MATH 422
Lia Vas

Empirical Models

Data Collection.

It is possible to built a mathematical model solely out of the abstract concepts. However, if the models are to be made to confront reality it is through the data that the confrontation happens. By **data** we mean measurements or observations collected in the real world. Interaction between data and models occurs in a couple of ways:

1. Data are needed to **suggest** a right model. The models called **empirical** are based entirely on data.

2. Data are needed to estimate the values of the parameters appearing in a model. This is sometimes called **calibrating** a model.

3. Data are needed to **test** a model.

It happens very often that the data given at the beginning is not sufficient for making a good model. In these cases further data collection is needed. Considering the following questions might be useful:

- What is the relevant data? Exactly what kind of data is needed?

- How can the relevant data be obtained?

- In what form do you need the data?

Once the data is collected, you need to decide on the techniques you want to use in order to find an appropriate model. There are two major groups of techniques based on two different ideas

1. Interpolation – finding a function that contains all the data points.

2. Model fitting – finding a function that is as close as possible to containing all the data points. Such function is also called a **regression curve.**

Sometimes you would need to combine these methods since the interpolation curve might be too complex and the best fit model might not be sufficiently accurate.

Model Fitting. Modeling using Regressions.

Most of the technology used (e.g. Excel, graphing calculators, Matlab) can be used to find regression curves and a variable monitoring the validity of the model, the **coefficient of determination** usually denoted by R^2. This coefficient takes values in interval [0,1] and indicates how close the data points are to be exactly on the regression curve. If R^2 is close to 1, the model is reliable. If R^2 is close to 0, other model should be considered.

- **Linear** $y = ax+b$. Easiest, simplest, used very frequently. A simple test can be performed in order to determine if data is linear: if independent variable values are equally spaced, simply check if difference of consecutive y-values is the same.

  If b is sufficiently small, y is said to be **proportional** to x ($y \propto x$)

- **Quadratic** $y = ax^2 + bx + c$. Appropriate for fitting data with one minimum or one maximum. To find out if equally spaced data is quadratic, check if the differences of the successive differences of consecutive y-values are constant.

  If $a > 0$, this function is concave up, if $a < 0$, it is concave down.

- **Cubic** $y = ax^3 + bx^2 + cx + d$. Appropriate for fitting data with one minimum and one maximum.

- **Quartic** $y = ax^4 + bx^3 + cx^2 + dx + e$. Convenient for fitting data with two minima and one maximum or two maxima and one minimum. When working with polynomial models a thing to keep in mind: balance between complexity and precision. For examples, see section Testing the effectiveness of a model. Also, monitor the long term behavior and check how realistic it is. For example, see sections Testing the validity and Choosing the right model.

- **Exponential** $y = ab^x$ or $y = ae^{kx}$. While a linear function has constant average rate of change (a constant difference of two consecutive y-values), an exponential function has *constant percent (relative) rate of change* (a constant quotient of two consecutive y-values). Thus, an easy test to check if data is linear: if independent variable values are equally spaced, simply check if quotient of consecutive y-values is the same.

  If $k > 0$, then the function is increasing and concave up. If $k < 0$, then the function is decreasing and concave up. This model is appropriate if the increase is slow at first but then it speeds up (or, if $k < 0$ if the decrease is fast at first but then slows down).

- **Logarithmic** $y = a + b \ln x$. If $b > 0$, then the function is increasing and concave down. If $b < 0$, then the function is decreasing and concave up. If the data indicates an increase, this model is appropriate if the increase is fast at first but then it slows down.

- **Logistic** $y = \frac{c}{1+ae^{-bx}}$. Increasing for $b > 0$. In this case, the increase is slow at first, then it speeds up and then it slows down again and approaches the y-value c for $x \to \infty$.

- **Power** ax^b. If $a > 0$,I it is increasing for $b > 0$ and decreasing for $b < 0$. It is called a power model since an increase of x by factor of t causes an increase of y by the power t^b of t (for $b > 0$). Increasing power function will not increase as rapidly as an increasing exponential function.

  Connection with linear model: If $y = ax^b$, then $\ln y = \ln a + b \ln x$. So, if y is a power function, $\ln y$ is a linear function of $\ln x$. Note: If $y = ab^x$, then $\ln y = \ln a + x \ln b$. So, if y is an exponential function, $\ln y$ is a linear function of x. In conclusion:

Linear	y depends linearly on x
Power	$\ln y$ depends linearly on $\ln x$
Exponential	$\ln y$ depends linearly on x
Logarithmic	y depends linearly on $\ln x$

- **Sine** $a\sin(bx+c)+d$. Appropriate for periodic data. In the formula, a is the amplitude, $b/360$ the period, c/b the horizontal shift, and d the vertical shift.

