Chapter 3: Describing Relationships

Section 3-2: Least-Squares Regression
3-2 Line of Best Fit

Let’s go back to the GPA vs ACT data from the sample of 6 students:

GPA (x) 3.0 3.5 3.1 4.0 3.7 3.1  
ACT (y) 19 27 22 31 26 19

Make a scatterplot on the calculator. Then find the line of best fit and show it on the scatterplot.
The equation of the best fit line is: $y = -15.1 + 11.5x$

but we will write is as: $\overline{ACT} = -15.1 + 11.5(GPA)$

What form is this in? Remember Alg. 1

\[ y = -15.1 + 11.5x \]

\[ \overline{ACT} = -15.1 + 11.5(GPA) \]

What is the slope of the regression line? 11.5

Interpret: As the GPA increases by one, we would predict the ACT scores to go up by 11.5.

What is the y-intercept of the regression line? -15.1

Interpret: This is the prediction score for a student with a GPA of 0.
\[ \overline{ACT} = -15.1 + 11.5(GPA) \]

Predict the ACT score of a student with a 3.4 GPA and a student with a 4.5 GPA.

\[ \overline{ACT} = -15.1 + 11.5(3.4) \]
\[ = -15.1 + 39.1 \]
\[ = 24 \]

So, the predicted ACT score for a student with a 3.4 GPA is 24.

\[ \overline{ACT} = -15.1 + 11.5(4.5) \]
\[ = -15.1 + 51.75 \]
\[ = 36.65 \]

So, the predicted ACT score for a student with a 4.5 GPA is 36.65.

Notice the range of the GPA values: 3.0 to 4.0. So, using the regression line to predict the ACT score for a student that lies outside this range is not going to be accurate.

Using a regression line to make a prediction that is outside the interval of values of the explanatory variable x is called **extrapolation** and is often inaccurate.
1) The slope is 40. We predict that a rat will gain 40 grams of weight per week.

2) The y intercept is 100. This suggests that we expect a rat at birth to be 100 grams.

3) We predict the rat’s weight to be 740 grams.

4) The time is measured in weeks for this equation, so 2 years becomes 104 weeks. We then predict the rat’s weight to be 4260 grams, which is equivalent to 9.4 pounds (about the weight of a large newborn human). This is unreasonable and is the result of extrapolation.
<table>
<thead>
<tr>
<th>GPA</th>
<th>3.0</th>
<th>3.5</th>
<th>3.1</th>
<th>4.0</th>
<th>3.7</th>
<th>3.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>19</td>
<td>27</td>
<td>22</td>
<td>31</td>
<td>26</td>
<td>19</td>
</tr>
</tbody>
</table>

Difference between the observed $y$ and predicted $\hat{y}$:

$$y - \hat{y} = -15.1 + 11.5(GPA)$$

Make a prediction for the ACT score of a student with a GPA of 3.5.

$$\hat{ACT} = 25.15$$

What is a residual? **Difference between the observed $y$ and predicted $\hat{y}$**

Are residuals positive or negative? DEPENDS because the observed can be less than or greater than the predicted.

But why is this the best fit line?

The **least squares regression line (LSRL)** of $y$ on $x$ is the line that makes the sum of the squares of the residuals as small as possible.
How does the calculator find the equation of the least squares regression line \( (\hat{y} = a + bx) \)?

slope: \( b = r \cdot \frac{s_y}{s_x} \)

y-intercept: \( a = \bar{y} - b\bar{x} \)

These formulas are on the AP Exam Formula Sheet.

The formula for slope tells us that a change of one standard deviation in \( x \) corresponds to a change of \( r \) standard deviations in \( y \).

\[
b = r \cdot \frac{s_y}{s_x} = \frac{r \cdot s_y}{\bar{x} \cdot s_x}
\]
Now let’s find the least squares regression line for our GPA/ACT data by hand. Then verify that \((\bar{x}, \bar{y})\) is on the LSRL.

<table>
<thead>
<tr>
<th>GPA</th>
<th>3.0</th>
<th>3.5</th>
<th>3.1</th>
<th>4.0</th>
<th>3.7</th>
<th>3.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>19</td>
<td>27</td>
<td>22</td>
<td>31</td>
<td>26</td>
<td>19</td>
</tr>
</tbody>
</table>

\[ r = 0.995 \quad \bar{x} = 3.4 \quad \bar{y} = 24 \]

\[ s_x = 0.4 \quad s_y = 4.81664 \]

Slope: \[ b = r \cdot \frac{s_y}{s_x} = (0.995) \cdot \frac{4.81664}{0.4} = 11.5 \]

\[
\begin{align*}
\hat{y} &= a + bx \\
\overline{\text{ACT}} &= -15.1 + 11.5(GPA)
\end{align*}
\]

\[
\begin{align*}
y\text{-int: } a &= \bar{y} - b\bar{x} = 24 - (11.5)(3.4) = -15.1 \\
\end{align*}
\]

Check: \((\bar{x}, \bar{y}) = (3.4, 24)\)

24 = -15.1 + (11.5)(3.4)
24 = -15.1 + 39.1
24 = 24

“center of gravity” \((\bar{x}, \bar{y})\) always on the LSRL

24 = 24 \checkmark
What does $-0.4$ mean? The LSRL overpredicted the ACT score by 0.4.

