Regression

• Chapter 10 (omit 10.6, 10.7)
• Aplia (week 6 – Regression)

• Supplemental reading on regression
  • excerpt from May, Masson, & Hunter (1990)
Regression

• Using the linear relationship between variables to make predictions
  • predict value of one variable, given a specific score on another variable
    • success in graduate school based on GRE score
    • marital success based on degree of shared political beliefs
Regression

• The concept of prediction – an intuitive approach
  • a single score sampled at random from a distribution
  • no other information available
  • what score would you predict?

• How do we measure accuracy/error of prediction?
Regression

• Best prediction defined as minimum squared deviation from actual score

\[ (X - ?)^2 \]

predicted value

• For a set of scores, best prediction, in absence of any other relevant information, is the mean of the scores

\[ \sum (X - ?)^2 \] is a minimum when \( ? = M \)

• example: 5 scores, consider all possible outcomes

\[
\begin{align*}
X & & M & & (X - M) & & (X - M)^2 \\
2 & & 7 & & (2 - 7) & & 25 \\
5 & & 7 & & (5 - 7) & & 4 \\
8 & & 7 & & (8 - 7) & & 1 \\
8 & & 7 & & (8 - 7) & & 1 \\
12 & & 7 & & (12 - 7) & & 25
\end{align*}
\]

\[ \sum (X - M)^2 = 56 \]

Using the mode:

\[ \sum (X - 8)^2 = 61 \]
Regression

• Improve prediction by using additional information
  • predict worker efficiency
  • prediction for a random applicant with no other information = ?

• now, assume attitude toward company's products and work efficiency are positively related

• for an arbitrarily selected applicant with an attitude score that is above the mean, what is your prediction for this applicant's efficiency, relative to the mean for efficiency?
Regression

• How can we make more precise predictions using the relationship between attitude toward company products and work efficiency?

• ideally, we would have data on subgroups of people, where each subgroup consists of people with the same attitude score

• determine the mean efficiency score for each of these subgroups

• with a new applicant, use his or her attitude score to place him or her in one of the subgroups and take the mean efficiency score for that subgroup as the prediction
Regression

$M_Y = 12.67$
$M_Y = 12.67$

- But data like these are sparse and idiosyncratic
Regression

• Available data treated as a sample from a much larger population

• Assume the relationship in the population is linear and treat available data as a sample that can be used to estimate that linear relationship
Regression

• What does the population of scores look like?
  • bivariate normal distribution

• what would this distribution look like from a top-down view?
Regression

• Bivariate normal distributions
  • what is the correlation between $X$ and $Y$ here?

• What would this distribution look like if the correlation were positive?
Regression

• A 3-D view of subgroups sharing a common $X$ score
Regression

- Definition of a linear relationship
  - subgroup (conditional) means form a straight line
Regression

- Formula for a straight line
  \[ Y = a + bX \]
  \( a \) = intercept (value of \( Y \) when \( X = 0 \))
  \( b \) = slope (change in \( Y \) relative to change in \( X \))

\[
b = \frac{Y_2 - Y_1}{X_2 - X_1}
\]

\[
a = 2
\]

\[
b = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{4 - 3}{8 - 4} = \frac{1}{4} = 0.25
\]
Regression

• Formula for a straight line – consider a negative slope

\[ Y = a + bX \]

\( a = \) intercept (value of \( Y \) when \( X = 0 \))

\( b = \) slope (change in \( Y \) relative to change in \( X \))

\[ b = \frac{Y_2 - Y_1}{X_2 - X_1} \]

\( a = 6 \)

\[ b = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{2 - 4}{8 - 4} \]

\[ = \frac{-2}{4} = -0.50 \]
Regression

• Best fitting line for predicting $Y$ from $X$

$\hat{Y} = a + bX$

• line is defined so that it minimizes squared deviations between predicted and observed values
Regression

• Best fitting line for predicting $Y$ from $X$

$$\hat{Y} = a + bX$$

• line is defined so that it minimizes squared deviations between predicted and observed values

$$b = r \frac{s_Y}{s_X} \quad a = M_Y - bM_X$$

$$\hat{Y} = M_Y - bM_X + bX$$

$$= M_Y + b(X - M_X)$$

$\hat{Y}$ built from $M_Y$ and deviation of $X$ from $M_X$ weighted by slope
Regression

• Best fitting line for predicting $Y$ (efficiency) from $X$ (attitude)

$$\hat{Y} = a + bX$$

$\begin{array}{c|c}
X & Y \\
1 & 8 \\
1 & 12 \\
2 & 6 \\
2 & 12 \\
... \\
7 & 12 \\
7 & 16 \\
\end{array}$
Regression

