Terms to Know

scatterplot	display the relationship between two numerical variables
correlation coefficient "-r"	the strength, direction, and linear relationship between the x-variable and y-variable
least square regression line	line of best fit for the scatterplot; minimizes the sum of the square of the deviations from a line
explanatory variable	explains the other variable; causes the response variable to change
response variable	response to the other variable; dependant
extrap-olation	not right; using LSRL to predict values outside of the range of the original data set
outliers	points that are far away from the LSRL relative to other points
influential points	points that significantly impacts the slope of the LSRL
lurking variable	different outside variables that causes both x and y to change
residual	y - ŷ
coefficient of determ-ination "-r^2"	r^2% of the variation in y-variable can be explained by the approximate linear relationship between x-variable and y-variable

Strength of "r" (Correlation Coefficient)

legitimate values	[-1,1]
none	0
weak	(-0.5,0) U (0, 0.5)
moderate	(-0.8, -0.5) U (0.5, 0.8)
strong	[-1, -0.8) U 90.8, 1]

LSRL Example

Predictor	Coef	SE Coef	T	P
Constant	2.544	0.134	18.955	0.000
Caffeine (mg)	0.164	0.057	2.862	0.005

$S = 1.532$ R-Sq $= 60.032\%$ R-Sq(adj) $= 58.621\%$

Desiree is interested to see if students who consume more caffeine tend to study more as well. She randomly selects 202020 students at her school and records their caffeine intake (mg) and the number of hours spent studying. A scatterplot of the data showed a linear relati-onship.

This is computer output from a least-squares regression analysis on the data.

LSRL Example Interpretations

find the LSRL	ŷ = 2.544 + 0.164x
identify the variables	x = amount of caffeine intake (mg); y = number hours spent studying
interpret the slope	when the amount of caffeine intake increases by one, the number of hours spent studying increase by 0.164
identify the coefficient of determination	r^2 = 60.032
interpret the coefficient of determination	60.032% of the variation in the amount of hours spent studying can be explained by the approximate linear relationship with caffeine intake
find the correlation coefficient	r = 0.7748
interpret the correlation coefficient	there is a moderately strong, positive, linear relati-onship between the intake of caffeine and the amount of time spent studying

Interpretations

slope of LSRL	for each increase in the "x-variable" of one "x-unit", there is a predicted "increase/decrease" in the "y-variable" of "b constant" "y-units"
correlation coefficient	there is a "strength", "direction", linear relationship between "x-variable" and "y-variable"
correlation of determination	"r^2"% of the variation in the "y-variable" can be explained by the approximate linear relationship between "x-variable" and "y-variable"
residual	the actual "y-variable" is "residual" "y-unit" "above/below" the predicted "y-variable"

Residuals and Residual Plots

the sum of the residual is always zero

error = observed - predicted

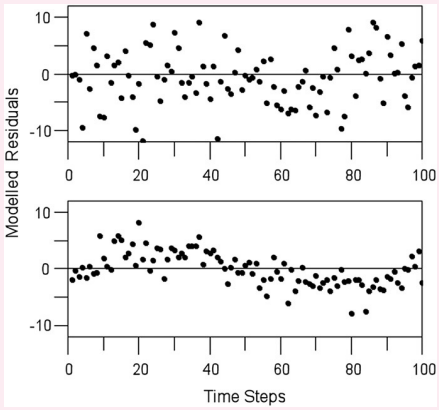
residual plots show if the model is appropriate or not between two variables

if there is no pattern between the points on the residual plot, the model is appropriate

if there is a pattern between the points on the residual plot, the model is not appropriate

when the residual plot is not appropriate, you can transform the data points until the plot turns random

Residual Plot Examples



the top residual plot is appropriate because the points are random while the bottom residual plot is not appropriate because there is a pattern between the points

Non-Linear Transform Data

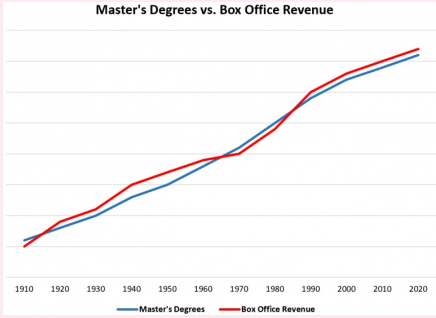
x & log y

log x & log y

x & sqrt y

x & 1/y

Correlation Doesn't Imply Causation



If we collect data for the total number of Master's degrees issued by universities each year and the total box office revenue generated by year, we would find that the two variables are highly correlated.

Correlation Doesn't Imply Causation Explanation

Does this mean that issuing more Master's degrees is causing the box office revenue to increase each year? Not quite. The more likely explanation is that the global population has been increasing each year, which means more Master's degrees are issued each year and the sheer number of people attending movies each year are both increasing in roughly equal amounts. Although these two variables are correlated, one does not cause the other.