Sections 3.9 and 6.8: Transformations

Adapted from Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Transformations of variables (Section 3.9 & p. 236)

- Some violations of our model assumptions may be fixed by transforming one or more predictors $x_1, \ldots, x_k$ or Y.
- If the *only* problem is a nonlinear relationship between Y and the predictors, i.e. constant variance seems okay, a transformation of one or more of the $x_1, \ldots, x_k$ is preferred.
- If non-constant variance appears in one or more plots of Y versus the predictors, a transformation in Y can help...or make it worse!
- *Data analysis is an art.* The best way to learn how to analyze data is to analyze data.
- A nonlinear relationship *could* manifest itself the scatterplot matrix of Y_i versus x_{ij} for $j = 1, \ldots, k$, or the residuals e_i versus x_{ij} from an initial fit.
- The chosen transformation should roughly mimic the relationship seen in the plot.

Transformations for $x_{i1}, \ldots, x_{ik}$

Examples of transformations for predictors are:

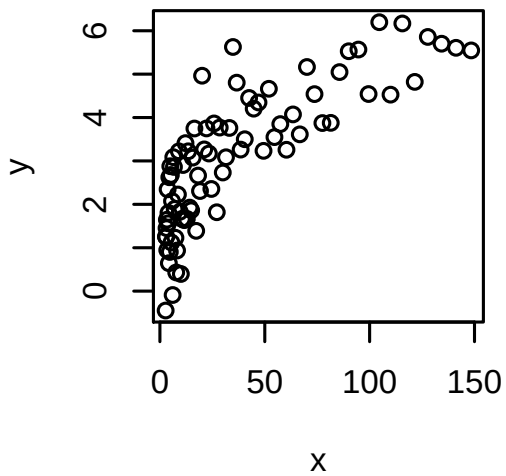
- $x^* = \log(x)$
- $x^* = \sqrt{x}$
- $x^* = 1/x$
- $x^* = \exp(x)$ or $x^* = \exp(-x)$

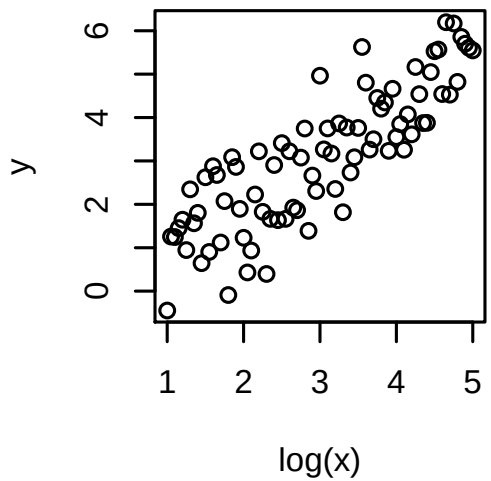
See Figure 3.13, page 130.

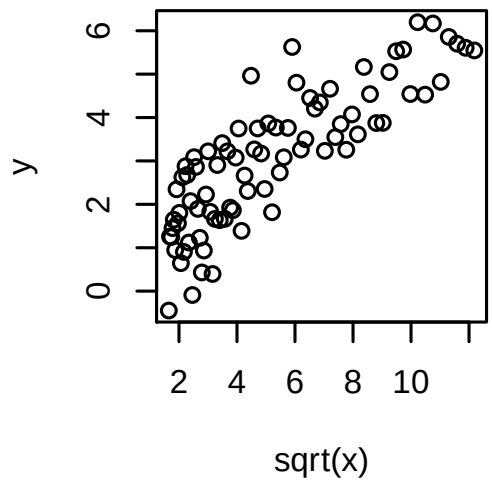
For how many of these do we implicitly assume $x > 0$?

We will examine *marginal* relationships and transformation "fixes." For multiple regression these might better be residual plots versus predictors, or better yet added variable plots.

