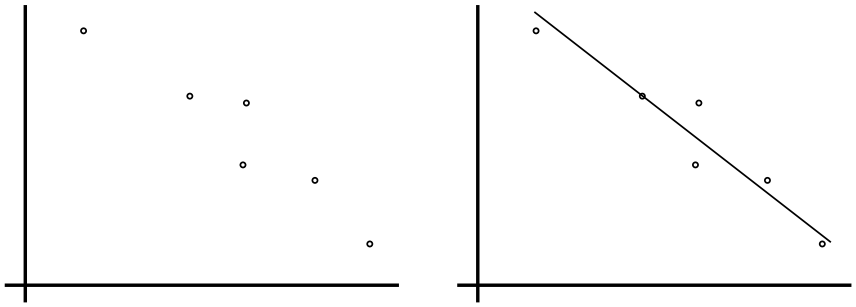
Dealing with Data and Fitting Empirically

Notes by Holly Hirst

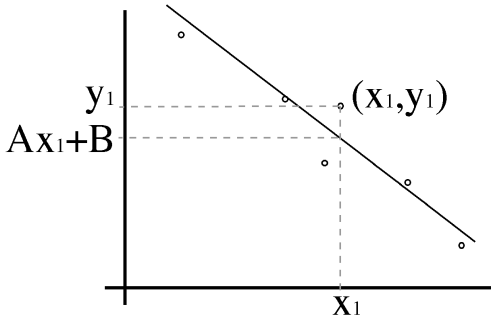
Fitting Functions to Data: Regression

When working with a situation for which we don't have a physical law (and hence an equation, function or inequality), we observe the system, recording data from our observations. We are then left with the task of making predictions and drawing conclusions from the data. To do this we start by determining the trend of the data, and then fitting an appropriate function. If the data from the observations are ordered pairs, we can look at a graph of the data to determine the trend, and hence the form of the function to fit. Common practice puts the variable we are selecting to measure (input) on the x axis – statisticians call it the **predictor** – and the variable we are hoping to predict (output) on the y axis – the **response**.

In the graph below on the left, the trend appears to be roughly linear with a downward slope. We can fit the line "by eye" as in the graph on the right, simply drawing the line so that it looks like no one data point is too far from the line. If an equation is needed, we could choose two points on the line and use simple algebra to find the equation.



While fitting by eye is appealing because it is so simple, a slightly more mathematical approach yields a line for which we can quantify the fact that no point is too far from the line. Let the formula for the line we are looking for be $y = Ax+B$, where A is the unknown slope and B is the unknown y-intercept. Look at the graph in the picture below.



The actual data point value for x_1 is y_1 and the response value for the predictor x_1 from the line is Ax_1+B. We'll call the difference between these values the **deviation**. One way to fit a line to the data is to minimize the sum of these deviations,

$$\text{minimize } \sum_{i=0}^{n} \left| Ax_i + B - y_i \right|,$$

where we have assumed that there are n+1 data points numbered 0 through n. Also note that we have used the absolute values of the deviations. We do this so that deviations for points above the line don't cancel out deviations for points below the line.

So what next? To minimize this expression, we need to take the partial derivatives with respect to A and B (the unknowns), set the derivatives equal to zero, and then solve the resulting equations simultaneously. Unfortunately, the absolute values are difficult to deal with when taking derivatives, since absolute values have discontinuous derivatives. Note, however, that a linear programming approach can be taken to get a "Chebychev line of best fit." We will not pursue that here.

How do we fix this? Instead of minimizing the sum of the deviations, we will minimize the sum of the *squares* of the deviations:

$$\text{minimize } \sum_{i=0}^{n}\left(Ax_i + B - y_i\right)^2$$

The squares of the deviations are always non-negative – which is why we used the absolute values – but avoid the problems with derivatives. In addition, minimizer for this sum is equal to the minimizer for the square root of this sum – and the square root of the sum is closely related to Euclidean distance.

