Lecture 13: Transformations

Reading Assignment:

- Muller and Fetterman, Chapter 10: "Transformations"
- Weisberg, Chapter 8: "Transformations"

Transformation of the response and/or predictor variables may correct violations of homogeneity, linearity, and Gaussian distribution of errors (often these three assumptions stand or fall as a group).
Transformation may also simplify a model by linearizing a relationship.

Using transformations often boils down to trial and error, and we use diagnostics on the residuals to gauge the value of a transformation.

Always bear in mind that nonlinear models are part of the complete picture and may be the best alternative.

For regression, we consider three groups of transformations: linear, monotone, or non-monotone. This grouping is used to indicate invariance of statistical properties. For a T test, means and variances change under linear transformations, but the test statistic and p-value do not. Test statistics and p-values for many non-parametric methods do not change with monotone transformations of the data.

A transformation has statistical benefit or cost only if it changes probabilities (and hence inferences).

Transformations may help in model fitting and may provide scientific insight. If height^2 works better than height, then perhaps the true relationship involves surface area.

Variance Stabilizing Transformations of the Response

Variance stabilizing transformations are useful in some cases for treating heteroscedasticity.

When observations are amounts or measurements (think of a ratio scale variable), the standard deviation is often proportional to the mean. As an example, think of counting the money in a coke machine versus counting all the money in a Wachovia office.

Standard variance stabilizing transformations include the following.

Data	Distribution	Transformation
Count	Poisson	$\sqrt{y_i}$
Amount	Gamma	$\log y_i$
Proportion	Binomial/n	$\sin^{-1}(\sqrt{y_i})$

Despite these transformations, there are much better modeling strategies for count data and proportions. *Generalized Linear Models* are more appropriate for counts and proportions/percentages since they accommodate data from distributions other than the normal distribution.

We will devote most of our energy to discussion of *linearizing* transformations.

log transformation

The natural logarithm is useful when one expects the effect to be proportional to the response. For example, consider a model with one predictor, x in which the response is expected to increase 100ρ percent for each 1-unit increase in x. In addition, suppose that the error, δ, is multiplicative. Then we can write the model

$$y = \gamma(1+\rho)^x\delta.$$

As you can see, when $x = 0$, $y = \gamma\delta$, and when $x = 1$, $y = \gamma\delta + \gamma\rho\delta$, which is a $100\rho\%$ increase!

Taking logs on both sides, we have

$$\begin{aligned} \log(y) &= \log(\gamma) + x\log(1+\rho) + \log(\delta) \\ &= \alpha + \beta x + \varepsilon, \end{aligned}$$

where $\alpha = \log(\gamma)$, $\beta = \log(1+\rho)$, and $\varepsilon = \log(\delta)$. Thus taking logs transforms the complex multiplicative model into a simple linear model.

Solving for ρ in terms of β, we have $\rho = \exp(\beta) - 1$. So a one-unit increase in x corresponds to a $100(\exp(\beta) - 1)\%$ increase in y. For small β, say $|\beta| < 0.10$, you can interpret $\widehat{\beta}$ as a $100\widehat{\beta}$ percent effect on the response.

Power Transformation of the Response

Box-Cox Transformations

The square root transformation involves taking $y^{\frac{1}{2}}$. We may also wish to consider other powers of y in models of the form $y_i^{\pi} = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$. Although we can fit this model for various values of π, it is difficult to decide which transformation to use since the models are not directly comparable (the SSE's are not in the same units). To solve this problem, Box and Cox (1964) introduced a family of transformations of the response variable:

$$Y_i(\pi) = \begin{cases} \frac{Y_i^{\pi} - 1}{\pi \cdot Y^{*(\pi-1)}} & \pi \neq 0 \\ Y^* \cdot \ln(Y_i) & \pi = 0\,. \end{cases}$$

where $Y^* = \left(\prod_{i=1}^{N} Y_i\right)^{1/N}$, the geometric mean of the response, and the natural logarithm for $\pi = 0$ reflects the limit as $\pi \longrightarrow 0$.

This basically corresponds to the transform y^π for $\pi \neq 0$ and $\log(y)$ for $\pi = 0$, except that the Box-Cox transformation puts the SSE's on the same scale so that they can be compared.

Box and Cox recommended choosing π to minimize the SSE.

