THE UNIVERSITY OF AUCKLAND NEW ZEALAND DEPARTMENT OF STATISTICS

r and R^2

- *What are they?*
- *When should I use them?*
- *How do Excel and GCs handle them?*
- *Why should I be careful with the Nulake text?*

Matt Regan (*et al*)
Department of Statistics
The University of Auckland

THE UNIVERSITY OF AUCKLAND NEW ZEALAND DEPARTMENT OF STATISTICS

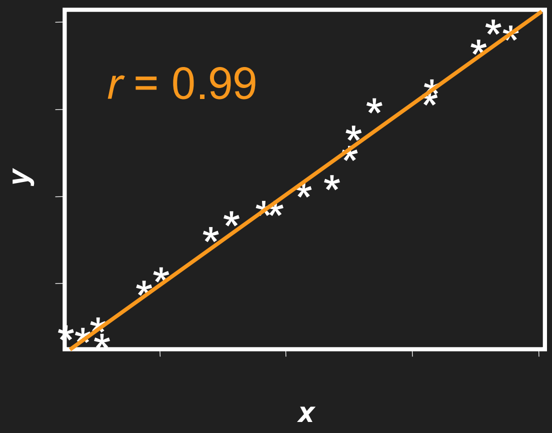
r – little r – what is it?

- r is the correlation coefficient between y and x
- r measures the strength of a linear relationship
  - r is a multiple of the slope

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
DEPARTMENT OF STATISTICS

r – when can it be used?

- Only use *r* if the scatter plot is linear

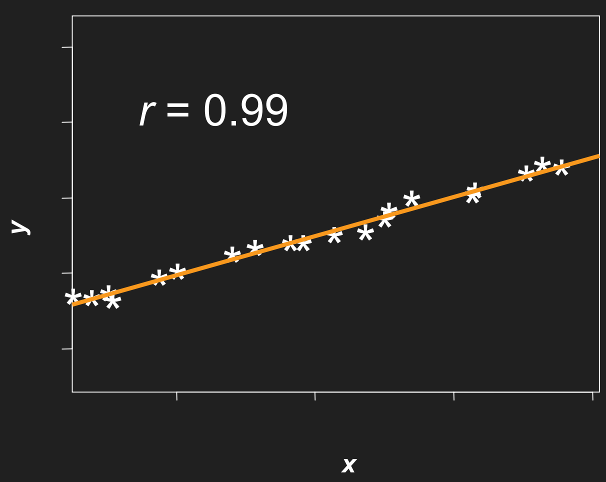


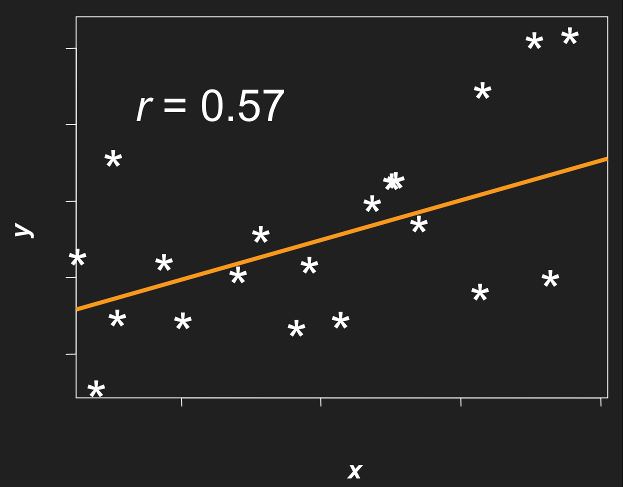
- Don't use *r* if the scatter plot is non-linear!

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
DEPARTMENT OF STATISTICS

r – what does it tell you?

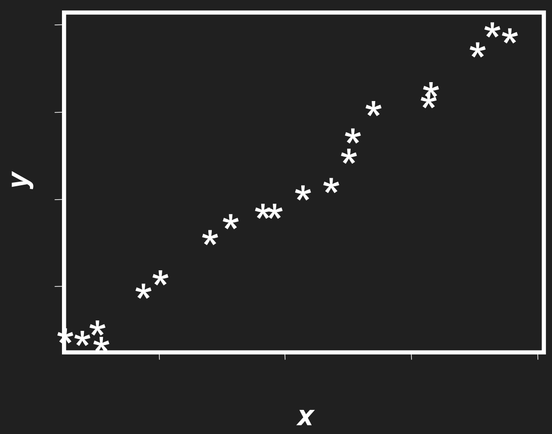
- How close the points in the scatter plot come to lying on the line

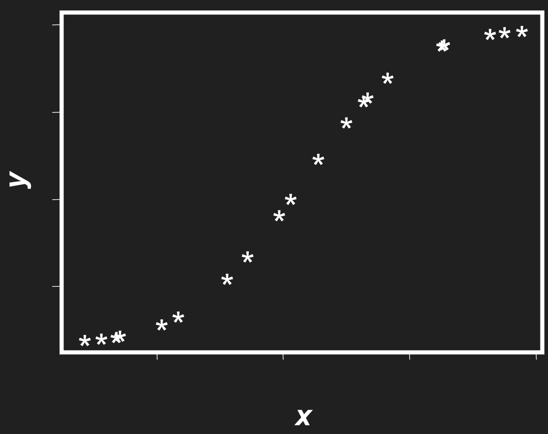




R^2 – *big* R^2 – what is it?

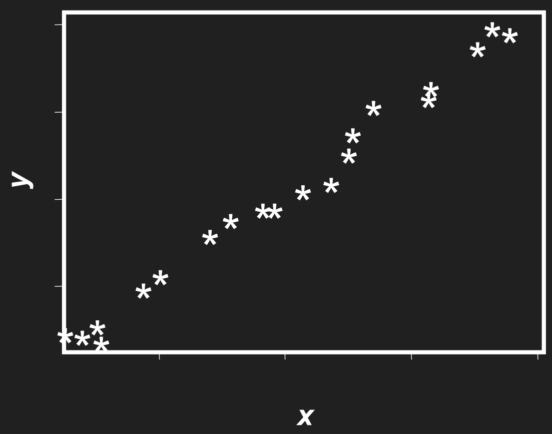
- R^2 is the coefficient of determination
  - Measures how close the points in the scatter plot come to lying on the fitted line or curve

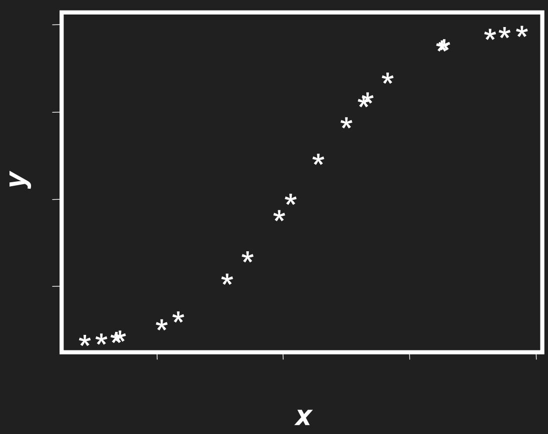




R^2 – *big* R^2 – when can it be used?