To finish this, we need to take the two partials and set them equal to zero:

$$\frac{\partial}{\partial A}\sum_{i=0}^{n}\left(Ax_i + B - y_i\right)^2 = \sum_{i=0}^{n} 2\left(Ax_i + B - y_i\right)\left(x_i\right) = 2\sum_{i=0}^{n}\left(Ax_i^{\,2} + Bx_i - x_i y_i\right) = 0 \qquad (1)$$

$$\frac{\partial}{\partial B}\sum_{i=0}^{n}\left(Ax_i + B - y_i\right)^2 = \sum_{i=0}^{n} 2\left(Ax_i + B - y_i\right)(1) = 2\sum_{i=0}^{n}\left(Ax_i + B - y_i\right) = 0 \qquad (2)$$

Yuk! Now what? We need to simplify these two equations and then solve them for A and B. To do this we need to use the fact that we can rearrange sums of sums as follows:

$$\sum\left(T_i + R_i + S_i\right) = \sum T_i + \sum R_i + \sum S_i$$

The left side of this equation indicates to add the ith T, R and S values first and then add those sums; the right side says to add the T's, add the R's, add the S's and then sum those sums. Either way, we have added all the T's, R's and S's together. Using this fact, we can simplify (1) to get:

$$\sum_{i=0}^{n} Ax_i^{\,2} + \sum_{i=0}^{n} Bx_i - \sum_{i=0}^{n} x_i y_i = A\sum_{i=0}^{n} x_i^{\,2} + B\sum_{i=0}^{n} x_i - \sum_{i=0}^{n} x_i y_i = 0 \qquad (3)$$

Where did the factor of 2 go? We divided both sides of the equation by 2. Similarly, (2) simplifies to:

$$\sum_{i=0}^{n} Ax_i + \sum_{i=0}^{n} B - \sum_{i=0}^{n} y_i = A\sum_{i=0}^{n} x_i + Bn - \sum_{i=0}^{n} y_i = 0 \qquad (4)$$

We are finally ready to solve these two equations for A and B. We will rewrite these without the subscripts and limits for simplicity as:

$$A\sum x^2 + B\sum x = \sum xy \quad \text{and} \quad A\sum x + Bn = \sum y$$

Multiplying the first one by n and multiplying second one by $\sum x$ and then subtracting gives:

$$A\left(n\sum x^2 - \sum x \sum x\right) = n\sum xy - \sum x \sum y \qquad (5)$$

Solving for A gives:

$$A = \frac{n\sum_{i=0}^{n} x_i y_i - \sum_{i=0}^{n} x_i \sum_{i=0}^{n} y_i}{n\sum_{i=0}^{n} x_i{}^2 - \left(\sum_{i=0}^{n} x_i\right)^2}$$

Plugging this expression for A into (4) gives a formula for B:

$$B = \frac{\sum_{i=1}^{n} y_i - A\sum_{i=1}^{n} x_i}{n}$$

Here is an example done using these formulas. We have used a table to organize the calculations of all of the sums. (FYI: These data are from a physics experiment on springs.)

	X	Y	X^2	XY
	0	5.3	0	0
	2	7.0	4	14
	4	9.4	16	37.6
	5	11.1	25	55.5
	6	12.3	36	73.8
	8	14.2	64	113.6
sums:	25	59.3	145	294.5

A = (6*294.5 – 25*59.3) / (6*145 – 25*25) = 1.61122
B = (294.5 – 1.61122*25) / 6 = 5.044898

So the line we might use to predict y using an x value is: $y = 1.61\,x + 5.04$.

Goodness of Fit

So how can we tell the quality of the fit? First we need to look at the data and a graph showing the line and the data, asking whether the conditions needed for linear regression hold. If we are satisfied with the fit, then we can look at a statistical measure of the strength of the fit called the **coefficient of determination**.

As usual, let the actual data points be (x_i, y_i), i = 1, ..., n. Also let the predicted value for x_i from the fitted line be $\hat{y}_i$. So we have

$$\hat{y}_i = Ax_i + B.$$

Conditions for Regression

The assumptions we need for regression information to make sense (be unbiased estimates of coefficients and give reliable estimates of variance) are:

1. The x values are fixed, not random.

2. The *y* data values are independent of each other and normally distributed. This is usually the case if the data were collected carefully. One notable exception is values that are serially collected for which the independent variable is not time. These time series data may depend on time and hence on each other.

3. The individual **residuals** ($\hat{y}_i - y_i$) are independent of each other and normally distributed with mean 0.

How do we recognize when these assumptions are violated?

1. The single best way to tell if the regression is good is: LOOK AT THE GRAPH! Look for the following things:

   - Does the line go through the middle of the data? no = bad (might be influenced by an outlier)
   - Is there a pattern to the deviations? yes = bad (might be non-linear)
   - Is the shape of the data rectangular (rather than wedge shaped)? no = bad (might need a transformation of the *y* values or a weighted regression).

