STAT 8230 — Applied Nonlinear Regression Lecture Notes

Linear vs. Nonlinear Models

Linear regression, analysis of variance, analysis of covariance, and most of multivariate analysis are concerned with linear statistical models.

These models describe the dependence relationship between one or more continuously distributed response random variables and a set of explanatory variables or factors.

- These models are **parametric** because, when fully specified, they assume that the probability distribution of the response variable(s), including a model for the dependence between response and explanatory variables, is known except for the values of a small number of unknown constants called parameters.

- These models are **linear** in the sense that the regression parameters (the parameters that describe the dependence of the mean response on explanatory variables) enter into the models linearly.

    - The model is linear in the parameters, not the explanatory variables.

For example, the following is the general form of the classical multiple regression model:

$$\begin{aligned} & y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i, \quad i = 1, \ldots, n, \\ & \text{and} \quad e_1, \ldots, e_n \overset{iid}{\sim} N(0, \sigma^2) \end{aligned} \tag{$*$}$$

- Here, we assume that we have a random sample of n observations $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, of a response variable Y and a set of p explanatory variables $X_1, \ldots, X_p$.

- In addition, the notation $\overset{iid}{\sim} N(0, \sigma^2)$ means, "are independent, identically distributed random variables each with a normal distribution with mean 0 and variance σ^2."

- Typically, x_{i1} is equal to one for all i in multiple linear regression models, but this need not be so.

- In model (*) the parameters are $\beta_1, \ldots, \beta_p, \sigma^2$. The regression parameters are $\beta_1, \ldots, \beta_p$.

In **regression models**, the explanatory variables ($x_{i1}, \ldots, x_{ip}$, $i = 1, \ldots, n$, above) are treated as nonrandom, either by assumption that they have been set to their observed values by design or some other nonrandom mechanism, or, more generally, by making the model conditional on the observed X vales.

- That is, regression models specify the conditional distribution of $Y|X_1, \ldots, X_p$. In particular, regression models are primarily concerned with the mean of this distribution: $\mathrm{E}(Y|X_1, \ldots, X_p)$.

- $\mathrm{E}(Y|X_1, \ldots, X_p)$, the conditional expectation of the response given the values of the explanatory variables, is known as the **regression function.**

  - Since we always condition on the explanatory variables in linear *and* nonlinear regression models, we will often drop the conditioning from the notation for convenience and write $E(Y)$ in place of $\mathrm{E}(Y|X_1, \ldots, X_p)$.

In the multiple linear regression model (*), the regression function for the i^{th} subject (unit of observation), call it μ_i, is

$$\begin{aligned}\mu_i \equiv \mathrm{E}(y_i) &= \mathrm{E}(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i) \\ &= \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \underbrace{\mathrm{E}(e_i)}_{=0} \\ &= \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\end{aligned}$$

Notice that the regression parameters $\beta_1, \ldots, \beta_p$ enter into the regression function in a linear fashion.

- Recall that a linear combination of $z_1, \ldots, z_k$ is a weighted sum $a_1 z_1 + a_2 z_2 + \cdots + a_k z_k$ of the z_j's with coefficients $a_1, \ldots, a_k$.
- Of course, the multiple linear regression model is linear in the β_j's and in the x_{ji}'s, but the fact that it is linear in the β_j's is what makes it a linear model.

In this course, a **nonlinear regression model** is still going to be a regression model describing the relationship between a continuously distributed response variable y_i and explanatory variables $x_{i1}, \ldots, x_{ip}$, but now we drop the linearity assumption,

$$\mu_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$

and allow the parameters $\theta_1, \ldots, \theta_p$ and explanatory variables $x_{i1}, \ldots, x_{ik}$ to enter into the regression function in a nonlinear way.

- Notice we've switched to calling the parameters θ's instead of β's. In addition, the number of explanatory variables, k, is not necessarily equal to the number of parameters, p.

That is, in the nonlinear regression models under study in this course, $\mu_i = f(x_{i1}, \ldots, x_{ik}, \theta_1, \ldots, \theta_p)$, where $f(\cdot)$ is a function not necessarily linear in the θ's. Otherwise, the model is the same:

$$y_i = f(x_{i1}, \ldots, x_{ik}, \theta_1, \ldots, \theta_p) + e_i, \quad i = 1, \ldots, n, \qquad (**)$$

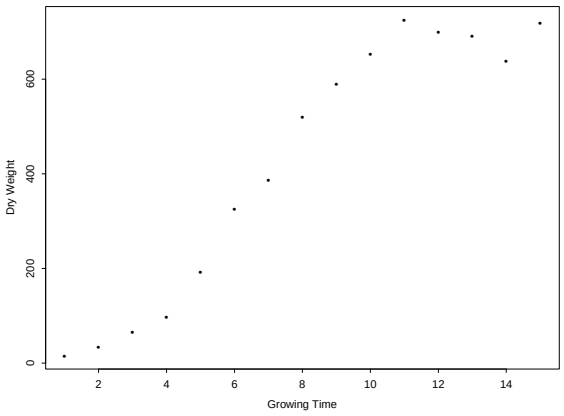
where $e_1, \ldots, e_n \overset{iid}{\sim} N(0, \sigma^2)$

- Note that we've restricted attention here somewhat in the class of all possible not-linear models. We retain the assumption of a continuous response, additive error, and (usually) normally distributed errors.
- Excluded from consideration are several important classes of regression models that are nonlinear in some sense.
  - In particular, we exclude generalized linear models (GLMs, e.g., logistic regression models, Poisson loglinear models, etc.) Many cases of GLMs are for discrete data, and GLMs retain a (modified) linearity in the parameters assumption.

Example — Onion Data:

The following table and scatterplot display data on the dry weight (Y) of 15 onion bulbs randomly assigned to 15 growing times (X) until measurement.

Growing Time	Dry Weight	Growing Time	Dry Weight
1	16.08	9	590.03
2	33.83	10	651.92
3	65.8	11	724.93
4	97.2	12	699.56
5	191.55	13	689.96
6	326.20	14	637.56
7	386.87	15	717.41
8	520.53		



Suppose we wanted to fit a model to describe how the mean dry weight of onions depends upon growing time. From the data and scatterplot, it is clear that weight tends to increase with growing time in a nonlinear (in growing time) fashion.

However, a linear (in the parameters) model can still be used to capture this nonlinear pattern of growth by considering polynomial models in growing time. That is, consider models of the form

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \cdots + \beta_p x_i^{p-1} + e_i, \quad i = 1, \ldots, n,$$

where $e_1, \ldots, e_n$ are i.i.d. each with mean 0, and variance σ^2 (constant variance).

Alternatively, we might consider a nonlinear model. In particular, consider the **3-parameter logistic model** (a.k.a. simple logistic model):

$$y_i = \frac{\theta_1}{1 + \exp\{(\theta_2 - x_i)/\theta_3\}} + e_i, \quad i = 1, \ldots, n,$$

with the same assumptions on the e_i's.

How do we choose between a linear model (e.g., polynomial in growing time) and a nonlinear model (e.g., simple logistic model) in this problem?

From a purely empirical point of view, we might choose the model that fits the data most closely.

However, we need to be a little bit careful here to balance fit against parsimony and generality. If we include enough terms in a polynomial model we can fit the data perfectly.

In particular, a $(n-1)^{\text{th}}$ degree polynomial can fit n points exactly. Such a model has n parameters and is equivalent to (is just a reparameterization of) the model

$$y_i = \beta_i + e_i, \quad i = 1, \ldots, n. \qquad (\dagger)$$

Such a model clearly doesn't *summarize* or *simplify* the data at all and can't be expected to generalize beyond the particular features of this one randomly drawn data set.

- In addition, a model such as (†) leaves no degrees of freedom left to estimate error variance $\Rightarrow$ can't do inference (test hypotheses, form confidence intervals) based on the model.

If we think of each of the n observations as an independent piece of information (or degree of freedom) from which to fit a model, then we use up one of these pieces of information (degrees of freedom) for every (nonredundant) parameter estimated in the model.

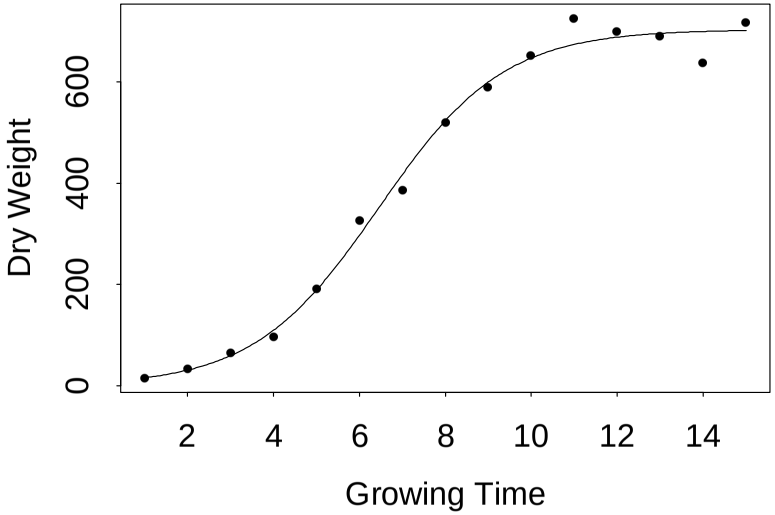
- n regression parameters to be estimated $\Rightarrow$ n d.f. used to estimate the model (model d.f.) $\Rightarrow$ $n - n = 0$ d.f. left to estimate the error variance parameter σ^2 (0 d.f. for error).

The smaller the number of regression parameters in the model, the more d.f. available to estimate error variance $\Rightarrow$ more power for hypothesis tests, more precision in confidence intervals.

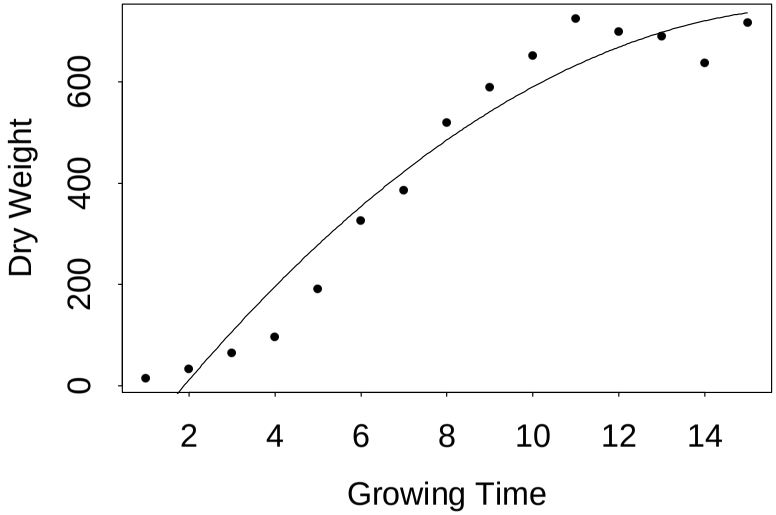
- Parsimonious models that fit the main features of the data are preferred.

Consider the fits of the simple logistic model and polynomial models of order 2, 3, and 4 to the onion data on the following page.

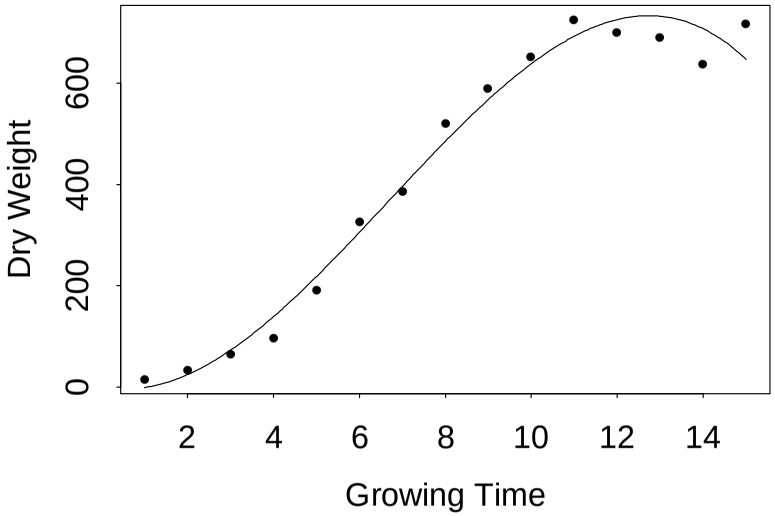
3-parameter logistic model (nonlinear



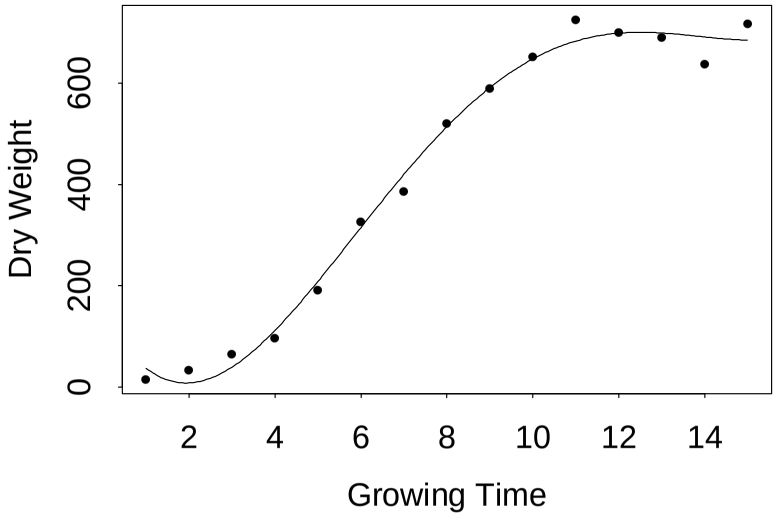
3-parameter linear model



4-parameter linear model



5-parameter linear model



- Notice that it requires a less parsimonious (more parameters) linear model to fit the main features of the data than for a nonlinear model.
- In addition, while the quadratic (3 parameter linear) model clearly underfits the general shape of the curve, the cubic and quartic linear models appear to overfit the data.
- So, from a purely empirical point of view, the logistic model appears preferrable.
- In addition, the logistic model has a big advantage in terms of parameter interpretability for a growth model such as this one.

Interpretations of the parameters in the simple logistic model:

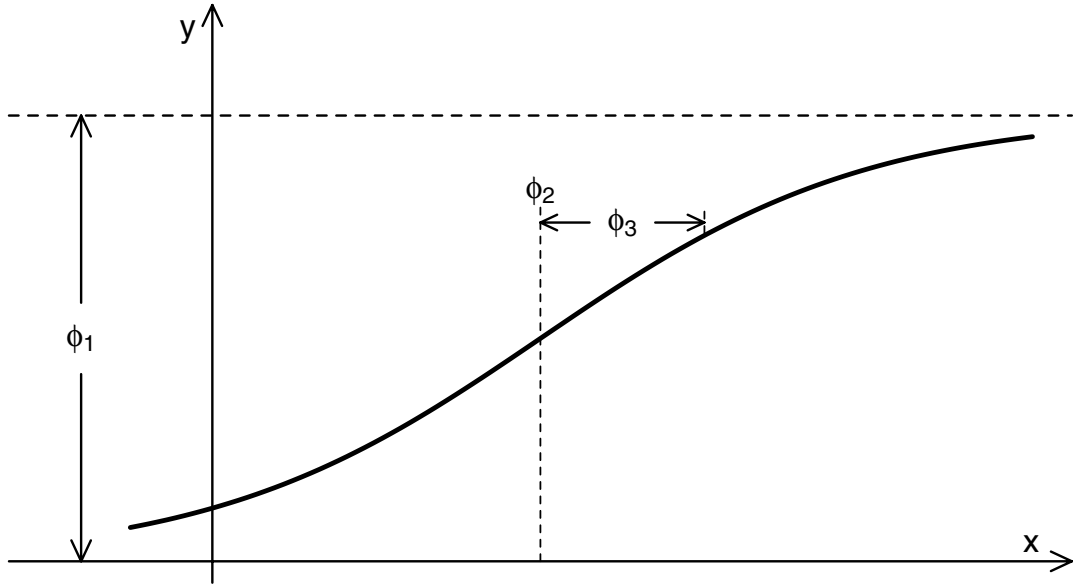


FIGURE C.7. The simple logistic model showing the parameters ϕ_1, the horizontal asymptote as $x \to \infty$, ϕ_2, the value of x for which $y = \phi_1/2$, and ϕ_3, a scale parameter on the x-axis. If $\phi_3 < 0$ the curve will be monotone decreasing instead of monotone increasing and ϕ_1 will be the horizontal asymptote as $x \to -\infty$.

- θ_1 (ϕ_1 in the plot on the previous page) represents the asymptote of the curve (limit of onion weight as growth time increases toward its maximum value).

- θ_2 represents the x-value at which y is equal to $\theta_1/2$, one-half of its asymptotic value. The growth time at which onions have achieved half of their total potential weight.

- θ_3 is a scale parameter that does not have as natural of an interpretation.

- In contrast, the polynomial parameters are not as meaningful to the context of the problem.

So fit, parsimony, and parameter interpretability can point to using nonlinear models over linear ones. A further important motivation for using nonlinear models over linear ones is subject matter theory.

- It may be that we have a *mechanistic* theory that explains the nature of onion growth and which implies a nonlinear functional form for the relationship between weight (or some other measure of size) and growing time.

- We will see plenty of examples of theoretically motivated nonlinear models as the course progresses.

Review of Linear Regression Models

Before discussing nonlinear regression we need to review linear regression.

Why?

- Because many of the ideas and methods from linear regression transfer directly or with minor modification to the nonlinear case.
- And because many of the methods of estimation and inference in NLMs are linear methods applied to a linear approximation to the NLM.

Again, we assume that we observe a sample of independent pairs, $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ where y_i is a response variable and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is a $p \times 1$ vector of explanatory variables (assumed fixed).

The classical linear model can be written

$$\begin{aligned} y_i &= \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i, \quad i = 1, \ldots, n, \\ &= \mathbf{x}_i^T \boldsymbol{\beta} + e_i, \end{aligned}$$

where $e_1, \ldots, e_n \overset{iid}{\sim} N(0, \sigma^2)$. Equivalently, we can stack these n equations and write the model as follows:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

$$\text{or} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Our assumptions on $e_1, \ldots, e_n$ can be equivalently restated as

$$\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

where $N_n(\boldsymbol{\mu}, \mathbf{M})$ denotes the $n-$dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{M}$.

Multivariate normal distribution:

The multivariate normal distribution is to a random vector (vector of random variables) as the univariate (usual) normal distribution is to a random variable. It is the version of the normal distribution appropriate to the joint distribution of several random variables (collected and stacked as a random vector) rather than the distribution of a single random variable.

- E.g., for a bivariate random vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ that has a bivariate normal distribution, the density function of $\mathbf{x}$ maps out a bell over the (x_1, x_2) plane.

The $N_n(\boldsymbol{\mu}, \Sigma)$ distribution is completely described by the two parameters $\boldsymbol{\mu}$, the mean of the distribution, and Σ, the **variance-covariance matrix** of the distribution.

- That is, for $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \sim N_n(\boldsymbol{\mu}, \Sigma)$,

$$\boldsymbol{\mu} = \begin{pmatrix} \mathrm{E}(x_1) \\ \mathrm{E}(x_2) \\ \vdots \\ \mathrm{E}(x_n) \end{pmatrix}, \quad \text{and} \quad \Sigma = \begin{pmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) & \cdots & \mathrm{cov}(x_1, x_n) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) & \cdots & \mathrm{cov}(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(x_n, x_1) & \mathrm{cov}(x_n, x_2) & \cdots & \mathrm{var}(x_n) \end{pmatrix}$$

  describe the location and dispersion (spread), respectively, of the bell-shaped distribution of possible values for $\mathbf{x}$.

The probability density function (p.d.f.) of $\mathbf{x}$ generalizes the univariate normal p.d.f.

- Recall for $X \sim N(\mu, \sigma^2)$ the p.d.f. of X is

$$f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}, \quad -\infty < x < \infty$$

- In the multivariate case, for $\mathbf{x} \sim N_k(\boldsymbol{\mu}, \Sigma)$, the p.d.f. of $\mathbf{x}$ is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \quad \mathbf{x} \in R^k.$$

  - Here $|\Sigma|$ denotes the *determinant* of the var-cov matrix Σ.

In the CLM, since $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ and $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, it follows that $\mathbf{y} \sim N_n$ too, with mean

$$\mathrm{E}(\mathbf{y}) = \mathrm{E}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = \mathbf{X}\boldsymbol{\beta} + \underbrace{\mathrm{E}(\mathbf{e})}_{=\mathbf{0}} = \mathbf{X}\boldsymbol{\beta}$$

and var-cov matrix

$$\begin{aligned}\mathrm{var}(\mathbf{y}) &= \mathrm{var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) \\ &= \mathrm{var}(\mathbf{e}) = \sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}. \qquad (*)\end{aligned}$$

- Assumption (*) says that the y_i's are uncorrelated ($\mathrm{cov}(y_i, y_{i'}) = 0$, for $i \neq i'$) and have constant variance ($\mathrm{var}(y_1) = \cdots = \mathrm{var}(y_n) = \sigma^2$).

