Statistical Techniques II

EXST7015

Curvilinear Regression

Curvilinear Regression

- As the name implies, these are regressions that fit curves.
- However, the regressions we will discuss are also linear models, so most of the techniques and SAS procedures we have discussed will still be relevant.

Curvilinear Regression *(continued)*

- We will discuss two basic types of curvilinear model.
  - The first are models that are not linear, but that can be "linearized" by transformation. These models are referred to as "intrinsically linear", because after transformation they are linear, often SLR.
  - Later we will cover polynomial regressions. These are an extraordinarily flexible family of curves that will fit almost anything. Unfortunately, they rarely have a good, interpretation of the parameter estimates.

Curvilinear Regression *(continued)*

- Intrinsically linear models
  - These are models that contain some transformed variable, logarithms, inverses, square roots, sines, etc.
  - We will concentrate on logarithms, since these models are some of the most useful.
- What is the effect of taking a logarithm of a dependent or independent variable? For example, instead of $Y_i=b_0+b_1X_i+e_i$, fit $\log(Y_i)=b_0+b_1X_i+e_i$

Curvilinear Regression *(continued)*

- If we fit $\log(Y_i) = b_0 + b_1X_i + e_i$
  - Then the original model, before we took logarithms, must have been $Y_i=b'_0\textit{exp}^{b1Xi}e_i$
    - Where "*exp*" is the base of the natural logarithm (2.718281828)
  - This model is called the "Exponential Growth model" if b_1 is positive, or the exponential decay model if it is not.
  - It is used in the biological sciences to fit exponential growth (+b_1) or mortality (-b_1).

Curvilinear Regression *(continued)*

- Exponential model

$$Y_i = b_0 \exp^{b_1 X_i} e_i$$

Blue

$b_0=34$

$b_1=-0.0953$

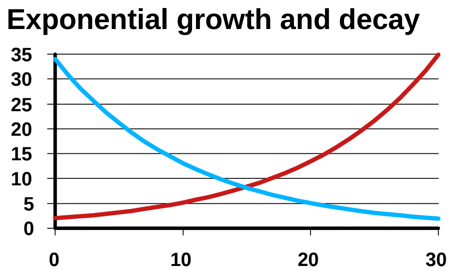
$e^{b1}=0.909$

Red

$b_0=2$

$b_1=+0.0953$

$e^{b1}=1.1$



Curvilinear Regression *(continued)*

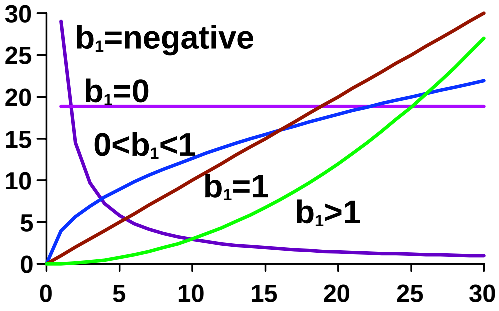
- Other examples of curvilinear models.
  - $\log(Y_i = b_0X_i^{b1}e_i)$ produces $\log(Y_i)=b_0+b_1\log(X_i) + \log(e_i)$
- This model is used to fit many things, including morphometric data,
- A model with an inverse ($1/X_i$) will fit a "hyperbola", with it's asymptote.
  - $Y_i = b_0 + b_1(1/X_i) + e_i$

Curvilinear Regression *(continued)*

- Power model

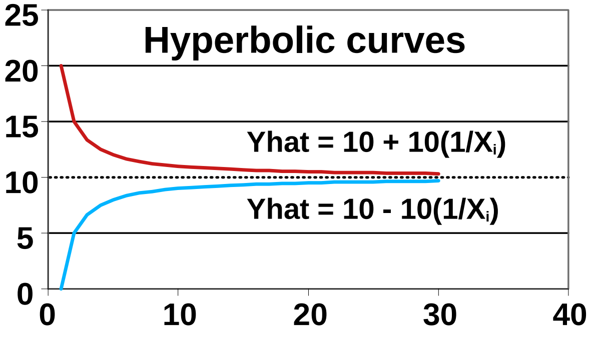
$$Y_i = b_0 X_i^{b_1} e_i$$

b0, b1
29, -1
19,0
4, 0.5
1,1
0.03, 2



Curvilinear Regression *(continued)*

- Hyperbolic model: $Y_i = b_0 + b_1(1/X_i) + e_i$
  - note that b_0 fits the asymptote