2. One can also look at some other graphs:

   - Plot the *y* values on a histogram to look for normality
   - Plot the residuals ($\hat{y}_i - y_i$) to look for normality with mean = 0.
   - Plot residuals against fitted values (Y = ($\hat{y}_i - y_i$) versus X = $\hat{y}_i$). This should be a horizontal band across the graph. If there's a (sloped) linear pattern there might be an underlying independent variable that should replace the *x* that was used. If there is a wedge shape, a transformation of *y* or weighted regression might be needed.

Coefficient of Determination

Once we are sure that none of the conditions above are violated, we can look at a value called the **coefficient of determination**, aka R-Squared. This will give us a statistical measure of the quality of the fit.

What did we do to find the fits? We solved the problem

$$\text{minimize} \sum_{i=0}^{n} \left(Ax_i + b - y_i\right)^2$$

or, using the hat notation:

$$\text{minimize} \sum_{i=0}^{n} \left(\hat{y}_i - y_i\right)^2$$

If we want to find out how well the line captures the relationship in the data, we can examine the following quantities that measure variance:

SSTO (total sum of squares): $\sum_{i=0}^{n}(y_i-\bar{y})^2$, which measures the overall deviation from the mean ($\bar{y}$) of the response variables without regard to the predictor variable. If this were 0, then there would be no need to know any x values, i.e., no dependence on the x values.

SSE (error sum of squares): $\sum_{i=0}^{n}(y_i-\hat{y}_i)^2$, which measures the deviation of the line values from the observed values for the response, i.e., taking into consideration the predictor variable. If this were 0, then we would know that the relationship was exactly linear, i.e., all of the data points are on the regression line.

SSR (regression sum of squares): $\sum_{i=0}^{n}(\hat{y}_i-\bar{y})^2$, which measures the overall deviation from the mean ($\bar{y}$) of the fitted variables. If this were 0, then there would be no need to know any x values, i.e., no dependence on the x values and the regression line would be horizontal.

Using basic ideas from linear algebra, we can show that

$$\sum_{i=0}^{n}(y_i-\bar{y})^2=\sum_{i=0}^{n}(y_i-\hat{y}_i)^2+\sum_{i=0}^{n}(\hat{y}_i-\bar{y})^2, \qquad (1)$$

i.e., SSTO = SSE + SSR. This equality is the fundamental reason for choosing to minimize the sum of the squared deviations. In words, this equality says the following:

> *The overall variation in the data is equal to the sum of the overall deviation of the data from the fitted line plus the deviation of the line values from the mean.*

The coefficient of determination is defined as the proportion of the error not attributable to the line (SSTO-SSE) out of the total error: $\frac{SSTO-SSE}{SSTO}$, and thus calculated as

$$1-\frac{\sum_{i=0}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=0}^{n}(y_i-\bar{y})^2}$$

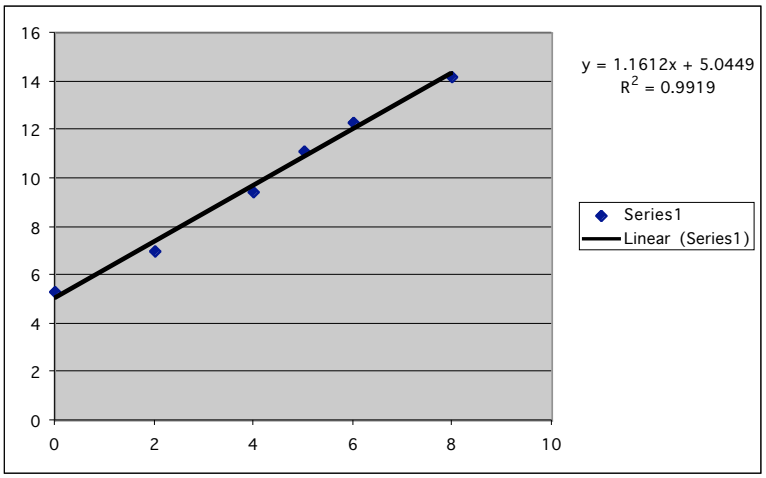
or, in words, this is the proportion of the total variation that can be accounted for by the line fit. It can also be calculated as:

$$\frac{\sum_{i=0}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=0}^{n}(y_i-\bar{y})^2}$$

That these two values are the same can be shown easily from the fundamental relationship (1). Also from (1) we know that this number is always between 0 and 1, and the closer to 1 it is, the

more the line "explains" the overall variation in the data. In fact, the coefficient of determination can be thought of as the %-goodness of fit of the line.