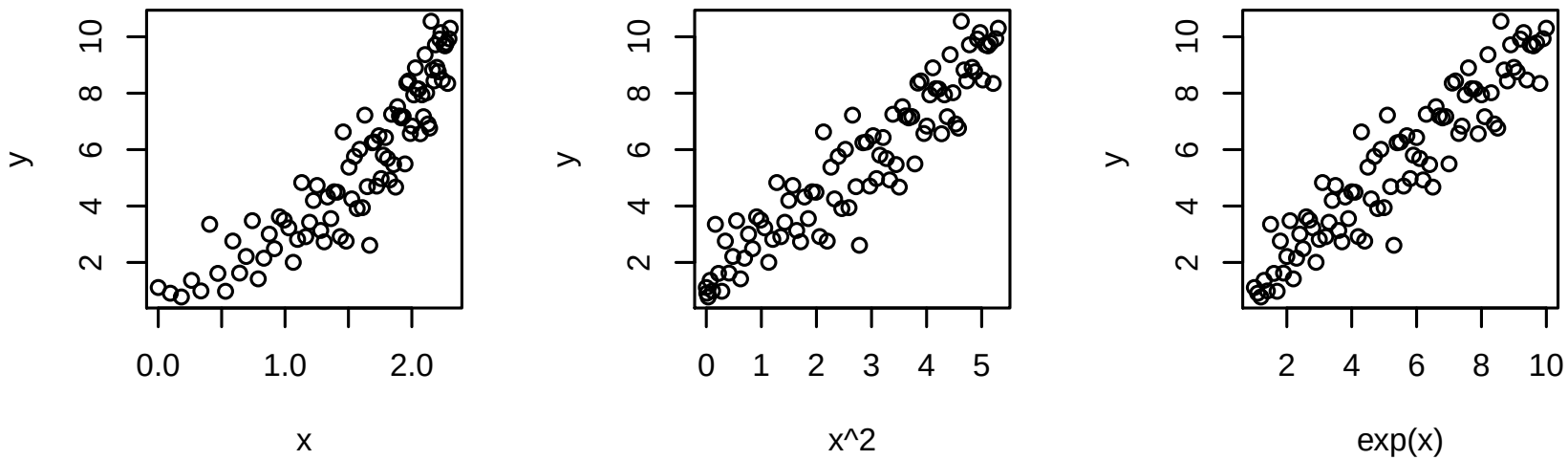
Example 1: transforming a predictor



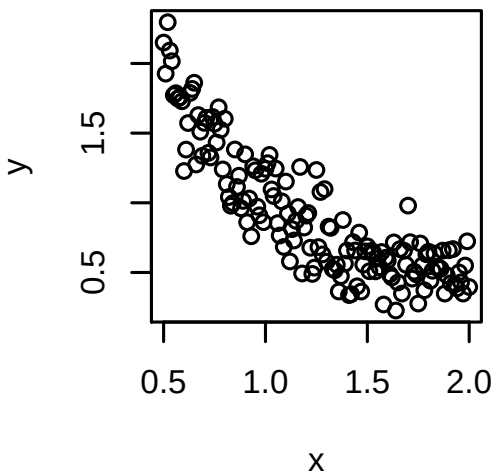


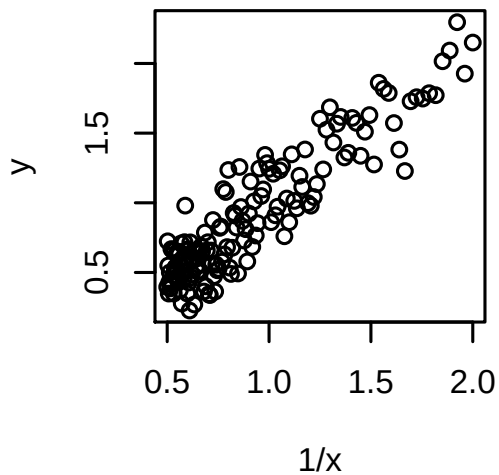


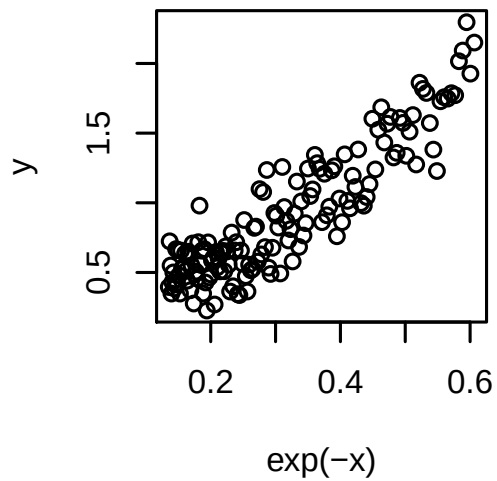
Example 2: transforming a predictor



Example 3: transforming a predictor







Transforming the response

If there is evidence of nonconstant error variance, a transformation of Y can often fix things. Examples include:

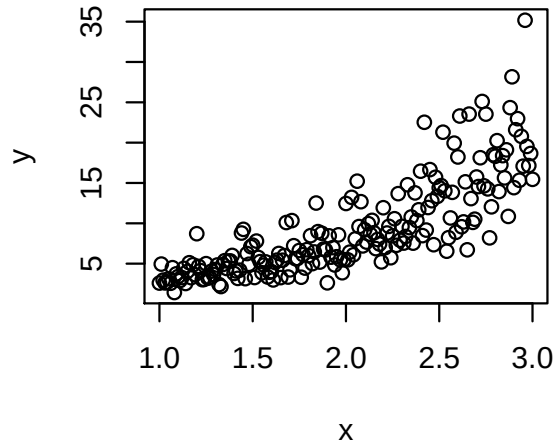
- $Y^* = \log(Y)$
- $Y^* = \sqrt{Y}$
- $Y^* = 1/Y$

