Regression and Time Series

Jianqing Fan

Princeton University

https://fan.princeton.edu



2. Generalized Linear Models

2.1 Binary Responses

- **Read materials and R-implementations here**

Binary Response

Dichotomized response: Very frequently

Example: (Gene expression and autism) Over 60K gene expression profiles (Next Generation Sequence) are measured among 104 samples: 47 autisms and 57 healthy controls, along with gender, brain region, age, and sites. Of interest is to find the genes that are associated with autism. We select top 5 differently expressed (**feature screening**) by using two-sample t-test and would like to examine their effect on the response along with other variables.

Response: $Y = 1$ and 0, indicating autism or not.

Question: How to model $p(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$?

Modeling Binary Data

If a latent response (e.g. severity, distance to default) follows

$$Z = \alpha + \beta^T \mathbf{X} - \varepsilon, \qquad \textbf{linear model},$$

but we only get $Y = I(Z > c)$ for an unknown c.

Conditional probability: If $\varepsilon \sim F$, we have

$$p(\mathbf{x}) = P(Y = 1 | X = x) = P(\alpha + \mathbf{x}^T\beta - \varepsilon > c | \mathbf{x}) = F(\beta_0 + \mathbf{x}^T\beta)$$

where $\beta_0 = \alpha - c$. ★Same as $p(\mathbf{x}) = F(\mathbf{x}^T\beta)$ by dropping β_0 but with $x_1 = 1$.

Link function = $F^{-1}(\cdot)$ —maps $p(\mathbf{x})$ into a linear space spanned by $\mathbf{x}$

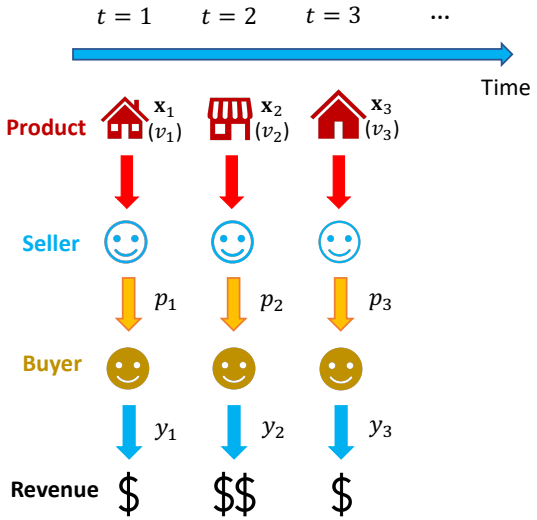
- ★ **probit link**: $F(x) = \Phi(x)$, normal cdf. ⟹ $p(\mathbf{x}) = \Phi(\mathbf{x}^T\beta)$
- ★ **logit link**: $F(x) = \frac{\exp(x)}{1+\exp(x)}$, ⟹ $p(\mathbf{x}) = \frac{\exp(\mathbf{x}^T\beta)}{1+\exp(\mathbf{x}^T\beta)}$ *(softmax)*

Dynamic pricing – another application

Price: $v(\mathbf{x}) = \mathbf{x}^T\theta - \varepsilon$, $\mathbf{x}$ = attributes (e.g. Airbnb), $\varepsilon \sim F$.

Observe: $Y = 1$ if $v(\mathbf{x}) > p$, p asked price.

$$P(Y = 1 | \mathbf{X} = \mathbf{x}) = F(\mathbf{x}^T\theta - p)$$



Optimal price: $p^*(\mathbf{x}) = \text{argmax}_p \; pF(\mathbf{x}^T\theta - p)$

expected rev.

Goal: learn θ and F from data $\{(\mathbf{x}_t, p_t, y_t)\}$ dynamically with min regret.

★Bernoulli regression with unknown link.

2.2. Maximum Likelihood Estimation

Statistical Framework —Generative Models

Statistical Modeling: Regard data $\mathbf{x}$ as a realization from a density $f(\mathbf{x}; \theta_0)$.

