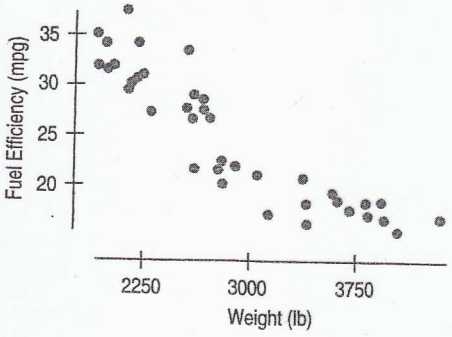
Linearising Transformations.

Q: How does a car's weight its fuel efficiency?

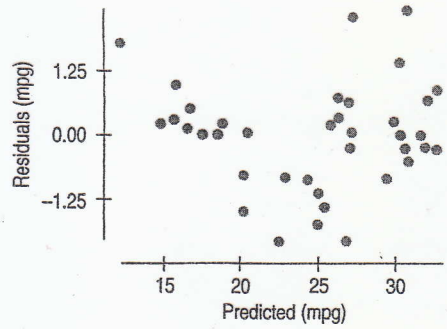
Below we display the reported weight and fuel efficiency for 38 cars:



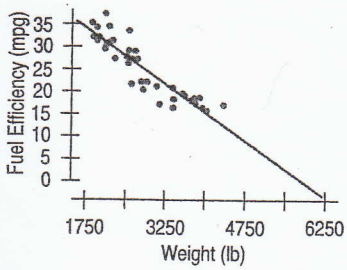
The plot shows a negative direction, roughly linear shape, and strong relationship.

Let's examine the residuals:

The residuals reveal a bent shape when plotted against the predicted values.

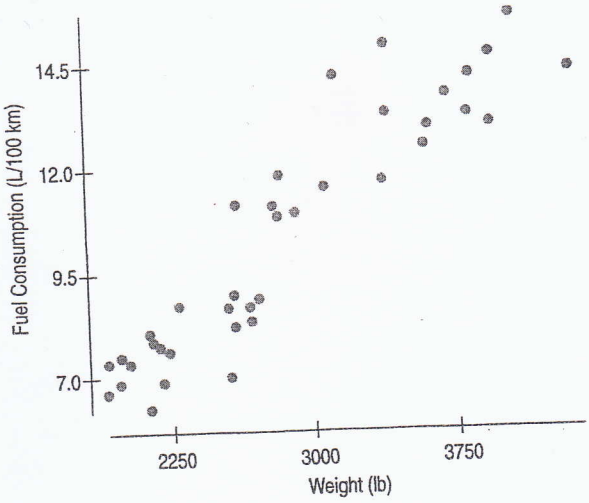


Now let's think about the regression line through the points:



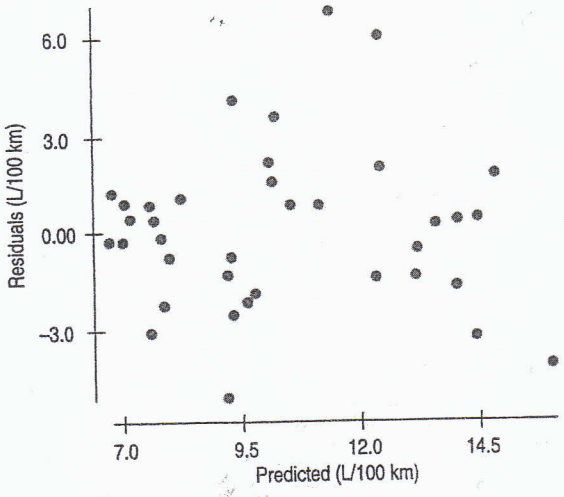
It looks as though the fuel efficiency would go negative at about 6000 pounds, which is absurd.

FIGURE 10.3
Extrapolating the regression line gives an absurd answer for vehicles that weigh as little as 6000 pounds.



we multiplied by 235 to get litres / 100 km, which is more nearly linear against weight than the original variable, but the re-expression changes the direction of the relationship

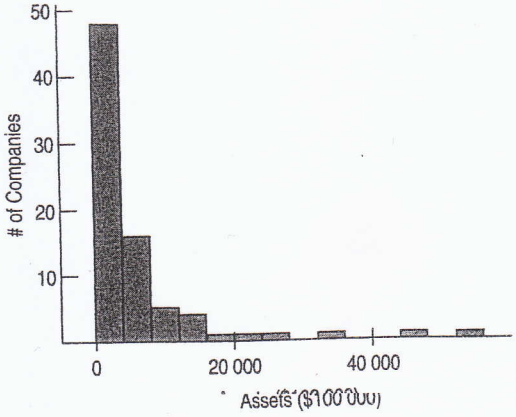
The residuals show less of a pattern than before.



Why do we need to re-express data? There are several reasons for that.

