# 36-617: Applied Linear Models

Transformation of Variables

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

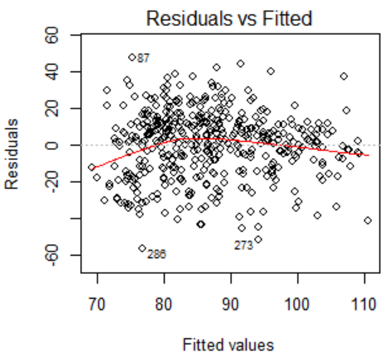
# Announcements

- Quiz 02 available at 5pm – due Tues at 7pm
- HW02 – Due tonight at 11:59pm
- Reading
  - This week: Sheather Ch 6 (diagnostics & transformations)
    - (supplemental: ISLR 3.3.3; G&H Ch 4)
  - Next week: Sheather Ch 7 (variable selection)
    - (supplemental: ISLR Ch 6; G&H Ch 4)
- HW 03 out on Canvas
  - Due Mon 1159pm

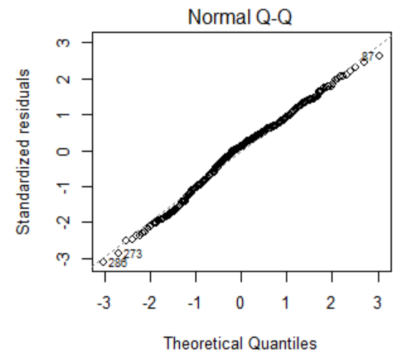
# Outline

- Quick Review of Casewise Diagnostic Plots
- Transformations -- Why & How for X and Y
  - Aside: the approach I suggested in HW01 solutions
- Substantive (investigator-driven) considerations
- Variance Stabilization for Y
- Box-Cox for X or Y: Fix distribution(s)
- Inverse Response Plot for Y
- Perspective and recommendations
  - Aside: the approach I suggested in HW01 solutions
- Can residual plots distinguish $y^{(1)} = \beta_0 + \beta_1 x^2 + \varepsilon$, vs. $y^{(2)} = (\beta_0 + \beta_1 x + \varepsilon)^2$ ?

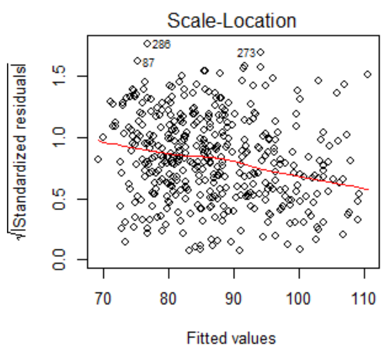
# Casewise Diagnostics and Patterns



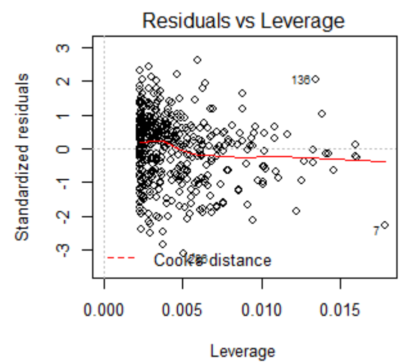

- Mean zero?
- Outliers?
- Functional dependence on $\hat{y}_i$?




- Normal?
- Outliers?
- Large enough sample?




- Constant variance?
- Outliers?
- Functional dependence on $\hat{y}_i$?




- NE & SE corners:
  - High leverage $h_{ii}$
  - High std resid $r_i$
- $D_i$ > 0.5 or so?

- Generally these are conversation points
  - Could reveal things investigator cares about!
  - Otherwise, look for data collection/recording errors
- Delete data only with a good justification!

