Expert Journal of Business and Management, Volume 3, Issue 2, pp. 205-228, 2015
 ISSN 2344-6781
http://Business.ExpertJournals.com

Linear and Non-Linear Regression: Powerful and Very Important Forecasting Methods

Athanasios VASILOPOULOS*

St. John's University, United States

Regression Analysis is at the center of almost every Forecasting technique, yet few people are comfortable with the Regression methodology. We hope to improve the level of comfort with this article. In this article we briefly discuss the theory behind the methodology and then outline a step-by-step procedure, which will allow almost everyone to construct a Regression Forecasting function for both the linear and some non-linear cases. Also discussed, in addition to the model construction mentioned above, is model testing (to establish significance) and the procedure by which the Final Regression equation is derived and retained to be used as the Forecasting equation. Hand solutions are derived for some small-sample problems (for both the linear and non-linear cases) and their solutions are compared to the MINITAB-derived solutions to establish confidence in the statistical tool, which can be used exclusively for larger problems.

Keywords: *Linear Regression, Non-Linear Regression, Best-Fitting Model, Forecasting*

JEL Classification: *M10*

1. Introduction and Model Estimation for the Linear Model

Regression analysis, in which an equation is derived that connects the value of one dependent variable (Y) to the values of one independent variable X (linear model and some non-linear models), starts with a given bivariate data set and uses the Least Squares Method to assign the best possible values to the unknown multipliers found in the models we wish to estimate. The bivariate data, used to estimate the linear model and some non-linear models, consists of n ordered pairs of values:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

The linear model we wish to estimate, using the given data, is:

$$y = a + bx \quad (1)$$

* Corresponding Author:
Athanasios Vasilopoulos, Ph.D., St. John's University, The Peter J. Tobin College of Business, CIS/DS Department, 8000 Utopia Parkway Jamaica, N.Y. 11439

Article History:
Received 2 November 2015 | Accepted 23 November 2015 | Available Online 19 December 2015

Cite Reference:
Vasilopoulos, A., 2015. Linear and Non-Linear Regression: Powerful and Very Important Forecasting Methods. *Expert Journal of Business and Management*, 3(2), pp.205-228

while the non-linear models of interest are given by

$$y = ke^{cx} \text{ (Exponential Model)} \quad (2)$$
$$y = ax^{b} \text{ (Power Model)} \quad (3)$$

and

$$y = a + bx + cx^2 \text{ (Quadratic Model)} \quad (4)$$

To estimate model (1) we use the Least Squares Methodology, which calls for the formation of the quadratic function:

$$Q(a,b) = \sum_{i=1}^{n}\left[y_{actual} - y_{linearequation}\right]^2 = \sum_{i=1}^{n} y_i^2 - 2a\sum_{i=1}^{n} y_i - 2b\sum_{i=1}^{n} x_i y_i + 2ab\sum_{i=1}^{n} x_i + b^2\sum_{i=1}^{n} x_i^2 \quad (5)$$

To derive the "normal" equations for the linear model from which the values of **a** and **b** of the linear model are obtained, we take the partial derivative of Q(a,b) of equation (5) with respect to a and b, set each equal to zero, and then simplify:

The result is:

$$\frac{\partial Q(a,b)}{\partial a} = -2\sum_{i=1}^{n} y_i + 2b\sum_{i=1}^{n} x_i + 2an \quad (6)$$

and

$$\frac{\partial Q(a,b)}{\partial b} = -2\sum_{i=1}^{n} X_i y_i + 2a\sum_{i=1}^{n} x_i + 2b\sum_{i=1}^{n} x_i^2 \quad (7)$$

When (6) and (7) are set equal to zero and simplified, we obtain the "Normal" equations for the linear model:

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (8)$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad (9)$$

The only unknowns in equation (8) and (9) are **a** and **b** and they should be solved for them simultaneously, thus deriving (or estimating) the linear model. This is so because all the other values of equations (8) and (9) come from the given data, where:

n = number of ordered pairs (x_i, y_i)

$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$ = sum of the x values

$\sum_{i=1}^{n} y_i = y_1 + y_2 + \ldots + y_n$ = sum of the y values

$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \ldots + x_n^2$ = sum of the given x values, which are first squared

$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$ = sum of the products of the x_i and y_i values in each ordered pair.

Note: The values of (a) and (b) obtained from the Normal equations correspond to a minimum value for the Quadratic function Q(a,b) given by equation (5), as can be easily demonstrated by using the

Optimization methodology of Differential Calculus for functions of 2 independent variables.

To complete the Estimation of the Linear model we need to find the standard deviation for a, σ(a), and b, σ(b), which are needed for testing of the significance of the model. The standard deviations, σ(a), and σ(b), are given by:

$$\sigma(a)=\frac{\hat{\sigma}}{\sqrt{n}}\left[\frac{\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]^{1/2}=\frac{\hat{\sigma}}{\sqrt{n}}\left[\frac{\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}\right]^{1/2} \quad (10)$$

and

$$\sigma(b)=\frac{\hat{\sigma}}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{1/2}}, \quad (11)$$

where:

$$\hat{\sigma}=\left[\frac{\sum_{i=1}^{n}y_i^2-a\sum_{i=1}^{n}y_i-b\sum_{i=1}^{n}x_iy_i}{n-2}\right]^{1/2} \quad (12)$$

The **a** and **b** in equation (12) come from the solution of equations (8) and (9) while $\sum_{i=1}^{n}y_i^2$, $\sum_{i=1}^{n}y_i$, and $\sum_{i=1}^{n}x_iy_i$ come directly from the given bivariate data.

2. Model Testing

Now that our model of interest has been estimated, we need to test for the significance of the terms found in the estimated model. This is very important because the results of this testing will determine the final equation which will be retained and used for Forecasting purposes.

Testing of the linear model consists of the following steps:

2.1. Testing for the significance of each term separately

Here we test the hypotheses:

1. H_0: $\beta = 0$ vs H_1: $\beta \neq 0$, and
2. H_0: $\alpha = 0$ vs H_1: $\alpha \neq 0$, based on our knowledge of b, σ(b), a, and σ(a).

If $n \geq 30$, we calculate

$$Z_b^*=\frac{b}{\sigma(b)}$$

and

$$Z_a^*=\frac{a}{\sigma(a)}$$

and compare each to $Z_{\alpha/2}$ (where $Z_{\alpha/2}$ is a value obtained from the standard Normal Table when α, or 1 - α, is specified).

For example if $\alpha = 0.05$, $Z_{\alpha/2} = Z_{0.025} = 1.96$; if $\alpha = 0.10$, $Z_{\alpha/2} = Z_{0.05} = 1.645$; if $\alpha = 0.02$, $Z_{\alpha/2} = Z_{0.01} = 2.33$ and if $\alpha = 0.01$, $Z_{\alpha/2} = Z_{0.005} = 2.58$).

If $Z^*_b > Z_{\alpha/2}$ (or $Z^*_b < -Z_{\alpha/2}$), the hypothesis H_0: $\beta = 0$ is rejected and we conclude that $\beta \neq 0$ and the term bx (in the estimated model $\hat{y}=a+bx$) is important for the calculation of the value of y. Similarly, if $Z^*_a > Z_{\alpha/2}$ (or $Z^*_a < -Z_{\alpha/2}$), H_0: $\alpha = 0$ is rejected, and we conclude that the linear equation $\hat{y}=a+bx$ does not go through the origin.