Curvilinear Regression *(continued)*

- These are a few of many possible curvilinear regressions. Models including power terms, exponents, logarithms, inverses, roots, and trigonometric functions fit may be curvilinear.
-

Curvilinear Regression *(continued)*

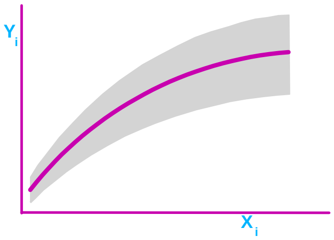
- However, not all are curves can be fitted by linear models with transformations. Some are nonlinear, and require nonlinear curve fitting techniques.
- For example,
  - $Y_i = b_0X_i^{b1}e_i$ is curvilinear
  - $Y_i = b_0X_i^{b1}+e_i$ is nonlinear
  - $Y_i = b_0 + b_1X_i + b_2X^2_i + e_i$ is linear (polynomial)
  - $Y_i = b_0 + b_1X_i + b_2X_i^{b3} + e_i$ is nonlinear

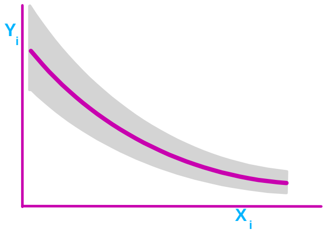
Curvilinear Regression *(continued)*

- Note that $Y_i = b_0X_i^{b1}e_i$ has an error multiplied by X_i. This is interesting because when the error is multiplied by the independent variable, the variance about the regression line should appear to increase as X_i increases.
- The log transformation (of Y_i) should remove this nonhomogeneous variance.
- This is not true for the log transformation of X_i.

Curvilinear Residual Patterns

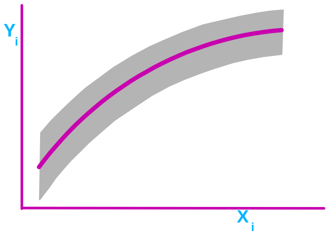
- Transformations of Y_i, like log transformations, will affect homogeneity of variance. The raw data should actually appear nonhomogeneous.

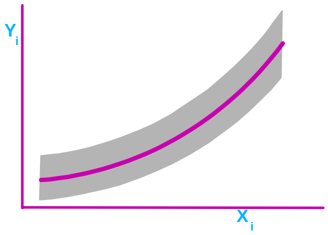




Curvilinear Residual Patterns *(continued)*

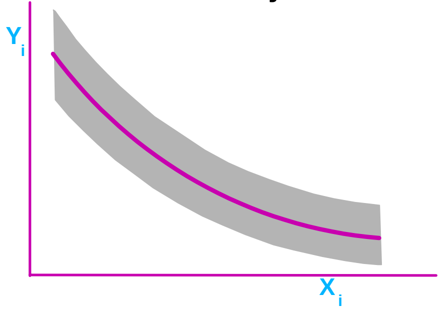
- Transformations of X_i will not.





Curvilinear Residual Patterns *(continued)*

- Polynomials assume homogeneous variance and will not adjust variance.



Curvilinear Regression Examples

- Air speed example.
- A small example from The Science of Flight by Peter P. Wagener, Am Sci, volume 74,(3),May-June 1986, page 274.
- The author fitted a quadratic model to this data (we will later). However, many examples of technological development over time follow an "exponential" model. So, we will fit an exponential model to this example.

Curvilinear Regression Examples

- Air speed example. I digitized the following data from a graph.
- Like the author, I omit values after 1963 (speed changes little once jet age reached).

YEAR	SPEED	AIRCRAFT
1926	108	Ford 5-AT
1932	150	247D
1935	179	DC-3
1939	200	307 Strat
1941	204	DC-4
1942	292	L-749
1946	304	DC-6
1947	283	Convair 2
1947	292	377 strat
1950	308	DC-6B
1952	354	DC-7
1954	304	Viscount
1951	458	Comet
1958	404	L188A Ele
1957	550	707/DC-8
1964	500	BAC1-11-2
1963	571	727