# Transformations

- Why to transform
  - ***Substantive (investigator-driven) reasons***
  - Improving fit of data to modelling assumptions; makes formal (and informal) inference more valid
- Why not to transform
  - ***Substantive (investigator-driven) reasons!***
- What to transform
  - X: often trying to reduce leverage; normality is an ***<u>informal</u>*** target
  - Y: really trying to improve distribution of $\epsilon_i$ , but access is indirect
  - X and/or Y: linearity wrong; improve functional form *y= f(x)*
- How to transform
  - We will concentrate on power-function methods for now
  - Nonparametric function estimation (e.g. gam() in R) provides another approach

# Aside: Approach in HW01 Solutions

- In the solutions to HW01, I recommended trying each of these approaches:
  - For all the variables (y as well as the x's), find any that are asymmetric (skewed) and transform each one to make it more symmetric

  OR

  - Use the scatterplot matrix to find x's that are nonlinearly related to y, and then transform the x so that it is more linearly related to y.
- These are a heuristic one-variable-at-a-time approach to the more general recommendations in this lecture.
  - They often get you pretty far, and are almost always worth a try.
  - Limit yourself to simple transformations: $x^2, \sqrt{y}, \sqrt{x}, \log(y)\,, \log(x)\,, \exp(y)\,, \exp(x)\,, \ldots$
- Substantive (investigator) considerations $\gg$ math, always!!

# Transformations of *X*

- If *X* is discrete or a design variable, there is usually no sensible transformation to make!
- If *X* is continuous, it has an (empirical) distribution. We might want to transform *X* for any of three reasons
  - *Substantive*: we know *Y* is a nonlinear function of *X*, or we want a particular interpretation
  - *Leverage*: bring the (empirical) distribution of *X* closer to normality; reduces high-leverage points
  - *Functional*: $y = f(X)$ is not linear and we want to find a better functional form for $f()$

# Substantive Transformation of $X$

- There might be substantive knowledge.
  - E.g. in physics if $Y$ is the intensity of an effect at distance $X$, often an inverse-square law applies, so we might replace $X$ with $X' = 1/X^2$.

- A better interpretation might be available
  - Recenter $X$ so that the intercept $\beta_0$ is interpretable
  - Rescale $X$ to change units of slope $\beta_1$ (e.g. to SD's of $X$)

- Percent change in $X$ matters more than additive change: logarithms…

# A Substantive reason for log transform: effect of percent change

- For the model $y = \beta_0 + \beta_1 x + \epsilon$:

We consider a small change in x, instead of a 1 unit change

$$\begin{aligned} E[y|x+1] &= \beta_0 + \beta_1(x+1) \\ E[y|x] &= \beta_0 + \beta_1 x \\ \Delta E[y] = E[y|x+1] - E[y|x] &= \beta_1 \cdot 1 \end{aligned}$$

$\beta_1$ is the change in E[*y*] for a 1 unit change in *x*

- For the model $y = \beta_0 + \beta_1 \log(x) + \epsilon$:

Δx is only a 1% change in x

$$\begin{aligned} E[y|x+\Delta x] &= \beta_0 + \beta_1 \log(x+\Delta x) \\ E[y|x] &= \beta_0 + \beta_1 \log x \\ \Delta E[y] = E[y|x+\Delta x] - E[y|x] &= \beta_1 \cdot \log\left(1 + \frac{\Delta x}{x}\right) \approx \beta_1 \cdot \left(\frac{\Delta x}{x}\right) \quad (*) \\ \text{Putting } \Delta x = 0.01x,\ \Delta E[y] &\approx \beta_1(0.01) \end{aligned}$$

$\boxed{(0.01)\beta_1 \text{ is the change in E}[y] \text{ for a 1\% change in } x}$

(*) Since *log(1+x)* = *x* – *x*$^2$/2 +/- … (Taylor series)
See also “log xform and percent interpretation.pdf”

