Stat 202 - 2015 ~~20~~ - W3 - Tues

Properties

* r does not depend on units
  change data with a linear transformation
  to change unit cm to inches or lightyears
  doesn't change correlation
* r has no units
* $r > 0$ means pos association
* $r < 0$ means negative association
* [redacted]
* no distinction between explanatory / response vars
* Both variables must be quantitative
* $-1 \leq r \leq 1$
* values near zero indicate a weak linear relationship.
* values close to ± 1 indicate data like close to a straight line
* $r = \pm 1$ exactly means points lie exactly on a line
* measures only strength of linear relationship not curved relationship no matter how strong.

Oasis
Bioclocks
Fidget
] quiz

Regression

Regression vs Correlation

* Correlation is a measure which quantifies the strength and direction of the linear relationship between two variables.

* Regression finds the best fitting line

$y = mx + b$ Finds m and b

Statisticians use other notation

Slope and intercept

$$y = b_0 + b_1 x$$

m or b_1 $\quad$ Slope $= r\dfrac{S_y}{S_x}$ $\quad$ b or b_0 $\quad$ intercept $= b_0 = \bar{y} - b_1\bar{x}$ $\quad$ $\bar{x}$ is mean of x

Show Statcrunch

* Regression line passes through $(\bar{x}, \bar{y})$

$$\left[y = (\bar{y} - b_1\bar{x}) + b_1 x \right]$$

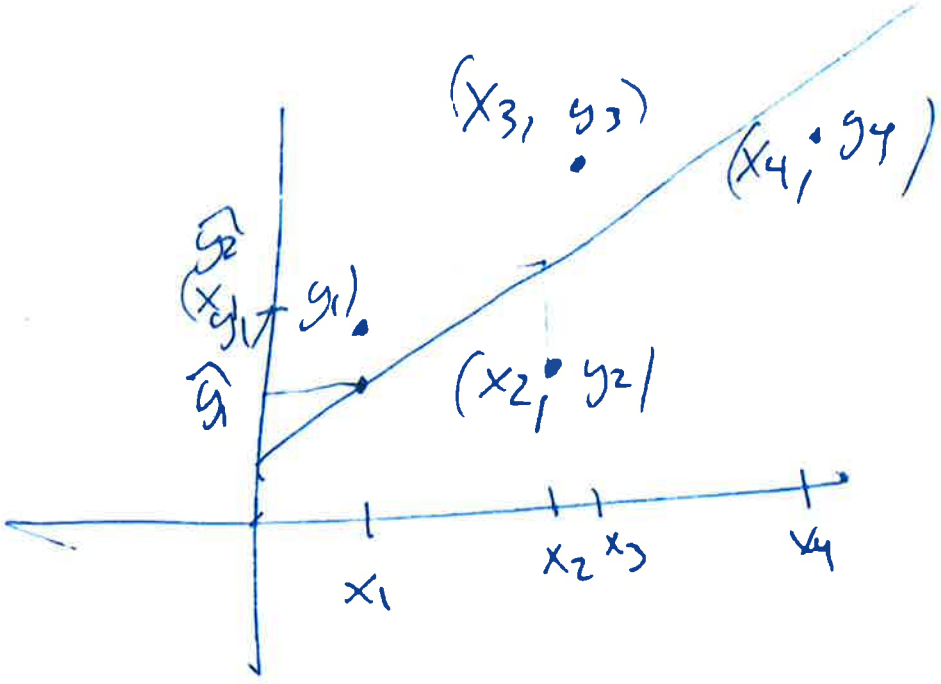
(b_0; plug $x = \bar{x}$ get $y = \bar{y}$)

Correlation: Line on scatterplot intercept is blue

* Has different slope depending on what is explanatory/response. even when plotted on same axes,
* must decide which is explanatory which is response.

New Mon 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 predicted y (regression)

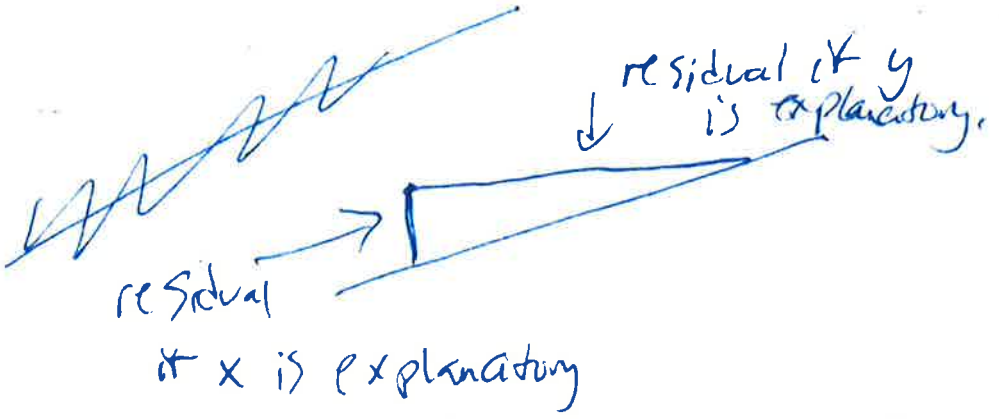
residual = observed − predicted

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

Residuals will be used in your homework

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

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



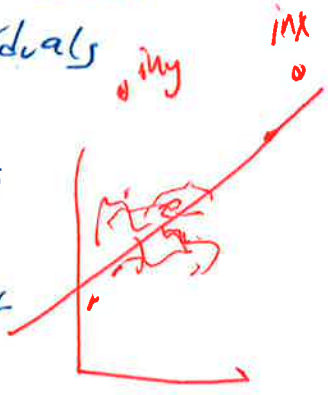
A residual plot is a scatter plot of the regression residuals against the explanatory variable.

