DP IB Maths: AI HL

4.3 Further Correlation & Regression

Contents

4.3.1 Non-linear Regression

Non-linear Regression

What is non-linear regression?

- You have already seen that **linear regression** is when you can use a straight line to fit bivariate data
- **Non-linear regression** is when you can use a **curve** (rather than a straight line) to fit bivariate data
- In your exam the regression could be:
  - Linear: $y = ax + b$
  - Quadratic: $y = ax^2 + bx + c$
  - Cubic: $y = ax^3 + bx^2 + cx + d$
  - Exponential: $y = ab^x$ or $y = ae^{bx}$
  - Power: $y = ax^b$
  - Sine: $y = a\sin(bx + c) + d$

How do I find the equation of the non-linear regression model?

- Using your GDC:
  - Type the **two sets** of the data into your GDC
  - Select the **relevant model**
    - The exam question will tell you which model to use
  - Your GDC will calculate the **constants**
- You can use **logarithms** to **linearise exponential and power** relationships
  - Power: $y = ax^b$ then $\ln y = \ln a + b\ln x$
    - $\ln y$ and $\ln x$ will have a linear relationship
  - Exponential: $y = ab^x$ then $\ln y = \ln a + x\ln b$
    - $\ln y$ and x will have a linear relationship

Worked example

Scarlett and Violet collect data on the length of a film (x minutes) and the audience rating (y %).

x	75	93	101	107	115	124	132	140	171
y	83	75	51	38	47	56	76	91	70

a) Scarlett claims that there is a cubic relationship. Find the equation of a cubic regression model of the form $y = ax^3 + bx^2 + cx + d$.

Type the data into GDC and choose the cubic regression model

$a = -0.0005291...$ $b = 0.2030...$ $c = -24.89...$ $d = 1037.7...$

$$\boxed{y = -0.000529x^3 + 0.203x^2 - 24.9x + 1040}$$

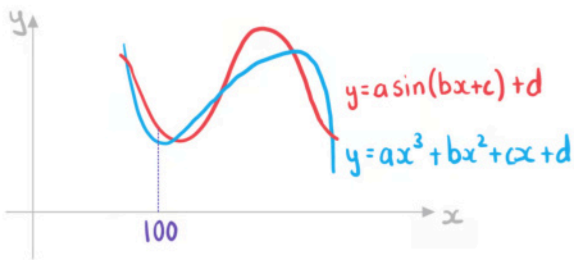
b) Violet claims that there is a sine relationship. Find the equation of a sine regression model of the form $y = a\sin(bx + c) + d$.

Type the data into GDC and choose the sine regression model

$a = 24.74...$ $b = 0.08030...$ $c = 2.086...$ $d = 69.49...$

$$\boxed{y = 24.7\sin(0.0803x + 2.09) + 69.5}$$

c) Whose model predicts a higher audience rating for a film which is 100 minutes long?



Using the cubic model $y = 49.640...$

Using the sine model $y = 53.690...$

$\boxed{\text{Violet's model predicts a higher rating.}}$

Least Squares Regression Curves

What is a residual?

- Given a set of *n* pairs of data and a **regression model** *y* = *f*(*x*)
- A **residual** is the **actual *y*-value** (from the data) **minus** the **predicted *y*-value** (using the regression model)
  - $y_i - f(x_i)$
- The **sum of the square residuals** is denoted by SS_{res}
  - $SS_{res} = \sum_{i=1}^{n} (y_i - f(x_i))^2$
- If you have two regression models using the **same data** then the one with the **smaller** SS_{res} **fits the data better**

What is a least squares regression curve?

- The **least squares regression curve** can be thought of as a "**curve of best fit**" *y* = *f*(*x*)
- For a **given type of model** the least squares regression curve **minimises the sum of the square residuals**
  - Your GDC calculates the constants for the least squares regression curves

Why is the sum of the square residuals not always a good measure of fit?

