Regression.

Percentile Ranks.

Continuing the previous example, consider a student with 90th %ile rank for SAT. What is the predicted %ile rank for GPA?

\[ \text{90th } %\text{-ile} = 1.3 \text{ SD above mean} \]

The regression method predicts that this student will be \[ 0.4 \times 1.3 = 0.5 \text{ SD}_{\text{GPA}} \text{ above the mean in GPA} \].

%ile rank for GPA is 69%.

%iles are connected with standard units.
Why does a 90th %-ile student get pulled down to being a 70th %-ile student?

- because correlation between SAT + GPA is not perfect.

→ if correlation was perfect, \( r = 1 \), then 90th % SAT \( \rightarrow \) 90th %ile GPA.

If correlation was \( r = 0 \):

90th %ile SAT \( \rightarrow \) 50th %ile GPA

(actually, any %ile GPA SAT).

Of all the students with 90th %ile SAT scores, some will do better, some will do worse in terms of GPA, and on average, they will do worse.

**Regression Effect**

It is commonly observed in test-retest situations that the bottom group in the first test will, on average, show some improvement on the second test; conversely the top group on the first test will show some reduction.
where does this come from
- random fluctuations - imperfect correlation.
and nothing else.
impl ications - makes it hard to judge the
effectiveness of interventions.
(eg give tutoring to students who performed
worst on 1st quiz - is their improvement
on 2nd quiz due to the tutoring or
the regression effect?)

quiz vs midterm.
mean quiz score 18.9 SD 2.9 r = 0.63
mean midterm score 16.6 SD 2.8

Mean quiz score is 2.3 points above mean
midterm score.
Mid-term score of 2.5 points below quiz score.

For quiz's score of 22, most mid-term scores are below the line.

Average mid-term score will also be below the red line.
How to explain the regression effect.

Individual.

Model

observed score = true score + chance error.

Suppose true scores have mean = 100, SD = 15

Consider people who score 140
- split into 2 categories
  1) true score below 140 & positive chance error ("good day")
  2) true score above 140 & negative chance error ("bad day")

Which is more likely?
Which contains more students?
Group 1 - because there are more students with true scores where a true chance error gives observed score of 140.

By more people have true score = 135 than have true score = 145

so for the same sized chance error an observed score of 145 is more likely to correspond to a true score below 145.

ie if score above average on one test, chances are the true score is lower than the observed score.

If take a similar test, expect that the score will decrease

and vice versa by symmetry.
Regression Errors.

Regression predicts the average value.

eg predict average temp given altitude
average price (€/gram) given TME concentration

Actual data will differ from the predictions

"regression errors"

error = actual value - predicted value

= vertical distance from
the data to the regression line.

"Size" of the error (or average) is given by

\[ \text{RMS of all errors} = \sqrt{\frac{(\text{error}_1)^2 + (\text{error}_2)^2 + \ldots + (\text{error}_n)^2}{n}} \]

This tells us how far a typical point is from the regression line.
68% of the data lie within 1 RMS error of the regression line.
95% of data within ±2 RMS errors of regression line.
How much better, in terms of the size of the expected error, in predicting using the regression rather than predicting the mean value of the unknown variable?

- if ignore $x$, RMS error is SD of $y$
- if use regression, RMS error must be smaller.

$$\text{RMS error} = \sqrt{1 - r^2} \times SD_y$$

Units of RMS error are the same as the units of what's being predicted (of, $\text{e}/\text{g}$, etc).

Perfect perfect correlation, $r = 1$, RMS error = 0

No correlation $r = 0$, RMS error = SD of what you are predicting.
**Example:** predicting midterm grades.

SD of midterm grades = 2.8

\[ r = 0.63 \]

\[
\text{RMS Error} = \sqrt{1 - r^2} \times SD_y
\]
\[
= \sqrt{1 - 0.63^2} \times 2.8
\]
\[
= 2.17
\]

Using regression, expected error is reduced from 2.8 to 2.17.

**Temp vs altitude.**

SD temp 6.5°F

\[ r = -0.76 \]

\[
\text{RMS Error using regression}
\]
\[
= \sqrt{1 - r^2} \times SD_y
\]
\[
= \sqrt{1 - (-0.76)^2} \times 6.5
\]
\[
= 4.22
\]

Knowing the altitude makes your prediction a 2°F more accurate.
Regression errors are also called **Residuals**.

Examining the residuals is a useful way to check whether the assumptions needed for regression to be valid actually hold.

(mostly "football shaped" scatter diagram")
Plot residuals.

plot (x vs y - predicted value from the regression)

looking for no obvious patterns in the residuals.

- average is always zero
- no tendency to drift up/down.
- no trends.
- about the same number of +ve as -ve residuals.

If see trends/patterns, the regression may not be so useful.