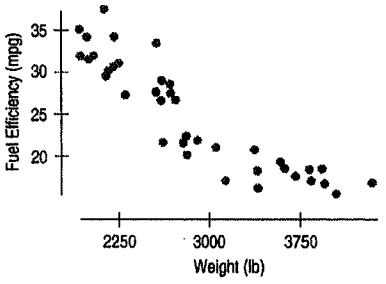
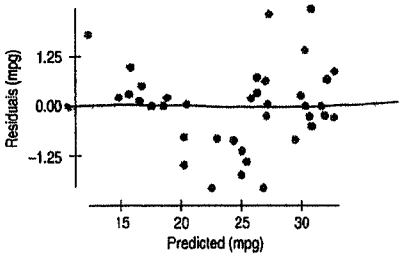
Chapter 10: Re-expressing Data: Get It Straight!

Re-expressing Data

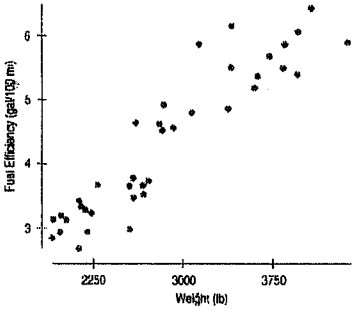
We cannot use a linear model unless the relationship between the two variables is linear. If the relationship is nonlinear (which we can verify by examining the residual plot) we can try reexpressing the data. Then we can fit and use a simple linear model. To re-express the data, we perform some mathematical operation on the data values such as taking the reciprocal, taking the logarithm, or taking the square root

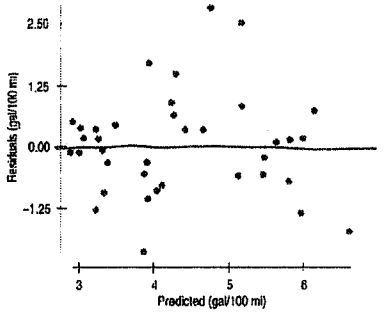
For example, consider the relationship between the weight of cars and their fuel efficiency (miles per gallon). What do the scatterplot and residual plots reveal? linear model is not appropriate





If we take the reciprocal of the *y*-values, we get the following scatterplot and residual plot. What do these plots reveal? relationship between weight and gal/100 miles is linear





There are several reasons we may want to re-express our data:

1. To make the distribution of a variable more symmetric
2. To make the spreads of several groups more alike.
3. To make the form of a scatterplot more linear.
4. To make the scatter in a scatterplot more spread evenly.

Re-expressing Data Using Logarithms

An equation of the form $y = a + bx$ is used to model linear data.

The process of transforming nonlinear data into linear data is called linearization. In order to linearize certain types of data we use properties of logarithms.

PROPERTIES OF LOGARITHMS:

1. $\log ab =$ log a + log b
2. $\log \frac{a}{b} =$ log a − log b
3. $\log x^p =$ p log x

Case 1: Consider the following set of Linear Data representing an account balance as a function of time:

x: time (months)	0	48	96	144	192	240
y: account balance ($)	100	580	1060	1540	2020	2500

Describe the pattern of change... balance increases by $480 per month

The relationship between *x* and *y* is linear if, for equal increments of *x*, we add a fixed increment to *y*.

Case 2: Consider the following set of Nonlinear Data representing an account balance as a function of time:

x: time (months)	0	48	96	144	192	240
y: account balance ($)	100	161.22	259.93	419.06	675.62	1089.30

Describe the pattern of change... balance increases by 61.22%

The relationship between *x* and *y* is exponential if, for equal increments of *x*, we multiply a fixed increment by *y*. This increment is called the common ratio

We want to find the best fitting model to represent our data.

- For the linear data, we use least-squares regression to find the best fitting line.
- For the nonlinear data, the best fitting model would be an exponential curve.

PROBLEM: We cannot use least-squares regression for the nonlinear data because least-squares regression depends upon correlation, which only measures the strength of linear relationships.

SOLUTION: We transform the *nonlinear data* into *linear data*, and then use least-squares regression to determine the best fitting line for the transformed data.

Finally, do a reverse transformation to turn the linear equation back into a nonlinear equation which will model our original *nonlinear data*.

Linearizing Exponential Functions:

(We want to write an exponential function of the form $y = a \cdot b^x$ as a function of the form $y = a + bx$).

$y = a \cdot b^x$ (x, y are variables and a, b are constants)

$\log y = \log(a \cdot b^x)$

$\log y = \log a + \log b^x$

$\log y = \log a + x \log b$

Var2 = con1 + (var1)(con1)

This is in the general form $y = a + bx$, which is linear.

So, the graph of (var1, var2) is linear. This means the graph of $(x, \log y)$ is linear.

CONCLUSIONS:

1. If the graph of x, $\log y$ is linear, then the graph of x, y is exponential.

2. If the graph of x, y is exponential, then the graph of x, $\log y$ is linear.

Once we have linearized our data, we can use least-squares regression on the transformed data $(x, \log y)$ to find the best fitting linear model.

PRACTICE:

Linearize the data for Case 2 and find the least-squares regression line for the transformed data.

Then, do a reverse transformation to turn the linear equation back into an exponential equation.

$\log \hat{y} = 2 + 0.0043x$ ← linear model

$10^{\log \hat{y}} = 10^{2+0.0043x}$

$\hat{y} = (10^2)(10^{0.0043})^x$

$\hat{y} = 100(1.01)^x$ ← exponential model

Compare this to the equation the calculator gives when performing exponential regression on the Case 2 data.

Linearizing Power Functions:

(We want to write a power function of the form $y = ax^b$ as a function of the form $y = a + bx$).

$y = ax^b$ (x , y are variables and a , b are constants)

$\log \hat{y} = \log(a x^b)$

$\log \hat{y} = \log a + b \log x$

Var2 = con 1 + con2 var1

This is in the general form $y = a + bx$, which is linear.

So, the graph of (var1, var2) is linear. This means the graph of $(\log x, \log y)$ is linear.

Case 3: Consider the following set of Nonlinear Data representing the average length and weight at different ages for Atlantic Ocean rockfish:

x: age (years)	0	4	8	12	16	20
y: weight (grams)	0	48	192	432	768	1200

PRACTICE:

Linearize the data for Case 3 and find the least-squares regression line for the transformed data.

Then, do a reverse transformation to turn the linear equation back into a power equation.

$$\log \hat{y} = 0.726 + 1.762 \log x$$

$$10^{\log \hat{y}} = 10^{0.726 + 1.762 \log x}$$

$$\hat{y} = (10^{0.726})(10^{\log x^{1.762}})$$

$$\hat{y} = (10^{0.726}) x^{1.762}$$

$$\hat{y} = 5.321 x^{1.762}$$

Compare this to the equation the calculator gives when performing power regression on the Case 3 data.