Chapter 10: Correlation and Regression (Linear Regression)

Section 10.1: Correlation

Stat 50

CORRELATION

Def A correlation exists between two variables when the values of one variable are somehow associated with the values of the other variable.

relationship

Ex from algebra: $y = x^2$, $x=0 \rightarrow y=0$, $x=1 \rightarrow y=1$, $x=2 \rightarrow y=4$

EXPLANATORY VS RESPONSE VARIABLE

Def Explanatory Variable (EV)

the independent variable used to predict or explain differences in a response variable

Def Response Variable (RV)

the dependent variable found as an outcome that is measured following a manipulation of the explanatory var.

The way to distinguish the difference between these two variables is by asking, "Which statement makes sense?"

Ex: *A researcher wants to examine whether babies fed on breast milk are more or less likely to be ill.*

(a) *Feeding a baby on breast milk causes resistance to disease.* plausible

(b) ~~*Resistance to disease causes a baby to feed on breast milk.*~~ nonsense

Can you identify which variable is which? – presence or absence of breast milk (EV) – resistance to disease (RV)

Ex: **Identify the explanatory and response variables in the following examples.**

(a) An experiment was conducted to test the effects of sleep deprivation on human response times.

EV RV

(b) Researcher Penny Gordon Larson and her associate wanted to determine whether young couples who marry or cohabitate are more likely to gain weight than those who stay single.

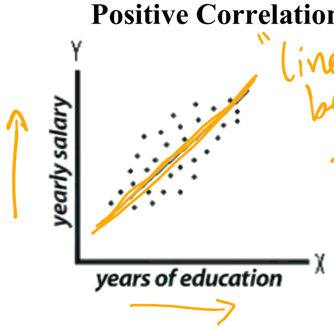
EV RS

SCATTERPLOTS

Def A **scatterplot** is a plot of paired (x, y) quantitative data.

Note: A scatter diagram is often helpful in determining whether there is a relationship between the two variables. "correlation does not imply causation"

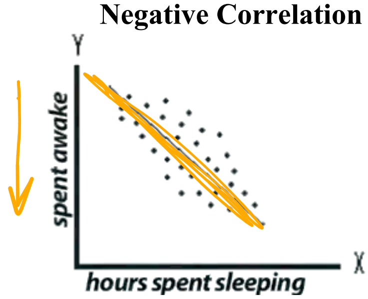
Positive Correlation



Variables move in the SAME *direction.*

⊕ "positive"

Negative Correlation

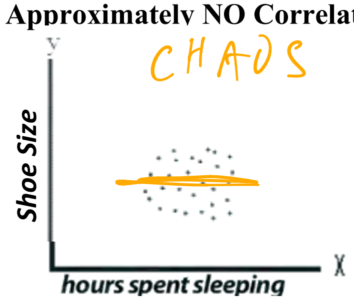


Variables move in opposite *directions.*

"negative"

Approximately NO Correlation

CHAOS



No apparent relationship between the two variables.

To test for ___LINEAR (relationship)___ correlation we attempt to draw a line that best fits the data. Every linear correlation is expressed by two features: when there is linear correlation:

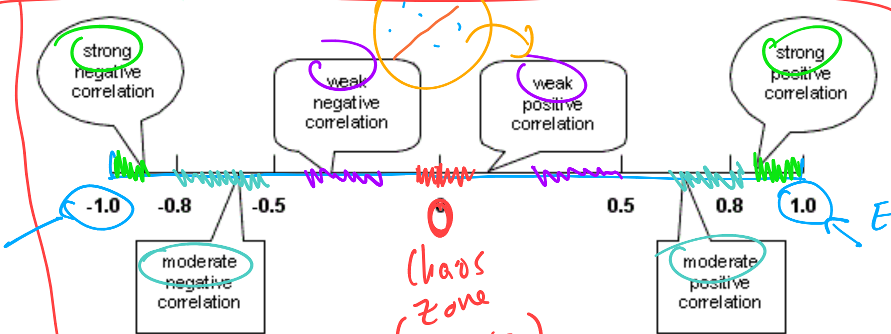
STRENGTH –	DIRECTION –
how close the data points fall near a line	As x (EV) increases, y increases (+) As x increases, y decreases (−)

The strength of the linear correlation is represented by a numerical value called the ___correlation coefficient (r)___

Properties of the Linear Correlation Coefficient

1. The value of r is always between -1 and 1, inclusive. $-1 \le r \le 1$
2. The value of r does not change if all values of either variable are converted to a different scale.
3. r measures the strength of a linear relationship only. (*It does not measure nonlinear relationships.*)