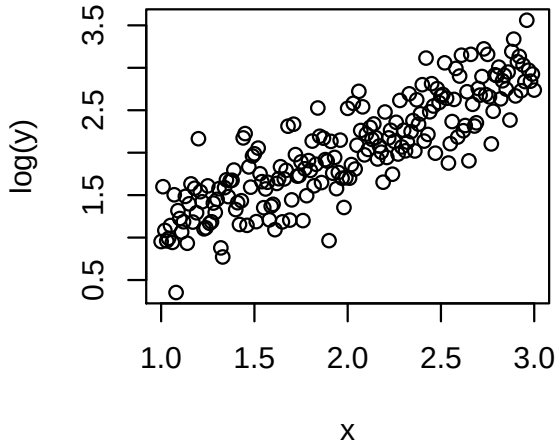
See Figure 3.15, page 132.

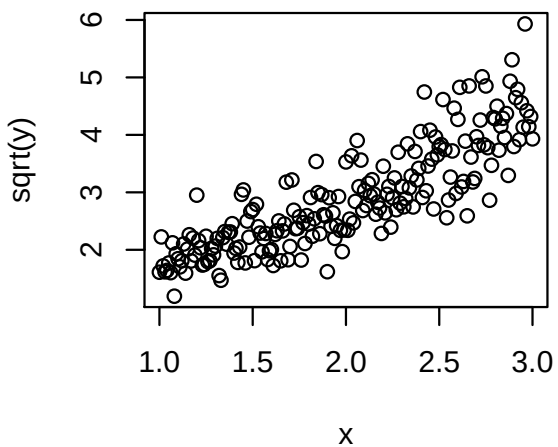
All of these are included in the Box-Cox family of transformations.

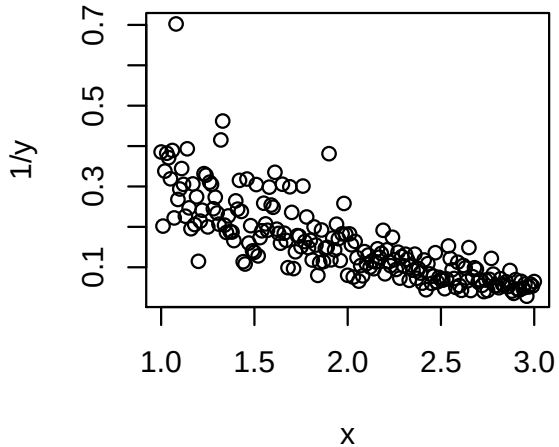
For some data, a transformation in Y may be followed by one or more transformations in the $x_{i1}, \ldots, x_{ik}$.

