AP Statistics

2019-01-24 3.2 Least-Squares Regression

Notes taken by: **Noah Overcash**

Warm-up

Consider each of the following relationships: the heights of fathers and the heights of their adult sons, the heights of husbands and the heights of their wives, and the heights of women at age 4 and their heights at age 18. Rank the correlations between these pairs of variables from highest to lowest. Explain your reasoning.

Highest: heights from age 4 to 18

The heights of fathers and their adult sons

Lowest: the heights of husbands and their wives

This is because the heights from age 4-18 are of the same person, the heights of fathers and adult sons are genetically related, and husbands and wives are are related.

A **regression line** (line of best fit) describes how a response variable y changes with an explanatory value x.

These are often used to predict a value of y for a given value of x

A regression line has an equation of the form y = a + bx

ˆy (y-hat) is the predicted value of y for a x

b is the slope

a is the y-intercept

The regression line helps us predict y for certain values of x

The accuracy depends on the scatter about the line

Don't make predictions of *x* way outside the interval we have data for (extrapolation)

Residuals are the difference between the actual value and the predicted value

r = y - ˆy

Least-squares regression line

Minimizes the residuals^2

The slope and correlation are closely related

When the variables are perfectly correlated, the change in ^y is the same as the change in x

Residual plots show the residuals for each point and make it easy to find outlines

Switching X and Y does not affect r