Below is a simple example to illustrate the Box-Cox transformation. Data were generated from the model $y = e^{x+\epsilon}$ where $\epsilon \sim N(0, 1)$. The transformed data can be fit with a linear model $\log(y) = x + \epsilon$. SAS code is below:

```
data x;
   do x = 1 to 8 by 0.025;
      y = exp(x + normal(7));
      output;
      end;
   run;

proc transreg data=x details;
   title2 Defaults;
   model boxcox(y) = identity(x);
   run;
```

```
   The TRANSREG Procedure
Transformation Information for BoxCox(y)
Lambda      R-Square    Log Like
-3.00          0.03    -4601.01
-2.75          0.04    -4266.08
-2.50          0.04    -3934.11
-2.25          0.05    -3605.75
-2.00          0.06    -3281.88
-1.75          0.07    -2963.74
-1.50          0.10    -2653.14
-1.25          0.14    -2352.72
-1.00          0.21    -2066.32
-0.75          0.34    -1799.25
-0.50          0.52    -1558.55
-0.25          0.71    -1360.28
0.00           0.79    -1275.31 <- best lambda
0.25           0.70    -1382.62
0.50           0.51    -1589.03
0.75           0.34    -1834.53
1.00           0.22    -2105.88
1.25           0.15    -2397.35
1.50           0.11    -2704.64
1.75           0.08    -3024.24
2.00           0.06    -3353.38
2.25           0.05    -3689.91
2.50           0.04    -4032.18
```

2.75	0.03	-4378.97
3.00	0.03	-4729.37

Regression assumption diagnostics must always be examined as well.

Note: all of these transformations are not well-defined if y is not strictly positive. In such a case, you may encounter problems for $\pi \leq 0$.

Ladder of Power Transformations

Alternatively, the "ladder of power transformations" below can be used to guide the choice of transformations.

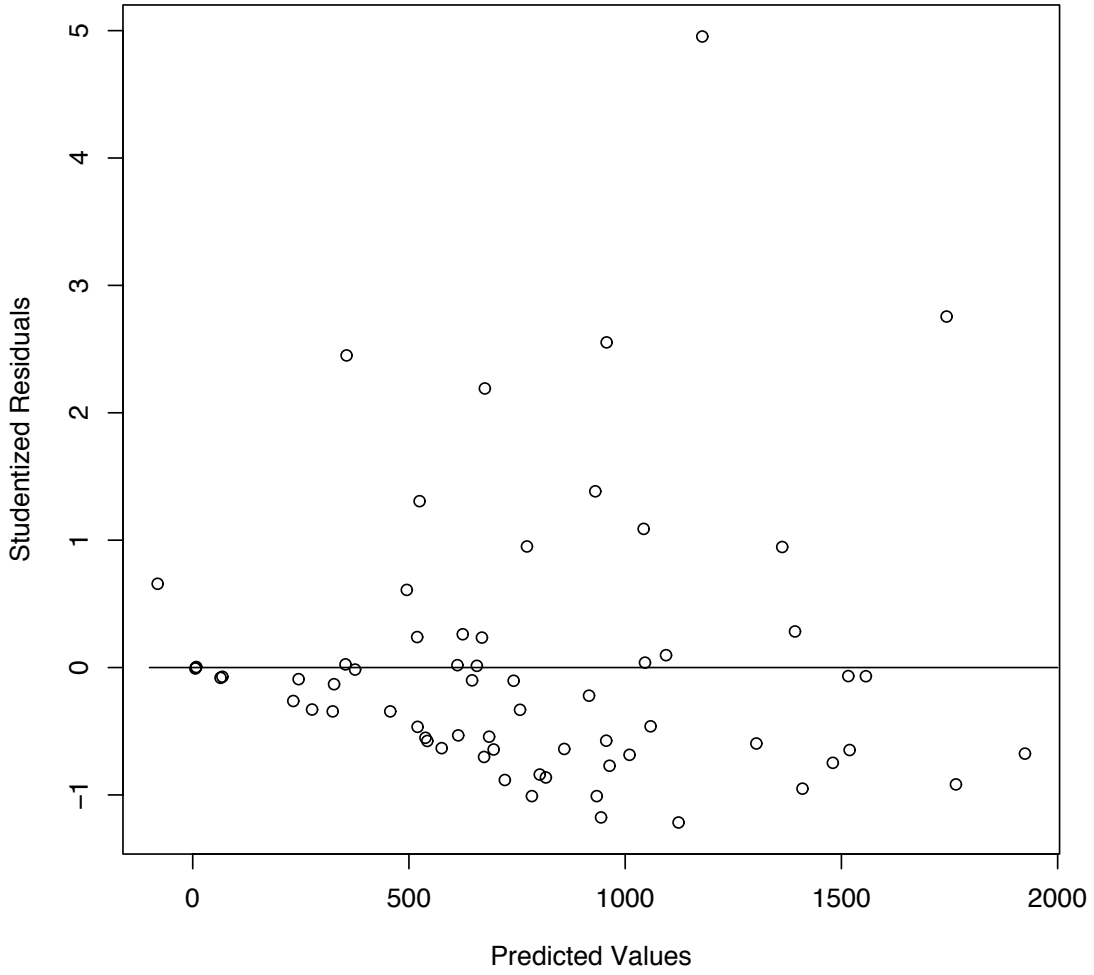
Half-Steps on the Ladder of Power Transformations

