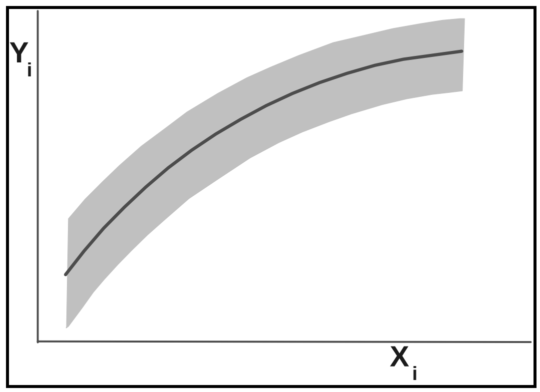
Transformations - your textbook takes a rather simplistic approach, does not get into modelling aspects or interpretation

Why transform? Data appears curved, LOF shows linear fit not adequate. Is the variance homogeneous? (Probably not).

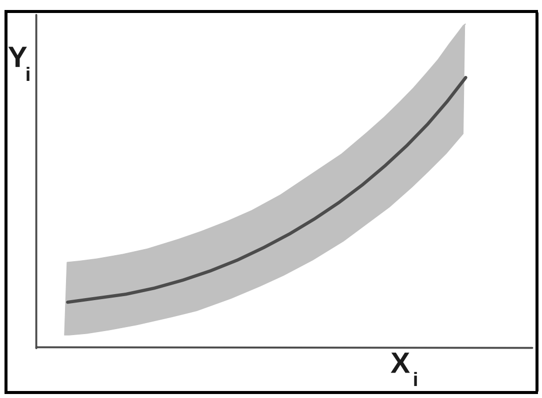
MAJOR CONSIDERATION - do you have any reason to suspect any particular curved, functional relationship? Your textbook mentions theoretical considerations late as a "comment".

Transformations of X, used to fit curvature when variance is homogeneous and normal

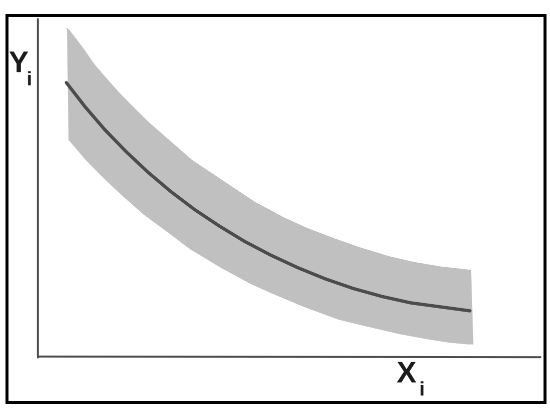
a) $\log(X_i)$ or $\ln(X_i)$ or $\sqrt{X_i}$



