Sections 3.9 and 6.8: Transformations

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

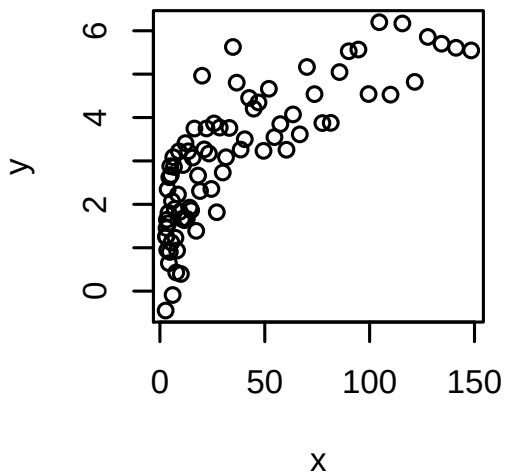
Examples of transformations for predictors are:

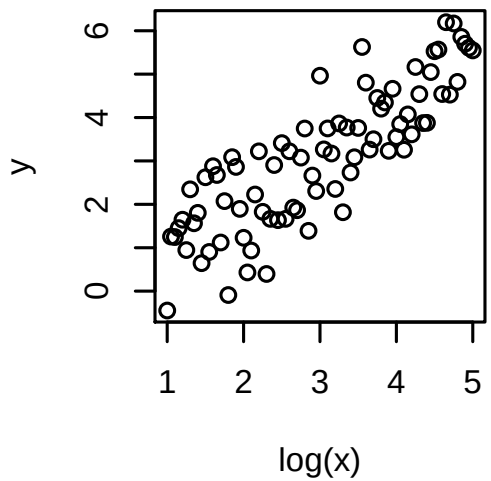
- $x^* = \log(x)$
- $x^* = \sqrt{x}$
- $x^* = 1/x$
- $x^* = \exp(x)$ or $x^* = \exp(-x)$

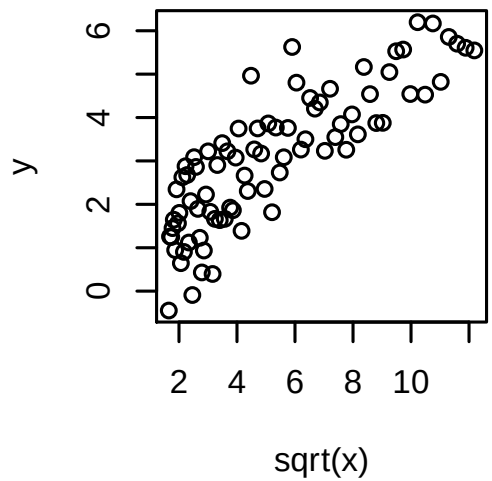
See Figure 3.13, page 130.

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.

