Math 1000
Mathematical Literacy in Today’s World

Lecture 8
Winter 2010
# Back to Course Total vs. Final

<table>
<thead>
<tr>
<th>46</th>
<th>41</th>
<th>99</th>
<th>69</th>
<th>99</th>
<th>63</th>
<th>50</th>
<th>77</th>
<th>81</th>
<th>77</th>
</tr>
</thead>
<tbody>
<tr>
<td>341</td>
<td>282</td>
<td>495</td>
<td>376</td>
<td>478</td>
<td>371</td>
<td>319</td>
<td>350</td>
<td>451</td>
<td>415</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>56</th>
<th>22</th>
<th>38</th>
<th>67</th>
<th>77</th>
<th>18</th>
<th>68</th>
<th>59</th>
<th>68</th>
<th>78</th>
</tr>
</thead>
<tbody>
<tr>
<td>327</td>
<td>233</td>
<td>282</td>
<td>392</td>
<td>403</td>
<td>199</td>
<td>362</td>
<td>336</td>
<td>401</td>
<td>411</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>61</th>
<th>59</th>
<th>79</th>
<th>21</th>
<th>52</th>
<th>51</th>
<th>Final Exam Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>364</td>
<td>304</td>
<td>372</td>
<td>221</td>
<td>280</td>
<td>336</td>
<td>Final Course Totals</td>
</tr>
</tbody>
</table>
Overall Pattern – almost linear
For almost linear relationships

- We say the values are “well correlated”
- And are (in this case) positively associated
- The numerical measure $r$ of correlation will be close to $r = +1$ in this case.
- In fact $r = 0.95$ in this case.
About the Correlation Coefficient $r$

- Measures strength of relationship between variables
- Ignores difference between explanatory and response variables
- Has values between $-1$ and $+1$
- Has no units (not changed by scaling)
- Strongly affected by outliers
Scatterplot: Times 0-60 MPH against Horsepower
Times vs. Horsepower

- Again the values are “well correlated”
- And are (in this case) negatively associated
- The numerical measure $r$ of correlation will be close to $r = -1$ in this case.
- In fact $r = -0.82$ in this case.
Now Times against Weight
Times vs. Weight

- Here the values are “badly correlated”
- And are (maybe in this case) positively associated
- The numerical measure r of correlation will be close to \( r = 0 \) in this case.
- In fact \( r = +0.15 \) in this case, showing weak positive correlation
Supposedly no Correlation

Random Pairs

0 200 400 600 800 1000
0 10000 8000 6000 4000 2000 0
Random Pairs

- Here the values are “badly correlated”
- Actual association is unclear
- The numerical measure r of correlation will be close to $r = 0$ in this case.
- In fact $r = -0.10$ in this case, showing weak negative correlation
Regression gives
\[ y = -(0.62)x + 5288 \]
But $r = -0.10$, so not really a useful relationship.
How is Correlation \( r \) computed?

- Need to compute means \( \bar{X} \), \( \bar{Y} \) and standard deviations \( s_x, s_y \) for both \( X \) and \( Y \) data.
- Then add up all the “scaled deviations” \( \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{s_x s_y} \).
- Then divide by \( n-1 \).
How is Correlation $r$ computed?

- Or:
- Still need to compute means $X_{\text{bar}}$, $Y_{\text{bar}}$ and standard deviations $s_x$, $s_y$ for both $X$ and $Y$ data
- Then add up the $(X_i)(Y_i)$, subtract $n(X_{\text{bar}})(Y_{\text{bar}})$, then divide by $(n-1)(s_x)(s_y)$
How is Correlation $r$ computed?

- Formula in book; or do it by calculator
- We give some idea why it works in what follows
## Back to Course Total vs. Final

<table>
<thead>
<tr>
<th>46</th>
<th>41</th>
<th>99</th>
<th>69</th>
<th>99</th>
<th>63</th>
<th>50</th>
<th>77</th>
<th>81</th>
<th>77</th>
</tr>
</thead>
<tbody>
<tr>
<td>341</td>
<td>282</td>
<td>495</td>
<td>376</td>
<td>478</td>
<td>371</td>
<td>319</td>
<td>350</td>
<td>451</td>
<td>415</td>
</tr>
<tr>
<td>56</td>
<td>22</td>
<td>38</td>
<td>67</td>
<td>77</td>
<td>18</td>
<td>68</td>
<td>59</td>
<td>68</td>
<td>78</td>
</tr>
<tr>
<td>327</td>
<td>233</td>
<td>282</td>
<td>392</td>
<td>403</td>
<td>199</td>
<td>362</td>
<td>336</td>
<td>401</td>
<td>411</td>
</tr>
<tr>
<td>61</td>
<td>59</td>
<td>79</td>
<td>21</td>
<td>52</td>
<td>51</td>
<td>Final Exam Scores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>364</td>
<td>304</td>
<td>372</td>
<td>221</td>
<td>280</td>
<td>336</td>
<td>Final Course Totals</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Find Means for each variable