# Reducing leverage – power transforms

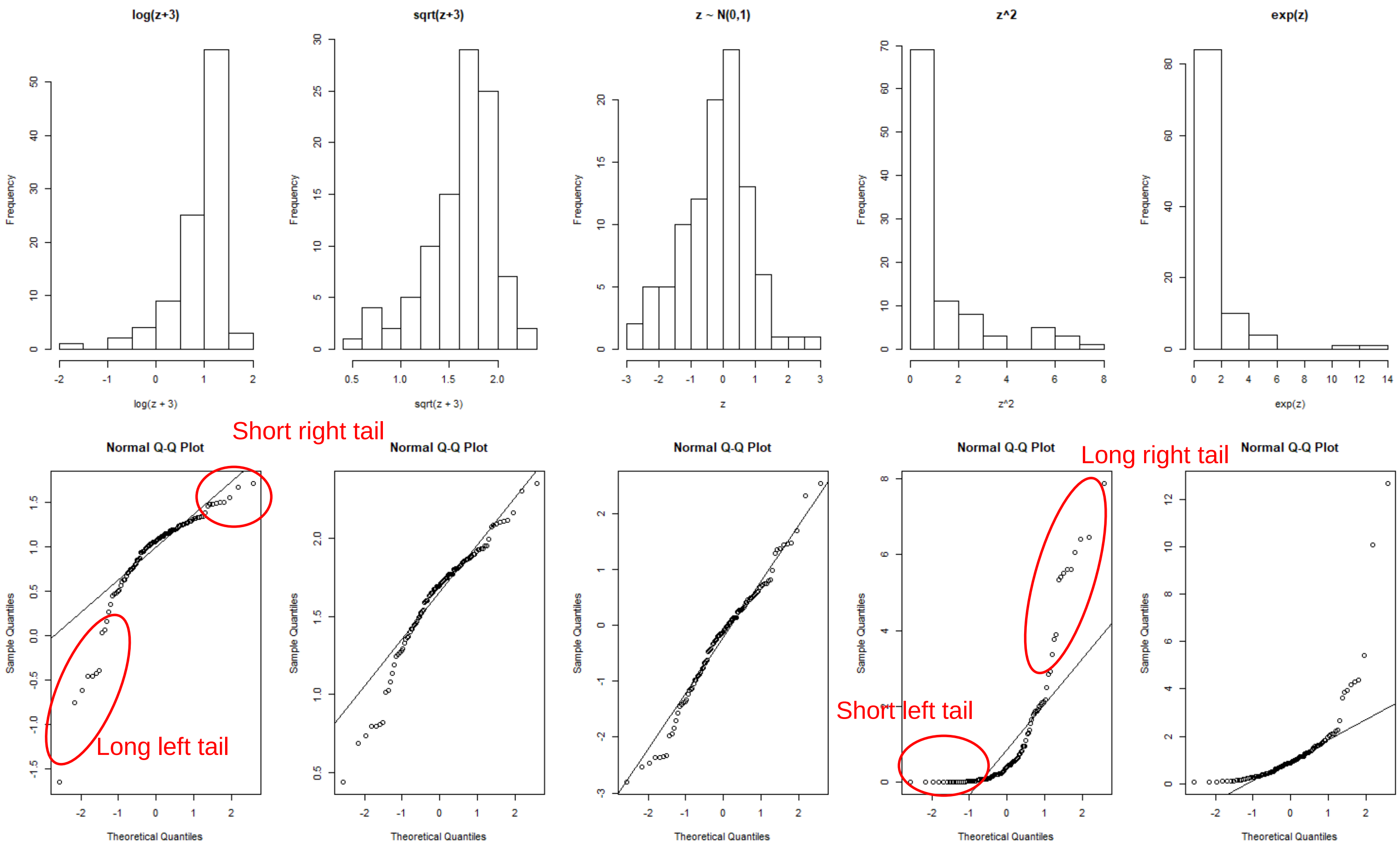
- In regression, we are conditioning on *X*:

$$Y|X \sim N(X\beta, \sigma^2)$$

so "officially" the distribution of *X* doesn't matter

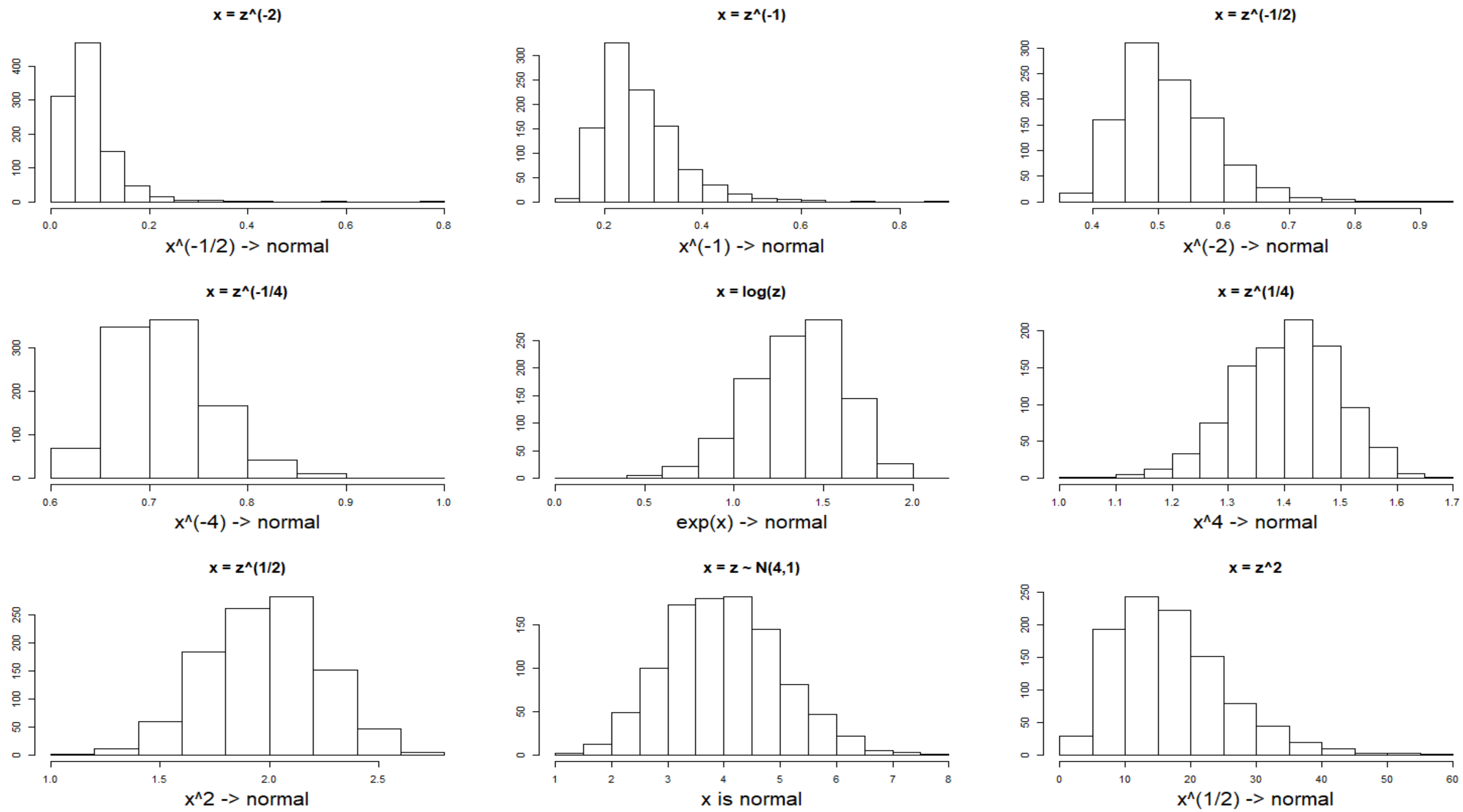
- However, if the (empirical) distribution of *X* is skewed, many *X*'s will have high leverage.
- Helps to make empirical distribution of *X* more symmetric – pull tails in
  - If X is skewed left (long left tail), $X^{\lambda}$, $\lambda>1$, pulls in tail
  - If X is skewed right (long right tail), $X^{\lambda}$, $\lambda<1$, pulls in tail
- Since $\log(x) = \lim_{\lambda \to 0} \frac{x^{\lambda} - 1}{\lambda}$, useful to think "$x^0 = \log(x)$"

# Aside: Reminder of distribution shapes




Transform back to symmetry: ? ? ? ? ?

