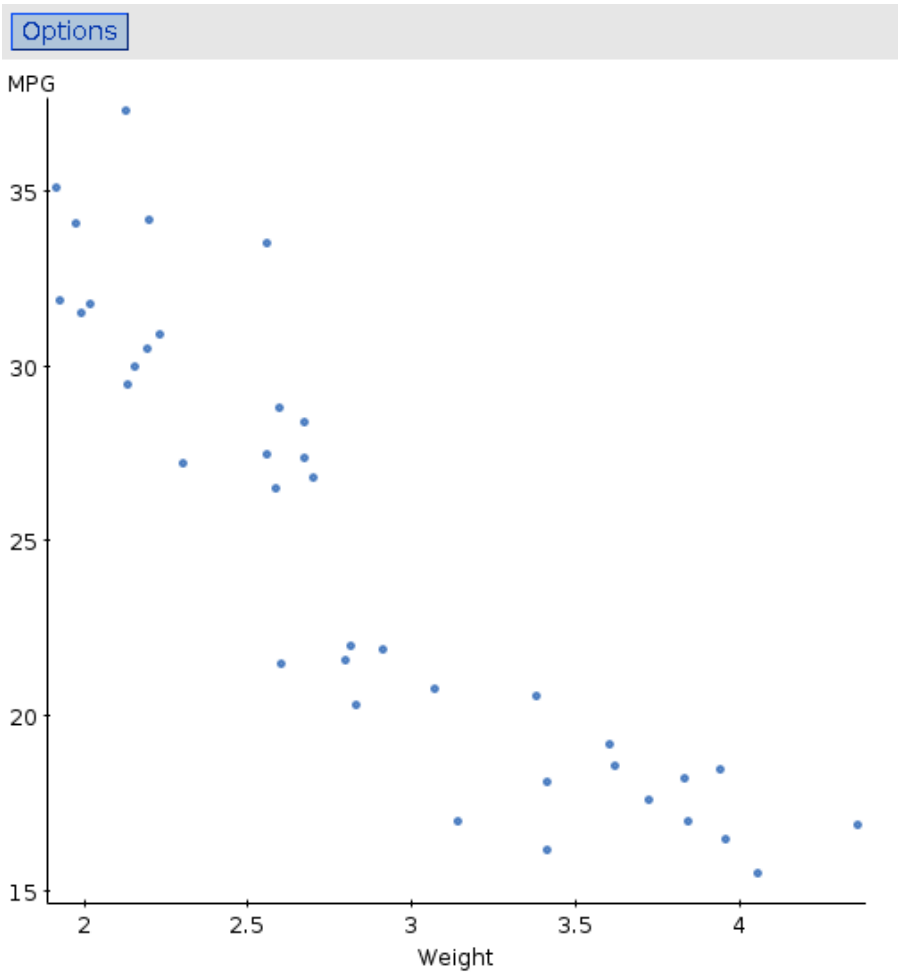
Chapter 9: Regression Wisdom p231

Patterns on residual plots p231

- If a car is heavier, what effect does that have on gas mileage?
- Data for 38 cars from magazine survey.
- Direction, strength, *form*?



Do the regression for predicting MPG from weight:

Options

Simple linear regression results:
Dependent Variable: MPG
Independent Variable: Weight
MPG = 48.707497 - 8.3646 Weight
Sample size: 38
R (correlation coefficient) = -0.9031
R-sq = 0.8155369
Estimate of error standard deviation: 2.850805

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	48.707497	1.9536817	≠ 0	36	24.931131	<0.0001
Slope	-8.3646	0.6630203	≠ 0	36	-12.615903	<0.0001

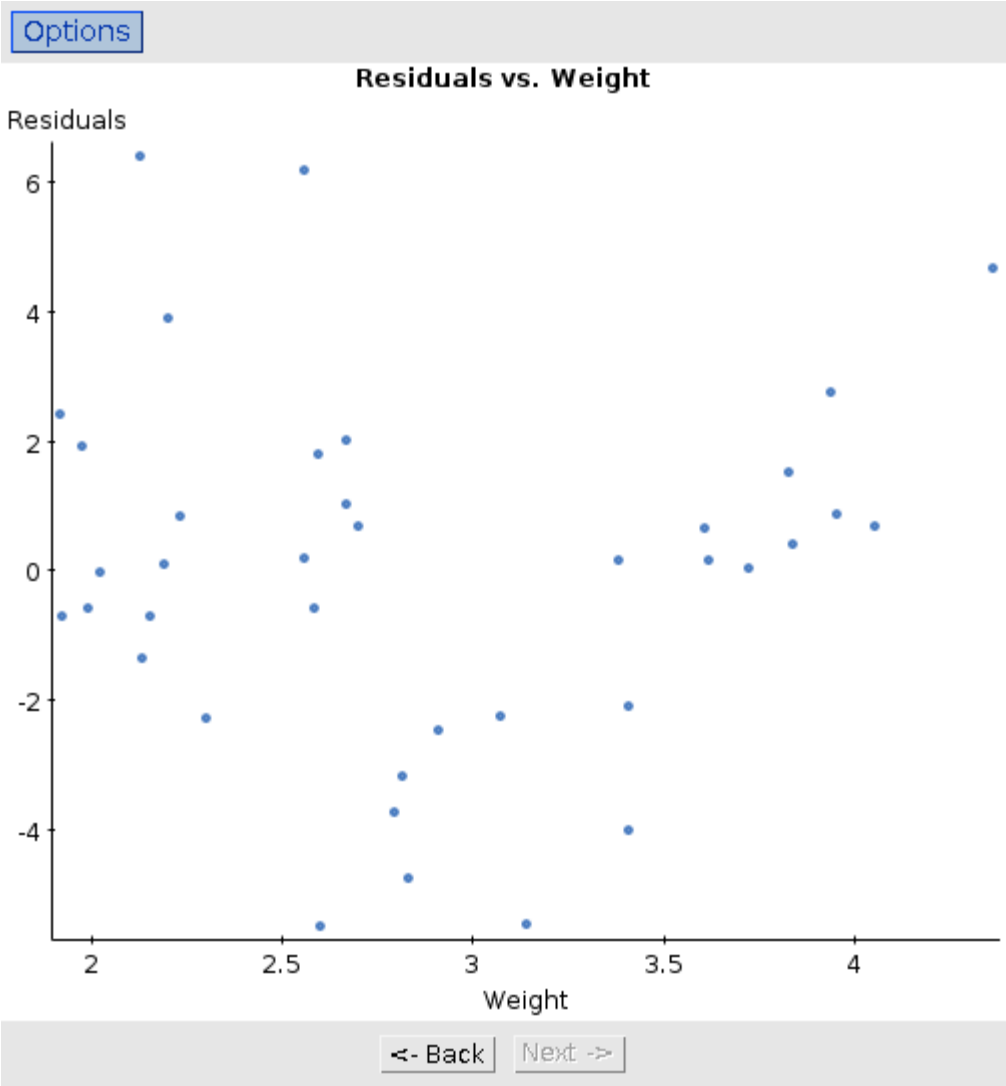
Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	1293.5156	1293.5156	159.16101	<0.0001
Error	36	292.5752	8.127089		
Total	37	1586.0908			

- high R-squared
- strongly negative correlation and slope.
- Next stage: plot the *residuals.* I also got a scatterplot with fitted line on it.

