Chapter 4: Transforming Data

Aims:

- make distributions (more) symmetric
- linearize scatterplot relationships (why?)
- stabilize (make equal) variances across multiple groups

4.1 The Family of Powers and Roots

Power transformations: $X \to X^p$

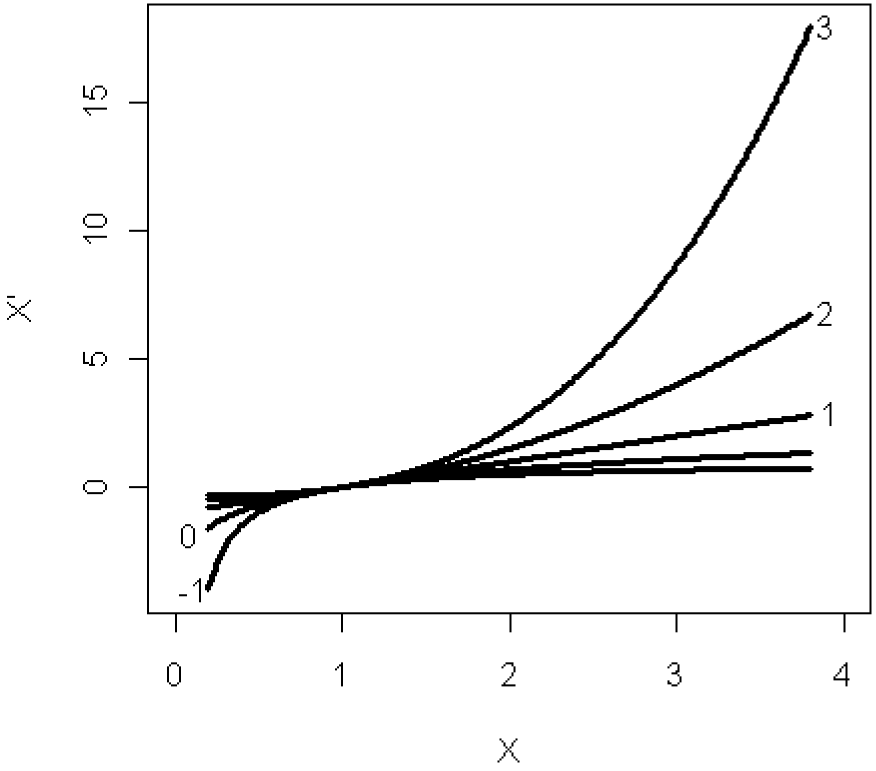
It turns out to be useful to have a family of power transformations defined slightly differently so that we can consider the limit of power transformations for $p \to 0$. We introduce the

Box-Cox family of transformations: $X \to X^{(p)} = \dfrac{X^p - 1}{p}$

Note that all transformations at the value $X = 1$

- have value 0
- have slope 1
- dividing by p preserves the direction of X when p is negative

Fig. 4.1 Family of power transformation $X^{(p)}$



Why do we only plot x>0?

What do we mean by $X^{(0)}$?

Fox calls this the "ladder of transformations"

*** $\lim\limits_{p\to\infty} X^{(p)} = \log_e X$ ***

[Note: Fox likes $\log_{10} X = \log_{10} e \times \log_e X$

In practice, the most common power transformations are:

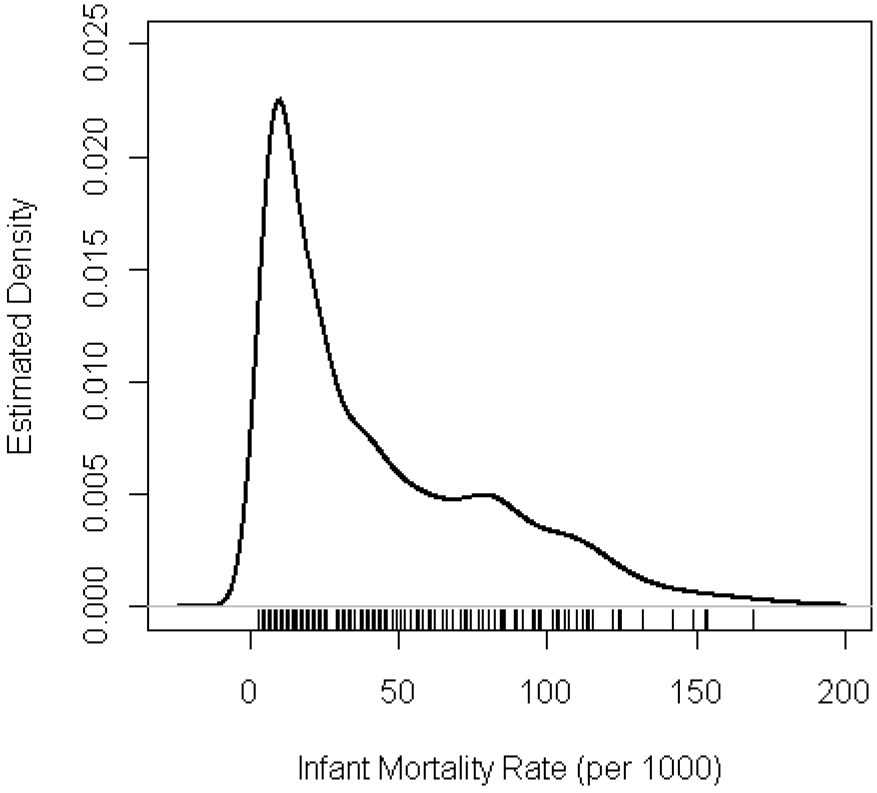
$p=-1$	$1/X$
$p=0$	$\log(X)$
$p=1/3$	$\sqrt[3]{X}$
$p=1/2$	$\sqrt{X}$

All of these deal with positively skewed distributions.