- When the scatter plot of y versus x is linear or non-linear



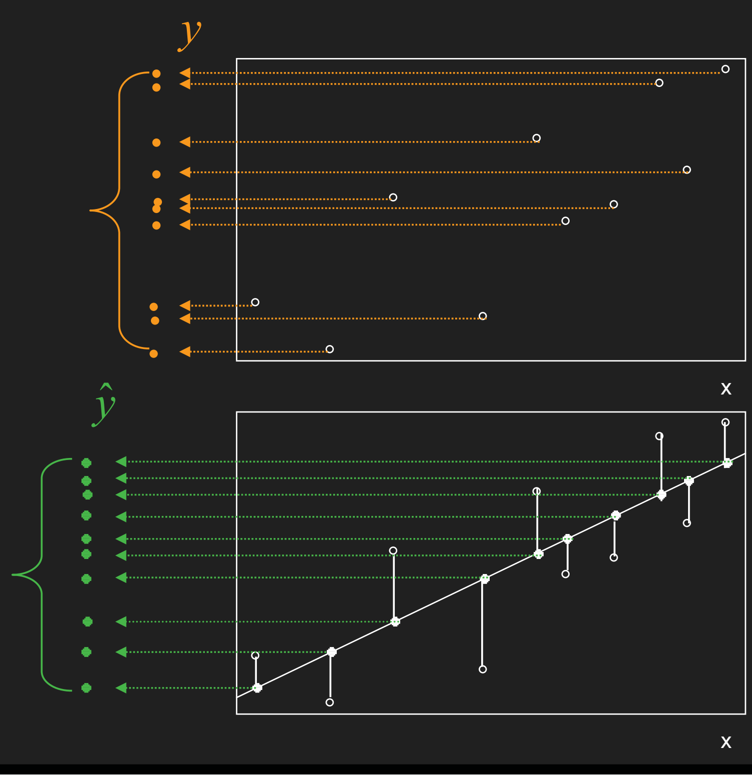


THE UNIVERSITY OF AUCKLAND NEW ZEALAND DEPARTMENT OF STATISTICS

R^2 – what does it tell you?

Dotplot of the y's

Shows the variation in the y's



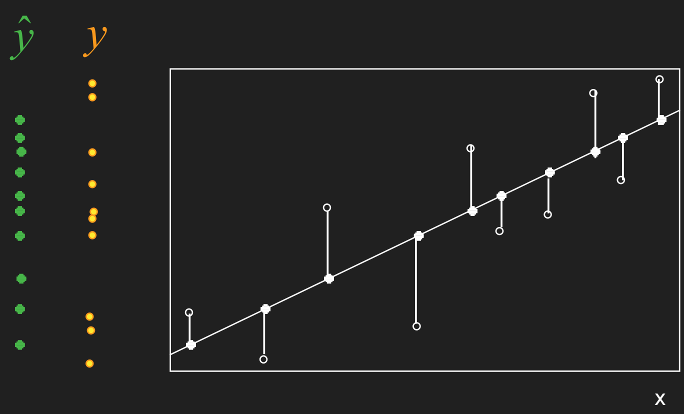
Dotplot of the $\hat{y}$'s
Shows the variation in the $\hat{y}$'s

THE UNIVERSITY OF AUCKLAND NEW ZEALAND DEPARTMENT OF STATISTICS

R^2 – what does it tell you?

Variation in the $\hat{y}$'s:

This amount of variation can be explained by the model



We see some additional variation in the y's.

The excess is not explained by the model.

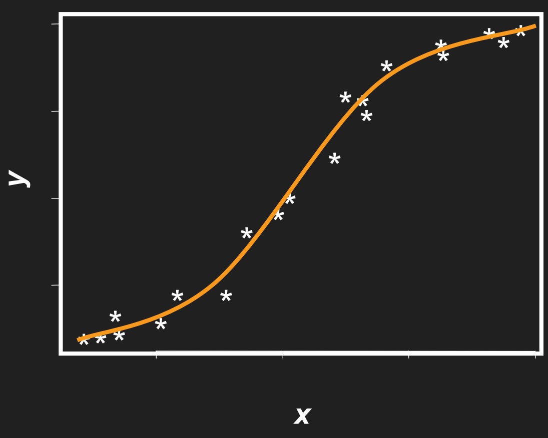
$$R^2 = \frac{\text{Variation in fitted values}}{\text{Variation in } y \text{ values}} = \frac{\text{Variation in } \hat{y}\text{'s}}{\text{Variation in } y\text{'s}}$$

R^2 – what does it tell you?

- When expressed as a percentage, R^2 is the percentage of the variation in Y that our regression model can explain
  - R^2 near 100% $\Rightarrow$ model fits well
  - R^2 near 0% $\Rightarrow$ model doesn't fit well

R^2 – what does it tell you?

R^2 = 90%



- 90% of the variation in Y is explained by our regression model.

R^2 – pearls of wisdom!

- R^2 and r^2 have the same value **ONLY** when using a linear model
- **DON'T** use R^2 to pick your model
  - **Use your eyes!**

R^2 and *Excel* & Graphics Calculators

Department of Statistics

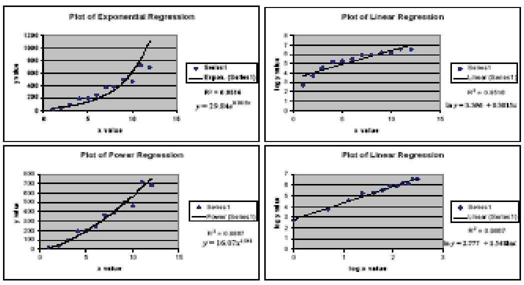
R^2 and *Excel* & Graphics Calculators

Excel and two graphics calculators were used on the following data. Their printouts are given below.