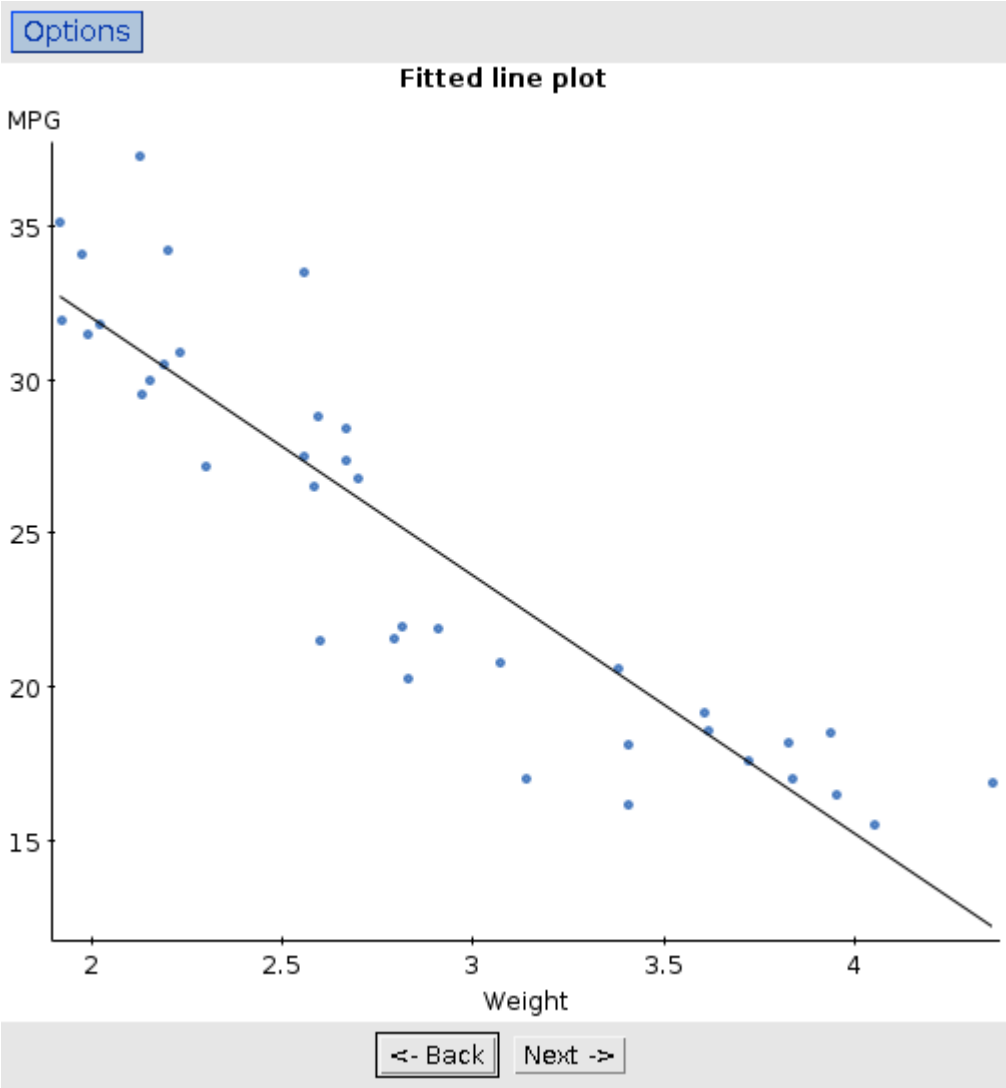
The residual plot:

- Doesn't look completely random, but a bit *curved.*



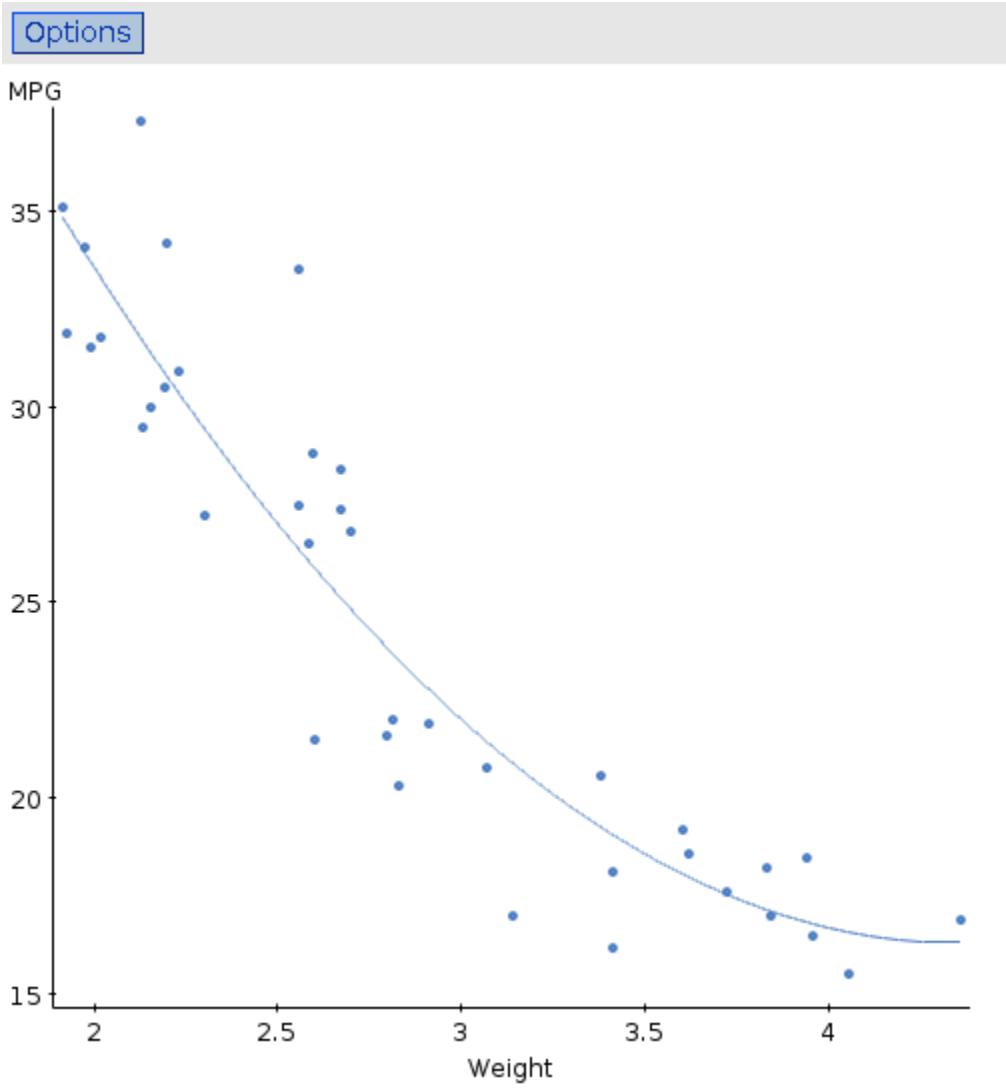
Look at scatterplot with fitted line:

- Looking more carefully, the curve *does* show up.
- See how to fix this in chapter 10.