We can calculate this by hand, but luckily for us, many software packages can calculate this automatically when the trendline calculation is done.



Notice that the line is a pretty good fit (0.9919 is very close to 1).

We need to use caution when looking at the coefficient of determination aka R-Squared. When fitting a line, we can interpret the number as a percent goodness of fit. So for the line fitted above,

99% of the variation in the data is explained by the linear relationship y = 1.16 x — 5.04.

What shouldn't we say? None of the following statements are true!

~~The line will predict the *y* value 99% of the time.~~

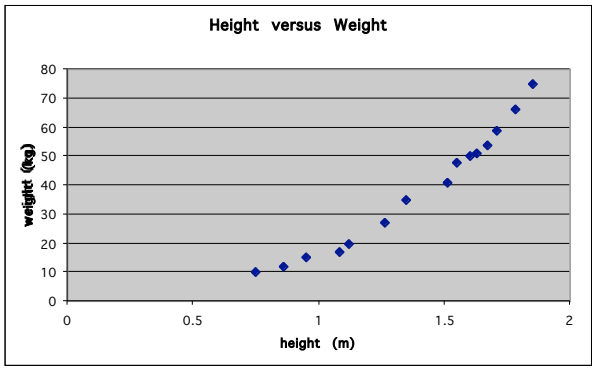
~~The *y* value predicted by the line will be 99% of the actual value.~~

Non-Linear Fits

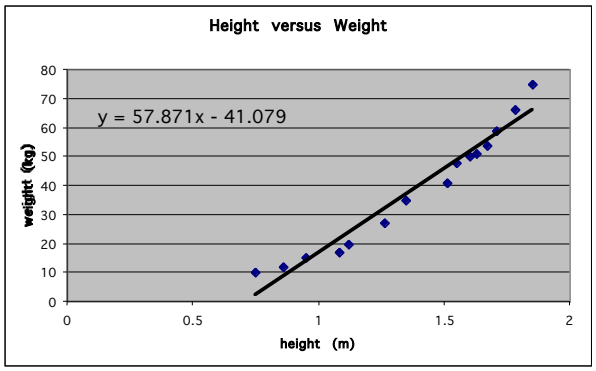
Consider the example in Chapter 5 of the Edward's and Hamson's *Guide to Modelling* – data collected to see if there is a relationship between height in meters and weight in kilograms. (Table 6.1 on page 103)

x(height)	y(weight)
0.75	10
0.86	12
0.95	15
1.08	17
1.12	20
1.26	27
1.35	35
1.51	41
1.55	48
1.6	50
1.63	51
1.67	54
1.71	59
1.78	66
1.85	75

These data, when plotted, give the following graph:



These data seem a little curved, but let's fit a line to it and take a look:



The linear fit isn't bad, but notice that there is a pattern to the deviations – below 1 meter and above 1.75 meters, the data points are above the line; in between 1 and 1.75 meters, the data points lie below the line. As mentioned in the previous section, patterns to the deviations are indicative of a non-linear fit.

So what non-linear function should we try to fit? If we take the approach from MAT 5950, we would think about what the data represent: Height versus weight in humans. Let's make some assumptions:

- Humans are geometrically similar
- Weight is proportional to volume
- Volume is a 3-D measurement whereas height is 1-D.

From these assumptions, it seems reasonable to try $weight \propto height^3$. So let's repeat the steps we used to find the formula for linear regression to find a formula for simple cubic regression. We start with minimizing the sum of the squared deviations:

$$\text{minimize } \sum_{i=0}^{n}\left(Ax_i^{\,3}-y_i\right)^2$$

Next we take the derivative with respect to the variable (A) and set the result equal to zero.

$$\frac{d}{dA}\sum_{i=0}^{n}\left(Ax_i^{\,3}-y_i\right)^2=\sum_{i=0}^{n}2\left(Ax_i^{\,3}-y_i\right)\left(x_i^{\,3}\right)=2A\sum_{i=0}^{n}x_i^6-2\sum_{i=0}^{n}x_i^3y_i=0 \qquad (1)$$