Data Set

x value	y value	log x value	log y value
1	15	0	2.70805
2	41	0.693147	3.713572
3	93	1.098612	4.532599
4	188	1.386294	5.236442
5	196	1.609438	5.278115
6	247	1.791759	5.509388
7	366	1.94591	5.902633
8	386	2.079442	5.955837
9	488	2.197225	6.190315
10	460	2.302585	6.131226
11	722	2.397895	6.582025
12	685	2.484907	6.529419

Excel



Note:

1. The R^2 values printed with the **exponential** and **power** plot curves for *Excel* are **wrong**. Do **not** use them.
2. *Excel* has based its R^2 values for the exponential and power plots on the linearly transformed data. For example, for the exponential curve, $y = ae^{bx}$, *Excel* computes R^2 using $\ln y = \ln a + bx$. It is wrong to do this. (*Excel's* R^2 values for the linear (semi-log and log-log) plots are correct.)

The University of Auckland

Graphics Calculators

Casio fx-9750G PLUS

Exponential Regression
y = a.e^bx

a = 29.8404023
b = 0.30152902
r = 0.92283291
r^2 = 0.85162058

Linear Regression
lny = bx + lna

b = 0.30152902
lna = 3.39586326
r = 0.92283291
r^2 = 0.85162058

Power Regression
y = a.x^b

a = 16.070768
b = 1.54826975
r = 0.9933374238
r^2 = 0.98671923

Linear Regression
lny = blnx + lna

b = 1.54826975
lna = 2.77700197
r = 0.99333742
r^2 = 0.98671923

Texas Instruments TI-83

Exponential Regression
y = a.b^x

a = 29.84040238
b = 1.351924358
r^2 = 0.8516205865
r = 0.9228329136

Linear Regression
lny = lna + (lnb)x

lna = 3.39586326
lnb = 0.3015290279
r^2 = 0.8516205865
r = 0.9228329136

Power Regression
y = a.x^b

a = 16.07076801
b = 1.548269753
r^2 = 0.9867192375
r = 0.9933374238

Linear Regression
lny = lna + blnx

lna = 2.77700197
b = 1.548269753
r^2 = 0.9867192375
r = 0.9933374238

Note:

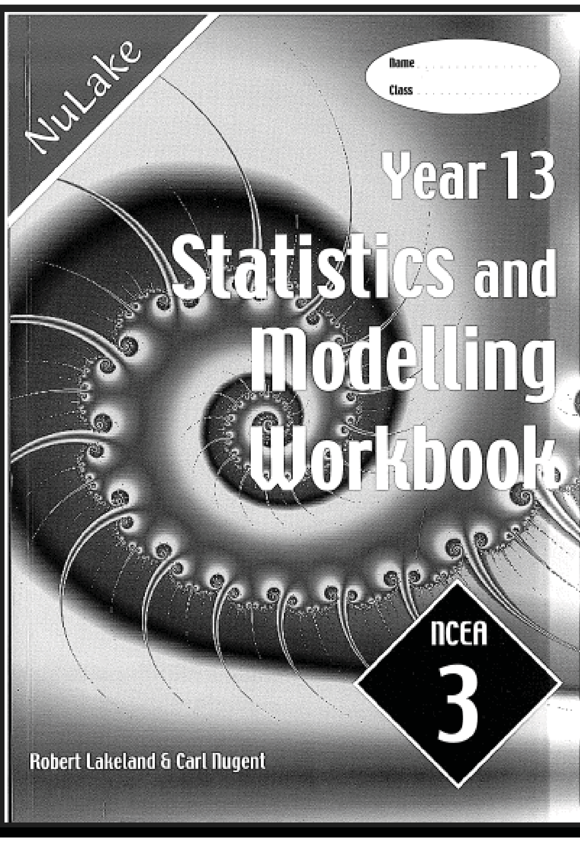
1. The r and r^2 values printed with the **exponential** and **power** regressions for these two graphics calculators are **wrong**. Do **not** use them. The sample correlation coefficient, r, should only be used with **linear** relationships.
2. The given r^2 values printed with the exponential and power relationships are **not** equal to their R^2 values. (These r^2 values have been based on and computed from the linearly transformed data, i.e., linear relationships.)
3. The r and r^2 values printed with the **linear** models are correct. In these linear cases, the given r^2 values are equal to R^2 values.

Use r, r^2, and R^2 values from *Excel* and graphics calculators with caution!

References:
Scott, A.J. and Wild, C.J. (1991) "Transformations and R^2", *The American Statistician*, 45, 127—129.
Kvalseth, T.O. (1985) "Cautionary Note About R^2", *The American Statistician*, 39, 279—285.

THE UNIVERSITY OF AUCKLAND NEW ZEALAND
DEPARTMENT OF STATISTICS

Year 13 Statistics and Modelling Workbook – Nulake



THE UNIVERSITY OF AUCKLAND NEW ZEALAND
DEPARTMENT OF STATISTICS

5.0 Statistical Investigation: page 224 – 261

5.0 Statistical Investigation

NCEA Achievement Standard
Achievement

- Select and analyse continuous bi-variate data.
- Data may be collected by candidates or provided. It should be data for which a linear model is appropriate. Where the data is provided it will involve more than one pair of variables from which the candidate selects a pair.
- The analysis will involve:
  - developing a purpose statement from the data selected
  - graphing data
  - using regression to establish a linear relationship between a pair of variables
  - describing the relationship between at least one pair of variables in context.

Merit

- Carry out an in-depth analysis of bi-variate data.

  The analysis will include some of the following:
  - comparing the relationship between more than one pair of variables
  - discussing the appropriateness of the model
  - interpreting correlation coefficients, r, and coefficients of determination, R^2, when appropriate
  - making predictions from regression equations (interpolation and/or extrapolation)
  - use of residuals
  - discussing the difference between correlation and causality when appropriate.

Excellence

- Report on the validity of the analysis.
- The report will include justified comments on some of the following:
  - method(s) of analysis
  - assumptions made
  - limitations
  - improving regression models eg discussing the effect of outliers, fitting piecewise or non-linear models
  - alternative approaches
  - data source or data collection method if the student collects own data
  - potential sources of bias
  - relevance and usefulness of evidence
  - how widely the findings can be applied.

5.1 Planning an Investigation

Planning an Investigation

The assessment for this topic requires you to carry out an investigation that involves the use of bivariate continuous data.

You are required to use the techniques that are described in this chapter to analyse the data and draw conclusions from the results.

You may be given data or will need to find appropriate data that will enable you to identify and formulate questions to help investigate relationships between at least 3 pairs of continuous variables.

There are a large number of sources for bivariate data e.g. Statistics NZ and the package PCINFOS for Schools to the internet and sites such as UNICEF (www.unicef.org).

You should spend a reasonable amount of time studying the data and identifying a purpose statement i.e. what are you trying to measure or show.

