

Notes on logarithms

Ron Michener

Revised January 2003

In applied work in economics, it is often the case that statistical work is done using the logarithms of variables, rather than the raw variables themselves. This seems mysterious when one first encounters it, but there are good reasons for it.

Reviewing some basics.

Since many of you have not seen logs in many years, let me review a few basic facts about them. If you have an expression such as

$$x = a^y$$

y is said to be the “base a logarithm of x .”

Practically speaking, there are only two bases that are ever used – base 10 (which is the one most commonly taught in high school) and base e . ($e \approx 2.718$ is a natural constant. Like π it is an infinite decimal.) A base e log is also known as a “natural log,” and it is the most commonly used log in advanced applications (including economics).

Here are some examples:

$$100 = 10^2$$

so that two is the base 10 log of 100.

$$4 = e^{1.386}$$

so that 1.386 is the natural log of 4. Since natural logs are most commonly used in economics, I will use them exclusively. When I say *log*, I will mean the natural log.

One useful property of logarithms is that they simplify certain arithmetic calculations. This used to be very important before the days of pocket calculators. For example, consider the product of two numbers, x_1 and x_2 .

$$z = x_1 x_2 = e^{y_1} e^{y_2} = e^{(y_1 + y_2)}$$

As you can see, y_1 is the log of x_1 , and y_2 is the log of x_2 . The log of the product, z , is the sum $y_1 + y_2$. *Multiplication* in the original numbers becomes *addition* in the logs, an easier operation. Similarly, if one raises x to the power p

$$z = x^p = (e^{y_1})^p = e^{py_1}$$

the log of the result is just p times the log of x . *Exponentiation* in the original numbers becomes *multiplication* in logs.

Why do natural logs appear in economics?

Natural logs appear in economics for several reasons.

First: If you have a variable – be it sales, population, prices, or whatever – that grows at a constant percentage rate, the log of that variable will grow as a linear function of time. The demonstration requires some calculus. Constant percentage growth means that

$$\frac{dy}{dt} = a y$$

The numerator on the left-hand side is the rate of growth of the variable y ; by dividing by y the left-hand side becomes the percentage rate of growth, which is equal to a constant, a . To solve this equation, we must rearrange and integrate.

$$\begin{aligned}\frac{dy}{y} &= a dt \\ \int \frac{dy}{y} &= \int a dt \\ \log(y) &= c + at\end{aligned}$$

where the constant c is a constant of integration. What happens to y itself (as opposed to $\log(y)$) can be found by exponentiating, which undoes the log function.

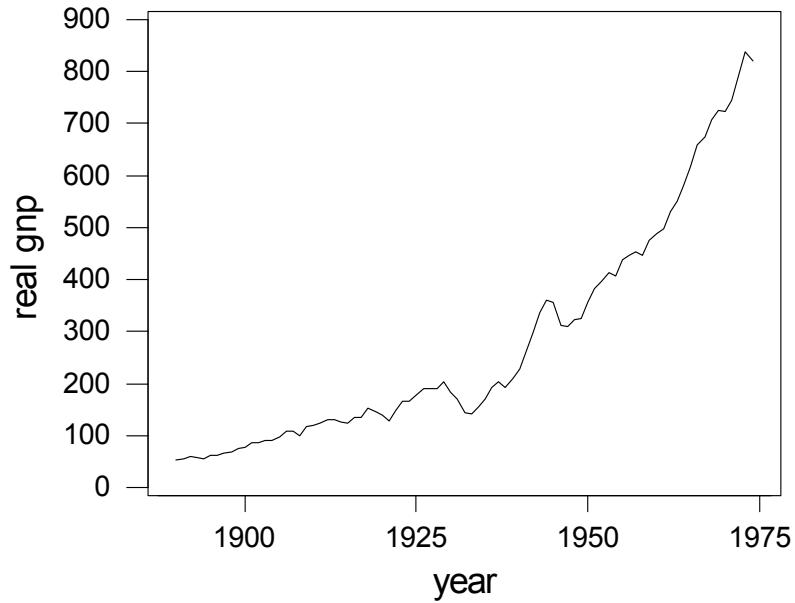
$$\begin{aligned}\log(y) &= c + at \\ e^{\log(y)} &= e^{c+at} \\ y &= e^c e^{at} \equiv c^* e^{at}\end{aligned}$$

To summarize what this derivation shows: If a variable grows at a constant percentage rate, the log of that variable will be a linear function of time, and the coefficient of time will be the percentage growth rate. The variable itself will exhibit exponential growth.