Solving this for A gives:

$$A = \frac{\sum_{i=0}^{n} x_i^{\,3} y_i}{\sum_{i=0}^{n} x_i^{\,6}}$$

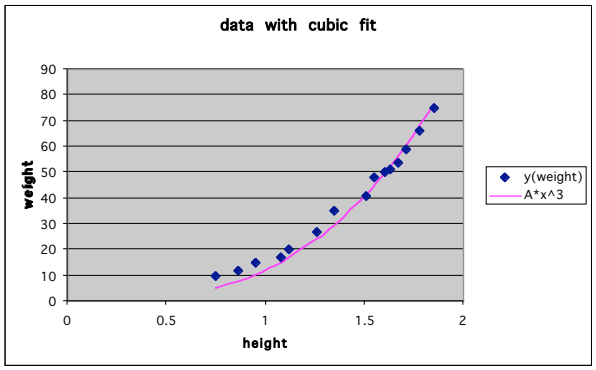
Organizing the calculations in table form:

x(height)	y(weight)	x^6	x^3*y
0.75	10	0.17797852	4.21875
0.86	12	0.40456724	7.632672
0.95	15	0.73509189	12.860625
1.08	17	1.58687432	21.415104
1.12	20	1.97382269	28.09856
1.26	27	4.00150414	54.010152
1.35	35	6.05344514	86.113125
1.51	41	11.8539116	141.160991
1.55	48	13.867245	178.746
1.6	50	16.777216	204.8
1.63	51	18.7553696	220.868097
1.67	54	21.6919616	251.503002
1.71	59	25.00211	295.012449
1.78	66	31.8068026	372.223632
1.85	75	40.0894751	474.871875
	sum:	194.777376	2353.53503
		A=	12.0832054

Building a new table with *x*, *y* and *A***x*^3 gives:

x(height)	y(weight)	A*x^3
0.75	10	5.09760228
0.86	12	7.68559529
0.95	15	10.3598382
1.08	17	15.2213588
1.12	20	16.9760336
1.26	27	24.1709541
1.35	35	29.7292165
1.51	41	41.6018841
1.55	48	44.9963465
1.6	50	49.4928093
1.63	51	52.3293055
1.67	54	56.2770821
1.71	59	60.4185766
1.78	66	68.1462818
1.85	75	76.5063254

Plotting this:



Note from the graph that this fit appears to be bad at the bottom. Why might our assumptions have let us down? The lower values in our data set come from people under 1 meter tall – namely children. Are children geometrically similar to adults? Not really.

Let's try some other models. In the last attempt we chose the power on the height to be 3. What if we let that be unknown as well – letting the data drive the power empirically?

$$\text{minimize} \sum_{i=0}^{n} \left(A x_i^{\,B} - y_i \right)^2$$

Next we take the derivative with respect to the variables (A and B) and set the results equal to zero. Here is the partial derivative with respect to A.

$$\frac{\partial}{\partial A} \sum_{i=0}^{n} \left(A x_i^{\,B} - y_i \right)^2 = \sum_{i=0}^{n} 2\left(A x_i^{\,B} - y_i \right)\left(x_i^{\,B} \right) = 2A \sum_{i=0}^{n} x_i^{\,2B} - 2 \sum_{i=0}^{n} x_i^{\,B} y_i = 0 \qquad (1)$$

Notice that the resulting equation is not going to be linear (B is in the exponent!), leading to a system that is not easy to handle. Is there another way? Yes! We can use a log trick to get the B out of the exponent:

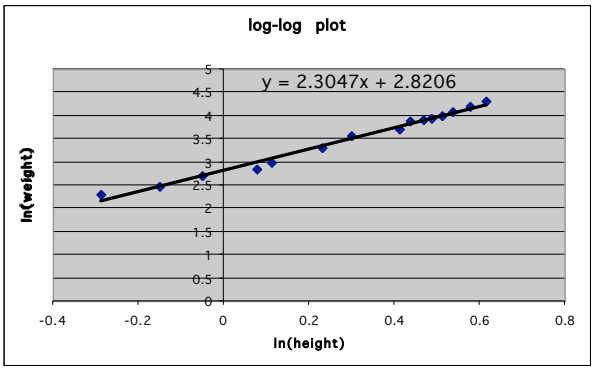
$$y = Ax^B \Rightarrow \ln(y) = \ln\left(Ax^B\right) \Rightarrow \ln(y) = \ln(A) + \ln\left(x^B\right)$$
$$\Rightarrow \ln(y) = \ln(A) + B\ln(x)$$