Once a purpose statement has been formulated then it is important to identify the variables within the data that will enable you analyse it effectively and to draw conclusions.

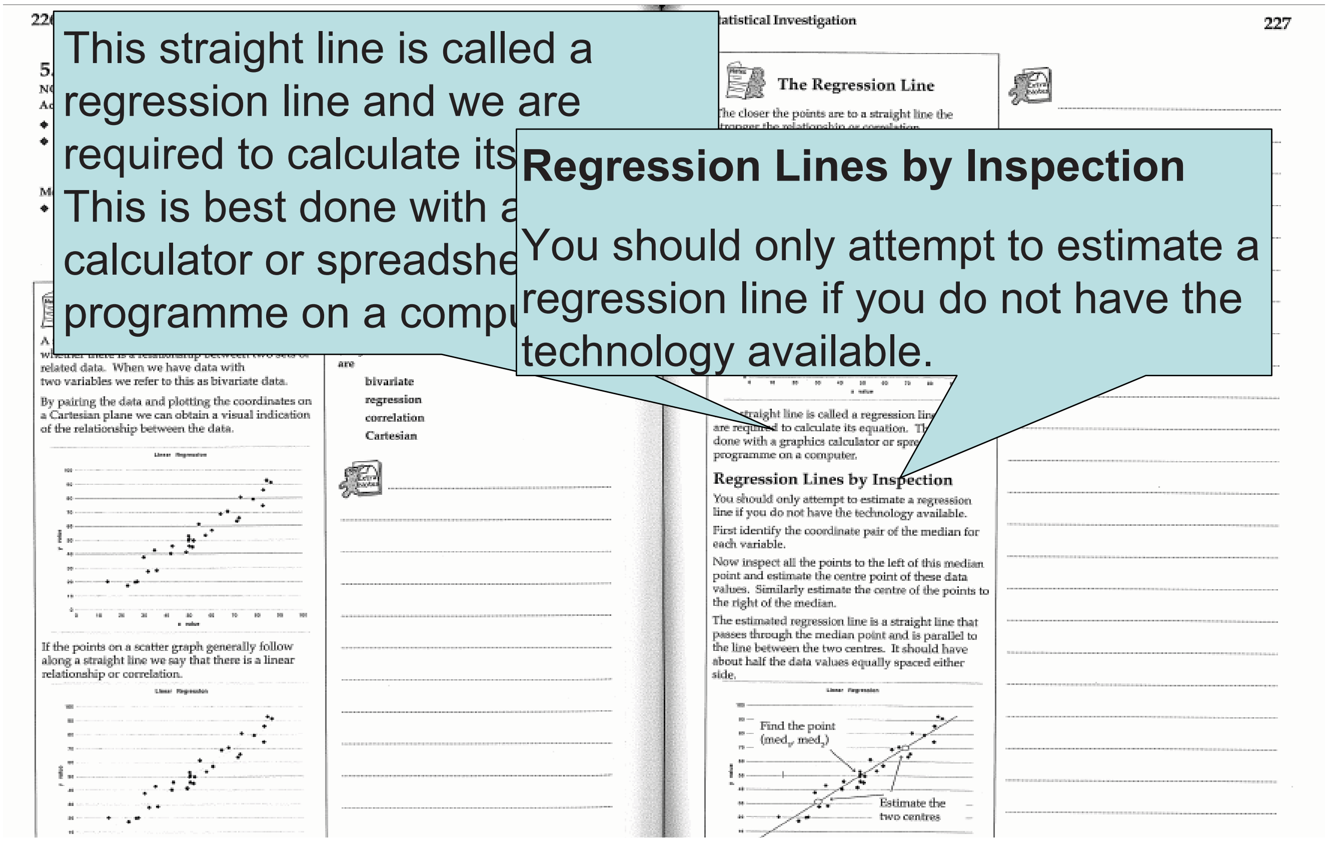
After formulating a purpose statement and identifying appropriate variables within the data it is necessary to describe your data collection or source and then explain how you intend to display and analyse the data.

The next step is the analysis itself. This should include the use of several model fitting methods, identification of trends, clusters, outliers etc. as well as the 'goodness of fit' of each of your models and interpretation of correlation coefficients. This may mean making predictions from regression equations (interpolation and extrapolation), improving regression models e.g. minimising the effect of outliers, fitting piece-wise or non-linear models.

Following the analysis you will need to draw conclusions, which could include predictions that can be made from the analysis you have conducted as well as the connection between correlation and causality and then relate this to your findings.

The final part of the investigation should be the critical evaluation. This should include such items as recommendations, limitations, possible

