Ch 12 Linear regression and correlation

1) Analyze matched pair data to determine if there is "correlation" between the matched pair (x, y).
2) Determine if the correlation is linear.
3) Use the linear correlation result to predict y value give a x value.

Ex1. Given the matched pair sample data below. Can we conclude correlation between height and shoe size?

Adult	A	B	C	D	E
height	67"	70"	71"	65"	73"
shoe size	8	12	10.5	7.5	11

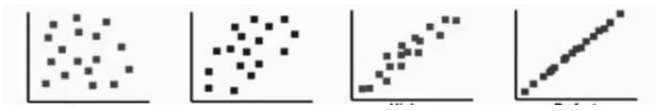
Ex2. Given the matched pair sample below, can we conclude correlation between shoe size and math scores?

children	A	B	C	D	E	F
shoe size	8	10	8.5	9	11	8
scores	65	92	71	82	95	80

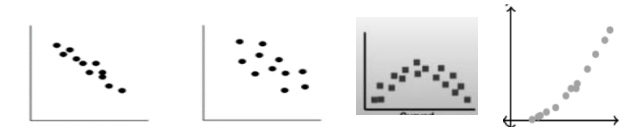
Ch 12.2 and 12.4 Scatter plot and correlation

Terms:
Correlation: Correlation between matched pair data (x, y) exists when values of y are associated with the values of x. Note: **correlation does not imply causation.**

Tools to study correlation:
1**) Graphical**: scatter plot. Each pair of (x, y) is plotted as one point on a graph. If a systematic pattern exists, there is correlation between x and y. Note: the pattern can be linear or non-linear.
2) **Mathematical**: use (x, y) sample data to calculate a correlation coefficient (r) . Value of r is used to determine if linear correlation exists and the strength and type of linear correlation.
Scatter plot examples



no correlation weak positive strong positive prefect positive



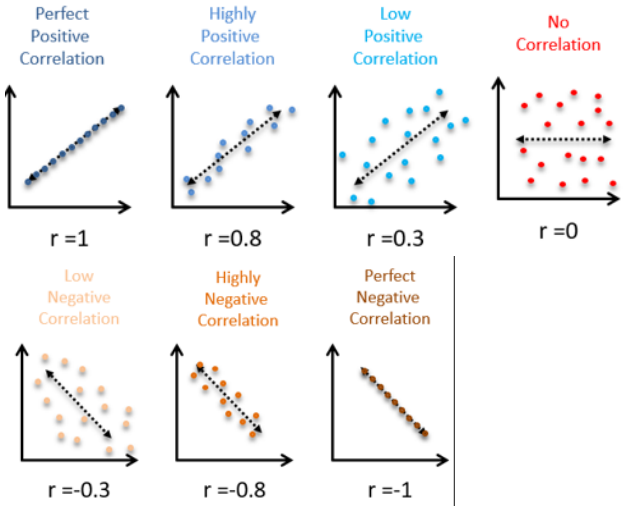
strong negative weak negative non-linear correlations

Construct scatter plot: Enter x and y data to statdisk in two different columns. Data/scatter plot/
Select x and y columns. Uncheck show regression line.
copy the scatter plot by labeling the axis and axis title.
Correlation coefficient (r)

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \times \sqrt{n(\sum y^2) - (\sum y)^2}}$$

The value shows how strongly the matched pair data x, y related to each other linearly.
Use Statdisk, enter data to 2 different columns. Analysis/Correlation and Regression. Enter significance, select x and y columns, Evaluate. output under "correlation result"
r = is the correlation coefficient, critical r is the critical threshold for evidence of linear correlation.
p-value is the probability of getting the sample under the H0 assumption of no linear correlation.

Properties of r:
1) between -1 and 1. r = 0 means no linear correlation. r =1 means perfect linear correlation.
2) If |r| is close to 1, there is strong linear correlation.
If |r| is close to 0, there is weak linear correlation.
3) r > 0, correlation is positive, x increase, y increase.
r < 0, correlation is negative, x increase, y decrease.
Relationship between scatter plot and correlation coefficient r.



