Applied Statistical Methods II

Introduction to Nonlinear Regression

Tuesday, January 19, 2021

What we will cover:

- Nonlinear regression (reference: Chapter 13 in KNNL)
    - Mean model formulation
    - Assumptions about error (i.e. the mean-variance relationship)
    - How to fit them (i.e. Gauss-Newton)
    - Inference
- General linear models (reference: McCullagh and Nelder)
    - Logistic regression, Poisson regression, other common examples
    - General formulation in terms of cumulant generating function
    - Mean-variance relationship
    - Quasi-likelihood

What we will cover (cont.):

- Mixed models, ANOVA, ANCOVA (reference: McCullagh and Nelder)
    - Modeling dependencies between observations
    - Estimation and inference
    - Lots of data examples
- Advanced topics, if time permits (lecture notes & academic papers)
    - High dimensional factor analysis
    - Dimension reduction
    - Missing data
    - Challenges in modern scientific data

What do each of these topics do?

- Nonlinear Regression
    - The regression function is not linear in the parameters.
    - Still assume Gaussian errors.
    - Parameters often have a nice physical meaning that drives the shape of the regression function.
    - The models we will consider here are parametric models, i.e. the regression function is known up to a parameter γ with fixed dimension.
- GLMs
    - Your data might not be normal.
    - Binomial (developed cancer or did not) or count data (number of murders in a city).
        - First: logistic regression for binomial data.
        - Then: log-linear/ Poisson regression for count data.
        - Finally: tie them together with linear Gaussian regression in the framework of generalized linear models. (We may do this before Poisson regression)
        - We will rely on the moment/cumulant generating functions here.

- Analysis of designed studies
    - ANOVA
    - ANCOVA
    - Balanced and unbalanced designs
- Random and Mixed Effects Models
    - mixed model for ANOVA
    - mixed model for repeated measures

Linear Regression

Last semester mostly looked at models of the form:

$$Y_i \quad = \quad \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)} + \epsilon_i$$

- ϵ_i iid are normal.
- $\mathbf{X}_i = (1, X_{i1}, \ldots, X_{i(p-1)})^T$
- $\beta = (\beta_0, \ldots, \beta_{p-1})^T$
- Recall that we can have that several Xs correspond to one categorical variable or $X_{i2} = X_{i1}^2$ (or other polynomial).
- Model is linear in the parameters β.
- $Y_i = \mathbf{X}_i^T \beta + \epsilon_i$
    - All linear functions of vectors can be written as matrix operations.

Non-linear function

In many of the physical sciences and population studies:

- The science tells you that the data should take a certain non-linear shape.
    - Exponential decay in physics.
    - Logistic population growth model in biology.
- It can be parameterized as a non-linear function of unknown parameters.
- These parameters are the interpretable coefficients we want to estimate.

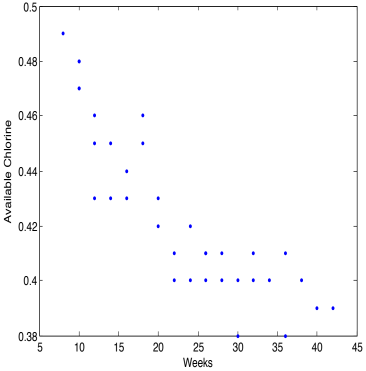
Real Example

- Proctor & Gamble are manufacturing a certain product.
- The amount of available chlorine in the product decreases over time.
    - It is known that available chlorine is expected to be **0.49 at 8 weeks**.
    - Its dynamics after 8 weeks can be described by a starting fraction at 8 weeks, a plateaued fraction, and the rate in between.

The Data

- Several products are held in the factory for some time and their available chlorine is measured.
- The data consist of
  - X_i - amount of time from manufacturing
  - Y_i - available chlorine.
- Have n=44 observations.
  - X_i is between 8 and 42 weeks.
  - Y_i is between 0.49 and 0.39.

The Data



Possible model class:
$Y_i = \gamma_0 + \gamma_1 \exp(\gamma_2 X_i) + \epsilon_i$. (13.8 in KNNL)

Proctor and Gamble Example

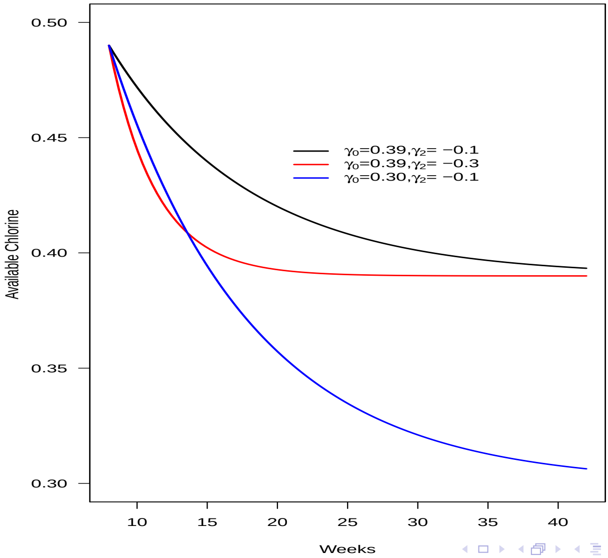
Use a general exponential function:

- $Y_i = \gamma_0 + \gamma_1 \exp(\gamma_2 X_i) + \epsilon_i$. (13.8 in KNNL)
- $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$.
  - Why might this be a problem? Do we need to worry about this?
- $E(Y_i) = \gamma_0 + \gamma_1 \exp(\gamma_2 X_i)$
- At $X_i = 0$, $E(Y_i) = \gamma_0 + \gamma_1$.
  - $\gamma_0 + \gamma_1$ can be seen as the expected value when $X_i = 0$.
- γ_2 is usually restricted to be negative.
  - $\exp(\gamma_2 X_i)$ gets small as $X_i \to \infty$.
  - $\gamma_0 = \lim_{X_i \to \infty} E(Y_i \mid X_i)$, i.e. the asymptote of the mean function.
  - $\gamma_1 = E(Y_i \mid X_i = 0) - \lim_{X_i \to \infty} E(Y_i \mid X_i)$.
  - γ_2 is the rate of decay in $E(Y_i)$.

In our example

- We know the expected value at $X_i = 8$ weeks.
  - Reduces the model to only 2 parameters.
- $Y_i = \gamma_0 + (0.49 - \gamma_0)\exp[\gamma_2(X_i - 8)] + \epsilon_i$
- We assume that $0 < \gamma_0 < 0.49$ and $\gamma_2 < 0$.
- $E(Y_i)$ at $X_i = 8$ will be 0.49.
- $E(Y_i) \rightarrow \gamma_0$ as $X_i \rightarrow \infty$.
- γ_2 describes how quickly Y_i approaches its minimum γ_0.
- The parameters are chosen to be interpretable.
  - The function is non-linear in the parameters
  - You no longer have the interpretation that “one unit increase in X_{ij} is associated with an expected increase in Y_i by β_j units.”

Three Examples



Idea of Non-Linear Regression

$$Y_i = f(\mathbf{X}_i, \gamma) + \epsilon_i$$

- γ is a vector of unknown parameters.
- The function f is assumed to be known, unlike non-parametric regression.
  - In non-linear regression, convention is to use γ instead of β.
- f is some function of $\mathbf{X}_i$ and γ.
- ϵ_i are the error terms.
- All of the stochastic information comes from ϵ_i.
- When you say “non-linear regression," it is usually assumed that $\epsilon_i \overset{i.i.d}{\sim} N\left(0, \sigma^2\right)$.
  - **Critical assumption:** variance σ^2 is NOT a function of the mean $f(\mathbf{X}_i, \gamma)$
  - Will talk about this more when compared to GLM.

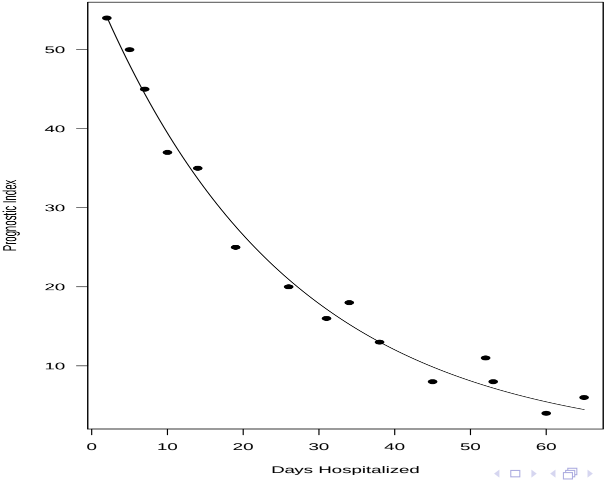
Popular Non-Linear Functions

Exponential Regression Model:

- $Y_i = \gamma_0 \exp(\gamma_1 X_i) + \epsilon_i$
- ϵ_i are iid $N(0, \sigma^2)$
- $E(Y_i) = \gamma_0 \exp(\gamma_1 X_i)$
- As $X_i \to 0$, $E(Y_i) \to \gamma_0$.
  - γ_0 can be seen as the expected value when $X_i = 0$.
- γ_1 is usually restricted to be negative when X_i must be positive.
  - $\exp(\gamma_1 X_i)$ gets small as $X_i \to \infty$.
  - $E(Y_i) \to 0$ as $X_i \to \infty$
  - γ_1 is the rate of decay.
  - Used a lot for radioactive or chemical decay.