5.2 The Regression Line – p227

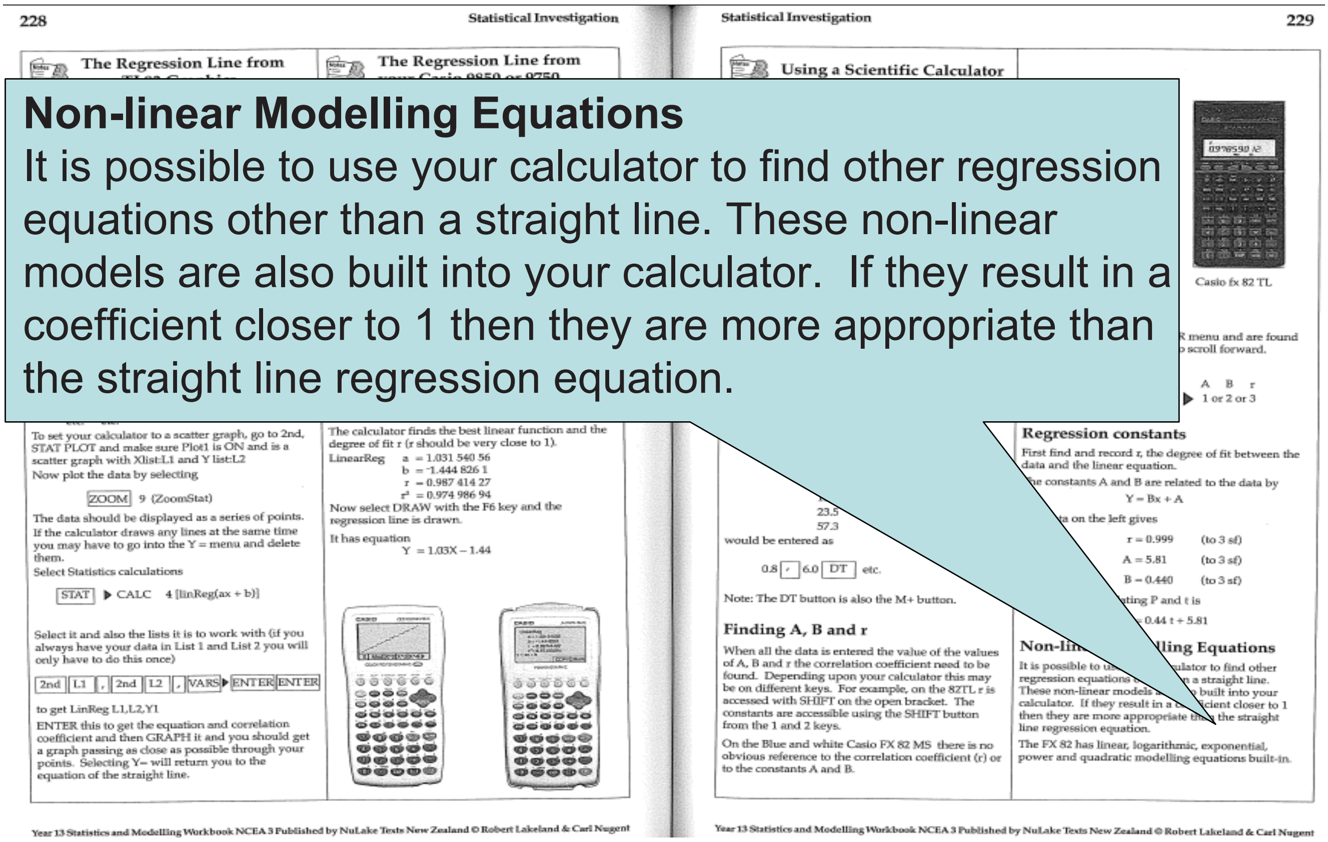


This straight line is called a regression line and we are required to calculate its equation. This is best done with a graphics calculator or spreadsheet programme on a computer.

Regression Lines by Inspection

You should only attempt to estimate a regression line if you do not have the technology available.

5.2 The Regression Line – p229



Non-linear Modelling Equations

It is possible to use your calculator to find other regression equations other than a straight line. These non-linear models are also built into your calculator. If they result in a coefficient closer to 1 then they are more appropriate than the straight line regression equation.

5.4 Non-linear Regression Models – p241

THE UNIVERSITY OF AUCKLAND NEW ZEALAND
DEPARTMENT OF STATISTICS

The United Nations social indicators web site records the following information about nine different countries. *Source http://www.unicef.org.*

	Argentina	Bolivia	Cuba	Dom. Rep.	Egypt	Fiji	Greece	Hungary	Ireland
Income (US $000)	5.3	0.9	2.5	2.5	1.4	2	10.7	5.2	26.7
Life Expectancy	73	64	76	66	69	70	78	72	77
Infant mortality	20	56	7	36	41	18	6	9	6
Sanitation	76	70	98	67	98	43	98	99	100

5.4 Non-linear Regression Models

NCEA Achievement Standard
Excellence

- Report on the validity of the analysis.
- The report will include justified comments on some of the following:
  - improving regression models eg discussing the effect of outliers, fitting piecewise or non-linear models.

Non-linear Regression Models

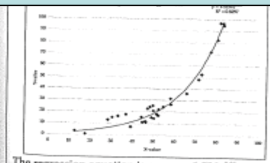
It is possible that the best model for some data is a non-linear regression line (e.g. a curve) such as an exponential or power function. All the technology aids (calculator and spreadsheets programmes) are able to model data with different models and by inspecting the correlation coefficient, the researcher should be able to determine the best model.

13. a)
b)
c)
d)
e)

Two s
a)
c)
d)
e)

14. a)
b)
c)
d)
e) poorly related?

Two statistics:
a)
c) r =
d)
e)



The regression equation is now $y = 1.526e^{0.03x}$ and $r = 0.8998$ ($r^2 = 0.8097$). The correlation coefficient has improved only slightly but our end points are now on the regression line.

Extrapolation and Interpolation

In regression you can be required to extrapolate (find values beyond the range of data you already have) and interpolate (find data values between data values you already have).

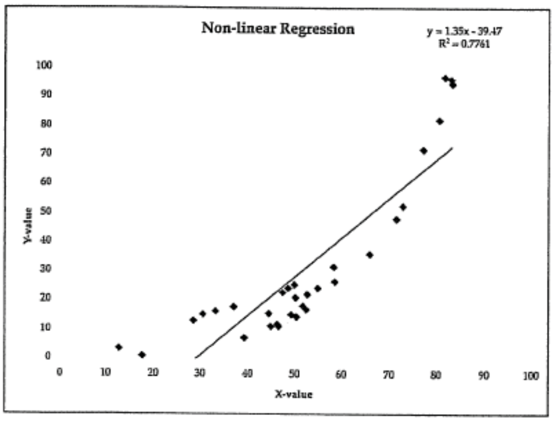
Find the best regression equation for your data and use this to do your extrapolation and interpolation.

Do not attempt to predict values on the basis of the scatter graph.

5.4 Non-linear Regression Models – p241

THE UNIVERSITY OF AUCKLAND NEW ZEALAND
DEPARTMENT OF STATISTICS

The manufactured data on this scatter graph follows a curve. If we had attempted to model this as a linear regression line we would get an acceptable correlation coefficient $r = 0.8810$ ($r^2 = 0.7761$) but the extreme values do not fit on the line.



poorly related?

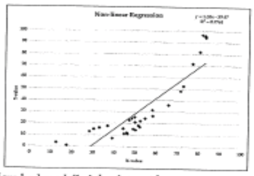
Two statistics:
a)
c) r =
d)
e)

Statistical Investigation 241

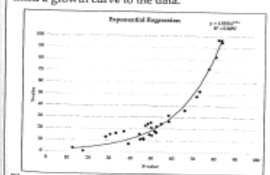
5.4 Non-linear Regression Models
NCEA Achievement Standard
Excellence
Report on the validity of the analysis.
The report will include justified comments on some of the following:
improving regression models eg discussing the effect of outliers, fitting piecewise or non-linear models.

Non-linear Regression Models
It is possible that the best model for some data is a non-linear regression line (e.g. a curve) such as an exponential or power function. All the technology aids (calculator and spreadsheet programmes) are able to model data with different models and by inspecting the correlation coefficient, the researcher should be able to determine the best model.

The manufactured data on this scatter graph follows a curve. If we had attempted to model this as a linear regression line we would get an acceptable correlation coefficient r =0.8810 (r² = 0.7761) but the extreme values do not fit on the line.



