Linear & Generalized Linear Models/ Advanced Regression for Independent Data

Andrew J. Spieker, Ph.D.

Associate Professor of Biostatistics
Vanderbilt University

Set 14: Nonlinear least squares

Version: 07/03/2026

Table of Contents

Linear model

What is a linear model?

- In this course, we have considered models that—even if not linear on the original scale of measurement—embedded linearity on some scale.
  - Simple linear regression (a single line).
  - Linear regression with adjustment (multiple parallel lines).
  - Linear regression with interactions (lines needn't be parallel).
  - Linear regression with log-transformed outcomes (still linear).
  - Logistic regression (linear in log-odds).
  - Proportional odds (linear in log-odds).
  - Proportional hazards (linear in log-hazard).
  - Polynomial models (linear in $\boldsymbol{\beta}$, as with all of the above).
- When we speak of linearity assumptions, we're usually talking about linearity in the *predictor*, X (at least on some scale).
- As you may know, linear models are so named due to linearity in $\boldsymbol{\beta}$!
- This supplemental set of notes handles estimation for models that are not linear in its parameters.

Table of Contents

Nonlinear model

Nonlinear models:

- A nonlinear model takes the following form:

$$\begin{aligned} Y &= f(\mathbf{x};\boldsymbol{\beta}) + \epsilon, \quad \mathrm{E}[\epsilon|\mathbf{x}] = 0; \quad \text{or} \\ \mathrm{E}[Y|\mathbf{x}] &= f(\mathbf{x};\boldsymbol{\beta}). \end{aligned}$$

- Of course, $f(\mathbf{x};\boldsymbol{\beta}) = \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_{K-1}$ is just a linear regression model.
- Short of just specifying $f(\cdot)$ to be linear, how do we estimate $\boldsymbol{\beta}$ under this new, more general kind of model?

Table of Contents

Nonlinear models:

- Loss function:

$$L(\boldsymbol{\beta}; \mathbf{X}) = \sum_{i=1}^{N} (Y_i - \mathrm{E}[Y_i|\mathbf{x}_i])^2 = \sum_{i=1}^{N} (Y_i - f(\mathbf{x}_i; \boldsymbol{\beta}))^2.$$

- For notational convenience, let $f_i(\boldsymbol{\beta}) = f(\mathbf{x}_i; \boldsymbol{\beta})$. Then, the loss function is written as:

$$L(\boldsymbol{\beta}; \mathbf{X}) = \sum_{i=1}^{N} (Y_i - f_i(\boldsymbol{\beta}))^2.$$

- We seek to *minimize* this quantity (least squares) with respect to $\boldsymbol{\beta}$.

Objective function:

- With some matrix calculus, minimizing the sum of squares gives us the following estimating equations:

$$\begin{pmatrix} \sum_{i=1}^{N} \frac{\partial f_i}{\partial \beta_0}(Y_i - f_i(\boldsymbol{\beta})) \\ \vdots \\ \sum_{i=1}^{N} \frac{\partial f_i}{\partial \beta_{K-1}}(Y_i - f_i(\boldsymbol{\beta})) \end{pmatrix} = \mathbf{D}^{\mathsf{T}}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{0},$$

  where $\mathbf{D}^{\mathsf{T}} \equiv \mathbf{D}^{\mathsf{T}}(\boldsymbol{\beta})$ is the Jacobian (matrix of partial derivatives, of dimension $K \times N$) for $\boldsymbol{\mu}(\boldsymbol{\beta}) : \mathbb{R}^K \to \mathbb{R}^N$. Here, $\boldsymbol{\mu} = \text{vec}(f_i(\boldsymbol{\beta}))$.

