Regression Methods

Myrto Limnios

Autumn 2025 - Week 5

General Models - Revisions

Likelihood

Definition 1

Let y be a data set, assumed to be the realisation of a random variable $Y \sim f(y;\theta)$, where the unknown parameter θ lies in the parameter space $\Omega_\theta \subset \mathbb{R}^p$.

Then the **likelihood** *(for θ based on y) and the corresponding* **log likelihood** *are*

$$L(\theta) = L(\theta; y) = f_Y(y;\theta), \quad \ell(\theta) = \log L(\theta), \quad \theta \in \Omega_\theta.$$

The **maximum likelihood estimate** *(MLE) $\widehat{\theta}$ satisfies $\ell(\widehat{\theta}) \geq \ell(\theta)$, for all $\theta \in \Omega_\theta$.*

Often $\widehat{\theta}$ is unique and in many cases it satisfies the **score (or likelihood) equation**

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0,$$

which is interpreted as a vector equation of dimension $p \times 1$ if θ is a $p \times 1$ vector.

The **observed information** *and* **expected (Fisher) information** *are defined as*

$$J(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}, \quad I(\theta) = \mathrm{E}\{J(\theta)\};$$

these are $p \times p$ matrices if θ has dimension p.

Regular model

We say that a statistical model $f(y;\theta)$ is **regular (for likelihood inference)** if

1. the true value θ^0 of θ is interior to the parameter space $\Omega_\theta \subset \mathbb{R}^p$;
2. the densities defined by any two different values of θ are distinct;
3. there is an open set $\mathcal{I} \subset \Omega_\theta$ containing θ^0 within which the first three derivatives of the log likelihood with respect to elements of θ exist almost surely, and
$$\left| \frac{\partial^3 \log f(Y_j;\theta)}{\partial\theta_r \partial\theta_s \partial\theta_t} \right| \leq g(Y_j)$$
uniformly for $\theta \in \mathcal{I}$, where $0 < \mathrm{E}_0\{g(Y_j)\} = K < \infty$; and
4. for $\theta \in \mathcal{I}$ we can interchange differentation with respect to θ and integration, that is,
$$\frac{\partial}{\partial\theta}\int f(y;\theta)\, dy = \int \frac{\partial f(y;\theta)}{\partial\theta}\, dy, \quad \frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\int f(y;\theta)\, dy = \int \frac{\partial^2 f(y;\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\, dy.$$

The results are also true under weaker conditions, for non-identically distributed and dependent data.

Maximum likelihood estimator

- In large samples from a **regular model** in which the true parameter is $\theta^0_{p\times 1}$, the maximum likelihood estimator $\widehat{\theta}$ has an approximate normal distribution,

$$\widehat{\theta} \dot{\sim} \mathcal{N}_p \left\{ \theta^0, J(\widehat{\theta})^{-1} \right\},$$

  so we can compute an approximate $(1-2\alpha)$ confidence interval for the rth parameter θ^0_r as

$$\widehat{\theta}_r \pm z_\alpha v_{rr}^{1/2},$$

  where v_{rr} is the rth diagonal element of the matrix $J(\widehat{\theta})^{-1}$.

- This is easily implemented:
  - we code the negative log likelihood $-\ell(\theta)$ (and check the code carefully!);
  - we minimise $-\ell(\theta)$ numerically, ensuring that the minimisation routine returns $\widehat{\theta}$ and the Hessian matrix $J(\widehat{\theta}) = -\partial^2 \ell(\theta)/\partial\theta\partial\theta^{\mathrm{T}}|_{\theta=\widehat{\theta}}$
  - we compute $J(\widehat{\theta})^{-1}$, and use the square roots of its diagonal elements, $v_{11}^{1/2}, \ldots, v_{dd}^{1/2}$, as standard errors for the corresponding elements of $\widehat{\theta}$.

Comments on regular models

Condition

1. is needed so that $\widehat{\theta}$ can lie 'on both sides' of θ^0 and hence can have a limiting normal distribution, once standardized—**fails**, for example, if θ has a discrete component (e.g. changepoint $\gamma \in \{1, \ldots, n\}$);
2. is needed to be able to identify the model on the basis of the data;
3. ensures the validity of Taylor series expansions of $\ell(\theta)$—not usually a problem;
4. ensures that the score statistic has a limiting normal distribution—can **fail** in some models — sometimes good news, leading to faster convergence than $n^{-1/2}$.

All the above assumes the postulated model is correct! — there is a literature on what happens when we fit the wrong model, or if the parameter dimension increases with n for example, but usually there are no generic results for such cases.