- If two models are formed using the **same number of pairs** of data then the sum of the square residuals is a **good measure of fit**
- If two models use **different number of pairs** of data then SS_{res} is **not always a good measure of fit**
  - The sum will increase with more pairs of data and so can no longer be compared against a data set with a different number of pairs
  - Compare the two scenarios
    - 10 pairs of data and the absolute value of each residual is 15 then $SS_{res} = 10 \times 15^2 = 2250$
    - 2250 pairs of data and the absolute value of each residual is 1 then $SS_{res} = 2250 \times 1^2 = 2250$
  - They have the same value of SS_{res} but the residuals in the second scenario are much smaller
- Your GDC may give you the **mean squared error**
  - $MSe = \frac{1}{n} SS_{res} = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$
  - This is a **better measure of fit**
  - You **do not need to know this** for your exam but it might help with your understanding

Worked example

Jet is the owner of a gym and he is testing different prices options. The table below shows the number of new members per month (M) and the price of a monthly membership (£p).

p	10	20	30
M	97	68	55

Jet believes that he can fit the data with either the model $M_1(p) = \dfrac{2700}{p+20}$ or the model $M_2(p) = \dfrac{2100}{p+10}$.

Jet wants to choose the model with the smallest value for the sum of square residuals.

Determine which model Jet should choose.

Calculate the predicted values

p	M	M_1	M_2
10	97	90	105
20	68	67.5	70
30	55	54	52.5

For M_1 : $SS_{res} = (97-90)^2 + (68-67.5)^2 + (55-54)^2 = 50.25$

For M_2 : $SS_{res} = (97-105)^2 + (68-70)^2 + (55-52.5)^2 = 74.25$

$\boxed{\text{Jet should choose model } M_1}$

The Coefficient of Determination

What is the coefficient of determination?

- The **coefficient of determination** is a **measure of fit** for a model
  - If the coefficient of determination is 0.57 this means 57% of the variation of the *y*-variable can be explained by the variation in the *x*-variable
  - The other 43% can be explained by other factors
  - The higher this proportion the more the model fits the data
- The coefficient of determination is **denoted by R^2**
  - $R^2 \le 1$
  - $R^2 = 1$ means the model is a **perfect fit** for the data
  - The closer to 1 the better the fit
  - R^2 is usually greater than or equal to zero
    - R^2 can be negative but this is outside the scope of this course
- If the regression model is linear then the coefficient of determination is **equal to square of the PMCC**
  - $R^2 = r^2$ for linear models
  - Some GDCs will simply denote R^2 as r^2 due to its connection to the PMCC for linear models

How do I calculate the coefficient of determination?

- When finding the constants for regression models your **GDC might give you the value** of R^2
  - You will only be asked to calculate the coefficient of determination for models for which GDCs give the value of R^2
- The coefficient of determination can be calculated by
  - $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$
    - Where $SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$
  - You **do not need to know this** formula but it might help with your understanding

Does the coefficient of determination determine the validity of a model?

- If R^2 is close to 1 then the model fits the data well
  - However this alone **does not guarantee** that it is a **good model for the relationship** between the two variables
- Consider the scenario where there are big gaps between data points and a model which fits the data well
  - The model only fits the data at the data points
  - As there are gaps between the data points the model might not be a good fit for these areas
- Different types of models have **different number of parameters**
  - Therefore using different types of models to fit the same data will have **different levels of accuracy**
  - Linear models need **at least two pairs** of data

- Quadratic models need **at least three pairs** of data
- Cubic models need **at least four pairs** of data
  - Using four pairs of data will mean the cubic model will have $R^2 = 1$
    This is because the cubic graph will go through all four pieces of data – the value is likely to decrease as extra pairs of data are included
  - However this does not mean it is a better fit than the quadratic model
  - The quadratic model could be more accurate as it has one more pair of data than is needed

Worked example

Data is collected on the lengths of cheetahs (x metres) and their average running speeds (y ms^{-1}).

x	1.21	1.33	1.12	1.45	1.42	1.39	1.24	1.19	1.32
y	24.3	25.1	22.2	35.1	35.1	33.4	27.1	23.1	24.8

a) Find the equation of the least squares regression curve using:

(i) a quadratic model $y = ax^2 + bx + c$.

(ii) an exponential model $y = ab^x$.