This says that Y = ln(y) and X = ln(x) have a linear relationship whenever y and x have a power relationship, with the slope equal to the power, B, and the y-intercept equal to ln(A) -- the log of the constant. So what should we do? Fit a line between ln(x) and ln(y)...

x(height)	y(weight)	ln(x)	ln(y)
0.75	10	-0.2876821	2.30258509
0.86	12	-0.1508229	2.48490665
0.95	15	-0.0512933	2.7080502
1.08	17	0.07696104	2.83321334
1.12	20	0.11332869	2.99573227
1.26	27	0.23111172	3.29583687
1.35	35	0.30010459	3.55534806
1.51	41	0.41210965	3.71357207
1.55	48	0.43825493	3.87120101
1.6	50	0.47000363	3.91202301

1.63	51	0.48858001	3.93182563
1.67	54	0.51282363	3.98898405
1.71	59	0.53649337	4.07753744
1.78	66	0.57661336	4.18965474
1.85	75	0.61518564	4.31748811

Here is a line fit to ln(x) and ln(y):

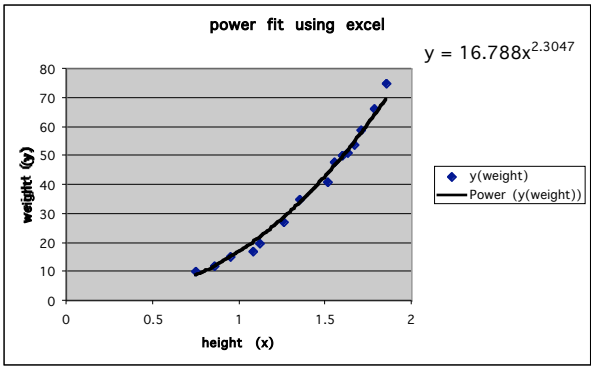


This is often referred to as log-log regression. This looks linear, and the deviations are more randomly placed. How do we recover the original function from this? We'll use the statement we made two paragraphs up: *$Y = ln(y)$ and $X = ln(x)$ have a linear relationship whenever y and x have a power relationship, with the slope equal to power, B, and the y-intercept equal to ln(A) -- the log of the constant*.

So 2.3047 = B = power and 2.8206 = ln(A) or exp(2.8206) = 16.79 = constant, giving

$$y = 16.79x^{2.305}$$

Here is the graph using the power curve:



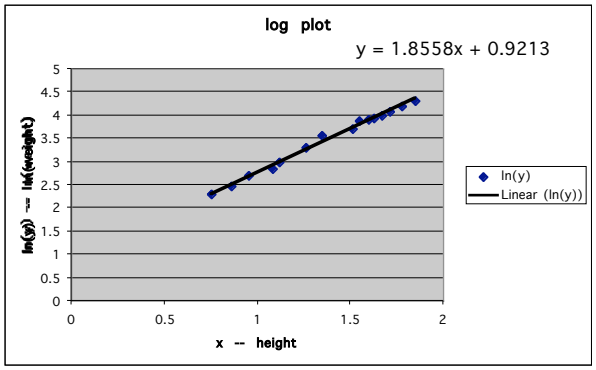
One last thing to try on this height-weight data set is **exponential regression**. A*exp(Bx) is another function that has a similar shape for $x > 0$. If we tried again with the naive approach to regression – minimizing the sum of the squared deviations, we would run into algebraic problems again, just like with the power curve calculations above. We can transform this in a similar way:

$$y = Ae^{Bx} \Rightarrow \ln(y) = \ln\left(Ae^{Bx}\right) \Rightarrow \ln(y) = \ln(A) + \ln\left(e^{Bx}\right)$$
$$\Rightarrow \ln(y) = \ln(A) + Bx\ln(e)$$
$$\Rightarrow \ln(y) = \ln(A) + Bx$$

This says that Y = ln(y) and X = x have a linear relationship whenever y and x have an exponential relationship, with the slope equal to the power, B, and the y-intercept equal to ln(A) -- the log of the constant. So what should we do? Fit a line between x and ln(y)...