b) X_i^2 or e^{X_i}



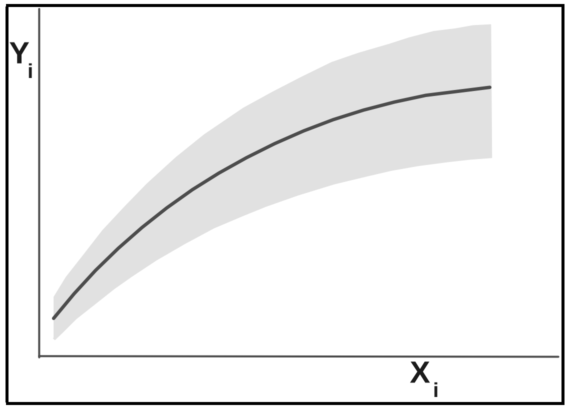
c) $\frac{1}{X_i}$ or e^{-X_i}

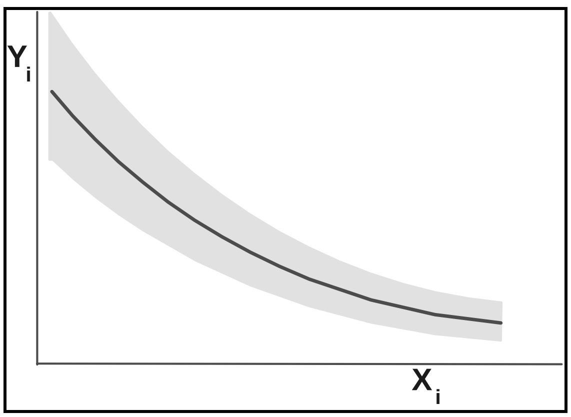


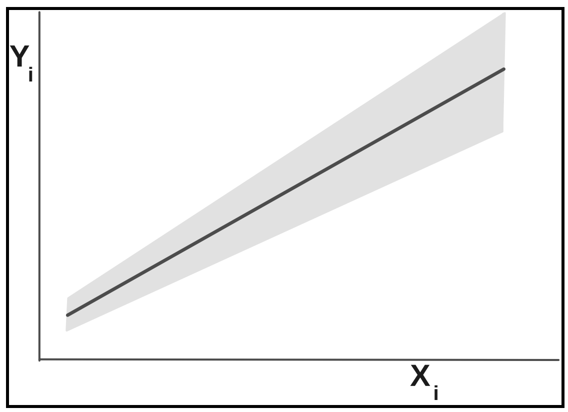
Transformations of Y, used to fit curvature when variance is NOT homogeneous and NOT normal

Do you ever EXPECT to see nonhomogeneous variance? If you use these models you do.

a) $\log(Y_i)$ or $\ln(Y_i)$ or $\sqrt{Y_i}$ or $\frac{1}{Y_i}$







The transformation of X and Y can also be done simultaneously. Generally there would be some rhyme and reason to this. Don't use a square root transformation of Y and a log transformation of X.

Box-Cox transformations : Choose the power of Y which minimizes the regression of Y^λ on X.

when

λ	= 2, then the transformation is	Y^2
	= 1	Y
	= 0.5	$\sqrt{Y}$
	= 0	log(Y)
	= $-$ 0.5	$\frac{1}{\sqrt{Y}}$
	= $-$ 1	$\frac{1}{Y}$
	= $-$ 2	$\frac{1}{Y^2}$

Note that since the Y values are transformed, the total SS and other SS are of differing magnitudes. The best way to compare is using a standardized values of some for. The recommended form is

$$W = K_1(Y^\lambda - 1) \text{ for } \lambda \neq 0 \quad \text{and} \quad K_2(\log_e Y) \text{ for } \lambda = 0$$

where

$$K_2 = \left[\prod_{i=1}^{n} Y_i\right]^{\frac{1}{n}} \quad \text{and} \quad K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$$

This approach is intended primarily as an **aid in model selection**, not as a way to find the perfect model.

In my judgment, theoretical considerations take precedence.

CURVILINEAR AND NONLINEAR REGRESSION

We will examine a number of Curvilinear techniques. POLYNOMIALS will come later.

We will also look at NONLINEAR techniques later, particularly to see how they differ from curvilinear techniques.

Nonlinear techniques are iterative. There is no simple, unique least squares solution, so different values of the parameters to be estimated are tried and adjusted until the best combination is found.