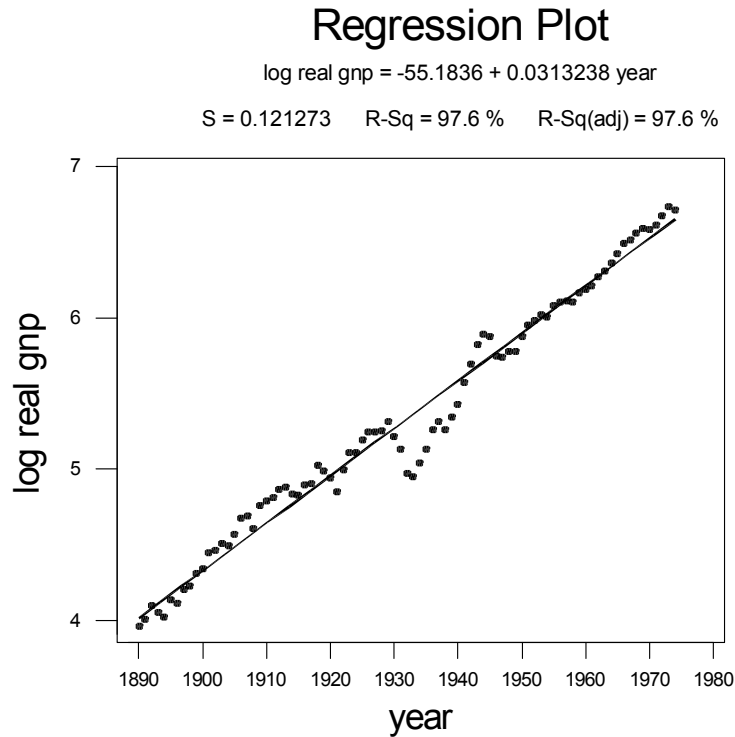
An Important Example: US GNP

Here is a plot of US real gross national product against time, for 1896 – 1974.

Real US Gross National Product by Year
1896-1974



If you examine the plot, the relationship is obviously nonlinear. However, it is sensible to think that growth in gross national product might occur at an approximately constant *percentage* rate. Therefore, we consider the plot of *the log* of real GNP against time.



Not only is the relationship very nearly linear, but also the deviations from the fitted line are meaningful. The dramatic dip in the middle of the graph corresponds to the Great Depression, and the peak that immediately follows is World War II. As you can see in the fitted line plot, the estimated coefficient of year is .0313. This means the estimated annual percentage growth rate of GNP is 3.13%.¹

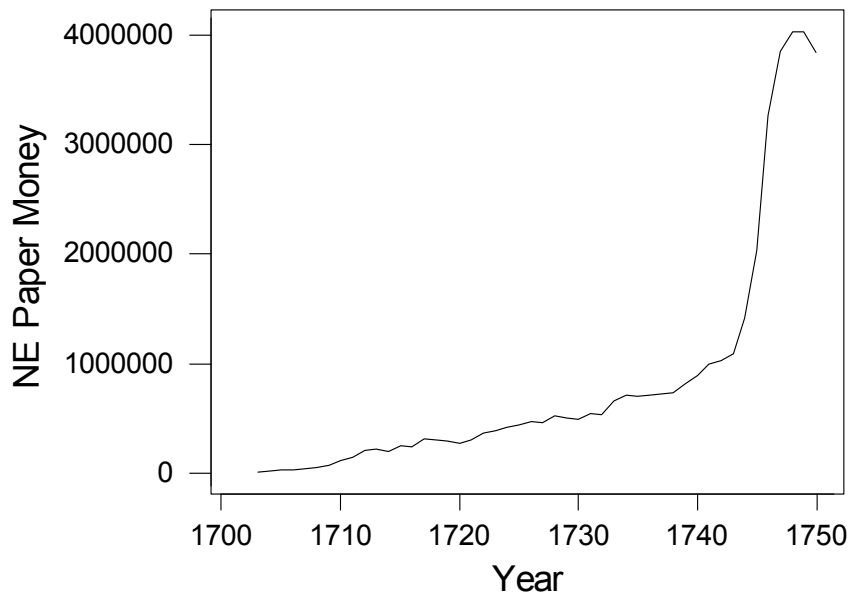
Many economic time series exhibit constant or near constant percentage growth, as this GNP series does. This is true of economic time series of interest to your future employers, as well as those of interest to academics. If you were forecasting future sales of *Tide* detergent for Procter and Gamble you might find the logarithmic transformation useful.

¹ If you examine the scatter of points around the line, you can see that the residuals from the regression line are not independent. Good years are almost invariably followed by good years, and bad years by bad. Correcting for this lack of independence would permit us to refine our estimate of the growth rate, but is a subject beyond the scope of this course.

Another example from economic history

During the early 18th century, the New England colonies conducted one of the first modern trials of paper money.² The experiment ended badly. During the late 1740s, the New England colonies financed major military expeditions against the French in Canada with freshly printed paper money, and the result was what was known at the time as “The Great Inflation.” By modern standards, the inflation wasn’t all that terrible, being at worst only about 35% annually. However, it led to paper money being phased out in New England after 1750. This sketch of New England’s monetary history is a prelude to showing you some data on the quantity of paper money circulating in New England over the period.