x(height)	y(weight)	x(height)	ln(y)
0.75	10	0.75	2.30258509
0.86	12	0.86	2.48490665
0.95	15	0.95	2.7080502
1.08	17	1.08	2.83321334
1.12	20	1.12	2.99573227
1.26	27	1.26	3.29583687
1.35	35	1.35	3.55534806
1.51	41	1.51	3.71357207
1.55	48	1.55	3.87120101
1.6	50	1.6	3.91202301
1.63	51	1.63	3.93182563
1.67	54	1.67	3.98898405
1.71	59	1.71	4.07753744
1.78	66	1.78	4.18965474
1.85	75	1.85	4.31748811

Here is a graph of the linearized data:

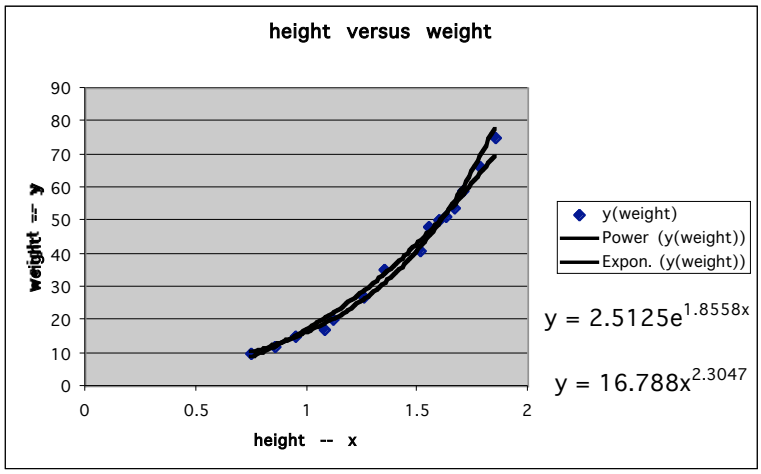


This is often referred to as log regression. This looks linear, and as in the power fit the deviations are more randomly placed. How do we recover the original function from this? We'll use the statement we made two paragraphs up: *Y = ln(y) and X = x have a linear relationship whenever y and x have an exponential relationship, with the slope equal to power, B, and the y-intercept equal to ln(A) -- the log of the constant*.

So 1.8558 = B = power and 0.9213 = ln(A) or exp(0.9213) = 2.513 = constant, giving

$$y = 2.513e^{1.856x}$$

Here is the graph using the power curve and the exponential curve and showing the functions together:



Checking for Goodness of Fit in the Non-Linear Case

Since these fits are based upon a linearization of the original data, the same "rules" apply when deciding if the fit is good. Look at the transformed data and check:

1. The *x* values are fixed, not random.
2. The *y* data values are independent of each other and normally distributed. This is usually the case if the data were collected carefully. One notable exception is values that are serially collected for which the independent variable is not time. These time series data may depend on time and hence on each other.
3. The residuals are independent of each other and normally distributed with mean 0.

How do we recognize when these assumptions are violated? The single best way to tell if the regression is good is – LOOK AT THE GRAPH! Look for the following things both with the original and with the transformed data:

- Does the curve go through the middle of the data? no = bad
- Is there a pattern to the deviations? yes = bad
- Is the shape of the data rectangular (rather than wedge shaped)? In the case of the original data, look for a rectangle that is bent to follow the curve. no = bad

Coefficient of Determination as used with Non-Linear Fits

Note that in each case we could get an R-squared value. When fitting non-linear curves to the data, in particular the power and exponential fits we have seen already, caution must be used when looking at this number. It tells us how good the fit is for the transformed data — NOT the original data. In particular, logarithmic transformations change the distances between the *y* data values and the mean ($\overline{y}$) in a non-uniform way. This implies that the traditional explanation for the meaning of the coefficient of determination doesn't work if the data have been transformed in this way.

We can still look at the coefficient of determination in a comparative way to choose between two curves of the same kind, such as two different exponentials, two power fits, etc., but not as a way to choose between the fits from different kinds of curves. This mis-use is common in applications of statistics to the social sciences — remember to avoid this!!!

So, bottom line, what is the best way to tell if we have chosen the best curve? Use your eyes! Look at the data with the various fits and choose the one that appears to come closest to all of the points and gives a random pattern in the deviations.