- When $\boldsymbol{\mu}$ is linear, then $\mathbf{D}^{\mathsf{T}} = \mathbf{X}^{\mathsf{T}}$, and this problem reduces to the ordinary least squares linear regression problem we know and love:

$$\mathbf{X}^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \Rightarrow \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}.$$

Notation/key quantities:

- We can define key quantities analogously to those for GLMs:

$$\begin{aligned}
\mathbb{G}_N(\boldsymbol{\beta}) &= \mathbf{D}^\mathsf{T}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{0}. \\
\mathbb{A}_N^{\text{obs}}(\boldsymbol{\beta}) &= -\frac{\partial}{\partial \boldsymbol{\beta}}\mathbb{G}_N(\boldsymbol{\beta}) = \mathbf{D}^\mathsf{T}\mathbf{D} - \sum_{i=1}^{N}(Y_i - f_i(\boldsymbol{\beta})) \otimes \mathbf{H}_i. \\
\mathbb{A}_N(\boldsymbol{\beta}) &= \mathsf{E}\left[-\frac{\partial}{\partial \boldsymbol{\beta}}\mathbb{G}_N(\boldsymbol{\beta})\right] = \mathbf{D}^\mathsf{T}\mathbf{D}. \\
\mathbb{B}_N^{\text{obs}}(\boldsymbol{\beta}) &= \mathbf{D}^\mathsf{T}\text{diag}(Y_i - f_i(\boldsymbol{\beta}))^2\mathbf{D}.
\end{aligned}$$

- Note: "$\otimes$" denotes the Kronecker product and $\mathbf{H}_i$ is the Hessian (symmetric matrix of second partial derivatives for f_i) for the i^{th} observation. Without further information on $f(\cdot)$, we can't reduce these expressions down further.

Nonlinear least squares

Estimation:

- Estimating equations: $\mathbb{G}_N(\boldsymbol{\beta}) = \mathbf{D}^\top(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{0}$.
- A very large segment of the case dealt with GLMs, whereby the associated estimating equations (which could emerge from maximum likelihood, quasilikelihood) took a similar form to those above. However, we're not able to reduce the expression $\mathbf{D}^\top$ down more generally without more knowledge of $f(\cdot)$.
- There is typically no general closed-form expression for $\widehat{\boldsymbol{\beta}}$. We need numerical techniques for nonlinear least squares (e.g., Gauss-Newton):

$$\begin{aligned}\boldsymbol{\beta}^{(j+1)} &= \boldsymbol{\beta}^{(j)} + \left[\mathbb{A}_N(\boldsymbol{\beta}^{(j)})\right]^{-1}\mathbb{G}_N(\boldsymbol{\beta}^{(j)}) \\ &= \boldsymbol{\beta}^{(j)} + (\mathbf{D}^\top(\boldsymbol{\beta}^{(j)})\mathbf{D}(\boldsymbol{\beta}^{(j)}))^{-1}\mathbf{D}^\top(\boldsymbol{\beta}^{(j)})(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(j)})).\end{aligned}$$

- Initialization is more challenging...

Variance:

- Variance estimator based on correct mean model and homoscedasticity:
  - $\widehat{\text{Cov}}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2 \left[\mathbb{A}_N(\widehat{\boldsymbol{\beta}})\right]^{-1}$, where $\widehat{\sigma}^2 = (N-K)^{-1}\sum_{i=1}^{N}(Y_i - f_i(\widehat{\boldsymbol{\beta}}))^2$.
- Variance estimator based on correct mean model, with protection against heteroscedasticity:
  - $\widehat{\text{Cov}}(\widehat{\boldsymbol{\beta}}) = \left[\mathbb{A}_N(\widehat{\boldsymbol{\beta}})\right]^{-1} \mathbb{B}_N^{\text{obs}}(\widehat{\boldsymbol{\beta}}) \left[\mathbb{A}_N(\widehat{\boldsymbol{\beta}})\right]^{-1}$.
- Variance estimator with protection against heteroscedasticity and mean-model misspecification:
  - $\widehat{\text{Cov}}(\widehat{\boldsymbol{\beta}}) = \left[\mathbb{A}_N^{\text{obs}}(\widehat{\boldsymbol{\beta}})\right]^{-1} \mathbb{B}_N^{\text{obs}}(\widehat{\boldsymbol{\beta}}) \left[\mathbb{A}_N^{\text{obs}}(\widehat{\boldsymbol{\beta}})\right]^{-1}$.
- The theory necessary to verify this is analogous to what we covered for GLMs.