New England Paper Money in Circulation 1703 - 1750

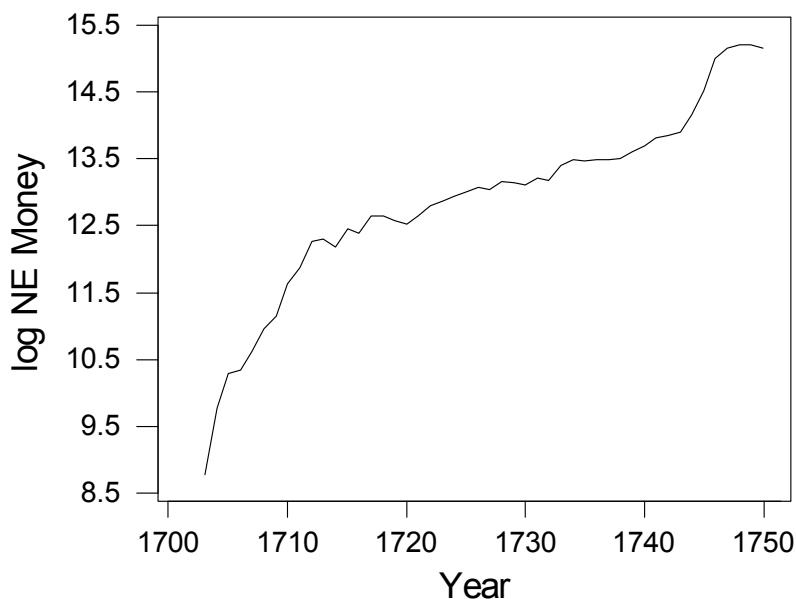


The explosion in the money supply during King George’s War, the period of the Great Inflation, is readily apparent.

What use are logs to this example? Well, look at the corresponding plot of the logs.

² The earlier trials were in China in the middle ages. The colony of Massachusetts Bay is often credited with being the first western experiment with government-issued paper money.

Log of New England Paper Money in Circulation 1703 - 1750



This plot looks rather different. You can see that there are crudely three periods to the history. From 1703 to about 1712, the money supply grew very rapidly (remember, the slope of the line is the percentage growth rate). Then from about 1713 to the mid-1740s the money supply grew slowly at a roughly constant rate. In the late 1740s there was another short burst of rapid money growth. The first episode of rapid money growth was completely hidden by the scaling of the original diagram. There is a straightforward explanation for what was going on. The period from 1703 to 1711 was Queen Anne's War, another of the colonial conflicts involving England and France. The period of slower growth that followed was a period of peace (except for a relatively minor expedition against the Spanish in 1740). The last growth spurt was King George's war.

Second: Many theoretical relationships in economics are nonlinear in the levels of variables, but linear in logs.

An Example: the Cobb-Douglas Production function

In your microeconomics classes you have probably been introduced to the Cobb-Douglas production function. The Cobb-Douglas function is one of the simplest functional forms relating capital and labor inputs to output that produces well-behaved isoquants. The Cobb-Douglas production function is given by $Q = AK^\alpha L^\beta$, where K is the capital input, L is the labor input, A , α and β are constants, and Q is the quantity of output produced. (If $\alpha + \beta = 1$, there are constant returns to scale. In some presentations this restriction is imposed and the production function is written as $Q = AK^\alpha L^{1-\alpha}$.) Suppose

you were given data on labor inputs, capital inputs, and output. How could you estimate α and β ? This would appear to be a difficult question, but all that is involved is a log transformation.

$$Q = AK^\alpha L^\beta$$
$$\log(Q) = \log(A) + \alpha \log(K) + \beta \log(L)$$

If you transform the data by taking logs of output, capital and labor, and then perform a multiple regression with the log of output as the dependent variable and the log of capital and the log of labor input as the independent variables, the coefficient of log (K) will be your estimate of α and the coefficient of log (L) will be your estimate of β .

An Example: the Baumol Money Demand Function

This is one of the most famous theoretical models explaining why the transactions demand for money can depend on the interest rate. In the current (updated 6th) edition of Mishkin's *Economics of Money, Banking, and Financial Markets*, this model is discussed on pages 547-50. Mishkin avoids discussing the formula I am about to present, but other, more rigorous money and banking textbooks, such as Meir Kohn's *Money, Banking, and Financial Markets* (2nd Edition, pp. 604-5), do present the formula.

My interest here is in the formula and the log transformation, not monetary theory. But if I simply write the formula down without explanation, it will look very mysterious, so I am going to include the derivation.

The idea of the model is this. Imagine a world in which people receive income, Y , periodically. A concrete example is someone who receives a paycheck on the first of each month. The person could make a single trip to the bank, put the money in a checking account (assumed to be non-interest bearing) and then spend the money at a constant rate (another assumption) throughout the period, until it was gone just as the next paycheck arrived. In this case, the individual's money holding would vary over time smoothly between Y on payday to zero just before the next payday, with average money balances held of $Y/2$. However, there is another strategy. One could take half one's pay, invest it in bonds paying an interest rate i , put the other half in the bank. At the middle of the period (that is, on the 15th of the month) when your bank account is exhausted, you liquidate the bond, put your money back in the bank, and live off that until payday. But there are lots of strategies like this. One could take 2/3rds of one's pay, invest it in interest-bearing bonds, live off the remaining third until the 10th of the month, then sell half your bonds, live until the 20th, then sell the remaining bonds to make it to payday. The cost of following such a strategy is the transactions cost of making all those trips to the bank, assumed to be $\$b$ per trip. The cost of foregoing such a strategy is the foregone interest. There is a trade-off. If your strategy involves making n trips to the bank each period, your transactions costs are bn . If you make n trips to the bank each period, your average money balances held are $Y/2n$, and the foregone interest on these money balances is $Yi/2n$. Therefore your total cost is

$$\text{Total cost} = bn + \frac{Yi}{2n}.$$

By using calculus, you can find the value of n (the number of trips to the bank) that minimizes this total cost. That value is $n^* = \sqrt{\frac{Yi}{2b}}$ and since average money holdings are $\frac{Y}{2n}$, we simply substitute for the optimal number of trips to the bank to derive money demand.³

$$\text{Money Demand} = \sqrt{\frac{Yb}{2i}}.$$

This is a good example of a well-known economic model that gives rise to a functional form that is non-linear. Note, however, what happens in this example if you take logs.