Type the data into GDC and choose the :

quadratic regression model

$a = 140.9...$

$b = -322.6...$

$c = 207.5...$

$\boxed{y = 141x^2 - 323x + 208}$

exponential regression model

$a = 4.193...$

$b = 4.250...$

$\boxed{y = 4.19 \times 4.25^x}$

b) Based solely on the coefficients of determination, suggest which model is better fit for the data.

Find the coefficients of determination using GDC

Quadratic $R^2 = 0.86429...$

Exponential $R^2 = 0.80157...$

Based on the coefficients of determination, the quadratic regression model as its R^2 value is bigger.

4.3.2 Logarithmic Scales

Logarithmic Scales

What are logarithmic scales?

- **Logarithmic scales** are scales where intervals **increase exponentially**
  - A normal scale might go 1, 2, 3, 4, ...
  - A logarithmic scale might go 1, 10, 100, 1000, ...
- Sometimes we can keep the scales with **constant intervals** by **changing the variables**
  - If the values of x increase exponentially: 1, 10, 100, 1000, ...
  - Then you can use the variable **log *x*** instead which will have the scale: 1, 2, 3, 4, ...
  - This will change the shape of the graph
    - If the graph transforms to a straight line then it is easier to analyse
- **Any base** can be used for logarithmic scales
  - The most common bases are 10 and e

Why do we use logarithmic scales?

- For variables that have a **large range** it can be difficult to plot on one graph
  - Especially when a lot of the values are **clustered in one region**
  - For example: populations of countries
    - This can range from 800 to 1 450 000 000
- If we are interested in the **rate of growth** of a variable rather than the actual values then a logarithmic scale is useful

log-log & semi-log Graphs

What is a log-log graph?

- A **log-log graph** is used when **both scales** of the original graph are logarithmic
  - You transform both variables by taking logarithms of the values
- log y & log x will be used instead of y & x
- **Power graphs** ($y = ax^b$) look like **straight lines** on log-log graphs

What is a semi-log graph?

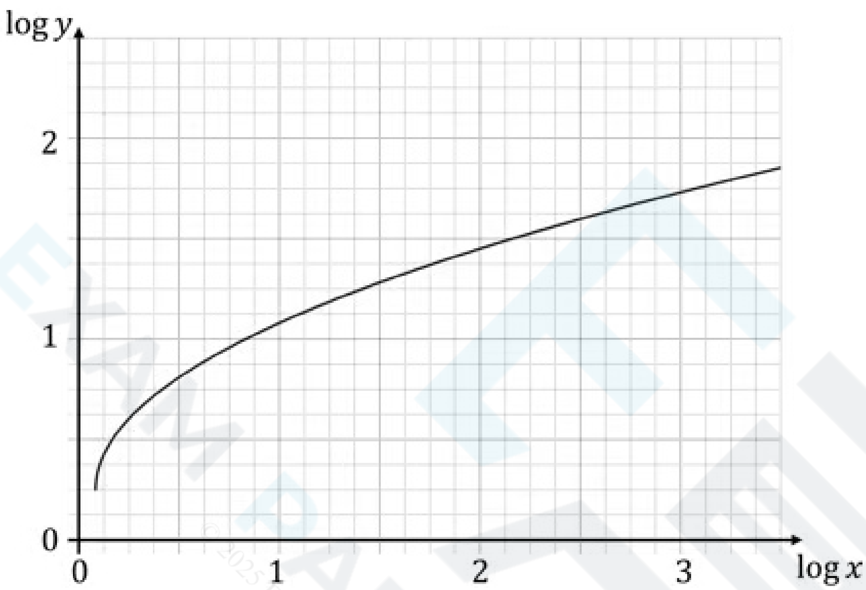
- A **semi-log graph** is used when **only one scale** (the y-axis) of the original graph is logarithmic
  - You transform only the y-variable by taking logarithms of those values
- log y will be used instead of y
- **Exponential graphs** ($y = ab^x$) look like **straight lines** on semi-log graphs

How can I estimate values using log-log and semi-log graphs?