π	Transform			Description
	$\vdots$			
-2	y^{-2}			
$-3/2$	$y^{-3/2}$			
-1	y^{-1}			reciprocal
$-1/2$	$y^{-1/2}$	$=$	$1/\sqrt{y}$	
"0"	$\lim_{\pi \to 0} y^{\pi}$	$=$	$\ln y$	
$1/2$	$y^{1/2}$	$=$	$\sqrt{y}$	square root
1	y^{1}			identity
$3/2$	$y^{3/2}$			
2	y^{2}	$=$		square
	$\vdots$			

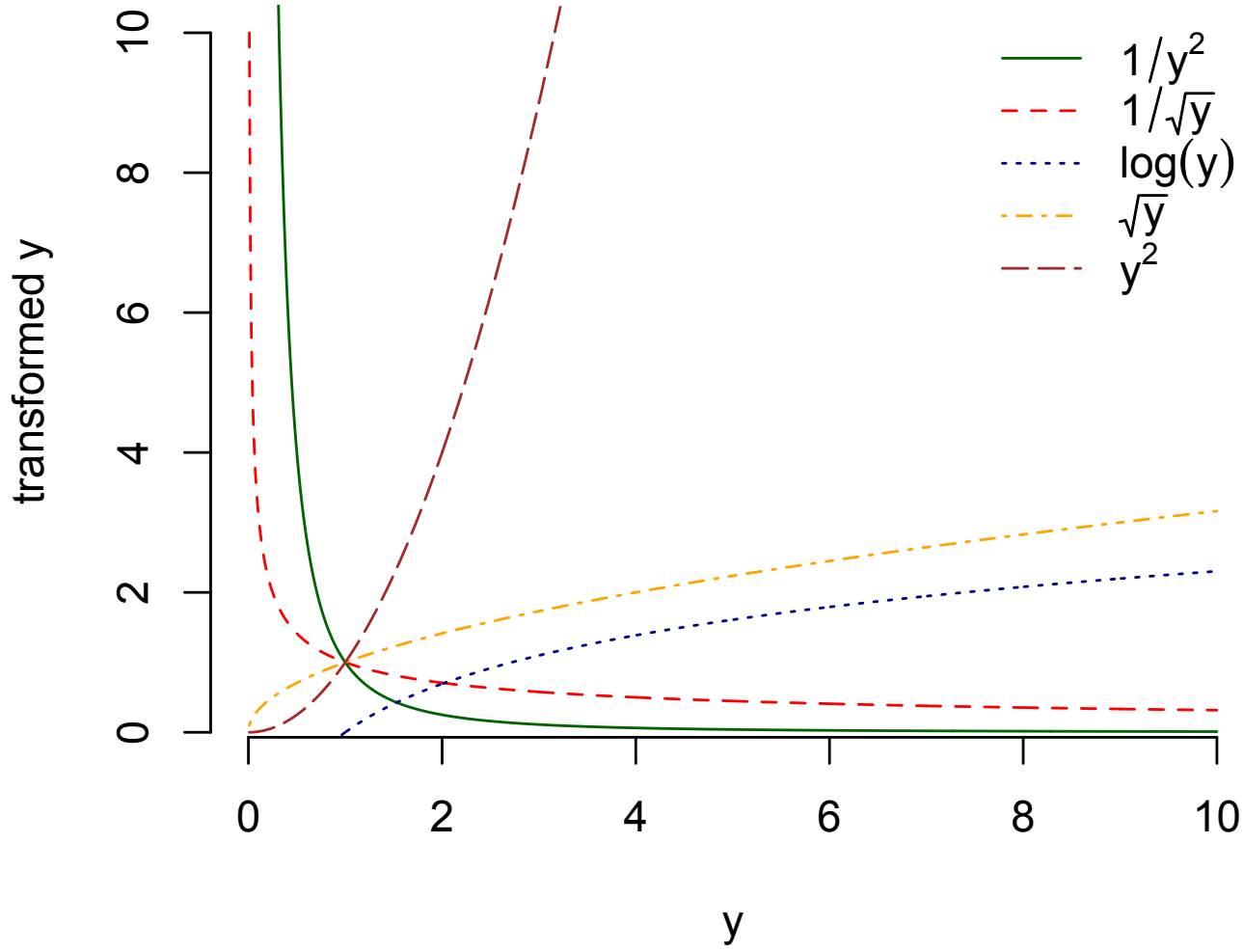
A crude but effective selection strategy is outlined below.

