Model-Fitting with Linear Regression: Exponential Functions

In class we have seen how least squares regression is used to approximate the linear mathematical function that describes the relationship between a *dependent* and an *independent* variable by minimizing the variation on the *y* axis. Linear regression is a very powerful statistical technique as it can be used to describe more complicated functions (such as exponential or power functions) by linearizing the data sets in question.

In this example we will look at the macroecological relationship between the size of the home-range (km^2) of a hunter-gatherer group, and the contribution (%) of hunted foods to the diet. We are interested in 1) *describing* this functional relationship mathematically, 2) *explaining* why this relationship holds as it does, and 3) *testing* the strength of this relationship using an alternative, independent data set. We will be using data from Kelly (1995) and Binford (2001).

If we were interested in developing a model that predicts the annual territory size (or home-range) of a hunter-gatherer group, a logical place to start would be to think about environmental constraints. Kelly (1995, chapter 4) hypothesizes that territory size should be related to the amount of hunted foods in the diet (% hunting) as the greater the reliance on mobile food resources, the greater the required area for hunting (holding all else equal). We can expand on Kelly's hypothesis by noting that area should increase exponentially, not linearly, with % hunting, as area is measured in km^2, not linear km, such that as % hunting increases, area should increase by a factor greater than 1: that is we expect the slope of the relationship $\beta_{x*y} > 1$.

Let *% Hunt* = the percentage contribution of hunted foods to the diet, *Area* = the home-range or area of the annual territory size of a hunter-gatherer group measured in square kilometers (km^2), and β_{H*A} = the slope of the relationship between *% Hunt* on *Area*. We wish to test the following hypothesis at the $a = 0.05$ (95%) confidence level:

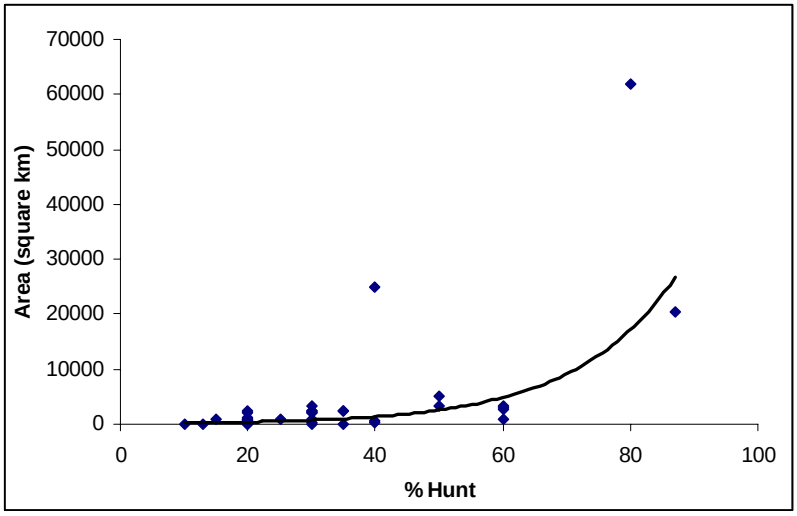
H_O: $\beta_{H*A} = 0$
H_A : not H_O

Kelly's data are as follows ($n = 39$):