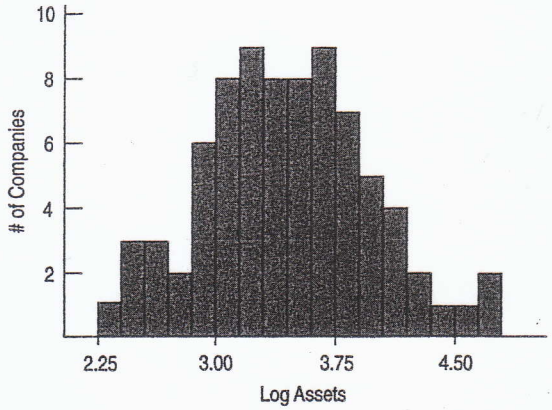
Reason 1:

Ex: Let's look at the assets of 77 companies:

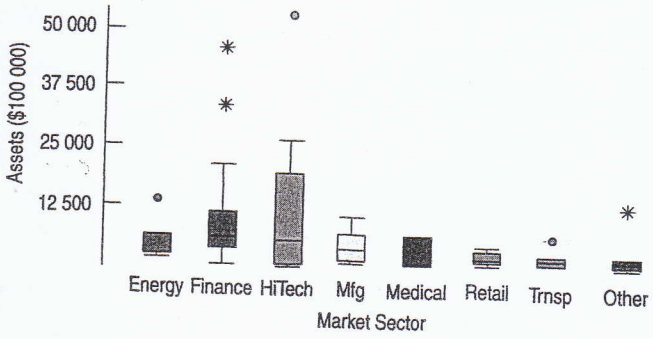


The dist'n is skewed to the right. Data on wealth often look like this

Taking logs produces a more nearly symmetric dist'n.

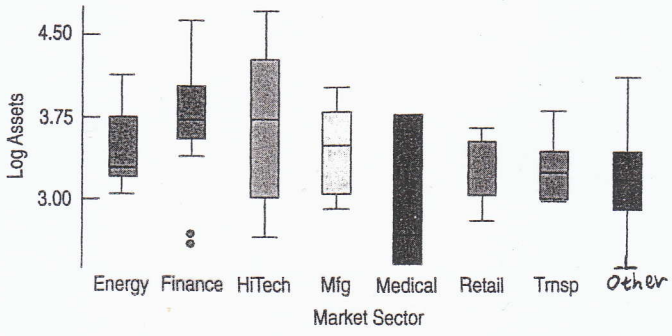


Reason 2:



Assets of large companies by market sector. It's hard to compare centres or spreads, and there seems to be a number of high outliers.

After re-expressing by logs. It's much easier to compare across market sectors.



Reason 3:

Reason 4:

How can we pick a re-expression to use?

The re-expressions line up in order.

Here are some scatterplot shapes that may arise;

Possible transformations are:

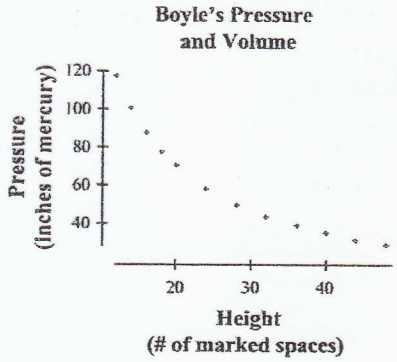
Ex.

Pressure. Scientist Robert Boyle examined the relationship between the volume in which a gas is contained and the pressure in its container. He used a cylindrical container with a moveable top that could be raised or lowered to change the volume. He measured the height in inches by counting equally spaced marks on the cylinder, and measured the pressure in inches of mercury (as in a barometer). Some of his data are listed in the table. Create an appropriate model.

Height	48	44	40	36	32	28
Pressure	29.1	31.9	35.3	39.3	44.2	50.3
Height	24	20	18	16	14	12
Pressure	58.8	70.7	77.9	87.9	100.4	117.6

Pressure. The scatterplot at the right shows a strong, curved, negative association between the height of the cylinder and the pressure inside. Because of the curved nature of the association, a linear model is not appropriate.

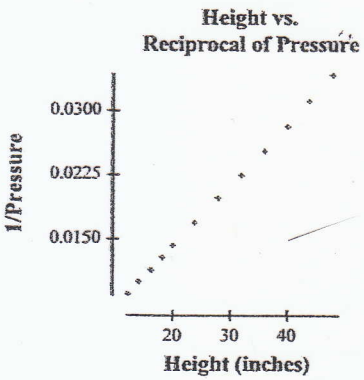
Re-expressing the pressure as the reciprocal of the pressure produces a scatterplot (below the first) that is much straighter. Computer regression output for the height versus the reciprocal of pressure is below.



