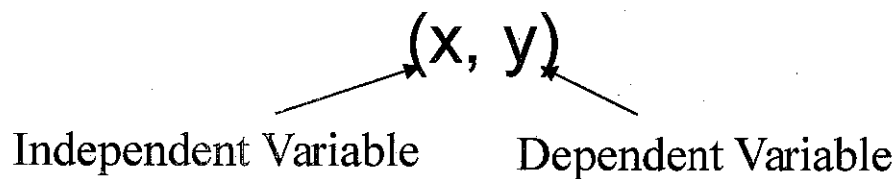
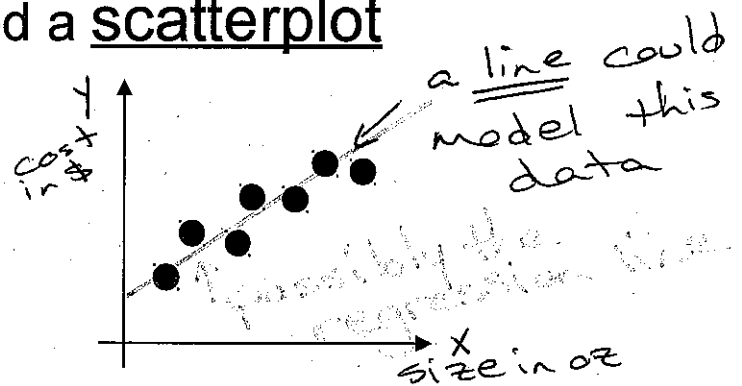
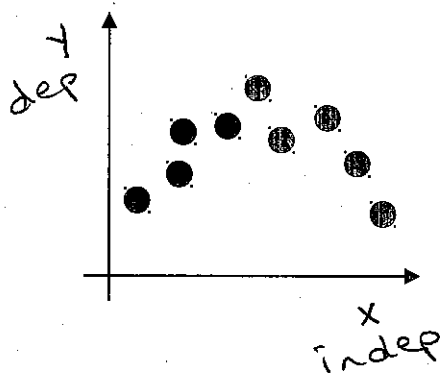


Chapter 12: Bivariate Data



Bivariate Data can be graphed on a set of axes.
This yields a graph called a scatterplot



If the data is linear we can fit a line to the data

This process is called linear regression

The regression line is a straight line
(i.e. a linear function) that best
models our data

1. The cost of a box of laundry detergent is compared to the number of ounces the box contains.

Indep: size in ounces Dep: cost

2. The time it takes an assembly line worker to assemble a chip board is compared to the number of boards assembled per day

Indep: Dep:

3. A study is done of workers to see if years of education increases their average annual salary

Indep: Dep:

4. Your long distance telephone bill varies according to the number of minutes of long distance phone calls made.

Indep: Dep:

5. A study is done of people living near a former industrial site to see if cancer rates are related to the level of a certain pollutant found in the ground water

Indep: Dep:

y depends
on x

Example: In an effort to get people to see the effect of drinking on a person's driving, a local police precinct solicited volunteers for the following activity. The activity consisted of a person attempting to place 18 geometric shaped blocks into matching slots in a 6-sided cube. The person first completed the task while sober. The person then began consuming one screwdriver every 15 minutes. (A screwdriver is an alcoholic drink made of orange juice and a shot of vodka.) After each drink, the person would again attempt to place the 18 blocks in the cube. The time it took to complete the task was recorded each time.

$x \leftarrow \text{indep. var.}$

$y \leftarrow \text{dep. var.}$

Number of Drinks	Time (seconds)
0	80
1	84
2	95
3	102
4	105
5	111
6	117
7	120
8	126
9	135
10	184

$$\hat{y} = 75.1364 + 7.8636x$$

Interpreting the Regression Line

Domain = possible x -values

In our # of drinks example, the domain is $0 \leq x \leq 10$

* Even if data is discrete, we write the domain this way b/c it's the domain of the line \hat{y}

If you use \hat{y} to predict \hat{y} for x values outside the domain, the results are unreliable.

a = y -int for \hat{y}
= the predicted value for y when $x=0$

b = Slope of \hat{y}

= change in \hat{y} per unit change in x

In our ex. $b \approx 7.9$, so we predict it will take an extra 7.9 sec. to complete the puzzle after each drink

Example: The drunk driving data gave the following regression line:

$x = \# \text{ drinks}$

$y = \text{time to finish cube}$

Predict how long it should take, on average, for someone to complete the cube if they have:

a. 5 drinks

$$\hat{y}(5) = 75.1364 + 7.8636(5) \approx 114.5 \text{ sec}$$

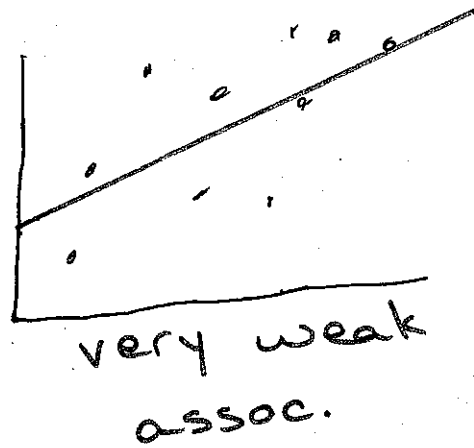
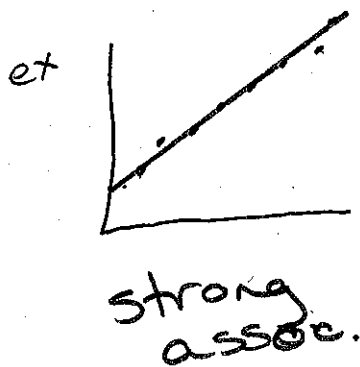
b. 7.5 drinks $\approx 134.1 \text{ sec}$

c. 15 drinks $\approx 193.1 \text{ sec}$

*Our domain for \hat{y} to be reliable is only $0 \leq x \leq 10$, so this prediction might not be reliable

Corr = r = Correlation Coefficient

- ① The correlation coefficient tells you how strong the association is between x and y
- ② It also tells you how the assoc. works




Ch12: Correlation Coefficient

$r =$ correlation coeff.


if $r > 0$

then y increases as x increases


i.e. \hat{y} has pos. sl.