Problems

1. The following data table contains a sample of 40 individuals out of over 7000 who participated in a study in Honolulu, HI in 1969. Answer each question, and explain your conclusions. Is there a link between weight and cholesterol levels? Please note: Lower total cholesterol is considered healthier. Does age matter? Does physical activity matter? Does smoking affect blood pressure?

Some experts propose that comparing weight alone to blood pressure is not as good as comparing weight per cm of height to blood pressure, i.e., create a new variable that is weight divided by height for each person. Can this new variable be used to predict blood pressure?

Honolulu Study Data (1972)

Education (Highest completed)	Weight (kg)	Height (cm)	Age	Smoker	Physical Activity	Chole-sterol	Systolic Blood Pressure
primary	70	165	61	y	moderate	199	102
none	60	162	52	n	heavy	267	138
none	62	150	53	y	moderate	272	190
primary	66	165	51	y	moderate	166	122
primary	70	162	51	n	heavy	239	128
high school	59	165	53	n	moderate	189	112
none	47	160	61	n	heavy	238	128
intermediate	66	170	48	y	moderate	223	116
college	56	155	54	n	heavy	279	134
primary	62	167	48	n	moderate	190	104
high school	68	165	49	y	heavy	240	116
none	65	166	48	n	moderate	209	152
none	56	157	55	n	heavy	210	134
primary	80	161	49	n	moderate	171	132
intermediate	66	160	50	n	heavy	255	130
high school	91	170	52	n	heavy	232	118
intermediate	71	170	48	y	moderate	147	136
college	66	152	59	n	heavy	268	108
none	73	159	59	n	moderate	231	108
high school	59	161	52	n	moderate	199	128
none	64	162	52	y	moderate	255	118
intermediate	55	161	52	y	moderate	199	134
primary	78	175	50	y	moderate	228	178
primary	59	160	54	n	moderate	240	134
intermediate	51	167	48	y	heavy	184	162
intermediate	83	171	55	n	moderate	192	162
primary	66	157	49	y	heavy	211	120
high school	61	165	51	n	moderate	201	98
primary	65	160	53	n	moderate	203	144
intermediate	75	172	49	n	moderate	243	118
high school	61	164	49	n	heavy	181	118
none	73	157	53	y	heavy	382	138
primary	66	157	52	n	moderate	186	134
none	73	155	48	n	heavy	198	108
primary	61	160	53	n	moderate	165	96
intermediate	68	162	50	n	heavy	219	142
primary	52	157	50	n	heavy	196	122
college	73	162	50	n	moderate	239	146
none	52	165	61	y	heavy	259	126
none	56	162	53	y	moderate	162	176

2. Wild black bears were anesthetized, and their bodies were measured and weighed. One goal of the study was to look at whether the distribution of the weights is different for male and female bears.

The second goal of the study was to determine the kind of relationship that explained best the association between length and weight. The third goal of the study was to find a linear equation for weight in terms of some other measurable characteristic for forest rangers, so they could estimate the weight of a bear based on that one measurement. This would be useful because in the field it is easier to measure a length than it is to weigh a bear with a scale. Use the dataset to investigate these goals.

link to data: http://mathsci2.appstate.edu/~hph/SageMath/bears.csv

3. In order to avoid the expense and inconvenience of using a water tank to determine body fat, we wish to find a measurement that can predict body fat. The data set provided below consists of estimates of body fat determined by underwater weighing along with various body circumference measurements for 252 men. Which of these measurements is the best predictor of body fat?

link to data: http://mathsci2.appstate.edu/~hph/SageMath/bodyfat.csv

4. Is there a relationship between brain and body weight in mammals? Use the data provided for 53 species' average adult male body weight to investigate this question.

link to data: http://mathsci2.appstate.edu/~hph/SageMath/brain-body.csv

5. World Rankings: Data Source: http://www.photius.com/rankings/

This problem is designed to let you grub around with real data. Real data are not always "nice," so don't be surprised if you run into less than satisfying results. Your assignment is to try various fits and use the ideas from class to choose "the best" relationship or to say why you think there is no relationship. Some examples of relationships you might investigate:

- Explore the relationship between life expectancy and GDP.
- Explore the relationship between energy use and GDP.
- Explore the relationship between literacy and infant mortality rate.
- Explore the relationship between CO2 emission and GDP.
- Explore the relationship between population and energy use.
- Explore the relationship between literacy and life expectancy.
- Explore the relationship between CO2 emissions and energy use.
- Explore the relationship between fertility rate and % over 60.