Scatterplot with fitted curve

- curve *does* seem to go through points better
- as weight increases, MPG does not decrease so fast.



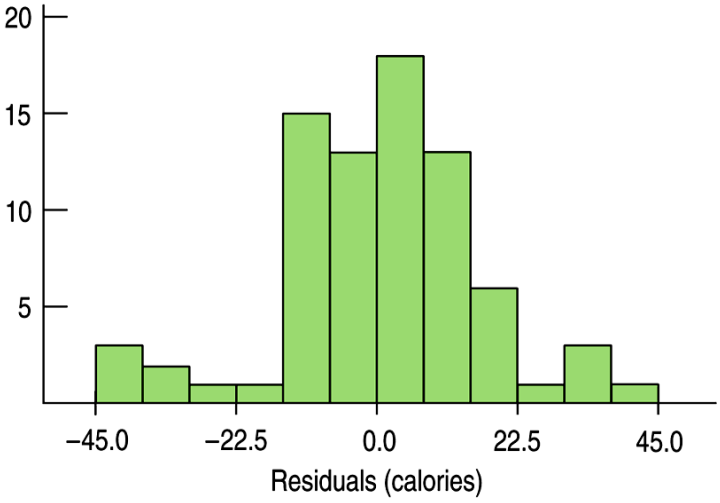
Sifting Residuals for Groups p233

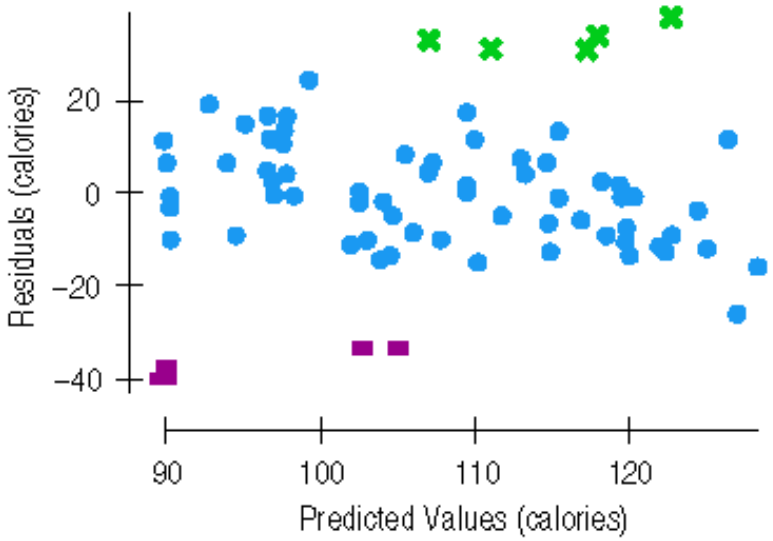
No regression analysis is complete without a display of the residuals to check that the linear model is reasonable.

Residuals often reveal subtleties that were not clear from a plot of the original data

Sometimes they reveal violations of the regression conditions that require our attention

- It is a good idea to look at both a histogram of the residuals and a scatterplot of the residuals vs. predicted values:





- The small modes in the histogram are marked with different colors and symbols in the residual plot above. What do you see?

- An examination of residuals often leads us to discover groups of observations that are different from the rest.

- When we discover that there is more than one group in a regression, we may decide to analyze the groups separately, using a different model for each group.

Outliers, Leverage, and Influence p234

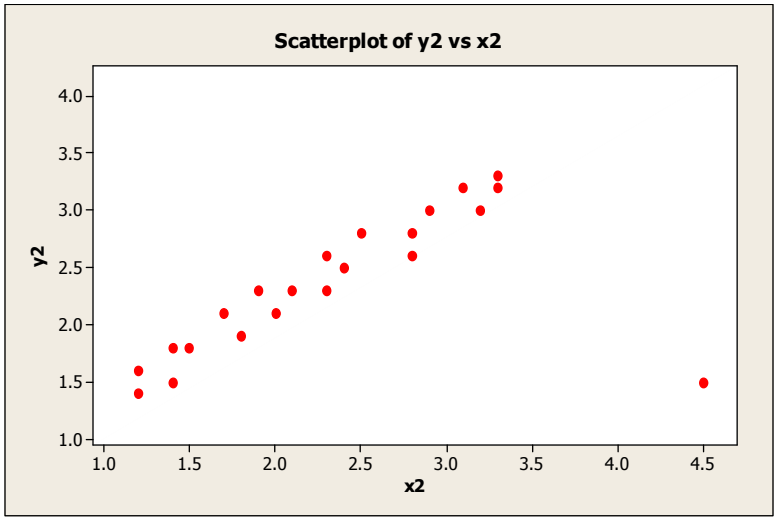
Any point that stands away from the others can be called an outlier and deserves your special attention.

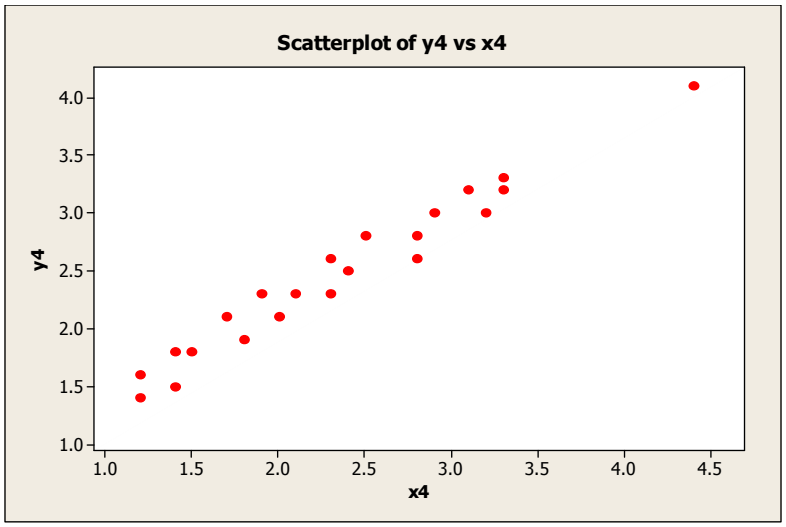
Outlying points can strongly influence a regression. Even a single point far from the body of the data can dominate the analysis.

High Leverage point p236

A data point that has an x-value far from the mean of the x-values is called a high leverage point.