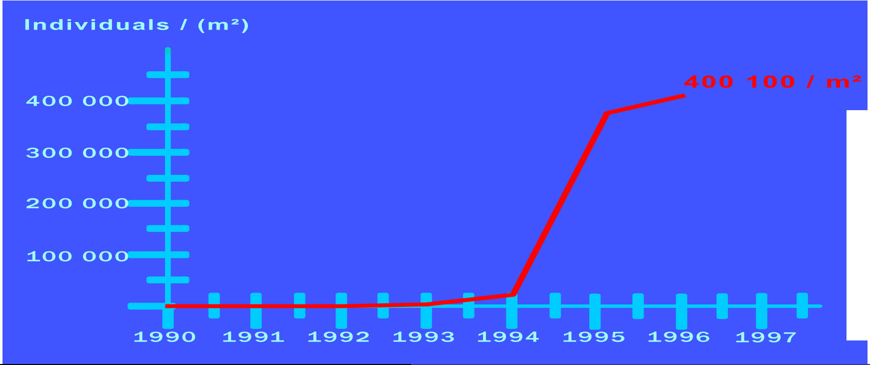
Exponential Regression Model (Continued):

- γ_1 is usually restricted to be negative when X_i must be positive.
  - e.g. prognostic index vs. days hospitalized in severly injured patients (fig 13.2)



Exponential Regression Model (Continued):

- Less common is to have $\gamma_1 > 0$.
  - Would imply exponential growth.
  - Population explosion of invasive species
  - e.g. zebra mussels in Ontario's Rideau River and Canal.
  - Data: $\sim$ 2000 mussels were first found in 1990
    24 mussels per m^2 in 1993
    23,000 per m^2 in 1994
    383,100 per m^2 in 1995



Logistic Regression Models (might not be a good name)

- $Y_i = \frac{\gamma_0}{1+\gamma_1 \exp(\gamma_2 X_i)} + \epsilon_i$.
- Usually, $\gamma_1 \geq 0$ and $\gamma_2 \leq 0$ while $X_i \geq 0$.
  - γ_0 is the maximum $E(Y_i)$.
  - $E(Y_i) \to \gamma_0$ as $X_i \to \infty$.
  - $X_i \to 0$, then $E(Y_i) \to \frac{\gamma_0}{1+\gamma_1}$.
  - γ_2 is the rate between $\frac{\gamma_0}{1+\gamma_1}$ when $X_i = 0$ and γ_0 as $X_i \to \infty$.
- Popular in modeling animal populations:
  - If P is expected population and X is time, $dP/dt = (-\gamma_2) P \left(1 - \frac{P}{\gamma_0}\right)$.
  - $-\gamma_2$ is the growth rate.
  - nice conditions allow a population to thrive (P small, γ_0 large).
  - slows when they start to compete for resources.
  - too much growth eventually inhibits the rate of growth (P approaches γ_0).

Do not confuse non-linear regression with GLM

- This logistic regression model is not what is usually thought of when a statistician says logistic regression.
- What we discussed is better referred to as non-linear regression with a logistic regression function.
- Logistic regression is part of what is known as generalized linear models (GLM)
- Logistic regression is used when the responses are binary
  - ie. person died or they didn't
  - ie. person got cancer or they didn't

Parameterizations

- In linear regression:
  - If we have $X_{i1}, \ldots, X_{i(p-1)}$...
  - we have $\beta_0, \ldots, \beta_{p-1}$
- In the exponential regression example:
  - We have X_{i1} only ...
  - we have γ_0, γ_1.
- In the logistic example:
  - We have X_{i1} only ...
  - we have $\gamma_0, \gamma_1, \gamma_2$.
- In general non-linear regression, you can have more/less parameters than covariates.
  - q - number of covariates $X_{i1}, \ldots, X_{iq}$
  - p - number of parameters $\gamma_0, \ldots, \gamma_{p-1}$

Fitting The Model

- Two equivalent approaches:
  1. least squares
  2. maximum likelihood.
- These are equivalent because we assume ϵ_i are normal.
- These two approaches minimize/maximize the functions:
  1. $Q = \sum_{i=1}^{n} \left[Y_i - f(\mathbf{X}_i, \gamma)\right]^2$
  2. $L(\gamma, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[Y_i - f(\mathbf{X}_i, \gamma)\right]^2\right]$
- To see the equivalence:
  - Maximizing L is equivalent to minimizing $-2 \log L$.
  - This is what is done in numerical packages.

To Minimize Q

- Take the derivative with respect to each parameter and set equal to zero.
- By the chain rule, for $k = 0, \ldots, p-1$:

$$\frac{\partial Q}{\partial \gamma_k} = \sum_{i=1}^{n} -2\left[Y_i - f(\mathbf{X}_i, \gamma)\right]\left[\frac{\partial f(\mathbf{X}_i, \gamma)}{\partial \gamma_k}\right].$$

- when you know f, you can compute $\left[\frac{\partial f(\mathbf{X}_i,\gamma)}{\partial \gamma_k}\right]$.
- This will give you p normal equations which you must solve.
- Problem: these are non-linear functions in $\gamma_0, \ldots, \gamma_{p-1}$ and there is (almost always) no nice closed form.