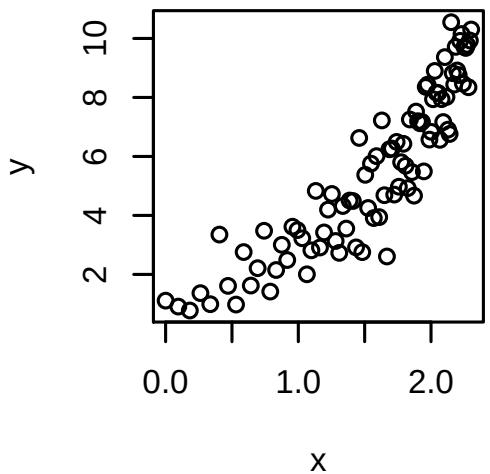
Example 1: transforming a predictor

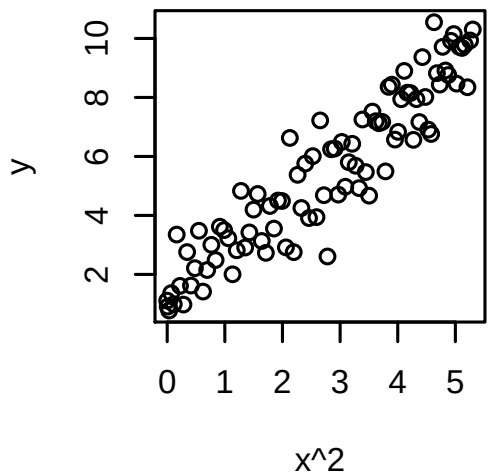


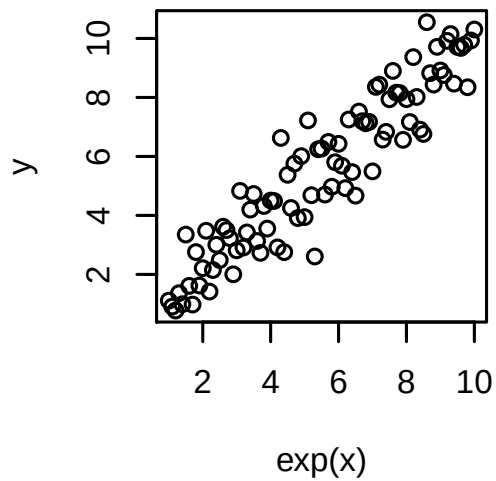




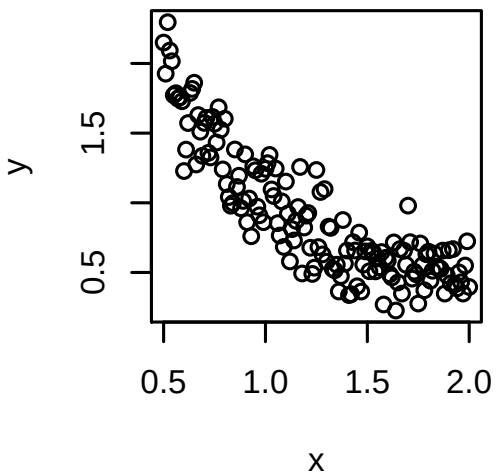
Example 2: transforming a predictor

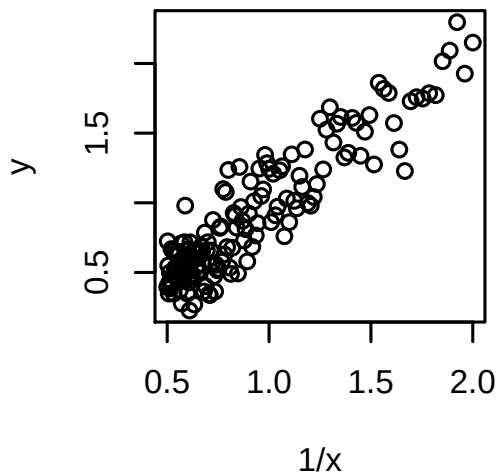


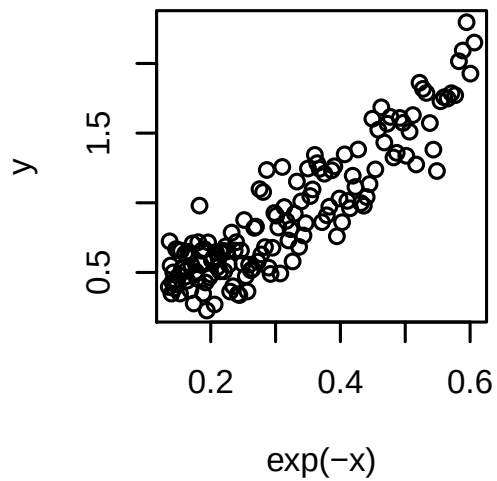




Example 3: transforming a predictor







Transforming the response

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

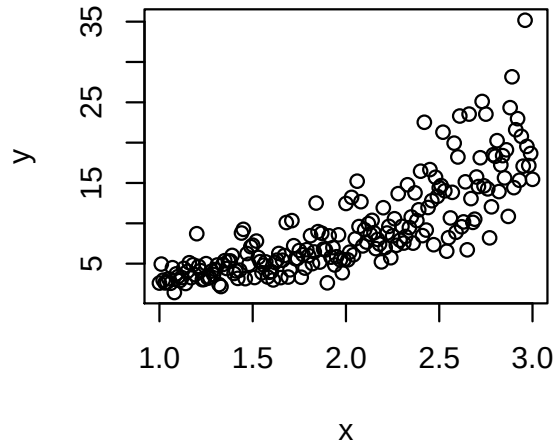
- $Y^* = \log(Y)$
- $Y^* = \sqrt{Y}$
- $Y^* = 1/Y$

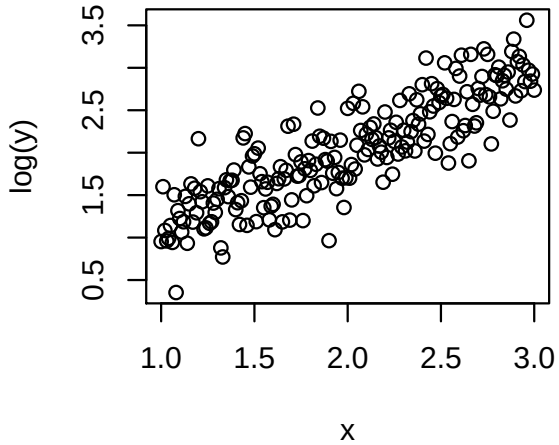
See Figure 3.15, page 132.

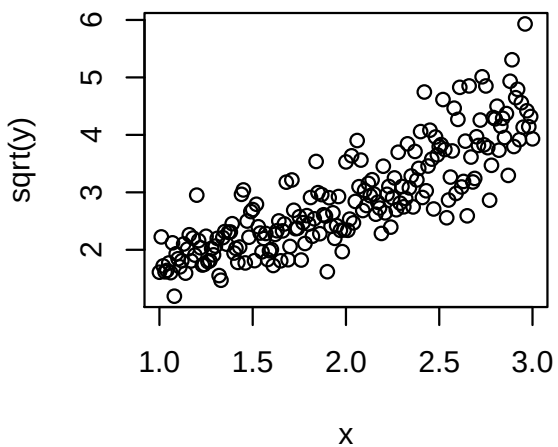
All of these are included in the Box-Cox family of transformations.

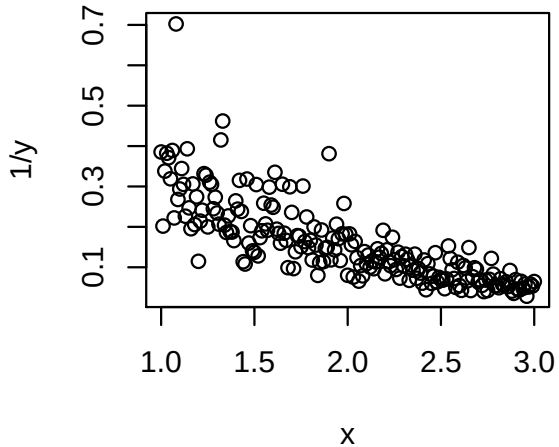
For some data, a transformation in Y may be followed by one or more transformations in the $x_{i1}, \ldots, x_{ik}$.

