

Transformation of Variables

Supplement to Section 7.8

Brian Habing – University of South Carolina

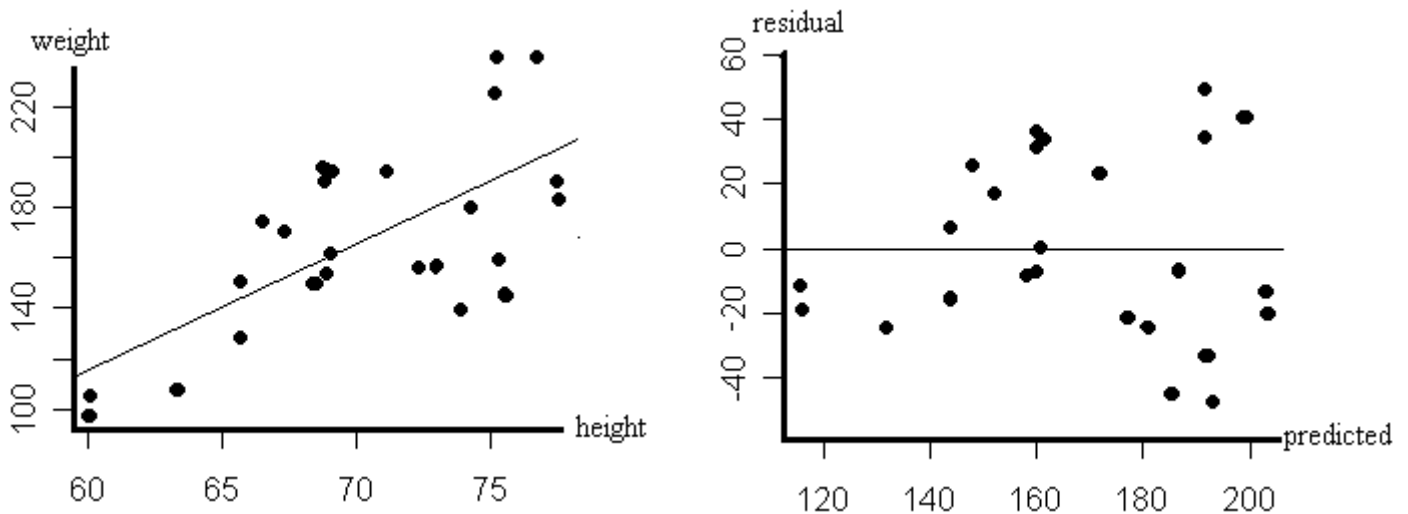
Last Updated: July 11, 2004

The great advantage of the simple linear regression is that it is very straightforward in terms of mathematics and interpretation. Unfortunately, even though you can always use the linear regression formulas to estimate the regression line and construct the ANOVA table, it won't necessarily be valid. This is because the regression assumptions might not be met. Below we deal with two particular violations: the residual versus predicted plot is fan-shaped because the residuals do not have a constant variance, and the residual versus predicted plot has a curved shape because a linear form is not appropriate.