Use Guess correlation game to understand relationship between r and scatter plot.
https://istics.net/Correlations/

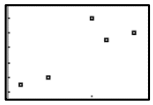
To determine if matched pair (x, y) has linear correlation:
Step 1: Check scatter plot, If non -linear pattern exists, conclude no linear correlation.
Step 2:
Method 1: Use Hypothesis test method with a given α.
ρ = correlation coefficient for population.
r = correlation coefficient for sample.
H0: ρ = 0 (no linear correlation) Ha: ρ ≠ 0
Use Analysis/Correlation and Regression to find p-value.
P-value ≤ α Reject H0, conclude linear correlation

p-value > α Fail to reject H0, conclude no linear correlation.
Method2: Compare r and critical value.
Use Analysis/Correlation and Regression to find r and critical r.
If – critical r ≤ r ≤ +critical r , conclude no linear correlation.
If r < - critical r or r > + critical value of n and α, conclude linear correlation.
Note: Check scatter plot for non-linear correlation before deciding linear correlation. Do not depend on r only or p-value only.

Ex1. Determine if linear correlation exists between the following pairs of r and p-value given n and α. Assume scatter plots do not show any non-linear patterns.
a) r = – 0.823, critical r = ±754
Since r is < - 754, conclude there is linear correlation.
b) α = 0.05, p-value = 0.012
Since only p-value is given, use hypothesis testing method
0.012 < 0.05, Reject H0, conclude there is linear correlation.

Ex2. Determine if linear correlation exists between height and shoe size in the given matched pair data. Use α= 0.05

Adult	A	B	C	D	E
height	67"	70"	71"	65"	73"
shoe size	8	12	10.5	7.5	11

Enter data Statdisk data columns.

Data/Scatter plot/uncheck regression line.
Graph does not show non-linear pattern.
Analysis/Correlation and Regression/
Select height and shoe size columns.
Output: r = 0.8485, critical r = ±0.8783, p-value = 0.0692
Method 1: (p-value method)
0.0691 > 0.05, fail to reject H0, conclude no linear correlation.
Method 2: (critical value method)
0.8485 < 0.878, conclude no linear correlation.
Ex3. Determine if linear correlation exists between shoe size and math scores for 6 children. use α = 0.05

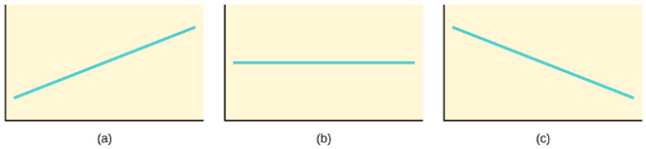
children	A	B	C	D	E	F
shoe size	8	10	8.5	9	11	8
scores	65	92	71	82	95	80


Enter to statdisk data columns.
Data/scatter plot/ select data columns, uncheck show regression line.
Scatter plot does not show non-linear pattern.
Analysis/Correlation and Regression/enter significance.
Select data columns. Evaluate
Output: r = 0.8758, critical r =±0.8114, p-value = 0.0222
Method 1: (p-value method)
p-value 0.0222 < 0.05 Reject H0, conclude linear correlation.
Method 2: (critical value method)
0.876 > 0.811, conclude linear correlation

Other properties of r:
1) r does not change if x and y switches.
2) r does not change when different units are used in x and/or y.

Ch 12.1 Linear Equations

A linear equation has the form of $y = a + bx$ where the graph is in the form of a line.
y is the dependent variable or explanatory variable.
x is the independent variable or predictor variable.
The goal is to use x to predict y.



when b > 0, the line has a positive slope, graph (a).
when b < 0, the line has a negative slope, graph (c).
when b = 0, the line is a horizontal line, graph (b).
The value of a is called y-intercept, it is the value of y when x = 0.

Ex1. The cost of ordering x items includes a fix shipping cost of $4.99 and $3.20 per item. Write the cost y and number of item x. Interpret the slope and y intercept of the equation.