Examples



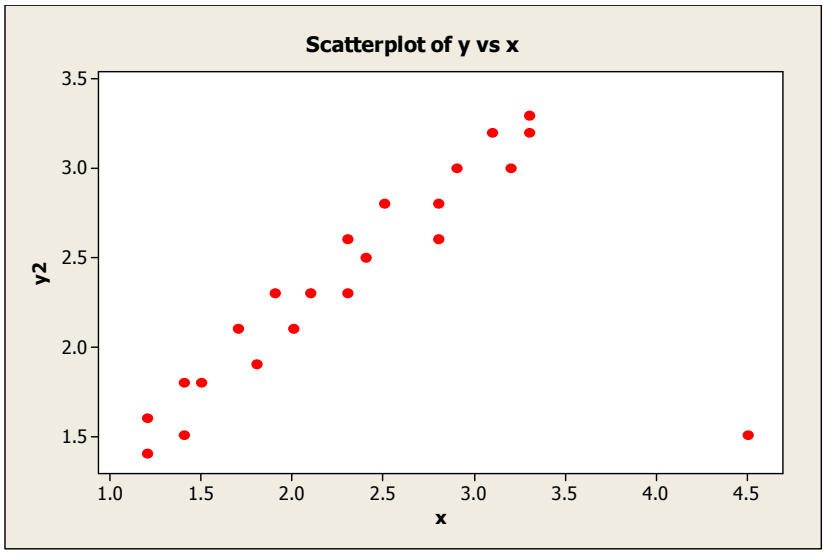


Influential observations

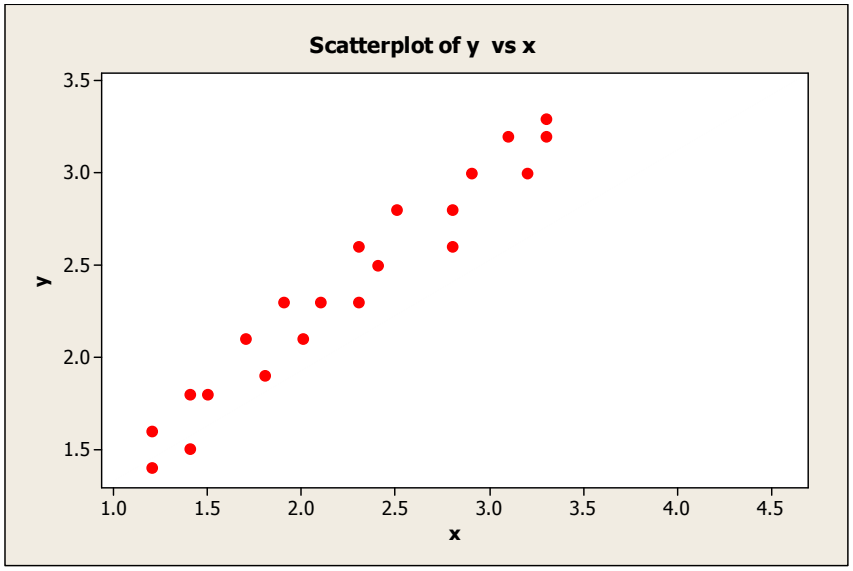
A data point is influential if omitting it from the analysis gives a very different model.

P 236

Example

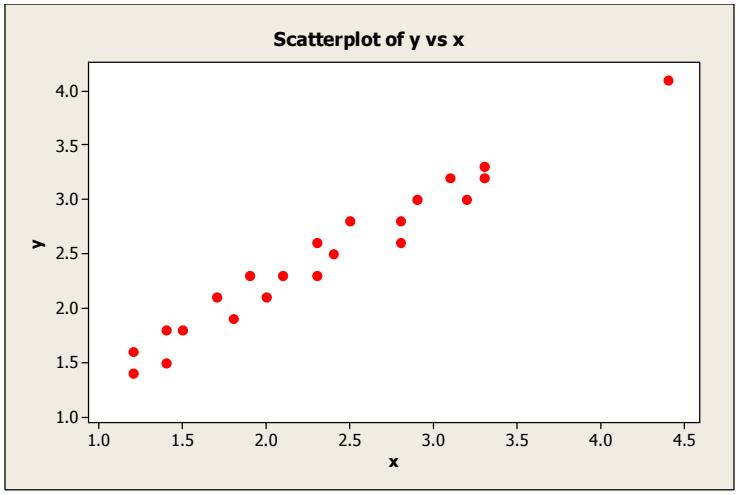


y = 1.38 + 0.414 x,
R-Sq = 33.2%

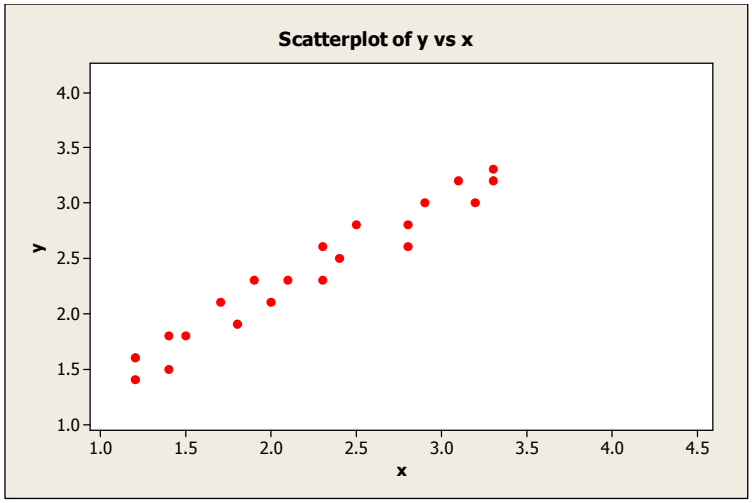


y = 0.567 + 0.811 x
R-Sq = 94.8%

Example (A high leverage point that is not influential) p236



y = 0.577 + 0.806 x
R-Sq = 96.3%



y = 0.567 + 0.811 x
R-Sq = 94.8%

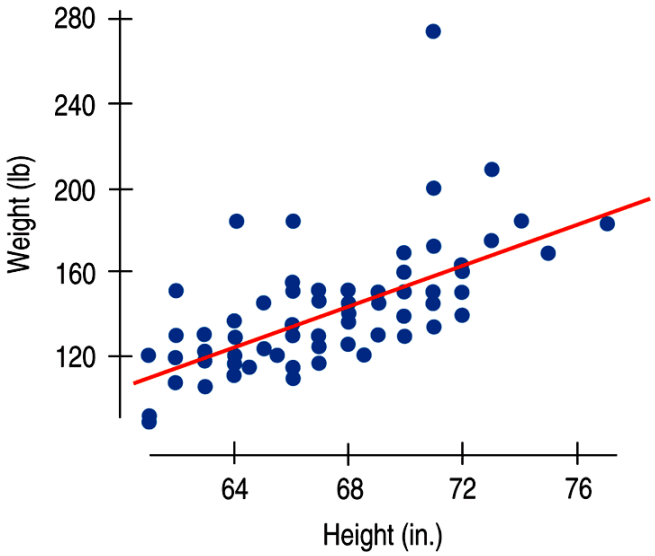
(Note R=sq is a bit less)