Variance Stabilizing Transformations – $\log(y)$ or \sqrt{y}

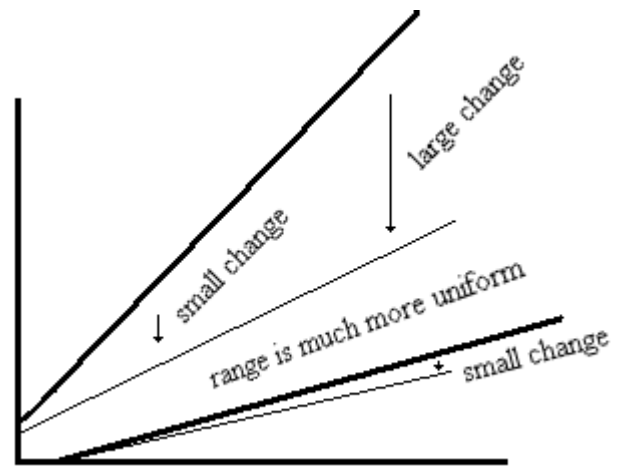
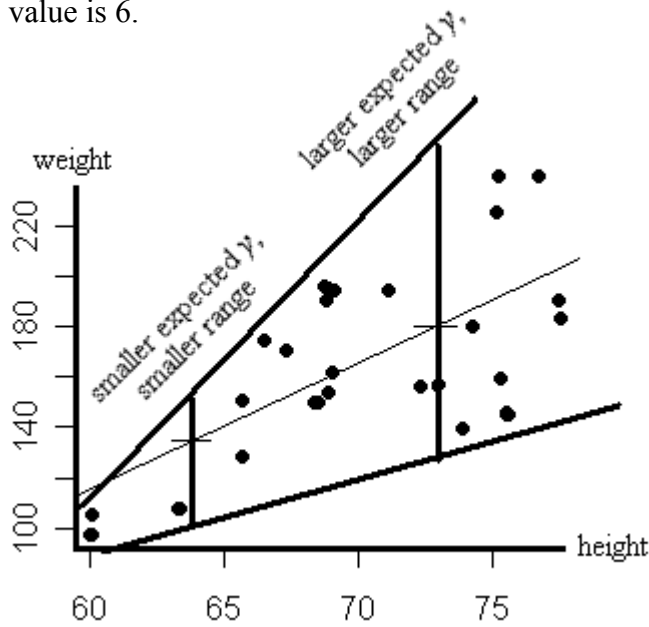
Figure 7.12 on page 322 is an example of the case where the assumption of equal variances does not hold in a very specific way: the variances of the errors increases with increasing predicted y . This is actually a fairly common occurrence. Consider height and weight; infants are expected to be shorter than professional football players and they also have a much smaller variation in their weights. Similarly, consider the variability in prices when comparison shopping; you are likely to find a much smaller range of prices when shopping for groceries than you would for cars.

The graphs below are an example of heights (in inches) and weights (in pounds) for a sample of adults.

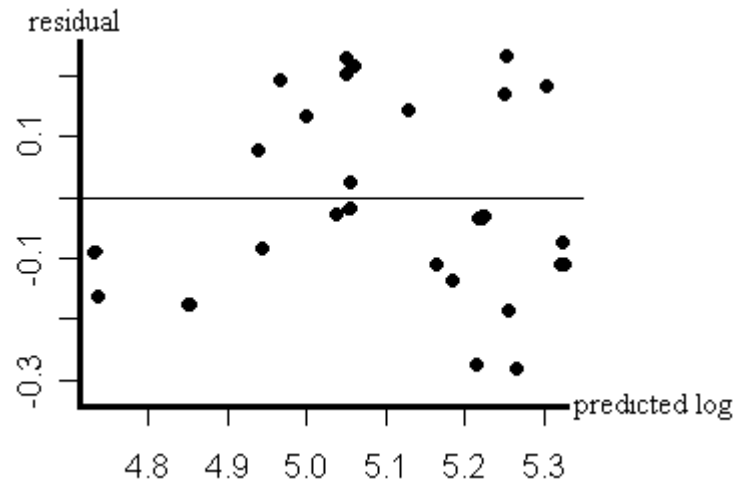


Notice that the residual plot makes a fan shape (<) opening to the right, where the taller people (who are predicted to be heavier) have a wider range of estimated errors.

What is needed is a transformation that shrinks values of y in such a way that large values of y are affected much more than small values are. Two functions that have this effect are the square root and the natural logarithm. For example, consider the values $y = 1, 4, 9$ and their square roots $= 1, 2, 3$. Notice that the change for the small y value is 0, the change for the middle y -value is 2, and the change for the largest y -value is 6.