Example 1. In order to conduct an experiment to measure the stretch of a spring as a function of mass, a spring-mass system is considered. The following data is obtained

mass (grams)	50	100	150	200	250	300	350	400	450	500	550
elongation (cm)	1	1.875	2.75	3.25	4.375	4.875	5.765	6.5	7.25	8	8.75

The data suggest that the elongation is proportional to the mass. Finding linear regression results in $y = .0154x + .324$ with coefficient of determination $R^2 = .9985$. Note that the initial value is not that small compared with the size of first y-value. Thus suggest that .324 is the elongation of spring due to its own weight and that the data is better to be modeled with linear regression than just with proportionality. Another option is to consider a power model. $y = .029x^{.901}$ is obtained. $R^2 = .9984$ suggests that this also fits data well. Note that the exponent .901 is close to 1 and so it can be argued that the data is close to being proportional.

Example 2. The relation between the radius and the volume of a sphere is measured to be as table below indicates.

radius r	1	2	3	4	5
volume V	4.19	33.51	113.10	268.08	523.60

Use regressions to find the formula for the volume as a function of the radius.

The data does not suggest a linear model. The exponential model does not have a particularly high coefficient of determination. The power model gives us $y = 4.1897x^{2.9998} \approx 4.19x^3$. Note that $4\pi/3 = 4.1887 \approx 4.19$.

A note about the simplicity. Note that linear regression is the only one that can be relatively easily done "by hand" (details to be presented later). Let us consider how this problem can be converted to finding the linear regression. Recall that if $y = ax^b$, then $\ln y = \ln a + b\ln x$. Note that a is $e^{y\text{–intercept}}$ and b is the slope. So, we can find the linear regression for the following data

$\ln r$	0	.693	1.0986	1.3863	1.6094
$\ln V$	1.4327	3.5118	4.7283	5.5913	6.2607

On TI83, you can enter r-values into L_3 and V-values into L_4. Then the commands "$\ln(L_3) \to L_1$ and $\ln(L_4) \to L_2$ will place the data from the second table into the first two list allowing you do a linear regression in a usual way. The linear regression turns out to be $y = 2.9998x + 1.432 \approx 3x + 1.432$. As $e^{1.432} = 4.187 \approx 4.19$, this gives us the same result that $V = 4.19r^3$.

A word on Empirical Modeling. Empirical models are those that are based entirely on data. The important distinction between empirical models and examples from the previous section is that the empirical models are not derived from assumptions concerning the relationship between variables and they are not based on physical principles.

The first step in deriving an empirical models is to get the scatterplot of the data. If the data does not seem to be linear, try to plot one or both variables as logarithms so that you can check if an exponential or power models are good fits. The idea is to get a graph that looks reasonably linear and then to get a linear model. Keep in mind that:

Linear model	y depends linearly on x
Power model	$\ln y$ depends linearly on $\ln x$
Exponential model	$\ln y$ depends linearly on x
Logarithmic	y depends linearly on $\ln x$

Practice Problems

1. The size of population of US in 1800s has been measured and given in the table below. $t = 0$ denotes year 1800.

time t	0	10	20	30	40	50	60
population P (millions)	5.31	7.24	9.64	12.87	17.07	23.19	31.44

   Considering a logarithm of P as a function of t, find a linear model for $\ln P$ and t and deduce that the population was increasing exponentially. Write down the exponential model.

2. In the table below, the height and weight of a sample of people of various ages is recorded.

Height H (m)	0.75	0.95	1.12	1.35	1.55	1.63	1.71	1.85
Weight W (kg)	10	15	20	35	48	51	59	75

   Considering logarithms of both variables, find a linear model for $\ln H$ and $\ln W$ and deduce which power model that can be used for describing the relation of H and W. Note that the exponential model is not a good fit here since for exponential model zero H would give you nonzero W. A logarithmic model is also not a good fit since the data is concave up and increasing.

Solutions 1. $\ln P = 1.6747 + .0294t$. Thus, $P = e^{1.6747}e^{.0294t} = 5.337e^{.0294t}$ or $P = 5.337(1.0298)^t$. 2. $\ln W = 2.862 + 2.26 \ln H$, $R^2 = .992$. Thus, $W = e^{2.862}e^{\ln H^{2.26}} = 17.49H^{2.26}$.

Testing the validity of a model

Example 3. A projectile is fired upwards from the ground. The height of the projectile above the ground is shown in the following table:

Time (seconds)	0	0.5	1	1.5	2	2.5
Height (feet)	0	20.5	31.36	36.25	30.41	28.23

a) Find a good model to fit this data. b) Find the time at which the projectile hit the ground.

Discussion. The linear, exponential and logarithmic models are not good fits as they are always increasing or decreasing and the data is first increasing and then decreasing. So, you can start by finding the quadratic, cubic and quartic models. Look at the graphs. Note that the cubic and quartic polynomials start increasing instead of decreasing after the maximum height is reached. Because of this, these models are not very appropriate. The quadratic model seems to fit the data and the reality of the situation the best. For this model $R^2 = 0.9738$ which is pretty close to 1 so this supports the decision to use this model.