Dependent variable is: recip pressure
No Selector
R squared = 100.0% R squared (adjusted) = 100.0%
s = 0.0001 with 12 - 2 = 10 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	0.000841	1	0.000841	75241
Residual	0.000000	10	0.000000	

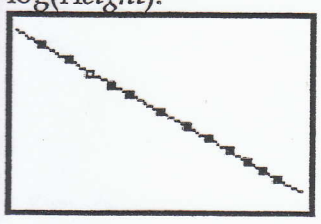
Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-7.66970e-5	0.0001	-0.982	0.3494
Height	7.13072e-4	0.0000	274	Ŝ 0.0001



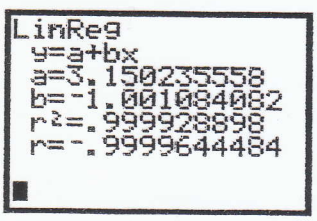
The reciprocal re-expression is very straight (perfectly straight, as far as the statistical software is concerned!). $R^2 = 100\%$, meaning that 100% of the variability in the reciprocal of pressure is explained by the model. The equation of the model is: $\frac{1}{\hat{Pressure}} = -0.000077 + 0.000713(Height)$.

Re-expressing each variable using logarithms is also a good model. TI-83 regression output for the log-log re-expression is given below.

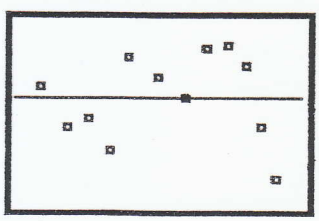
There is a strong, negative, linear relationship between log(*Pressure*) and log(*Height*).



Scatterplot of Log(pressure) vs. Log(height)



Regression Output



Residuals Plot

$\log(\hat{Pressure}) = 3.150 - 1.001(\log(Height))$ models the situation well, explaining nearly 100% of the variability in the logarithm of pressure. The residuals plot is fairly scattered, indicating an appropriate model.

Analysis of Variance.

The statistical methodology for comparing several means is called analysis of variance, or ANOVA. We will consider two ANOVA techniques:

One-way ANOVA:

Two-way ANOVA

One-way ANOVA.

Ex. 1

EXAMPLE

Choosing the best magazine layout. A magazine publisher wants to compare three different layouts for a magazine that will be offered for sale at supermarket checkout lines. She is interested in whether there is a layout that better catches shoppers' attention and results in more sales. To investigate, she randomly assigns each of 60 stores to one of the three layouts and records the number of magazines that are sold in a one-week period.

Ex. 2

EXAMPLE

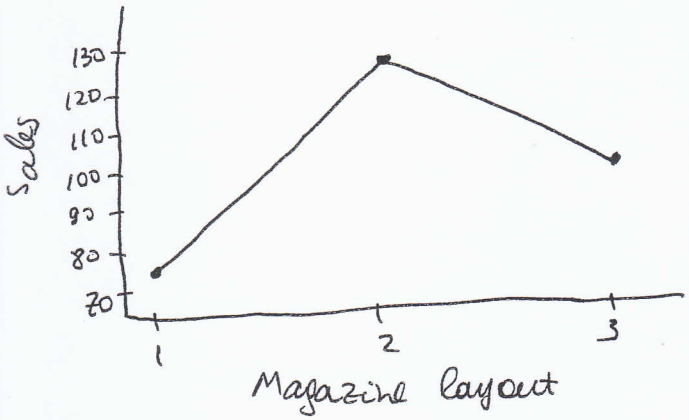
Average age of bookstore customers. How do five bookstores in the same city differ in the demographics of their customers? Are certain bookstores more popular among teenagers? Do upper-income shoppers tend to go to one store? A market researcher asks 50 customers of each store to respond to a questionnaire. Two variables of interest are the customer's age and income level.

What the difference between these two examples?

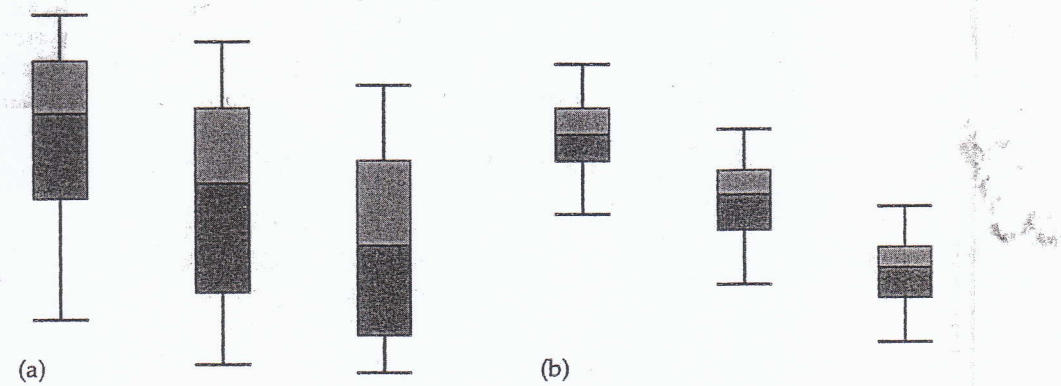
Comparing means.

Question: Do all groups have the same population mean?

Let's look at the dispay of the sample means for Ex 1.



Is the observed difference in sample means are statistically significant? Is it the result of chance variation? ANOVA answers this question.



a) Side-by-side boxplots for three groups with large within-group variation. The differences among centers may be just chance variation. (b) Side-by-side boxplots for three groups with the same centers as in Figure (a) but with small within-group variation. The differences among centers are more likely to be significant.