- Identify whether **one or both of the scales** are logarithmic
- Identify the variable so that the scales have **equal intervals**
  - x : 1, 10, 100, 1000, ... use log x
  - For x : 1, e, e², e^3, ... use ln x
- If you are asked to estimate a value:
  - First find the value of any logarithms
    - For example: log y, ln x, etc
  - Use the graph to read off the value
  - If it is a value for a logarithm find the actual value using:
    - $\log x = k \Rightarrow x = 10^k$
    - $\ln x = k \Rightarrow x = e^k$

Worked example

The function $y = f(x)$ is drawn below using a log-log graph.

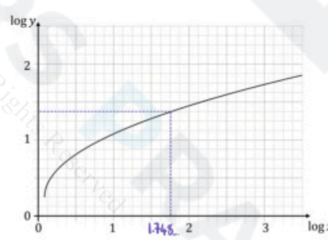


Show that when $x = 56$ the value of y is approximately 24.

Find $\log x$

$x = 56 \Rightarrow \log 56 = 1.7481...$

Use graph to find $\log y$



$\log y \approx 1.375$

Find y

Exponents & logarithms	$a^x = b \Leftrightarrow x = \log_a b$

$\boxed{y = 10^{1.375} = 23.71... \approx 24}$

4.3.3 Linearising using Logarithms

Exponential Relationships

How do I use logarithms to linearise exponential relationships?

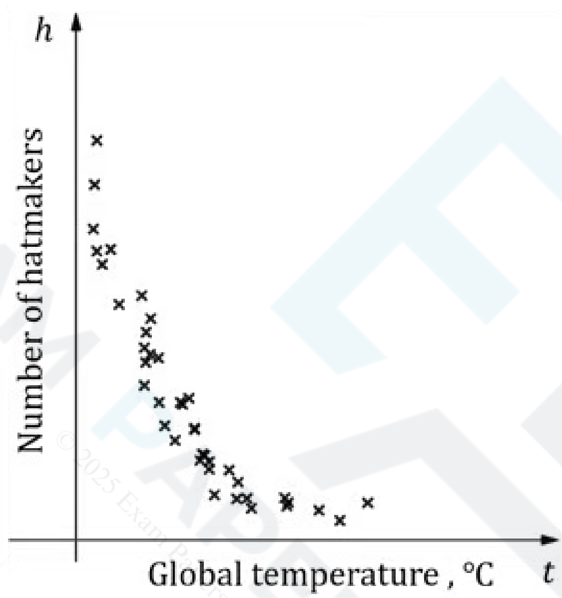
- Graphs of **exponential functions** appear as straight lines on **semi-log graphs**
- Suppose $y = ab^x$
    - You can take logarithms of both sides
        - $\ln y = \ln(ab^x)$
    - You can split the right hand side into the sum of two logarithms
        - $\ln y = \ln a + \ln(b^x)$
    - You can bring down the power in the final term
        - $\ln y = \ln a + x \ln b$
- $\ln y = \ln a + x \ln b$ is in linear form $Y = mX + c$
    - $Y = \ln y$
    - $X = x$
    - $m = \ln b$
    - $c = \ln a$

How can I use linearised data to find the values of the parameters in an exponential model *y* = *ab*^{x}?

- **STEP 1**: **Linearise** the data using $Y = \ln y$ and $X = x$
- **STEP 2**: Find the equation of the **regression line** of *Y* on *X*: $Y = mX + c$
- **STEP 3**: **Equate coefficients** between $Y = mX + c$ and $\ln y = \ln a + x \ln b$
    - $m = \ln b$
    - $c = \ln a$
- **STEP 4**: **Solve** to find *a* and *b*
    - $a = e^c$
    - $b = e^m$

Worked example

Hatter has noticed that over the past 50 years there seems to be fewer hatmakers in London. He also knows that global temperatures have been rising over the same time period. He decides to see if there could be any correlation, so he collects data on the number of hatmakers and the global mean temperatures from the past 50 years and records the information in the graph below.



Hatter suggests that the equation for h in terms of t can be written in the form $h = ab^t$. He linearises the data using $x = t$ and $y = \ln h$ and calculates the regression line of y on x to be $y = 4.382 - 1.005x$.

Find the values of a and b.