Example 4: transforming the response

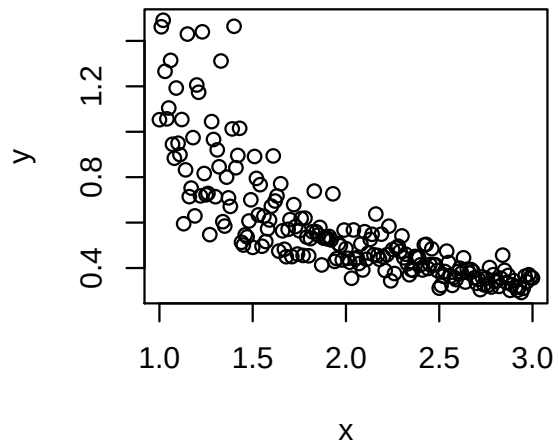


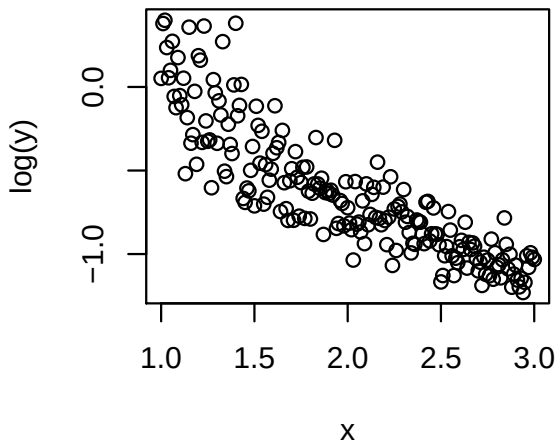


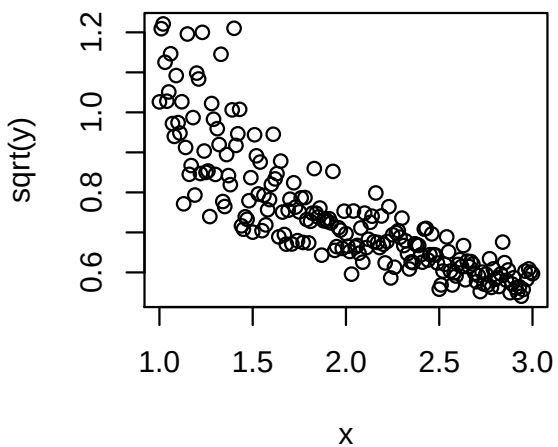


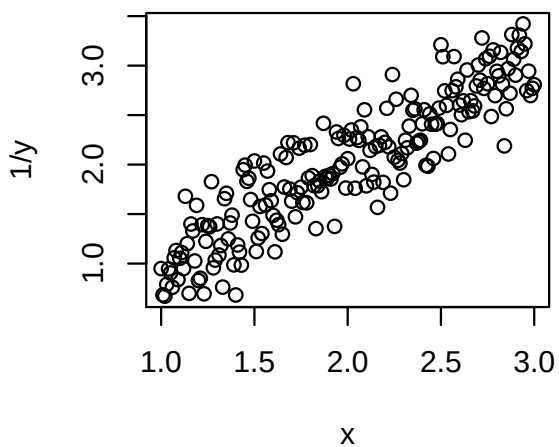


Example 5: transforming the response









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.

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.