So, the assumptions of the CLM can be stated quite succinctly as:

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n). \qquad (\dagger)$$

Therefore, in the CLM $\mathbf{y}$ is assumed to have the joint p.d.f.

$$\begin{aligned} f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) &= \frac{1}{(2\pi)^{n/2}|\sigma^2\mathbf{I}_n|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\sigma^2\mathbf{I}_n)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\| \underbrace{\mathbf{y}-\mathbf{X}\boldsymbol{\beta}}_{=\mathbf{e}} \|^2\right\}, \quad \mathbf{y} \in R^k. \end{aligned}$$

- Here, $\|\mathbf{v}\| = \sqrt{\mathbf{v}^T\mathbf{v}}$ denotes the norm, or length, of the vector $\mathbf{v}$. Therefore, $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|$ denotes the length of the difference between $\mathbf{y}$ and $\mathbf{X}\boldsymbol{\beta}$; i.e., the (Euclidean) distance between $\mathbf{y}$ and $\mathbf{X}\boldsymbol{\beta}$.

- (Note that we've used here the fact that the determinant of a diagonal matrix (a matrix whose off-diagonal elements are all 0) is the product of the diagonal elements $\Rightarrow |\sigma^2\mathbf{I}_n| = \sigma^{2n}$.)

- Actually, many of the results in the theory of linear models can be established with the weaker assumptions obtained by dropping normality from (†); that is, under the assumptions $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{I}_n$ and $y_1, \ldots, y_n$ are independent.

- As mentioned previously, the mean of $\mathbf{y}$ (conditional on $\mathbf{X}$) is known as the regression function or (as we'll call it) the **expectation function** of the model. In the linear regression model, the expectation function is $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$.

Notice that in the linear model,

$$\frac{\partial \mathrm{E}(y_i)}{\partial \beta_j} = \frac{\partial(\mathbf{x}_i^T \boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial(x_{i1}\beta_1 + \cdots + x_{ip}\beta_p)}{\partial \beta_j} = x_{ij}$$

To denote the derivative of the vector $\boldsymbol{\mu} = \mathrm{E}(\mathbf{y})$ with respect to the vector $\boldsymbol{\beta}$ we use the notation

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \boldsymbol{\beta}^T} \\ \frac{\partial \mu_2}{\partial \boldsymbol{\beta}^T} \\ \vdots \\ \frac{\partial \mu_n}{\partial \boldsymbol{\beta}^T} \end{pmatrix} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \frac{\partial \mu_1}{\partial \beta_2} & \cdots & \frac{\partial \mu_1}{\partial \beta_p} \\ \frac{\partial \mu_2}{\partial \beta_1} & \frac{\partial \mu_2}{\partial \beta_2} & \cdots & \frac{\partial \mu_2}{\partial \beta_p} \\ \cdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_n}{\partial \beta_1} & \frac{\partial \mu_n}{\partial \beta_2} & \cdots & \frac{\partial \mu_n}{\partial \beta_p} \end{pmatrix}$$

- So we see that the derivative of $\mathbf{X}\boldsymbol{\beta}$ with respect to $\boldsymbol{\beta}$ gives the matrix $\mathbf{X}$. For this reason $\mathbf{X}$ is called the **derivative matrix**.
  - Note that $\mathbf{X}$ is also sometimes called the *model matrix*, or *design matrix* in linear models.
  - In linear regression notice that the derivative matrix $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T}$ does not depend on $\boldsymbol{\beta}$. This will not be the case in nonlinear regression.

Estimation of $\boldsymbol{\beta}$ and σ^2:

Maximum likelihood estimation:

In general, the **likelihood function** is just the probability density function, but thought of as a function of the parameters rather than of the data.

- For Example, in the CLM, the p.d.f. of the response variable $\mathbf{y}$ is

$$f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\right\}$$

  - This p.d.f. is a function of the observed response $\mathbf{y}$ and the parameters $\boldsymbol{\beta}$ and σ^2, but we think of it primarily as a function of $\mathbf{y}$.
  - In the discrete case, the p.d.f. gives the probability of observing its argument, the data ($\mathbf{y}$ above), for given values of the parameters. In the continuous case the interpretation is very similar, but slightly more complicated.

Since the p.d.f. involves both the parameters ($\boldsymbol{\beta}$ and σ^2) and the data ($\mathbf{y}$), once the data are observed, we can think of it as a function of the parameters given the data.

This re-interpretation of the density function is given a new name, the likelihood function, and written as primarily a function of the parameters.

E.g., in the CLM the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\right\}$$

- The idea behind maximum likelihood estimation is to find the values of $\boldsymbol{\beta}$ and σ^2 under which the data are most likely. That is, we find the $\boldsymbol{\beta}$ and σ^2 that maximize the likelihood function (and p.d.f.) for the value of $\mathbf{y}$ actually observed. These values are the maximum likelihood estimators (MLEs) of the parameters.

- Note that since the natural logarithm is an increasing function, maximizing $L(\boldsymbol{\beta}, \sigma; \mathbf{y})$ with respect to the parameters is equivalent to (produces the same answer as) maximizing $\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) \equiv \log\{L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})\}$. Since taking logarithms is often mathematically convenient and it doesn't change the probelm, we'll typically work with this **loglikelihood function** rather than the likelihood function.

For the CLM, the loglikelihood is

$$\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{a constant}} \underbrace{-\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2}_{\text{kernel of } \ell}.$$

- Note that it is equivalent to maximize the **kernel** of the loglikelihood — that portion of the loglikelihood depending upon the parameters. Terms not involving the parameters can be ignored.

Obtaining the MLEs in the CLM: This can be done in two steps

1. Maximize $\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$ with respect to $\boldsymbol{\beta}$, treating σ^2 as known. Call the resulting estimator $\hat{\boldsymbol{\beta}}$.

2. Then maximize $\ell(\hat{\boldsymbol{\beta}}, \sigma^2; \mathbf{y})$ with respect to σ^2. Call the resulting estimator $\hat{\sigma}^2$.

Then $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$ will be the MLEs of $\boldsymbol{\beta}, \sigma^2$.

1. In step 1 we treat σ^2 as known and maximize $\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$ with respect to $\boldsymbol{\beta}$. Note that this is equivalent to maximizing the third term,

$$-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

which is equivalent to minimizing

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \|\mathbf{e}\|^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 \equiv S(\boldsymbol{\beta}). \qquad (*)$$

- Thus the MLE of $\boldsymbol{\beta}$ minimizes $S(\boldsymbol{\beta})$, which is known as the least squares criterion.
  - So, the estimators of $\boldsymbol{\beta}$ given by ML and least squares coincide.

  We can obtain the MLE/LSE, $\hat{\boldsymbol{\beta}}$, by solving the normal equations which are obtained by differentiating $S(\boldsymbol{\beta})$ and setting the result equal to 0.

  $S(\boldsymbol{\beta})$ can be written

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \underbrace{\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}}_{=\mathbf{y}^T\mathbf{X}\boldsymbol{\beta}} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

$$= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

- Need $\frac{\partial S}{\partial \boldsymbol{\beta}^T}$.

To take the necessary derivatives, we need some results on matrix differentiation. For $\mathbf{x}$ a vector and $\mathbf{A}$ a matrix,

i. $\frac{\partial \mathbf{Ax}}{\partial \mathbf{x}^T} = \mathbf{A}$.

ii. $\frac{\partial \mathbf{x}^T\mathbf{Ax}}{\partial \mathbf{x}^T} = 2\mathbf{x}^T\mathbf{A}$.

Using (i) and (ii) we get

$$\frac{\partial S}{\partial \boldsymbol{\beta}^T} = -2\mathbf{y}^T\mathbf{X} + 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}$$

so that the normal equations become

$$\begin{aligned} 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X} &= 2\mathbf{y}^T\mathbf{X} \\ \text{or} \qquad \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^T\mathbf{y}. \end{aligned}$$

If $\mathbf{X}^T\mathbf{X}$ is invertible (nonsingular) we can multiply through on both sides by $(\mathbf{X}^T\mathbf{X})^{-1}$ to give the MLE/Least squares estimator of $\boldsymbol{\beta}$ as:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \qquad (\ddagger)$$

- We'll assume $(\mathbf{X}^T\mathbf{X})$ is invertible henceforth unless stated otherwise. Note that if $(\mathbf{X}^T\mathbf{X})$ is not invertible a (no longer unique) estimator of $\boldsymbol{\beta}$ is obtained simply by replacing the matrix inverse in $(\ddagger)$ with a *generalized matrix inverse.*

2. Now we maximize $\ell(\hat{\boldsymbol{\beta}}, \sigma^2; \mathbf{y})$ with respect to σ^2. Taking derivatives with respect to σ^2 we get

$$\frac{\partial \ell}{\partial(\sigma^2)} = \frac{-n/2}{\sigma^2} + \frac{(1/2)\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{\sigma^4} = 0,$$

which has solution,

$$\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2.$$

- We will see that $\hat{\boldsymbol{\beta}}$ has a number of desirable properties including some optimality properties. However, $\hat{\sigma}^2$ is not typically the preferred estimator.

  One fault with $\hat{\sigma}^2$ is that it is biased. It can be shown that

$$\mathrm{E}(\hat{\sigma}^2) = \frac{n-p}{n}\sigma^2$$

  Therefore, an unbiased estimator that is generally superior to σ^2 can be formed by taking

$$\frac{n}{n-p}\hat{\sigma}^2 = \frac{1}{n-p}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2 \equiv s^2$$

- We call this estimator s^2, the mean squared error.

Properties of $\hat{\boldsymbol{\beta}}$ and methods of inference on $\boldsymbol{\beta}$:

Several properties of $\hat{\boldsymbol{\beta}}$ follow from the fact that it is a linear function of $\mathbf{y}$.

> (That is, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is of the form $\mathbf{My}$ for $\mathbf{M}$ a matrix of constants.)

This linearity combined with the model equation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ leads to some nice, simple properties.

Notice,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e} \\ &= \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}\end{aligned}$$

It follows that

1. $\hat{\boldsymbol{\beta}}$ is unbiased, since

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \mathrm{E}\{\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}\} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underbrace{\mathrm{E}(\mathbf{e})}_{=\mathbf{0}} = \boldsymbol{\beta}.$$

2. $\hat{\boldsymbol{\beta}}$ has var-cov matrix

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{\beta}}) &= \mathrm{var}\{\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}\} = \mathrm{var}\{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}\} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underbrace{\mathrm{var}(\mathbf{e})}_{=\sigma^2\mathbf{I}}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\end{aligned}$$

> (Here we've used that fact that if $\mathbf{w}$ is a $n \times 1$ random vector with var-cov matrix Σ and $\mathbf{B}$ is an $m \times n$ matrix of constants, then $\mathrm{var}(\mathbf{Bw}) = \mathbf{B}\mathrm{var}(\mathbf{w})\mathbf{B}^T = \mathbf{B}\Sigma\mathbf{B}^T$.)

3. (normality) $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ (if $\mathbf{e}$ is assumed normal).

4. $\hat{\boldsymbol{\beta}}$ is the Best (minimum variance) estimator in the class of all Linear Unbiased Estimators ($\hat{\boldsymbol{\beta}}$ is BLUE). This result doesn't require the assumption of normality on the errors of the CLM.

5. Under the assumption of normal errors, $\hat{\boldsymbol{\beta}}$ and s^2 are minimum variance unbiased estimators (best in the class of unbiased, but not-necessarily linear estimators).

6. Since $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ (by (2)), $\text{var}(\hat{\beta}_j) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}$. Since σ^2 is typically unknown, we must estimate it with s^2 to get an estimate of $\text{var}(\hat{\beta}_j)$. The square root of this estimated variance is the **standard error** of $\hat{\beta}_j$:

$$\text{s.e.}(\hat{\beta}_j) = s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}.$$

With this standard error in hand, methods of inference (hypothesis tests, confidence intervals) follow from the fact that

$$\frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} \sim \underbrace{t(n-p)}_{\text{the } t \text{ distribution with } n-p \text{ d.f.}}$$

   - $\Rightarrow \hat{\beta}_j \pm t_{1-\alpha/2}(n-p)\text{s.e.}(\hat{\beta}_j)$ forms a $100(1-\alpha)\%$ **marginal confidence interval** for β_j.

   - For an α-level test of $H_0 : \beta_j = \beta_0$ versus $H_1 : \beta_j \neq \beta_0$ we use the rule: reject H_0 if

$$\frac{|\hat{\beta}_j - \beta_0|}{\text{s.e.}(\hat{\beta}_j)} > t_{1-\alpha/2}(n-p)$$

7. Inference on the entire vector $\boldsymbol{\beta}$ is based on the fact that

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{ps^2} \sim \underbrace{F(p,n-p)}_{\text{the } F \text{ distribution with } p \text{ and } n-p \text{ d.f.}}$$

   – A $100(1-\alpha)\%$ **joint confidence region** for $\boldsymbol{\beta}$ is given by the set of all $\boldsymbol{\beta}$ such that

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{ps^2} \leq F_{1-\alpha}(p,n-p)$$

   This region consists of the surface and interior of an ellipsoid (p-dimensional ellipse; e.g., a watermelon for $p=3$).

   – For an α-level test, we reject $H_0: \boldsymbol{\beta}=\boldsymbol{\beta}_0$ in favor of $H_1: \boldsymbol{\beta}\neq\boldsymbol{\beta}_0$ if

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T(\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)}{ps^2} > F_{1-\alpha}(p,n-p).$$

   – The previous result is sometimes of use when $\boldsymbol{\beta}_0=\mathbf{0}$, but often we want to test that some linear function of $\boldsymbol{\beta}$ (e.g., a subvector of $\boldsymbol{\beta}$) is equal to $\mathbf{0}$ or some other null value $\mathbf{b}$. In that case, it is useful to have a generalization of the above F test for $H_0: \mathbf{A}\boldsymbol{\beta}=\mathbf{b}$ versus $H_1: \mathbf{A}\boldsymbol{\beta}\neq\mathbf{b}$ where $\mathbf{A}$ is a $k\times p$ matrix of constants and $\mathbf{b}$ is a $k\times 1$ vector of constants. The appropriate test has rejection rule: reject if

$$F=\frac{(\mathbf{A}\hat{\boldsymbol{\beta}}-\mathbf{b})^T\{\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}\}^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}}-\mathbf{b})}{ks^2} > F_{1-\alpha}(k,n-p).$$

   (Note that not every such hypothesis is testable. We require that $\mathbf{A}$ has full row rank. Essentially, this means there is no redundancy in the statement of the null hypothesis.)

8. A $100(1-\alpha)\%$ C.I. for the expected response at a given value of the vector of explanatory variables $\mathbf{x}_o$ is given by

$$\mathbf{x}_0^T\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2}(n-p)\sqrt{s^2\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}.$$

   – Similarly, we can form a C.I. for any linear combination of the β's that can be written in the form $\mathbf{c}^T\boldsymbol{\beta}$ for $\mathbf{c}$ a vector of constants by replacing $\mathbf{x}_0$ with $\mathbf{c}$. A $100(1-\alpha)\%$ C.I. for $\mathbf{c}^T\boldsymbol{\beta}$ is given by

$$\mathbf{c}^T\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2}(n-p)\sqrt{s^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}.$$

9. A $100(1-\alpha)\%$ C.I. for the predicted response (not the mean response over the population, but a single new observation of the response variable) at a given value of the vector of explanatory variables $\mathbf{x}_o$ is given by

$$\mathbf{x}_0^T\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2}(n-p)\sqrt{s^2[1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0]}.$$

   • Such an interval is usually called a **prediction interval** rather than a confidence interval.

10. A $100(1-\alpha)\%$ **confidence band** for the response function at any $\mathbf{x}$ is given by

$$\mathbf{x}^T\hat{\boldsymbol{\beta}} \pm \sqrt{F_{1-\alpha}(p,n-p)}\sqrt{ps^2\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}.$$

   • Result (8) gives a C.I. at a single given point ($\mathbf{x}_0$) whereas result (10) gives a confidence band that holds for all values of $\mathbf{x}$ considered simultaneously.

Example – PCB in Trout (fitting a linear model in R):

The data below consist of PCB (polychlorinated biphenyls, a toxin) concentrations in Lake Cayuga (NY) trout of various ages.

Age (years)	PCB Conc. (ppm)	Age (years)	PCB Conc. (ppm)
1	0.6	6	3.4
1	1.6	6	9.7
1	0.5	6	8.6
1	1.2	7	4.0
2	2.0	7	5.5
2	1.3	7	10.5
2	2.5	8	17.5
3	2.2	8	13.4
3	2.4	8	4.5
3	1.2	9	30.4
4	3.5	11	12.4
4	4.1	12	13.4
4	5.1	12	26.2
5	5.7	12	7.4

- See the handout labelled trout1. The first page of this handout contains R commands contained in the R script file trout1.R. Pages 2–3 contain the text output of these commands and p.4 the graphics output.

- The first plot p.4 of trout1 contains a scatterplot of PCB concentration versus age. From the plot there appears to be some nonlinearity in age and heteroscedasticity (nonconstant, in this case increasing with age, variance).

- From these observations it appears that the assumptions of the CLM preclude its use here. However, transformations of the response and explanatory variables can often induce linearity, normality and constant variance thereby making the CLM an appropriate tool.

- One useful class of transformations is the Box-Cox family of power transformations:

$$g(Y;\lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log Y, & \text{if } \lambda = 0 \end{cases}$$

  - For CLMs with an intercept term, this family is equivalent to the "simple" family of power transformations given by

$$g_S(Y;\lambda) = \begin{cases} Y^\lambda, & \text{if } \lambda \neq 0 \\ \log Y, & \text{if } \lambda = 0 \end{cases}$$

    but the Box-Cox family is slightly more convenient mathematically.

- The specific transformation in the Box-Cox family can be chosen by estimating λ by ML estimation. Box and Cox showed that the ML estimator (MLE) of λ can be obtained as the maximizer of the function

$$-\frac{n}{2}\log \text{SSE}\{\mathbf{z}(\lambda)\},$$

  where $\mathbf{z}(\lambda)$ has i^{th} element $g(y_i;\lambda)/\dot{y}^{\lambda-1}$, $\dot{y}$ is the geometric mean of the elements of $\mathbf{y}$ and $\text{SSE}\{\mathbf{z}(\lambda)\}$ denotes the error sum of squares for the regression of $\mathbf{z}(\lambda)$ on $\mathbf{X}$.

  - This function is known as the **profile likelihood** for λ.

- Therefore, it is possible to obtain the MLE of λ by plotting

$$-\frac{n}{2}\log \mathrm{SSE}\{\mathbf{z}(\lambda)\}$$

  over a range of λ values and selecting the λ-value that maximizes this function. This is automated in the boxcox macro in R (part of the MASS package).

- See trout1.R. trout1.R contains R commands to select the appropriate transformation of PCB concentration and then to fit a linear regression model to the transformed data. These commands also produce the plots on p.4 of trout1.

- The boxcox macro is part of the MASS (Modern Applied Statistics with S-PLUS — the book by Venables and Ripley mentioned in the syllabus) library. This library comes standard with R, but must be loaded into an R session with the library(MASS) command.

- The par(mfrow=c(2,3)) command sets the graphical parameter mfrow so that plots are laid out 2×3 on a page.

- boxcox plots the profile likelihood for λ (see p.4 of trout1 handout, top-middle plot). From this plot we see that the MLE $\hat{\lambda}$ is close to 0. Rather than using the exact MLE, its preferable to round $\hat{\lambda}$ to the nearest interpretable value (e.g., 0, 1/4, 1/3, 1/2). In this case we take $\lambda = 0$ and use the log transformation.

- A plot of log(PCB Conc.) against age looks much more linear and homoscedastic. A cube-root transformation of age improves the situation even further (we omit discussion of selecting transformations of the explanatory variables, but this subject is discussed in Box and Tidwell (1962) and elsewhere).

- A linear regression of the form

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, 28,$$

where $e_1, \ldots, e_{28} \overset{iid}{\sim} N(0, \sigma^2)$ and $y_i = \log(\text{PCB})$, $x_i = \text{age}_i^{1/3}$ is fit using the lm function. The "data frame" trout contains the variables on their original scale. The $I()$ function just allows the computation of the transformation to be done within the call to lm.

- R is an object-oriented language. This means that quite complicated entities like model fits can be stored and operated on as a single object. E.g., m1trout.lm is assigned the entire model fit. It is stored as a list containing all of the results of the model fit listed by the function names(m1trout.lm). The fitted model can be summarized with the command summary(m1trout.lm) and various other functions (like coef) exist for extracting results from m1trout.lm.

- The remainder of the code in trout1.R computes confidence intervals and regions for $\boldsymbol{\beta}$ and produces the rest of the plots on p.4 of trout1.