- Best fitting line for predicting $Y$ (efficiency) from $X$ (attitude)

$$\hat{Y} = a + bX$$

$M_X = 4.06 \quad s_X = 1.98$

$M_Y = 12.67 \quad s_Y = 3.50$

$r = .48$

$$b = r \frac{s_Y}{s_X} = .48 \frac{3.50}{1.98} = 0.85$$

$$a = M_Y - bM_X$$

$$= 12.67 - (.85)4.06$$

$$= 9.22$$
Regression

- Best fitting line for predicting $Y$ (efficiency) from $X$ (attitude)

\[ \hat{Y} = 9.22 + (0.85)X \]

Find two points using regression equation

$X = 0, \quad \hat{Y} = 9.22 + (0.85)0$

\[ = 9.22 \]

$X = 5, \quad \hat{Y} = 9.22 + (0.85)5$

\[ = 13.47 \]
Regression

• Prediction error
  • predictions are not perfectly accurate

• Two applications using prediction error
  • assess individual predictions
  • illustrate an important relation between correlation and regression
Regression

- Prediction error

\[ \hat{Y} = 9.22 + (0.85)X \]

\[ X = 1, \hat{Y} = 9.22 + (0.85)1 \]
\[ = 10.07 \]

\[ X = 2, \hat{Y} = 9.22 + (0.85)2 \]
\[ = 10.92 \]

\[ X = 4, \hat{Y} = 9.22 + (0.85)4 \]
\[ = 12.62 \]
Regression

- Prediction error

\[ \hat{Y} = 9.22 + (0.85)X \]

\[ X = 1, \hat{Y} = 9.22 + (0.85)1 \]
\[ = 10.07 \]

\[ X = 2, \hat{Y} = 9.22 + (0.85)2 \]
\[ = 10.92 \]

\[ X = 4, \hat{Y} = 9.22 + (0.85)4 \]
\[ = 12.62 \]

\[ (Y - \hat{Y})^2 = (12 - 10.07)^2 = 3.72 \]
• Prediction error
\[ \hat{Y} = 9.22 + (0.85)X \]

\[ X = 1, \hat{Y} = 9.22 + (0.85)1 \]
\[ = 10.07 \]

\[ X = 2, \hat{Y} = 9.22 + (0.85)2 \]
\[ = 10.92 \]

\[ X = 4, \hat{Y} = 9.22 + (0.85)4 \]
\[ = 12.62 \]

\[ (Y - \hat{Y})^2 = (6 - 10.92)^2 \]
\[ = 24.21 \]
Regression

• Prediction error

\[ \hat{Y} = 9.22 + (0.85)X \]

\[ X = 1, \; \hat{Y} = 9.22 + (0.85)1 \]
\[ = 10.07 \]

\[ X = 2, \; \hat{Y} = 9.22 + (0.85)2 \]
\[ = 10.92 \]

\[ X = 4, \; \hat{Y} = 9.22 + (0.85)4 \]
\[ = 12.62 \]
• Squared error of prediction for each subject

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Ŷ</th>
<th>Y – Ŷ</th>
<th>(Y – Ŷ)²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>10.07</td>
<td>-2.07</td>
<td>4.28</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>10.07</td>
<td>1.93</td>
<td>3.72</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>10.92</td>
<td>-4.92</td>
<td>24.21</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>10.92</td>
<td>1.08</td>
<td>1.17</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>10.92</td>
<td>4.08</td>
<td>16.65</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>11.77</td>
<td>-2.77</td>
<td>7.67</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>11.77</td>
<td>1.23</td>
<td>1.51</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>12.62</td>
<td>-5.62</td>
<td>31.58</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>12.62</td>
<td>-0.62</td>
<td>0.38</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>12.62</td>
<td>1.38</td>
<td>1.90</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>12.62</td>
<td>6.38</td>
<td>40.70</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>13.47</td>
<td>-0.47</td>
<td>0.22</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>14.32</td>
<td>-2.32</td>
<td>5.38</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>14.32</td>
<td>0.68</td>
<td>0.46</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>14.32</td>
<td>1.68</td>
<td>2.82</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>14.32</td>
<td>2.68</td>
<td>7.18</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>15.17</td>
<td>-3.17</td>
<td>10.05</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>15.17</td>
<td>0.83</td>
<td>0.69</td>
</tr>
</tbody>
</table>

• Prediction errors estimate how much variability there is around conditional means in population

• Assumption: each conditional distribution of Y scores in the population has equal variance
Regression