- $X_{bar} = \frac{(46 + 41 + \ldots + 51)}{26} = 60.6$
- $Y_{bar} = \frac{(341 + 282 + \ldots + 336)}{26} = 350.0$
- The corresponding point $(60.6, 350.0)$ serves as a “pivot” around which possible regression lines can be drawn
Plot \((X_{\text{bar}}, Y_{\text{bar}}) = (60.6, 350.0)\)
Regression Line goes through
Examine “Quadrants”
Two Quadrants count Positive

![Graph showing two quadrants with positive counts.](image-url)
Other Two count Negative
Total for Correlation

- Each data point adds in some positive or negative part in correlation
- More positive parts means that correlation is positive
- More negative parts means that correlation is negative
- Some of each means correlation is close to zero
Scatterplot: Times 0-60 MPH against Horsepower
Xbar = 252.6, Ybar = 6.8
Mostly Negative parts, \( r = -0.82 \)
Back to Random Pairs
Here Xbar = 397, Ybar = 5540
Positives and Negatives almost cancel out, leaving $r = -0.10$
When Correlation is close to +1 or -1

- Relationship between variables is strong
- Data points are close to the regression line
- Predictions are more likely to be useful
When Correlation is close to 0

- Relationship between variables is weak
- Data points are often far from the regression line
- Predictions are less likely to be useful
About the Correlation $r$

- Measures strength and direction of relationship between variables
- Ignores difference between explanatory and response variables
- Has values between $-1$ and $+1$
- Has no units (not changed by scaling)
- Strongly affected by outliers
Samples of the Correlation $r$

- Correlation $r = 0$
- Correlation $r = -0.3$
- Correlation $r = 0.5$
- Correlation $r = -0.7$
- Correlation $r = 0.9$
- Correlation $r = -0.99$
With Outlier Added
With Outlier $y = 2.85x + 170.3$ with $r = 0.82$
Without Outlier, \( y = 3.29x + 150.52 \) and \( r = 0.95 \).
One More Data Set

First-year college GPA versus ACT score

Which is explanatory variable?

<table>
<thead>
<tr>
<th>ACT</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>2.4</td>
</tr>
<tr>
<td>25</td>
<td>3.67</td>
</tr>
<tr>
<td>29</td>
<td>3.1</td>
</tr>
<tr>
<td>23</td>
<td>3.03</td>
</tr>
<tr>
<td>15</td>
<td>2.51</td>
</tr>
<tr>
<td>30</td>
<td>3.07</td>
</tr>
<tr>
<td>23</td>
<td>2.98</td>
</tr>
<tr>
<td>16</td>
<td>2.28</td>
</tr>
<tr>
<td>28</td>
<td>3.93</td>
</tr>
<tr>
<td>24</td>
<td>2.45</td>
</tr>
<tr>
<td>27</td>
<td>1.91</td>
</tr>
<tr>
<td>26</td>
<td>3.83</td>
</tr>
<tr>
<td>23</td>
<td>1.69</td>
</tr>
<tr>
<td>27</td>
<td>3.78</td>
</tr>
<tr>
<td>26</td>
<td>2.88</td>
</tr>
<tr>
<td>21</td>
<td>2.57</td>
</tr>
<tr>
<td>20</td>
<td>2.6</td>
</tr>
<tr>
<td>19</td>
<td>3.42</td>
</tr>
</tbody>
</table>
Scatterplot – Relationship is?
Relationship is weakly positively associated
Regression gives \( y = 0.06x + 1.49 \)
Using $y = 0.06x + 1.49$

- For $x = 0$, $y = 0.06(0) + 1.49$, so point (0, 1.49)
- For $x = 40$, $y = 0.06(40) + 1.49 = 2.4 + 1.49 = 3.89$, so point (40, 3.89)
Plot (0, 1.49) and (40, 3.89)
Estimate for $r$?
Estimate for r?
Actual $r = 0.38$
Actual $r = 0.38$
Book’s Website Applet

- Correlation and Regression
- At [www.whfreeman.com/fapp8e](http://www.whfreeman.com/fapp8e)
- Allows you to see how the correlation coefficient $r$ gives only some information about the strength of the relationship between the variables.
Reminder: Unfortunately

- We say $y = mx + b$
- The book says $y = a + bx$ (so $b$ is slope and $a$ is $y$-intercept)
- The TI-30X(II) calculator says $y = ax + b$ (so $a$ is slope and $b$ is $y$-intercept)
The End for Today