1. Fit the proposed model and check studentized residuals, including the R/P plot, the frequency histogram for skewness, and a test of Gaussian distribution. Multi-modality suggests an important predictor (or predictors) is missing from the model. If the residuals look good, keep the current model.
2. Based on the R/P plot and frequency histogram, select a grid of π values. Positive skewness in the histogram or a fan shape opening to the right on the R/P plot implies moving up the ladder from the value 1 (choosing values of $\pi < 1$), while negative skewness or a fan shape opening to the left on the R/P plot implies moving down ($\pi > 1$). Ratio scale variables often require $\pi < 1$.
3. Check studentized residuals for each value of π and examine other appropriate diagnostics, including tests of the Gaussian distribution of residuals.
4. Note that if you pick $\pi < 0$, the ladder direction reverses due to

the reciprocal transformation in addition to the power transformation. So if $\pi < 0$ for the current model, an R/P plot with a fan opening to the right implies moving down the ladder, and a fan opening to the left implies moving up the ladder. For example, consider the following R/P plot for $\pi = 2$ in the ozone data.



In this case, the fan opens to the right, and we move up the ladder.



5. To select among models with residuals that look ok, choose the

model with the smallest SSE.

6. Failure to obtain an acceptable model implies a more serious problem such as inadequate predictors or an entirely inappropriate model.

Example: Power Transformations for Ozone Data

The SAS code below can be used to program the Box-Cox transformation for the ozone data for selected values of π in the range $[-2, 2]$.

```
data ozonex;
merge gmean ozone;
y_2=((personal**2)-1)/(2*(geomeany**(2-1)));
y_1_5=((personal**1.5)-1)/(1.5*(geomeany**(1.5-1)));
y_1=((personal**1)-1)/(1*(geomeany**(1-1)));
y_5=((personal**0.5)-1)/(0.5*(geomeany**(0.5-1)));
y_0=geomeany*log(personal);
y_m5=((personal**(-0.5))-1)/(-0.5*(geomeany**(-0.5-1)));
y_m1=((personal**(-1))-1)/(-1*(geomeany**(-1-1)));
y_m1_5=((personal**(-1.5))-1)/(-1.5*(geomeany**(-1.5-1)));
y_m2=((personal**(-2))-1)/(-2*(geomeany**(-2-1)));
run;
```

```
proc glm data=ozonex;
model y_2=outdoor home time_out;
run;

proc glm data=ozonex;
model y_1_5=outdoor home time_out;
run;

proc glm data=ozonex;
model y_1=outdoor home time_out;
run;

proc glm data=ozonex;
model y_5=outdoor home time_out;
run;

proc glm data=ozonex;
model y_0=outdoor home time_out;
run;


proc glm data=ozonex;
```