• Squared error of prediction for each subject

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Ŷ</th>
<th>Y - Ŷ</th>
<th>(Y - Ŷ)^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>10.07</td>
<td>-2.07</td>
<td>4.28</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>10.07</td>
<td>1.93</td>
<td>3.72</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>10.92</td>
<td>-4.92</td>
<td>24.21</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>10.92</td>
<td>1.08</td>
<td>1.17</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>10.92</td>
<td>4.08</td>
<td>16.65</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>11.77</td>
<td>-2.77</td>
<td>7.67</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>11.77</td>
<td>1.23</td>
<td>1.51</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>12.62</td>
<td>-5.62</td>
<td>31.58</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>12.62</td>
<td>-0.62</td>
<td>0.38</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>12.62</td>
<td>1.38</td>
<td>1.90</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>12.62</td>
<td>6.38</td>
<td>40.70</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>13.47</td>
<td>-0.47</td>
<td>0.22</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>14.32</td>
<td>-2.32</td>
<td>5.38</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>14.32</td>
<td>0.68</td>
<td>0.46</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>14.32</td>
<td>1.68</td>
<td>2.82</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>14.32</td>
<td>2.68</td>
<td>7.18</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>15.17</td>
<td>-3.17</td>
<td>10.05</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>15.17</td>
<td>0.83</td>
<td>0.69</td>
</tr>
</tbody>
</table>

• So prediction errors can be averaged together to obtain one estimate of variance instead of estimating variance separately for each conditional distribution

\[ \sum (Y - \hat{Y})^2 = 160.60 \]

• Variance of errors of prediction

\[ s^2_{Y - \hat{Y}} = \frac{\sum (Y - \hat{Y})^2}{N - 2} = \frac{160.60}{18 - 2} = 10.04 \]
Regression

- Squared error of prediction for each subject

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>$\hat{Y}$</th>
<th>$Y - \hat{Y}$</th>
<th>$(Y - \hat{Y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>10.07</td>
<td>-2.07</td>
<td>4.28</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>10.07</td>
<td>1.93</td>
<td>3.72</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>10.92</td>
<td>-4.92</td>
<td>24.21</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>10.92</td>
<td>1.08</td>
<td>1.17</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>10.92</td>
<td>4.08</td>
<td>16.65</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>11.77</td>
<td>-2.77</td>
<td>7.67</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>11.77</td>
<td>1.23</td>
<td>1.51</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>12.62</td>
<td>-5.62</td>
<td>31.58</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>12.62</td>
<td>-0.62</td>
<td>0.38</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>12.62</td>
<td>1.38</td>
<td>1.90</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>12.62</td>
<td>6.38</td>
<td>40.70</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>13.47</td>
<td>-0.47</td>
<td>0.22</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>14.32</td>
<td>-2.32</td>
<td>5.38</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>14.32</td>
<td>0.68</td>
<td>0.46</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>14.32</td>
<td>1.68</td>
<td>2.82</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>14.32</td>
<td>2.68</td>
<td>7.18</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>15.17</td>
<td>-3.17</td>
<td>10.05</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>15.17</td>
<td>0.83</td>
<td>0.69</td>
</tr>
</tbody>
</table>

- Variance of errors of prediction

$$s_{Y-\hat{Y}}^2 = \frac{\sum (Y - \hat{Y})^2}{N - 2} = \frac{160.60}{18 - 2} = 10.04$$

- Standard error of estimate

$$s_{Y-\hat{Y}} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{N - 2}} = \sqrt{\frac{160.60}{18 - 2}} = 3.17$$
Regression

- Population

\[ \hat{Y} = 2 + 2X \]
\[ \rho = .82 \]

Standard deviation of conditional distributions = 2.00

\[ \sigma_Y = 3.46 \]
• Sample \((N = 8)\)

\[
\hat{Y} = 2.81 + 1.89X
\]

\(r = .83\)

\(s_{Y-\hat{Y}} = 2.51\)

\(s_Y = 4.13\)
Regression

• Prediction error

  • *standard error of estimate* is an estimate of variation around $\hat{Y}$ values in the population

- what shape do we assume for the distribution of scores around each $\hat{Y}$ in the population?
Regression

- Prediction error
  - assume a normal distribution
Regression

- Cross-section of an example bivariate normal distribution ($\rho = 0$)

\[ \mu_Y = 11 \]
\[ \sigma_Y = 3 \]
Regression

- Prediction error
  - use these assumptions to make inferences about predicted values

- what proportion of people in the population with an attitude score of 6 have efficiency scores of 17 or higher?

\[ \hat{Y} = 9.22 + (.85)X \]