In Our Data Example

$$
\begin{aligned}
f(X,\gamma_0,\gamma_2) &= \gamma_0 + (0.49 - \gamma_0)\exp\left[\gamma_2(X-8)\right] \\
\frac{\partial f}{\partial \gamma_0} &= 1 - \exp\left[\gamma_2(X-8)\right] \\
\frac{\partial f}{\partial \gamma_2} &= (.49 - \gamma_0)(X-8)\exp\left[\gamma_2(X-8)\right]
\end{aligned}
$$

The normal equations become:

$$\sum_{i=1}^{n} \{Y_i - \gamma_0 - (.49 - \gamma_0)\exp[\gamma_2(X_i - 8)]\}\{1 - \exp[\gamma_2(X_i - 8)]\} = 0$$

$$\begin{array}{cl} \sum_{i=1}^{n} & \{Y_i - \gamma_0 - (.49 - \gamma_0)\exp[\gamma_2(X_i - 8)]\} \\ \times & (.49 - \gamma_0)(X_i - 8)\exp[\gamma_2(X_i - 8)] = 0 \end{array}$$

There is no nice close-form solution.

Solving Non-Linear Least Squares

$\hat{\gamma} = \text{argmin}_{\gamma} Q(\gamma) = \sum_{i=1}^{n} \left[Y_i - f(\mathbf{X}_i, \gamma)\right]^2$

- We have to use a numerical method to get $\hat{\gamma}$.
- There are several methods: Newton-Raphson's method, Gradient Descent, etc.
- For non-linear least squares problem, a common method is Gauss-Newton's Method (a simpler version of Newton's method).
    - Start with some initial value for γ.
    - Locally approximate the non-linear function f with a linear function. Equivalent descriptions:
        - Approximate Q with quadratic
        - Approximate the Hessian matrix in Newton's method with a function of Jacobian matrix.
    - based on this approximation to update the parameters.
    - Repeat until convergence.

First Order Taylor's Theorem

- Assume that the function $f(\mathbf{X}, \gamma)$ is well-behaved around the true γ.
  - All second order partial derivatives exist and are continuous at true value γ.
- For some initial value $\gamma^{(0)}$ close to γ:
  - $f(\mathbf{X}_i, \gamma) \approx f(\mathbf{X}_i, \gamma^{(0)}) + \sum_{k=0}^{p-1} \left[\frac{\partial f(\mathbf{X}_i, \gamma)}{\partial \gamma_k}\right]_{\gamma=\gamma^{(0)}} \left(\gamma_k - \gamma_k^{(0)}\right) = f(\mathbf{X}_i, \gamma^{(0)}) + \underbrace{\mathbf{J}_i^T}_{\mathbf{J}_i = \mathbf{J}_i\left(\gamma^{(0)}\right)} \left(\gamma - \gamma^{(0)}\right)$
- Conditional on $\gamma^{(0)}$, right side is a linear function in γ.
- $$Q(\gamma) = \sum_{i=1}^{n} [Y_i - f(\mathbf{X}_i, \gamma)]^2$$
  $$\approx \sum_{i=1}^{n} \left[\left(Y_i - f(\mathbf{X}_i, \gamma^{(0)})\right) - \mathbf{J}_i^T \left(\gamma - \gamma^{(0)}\right)\right]^2$$

First Order Taylor's Theorem

- Assume that the function $f(\mathbf{X}, \gamma)$ is well-behaved around the true γ.
  - All second order partial derivatives exist and are continuous at true value γ.
- For some initial value $\gamma^{(0)}$ close to γ:
  - $f(\mathbf{X}_i, \gamma) \approx f(\mathbf{X}_i, \gamma^{(0)}) + \sum_{k=0}^{p-1} \left[\frac{\partial f(\mathbf{X}_i, \gamma)}{\partial \gamma_k}\right]_{\gamma=\gamma^{(0)}} \left(\gamma_k - \gamma_k^{(0)}\right) = f(\mathbf{X}_i, \gamma^{(0)}) + \underbrace{\mathbf{J}_i^T}_{\mathbf{J}_i=\mathbf{J}_i\left(\gamma^{(0)}\right)} \left(\gamma - \gamma^{(0)}\right)$
- Conditional on $\gamma^{(0)}$, right side is a linear function in γ.
- $$\begin{aligned} Q(\gamma) &= \sum_{i=1}^{n} [Y_i - f(\mathbf{X}_i, \gamma)]^2 \\ &\approx \sum_{i=1}^{n} \left[\left(Y_i - f(\mathbf{X}_i, \gamma^{(0)})\right) - \mathbf{J}_i^T \left(\gamma - \gamma^{(0)}\right)\right]^2 \end{aligned}$$

First Order Taylor's Theorem (cont.)

$$Q(\gamma) \approx \sum_{i=1}^{n} \left[\left(Y_i - f(\mathbf{X}_i, \gamma^{(0)}) \right) - \mathbf{J}_i^T \left(\gamma - \gamma^{(0)} \right) \right]^2$$

- Let $r_i = Y_i - f(\mathbf{X}_i, \gamma^{(0)})$ be the current residuals, $\mathbf{J} = \begin{pmatrix} \mathbf{J}_1^T \\ \vdots \\ \mathbf{J}_n^T \end{pmatrix}$
- Optimize Q with OLS: $\gamma^{(1)} = \gamma^{(0)} + \left(\mathbf{J}^T \mathbf{J} \right)^{-1} \mathbf{J}^T \boldsymbol{r}$.
- Intuition: $\boldsymbol{J} \in \mathbb{R}^q$ acts like the design matrix!