Solution. a) Quadratic model is $y = -12.87x^2 + 42.22x + 1.177$ and $R^2 = 0.9738$. b) The object falls to the ground 3.31 second after it is thrown up.

Testing the effectiveness of a model

Example 4. Healthcare costs have been increasing over the years. The following data shows the average cost of healthcare per person from 1976 to 1998:

Year	1976	1980	1987	1993	1998
Cost (per person)	618	860	1324	1865	2256

a) Find a model that fits the data well. b) Find the time the average healthcare cost will reach \$2800 per person.

Discussion. Let $x = 0$ denotes the year 1970. Find the quadratic, cubic and quartic model. Look at the graphs. Compare R^2 for all models. Note that cubic model increases complexity of the equation without changing the value of R^2 significantly (also it starts to decrease at some point and the data does not indicate that). The quartic model, although accurate because $R^2 = 1$ also starts decreasing which does not fit the data. The exponential model has $R^2 = .988$ so a good case could be made for choosing it. The logarithmic model is not appropriate as the scatterplot indicates concave up data, not concave down. Thus, either quadratic or exponential models are appropriate to use.

Solution. 3. a) Quadratic Model: $y = .7686x^2 + 49.204x + 290.8067$, $R^2 = .9988$ b) Using quadratic model, $y = 2800$ when $x = 33.48$. So, by year 2004, the healthcare cost will be over \$2800 per person.

Choosing the right model

Practice Problems

1. The population present in a bacteria culture over 5 days is given in the table below:

time (days)	0	1	2	3	4	5
population	30	133	214	337	527	819

   a) Find a good model for the data. b) Estimate the population after 7 days.

2. A company decided to develop a cost equation based on the quantity of the product produced in a day. They collected the following data:

quantity produced	20	35	50	65	80	95	110
cost	642.35	766.48	858.82	928.83	1005.32	1078.82	1140.79

a) Find a good model for the data. b) According to the model, find how many units could be produced for \$800.

3. The table below shows the yield (in mg) of a chemical reaction in the first 6 minutes.

time (minutes)	1	2	3	4	5	6
yield (mg)	1.2	6.9	9.3	12.7	14.1	15.7

a) Use the scatterplot to find the best model to fit this data. b) Using that model, determine in how many minutes will the yield be 20 mg.

Solutions.

1. a) The cubic model is a slightly better fit than quadratic. For cubic $R^2 = .9999$. The quartic model has almost the same R^2 as cubic so it increases complexity without adding much accuracy. Logarithmic is concave down instead of up and exponential does not have very high R^2. b) 1886 bacteria.

2. a) The leading coefficient of cubic model is $3.25 \cdot 10^{-4}$. Since this number is relatively small, this means that the cubic curve is almost a quadratic curve. R^2 of quadratic is .997 and the cubic has $R^2 = .9993$ which suggests that they are almost equally efficient. Considering effectiveness, we are leaning towards quadratic. However, the quadratic curve starts decreasing after about $x = 200$. Thus, if it is to predict the cost if more than 200 items are produced, the cubic model is better. If the production is not expected to exceed 200 items, the quadratic is better since it is more efficient.

   The leading coefficient of quartic model is $-9.7 \cdot 10^{-6}$. As this number is very small, this means that this curve is almost a cubic. Thus, when deciding between these two models, cubic seems to be a better choice. Also, the quartic model starts decreasing and the data does not indicate that should happen.

   b) With quadratic: about 42 units. With cubic: about 41 units.

3. The quartic and logarithmic model have almost the same R^2. The other models are not very accurate. As logarithmic is simpler than quartic (especially considering that taking $\ln x$-values instead of x-values would give you a linear model), we can choose logarithmic. b) 10.34 min.

Modeling with piecewise defined functions

Recall that a piecewise defined function is a function defined by different formulas for different values of the independent variable. Sometimes, it is better to model given data using one piecewise defined function than with a single function.

Practice problems

1. The size of a population of rabbits in a certain habitat is described by a table below.

year	2000	2001	2002	2003	2004	2005	2006
number of rabbits	30	45	68	60	96	154	247

We can see that the number is increasing from 2000 to 2003. There was a decrease in number of rabbits in 2004 due to a flood in the habitat but after 2004, the number of rabbits is increasing again. Assuming that the number of rabbits is increasing exponentially both before the flood and after the flood, find the two exponential regressions that will best fit the data before and after the flood. Using the two formulas, write down a piecewise function that will describe the number of rabbits from 2000 to 2006. Estimate the number of rabbits in 2010.