If we had modelled the data with an exponential function, the computer or calculator would have fitted a growth curve to the data.



The regression equation is now y = 1.526e^0.03x and r = 0.8998 (r² = 0.8097). The correlation coefficient has improved only slightly but our end points are now on the regression line.

New vocabulary used here

New words should be identified and defined (maybe accent marked). Words mentioned here are

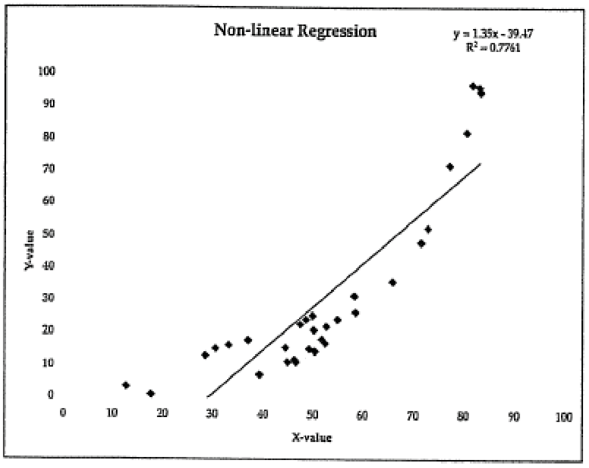
- extrapolation and extrapolate
- interpolation and interpolate
- non-linear
- piecewise
- outliers

Extrapolation and Interpolation
In regression you can be required to extrapolate (find values beyond the range of data you already have) and interpolate (find data values between data values you already have).
Find the best regression equation for your data and use this to do your extrapolation and interpolation.
Do not attempt to predict values on the basis of the scatter graph.

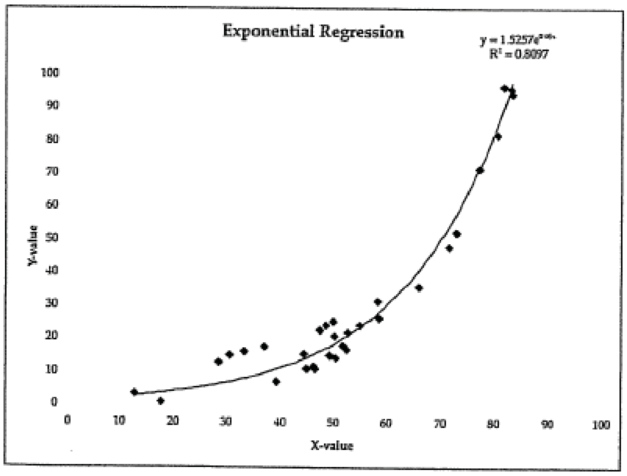
THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
DEPARTMENT OF STATISTICS

5.4 Non-linear Regression Models – p241

The manufactured data on this scatter graph follows a curve. If we had attempted to model this as a linear regression line we would get an acceptable correlation coefficient r =0.8810 ($r^2 = 0.7761$) but the extreme values do not fit on the line.



If we had modelled the data with an exponential function, the computer or calculator would have fitted a growth curve to the data.



The regression equation is now $y = 1.526e^{0.05x}$ and r = 0.8998 ($r^2 = 0.8097$). The correlation coefficient has improved only slightly but our end points are now on the regression line.

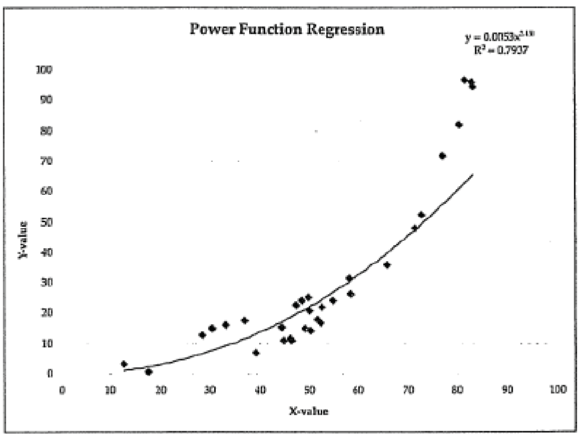
THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
DEPARTMENT OF STATISTICS

5.4 Non-linear Regression Models – p242

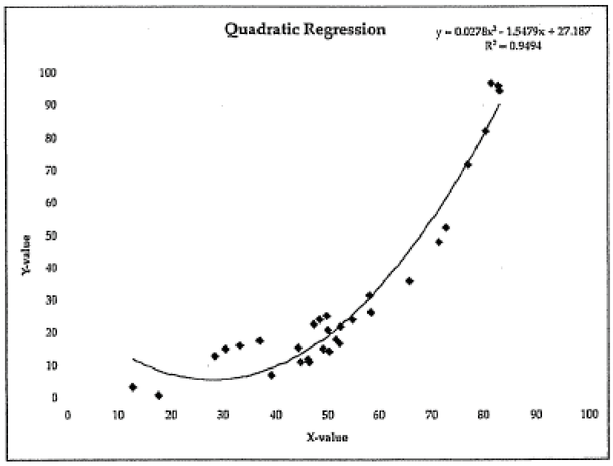
Non-linear Regression Models cont...

We should try other models watching both the correlation coefficient and the points particularly at the extremes.

Using the same manufactured data, a power function model comes up with the equation $y = 0.0053x^{2.131}$ and r = 0.8909 ($r^2 = 0.7937$) but the last few points are a long way from the curve so this model would be poor for extrapolation.



A quadratic model has a very good correlation coefficient for this data r = 0.9744 ($r^2 = 0.9494$) with the equation $y = 0.0278x^2 – 1.548x + 27.19$.



It follows most points closely but this model has problems with the points x < 40. You could fit other polynomial models or a logarithmic model.

Modelling: Drawing Graphs and adding a Trendline using a Spreadsheet – page 337

7.7 Modelling using Spreadsheets

NCEA Achievement Standard
Achievement

- Use a mathematical model involving curve fitting to solve a problem.
  Model will involve:
  $y = Ax^n$ or $y = Am^x$ or $y = Ae^{kt}$
  Data will be provided and the form of the model will be given.

Merit

- Use curve fitting to determine an appropriate model and solve a practical problem.
  Data will be raw generated by a practical situation.
  The choice of model will be either a power function or an exponential function or a piecewise function involving a combination of both, and will require:
  - selecting a model to test
  - constructing an appropriate graph
  - determining the equation of the model.

Excellence

- Justify the choice of mathematical model to solve a problem.