Some properties of Gauss-Newton

$$Q(\gamma) = \sum_{i=1}^{n} [Y_i - f(\boldsymbol{X}_i, \gamma)]^2.$$

- The step $\boldsymbol{s} = \gamma^{(1)} - \gamma^{(0)} = \left(\mathbf{J}^T\mathbf{J}\right)^{-1}\mathbf{J}^T\boldsymbol{r}$ is a **descent direction**, i.e. it tends to decrease the objective.
  - $$\begin{aligned}\boldsymbol{s}^T\nabla Q_\gamma\left(\gamma^{(0)}\right) &= -2\boldsymbol{s}^T\sum_{i=1}^{n}\left[Y_i - f\left(\boldsymbol{X}_i, \gamma^{(0)}\right)\right]\boldsymbol{J}_i = -2\boldsymbol{s}^T\boldsymbol{J}^T\boldsymbol{r}\\ &= -2\boldsymbol{r}^T\boldsymbol{J}\left(\mathbf{J}^T\mathbf{J}\right)^{-1}\boldsymbol{J}^T\boldsymbol{r} \le 0\end{aligned}$$
  - $= 0$ if and only if $\boldsymbol{r} \in \ker\left(\boldsymbol{J}^T\right)$, which is true if and only if $\nabla Q_\gamma\left(\gamma^{(0)}\right) = 0$.
- By Taylor's theorem, taking a step in $\alpha\boldsymbol{s}$ for $\alpha > 0$ small enough is guaranteed to decrease the objective.
- More sophisticated algorithms can be designed to properly choose α (trust region, Wolfe conditions, etc.). This is beyond the scope of this course.

Some properties of Gauss-Newton

$$Q(\gamma) = \sum_{i=1}^{n} [Y_i - f(\boldsymbol{X}_i, \gamma)]^2.$$

- The step $\boldsymbol{s} = \gamma^{(1)} - \gamma^{(0)} = \left(\mathbf{J}^T\mathbf{J}\right)^{-1} \mathbf{J}^T \boldsymbol{r}$ is a **descent direction**, i.e. it tends to decrease the objective.
  -
    $$\begin{aligned} \boldsymbol{s}^T \nabla Q_\gamma\left(\gamma^{(0)}\right) &= -2\boldsymbol{s}^T \sum_{i=1}^{n} \left[Y_i - f\left(\boldsymbol{X}_i, \gamma^{(0)}\right)\right] \boldsymbol{J}_i = -2\boldsymbol{s}^T \boldsymbol{J}^T \boldsymbol{r} \\ &= -2\boldsymbol{r}^T \boldsymbol{J} \left(\mathbf{J}^T\mathbf{J}\right)^{-1} \boldsymbol{J}^T \boldsymbol{r} \leq 0 \end{aligned}$$
  - $= 0$ if and only if $\boldsymbol{r} \in \ker\left(\boldsymbol{J}^T\right)$, which is true if and only if $\nabla Q_\gamma\left(\gamma^{(0)}\right) = 0$.
- By Taylor's theorem, taking a step in $\alpha \boldsymbol{s}$ for $\alpha > 0$ small enough is guaranteed to decrease the objective.
- More sophisticated algorithms can be designed to properly choose α (trust region, Wolfe conditions, etc.). This is beyond the scope of this course.

Accuracy of approximation

Gauss-Newton relies on the approximation $f\left(\boldsymbol{X}_i,\gamma\right) \approx f\left(\boldsymbol{X}_i,\gamma^{(0)}\right) + \boldsymbol{J}_i^T\left(\gamma-\gamma^{(0)}\right)$. How accurate is it?

- Let $\boldsymbol{M}\left(\gamma\right) = \nabla_\gamma^2 f\left(\boldsymbol{X}_i,\gamma\right)$. By Taylor's Theorem:

$$\begin{aligned} f\left(\boldsymbol{X}_i,\gamma\right) =& f\left(\boldsymbol{X}_i,\gamma^{(0)}\right) + \boldsymbol{J}_i^T\left(\gamma-\gamma^{(0)}\right) \\ &+ \frac{1}{2}\left(\gamma-\gamma^{(0)}\right)^T \boldsymbol{M}\left(\tilde{\gamma}\right)\left(\gamma-\gamma^{(0)}\right), \quad \tilde{\gamma}\in\ell\left(\gamma^{(0)},\gamma\right) \end{aligned}$$

- Error of approximation: $\leq \frac{1}{2}M_{\max}\|\gamma-\gamma^{(0)}\|^2$
  - $M_{\max} = \sup\left\{\lambda_{\max}\left(\boldsymbol{M}\left(\tilde{\gamma}\right)\right) : \tilde{\gamma}\in\ell\left(\gamma^{(0)},\gamma\right)\right\}$
- Gauss-Newton's method is very sensitive to selection of initial values.
- Can use some type of search for good initial values.
- We will work mostly with selecting reasonable values from the data.