KEY TO CH10

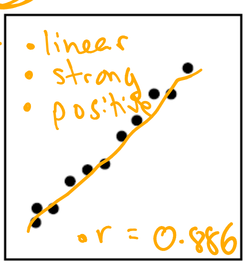


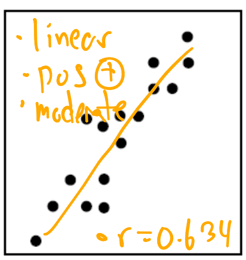
Exact −1 (Perfect line − fit)

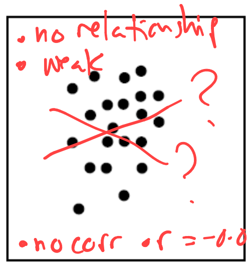
Exact 1 (Perfect + line fit)

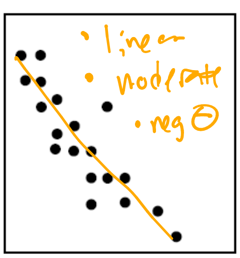
*Note: Correlation does not imply ___CAUSATION___, just ___RELATION/ASSOCIATION___.

Ex: Choose from the following word bank to determine the features of each scatter plot below.

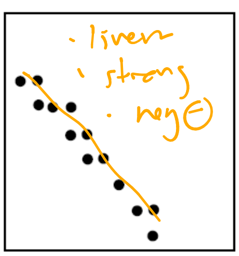




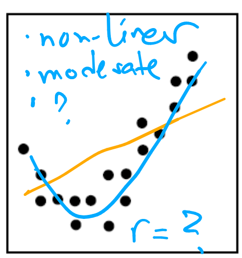




r = −0.735



r = −0.89



Choose a word from each category to describe each scatter plot shown to the left and write near/next to each plot:

- Linear relationship
- Non-Linear Relationship
- No Relationship

- Perfect
- Strong
- Moderate
- Weak

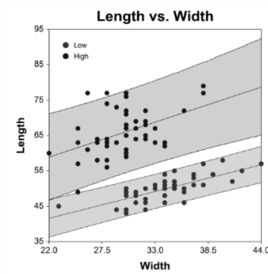
- Positive Association
- Negative Association
- No Correlation

- $r = -0.735$
- $r = 0.634$
- $r = -0.02$
- $r = 0.886$
- $r = -0.89$
- *no r* value

Explained Variation (r^2)

The value of r^2 is the proportion of the variation in y that can be explained by the linear relationship between x and y.

Different samples will produce different scatterplots, and thus, different r values.



Our job is to take a large enough sample to get close to the *true* population linear correlation coefficient called $\underline{\ \rho\ }$.

← new parameter

ρ = "rho" (r sample corr. coeff)

We are going to assume there is no correlation ($\rho = \underline{\ 0\ }$) and try to prove otherwise based on our sample.

Steps for Hypothesis Test when Applied to testing ρ		
Check Requirements • Simple Random Sample (SRS) • Visual examination shows straight line pattern • Remove any outliers (optional)	***Step 1:*** *Hypotheses* $H_0: \rho = 0$ (there is no linear correlation) $H_1: \rho \neq 0$ (there **is** a linear correlation) (Two Tailed Test - ALWAYS!)	***Step 2:*** *Level of Significance*

Step 3: *Test Statistic*

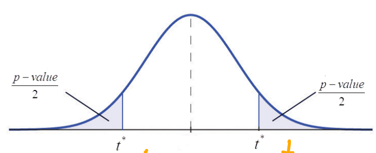
$$\boxed{t_0 = \frac{r}{\sqrt{\frac{1-r^2}{df^*}}}}$$

where $df^* = n - 2$

Step 4: *Find a Critical Value or P-Value*

P-VALUE METHOD

DECISION 2-Tailed

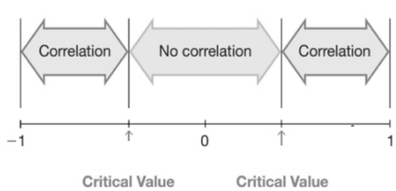


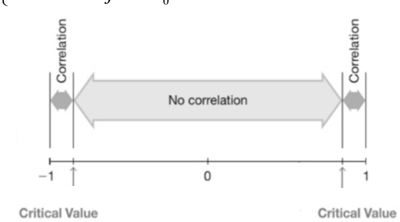
Reject H_0 ~ if $P\text{-value} \leq \alpha$
Fail to Reject H_0 ~ if $P\text{-value} > \alpha$