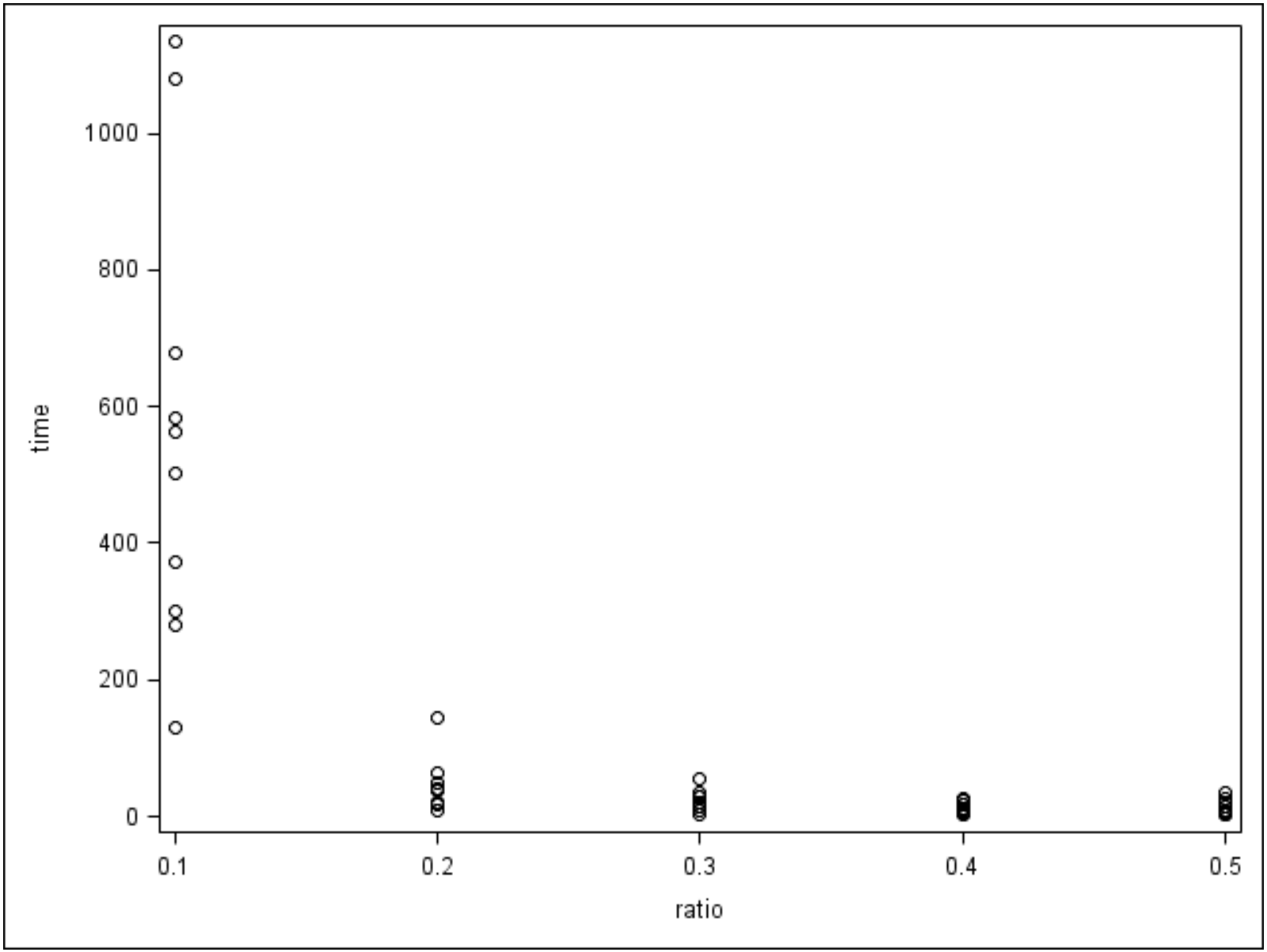
Example: Electrical components

- 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.