Recall Non-Linear Regression

- $Y_i = f(\mathbf{X}_i, \gamma) + \epsilon_i$
  - $\gamma = (\gamma_0, \ldots, \gamma_{p-1})$ is a vector of parameters.
  - f is some function of $\mathbf{X}_i$ and γ.
  - ϵ_i are the error terms.
- Estimate parameters γ through least squares. Minimize:
  - $Q(\gamma) = \sum_{i=1}^{n} \left[Y_i - f(\mathbf{X}_i, \gamma)\right]^2$
- Problem: there is no closed form solution for γ that minimizes $Q(\gamma)$.
  - It exists.
  - Can not write it out algebraically.
- Solution: We use Gauss-Newton.
  - Updates took the form $\gamma^{(1)} = \gamma^{(0)} + \alpha \left(\boldsymbol{J}^T \boldsymbol{J}\right)^{-1} \boldsymbol{J}^T \boldsymbol{r}$. In simple GN, $\alpha = 1$. Can also choose α at each iteration for better convergence properties.

Inference

- Nonlinearity inhibits an exact distribution for our estimates.
  - To be expected seeing as how we don't even have a closed form solution.
- Asymptotic distributions are known.
- Assumes n is large and ϵ_i are i.i.d. By CLT, normality is not necessary, but we will assume it for convenience.

What is the limiting dist'n of $\hat{\gamma}$?

We want something of the form $n^{1/2}(\hat{\gamma} - \gamma) \to N(0, \boldsymbol{A})$.

- No closed form for $\hat{\gamma}$, so we have to rely on Taylor's Theorem.
- If $\hat{\gamma} \approx \gamma$, expand $\nabla Q(\hat{\gamma})$ around the true γ. Let $\hat{\boldsymbol{J}} = \boldsymbol{J}(\hat{\gamma})$, $\boldsymbol{J} = \boldsymbol{J}(\gamma)$, $\hat{\boldsymbol{r}} = \boldsymbol{r}(\hat{\gamma})$, $\boldsymbol{r} = \boldsymbol{r}(\gamma)$.
- Ideas behind the derivation of the asymptotic distribution:
  - Use a Taylor expansion
  - $\hat{\boldsymbol{J}} \approx \boldsymbol{J}$ if $\hat{\gamma} \approx \gamma$ (i.e. as a function, $\boldsymbol{J}$ is continuous).
  - $\boldsymbol{J}$ is full rank and $\lim_{n\to\infty} \Lambda_{\min}(n^{-1}\boldsymbol{J}^T\boldsymbol{J}) > 0$ (analogous to assumptions on design matrix)

- $$0 = \nabla Q(\hat{\gamma}) = \hat{\boldsymbol{J}}^T \hat{\boldsymbol{r}} = \boldsymbol{J}^T \boldsymbol{r} - \boldsymbol{J}^T \boldsymbol{J}(\hat{\gamma} - \gamma) + o_P(\|\hat{\gamma} - \gamma\|)$$
- $\Rightarrow n^{1/2}(\hat{\gamma} - \gamma) \approx n^{1/2}(\boldsymbol{J}^T \boldsymbol{J})^{-1} \boldsymbol{J}^T \boldsymbol{r} \underbrace{\approx}_{n \text{ large}} N\left(0, \sigma^2 (n^{-1} \boldsymbol{J}^T \boldsymbol{J})^{-1}\right)$
- Under suitable regularity conditions, $\left(n^{-1} \hat{\boldsymbol{J}}^T \hat{\boldsymbol{J}}\right)^{-1} \approx (n^{-1} \boldsymbol{J}^T \boldsymbol{J})^{-1}$

Large Sample Sampling Distb'n

- $\hat{\boldsymbol{J}}$ is the $n \times p$ matrix of first derivatives evaluated at $\hat{\gamma}$
  - $\hat{\boldsymbol{J}}_{ij} = \frac{\partial f(\mathbf{X}_i, \gamma)}{\partial \gamma_j} |_{\gamma = \hat{\gamma}}$
- Assume that there is a positive definite matrix $\boldsymbol{A}$ such that $n^{-1}\hat{\boldsymbol{J}}^T\hat{\boldsymbol{J}} \to \boldsymbol{A}$.
- $n^{1/2}(\hat{\gamma} - \gamma) \to N(0, \sigma^2 \boldsymbol{A}^{-1})$
  - Exact proof is tedious and not the focus of this class. Provided a reference by Jennrich 1969 on Canvas.
  - Note that $Var(\hat{\gamma}) \approx \sigma^2 (n\boldsymbol{A})^{-1}$
  - Goal: get estimates of σ^2 and $n\boldsymbol{A}$ so that we can do large sample inference.

