Regression Analysis I

Nonlinear Regression

Hossein Moradi Rekabdarkolaee

South Dakota State University

email: *hossein.moradirekabdarkolaee@sdstate.edu*.

Office: Arch, Math, and Engineering Building, *Room: 254*.

Multiple Linear Regression

- A multiple linear regression model can be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p] = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Nonlinear Regression

- The basic assumption in linear regression model: model is linear in terms of θ's.
- Consider a model
$$y_i = f(\mathbf{x}_i; \theta) + \epsilon_i$$
where parameters θ in the model is nonlinear.
- **Example**:
  1. Exponential model:
  $$y_i = \beta_0 \exp\{\beta_1 \mathbf{x}_i\} + \epsilon_i.$$
  2. $$y_i = \beta_0 + \beta_1 \exp\{\beta_2 \mathbf{x}_i\} + \epsilon_i.$$
  3. Generalized logistic model:
  $$y_i = \frac{\beta_0}{1 + \exp\{-\mathbf{x}_i^T \beta\}} + \epsilon_i.$$
- The data structure is the same as linear regression.

Nonlinear Regression

- LS involves finding $\hat{\theta}$ which minimizes

$$\sum_{i=1}^{n}[y_i - f(\mathbf{x}_i;\hat{\theta})]^2.$$

- If we assume normality assumption for error, i.e. $\epsilon_i \sim N(0,\sigma^2)$ for all i, the likelihood function is

$$L(\theta,\sigma^2;\mathbf{x}_i) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - f(\mathbf{x}_i;\hat{\theta})]^2\right\}$$

- ML involves maximizing $L(\theta,\sigma^2,\mathbf{x}_i)$ which is equivalent to LS.

Nonlinear Regression

- Since these are nonlinear functions, finding an analytical solution for them can be difficult and sometimes infeasible.
- There are iterative procedures that can be used to find LS and ML estimates for θ including, but not restricted to,
    - IRWLS (we talked about this in detail).
    - Gauss-Newton procedure.
    - Marquardt procedure.

Gauss-Newton procedure

- The Gauss-Newton procedure uses Taylor series around an initial value.
- We choose a set of starting values $\theta_0 = (\theta_{1,0}, \theta_{2,0}, \ldots, \theta_{p,0})^T$.
- The first order Taylor Series expansion of $f(\mathbf{x}_i, \theta)$ around $\theta = \theta_0$, only contains linear terms:

$$f(\mathbf{x}_i;\theta) \approx [f(\mathbf{x}_i;\theta)]_{\theta=\theta_0} + \left(\frac{\partial f(\mathbf{x}_i,\theta)}{\partial \theta_1}\right)_{\theta=\theta_0} (\theta_1 - \theta_{1,0}) + \ldots + \left(\frac{\partial f(\mathbf{x}_i,\theta)}{\partial \theta_p}\right)_{\theta=\theta_0} (\theta_p - \theta_{p,0})$$

- Letting $w_{ji} = \left(\frac{\partial f(\mathbf{x}_i,\theta)}{\partial \theta_j}\right)_{\theta=\theta_0}$ and $\gamma_j = (\theta_j - \theta_{j,0})$ then we have:

$$f(\mathbf{x}_i,\theta) - f(\mathbf{x}_i,\theta_0) \approx \gamma_1 w_{1i} + \ldots + \gamma_p w_{pi}.$$

- Thus, γ_j plays the regression coefficient role and w_{ji} are regressors and we have a linear regression structure.

Gauss-Newton procedure

- This procedure is as follows:

  1. Use OLS to regress $y_i - f(\mathbf{x}_i; \theta_0)$ versus w_{ji} to obtain $\{\hat{\gamma}_{1,1}, \dots, \hat{\gamma}_{p,1}\}$.
  2. Since $\gamma_j = \theta_j - \theta_{j,0}$, then $\hat{\theta}_{j,1} = \hat{\gamma}_j + \theta_{j,0}$. These are first iteration estimates of θ.
  3. Use the $\hat{\theta}_1$ as new starting value and repeat previous steps.
  4. Continue this procedure until convergence.

- Any time we do nonlinear regression, we have collinearity problem. (Why??).
- At s^{th} stage, we have $\hat{\theta}_s = \hat{\theta}_{s-1} + \hat{\gamma}_s$.
- If $\hat{\gamma}_s$ are poorly estimated, then the whole procedure can move to the wrong direction. (does this happen??)