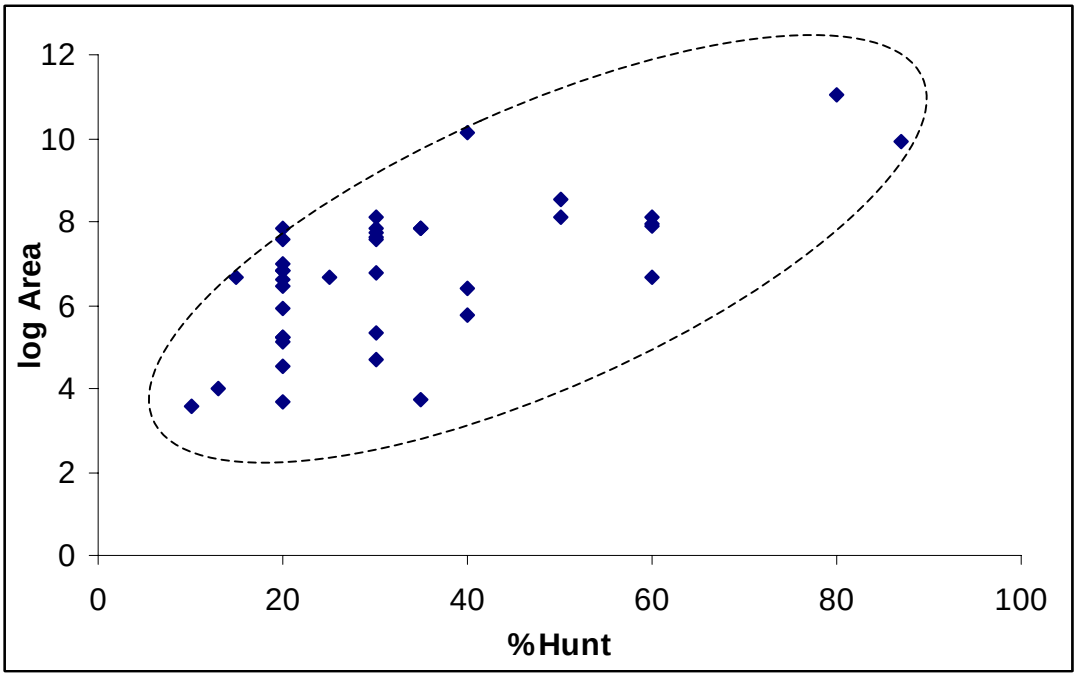
	% Hunt	*Area*		*% Hunt*	*Area*
yurok	10	35	walapai	40	588
andamanese	20	40	bella coola	20	625
vedda	35	41	s. kwakiutl	20	727
anbarra	13	56	siriono	25	780
tolowa	20	91	gwi	15	782
quinault	30	110	penan	30	861
ainu	20	171	kade gwi	20	906
makah	20	190	haida	20	923
puyallup	20	191	klamath	20	1058
twana	30	211	s. tlingit	30	1953
ojibwa	40	320	n. paiute	20	1964
nootka	20	370.5	nez perce	30	2000

washo	30	2327	polar inuit	40	25000
semang	35	2475	crow	80	61880
n. tlingit	30	2500	micmac	50	5200
hadza	35	2520	mbuti	60	780
monyagnais	60	2700	dobe	20	2500
plains cree	60	2890	maidu	30	3255
aeta	60	3265	nunamiut	87	20500
mistassini cree	50	3385			

We first need to see whether is some kind of consistent relationship between *% Hunt* and *Area*. To do this we can produce scatter plots in either EXCEL or MINITAB.



We can see from this EXCEL scatter plot that there does seem to be a trend to the data, only the trend is curvilinear rather than linear. Also we note that as *% Hunt* increases, *Area* seems to increase exponentially, as we hypothesized. The black line on the plot is a fitted exponential function. How do we describe mathematically an exponential function without a lot of math? Well, first we can try to linearize the relationship between *% Hunt* and *Area*. With an exponential relationship like this, we log transform the data on the *y* axis, that is for each y_i data point (*Area*) we take the base of the natural logarithms $log_e(y_i)$, or the command =ln(*y*) in EXCEL. We can then plot out this data.



We can see that by log-transforming the *y*-axis we have now linearized the trend in the data. This means that we can now use a simple linear regression model to describe the relationship between our variables of interest, remembering that we are now actually calculating the linear equation $\log_e Y = f(X)$, that is $log\ Y = \alpha + \beta X$. To convert $\log_e Y$ into Y we use some simple algebra with our final regression equation.

First, let's calculate the regression equation:

$\overline{X} = 33.205$
$\overline{Y} = 6.801$ (remember this is the mean of $\log_e Y$, *not* the mean of Y logged)