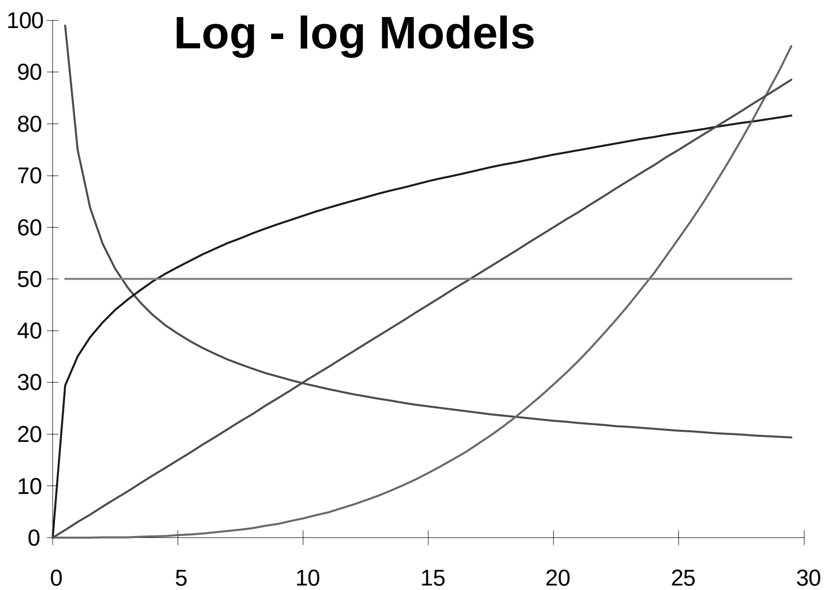
CURVILINEAR MODELS

1) $Y_i = b_0 X_i^{b_1} e_i$ log – log model

this model is used to fit meristic or morphometric relationships.
they are linearized by taking logarithms

$$\ln(Y_i) = \ln(b_0) + b_1 \ln(X_i) + \ln(e_i)$$

which is actually a simple linear regression



There are 4 models commonly used for morphometric relationships

Linear, “Direct proportion", Log-log and Polynomials

The correct model can often be determined by hypothesis testing.

a) Linear: if H_0: $\beta_0 = 0$ is not rejected then use Direct Proportion

b) Log-log: if H_0: $\beta_1 = 1$ is not rejected then use Direct Proportion

c) Polynomial: if H_0: $\beta_2 = 0$ is not rejected then use Linear

d) Polynomial: if H_0: $\beta_1 = 0$ is not rejected then use Direct Proportion

However, determination between Linear and Log-log cannot be made by hypothesis testing. Use other CONSIDERATIONS

a) Is the variance homogeneous? Yes → Linear; No → Log-log

b) Should the line go through the origin? No → Linear; Yes → Log-log

c) Does the line appear curved? No → Linear; Yes → Log-log

1) LENGTH - WEIGHT RELATIONSHIP

a) always use the model

$$W_i = \beta_0 L_i^{\beta_1} \epsilon_i$$

$$\text{Log}(W_i) = \text{Log}(\beta_0) + \beta_1 * \text{Log}(L_i) + \text{Log}(\epsilon_i)$$

b) in SAS

```
INPUT ... LT WT ...;
    LWT = LOG(WT);
    LLT = LOG(LT);
PROC GLM:
    MODEL LWT = LLT;
```

c) statistically all assumptions apply except,

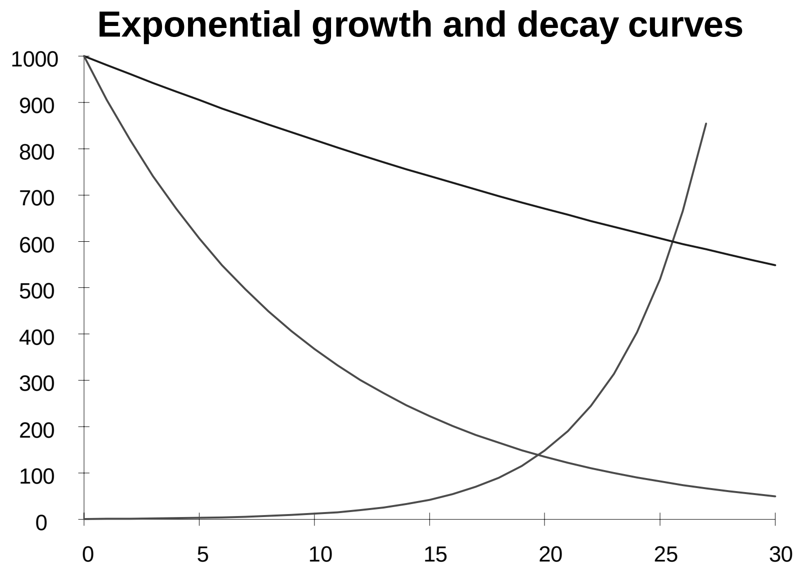
(a) multiplicative (non-homogeneous) error is implied for raw data because of the original model

(b) Ricker (1973) discusses the fact that length is probably not measured without error

