Case Study 18

The regression line:

- The best way to predict $y$ from $x$.
- Galton found that the line that was a smooth version of the graph of avg's and called it the regression line.
- The slope of the regression line is equal to:

\[ r \times \frac{s_y}{s_x} = b \]

- The equation of the regression line:

\[ \hat{y} = a + b \times x \]

predicted $y$ value

$y$ intercept

slope

2 ways to make regression predictions

① If a guy (from the sample) is 70.5" tall and the mean is 68" with an SD of ± 2.5", he is about 1 SD above mean in height.

\[ \frac{70.5 - 68}{2.5} = +1 \]

- We predict, because he is one SD above mean in $x$, he will be $r \times 1 = 0.36 \times$ SDs above mean $y$.
- mean $y = 158$ lbs (given)
- $SD = 25$ lbs
- $0.36 \times (25 \text{ lbs}) = 9 \text{ lbs above avg weight}$
- So, a guy who is 70.5" tall is predicted to weigh 167 lbs (158 + 9)

2nd method: 2) Work out slope and intercept of regression line and plug x into the equation.

\[
\text{Slope} = \hat{b} = r \frac{\text{Sy}}{\text{Sx}}
\]

\[
\hat{a} = ?
\]

Math fact: Regression line always goes through pt. of avg.
- So we use pt of avg. to fill in equation.

\[
(x, \bar{y}) : \quad \hat{y} = \hat{a} + \hat{b} \times (\text{in general})
\]

\[
\bar{y} = \hat{a} + \hat{b} \times (\text{so passing through } (x, \bar{y}))
\]

Rearrange:

\[
\hat{a} = \bar{y} - \hat{b} \times \bar{x}
\]

* For predicting y from x, the best slope is

\[
\hat{b} = r \frac{\text{Sy}}{\text{Sx}}
\]

And the best y intercept is \[
\hat{a} = \bar{y} - \hat{b} \times \bar{x}
\]

Now plug into eq. for regression line:

\[
y = \hat{a} + \hat{b} \times x
\]

If \( x = 70.5 \)"

\[
\hat{b} = r \frac{\text{Sy}}{\text{Sx}} = 0.36 \cdot \frac{25 \text{lb}}{2.5 \text{in}} = 3.6 \text{ lb/in}
\]

\[
\hat{a} = \bar{y} - \hat{b} \times \bar{x} = 158 \text{lb} - (3.6 \text{lb/in})(6.8 \text{in}) = -87 \text{lb}
\]

\[
\frac{\bar{y}}{\hat{y}} = (-87 \text{lb}) + (3.6 \text{lb/in})(70.5 \text{in})
\]

\[
\hat{y} = 167 \text{lb}
\]
Which method is better?

With only 1 prediction to make, method 1 is slightly faster, but with 2 or more predictions, method 2 is probably faster.

What the correct give/take for a regression prediction?

I predict that a guy 70.5" tall will weigh about 167 lbs, but I give or take how much?

\[
\text{reg line}
\]

\[
\text{If one ignores x and tries to predict y, the best prediction would be } \bar{y}_i \text{, give or take } SE(y) = SD \text{ of } y.
\]

If instead one uses x to predict y:

\[
x = x^* \Rightarrow \hat{y} = a + b x^* \text{ is best prediction.}
\]

Give or take should be smaller (give or take within the vertical strip is smaller than it is at the point of augs).

\[
** \frac{SE(\hat{y})}{\bar{y}} \leq \frac{SD(y)}{\bar{y}} \cdot \sqrt{1-r^2}
\]

Is approximate

Check: \( r = 0 \Rightarrow SE(\hat{y}) = SD(y) \)

Check: \( r = +1 \) or \( r = -1 \Rightarrow SE(\hat{y}) = 0 \)

These make sense

Ex. C518

If I don't know the guy's height, I would predict his weight to be \( \bar{y} = 158 \text{ lbs} \pm 25 \text{ lbs.} \)
If I do know he is 70.5" tall, then I predict he weighs 167 lbs with a give/take of $SE(y) = sy \sqrt{1-r^2}$ (for this prediction).

\[
SE(y) = 25 \text{ lbs} \cdot \sqrt{1-(0.36)}
\]

\[
SE(y) = 23 \text{ lbs}
\]

This is not much smaller than the situation where we do not know his height because height and weight are not very strongly correlated in this study.

Why did Galton call it regression?

Tall fathers have tall sons, but if the father is more than 2 S.D.'s above the avg in height, we only predict his son will be $r \cdot 2$ S.D.'s above avg = $(0.5)(2)\text{S.D.} = 1\text{ S.D.}$ above avg. So, tall fathers have tall sons, but not as tall (relatively) as their fathers were.

Similarly, short fathers tend to have short sons, but not as short as their fathers were.

This is called the regression effect.

- Galton called it "Regression to mediocrity."

Ex: Find the best line for predicting $y$ from $x$:

- This line has about vertical mistakes = BAD
- Want line with smallest sum of squared vertical errors.
This line is much better because it has a much smaller sum of squared vertical errors.

Find \((a, b)\) to make:

\[
\sum_{i=1}^{n} (Y_i - (a + bX_i))^2
\]

as small as possible. This is called the least squares line.

Math fact: regression line = least squares line.