Likelihood ratio statistic

- Model $f_B(y)$ is **nested** within model $f_A(y)$ if A reduces to B on restricting some parameters:
  - for example, the model $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ is nested within the model $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, because the first is obtained from the second by setting $\mu = 0$;
  - the maximised log likelihoods satisfy $\widehat{\ell}_A \geq \widehat{\ell}_B$, because the more comprehensive model A contains the simpler model B.
- The **likelihood ratio statistic** for comparing them is

$$W = 2(\widehat{\ell}_A - \widehat{\ell}_B).$$

- If the model is regular, the simpler model is true, and A has q more parameters than B, then

$$W \dot{\sim} \chi^2_q.$$

B is true and depends on say 2 parameters then $q = p - 2$

- This implicitly assumes that ML inference for model A is OK, so that the approximation $\widehat{\theta}_A \dot{\sim} \mathcal{N}\{\theta_A, J_A(\widehat{\theta}_A)^{-1}\}$ is adequate.

Profile log likelihood

- Consider a regular log likelihood $\ell(\psi, \lambda)$, where the **parameter of interest** ψ is variation independent of the **nuisance parameter** λ, i.e., $(\psi, \lambda) \in \Omega_\psi \times \Omega_\lambda$, and the overall MLE is $(\widehat{\psi}, \widehat{\lambda})$.
- For a confidence set for ψ, without reference to λ, we use the **profile log likelihood**

$$\ell_{\mathrm{p}}(\psi) = \max_{\lambda \in \Omega_\lambda} \ell(\psi, \lambda) = \ell(\psi, \widehat{\lambda}_\psi),$$

  say, and, based on the limiting distribution of the likelihood ratio statistic, take as $(1 - 2\alpha)$ confidence region the set

$$\left\{ \psi \in \Omega_\psi : 2\{\ell(\widehat{\psi}, \widehat{\lambda}) - \ell(\psi, \widehat{\lambda}_\psi)\} \leq \chi^2_{\dim \psi}(1 - 2\alpha) \right\}.$$

- When ψ is scalar, this yields

$$\left\{ \psi \in \Omega_\psi : \ell(\psi, \widehat{\lambda}_\psi)\} \geq \ell(\widehat{\psi}, \widehat{\lambda}) - \tfrac{1}{2}\chi^2_1(1 - 2\alpha) \right\},$$

  and $\frac{1}{2}\chi^2_1(0.95) = 1.92$.
- Such intervals are generally better than the standard interval $\widehat{\psi} \pm z_\alpha \mathrm{SE}$, particularly when the distribution of $\widehat{\psi}$ is asymmetric, but require more computation, since they involve many maximisations of ℓ.

Model setup

- Independent random variables $Y_1, \ldots, Y_n$, with observed values $y_1, \ldots, y_n$, and covariates $x_1, \ldots, x_n$.
- Suppose that probability density of Y_j is $f(y_j; \eta_j, \phi)$, where $\eta_j = \eta(\beta, x_j)$, and ϕ is common to all models.
- We will later refer to η as the **link function**, and to ϕ as the **dispersion**.
- Log likelihood is

$$\ell(\beta, \phi) = \sum_{j=1}^{n} \ell_j(\beta, \phi) = \sum_{j=1}^{n} \log f\{y_j; \eta(\beta, x_j), \phi\}.$$

- More generally, just let $\ell_j(\beta, \phi)$ denote the log likelihood contribution from the jth observation.
- Suppose ϕ known (for now), suppress it, and estimate β.

Example 9 (Normal regression model)

Express the normal regression model in the terms above.

Here $Y_j \overset{\text{ind}}{\sim} \mathcal{N}(\mu_j, \sigma^2)$ *with* $\mu_j = \eta_j = \eta(x_j; \beta)$, *so obviously*

$$\eta_j = \eta(x_j; \beta), \quad \phi = \sigma^2, \quad \ell_j \equiv -\tfrac{1}{2}\{(y_j - \eta_j)^2/\phi + \log \phi\}.$$

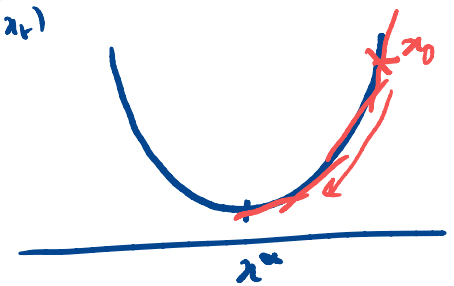
Iterative weighted least squares (IWLS)