# Effect of positive & negative powers: $\lambda \in (-2,-1,-1/2, -1/4, 0^*, 1/4, 1/2, 1, 2)$



# Reducing Leverage: Powers of X

- Check for symmetry after trying simple powers
- More formally, try to maximize likelihood

$$L(\lambda, \mu, \sigma^2) = [\lambda \cdot gm(x)^{(\lambda-1)}]^n \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x_i^\lambda - \mu}{\sigma}\right)^2\right]$$

  - **Box-Cox**: Likelihood simplifies if we replace $x^\lambda$ with

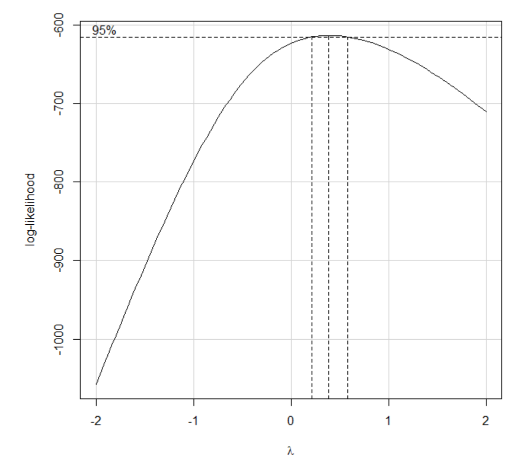
$$\Psi_M(x, \lambda) = gm(x)^{(1-\lambda)} \cdot \frac{x^\lambda - 1}{\lambda}, \quad gm(x) = \left[\prod_{i=1}^{n} x_i\right]^{1/n}$$

  - Usually suggests awkward values ($\lambda$ = 0.33453) that should be "rounded" to a simpler power ($\lambda$ = 1/3)
  - *x is assumed to be positive!*

# Implementing Box-Cox for *X* in R

- `library(car)`
  ("Companion to Applied Regression(*)")
  - `boxCox()`: show Box-Cox likelihood as a function of $\lambda$ ("profile likelihood")
  - `powerTransform()`: compute optimal $\lambda$ using the Box-Cox likelihood

```
> z <- rnorm(100,4,1)
> x <- z^3
> boxCox(x~1)
> powerTransform(x~1)
Estimated transformation parameter
       x
0.390494
```




(*) for Weissberg's Applied Linear Regression text.

# Functional: y = f(x) is not linear

- We can replace

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

with

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x^p + \epsilon_i \tag{2}$$

- This is also a good idea…

- N.b., model (2) still assumes equal additive errors!

# Transformations of Y

- We might want to transform Y for any of three reasons:
  - *Substantive*: we know *Y* is a nonlinear function of *X*, or we want a particular interpretation
  - *Improve residuals*: bring the (empirical) distribution of $\varepsilon_i$ closer to normality; makes inferences more valid
  - *Functional*: $y = f(X)$ is not linear and we want to find a better functional form for $f()$

# Substantive Transformation of *Y*

- There might be substantive knowledge.
  - If we know 0< *Y<100* (e.g. a test or hw score) we may need to transform *Y* before building a linear predictor for it: e.g. replace *Y* with *log[Y/(100-Y)]* ….
- Percent change in *Y* :
  - For $\log y = \beta_0 + \beta_1 x + \epsilon$, let $\Delta y = y' - y$, then

$$\begin{aligned} E[\log(y + \Delta y)] &= \beta_0 + \beta_1 (x+1) \\ E[\log(y)] &= \beta_0 + \beta_1 x \\ \Delta E[\log y] = E[\log(y + \Delta y)] - E[\log(y)] &= \beta_1 \cdot 1 \\ \text{So, } \beta_1 = E[\log(y + \Delta y)] - E[\log(y)] &= E\left[\log\left(1 + \frac{\Delta y}{y}\right)\right] \approx E\left[\frac{\Delta y}{y}\right] \quad (*) \end{aligned}$$

$$\boxed{100 \times \beta_1 = \text{expected pct change in y per unit change in } x}$$

(*) Since *log(1+x)* = *x* – *x*²/2 +/- … (Taylor series)
See also “log xform and percent interpretation.pdf”

# Improve Error (residual) Distribution

- We want to replace

$$y_i = X_i\beta + \epsilon_i$$

$i^{th}$ row of X matrix

with

$$y_i^\lambda = X_i\beta + \epsilon_i$$

to improve the distribution of $\epsilon_i$ (or $\hat{e}_i$ ).