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 centred 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 Even if residuals are small

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 interpretation:

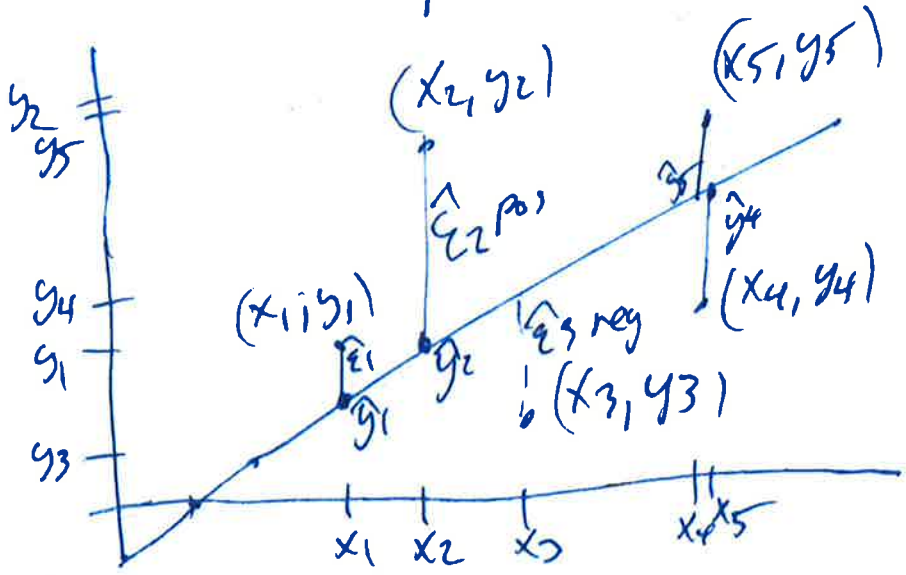
it is the fraction of variation in the values of y that is explained by the ~~least squares~~ by the regression line.

$$0 \leq r^2 \leq |r|$$

Stat 202-20155-W7-Wed

Review

Given data points



And the regression line $y = mx + b$ Standard notation

$\hat{y} = b_0 + b_1 x$ Stat notation

the explanatory variable is x

the response variable is y

The predicted y_i or $\hat{y}_i$ is what is predicted from the regression line

$\hat{y}_i = m x_i + b$ Standard notation

$\hat{y}_i = b_0 + b_1 x_i$ Stat notation

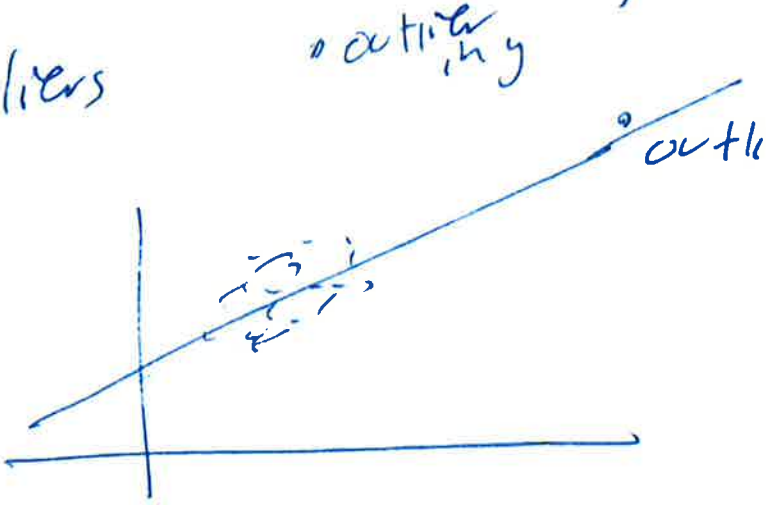
$\varepsilon_i = y_i - \hat{y}_i$ residual

Residual plot – plot residuals against explanatory variable

Should be no pattern if regression line captures overall pattern of data

(Specifically an unstructured horizontal band centered at zero)

Outliers



Influential – Removing obs has a large effect.

Outliers in x tend to be influential

~~Outliers in y too, I think~~

Fact

$$r^2 = \frac{\text{variance of predicted values } \hat{y}}{\text{variance of observed values } y}$$

This is what is meant by saying

r^2 is the fraction of the variance in the values of y that is explained by the regression line.