Restricted-range problem p242

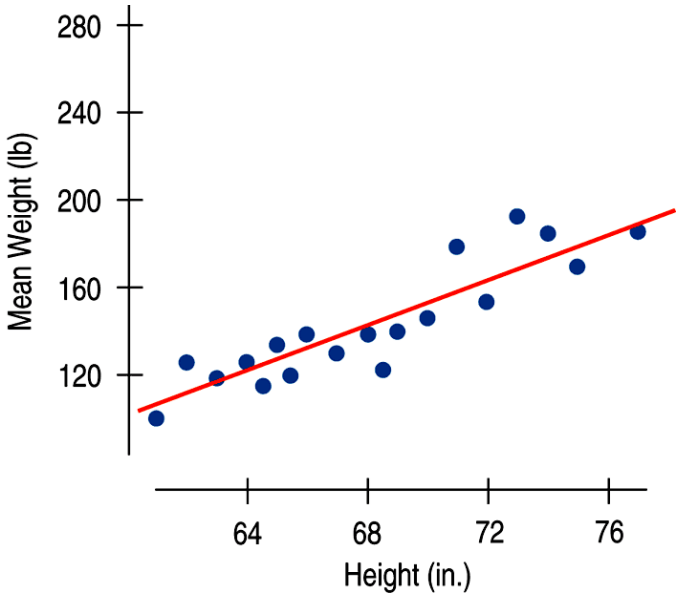
When one of the variables is restricted (you only look at some of the values), the correlation can be surprisingly low.

Working with summary statistics p242

There is a strong, positive, linear association between *weight* (in pounds) and *height* (in inches) for men



If instead of data on individuals we only had the mean weight for each height value, we would see an even stronger association



Chapter 10: Re-expressing data (Transformations) – Get it Straight! P263

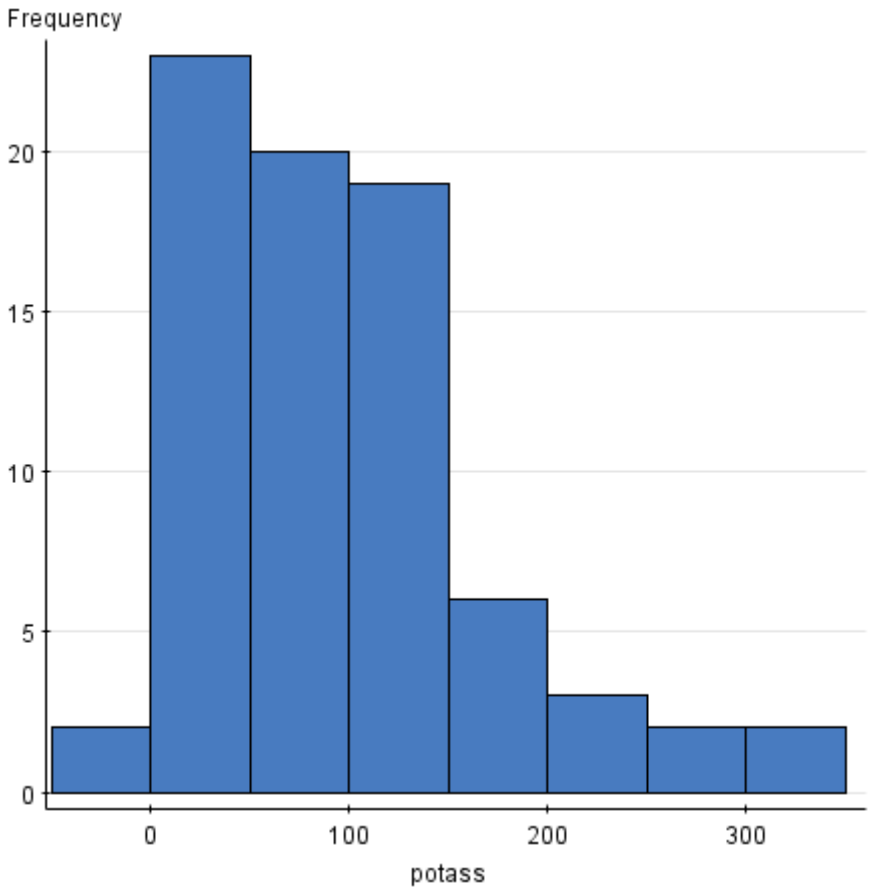
- Take a simple function (a transformation) of the data to achieve:
  - make the distribution more symmetric
  - make spreads of several groups more similar
  - make a scatterplot more linear
  - make spread in a scatterplot same all the way along

Example Cereal data (Potassium)

What would you like to fix?

Decimal point is 2 digit(s) to the right of the colon.

0 : 002233333333444444444
0 : 55555666666778999999
1 : 00000001111111222233444
1 : 667799
2 : 034
2 : 68
3 : 23

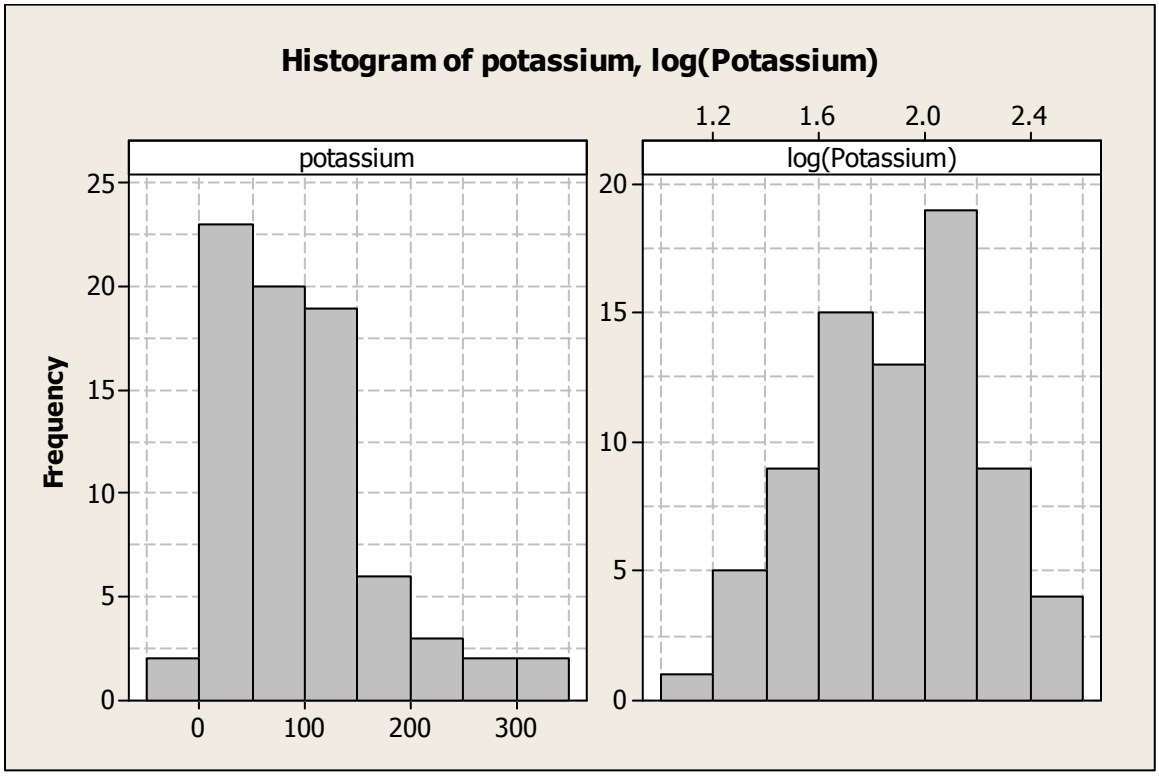


Try log of potassium:

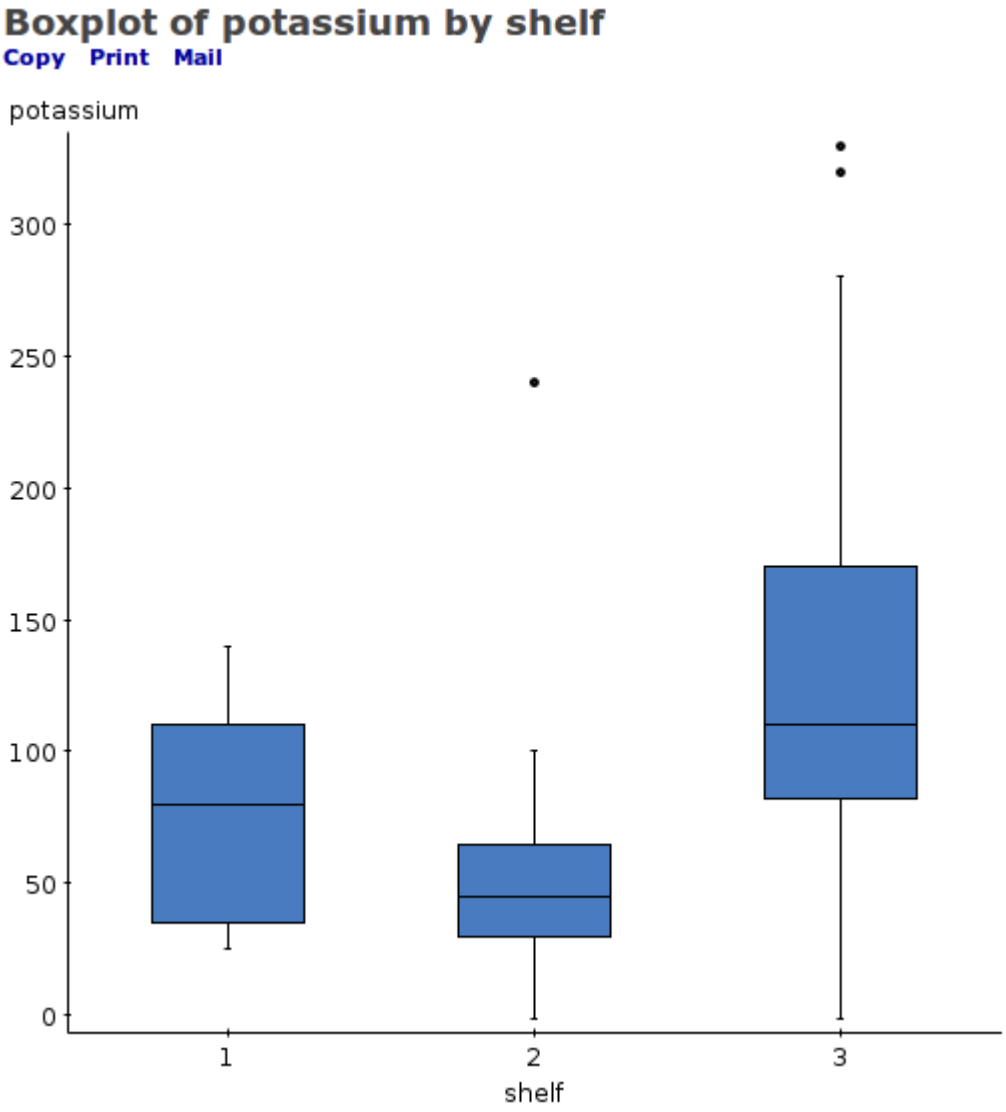
Variable: log(potass)

Decimal point is at the colon.

2 : 7
3 : 022224444
3 : 66666777788889
4 : 0001112244
4 : 55555666666677777777788889999
5 : 11112234
5 : 56688

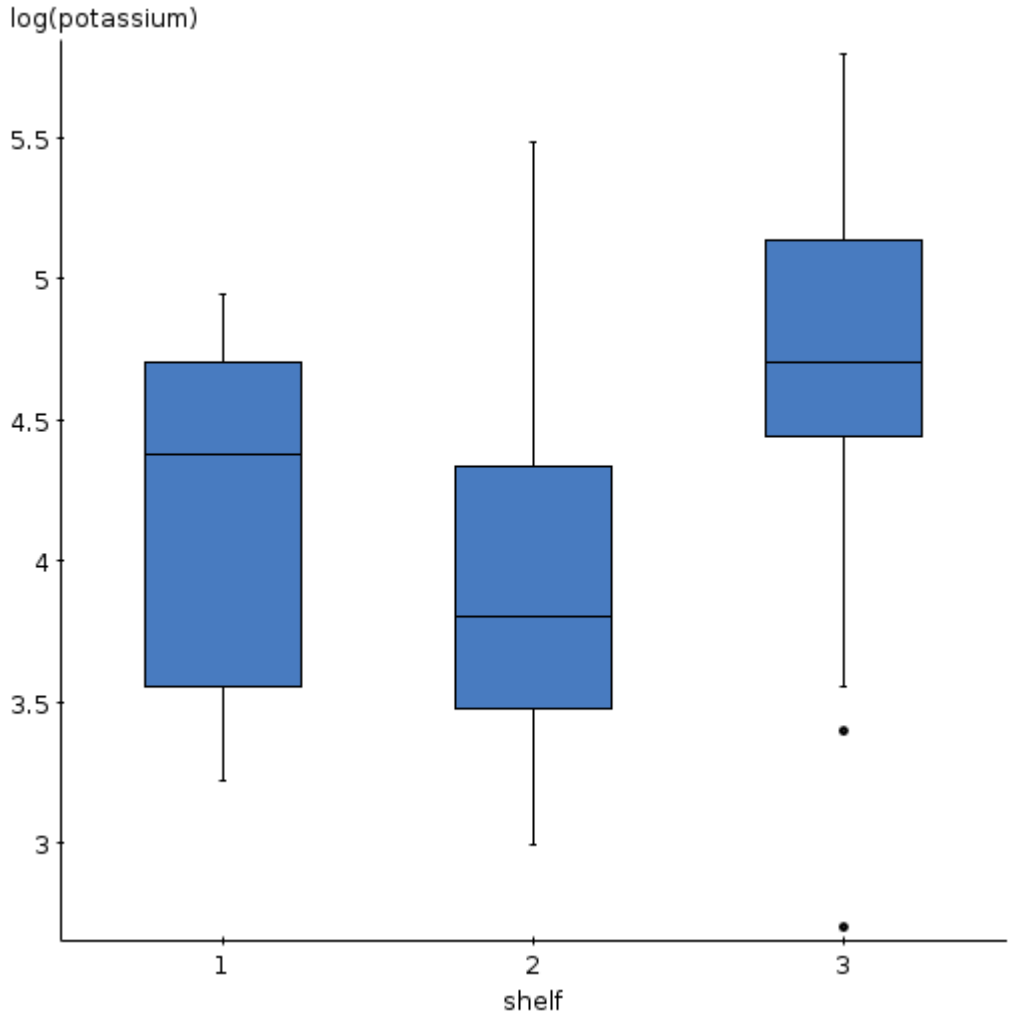


What about boxplots of potassium by shelf?



