Stat 202 - 2015S - W7 - Tuesday

Review

Correlation - quantifies direction and strength of the linear relationship between two variables.

Direction: pos association $r > 0$
neg association $r < 0$

Strength: perfect line $r = \pm 1$
strong r close to 1 or -1
weak r close to 0

Regression Finds line $y = mx + b$

Or in statisticians notation

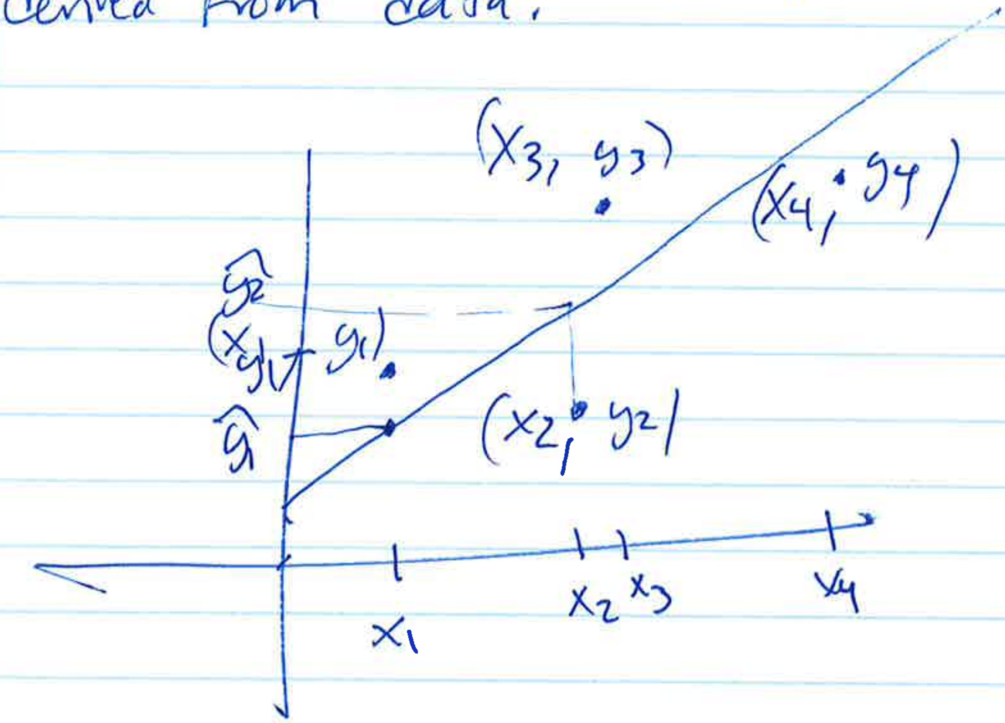
$y = b_0 + b_1 x$

Statcrunch just calls it

Slope
and intercept

NEW More on regression

Suppose I have a regression line derived from data:



line $y = mx + b$

If I plug in x_1 I get a prediction for y_1, specifically

$$\hat{y}_1 = mx_1 + b$$

The residual is the difference between the observed y (data) and the ~~[illegible]~~ predicted. y (regression)

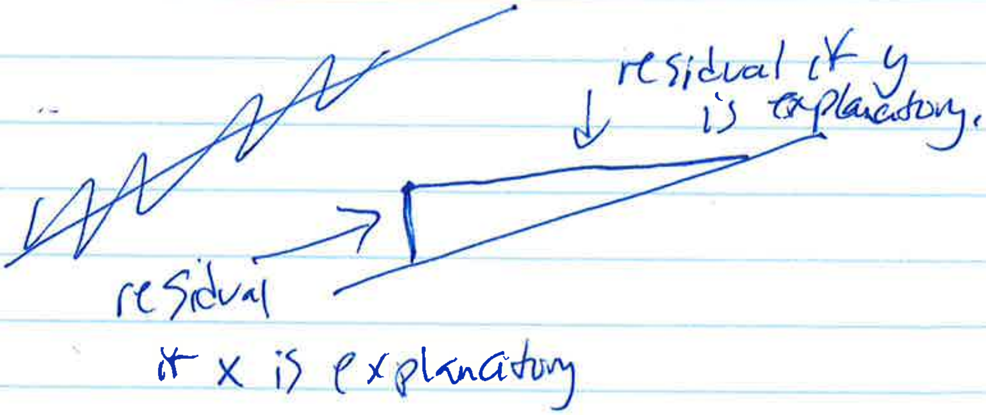
residual = observed − predicted

$$\boxed{\varepsilon_i = y_i - \hat{y}_i}$$

Residuals will be used in your homework

The regression line is the line that minimizes the squares of the residuals.

Note ~~[illegible]~~ why explanatory var can't be swapped with response var without getting ~~[illegible]~~ a different answer



A residual plot is a scatter plot of the regression residuals against the explanatory variable.

If the regression line catches the overall pattern of the data there should be no pattern in residual plot

(it should be an unstructured horizontal band centered at zero)

An outlier is an observation that lies outside the overall pattern of the other observations

Points that are outliers in y direction have large regression residuals but other outliers need not have large residuals

An observation is influential for a statistical calculation if removing it would markedly change the result of the calculation

Points that are outliers in the x direction of a scatterplot are often influential for regression

Some cautions

Caution

Extrapolation - using the regression line to predict response outside of range of data

Often Unreliable ↗

Caution

A lurking variable is a variable that is not among the explanatory or response variables and yet may influence the interpretation of the relationships among those variables

Caution Averaging data leads to greater correlation,

Finally the square of the correlation

has a handy intepretation:

it is the fraction of variation in the values of y that is explained by the ~~least squares~~ by the regression line

$$r^2 < |r|$$