$$\log(M^d) = -.5 \log(2) + .5 \log(b) + .5 \log(Y) - .5 \log(i)$$

The relationship is linear in logs, and in addition, the theory predicts what the coefficients of income and the interest rate ought to be – oft-tested hypotheses in monetary theory.

Example: Demand Functions derived from a Cobb-Douglas Utility Function

You may have encountered the Cobb-Douglas function in microeconomics not as a production function, but as a utility function. It is a nice utility function to use as a class room example because it is perhaps the simplest functional form with well-behaved indifference curves. If you saw the utility function in intermediate microeconomics, it was in the theory of consumer choice. A utility maximizing consumer might be thought to maximize a Cobb-Douglas utility function that depends on consumption of two goods, x and y . This maximization is subject to a budget constraint: $I = p_x x + p_y y$, where p_x is the price of x , p_y is the price of y , and I is (nominal) income. Formally:

$$\begin{aligned} \max_{\{x,y\}} x^\alpha y^\beta \\ \text{s.t. } I = p_x x + p_y y \end{aligned}$$

The result of this exercise is that you derive demand functions for x and y . In particular, the demand function for x that results is:

$$x^D = \frac{\alpha I}{(\alpha + \beta) p_x}.$$

Suppose you were interested in using data on sales of x , income, and prices to estimate this demand function. One difficulty is that it is non-linear. However, voila! Look at the log transform.

$$\log(x^D) = \log\left(\frac{\alpha}{\alpha + \beta}\right) + \log(I) - \log(p_x).$$

³ In differentiating, we are treating n as if it varies continuously, although in fact, n must be an integer. This is one criticism of the model. However, as I said, my interest here is in the log transformation, not monetary theory.

The functional form even predicts the coefficients in front of the variables $\log(I)$ and $\log(p_x)$ --- they ought to be +1 and -1 respectively!

Third – Logs are used in economics because the estimated coefficients in log regressions have a nice interpretation.

Consider the following functional form:

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) \dots + \beta_p \log(x_p).$$

What is the interpretation of β_i ?

Well, mathematically, β_i is:

$$\frac{\frac{\partial y}{\partial x_i}}{y} = \frac{\beta_i}{x_i}$$

rearranging

$$\beta_i = \frac{y}{x_i} \frac{\partial y}{\partial x_i}$$

This should look familiar to anyone who mastered introductory price theory! The coefficient is a measure of the *percentage change* in y that occurs as a result of a certain *percentage change* in x . It is an *elasticity* (holding the other x 's constant, as usual).

Therefore, if you look at the Cobb-Douglas utility function example again, you will see that with Cobb-Douglas preferences the income elasticity of demand is one, and the price elasticity of demand is either one or minus one, depending on the convention you use for defining price elasticity. {Assuming demand curves slope downwards, the number one gets here will be negative. However, price elasticity is defined by most textbook writers to be the *absolute value* of the percentage change in demand associated with a percentage change in price.} The Cobb-Douglas utility function is not used much in advanced statistical work, because it is unnecessarily restrictive to impose these unitary elasticities.

If you look at the Baumol money demand function, you will see it predicts the income elasticity of money demand will be $\frac{1}{2}$ and the interest elasticity of money demand will also be $\frac{1}{2}$ (or minus $\frac{1}{2}$, depending again on the convention you use.)

Convenience, then, is one important reason economists use the log transformation so frequently. Economists often think in terms of elasticities, and log regressions have coefficients that estimate elasticities.

Are there other considerations?

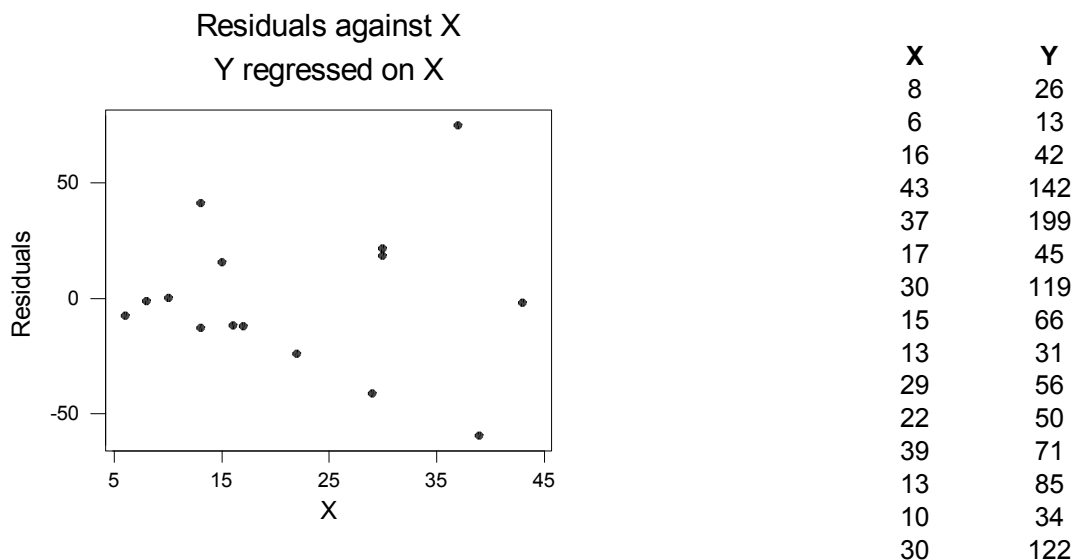
I can think of two.