CRITICAL VALUE METHOD

DECISION

Reject H_0 ~ if r^* lies in the critical region
Fail to Reject H_0 ~ if r^* doesn't lie in the critical region





Step 5: Write a *CONCLUSION* either rejecting or failing to reject H_0

n	$\alpha = .05$	$\alpha = .01$
4	.950	.990
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	.330
70	.236	.305
80	.220	.286
90	.207	.269
100	.196	.256

Critical Values of the Pearson Correlation Coefficient r

GRAPHING CALCULATOR (TI-83 OR 84)

Instructions:

STAT ⇒ TESTS ⇒ LinRegTTest

STATS → Tests → LinRegTTest

Ex: The following table gives information on average saturated fat (in grams) consumed per day and cholesterol level (in milligrams per centiliters) of ten men taken from a simple random sample. (Note Matched Pairs)

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Fat Consumption (in grams)	55	68	50	34	43	58	77	36	60	39
Cholesterol level (in mg/cL)	180	215	195	165	170	204	235	150	190	185

Use a 0.01 significance level to determine if there is a linear correlation between saturated fat consumption and cholesterol level.

Check Req
① SRS ✓
② Visual check (moderate Lin)
③ remove out? (none)

Null and Alternative Hypothesis

$$\begin{cases} H_0: \rho = 0 & (\text{no lin correlation}) \\ H_A: \rho \neq 0 & (\text{is lin. corr}) \end{cases}$$

Test Statistic

Calc
$r = 0.908$
$r^2 = 0.824$
(Rounding Rule 3 decimals or 3 sig fig)

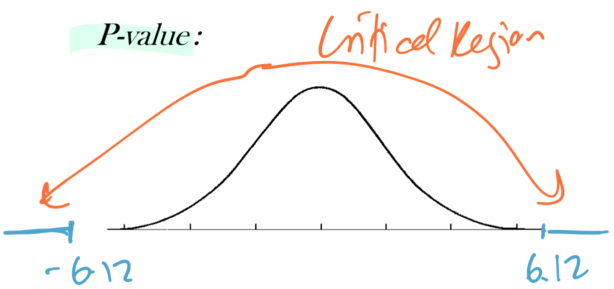
$$t_0 = \frac{r}{\sqrt{\frac{1-r^2}{df^*}}} = \frac{0.908}{\sqrt{\frac{1-0.824}{8}}} = 6.12$$

$df^* = 8 = 10-2$

$\boxed{t_0 = 6.12}$

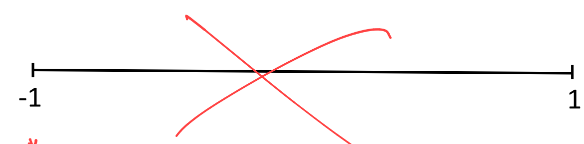
P-Val
$P = 2.81E-4$
$= 0.000281$

P-value:



$P(\text{Crit Reg}) = 1 - \text{tcdf}(\overset{a}{-6.12}, \overset{b}{6.12}, \overset{df^*}{8})$
$= 2.83E-4$
$= 0.000283$

~~*Critical Value:*~~



Decision about Null Hypothesis

$p = 0.000281 < \alpha = 0.01$
P low, null go $\boxed{\text{Reject } H_0}$

Conclusion

"We found enough statistical evidence to support the claim that their is a linear relationship b/w fat consumption & cholesterol level in Men."

Ex: The following table gives the total 2004 payroll (on the opening day of the season, rounded to the nearest million dollars) and the percentage of games won in 2004 by each National League team.

	D'backs	*Braves*	*Cubs*	*Reds*	*Rockies*	*Marlins*	*Astros*	*Dodgers*
Payroll (in millions)	70	90	91	47	65	42	75	93
Percentage of Wins	31.5	59.3	54.9	46.9	42.0	51.2	56.8	57.4

keep it
outlier?

	Brewers	*Expos*	*Mets*	*Phillies*	*Pirates*	*Cards*	*Padres*	*Giants*
Payroll (in millions)	28	41	97	93	32	83	55	82
Percentage of Wins	41.6	41.4	43.8	53.1	44.7	64.8	53.7	56.2

Use a 0.05 significance level to determine if there's a correlation between payroll and percentage of games won.