If n < 30, we calculate

$$t_b^* = \frac{b}{\sigma(b)}$$

and

$$t_a^* = \frac{a}{\sigma(b)}$$

and compare each to $t_{n-2(\alpha/2)}$, for a given α value, where $t_{n-2(\alpha/2)}$ is obtained from the t-distribution table, with the same interpretation for H_0: $\beta = 0$ and H_0: $\alpha = 0$ as above.

But, instead of hypothesis testing, we can construct Confidence Intervals for β and α using the equations:

$$P\left[b - Z_{\propto/2}\sigma(b) \leq \beta \leq b + Z_{\propto/2}\sigma(b)\right] = 1-\propto \quad (13)$$

and, if $n \geq 30$,

$$P\left[a - Z_{\propto/2}\sigma(a) \leq \propto \leq a + Z_{\propto/2}\sigma(a)\right] = 1-\propto \quad (14)$$

or

$$P\left[b - t_{n-2(\propto/2)}\sigma(b) \leq \beta \leq b + t_{n-2(\propto/2)}\sigma(b)\right] = 1-\propto \quad (15)$$

and, if $n < 30$,

$$P\left[a - t_{n-2(\propto/2)}\sigma(a) \leq \propto \leq a + t_{n-2(\propto/2)}\sigma(a)\right] = 1-\propto \quad (16)$$

If the hypothesized values: $\beta = 0$ falls inside the Confidence Intervals given by equations (13) or (15), or $\alpha = 0$ falls inside the Confidence Intervals given by equations (14) or (16), the corresponding hypotheses H_0: $\beta = 0$ and H_0: $\alpha = 0$ are not rejected and we conclude that $\beta = 0$ (and b = 0 and the term bx is not important for the calculation of y) and $\alpha = 0$ (i.e. a = 0 and the line goes through zero). If for a given data set, we performed the above-discussed tests, we will obtain one of 4 possible conclusions:

A) H_0: $\beta = 0$ and H_0: $\alpha = 0$ are both rejected; Therefore $\beta \neq 0$, and $\alpha \neq 0$, and both the terms **a** and **bx** are important to the calculation of y. In this case the final equation is $\hat{y}=a+bx$, with both terms staying in the equation.

B) H_0: $\beta = 0$ is rejected, but H_0: $\alpha = 0$ is not rejected. Therefore $\beta \neq 0$ but $\alpha = 0$ and the term **a** is not important to the calculation of y. In this case the final equation is $\hat{y}=a+bx$, with the term **a** dropping out of the equation.

C) H_0: $\beta = 0$ is not rejected but H_0: $\alpha = 0$ is rejected. Therefore $\beta = 0$ and the term **bx** is not important for the calculation of y, while $a \neq 0$ and is important to the calculation of y. In this case the final equation is $\hat{y} = a$, with the term **bx** dropping out of the equation

D) H_0: $\beta = 0$ and H_0: $\alpha = 0$ are both not rejected; Therefore $\beta = 0$, and $\alpha = 0$, and both terms **a** and **bx** are not important to the calculation of y. In this case the final equation will be $\hat{y} = 0$, with both terms **a** and **bx** dropping out of the equation.

2.2. Testing for the Significance of the Entire Linear Equation

This test consists of testing the hypothesis:

1. H_0: $\alpha = \beta = 0$ vs H_0: α and β are not both equal to 0, or
2. H_0: The Entire Regression equation is not significant vs H_1: The Entire Regression equation is significant

For a given bivariate data set and a given α value, we need to first calculate:

$$\text{Total Sum of Squares} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{17}$$

$$\text{Regression Sum of Squares} = RSS_b = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = b^2 \left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] \tag{18}$$

$$\text{Error Sum of Squares} = ESS = \sum_{i=1}^{n}(y_i - \hat{y})^2 = Q^* = \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i \tag{19}$$

$$\text{Sum of Squares Due to the Constant} = SS_a = \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{20}$$

Then we calculate:

$$F^*_{Total} = \frac{(RSS_b + SS_a)/2}{ESS/n-2} \tag{21}$$

and compare F^*_{Total} to $F^2_{n-2}(\alpha)$, which is a tabulated value, for a specified α value. If $F^*_{Total} > F^2_{n-2}(\alpha)$, we reject H_0 and conclude that the entire regression equation (i.e. $\hat{y}$=a+bx) or that both the constant term **a**, and the factor x (and term **bx**) are significant to the calculation of the y value, simultaneously.

Note 1:

When TSS, RSS_b, and ESS are known, we can also define the coefficient of determination R^2, where:

$$R^2 = \frac{RSS_b}{TSS} = 1 - \frac{ESS}{TSS} \tag{22}$$

where $0 \leq R^2 \leq 1$, which tells us how well the regression equation $\hat{y} = a + bx$ fits the given bivariate data. A value of R close to 1 implies a good fit.

Note 2:

$$r = correlation\ coefficient = \sqrt{R^2} \tag{23}$$

2.3. A Bivariate Example

A sample of 5 adult men for whom heights and weights are measured gives the following results (Table 1):

Table 1. *Given bivariate data set (n =5)*

x = H	y = W	$x^2=H^2$	$y^2=W^2$	xy = HW
64	130	64^2	130^2	64 x 130
65	145	65^2	145^2	65 x 145
66	150	66^2	150^2	66 x 150
67	165	67^2	165^2	67 x 165
68	170	68^2	170^2	68 x 170

For this Bivariate Data set we have: n = 5

$$\sum_{i=1}^{5} x_i = 64 + 65 + 66 + 67 + 68 = 330$$

$$\sum_{i=1}^{5} x_i^2 = 64^2 + 65^2 + 66^2 + 67^2 + 68^2 = 21{,}790$$

$$\sum_{i=1}^{5} y_i = 130 + 145 + 150 + 165 + 170 = 760$$

$$\sum_{i=1}^{5} y_i^2 = 130^2 + 145^2 + 150^2 + 165^2 + 170^2 = 116{,}550$$

$$\sum_{i=1}^{5} x_i y_i = (64 \times 130) + (65 \times 145) + (66 \times 150) + (67 \times 165) + (68 \times 170) = 50{,}260$$

To obtain the linear equation $\hat{y} = a + bx$, we substitute the values of n, $\sum_{i=1}^{5} x_i$, $\sum_{i=1}^{5} x_i^2$, $\sum_{i=1}^{5} x_i y_i$ to equations (8) and (9) and obtain:

$$\begin{cases} 5a + 330b = 760 \\ 330a + 21{,}790b = 50{,}260 \end{cases}$$

When these equations are solved simultaneously we obtain: a = -508 and b = 10, and the regression equation is

$$\hat{y} = a + bx = -508 + 10x.$$

Then, using the values of a = -508, b=10, and $\sum_{i=1}^{5} y_i$, $\sum_{i=1}^{5} y_i^2$ and $\sum_{i=1}^{5} x_i y_i$ we obtain from equation (12):

$$\hat{\sigma} = \left[\frac{116{,}550 - (-508)(760) - (10)(50{,}260)}{5-2}\right]^{1/2} = \left[\frac{30}{3}\right]^{1/2} = \sqrt{10} = 3.16228$$

and from equations (10) and (11):

$$\sigma(a) = \frac{\sqrt{10}}{\sqrt{5}}\left[\frac{21{,}790}{21{,}790 - \frac{(330)^2}{5}}\right]^{1/2} = \sqrt{2}\left[\frac{21{,}790}{10}\right]^{1/2} = \left[\frac{2x21{,}790}{10}\right]^{1/2} = \sqrt{4358} = 66.015$$

$$\sigma(b) = \frac{\sqrt{10}}{\left[21{,}790 - \frac{(330)^2}{5}\right]^{1/2}} = \frac{\sqrt{10}}{[10]^{1/2}} = \frac{\sqrt{10}}{\sqrt{10}} = 1$$

Since n=5<30, **a** and **b** are distributed as $t_{n-2} = t_3$variables and when $\alpha = 0.05$, $t_3(\alpha/2) = t_3(0.025) =$ $= \pm 3.1824$.