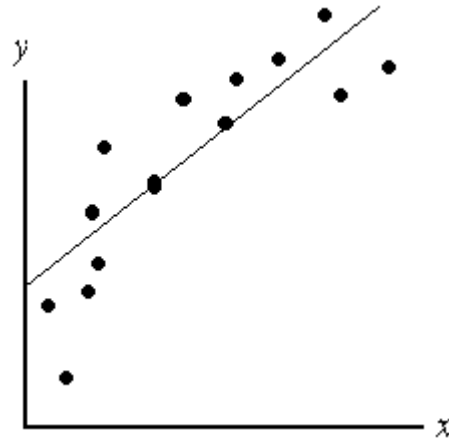
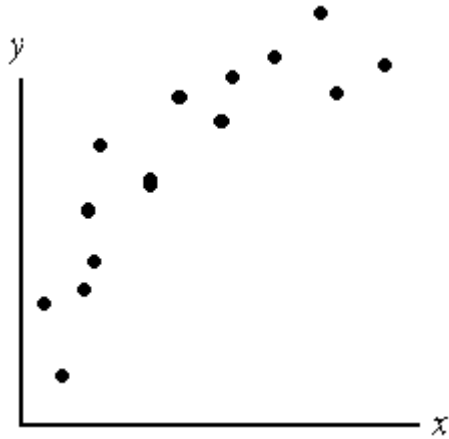
While $\log(y)$ or square root of y often corrects this problem with the residuals, it doesn't always work as well as we would like. That is the case in this example after we take $\log(\text{weight})$. The residual plot still has a fan shape, however it is not as severe as before. The relative spread for the first two points is double what it was previously, the top of the middle and right end are now even, and the bottom of the right and left are much more similar.



Still, in this case a second transformation might be needed both to finish removing the change in error variances as well as to remove the up-side down U shape in the residual plot.

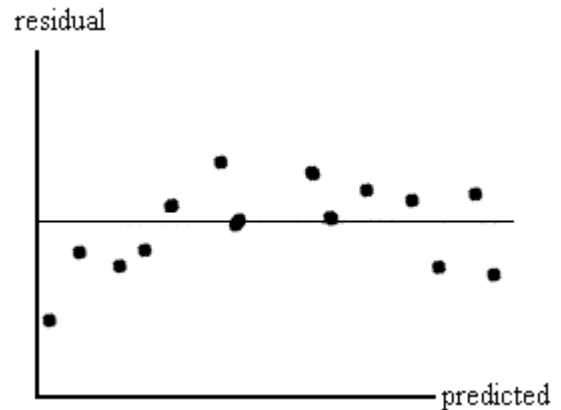
Transformations for Nonlinearity

Figures 7.11 on page 321 and 7.14 on page 323 both show examples of “specification error” where the errors do not have zero mean because a line is not appropriate. This is not always immediately obvious from the scatter plot, even if the regression line is drawn on. Consider the example below:

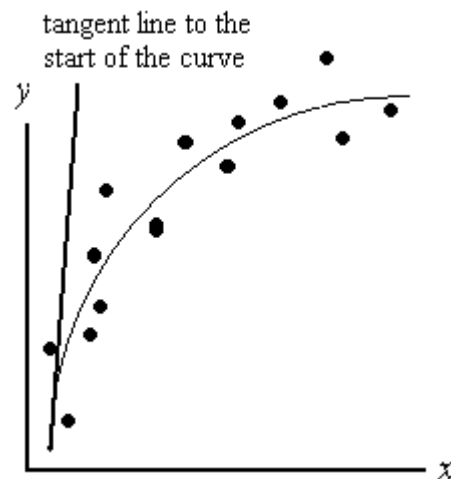
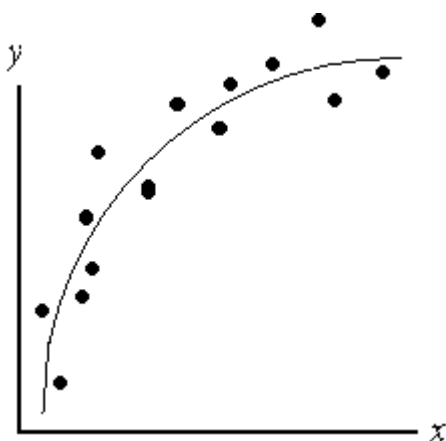


It is often much easier to see the pattern in the points when observing the residual versus predicted plot. This is because you no longer have to focus on the linear trend in the data, and can instead focus on simply the lack of a horizontal pattern.

Here we can see that the residuals have a \cap pattern, with the mean of the residuals being negative, then positive, and finally slightly negative again.