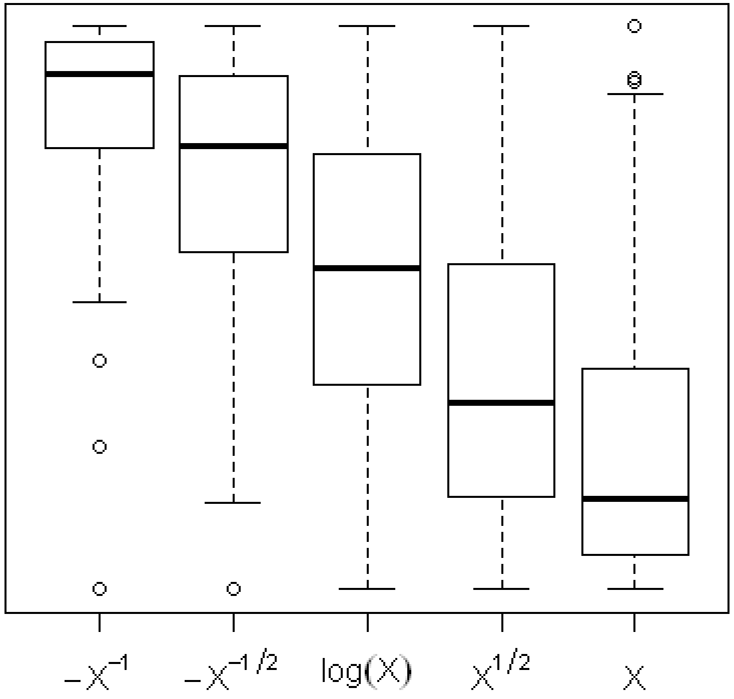
4.2 Transforming Skewness

The text discusses first transforming skewness (4.2), then transforming nonlinearity (4.3) and finally transforming nonconstant spread (4.4). The latter is most important, but we'll quickly review the figures for the first skewness and nonlinearity.

Fig 4.2 Infant mortality data (which we have already examined with histograms and qq-plots)



Figs. 4.3 and 4.4: Trial and error choice of transformation



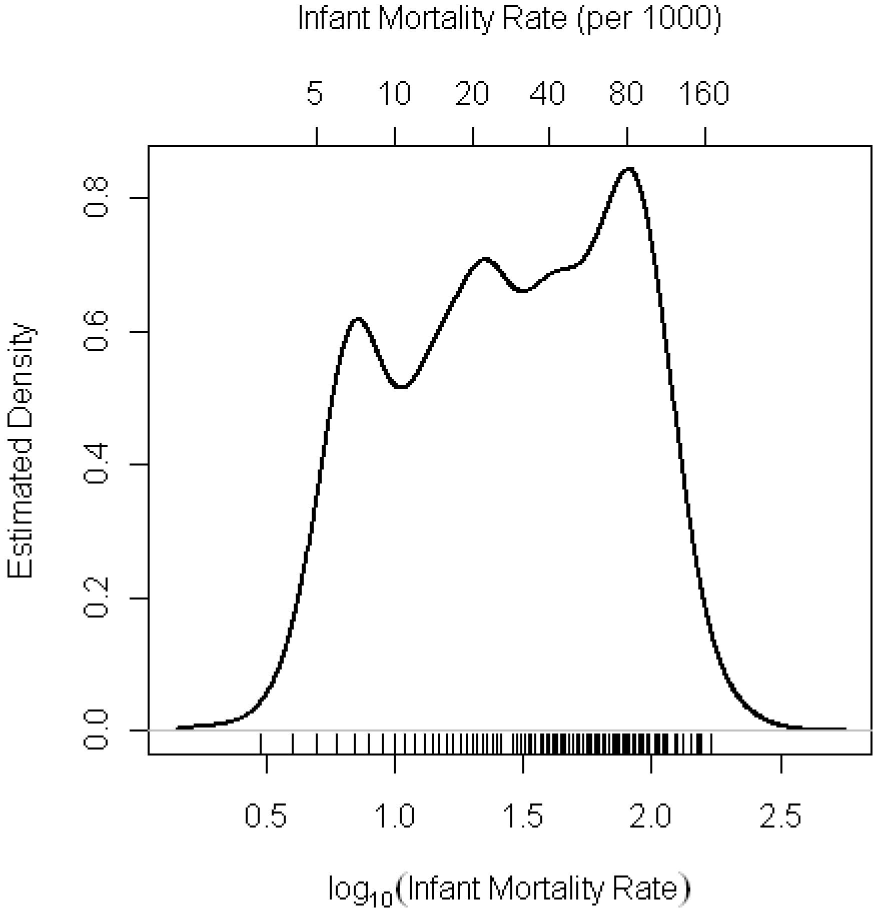


Fig. 4.3

Note that for Fig 4.4 we plotted the log of Infant Mortality rate, added a “rug” of data values below it, and added an axis of the original scale on top using the function `power.axis` from the `car` package. This is a cool thing to do, but I won’t expect you to figure out how to do it.

4.2 Transforming Nonlinearity

Why bother?