Table of Contents

Example: Michaelis-Menten

Model:

- Consider the following model:

$$Y = \frac{\beta_0 X}{\beta_1 + X} + \epsilon,$$

  where $\mathrm{E}[\epsilon | X = x] = 0$ for all x.
- Another way to express this:

$$\mathrm{E}[Y | X = x] = \frac{\beta_0 x}{\beta_1 + x}.$$

- This is an example of a nonlinear model (in particular, nonlinear in $\boldsymbol{\beta}$).
- In particular, this represents the Michaelis-Menten reaction kinetics model, where Y represents a formulation rate and X represents the concentration of a substrate.
- Coefficient interpretation:
  - β_0: The limiting (expected) reaction velocity.
  - β_1: Substrate concentration at which the expected reaction velocity is half its maximum.

Example: Michaelis-Menten

Estimating equations:

- Here, $f_i(\boldsymbol{\beta}) = \beta_0 x_i/(\beta_1 + x_i)$. Note the following:

$$\frac{\partial f_i(\boldsymbol{\beta})}{\partial \beta_0} = \frac{X_i}{\beta_1 + X_i}, \text{ and}$$

$$\frac{\partial f_i(\boldsymbol{\beta})}{\partial \beta_1} = -\frac{\beta_0 X_i}{(\beta_1 + X_i)^2}.$$

- This is to say:

$$\mathbf{D}^\intercal = \begin{bmatrix} X_1(\beta_1 + X_1)^{-1} & \cdots & X_N(\beta_1 + X_N)^{-1} \\ -\beta_0 X_1(\beta_1 + X_1)^{-2} & \cdots & -\beta_0 X_N(\beta_1 + X_N)^{-2} \end{bmatrix}.$$

- Estimating equations:

$$\mathbb{G}_N(\boldsymbol{\beta}) = \begin{bmatrix} \sum_{i=1}^N \frac{X_i}{\beta_1 + X_i}(Y_i - f_i(\boldsymbol{\beta})) \\ \sum_{i=1}^N -\frac{\beta_0 X_i}{(\beta_1 + X_i)^2}(Y_i - f_i(\boldsymbol{\beta})) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Example: Michaelis-Menten

Gauss-Newton algorithm: Estimation of $\boldsymbol{\beta}$

- Step 1: Initialize $\boldsymbol{\beta}^{(0)}$.
  - Not always as straightforward as with GLMs.
- Step 2: Set tolerance threshold, iterating the following until apparent convergence:

$$\begin{aligned}\boldsymbol{\beta}^{(j+1)} &= \boldsymbol{\beta}^{(j)} + \left[\mathbb{A}_N(\boldsymbol{\beta}^{(j)})\right]^{-1}\mathbb{G}_N(\boldsymbol{\beta}^{(j)}) \\ &= \boldsymbol{\beta}^{(j)} + \left[\mathbf{D}^{\mathsf{T}}(\boldsymbol{\beta}^{(j)})\mathbf{D}(\boldsymbol{\beta}^{(j)})\right]^{-1}\mathbf{D}^{\mathsf{T}}(\boldsymbol{\beta}^{(j)})(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta}^{(j)})).\end{aligned}$$

- Step 3: $\widehat{\boldsymbol{\beta}}$ is the final iteration.

Example: Michaelis-Menten

Simulation:

- Let's conduct a simulation to evaluate the behavior of nonlinear least squares (model-based and sandwich-based estimator protecting only against heteroscedasticity).
- Data generation mechanism:
  - Sample size: $N = 100$.
  - $X_i \sim \text{Uniform}(1, 10)$
  - $Y_i = 4X_i/(5 + X_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(\mu = 0, \sigma^2 = 0.15^2)$.
- Number of simulations: $M = 5000$.