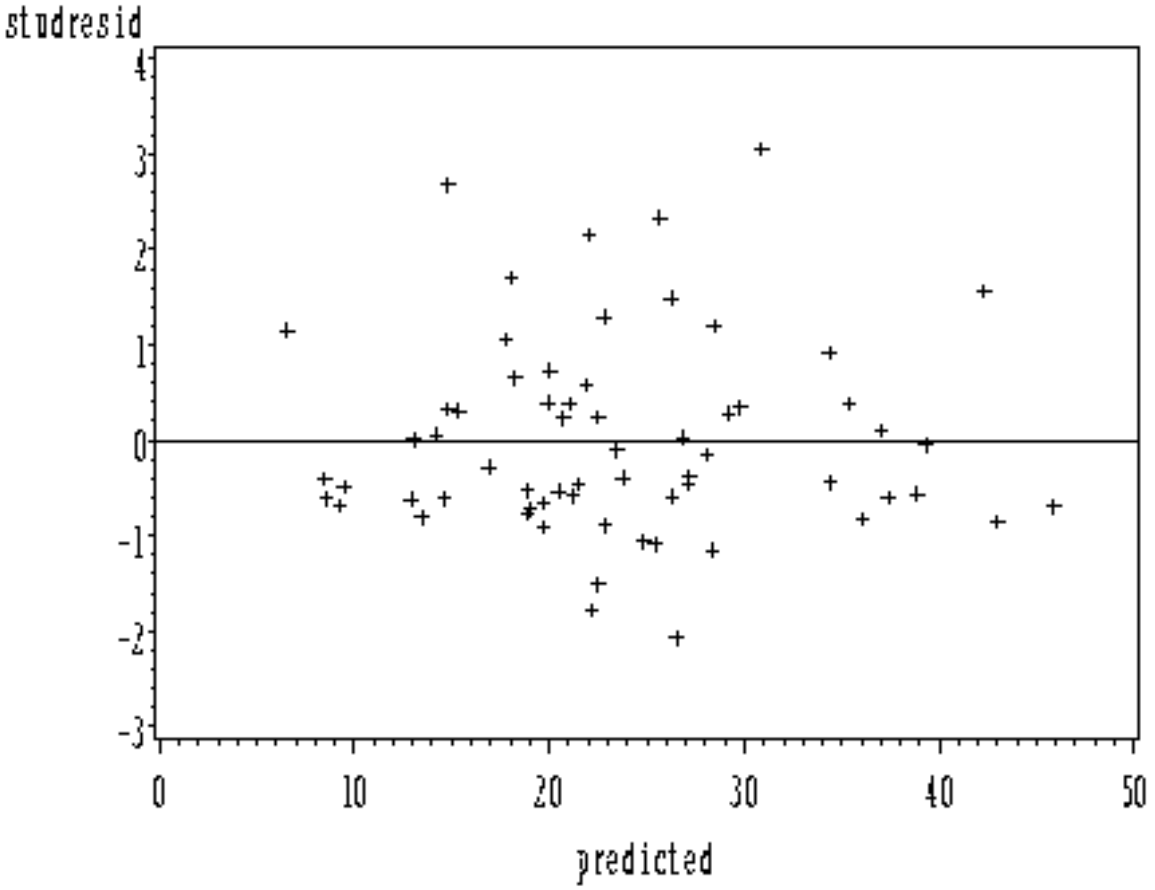
```
model y_m5=outdoor home time_out;
run;

proc glm data=ozonex;
model y_m1=outdoor home time_out;
run;

proc glm data=ozonex;
model y_m1_5=outdoor home time_out;
run;

proc glm data=ozonex;
model y_m2=outdoor home time_out;
run;
```

In the interest of time, we just look at the SSE for each model in the table below.