- For example, we can obtain the 95% confidence band for the average log(PCB) value at all x (values of age$^{1/3}$) by plotting

$$\begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \pm \sqrt{F_{.95}(2, 28-2)} \sqrt{2s^2 \begin{pmatrix} 1 & x \end{pmatrix} (\mathbf{X}^T\mathbf{X})^{-1} \begin{pmatrix} 1 \\ x \end{pmatrix}}$$

or $\quad (-2.391 + 2.300x) \pm \sqrt{3.369}\sqrt{2(.246)(.637 - .718x + .214x^2)}$

over the range of x-values observed in the data.

The Geometry of Linear Least Squares:

Again, let $S(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, the squared (Euclidean, or straight-line) distance from $\mathbf{y}$ to its mean according to the model, $\mathbf{X}\boldsymbol{\beta}$.

- The linear least squares estimator $\hat{\boldsymbol{\beta}}$ minimizes $S(\boldsymbol{\beta})$.

Calculating $S(\boldsymbol{\beta})$ consists of 2 steps:

1. Using the $n \times p$ derivative matrix $\mathbf{X}$ and the $p \times 1$ parameter vector $\boldsymbol{\beta}$ to form the **expected response vector** $\boldsymbol{\eta}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$.

2. Calculating the squared distance from the expected response $\boldsymbol{\eta}(\boldsymbol{\beta})$ to the observed response $\mathbf{y}$.

- Though $\boldsymbol{\eta}(\boldsymbol{\beta})$ lies in $n-$space (has n components), the set of all possible values it can take is not $n-$space. We can only vary the p parameters in $\boldsymbol{\beta}$ to get different values of $\boldsymbol{\eta}(\boldsymbol{\beta})$. That is, $\boldsymbol{\eta}(\boldsymbol{\beta})$ lies in a p-dimensional subspace of $n-$dimensional space.

- We call the set of all possible values of $\boldsymbol{\eta}(\boldsymbol{\beta})$ the **expectation surface** of the model.

  - In a linear model, $\boldsymbol{\eta}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$ is a linear combination of the columns of $\mathbf{X}$ ($\mathbf{X}\boldsymbol{\beta} = \beta_1\mathbf{x}_1 + \cdots + \beta_p\mathbf{x}_p$ where $\mathbf{x}_j$ is the j^{th} column of $\mathbf{X}$) so we call the $\mathbf{X}\boldsymbol{\beta}$ the **expectation plane** of the model.

Very Simple Example — $n = 2, p = 1$**:**

Suppose we have a response vector with just two components: $\mathbf{y} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$ to which we'd like to fit the linear model

$$y_i = \beta + e_i, \quad i = 1, 2$$

or

$$\mathbf{y} = \beta\mathbf{x} + \mathbf{e}, \quad \text{where} \quad \mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The response vector $\mathbf{y}$ falls in two-dimensional space. We can plot $\mathbf{y}$ as follows:

- Least-squares estimate of β: the $\hat{\beta}$ so that $\boldsymbol{\eta}(\beta)$ is the closest point on the expectation plane to $\mathbf{y}$.
  - Since $\boldsymbol{\eta}(\hat{\beta}) = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$ is the closest point to $\mathbf{y} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$ it is easy to find that the $\hat{\beta}$ that yields $\hat{\beta} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$ is $\hat{\beta} = 3$.

A Slightly Less Simple Example — $n = 3, p = 2$:

Consider again the PCB in trout data and suppose we want to fit our simple linear regression model as before

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, n$$

for $y = \log(\text{PCB})$ and $x = \text{age}^{1/3}$, but now suppose we have only $n = 3$ observations:

$\text{age}^{1/3}$	$\log(\text{PCB})$
1.26	0.92
1.82	2.15
2.22	2.52

The derivative matrix here is

$$\mathbf{X} = \begin{pmatrix} 1 & 1.26 \\ 1 & 1.82 \\ 1 & 2.22 \end{pmatrix}$$

with columns $\mathbf{x}_1 = (1, 1, 1)^T$ and $\mathbf{x}_2 = (1.26, 1.82, 2.22)^T$.

- Since the response has $n = 3$ components, the response space is 3−dimensional space. We can plot $\mathbf{x}_1$ and $\mathbf{x}_2$ in that space (below, plot (b)).

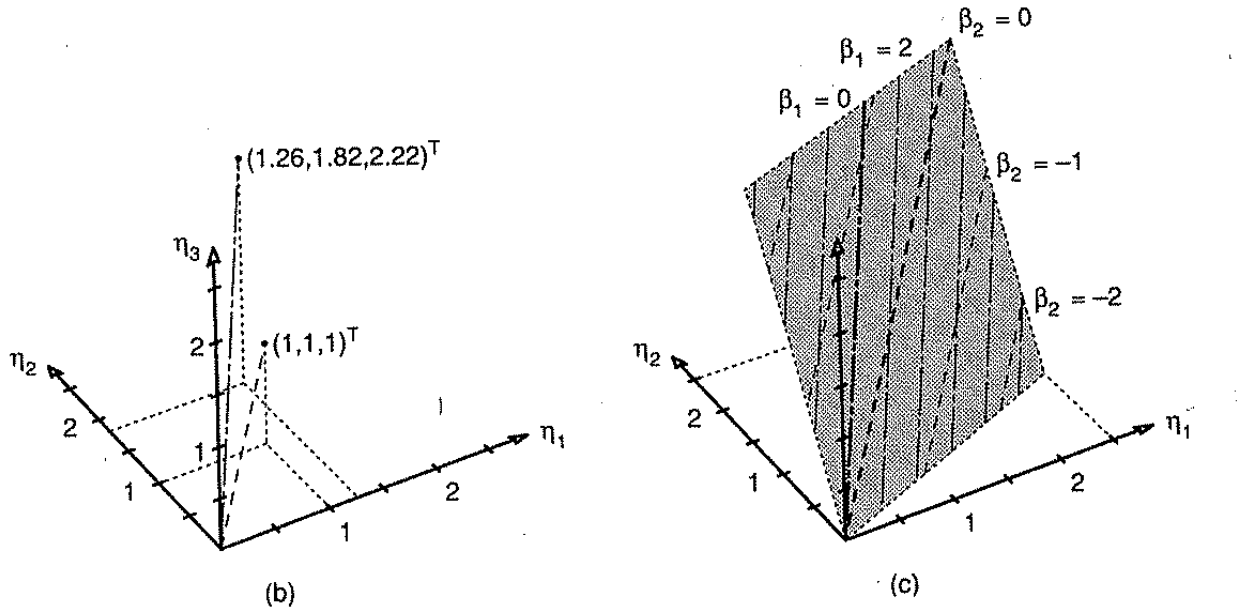


Figure 1.4 Expectation surface for the 3-case PCB example. Part *a* shows the parameter plane with β_1 parameter lines (dashed) and β_2 parameter lines (dot–dashed). Part *b* shows the vectors $\mathbf{x}_1$ (dashed line) and $\mathbf{x}_2$ (dot–dashed line) in the response space. The end points of the vectors correspond to $\boldsymbol{\beta}=(1, 0)^T$ and $\boldsymbol{\beta}=(0, 1)^T$ respectively. Part *c* shows a portion of the expectation plane (shaded) in the response space, with β_1 parameter lines (dashed) and β_2 parameter lines (dot–dashed).

- The expectation plane is the set of all $\boldsymbol{\eta}$ such that $\boldsymbol{\eta} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2$ for some constants β_1 and β_2. This plane is depicted in plot (c) above.

- $\boldsymbol{\eta}(\hat{\boldsymbol{\beta}})$ is the point on this plane that is closest to $\mathbf{y} = (0.92, 2.15, 2.52)^T$ (it is the (Euclidean) **projection** of $\mathbf{y}$ onto that plane).

- $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that yields this closest point to $\mathbf{y}$.

So, to find $\hat{\boldsymbol{\beta}}$ we

1. Find $\hat{\boldsymbol{\eta}}$ that is closest to $\mathbf{y}$; then
2. find $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{\eta}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\eta}}$.

We know from differentiating the least squares criterion that $\hat{\boldsymbol{\beta}}$ solves the normal equation

$$(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} \qquad (*)$$

- Another way we can derive (*) is from the geometry of the problem:
  - The expectation plane is the set of all $n \times 1$ vectors that can be written as $\mathbf{Xa}$ for some $p \times 1$ vector $\mathbf{a}$ (the set of all possible linear combinations of the columns of $\mathbf{X}$).
  - We know that the residual vector $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ must be **orthogonal** (i.e., perpendicular) to the expectation plane, so the angle between $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{Xa}$ must be $90°$ for any $p \times 1$ vector $\mathbf{a}$.
  - Algebraically, two vectors $\mathbf{b}$ and $\mathbf{c}$ form a $90°$ angle if and only if their inner product $\mathbf{b}^T\mathbf{c}$ equals 0.
  - Therefore, the geometry of least squares implies that the least squares estimator of $\boldsymbol{\beta}$ satisfies

$$\begin{aligned}
(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Xa}) = 0, \quad & \text{for all } \mathbf{a} \\
\text{or} \quad \mathbf{y}^T\mathbf{Xa} = \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Xa}, \quad & \text{for all } \mathbf{a} \\
\text{which implies} \quad \mathbf{y}^T\mathbf{X} = \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}, & \\
\text{or, equivalently,} \quad \mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}, \quad & \text{(the normal equation)}
\end{aligned}$$

  and again we obtain that $\hat{\boldsymbol{\beta}}$ must satisfy the normal equation.

Assumptions of the CLM:

Model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + e, \quad \mathbf{e} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

1. *Expectation function is correctly specified as* $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. Expectation form assumed linear in $\boldsymbol{\beta}$ and contains all important predictor variables each on the right scale.

2. *Additive error.* We assume $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ rather than, for example, $y_i = (\mathbf{x}_i^T \boldsymbol{\beta})^{e_i}$. This assumption implies

   - Distribution of $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ doesn't depend on $\boldsymbol{\beta}$.

   - $\boldsymbol{\beta}$ should be estimated to make $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ "small."

3. *The distribution of* $\mathbf{e}$ *does not depend on* $\mathbf{X}$. That is, the effect of $\mathbf{X}$ on $\mathbf{y}$ is completely captured by $\mathbf{X}\boldsymbol{\beta}$. Often justified by randomization.

4. *Each* e_i *has mean 0.* This is a consequence of (1) and (2). Not at all restrictive in a linear model with intercept, but deserves some attention in a nonlinear model with no intercept.

5. *Homoscedasticity.* Each e_i has the same variance, σ^2. Implies $y_1, \ldots, y_n$ all have the same variance, σ^2.

6. *Independence.* The e_i's are independent $\Rightarrow$ the y_i's are independent. Often justified by randomization.

7. *Normality.* Assume each e_i follows a normal (Gaussian) distribution.

- Assumptions (4)–(7) imply that the length of $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ should be measured with Euclidean distance: $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|$.

- Assumption (7) is not necessary to motivate least-squares and establish its optimality (BLUE-ness). However, classical inference methods rely on (7).

Model fitting is an iterative process: Make assumptions. Fit model. Check assumptions. Revise model. Check revised model's assumptions. Etc.

Verifying the Assumptions:

Since most of the assumptions are made on the error terms, it makes sense to check whether these assumptions appear to hold for the estimated error terms, or **residuals**.

Let $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ be the vector of **fitted values** (a.k.a. **predicted values**) and let

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} \qquad \textbf{(raw residuals)}$$

Most of the strongest and/or most commonly violated assumptions of the CLM can be checked by plotting the residuals. E.g., to check (1) we could plot $\hat{e}_1, \ldots, \hat{e}_n$ versus the sample values of other potential explanatory variables.

- For most residual plots presence of any pattern indicates violation. In addition, we look for large residuals.
- Since the size of raw residuals depends upon the units of $\mathbf{y}$, its useful to standardize the residuals in some way.
- Several possibilities: one simple way is to look at *studentized residuals* which are simply the raw residuals divided by their estimated standard deviation.

Raw residuals:

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\equiv \mathbf{H}}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

- Here $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is known as the **hat matrix** because $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ ($\mathbf{H}$ is the matrix that puts the "hat" on $\mathbf{y}$).

To studentize the elements of $\hat{\mathbf{e}}$ we divide each element by its estimated standard deviation. It can be shown that

$$\text{var}(\hat{\mathbf{e}}) = \sigma^2(\mathbf{I} - \mathbf{H}),$$

which we estimate with

$$\hat{\text{var}}(\hat{\mathbf{e}}) = s^2(\mathbf{I} - \mathbf{H}). \qquad (*)$$

Therefore, the estimated standard deviation of $\hat{e}_i$ is the i^{th} diagonal element of (*), or $s\sqrt{1 - h_{ii}}$, where h_{ii} is the i^{th} diagonal element of $\mathbf{H}$. Thus, the studentized residuals are

$$\frac{y_i - \hat{y}_i}{s\sqrt{1 - h_{ii}}}, \qquad i = 1, \ldots, n.$$

- A simpler way to standardize residuals is to use the **Pearson residuals:**

$$\frac{y_i - \hat{y}_i}{\sqrt{\hat{\text{var}}(y_i)}} = \frac{y_i - \hat{y}_i}{s}, \quad i = 1, \ldots, n.$$

- Other types of standardized are possible, but different choices usually lead to the same conclusions and, for many purposes, there's little reason to prefer one definition over another. Unfortunately, there's considerable variability in the terminology used for various types of residuals.

Residual plots (see, e.g., Draper & Smith, Ch. 3):

1. versus the fitted values. Should see no pattern, few large (in absolute value) residuals. Violations can indicate heteroscedasticity, incorrect specification of expectation functional form, outliers, correlation.

2. versus the predictor variables. Should see no pattern. Patterns can indicate heteroscedasticity, need for extra terms (e.g., a term quadratic in the predictor).

3. versus potential predictor variables. Should indicate no pattern; otherwise, potential predictor should be included.

4. versus time (or some other potentially ordering index of the responses). Should see no pattern. Pattern can reveal autocorrelation (dependence through time), heteroscedasticity, or the need to include time as predictor variable.

5. quantile-quantile plot (normal probability plot). Plot the sample quantiles versus the expected quantiles under the assumption that the e_i's are normally distributed. Should be a straight line. For moderate to large sample sizes, non-straight plots indicate non-normality. Not useful in small samples ($n < 30$ or so).

Example — Scottish Hill Races Data:

The S-PLUS library MASS contains a data set containing record fastest times in 35 Scottish hill races (running races) against distance and total height climbed in the race.

- See handout, hills1. This handout contains hills1.SSC, a file containing S-PLUS commands to analyze these data; the associated output; and associated graphics from the analysis.

- On line 2 of hills1.SSC we print out the data. The par(mfrow=c(2,2)) command sets up 8 plots per page in a 2×2 grid. In the first plot (labelled "(a)") we simply plot time vs. dist. As we should expect, there appears to be an increasing relationship between time and distance.

- We first fit a simple linear regression model of the form

$$\text{time}_i = \beta_0 + \beta_1 \text{dist}_i + e_i, \quad i = 1, \ldots, 35. \tag{m1}$$

  and add the fitted regression line to plot (a). This model appears to fit reasonably well, but we should check residuals.

- The functions fitted() and stdres() extract the fitted values and studentized residuals (as I've defined them) from the fitted model. We plot these residuals vs. fitteds in plot (b). Notice that there appear to be several outliers and, perhaps, some increasing variance. In addition, the residuals don't appear to be centered around zero as much as would be desirable. This is probably an effect of fitting outliers.

- Next we consider plot (c), a plot of residuals vs. the potential predictor, climb. There appears to be an increasing pattern, suggesting that climb belongs in the model.

- So, next we fit model m2hills.lm which is of the form

$$\text{time}_i = \beta_0 + \beta_1 \text{dist}_i + \beta_2 \text{climb}_i + e_i, \quad i = 1, \ldots, 35. \qquad \text{(m2)}$$

- These models are examples of **nested models.** Note that m1hills.lm is nested in model m2hills.lm in the sense that m1 is a special case of m2 that occurs when β_2 is fixed at 0. We can test m1 versus m2 by testing $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$ in model m2. Notice that this testing situation is a special case of that described on bottom of p.23 with

$$\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{b} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - 0.$$

- In nested models testing situations such as this, the F test statistic on the bottom of p.23 has an algebraically equivalent, and more convenient, form in terms of the degrees of freedom and sums of squares for error for the two models:

$$F = \frac{(\mathrm{SS_{E0}} - \mathrm{SS_E})/(\mathrm{dfE_0} - \mathrm{dfE})}{\mathrm{SS_E}/\mathrm{dfE}}$$

  where $\mathrm{SS_E}0$ and $\mathrm{SS_E}$ are the sums of squares for error associated with the null model (model in which H_0 holds) and the alternative models, respectively, and $\mathrm{dfE_0}$ and dfE are the degrees of freedom for these two models.

  - In general, the $\mathrm{SS_E}$ and dfE for a (full rank) model with $p \times 1$ regression parameter $\boldsymbol{\beta}$ are given by

$$\mathrm{SS_E} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \sum_{i=1}^{n} (y_i - \underbrace{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}}_{=\hat{y}_i})^2, \quad \mathrm{dfE} = n - p.$$

  - We reject H_0 at significance level α if $F > F_{1-\alpha}(\mathrm{dfE_0} - \mathrm{dfE}, \mathrm{dfE})$.

- The anova() function in S-PLUS automates the testing of nested models using the test described above. It takes as arguments, two fitted model objects and tests the null hypothesis that the smaller (null) model holds versus the alternative that it does not under the maintained hypothesis that the larger model holds. In the example, we reject H_0 : {model m1 holds} in favor of model m2 ($F = 29.02$, $p < .0001$).

- After adding climb to our model, we recheck the residuals versus climb plot (plot (d)). There still appears to be a pattern to this plot, although now it looks different — a convex curve. This suggestes adding a climb^2 term.

- In m3hills.lm we fit

$$\text{time}_i = \beta_0 + \beta_1 \text{dist}_i + \beta_2 \text{climb}_i + \beta_3 \text{climb}_i^2 + e_i, \quad i = 1, \ldots, 35. \quad \text{(m3)}$$

  Again, using the anova() function, we see that m3 fits significantly better than m2.

- However, the residuals versus climb plot (plot (e)) and the residuals versus fitteds plot (plot (f)) still don't look particularly good. In plot (f) we can identify the outlier using the identify() function. Type ?identify in S-PLUS to get a description on this function.

- We print out the predicted (based on model m3) and observed data for this outlier using the predict function. Notice that the predicted and observed times are about an hour apart. It is possible that this data point was misrecorded. We complete the analysis under this assumption, omitting this point from further models.

- Model m4hills.lm refits m3 with the outlier removed. Plot (g) displays the residuals versus fitteds from this model. These residuals look fairly good.

- Although there doesn't appear to be any heteroscedasticity in plot (g), it seems intuitively reasonable that variability in race times should increase with the length of the race. If this were the case, we might account for it be transforming the response so that it had constant variance on the transformed scale.

- Alternatively, we might consider a model such as m5hills.lm. This model is identical to m4, but instead of the constant variance assumption $\text{var}(e_i) = \sigma^2$, $i = 1, \ldots, n$, we assume the error variance is proportional to race length squared: $\text{var}(e_i) = \sigma^2 \text{dist}_i^2$ (i.e., the error standard deviation is proportional to dist).

- This is an example of a linear model fit with **weighted least squares.**

Weighted Least Squares

Suppose we have a linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad (\dagger)$$

where $\mathbf{e}$ has variance $\text{var}(\mathbf{e}) = \sigma^2\mathbf{V}$ where $\mathbf{V}$ is a known positive-definite matrix not necessarily equal to $\mathbf{I}_n$.

- For such a $\mathbf{V}$ it is always possible to find a square-root matrix $\mathbf{V}^{1/2}$ that has the property $(\mathbf{V}^{1/2})^T\mathbf{V}^{1/2} = \mathbf{V}$.

