Interpreting R^2 in context! (*APstatsguy.com)

The squared correlation, R^2, gives the fraction of the data's variation accounted for by the model!

Or the fraction of the variation in y (the response variable) accounted for by x (the explanatory variable)

Just Checking (pg. 180): regression of house price (in thousands of dollars) on house size (in thousands of square feet)

$\widehat{Price} = 9.564 + 122.74(size)$

the $R^2 = 71.4\%$

① What does R^2 mean about the relationship of price & size? **A:** Differences in the size of houses account for about 71.4% of the variation in the house prices?

WHAT ABOUT THE OTHER 28.6%?

Maybe some of the houses have toilets made of gold, or are already furnished, etc.

The variation in price is NOT solely based on size (only about 71.4%)

② Is correlation of price & size positive or negative?

A: positive, because the slope is positive!

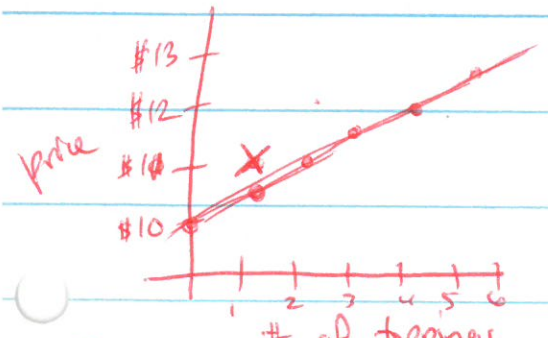
③ If we measured the size in thousands of square meters instead of thousands of square feet, would the R^2 value change? How about the slope?

A:
- r is not affected by scaling, therefore R^2 is not affected by scaling.
- slope would change to fit the new units.

A tale of two regressions ((PIZZA SHOPS))

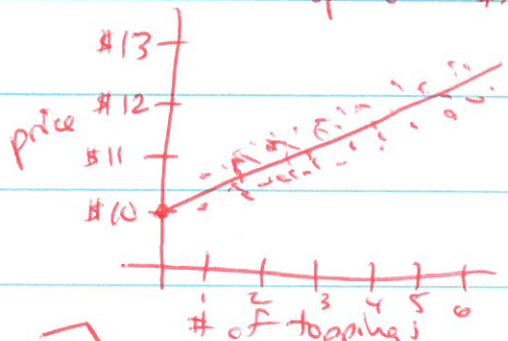
(APstatsguy.com)

Pizza Hut	( )'s Pizza
Plain Pizza: $10 toppings: $.50	Plain Pizza: $10 toppings - Extra cheese $1.00 pepperoni $.50 mushrooms $.25 special $1.50



$\widehat{Price} = 10 + .50(\#\ toppings)$

$r = 1$ $R^2 = 100\%$



$\widehat{Price} = 10 + .64(\#\ toppings)$

$r = .8485$ $R^2 = 72\%$

- If you notice, you can interpret both lines, and in both pizza shops, the more toppings you get, the higher the price of the pizza.

~~HOWEVER, some things are different.~~

- In both of them, the equations are very similar.
- Big difference is R^2 is 100% vs 72%.

R^2 is NOT:

- R^2 does not tell you the % of times the model will predict perfectly
- R^2 does not tell you the % of points that lie right on our model. (line goes through)

It's more complicated than that.

You'll never actually get R^2 to be 100%.

The model will never be perfect.

Even with an R^2 of 72%, you could use the model a million times & never predict perfectly.

We discuss variability - how things differ.

- We know the prices of pizza vary.
- We also know the # of toppings vary.

What this R^2 means is that 100% of the differences in prices of pizza can be explained by the differences in # of toppings.

OR 100% of the variability in prices can be explained by the variability in # of toppings.

* The pizza price is based SOLELY on the # of toppings!! Knowing the toppings alone is the only reason why there are differences in pizza prices.

OR 100% of the variability in prices is explained BY THE MODEL.

BUT, it's a different story at the other pizzeria. While the model does explain some of the price differences,

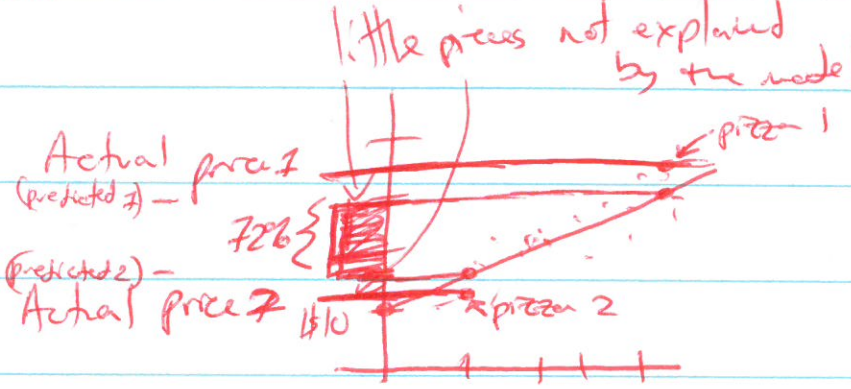
NOTICE here you can have the same # of toppings & TWO Different prices!

- So while the model does explain 72% of the price differences, it doesn't explain ALL of the price differences. There's other reasons why two pizzas may be different prices!
  - for example, someone wants Pepperoni, someone else wants Oreos.
  - Only 7 toppings per pizza, but different prices.

72% of the variability in price is explained by the model (or the variability in # of toppings)

- What about the other 28%?
  - That's in the residuals, not explained by toppings.

The model doesn't explain that part. The residuals are what's left over AFTER the model does its job.



pizza 1 and pizza 2 have different prices AND different # of toppings, but let's see what PART of that difference the model explained.

- R^2 tells you the percent in of variability of y explained by the model!

multiple choice

R^2 only explains differences. Think about that initial $10 base price for a pizza (NOT INCLUDED)

a test might say: "R^2 ~~explains~~ of .72 means that 72% of the price of pizza is explained by the model"

FALSE! 72% of differences in price, very important