= 9.22 + (.85)6

= 14.32
Regression

• Prediction error
  • use these assumptions to make inferences about predicted values
• what proportion of people in the population with an attitude score of 6 have efficiency scores of 17 or higher?

\[ \hat{Y} = 9.22 + (0.85)X \]
\[ = 9.22 + (0.85)6 \]
\[ = 14.32 \]

\[ s_{Y-Y} = 3.17 \]

Area in smaller portion = 0.1977

\[ z = \frac{17 - 14.32}{3.17} \]
\[ = 0.85 \]
Regression

\[ \hat{Y} = 9.22 + (0.85)X \]

\[ s_{Y-\hat{Y}} = 3.17 \]

\[ z = \frac{X - \bar{M}}{s} \]

\[ X = \bar{M} + z(s) \]

Among workers with an attitude score of 3, what raw score on efficiency separates the 20% most efficient workers in this group from the rest of the workers in this group?

<table>
<thead>
<tr>
<th>( z )</th>
<th>Mean to ( z )</th>
<th>Larger Portion</th>
<th>Smaller Portion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.80</td>
<td>0.2881</td>
<td>0.7881</td>
<td>0.2119</td>
</tr>
<tr>
<td>0.81</td>
<td>0.2910</td>
<td>0.7910</td>
<td>0.2090</td>
</tr>
<tr>
<td>0.82</td>
<td>0.2939</td>
<td>0.7939</td>
<td>0.2061</td>
</tr>
<tr>
<td>0.83</td>
<td>0.2967</td>
<td>0.7967</td>
<td>0.2033</td>
</tr>
<tr>
<td>0.84</td>
<td>0.2995</td>
<td>0.7995</td>
<td>0.2005</td>
</tr>
<tr>
<td>0.85</td>
<td>0.3023</td>
<td>0.8023</td>
<td>0.1977</td>
</tr>
<tr>
<td>0.86</td>
<td>0.3051</td>
<td>0.8051</td>
<td>0.1949</td>
</tr>
<tr>
<td>0.87</td>
<td>0.3078</td>
<td>0.8078</td>
<td>0.1922</td>
</tr>
<tr>
<td>0.88</td>
<td>0.3106</td>
<td>0.8106</td>
<td>0.1894</td>
</tr>
</tbody>
</table>
Regression

• Prediction error

• consider sum of squared errors of prediction had we used $M_Y$ for each prediction

$$\sum (Y - M_Y)^2 = 208.02$$

• and this can be used to compute variance and standard deviation

$$s_Y^2 = \frac{\sum (Y - M_Y)^2}{N - 1} = \frac{208.02}{17} = 12.24$$

$$s_Y = \sqrt{\frac{\sum (Y - M_Y)^2}{N - 1}} = 3.50$$

$$s_{Y-Y} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{N - 2}} = \sqrt{\frac{160.60}{18 - 2}} = 3.17$$

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>$M_Y$</th>
<th>$Y - M_Y$</th>
<th>$(Y - M_Y)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>12.67</td>
<td>-4.67</td>
<td>21.81</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>12.67</td>
<td>-0.67</td>
<td>0.45</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>12.67</td>
<td>-6.67</td>
<td>44.49</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>12.67</td>
<td>-0.67</td>
<td>0.45</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>12.67</td>
<td>2.33</td>
<td>5.43</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>12.67</td>
<td>-3.67</td>
<td>13.47</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>12.67</td>
<td>0.33</td>
<td>0.11</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>12.67</td>
<td>-5.67</td>
<td>32.15</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>12.67</td>
<td>-0.67</td>
<td>0.45</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>12.67</td>
<td>1.33</td>
<td>1.77</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>12.67</td>
<td>6.33</td>
<td>40.07</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>12.67</td>
<td>0.33</td>
<td>0.11</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>12.67</td>
<td>-0.67</td>
<td>0.45</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>12.67</td>
<td>2.33</td>
<td>5.43</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>12.67</td>
<td>3.33</td>
<td>11.09</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>12.67</td>
<td>4.33</td>
<td>18.75</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>12.67</td>
<td>-0.67</td>
<td>0.45</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>12.67</td>
<td>3.33</td>
<td>11.09</td>
</tr>
</tbody>
</table>
Regression

- Variability explained and not explained
Regression

• Variability explained and not explained
Regression

- Prediction error – proportion of variability in $Y$ that cannot be explained by differences in $X$ scores

\[ Y - M_Y = (Y - \hat{Y}) + (\hat{Y} - M_Y) \]

\[ 15 - 12.67 = 2.33 \]
\[ 15 - 14.32 = 0.68 \]
\[ 14.32 - 12.67 = 1.65 \]
\[ 2.33 = 0.68 + 1.65 \]
Regression