What does $1.85$ mean? The LSRL underpredicted the ACT score by 1.85.
Now do all of these calculations in LISTS on the calculator.

\[
PRED = \frac{Y_1(L_1)}{y - \hat{y} = L_2 - LPRED}
\]

Create a list called PRED

Create another list called RES

Now create a residual plot. Sketch it to the right:

Horizontal axis: explanatory variable

Vertical axis: residuals

What to look for on the residual plot:

- If the residual plot is scattered evenly above and below 0, the LSRL is **a good fit**.
- If the residual plot is curved, the LSRL is **a bad approximation**.
- Look for residuals that increase as **x increases**.
- Large residual = **outlier**.
Examples of Residuals Plots

• The line at $y = 0$ marks the sum (and mean) of the residuals.
• This “residual = 0” line corresponds with the regression line.

Figure 3.14 The response variable $y$ has more spread for larger values of the explanatory variable $x$, so prediction will be less accurate when $x$ is large.
Find the average of all the residuals for our GPA/ACT example: **0 (always)**

This is not very helpful. How could we get a good idea for, on average, how far off the LSRL is in predicting the ACT scores?

**Standard deviation of the residuals** $(s) = \sqrt{\frac{\sum resid^2}{n-2}} = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}}$

\[
= \sqrt{\frac{(-0.4)^2 + (1.85)^2 + (1.45)^2 + (0.1)^2 + (-1.45)^2 + (-1.55)^2}{6 - 2}}
\]

\[
= \sqrt{\frac{10.2}{4}}
\]

\[
= 1.60
\]

Using a calculator:

1) Make a list of the residuals
2) Do 1-Var Stats
3) Use $\sum x^2$ value for $\sum resid^2$

Interpret $s$ for our GPA/ACT example: ______________________________________

The average error (residual) in predicting ACT scores is 1.60 using the least-squares regression line (LSRL).
1) The residual plot does not show a random scatter. Describe the pattern you see.

2) For this regression, $s = 2.27$. Interpret this value in context.

1) There is a moderate positive linear relationship with one outlier in the bottom-right corner of the plot.

2) The average error (residual) in predicting the backpack weight is 2.27 lb using the least-squares regression line (LSRL).
In today’s lesson, you will learn new stuff by going through this worksheet. There will be a review of some old ideas and presentation of some new ideas.

**Example:** Mr. Wilcox is extremely interested in predicting his two children’s aptitude later in life. One way to predict later aptitude in children is by knowing the age (in months) when the child spoke their first word (daddy). A recent study of the development of young children recorded the age in months at which each of 21 children spoke their first word and their Gesell Adaptive Score, the result of an aptitude test taken much later in life. Here are the data:

<table>
<thead>
<tr>
<th>Child #</th>
<th>Age</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>95</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>71</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>83</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>91</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>102</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>87</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
<td>93</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>104</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>94</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>113</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Child #</th>
<th>Age</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>9</td>
<td>96</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>83</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
<td>84</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
<td>102</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>17</td>
<td>12</td>
<td>105</td>
</tr>
<tr>
<td>18</td>
<td>42</td>
<td>57</td>
</tr>
<tr>
<td>19</td>
<td>17</td>
<td>121</td>
</tr>
<tr>
<td>20</td>
<td>11</td>
<td>86</td>
</tr>
<tr>
<td>21</td>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>
Here is the output from Minitab – a statistical software $\hat{y} = a + bx$

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>109.87</td>
<td>22.71</td>
<td>-3.63</td>
<td>0.0018</td>
</tr>
<tr>
<td>Age</td>
<td>-1.13</td>
<td>0.04</td>
<td>4.27</td>
<td>0</td>
</tr>
</tbody>
</table>

Always $x$ variable

S = 11.02

R-sq = 41.0%

Don’t use!

R-sq (adj) = 36.4%

1. What is the equation of the least squares regression line? $\text{score} = 109.87 - 1.13(\text{age})$

2. Interpret the slope in the context of the problem. For each increase in one month for age at first word, we predict the Gesell score to drop 1.13 points.