Example: Michaelis-Menten

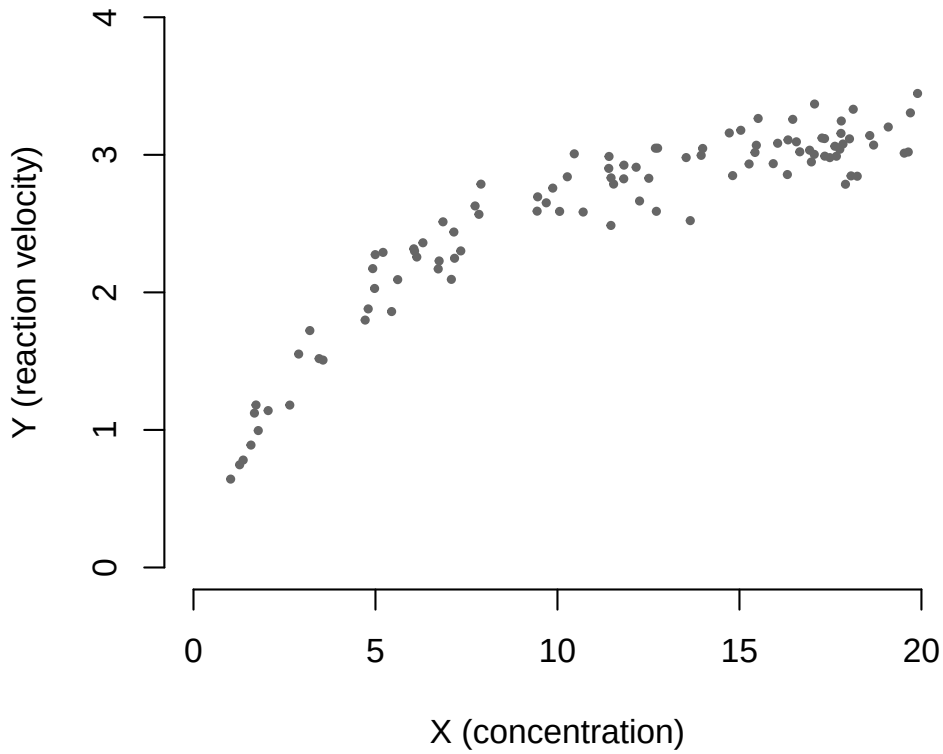

Figure 14.1: Example data from the Michaelis-Menten setup.

Example: Michaelis-Menten

Simulation:

- To initialize, note that:

$$\frac{1}{Y} = \alpha_0 + \alpha_1 \frac{1}{X},$$

  with $\beta_0 = 1/\alpha_0$ and $\beta_1 = \alpha_1/\alpha_0$.
- This suggests letting $V = 1/Y$ and $U = 1/X$, fitting the following model by OLS:

$$\mathrm{E}[V|U = u] = \alpha_0 + \alpha_1 u,$$

  and initializing as $\boldsymbol{\beta}^{(0)} = (1/\widehat{\alpha}_0, \widehat{\alpha}_1/\widehat{\alpha}_0)^\intercal$.
- This is known as the Lineweaver-Burk method. Note that $\boldsymbol{\beta}^{(0)}$ is not a consistent estimator of $\boldsymbol{\beta}$ (mind Jensen's inequality).

Example: Michaelis-Menten

Simulation setup:

```r
## Set seed for reproducibility
set.seed(7345)

## Set sample size
n <- 500

## Set parameters
b0 <- 4
b1 <- 5

## Number of simulation replicates
nsim <- 5000

## Store results
ests <- matrix(0, nrow = nsim, ncol = 2)
ses1 <- matrix(0, nrow = nsim, ncol = 2)
ses2 <- matrix(0, nrow = nsim, ncol = 2)

## Concentrations (assumed fixed, but not necessary)
X <- runif(n, 1, 20)
```

