CHAPTER 7 — REGRESSION

How can we use the correlation coefficient to estimate the \( y \) value for a given \( x \) value?

If there is a perfect correlation (that is, \( r = 1 \) or \( r = -1 \)) then we can perfectly predict \( y \) from \( x \).

- \( r = 1 \): If \( x \) goes up 1 SD from average, then \( y \) goes up 1 SD from average. All points lie on the SD line.

- \( r = .8 \): If \( x \) increases 1 SD, then \( y \) increases only .8 SD on the average.

- \( r = .5 \): If \( x \) increases 1 SD, then \( y \) increases only .5 SD on the average.

- \( r = -.3 \): If \( x \) increases 1 SD, then \( y \) decreases only .3 SD on the average.

- \( r = 0 \): If \( x \) increases 1 SD, then \( y \) decreases only 0 SD on the average.

Definition: The regression line estimates the average value for the dependent variable (\( y \)) corresponding to each value of the independent variable (\( x \)).

**1 SD** increase in \( x \) means only an \( r \times \text{SD} \) increase in \( y \).

The slope of the SD line is

\[
\frac{\text{SD}_y}{\text{SD}_x} = \frac{1}{r} = 1
\]

(If \( r \) is negative, the slope is negative and the line slopes down.)

The slope of the regression line is

\[
\frac{r \times \text{SD}_y}{\text{SD}_x} = r \left( \frac{1}{r} \right) = r
\]

So the regression line is always less steep than the SD line (unless \( r = \pm 1 \), in which case they are the same line.)
The **Graph of Averages** gives the actual average of \( y \) for each value of \( x \). The **regression line** estimates the average of \( y \) for each value of \( x \).
- If graph of averages forms a line, then it is the regression line.
- If the graph of averages is not a line, then the regression line is a smooth version of it.

Below is the scatter plot for the heights and weights of a group of 494 Stat 100 females. The scatter plot is divided into vertical strips. Within each vertical strip, the girls are the same height but have different weights. The little horizontal lines represent the average weight for girls of the same height. For example, there are 7 girls who are in the 71" strip (the third strip from the right). They range in weight from about 135 lbs. to 220 lbs. The horizontal line within the strip is their average height. The regression line is the line that fits best through the horizontal lines.
Regression Method for Individuals

Example 1: Here's the same example we did in Chapter 8 but this time let's include $r = 0.5$.

<table>
<thead>
<tr>
<th>Height</th>
<th>Avg</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>65 inches</td>
<td>3 inches</td>
</tr>
<tr>
<td>Weight</td>
<td>135 lbs</td>
<td>25 lbs</td>
</tr>
</tbody>
</table>

The heights and weights of 3 Stat 100 females are given below. What is your best prediction for how much each one weighs.

In Chapter 6 we calculated how much each would weigh if they fell on the SD line that is if they were exactly as heavy as they are tall (in terms of Standard Units).
Now we'll calculate how much each would weigh if she fell on the regression line. This gives us a better estimate of their true weight.

<table>
<thead>
<tr>
<th>Height in inches</th>
<th>$Z_{ht}$</th>
<th>$r = 0.5$</th>
<th>$Z_{wt}$</th>
<th>Weight in lbs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>71&quot;</td>
<td>$\frac{71-65}{3} = 2$</td>
<td>$0.5 \times 2 = 1$</td>
<td>$\frac{135}{5} + 1(25) = 160$</td>
<td>160 lbs</td>
</tr>
<tr>
<td>65&quot;</td>
<td>0</td>
<td>0</td>
<td>135 lbs</td>
<td></td>
</tr>
<tr>
<td>62&quot;</td>
<td>$\frac{62-65}{3} = -1$</td>
<td>$0.5 \times -1 = -0.5$</td>
<td>$135 - 0.5(25) = 122.5$</td>
<td>122.5 lbs</td>
</tr>
</tbody>
</table>

weight

$\text{height: gender} \geq 0.5$

$\text{reg line: } r = 0$ $\text{reg line: } r = 0.5$

$\text{reg line: } r = 0.9$
Example 2:
A large class took two exams with the following 5 summary statistics:
\[
\begin{align*}
\text{Ave}_{\text{Exam}1} &= 80, & \text{SD}_{\text{Exam}1} &= 10 \\
\text{Ave}_{\text{Exam}2} &= 70, & \text{SD}_{\text{Exam}2} &= 15 \\
& & r &= .6
\end{align*}
\]

a) Suppose a student is picked at random. What would be your best guess for what he got on Exam 2?