Null and Alternative Hypothesis

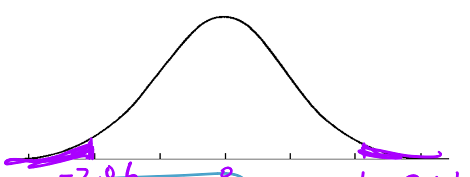
$$\begin{cases} H_0: \rho = 0 \\ H_A: \rho \neq 0 \end{cases}$$

Calculator
$t_0 = 2.06$
$p = 0.0585$
$r = 0.482$
$r^2 = 0.232$

Test Statistic

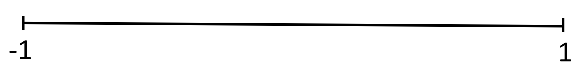
$\boxed{t_0 = 2.06}$

P-value:



$\boxed{P = 0.0585}$

Critical Value:



Decision about Null Hypothesis

$\alpha = 0.05$
$p = 0.0585$

$P > \alpha \longrightarrow$ P high, null fly

$\boxed{\text{Fail to Reject } H_0}$

Conclusion

"We do not have enough statistical evidence to support that the is a linear relationship b/w Baseball (NL teams) payroll & percentage of games won in 2004."

REGRESSION

"line of best fit"

Def Given a collection of paired sample data, the **regression equation** $\boxed{\hat{y} = b_0 + b_1 x}$ algebraically describes the relationship between the two variables.

Calculator Notation $y = a + bx$

*Note: The graph of the regression equation is called the **regression line** or line of best-fit*

TERMINOLOGY

x is referred to as the explanatory variable, predictor variable, or the independent variable.

y is referred to as the response variable or the dependent variable.

REGRESSION LINE A.K.A. LEAST-SQUARES LINE A.K.A. "LINE OF BEST FIT"	The line that minimizes the distance between the points and the line. It BEST FITS the data points. $\hat{y} = b_0 + b_1 x$ where the point $(\bar{x}, \bar{y})$ will always be on the least-squares line. $b_1 =$ slope $b_0 =$ y-intercept.

FORMULAS

Slope $\boxed{b_1 = r \cdot \dfrac{s_y}{s_x}}$ ***y-intercept*** $\boxed{b_0 = \bar{y} - b_1 \bar{x}}$

☆ *Round-Off Rule: Round the slope and y-intercept to three significant digits.* (3 sig fig) ☆

Ex: Below is a sample of five patients at a hospital with the information regarding their height and weight.

(a) Describe the relationship of the data using the scatterplot given.

linear relationship, positive correlation, ~ moderate

(b) Find the correlation coefficient/test statistic and determine whether a correlation exists.

$\boxed{r = 0.903}$

$t_0 = 3.64$

$p = 0.0357$

$\alpha = 0.05$

$p < \alpha \rightarrow$ null go

reject H_0 →

there is linear relationship.

(c) Find the line of best fit and sketch it below.

$y = b_0 + b_1 x$ or $y = a + bx$

$\boxed{y = -360 + 7.75x}$

2 points
(66, 151.5)
(78, 244.5)

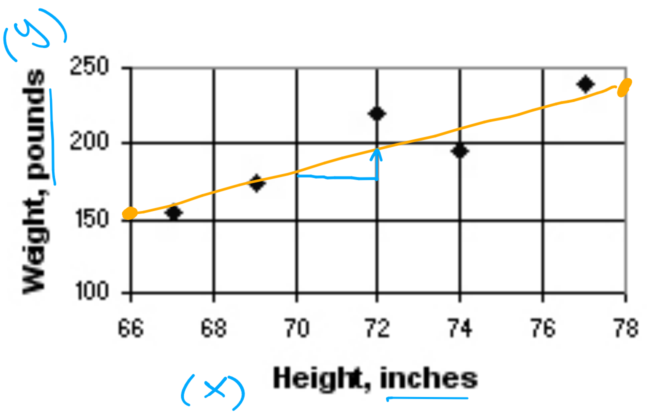
range of reasonableness:
$x = 0$, $y = -360$ (nonsense)

(c) Interpret the slope.

slope: $b_1 = b = 7.75$

"for one inch increase in Height, the weight increases by 7.75 pounds"

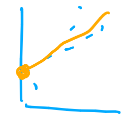
Height (in)	Weight (lbs)
67	155
72	220
77	240
74	195
69	175



Interpretation of slope: *For every unit increase in the explanatory variable, on average, there is an increase/decrease of "slope" units on the response variable.*

b_1

Interpretation of y-intercept: *When the explanatory variable is 0 units, on average, the response variable is* ___b_0___ *units.*