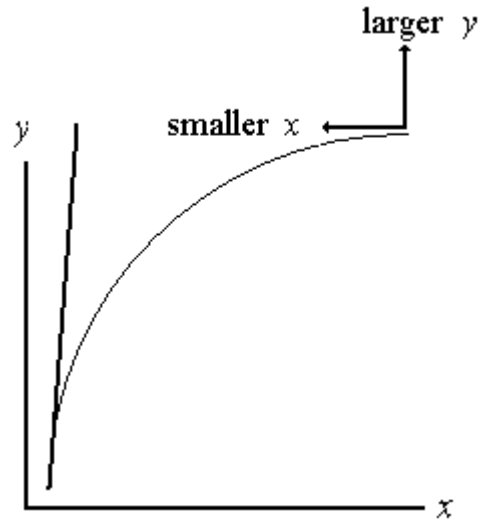


If we try to imagine a curve going through the scatter-plot of the data, it would look an arc from a circle, when what we really want is a straight line.

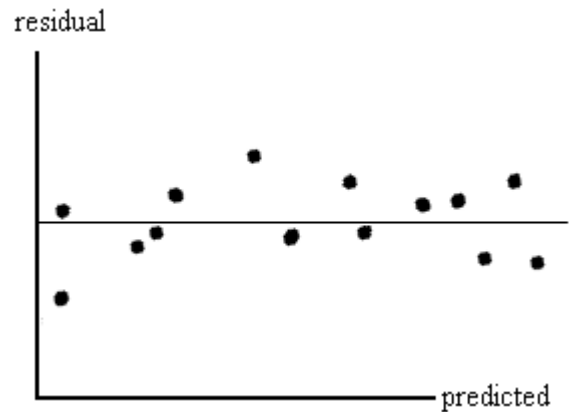
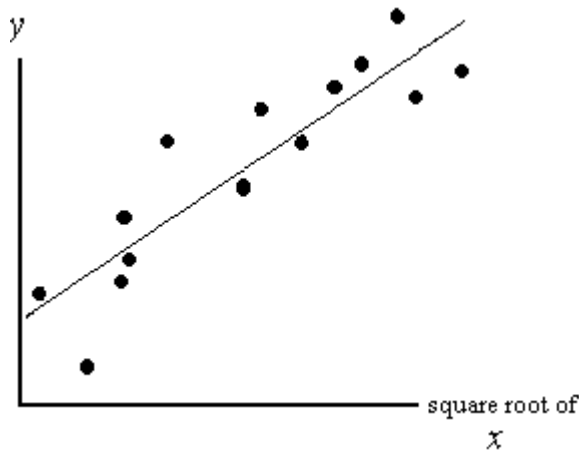


One way of making the observed curve more like the tangent line would be to shrink the x values in such a way that the larger the x -value the larger the shrinkage. Functions that do this include the square root of x , the logarithm of x , and $-1/x$.

On the other hand, y could be expanded in such a way that the larger the y -value the larger the expansion. Examples of functions that do this include y^2 , y^3 , and exponent of y .