Estimation of the Variance

- $MSE = \frac{1}{n-p} \sum [Y_i - f(\mathbf{X}_i, \hat{\gamma})]^2$.
  - MSE is used to estimate σ^2.
  - Why do we divide by $n - p$ and not n. Can you motivate this mathematically?
  - MSE is not unbiased in the non-linear regression setting.
  - It is asymptotically unbiased.
- Obvious estimate of $n\boldsymbol{A}$ is $\hat{\boldsymbol{J}}^T\hat{\boldsymbol{J}}$.
- Estimate $Var(\hat{\gamma})$ with $s^2(\hat{\gamma}) = MSE \times (\hat{\boldsymbol{J}}^T\hat{\boldsymbol{J}})^{-1}$

Inference

- Inference on a single parameter γ_j uses the approximate distribution
  - $\frac{\hat{\gamma}_j - \gamma_j}{s(\hat{\gamma}_j)} \sim t_{n-p}$.
- Hypothesis testing and confidence intervals follow.
- For simultaneous confidence intervals, can use Bonferroni.
- Hypothesis testing for multiple parameters: Approximate F-test.
  - Fit a full and reduced model to obtain the sums-of-squares SSE(F) and SSE(R).
  - $F^* = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F}$
    - When $\epsilon_i \sim N\left(0, \sigma^2\right)$, this is the likelihood ratio statistic.
  - When the reduced model fits as well as the full, F^* is asymptotically distributed as $F_{df_R - df_F, df_F}$.
  - Just like linear regression, need reduced model to be a submodel of the original model.

General hypothesis testing

- Suppose $H_0 : \gamma \in \mathcal{S}_R$ for some subset $\mathcal{S}_R$ of the full parameter space is true.
- Let $\boldsymbol{J}_R \in \mathbb{R}^{n \times q_R}$ and $\boldsymbol{J}_F \in \mathbb{R}^{n \times q_F}$ are the Jacobians evaluated at the true γ

An overview of inference with large sample sizes

$$\boldsymbol{Y} = f(\boldsymbol{X};\gamma) + \boldsymbol{\epsilon}, \quad \epsilon_i \overset{i.i.d}{\sim} (0, \sigma^2)$$

Here, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\gamma \in \mathbb{R}^q$. Note we typically have $q \neq p$.

- Fundamental idea: for n large and $\boldsymbol{J} = \nabla_\gamma f(\boldsymbol{X};\gamma) \in \mathbb{R}^{n \times q}$,

  $$\hat{\gamma} \underbrace{\approx}_{\substack{\text{GN w/} \\ \gamma^{(0)}=\gamma}} \gamma + (\boldsymbol{J}^T\boldsymbol{J})^{-1}\boldsymbol{J}^T \underbrace{\{\boldsymbol{Y} - f(\boldsymbol{X};\gamma)\}}_{\boldsymbol{\epsilon}} \underbrace{\approx}_{\boldsymbol{J}\approx\hat{\boldsymbol{J}}} \gamma + (\hat{\boldsymbol{J}}^T\hat{\boldsymbol{J}})^{-1}\hat{\boldsymbol{J}}^T\boldsymbol{\epsilon}$$

  Therefore,

  $$\boldsymbol{Y} - f(\boldsymbol{X};\hat{\gamma}) \underbrace{\approx}_{\substack{\text{Taylor's} \\ \text{Theorem}}} \boldsymbol{\epsilon} - \hat{\boldsymbol{J}}(\hat{\gamma} - \gamma) \approx (I_n - H_{\hat{J}})\boldsymbol{\epsilon}$$

- If $f(\boldsymbol{X};\gamma) = \boldsymbol{X}\gamma$, check that the approximation is exact with $\boldsymbol{J} = \hat{\boldsymbol{J}} = \boldsymbol{X}$
- For large n, all inference proceeds as if we are using ordinary least squares with design matrix $\hat{\boldsymbol{J}} \in \mathbb{R}^{n \times q}$

Diagnostics

- Asymptotic normality relies on:
  1. Errors ϵ_i are i.i.d with mean 0 and variance σ^2.
  2. If ϵ_i is skewed, convergence is slow. Why?
  3. While not essential, it would be nice is ϵ_i were approximately normal, as this would lead to faster convergence and more accurate inference with smaller sample sizes.
- Easy checks:
  - the residuals vs. fitted values. This checks for accuracy of the mean function and constant variance.
  - the qq-plots. This checks normality.

Example in R

- Let’s analyze the example from last class in R.
- Determine the relationship between the amount of time from a cleaning product being manufactured and the fraction of available chlorine.
- R function "nls" estimates parameters with GN.
- Will run three statements:
    1. Using starting values close to the optimum.
    2. Using a poor starting value.
    3. Having R choose the starting values.
- In SAS: Proc NLIN.