Then the hypotheses H_0: $\beta = 0$ vs. H_1: $\beta \neq 0$, and H_0: $\alpha = 0$ vs. H_0: $\alpha \neq 0$ are both rejected because:

$$t_a^* = \frac{a}{\sigma(a)} = \frac{-508}{66{,}015} = -7.695 < -3.1824$$

and

$$t_b^* = \frac{b}{\sigma(b)} = \frac{10}{1} = 10 > 3.1824$$

Therefore, the final equation is

$$\hat{y} = a + bx = -508 + 10x.$$

To test for the significance of the entire equation, and to calculate the coefficient of determination, we first evaluate, TSS, RSS_b, ESS, SS_a using equations (17) – (20) and obtain:

$$TSS = 116{,}550 - \frac{(760)^2}{5} = 1030$$

$$RSS_b = 10^2\left[21{,}790 - \frac{(330)^2}{5}\right] = 10^2(10) = 1000$$

$$ESS = 116{,}550 - (-508)(760) - 10(50{,}260) = 30$$

$$SS_a = \frac{(760)^2}{5} = 115{,}520$$

From equation (22), we obtain $R^2 = 1000/1030 \approx 0.971$, which tells us that 97% of the variation in the values of Y can be explained (or are accounted for) by the variable X included in the regression equation and only 3% is due to other factors. Since R^2 is close to 1, the fit of the equation to the data is very good.

Note:

The correlation coefficient r, which measures the strength of the linear relationship between Y and X is related to the coefficient of determination by:

$$r = \sqrt{R^2} = \sqrt{0.97} = 0.985$$

for this example. Clearly X and Y are very strongly linearly related.

Using equation (21) we obtain:

$$F^*_{Total} = \frac{(RSS_b + SS_a)/2}{ESS/n-2} = \frac{(1000+115{,}520)/2}{30/3} = \frac{58{,}260}{10} = 5{,}826$$

when F^*_{Total} is compared to

$$F^2_{n-2}(\alpha) = F^2_3(\alpha) = \begin{cases} 10.13 & if\ \alpha = 0.05 \\ 34.12 & if\ \alpha = 0.01 \end{cases}$$

H_0 (The entire regression equation is not significant) is rejected, and we conclude that the entire regression equation is significant.

3. MINITAB Solution to the Linear Regression Problem

We enter the given data and issue the regression command as shown in Table 2.

Table 2. *Data set in MINITAB*

MTB > Set C1
DATA> 64 65 66 67 68
DATA> end
MTB > set C2
DATA> 130 145 150 165 170
DATA> end
MTB > Name C1 'X' C2 'Y'
MTB > REGRESS 'Y' 1 'X'

and obtain the MINITAB output presented in Table 3, Table 4, and Figure 1.

Table 3. *Regression Analysis: Y versus X*

Regression equation:			Y = - 508 + 10.0 X		
Predictor	**Coef**	**SE Coef**	**T**	**p**	
Constant	-508.000	66.020	-7.700	0.005	
X	10.000	1.000	10.000	0.002	
Regression fit:	**S**	**R-Sq**	**R-Sq (adj)**		
	3.162	97.1%	96.1%		
Analysis of Variance:					
Source	**DF**	**SS**	**MS**	**F**	**p**
Regression	1	1000.0	1000.0	100.0	0.002
Residual Error	3	30.0	10.0		
Total	4	1030.0			

Table 4. *Correlations: Y, X*

Pearson correlation of Y and X	0.985
P-Value	0.002

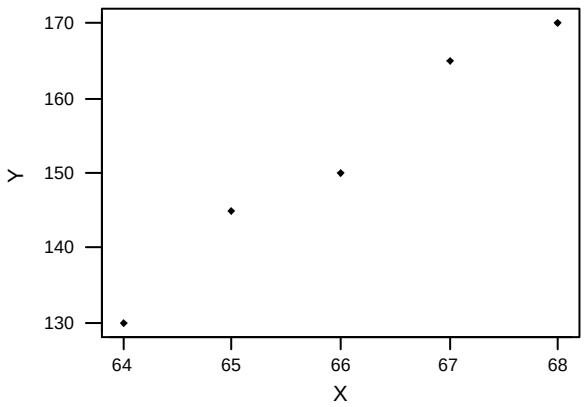

Figure 1. *Plot Y * X*

When we compare the MINITAB and hand solutions, they are identical. We obtain the same equation $\hat{y}$ = -508 + 10x, the same standard deviations for **a** and **b** (under SE Coefficient) and the same t values, the same R^2, the same s = σ and $\sigma^2 = 10$. Notice also that an Analysis of Variance table provides the values for RSS_b, ESS, and TSS. The only value missing is SS_a, which can be easily calculated from

$$SS_a = \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$$

The MINITAB solution also gives a p-value for each coefficient. The p-value is called the "Observed Level of Significance" and represents the probability of obtaining a value more extreme than the value of the test statistic. For example the p-value for the predictor X is calculated as p = 0.002, and it is given by:

$$\text{p - value} = \text{P}(\text{t} > \text{t}^* = 10) = \int_{10}^{\infty} f(t)dt = 0.002 \qquad (24)$$

The p-value has the following connection to the selected α-value.
If $p \geq \alpha$, do not reject H_0

If $p < \alpha$, reject H_0
Since $p = 0.002 < \alpha = 0.05$, H_0: $\beta = 0$ will be rejected.

4. Introduction and Model Estimation for Some Non-Linear Models of Interest

Sometimes two variables are related but their relationship is not linear and trying to fit a linear equation to a data set that is inherently non-linear will result in a bad-fit. But, because non-linear regression is, in general, much more difficult than linear regression, we explore in this part of the paper estimation methods that will allow us to fit non-linear equations to a data set by using the results of linear regression which is much easier to understand and analyze.

This becomes possible by first performing logarithmic transformations of the non-linear equations, which change the non-linear into linear equations, and then using the normal equations of the linear model to generate the normal equations of the "linearized" non-linear equations, from which the values of the unknown model parameters can be obtained. In this paper we show how the exponential model, $\hat{y} = ke^{cx}$, and the power model, $\hat{y} = ax^b$ (for $b \neq 1$) can be easily estimated by using logarithmic transformations to first derive the linearized version of the above non-linear equations, namely:

$$\ln \hat{y} = \ln k + cx$$

and

$$\ln \hat{y} = \ln a + b \ln x,$$

and then comparing these to the original linear equation, $\hat{y} = a + bx$, and its normal equations (see equations (8) and (9)).

Also discussed is the quadratic model, $\hat{y} = a + bx + cx^2$ which, even though is a non-linear model, can be discussed directly using the linear methodology. But now we have to solve simultaneously a system of 3 equations in 3 unknowns, because the normal equations for the quadratic model become:

$$\begin{aligned} na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 &= \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 &= \sum_{i=1}^{n} x_i^2 y_i \end{aligned} \quad (25)$$

A procedure is also discussed which allows us to fit these four models (i.e. linear, exponential, power, quadratic), and possibly others, to the same data set, and then select the equation which fits the data set "best". These four models are used extensively in forecasting and, because of this, it is important to understand how these models are constructed and how MINITAB can be used to estimate such models efficiently.