Curvilinear Example 1 *(continued)*

- The exponential is a logical and interesting model for this data.

```
Dependent Variable: LOGSPEED
                               Sum of           Mean
Source             DF         Squares         Square    F Value      Pr>F
Model               1      3.11456238     3.11456238     145.18    0.0001
Error              15      0.32179173     0.02145278
Corrected Total    16      3.43635410

      R-Square             C.V.          Root MSE         LOGSPEED Mean
      0.906357         2.579479          0.146468              5.678188

Source      DF        Type I SS     Mean Square    F Value      Pr > F
YR           1       3.11456238      3.11456238     145.18      0.0001
Source      DF      Type III SS     Mean Square    F Value      Pr > F
YR           1       3.11456238      3.11456238     145.18      0.0001
                                 T for H0:     Pr > |T|    Std Error of
Parameter          Estimate     Parameter=0                  Estimate
INTERCEPT      4.750697794           56.04       0.0001      0.08477711
YR             0.041602463           12.05       0.0001      0.00345273
```

Curvilinear Example 1 *(continued)*

- For exponential models the slope is interpreted as a "proportional" or percentage increase per X variable unit.
- To find the percentage value per X_i unit, assess EXP(b_1) = exp(0.0416) = 1.0425. So there was an average annual increase in speed of 4.25%.

Curvilinear Example 1 *(continued)*

- Note that I adjusted years (YR = YEAR - 1925) so that 1925 is year zero. Otherwise the zero value would be 1 BC.
- Another use of exponential models is to calculate doubling times or half-life values. $Y_i = b_0$ at time = $X_i = 0$, so how long does it take to get to speed=$2b_0$?
- Set $2b_0=b_0\textit{exp}^{b1X'}$, $2=\textit{exp}^{b1X'}$, $\log(2)=b_1X'$, $\log(2)/b_1 = X'$, $X' = 0.693/0.0416 = 16.67$.
- So speed doubled every 16.67 years.

Curvilinear Example 1 *(continued)*

- Exponential models. What can I say?
  - Good fit,
  - Few d.f. (basically a SLR),
  - clear interpretation.
- I like them!
- A note on logarithms. This model requires natural logs. In SAS the function "LOG()" gives natural logs (LOG10 gives log base 10). In EXCEL the natural log function is "LN()".

Curvilinear Example 2

- Remember our SLR example about the amount of wood harvested from trees, predicted on the basis of DBH (diameter at breast height)?
- Remember that it looked a little curved, and maybe even had NONHOMOGENEOUS variance?
- Let's take another look at that model.

Curvilinear Example 2 *(continued)*

- Typically, morphometric relationships (between parts of an organism) are best fitted with Log(Y),Log(X) models.
  - Fish length - scale length
  - Fish total length - fish fork length
  - Crab width - crab length
  - Fish length - fish weight, etc.
- Here we have tree diameter and tree weight. Lets try a log-log model (using natural logs).

Curvilinear Example 2 *(continued)*

- Since we are fitting a linear measurement to a volumetric or weight measurement, I expect the following relationship to apply.
  - 1 gm = 1 c^3, for the metric system if the material has the same density as water (specific gravity = 1).
- In other words, I expect
  - Wood weight = b_0(wood length)^3
  - log(Wood weight) = log(b_0) + 3(wood length)

Curvilinear Example 2 *(continued)*

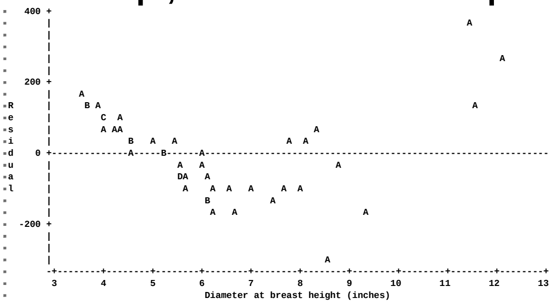
- Therefore, I will test the coefficient of DBH to see if it equals 3.
- Results for the SLR were,

```
Analysis of Variance     Sum of         Mean
Source      DF      Squares       Square      F Value     Prob>F
Model        1  6455979.821  6455979.821      433.487     0.0001
Error       45 670190.73220  14893.12738
C Total     46 7126170.5532                 (R-square = 0.9060)

Parameter Estimates
                Parameter      Standard    T for H0:
Variable  DF     Estimate         Error   Parameter=0    Prob>|T|
INTERCEP   1  -729.396300    55.69366336       -13.097      0.0001
DBH        1   178.563714    8.57640103         20.820      0.0001
```

Curvilinear Example 2 *(continued)*

- Reasons for concern (of model adequacy) were the large negative intercept, low R^2 and residual plot.