Calculation for $\hat{Y}$:

$$\beta = \frac{\sum xy}{x^2} = \frac{778.558}{12362.36} = 0.062978$$

$$\alpha = \overline{Y} - \beta\overline{X} = 6.801 - 0.062978 * 33.205 = 4.709847$$

$$\hat{Y} = \alpha + \beta X = 4.709847 + 0.062978X$$

%H	Area	ln Area	x	y	x2	xy	y2	Yhat	dXY	d2XY	yhat	y2hat
10	35	3.56	-23.21	-3.25	538.48	75.32	10.53	5.34	-1.78	3.18	-1.46	2.14
20	40	3.69	-13.21	-3.11	174.38	41.10	9.69	5.97	-2.28	5.20	-0.83	0.69
35	41	3.71	1.79	-3.09	3.22	-5.54	9.53	6.91	-3.20	10.24	0.11	0.01
13	56	4.03	-20.21	-2.78	408.25	56.08	7.70	5.53	-1.50	2.26	-1.27	1.62
20	91	4.51	-13.21	-2.29	174.38	30.24	5.24	5.97	-1.46	2.13	-0.83	0.69
30	110	4.70	-3.21	-2.10	10.27	6.73	4.41	6.60	-1.90	3.61	-0.20	0.04
20	171	5.14	-13.21	-1.66	174.38	21.91	2.75	5.97	-0.83	0.69	-0.83	0.69
20	190	5.25	-13.21	-1.55	174.38	20.52	2.41	5.97	-0.72	0.52	-0.83	0.69
20	191	5.25	-13.21	-1.55	174.38	20.45	2.40	5.97	-0.72	0.51	-0.83	0.69
30	211	5.35	-3.21	-1.45	10.27	4.64	2.10	6.60	-1.25	1.56	-0.20	0.04
40	320	5.77	6.79	-1.03	46.17	-7.02	1.07	7.23	-1.46	2.13	0.43	0.18
20	370.5	5.91	-13.21	-0.89	174.38	11.70	0.79	5.97	-0.05	0.00	-0.83	0.69
40	588	6.38	6.79	-0.42	46.17	-2.88	0.18	7.23	-0.85	0.73	0.43	0.18
20	625	6.44	-13.21	-0.36	174.38	4.80	0.13	5.97	0.47	0.22	-0.83	0.69
20	727	6.59	-13.21	-0.21	174.38	2.80	0.04	5.97	0.62	0.38	-0.83	0.69
25	780	6.66	-8.21	-0.14	67.32	1.16	0.02	6.28	0.37	0.14	-0.52	0.27
15	782	6.66	-18.21	-0.14	331.43	2.53	0.02	5.65	1.01	1.01	-1.15	1.31
30	861	6.76	-3.21	-0.04	10.27	0.14	0.00	6.60	0.16	0.03	-0.20	0.04
20	906	6.81	-13.21	0.01	174.38	-0.11	0.00	5.97	0.84	0.70	-0.83	0.69
20	923	6.83	-13.21	0.03	174.38	-0.35	0.00	5.97	0.86	0.74	-0.83	0.69
20	1058	6.96	-13.21	0.16	174.38	-2.15	0.03	5.97	0.99	0.99	-0.83	0.69
30	1953	7.58	-3.21	0.78	10.27	-2.49	0.60	6.60	0.98	0.96	-0.20	0.04
20	1964	7.58	-13.21	0.78	174.38	-10.32	0.61	5.97	1.61	2.60	-0.83	0.69
30	2000	7.60	-3.21	0.80	10.27	-2.56	0.64	6.60	1.00	1.00	-0.20	0.04
30	2327	7.75	-3.21	0.95	10.27	-3.05	0.90	6.60	1.15	1.33	-0.20	0.04
35	2475	7.81	1.79	1.01	3.22	1.82	1.03	6.91	0.90	0.81	0.11	0.01
30	2500	7.82	-3.21	1.02	10.27	-3.28	1.05	6.60	1.22	1.50	-0.20	0.04
35	2520	7.83	1.79	1.03	3.22	1.85	1.06	6.91	0.92	0.84	0.11	0.01
60	2700	7.90	26.79	1.10	717.97	29.47	1.21	8.49	-0.59	0.35	1.69	2.85
60	2890	7.97	26.79	1.17	717.97	31.30	1.36	8.49	-0.52	0.27	1.69	2.85
60	3265	8.09	26.79	1.29	717.97	34.56	1.66	8.49	-0.40	0.16	1.69	2.85
50	3385	8.13	16.79	1.33	282.07	22.27	1.76	7.86	0.27	0.07	1.06	1.12
40	25000	10.13	6.79	3.33	46.17	22.60	11.06	7.23	2.90	8.40	0.43	0.18
80	61880	11.03	46.79	4.23	2189.76	198.03	17.91	9.75	1.28	1.65	2.95	8.69
50	5200	8.56	16.79	1.76	282.07	29.48	3.08	7.86	0.70	0.49	1.06	1.12
60	780	6.66	26.79	-0.14	717.97	-3.80	0.02	8.49	-1.83	3.35	1.69	2.85
20	2500	7.82	-13.21	1.02	174.38	-13.51	1.05	5.97	1.85	3.44	-0.83	0.69
30	3255	8.09	-3.21	1.29	10.27	-4.12	1.66	6.60	1.49	2.22	-0.20	0.04
87	20500	9.93	53.79	3.13	2893.89	168.22	9.78	10.19	-0.26	0.07	3.39	11.48
1295	**156171**	**265.2407**	**0**	**0**	**12362.36**	**778.56**	**115.501**	**265.241**	**0**	**66.4689**	**0**	**49.032**