2. The concentration of certain medication is decreasing as time is passing. The measurements of the concentration are given in a table below.

time (hours)	0	1	2	3	4	5	6	7	8
concentration (mg/cm^3)	4	3.2	2.5	1.4	0.9	0.4	0.2	0.15	0.09

   a) Assume that in the first four hours a polynomial function is the best fit. Find quadratic and cubic regression (record both curves) and choose the one that fits the best.

   b) After the fourth hour, assume that the concentration is decreasing exponentially. Find the exponential regression that fits the data recorded after the fourth hour.

   c) Write down a piecewise function that will describe the concentration within first 8 hours. By looking at the graph of the curve that describes the concentration in the first four hours, explain why that part is not a good fit for the concentration after four hours and why the piecewise function should be used.

Solutions.

1. The exponential regression for the first three given points gives us the formula $y = 29.96(1.5055)^x$. The exponential regression for the next four points gives us $14.56(1.6028)^x$. Thus, the piecewise function is $y = \begin{cases} 29.96(1.5055)^x & 0 \le x \le 2 \\ 14.56(1.6028)^x & x \ge 3 \end{cases}$ $y(10) = 1629.47 \approx$ 1629 rabbits.

2. a) Cubic is a better fit. $y = .0417x^3 - .236x^2 - .4988x + 3.979$. b) $y = 7.13(.572)^x$ If you start from $x = 5$, $y = 4.01(.621)^x$. c) $y = \begin{cases} .0417x^3 - .236x^2 - .4988x + 3.979 & 0 \le x \le 4 \\ 7.13(.572)^x & x > 4 \end{cases}$ Piecewise function is better as cubic curve starts to increase after 4th hour.

Analytic Methods of Model fitting

In this section, we look into the mathematics behind obtaining regression models and curve fitting.

Least-Squares Criterion

This criterion is the most frequently used curve-fitting criterion. It is based on an idea that in order to fit a set of data points (x_i, y_i), $i = 1, \ldots, m$ onto a curve $y = f(x)$, we again want the differences between y_i and $f(x_i)$ to be as small as possible. Recall that the square

of distance of the vector between the observed value y_i and predicted value $f(x_i)$ is given by $\sum_{i=1}^{m}(y_i - f(x_i))^2$. Thus, a way to make the differences $y_i - f(x_i)$ small is to

$$\text{minimize the sum } \sum_{i=1}^{m}(y_i - f(x_i))^2$$

Least-Squares Line Fit.

Suppose that we would like to fit the points (x_i, y_i), $i = 1, \ldots, m$ onto a line $y = ax + b$. So, we need to minimize the sum

$$S = \sum_{i=1}^{m}(y_i - ax_i - b)^2$$

This will be satisfied when $\partial S/\partial a = 0$ and $\partial S/\partial b = 0$.

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{m} 2(y_i - ax_i - b)(-x_i) = 0$$

$$\frac{\partial S}{\partial b} = \sum_{i=1}^{m} 2(y_i - ax_i - b)(-1) = 0$$

We can view the equations above as two linear equations in the unknowns a and b. Solving them for a and b, we obtain

$$a = \frac{m\sum x_i y_i - \sum x_i \sum y_i}{m\sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i}{m\sum x_i^2 - (\sum x_i)^2}$$

The two equations above are called the **normalizing equations.** It is not hard to write a computer code that will compute these values.

Least-Squares Power Fit.

Suppose that a positive integer n is fixed and that we would like to fit the points (x_i, y_i), $i = 1, \ldots, m$ onto a curve of the form $y = ax^n$. So, we need to minimize the sum

$$S = \sum_{i=1}^{m}(y_i - ax_i^n)^2$$

This will be satisfied when

$$0 = \frac{\partial S}{\partial a} = \sum_{i=1}^{m} 2(y_i - ax_i^n)(-x_i^n) = 0$$

Solving this equation for a, we obtain

$$a = \frac{\sum x_i^n y_i}{\sum x_i^{2n}}$$

Example 5. Let us find the quadratic power fit for the data in the table below.

x	0.5	1	1.5	2	2.5
y	0.7	3.4	7.2	12.4	20.1

Compute $\sum x_i^2 y_i$ to be 195.0 and $\sum x_i^4$ to be 61.1875. Thus $a = \frac{195.0}{61.1875} = 3.1869$ and so $y = 3.1869x^2$.

Least-Squares Fits in Matlab.

If we need to fit the data (x_i, y_i), $i = 1, 2, \ldots, m$ to a curve $y = f(x)$ using the least-square, we need to minimize the function $S = \sum_{i=1}^{m}(y_i - f(x_i))^2$. For an M-file calculating the parameters of the curve $f(x)$, the input should be a vector $\mathbf{X}$ whose entires are values x_i, for $i = 1, 2, \ldots, m$ and a vector $\mathbf{Y}$ whose entries are the values y_i for $i = 1, 2, \ldots, m$. The size m of these vectors also needs to be recorded.

The command **size(X, 2)** calculates the length of a vector $\mathbf{X}$. In general if A is an $m \times n$ matrix, the command **size(A, 1)** returns the number of rows m and the command **size(A, 2)** returns the number of columns n.