70 bc it's the average

b) Now suppose you're told that the student got a 90 on Exam 1, what would be your best guess for his score on Exam 2?

\[
\begin{align*}
\text{Exam 1 score} & \quad Z_{\text{score}} \quad Z \quad \text{Exam 2 score} \\
90 & \quad \frac{90-80}{10} = 1 \quad \times .6 = .6 \quad 70 + .6(15) = 79
\end{align*}
\]

c) A student got a 65 on Exam 1, but hasn't taken Exam 2 yet. What would you predict his score to be?

\[
\begin{align*}
65 & \quad \frac{65-80}{10} = -1.5 \quad \times .6 = - .9 \quad 70 - .9(15) = 56.5
\end{align*}
\]

Important Note: Predicted Z is always closer to the average (0) be r is always between -1 and 1
Regression Estimates for Percentiles

If 2 sets of numbers are both *normally distributed* then all you need is the correlation coefficient to make regression estimates for percentiles, you don’t need to know the averages or the SD’s. The key idea once again is to convert to Z scores and multiply by r.

**Example 3:**
Scores on the Math SAT and Verbal SAT follow the normal curve with $r = 0.6$. If a student was in the 80th percentile on the Math SAT I, predict what the student’s percentile on Verbal SAT.

*Note:* If $r = 1$ then the answer would be 80th percentile. If $r = 0$ then the answer would be the median (the 50th percentile). So, for $r = 0.6$ we know it will be somewhere in between. Where?

On the table below you’re given the percentile for one exam, make a regression estimate for the percentile for the other exam.

<table>
<thead>
<tr>
<th>Math SAT percentile</th>
<th>$Z_m$</th>
<th>$r = 0.6$</th>
<th>$Z_v$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$80^{th}$</td>
<td>$Z_m = 0.85$</td>
<td>$0.85 \times 0.6 = 0.51$</td>
<td>$Z_v$</td>
</tr>
<tr>
<td>$20^{th}$</td>
<td>$-0.85$</td>
<td>$-0.85 \times 0.6 = -0.51$</td>
<td>$Z_v$</td>
</tr>
</tbody>
</table>

**Important:** Predicted percentile is always closer to the 50th percentile (the average) since Z is always closer to 0.
Regression Effect—In test-retest situations, bottom groups improve on average and top groups fall back. The correlation is not perfect between the first and second tests. If it were, then the standard scores would be exactly equal. If there were no correlation between the first and second tests then the best prediction for everyone's second score would be the average. So with correlations in between 0 and 1 there is a drift toward the average. This is called the regression effect.

The regression effect is simply due to the football-shaped cloud around the SD line in the scatter diagram.

Regression Fallacy is thinking that the regression effect must be due to something more than just the shape of the cloud.

More Regression Estimate Problems

The basic procedure is always this:
1. Convert the value (or percentile) of the independent variable to standard units (z scores).
2. Multiply by the correlation coefficient
3. Convert this result back to the units of the dependent variable.

Example 4: A study gives the following data:

- Average height of mother = 65"
- Average height of son = 69"
- SD = 2"
- SD = 3"
- \( r = 0.5 \)

a) Estimate the average height of those sons whose mothers are 69" tall.

\[
\begin{align*}
69" &\quad \frac{69-65}{2} = 2 \\
&\quad \times 0.5 = 1 \\
&\quad 69" + 1/2 = 72'
\end{align*}
\]

b) Estimate the average height of those mothers whose sons are 72" tall.

\[
\begin{align*}
72" &\quad \frac{72-69}{3} = 1 \\
&\quad \times 0.5 = 0.5 \\
&\quad 65" + 0.5(2) = 66"'
\end{align*}
\]

see picture →
Chapter 7- Regression

Summary:

- The regression line can be used to make predictions for individuals by estimating the average value of Y for a given X.

- Associated with each increase of one SD in X, there is an increase of only r SD’s in y on the average.

- The slope of the regression line = r (SDy/SDx)

- The regression line is always flatter than the SD line. It's flatter by a factor of r.

- When r = 1 or r = -1 the regression line and the SD line are the same.

- To make a regression estimate you must always first convert the X value or percentile to a Z-score, multiply by r to get the Z-score for Y and then convert ZY to either a value or percentile.