Write $h = ab^t$ in linearised form

$\ln(h) = \ln(ab^t) \Rightarrow \ln h = \boxed{\ln a} + t\ln b$

Compare coefficients

$y = 4.382 - 1.005x \Rightarrow \ln h = \boxed{4.382} - 1.005t$

$\ln a = 4.382 \Rightarrow a = e^{4.382} = 79.997...$ $\boxed{a = 80.0 \ (3sf)}$

$\ln b = -1.005 \Rightarrow b = e^{-1.005} = 0.36604...$ $\boxed{b = 0.366 \ (3sf)}$

Power Relationships

How do I use logarithms to linearise power relationships?

- Graphs of **power functions** appear as straight lines on **log-log graphs**
- Suppose $y = ax^b$
  - You can take logarithms of both sides
    - $\ln y = \ln(ax^b)$
  - You can split the right hand side into the sum of two logarithms
    - $\ln y = \ln a + \ln(x^b)$
  - You can bring down the power in the final term
    - $\ln y = \ln a + b\ln x$
- $\ln y = \ln a + b\ln x$ is in linear form $Y = mX + c$
  - $Y = \ln y$
  - $X = \ln x$
  - $m = b$
  - $c = \ln a$

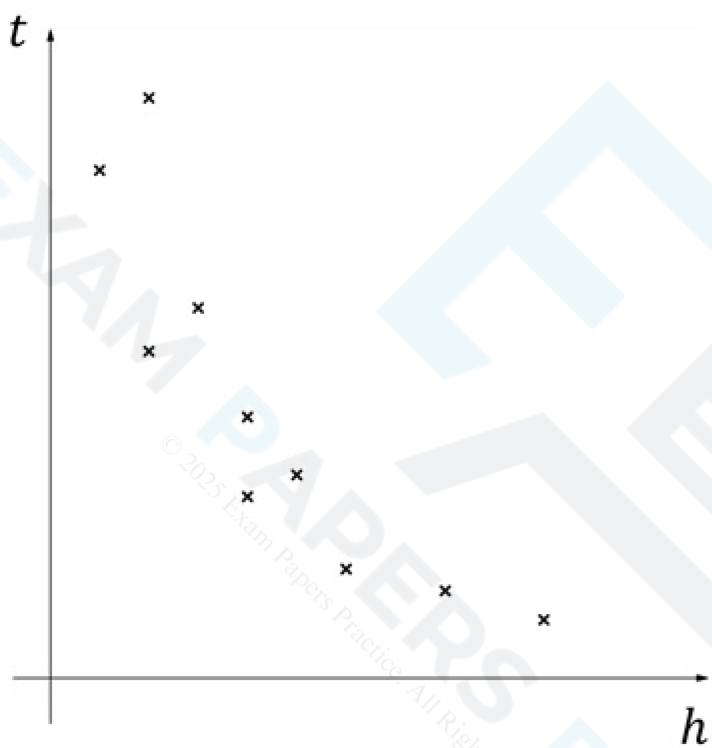
How can I use linearised data to find the values of the parameters in an power model *y* = *ax*^{b}?

- **STEP 1**: **Linearise** the data using $Y = \ln y$ and $X = \ln x$
- **STEP 2**: Find the equation of the **regression line** of *Y* on *X*: $Y = mX + c$
- **STEP 3**: **Equate coefficients** between $Y = mX + c$ and $\ln y = \ln a + b\ln x$
  - $m = b$
  - $c = \ln a$
- **STEP 4**: **Solve** to find *a* and *b*
  - $a = e^c$
  - $b = m$

Worked example

The graph below shows the heights, h metres, and the amount of time spent sleeping, t hours, of a group of young giraffes. It is believed the data can be modelled using $t = ah^b$

.



The data are coded using the changes of variables $x = \ln h$ and $y = \ln t$. The regression line of y on x is found to be $y = 0.3 - 1.2x$.

Find the values of a and b.

Write $t = ah^b$ in linearised form

$\ln(t) = \ln(ah^b) \Rightarrow \ln t = \boxed{\ln a} + b \ln h$

Compare coefficients

$y = 0.3 - 1.2x \Rightarrow \ln t = \boxed{0.3} - 1.2 \ln h$

$\ln a = 0.3 \Rightarrow a = e^{0.3} = 1.3498...$ $\boxed{a = 1.35 \ (3sf)}$

$\boxed{b = -1.2}$