \beta_1 \) would be the ratio of standard deviations:

\[
\beta_1 = \frac{s_{\text{son}}}{s_{\text{father}}}
\]
Son’s Heights from Their Fathers

Galton Data with line of slope $\frac{S_{\text{Son}}}{S_{\text{Father}}}$ added in blue.
Son’s Heights from Their Fathers

The fact that the red “best fitting line” (termed the line of regression) actually has less slope than the blue line is one form of Galton’s “regression to the mean.”

Tall fathers tend to have tall sons, but the sons are typically not as extreme in their tallness as their fathers were.
Son’s Heights from Their Fathers

The fact that the red “best fitting line” (termed the line of regression) actually has less slope than the blue line is one form of Galton’s

“regression to the mean.”

Tall fathers tend to have tall sons, but the sons are typically not as extreme in their tallness as their fathers were. In modern terms, we see two usages of the word “regression” here.
Son’s Heights from Their Fathers

Galton Data with Other Line of Regression in pink:
(predict father’s height from son’s)
Association

Given: paired data \((x_1, y_1), \ldots, (x_n, y_n)\)
Given: paired data \((x_1, y_1), \ldots, (x_n, y_n)\)

Scatterplot: Plot \(x\) horizontally, \(y\) vertically.
Association

Given: paired data \((x_1, y_1), \ldots, (x_n, y_n)\)
If one variable is potentially *explanatory* for the *response* of the other, choose the explanatory variable as \(x\).
Given: paired data \((x_1, y_1), \ldots, (x_n, y_n)\)

Association is what we care about.
If we know the \(x\) value of a point, does it tell us something about the likely \(y\) value?
Given: paired data \((x_1, y_1), \ldots, (x_n, y_n)\)
Association is what we care about.
If we know the \(x\) value of a point, does it tell us something about the likely \(y\) value?
Correlation (a number) and regression (a line) are just techniques to study association.
Association

Given: paired data \((x_1, y_1), \ldots, (x_n, y_n)\)

Principal Aspects of Association:

- **Direction:**

- **Strength:**

- **Form:**
Association

Given: paired data \((x_1, y_1), \ldots, (x_n, y_n)\)

Principal Aspects of Association:

- **Direction**: positive or negative
- **Strength**: 
- **Form**: 
Association

Given: paired data \((x_1, y_1), \ldots, (x_n, y_n)\)

Principal Aspects of Association:

- **Direction**: positive or negative
- **Strength**:
- **Form**:

Negative means as \(x\) increases, \(y\) generally decreases.
Association

Given: paired data \((x_1, y_1), \ldots, (x_n, y_n)\)
Principal Aspects of Association:

- **Direction:**
- **Strength:** strong, moderate, or weak
- **Form:**
Given: paired data \((x_1, y_1), \ldots, (x_n, y_n)\)

Principal Aspects of Association:

- **Direction:**
- **Strength:**
- **Form:** linear, curved, or clustered
Association Examples

Example b

![Graph showing Exam 1 and Exam 2 scores with points scattered across the chart.](image)
Association Examples

Example b

My call:

- **Direction**: positive
- **Strength**: moderate
- **Form**: curved
Association Examples

Example b using Data Desk

Pearson Product-Moment Correlation

No Selector

Exam2  Exam1
Exam2  1.000
Exam1  0.496  1.000

Exam2
75.0
67.5
60.0
52.5

Exam1
60.0  67.5  75.0  82.5
Example c
Association Examples

Example c

My call:

- **Direction**: negative
- **Strength**: strong
- **Form**: linear
Association Examples

Example d
Association Examples

Example d

My call:

- **Direction**: negative
- **Strength**: moderate
- **Form**: (perhaps some outliers)
Association Examples

Example d using Data Desk

Pearson Product-Moment Correlation
No Selector
18 total cases of which 3 are missing

<table>
<thead>
<tr>
<th></th>
<th>Exam2d</th>
<th>Exam1D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam2d</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Exam1D</td>
<td>-0.694</td>
<td>1.000</td>
</tr>
</tbody>
</table>

![Graph showing the relationship between Exam2d and Exam1D](image_url)
Association Examples

Example e

![Graph of Exam 2 E vs Exam 1 E](image-url)
Association Examples

Example e

My call:
- **Direction**: positive
- **Strength**: strong
- **Form**: linear
Association Examples

Example f
Association Examples

Example

My call:

- Direction: negative
- Strength: weak
- Form: curved, maybe an outlier
Correlation Properties

\[ r = \frac{1}{n - 1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]
Correlation Properties

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

Positive association means as \( x \) increases, so does \( y \)
(And similarly when \( x \) decreases.)
Correlation Properties

$$r = \frac{1}{n - 1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

Positive association means as $x$ increases, so does $y$ (And similarly when $x$ decreases.)
So for positive association, most terms in the sum are either $(+)(+)$ or $(-)(-)$.
Correlation Properties

$$r = \frac{1}{n - 1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

Positive association means as $x$ increases, so does $y$ (And similarly when $x$ decreases.)
So for pos. association, most terms in the sum are either $(+) \cdot (+)$ or $(-) \cdot (-)$.
Thus with pos. association, $r$ tends to be positive.
Correlation Properties