First, logs aren't even defined for non-positive numbers. Therefore, regressions can only be done in logs when the variables you wish to log are strictly positive. Sometimes this is useful. Suppose you are fitting a regression to predict a variable (such as sales or GNP) that simply can't be negative. If you fit the regression using the level of the variable as the dependent variable, there are almost always possible values of the independent variables that would make the predicted value of y negative! One way to guarantee that predicted values of y are never negative is to use $\log(y)$ as your dependent variable.

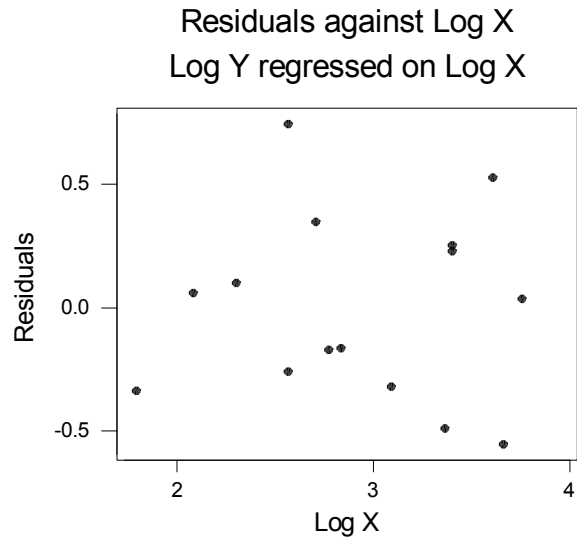
Whatever the predicted value of $\log(y)$, the corresponding value of y will always be positive.

Second, statisticians sometimes use logs as a *variance-stabilizing transformation*. If you are using a statistical technique, such as ordinary least squares regression, where constant error variances are important, it sometimes happens in practice that a log transform will result in estimated errors with a constant or near constant variance, while the same calculation done in levels produces estimated errors whose variance is obviously not constant. This most often occurs the standard deviations of the error are roughly constant in percentage terms, but not in absolute terms. A statistician might then be inclined to use a log transformation so that the statistical assumptions of the procedure are more nearly met.

An Example:



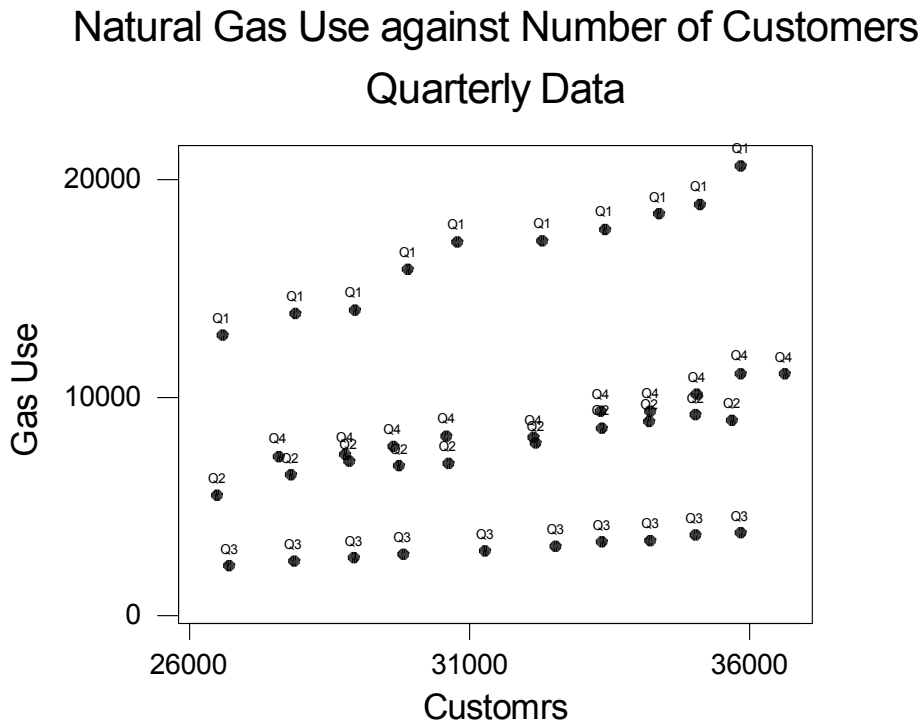
Consider the data set to the right. If you regress Y on X and plot the residuals, you see a clear pattern suggesting that the variance of the error increases with increasing values of x .