For example, if **X=[2 5 6 0]**, Matlab registers it as an 1x4 matrix. Thus, the outcome of the command **size(X, 2)** would be 4.

Example 6. Let us consider the Matlab code for calculating the best fit curve of the form $y = ax$. In this case $S = \sum_{i=1}^{m}(y_i - ax_i)^2$ and and $\frac{\partial S}{\partial a} = \sum_{i=1}^{m} 2(y_i - ax_i)(-x_i)$. Solving for a gives us that

$$a = \frac{\sum_{i=1}^{m} x_i y_i}{\sum_{i=1}^{m} x_i^2}$$

The following Matlab code calculates a for a data set (x_i, y_i), $i = 1, 2, \ldots, m$. Before executing this code, the x-values should be entered as a vector **X=[x1, x2, ... ,xm]** and the y-values should be entered as a vector **Y=[y1, y2, ... ,ym]**. Note that $\mathbf{X}$ and $\mathbf{Y}$ need to have the same size. The command **X(i)** returns the i-th coordinate x_i.

In the code below, **sxsq** denotes a variable for sum of squares of x-values and **sxy** denotes a variable for sum of products of x and y-values. The variable m denotes the number of x and y values.

```
function a=proportion(X, Y)
sxsq=0;
sxy=0;
m=size(X, 2);
for i=1:m
sxsq=sxsq+X(i)^2;
sxy=sxy+X(i)*Y(i);
end
a=sxy/sxsq;
```

Example 7. Write a Matlab code for calculating a linear least-squares fit $y = ax + b$, for a data set (x_i, y_i), $i = 1, 2, \ldots, m$. Using the formulas from the previous section, we have that

$$a = \frac{m\sum x_i y_i - \sum x_i \sum y_i}{m\sum x_i^2 - (\sum x_i)^2} \qquad b = \frac{\sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i}{m\sum x_i^2 - (\sum x_i)^2}$$

Let **sx** denote a variable for sum of x-values,
sy denote a variable for sum of y-values,
sxy denote a variable for sum of products of x and y-values,
sxsq denote a variable for sum of squares of x-values,
sysq denote a variable for sum of squares of y-values, and
m is the number of x and y values.

```
function [a, b]=linear_fit(X, Y)
sx=0; sy=0;
sxy=0; sxsq=0;
sysq=0; m=size(X, 2);
for i=1:m
sx=sx+X(i);
sy=sy+Y(i);
sxsq=sxsq+X(i)^2;
sysq=sysq+Y(i)^2;
sxy=sxy+X(i)*Y(i);
end
a=(m*sxy-sx*sy)/(m*sxsq-sx^2);
b=(sxsq*sy-sxy*sx)/(m*sxsq-sx^2);
```

Transformed Least-Squares Fit.

The formula computing the coefficients of the linear and power least-squares fits is not too complex because the partial derivatives considered were linear functions of the unknown coefficients. With exponential, logarithmic or power model (with n not fixed in $y = ax^n$), the equations determined by the partial derivatives are not linear equations in unknown coefficients. In cases like this, one may modify the data (considering ln of one or both of the variables instead of the original data) in order to reduce a non-linear model to a linear one. Thus

i) **Finding an exponential model $y = be^{ax}$ using the transformed data.** Taking ln of $y = be^{ax}$, obtain that $\ln y = ax + \ln b = ax + B$ for $\ln b = B$. Thus, you can find a linear model for $(x_i, \ln y_i)$ instead of (x_i, y_i), $i = 1, \ldots, m$. Note that in this case

$$a = \frac{m \sum x_i \ln y_i - \sum x_i \sum \ln y_i}{m \sum x_i^2 - (\sum x_i)^2}$$

$$\ln b = B = \frac{\sum x_i^2 \sum \ln y_i - \sum x_i \ln y_i \sum x_i}{m \sum x_i^2 - (\sum x_i)^2}$$

ii) **Finding a logarithmic model $y = a \ln x + b$ using the transformed data.** Simply consider $(X_i, y_i) = (\ln x_i, y_i)$ instead of (x_i, y_i), $i = 1, \ldots, m$ and find the linear model for the transformed set of data. The linear model for $y = aX + b$ is the logarithmic model that we are looking for. Note that in this case

$$a = \frac{m \sum \ln x_i y_i - \sum \ln x_i \sum y_i}{m \sum (\ln x_i)^2 - (\sum \ln x_i)^2}$$

$$b = \frac{\sum(\ln x_i)^2 \sum y_i - \sum \ln x_i y_i \sum \ln x_i}{m\sum(\ln x_i)^2 - (\sum \ln x_i)^2}$$

iii) **Finding a power model $y = bx^a$ using the transformed data**. Taking ln of $y = bx^a$, gives you $\ln y = a \ln x + \ln b$. Consider $(\ln x_i, \ln y_i)$ instead of (x_i, y_i), $i = 1, \ldots, m$ and find the linear model for this transformed set of data. Denote $\ln b = B$. The linear model for $Y = \ln y = a \ln x + \ln b = aX + B$ can be found as:

$$a = \frac{m \sum \ln x_i \ln y_i - \sum \ln x_i \sum \ln y_i}{m\sum(\ln x_i)^2 - (\sum \ln x_i)^2}$$

$$\ln b = B = \frac{\sum(\ln x_i)^2 \sum \ln y_i - \sum \ln x_i \ln y_i \sum \ln x_i}{m\sum(\ln x_i)^2 - (\sum \ln x_i)^2}$$