Notice that we can multiply both sides of (†) by $\mathbf{V}^{-T/2} \equiv \{(\mathbf{V}^{1/2})^T\}^{-1}$ to obtain an equivalent transformed model,

$$\underbrace{\mathbf{V}^{-T/2}\mathbf{y}}_{\equiv\mathbf{y}^*} = \underbrace{\mathbf{V}^{-T/2}\mathbf{X}}_{\equiv\mathbf{X}^*}\boldsymbol{\beta} + \underbrace{\mathbf{V}^{-T/2}\mathbf{e}}_{\equiv\mathbf{e}^*}$$

$$\text{or} \qquad \mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{e}^* \qquad (*)$$

Notice that the *'ed model satisfies the CLM assumptions because

$$\text{E}(\mathbf{e}^*) = \text{E}(\mathbf{V}^{-T/2}\mathbf{e}) = \mathbf{V}^{-T/2}\underbrace{\text{E}(\mathbf{e})}_{=\mathbf{0}} = \mathbf{0}$$

and

$$\begin{aligned}\text{var}(\mathbf{e}^*) = \text{var}(\mathbf{V}^{-T/2}\mathbf{e}) &= \mathbf{V}^{-T/2}\underbrace{\text{var}(\mathbf{e})}_{=\sigma^2\mathbf{V}}\mathbf{V}^{-1/2} \\ &= \sigma^2\mathbf{V}^{-T/2}\underbrace{\mathbf{V}}_{=\mathbf{V}^{T/2}\mathbf{V}^{1/2}}\mathbf{V}^{-1/2} = \sigma^2\mathbf{I}_n.\end{aligned}$$

Therefore, the (ordinary) least squares estimator of $\boldsymbol{\beta}$ in (*),

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \{(\mathbf{X}^*)^T\mathbf{X}^*\}^{-1}(\mathbf{X}^*)^T\mathbf{y}^* \\ &= \{(\mathbf{V}^{-T/2}\mathbf{X})^T\mathbf{V}^{-T/2}\mathbf{X}\}^{-1}(\mathbf{V}^{-T/2}\mathbf{X})^T\ \mathbf{V}^{-T/2}\mathbf{y} \\ &= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}\end{aligned}$$

is the optimal (BLUE) estimator of $\boldsymbol{\beta}$.

- Because this estimator differs from the ordinary least squares estimator $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ by the inclusion of a weight matrix $\mathbf{V}^{-1}$ in the formula, we call this estimator the **weighted least squares** estimator of $\boldsymbol{\beta}$.

- It can be shown that the WLS estimator $\hat{\boldsymbol{\beta}}$ minimizes

  $$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \qquad \textbf{(WLS Criterion)}$$

  instead of the OLS criterion $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{I}_n(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.

  - This is to say that the OLS estimator minimizes the length of $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ with respect to Euclidean distance and the WLS estimator minimizes the length of $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ with respect to a more general statistical distance metric.

  - The statistical distance used in WLS is *weighted Euclidean distance* or *Karl Pearson distance* in the case that $\mathbf{V}$ is a diagonal matrix. In this case, we are just accounting for heteroscedasticity.

  - The statistical distance used in WLS is *Mahalanobis distance* in the case that $\mathbf{V}$ is non-diagonal (a.k.a. generalized least squares). In this case, we are accounting for heteroscedasticity and correlation among the e_i's.

- It's also straight-forward to show that the WLS estimator $\hat{\boldsymbol{\beta}}$ maximizes the log-likelihood of model (†) under the assumption $e \sim N_n(\mathbf{0}, \sigma^2\mathbf{V})$ so that WLS estimation = ML estimation in this model.

Back to the Scottish Hills Races Example:

- In model m5hills.lm, we use the weight option in function lm() to fit the model using WLS.

- The residuals from this model (plot (h)) don't look any better than those from model m4. It is not possible to test m4 versus m5 using an F test for nested models (e.g., using the anova() function) because these are not nested models. They have the same linear predictor and same total number of parameters.

- However, it is possible to informally compare the two models using **information criteria.** Two of the most popular information criteria are **AIC** (Akaike's Information Criterion) and **BIC** (Bayesian Information Criterion, a.k.a. Schwarz's Bayesian Criterion).

- Both of these quantities are penalized version of the maximized loglikelihood function. That is, they measure how likely the data are according to the model (as quantified by the loglikelihood function evaluated at the MLEs of the model parameters) but then penalize this quantity by an amount related to the complexity of the model. (This same penalization for lack of parsimony idea is the idea behind adjusted R^2.)

- For a model with $k \times 1$ parameter vector $\boldsymbol{\theta}$ (including all parameters, not just regresison parameters) AIC and BIC are defined as

$$\begin{aligned} \text{AIC} &= -2\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) + 2k \\ \text{BIC} &= -2\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) + k\log(n) \end{aligned}$$

  where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$.

  - AIC and BIC are sometimes given in other forms, but in this form, the model with the smallest value of AIC (or BIC, if that criterion is used) is the winner.

  - Its hard to say which criterion is best, but BIC tends to lead to more parsimonious models than AIC.

- The S-PLUS function AIC() and BIC() extract these information criteria from fitted model objects. The function logLik() can be used to obtain the maximized log likelihood values.

- According to both AIC and BIC, m4 is preferred to m5, and we abandon the idea of accounting for heteroscedasticity in this example.

- Although m4 fits pretty well, it is possible to obtain a more parsimonious model for these data that fits even better. Venables and Ripley (1999, Ch. 6) consider regressing inverse speed (time/distance) on the race course gradient (climb/distance). We fit this model,

$$\text{speed}_i^{-1} = \beta_0 + \beta_1 \text{grad}_i + e_i, \qquad i = 1, \ldots, n, \qquad \text{(m6)}$$

  as model m6hills.lm.

- The residuals versus fitted values for m6 (see plot (i)) look as good or better than any previous model.

- We produce a Q-Q plot for model m6 using the qqnorm() and qqline() function. This plot (plot (j)) indicates there are more extreme values in the data set than expected according to a normal distribution. This suggests that a more appropriate distribution for the errors in model (m6) might be a distribution with fatter tails than the normal (e.g., the $t(\nu)$-distribution with ν small). See Venables and Ripley (1999, Ch.6) for discussion of such a robust regression approach to analyzing these data.

- The Scottish hill race data are also analyzed in Ch.6 of Maindonald & Braun's book, *Data Analysis and Graphics Using R*, which was distributed in class.

Nonlinear Regression (Ch.2 of Bates & Watts)

We assume that we observe data: $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ where y_i is a scalar response variable and $\mathbf{x}_i$ is an $m \times 1$ vector of explanatory variables.

Model:

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + e_i, \quad i = 1, \ldots, n,$$

$$\text{where} \quad e_1, \ldots, e_n \overset{iid}{\sim} N(0, \sigma^2)$$

where

$f(\cdot)$ is a known function (the expectation or regression function)

$\boldsymbol{\theta}$ is a $p \times 1$ parameter vector

e_i's are i.i.d. error terms

$f(\cdot)$ is a nonlinear function of $\boldsymbol{\theta}$. That is,

$$\frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_j} \quad \text{depends on } \boldsymbol{\theta} \text{ for some } j.$$

- If $f(\cdot)$ is nonlinear in any component of $\boldsymbol{\theta}$, it is a nonlinear model.

Let $\eta_i(\boldsymbol{\theta}) = f(\mathbf{x}_i, \boldsymbol{\theta})$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$. Then we can write our model equivalently as

$$\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \mathbf{e}, \quad \mathbf{e} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Some examples of expectation functions:

$$f(x, \boldsymbol{\theta}) = \frac{\theta_1 x}{\theta_2 + x} \qquad \text{(Michaelis-Menten Model)}$$

$$f(x, \boldsymbol{\theta}) = \frac{\theta_1}{1 + \exp\{(\theta_2 - x)/\theta_3\}} \qquad \text{(Simple Logistic Model)}$$

$$f(x, \boldsymbol{\theta}) = \theta_1 + (\theta_2 - \theta_1)\exp\{-\exp(\theta_3)x\} \qquad \text{(Asymptotic Regression Model)}$$

$$f(\mathbf{x}, \boldsymbol{\theta}) = \theta_1 + \theta_2 x_1 + \theta_3 x_2^{\theta_4} \qquad \text{(regression w/ power trans. of } x_2\text{)}$$

How do we choose $f(\cdot)$?

1. Mechanistic models. Often there is some scientific theory available that describes the data-generating mechanism. This theory suggests the form of f.

2. Empirical models. At other times no theory is available or we simply want to describe the relationship between y and $\mathbf{x}$ in a simple model or develop a model that produces good predictions (unconcerned by how those predictions come about). In such cases we simply try to "fit the data". For data that follow certain general shapes of curves (e.g., sigmoidal, parabolic, etc.) "promising candidates" for nonlinear expectation functions are available (e.g., Ratkowsky, 1990) that can be tried.

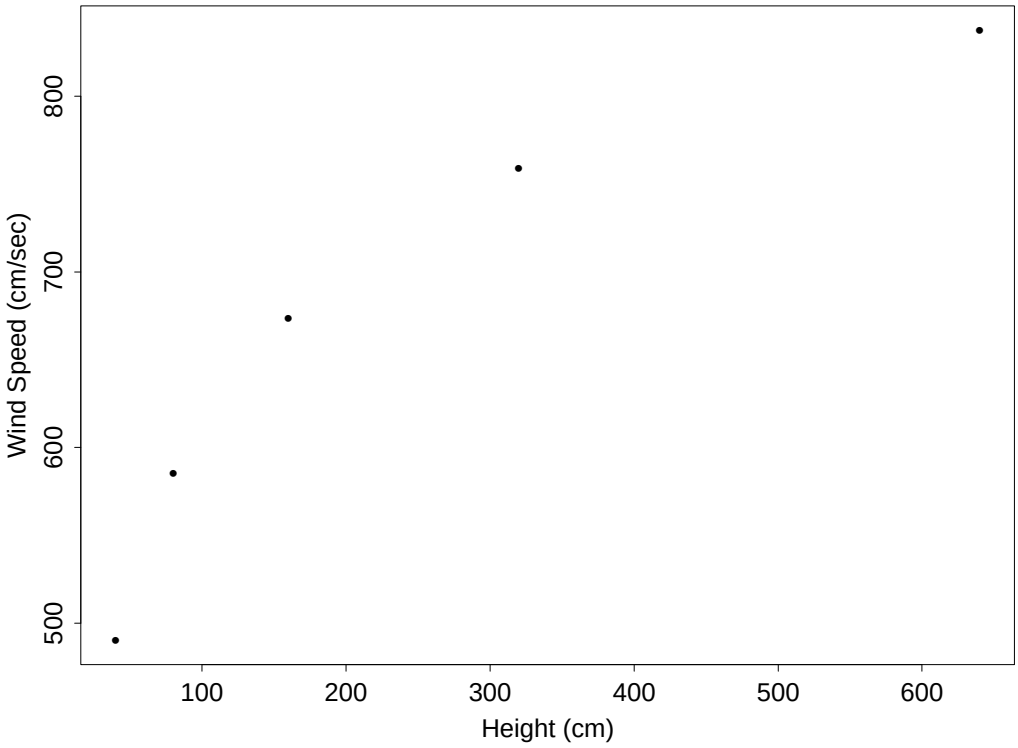
- In linear modelling, empirical models are most common. In nonlinear modelling, mechanistic models are more common.

Example — Wind Speed:

Consider the following five measurements of wind speed (y, in cm/sec) at various heights (x in cm):

x	y
40	490.2
80	585.3
160	673.7
320	759.2
640	837.5

These data are plotted below



- See handout, "wind1".
- For these data we might consider fitting linear models.

Model 1 (simple linear regression):

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, 5$$

yields $R^2 = 0.847$, and residual standard deviation $s = \sqrt{\text{mse}} = 62.13$.

Model 2 (quadratic regression):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \quad i = 1, \ldots, 5$$

yields $R^2 = 0.975$, and residual standard deviation $s = \sqrt{\text{mse}} = 30.52$.

- The fitted regression lines for these models are plotted on the last page of wind1. Clearly, model 1 is inadequate. Model 2 fits ok, but not great. We've already used 3 out of 5 d.f. to fit this model, so going to a cubic to improve the fit is not particularly attractive.

Theory: Under *adiabatic* conditions, wind speed is related to height as

$$\text{windspeed} = \theta_1 \log\{\text{height}(1 - \theta_2/\theta_3) - 1/\theta_3\}, \qquad (*)$$

where

$$\begin{aligned} \theta_1 &= \text{friction velocity} \\ \theta_2 &= \text{zero point displacement} \\ \theta_3 &= \text{roughness length} \end{aligned}$$

This relationship is not likely to hold exactly for our data due to measurement error and deviations from the ideal conditions under which the relationship is theorized to hold. Therefore, we fit a stochastic version of (*):

$$y_i = \theta_1 \log\{x_i(1 - \theta_2/\theta_3) - 1/\theta_3\} + e_i, \quad i = 1, \ldots, 5. \qquad (**)$$

- Notice the model (**) is nonlinear since $\partial f(x, \boldsymbol{\theta})/(\partial \theta_j)$ depends upon $\boldsymbol{\theta}$ for all $j = 1, 2, 3$. E.g.,

$$\frac{\partial f(x, \boldsymbol{\theta})}{\partial \theta_1} = \log\{x_i(1 - \theta_2/\theta_3) - 1/\theta_3\}.$$

- For now, we skip the details of how to fit a nonlinear such as (**), but in R it can be done using nonlinear least squares with the nls() function. The parameter estimates (standard errors) turn out to be

$$\hat{\theta}_1 = 115.1(2.04), \quad \hat{\theta}_2 = -.0595(.00546), \quad \hat{\theta}_3 = .0454(.0132).$$

- The fitted regression line for model (**) fits the observed data much more closely than that of either linear model. In addition, the residual standard deviation of the nonlinear model, $s = 1.87$, is much smaller than for the linear models.

- Perhaps most importantly, this model reflects what is know about the relationship between wind speed and the height at which it is measured, and the parameter estimates have specific, useful interpretations in terms of the physics/meteorology of the problem.

Classes of Nonlinear Models:

1. Yield-Density Models.

   – Common in agriculture (e.g., forestry).

   – Models describe the relationship between the yield of a crop and the density of planting.

   Let

   $$X = \text{plant density in plants/unit area}$$
   $$R = \text{yield/plant.}$$

   Then

   $$W = XR = \text{total yield per unit area (e.g., acre).}$$

   Two common yield density relationships:

i. "asymptotic relationship" (most crops. e.g., carrots, beans, tomatoes)

ii. "parabolic relationship" (e.g., sweet corn, cotton)

A common model for the asymptotic case is

$$R = (\theta_1 + \theta_2 X)^{-1}.$$

Notice that as $X \to 0$, $R \to 1/\theta_1 \Rightarrow \theta_1^{-1}$ has interpretation as the "genetic potential" of a crop uninhibited by competition. In addition, as $X \to \infty$, $W = XR \to 1/\theta_2$, $\Rightarrow \theta_2^{-1} =$"environmental potential".

For observed data, the model

$$R_i = (\theta_1 + \theta_2 X_i)^{-1} \exp(e_i)$$

is often used; or, transforming,

$$Y_i = \log(\theta_1 + \theta_2 X_i) + e_i,$$

where $Y_i = -\log(R_i)$.

- This model is not based on any theory, its simply an empirical model whose parameters have meaningful interpretations in this context.

An important model that is used in the asymptotic case of the yield-density curve but also in a wide variety of other applications is the **Asymptotic Regression Model:**

$$Y_i = \alpha - \beta\gamma^{X_i} + e_i, \quad i = 1, \ldots, n.$$

This model yields a curve that has the following shape:

- As with many named classes of nonlinear models, a variety of different parameterizations of the asymptotic regression models are used:

$$\begin{aligned} Y_i &= \theta_1 - \theta_2 e^{-\theta_3 X_i} + e_i, \qquad (\text{here, } e^{-\theta_3} = \gamma) \\ Y_i &= \theta_1 - \theta_1 e^{-(X_i+\theta_2)\theta_3} + e_i, \\ Y_i &= \theta_1 - e^{-(\theta_2+\theta_3 X_i)} + e_i \end{aligned}$$

- In nonlinear regression, the parameterization you use affects the algorithms used to solve the models, the properties of the estimators, and the accuracy of approximations used for inference. To a much greater degree than in linear models, it is important to choose the right parameterization! We will return to this point later.

2. Growth Models.

   – These models relate the size or change in size of some entity to time.

   The most common shape for growth curves is a "sigmoidal curve":

   Let R =a measure of size, and X =time. There are several commonly used growth curves that capture the sigmoidal shape for R as a function of X; e.g.,

$$R = \theta_1 \exp\{-\exp(\theta_2 - \theta_3 X)\}, \qquad \text{(Gompertz)}$$

$$R = \frac{\theta_1}{1 + \exp(\theta_2 - \theta_3 X)}, \qquad \text{(simple logistic)}$$

$$R = \frac{\theta_1 X^{\theta_2} + \theta_3 \theta_4}{X^{\theta_2} + \theta_3}, \qquad \text{(Morgan-Mercer-Flodin)}$$

$$R = \theta_1 \{1 - \exp(-\theta_2 X)\}^{\theta_3}, \qquad \text{(Chapman-Richards)}$$

- In fitting these models to data, either an additive or multiplicative error term may be appropriate. E.g., in the case of the logistic model we might consider an additive version:

$$Y_i = \frac{\theta_1}{1 + \exp(\theta_2 - \theta_3 X_i)} + e_i, \quad \text{where } Y_i = R_i$$

  or a multiplicative version:

$$Y_i = \log(\theta_1) - \log\{1 + \exp(\theta_2 - \theta_3 X_i)\} + e_i, \quad \text{where } Y_i = \log(R_i).$$

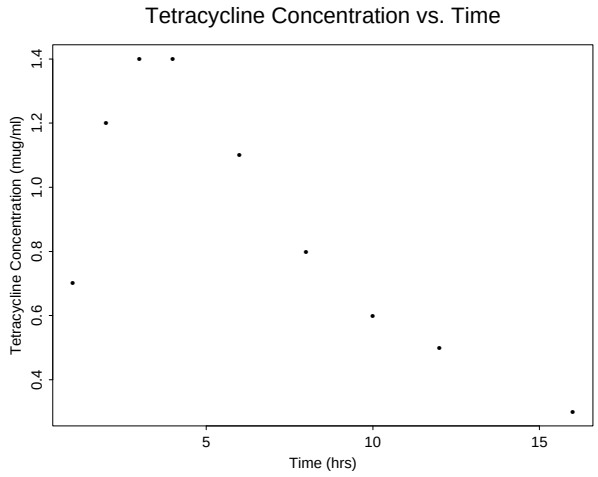
3. Compartmental Models and Other Models Based on Systems of Differential Equations.

   – Compartmental models are mechanistic models where one or more measurements of some physical process is related to time, inputs to the system, and other explanatory variables through a compartmental system.

   – A compartmental system is "a system which is made up of a finite number of macroscopic subsystems, called compartments or pools, each of which is homogeneous and well mixed, and the compartments interact by exchanging materials. There may be inputs from the environment into one or more of the compartments, and there may be outputs (excretion) from one or more the compartments to the environment." (Seber & Wild, p.367)

   – Compartmental models are common in chemical kinetics, toxicology, hydrology, geology, and pharmacokinetics.

   As an example from pharmacokinetics, consider the data in the following scatterplot on tetracycline concentration over time.



The data come from a study in which a tetracycline compound was administered to a subject orally, and the concentration of tetracycline hydrochloride in the blood serum was measured over a period of 16 hours (the data are in Appendix A1.14 of Bates & Watts).

A simple compartmental model for the biological system determining tetracycline concentration in serum is one that hypothesizes

a. a gut compartment into which the chemical is introduced,

b. a blood compartment which absorbs the chemicals from the gut, and

c. an elimination path.

This simple two-compartment open model can be represented in a compartment diagram as follows:

Here, γ_1 and γ_2 represent the the concentrations of the chemical in compartments 1 and 2, respectively, and θ_1 and θ_2 represents rates of transfer into and out of compartment 2, respectively.

Under the assumption of *first-order (linear) kinetics*, it is assumed that at time t, the rate of elimination from any compartment is proportional to $\gamma(t)$, the concentration currently in that compartment.

Thus the rates of change in the concentrations in the two compartments in the model represented above are

$$\frac{\partial \gamma_1(t)}{\partial t} = -\theta_1 \gamma_1(t)$$
$$\frac{\partial \gamma_2(t)}{\partial t} = \theta_1 \gamma_1(t) - \theta_2 \gamma_2(t)$$

Differential equations solutions for linear compartmental models generally take the form of linear combinations of exponentials. Therefore, these models are nonlinear models that can be fit using methods similar to those used for yield-density models, growth curve models, etc.

For example, the solution for $\gamma_2(t)$, the concentration in blood serum at time t is

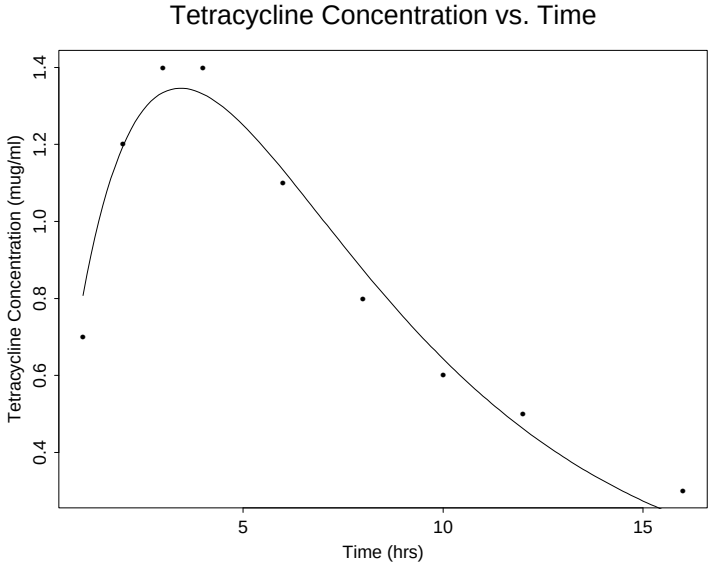
$$\gamma_2(t) = \frac{\theta_3 \theta_1 (e^{-\theta_1 t} - e^{-\theta_2 t})}{\theta_2 - \theta_1}.$$

- Here, θ_3 is the amount of drug (tetracycline) ingested initially (at time $t = 0$).