- Linear relationships, $\hat{Y} = A + BX$, are simple and easily interpreted
- When dealing with multple explanatory variables, it is quite difficult to deal multiple nonlinear relationships

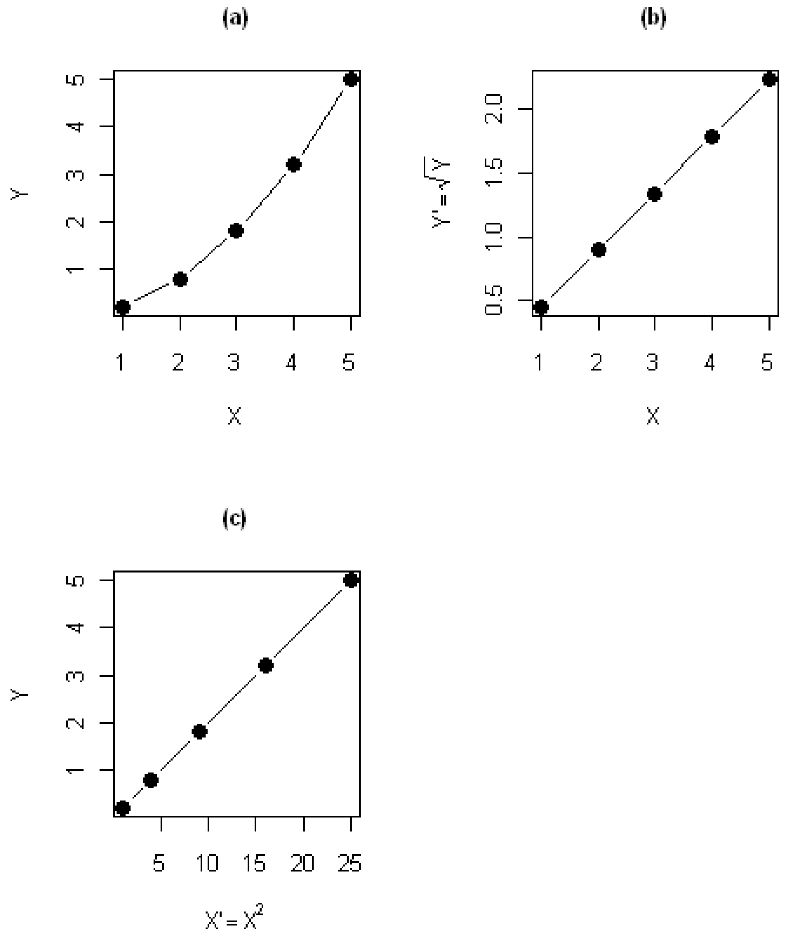
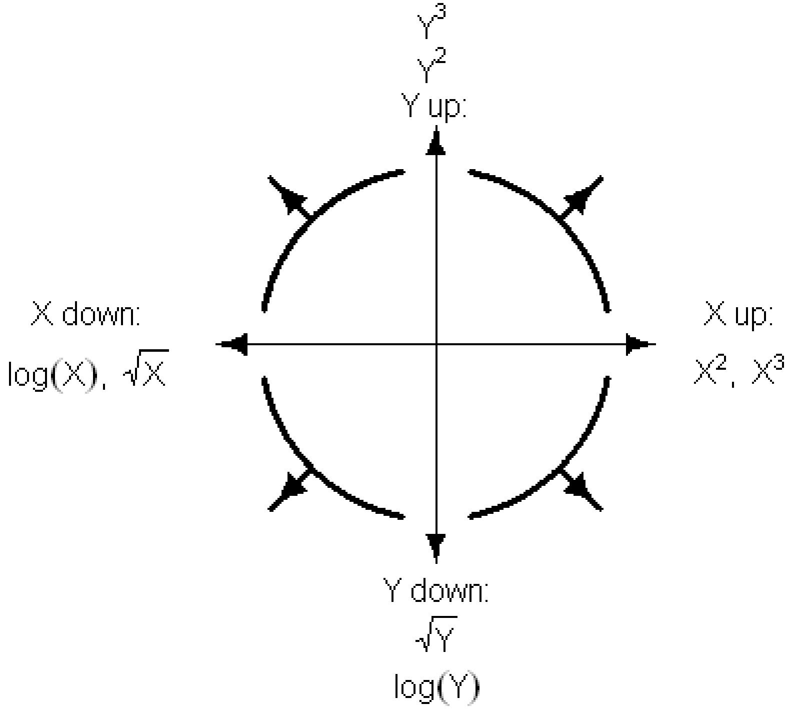
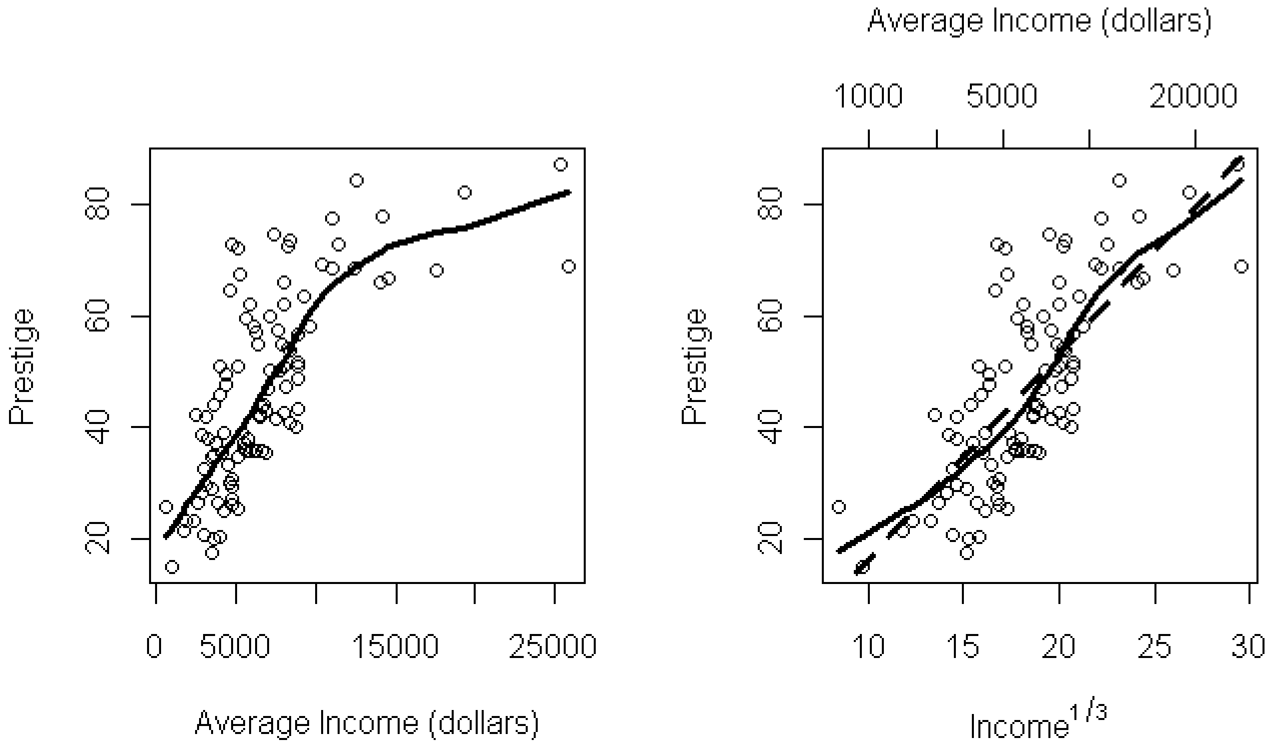