Ans: equation is $y = 4.99 + 3.2x$
slope is 3.2 that is a cost of $3.2 per item.
y-intercept is 4.99 which is the cost when 0 item are ordered. 4.99 is not meaningful in real life.

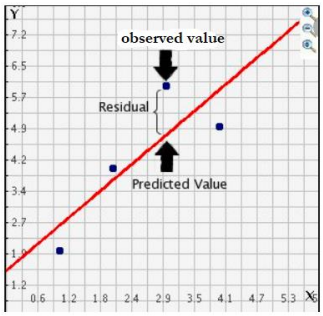
Ch 12.3 The regression equation

Match pairs sample can be used to find the equation of the "best fit line" also known as "linear regression line" or "least-squares line".
The line of best fit is used to predict y given a known value of x. (note: the prediction is a point estimate.)
Terms: Given a matched pair data (x, y)
x – explannatory variable, independent variable
y – response variable, predictor variable, dependent variable.
Line of best fit is $\hat{y} = b_0 + b_1 x$
where $\hat{y}$ is the predicted value of y. (y is the observed value in the data.)
where b0 is the y-intercept (predicted y value when x = 0)
b1 is the slope (rate of change of y per change of x)

Coefficient of determination (r^2)

r^2 shows the proportion of variation of y that can be predicted by change of x. It tells how good linear prediction is.

How to determine the line of best fit?
The criterion to determin the line that is better than all others is based on the vertical distances between the original data points and the regression line. The distance is also known as residuals.
Residual (ε)= observed y – predicted y = $y - \hat{y}$.



The best fit line is the line that satisify the "least-squares proprety" if the sum of squares of the residuals (SSE) is the smallest sum possible. (Calculus are used to build this.)

This also results in $(\bar{x}, \bar{y})$ always on the line.

Find equation of line of best fit:
Method 1: (use Statdisk)
- Enter match data to two columns of Statdisk.
- Analysis/Correlation and Regression/
Enter significance, select data columns, evaluate
output: b0 and b1 for equation is $\hat{y} = b_0 + b_1 x$
x is dependent variable, $\hat{y}$ = predicted value of y.
Method 2: use formula $b_1 = r\frac{s_y}{s_x}$, $b_0 = \bar{y} - b1\bar{x}$
Note: slope has the same sign as r.
use x to predict or estimate y.
the line of best fit is different if x and y switch.

Ex1. Given following matched pair data:

children	A	B	C	D	E	F
shoe size	8	10	8.5	9	11	8
scores	65	92	71	82	95	80

a) Find the best fit-line. Interpret slope.
Enter shoe size and scores to Statdisk.
- Analysis/Correlation and Regression/
Enter significance, select data columns, evaluate.
output: b_0 = 3.861, b_1=8.474 (round to 3 dec. places)
r^2 = 0,767 or 76.7%
$\hat{y} = 3.861 + 8.474x$ is the linear regression line.
The score will increase 8.574 points for every increase in shoe size of the child.
b) Find correlation of determination. Interpret in the context of this problem.
Ans:
r^2 = 76.7% means 76.7% of variation in math scores can be predicted by the variation in shoe size.

Ch 12.5 Prediction

Criteria for using the line of best fit to predict y:

1) Scatte plot indicates a linear pattern with no other non-linear patterns or outliers.
2) (x, y) are matched-pair and linearly correlated. Scatter plot does not show non-linear patterns.
3) x is within the prediction domain for intrapolation. The range of x values in the sample is the appropriate domain.
4) For each fixed value of x, the corresponding values of y have a normal distribution. (loose requirement)

Find best prediction for y.

Step1: Find the Linear regression equation, p-value and r and the scatter plot.
Enter match pairs data to statdisk columns.
Analysis/Correlation and Regression/ Enter significance, select data columns.
Output: r, critical r, p-value, b0 and b1, r^2 and scatter plot. $\hat{y} = b0 + b1x$ is the regression line.
Step 2: check scatter plot linear pattern and inspect if there is any non-linear pattern or outliers.
Step 3: Determine if (x, y) are linear related.
If p-value ≤ α , reject H0, conclude x, y are linear correlated
If p-value > α, conclude x, y are not linear correlated.
OR
If r outside the range of -critical r and +critical r, conclude x, y are linear correlated.