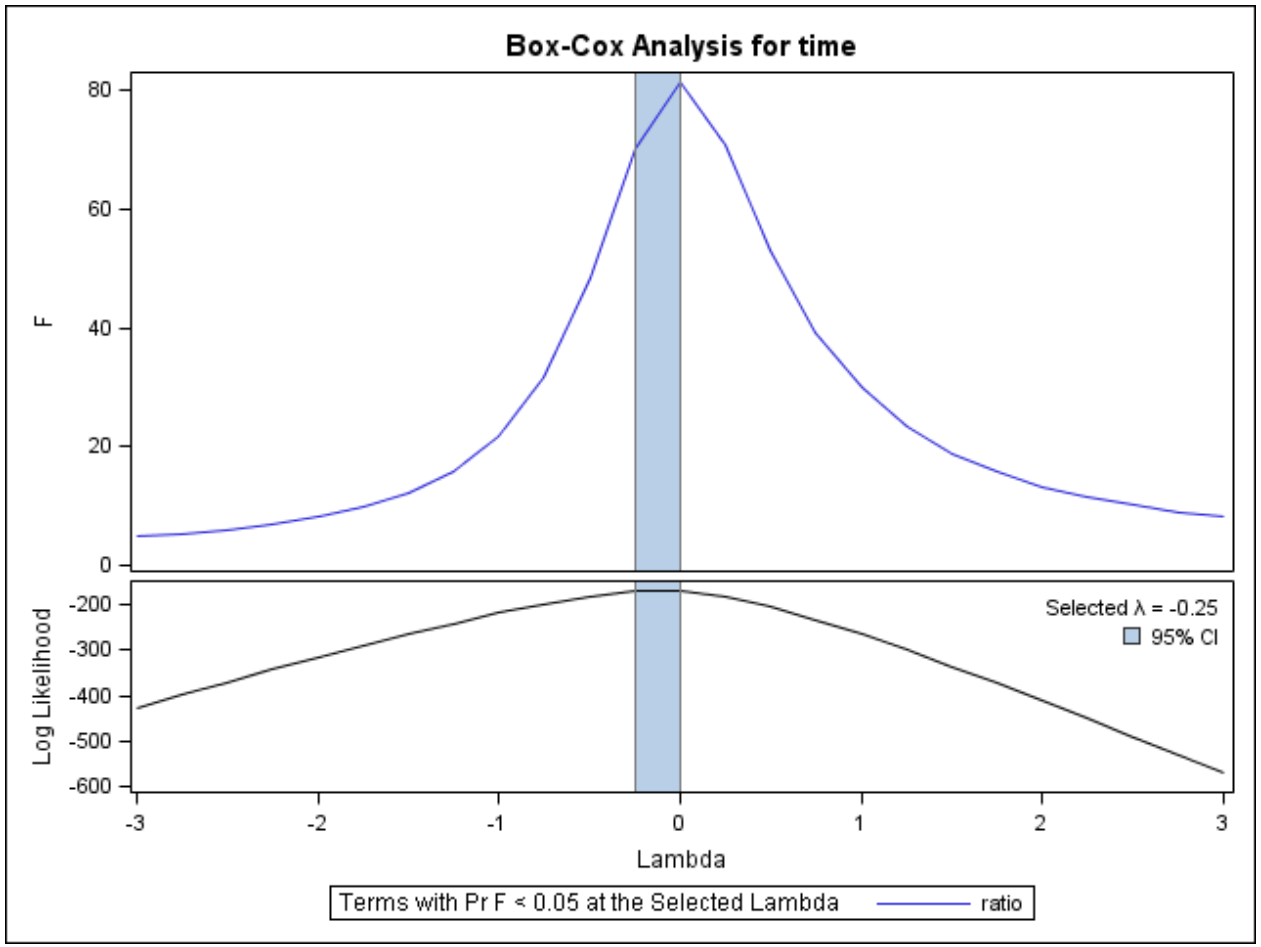
SAS code: data & plot

```
data elec;
input ratio time @@;
datalines;
0.5     34.9 0.5      9.3 0.5      6.0 0.5      3.4 0.5     14.9
0.5      9.0 0.5     19.9 0.5      2.3 0.5      4.1 0.5     25.0
0.4     16.9 0.4     11.3 0.4     25.4 0.4     10.7 0.4     24.1
0.4      3.7 0.4      7.2 0.4     18.9 0.4      2.2 0.4      8.4
0.3     54.7 0.3     13.4 0.3     29.3 0.3     28.9 0.3     21.1
0.3     35.5 0.3     15.0 0.3      4.6 0.3     15.1 0.3      8.7
0.2      9.3 0.2     37.6 0.2     21.0 0.2    143.5 0.2     21.8
0.2     50.5 0.2     40.4 0.2     63.1 0.2     41.1 0.2     16.5
0.1    373.0 0.1    584.0 0.1   1080.1 0.1    300.8 0.1    130.8
0.1    280.2 0.1    679.2 0.1    501.6 0.1   1134.3 0.1    562.6
;

proc sgscatter; plot time*ratio; run; * non-constant variance;

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

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


time
1000
800
600
400
200
0
0.1
0.2
0.3
0.4
0.5
ratio

Box-Cox plot


Box-Cox Analysis for time
F
80
60
40
20
0
Log Likelihood
-200
-300
-400
-500
-600
Selected λ = -0.25
95% CI
-3
-2
-1
0
1
2
3
Lambda
Terms with Pr F < 0.05 at the Selected Lambda
ratio

Box Cox output chooses log(Y) as "convenient"

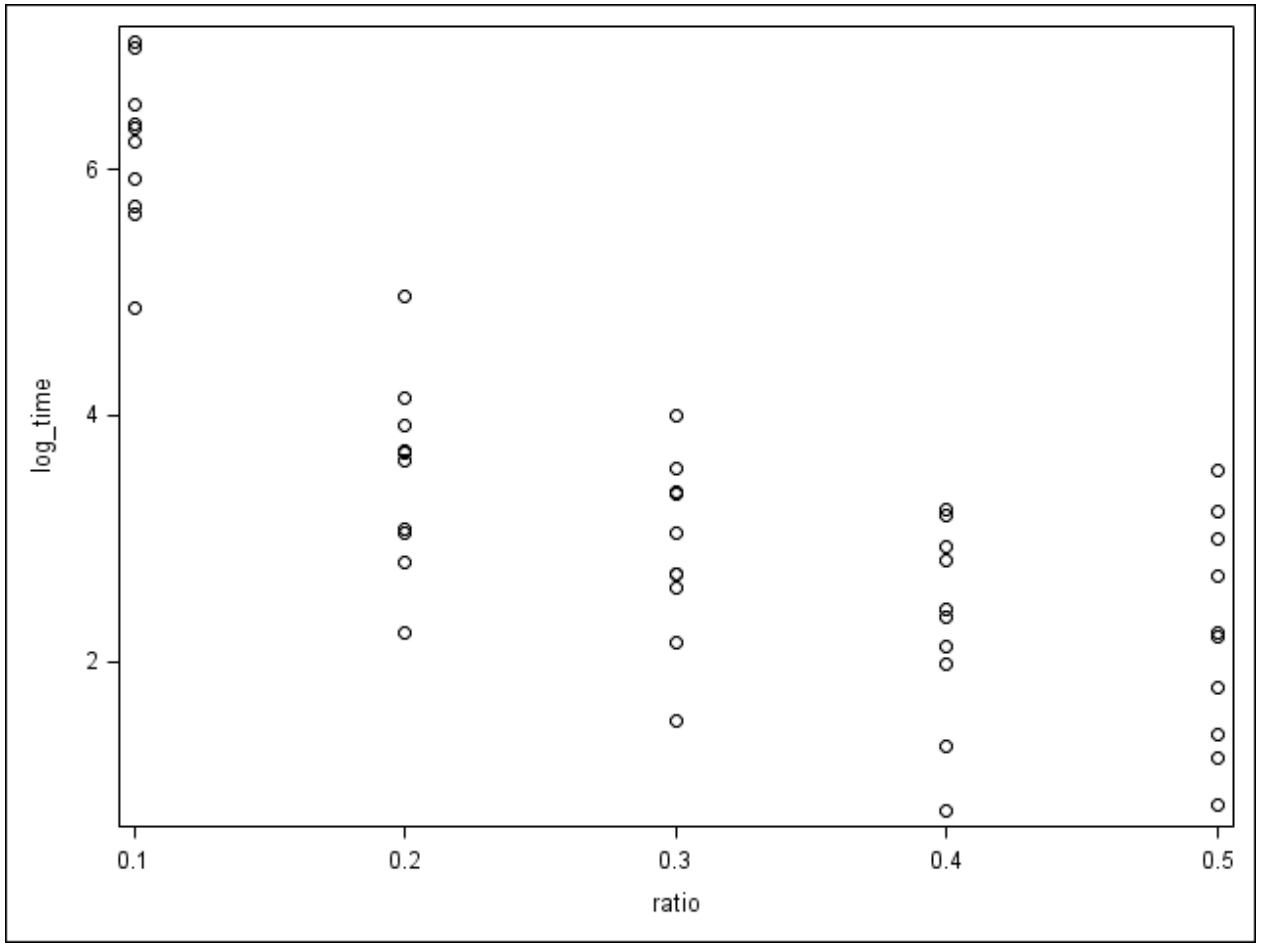
```
Box-Cox Transformation Information for time

 Lambda      R-Square    Log Like
  -3.00          0.09    -426.733
  -2.75          0.10    -398.478
  -2.50          0.11    -370.675
  -2.25          0.13    -343.407
  -2.00          0.14    -316.779
  -1.75          0.17    -290.917
  -1.50          0.20    -265.983
  -1.25          0.25    -242.186
  -1.00          0.31    -219.829
  -0.75          0.40    -199.453
  -0.50          0.50    -182.292
  -0.25          0.59    -171.350 <
   0.00 +        0.63    -171.615 *
   0.25          0.60    -184.969
   0.50          0.53    -207.524
   0.75          0.45    -235.373
   1.00          0.38    -266.617
   1.25          0.33    -300.324
   1.50          0.28    -335.898
   1.75          0.25    -372.902
   2.00          0.22    -411.006
   2.25          0.19    -449.964
   2.50          0.17    -489.595
   2.75          0.16    -529.764
   3.00          0.15    -570.371