Drawing Graphs using a Spreadsheet

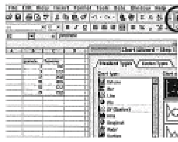
A computer spreadsheet programme can draw a scattergraph of any raw data and attempt to draw the line of best fit through the data. As well as being able to draw a straight line of best fit, the spreadsheet can also draw exponential and power functions that model the data.

It is important that a decision is made as to which relationship best models the data. At NCEA 3 this can be done by seeing how closely the curve is to the points.

It will also find the equation of this line of best fit.

The instructions in this workbook are for Microsoft Excel but other spreadsheet programmes have similar functions.

Enter the data on a blank sheet. Select just the data you are modelling then select the Chart wizard.



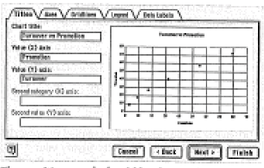
Select the Chart Wizard and then a Scattergraph

A spreadsheet programme is a tool that enables you to identify an appropriate model and find the equation of best fit.

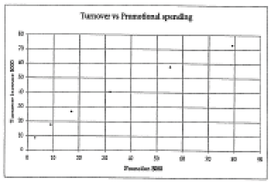
A spreadsheet can plot many types of trend lines. For this section we are only interested in exponential and power functions.

Drawing Graphs and adding a Trendline using a Spreadsheet Programme

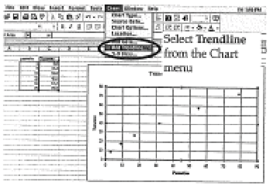
Make sure you label the graph and axes.



The resulting graph should be clear and appropriately labelled.

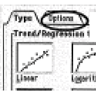


Now select the Trendline from the Chart menu.



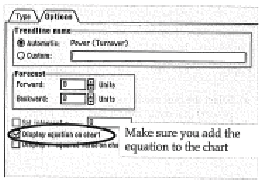
Select Trendline from the Chart menu

Select the function that you believe will best model the data (check how well it fits the data) and select Option to add the equation to the graph.



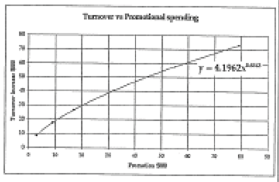
Select the desired function then **Options**

You also have the option of adding the regression coefficient (measure of fit) to the graph.



Make sure you add the equation to the chart

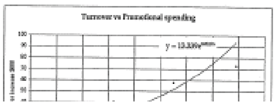
The resulting equation is the best fit for the particular model selected. If an inappropriate model is selected the programme will still calculate an equation of best fit.



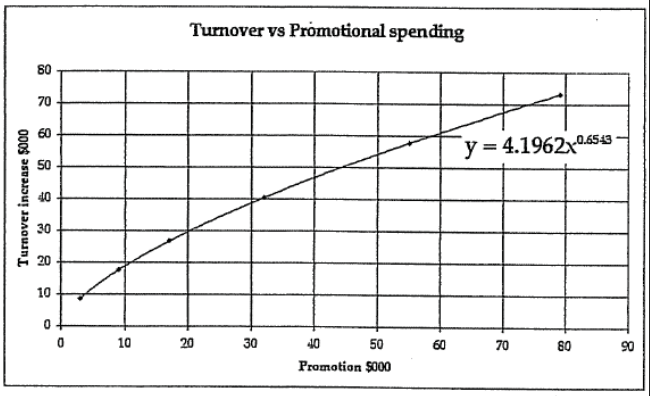
If the modelling function is not appropriate the graph will not pass through all the points.

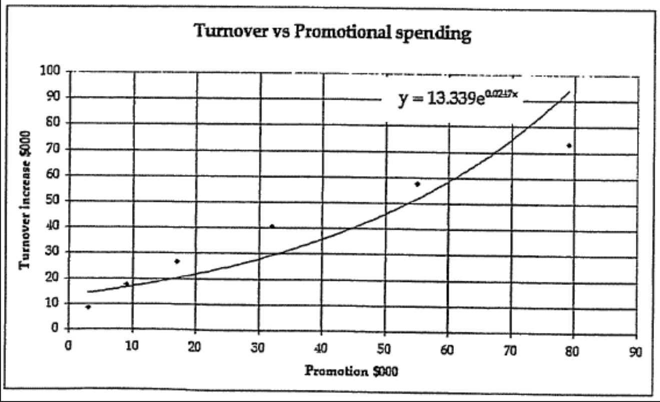
If only a few points are off the trendline then it may be just natural variation. If the trendline does not fit the data then it may be a case of an incorrect selection of the modelling function.

In this case the correct modelling function is a power function Turnover = $4.2\,x^{0.654}$. If we asked for an exponential trendline then the answer would obviously be a poor fit.



Modelling: Drawing Graphs and adding a Trendline using a Spreadsheet – page 337





If the modelling function is not appropriate the graph will not pass through all the points.

If only a few points are off the trendline then it may be just natural variation. If the trendline does not fit the data then it may be a case of an incorrect selection of the modelling function.

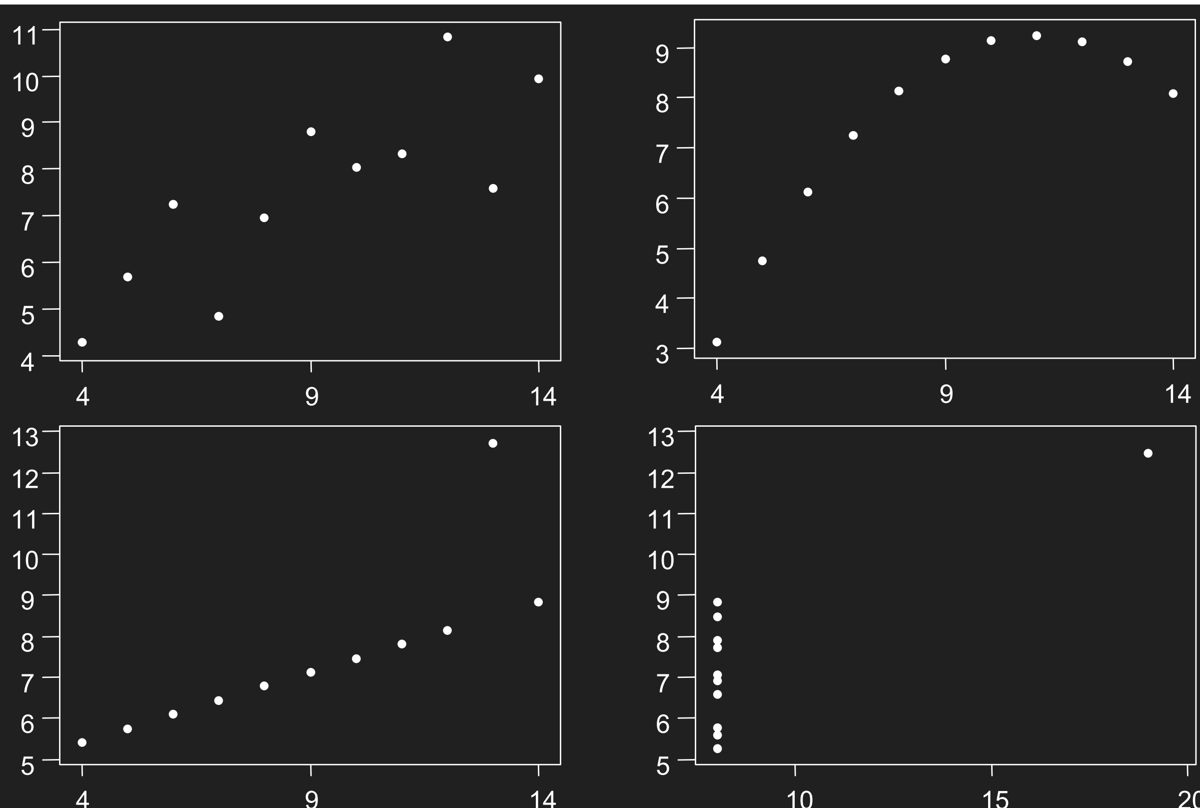
In this case the correct modelling function is a power function Turnover = $4.2\,x^{0.654}$. If we asked for an exponential trendline then the answer would obviously be a poor fit.