Example 8. Let us find a power fit for the data from Example 5. Computing a and b using the formulas from iii) above, we obtain that $a = 2.062$ and $b = 1.1266$ so that $B = e^b = 3.085$. So, $y = 3.085x^{2.0628}$.

An important fact to note is that if the partial derivative equations are used to find the unknown coefficients a and B in $y = Be^{ax}$, $y = a \ln x + b$ or $y = Bx^a$ *without* transforming the original data, we would end up with *different* models than the models that we get when we transform thee data and find a linear regression for the new data. It is important to keep in mind how different technology handles the data. For example, TI83 is programmed to do the transformed least-squares linear fit when finding the power model.

Transforming the original data often simplifies the formulas even for models with less parameters than exponential, power or logarithmic. The next example illustrates this.

Example 9. Find the formula computing the coefficient a for the quadratic fit $y = ax^2$ and use it to find a transformed quadratic fit for the data from Example 5.

Take ln of $y = ax^2$. Obtain than $\ln y = \ln a + 2 \ln x$. Let us denote $\ln a$ by A. Thus, we need to consider $\ln y = A + 2 \ln x$ and to to minimize $S = \sum(\ln y_i - A - 2 \ln x_i)^2$. The partial derivative is $\frac{\partial S}{\partial A} = \sum_{i=1}^m 2(\ln y_i - A - 2 \ln x_i)(-1)$. Setting this derivative to 0 and solving for A gives us

$$\ln a = A = \frac{1}{m} \sum_{i=1}^{m} (\ln y_i - 2 \ln x_i)$$

which can be easily solved for a.

Using the data from Example 5, we obtain that $A = 1.1432$ and so $a = 3.1368$. Compare this model $y = 3.1368x^2$ with the model $y = 3.1869x^2$ obtained using the direct power curve fit.

Example 10. Write Matlab code for determining the power curve of the form $y = ax^2$ using the transformed least-squares fit for a data set (x_i, y_i), $i = 1, 2, \ldots, m$.

Use the formula from previous example $A = \frac{1}{m} \sum_{i=1}^m (\ln y_i - 2 \ln x_i)$ and $a = e^A$.

The following Matlab code calculates the coefficient a. In the program the variable **S** denotes the sum of $\ln y_i - 2 \ln x_i$ values

```
function a=transformed_squares(X, Y)
m=size(X, 2);
S=0;
for i=1:m
S=S+log(Y(i))-2*log(X(i));
end
A=S/m;
a=exp(A);
```

Chebyshev Approximation Criterion.

Let us assume that m points (x_i, y_i), $i = 1, \ldots, m$ are given and that we need to fit them on a curve $y = f(x)$. As we would like the differences between y_i and $f(x_i)$ to be as small as possible, the idea is to minimize the largest absolute value $|y_i - f(x_i)|$. Thus, the only difference between the least square fit and Chebyshev method is that instead of sum of squares of $y_i - f(x_i)$ we are minimizing the maximum of the absolute values of $y_i - f(x_i)$.

Following steps achieve this goal.

1. Let $r_i = y_i - f(x_i)$ for $i = 1, \ldots, m$. The variables r_i are called **residuals**.
2. Find the maximum of $|r_i|$ for $i = 1, \ldots, m$. Let us call this quantity r.
3. Solve the following optimization problem: minimize r subject to constraints $-r \leq r_i \leq r$ for $i = 1, \ldots, m$ i.e.

$$r - r_i \geq 0, \quad r + r_i \geq 0, \quad \text{for } i = 1, \ldots, m.$$

Example 11. Suppose that a point B is somewhere inside a line segment AC. Assume that measuring the distances AB, BC and AC, we obtain that $AB = 13$, $BC = 7$ and $AC = 19$. As $13 + 7 = 20 \neq 19$, we need to resolve the discrepancy. We can do that using the Chebyshev approximation criterion. Let x_1, x_2 and x_3 stand for exact lengths of AB, BC and AC respectively. Then $r_1 = x_1 - 13$, $r_2 = x_2 - 7$, and $r_3 = x_1 + x_2 - 19$. Let r stands for the maximum of the absolute values of r_1, r_2 and r_3 (note that we don't know which one is r). We would like to minimize r subject to

$$r - x_1 + 13 \geq 0, \;\; r + x_1 - 13 \geq 0, \;\; r - x_2 + 7 \geq 0, \;\; r + x_2 - 7 \geq 0$$

$$r - x_1 - x_2 + 19 \geq 0 \quad r + x_1 + x_2 - 19 \geq 0$$

Using the methods of linear programming (see section on Optimization), we obtain that $r = \frac{1}{3}$, $x_1 = 12\frac{2}{3}$ and $x_2 = 6\frac{2}{3}$.