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

pos. \( r \) \( \longleftrightarrow \) pos. association
neg. \( r \) \( \longleftrightarrow \) neg. association
Correlation Properties

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

\(-1 \leq r \leq 1 \quad (= \pm 1 \text{ only for perfect linear association})\)
Correlation Properties

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

\[-1 \leq r \leq 1 \quad (= \pm 1 \text{ only for perfect linear association})\]
Correlation Properties

\[ r = \frac{1}{n - 1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

\(-1 \leq r \leq 1 \) (= \( \pm 1 \) only for perfect linear association)

To see \( = \pm 1 \) for perfect linear association

\( y = \beta_1 x + \beta_0 \) means \( s_y = \beta_1 s_x \) and \( \bar{y} = \beta_1 \bar{x} + \beta_0 \).
Correlation Properties

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

To see \( = \pm 1 \) for perfect linear association

\( y = \beta_1 x + \beta_0 \) means \( s_y = \beta_1 s_x \) and \( \bar{y} = \beta_1 \bar{x} + \beta_0 \).

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

\[ = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \beta_1 \left( \frac{x_i - \bar{x}}{|\beta_1| s_x} \right) \]

\[ = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right)^2 \frac{\beta_1}{|\beta_1|} \]

\[ = \frac{\beta_1}{|\beta_1|} \]

where the last line used the definition of the variance.
Correlation Properties

\[
    r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)
\]

If you’ve studied vectors, the fact \(-1 \leq r \leq 1\) comes from the same mathematics which explains why

\[
    \cos \theta = \frac{\vec{v} \cdot \vec{w}}{||\vec{v}|| ||\vec{w}||}
\]

has right hand side between \(-1\) and \(+1\).
Correlation Properties

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

*r* is unchanged if *x* and *y* are exchanged.
Correlation Properties

\[
r = \frac{1}{n - 1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)
\]

invariant under rescaling
Correlation Properties

\[ r = \frac{1}{n - 1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

Curved association and \( r=0 \) are consistent!
Correlation Properties

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

Curved association and \( r=0 \) are consistent!
Five points along \( y = x^2 \). (\( \bar{x} = 0 \) and \( \bar{y} = 2 \).)

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>(x-0)</th>
<th>(y-2)</th>
<th>(x-0)(y-2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>-2</td>
<td>4</td>
<td>-2</td>
<td>2</td>
<td>-4</td>
</tr>
</tbody>
</table>
Correlation Properties

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

Curved association and \( r=0 \) are consistent!
Five points along \( y = x^2 \). (\( \bar{x} = 0 \) and \( \bar{y} = 2 \).)

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>(x-0)</th>
<th>(y-2)</th>
<th>(x-0)(y-2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>-2</td>
<td>4</td>
<td>-2</td>
<td>2</td>
<td>-4</td>
</tr>
</tbody>
</table>

\( r \) is exactly zero even though the association is very strong.
Correlation Properties

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

\( r \) is strongly affected by outliers.
Correlation Properties

\[ r = \frac{1}{n - 1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

samples from independent RV's \( \Rightarrow \) \( r \sim 0 \)
Correlation Properties

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

X, Y indep std normal RV's; set \( Y^* = \rho X + \sqrt{1 - \rho^2} Y \) Then (X, Y*) will tend to generate data with \( r \sim \rho \). (e.g. \( \rho = 0.99 \)

\[ \Rightarrow \frac{\sqrt{1-\rho^2}}{\rho} = 0.14 \)
What does Best Fitting Mean?

Given any point \((x_i, y_i)\)
What does Best Fitting Mean?

Given any point \((x_i, y_i)\)

and any line \(y = c_0 + c_1 x\)
What does Best Fitting Mean?

Given any point \((x_i, y_i)\)
and any line \(y = c_0 + c_1x\)
we can use the line to get a predicted value

\[ \hat{y}_i = c_0 + c_1x_i \]
Given any point \((x_i, y_i)\) and any line \(y = c_0 + c_1 x\), we can use the line to get a predicted value
\[
\hat{y}_i = c_0 + c_1 x_i;
\]
and define the residual \(d_i\) by
\[
d_i = y_i - \hat{y}_i.
\]
What does Best Fitting Mean?

Best fitting means the line minimizing the sum $\sum d_i^2$ of the squares of the vertical distances.
What does Best Fitting Mean?

Best fitting means the line minimizing the sum $\Sigma d_i^2$ of the squares of the vertical distances.

Answer: We’ll show the best fitting line $\hat{y} = b_1 x + b_0$ is given by

$$b_1 = r \left( \frac{s_y}{s_x} \right)$$
Best fitting means the line minimizing the sum $\sum d_i^2$ of the squares of the vertical distances.

Answer: We’ll show the best fitting line $\hat{y} = b_1 x + b_0$ is given by

$$b_1 = r \left( \frac{s_y}{s_x} \right)$$

i.e.: **Slope is $r$ in standard deviation units.**

And

$$b_0 = \bar{y} - b_1 \bar{x}$$

i.e.: **The point $(\bar{x}, \bar{y})$ lies on the line.**
What does Best Fitting Mean?

Strictly speaking, the word residual refers to this vertical distance $d_i$ JUST in the case that the line is the “answer line”

$$\hat{y} = b_0 + b_1x.$$