Curvilinear Example 2 *(continued)*

- The residual plot shows possible curvature and nonhomogeneous variance.
- Tests of normality were not a cause of concern. From SAS version 8, we get.

```
Shapiro-Wilk       W     0.973389    Pr < W      0.3544
Kolmogorov-Smirnov  D     0.084574    Pr > D     >0.1500
Cramer-von Mises    W-Sq  0.044081    Pr > W-Sq  >0.2500
Anderson-Darling    A-Sq  0.354877    Pr > A-Sq  >0.2500
```

- Now lets look at the log-log model results (called a "power model" in some disciplines).

Curvilinear Example 2 *(continued)*

- PROC REG output. Good fit, higher R^2.

```
                         Analysis of Variance
                                Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     1       35.94979       35.94979    1236.37    <.0001
Error                    45        1.30846        0.02908
Corrected Total          46       37.25825

Root MSE            0.17052    R-Square     0.9649
Dependent Mean      5.49466    Adj R-Sq     0.9641
Coeff Var           3.10337

                      Parameter Estimates
                   Parameter       Standard
Variable     DF     Estimate          Error    t Value    Pr > |t|
Intercept     1      0.55219        0.14275       3.87      0.0004
ldbh          1      2.79854        0.07959      35.16      <.0001
```

- Note that this line will go through the origin, no problem there. $Y_i=b_0X_i^{b1}e_i$

Curvilinear Example 2 *(continued)*

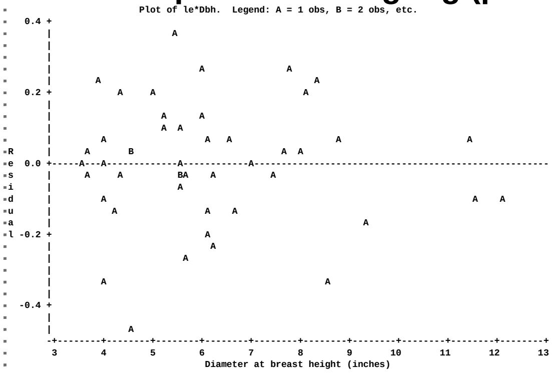
- The test of the slope against 3

```
Test 1 Results for Dependent Variable lweight
                         Mean
Source        DF     Square  F Value    Pr>F
Numerator      1    0.18629     6.41  0.0149
Denominator   45    0.02908
```

- Shows a significant difference, but not too far off.
- So how about the residual plot? It showed curvature and nonhomogeneous variance for the linear model.

Curvilinear Example 2 *(continued)*

- Residual plot for the log-log (power model).



Curvilinear Example 2 *(continued)*

- Residual from the transformed model were also tested for normality.

```
Tests for Normality
Test                  --Statistic---    -----p Value------
Shapiro-Wilk          W     0.979294    Pr < W      0.5634
Kolmogorov-Smirnov    D     0.128993    Pr > D      0.0483
Cramer-von Mises      W-Sq  0.069887    Pr > W-Sq  >0.2500
Anderson-Darling      A-Sq  0.396238    Pr > A-Sq  >0.2500
```

- The hypothesis of normality is not rejected for these results.

Curvilinear Example 2 *(continued)*

- The residual plot for the log-log model appears to show no curvature, no nonhomogeneous variance, no obvious outliers, and no significant departure from the normal distribution.
- In short, it is much improved, and probably fits better than the linear model.
- And it is interpretable.
- Geometrically, the model should be
  - Weight = C*π*(specific gravity)*$(D/2)^2$*H
  - where C=1 for a cylinder and 1/3 for a cone.

Curvilinear Regression Notes and Summary

- For transformed models,
  - The usual regression assumptions must be met for the transformed model, not the raw data (homogeneity, normality, etc.).
  - Estimates, hypothesis tests and confidence intervals would be calculated for the transformed model. The estimates and limits can then be detransformed.

Curvilinear Regression Summary *(continued)*

- A wide range of biometrics situations call for established curvilinear models. These would include, exponential growth, mortality, morphometric models, instrument standardization, some other growth models (power models and quadratics have been used), recruitment models.
- Check the literature in your field to see what models are used.