4.1. The Linear Model and its Normal Equations

The linear model and the normal equations associated with it as explained above, are given by:
Linear Model

$$y = a + bx \quad (1)$$

Normal Equations

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (8)$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad (9)$$

4.2. The Exponential Model

The *exponential model* is defined by the equation:

$$\hat{y} = ke^{cx} \tag{26}$$

Our objective is to use the given data to find the best possible values for k and c, just as our objective in equation (1) was to use the data to find the best (in the least-square sense) values for a and b.

Taking natural logarithms (i.e. logarithms to the base e) of both sides of equation (26) we obtain

$$\ln(\hat{y} = ke^{cx})$$

or

$$\ln \hat{y} = \ln(ke^{cx}) \tag{27}$$

4.2.1. Logarithmic Laws

To simplify equation (27), we have to use some of the following laws of logarithms:

i) $\log (A \cdot B) = \log A + \log B$ (28)

ii) $\log (A/B) = \log A - \log B$ (29)

iii) $\log (A^n) = n \log A$ (30)

Then, using equation (28) we can re-write equation (27) as:

$$\ln y = \ln k + \ln e^{cx} \tag{31}$$

and, by applying equation (30) to the second term of the right hand side of equation (31), equation (31) can be written finally as:

$$\ln y = \ln k + cx(\ln e)$$

or

$$\ln y = \ln k + cx \text{ (because } \ln e = \log_e e = 1) \tag{32}$$

Even though equation (26) is non-linear, as can be verified by plotting y against x, equation (32) is linear (i.e. the logarithmic transformation changed equation (26) from non-linear to linear) as can be verified by plotting: ln y against x.

But, if equation (32) is linear, it should be similar to equation (1), and must have a set of normal equations similar to the normal equations of the linear model (see equations (8) and (9)).

Question: How are these normal equations going to be derived?

Answer: We will compare the "transformed linear model", i.e. equation (32), to the actual linear model (equation (1)), note the differences between these two models, and then make the appropriate changes to the normal equations of the linear model to obtain the normal equations of the "transformed linear model".

4.2.2. Comparison of the Logarithmic Transformed Exponential Model to the Linear Model

To make the comparison easier, we list below the 2 models under consideration, namely:

a) Original Linear Model:

$$y = a + bx \tag{1}$$

b) Transformed Linear Model:

$$\ln y = \ln k + cx \tag{32}$$

Comparing equations (1) and (32), we note the following three differences between the two models:

i. ***y*** in equation (1) has been replaced by ***ln y*** in equation (32)

ii. ***a*** in equation (1) has been replaced by ***ln k*** in equation (32)

iii. ***b*** in equation (1) has been replaced by ***c*** in equation (32)

4.2.3. Normal Equations of Exponential Model

When the three changes listed above are applied to the normal equations of the actual linear model (equations (8) and (9)), we will obtain the normal equations of the "transformed model". The normal equations of the "transformed linear model" are:

$$n(\ln k) + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \ln y_i \tag{33}$$

$$(\ln k)\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i(\ln y_i) \tag{34}$$

In equations (33) and (34) all the quantities are known numbers, derived from the given data as will be shown later, except for: ln *k* and c, and equations (33) and (34) must be solved simultaneously for ln *k* and c.

Suppose that for a given data set, the solution to equations (33) and (34) produced the values:

$$\ln k = 0.3 \quad \text{and} \quad c = 1.2 \tag{35}$$

If we examine the exponential model (equation (26)), we observe that the value of c = 1.2 can be substituted directly into equation (26), but we do not yet have the value of k; instead we have the value of ln *k* = 0.3!

Question: If we know: ln *k* = 0.3, how do we find the value of k?

Answer: If ln *k* = 0.3, then: $k = e^{0.3} \approx (2.718281828)^{0.3} \approx 1.349859$

Therefore, now that we have both the ***k*** and ***c*** values, the non-linear model, given by equation (26), has been completely estimated.

4.3. The Power Model

Another non-linear model, which can be analyzed in a similar manner, is the Power Model defined by the equation:

$$\hat{y} = ax^b \tag{36}$$

which is non-linear if $b \neq 1$ and, as before, we must obtain the best possible values for a and b (in the least-square sense) using the given data.

4.3.1. Logarithmic Transformation of Power Model

A logarithmic transformation of equation (36) produces the "transformed linear model"

$$\ln y = \ln a + b \ln x \tag{37}$$

When equation (37) is compared to equation (1), we note the following 3 changes:

i. ***y*** in equation (1) has been replaced by ***ln y*** in equation (37)
ii. ***a*** in equation (1) has been replaced by ***ln a*** in equation (37) (38)
iii. ***x*** in equation (1) has been replaced by ***ln x*** in equation (37)

When the changes listed in (38) are substituted into equations (8) and (9), we obtain the normal equations for this "transformed linear model" which are given by equations (39) and (40) below:

4.3.2. Normal Equations of Power Model

$$n(\ln a) + b\sum_{i=1}^{n} \ln x_i = \sum_{i=1}^{n} \ln y_i \tag{39}$$

$$(\ln a)\sum_{i=1}^{n} \ln x_i + b\sum_{i=1}^{n} (\ln x_i)^2 = \sum_{i=1}^{n} (\ln x_i)(\ln y_i) \tag{40}$$

Equations (39) and (40) must be solved simultaneously for (ln *a*) and b.

If ln *a* = 0.4, then a = $e^{0.4} \approx (2.718251828)^{0.4} \approx 1.491825$ and, since we have numerical values for both ***a*** and ***b***, the non-linear model defined by equation (36) has been completely estimated.

4.4. Derivation of the normal equations for the Quadratic model, $y = a + bx + cx^2$

To derive the normal equations of the quadratic model, first form the function

$$Q(a,b,c) = \sum_{i=1}^{n}\left[y_i - a - bx_i - cx_i^2\right]^2 \tag{41}$$

Then take the partial derivatives: $\frac{\partial Q}{\partial a}$, $\frac{\partial Q}{\partial b}$, $\frac{\partial Q}{\partial c}$, and set each equal to 0, to obtain the 3 equations needed to solve for a, b, c.

We obtain:

$$\frac{\partial Q}{\partial a} = +2\sum_{i=1}^{n}\left[y_i - a - bx_i - cx_i^2\right](-1) = 0,$$

or:

$$na + b\sum_{i=1}^{n}x_i + c\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}y_i \tag{42}$$

$$\frac{\partial Q}{\partial b} = +2\sum_{i=1}^{n}\left[y_i - a - bx_i - cx_i^2\right](-x_i) = 0,$$

or:

$$a\sum_{i=1}^{n}x_i + b\sum_{i=1}^{n}x_i^2 + c\sum_{i=1}^{n}x_i^3 = \sum_{i=1}^{n}x_i y_i \tag{43}$$

$$\frac{\partial Q}{\partial c} = +2\sum_{i=1}^{n}\left[y_i - a - bx_i - cx_i^2\right](-x_i^2) = 0,$$

or:

$$a\sum_{i=1}^{n}x_i^2 + b\sum_{i=1}^{n}x_i^3 + c\sum_{i=1}^{n}x_i^4 = \sum_{i=1}^{n}x_i^2 y_i \tag{44}$$

Equations (42), (43), and (44) are identical to equation (25).

4.5. Data Utilization in Estimating the 4 Models

To generate the quantities needed to estimate the 4 models:

a. The Linear Model
b. The Exponential Model
c. The Power Model,
d. The Quadratic Model,

the given (x, y) bivariate data must be "manipulated" as shown in Tables: 5, 6, 7, and 8, respectively.