So, our regression equation at this stage is $\log_e\left(\hat{Y}\right) = \alpha + \beta X = 4.709847 + 0.062978X$.
However, we are really interested in $\hat{Y}$, not $\log_e(\hat{Y})$, so we use some algebra to get us there:

$$\log_e\left(\hat{Y}\right) = \alpha + \beta X$$
$$e^{\log_e\left(\hat{Y}\right)} = e^{(\alpha+\beta X)}$$
$$\hat{Y} = e^{\alpha} e^{\beta X}$$
$$\hat{Y} = \mathrm{A}e^{\beta X}$$

So our final regression equation is,

$Y = 111.04e^{0.063X}$

This is an exponential function where the Y intercept is the same as usual (a) but Y increases as an exponential function of X. In this case our $\beta_{H*A} = e^{0.063} = 1.065$, which is as we hypothesized, $\beta_{H*A} > 1$.

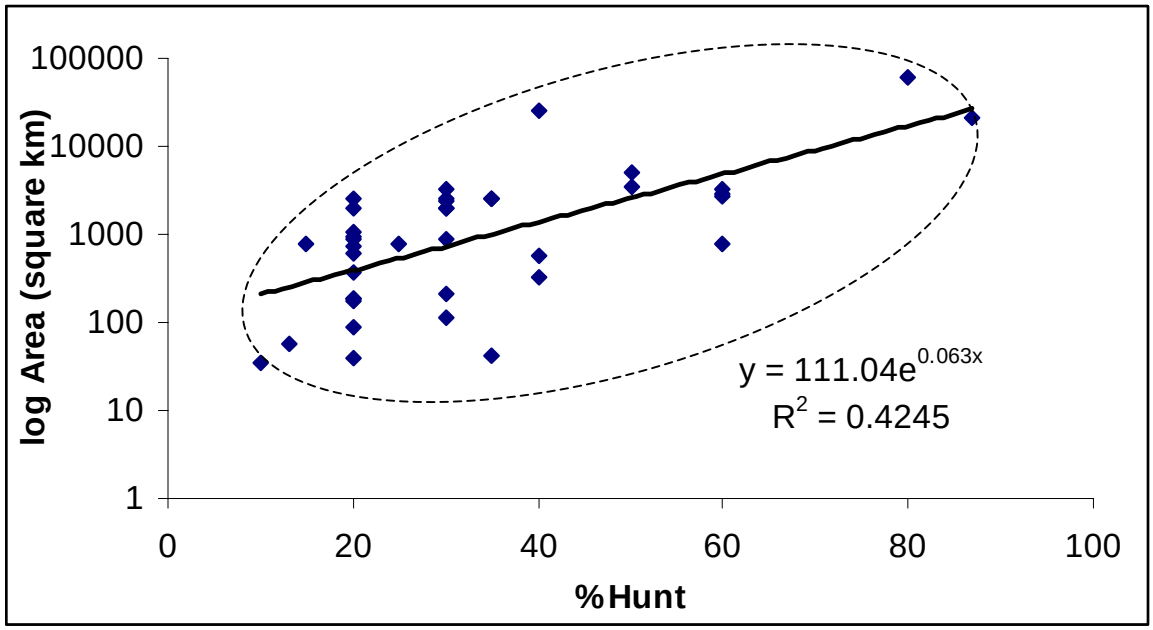
But, we are far from finished! We still need to calculate our ANOVA table, and calculate the resulting significance. So, calculating the quantities:

$\Sigma X = 1295$
$\Sigma Y = 265.241$
$\Sigma X^2 = 55363$
$\Sigma Y^2 = 1919.415$
$\Sigma XY = 9585.909$
$\overline{X} = 33.205$
$\overline{Y} = 6.801$
$\Sigma x^2 = 12362.35$
$\Sigma y^2 = 115.501$
$\Sigma xy = 778.558$
$\Sigma y^2 = 49.032$
$\Sigma d^2_{YX} = 66.467$

The resulting ANOVA table is:

Source of Variation	***df***	***SS***	***MS***	F_{STAT}	***p***
Explained	1	49.032	49.032	27.295	<0.001
Error	37	66.467	1.796		
Total	38	115.499			

To estimate our p value, we look up the F_{CRIT} {1,37,0.05}. In the table, they do not give a value for 37, so the closest value is 40. Therefore our $F_{CRIT} \approx 4.08$ which is much less than our F_{STAT} suggesting we reject the null hypothesis in favor of the alternative. Looking down the F_{CRIT} values, our F_{STAT} is greater than the smallest value given, so that we can be safe in saying our $p < 0.001$.



The regression coefficient of this test is r^2 = ExSS/TotalSS = 49.032/115.499 = 0.425, that is to say the amount of hunting in the diet explains about 42.5% of the variation in annual territory size. Is this a good result? Well first of all I would note that the relationship is highly significant, as the *p* value is infinitesimally small, and further, if you consider all the factors that might go into the size of a territory (environmental/ecological variation, competition with neighboring groups, different types of hunted prey, different technologies, group sizes etc.), being able to explain over 40% of the variation with a single variable is pretty powerful.

To run this test in MINITAB:

>STAT
>REGRESSION
>REGRESSION
>RESPONSE is log AREA
>PREDICTOR is % HUNT
>STORAGE
>RESIDUALS
>OK

The output looks like:

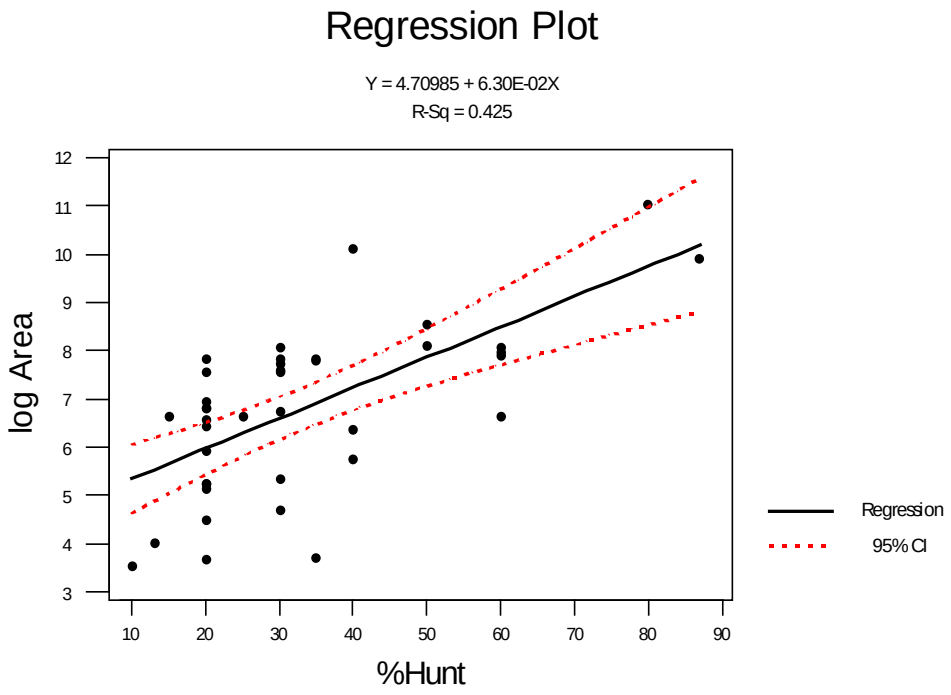
Regression Analysis

```
The regression equation is
log Area = 4.71 + 0.0630 %Hunt

Predictor       Coef       StDev          T        P
Constant      4.7098      0.4542      10.37    0.000
%Hunt        0.06298     0.01205       5.22    0.000

S = 1.340       R-Sq = 42.5%     R-Sq(adj) = 40.9%

Analysis of Variance

Source       DF          SS          MS          F        P
Regression    1      49.032      49.032     27.29    0.000
Error        37      66.469       1.796
Total        38     115.501

Unusual Observations
Obs     %Hunt   log Area        Fit  StDev Fit   Residual    St Resid
  3      35.0      3.714      6.914      0.216     -3.201      -2.42R
 33      40.0     10.127      7.229      0.230      2.898       2.19R
 34      80.0     11.033      9.748      0.604      1.285       1.07 X
 39      87.0      9.928     10.189      0.683     -0.261      -0.23 X
```