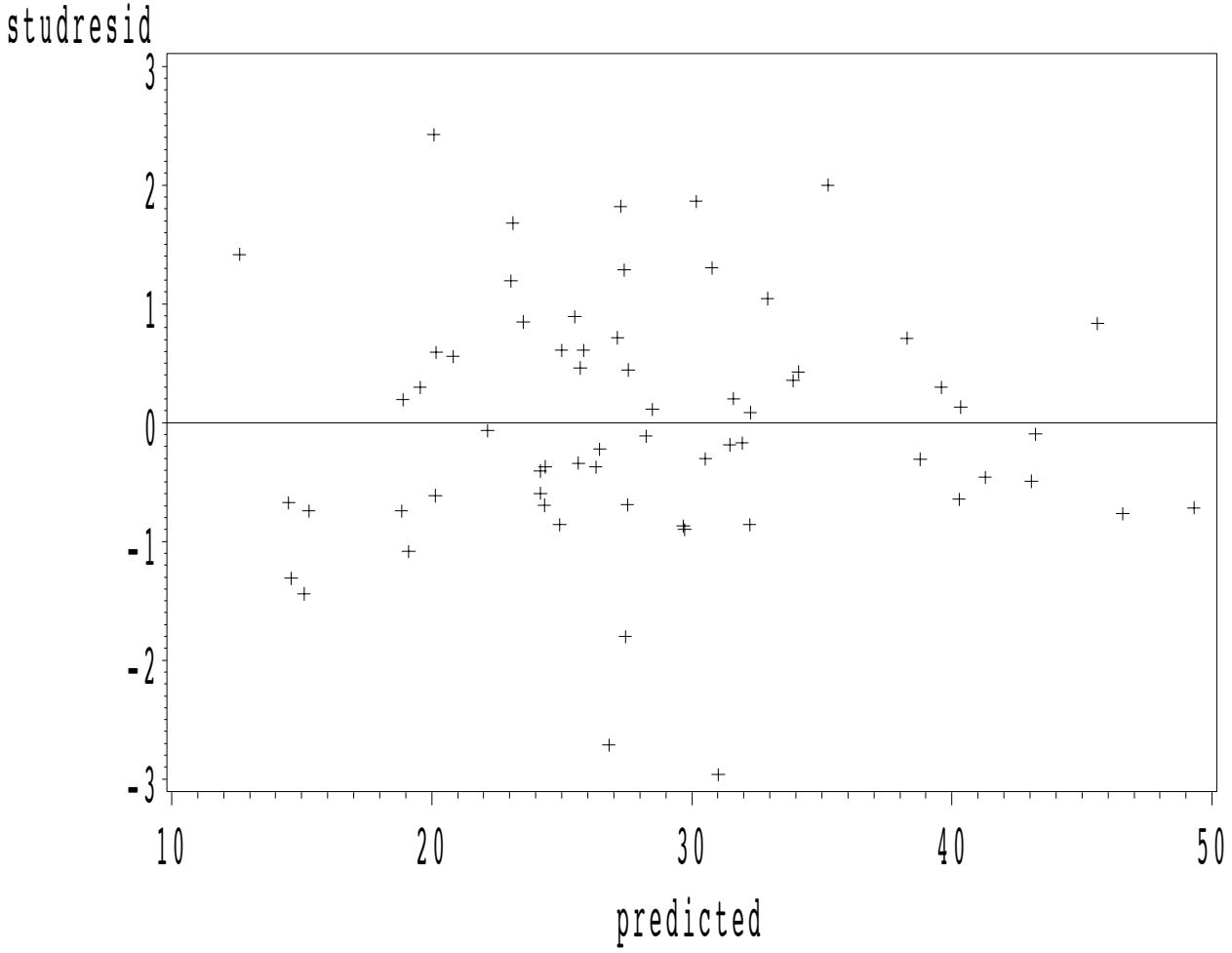
SSE for Ozone Data

π	SSE
2.0	37788
1.5	17671
1.0	10148
0.5	8578
0.0	15126
-0.5	69083
-1.0	637807
-1.5	8767407
-2.0	152261093

Let's examine the residuals for the untransformed data and for the square root transformed data. For the untransformed data, we have



and for the square root transformation, we have



Which transformation is best?

Alternatively, we can try the following SAS code:

```
proc transreg data=ozone ss2 details;
title2 Defaults;
model boxcox(personal) =
                identity(outdoor home time_out);
run;

Transformation Information
for BoxCox(personal)

Lambda      R-Square    Log Like
-3.00          0.05    -667.601
-2.75          0.05    -617.486
-2.50          0.05    -568.075
-2.25          0.05    -519.492
-2.00          0.05    -471.896
-1.75          0.05    -425.492
-1.50          0.05    -380.551
-1.25          0.06    -337.442
-1.00          0.07    -296.686
-0.75          0.09    -259.033
```

```
-0.50          0.14    -225.559
-0.25          0.19    -197.696
0.00           0.26    -176.954
0.25           0.31    -164.160
0.50           0.34    -158.805 <
0.75           0.34    -159.364
1.00           0.33    -164.183
1.25           0.31    -171.987
1.50           0.29    -181.931
1.75           0.27    -193.471
2.00           0.25    -206.253
2.25           0.23    -220.036
2.50           0.21    -234.650
2.75           0.19    -249.966
3.00           0.17    -265.885
```