$f(y_j, \eta_j, \phi)$ with $\eta_j = \eta(x_j, \beta)$

- General approach for estimation in regression models, based on Newton–Raphson iteration
- Assume that ϕ is fixed, and write

$$\ell(\beta) = \sum_{j=1}^{n} \ell_j\{\eta_j(\beta)\}.$$

$\rightsquigarrow \; x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$ until $\|x_{t+1} - x_t\| \leq tol$



- MLEs $\widehat{\beta}$ usually satisfy

$$\frac{\partial \ell(\widehat{\beta})}{\partial \beta_r} = 0, \qquad r = 1, \ldots, p,$$

or equivalently

$$\frac{\partial \ell(\widehat{\beta})}{\partial \beta} = \frac{\partial \eta^{\mathrm{T}}}{\partial \beta}\frac{\partial \ell}{\partial \eta} = \frac{\partial \eta^{\mathrm{T}}}{\partial \beta} u(\widehat{\beta}) = \sum_{j=1}^{n} \frac{\partial \eta_j}{\partial \beta}\frac{\partial \ell_j\{\eta_j(\beta)\}}{\partial \eta_j} = 0, \qquad (1)$$

$u(\hat{\beta})$

where $u(\beta)$ is $n \times 1$ vector with jth element $\partial \ell / \partial \eta_j$.

Derivation

- To find the MLE $\widehat{\beta}$ starting from a trial value β, we make a Taylor series expansion in (1), to obtain

$$\frac{\partial \eta^{\mathrm{T}}(\beta)}{\partial \beta} u(\beta) + \left\{ \sum_{j=1}^{n} \frac{\partial \eta_j(\beta)}{\partial \beta} \frac{\partial^2 \ell_j(\beta)}{\partial \eta_j^2} \frac{\partial \eta_j(\beta)}{\partial \beta^{\mathrm{T}}} + \sum_{j=1}^{n} \frac{\partial^2 \eta_j(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}} u_j(\beta) \right\} (\widehat{\beta} - \beta) \doteq 0. \tag{2}$$

If we denote the $p \times p$ matrix in braces on the left by $-J(\beta)$, assumed invertible, we can rearrange (2) to obtain

$$\widehat{\beta} \doteq \beta + J(\beta)^{-1} \frac{\partial \eta^{\mathrm{T}}(\beta)}{\partial \beta} u(\beta). \tag{3}$$

This suggests that maximum likelihood estimates may be obtained by starting from a particular β, using (3) to obtain $\widehat{\beta}$, then setting β equal to $\widehat{\beta}$, and iterating (3) until convergence. This is the Newton–Raphson algorithm applied to our particular setting. In practice it can be more convenient to replace $J(\beta)$ by its expected value

$$I(\beta) = \sum_{j=1}^{n} \frac{\partial \eta_j(\beta)}{\partial \beta} \mathrm{E}\left(-\frac{\partial^2 \ell_j}{\partial \eta_j^2} \right) \frac{\partial \eta_j(\beta)}{\partial \beta^{\mathrm{T}}};$$

the other term vanishes because $\mathrm{E}\{u_j(\beta)\} = 0$. We write

$$I(\beta) = X(\beta)^{\mathrm{T}} W(\beta) X(\beta), \tag{4}$$

where $X(\beta)$ is the $n \times p$ matrix $\partial \eta(\beta)/\partial \beta^{\mathrm{T}}$ and $W(\beta)$ is the $n \times n$ diagonal matrix whose jth diagonal element is $\mathrm{E}(-\partial^2 \ell_j / \partial \eta_j^2)$.

- If we replace $J(\beta)$ by $X(\beta)^{\mathrm{T}}W(\beta)X(\beta)$ and reorganize (3), we obtain

$$\widehat{\beta} = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W(X\beta + W^{-1}u) = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}Wz, \qquad (5)$$

  say, where the dependence of the terms on the right on β has been suppressed. That is, starting from β, the updated estimate $\widehat{\beta}$ is obtained by weighted linear regression of the $n \times 1$ vector **adjusted dependent variable**

$$z = X(\beta)\beta + W(\beta)^{-1}u(\beta)$$

  on the columns of $X(\beta)$, using weight matrix $W(\beta)$.
- The MLEs are obtained by repeating this step until the log likelihood, the estimates, or more often both, are essentially unchanged.
- The variable z plays the role of the response or dependent variable in the weighted least squares step.
- Often the structure of a model simplifies the estimation of an unknown value of ϕ. It may be estimated by a separate step between iterations of $\widehat{\beta}$, by including it in the step (3), or from the profile log likelihood $\ell_{\mathrm{p}}(\phi)$.