3. $r$ is called the correlation and it tells us the strength of the linear relationship between $x$ and $y$. Find $r$ and then interpret it in context. $r = -0.64$  
   $\sqrt{r^2} = \sqrt{0.41} = -0.64$

   There is a fairly strong negative linear relationship between age at first word and Gesell score. Be careful! $r$ is neg because slope is neg
4. $r^2$ is called the **coefficient of determination**. It tells us “the percent of the variation in $y$ that can be accounted for by the LSRL of $y$ on $x$”. Find $r^2$ and interpret it. $r^2 = 0.410$

The LSRL accounts for 41% of the variation in Gesell score.

About 41% of the variation in $y$ among the individual children is due to the straight-line relationship between $y$ and $x$. The other 59% is individual variation among children that is not explained by the linear relationship. What other factors besides the age of when a child spoke his/her first word might help explain their Gesell score?

The **coefficient of determination** $r^2$ is the fraction of the variation in the values of $y$ that is accounted for by the least-squares regression line of $y$ on $x$. We can calculate $r^2$ using the following formula:

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

where $\text{SSE} = \sum \text{residual}^2$ and $\text{SST} = \sum (y_i - \bar{y})^2$.

- $r^2$ tells us how well the least-squares regression line (LSRL) predicts the values of the response variable $y$.
- If all of the points fall directly on the least-squares line, then $r^2 = 1$. 

$SSE$ – sum of squared errors of predictions

$SST$ – sum of squared errors about the mean
An example of a perfect positive correlation \((r = 1 \rightarrow r^2 = 1)\) is when comparing the number of people who go to see a movie and the total spent money on tickets, when plotted on a graph, it equals to 1. This means that every time a number of people \((x)\) go, an amount of money \((y)\) is spent without variation on the tickets:

The LSRL accounts for 100% of the variation in money spent on tickets. Nothing else explains the amount of money spent on tickets besides the number of people that go.
5. Make a scatterplot of the data. Then make a residual plot. Sketch both below.

6. Predict the Gesell score for a kid who spoke his first word at exactly 1 year. 96.31

7. Find the residual for this prediction. Then interpret.

   \[
   \text{Residual} = \frac{105 - 96.31}{5.41} = 8.69
   \]
   Interpret: The LSRL underpredicted the score by almost 9 points.

8. An outlier is an observation that lies outside the overall pattern of the other observations. There are two possible outliers. Identify which two child numbers. Child 18 and 19
   Circle these two children on the scatterplot and the residual plot.

9. Find the residual for each of the outliers.

   Child 18 has residual \(-5.41\). Child 19 has residual 30.34.

11. An influential observation is one that, if removed, would markedly change the LSRL. Let’s see if either of the two outliers should be considered influential.

Points that are outliers in the $x$ direction of a scatterplot are often influential for the LSRL. A point that is extreme in the $x$ direction with no other points near it pulls the line toward itself.

LSRL for all the data: \[ \text{score} = 109.87 - 1.13(\text{age}) \]

LSRL with only Child 19 removed: \[ \text{score} = 109.3 - 1.19(\text{age}) \]

LSRL with only Child 18 removed: \[ \text{score} = 105.63 - 0.78(\text{age}) \]

Which child is most influential on the LSRL? \underline{18} Why? Child 18 is an outlier in the $x$ direction. Notice it has a small residual.
12. TRUE/FALSE

a. **true** Knowing which variable is explanatory and which is response is important for regression.

b. **false** Correlation and regression lines describe the strength of any scatterplot.

c. **false** Correlation and regression are resistant to outliers.

d. **false** A high correlation means that one variable is causing changes in another variable.

**Association ≠ Causation**
Association = Causation?

• **Association** (correlation) means that two variables are in some way related to each other.

• **Causation** means that a change in one variable cause changes in another variable.
Outdoor temperature and gas consumption

- 1) There is a strong association between the outdoor temperature and the amount of gas consumed by households.
- Therefore, an increase in outdoor temperature causes a decrease in gas consumption.
Ice cream sales and shark attacks

- 2) Ice cream sales and the number of shark attacks on swimmers are strongly correlated.
- Therefore, an increase in ice cream sales causes an increase in shark attacks.
Family income and SAT scores

• 3) The amount of family income and scores on SAT tests are very strongly correlated.
• Therefore, an increase in family income causes an increase in SAT scores.
Vocabulary size and # of cavities

4) The vocabulary size and number of cavities in elementary school children have a strong positive correlation.

Therefore, increasing vocabulary size causes the number of cavities to increase.