4.5.1. Given Data to Evaluate the Linear Model

Table 5. *Manipulation of Given Data to Evaluate the Linear Model*

x	***y***	***xy***	$\mathbf{x^2}$
x_1	y_1	x_1y_1	x_1^2
x_2	y_2	x_2y_2	x_2^2
x_3	y_3	x_3y_3	x_3^2

…	…	…	…
x_n	y_n	$x_n y_n$	x_n^2 x_n^2
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} x_i y_i$	$\sum_{i=1}^{n} x_i^2$
N_1	N_2	N_3	N_4

To evaluate y = a + bx, substitute: N_1, N_2, N_3, N_4 into equations (8) and (9) and solve for a and b simultaneously.

4.5.2. Given Data to Evaluate the Exponential Model

Table 6. *Manipulation of Given Data to Evaluate the Exponential Model*

x	***y***	$\mathbf{x^2}$	***ln y***	***x ln y***
x_1	y_1	x_1^2	$\ln y_1$	$x_1 \cdot \ln y_1$
x_2	y_2	x_2^2	$\ln y_2$	$x_2 \cdot \ln y_2$
x_3	y_3	x_3^2	$\ln y_3$	$x_3 \cdot \ln y_3$
…	…	…	…	…
x_n	y_n	x_n^2	$\ln y_n$	$x_n \cdot \ln y_n$
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} \ln y_i$	$\sum_{i=1}^{n} x_i \ln y_i$
N_5	N_6	N_7	N_8	N_9

To evaluate $y = ke^{cx}$, substitute N_5, N_7, N_8, N_9 into equations (33) and (34) and solve for ln *k* and c simultaneously.

4.5.3. Given Data to Evaluate the Power Model

Table 7. *Manipulation of Given Data to Evaluate the Power Model*

x	***y***	***ln x***	$\mathbf{(\ln x)^2}$	***(ln x)(ln y)***	***ln y***
x_1	y_1	$\ln x_1$	$(\ln x_1)^2$	$(\ln x_1)(\ln y_1)$	$\ln y_1$
x_2	y_2	$\ln x_2$	$(\ln x_2)^2$	$(\ln x_2)(\ln y_2)$	$\ln y_2$
x_3	y_3	$\ln x_3$	$(\ln x_3)^2$	$(\ln x_3)(\ln y_3)$	$\ln y_3$
…	…	…	…	…	…
x_n	y_n	$\ln x_n$	$(\ln x_n)^2$	$(\ln x_n)(\ln y_n)$	$\ln y_n$
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} \ln x_i$	$\sum_{i=1}^{n} (\ln x_i)^2$	$\sum_{i=1}^{n} (\ln x_i)(\ln y_i)$	$\sum_{i=1}^{n} \ln y_i$
N_{10}	N_{11}	N_{12}	N_{13}	N_{14}	N_{15}

To evaluate $\hat{y} = ax^b$, substitute N_{12}, N_{13}, N_{14}, N_{15} into equations (39) and (40) and solve simultaneously for (ln *a*) and b.

4.5.4. Given Data to Evaluate the Quadratic Model

Table 8. *Manipulation of Given Data to Evaluate the Quadratic Model*

x	y	x^2	x^3	xy	x^4	x^2y
x_1	y_1	x_1^2	x_1^3	x_1y_1	x_1^4	$x_1^2y_1$
x_2	y_2	x_2^2	x_2^3	x_2y_2	x_2^4	$x_2^2y_2$
x_3	y_3	x_3^2	x_3^3	x_3y_3	x_3^4	$x_3^2y_3$
…	…	…	…	…	…	…
x_n	y_n	x_n^2	x_n^3	x_ny_n	x_n^4	$x_n^2y_n$
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} x_i^3$	$\sum_{i=1}^{n} x_iy_i$	$\sum_{i=1}^{n} x_i^4$	$\sum_{i=1}^{n} x_i^2y_i$
N_{16}	N_{17}	N_{18}	N_{19}	N_{20}	N_{21}	N_{22}

To evaluate $y = a + bx + cx^2$, substitute N_{16}, N_{17}, N_{18}, N_{19}, N_{20}, N_{21}, N_{22} into equations (42), (43), and (44), and solve simultaneously for a, b, and c.

5. Selecting the Best-Fitting Model

5.1. The Four Models Considered

Given a data set (x_i, y_i), we have shown how to fit to such a data set four different models, namely:

a. Linear:

$$\hat{y}_i = a + bx_i \tag{45}$$

b. Exponential:

$$\hat{y}_i = ke^{cx_i} \tag{46}$$

c. Power:

$$\hat{y}_i = ax_i^b \tag{47}$$

d. Quadratic:

$$\hat{y}_i = a + bx_i + cx_i^2 \tag{48}$$

We might decide to fit all four models to the same data set if, after examining the scatter diagram of the given data set, we are unable to decide which of the "4 models appears to fit the data BEST."

But, after we fit the 4 models, how can we tell which model fits the data best?

To answer this question, we calculate the "variance of the residual values" for each of the models, and then "select as the best model" the one with the smallest variance of the residual values.

5.2. Calculating the Residual Values of Each Model and Their Variance

Use each x_i value, of the given data set (x_i, y_i), to calculate the $\hat{y}_i$ value, from the appropriate model, and then for each i, form the residual:

$$\text{Residual of observations} = (y_i - \hat{y}_i), \tag{49}$$

for each i.

Then the variance of the residual values is defined by:

$$\text{V (Residual)} = \frac{1}{DOF}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{50}$$

where DOF = Degrees of Freedom.

Note: The DOF are DOF = n – 2 for the first three models (Linear, Exponential, Power) due to the fact that each of these 3 models has 2 unknown quantities that need to be evaluated from the data (***a*** and ***b***, ***k*** and ***c***, and ***a*** and ***b***, respectively) and, as a consequence, 2 degrees of freedom are lost. For the Quadratic model, DOF = n – 3 because the model has 3 unknown quantities that need to be estimated and, as a consequence, 3 degrees of freedom are lost.

Using equation (50) to calculate the variance of the residuals for each of the 4 models, we obtain:

$$V(\text{Residual})_{\text{Linear}} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2 \quad (51)$$

$$= \frac{1}{n-2}\left[(y_1 - a - bx_1)^2 + (y_2 - a - bx_2)^2 + ... + (y_n - a - bx_n)^2\right] \quad (52)$$

$$V(\text{Residual})_{\text{Exponential}} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - ke^{cx_i})^2 \quad (53)$$

$$= \frac{1}{n-2}\left[(y_1 - ke^{cx_1})^2 + (y_2 - ke^{cx_2})^2 + \cdots + (y_n - ke^{cx_n})^2\right] \quad (54)$$

$$V(\text{Residual})_{\text{Power}} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - ax_i^b)^2 \quad (55)$$

$$= \frac{1}{n-2}\left[(y_1 - ax_1^b)^2 + (y_2 - ax_2^b)^2 + \cdots + (y_n - ax_n^b)^2\right] \quad (56)$$

$$V(\text{Residual})_{\text{Quadratic}} = \frac{1}{n-3}\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)^2 \quad (57)$$

$$= \frac{1}{n-3}\left[(y_1 - a - bx_1 - cx_1^2)^2 + (y_2 - a - bx_2 - cx_2^2)^2 + ... + (y_n - a - bx_n - cx_n^2)^2\right] \quad (58)$$

After the calculation of the 4 variances from equations: (52), (54), (56), and (58), the model with the "smallest" variance is the model which fits the given data set "best".

We will now illustrate, through an example, how the 4 models we discussed above can be fitted to a given bivariate data set, and then how the "best" model from among them is selected.