Fig 4.5 How should we choose whether to linearize by plotting $\sqrt{Y}$ vs. X or Y vs. X^2?

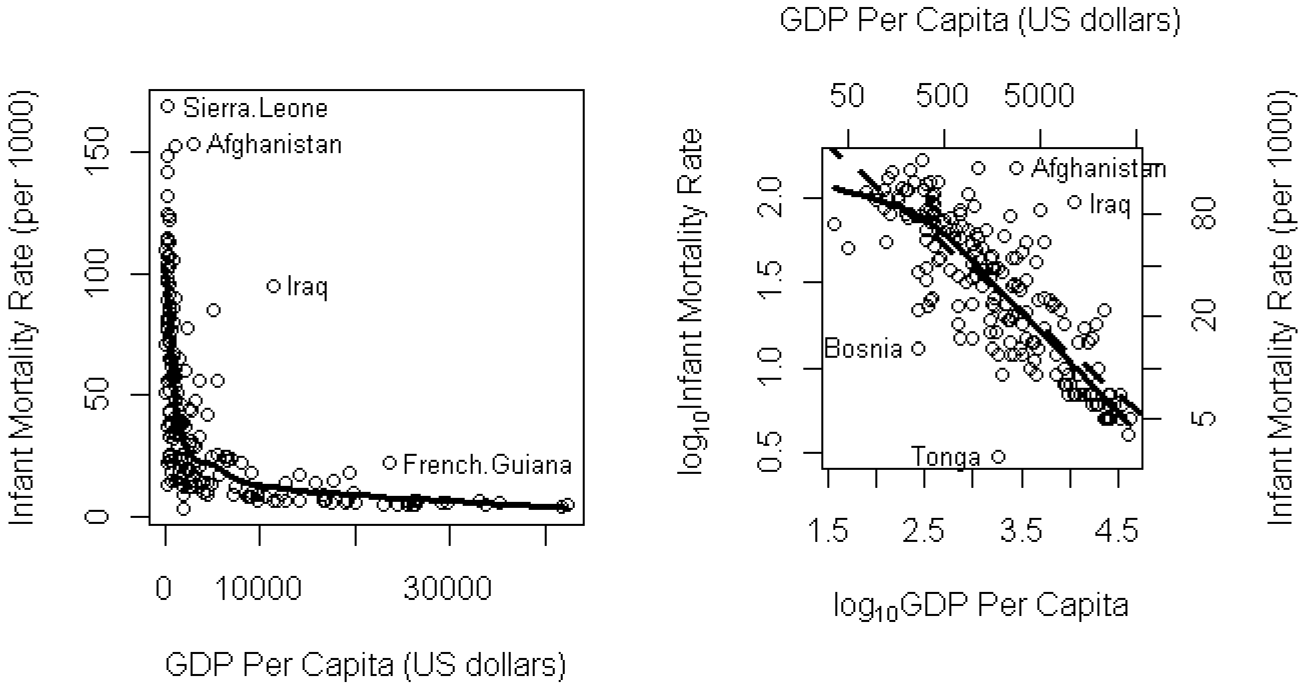
Fig 4.7. Tukley and Mosteller's bulging rule: The direction of the bulge indicates the directin of the power transformation ofY and/or X to straighten the relationship between them.



Figs. 4.8 & 4.9   Prestige vs. Income: tranform Prestige “up” or Income “down”



Figs. 4.10 & 4.11   Infant mortality vs GDP



With both variables expressed on a log scale, the slope of -0.49 means that a 1% increase in per capita GDP is associated with a 0.49% decrease in infant mortality rate. A regression coefficient for this type of log-log relationship is called an “elasticity”.