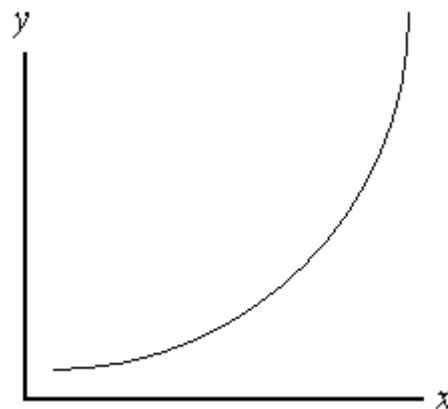
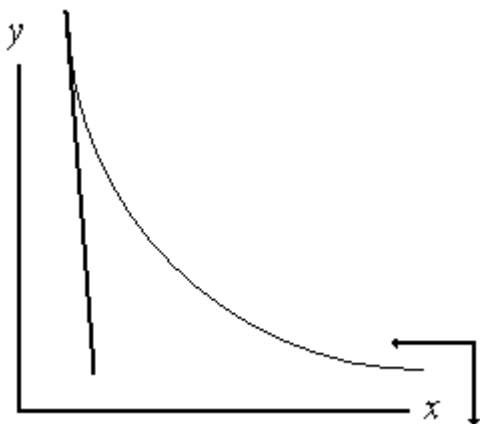
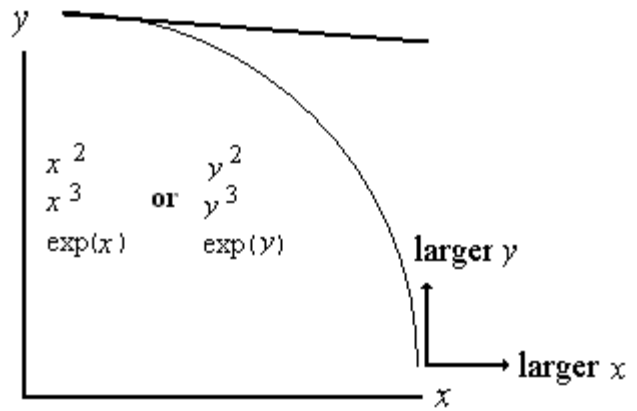
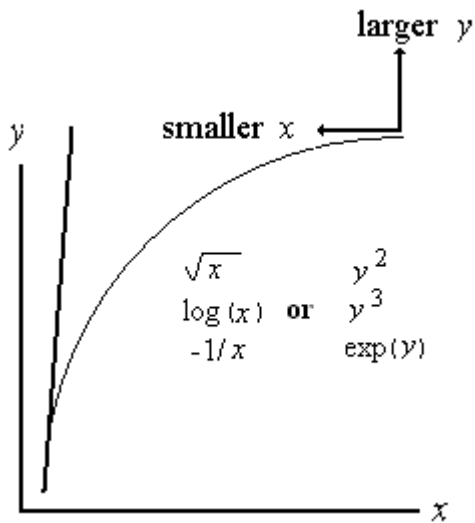
Returning to the data on the previous page, the square root of x gives a much-improved residual versus predicted plot (using y^2 would also have greatly improved the plot).



The general method for seeing which transformation to try is to draw the tangent line to the curve that we saw above, and to see what types of transformations would change the data appropriately. The figure on the next page demonstrates the basic idea for two of the possible curve shapes, and it would be good practice for you to fill in the appropriate transformations for the other two.

It is also important to note that these transformations will not always be successful. In those cases a statistical consultant may be able to apply results like the polynomial model in section 8.6 (pg. 370) that are beyond the scope of this course.

The following table will show the appropriate transformation for four basic situations. You should be able to complete the bottom half of the diagram.



Interpretation

One complication in using the above transformations is that you are no longer predicting y from x , and it might not be clear what it means to have a regression predicting the logarithm of weight, for example (what does a $\log(\text{pound})$ mean?). In such cases it is important to remember to undo the transformation to get a predicted value, or prediction interval (if the prediction is 5.1929 $\log(\text{pounds})$ then the exponent gives a prediction of 180 pounds). It is also important to remember that r^2 does not give the percentage of variation of y explained by x it gives the percentages for the transformed variables. Unlike predictions, there is no easy transformation for the correlation coefficient.

One case that is fairly easy to interpret is when both the logarithm of y and x are taken. In this case the model becomes multiplicative as discussed on pages 374-377 in the context of multiple regression.

In the context of the height and weight example on pages 1-2 of this supplement, if we take the logarithm of both height and weight the estimated regression equation is:

$$\log(\text{weight}) = -4.792 + 2.326 \log(\text{height}) + \text{error}$$

Taking the exponent of both sides of the equation gives:

$$\text{weight} = e^{-4.792} e^{2.326 \log(\text{height})} e^{\text{error}}$$

which is:

$$\text{weight} = 0.008 \text{ height}^{2.326} e^{\text{error}}$$

At first this may seem out of nowhere, but note that the formula for BMI relates weight to height-squared! By taking the logarithm of both x and y we could thus use the methods in section 7.5 to perform a test of hypotheses that the exponent is indeed equal to 2, or to form a confidence interval for it.