5.3. A Considered Example

A sample of 5 adult men for whom heights and weights are measured gives the following results (Table 9).

Table 9. *Sample of 5 adult men*

#	***X = Height***	***Y = Weight***
1	64	130
2	65	145
3	66	150
4	67	165
5	68	170

Problem: Fit the linear, exponential, power, and quadratic models to this bivariate data set and then select as the "best" the model with the smallest variance of the residual values.

5.3.1. Fitting the Linear Model $\hat{y}$ = a + bx

To fit the linear model, we must extend the given bivariate data so that we can also calculate $\sum_{i=1}^{n} x_i^2$ and $\sum_{i=1}^{n} x_i y_i$, as shown below, in Table 10:

Table 10. *Calculations for bivariate data of 5 adults for the linear model*

x^2	x	y	Xy
4096	64	130	8320
4225	65	145	9425
4356	66	150	9900
4489	67	165	11055
4624	68	170	11560

$\sum_{i=1}^{5} x_i^2 = 21{,}790$	$\sum_{i=1}^{5} x_i = 330$	$\sum_{i=1}^{5} y_i = 760$	$\sum_{i=1}^{5} x_i y_i = 50{,}260$

We then substitute the generated data into the normal equations for the linear model, namely equations (8) and (9):

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$
$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \ ,$$

and obtain the equations:

$$\begin{cases} 5a + 330b = 760 \\ 330a + 21{,}790b = 50{,}260 \end{cases}$$

When these equations are solved simultaneously for *a* and *b* we obtain:

$$\begin{cases} a = -508, and \\ b = 10 \end{cases}$$

Therefore, the linear model is:

$$\hat{y} = a + bx = -508 + 10x$$

The variance of the residual values for the linear model is calculated as shown below, in Table 11:

Table 11. *Variance of the residual values for the linear model*

Given X	**Given Y**	**Calculated Y**	**Residual**	**(Residual)^2**
x	y	$\hat{y}$ = -508 + 10x	$y - \hat{y}$	$(y - \hat{y})^2$
64	130	-508 + 10 (64) = 132	-2	$(-2)^2 = 4$
65	145	-508 + 10 (65) = 142	+3	$(+3)^2 = 9$
66	150	-508 + 10 (66) = 152	-2	$(-2)^2 = 4$
67	165	-508 + 10 (67) = 162	+3	$(+3)^2 = 9$
68	170	-508 + 10 (68) = 172	-2	$(-2)^2 = 4$
				$\sum_{i=1}^{5}(y_i - \hat{y}_i)^2 = 30$

Therefore, the variance of the residual values, for the linear model is:

$$V(\text{Residual})_{\text{Linear}} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{5-2}(30) = \frac{30}{3} = 10$$

5.3.2. Fitting the Exponential Model $\hat{y} = ke^{cx}$

To fit the exponential model we need to extend the given bivariate data so that we can calculate, in addition to $\sum_{i=1}^{5} x_i = 330$ and $\sum_{i=1}^{5} x_i^2 = 21{,}790$, $\sum_{i=1}^{5} \ln y_i$ and $\sum_{i=1}^{5}(x_i \ln y_i)$ as shown below, in Table 12:

Table 12. *Calculations for bivariate data of 5 adults for the exponential model*

x^2	x	y	$\ln y$	$x \cdot \ln y$
4096	64	130	4.8675	311.5200
4225	65	145	4.9767	323.4855

4356	66	150	5.011	330.726
4489	67	165	5.1059	342.0953
4624	68	170	5.1358	349.2344
$\sum_{i=1}^{5} x_i^2 = 21{,}790$	$\sum_{i=1}^{5} x_i = 330$	$\sum_{i=1}^{5} y_i = 760$	$\sum_{i=1}^{5} \ln y_i = 25.0966474$	$\sum_{i=1}^{5} x_i \cdot \ln y_i = 1657.04468$

We then substitute the generated data into the normal equations for the exponential model (i.e. equations (33) and (34)):

$$n(\ln k) + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \ln y_i$$

$$(\ln k)\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i(\ln y_i) \ ,$$

and obtain the equations:

$$\begin{cases} 5\ln k + 330c = 25.0967 \\ 30\ln k + 21{,}790c = 1657.0447 \end{cases}$$

When these equations are solved simultaneously for ln k and c, we obtain: c = 0.06658 and lnk = 0.6251, or: $k = e^{0.6251} = 1.868432$

Therefore, the exponential model is:

$$\hat{y} = ke^{cx} = 1.868432e^{0.06658x}$$

or

$$\ln y = \ln k + cx = 0.6251 + 0.06658x$$

Then, the variance of the residual values, for the exponential model, is calculated as shown below, in Table 13:

Table 13. *Variance of the residual values for the exponential model*

x	***y***	$\hat{y} = ke^{cx} = 1.868432\ e^{0.06658x}$	$y - \hat{y}$	$(y - \hat{y})^2$
64	130	$1.868432\ e^{0.06658(64)} = 132.4515$	-2.4515	6.0099
65	145	$1.868432\ e^{0.06658(65)} = 141.5703$	3.4297	11.7628
66	150	$1.868432\ e^{0.06658(66)} = 151.3169$	-1.3169	1.7324
67	165	$1.868432\ e^{0.06658(67)} = 161.7346$	3.2654	10.6630
68	170	$1.868432\ e^{0.06658(68)} = 172.8694$	-2.8694	8.2336
				$\sum_{i=1}^{5} (y_i - \hat{y}_i)^2 = 38.4035$

Therefore, the variance of the residual values, for the exponential model is:

$$V(\text{Residual})_{\text{Exponential}} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - ke^{cx_i})^2$$

$$= \frac{1}{5-1}\sum_{i=1}^{5}\left[y_i - 1.868432e^{0.06658x_i}\right]^2$$

$$= \frac{38.4035}{3} = 12.8017$$

5.3.3. Fitting the Power Model, $\hat{y} = ax^b$

To fit the power model we need to extend the given bivariate data set to generate the quantities: $\sum_{i=1}^{n} \ln x_i$, $\sum_{i=1}^{n} (\ln x_i)^2$, $\sum_{i=1}^{n} \ln y_i$ and $\sum_{i=1}^{n} (\ln x_i)(\ln y_i)$, and this is accomplished as shown below, in Table 14:

Table 14. *Calculations for bivariate data of 5 adults for the power model*

x	***y***	***ln x***	***(ln x)²***	***ln y***	***(ln x)(ln y)***
64	130	4.158883	17.2963085	4.867553	20.2435
65	145	4.1738727	17.42550908	4.976734	20.7723
66	150	4.189654742	17.55320686	5.010635	20.9928
67	165	4.204692619	17.67944002	5.105945	21.4689
68	170	4.219507705	17.80424527	5.135798	21.6705
		$\sum_{i=1}^{5} \ln x_i$ =20.9471	$\sum_{i=1}^{5} (\ln x_i)^2$ =87.7581	$\sum_{i=1}^{5} \ln y_i$ =25.0967	$\sum_{i=1}^{5} (\ln x_i)(\ln y_i)$ =105.1505

We then substitute the generated data into the normal equations of the power model, namely equations (39) and (40):

$$n(\ln a) + b\sum_{i=1}^{n} \ln x_i = \sum_{i=1}^{n} \ln y_i$$

$$(\ln a)\sum_{i=1}^{n} \ln x_i + b\sum_{i=1}^{n} (\ln x_i)^2 = \sum_{i=1}^{n} (\ln x_i)(\ln y_i)$$

and obtain the equations:

$$\begin{cases} 5\ln a + 20.9471b = 25.0967 \\ 20.9471\ln a + 87.7581b = 105.1505 \end{cases}$$