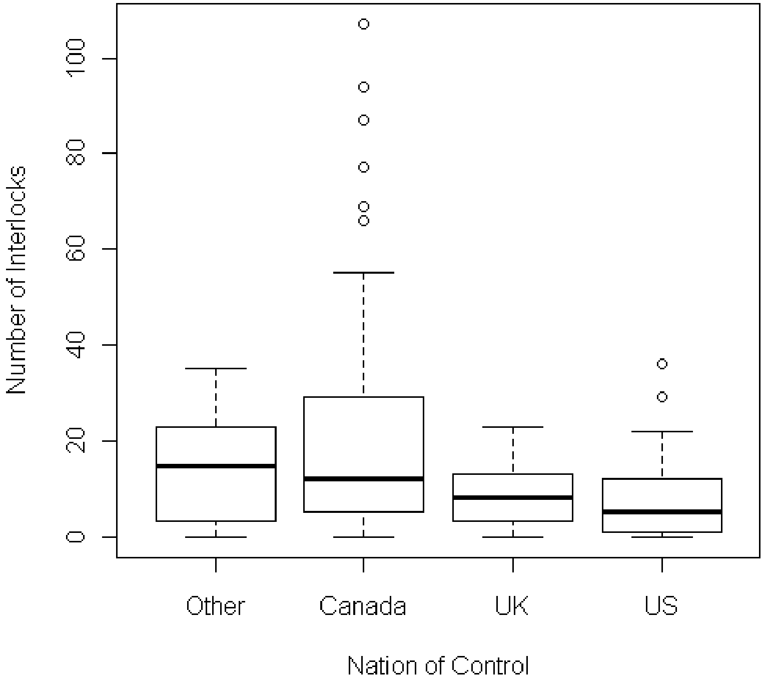
4.3 Transforming Nonconstant Spread

Differences in spread (variance) are often systematically related to differences in level (mean).

Suggested: *For grouped data*, plot log(hinge-spread) vs log(median). [The "hinge-spread" is essentially the inter-quartile range (IQR), an alternative "robust" estimate of spread in comparison with the standard deviation.]

If the plot is reasonably characterized as a line (plus noise): $\log spread \approx a + b \log level$, the spread-stabilizing transformation uses the power $p = 1 - b$. Let's compute this in R. *Note*: Fox adds 1 to interlocks because there are some zero values. However, it is not necessary to do so to compute medians and IQR, only to compute logs of observations.

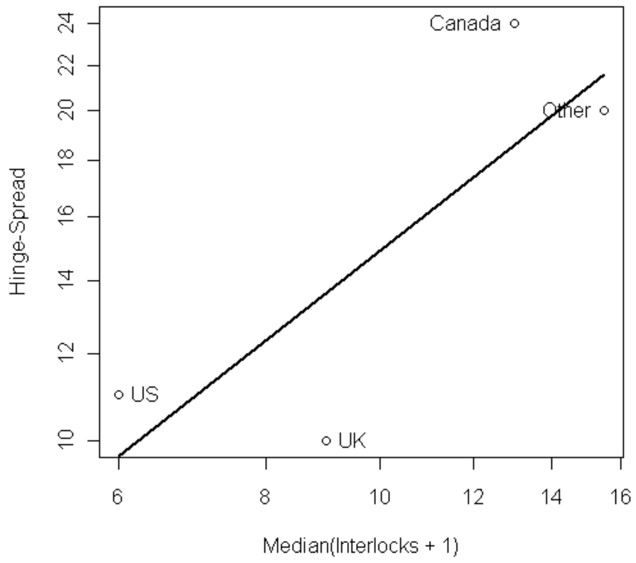
Fig 4.12

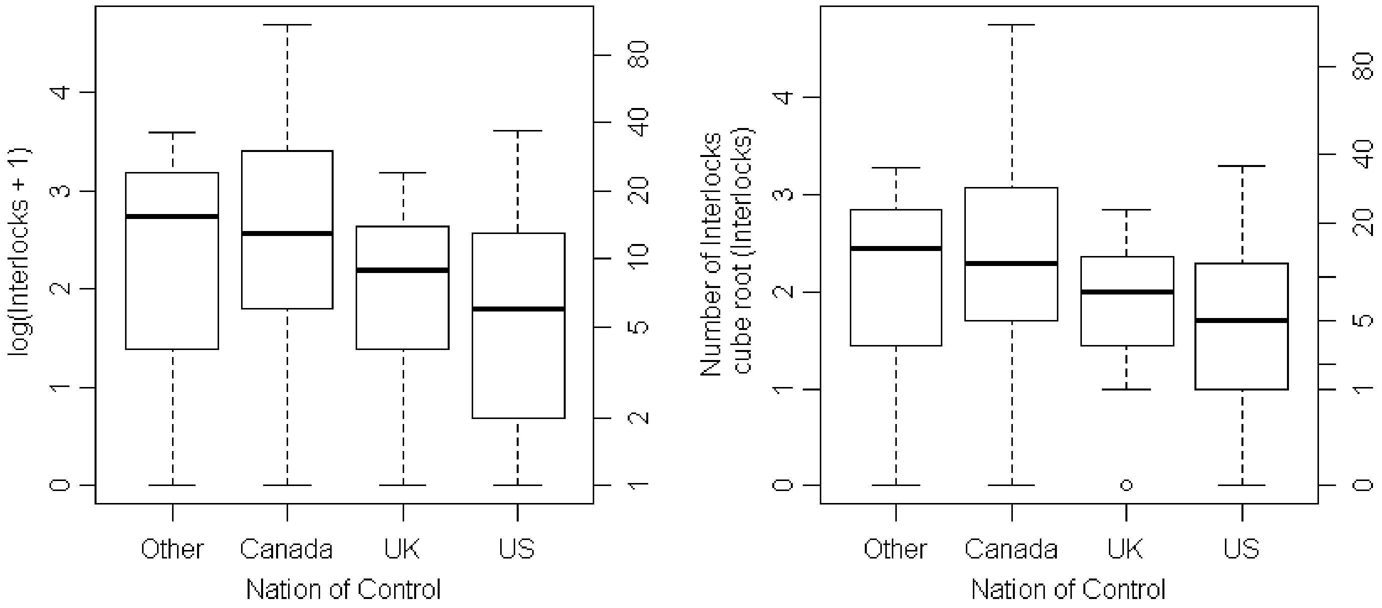


```
> nation.med <- tapply(interlocks+1,nation,median)
> nation.lq <-
    tapply(interlocks+1,nation,quantile,.25)
> nation.uq <-
    tapply(interlocks+1,nation,quantile,.75)
> nation.iqr <- nation.uq-nation.lq
> plot(log(nation.med),log(nation.iqr))
> coef(lm(log(nation.iqr) ~ log(nation.med)))
   (Intercept) log(nation.med)
     0.8530828       0.7958663
> abline(coef(lm(log(nation.iqr) ~
    log(nation.med))))
```