< - Best Lambda
* - 95% Confidence Interval
+ - Convenient Lambda
```

Transformed response log(Y) fixed non-constant variance

Add

```
log_time=log(time);
```

in data step and plot again.

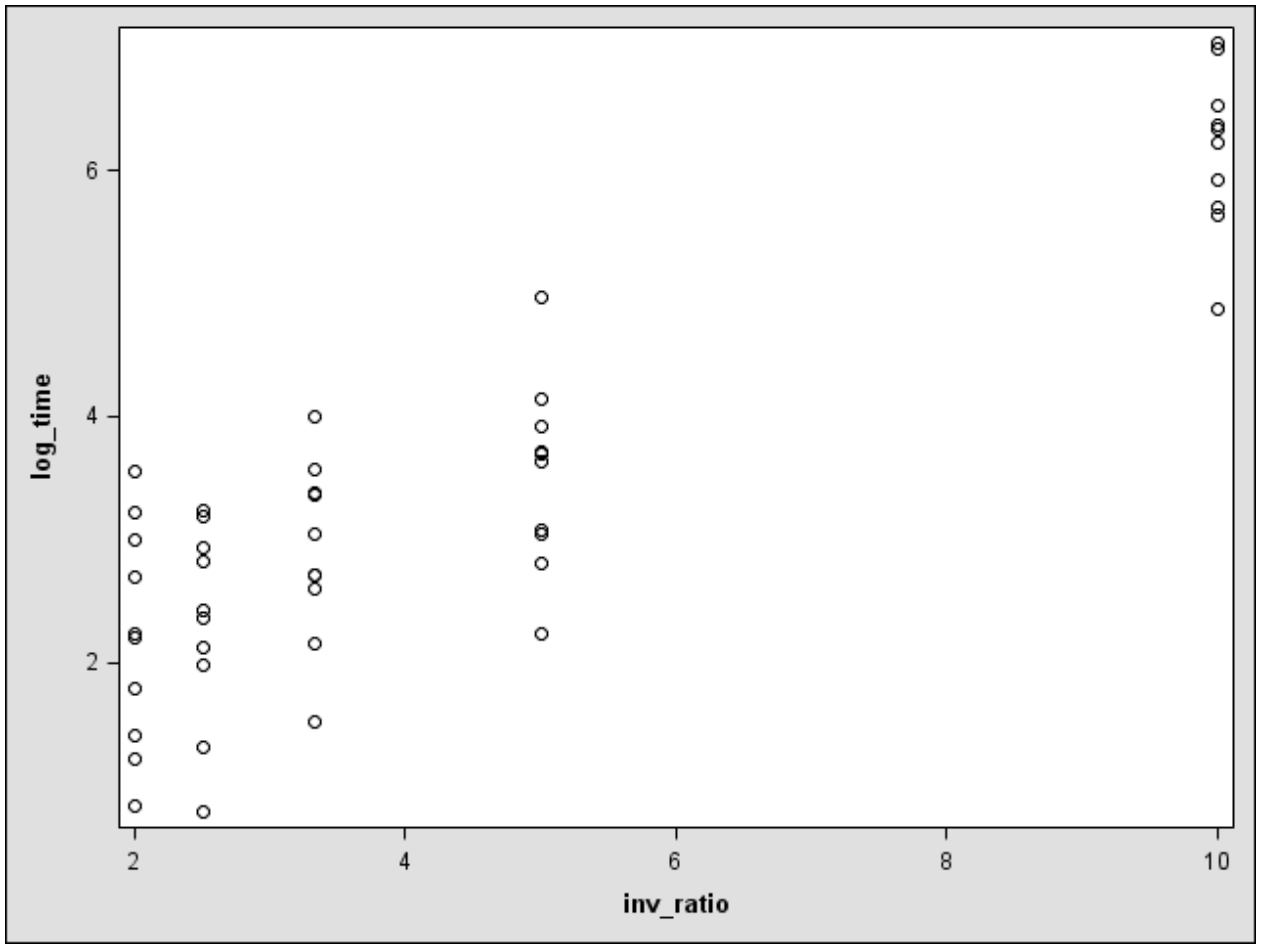


Try transforming the predictor using $1/x$

Add