So in the MINITAB output we get the regression equation, the r^2 value, and the ANOVA table, all of which should agree with our hand calculations. The *unusual observations* at the bottom of the output is a list of variables that have a large influence on the relationship. What does this mean? This means that, depending on your time, interest, or the question at hand, you may choose to run the regression analysis with all or none of these variables included. By omitting these variables it is possible to weed out those observations that have a large influence on the end result. There is no cut and dried formula as to whether you should do this: it is really up to you to decide how you want to manage your data.

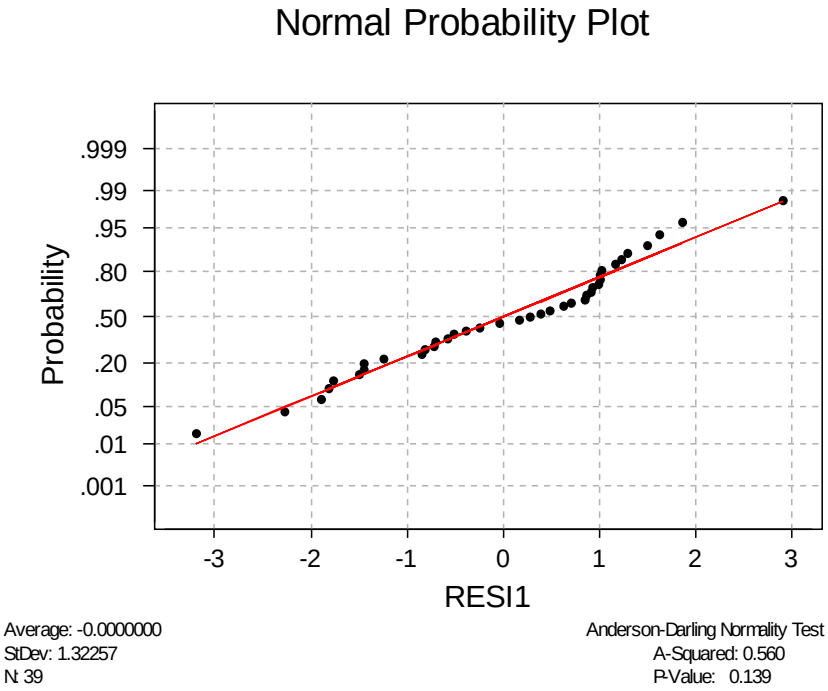
In MINITAB you cannot produce a regression scatter plot from the GRAPH option in the analysis, but you can produce one under FITTED LINE PLOTS under the regression options dialogue plot.

Here is the MINITAB version of the graphical output:



Notice that the regression equation given here is *before* we transformed it back into the original unlogged version (see above).

One last thing we have to do is check our normality assumption as it relates to the residuals. To do this we run a normality test on our residuals and we find:



Our *p* is greater than our unstated *a*, and so we can say our normality assumption in met.

An interesting second stage of this analysis would be to test the relationship we have documented here against an independently derived data set. If we could do this, it would help us establish whether the relationship we have documented above is simply an artifact of the data Kelly gathered, or whether this is due to a real, robust relationship between territory size and hunting. Luckily, Binford's data can be used to test this idea.

The Binford data set includes a very large sample size ($n = 339$) suggesting any relationships we find should be pretty powerful, and is therefore an ideal data set with which to test the Kelly relationship. I will not list all 339 hunter-gatherer groups!