Although in this example linear programing is used, note that in some cases the constraints will not be linear equations and other optimization methods should be used.

Measuring the validity of a model

In the previously seen examples, two different quadratic power models $y = 3.1368x^2$ and $y = 3.1869x^2$ are obtained for the same set of data. If Chebyshev criterion is to be used for the

same set of data, it would yield yet another model $y = 3.17073x^2$ for the same set of data. A natural question is: how can we choose the best model?

Before we we attempt to answer, let us introduce some notation. If the data (x_i, y_i), $i = 1, \ldots, m$ is to be fit on the curve $y = f_1(x)$ obtained using the Chebyshev criterion, let c_i denote the absolute deviations $|y_i - f_1(x_i)|$. If the same data is to be fit on the curve $y = f_2(x)$ obtained using the least-squares criterion, let d_i denote the absolute deviations $|y_i - f_2(x_i)|$. Also, let c_{max} denote the largest of c_i and d_{max} denote the largest of d_i.

Since the least-squares criterion is such that the squares of the deviations are minimal, we have that

$$\sum d_i^2 \leq \sum c_i^2 \leq m c_{max}^2$$

Thus,

$$D = \sqrt{\frac{\sum d_i^2}{m}} \leq c_{max}$$

Also, since the Chebyshev criterion is such that the maximum of the absolute deviations is minimal, we have that $c_{max} \leq d_{max}$. Thus,

$$D \leq c_{max} \leq d_{max}$$

This last equation can help us determine which model to use: the least-squares criterion is convenient to use but there is always to concern that the difference between D and d_{max} might be too large. Balancing these two conditions might help us decide which model to use.

Let us go back to the example with three different quadratic power models. Computing the deviations for all three models we conclude

1. For the model $y = 3.1869x^2$ obtained using the least-squares fit. The largest absolute deviation is 0.347. The smallest is 0.0976.

2. For the model $y = 3.1368x^2$ obtained using transformed least-squares fit. The largest absolute deviation is 0.495. The smallest is 0.0842. This model was the easiest to compute but has the largest absolute deviation.

3. For the model $y = 3.17073x^2$ obtained using Chebyshev criterion. The largest absolute deviation is 0.282. The smallest is 0.0659. Computationally, this model was the hardest to get but has the smallest absolute deviation. However, considering the sum of squares of deviations instead of the maximum of the absolute deviations, the least-squares fit model is better than the Chebyshev (.2095 for the least squares versus .2256 for Chebyshev).

The conclusion that we can draw from this example is that there is not a single right answer when trying to decide which model is the best. Thus, in each of the following cases, our choice of models would be different.

1. If it is more important to minimize the sum of squares of deviations than the maximal absolute deviation, the least-squares criterion should be used.

2. If it is more important that the model is computationally simple than to have small maximum or sum of squares of absolute deviations, the transformed least-squares criterion should be used.

3. If it is more important to minimize the maximal absolute deviation than the sum of squares of deviation, the Chebyshev criterion should be used.

So, we need to decide which model is the best on case-by-case basis, taking all the specifics into account (what is the purpose of the model, how precise should it be, how accurate is the data, etc).

Linear Regression

We can gain further insight in the basic linear model when we consider some further parameters of the statistical analysis. For the data (x_i, y_i), $i = 1, \ldots, m$, consider the following values.

1. The **error (or explained) sum of squares** ESS

$$ESS = \sum_{i=1}^{m} [y_i - (ax_i + b)]^2$$

   ESS reflects the variation about the regression line.

2. The total (or corrected) sum of squares TSS

$$TSS = \sum_{i=1}^{m} (y_i - \overline{y})^2$$

   where $\overline{y}$ is the average of the y-values of the data points. TSS reflects the variation in the y values about the line $y = \overline{y}$.

3. The **regression sum of squares** is $RSS = TSS - ESS$.

4. The **coefficient of determination** R^2

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

Thus defined R^2 is independent on the units of x and y and is also independent on which of the two variables is labeled as independent and which as dependent (in other words, data (y_i, x_i), $i = 1, \ldots, m$, would produce the same value of R^2).

If $R^2 = .9$ for example, then 90% of the total variation of y-values from the line $y = \overline{y}$ is accounted for by a linear relationship with the values of x.

Example 12. Consider developing a model that will relate the human weight with the height. Use the model to predict the body weight.