- Can do "by hand" or by applying Box-Cox to

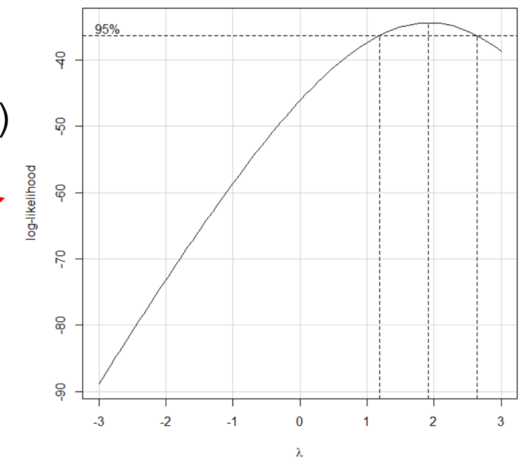
$y_i^\lambda - X_i\beta$ instead of $x_i^\lambda - \mu$

again, replace $y^\lambda$ with $\Psi_M(x, \lambda)$ ...

# Implementing Box-Cox for *Y* in R

- `library(car)`
  ("Companion to Applied Regression")
  - `boxCox()`: show Box-Cox likelihood as a function of $\lambda$ ("profile likelihood")
  - `powerTransform()`: compute optimal $\lambda$ using the Box-Cox likelihood

```
> z <- rnorm(100,4,1)
> y <- (1 + 3*z + rnorm(100,0,.25))^(1/2)
> lm.1 <-  lm(y ~ z)
> boxCox(lm.1,lambda=seq(-3,3,.1))
> powerTransform(lm.1)
Estimated transformation parameter
      Y1
1.922959
```



# Improve Error (residual) Distribution: Variance-stabilizing Transformations

- Suppose E[Y] = $\mu$, and Var(Y) = h($\mu$). We want a transformation Y*=g(Y) such that Var(Y*)=Const
- Taylor's Theorem says $g(y) \approx g(\mu) + g'(\mu)(y - \mu)$
- Therefore

$$\mathrm{Var}(Y^*) \approx \mathrm{Var}(g(\mu) + g'(\mu)(Y - \mu)) = [g'(\mu)]^2 h(\mu)$$

- We want this to be constant, i.e.

$$g'(\mu) = \frac{C}{\sqrt{h(\mu)}}; \quad \text{so} \quad g(\mu) = \int \frac{C}{\sqrt{h(\mu)}} d\mu$$

# Variance-Stabilizing Transform Example

- If Y ~ Poiss(μ), then we know E[Y]=μ and Var(Y)=μ
- So h(μ)=μ, and

"is proportional to"

$$g(\mu) = \int \frac{C}{\sqrt{\mu}}\, d\mu \propto \sqrt{\mu}$$

- Therefore $Y^* = \sqrt{Y}$ will have approximately constant variance (not depending on E[Y]).
- Nonconstant variance in a scale-location plot $\Longrightarrow$ consider a variance-stabilizing transformation.

# Functional form of Y: Inverse Response Plot

- Suppose

$$y_i = g(\beta_0 + \beta_1 x_i + \epsilon_i)$$

  then of course

$$g^{-1}(y) = \beta_0 + \beta_1 x_i + \epsilon_i$$

- It turns out[1] that if x has an elliptically symmetric distribution, then g can be estimated from a plot of $\hat{y}_i$ vs $y_i$, where $\hat{y}_i$ are predicted values from