• Proportion of variability not explained and proportion explained

\[
\sum (Y - M_Y)^2 \rightarrow SS_Y \quad \text{(total variability in Y scores)}
\]

\[
\sum (Y - \hat{Y})^2 \rightarrow SS_{error} \quad \text{(unexplained variability in Y scores)}
\]

\[
\sum (\hat{Y} - M_Y)^2 \rightarrow SS_{\hat{Y}} \quad \text{(variability in Y scores associated with differences in X scores)}
\]

\[
SS_Y = SS_{error} + SS_{\hat{Y}}
\]

\[208.02 \approx 160.60 + 48.37\]

Proportion of variability explained = \[
\frac{SS_{\hat{Y}}}{SS_Y} = \frac{48.37}{208.02} = .23
\]
Regression

• Variability in $Y$ scores associated with differences in $X$ scores

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$\hat{Y}$</th>
<th>$\hat{Y} - M_Y$</th>
<th>$(\hat{Y} - M_Y)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>10.07</td>
<td>-2.60</td>
<td>6.76</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>10.07</td>
<td>-2.60</td>
<td>6.76</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>10.92</td>
<td>-1.75</td>
<td>3.06</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>10.92</td>
<td>-1.75</td>
<td>3.06</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>10.92</td>
<td>-1.75</td>
<td>3.06</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>11.77</td>
<td>-0.90</td>
<td>0.81</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>11.77</td>
<td>-0.90</td>
<td>0.81</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>12.62</td>
<td>-0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>12.62</td>
<td>-0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>12.62</td>
<td>-0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>12.62</td>
<td>-0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>13.47</td>
<td>0.80</td>
<td>0.64</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>14.32</td>
<td>1.65</td>
<td>2.72</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>14.32</td>
<td>1.65</td>
<td>2.72</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>14.32</td>
<td>1.65</td>
<td>2.72</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>14.32</td>
<td>1.65</td>
<td>2.72</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>15.17</td>
<td>2.50</td>
<td>6.25</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>15.17</td>
<td>2.50</td>
<td>6.25</td>
</tr>
</tbody>
</table>

$M_Y = 12.67$

$\sum (\hat{Y} - M_Y)^2 = 48.37$
Regression

• Proportion of variability explained – *relation between correlation and regression*

• A way to interpret the size of the correlation coefficient, $r$
  • proportion of variability explained = $r^2$
  • from the attitude/efficiency example:
    
    $r = .48$, $r^2 = .23$

$$\frac{SS_{\hat{Y}}}{SS_Y} = \frac{48.37}{208.02} = .23$$
Regression

• Using $r^2$ to compute prediction error

• standard error of estimate

\[ s_{Y-Y} = \sqrt{ \frac{\sum (Y - \hat{Y})^2}{N - 2}} = \sqrt{ \frac{160.60}{18 - 2}} = 3.17 \]

\[ s_{Y-Y} = s_Y \sqrt{(1 - r^2) \left( \frac{N - 1}{N - 2} \right)} \]

\[ = 3.50 \sqrt{(1 - .477^2) \left( \frac{18 - 1}{18 - 2} \right)} \]

\[ = 3.50(0.906) = 3.17 \]
Regression

• Numerical example of regression, standard error of estimate, and proportion of variability explained

\[
SS_{\text{error}} = \sum (Y - \hat{Y})^2 = 2.196
\]

\[
s_{Y-\hat{Y}} = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N - 2}}
\]

\[
= \sqrt{\frac{2.196}{5 - 2}} = .856
\]

\[
s_{Y-\hat{Y}} = s_Y \sqrt{(1 - r^2) \left( \frac{N - 1}{N - 2} \right)}
\]

\[
= 1.140 \sqrt{(1 - .76^2) \left( \frac{5 - 1}{5 - 2} \right)} = .856
\]

\[
b = r \frac{s_Y}{s_X} = .760(1.140/1.673) = .518
\]

\[
a = M_Y - bM_X = 4.4 - .518(4.6) = 2.017
\]

\[\hat{Y} = 2.017 + (.518)X\]

\[
M_{\hat{Y}} = 4.4 \quad r_{X,Y-\hat{Y}} = 0
\]

\[
M_X = 4.6, \quad s_X = 1.673,
\]

\[
M_Y = 4.4, \quad s_Y = 1.140, \quad r = .760
\]

\[
SS_{\hat{Y}} = \sum (\hat{Y} - M_Y)^2 = 3.005
\]

\[
SS_{\hat{Y}}/SS_Y = \frac{3.005}{5.200} = .578
\]