Therefore, we might try the additive error model

$$y_i = \frac{\theta_3\theta_1(e^{-\theta_1 t_i} - e^{-\theta_2 t_i})}{\theta_2 - \theta_1} + e_i, \quad i = 1, \ldots, n,$$

to model the tetracycline data. The resulting fitted regression curve is displayed below.



4. Multiphase and Spline Regressions.

   In multiphase and spline regression models, the expectation function for the regression of y on x, $\mathrm{E}(y) = f(x; \boldsymbol{\theta})$, is obtained by piecing together different curves over different intervals.

That is, in multiphase and spline models,

$$f(x;\boldsymbol{\theta},\boldsymbol{\alpha}) = \begin{cases} f_1(x;\boldsymbol{\theta}_1), & x \le \alpha_1; \\ f_2(x;\boldsymbol{\theta}_2), & \alpha_1 < x \le \alpha_2; \\ \vdots & \vdots \\ f_D(x;\boldsymbol{\theta}_D), & \alpha_{D-1} < x. \end{cases}$$

  - Here, the expectation function $f(x;\boldsymbol{\theta},\boldsymbol{\alpha})$ is defined by different functions $(f_1, f_2, \ldots, f_D)$ on different intervals, and typically the endpoints of the intervals are unknown and must be estimated.

  - The D submodels are referred to as *phases*, and the α's are *changepoints.*

  - Multiphase models are intended for situations in which (a) the number of phases is small; (b) the behavior in each phase is well-described by a simple parametric function like a line or quadratic; and (c) there are fairly abrupt changes between regimes.

  - Spline models have the same piecewise form, but the individual phase models f_d, $d = 1, \ldots, D$, are always polynomials and the emphasis is on joining these "splines" to obtain a smooth and very flexible function to capture the underlying regression form.

- After presenting methodology for the general nonlinear regression model, we'll come back and spend some more time on certain special class of nonlinear models as time allows.

Special Types of Nonlinear Models:

1. Transformably Linear Models:

   Suppose we observe z_i, $\mathbf{x}_i$, $i = 1, \ldots, n$, where

   $$\mathrm{E}(z_i) = f(\mathbf{x}_i; \boldsymbol{\theta}).$$

   If it is possible to find a transformation $h(\cdot)$ such that $y_i = h(z_i)$ satisfies a linear regression model, then we say that the expectation function f is **transformably linear.**

- We must be careful about assumptions on error terms when using linearizing transformations!

  Suppose

  $$\mathrm{E}(z_i) = e^{\alpha+\beta x_i}.$$

  If the error in z_i is proportional to the expected magnitude of z_i but otherwise independent of x_i ("constant relative error") then we can write a model for z_i as

  $$z_i = e^{\alpha+\beta x_i}(1 + e_i),$$

  where

  $$\mathrm{E}(e_i) = 0, \quad \text{and} \quad \mathrm{var}(e_i) = \sigma^2 \quad \text{(constant variance)}.$$

  Equivalently,

  $$z_i = e^{\alpha+\beta x_i} + e_i^* \qquad (*)$$

  where $e_i^* = \mathrm{E}(z_i)e_i$ has mean 0 and variance $\mathrm{var}(e_i^*) = \sigma^2\{\mathrm{E}(z_i)\}^2$.

  If we transform to linearity by taking logs we get

  $$\begin{aligned} y_i &= \alpha + \beta x_i + \log(1 + e_i), \quad \text{where } y_i = \log(z_i) \\ &= \alpha + \beta x_i + \tilde{e}_i, \end{aligned}$$

  where $\tilde{e}_i = \log(1 + e_i)$ has mean $\mathrm{E}(\tilde{e}_i) \approx \mathrm{E}(e_i) = 0$ (for small e_i), and variance $\mathrm{var}(\tilde{e}_i)$ that is independent of x_i.

– That is, in (*) we had a model in the original scale with an additive error with variance proportional to the square of the mean. This transformed to a model with nearly constant variance on the log scale.

– Another way to say this is that if the original model had had a multiplicative error,

$$z_i = e^{\alpha+\beta x_i} \underbrace{e^{u_i}}_{\text{error term}} = e^{\alpha+\beta x_i+u_i},$$

where $\mathrm{E}(u_i) = 0$ and $\mathrm{var}(u_i) = \sigma^2$, then the transformed model would be

$$y_i = \alpha + \beta x_i + u_i, \text{where } \mathrm{E}(u_i) = 0, \mathrm{var}(u_i) = \sigma^2.$$

<u>However,</u> if instead of (*) we had additive homoscedastic errors on the original scale:

$$\begin{aligned} z_i &= e^{\alpha+\beta x_i} + e_i, \quad \text{where } \mathrm{E}(e_i) = 0, \mathrm{var}(e_i) = \sigma^2. \\ &= e^{\alpha+\beta x_i}\left(1 + \frac{e_i}{\mathrm{E}(z_i)}\right). \end{aligned}$$

Then

$$\begin{aligned} y_i &= \alpha + \beta x_i + \log\left(1 + \frac{e_i}{\mathrm{E}(z_i)}\right) \\ &= \alpha + \beta x_i + v_i, \end{aligned}$$

where now the additive error, $v_i = \log\left(1 + e_i/\mathrm{E}(z_i)\right)$ has mean $\mathrm{E}(v_i) \approx \mathrm{E}\{e_i/\mathrm{E}(z_i)\} = 0$ (for e_i small compared with $\mathrm{E}(z_i)$) and variance $\mathrm{var}(v_i)$ that varies with $\mathrm{E}(z_i)$ (heteroscedasticity).

To summarize:

- homoscedasticity on original scale generally induces heteroscedasticity on transformed scale. In this case, we should either fit a homoscedastic nonlinear model to the original data (NLS) or use WLS to fit a heteroscedastic linear model on the transformed scale.

- Certain types of heteroscedasticity on the original scale can lead to homoscedasticity on the transformed scale. In this case, either nonlinear WLS on the original scale or linear OLS on the transfored scale can be used.

- As long as the error variance is accounted for correctly, working on either the linear or nonlinear scale may be appropriate. Desirability of interpretable parameter estimates may argue for nonlinear model on original scale.

- We assume nonlinear model will be fit. Even in this case availability of a linear transformation can be very useful (e.g., for obtaining starting values).

- Note that transformations affect entire distribution of the error terms, not just their variance. $\Rightarrow$ normal additive errors on original scale lead to non-normal additive errors on the transformed scale.

2. Partially Linear Models:

Consider

$$y_i = \theta_1(1 - e^{-\theta_2 x_i}) + e_i.$$

If θ_2 is known (i.e., fixed) then the model is linear:

$$y_i = \theta_1 w_i + e_i, \qquad \text{where } w_i = 1 - e^{-\theta_2 x_i}.$$

θ_1 is a **conditionally linear parameter** if for fixed $\theta_2, \ldots, \theta_p$ the model is linear.

If there is at least one conditionally linear parameter in the model then the model is partially linear.

- For partially linear models, some procedures (e.g., obtaining starting values) are simplified.

Geometry of the Expectation Surface:

Consider the linear model with $n = 2$, $p = 1$ and model equation

$$y_i = \theta x_i + e_i, \quad i = 1, 2$$

where $x_1 = 1, x_2 = 2$.

The expectation plane is the set of all 2×1 vectors $\boldsymbol{\eta}(\theta) = \theta \begin{pmatrix} 1 \\ 2 \end{pmatrix}$. We plot this expectation plane below

Features:

1. Linearity (with respect to θ).

    a. The effect of changing θ by δ units is the same for all θ.

    b. θ points with equal spacing correspond to $\boldsymbol{\eta}$ points with equal spacing.

    c. Line segments in the parameter space correspond to line segments in the expectation plane.

    – (a)–(c) above are all essentially restatements of the same idea, linearity.

2. Expectation is of infinite extent.

Now consider the nonlinear model:

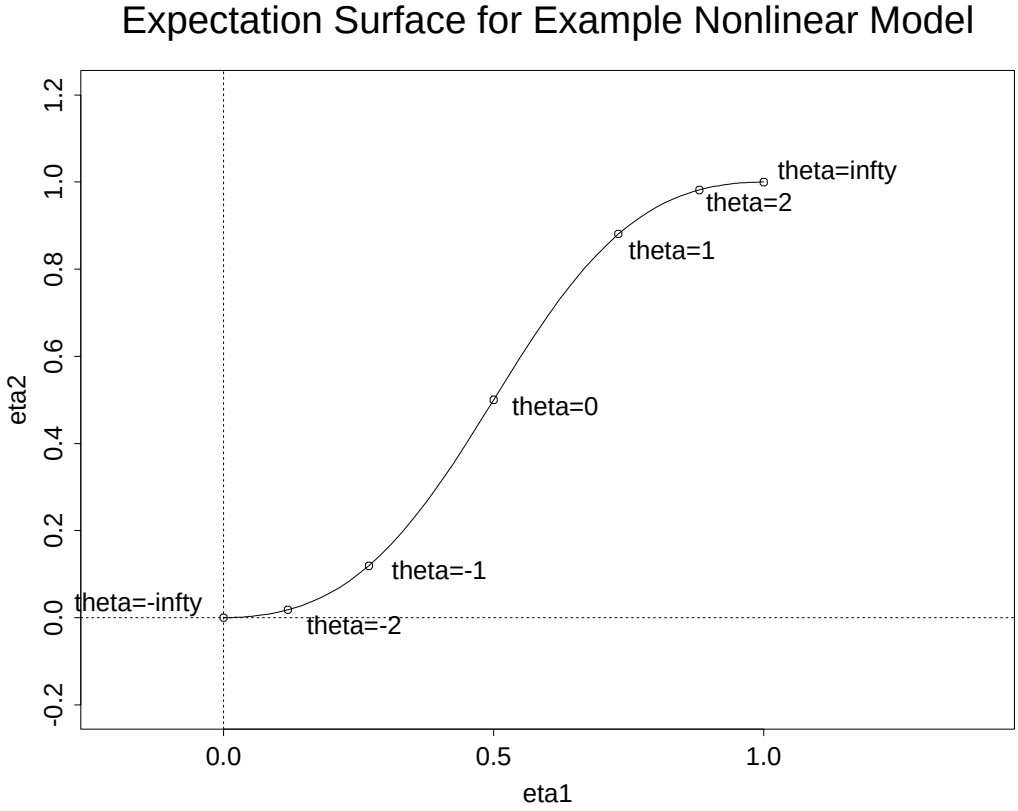
$$y_i = \frac{1}{1 + e^{-\theta x_i}} + e_i. \qquad (*)$$

Again, take $\mathbf{x} = (x_1, x_2)^T = (1, 2)^T$. Then the *expectation surface* is

$$\boldsymbol{\eta}(\theta) = \begin{pmatrix} (1 + e^{-\theta})^{-1} \\ (1 + e^{-2\theta})^{-1} \end{pmatrix}, \quad \text{(1-dim. surface in 2-space)}$$

We can plot this surface in 2-space by plugging a few values of θ into the formula for $\boldsymbol{\eta}(\theta) = \begin{pmatrix} \eta_1(\theta) \\ \eta_2(\theta) \end{pmatrix}$:

θ	η_1	η_2
$-\infty$	0	0
-2	.119	.0180
-1	.269	.119
0	.500	.500
1	.731	.881
2	.881	.982
∞	1	1



Features:

1. Nonlinearity:

    a. Effect of changing θ by 1 unit depends upon the value of θ.

    b. θ points with equal spacing correspond to $\boldsymbol{\eta}$ points with unequal spacing.

    c. Line segments in the parameter space correspond to curves in the expectation space.

2. The expectation surface may be of finite extent.

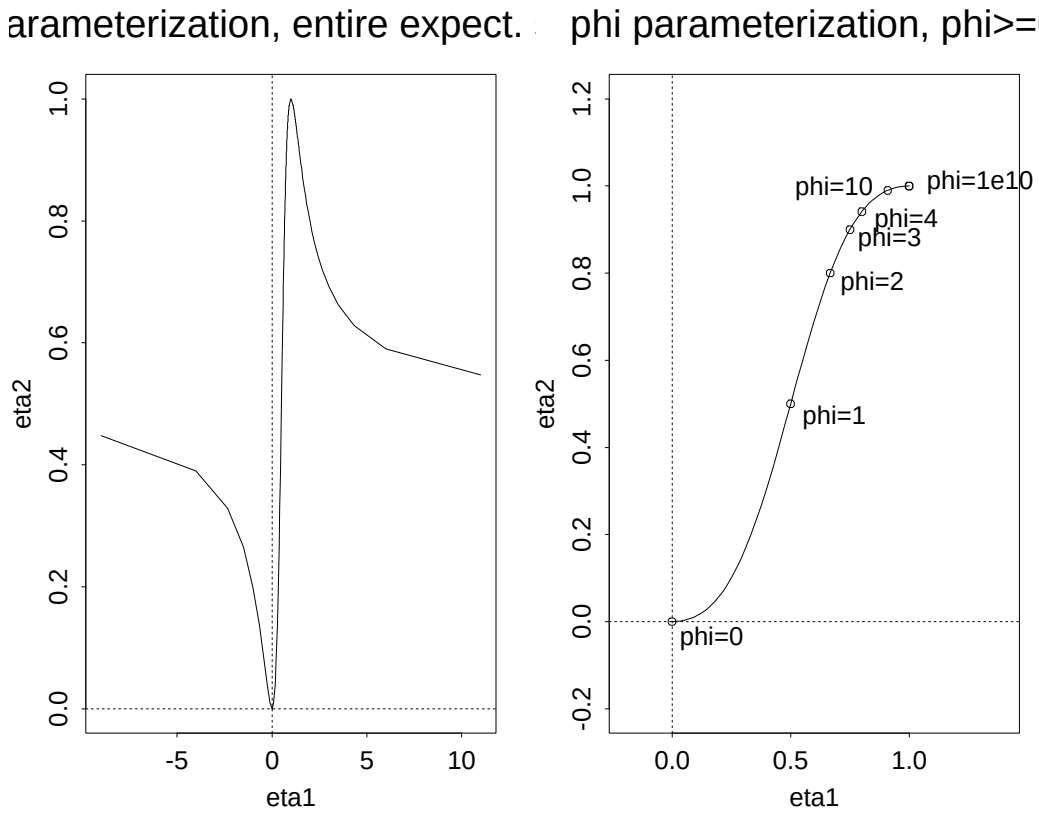
- The curved shape of the expectation surface is invariant to reparameterization. That is, the degree of curvature is the same no matter what parameterization is used. This aspect of nonlinearity is known as **intrinsic nonlinearity.**

- The extent to which equally-spaced $\boldsymbol{\theta}$ points map to unequally spaced $\boldsymbol{\eta}$-points is known a **parameter-effects nonlinearity.** This type of nonlinearity depends upon the parameterization chosen for the model. A good parameterization leads to small parameter-effects nonlinearity.

For example, consider the following reparameterization of model (*) on p.63:

$$y_i = \frac{1}{1+\phi^{-x_i}} + e_i, \qquad \text{here, } \phi = e^{\theta}.$$

We can again plot the expectation surface based on the new parameterization. The result is as follows:



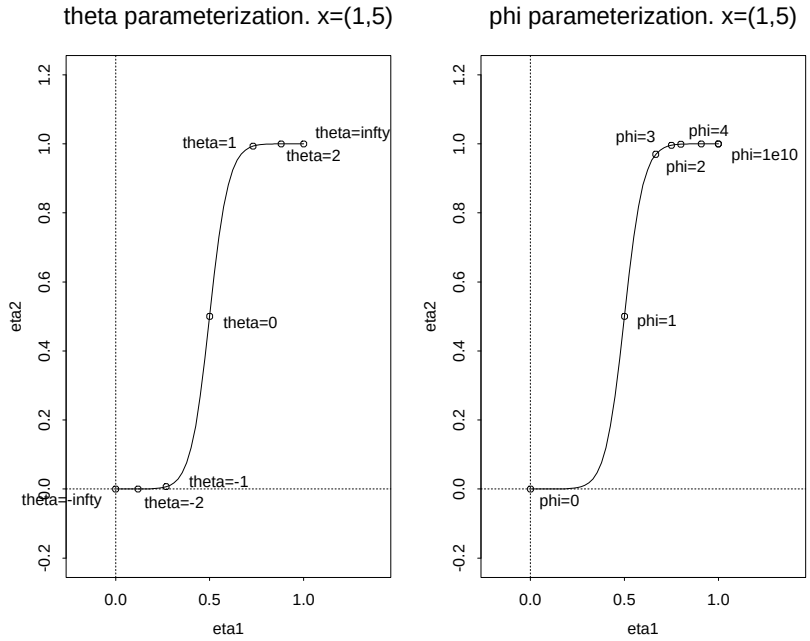
- Notice that under the reparameterization, the expectation surface is of infinite extent.

- Restricting attention to the region $\phi \geq 0$ (corresponding to the entire range of θ) depicted in the right-hand plot, we see that the θ-parameterization has the exact same intrinsic nonlinearity as the $\phi-$parameterization.

- However, the θ-parameterization has considerably less parameter-effects nonlinearity than the ϕ-parameterization. This is so because under the ϕ-parameterization, $\boldsymbol{\eta}$-curve segments corresponding to values of ϕ one unit apart differ in length to a much greater extent than under the θ-parameterization.

- Note that the degree of both intrinsic nonlinearity and parameter-effects nonlinearity in a particular model depend upon the values of $\mathbf{x}$. Two models with the same form and parameterization but different values of the explanatory variables will have different nonlinearities of both types.

E.g., if we change $\mathbf{x}$ from $(1, 2)^T$ to $(-1, 5)^T$ we get the following curves.



- Later in the course we will come back to the ideas of intrinsic and parameter-effects nonlinearity and discuss measures of these two-types of curvature. For now, suffice it to say that it is desirable to choose parameterizations with low parameter-effects nonlinearity.

Linear Approximations:

Let $h(\cdot)$ denote a (twice differentiable) scalar valued function of a scalar argument. For any fixed points u and u^*, **Taylor's Theorem** says

$$h(u) = h(u^*) + h'(u^*)(u - u^*) + \frac{1}{2}h''(u^{**})(u - u^*)^2,$$

where $h'(u^*) = \frac{\partial h(u)}{\partial u}\Big|_{u=u^*}$, $h''(u^{**}) = \frac{\partial^2 h(u)}{\partial u^2}\Big|_{u=u^{**}}$, and u^{**} is a point between u and u^*.

If u^* is close to u, then the last term in this expansion will be small relative to the rest and we have

$$h(u) \approx h(u^*) + h'(u^*)(u - u^*) \equiv \tilde{h}(u)$$

for u close to u^*.

- This is known as a **first-order (linear) Taylor series approximation**.

This approximation is

i. Linear. ($\tilde{h}(u)$ is a linear function of u.)

ii. Local. (Only valid for u near u^*.)

Now suppose h is a scalar-valued function of $\mathbf{u}$, a $p \times 1$ vector. For this situation the above linear Taylor series approximation generalizes to

$$h(\mathbf{u}) \approx h(\mathbf{u}^*) + \underbrace{\frac{\partial h(\mathbf{u}^*)}{\partial \mathbf{u}^T}}_{1\times p}\underbrace{(\mathbf{u} - \mathbf{u}^*)}_{p\times 1}$$

for $\mathbf{u}$ close to $\mathbf{u}^*$ (that is, when $\|\mathbf{u} - \mathbf{u}^*\|$ is small).

Here,

$$\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_p \end{pmatrix}, \mathbf{u}^* = \begin{pmatrix} u_1^* \\ \vdots \\ u_p^* \end{pmatrix}, \quad \text{and} \quad \frac{\partial h(\mathbf{u}^*)}{\partial \mathbf{u}^T} = \begin{pmatrix} \frac{\partial h(\mathbf{u})}{\partial u_1} & \frac{\partial h(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial h(\mathbf{u})}{\partial u_p} \end{pmatrix}$$

If we prefer a non-vector notation form, we can multiply out the second term and write this approximation in an equivalent form as follows:

$$h(\mathbf{u}) \approx h(\mathbf{u}^*) + (u_1 - u_1^*)\frac{\partial h(\mathbf{u})}{\partial u_1}\Big|_{\mathbf{u}=\mathbf{u}^*} + \cdots + (u_p - u_p^*)\frac{\partial h(\mathbf{u})}{\partial u_p}\Big|_{\mathbf{u}=\mathbf{u}^*}.$$

Consider

$$\begin{aligned} y_i &= f(\mathbf{x}_i, \boldsymbol{\theta}) + e_i, \\ &= \eta_i(\boldsymbol{\theta}) + e_i, \end{aligned} \qquad i = 1, \ldots, n,$$

or, in vector form,

$$\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \mathbf{e}.$$

Estimation and inference about $\boldsymbol{\theta}$ is easy if $f(\mathbf{x}_i, \boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$. This suggests using a linear Taylor series approximation of $f(\mathbf{x}_i, \boldsymbol{\theta})$.