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

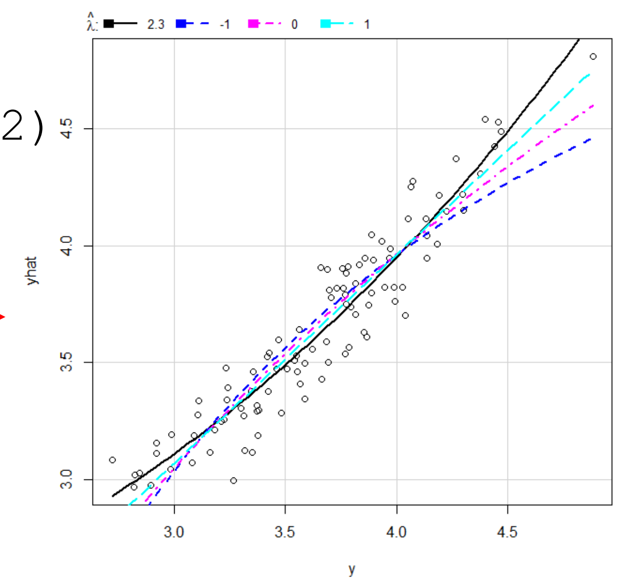
[1] Li and Duan (1989), Cook and Weisberg (1994)

# Implementing Inverse Response Plots In R

- `library(car)`
  ("Companion to Applied Regression")
  - `invResPlot()`: show inverse response plot ($\hat{y}_i$ vs. $y_i$) and calculate the power $\lambda$ for $y_i^\lambda$ by nonlinear least-squares(*)

```
> z <- rnorm(100,4,1)
> y <- (1 + 3*z + rnorm(100,0,.25))^(1/2)
> lm.1 <-  lm(y ~ z)
> invResPlot(lm.1)
  lambda      RSS
1  2.300377 1.711786
2 -1.000000 2.711177
3  0.000000 2.211967
4  1.000000 1.874950
```

← y' = y^(2.3)
← y' = 1/y
← y' = log(y)
← y' = y




(*) Specify particular lamdas to try with the lambda=c(…) argument.

# Perspectives and Recommendations

- Substantive (investigator-driven) considerations *always come first*
- Power transforms of X *to reduce leverage* & Power transforms of Y *to improve distribution of* $\epsilon_i$
  - By hand, or Box-Cox rounded to a simple power
- Inverse response plot for power transform of Y
  - Visually appealing, but Box-Cox probably better (directly addresses distribution of $\epsilon_i$ )
- There does not always exist a "perfect" transform!
- Transform for fcn form – depends on resid. plots!

# Aside: Approach in HW01 Solutions

- In the solutions to HW01, I recommended trying each of these approaches:
  - For all the variables (y as well as the x's), find any that are asymmetric (skewed) and transform each one to make it more symmetric