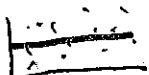
if $r < 0$

then y decreases as x increases


 \hat{y} has neg. slope

if $r = 0$

there is no correlation between x & y



if r close to ± 1

the correlation is strong or significant

if r close to 0

the correlation is not significant

How close to ± 1 does r have to be for strong?

Critical Values Table in

§ 12.10

Important
 $-1 \leq r \leq 1$

Correlation Coefficient (continued)

$$\hat{y} = 75.14 + 7.86x$$

$$r = .9120$$

$$df = n - 2 = 11 - 2 = 9$$

$n = \# \text{ of points } (x, y) = 11$

12.10 Critical Values Table (in text)

$$CV = 0.602$$

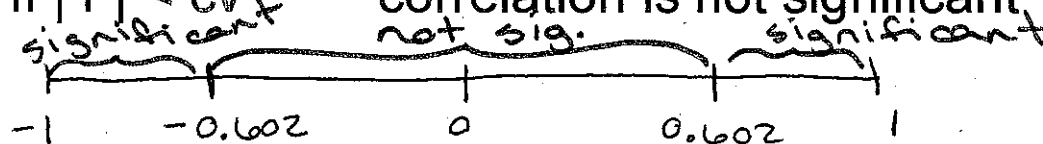
To determine if r is significant:

1. Calculate $df = n - 2$

2. Look up critical value (cv) in table

3. If $|r| > cv$ correlation is significant

If $|r| < cv$ correlation is not significant



Our corr. is
definitely significant

Chapter 12: Correlation Coefficient Significance

Practice Math 10

Below are given the number of data pairs and the correlation coefficient for a set of data. In each case, use the table on p 262 (Blue) to determine whether the correlation coefficient is significant. Draw a number line graph for each.

1. $n=16$ $\text{corr} = 0.31$

$$CV = 0.497$$

NOT sig.

2. $n=32$ $\text{corr} = -0.28$

$$CV = 0.349$$

NOT sig.

3. $n=5$ $\text{corr} = -0.83$

$$CV = 0.878$$

NOT sig

4. $n=10$ $\text{corr} = -0.83$

$$CV = 0.632$$

sig.

5. $n=20$ $\text{corr} = 0.47$

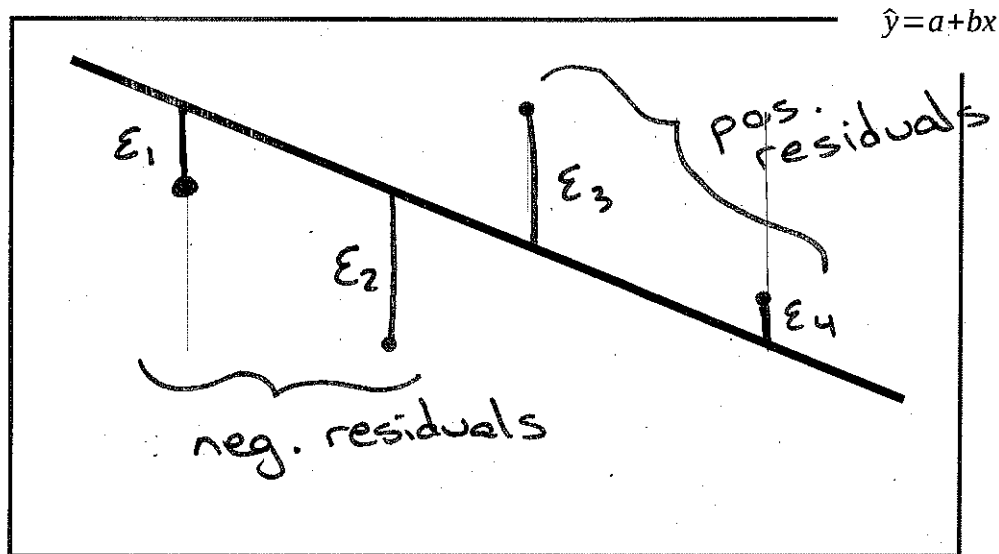
6. $n=12$ $\text{corr} = 0.47$

When does it make sense to use
to make predictions?

Making Predictions – Some points to remember

- If the correlation coefficient is significant,
you can use the line for prediction
- If you make predictions outside of the
domain, the results are unreliable

Outliers



The directed vertical distance of each data point from the line is called its **residual** (or **error**)

- If the point is above the line, the residual is positive
- If the point is below the line, the residual is negative

The calculator then calculates a critical value for the line.

We compare the residuals (absolute value) to the critical value. If any residual is greater than the critical value, the data value it belongs to is an outlier.

To calculate Outliers for Linear Regression

- Enter data in L1 and L2 **clear L3 & L4*

$$\begin{array}{cc} & x & y \end{array}$$
- Run OUTLIER program

Program

Arrow down to OUTLIER

ENTER

ENTER

- **Note: If you ever get an error message, select QUIT and then press ENTER.**
- Take abs value of residuals listed. Compare with given critical value.
- When done press ENTER to exit program
- Look up points in L1, L2 corresponding to identified residuals

Note: The Outlier Program can only be used in Chapter 12, not for Chapter 2 or your project