Example: Michaelis-Menten

Conduct simulation:

```
## Conduct simulation
for (j in 1:nsim)
{
  ## Generate outcome
  Y <- (b0 * X)/(b1 + X) + rnorm(n, 0, 0.15)

  ## Lineweaver-Burk linearization/initialization
  LB <- as.numeric(coef(lm(1/Y ~ 1 + I(1/X))))
  beta <- c(1/LB[1], LB[2]/LB[1])
  tol <- 1
  iter <- 1

  ## Iterate Gauss-Newton algorithm
  while (tol > 1e-12)
  {
    beta.p <- beta
    D <- cbind(X/(beta[2] + X), -beta[1]*X/(beta[2] + X)^2)
    G <- t(D) %*% c(Y - (beta[1] * X)/(beta[2] + X))
    beta <- beta.p + solve(t(D) %*% D) %*% G
    tol <- sum((beta.p - beta)^2)
    iter <- iter + 1
  }
  ## Variance estimation
  sigma.hat <- ((n - 2)^(-1))*sum(c(Y - (beta[1] * X)/(beta[2] + X))^2)
  DTD.inv <- solve(t(D) %*% D)
  vhat2 <- DTD.inv %*% t(D * c(Y - (beta[1] * X)/(beta[2] + X))^2) %*% D %*% DTD.inv
  ests[j,] <- beta
  ses1[j,] <- sqrt(diag(sigma.hat * DTD.inv))
  ses2[j,] <- sqrt(diag(vhat2))
}
```

Example: Michaelis-Menten

Simulation results:

```
## Average estimates
> colMeans(ests)
[1] 4.001016 5.003613

## Empirical standard errors
> apply(ests, 2, sd)
[1] 0.03148471 0.12238171

## Average model-based standard errors
> colMeans(ses1)
[1] 0.03136194 0.12131667

## Average sandwich-based standard errors
> colMeans(ses2)
[1] 0.03126413 0.12093693
```

Example: Michaelis-Menten

Additional considerations:

- The mean model was correctly specified, and these data were not generated with heteroscedasticity, so we expected both variance estimators to perform well.
- As you can imagine, computing $\mathbb{A}_N^{\text{obs}}(\boldsymbol{\beta})$ takes a bit more work, but it is manageable—and it should also perform well in this case.
- The `nls()` function in R will perform nonlinear least squares.

```
## Nonlinear least squares in R
zz <- nls(Y ~ b0*X/(b1 + X), start = list(b0 = 1/LB[1], b1 = LB[2]/LB[1]))

## Model-based variance (will match our answer very closely)
sqrt(diag(vcov(zz)))

## Sandwich-based variance (will match our answer very closely)
sqrt(diag(sandwich(zz)))
```

- There are other ways to initialize the Michaelis-Menten.

Table of Contents

Additional thoughts

Optimality:

- A statement of the Gauss-Markov variety can be made about the linear unbiased estimating equations that emerge from nonlinear least squares.
  - At this point in the semester, I want you to be able to formulate the statement without as much scaffolding from me. We can try it together!
- When the errors are assumed normally distributed and homoscedastic, it turns out (and I hope this is believable) that the nonlinear least squares estimating equations are nothing more or less than those that would emerge from the likelihood that presumes normal errors. In this very special case, the nonlinear least squares estimating equations—because they coincide with score equations—produce asymptotically efficient estimators (overall, and not just in the Gauss-Markov sense).

Additional thoughts

Precision and other matters:

- How might you imagine forming a confidence interval for $\mathrm{E}[Y|\mathbf{x}]$ in a nonlinear model?
  - Delta method, perhaps?
- We described a validity-stability trade-off in the setting of GLMs. A similar trade-off will exist for selection of a variance estimator for nonlinear least squares.
- Weights can easily be incorporated into nonlinear least squares.
- Recognize GLMs as a specific case of (possibly weighted) nonlinear least squares, in which the form for $\mathbf{D}^{\mathsf{T}}$ has a particular expansion.

Summary

This unit:

- Nonlinear least squares.

Summary: So far

- Random vectors and matrices; multivariate normal theory.
- Ordinary least squares.
- Hypothesis testing and ANOVA.
- Weighted least squares.
- Confidence regions and prediction.
- Diagnostics.
- Exponential families.
- Generalized linear models.
- Sandwich and bootstrap.
- Quasi-likelihood.
- Hypothesis testing for GLMs.
- Diagnostics for GLMs.
- Further considerations for binary outcomes.
- Nonlinear least squares.

Summary: Coming up

The end! Next steps:

- BIOS 7346/BIOS 8346 (longitudinal data analysis).
- Developing more skills focused on the implementation and practice of biostatistical methods.
- Research.
- Don't stop learning! :)