Gauss-Newton procedure

- One conservative alternative approach is to modify the general procedure
  1. Compute Gauss-Newton increment vector $\hat{\gamma}_s$ for s^{th} iteration.
  2. Compute $\hat{\theta}_s = \hat{\theta}_{s-1} + \hat{\gamma}_s$.
  3. Compute SSE at each stage and if:
     - $SSE_s < SSE_{s-1}$ continue to the next iteration using $\hat{\theta}_s$
     - $SSE_s > SSE_{s-1}$, go back to step 2 and compute $\hat{\theta}_s = \hat{\theta}_{s-1} + \frac{\hat{\gamma}_s}{2}$
  4. Continue step (3) until either SSE goes down or increment has been halved 10 times.
  5. Continue steps (1) to (4) until convergence.
- Most software use this approach.

Marquardt procedure

- Since Collinearity usually exists in nonlinear regression, the alternative approach to Gauss-Newton is the Marquardt procedure which is very similar to Ridge regression.
- Collinearity causes $W^T W$ to be ill-conditioned, i.e. large off-diagonal elements.
- The Marquardt procedure add a small positive constant (usually less than 2) to the diagonal elements of $W^T W$.
- This will reduce the size of $\hat{\gamma}_s$ and its variance.
- At each iteration of the Marquardt procedure

$$\hat{\gamma}_s = (W^T W + k\mathbf{I})^{-1} W^T \mathbf{z}; \;\; \mathbf{z} = \mathbf{y} - f(\mathbf{x}_i; \theta).$$

- We have to choose an optimal value for k

Marquardt procedure

- Since at each iteration, we want SSE to decrease, we require that $SSE(\hat{\theta}_s) < SSE(\hat{\theta}_{s-1})$.
- Since
$$\hat{\gamma}_s = (W^T W + k\mathbf{I})^{-1} W^T \mathbf{z},$$
- Thus
$$\frac{var(\hat{\gamma}_s)}{\sigma^2} = (W^T W + k\mathbf{I})^{-1} W^T W (W^T W + k\mathbf{I})^{-1}$$
- Using eigenvalue decomposition approach, there exists an orthogonal matrix V such that $V^T W^T W V = \Lambda$ and $V^T (W^T W + \mathbf{I}) V = \Lambda_k$ where $\Lambda = diag(\lambda_1, \ldots, \lambda_p$ and $\Lambda_k = diag(\lambda_1 + k, \ldots, \lambda_p + k)$.
- Then $W^T W = V \Lambda V^T$ and $W^T W + \mathbf{I} = V \Lambda_k V^T$.

Marquardt procedure

- Thus

$$\frac{var(\hat{\gamma}_s)}{\sigma^2} = (V\Lambda_k^{-1}V^T)V\Lambda V^T(V\Lambda_k^{-1}V^T) = V\Lambda_k^{-1}\Lambda\Lambda_k^{-1}V^T.$$

  and

$$\sum_{i=1}^{p}\frac{var(\hat{\gamma}_{i,s})}{\sigma^2} = \sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2}.$$

- Since in the nonlinear scenario, we do not have the intercept, we do not need to center the covariates **z**.

- However, we meed to scale so that the constant *k* is added to scale-free values.

Statistical Inference in Nonlinear Regression

- Consider

$$z_i = y_i - f(\mathbf{x}_i, \theta) = \mathbf{w}_i^T \gamma + \epsilon_i; \quad \mathbf{w}_i = \frac{\partial f(\mathbf{x}_i, \theta)}{\partial \theta}$$

- In order to do inference, we need to find the information matrix.
- Assuming normality of the errors, the likelihood is

$$L(\theta, \sigma^2; \mathbf{x}_i) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - f(\mathbf{x}_i; \hat{\theta})]^2 \right\}$$

- Thus

$$\begin{aligned} \frac{\partial ln[L(\theta, \sigma^2; \mathbf{x}_i)]}{\partial \theta} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} [y_i - f(\mathbf{x}_i; \hat{\theta})] \frac{\partial f(\mathbf{x}_i, \theta)}{\partial \theta} \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} [y_i - f(\mathbf{x}_i; \hat{\theta})] \mathbf{w}_i \end{aligned}$$

Statistical Inference in Nonlinear Regression

- And

$$\frac{\partial}{\partial\theta^T}\left(\frac{\partial ln[L(\theta,\sigma^2;\mathbf{x}_i)]}{\partial\theta}\right) = \frac{\partial}{\partial\theta^T}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}[y_i - f(\mathbf{x}_i;\hat{\theta})]\mathbf{w}_i\right)$$
$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left\{y_i\frac{\partial\mathbf{w}_i}{\partial\theta^T} - f(\partial\mathbf{x}_i;\hat{\theta})\frac{\partial\mathbf{w}_i}{\partial\theta^T} - \mathbf{w}_i\frac{\partial f(\mathbf{x}_i;\hat{\theta})}{\partial\theta^T}\right\}$$

where $\frac{\partial\mathbf{w}_i}{\partial\theta^T}$ is the Hessian Matrix.