For $\boldsymbol{\theta}$ near $\boldsymbol{\theta}^*$,

$$f(\mathbf{x}_i, \boldsymbol{\theta}) \approx f(\mathbf{x}_i, \boldsymbol{\theta}^*) + (\theta_1 - \theta_1^*)V_{i1} + \cdots + (\theta_p - \theta_p^*)V_{ip},$$

for each $i = 1, \ldots, n$, where

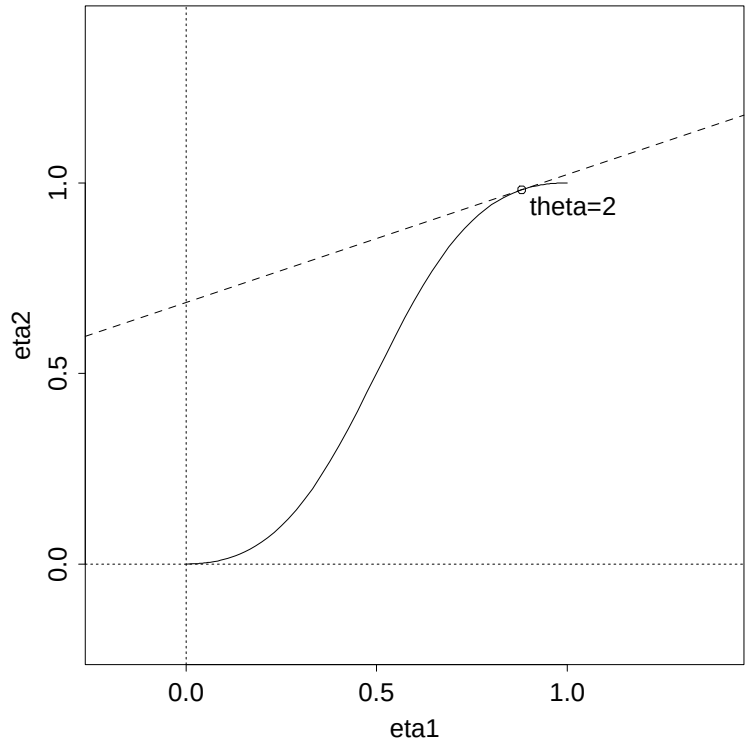
$$V_{ij} = \frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \theta_j}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}.$$

Stacking these n approximations on top of one another in vector form we have

$$\boldsymbol{\eta}(\boldsymbol{\theta}) \approx \boldsymbol{\eta}(\boldsymbol{\theta}^*) + \mathbf{V}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*),$$

where $\mathbf{V}(\boldsymbol{\theta}^*)$ is the $n \times p$ matrix with $(i,j)^{\text{th}}$ element V_{ij}. (I.e., $\mathbf{V}(\boldsymbol{\theta}^*)$ has i^{th} row $\frac{\partial f(\mathbf{x}_i, \theta)}{\partial \boldsymbol{\theta}^T}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$.)

A picture of the Taylor series linear approximation taken at $\theta^* = 2$:



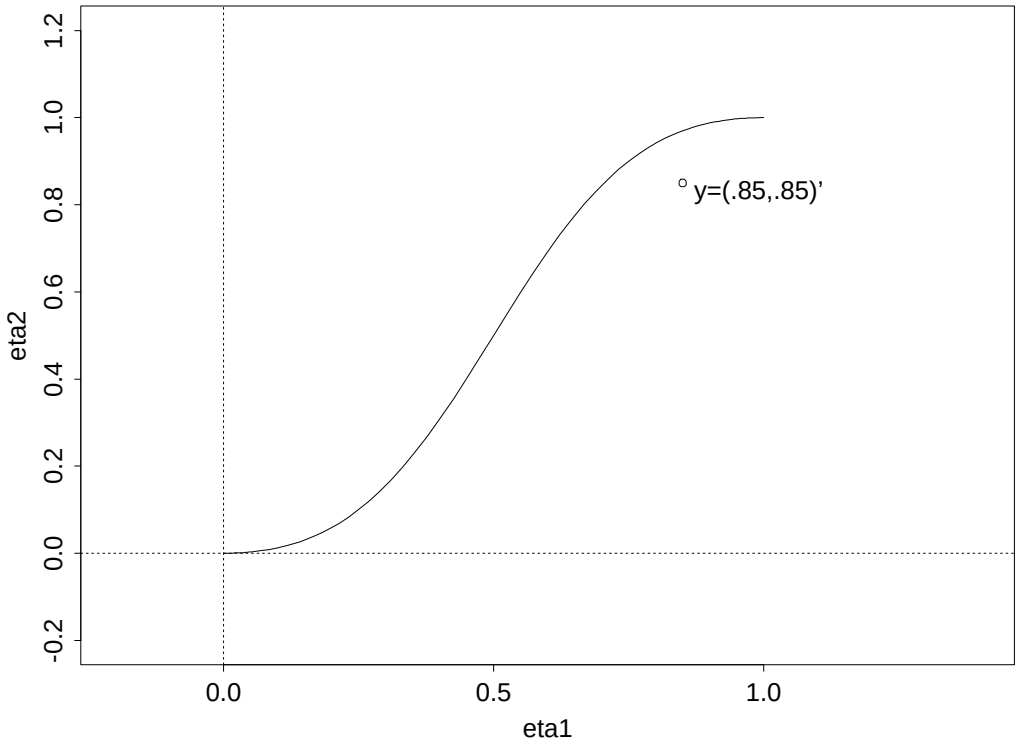
Estimation of $\boldsymbol{\theta}$:

Under the assumption that $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, the MLE of $\boldsymbol{\theta}$ is also the **nonlinear least-squares** estimator. I.e., it is the value of $\boldsymbol{\theta}$ that minimizes

$$\|\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2 = \sum_{i=1}^{n} \{y_i - f(\mathbf{x}_i, \boldsymbol{\theta})\}^2.$$

- This is still the (ordinary) least squares criterion. The only difference is that $f(\mathbf{x}_i, \boldsymbol{\theta}) \neq \mathbf{x}_i^T\boldsymbol{\theta}$ (nonlinearity).
- As in linear least squares, we need to find the point on the expectation surface, $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}})$, that is closest to $\mathbf{y}$ in terms of Euclidean distance.
- This is a harder problem because the expectation surface is no longer a plane.

E.g., Suppose $\mathbf{y} = (.85, .85)^T$ in our example from p.63–64.



- In the nonlinear case, finding the point $\hat{\boldsymbol{\eta}}$ on the expectation surface is hard because of intrinsic nonlinearity.
- Once we find $\hat{\boldsymbol{\eta}}$ we must find the value of $\boldsymbol{\theta}$ solving $\boldsymbol{\eta}(\boldsymbol{\theta}) = \hat{\boldsymbol{\eta}}$. In linear regression this step is easy because the mapping from $\boldsymbol{\beta}$ to $\boldsymbol{\eta}$ is linear and invertible. In nonlinear regression, this step is harder, because of both intrinsic and parameter-effects nonlinearity.

Nonlinear least squares: find the $\boldsymbol{\theta}$ to minimize

$$S(\boldsymbol{\theta}) = \|\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2 = \sum_{i=1}^{n} \{y_i - f(\mathbf{x}_i, \boldsymbol{\theta})\}^2.$$

Q: How do we minimize $S(\boldsymbol{\theta})$?

A: Solve normal equations.

Taking derivatives we get

$$\frac{\partial S(\boldsymbol{\theta})}{\partial \theta_j} = -2 \sum_{i=1}^{n} \{y_i - f(\mathbf{x}_i, \boldsymbol{\theta})\} \frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \theta_j}, \qquad j = 1, \ldots, p,$$

or, in matrix notation,

$$\frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = -2[\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})]^T \underbrace{\mathbf{V}(\boldsymbol{\theta})}_{\substack{n \times p \text{ derivative matrix}}}$$

Therefore, $\hat{\boldsymbol{\theta}}$ satisfies

$$[\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}})]^T \mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \qquad (\dagger)$$

- Recall that in linear regression, the derivative matrix was $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{X}$ and our normal equation took the form of the orthogonality condition:

  $$[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}]^T \mathbf{X} = \mathbf{0},$$

  which says, the residual vector $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is orthogonal to the expectation plane (all vectors of the form $\mathbf{X}\boldsymbol{\beta}$).

- In the nonlinear case, (†) says that the residual vector $\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}})$ is orthogonal to the "tangent plane" to the expectation surface at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

- Notice (†) is nonlinear in $\boldsymbol{\theta}$ and doesn't, in general, yield a closed form expressionfor the solution, $\hat{\boldsymbol{\theta}}$.

  Q: How do we solve a nonlinear set of equations like (†)?

  A: Usually requires an iterative method. For example,...

Gauss-Newton Method

The G-N method proceeds by

1. Obtaining a starting value $\hat{\boldsymbol{\theta}}^0$.
2. Using a linear approximation to $\boldsymbol{\eta}(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ near $\hat{\boldsymbol{\theta}}^0$.
3. Using linear regression methods to "estimate $\boldsymbol{\theta}$"; i.e., to update $\hat{\boldsymbol{\theta}}^0$ to $\hat{\boldsymbol{\theta}}^1$.
4. Repeat steps 2 and 3 until convergence.

Let $\mathbf{V}^0 = \mathbf{V}(\hat{\boldsymbol{\theta}}^0)$. Then,

$$\boldsymbol{\eta}(\boldsymbol{\theta}) \approx \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^0) + \mathbf{V}^0(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^0)$$

for $\boldsymbol{\theta}$ near $\hat{\boldsymbol{\theta}}^0$. It follows that

$$\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}) \approx \underbrace{\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^0)}_{\equiv \mathbf{z}^0} - \mathbf{V}^0 \underbrace{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^0)}_{\equiv \boldsymbol{\delta}} \equiv \mathbf{z}^0 - \mathbf{V}^0\boldsymbol{\delta}$$

Therefore, for $\boldsymbol{\theta}$ near $\hat{\boldsymbol{\theta}}^0$, choosing $\boldsymbol{\theta}$ to minimize $\|\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2$ is "approximately equivalent to" (should give nearly the same result as) the problem of choosing $\boldsymbol{\delta}$ to minimize $\|\mathbf{z}^0 - \mathbf{V}^0\boldsymbol{\delta}\|^2$.

- This minimization can be carried out using linear regression. E.g., we could use $\hat{\boldsymbol{\delta}} = \{(\mathbf{V}^0)^T\mathbf{V}^0\}^{-1}(\mathbf{V}^0)^T\mathbf{z}^0$. However, this formula, while correct, is not the best way to do the computations - either here, or in regular linear regression.

This process consists of two steps:

a. Obtaining the point $\hat{\boldsymbol{\eta}}^* = \mathbf{V}^0\hat{\boldsymbol{\delta}}$.

b. Determining $\hat{\boldsymbol{\delta}}$ from $\hat{\boldsymbol{\eta}}^*$.

Once we have $\hat{\boldsymbol{\delta}}$, we can easily translate to $\hat{\boldsymbol{\theta}}$. From its definition,

$$\boldsymbol{\delta} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^0 \qquad \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^0 + \boldsymbol{\delta}.$$

Therefore, we can update $\hat{\boldsymbol{\theta}}^0$ to $\hat{\boldsymbol{\theta}}^1$ via

$$\hat{\boldsymbol{\theta}}^1 = \hat{\boldsymbol{\theta}}^0 + \hat{\boldsymbol{\delta}} \qquad (*)$$

- Because of this updating formula, we call $\hat{\boldsymbol{\delta}}$ the **Gauss-Newton increment.**
- $\hat{\boldsymbol{\theta}}$ is updated this way until convergence.

Complication: There is no guarantee that the update will reduce the objective function $S(\boldsymbol{\theta})$. I.e., no guarantee that $S(\hat{\boldsymbol{\theta}}^1) < S(\hat{\boldsymbol{\theta}}^0)$.

We can fix this problem simply enough. If $S(\hat{\boldsymbol{\theta}}^1) \geq S(\hat{\boldsymbol{\theta}}^0)$ then replace the update (*) by

$$\hat{\boldsymbol{\theta}}^1 = \hat{\boldsymbol{\theta}}^0 + \lambda\hat{\boldsymbol{\delta}}, \qquad \text{where } 0 < \lambda \leq 1.$$

To choose λ, start with $\lambda = 1$ and then try λ values $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$ until we find a λ value that gives $S(\hat{\boldsymbol{\theta}}^1) < S(\hat{\boldsymbol{\theta}}^0)$.

- λ is called a **step factor.**

Each of the steps a and b in the description of the G-N algorithm above is made easier by using the **QR Decomposition**

The QR Decomposition:

- Our interest right now is in using the QR decomposition in the non-linear model, but just to keep things simple, let's return to the linear model for a moment.

In the linear model, the least squares problem is to find the value of $\boldsymbol{\beta}$ to minimize

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

We know that (if $\mathbf{X}^T\mathbf{X}$ is invertible) the answer is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

.

- However, the computation of $\hat{\boldsymbol{\beta}}$ via this formula can be computationally inefficient and error-prone. A better way is with the QR decomposition.

In general, an $n \times p$ ($n \geq p$) matrix $\mathbf{X}$ can be decomposed as

$$\mathbf{X} = \mathbf{QR}$$

where $\mathbf{Q}$ is an $n \times n$ **orthogonal matrix** (it has the property $\mathbf{QQ}^T = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_n$) and $\mathbf{R}$ is a $n \times p$ matrix of the form

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{pmatrix}$$

where $\mathbf{R}_1$ is a $p \times p$ upper-triangular matrix (it has zeros below the diagonal).

- From the fact that the last $n - p$ rows of $\mathbf{R}$ contain 0's we can write

$$\mathbf{X} = \mathbf{QR} = (\underbrace{\mathbf{Q}_1}_{n\times p}, \underbrace{\mathbf{Q}_2}_{n\times(n-p)}) \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{pmatrix} = \mathbf{Q}_1\mathbf{R}_1$$

where $\mathbf{Q}_1$ consists of the first p columns of $\mathbf{Q}$ and $\mathbf{R}_1$ contains the first p rows of $\mathbf{R}$.

Since we know that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, it also follows that the mean of $\mathbf{y}$, $\mathbf{X}\boldsymbol{\beta}$ has least squares estimate

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Applying the QR decomposition to $\mathbf{X}$ we get

$$\begin{aligned}
\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{Q}_1\mathbf{R}_1(\mathbf{R}_1^T \underbrace{\mathbf{Q}_1^T\mathbf{Q}_1}_{=\mathbf{I}} \mathbf{R}_1)^{-1}\mathbf{R}_1^T\mathbf{Q}_1^T\mathbf{y} \\
&= \mathbf{Q}_1\mathbf{R}_1(\mathbf{R}_1)^{-1}(\mathbf{R}_1^T)^{-1}\mathbf{R}_1^T\mathbf{Q}_1^T\mathbf{y} \\
&= \mathbf{Q}_1\mathbf{Q}_1^T\mathbf{y} \\
&= \mathbf{Q}\begin{pmatrix} \mathbf{Q}_1^T\mathbf{y} \\ \mathbf{0} \end{pmatrix} = \mathbf{Q}\begin{pmatrix} \mathbf{w}_1 \\ \mathbf{0} \end{pmatrix}
\end{aligned}$$

where $\mathbf{w}_1 = \mathbf{Q}_1\mathbf{y}$.

- So, we have that

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Q}\begin{pmatrix} \mathbf{w}_1 \\ \mathbf{0} \end{pmatrix} \qquad (\clubsuit)$$

  which allows us to get the least squares estimate of the mean of $\mathbf{y}$ without computing a matrix inverse - a computationally demanding an error-prone task.

All that's left to do is find $\hat{\boldsymbol{\beta}}$ once we have $\mathbf{X}\hat{\boldsymbol{\beta}}$. This is easy because applying the QR decomposition to $\mathbf{X}$ again in (♣), we get

$$\begin{aligned}
\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{Q}\begin{pmatrix}\mathbf{w}_1\\ \mathbf{0}\end{pmatrix}\\
\Rightarrow \qquad \mathbf{Q}\mathbf{R}\hat{\boldsymbol{\beta}} &= \mathbf{Q}\begin{pmatrix}\mathbf{w}_1\\ \mathbf{0}\end{pmatrix}\\
\Rightarrow \qquad \underbrace{\mathbf{Q}^T\mathbf{Q}}_{=\mathbf{I}}\mathbf{R}\hat{\boldsymbol{\beta}} &= \underbrace{\mathbf{Q}^T\mathbf{Q}}_{=\mathbf{I}}\begin{pmatrix}\mathbf{w}_1\\ \mathbf{0}\end{pmatrix}\\
\Rightarrow \qquad \mathbf{R}\hat{\boldsymbol{\beta}} &= \begin{pmatrix}\mathbf{w}_1\\ \mathbf{0}\end{pmatrix}\\
\Rightarrow \qquad \mathbf{R}_1\hat{\boldsymbol{\beta}} &= \mathbf{w}_1 \qquad (\diamondsuit)
\end{aligned}$$

- Solving for $\hat{\boldsymbol{\beta}}$ is now easy because $\mathbf{R}_1$ is upper-triangular!

E.g., suppose that in a simple linear regression problem we did the computations and found that $\mathbf{w}_1 = \begin{pmatrix}2\\ -1\end{pmatrix}$ and $\mathbf{R}_1 = \begin{pmatrix}1 & \frac{11}{3}\\ 0 & \frac{1}{3}\end{pmatrix}$ then to obtain $\hat{\boldsymbol{\beta}}$ we would solve

$$\begin{pmatrix}1 & \frac{11}{3}\\ 0 & \frac{1}{3}\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}2\\ -1\end{pmatrix}$$

$$\Rightarrow \qquad \begin{aligned}\beta_1 + \frac{11}{3}\beta_2 &= 2\\ \frac{1}{3}\beta_2 &= -1\end{aligned}$$

which solve easily to give

$$\hat{\beta}_2 = -3, \qquad \hat{\beta}_1 = 13.$$

Back to the G-N method for the nonlinear model:

Steps a and b above were aimed at finding the G-N increment $\hat{\boldsymbol{\delta}}$ by minimizing $\|\mathbf{z}^0 - \mathbf{V}^0\boldsymbol{\delta}\|^2$. The steps were:

a. Obtaining the point $\hat{\boldsymbol{\eta}}^* = \mathbf{V}^0\hat{\boldsymbol{\delta}}$.

b. Determining $\hat{\boldsymbol{\delta}}$ from $\hat{\boldsymbol{\eta}}^*$.

By utilizing the QR-decomposition $\mathbf{V}^0 = \mathbf{QR} = \mathbf{Q}_1\mathbf{R}_1$ in step a, we obtain the computationally efficient formula :

$$\hat{\boldsymbol{\eta}}^* = \mathbf{Q}_1\mathbf{Q}_1^T\mathbf{z}^0. \qquad \text{(compare p.76)}$$

(Note that $\hat{\boldsymbol{\eta}}^*$ lies on the tangent plane to the expectation surface, it's not on the expectation surface itself - hence the star notation. I.e., $\hat{\boldsymbol{\eta}}^*$ isn't equal to $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}})$ for any $\hat{\boldsymbol{\theta}}$.)

Then in step b, $\hat{\boldsymbol{\delta}}$ can be determined from $\hat{\boldsymbol{\eta}}^*$ by solving the linear equation, $\hat{\boldsymbol{\eta}}^* = \mathbf{V}^0\hat{\boldsymbol{\delta}}$. Or, since we can write this equation as $\mathbf{Q}_1\mathbf{R}_1\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\eta}}^* = \mathbf{Q}_1\mathbf{Q}_1^T\mathbf{z}^0$, this step reduces to solving

$$\mathbf{R}_1\hat{\boldsymbol{\delta}} = \mathbf{Q}_1^T\mathbf{z}^0, \qquad \text{(compare } (\diamondsuit)\text{)}$$

which is easy because $\mathbf{R}_1$ is upper triangular.