When these equations are solved simultaneously for b and ln *a* we obtain:

$$\begin{cases} b = 4.3766, and \\ \ln a = -13.316 \end{cases}$$

Therefore, the "linearized" power model becomes:

$$\ln \hat{y} = \ln a + b \ln x = -13.316 + 4.3766x$$

Then the variance of the residual values for the power model is obtained as shown below:

Table 15. *Variance of the residual values for the power model*

x	***y***	***ln x***	$\ln \hat{y} = \ln a + b \ln x$ = -13.316 + 4.3766x	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
64	130	4.158883	$\ln \hat{y}_1$ = 4.885768	132.3920	-2.3920	5.721664
65	145	4.173873	$\ln \hat{y}_2$ = 4.95623	141.6874	3.3126	10.973319
66	150	4.189655	$\ln \hat{y}_3$ = 5.020443	151.4784	-1.47843	2.185667
67	165	4.204693	$\ln \hat{y}_4$ = 5.086258	161.7833	3.2167	10.347159
68	170	4.219508	$\ln \hat{y}_5$ = 5.151097	172.6208	-2.6208	6.868592
						$\sum_{i=1}^{5} (y_i - \hat{y}_i)^2$ = 36.09640

Therefore, the variance of the residuals values for the power model is:

$$V(\text{Residual})_{\text{Power}} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - ax_i^b)^2 = \frac{36.0964}{3} = 12.0321$$

5.3.4. Fitting the Quadratic Model, $\hat{y} = a + bx + cx^2$

To fit the quadratic model, we need to use the given bivariate data set and extend it to generate the quantities:

$$\sum_{i=1}^{n} X_i = 330; \sum_{i=1}^{n} X_i^2 = 21,790; \sum_{i=1}^{n} X_i^3 = 1,439,460; \sum_{i=1}^{n} X_i^4 = 95,135,074;$$
$$\sum_{i=1}^{n} Y_i = 760; \sum_{i=1}^{n} X_i Y_i = 50,260; \sum_{i=1}^{n} X_i^2 Y_i = 3,325,270$$

We then substitute the generated data into the normal equations of the quadratic model (see equation (25)), and obtain:

$$\begin{cases} 5a + 330b + 21,790c = 760 \\ 330a + 21,790b + 1,439,460c = 50,260 \\ 21,790a + 1,439,460b + 95,135,074c = 3,325,270 \end{cases}$$

Solving these 3 equations simultaneously, we obtain a = -25,236/7, b = 730/7, c = -5/7. Therefore, the quadratic function $\hat{y} = f(x)$ is given by:

$$\hat{y} = a + bx + cx2 = \frac{1}{7}[-23,326 + 730x - 5x^2]$$

The variance of the residual values for the quadratic model is calculated as shown below, in Table 16:

Table 16. *Variance of the residual values for the quadratic model*

x	y	$\hat{y} = \frac{1}{7}[-25,326 + 730x - 5x^2]$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
64	130	$\hat{y}_1$ = 130.5714286	-0.5714286	0.326530644
65	145	$\hat{y}_2$ = 142.7142857	2.2857143	5.224489861
66	150	$\hat{y}_3$ = 153.4285714	-3.4285714	11.75510184
67	165	$\hat{y}_4$ = 162.7142857	2.2857143	5.224489861
68	170	$\hat{y}_5$ = 170.5714286	-0.5714286	0.326530644
				$\sum_{i=1}^{5}(y_i - \hat{y}_i)^2$ = 22.85714286

Therefore, the variance of the residual values for the quadratic model is:

$$V(\text{Residual})_{\text{Quadratic}} = \frac{1}{n-3}\sum_{i=1}^{5}(y_i - \hat{y}_i)^2 = \frac{22.85714286}{2} = 11.42857143 \approx 11.4286$$

5.3.5. Summary of Results and Selection of the "Best" Model

We have fitted the 4 models: linear, exponential, power, and quadratic models, calculated the respective residual variances, and have obtained the following results:

a) The linear model is:

$$\hat{y} = a + bx = -508 + 10x$$
with $V(\text{Residual})_{\text{Linear}} = 10$

b) The exponential model is:

$$\hat{y} = ke^{cx} = 1.868432e^{0.06658x}$$
with $V(\text{Residual})_{\text{Exponential}} = 12.8017$

c) The power model is:

$$\ln \hat{y} = \ln a + b \ln x = -13.316 + 4.3766 \ln x$$
with $V(\text{Residual})_{\text{Power}} = 12.0321$

d) The quadratic model is

$$\hat{y} = a + bx + cx^2 = \frac{1}{7}\left[-25{,}326 + 730x - 5x^2\right]$$
with $V(\text{Residual})_{\text{Quadratic}} = 11.4286$

Since the linear model has the smallest variance of the residual values of the 4 models fitted to the same bivariate data set, the linear model is the "best" model (but the other 3 values are very close). The linear model, therefore, will be selected as the "best" model and used for forecasting purposes.

6. MINITAB Solutions

To obtain the MINITAB solutions of the four models we discussed in this paper we do the following:

6.1. Finding the MINITAB Solution for the Linear Model

The data set used to find the MINITAB solution for the linear model is presented in Table 17.

Table 17. *Data set in MINITAB for the linear model*

MTB > Set C1
DATA> 64 65 66 67 68
DATA> end
MTB > set C2
DATA> 130 145 150 165 170
DATA> end
MTB > Name C1 'X' C2 'Y'
MTB > REGRESS 'Y' 1 'X'

The results of the regression analysis for the linear model is presented in Table 18.

Table 18. *Regression analysis: Y versus X for the linear model*

Regression equation:			Y = - 508 + 10.0 X		
Predictor	**Coef**	**SE Coef**	**T**	**p**	
Constant	-508.000	66.020	-7.700	0.005	
X	10.000	1.000	10.000	0.002	
Regression fit:	**S**	**R-Sq**	**R-Sq (adj)**		
	3.162	97.1%	96.1%		

Analysis of Variance:					
Source	**DF**	**SS**	**MS**	**F**	**p**
Regression	1	1000.0	1000.0	100.0	0.002
Residual Error	3	30.0	10.0		
Total	4	1030.0			

6.2. Finding the MINITAB Solution for the Exponential Model

The data set used to find the MINITAB solution for the exponential model is presented in Table 19.

Table 19. *Data set in MINITAB for the exponential model*

MTB > Set C1
DATA> 64 65 66 67 68
DATA> end
MTB > set C2
DATA> 130 145 150 165 170
DATA> end
MTB > Name C1 'X' C2 'Y'
MTB > REGRESS 'Y' 1 'X'

The results of the regression analysis for the exponential model is presented in Table 20.

Table 20. *Regression analysis: Y versus X for the exponential model*

Regression equation:			Y = 0.625 + 0.0666 X		
Predictor	**Coef**	**SE Coef**	**T**	**p**	
Constant	0.6251	0.4925	1.27	0.294	
X	0.066580	0.007460	8.92	0.003	
Regression fit:	**S**	**R-Sq**	**R-Sq (adj)**		
	0.0235917	96.4%	95.2%		
Analysis of Variance:					
Source	**DF**	**SS**	**MS**	**F**	**p**
Regression	1	0.044329	0.044329	79.65	0.003
Residual Error	3	0.001670	0.000557		
Total	4	0.045999			

6.3. Finding the MINITAB Solution for the Power Model

The data set used to find the MINITAB solution for the power model is presented in Table 21.