- Thus, the information matrix is

$$\mathbf{I}(\theta) = -E\left\{\frac{\partial}{\partial\theta^T}\left(\frac{\partial ln[L(\theta,\sigma^2;\mathbf{x}_i)]}{\partial\theta}\right)\right\}$$
$$= -E\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n}\left\{y_i\frac{\partial\mathbf{w}_i}{\partial\theta^T} - f(\partial\mathbf{x}_i;\hat{\theta})\frac{\partial\mathbf{w}_i}{\partial\theta^T} - \mathbf{w}_i\frac{\partial f(\mathbf{x}_i;\hat{\theta})}{\partial\theta^T}\right\}\right\}$$
$$= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}E(y_i)\frac{\partial\mathbf{w}_i}{\partial\theta^T} + \sum_{i=1}^{n}f(\partial\mathbf{x}_i;\hat{\theta})\frac{\partial\mathbf{w}_i}{\partial\theta^T} + \sum_{i=1}^{n}\mathbf{w}_i\mathbf{w}_i^T\right)$$
$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{w}_i\mathbf{w}_i^T = \frac{1}{\sigma^2}W^TW.$$

Statistical Inference in Nonlinear Regression

- Thus the asymptotic covariance matrix of θ is $\mathbf{I}^{-1}(\theta) = \sigma^2(W^T W)^{-1}$.

- Recall that in linear regression, $var(\hat{\beta}) = \sigma^2(X^T X)^{-1}$ exactly, not just asymptotically.

- Now consider properties of the estimated response. For some specific $\mathbf{x} = \mathbf{x}_0$, asymptotically

$$\frac{var[f(\mathbf{x}_0, \hat{\theta})]}{\sigma^2} = \mathbf{w}_0^T (W^T W)^{-1} \mathbf{w}_0$$

where $\mathbf{w}_0 = \left(\frac{\partial f(\mathbf{x}_0, \hat{\theta})}{\partial \theta_1}, \ldots, \frac{\partial f(\mathbf{x}_0, \hat{\theta})}{\partial \theta_p} \right)^T$

Statistical Inference in Nonlinear Regression

- To test $H_0 : \theta = 0$ versus $H_1 \neq 0$, an asymptotic test statistic is

$$t = \sqrt{\frac{\hat{\theta}^T(W^T W)\hat{\theta}}{\hat{\sigma}^2}}$$

  which approximately follow a t distribution with $n - p$ degrees of freedom.

- To test a single parameter $H_0 = \theta_j = \theta_{j,0}$, the approximate test statistic is

$$t = \frac{\hat{\theta}_j - \theta_{j,0}}{s_{\theta_j}}$$

  which has $n - p$ degrees of freedom and s_{θ_j} is the square root of j-thdiagonal element of $\hat{\sigma}^2(W^T W)^{-1}$.

Statistical Inference in Nonlinear Regression

- Approximate $100(1-\alpha)\%$ confidence interval on mean response at $\mathbf{x}_0$ is

$$f(\mathbf{x}_0, \hat{\theta}) \pm t_{n-p,1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{\mathbf{w}_0^T (W^T W)^{-1} \mathbf{w}_0}$$

  and for prediction of the individual response this interval is

$$f(\mathbf{x}_0, \hat{\theta}) \pm t_{n-p,1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{1 + \mathbf{w}_0^T (W^T W)^{-1} \mathbf{w}_0}$$

- For diagnostics, the Hat matrix is $\mathbf{H} = W(W^T W)^{-1} W^T$ with diagonal elements $h_{ii} = \mathbf{w}_i^T (W^T W)^{-1} \mathbf{W}_i$.
- For outlier diagnostics, studentized residuals are $r_i = \frac{y_i - f(\mathbf{x}, \hat{\theta})}{\hat{\sigma}(1-h_{ii}}$.
- For influence diagnostics a DFFITS-type statistic is $r_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$