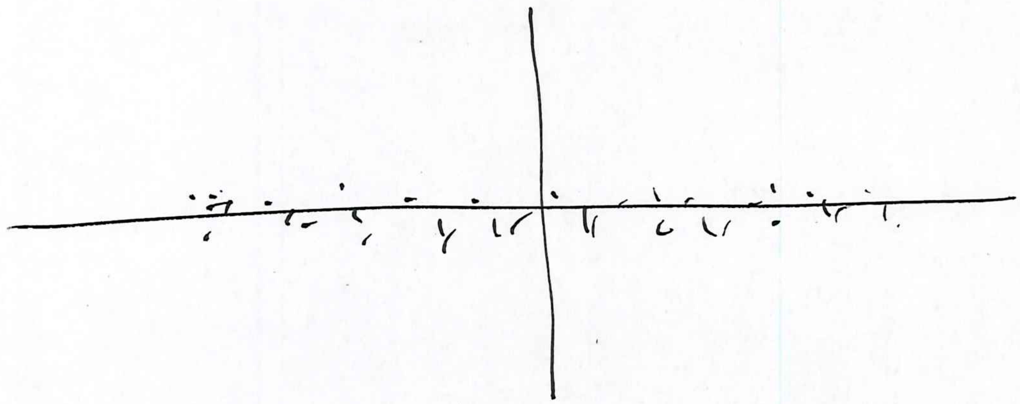
Math 455 Class 37 April 27

We considered a regression in which the scatter plot looked like this:



The regression line was $y = 0$ $[\hat{\beta}_0 = \hat{\beta}_1 = 0]$

The R^2 is 0.

Q: Why is $R^2 = 0$?

$R^2 = \mathrm{Cor}(Y, X)^2$. $\hat{\beta}_1 = \dfrac{\mathrm{Cor}(Y, X)\,\sigma_Y}{\sigma_X}$.

So if $\hat{\beta}_1 = 0$, $\mathrm{Cor}(Y, X) = 0$

(unless there is a data problem)

Q: R^2 is a measure of "goodness-of-fit".

The points are all close to the line.

Why is R^2 saying "bad fit"?

(R^2 close to 1 means "good fit".)

A: The predictor variable X does NOT help..

R^2 is saying that the "X" is irrelevant.

The null model $Y \sim \beta_0 + \varepsilon$ is just as good.

"Goodness-Of-Fit" is a fuzzy concept.

R^2 is a mathematical definition.

We might say that the model fits the data well because all points are close to the line.

The correspondence between fuzzy concepts and mathematical definitions is not perfect.

Important fact: the regression line goes through $(\bar{X}, \bar{Y})$. In the test question we had 100 points with $\bar{Y} = 0$, and added points 101, 102 with $Y_{101} = Y_{102} = 5$.

The NEW $\bar{Y}$ is

$$\bar{Y}_{new} = \frac{1}{102} 5 + \frac{1}{102} 5 + \frac{100}{102} \cdot 0$$

$$\approx 0.1$$

The new $\bar{X}$ is 0, as before.

Similar remarks apply to the case of adding just one point.