Suggested transformation is $p \approx 0.2$. So, take fourth root? cube root? log? R code in car library produces a slightly different slope estimate:

```
> spread.level.plot(interlocks + 1 ~ nat, robust=FALSE,
+     main="", xlab="Median(Interlocks + 1)",
+     ylab="Hinge-Spread", col="black")

       LowerHinge Median UpperHinge Hinge-Spread

US              2    6.0         13           11
UK              4    9.0         14           10
Canada          6   13.0         30           24
Other           4   15.5         24           20

Suggested power transformation:  0.1534487
```





```
windows(height=7, width=7)
par(mar=c(5,4,2,3),pty="s",mfrow=c(1,2),mgp=c(2,.75,0))

boxplot(log(interlocks + 1) ~ nat, xlab="Nation of Control",
    ylab=expression(paste(log, "(Interlocks + 1)")))

power.axis(0, base=exp(1), side="right", at=c(0, 1, 2, 5, 10, 20, 40, 80),
    axis.title="Number of Interlocks")

cr.interlocks <- interlocks^(1/3)

boxplot(cr.interlocks ~ nat, xlab="Nation of Control",
    ylab=expression(paste("cube root (Interlocks)")))

power.axis(0, power=(1/3), side="right", at=c(0, 1, 2, 5, 10, 20, 40, 80),
    axis.title="Number of Interlocks")
```

Note: For the example above, if you compute and plot log(sd) vs log(mean), you get a slope of 0.977, so a power estimate much closer to 0, suggesting the log scale more strongly.

Mathematical argument for variance-stabilizing transformations

The following couple of sections are from the Stat 421/502 notes.

Variance stabilizing transformations

Recall that one justification of the normal model for grouped data,

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

was that if the noise $\varepsilon_{ij} = X_{ij1} + X_{ij2} + \cdots$ was the result of the addition of unobserved **additive, independent** effects then by the central limit theorem ε_{ij} will be approximately normal.

However, suppose the effects are **multiplicative**, so that in fact:

$$Y_{ij} = \mu_i \times \varepsilon_{ij} = \mu_i \times (X_{ij1} \times X_{ij2} \times \cdots)$$

In this case the “noise” term ε_{ij} should be defined as a random variable close to (centered on) 1, the Y_{ij} will not be normal, and the variances will **not** be constant:

$$\mathrm{Var}[Y_{ij}] = \mu_i^2 \mathrm{Var}[X_{ij1} \times X_{ij2} \times \cdots]$$

Log transformation:

$$\begin{aligned}\log Y_{ij} &= \log \mu_i + (\log X_{ij1} + \log X_{ij2} + \cdots) \\ \mathrm{Var}\left(\log Y_{ij}\right) &= \mathrm{Var}\left(\log \mu_i + \log X_{ij1} + \log X_{ij2} + \cdots\right) \\ &= \mathrm{Var}\left(\log X_{ij1} + \log X_{ij2} + \cdots\right) \\ &= \sigma^2_{\log y}\end{aligned}$$

So that the variance of the log-data does not depend on the mean μ_i. Also note that by the central limit theorem the errors should be approximately normally distributed.

Other transformations:

For data having multiplicative effects , we showed above that

$$\sigma_i \propto \mu_i,$$

and taking the log stabilized the variances. In general, we may observe: $\sigma_i \propto \mu_i^b$ i.e. the standard deviation of a group depends on the group mean.

The goal of a variance stabilizing transformation is to find a transformation $y_{ij}^* = g(y_{ij})$ such that $\sigma_{y_{ij}^*} \propto (\mu_i^*)^0 = 1$, i.e. the standard deviation doesn't depend on the mean.

Consider the class of power transformations, transformations of the form $Y_{ij}^* = Y_{ij}^\lambda$. Based on a Taylor series expansion of $g_\lambda(Y) = Y^\lambda$ around μ_i, we have

$$\begin{aligned} Y_{ij}^{*} &= g_{\lambda}(Y_{ij}) \\ &\approx \mu_i^{\lambda} + (Y_{ij} - \mu_i)\lambda\mu_i^{\lambda-1} \\ \mathrm{E}\left(Y_{ij}^{*}\right) &\approx \mu_i^{\lambda} \\ \mathrm{Var}\left(Y_{ij}^{*}\right) &\approx \mathrm{E}\left((Y_{ij}-\mu_i)^2\right)(\lambda\mu_i^{\lambda-1})^2 \\ \mathrm{SD}\left(Y_{ij}^{*}\right) &\propto \mu_i^{b}\mu_i^{\lambda-1} = \mu_i^{b+\lambda-1} \end{aligned}$$

So if we observe $\sigma_i \propto \mu_i^b$, then $\sigma_i^* \tilde{\propto}\, \mu_i^{b+\lambda-1}$. So if we take $\lambda = 1 - b$ then we will have stabilized the variances to some extent. Of course, we typically don't know b, but we could try to estimate it from data.

Estimation of b:

$$\begin{aligned} \sigma_i \propto \mu_i^b &\Leftrightarrow \sigma_i = c\mu_i^b \\ \log\sigma_i &= \log c + b \times \log\mu_i, \\ \text{so } \log s_i &\approx \log c + b \log \bar{y}_{i\cdot} \end{aligned}$$

Thus we may use the following procedure:

- Plot $\log s_i$ vs. $\log \bar{y}_{i\cdot}$
- Fit a least squares line: `lm(` $\log s_i \sim \log \bar{y}_{i\cdot}$ `)`
- The **slope** $\hat{b}$ of the line is an estimate of b.
- Analyze $y_{ij}^* = y_{ij}^{1-\hat{b}}$.
-

4.5 Transforming Proportions

Proportions: $p \in [0,1]$

Transformations typically considered to have a score on the real line:

- $\text{logit}(p) = \log \frac{p}{1-p}$, "log odds"
- $\text{probit}(p) = \Phi^{-1}(p)$ Note: there are relatively subtle differences between logit and probit.
- $\sin^{-1}\left(\sqrt{p}\right) = \arcsin\left(\sqrt{p}\right)$, the variance-stabilizing transformation

Fox doesn't mention that $\arcsin\left(\sqrt{p}\right)$ is the transformation that approximately stabilizes the variance.