IWLS II: summary

- Newton–Raphson update step:

$$\widehat{\beta} = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W z,$$

where

$$\begin{aligned} X_{n\times p} &= \partial\eta/\partial\beta^{\mathrm{T}}, \quad \text{(design matrix)} \\ W_{n\times n} &= \mathrm{diag}\{\mathrm{E}(-\partial^2\ell_j/\partial\eta_j^2)\}, \quad \text{(weights)} \\ z_{n\times 1} &= X\beta + W^{-1}u, \quad \text{(adjusted dependent variable)} \end{aligned}$$

- Thus to obtain MLEs $\widehat{\beta}$ we use the **IWLS algorithm**:
- take an initial $\widehat{\beta}$. Repeat for $t = 1, \ldots, k$
  - compute X, W, u, z; → at step t
  - compute new $\widehat{\beta}$ and replace the preceding value; $\widehat{\beta}_{t+1} = (X_t^T W_t X_t)^{-1} X_t W_t z_t$

  until changes in $\ell(\widehat{\beta})$ (or, sometimes, $\widehat{\beta}$, or both) are lower than some tolerance. $|\ell(\widehat{\beta}_{t+1})| <$ threshold
- Sometimes a line search is added, if $\ell(\widehat{\beta}_{\mathrm{new}}) < \ell(\widehat{\beta}_{\mathrm{old}})$: i.e., we half the step length and try again.

Example

Example 10 (Normal nonlinear model)

Give the components of the IWLS algorithm for the normal nonlinear model.

Deviance

By analogy with standard linear regression, we want to check if the chosen model (and thus estimated optimal parameter) captures the characteristics of interest.

Recall that we estimated the residual sum of squares $RSS = \sum_{j \leq n} e_j^2$ *for explainability of the variance in* y*. Here we generalize it by considering the* **deviance** *of the model valued at two distinct points.*

Deviance-based statistics are used for log likelihood ratio tests.

- Let $\widehat{\eta}_j = \eta_j(\widehat{\beta}, x_j)$, where $\widehat{\beta}$ is MLE of β, giving maximised log likelihood $\ell(\widehat{\beta})$ and $\widehat{\eta}^{\mathrm{T}} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_n)$.
- Let $\tilde{\eta}_j$ be the value of η_j that maximises $\log f(y_j; \eta_j)$, and let $\tilde{\eta}^{\mathrm{T}} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_n)$. This corresponds to the **saturated model**, with

  $$\#\text{parameters in } \eta = \#\text{observations in } y,$$

  which will give the largest likelihood possible.
- Define the **scaled deviance**:

  $$D = 2 \sum_{j=1}^{n} \{\log f(y_j; \tilde{\eta}_j) - \log f(y_j; \widehat{\eta}_j)\} \geq 0.$$

- Small D implies $\widehat{\eta} \approx \tilde{\eta}$, so model fits well.
- Large D implies poor fit — like $SS(\widehat{\beta})$ in linear model.

Differences of deviances

- Consider two models:
  - Model A: $\beta^{\mathrm{T}} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ vary freely — MLEs $\widehat{\eta}^A = \eta(\widehat{\beta}^A)$;
  - Model B: $(\beta_1, \ldots, \beta_q) \in \mathbb{R}^q$ vary freely, but $\beta_{q+1}, \ldots, \beta_p$ are fixed — hence q free parameters, MLEs $\widehat{\eta}^B = \eta(\widehat{\beta}^B)$.
- Model B is **nested within** model A: B can be obtained by restricting A.
- Likelihood ratio statistic for comparing the models is

$$2(\widehat{\ell}_A - \widehat{\ell}_B) = 2\sum_{j=1}^{n}\left\{\log f(y_j; \widehat{\eta}_j^A) - \log f(y_j; \widehat{\eta}_j^B)\right\} = D_B - D_A,$$

  and this $\dot{\sim} \chi^2_{p-q}$ if the models are regular.
- If ϕ unknown, replace it by an estimate: same distributional approximations will apply.

Example 11 (Normal linear model)

Find the difference of deviances in the normal linear model.