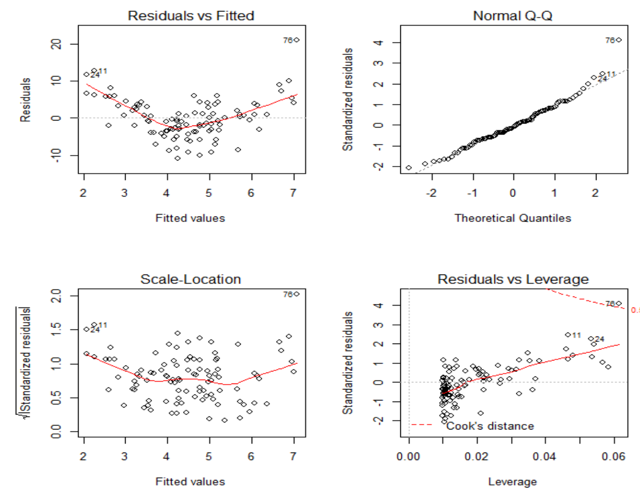
  OR

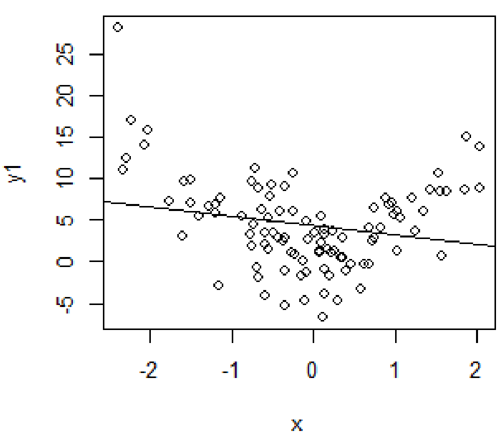
  - Use the scatterplot matrix to find x's that are nonlinearly related to y, and then transform the x so that it is more linearly related to y.
- These are a heuristic one-variable-at-a-time approach to the more general recommendations in this lecture.
  - They often get you pretty far, and are almost always worth a try.
  - Limit yourself to simple transformations: $x^2, \sqrt{y}, \sqrt{x}, \log(y), \log(x), \exp(y), \exp(x), \ldots$
- Substantive (investigator) considerations $\gg$ math, always!!

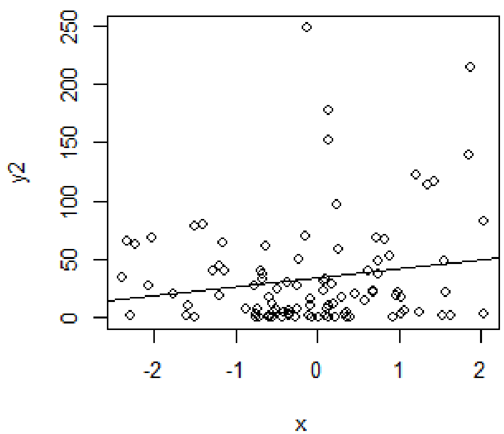
# Can residual plots distinguish $y^{(1)} = \beta_0 + \beta_1 x^2 + \varepsilon$, vs. $y^{(2)} = (\beta_0 + \beta_1 x + \varepsilon)^2$ ?

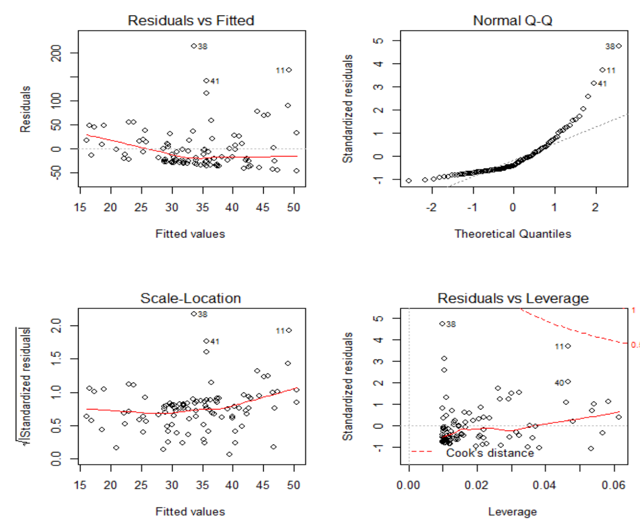
```
x <- rnorm(100,0,1)
y1 <- 1 + 3*x^2 + rnorm(100,0,4)
y2 <- (1 + 3*x + rnorm(100,0,4))^2
lm.1 <- lm(y1~x)
lm.2 <- lm(y2~x)
```

```
par(mfrow=c(2,2))
plot(x,y1); abline(lm.1)
plot(x,y2); abline(lm.2)
par(mfrow=c(2,2))
plot(lm.1)
plot(lm.2)
```



# Outline

- Quick Review of Casewise Diagnostic Plots
- Transformations -- Why & How for X and Y
  - Aside: the approach I suggested in HW01 solutions
- Substantive (investigator-driven) considerations
- Variance Stabilization for Y
- Box-Cox for X or Y: Fix distribution(s)
- Inverse Response Plot for Y
- Perspective and recommendations
  - Aside: the approach I suggested in HW01 solutions
- Can residual plots distinguish $y^{(1)} = \beta_0 + \beta_1 x^2 + \varepsilon$, vs. $y^{(2)} = (\beta_0 + \beta_1 x + \varepsilon)^2$ ?