Step 4: Find the best predicted y.
If x,y are linear correlated, use the linear regression equation to find the best predicted y, $\hat{y}$.
$\hat{y} = b0 + b1(x)$
If x, y are not linear correlated, use $\bar{y}$ (mean of y) as best predicted y.
To find $\bar{y}$, use Statdisk/ Explore Data/ to find mean of y.

How good is the prediction

The correlation of determination, r^2 descibes how good the linear regression is in predicating the variation of y. The higher the correlation of determination, the better is the prediction.

Ex 1:

Given the following matched pair data:

Screen time	2.5	3.5	1.3	1.5	2
sleep time	7	6	10	6	8

Use the information to find the best predicted value of sleep time if the screen time is 3 hours. Use α=0.05

1) Find regression line equation, p-value,r, scatter plot.

Enter data to 2 columns of statdisk.

Use Analysis/Correlation and Regression/ Enter significance, select data columns.

Output: r = -0.579, critical r = ±0.878, p-value=0.3061, b0= 9.774, b1=-1.099, scatter plot in other tab.

Linear regression equation is $\hat{y} = 9.774 + 1.099x$

2) Check Scatter plot

There is no systematic non-linear pattern. There seems to be a negative weak correlation.

3) Determine if x, y are linear correlated.

since r = -0.579 is between -0.878 and +0.878, conclude no linear correlation.

OR : since p-value (0.306) > 0.05, conclude no linear correlation.

4) Since x, y are not linearly correlated, the best predicted value is mean of y.

Statdisk/Data/Explore data/select sleep time column mean = 7.4. So best predicted y when x = 3 hour is 7.4 hours

Ex 3. Given matched pair data for 7 students' study hour and final exam scores.

Study hour	7	6	4	3	8	10	3
final scores	95	90	72	78	90	89	66

Use α= 0.05 to predict a student's final score based on study hour of 6 hours.

Step 1) Find linear regression equation, p-value ,r and scatter plot.

Enter study hour and final scores to statdisk.

Analysis/Correlation and Regression/enter significance = 0.05, select data columns.

Output: r = 0.789, critical r = ±0.754, p-value = 0.0351, b0 =64.017, b1 = 3.217.

So $\hat{y} = 64.017 + 3.217x$ is the linear regression line.

Step 2) Check scatter plot:

Check scatter plot tab.

There is no non-linear patterns or outliers. The correlation is not very strong.

Step 3) Determine if x and y are linearly correlated.

Since p-value < 0.05 reject H0, conclude x and y are linearly correlated. OR r = 0.789 is outside the range of -0.754 and +0.754, so there is linear correlation.

4) Since x, y are linearly correlated, use linear regression line to find best prediction of score.

$\hat{y} = 64.017 + 3.217(6) = 83.3$

Best predicted scores = 83.3

b) Can the line of best fit equation be used to find predicted scores when study hour is 0. Explain.

Since 0 is not within the prediction domain, the regression line should not be used for prediction.

Ex4:

Given x, y are matched pair data with no non-linear pattern in scatter plot with $\bar{x}$=3.3 and $\bar{y} = 3.1$

The line of best fit is $\hat{y} = 2.1 + 0.32x$.

Correlation r = 0.82, and critical r= 0.754, find the best predicted y when x is 2.5 at α= 0.05.

Since r = 0.82 > 0.754 so x, y are linear correlated.

The best predicted y is from the linear regression line = 2.1+ 0.32(2.5) =2.9.

Ex5:

Given x, y are matched pair data with no non-linear pattern in scatter plot with $\bar{x}$=6.5 and $\bar{y} = 1.9$

The line of best fit is $\hat{y} = 2.1 + 0.32x$.

Given correlation p-value = 0.11, find the best predicted y when x is 5 at α= 0.05.

Since p-value 0.11 > 0.05 so x, y are not linear correlated.

The best predicted y is the mean of y = 1.9 instead of using the linear regression line.