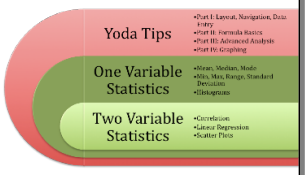
Statistical Analysis

Part II:
Two Variable Statistics

correlation (r) | linear regression (r^2) | scatter plots

...continued.



Analyse

r^2 +1 0

Two Variable Statistics: Relationships

Linear Regression or "line of best fit" (r^2)

Regression lines are "mathematical models". They are used for predictions and estimations. The "line of best fit" runs along a path, such that no other line can produce less area in its squares of error.

Excel Notation

=slope() =intercept()

How is it formed?

How is it used?

ACHTUNG!

Correlation ≠ Causation

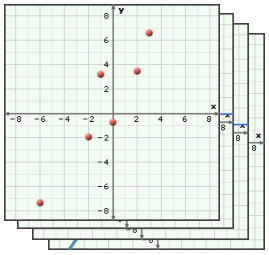
Analyse

r^2 +1 0

Two Variable Statistics: Relationships

"Goodness of fit"

What exactly does the value of r^2 measure?



Analyse

r^2 +1 0

Two Variable Statistics: Relationships

Wise regression analysis tips

Practice Assignment

Use the IRHS data file. Take the following grade 9 student at mid-term of semester one:

Courses	Absences
Math	8
English	6
Science	4
Business	3

If the second half of the semester proceeds just like the first did with the same rate of absences, use the IRHS data file to build a model that will predict this student's overall average at the end of this semester. Comment on the reliability or strength of the model, and note any significant findings.

tun

Linear Regression Exercises

Complete each of the questions below using both Fathom 2 and the TI-83.

1. The attendance and marks for ten students are shown below. Each column represents one student.

Absences	3	6	1	13	10	21	8	4	4	6
Grade	85	73	98	96	66	44	58	88	63	77

a. Complete a scatter plot of these data. Properly label all axes and indicate the scale. Circle the outlier.
b. Determine the correlation coefficient and the line of best fit.
r = _____ Line of best fit: _____
c. Determine the correlation coefficient and the line of best fit with the outlier removed.
r = _____ Line of best fit: _____
d. A student has 9 absences. Use both models to predict their grade.
Model b = _____ Model c: _____
e. Is your line accurate? Explain.

2. As his children got older, Mr. Merner's expenses (expressed below in hundreds of dollars per month) increased.

Time(yrs)	0	1	2	3	4	5	6	7	8	9
Expenses($)	5	6	8	11	23	19	22	27	35	41

a. Complete a scatter plot of these data. Properly label all axes and indicate the scale. Circle the outlier.
b. Determine the correlation coefficient and the line of best fit.
r = _____ Line of best fit: _____
c. Determine the correlation coefficient and the line of best fit with any outliers removed.
r = _____ Line of best fit: _____
d. Use both models to predict Merner's budget in the 23rd year.
Model b = _____ Model c: _____
e. Is your line accurate? Use the TI-83 to generate a different type of model (exponential, power, or quadratic). Is there another type that is more accurate?

3. Identify any outliers in the following sets of data and explain your choices.

a)

X	25	34	43	55	92	105	16
Y	30	41	52	66	18	120	21

b)

X	5	7	6	6	4	8
Y	304	99	198	205	106	9

4. Refer to question 3.
a) Perform a linear regression analysis to generate the line of best fit for each set of data in question 3.
b) Repeat the linear regressions in part a, leaving out any outliers.
c) Compare the lines of best fit in parts a and b.

5. The following table lists the heights and masses for a group of fire-department trainees.

Height (cm)	Mass (kg)
177	91
185	88
173	82
169	79
188	87
182	85
175	79

a) Create a scatter plot and classify the linear correlation.
b) Apply the method of least squares to generate the equation of the line of best fit.
c) Predict the mass of a trainee whose height is 165cm.
d) Predict the height of a 79 kg trainee.
e) Explain any discrepancy between your answer to part d and the actual height of the 79 kg trainee in the sample group.

6. Market research has provided the following data on the monthly sales of a licensed T-shirt for a popular rock band.

Price ($)	Monthly Sales
10	2500
12	2200
15	1600
18	1200
20	800
24	250

a) Create a scatter plot for these data.
b) Use linear regression to model these data.
c) Predict the sales if the shirts are priced at $19.
d) The vendor has 1500 shirts in stock and the band is going to finish its concert tour in a month. What is the maximum price the vendor can charge and still avoid having shirts left over when the band stops touring?

7. The Worldwatch Institute has collected the following data on concentrations of carbon dioxide (CO_2) in the atmosphere.

a) Use technology to produce a scatter plot of these data and describe any correlation that exists.
b) Use a linear regression to find the line of best fit for the data. Discuss the reliability of this model.
c) Use the regression equation to predict the level of atmospheric CO_2 that you would expect today.
d) Research current CO_2 levels. Are the results close to the predicted level? What factors could have affected the trend?

Year	CO2 Level (ppm)
1975	331
1976	332
1977	333.7
1978	335.3
1979	336.7
1980	338.5
1981	339.8
1982	341
1983	342.6
1984	344.3
1985	345.7
1986	347
1987	348.8
1988	351.4
1989	352.7
1990	354
1991	355.5
1992	356.2
1993	357
1994	358.8
1995	360.7