Correlation – What can go wrong?

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

N	11
Mean of x's	9.0
Mean of y's	7.5
Equation of regression line	y = 3 + 0.5x
Correlation coefficient (*r*)	0.82

Correlation – What can go wrong?



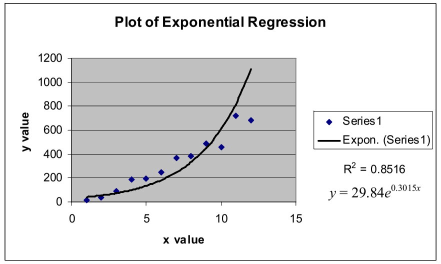
R^2 and *Excel* & Graphics Calculators

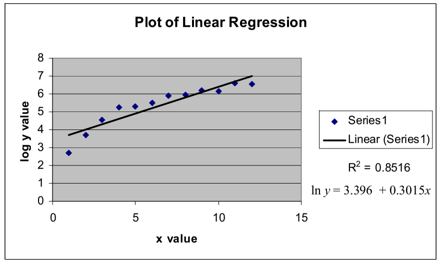
Excel and two graphics calculators were used on the following data. Their printouts are given below.

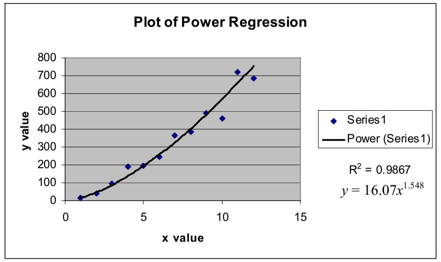
Data Set

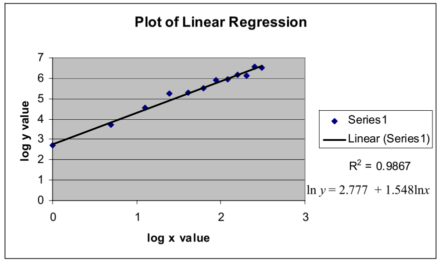
x value	y value	log x value	log y value
1	15	0	2.70805
2	41	0.693147	3.713572
3	93	1.098612	4.532599
4	188	1.386294	5.236442
5	196	1.609438	5.278115
6	247	1.791759	5.509388
7	366	1.94591	5.902633
8	386	2.079442	5.955837
9	488	2.197225	6.190315
10	460	2.302585	6.131226
11	722	2.397895	6.582025
12	685	2.484907	6.529419

Excel









Note:

1. The R^2 values printed with the **exponential** and **power** plot curves for *Excel* are **wrong**. Do **not** use them.
2. *Excel* has based its R^2 values for the exponential and power plots on the linearly transformed data. For example, for the exponential curve, $y = ae^{bx}$, *Excel* computes R^2 using ln$y$ = ln$a$ + bx. It is wrong to do this. (*Excel*'s R^2 values for the linear (semi-log and log-log) plots are correct.)

Graphics Calculators

Casio fx-9750G PLUS

Exponential Regression
y = a.e^bx

a = 29.8404023
b = 0.30152902
r = 0.92283291
r^2 = 0.85162058

Linear Regression
lny = bx + lna

b = 0.30152902
lna = 3.39586326
r = 0.92283291
r^2 = 0.85162058

Power Regression
y = a.x^b

a = 16.070768
b = 1.54826975
r = 0.9933374238
r^2 = 0.98671923

Linear Regression
lny = blnx + lna

b = 1.54826975
lna = 2.77700197
r = 0.99333742
r^2 = 0.98671923

Texas Instruments TI-83

Exponential Regression
y = a.b^x

a = 29.84040238
b = 1.351924358
r^2 = 0.8516205865
r = 0.9228329136

Linear Regression
lny = lna + (lnb)x

lna = 3.39586326
lnb = 0.3015290279
r^2 = 0.8516205865
r = 0.9228329136

Power Regression
y = a.x^b

a = 16.07076801
b = 1.548269753
r^2 = 0.9867192375
r = 0.9933374238

Linear Regression
lny = lna + blnx

lna = 2.77700197
b = 1.548269753
r^2 = 0.9867192375
r = 0.9933374238

Note:

1. The r and r^2 values printed with the **exponential** and **power** regressions for these two graphics calculators are **wrong**. Do **not** use them. The sample correlation coefficient, r, should only be used with **linear** relationships.
2. The given r^2 values printed with the exponential and power relationships are **not** equal to their R^2 values. (These r^2 values have been based on and computed from the linearly transformed data, i.e., linear relationships.)
3. The r and r^2 values printed with the **linear** models are correct. In these **linear** cases, the given r^2 values are equal to R^2 values.

Use r, r^2, and R^2 values from *Excel* and graphics calculators with caution!

References:
Scott, A.J. and Wild, C.J. (1991) "Transformations and R^2", *The American Statistician*, 45, 127—129.
Kvalseth, T.O. (1985) "Cautionary Note About R^2", *The American Statistician*, 39, 279—285.