How does spread compare as centre gets larger?

Try using log of potassium values:



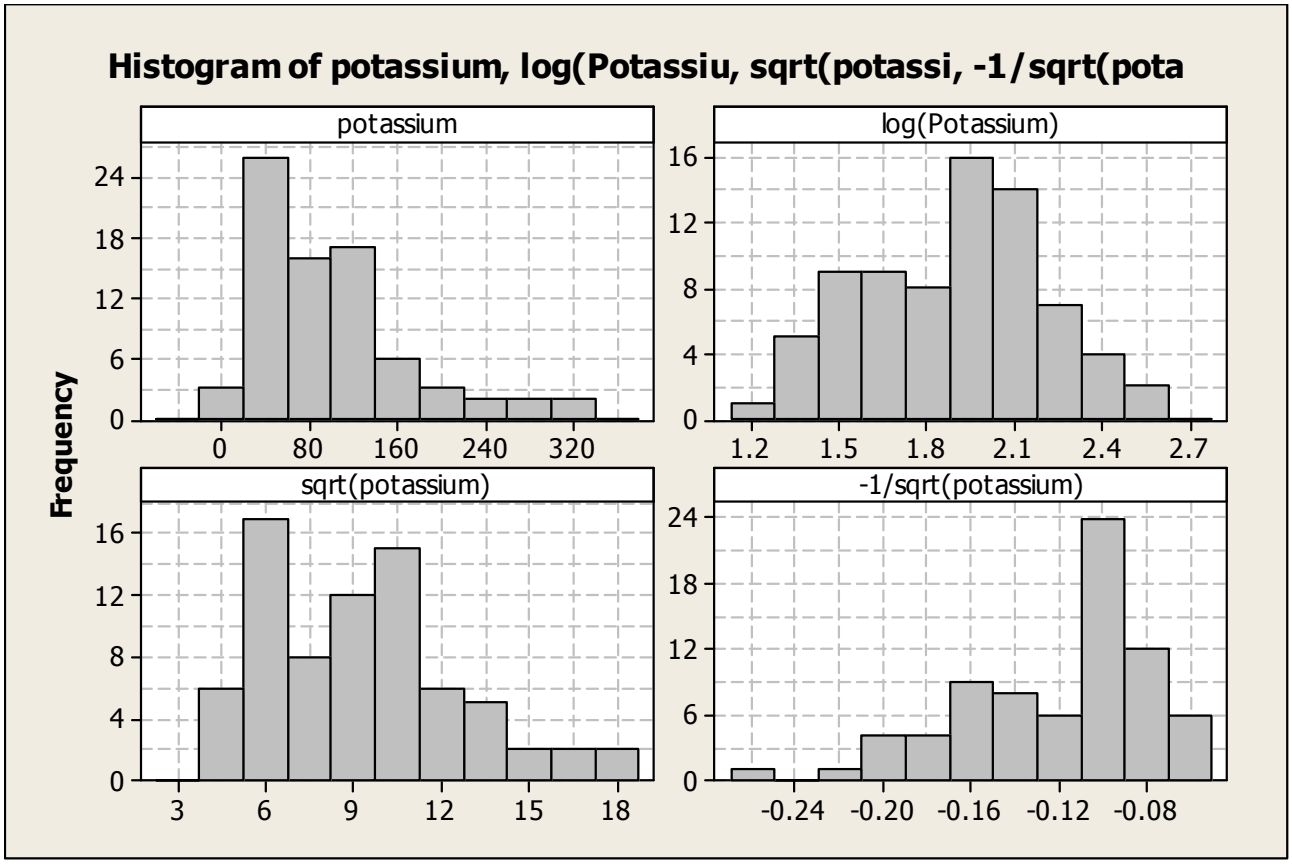
Spreads are more equal now, less dependent on centre.

Ladder of powers p269

Power	Name	Notes
2	Square of values	Left skewed
1	Unchanged data	
0.5	Square root	Counts
0	Logarithm	% change increases such as salaries. When in doubt start here
-0.5	-1/square root	Rare, but sometimes useful
-1	+/-1/data	Ratio “wrong way up” (eg hours per km)

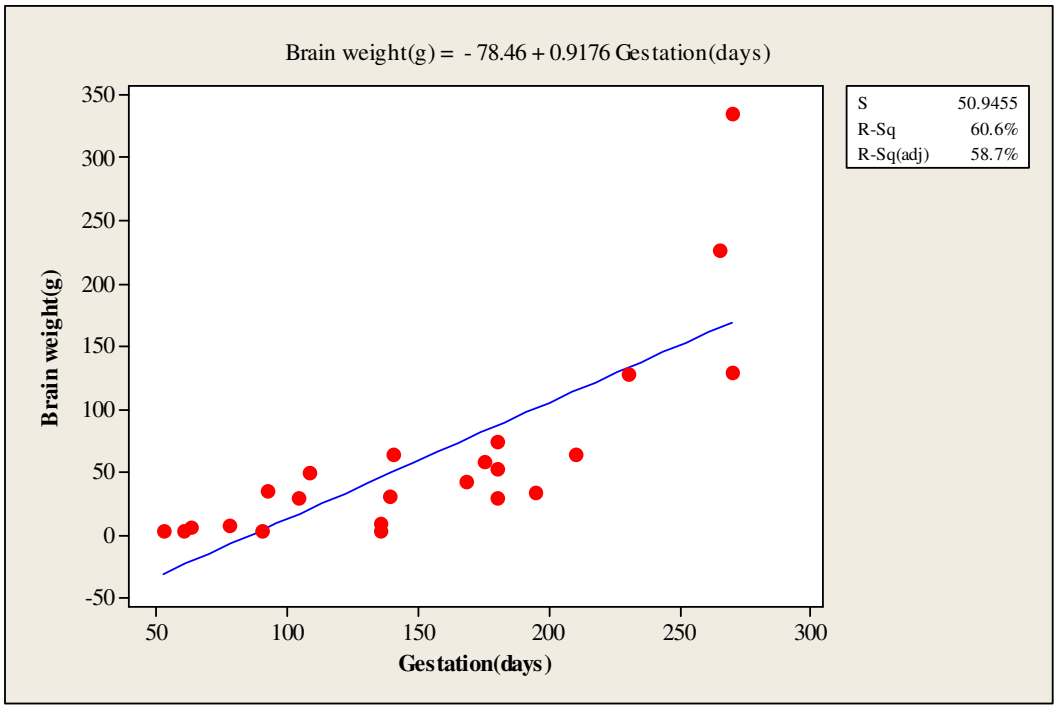
Cereal potassium data: log was good, but can we do better? Look at histograms.