Comments

- Regression assumption diagnostics should guide model choice.
- Remember that Gaussian distribution, homogeneity, and linearity often stand or fall together.
- Aesthetics often dictate using parallel transformations, as for a baseline covariate (if we choose to transform personal ozone exposures, then we may also wish to transform home and outdoor exposures using the same transformation).
- Choosing some transformations (say $y^{\frac{3}{2}}$) will make models difficult to interpret.
- Ratio scale variables often need transformation.
- The nature of any zeros affects the type of transformation. Log and reciprocal transformations often work well for blood, urine, or water assays of natural compounds with non-zero background levels. Values below a detection threshold that are recorded as

zero are instead informatively censored and not missing or equal to zero. Non-natural compounds, such as some pharmaceuticals in blood, have true zeros.

- Some transformations may not work well for some data (we do not wish to take the square root of a negative number or the log of zero).

Avoid back transformation because it may lead to potential bias. For example, a scientist was studying a population with a chronic disease. A depression scale was used as the response in a GLM. Using the strategy described above, the statistician chose $\sqrt{Y}$. The scientist complimented the statistician on the analysis. However, in the draft manuscript, the scientists included means, standard deviations, and back transformed predicted values (in the original metric). The statistician convinced the scientist to stick with transformed data. Why?

Nonlinear functions usually imply

$$E[f(X)] \neq f(E[X]).$$

For example, from Jensen's inequality we know that

$$E[Y^2] > (E[Y])^2.$$

Transforming Predictor Variables

Box and Tidwell suggest the following procedure to check whether a predictor should be transformed. To test whether a transform x^λ should be used in the model,

1. add the covariate $a_i = x_i \log(x_i)$ to a model already containing x_i
2. let $\widehat{\gamma}$ be the estimated coefficient of a_i and test to see if it is significantly different from zero (using a t-test say)
   (a) if not significant, no transform is needed
   (b) if significant, a preliminary guess at the proper transform is given by $\widehat{\lambda} = \frac{\widehat{\gamma}}{\widehat{\beta}} + 1$, where $\widehat{\beta}$ is the estimated coefficient of x in the model containing both x and a.

Why does this technique work?

Consider a true model,

$$y = \alpha + \beta x^{\lambda} + \varepsilon.$$

We will use Taylor's theorem to expand x^{λ} around $\lambda = 1$. Recall from calculus that Taylor's theorem says

$$f(b) = f(a) + f'(a)(b-a) + \frac{f''(a)}{2!}(b-a)^2 + \ldots.$$

In addition, recall that $\frac{\partial}{\partial x} c^x = c^x \log(c)$ for $c > 0$. Using these facts,

$$\begin{aligned} f(\lambda) = x^{\lambda}, \quad & \frac{\partial}{\partial \lambda} f(\lambda) = x^{\lambda} \log(x) \\ f(\lambda = 1) = x, \quad & f'(\lambda = 1) = x \log(x) \end{aligned}$$

Using a first-order Taylor series expansion with $b = \lambda$, $a = 1$,

$$x^{\lambda} \approx x + (\lambda - 1) x \log(x).$$

Substituting this into the above model, we have

$$\begin{aligned} y &= \alpha + \beta(x + \lambda x \log(x) - x \log(x)) + \varepsilon \\ &= \alpha + \beta x + \beta(\lambda - 1) x \log(x) + \varepsilon \\ &= \alpha + \beta x + \gamma x \log(x) + \varepsilon, \end{aligned}$$

where $\gamma = \beta(\lambda - 1)$. If γ is not significant, then either $\beta = 0$ (no effect of the covariate x) or $\lambda = 1$ (no need for a transformation). If γ is significant, then we solve to get $\lambda = \frac{\gamma}{\beta} + 1$.

Next: Model Selection

Reading Assignment:

- Muller and Fetterman, Chapter 11: "Selecting the Best Model"
- Weisberg, Chapter 10: "Model Selection"