GRAPHING CALCULATOR (TI-83 OR 84)

Instructions: STAT ⇒ TESTS ⇒ $\boxed{\text{LinRegTTest}}$

Plotting Regression Line & Data
Tests → LinRegTTest
Reg EQ: $\boxed{\text{VARS}}$ $\boxed{\text{Y-VARS}}$ $\boxed{\text{1. Function}}$ $\boxed{\text{Y1}}$

Ex: Recall the exercise from the last section in which we concluded there was a significant linear correlation between the average saturated fat consumed per day and the cholesterol level of ten men.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
(x) Fat Consumption (in grams)	55	68	50	34	43	58	77	36	60	39
(y) Cholesterol level (in mg/cL)	180	215	195	165	170	204	235	150	190	185

(a) Find the regression equation where x is the average daily fat consumption (in grams) of a man and y is the cholesterol level (in mg/cL).

$y = 106 + 1.59x$

(b) Interpret the slope in the context of the problem. (M-E)

"For every one gram increase in fat consumed, we expect a man's cholesterol to increase by 1.59 mg/cL"

(c) Predict the cholesterol level of a man who consumes 65 grams of saturated fat per day.

KEY to Regression Line EQ

y = cholesterol level
x = 65 g consumed

$y = 106 + 1.59(65)$
$= 209.35$

$\boxed{y = 209.4 \text{ mg/cL}}$

"We predict a man who consumes 65 g of fat to have a cholesterol level of 209.4 mg/cL"

PREDICTIONS

If there is a significant linear correlation between x and y, then use the ___LINE OF BEST FIT (Regression Line)___ to predict the value of y given a specific value of x.

If there is no significant linear correlation between x and y, then the best prediction of y is the ___MEAN ($\bar{y}$) of the ys___ for any given value of x.

i.e. all values of x

Ex: Are fat and sodium content related in fast food? Here are the fat and sodium content for several brands of burgers.

	1	2	3	4	5	6	7
(x) Fat (in grams)	19	31	34	35	39	39	43
(y) Sodium(mg)	920	1500	1310	860	1180	940	1260

Use a 0.05 significance level to determine if there is a linear correlation between fat and sodium content in burgers.

Null and Alternative Hypothesis

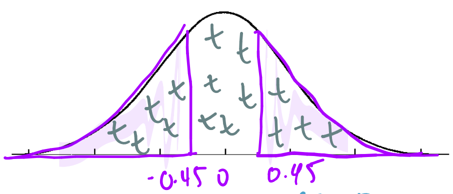
$$\begin{cases} H_0: \rho = 0 \\ H_A: \rho \neq 0 \end{cases}$$

Test Statistic (or correlation coefficient)

$t_0 = 0.45$ $\quad r = 0.199 \sim r = 0.2$

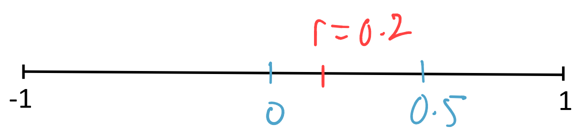
$p = 0.669$

P-value:



$\alpha = 0.05$ $\quad p = 0.669$

Critical Value:



Decision $\quad p > \alpha$ p high, null fly Fail to Reject H_0

Conclusion "There is not enough statistical evidence to support the claim that there's a linear correlation between fat & sodium content in burgers."

(a) What is the regression equation? Is it helpful in this situation? Why or why not?

$y = 930 + 6.08x$ ← USELESS! DONT USE! B/C no corr

(a) Predict the sodium level of a burger with 25 grams of fat.

↳ use mean $\bar{y}$! 1-VAR STATS $\boxed{\bar{y} = 1138.6 \text{ mg}}$

GRAPHING CALCULATOR (TI-83 OR 84)

To create and view a Scatterplot and Linear Regression Line

Instructions:

1) 2nd ⇒ 0 (catalog) ⇒ DiagnosticOn ⇒ Enter
2) STAT ⇒ EDIT (enter 1st Variable in L_1 and 2nd Variable in L_2)
3) STAT ⇒ CALC ⇒ 4: LinReg $(ax + b)$ ⇒ Store RegEQ: ⇒ Vars ⇒ Y-Vars ⇒ 1: Function ⇒ 1:Y_1 ⇒ Calculate
4) 2nd ⇒ $y =$ ⇒ 1: Plot1 ⇒ On ⇒ Zoom ⇒ 9: ZoomStat

Note can also do it from LINREGTTEST.