```
inv_ratio=1/ratio;
```

in data step and plot again.

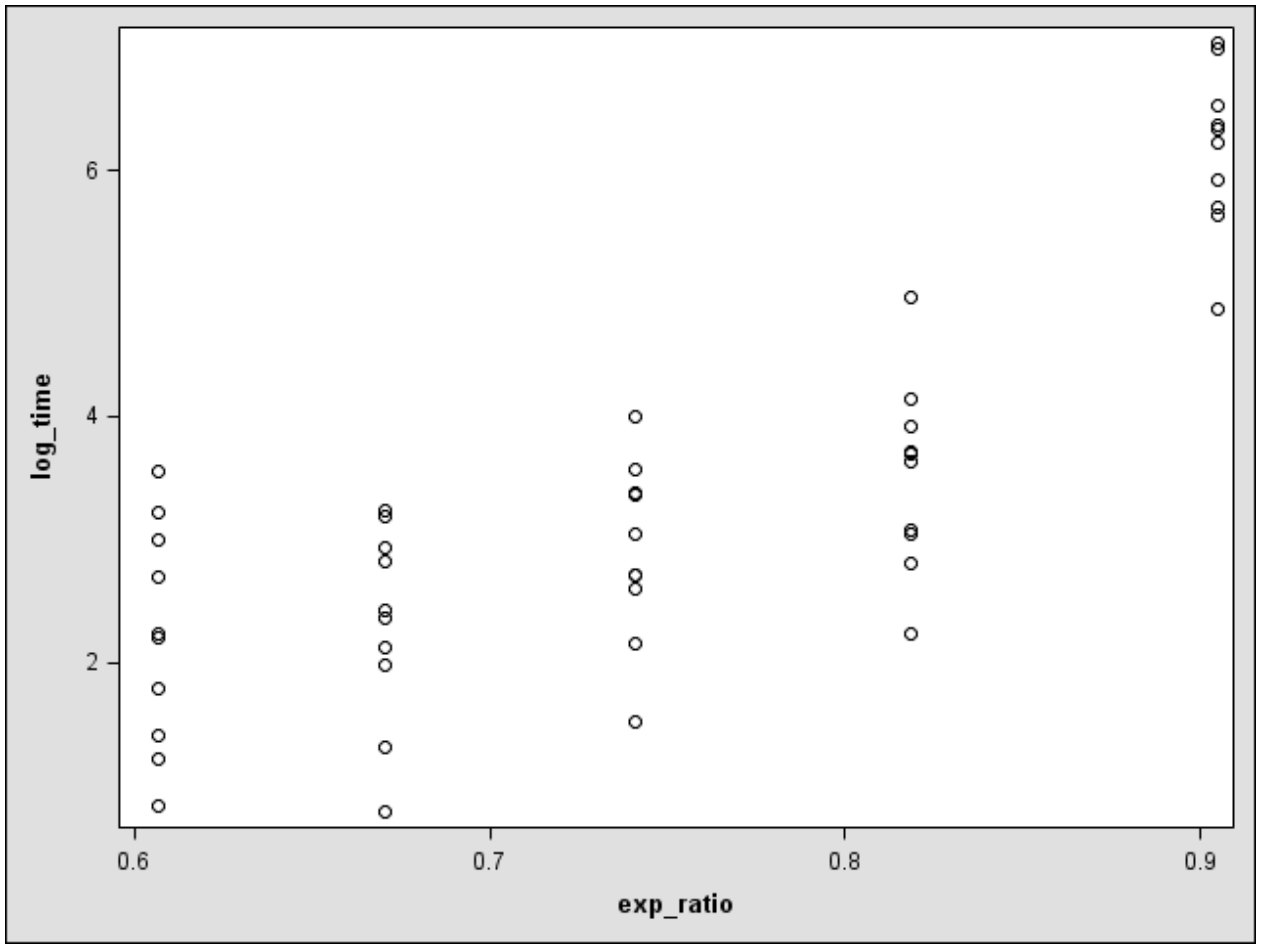


Try transforming the predictor using $\exp(-x)$

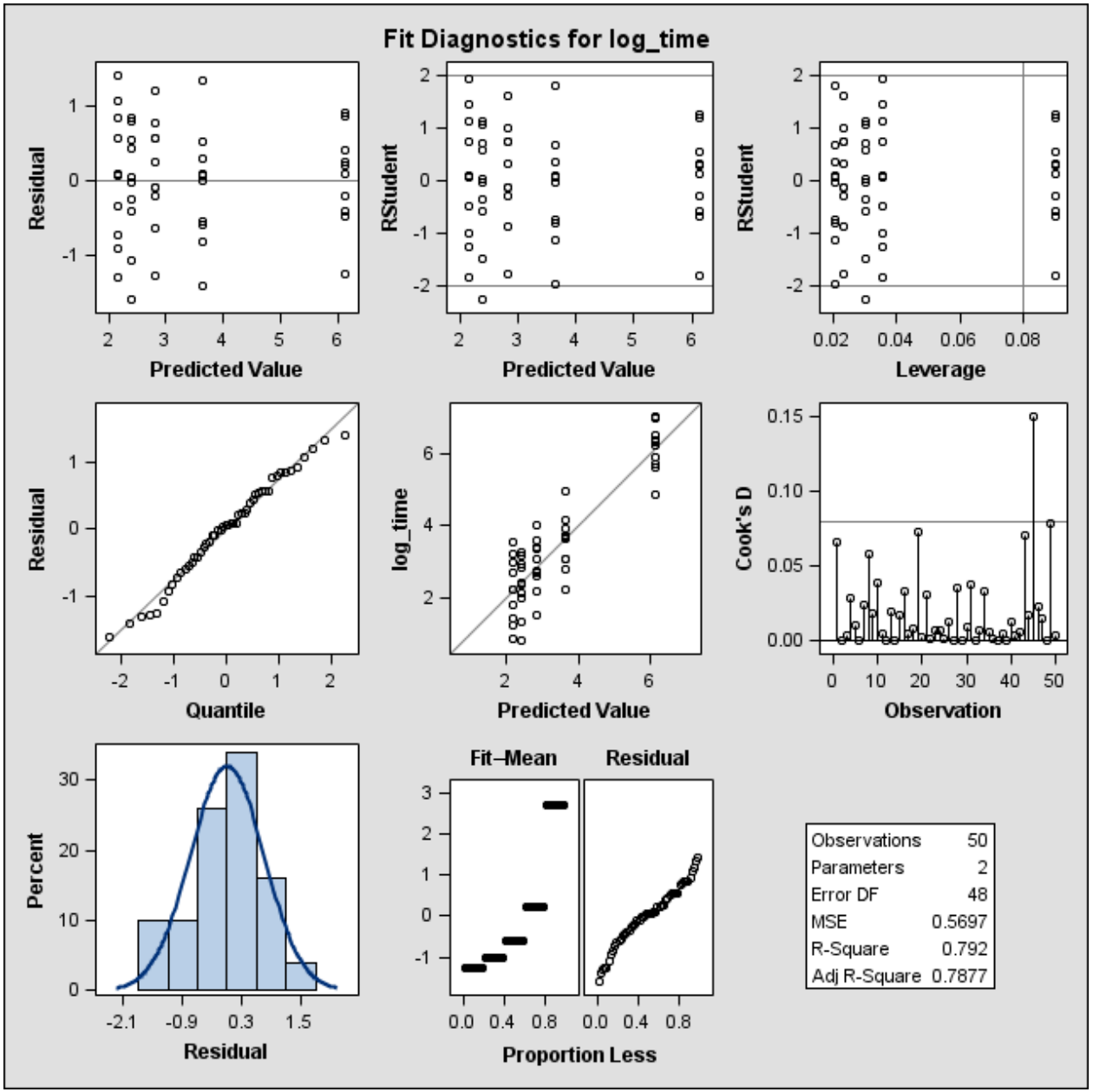
Add