Geometry of Nonlinear Least-Squares:

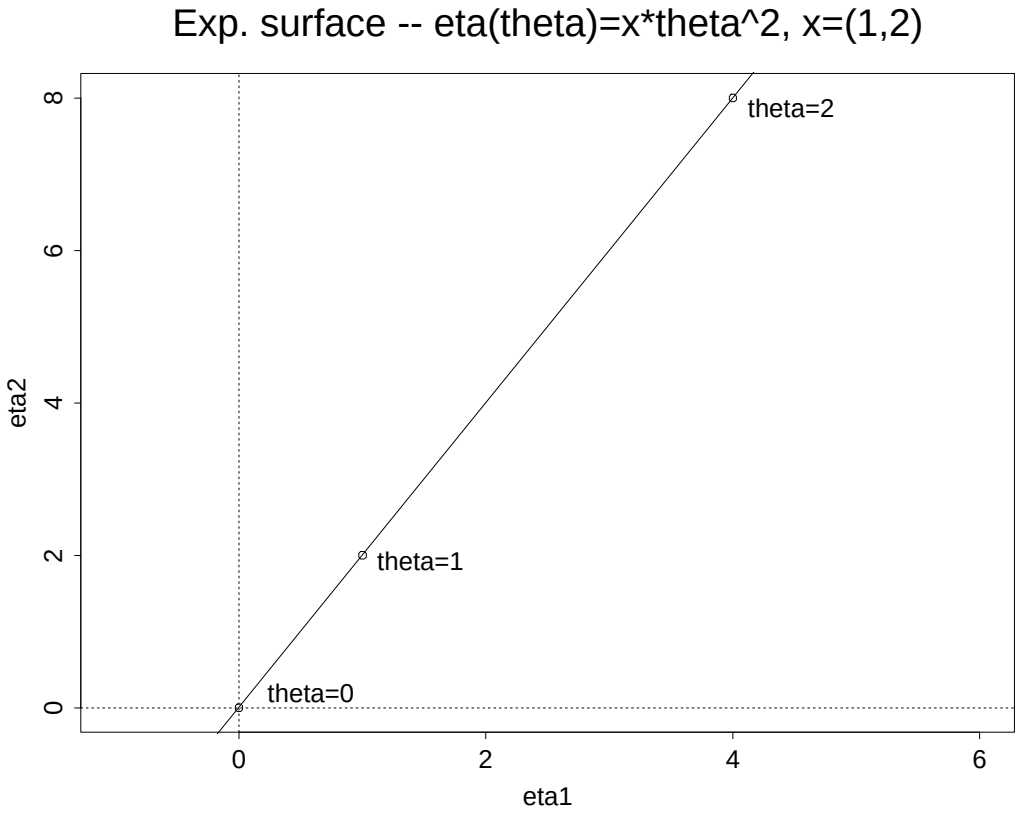
The $(k+1)^{\text{th}}$ G-N iteration consists of

1. approximation of $\boldsymbol{\eta}(\boldsymbol{\theta})$ by $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k) + \mathbf{V}^k(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^k)$ near $\hat{\boldsymbol{\theta}}^k$. This step consists of two parts:

   a. replacement of $\boldsymbol{\eta}(\boldsymbol{\theta})$ by the tangent plane (planar assumption).

   b. using a linear coordinate system $\mathbf{V}^k(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^k)$ on the tangent plane (uniform coordinate system).

   (a) and (b) are separate aspects of the linear approximation. That is, it is possible to have a linear (in some sense) approximation with non-uniform coordinates. E.g., if $f(x_i, \theta) = \theta^2 x_i$ then the expectation surface for this model is a plane with nonuniform coordinates with respect to θ:

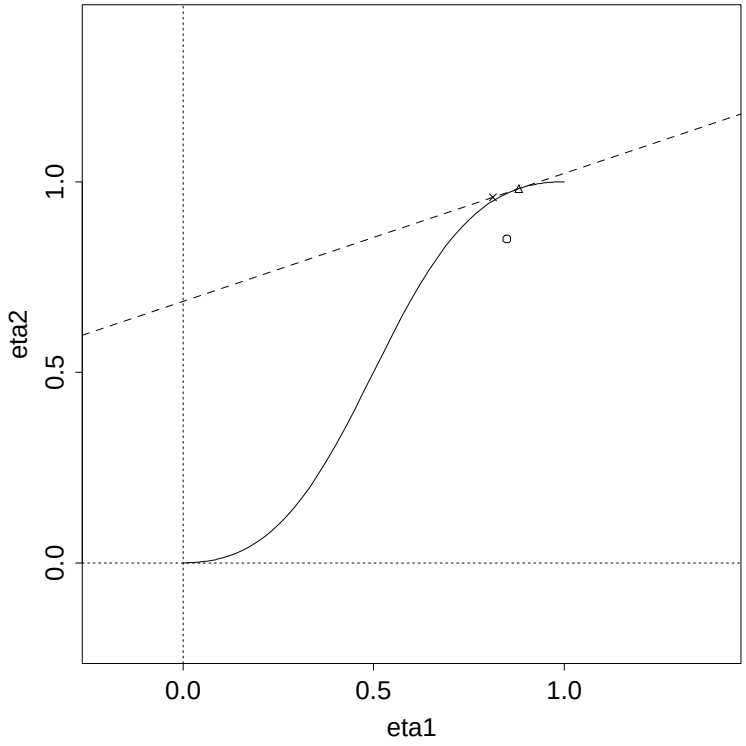


2. Finding the point on the tangent plane closest to $\mathbf{y}$. I.e., minimizing $\|\mathbf{z}^k - \mathbf{V}^k\boldsymbol{\delta}\|^2$.

3. Updating $\hat{\boldsymbol{\theta}}^k$ to $\hat{\boldsymbol{\theta}}^{k+1} = \hat{\boldsymbol{\theta}}^k + \hat{\boldsymbol{\delta}}$.

Example — $n = 2, p = 1$

Consider again the example from pp.63–64. Let $\mathbf{y} = (.85, .85)^T$ and suppose we take $\hat{\theta}^0 = 2$ as our starting value. Then the first ($k = 1$) G-N iteration looks like this geometrically:



- Once we obtain the point $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^0) + \mathbf{V}^0\hat{\boldsymbol{\delta}}$ on the tangent plane, we map back to the point $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^0 + \hat{\boldsymbol{\delta}}) = \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^1)$ on the expectation surface.

- If $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k + \hat{\boldsymbol{\delta}}) = \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k) + \mathbf{V}^k\hat{\boldsymbol{\delta}}$, then the linear approximation is exact and the MLE has been found. Unless the problem is linear, we will never obtain exact equality. However, we will obtain approximate equality when the G-N increment $\hat{\boldsymbol{\delta}}$ is small. When that happens, we say the algorithm has converged.

Convergence of the Algorithm:

Let $\hat{\boldsymbol{\theta}}^k$ denote the estimate from the k^{th} iteration. Convergence can be measured in several ways:

1. Look at change in the parameter estimates:

   If $\hat{\boldsymbol{\theta}}^k \approx \hat{\boldsymbol{\theta}}^{k-1}$ we may consider the sequence $\hat{\boldsymbol{\theta}}^0, \hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \ldots$ to have converged and take $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^k$ as the MLE/NLS estimator.

   The most common way to establish that $\hat{\boldsymbol{\theta}}^k \approx \hat{\boldsymbol{\theta}}^{k-1}$ is to look at

   $$\max_j \frac{|\hat{\theta}_j^k - \hat{\theta}_j^{k-1}|}{|\hat{\theta}_j^{k-1}|}.$$

   If this criterion is less than some tolerance value (1×10^{-6}, say) the algorithm stops.

2. Look at change in the objective function $S(\boldsymbol{\theta})$:

   If $S(\hat{\boldsymbol{\theta}}^k) \approx S(\hat{\boldsymbol{\theta}}^{k-1})$ then take $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^k$. To judge this we can use the criterion

   $$\frac{S(\hat{\boldsymbol{\theta}}^{k-1}) - S(\hat{\boldsymbol{\theta}}^k)}{S(\hat{\boldsymbol{\theta}}^k)}.$$

   Again, if this criterion is smaller than some tolerance value, then stop.

Q: Which approach, 1 or 2, is better?

A: It depends. If the main interest is in the mean response, $\boldsymbol{\eta}(\boldsymbol{\theta})$, then (2) may be more appropriate. If the regression parameters themselves are of more interest than perhaps (1) is better.

Both criteria have the disadvantage that they don't indicate when $\hat{\boldsymbol{\theta}}^k$ is close to $\hat{\boldsymbol{\theta}}$ or when $S(\hat{\boldsymbol{\theta}}^k)$ is close to $S(\hat{\boldsymbol{\theta}})$.

E.g., if we knew the minimum value of the least squares criterion $S(\hat{\boldsymbol{\theta}})$ was equal to 10, say, then we could stop the algorithm based on judging when $S(\hat{\boldsymbol{\theta}}^k) \approx 10$.

Of course, we would never know $S(\hat{\boldsymbol{\theta}})$ or $\hat{\boldsymbol{\theta}}$ to judge convergeence to the optimal value in this manner. However, we do know that $\hat{\boldsymbol{\theta}}$ must satisfy the normal equation

$$[\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}})]^T \mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$$

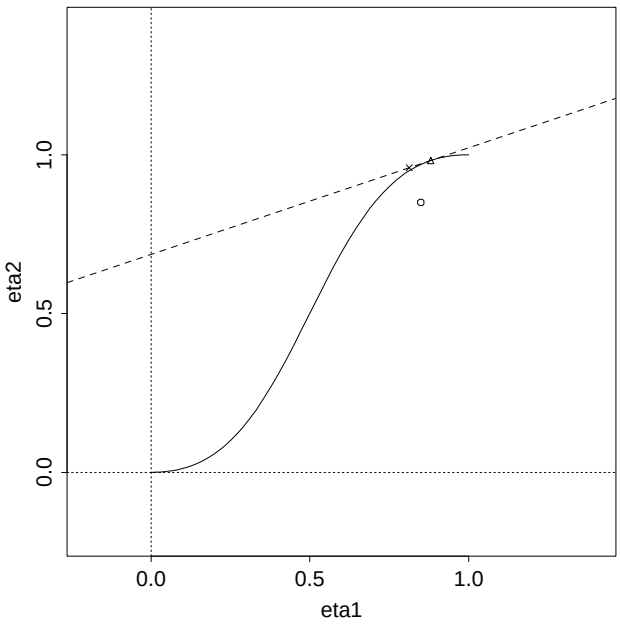
therefore we can stop iterating when

$$\underbrace{[\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k)]^T}_{\text{residual vector for } \hat{\boldsymbol{\theta}}^k} \mathbf{V}(\hat{\boldsymbol{\theta}}^k) \approx \mathbf{0}.$$

This suggests a third convergence criterion:

3. Relative offset criterion. Based on $\hat{\boldsymbol{\theta}}^k$, the residual vector $[\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k)]$, can be written as the sum of two components:
   - a vector in the tangent plane
   - a vector $\perp$ to the tangent plane

These two components can be seen in the following plot



This is to say that $[\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k)]$ can be written as

$$[\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k)] = \boldsymbol{\psi}_1 + \boldsymbol{\psi}_2$$

where ψ_1 is in the tangent plane, and ψ_2 is $\perp$ to the tangent plane.

Then one measure of "how orthogonal" $\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k)$ is to the tangent plane is

$$\frac{\|\boldsymbol{\psi}_1\|}{\|\boldsymbol{\psi}_2\|} \qquad \text{the relative offset criterion}$$

- When this criterion is close to zero, the $\boldsymbol{\psi}_1$ component of $\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k)$ is negligible, implying $\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k)$ is approximately $\perp$ to the tangent plane; i.e., $[\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k)]^T \mathbf{V}(\hat{\boldsymbol{\theta}}^k) \approx \mathbf{0}$.

- It can be shown that the components $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ have lengths

$$\|\boldsymbol{\psi}_1\| = \|\mathbf{Q}_1^T[\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k)]\|, \qquad \|\boldsymbol{\psi}_2\| = \|\mathbf{Q}_2^T[\mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^k)]\|$$

where $\mathbf{Q} = [\mathbf{Q}_1, \mathbf{Q}_2]$ is the Q matrix from a QR-decomposition of $\mathbf{V}(\hat{\boldsymbol{\theta}}^k)$.

Inference Using the Linear Approximation:

- The basic approach to inference in nonlinear models is to use the linear approximation to reduce the situation to the linear case and then use standard linear models results. Because of the approximation, these results will be approximate rather than exact.

Recall that the (nonlinear) least squares estimator $\hat{\boldsymbol{\theta}}$ minimizes

$$S(\boldsymbol{\theta}) = \|\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2$$

Let $\boldsymbol{\theta}^*$ denote the true value of $\boldsymbol{\theta}$. Using the linear Taylor series approximation, we can approximate $\boldsymbol{\eta}(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ close to $\boldsymbol{\theta}^*$ as

$$\boldsymbol{\eta}(\boldsymbol{\theta}) \approx \boldsymbol{\eta}(\boldsymbol{\theta}^*) + \mathbf{V}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$$

It follows that

$$\begin{aligned}\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}) &\approx \underbrace{\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}^*)}_{\equiv \mathbf{z}^*} - \mathbf{V}(\boldsymbol{\theta}^*) \underbrace{(\boldsymbol{\theta} - \boldsymbol{\theta}^*)}_{\equiv \boldsymbol{\delta}^*} \\ &= \mathbf{z}^* - \mathbf{V}(\boldsymbol{\theta}^*)\boldsymbol{\delta}^*\end{aligned}$$

so that

$$\begin{aligned}S(\boldsymbol{\theta}) &= \|\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2 \\ &\approx \|\mathbf{z}^* - \mathbf{V}(\boldsymbol{\theta}^*)\boldsymbol{\delta}^*\|^2 \equiv S^*(\boldsymbol{\theta}).\end{aligned}$$

We know from linear models that the value of $\overbrace{\boldsymbol{\theta} - \boldsymbol{\theta}^*}^{=\boldsymbol{\delta}^*}$ that minimizes $S^*(\boldsymbol{\theta})$ is

$$\{(\mathbf{V}^*)^T\mathbf{V}^*\}^{-1}(\mathbf{V}^*)^T\mathbf{z}^*,$$

where $\mathbf{V}^* \equiv \mathbf{V}(\boldsymbol{\theta}^*)$.

It follows that the value of $\boldsymbol{\theta}$ that minimizes $S^*(\boldsymbol{\theta})$ is

$$\{(\mathbf{V}^*)^T\mathbf{V}^*\}^{-1}(\mathbf{V}^*)^T\mathbf{z}^* + \boldsymbol{\theta}^*.$$

It's also true that the value of $\boldsymbol{\theta}$ that minimizes $S(\boldsymbol{\theta})$ is $\hat{\boldsymbol{\theta}}$.

Therefore, since $S(\boldsymbol{\theta}) \approx S^*(\boldsymbol{\theta})$ in a neighborhood of $\boldsymbol{\theta}^*$ (which we should expect to contain $\hat{\boldsymbol{\theta}}$, since $\hat{\boldsymbol{\theta}}$ is a "good estimator") these two minimizers should be approximately equal:

$$\hat{\boldsymbol{\theta}} \approx \{(\mathbf{V}^*)^T\mathbf{V}^*\}^{-1}(\mathbf{V}^*)^T\mathbf{z}^* + \boldsymbol{\theta}^*$$

or, equivalently,

$$\begin{aligned}\hat{\boldsymbol{\theta}} &\approx \boldsymbol{\theta}^* + \{(\mathbf{V}^*)^T\mathbf{V}^*\}^{-1}(\mathbf{V}^*)^T \underbrace{[\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}^*)]}_{=\mathbf{e}} \\ &= \boldsymbol{\theta}^* + \{(\mathbf{V}^*)^T\mathbf{V}^*\}^{-1}(\mathbf{V}^*)^T\mathbf{e}\end{aligned}$$

This result,

$$\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}^* + \{(\mathbf{V}^*)^T\mathbf{V}^*\}^{-1}(\mathbf{V}^*)^T\mathbf{e} \qquad (\dagger)$$

is an approximate version of the exact result

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e} \qquad \text{(see p.21)}$$

that we had for the linear model, where the derivative matrix has changed from $\mathbf{X}$ to $\mathbf{V}(\boldsymbol{\theta}^*)$.

From (†), analogues of all of the linear model inference results can be obtained for the nonlinear model, although they are now only approximate.

E.g.,

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\theta}}) &\approx \text{var}(\boldsymbol{\theta}^* + \{(\mathbf{V}^*)^T\mathbf{V}^*\}^{-1}(\mathbf{V}^*)^T\mathbf{e}) \\ &= \{(\mathbf{V}^*)^T\mathbf{V}^*\}^{-1}(\mathbf{V}^*)^T \underbrace{\text{var}(\mathbf{e})}_{=\sigma^2\mathbf{I}_n} \mathbf{V}^*\{(\mathbf{V}^*)^T\mathbf{V}^*\}^{-1} \\ &= \sigma^2\{(\mathbf{V}^*)^T\mathbf{V}^*\}^{-1} \equiv \text{avar}(\hat{\boldsymbol{\theta}})\end{aligned}$$

avar() for "asymptotic" variance-covariance matrix (approximate, based on a large sample approximation).

We estimate this asymptotic/approximate var-cov matrix as

$$\hat{\text{avar}}(\hat{\boldsymbol{\theta}}) = s^2(\hat{\mathbf{V}}^T\hat{\mathbf{V}})^{-1}, \quad \text{where } \hat{\mathbf{V}} \equiv \mathbf{V}(\hat{\boldsymbol{\theta}}),$$

and

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n}\{y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})\}^2$$

An asymptotic standard error for $\hat{\theta}_j$ is given by the square root of the j^{th} diagonal element of $\hat{\text{avar}}(\hat{\boldsymbol{\theta}})$:

$$\text{a.s.e.}(\hat{\theta}_j) = s\sqrt{(\hat{\mathbf{V}}^T\hat{\mathbf{V}})^{-1}_{jj}}$$

All other inference results 1–9 (except 4 & 5) on pp. 21–24 hold in an analogous fashion by replacing $\mathbf{X}^T\mathbf{X}$ by $\hat{\mathbf{V}}^T\hat{\mathbf{V}}$.

In particular, the following linear approximation methods of inference hold for the nonlinear model with NLS parameter estimator $\hat{\boldsymbol{\theta}}$:

1. Although $\hat{\boldsymbol{\theta}}$ is no longer *exactly* unbiased, it is *asymptotically* unbiased (i.e., **consistent**). This is to say that $\hat{\boldsymbol{\theta}}$ gets closer and closer to $\boldsymbol{\theta}^*$ (in probability) as the sample size increases.

2. $\hat{\boldsymbol{\theta}}$ has asymptotic/approximate var-cov matrix that can be estimated as
$$\hat{\text{avar}}(\hat{\boldsymbol{\theta}}) = s^2(\hat{\mathbf{V}}^T\hat{\mathbf{V}})^{-1}, \quad \text{where } \hat{\mathbf{V}} \equiv \mathbf{V}(\hat{\boldsymbol{\theta}}).$$

3. (Asymptotic normality) $\hat{\boldsymbol{\theta}} \dot{\sim} N_p(\boldsymbol{\theta}, \sigma^2\{(\mathbf{V}^*)^T\mathbf{V}^*\}^{-1})$ (if $\mathbf{e}$ is assumed normal). Here, "$\dot{\sim}$" mean approximately distributed as.

6. An asymptotic/approximate standard error of $\hat{\theta}_j$, the jth estimated regression parameter, is
$$\text{a.s.e.}(\hat{\theta}_j) = s\sqrt{(\hat{\mathbf{V}}^T\hat{\mathbf{V}})^{-1}_{jj}};$$
an approximate $100(1-\alpha)\%$ confidence interval for θ_j is given by
$$\hat{\theta}_j \pm t_{1-\alpha/2}(n-p)\text{a.s.e.}(\hat{\theta}_j);$$
and a test of $H_0 : \theta_j = \theta_0$ that has approximate level α has the rejection rule: reject H_0 if
$$\frac{|\hat{\theta}_j - \theta_0|}{\text{a.s.e.}(\hat{\theta}_j)} > t_{1-\alpha/2}(n-p)$$

7. A joint confidence region for $\boldsymbol{\theta}$ with approximate coverage probability $100(1-\alpha)\%$ is given by the set of all $\boldsymbol{\theta}$ such that
$$\frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\hat{\mathbf{V}}^T\hat{\mathbf{V}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{ps^2} \leq F_{1-\alpha}(p, n-p).$$
For an approximate α-level test, we reject $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ in favor of $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ if
$$F_1 \equiv \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T(\hat{\mathbf{V}}^T\hat{\mathbf{V}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)}{ps^2} > F_{1-\alpha}(p, n-p). \qquad (*)$$

Recall that in the linear model, the test statistic F_1 given in (*) was algebraically equivalent to the test statistic

$$F_2 = \frac{(\mathrm{SS_{E0}} - \mathrm{SS_E})/(\mathrm{dfE_0} - \mathrm{dfE})}{\mathrm{SS_E}/\mathrm{dfE}}$$

where $\mathrm{SS_{E0}}$ and $\mathrm{SS_E}$ are the sums of squares for error associated with the null model (model in which H_0 holds) and the alternative models, respectively, and $\mathrm{dfE_0}$ and dfE are the corresponding degrees of freedom for these two models.