Example 4: transforming the response

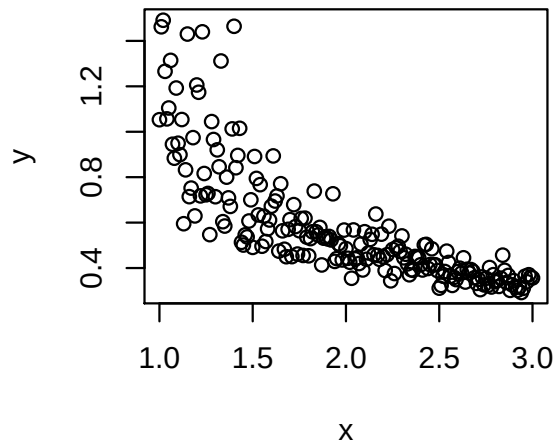


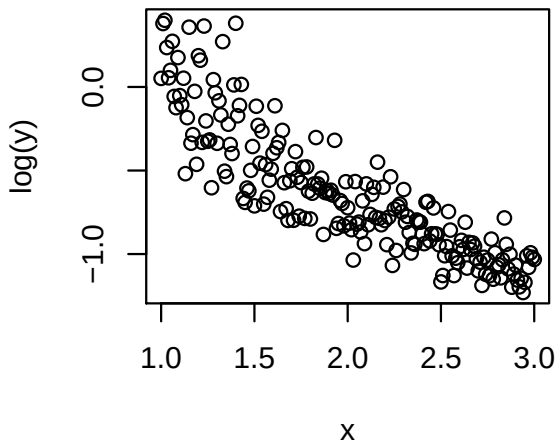


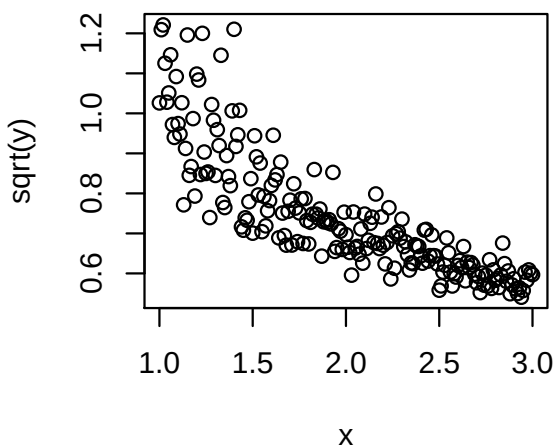


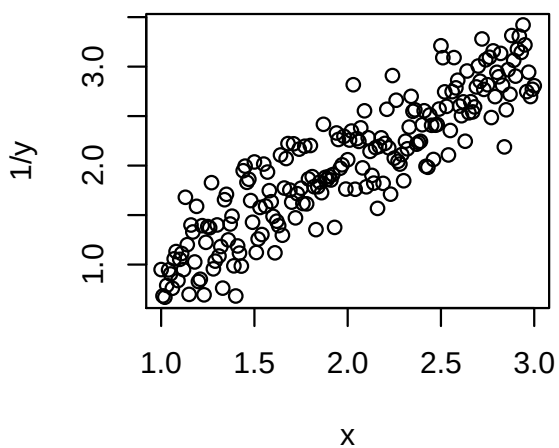


Example 5: transforming the response









Box-Cox transformations

Box-Cox transformations are of the type

$$Y^* = Y^\lambda$$

where λ is estimated from the data, typically $-3 \leq \lambda \leq 3$. These include

$\lambda = 2$	$Y^* = Y^2$	
$\lambda = 1$	$Y^* = Y$	no transformation!
$\lambda = 0$	$Y^* = \log(Y)$	by definition
$\lambda = -1$	$Y^* = 1/Y$	reciprocal
$\lambda = -2$	$Y^* = 1/Y^2$	

SAS will help you pick λ automatically in proc transreg. R uses `boxcox()` in the **MASS** package.

Interpretation changes with transformed data

Note: When working with transformed data, predictions and interpretations of regression coefficients are all in terms of the *transformed variables*.

To state the conclusions in terms of the original variables, we need to do a reverse transformation...carefully.

Example: Electrical components

- Consider time-to-failure in minutes of $n = 50$ electrical components.
- Each component was manufactured using a ratio of two types of materials; this ratio was fixed at 0.1, 0.2, 0.3, 0.4, and 0.5.
- Ten components were observed to fail at each of these manufacturing ratios in a designed experiment.
- It is of interest to model the failure-time as a function of the ratio, to determine if a significant relationship exists, and if so to describe the relationship simply.

SAS code: Plot & model

```
proc sgscatter data=elec; plot time*ratio; run; * non-constant variance;

proc transreg data=elec; * gets Box-Cox analysis;
 model boxcox(time / convenient) = identity(ratio)/pboxcoxtable; run;
```

Multiple predictors are included with, e.g., `identity(ratio temperature)`

Transformed response and explanatory variable

Transform response ($\log(Y)$) to fix non-constant variance. Try a couple different transformations of X to address nonlinearity. Add

```
log_time=log(time);
inv_ratio=1/ratio;
exp_ratio=exp(-ratio);
```

to data step and plot again.

PROC REG output

The fitted regression line is $\widehat{\log(\text{time})} = 1.15322 + \frac{0.49738}{\text{ratio}}$.

For a ratio of $x_h = 0.25$ we get
$\widehat{\log(\text{time})} = 1.15322 + \frac{0.49738}{0.25} = 3.14274$.

Exponentiating both sides we get $\widehat{\text{time}} = e^{3.143} = 23.2$ minutes.

Question: Is this the estimated mean failure time for the population with ratio $x_h = 0.25$? Is it the estimated *median* time?

Transforming back...

Question: How about a prediction interval? How would you get one?

Question: Let $g(Y) \sim N(\mu, \sigma^2)$ for some $g(x)$ monotone (and so invertible) function. What is the median of Y?

Question: Let $P(a < g(Y_h) < b) = 0.95$ (prediction interval for new $g(Y_h)$). How do you get a prediction interval for Y_h?

Question: What can you say about *any* Box-Cox transformation of the (positive) response (e.g. log, square root, reciprocal)?

Example: Water Quality data

Water quality for Congaree River watershed

When predicting E. Coli as a function of Fecal coliform and Enterococci bacteria, we used "transform both sides" without a formal evaluation of the transformation.

What does a more systematic evaluation show?