2) $Y_i = b_0 e^{bX_i} e_i = N_0 e^{rt} e_i$

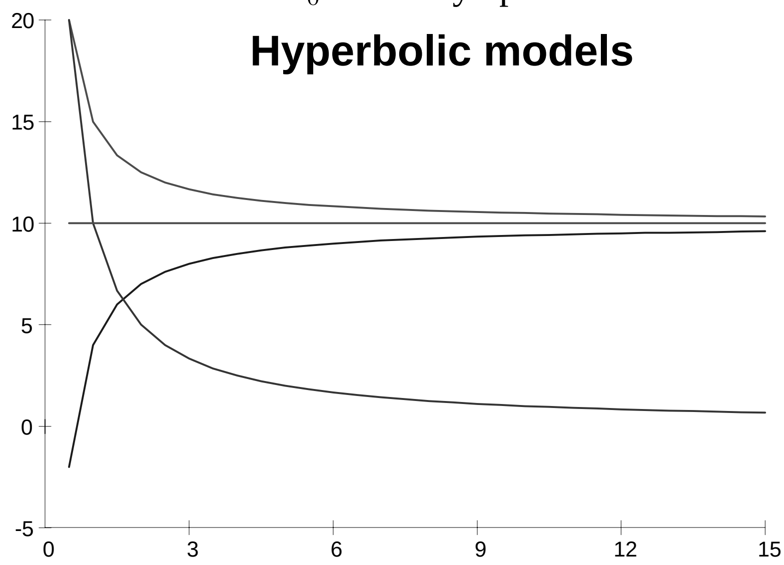
this models exponential growth (or decay => mortality)
used for modeling short term growth (+) or mortality (−)

linearize by logs $\Rightarrow$ $\ln(Y_i) = \ln(b) + bX_i + \ln(e_i)$



3) $Y_i = b_0 + b_1 \frac{1}{X_i} + e_i$

used for some recruitment models
hyperbolic model where b_0 is the asymptote



4) $Y_i = b_0 X_i e^{b_1 X_i} e_i$

This model is the Ricker Stock Recruit Model, it is linearized (best) by dividing through by X and then logs

$$\ln\left[\frac{Y_i}{X_i}\right] = \ln(b_0) + bX_i + e_i$$

NOTES on curvilinear models from transformations

Those curvilinear models shown above are
SIMPLE LINEAR REGRESSIONS AFTER TRANSFORMATION

The assumptions are made for the transformed models, not for the untransformed version, and the assumptions are the same as for any simple linear regression.

The tests of hypothesis and other diagnostics should be done on the transformed version of the model.

Confidence intervals are done on the transformed variables. Predicted values can be obtained from the transformed version and untransformed, or directly from the reconstructed original model. However, confidence limits should obtained from the transformed, linear model and detransformed.

Examples of Linear and Curvilinear Models

LINEAR MODELS -

$Y_i = b_0 + b_1X_i + e_i$ LINEAR

$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + e_i$ LINEAR

$Y_i = b + bX_{1i} + bX_{2i} + bX_{1i}X_{2i} + e_i$ LINEAR (interaction)

the cross product of independent variables, when used as an independent variable, for the interaction or the two variables. Since the value of both is known, this is also just another term in a multiple regression and still linear.

NONLINEAR MODELS	PARAMETER
$Y_i = b_0 + X_i^{b_1} e_i$	cannot take logarithms with addition sign
$Y_i = b_0 X_i^{b_1} + e_i$	log of addition, even for error term
$Y_i = b_0 e^{b_1 X_i} + e_i$	this is model fitted by NLIN, not GLM
$Y_i = L_\infty \left[1 - e^{-k(t-t_0)}\right]$	von Bertalanffy growth curve
$Y_i = L_0 e^{L_\infty (1 - e^{-kt})}$	Gompertz growth curve
$Y_i = \frac{L_\infty}{1 + \left(\frac{L_\infty - L_0}{L_0}\right) e^{-kt}}$	Logistic growth curve
$Y_i = b_0 + b_1X_i + b_2X_i^p + e_i$	with and UNKNOWN POWER

is a NONLINEAR model

$Y_i = b_0 + b_1X_i + b_2X_i^2 + b_3X_i^3 + e_i$ CURVILINEAR (polynomial)

However, if the power terms are known (eg. X^2, X^3, X^4, etc) then the model is linear (curvilinear), and can fit curves. Models containing powers from X=1 to any integer "k", including all values between are called POLYNOMIALS.