```
exp_ratio=exp(-ratio);
```

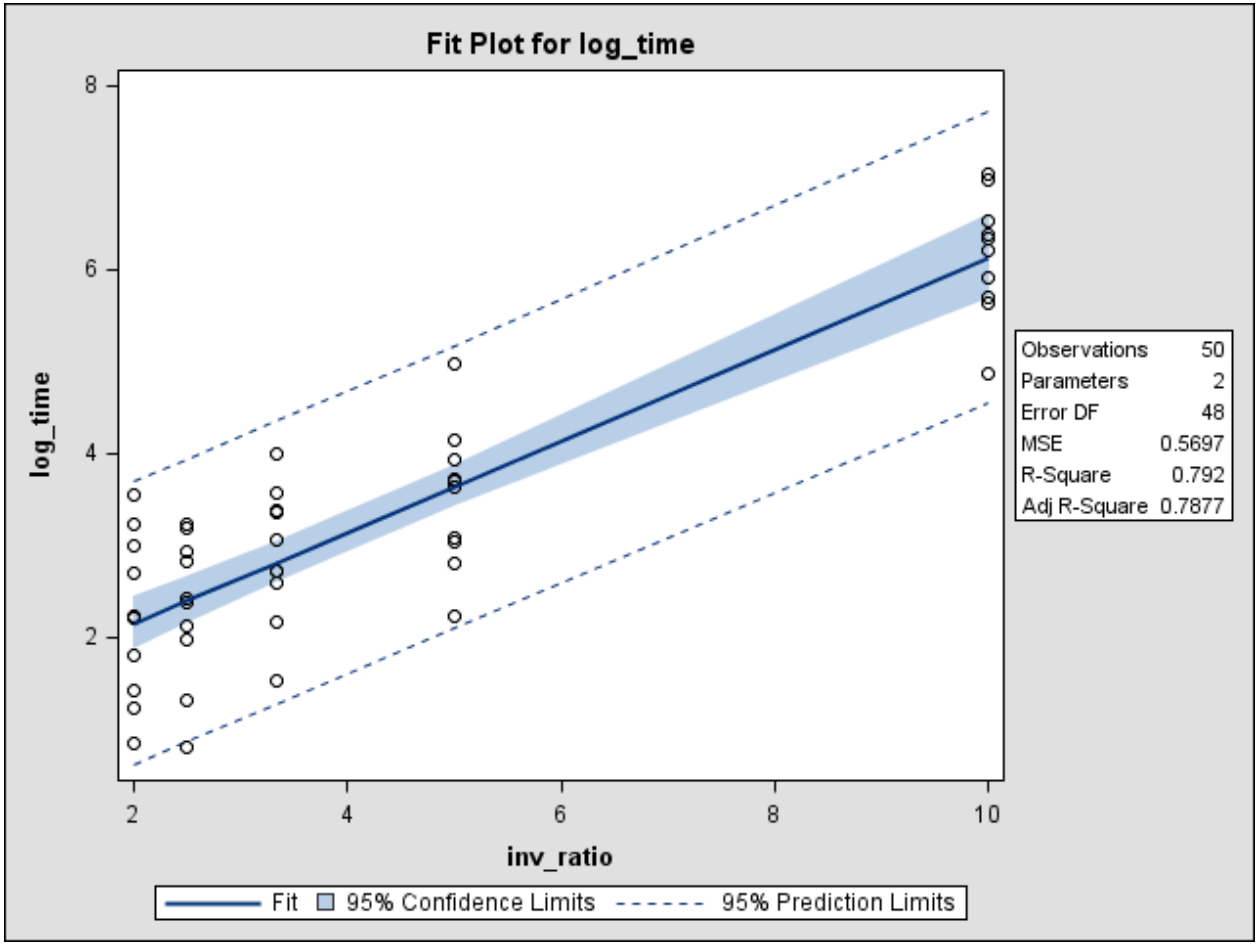
in data step and plot again.



Diagnostics from regressing $\log(Y)$ on $1/x$



Fit on transformed data



proc reg output

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.15322	0.19905	5.79	<.0001
inv_ratio	1	0.49738	0.03679	13.52	<.0001

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: Surgical unit example

Section 9.2: Surgical unit example (pp. 350–352)

- x_1 = blood-clotting score
- x_2 = prognostic index
- x_3 = enzyme function test score
- x_4 = liver function test score
- x_5 = age (years)
- x_6 = gender (0=male, 1=female)
- x_7 = alcohol use ($x_7 = 1$ indicates moderate)
- x_8 = alcohol use ($x_8 = 1$ indicates severe)
- Y = survival time (days?) after liver operation

For no alcohol use, $x_7 = x_8 = 0$ (baseline).