The simplest model could be derived from a simple assumption that weight is proportional to volume, and that volume is proportional to the cube of height.

$$W = aH^3.$$

In medicine, the BWI (body weight index) is calculated considering the quotient $\frac{W}{H^2}$ (in fact, $27 = \frac{W}{H^2}$ is considered a norm). This assumes that weight is proportional to the square of height.

$$W = aH^2.$$

In the table below, the height and weight of a sample of people of various ages is recorded.

Height H (m)	1.75	1.95	1.50	1.75	1.55	1.63	1.71	1.85
Weight W (kg)	65	85	45	70	48	51	59	75

If we were to find the quadratic or cubic model, this would yield more terms than necessary and we might not be able to test how well our hypothesis $W \propto H^2$ and $W \propto H^3$ match the data. Thus, we might want to find the linear model for the squares and the cubes of the x-values. This is called the **linearization** of data.

When we find a linear regression for the squares of the heights given, we obtain $W = 26.92H^2 - 17.12$ and $R^2 = .968$. When we find a linear regression for the cubes of the heights given, we obtain $W = 10.37H^3 + 9.23$ and $R^2 = .969$. Since the values of R^2 are approximately same, the cubic model is slightly better since the value of y-intercept is smaller so the model is closer to the form $W = aH^3$ than the quadratic model.

We could also calculate the power model. Considering logarithms of both variables, we can find a linear model for $\ln H$ and $\ln W$. We obtain that $\ln W = 2.749 + 2.547 \ln H$, with $R^2 = .968$. Thus, $W = e^{2.749} e^{\ln H^{2.547}} = 15.63H^{2.547}$. The benefit of this model is that it passes (0,0) (that corresponds to reality), but the exponent 2.547 is not as easy to work with as exponents 2 or 3.

Practice Problems

1. Fit the following data on a straight line using the equations for slope and y-intercept given in the section Least-Squares Line Fit.

x	1.0	2.3	3.7	4.2	6.1	7.0
y	3.6	3.0	3.2	5.1	5.3	6.8

   Calculate d_{max} and explain what your answer means.

2. In the table below, the height and weight of a sample of people of various ages is recorded.

Height H (m)	1.75	1.95	1.50	1.75	1.55	1.63	1.71	1.85
Weight W (kg)	65	85	45	70	48	51	59	75

   In an example solved previously, we have found the power model to be $W = 15.63H^{2.547}$. We have also seen that the models of the form $W = aH^2 + b$ and $W = aH^3 + b$ are $W = 26.92H^2 - 17.12$ and $W = 10.37H^3 + 9.23$. However, it is of interest also to consider the models $W = aH^3$ (starting from a simple assumption that weight is proportional to volume) and $W = aH^2$ (the model used in medicine for calculating the body weight index BWI). Find the model of the form:

   a) $W = aH^2$ b) $W = aH^3$

3. The following data represents the growth of population of fruit flies over a 6-week period.

time (in days) t	7	14	21	28	35	42
no. of fruit flies P	8	41	133	250	280	297

   Use the least-squares criterion either on data or on the transformed data to find the models of the following types

a) $P = kt$ b) $P = kt^2$ c) $P = k \ln t$

By considering the data plot and the graphs of the three models, conclude which models are appropriate.

4. In the following data, W represents the weight of a bass, l represents its length and g its girth.

Length (in in.) l	14.5	12.5	17.25	14.5	12.625	17.75	14.125	12.625
Girth (in in.) g	9.75	8.375	11.0	9.75	8.5	12.5	9.0	8.5
Weight (in oz.) W	27	17	41	26	17	49	23	16

Use the least-squares criterion to find the following models. a) $W = kl^3$ b) $W = klg^2$. Determine which model fits the data better.

5. Write a Matlab program that finds the least squares (or transformed least squares when appropriate) estimates of the coefficients of the following models: a) $y = ax^2$ b) $y = ax^n$ c) $y = ae^x$ d) $y = a \ln(x)$.

Solutions

1. $y = .5642x + 2.2149$; $d_{max} = 1.1$

2. a) $a = \frac{1517.9313}{71.3743} = 21.267 \Rightarrow W = 21.267H^2$. Note that the fact that a is smaller than the medical norm $a = 27$ suggests that the data was collected from people with BWI somewhat under the norm. b) $a = \frac{2675.1534}{221.531} = 12.07575 \Rightarrow W = 12.076H^3$.

3. a) $a = \frac{32697}{4459} \Rightarrow P = 7.33t$. b) $a = 11299895462275 = .2069 \Rightarrow P = 0.2069t^2$ c) $a = \frac{3467.33}{57.73} = 60.056 \Rightarrow P = 60.056 \ln t$. Neither of the models does not seem to be very appropriate.

4. a) $W = .00844l^3$ and $d_{max} = 2.305$. b) $W = .01868lg^2$ and $d_{max} = 2.794$. This suggests that the model $W = kl^3$ is better.

5. Write down the formulas that you would use first and then just translate them to appropriate M-files.