In the nonlinear model, it is important to note that the statistics F_1 and F_2 are no longer equal and yield different* tests of H_0!

Which is better?

- Since F_2 depends upon $\hat{\boldsymbol{\theta}}$ only through $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}})$, the estimated mean of $\mathbf{y}$, and not directly through $\hat{\boldsymbol{\theta}}$, this means that models with two different parameterizations will have the same value of F_2.
  - Hence, F_2 will be affected only by the intrinsic nonlinearity of the model and not by its parameter-effects nonlinearity.
  - In contrast, F_1 will be affected by both intrinsic and parameter-effects nonlinearity.
- The result is that the asymptotic (approximate) distribution of F_2 is often much more accurate than that of F_1 in finite samples, and therefore the test based on F_2 is superior.
- Both F_1 and F_2 can be generalized in an obvious way to test the more general hypothesis $H_0 : \mathbf{A}\boldsymbol{\theta} = \mathbf{b}$ for a matrix and vector of constants $\mathbf{A}$ and $\mathbf{b}$, respectively.

* (though asymptotically equivalent)

8. An approximate $100(1-\alpha)\%$ CI for the mean response at a given vector of explanatory variables $\mathbf{x}_0$ is given by

$$f(\mathbf{x}_0;\hat{\boldsymbol{\theta}}) \pm t_{1-\alpha/2}(n-p)\sqrt{s^2\hat{\mathbf{f}}_0^T(\hat{\mathbf{V}}^T\hat{\mathbf{V}})^{-1}\hat{\mathbf{f}}_0}$$

where

$$\hat{\mathbf{f}}_0^T = \left(\frac{\partial f(\mathbf{x}_0;\boldsymbol{\theta})}{\partial\theta_1}, \frac{\partial f(\mathbf{x}_0;\boldsymbol{\theta})}{\partial\theta_2}, \ldots, \frac{\partial f(\mathbf{x}_0;\boldsymbol{\theta})}{\partial\theta_p}\right)\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

($\mathbf{f}_0$ is the **gradient** of the expectation function at $\mathbf{x}_0$, $\mathbf{f}(\mathbf{x}_0;\boldsymbol{\theta})$, with respect to $\boldsymbol{\theta}$, evaluated at our estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$.)

An approximate $100(1-\alpha)\%$ CI for a linear combination of the $\boldsymbol{\theta}$'s of the form $\mathbf{c}^T\boldsymbol{\theta}$ for some vector of constants $\mathbf{c}$ is given by

$$\mathbf{c}^T\hat{\boldsymbol{\theta}} \pm t_{1-\alpha/2}(n-p)\sqrt{s^2\mathbf{c}^T(\hat{\mathbf{V}}^T\hat{\mathbf{V}})^{-1}\mathbf{c}}$$

9. An approximate $100(1-\alpha)\%$ prediction interval for the response y_0 at a given vector of explanatory variables $\mathbf{x}_0$ is given by

$$f(\mathbf{x}_0;\hat{\boldsymbol{\theta}}) \pm t_{1-\alpha/2}(n-p)s\sqrt{1+\hat{\mathbf{f}}_0^T(\hat{\mathbf{V}}^T\hat{\mathbf{V}})^{-1}\hat{\mathbf{f}}_0}$$

10. An approximate $100(1-\alpha)\%$ **confidence band** for the response function at that holds for all possible $\mathbf{x}_0$ is given by

$$f(\mathbf{x}_0;\hat{\boldsymbol{\theta}}) \pm \sqrt{F_{1-\alpha}(p,n-p)}\sqrt{ps^2\hat{\mathbf{f}}_0^T(\hat{\mathbf{V}}^T\hat{\mathbf{V}})^{-1}\hat{\mathbf{f}}_0}.$$

- The "approximateness" of most of the above approximate inference results in the nonlinear model depends upon the accuracy of the linear approximation. This depends upon both the planar assumption (intrinsic nonlinearity) and the uniform coordinates assumption (parameter-effects nonlinearity).
    - In particular, results 6, 7 (the use of the F statistic F_1, not F_2), 8, 9, 10 are collectively known as "linear approximation" inference methods and are affected by both the intrinsic and parameter-effects nonlinearity of the model.

A better approach to inference is based upon the **profile likelihood function** or, equivalently, the **profile t function**. The statistic F_2 is based on this approach. We will describe this approach in detail later, but for now, we can use linear approximation results for inference.

Practical Considerations in Nonlinear Regression

1. Model Specification:

There are two components of any model relating y to $\mathbf{x}$:

1. Deterministic component — the expectation function.
2. Stochastic component — the error (or disturbance) term.

Specification of the expectation function:

Ideally, the expectation function is implied by some contextual theory. I.e., the expectation function is based on a mechanistic model.

Examples:

- Wind Speed: Recall our wind speed model. Based on a theory of the relationship between wind speed and height under adiabatic conditions, we have

$$\text{windspeed} = \theta_1 \log\{\text{height}(1 - \theta_2/\theta_3) - 1/\theta_3\}, \qquad (*)$$

  where

$$\theta_1 = \text{friction velocity}$$
$$\theta_2 = \text{zero point displacement}$$
$$\theta_3 = \text{roughness length}$$

- Pressure and Temperature in Steam: The relationship between pressure and temperature in saturated steam can be written as

$$\text{pressure} = \theta_1(10)^{\theta_2 \text{temperature}/(\theta_3 + \text{temperature})}$$

  where $\theta_1, \theta_2, \theta_3$ are constants.

- Chemical Reaction: According to chemical theory, the reaction rate R for a certain chemical reaction is expected to be related to x_1, x_2, the partial pressures of reactant and product, respectively, according to

$$R = \frac{\theta_1 \theta_2 x_1}{1 + \theta_1 x_1 + \theta_2 x_2}$$

where θ_1 and θ_2 are absorption equilibrium constants for reactant and product, respectively, and θ_3 is the effective reaction rate constant.

- A Compartment Model: According to the two compartment open model introduced on pp.55–57, the concentration of tetracycline in blood serum over time (t) satisfies

$$\text{concentration}_t = \frac{\theta_1 \theta_3 (e^{-\theta_1 t} - e^{-\theta_2 t})}{\theta_2 - \theta_1}$$

- A Growth Model: A simple assumption leading to growth subject to an upper limit α on size is that growth is proportional to the size remaining; i.e., if x denotes time and f denotes size, this assumption can be expressed as the differential equation

$$\frac{\partial f}{\partial x} = \kappa(\alpha - f)$$

This diff eq has general solution

$$f(x) = \alpha - \beta e^{-\kappa x}$$

which can be reparameterized in a variety of ways including

$$f(x) = \theta_1(1 - e^{-\theta_2(x - \theta_3)}) \qquad \text{(monomolecular growth model)}$$

and

$$f(x) = \phi_1 + \phi_2 \phi_3^x \qquad \text{(asymptotic regression model)}$$

The above relationship is also Newton's law of cooling for a body cooling over time (in that context $f(x)$ represents temperature as a function of time, x).

Often, however, no mechanistic model is available, and an empirical approach must be taken. That is, we need to choose an appropriate expectation form that fits the data well. Preferrably, such a model will have interpretable parameters and low parameter-effects nonlinearity.

Approaches to choosing an empirical model:

- Search the literature. Is there a model (possibly a mechanistic one) that has been used in a similar context previously. If so, someone else has already done the work of selecting an empirical model and their choice *may be* appropriate for the data currently under consideration.

- Classify the problem. Are we modeling growth? Are we modeling yield? Are we modeling concentration over time? Are we modeling some kind of rate of change? For some general classes of problems, some "standard models" for these kinds of problems exist. (See Ratkowsky, *Handbook of Nonlinear Models*, for a collection of commonly used nonlinear models.)

- Examine the relationship between the response and the primary covariate(s).

  - Is the curve sigmoidal? Then consider a logistic model, or one of the other common sigmoidal forms: Gompertz, Von Bertalanffy, Richards, Weibull, Fletcher, or Morgan-Mercer-Flodin. We'll come back and talk about some of these in some detail when we discuss growth curves. These models are treated in detail in Seber & Wild, Ch. 7.

  - Does the curve rise monotonically to an asymptote and have a concave form? Then a Michaelis-Menten, asymptotic regression, asymptotic yield-density curve, or a simplified growth-curve model might be appropriate.

  - Is the curve parabolic? Then a Michaelis-Menten curve with a quadratic term in the denominator may be useful, or a parabolic yield density curve. Several yield-density models, both asymptotic and parabolic, are presented in Seber & Wild, §7.6.

– Does the curve show exponential decay? Then perhaps a single exponential model ($\mathrm{E}(y) = \theta_1 e^{-\theta_2 x}$) or biexponential form ($\mathrm{E}(y) = \theta_1 e^{-\theta_2 x} + \theta_3 e^{-\theta_4 x}$) would be useful.

Specification of the stochastic component:

The question of how the error term should enter into the model can be answered based on theory, but usually it is addressed empirically. The choice is often between an additive homoscedastic vs. a multiplicative homoscedastic error term, although later we will discuss a wider class of models that can accommodate heteroscedastic additive errors.

Let $\mathrm{E}(R) = h(\mathbf{x}, \boldsymbol{\theta})$ be the deterministic component relating some response R to explanatory variables $\mathbf{x}$. The additive situation is one in which

$$R = h(\mathbf{x}, \boldsymbol{\theta}) + u$$

where $\mathrm{E}(u) = 0$ and $\mathrm{var}(u) = \sigma^2$. In this case, we formulate our model by taking $y = R$, $f(\mathbf{x}, \boldsymbol{\theta}) = h(\mathbf{x}, \boldsymbol{\theta})$ and $e = u$ giving our standard form

$$y = f(\mathbf{x}, \boldsymbol{\theta}) + e, \quad \text{where } \mathrm{E}(e) = 0, \mathrm{var}(e) = \sigma^2.$$

The multiplicative situation is one in which

$$\begin{aligned} R &= h(\mathbf{x}, \boldsymbol{\theta})(1 + \tilde{u}) \\ &= h(\mathbf{x}, \boldsymbol{\theta}) z \end{aligned}$$

where $\tilde{u}$ has mean 0 so that $z = (1 + \tilde{u})$ has mean $\mathrm{E}(z) = 1$. In this case, we formulate our model by taking $y = \log(R)$, $f(\mathbf{x}, \boldsymbol{\theta}) = \log\{h(\mathbf{x}, \boldsymbol{\theta})\}$ and $e = \log(z)$ giving our standard additive-error form

$$y = f(\mathbf{x}, \boldsymbol{\theta}) + e \qquad (*)$$

- Note that the assumption $\mathrm{E}(e) = 0$ here is appropriate because a linear Taylor series approximation of $\log(z)$ about its mean 1 gives

$$\mathrm{E}(e) = \mathrm{E}\{\log(z)\} \approx \mathrm{E}\{\log(1) + 1(z - 1)\} = \mathrm{E}\{z - 1\} = 0.$$

- However, the appropriateness of the assumption $\text{var}(e) = \sigma^2$ depends upon the nature of the variance of R. Using the same Taylor linearization, we have $\log(z) \approx z - 1 = R/h(\mathbf{x}, \boldsymbol{\theta}) - 1$. Therefore,

$$\text{var}(e) = \text{var}\{\log(z)\} \approx \frac{\text{var}(R)}{\{h(\mathbf{x}, \boldsymbol{\theta})\}^2}$$

- Therefore, a homoscedastic error assumption on e in model (*) is appropriate in the multiplicative error case if R has standard deviation proportional to its mean. This is a special, but common, form of nonconstant variance.

- If the multiplicative-error model for R is not log-transformable to a homoscedastic additive-error model, or if we would prefer to work with an additive-error model for R rather than taking a transformation, then it will be necessary to drop the assumption of homoscedasticity. As in linear regression, it is possible to use weighted/generalized nonlinear least squares to fit a heteroscedastic nonlinear model.

- To judge the nature of the variability in R, a plot of R versus each of the covariates in x can be helpful. In addition, residual plots such as residuals versus fitteds can be used.

- In the case that there are multiple observations of R at each level of x we can calculate the sample standard deviation of R at each level and examine how these SDs change over increasing x.

Example — Age of Rabbits Measured "By Eye":

The European rabbit *Oryctolagus cuniculus* is a major pest in Australia. A reliable method of age determination for rabbits caught in the wild would be of importance in ecological studies. In a study by Dudzinski and Mykytowycz (1961), the dry weight of the eye lens was measured for 71 free-living wild rabbits of known age. Eye lens weight tends to vary much less with environmental conditions than does total body weight, and therefore may be a much better indicator of age

The rabbits were born and lived free in an experimental 1.7 acre enclosure at Gungahlin, ACT. The birth data and history of each individual were accurately known. Rabbits in the enclosure depended on the natural food supply. In this experiment, 18 of the eye lenses were collected from rabbits that died in the course of the study from various causes such as coccidiosis, bird predation or starvation. The remaining 53 rabbits were deliberately killed, immediately after being caught in the enclosure or after they had been kept for some time in cages. The lenses were preserved and their dry weight determined.

Here we take

$$R = \text{eye lens weight in mg.}$$

and $$x = \text{age in days}$$

Dudzinski and Mykytowycz suggest the deterministic relationship

$$\mathrm{E}(R) = \theta_1 \exp\left\{\frac{-\theta_2}{\theta_3 + x}\right\}.$$

- The data from this example are contained in file rabbiteye.dat available on the course web site.
- See handout Rabbit1. In this handout, we plot the data, fit some models, and check some residual plots to determine the appropriate scale for the error term.
- The first thing we do is plot both Lens (lens weight) and log(Lens) against Age. From these plots it is clear that a multiplicative error term is more appropriate than an additive one.
- To follow up on this conclusion, we fit the models

$$\text{Lens}_i = \theta_1 \exp\left\{\frac{-\theta_2}{\theta_3 + x_i}\right\} + e_i, \quad \text{var}(e_i) = \sigma^2 \qquad \text{(m1rabbit.nls)}$$

and

$$\log(\text{Lens}_i) = \theta_1 - \frac{\theta_2}{\theta_3 + x_i} + e_i, \quad \text{var}(e_i) = \sigma^2 \qquad \text{(m2rabbit.nls)}$$

- The fitted curves from these models are given as dotted lines on the bottom two plots on the first page of plots. Both fits look pretty good here, but the corresponding residual vs. fitted plots (top of last page of Rabbit1) show the failure of the homoscedasticity assumption in m1rabbit.nls.

- The function gnls() in S-PLUS will allow the user to fit a nonlinear model using weighted/generalized least squares. We use this function to fit m1rabbit.gnls. This model is as follows:

$$\text{Lens}_i = \theta_1 \exp\left\{\frac{-\theta_2}{\theta_3 + x_i}\right\} + e_i, \quad \text{var}(e_i) = \sigma^2 \text{Age}_i. \text{ (m1rabbit.gnls)}$$

- The fitted curve from m1rabbit.gnls is overlayed in the bottom-left plot on the first page of plots in Rabbit1. It is almost indistinquishable from the OLS fit, and the OLS and WLS parameter estimates are very similar. Note however, that the standard errors between m1rabbit.nls and m1rabbit.gnls have changed substantially (the gnls ones are more appropriate).

- In general, gnls will fit models for which

$$\text{var}(e_i) = \sigma^2 g^2(\mu_i, \mathbf{v}_i, \boldsymbol{\delta}),$$

  where $\mu_i = \text{E}(y_i)$, $\mathbf{v}_i$ is a vector of *variance covariates*, $\boldsymbol{\delta}$ is a vector of *variance parameters* to be estimated, and $g(\cdot)$ is a known *variance function.*

- In our example, there was just one variance covariate $v_i = \text{Age}_i$, and there were no variance paramters, so $g^2(\mu_i, v_i, \boldsymbol{\delta}) = g^2(\mu_i, \text{Age}_i) = \text{Age}_i$ and $\text{var}(y_i) = \sigma^2 \text{Age}_i$.

- Alternatively, we could have estimated the power of the variance covariate Age by taking $g^2(\cdot)$ as $g^2(\mu_i, \text{Age}_i, \delta) = \text{Age}_i^{2\delta}$, and treating δ as an unknown parameter to be estimated. This could be done in gnls() with the option, weights=varPower(form = ~Age).

  - Note that as soon as we include a parameter $\boldsymbol{\delta}$ to be estimated in the variance function $g^2(\cdot)$, this takes us out of the WLS context. Remember that in WLS, the variance is assumed proportional to a known value. Assuming $\text{var}(e_i) = \sigma^2\text{Age}_i^{2\delta}$ only corresponds to WLS if δ is known. Otherwise, we need ML estimation.

  - Furthermore, in the last two examples, $g^2(\mu_i, v_i, \boldsymbol{\delta})$ did not depend upon μ_i. If it does, then the model for the mean μ_i appears both in the expectation function and the variance function of the model. This takes us further afield, requiring *pseudo-likelihood estimation*. For example, the option, weights = varPower(form = ~fitted(.)), specifies that the error variance is proportional to some to-be-estimated power of the mean: $\text{var}(e_i) = \sigma^2|\mu_i|^{2\delta}$.

- In fact, gnls() can fit not only weighted least squares (heteroscedastic) models, but also generalized least squares (correlated errors) models. We will return to this capability later.

Typically, either an additive or multiplicate error structure is appropriate for most nonlinear models. However, these are not the only possibilities. gnls() provides several different choices for $g^2(\cdot)$ so that the variance assumption on an additive error term can be chosen to be appropriate for whatever scale a homoscedastic error enters the model.

- One example of a model in which a random component enters the model in neither a multiplicative nor an additive way, is an *errors-in-variables regression model*. In such a model we assume that one or more of the explanatory variables is observed subject to measurement (or some other kind of) error.

- E.g., in the rabbit eye example, the true age of the rabbits might have been estimated subject to error. An appropriate model for this situation is

$$\log(\text{Lens}_i) = \theta_1 - \frac{\theta_2}{\theta_3 + \underbrace{(x_i + e_i)}_{\text{“true x”}}} + e_i, \quad \text{var}(e_i) = \sigma_e^2, \text{var}(e_i) = \sigma_e^2$$

- Specialized techniques for errors-in-variables model are neceessary in both the linear and nonlinear regression context. See Seber & Wild, Ch.10, and the book-length treatments by Fuller and by Carroll et al.

2. Derivatives:

The G-N algorithm utilizes the derivative matrix $\mathbf{V}(\boldsymbol{\theta}) = \frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$.

Q: How do the functions nls() and gnls() know what these derivatives are?

A: By default, these routines use **numerical derivatives** rather than *analytic derivatives*. However, numerical derivatives can be inaccurate and slow (especially when second derivatives are taken, as in, e.g., the Newton-Raphson algorithm).

Forward Difference Approximations:

The forward difference method of computing numerical derivatives cost less computer time than central differences, but are less accurate. For many applications the accuracy of forward differences is sufficient, though.

Recall from calculus the definition of a derivative at some point x of a simple function $g(x)$ of a scalar argument x:

$$\lim_{h \to 0} \frac{g(x+h) - g(x)}{h}.$$

More generally, the partial derivative of a function $g(x, y)$ of two scalar variables x and y with respect to x is

$$\frac{\partial g}{\partial x} = \lim_{h \to 0} \frac{g(x+h, y) - g(x, y)}{h}.$$

- The forward difference approximation just replaces the limit as $h \to 0$ with a very small value of h.

Suppose we have a function $g(\boldsymbol{\theta})$ of a p-dimensional parameter $\boldsymbol{\theta}$ and (possibly) other variables or parameters (e.g., for g the expectation function of a nonlinear model, it also depends on $\mathbf{x}_i$, the vector of explanatory variables for the ith subject).

Let $\mathbf{j}_i$ be a $p \times 1$ vector with a 1 in the ith position, and 0s elsewhere. Then the forward difference approximation is

$$\frac{\partial g}{\partial \theta_i} \approx \frac{g(\boldsymbol{\theta} + h_i \mathbf{j}_i) - g(\boldsymbol{\theta})}{h_i}$$

where h_i is a very small number (as small as possible subject to the limits of computational precision of the computer).

- Note that $\boldsymbol{\theta} + h_i \mathbf{j}_i$ just adds h_i to the ith element of $\boldsymbol{\theta}$.

- The value of h_i can be taken to be the same for all elements of $\boldsymbol{\theta}$, but because some elements of $\boldsymbol{\theta}$ may have very different scales than others, h_i is usually scaled up or down depending on the magnitude of θ_i.

- Specifically, I recommend $h_i = \sqrt{\epsilon}(1 + |\theta_i|)$ where ϵ is the **machine precision** of the computer on which the calculations are being done.

- Machine precision is a constant specific to any given computer hardware that quantifies the limit of that machine's ability to distinguish small (in magnitude) double-precision floating point numbers.

- For the PC in my office, $\epsilon = 2.2204e-016$. Most statistical software can look up ϵ. E.g., in SAS, use the function: *constant('maceps')*; in S-PLUS and R ϵ is held in the constant *.Machine$double.eps*; in Matlab ϵ is held in the constant *eps*.