Table 21. *Data set in MINITAB for the power model*

MTB > Set C1
DATA> 4.158883; 4.1738727; 4.189654742; 4.204692619; 4.2195077
DATA> end
MTB > set C2
DATA> 4.867553; 4.976734; 5.010635; 5.105945; 5.135798
DATA> end
MTB > Name C1 'X' C2 'Y'
MTB > REGRESS 'Y' 1 'X'

The results of the regression analysis for the power model is presented in Table 22.

Table 22. *Regression analysis: Y versus X for the power model*

Regression equation:			Y = - 13.3 + 4.38X		
Predictor	**Coef**	**SE Coef**	**T**	**p**	
Constant	-13.316	2.069	-6.44	0.008	
X	4.3766	0.4939	8.86	0.003	
Regression fit:	**S**	**R-Sq**	**R-Sq (adj)**		
	0.0237507	96.3%	95.1%		
Analysis of Variance:					
Source	**DF**	**SS**	**MS**	**F**	**p**
Regression	1	0.044301	0.044301	78.53	0.003
Residual Error	3	0.001692	0.000564		
Total	4	0.045993			

6.4. Finding the MINITAB Solution for the Wuadratic Model

The data set used to find the MINITAB solution for the quadratic model is presented in Table 23.

Table 23. *Data set in MINITAB for the quadratic model*

MTB > Set C1
DATA> 64 65 66 67 68
DATA> end
MTB > set C2
DATA> 4096 4225 4356 4489 4624
DATA> end
MTB > SET C3
DATA> 130 145 150 165 170
DATA> END
MTB > NAME C1 'X1' C2 'X2' C3 'Y'
MTB > REGRESS 'Y' 2 'X1' 'X2'

The results of the regression analysis for the quadratic model is presented in Table 24.

Table 24. *Regression analysis: Y versus X1, X2 for the quadratic model*

Regression equation:			Y = - 3618 + 104 X1 - 0.714 X2		
Predictor	**Coef**	**SE Coef**	**T**	**p**	
Constant	-3618	3935	-0.92	0.455	
X1	104.3	119.3	0.87	0.474	
X2	-0.7143	0.9035	-0.79	0.512	
Regression fit:	**S**	**R-Sq**	**R-Sq (adj)**		
	3.38062	97.8%	95.6%		
Analysis of Variance:					
Source	**DF**	**SS**	**MS**	**F**	**p**
Regression	2	1007.14	503.57	44.06	0.022
Residual Error	2	22.86	11.43		
Total	4	1030.00			
Source	**DF**	**Seg SS**			
X1	1	1000.00			
X2	1	7.14			

7. Conclusions

Reviewing our previous discussion we come to the following conclusions:

The Linear Regression problem is relatively easy to solve and can be handled using algebraic methods.

The problem can also be solved easily using available statistical software, like MINITAB.

Even though the solution to Regression problems can be obtained easily using MINITAB (or other statistical software) it is important to know what the hand methodology is and how it solves these problems before you can properly interpret and understand MINITAB's output.

In general, non-linear regression is much more difficult to perform than linear regression.

There are, however, some simple non-linear models that can be evaluated relatively easily by utilizing the results of linear regression.

The non-linear models analyzed in this paper are: Exponential Model, Power Model, Quadratic Model.

A procedure is also discussed which allows us to fit to the same bivariate data set many models (such as: linear, exponential, power, quadratic) and select as the "best fitting" model the model with the "smallest variance of the residuals".

In a numerical example, in which all 4 of these models were fitted to the same bivariate data set, we found that the Linear model was the "best fit", with the Quadratic model "second best". The Power and Exponential models are "third best" and "fourth best" respectively, but are very close to each other.

The evaluation of these models is facilitated considerably by using the statistical software package MINITAB which, in addition to estimating the unknown parameters of the corresponding models, also generates additional information (such as the p-value, standard deviations of the parameter estimators, and R^2).

This additional information allows us to perform hypothesis testing and construct confidence intervals on the parameters, and also to get a measure of the "goodness" of the equation, by using the value of R^2. A value of R^2 close to 1 is an indication of a good fit.

The MINITAB solution for the linear model shows that both a and b (of $\hat{y} = a + bx = -508 + 10x$) are significant because the corresponding p-values are smaller than $\alpha = 0.05$, while the value of $R^2 = 97.1\%$, indicating that the regression equation explains 97.1% of the variation in the y-values and only 2.9% is due to other factors.

The MINITAB solution for the quadratic model shows that a, b, and c (of $\hat{y} = a + bx + cx^2 = -3{,}618 + 104.3x + 0.7143x^2$) are individually not significant (because of the corresponding high p-values, but b and c jointly are significant because of the corresponding p-value of $p = 0.022 < \alpha = 0.05$. The value of R^2 is: $R^2 = 97.8\%$.

The MINITAB solution for the power model shows that both a and b (of $\hat{y} = ax^b$ or $\ln y = \ln a + b \ln x = -13.3 + 4.3766 \ln x$) are significant because the corresponding p-values are smaller than $\alpha = 0.05$, while the value of $R^2 = 96.3\%$.

The MINITAB solution for the exponential model shows that the k (in $\hat{y} = ke^{cx} = 1.868432e^{0.06658x}$ or $\ln \hat{y} = \ln k + cx = 0.6251 + 0.06658x$) is not significant because of the corresponding high p-value, while the c is significant because of the corresponding p-value being smaller than $\alpha = 0.05$. The value of $R^2 = 96.4\%$.

References

Adamowski, J., H., Fung Chan, S.O., Prasher, B., Ozga-Zielinski, and Sliusarieva. A., 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting. *Water Resources*, 48, W01528, Montreal: Canada

Berenson, M.L., Levine, D.M. and Krehbiel, T.C., 2004. *Basic Business Statistics* (9th Edition). Upper Saddle River, N.J.: Prentice-Hall

Bhatia, N., 2009. Linear Regression: An Approach for Forecasting

Black, K., 2004. *Business statistics* (4th Edition). Hoboken, NJ: Wiley

Canavos, G.C., 1984. *Applied Probability and Statistical Methods*. Boston: Little Brown

Carlson, W.L. and Thorne, B., 1997. Applied Statistical Methods. Upper Saddle River, N.J.: Prentice-Hall

Chen, Kuan-Yu, 2011. Combining linear and nonlinear model in forecasting tourism demand. *Expert Systems with Applications*, 38(8), pp.10368–10376

Childress, R.L., Gorsky, R.D. and Witt, R.M., 1989. *Mathematics for Managerial Decisions*. Upper Saddle River, N.J.: Prentice-Hall

Chou, Ya-lun, 1992. *Statistical Analysis for Business and Economics*. New York: Elsevier

Freud, J.E. and Williams, F.J., 1982. *Elementary Business Statistics: The Modern Approach*. Upper Saddle River, N.J.: Prentice-Hall

McClave, J.T., Benson, G.P. and Sincich, T., 2001. *Statistics for Business and Economics* (8th Edition). Upper Saddle River, N.J.: Prentice-Hall

Pindyck, R. and Rubinfeld, D.L., 1981. *Econometric Models and Economic Forecasts* (2nd Edition). New York: McGraw-Hill

Vasilopoulos, A. and Lu, F.V., 2006. Quantitative Methods for Business with Computer Applications. Boston, MA: Pearson Custom Publishing

Vasilopoulos, A., 2005. Regression Analysis Revisited. *Review of Business*, 26 (3), pp.36-46

Vasilopoulos, A., 2007. Business Statistics – A Logical Approach. Theory, Models, Procedures, and Applications Including Computer (MINITAB) Solutions. Boston, MA: Pearson Custom Publishing