Dealing with Nonlinearity and Unequal Variance

22S.152 Applied Linear Regression

Lecture 15
Nov. 3, 2004

Kate Cowles, Ph.D.

Reasons why transformation of variables may be needed in linear regression

• Theory suggests a linear relationship when either the response or the predictor (or both) are transformed.
  — relationship between boiling point of water and barometric pressure
• Theory suggests that variance of response variable is a function of the mean.
  — this leads to a violation of which assumption of linear regression?
• no known theoretical basis, but examination of residuals suggests need for transformation to correct violation of linearity, homoscedasticity, or normality assumptions.

Linearizable simple regression functions

• $Y = \alpha X^\beta$

• $Y = \alpha e^{\beta X}$

• one-compartment pharmacokinetic model?

Example of fitting a model when the theoretical relationship is linearizable

• Bache, Serum, Youngs and Lisk (1072) reported data on concentrations of PCBs (a cancer-causing pollutant) in trout from Lake Cayuga, NY.

• variables
  — $Y = \text{PCB concentration in ppm}$
  — $X = \text{age of trout in years}$
• theoretical considerations suggest the model
  $$Y_i = e^{\beta_0} e^{\beta_1 X_i^{1/2}} e^{e_i}$$

• linearized form
  $$\log_e Y_i = \beta_0 + \beta_1 X_i^{1/2} + e_i$$
* trout.sas ;

options linesize = 75 ;
data trout ;
infile 'trout.dat' ;
input age pcb ;
logpcb = log(pcb) ;
sqrtage = sqrt(age) ;
run ;
proc print ;
run ;

proc plot ;
plot pcb * age = '.' / vpos = 20 hpos = 40 ;
run ;

proc reg ;
model pcb = age ;
run ;
The REG Procedure
Model: MODEL1
Dependent Variable: pcb

Analysis of Variance

Sum of Mean
Source      DF    Squares   Square    F Value  Pr > F
Model        1  822.54734  822.54734   30.80   <.0001
Error       26  694.42980  26.70884
Corrected Total 27 1516.97714

Root MSE  5.16806  R-Square  0.5422
Dependent Mean  7.17143  Adj R-Sq  0.5264
Coef Var     72.6647

Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|----------|----------------|---------|------|-------|
| Intercept|    | -1.48194 | 1.83535        | -0.79   | 0.4360 |      |
| age      |    | 1.55777  | 0.28071        | 5.55    | <.0001 |      |

NOTE: 6 obs hidden.
proc reg;
model logpcb = sqrtage;
run;

The REG Procedure
Model: MODEL1
Dependent Variable: logpcb

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>24.46336</td>
<td>24.46326</td>
<td>95.56</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>26</td>
<td>6.65630</td>
<td>0.26601</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>27</td>
<td>31.11966</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE = 0.5598
R-Square = 0.7861
Dependent Mean = 1.46893
Adj R-Sq = 0.7779
Coeff Var = 34.51563

Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|----------|----------------|---------|------|---|
| Intercept | 1  | -1.19475 | 0.28849        | -4.14   | 0.0003 |
| sqrtage  | 1  | 1.19861  | 0.12262        | 9.78    | <.0001 |
Plotting the Back Transformed Predicted Values

- To look at the nonlinear relationship between PCB concentration and age
- Transform predicted values from regression equation using transformations back to their original scale
  \[ \hat{y}^* = e^{\hat{y}} \]
- Plot these back transformed values vs. age (the untransformed predictor variable)

Example: the snow goose data

Aerial survey methods are regularly used to estimate the number of snow goose in their summer range areas west of Hudson Bay in Canada. To obtain estimates small aircraft fly over the range, and when a flock of goose is spotted, an experienced person estimates the number of goose in the flock. To investigate the reliability of this method of counting, an experiment was conducted in which an airplane carrying two observers flew over \( n = 45 \) flocks, and each observer made an independent estimate of the number of birds in each flock. Also, a photograph of the flock was taken so that an exact count of the number of birds in each flock could be made. The resulting data are given in the dataset "snowgees.dat". [Cooka ed Jacobson, 1978].

The variables are:
- photo = count based on photo
- obs1 = observer 1's count
- obs2 = observer 2's count

Variance stabilizing transformations

- Sometimes scientific or statistical theory indicates that the variance of the response variable varies with either
  - value of response variable itself
  - value of predictor
- Example: if response variable is counts of something then variance is proportional to expected value
  - square root transformation will stabilize variance
  - \( \sqrt{Y} \)
  - \( \sqrt{Y} + \sqrt{Y} + 1 \) if some \( Y_i \)'s are zero or very small
Regression without transformation

Plot of photo vs obs1. Symbol used is '.'.

```
proc reg;
model photo = obs1 ;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: photo

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>254769</td>
<td>254769</td>
<td>129.20</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>43</td>
<td>8790</td>
<td>202.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected</td>
<td>44</td>
<td>333560</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 44.4056
R-Square 0.7503
Dependent Mean 89.3111
Adj R-Sq 0.7445
Coeff Var 49.72%

Parameter Estimates

| Parameter | DF | Estimate | Std Error | t Value | Pr > |t| |
|-----------|----|----------|-----------|---------|------|---------|
| Intercept | 1  | 26.6497  | 8.6144    | 3.09    | <.005 |
| obs1      | 1  | 0.8850   | 0.0776    | 11.37   | <.001 |
Regression with square root transformations of response variable and predictor

```
\begin{verbatim}
proc reg;
    model y = x1 x2;
    run;

Dependent Variable: y
Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>123.45678</td>
<td>123.45678</td>
<td>234.56</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>44</td>
<td>56.78901</td>
<td>56.78901</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>45</td>
<td>180.12345</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>1.2345</td>
<td>0.0012</td>
<td>4.567</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>1</td>
<td>2.3456</td>
<td>0.0023</td>
<td>12.34</td>
<td>&lt;.001</td>
<td></td>
</tr>
</tbody>
</table>
```

---

Plot of y vs. y. Symbol used is '.'.