General Models - Model Checking

Model checking

- Two basic approaches:
  - overall tests either using generic statistic (e.g., chi-squared) or by **model expansion** (e.g., adding a term and testing for significance);
  - **regression diagnostics** for detecting a few possibly dodgy observations.
- Most widely used diagnostics in the linear model $y = X_{n\times p}\beta + \varepsilon$ are **residuals** $e_j = y_j - \widehat{y}_j$ and (much better) **standardized residuals**

$$r_j = \frac{y_j - \widehat{y}_j}{s(1-h_{jj})^{1/2}}, \quad j = 1, \ldots, n,$$

where the **leverage** h_{jj} is the jth diagonal element of the hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, and the **Cook statistic**

$$C_j = \frac{1}{ps^2}(\widehat{y} - \widehat{y}_{-j})^{\mathrm{T}}(\widehat{y} - \widehat{y}_{-j}) = \frac{r_j^2 h_{jj}}{p(1-h_{jj})},$$

which measures the effect of deleting the jth case (x_j, y_j) on the fitted model.

Diagnostics in general case

- Linear model ideas work as approximations (2nd order Taylor series, painful expansions).
- **Leverage** h_{jj} defined as jth diagonal element of

$$H = W^{1/2} X (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W^{1/2},$$

  depends in general on $\widehat{\beta}$, unlike in linear model.
- **Cook statistic** is change in deviance

$$C_j = 2p^{-1} \left\{ \ell(\widehat{\beta}) - \ell(\widehat{\beta}_{-j}) \right\} \doteq \frac{h_{jj}}{p(1-h_{jj})} r_{Pj}^2,$$

  where $\widehat{\beta}_{-j}$ is MLE when jth case (x_j, y_j) is dropped, and r_{Pj} is **standardized Pearson residual** (see below).
- There are several types of residual compared to the linear case, where we only had $r_j = e_j/(s(1-h_{jj})^{1/2})$ with mean 0 and approx. unit variance (recall that we plot the residuals against each covariates to check if there are unwanted patterns) (see next page).

Residuals in general case

- **Deviance residual**:

$$d_j = \text{sign}(\tilde{\eta}_j - \widehat{\eta}_j)[2\{\ell_j(\tilde{\eta}_j;\phi) - \ell_j(\widehat{\eta}_j;\phi)\}]^{1/2},$$

  for which $\sum d_j^2 = D$ is deviance.

- **Pearson residual**: $u_j(\widehat{\beta})/\sqrt{w_j(\widehat{\beta})}$.

- Standardized versions

$$r_{Dj} = \frac{d_j}{(1-h_{jj})^{1/2}}, \quad r_{Pj} = \frac{u_j(\widehat{\beta})}{\{w_j(\widehat{\beta})(1-h_{jj})\}^{1/2}},$$

  and (even better)

$$r_j^* = r_{Dj} + r_{Dj}^{-1}\log(r_{Pj}/r_{Dj}) \dot{\sim} N(0,1)$$

  for many models.

- These all reduce to usual standardized residual for normal linear model.

- So, we can compute these residuals and plot against the covariates to check for patterns.

Example

Example 12 (Gumbel linear model)

Give the components of the IWLS algorithm for fitting the linear model

$$y_j = \beta_0 + \beta_1(x_j - \overline{x}) + \tau\varepsilon_j, \quad j = 1, \ldots, n,$$

with Gumbel errors having density function

$$f(y_j; \eta_j, \tau) = \tau^{-1} \exp\left\{-\frac{y_j - \eta_j}{\tau} - \exp\left(-\frac{y_j - \eta_j}{\tau}\right)\right\},$$

where $\tau > 0$ and $\eta_j = \beta_0 + \beta_1(x_j - \overline{x})$; this distribution is natural for maxima; note that τ^2 is not the variance.

Summary

- For regression problems with independent responses y_j dependent on parameters β through parameter $\eta_j = \eta(x_j; \beta)$, generalise least squares estimation to maximum likelihood estimation, using iterative weighted least squares algorithm: iterate to convergence

$$\widehat{\beta} = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W z, \quad z = X\beta + W^{-1} u,$$

  where

$$X_{n\times p} \equiv X(\beta) = \frac{\partial \eta}{\partial \beta^{\mathrm{T}}}, \quad u_{n\times 1} \equiv u(\eta) = \frac{\partial \ell}{\partial \eta}, \quad W_{n\times n} \equiv W(\eta) = -\mathrm{E}\left\{ \frac{\partial^2 \ell}{\partial \eta \partial \eta^{\mathrm{T}}} \right\},$$

  with ℓ the log likelihood for the data.
- Standard likelihood theory is used for confidence intervals and model comparison.
- Linear model diagnostics (residuals, leverage, Cook statistics, ...) generalise to this setting.
- Next: generalized linear models (GLMs), wide class of models with exponential family-like response distributions.