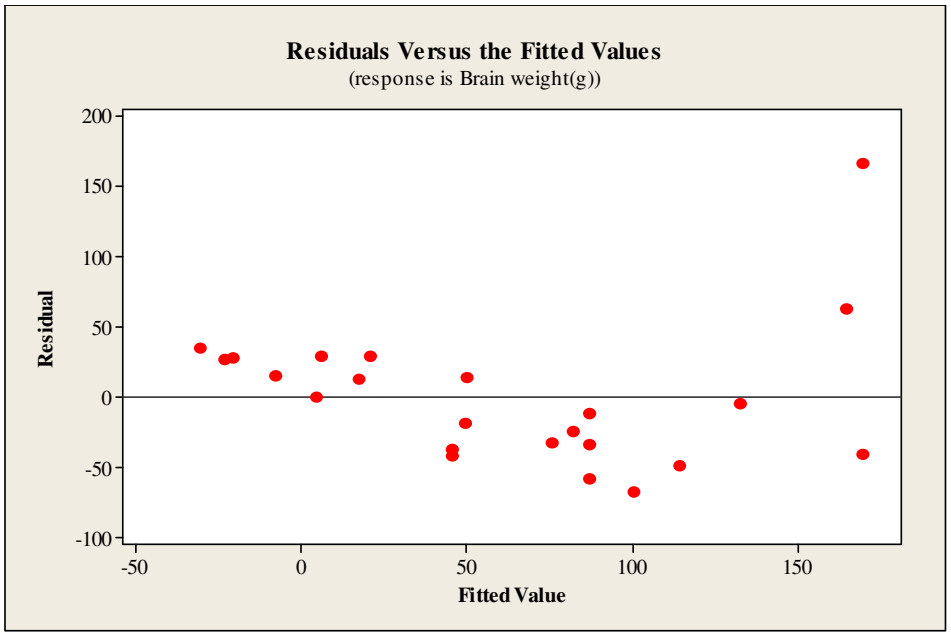
Where on the ladder of powers should we be to make the shape symmetric?



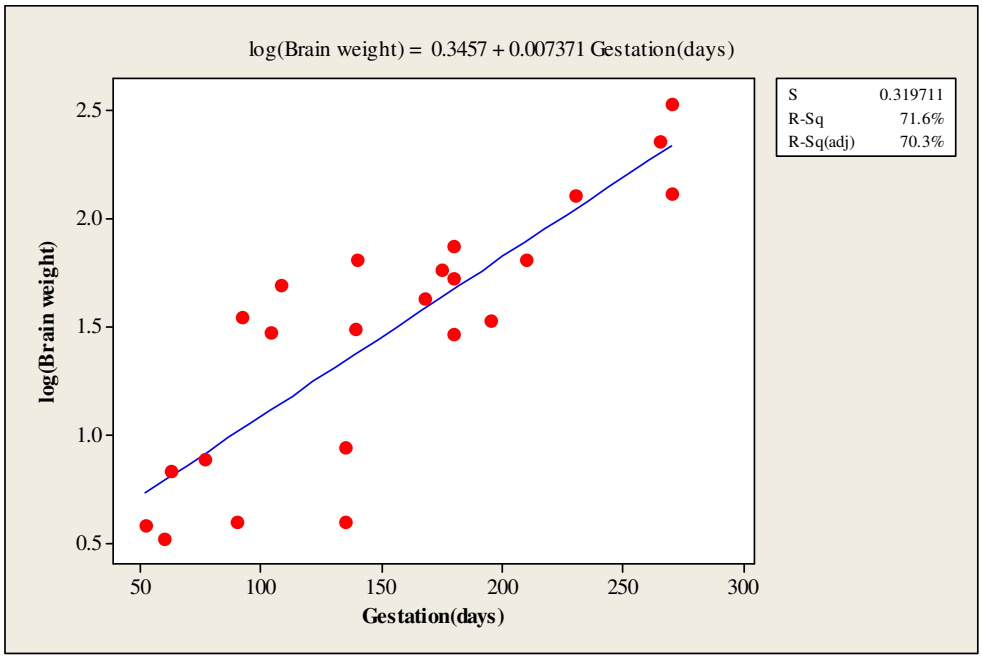
Example (regression)

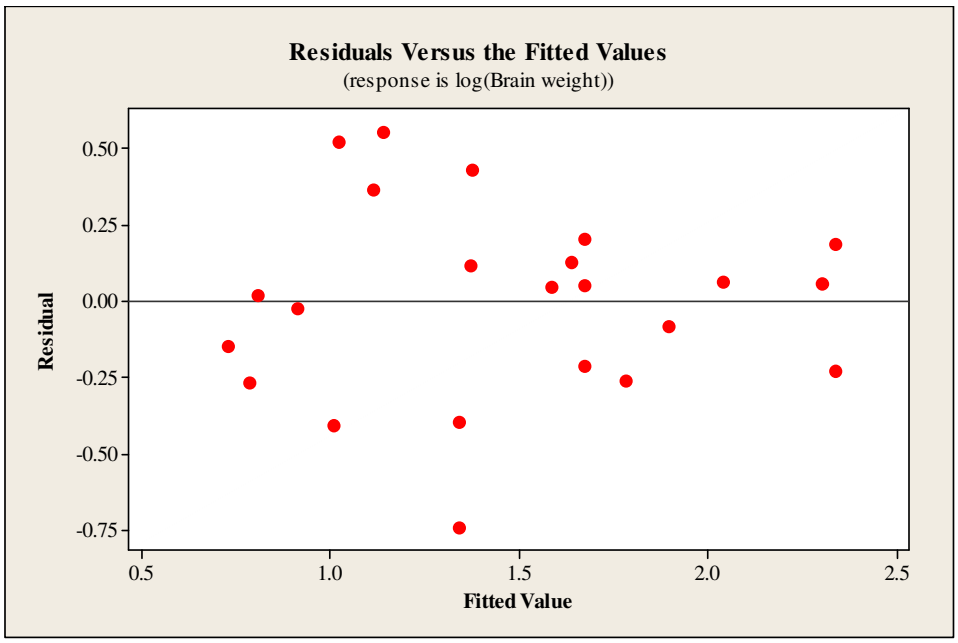
neonatal brain weight vs gestation (Ex 35p 286)





Try log transformation





The predicted log brain weight for a 200-day gestation period is $0.346 + 0.00737 \times 200 = 1.82$ and so the predicted brain weight is $10^{1.82} = 66.1$g.