Bootstrap confidence intervals

Bootstrapping cases to draw the *b*th data set for $b = 1, \ldots, B$

1. Select *n* numbers $\{m_{b1}, \ldots, m_{bn}\}$ from $\{1, \ldots, n\}$ with replacement.
2. The *b*th bootstrap data set is $\{\left(Y_{m_{bi}}, X_{m_{bi}1}, \ldots, X_{m_{bi}p-1}\right); i = 1, \ldots, n\}$

Bootstrapping Residuals

Fit the regression model to obtain the fitted values $\hat{Y}_i$ and the residuals $r_i = Y_i - \hat{Y}_i$. To draw the bth data set for $b = 1, \ldots, B$

1. Select n numbers $\{m_{b1}, \ldots, m_{bn}\}$ from $\{1, \ldots, n\}$ with replacement.
2. Letting $Y_i^b = \hat{Y}_i + r_{m_{bi}}$, the bth bootstrap data set is $\left\{\left(Y_i^b, X_{i1}, \ldots, X_{ip-1}\right); i = 1, \ldots, n\right\}$

Bootstrap Inference

- For either sampling scheme, obtain the parameter estimates $\hat{\gamma}_0^b, \ldots, \hat{\gamma}_{p-1}^b$ from the *b*th bootstrap data set.
- $(1 - \alpha)$ confidence intervals can be constructed as follows:
    - Let $\Gamma_j(p)$ be the $100 \times p$ percentile of $\hat{\gamma}_j^1, \ldots, \hat{\gamma}_j^B$.
    - The reflection confidence interval is $[2\hat{\gamma} - \Gamma_j(1 - \alpha/2), 2\hat{\gamma} - \Gamma_j(\alpha/2)]$, where $\hat{\gamma}$ is estimate from the original data. See also pp460 for the reflection confidence interval.
    - Percentile bootstrap confidence intervals are just the lower and upper $(1 - \alpha/2)$ percentiles.

Results

- Asymptotic inference:
  - The long term fraction of available chlorine γ_0 is estimated as .39 with a 95% CI of (.38, .40) and standard error 0.0052.
  - The growth rate γ_2 is estimated as -.10 with a 95% CI of (-.13, -.07) and standard error 0.013.
- Inference from 1000 bootstrap samples:
  - The 95% CI for the long term fraction of available chlorine is estimated as (0.38, 0.40).
  - The 95% CI for growth rate γ_2 is estimated as (-.13, -.08).
- We end up with the the same inference rounding to two significant digits for this example. This will not always the case.

The Cow data

- We want to know how serum dilution affects the presence of antibodies from cows.
- We take two samples from one cow: one in May and one in June.
- We separate each monthly sample into 16 equal parts. Each of these equal parts is then diluted and we observe optical densities Y.
- We have 8 different dilutions so that for each month, two observations are taken per dilution.
- Let $X = \log(\text{dilution})$.
- Main goal: is the relationship between X and Y the same for both months?

Our Model

- For one month, we will model the data as
  - $f(x_i, \theta) = \theta_1 + \frac{\theta_2 - \theta_1}{1 + \exp[\theta_3(x_i - \theta_4)]}$
  - Assume all parameters are positive and $\theta_2 > \theta_1$.
  - Has a reverse “S” shape as a function over x.
  - θ_1 is the smallest value.
  - θ_2 is the largest value.
  - θ_3 describes the rate of change.
  - θ_4 describes the inflection point.
- Our plan:
  - Look at each month individually.
  - Fit one single model to allow us to test equivalence of the curves.
  - If they are not equal, figure out what is different.

Starting Values

- Let's consider only the data from May.
- We know that θ_1 and θ_2 are the lower and upper bounds.
  - Take initial values to slightly smaller that the smallest and slightly larger than the largest observed Y.
- Conditional on these values, can we find a nice form for θ_3 and θ_4?
- Let $z_{i1} = \frac{Y_i - \theta_1}{\theta_2 - \theta_1}$
- $Ez_{i1} = \frac{1}{1+\exp(\theta_3(x_i - \theta_4))}$.
- Note that $\theta_3(x_i - \theta_4) = \log\left(\frac{1-Ez_{i1}}{Ez_{i1}}\right)$.

Starting Values Cont.

- Let $z_{i2} = \log\left(\frac{1-z_{i1}}{z_{i1}}\right)$
- Do a linear regression of z_{i2} on x_i.
- Slope term will be a starting estimate of θ_3.
- Minus intercept term over slope term will be a starting estimate for θ_4.

Testing

- We can use dummy variables to form a combined model for the two months.
- Let $M_i = 0$ for May and 1 for June.
- $f(x_i, M_i, \theta, \delta) = \theta_1 + \delta_1 M_i + \frac{\theta_2 + \delta_2 M_i - \theta_1 - \delta_1 M_i}{1 + \exp((\theta_3 + \delta_3 M_i)(x_i - \theta_4 - \delta_4 M_i))}$
- We want to test $H_0 : \delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$.
- Can do it via the F-test.