However, if you take *the same data*, and regress $\log(y)$ on $\log(x)$, the resulting plot of the residuals against $\log(x)$ is consistent with a constant error variance.

A Detailed Empirical example of the use of logs in Regression

The example I am going to use is quarterly data from the 1960s on the consumption (in the United States) of natural gas. There is data on the quantity of natural gas consumed, the number of natural gas customers, and the quarter of the year. Natural gas is commonly used for heating, so that the consumption of natural gas is highly seasonal. If one plots gas use against the number of customers, one gets the following plot.



You can see from the data labels that the observations at the top are all from the first quarter (winter), those along the bottom are all from the third quarter (summer), while those in the middle are second and fourth quarter observations (spring and fall). If there is a problem with using dummy variables to model this data, it is that the line through the first quarter observations obviously would not be parallel to a line through the third quarter observations – the absolute difference between the first and third quarters seems to be growing as the number of customers grows. If we proceeded naively, however, and fit a dummy variable model to explain Gas Use as a function of the number of Customers and quarter of the year, we would get the following:

Regression Analysis: Gas Use versus Customrs, Q1, Q2, Q3

The regression equation is

$$\text{Gas Use} = -5211 + 0.439 \text{ Customrs} + 8050 \text{ Q1} - 928 \text{ Q2} - 5573 \text{ Q3}$$

Predictor	Coef	SE Coef	T	P
Constant	-5211	1353	-3.85	0.000
Customrs	0.43924	0.04107	10.69	0.000
Q1	8050.4	349.8	23.01	0.000
Q2	-928.2	350.3	-2.65	0.012
Q3	-5572.5	349.7	-15.94	0.000

S = 778.2 R-Sq = 98.0% R-Sq(adj) = 97.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	1027681074	256920268	424.27	0.000
Residual Error	35	21194686	605562		
Total	39	1048875760			

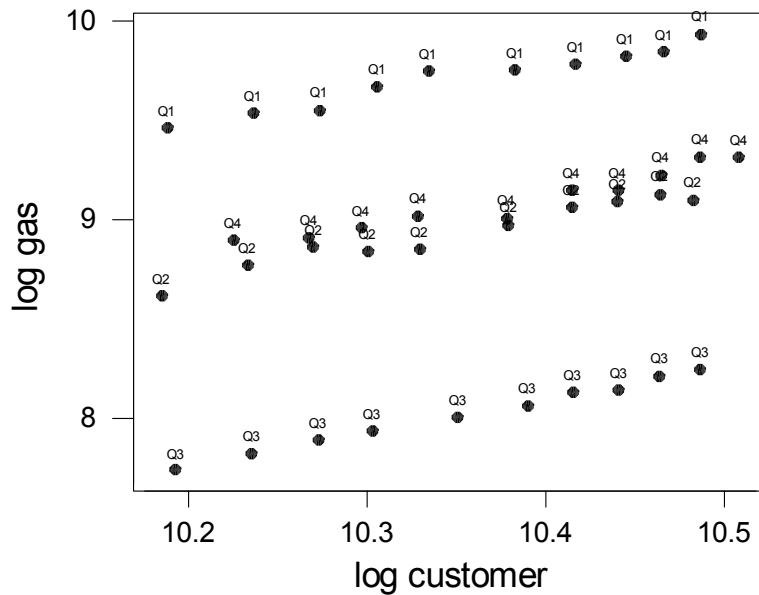
Superficially, this looks pretty good: Big t ratios, and a hefty R-square of 98%. However, the first and third quarter lines weren't parallel, so something isn't right. It is reasonable to guess that the *percentage* change in natural gas use from quarter to quarter is more likely to remain constant than the *absolute* change in natural gas use, as the number of customers increases. This insight would lead one to try a log transformation.

To compute the log of customers and log of natural gas use in Minitab, you'd use the LET command. The syntax would be

```
MTB > LET C8 = LOGE(C2)
```

This defines a new variable (the natural log of the variable located in C2) and stores the result in column eight. The command LOGE means "Log, base e." Note: You do *not* take the logs of the dummy variables. You use the same ones and zeros as before. After all, the log of zero isn't even defined. I named my newly defined variables and made the following scatter plot.

Log of Gas Use against Log of Customers Quarterly Data



These lines look much more nearly parallel, which we would expect them to be if the *percentage differences* between quarters were unchanging as the number of customers changed. Here is the result of the regression done in logs:

Regression Analysis: log gas versus log customer, Q1, Q2, Q3

The regression equation is

$$\log \text{ gas} = -7.26 + 1.58 \log \text{ customer} + 0.660 \text{ Q1} - 0.116 \text{ Q2} - 1.03 \text{ Q3}$$

Predictor	Coef	SE Coef	T	P
Constant	-7.2632	0.5828	-12.46	0.000
log cust	1.57578	0.05613	28.08	0.000
Q1	0.66014	0.01526	43.26	0.000
Q2	-0.11550	0.01528	-7.56	0.000
Q3	-1.03454	0.01525	-67.83	0.000