The MINITAB output is:

Regression Analysis

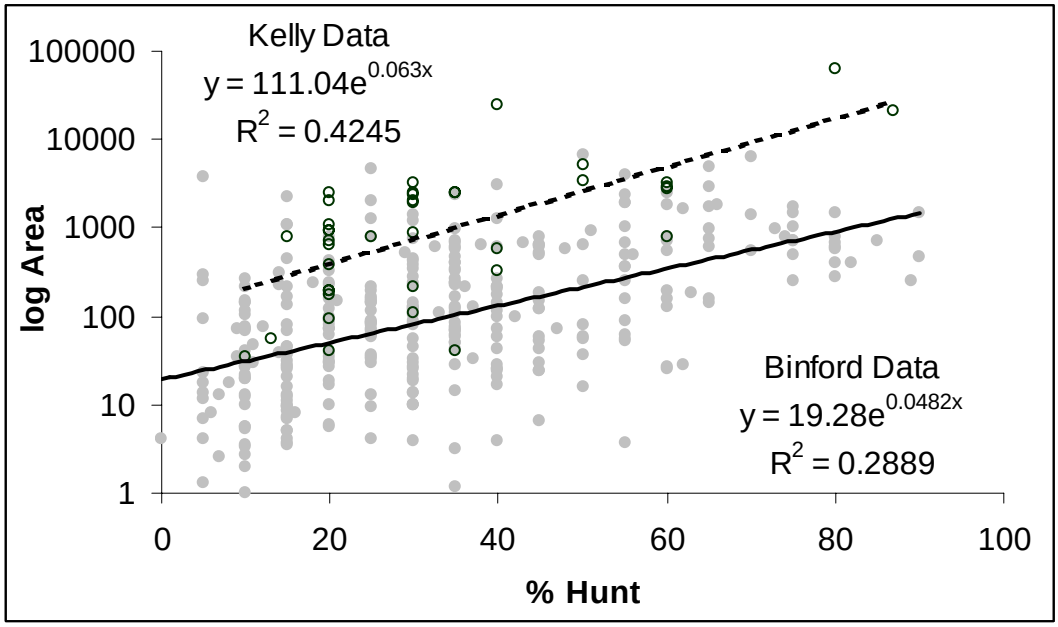
```
The regression equation is
log AREA = 2.96 + 0.0482 HUNT

Predictor       Coef       StDev           T         P
Constant      2.9591      0.1593       18.57     0.000
HUNT        0.048187    0.004118       11.70     0.000

S = 1.517       R-Sq = 28.9%     R-Sq(adj) = 28.7%

Analysis of Variance

Source       DF          SS          MS          F         P
Regression    1      314.97      314.97     136.91     0.000
Error       337      775.31        2.30
Total       338     1090.28
```

For consistency we'll convert the regression equation to reflect *Area*, not *log Area*.

$Y = 19.28e^{0.0482X}$

Notice that the $r^2 = 0.289$ or 28.9%, which is much less than the Kelly data, but still this is not bad considering we have about ten times more data from the Binford data set!

To see graphically how similar the relationships our in both of the data sets we can plot the two sets of data on the same scatter plot in EXCEL:



Now, the value we are interested in here is Binfords $\beta = \beta_B = \exp(0.0482)$, which on the face of it seems to be pretty close to the Kelly $\beta_K = \exp(0.063)$. This is illustrated on the above scatter plot. To test whether $\beta_B = \beta_K$ we can put confidence limits around β_K and see whether they encompass a value of β_B.

First we need to calculate the standard error of the regression coefficient:

$$S_\beta = \sqrt{\frac{S^2_{Y*X}}{\sum x^2}} = \sqrt{\frac{1.796}{12362.35}} = 0.0121$$

And the T_{CRIT} {0.05, 37} ≈ 2.021, therefore the upper and lower confidence limits are:

$CL_L = e^{0.063} - 0.0121*2.021 = 1.0406$
$CL_U = e^{0.063} + 0.0121*2.021 = 1.0895$

As $\beta_B = \exp(0.0482) = 1.0494$ we find that the slope from the Binford data falls within the 95% confidence limits of Kelly's slope suggesting they are statistically equal, and that the relationship we first documented between the contribution of hunted foods to the diet and home-range size holds. As such, we could go on to use the slope of this relationship as a robust model for predicting hunter-gatherer territory sizes from aspects of their diet.

To summarize all of the above, we can now use linear regression to approximate relatively complicated functions by simply log transforming the appropriate data.