If $X \sim \text{Bin}(n, p)$, and $\hat{p} = X/n$, then

$$E(\hat{p}) = p, Var(\hat{p}) = \frac{1}{n} p(1-p)$$

And $Var\left(\sin^{-1}\sqrt{\hat{p}}\right) \approx \frac{1}{4n}$ (radians).

Note, however, that there are many cases where the response variable being analyzed is a proportion (or fraction) that does not arise from a binomial sampling model, so that argument for $\arcsin\left(\sqrt{p}\right)$ may not apply.

Count data: A similar argument to be aware of:

If $Y \sim \text{Poisson}(\lambda)$

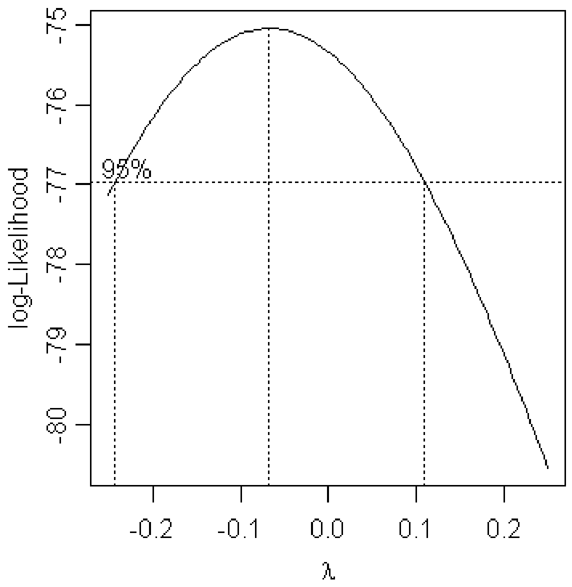
$$E(Y) = Var(Y) = \lambda$$

so $\sigma^2 = \mu$ and $\log \sigma = \frac{1}{2} \log \mu$, so $p = \left(1 - \frac{1}{2}\right)$ and $\sqrt{Y}$ is the variance stabilizing-transformation.

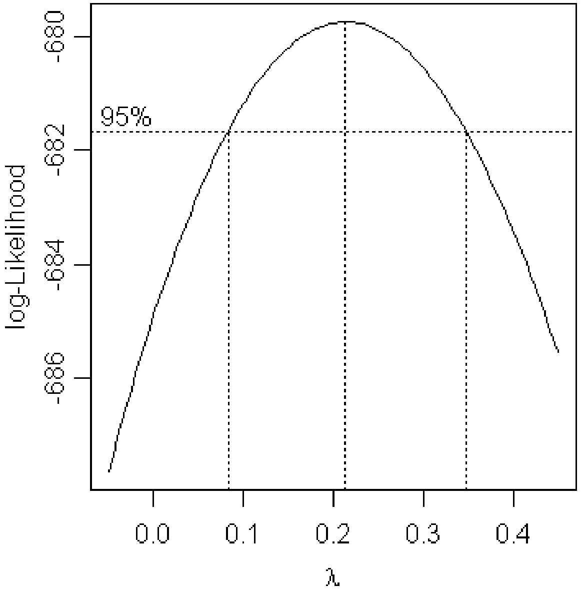
Remarks:

- Be aware of how transformations operate on the real line:
  - Fractional and negative powers "stretch out" values near zero and "pull in" large positive values
  - Powers greater than 1 do the opposite (and are rarely appropriate for variance stabilization)
- Transformations for proportions "stretch" values away from both zero and one.
  - Other types of outcomes that have both minimum (*floor*) and maximum (*ceiling*) values might need similar consideration
- How do you decide how to transform both independent (X) and dependent (Y) variables jointly. The Tukey-Mosteller "bulging" rule figure tells you what to try in order to straighten a relationship, but for statistical modeling, the most important (preferred) assumption is usually that of constant variance. Therefore:
  1. *Choose a transformation of the response variable Y to stabilize the variance*, being aware of the fact that the distribution of values on the horizontal (independent) variable axis does not influence whether the variance on the vertical (dependent) variable axis is constant. Then,
  2. *Choose a transformation of the independent variable to linearize the relationship* (if possible/feasible).

- The approach to determining a power transformation illustrated above assumes your data come in separate groups so that you can compute means and variances to be plotted in order to estimate a slope. *What do you do if you have a regression problem, say $Y \sim X$, where both variables take numeric (continuous) values and there are no grouping variables?*
  - Cut the data into groups according to the independent variable and apply the method above.
  - Try to transform (Y, X) jointly so that they have a 2-dimensional (bivariate) normal distribution using the method of maximum likelihood: `box.cox.powers` function in the `car` library.
  - For a specified linear model (which assumes you know the “right” scale or transformation for representing the independent variables), use the `boxcox` function of the MASS library. Examples for the boxcox help file:



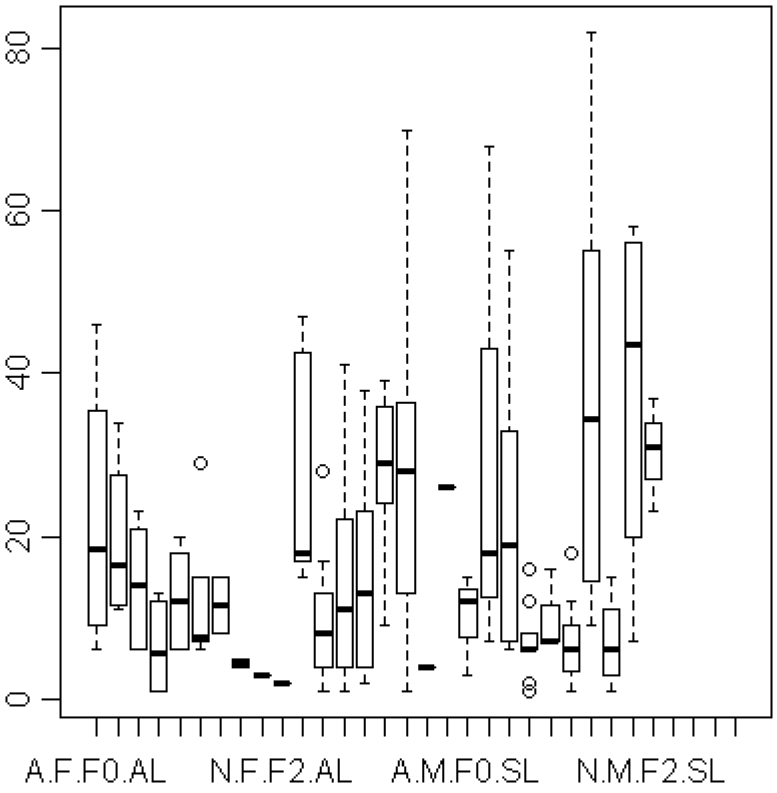
```
boxcox(Volume ~ log(Height) + log(Girth),
data = trees, lambda = seq(-0.25,0.25,length = 10))


> trees[1:5,]
Girth Height Volume
1   8.3     70   10.3
2   8.6     65   10.3
3   8.8     63   10.2
4  10.5     72   16.4
5  10.7     81   18.8
```

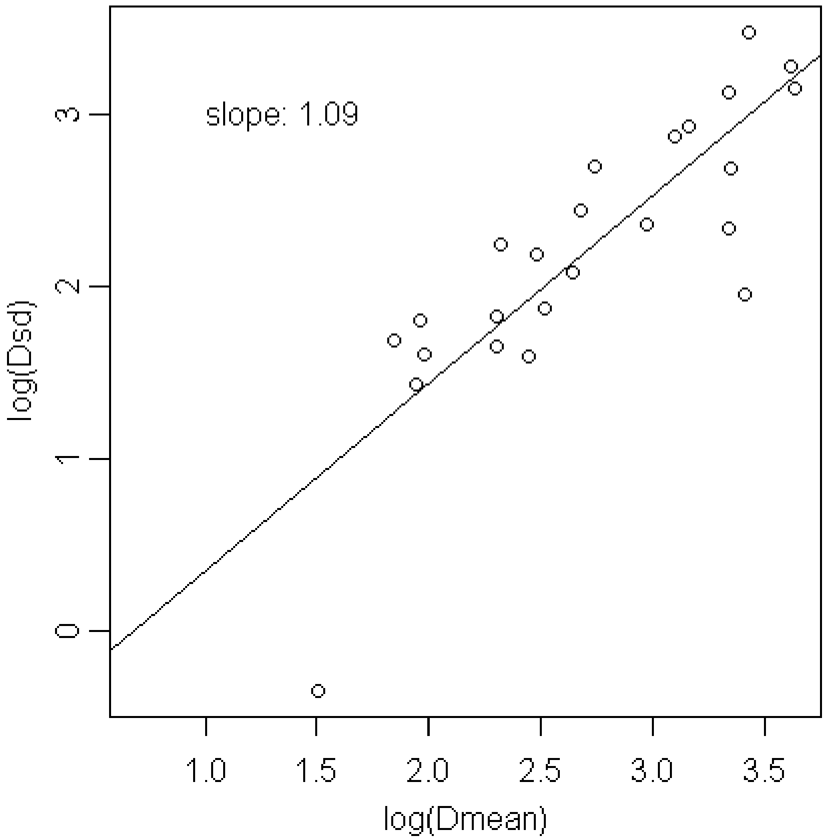


```
boxcox(Days+1 ~ Eth*Sex*Age*Lrn, data = quine,
       lambda = seq(-0.05, 0.45, len = 20))

> summary(quine)
 Eth    Sex    Age     Lrn          Days
 A:69   F:80   F0:27   AL:83   Min.   : 0.00
 N:77   M:66   F1:46   SL:63   1st Qu.: 5.00
               F2:40           Median :11.00
               F3:33           Mean   :16.46
                               3rd Qu.:22.75
                               Max.   :81.00
```

The response in this example is #days absent from school as a function of ethnic group (Aboriginal, Non-aboriginal), sex, age group (grade in school), and Learning group (average, slow). That is, these are grouped data. So let’s compare with variance stabilization.

```
attach(quine)
Days1 <- Days+1
boxplot(Days1 ~ Eth+Sex+Age+Lrn)

Dmean <- tapply(Days1,list(Eth,Sex,Age,Lrn),mean)
Dsd <- tapply(Days1,list(Eth,Sex,Age,Lrn),sd)

plot(log(Dmean),log(Dsd))
abline(coef(lm(log(Dsd)~log(Dmean))))
coef(lm(log(Dsd)~log(Dmean)))
```


0
20
40
60
80
A.F.F0.AL
N.F.F2.AL
A.M.F0.SL
N.M.F2.SL


slope: 1.09
log(Dsd)
0
1
2
3
1.0
1.5
2.0
2.5
3.0
3.5
log(Dmean)