S = 0.03395 R-Sq = 99.7% R-Sq(adj) = 99.7%

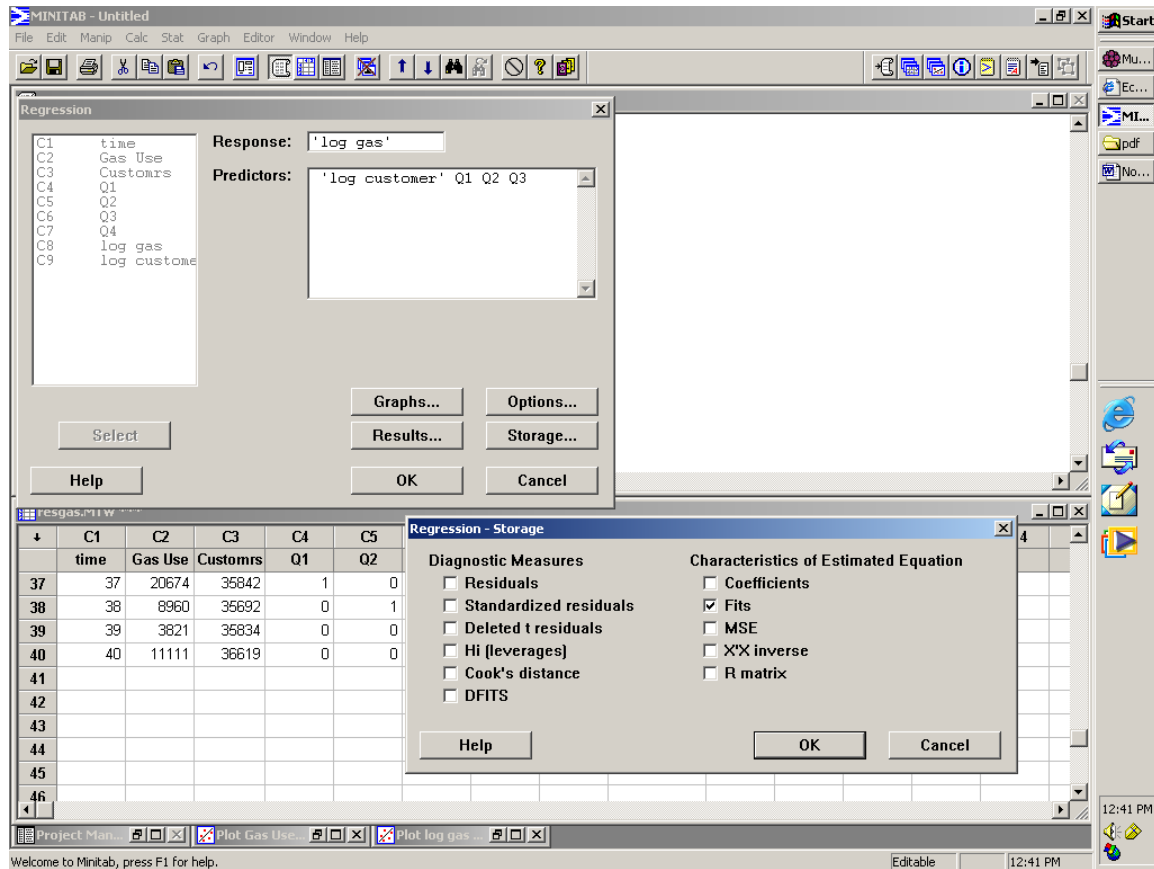
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	15.5790	3.8947	3379.07	0.000
Residual Error	35	0.0403	0.0012		
Total	39	15.6193			

Superficially, this looks really wonderful: Huge t ratios and an R-Square that jumped from 98% up to 99.7%! While the conclusion that the logs work much better proves to be correct, it is a bit premature to conclude so at this point. The R-Squares we are

comparing apples and oranges. In the original regression we explained 98% of the variability in natural gas use. In this regression we explained 99.7% of the variability in *the log of natural gas use*. To make these strictly comparable, we should use the log regression to recover predictions for natural gas use, and see what the sum of squared errors is in terms of natural gas use. Then we could make a valid comparison.

The first step is to save the predicted levels of the log of gas use for each of our observations. This is an option in Minitab, under STAT > REGRESSION > REGRESSION > STORAGE, where we select Fits.



When we perform the regression Minitab stores the fitted values for the log of gas use in our worksheet, with a column name of FITS1. For each observation, we need to compute predicted gas use from the predicted log of gas use. The log transform is “undone” by raising the log to the power e. Therefore, since my fitted values are stored in column eleven, the command

MTB > LET C12 = EXPO(C11)

raises each value in column 11 to the power e , and stores the result in column 12. Column 12, then, contains predicted values of gas use, based on the log regression.⁴ For instance, for our first observation, actual gas use was 12874. The prediction based on the log regression is 12722. By comparison, the prediction from our original (log free) regression was 14516.6. Of course, this is just one observation, and may not be representative. What we can do now, however, is compare the sum of squared errors predicting with the log regression to the sum of squared errors predicting with the regression in levels. In my worksheet, actual gas use is stored in column 2, and gas use predicted from the log regression is stored in column 12. We can compute squared prediction errors by simply using this command:

```
MTB > LET C13 = ( C2 - C12 )**2
```

Now Column 13 contains squared errors. We just need to add them up.

```
MTB > Sum C13
```

Sum of C13

```
Sum of C13 = 5203681
```

This is the sum of squared errors that one gets when predicting natural gas use (not the *log* of natural gas use) using the *log* regression. This can be directly compared to the sum of squared errors that one gets when not using logs. Flipping back to our original output, we discover that number is 21194686. Comparing these is easier if we add some commas. With logs – 5,203,681; without logs – 21,194,686. We reduced the sum of squared errors to less than a quarter of its original value by using the log transform and we did it with exactly the same number of parameters in the model!

Having decided that the log version works better, how do we use it to predict the future? Suppose we anticipate that we will have 37046 customers next quarter, and that next quarter will be the first quarter of the year. To predict gas use for this quarter, we need to first find the natural log of 37046, which is easy enough.

⁴ There is a complication here I am sweeping under the rug. The transformation is non-linear, so that we are not quite recovering the conditional expected value of gas use. There is an adjustment one can use in computing point estimates to correct for this: one can take $E(y) = \text{expo}\{\log_e(y) + \sigma_\epsilon^2/2\}$. However, the difference is generally small and the correction does not necessarily improve matters, since the coefficients in the log regression were selected to minimize squared errors in logs, not squared errors in levels. In this example, using the adjustment factor gives a SSE of 5,213,420, which is slightly worse than the result we achieve without the correction. For more discussion, see Terence C. Mills, *Time Series Techniques for Economists*, Cambridge University Press, 1990, pp. 338-39. I am using what Mills terms a “naïve retransformation.”

```
MTB > let k1=log(37046)
MTB > print k1
```

Data Display

K1 10.5199

In the first quarter of a year, the dummy variables take on the values 1, 0, 0. Therefore we predict for these values.

The screenshot shows the Minitab interface. The main window is titled 'MINITAB - Untitled'. The 'Regression' dialog box is open, showing the following settings:

- Response: 'log gas'
- Predictors: 'log customer', 'Q1', 'Q2', 'Q3'

The 'Regression - Options' dialog box is also open, showing the following settings:

- Weights: (empty)
- Fit intercept
- Display:
 - Variance inflation factors
 - Durbin-Watson statistic
 - PRESS and predicted R-square
- Lack of Fit Tests:
 - Pure error
 - Data subsetting
- Prediction intervals for new observations: '10.5119 1 0 0'
- Confidence level: '95'
- Storage:
 - Fits
 - SEs of fits
 - Confidence limits
 - Prediction limits

The background shows a data table with the following columns: C1 (time), C2 (Gas Use), C3 (Customrs), C4 (Q1), and C5 (Q2). The table contains data for rows 52 through 61.

When Minitab completes the regression, we obtain the following prediction information for the log of gas use.

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	9.96132	0.01394	(9.93303, 9.98962)	(9.88682, 10.03583)

We need to convert all these values from logs to levels by exponentiating them.

```
MTB > let k2=expo(9.96132)
MTB > let k3=expo(9.93303)
MTB > let k4=expo(9.98962)
MTB > let k5=expo(9.88682)
MTB > let k6=expo(10.03583)
MTB > print k2-k6
```

Data Display

K2	21190.7
K3	20599.7
K4	21799.0
K5	19669.4
K6	22830.0

In English, our point estimate of gas use is 21190.7, while the 95% Confidence interval for the conditional mean is (20599.7, 21799.0) and the 95% Prediction interval is (19669.4, 22830.0).

Interpreting regression coefficients in log regressions.

Returning to the estimated regression equation, what are we to make of the point estimates of the coefficients?

$$\log \text{ gas} = - 7.26 + 1.58 \log \text{ customer} + 0.660 \text{ Q1} - 0.116 \text{ Q2} - 1.03 \text{ Q3}$$

The coefficient of $\log \text{ customer}$, 1.58, is the elasticity of gas use with respect to the number of customers, holding season of the year constant. It is telling us that, holding season of the year constant, a 100% increase in customers is associated with a 158% increase in gas use. This seems a bit odd, since you might expect gas use to be proportional to the number of customers, in which case a 100% increase in the number of customers would yield a 100% increase in gas use, and the coefficient would be 1.00 instead of 1.58. It isn't a matter of sampling error, since the standard error of the estimated coefficient is only .056. The 95% confidence interval for the true value of this coefficient is $1.58 \pm (2.03)(.056) = 1.58 \pm .114$; statistically, the estimated coefficient is a *long way* from being 1.

The fact that as the number of customers grew over time, gas use grew more than proportionally may be a hint that something is missing in the model. Why did gas use rise so much faster than the number of customers? Perhaps new customers were more likely to use gas for heating instead of just cooking. Perhaps the new customers were in newer and bigger houses. Perhaps the new customers were disproportionately in cooler climates. We might want to look into these possibilities, because if one of them proved to be correct, we could improve our model. For instance, if the elasticity is so big because new customers were more likely to live in the north, the model might be improved by disaggregating customers into northern and southern customers.

Returning to interpretation, the coefficients of the dummy variables measure percentage shifts compared to the base (that is, the left out) quarter, holding customers constant. So the coefficient of Q1, namely, 0.660, says that holding number of customers constant, the first quarter uses 66% more gas than the fourth quarter. The other dummy variable coefficients have similar interpretations – for instance, the second quarter uses 11.6% less gas than the fourth quarter, holding number of customers constant.