Goal: To learn parameters θ_0 and its uncertainty from the data $\mathbf{x}$

Example (Regression): Let $\mathbf{X} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ and the observations are i.i.d. from a multiple regression (generative) model:

$$Y_i = \beta_0^T \mathbf{X}_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Then, given $\mathbf{X}_i$, $Y_i \sim \mathcal{N}(\beta_0^T \mathbf{X}_i, \sigma^2)$ with the conditional density

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_i - \beta_0^T \mathbf{X}_i)^2}{2\sigma^2}\right)$$

Assume that the density of $\mathbf{X}_i$ is $g(\cdot)$. Then,

$$f(\mathbf{X}; \beta) = \prod_{i=1}^n f(\mathbf{X}_i, Y_i) = \prod_{i=1}^n g(\mathbf{X}_i) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_i - \beta^T \mathbf{X}_i)^2}{2\sigma^2}\right)$$

Maximum Likelihood Methods

Density: $f(\mathbf{x};\theta_0)$ as a function of $\mathbf{x}$ —probability two sides of the same coin

Likelihood: $f(\mathbf{x};\theta)$ as a function of θ —statistics, denoted by $L(\theta)$.

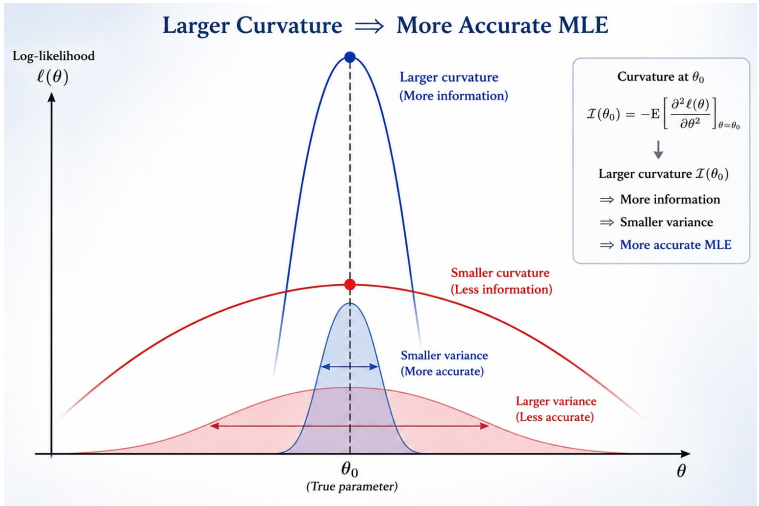
MLE: Find most probable θ to generate data:

$$\widehat{\theta}_{MLE} = \text{argmin}_\theta\, L(\theta) = \text{argmin}_\theta\, \ell(\theta), \qquad \ell(\theta) = \log L(\theta),$$

log-likelihood

satisfying the likelihood equation: $\ell'(\widehat{\theta}_{\mathbf{MLE}}) = \mathbf{0}$. The curvature indicates the uncertainty of MLE: $\text{var}(\widehat{\theta}) \approx -[\ell'(\widehat{\theta}_{MLE})]^{-1}$.

- ★ $\boxed{\text{var}(\widehat{\theta})^{-1/2}(\widehat{\theta}_{MLE} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_p)}$
- ★ MLE is most efficient if model is correct;
- ★ MLE is not robust to model misspectification.



Gaussian MLE

For multiple regression, 1^{st} term of the likelihood is indep of $\beta \Longleftrightarrow$ conditional MLE

$$\begin{aligned} \ell(\beta, \sigma) &= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(Y_i - \beta_0^T \mathbf{X}_i)^2}{2\sigma^2}\Big) \\ &= \text{const} - n \log \sigma - \sum_{i=1}^{n} \frac{(Y_i - \beta^T \mathbf{X}_i)^2}{2\sigma^2} \end{aligned}$$

MLE becomes the least-squares: $\widehat{\beta} = \text{argmin}_\beta \sum_{i=1}^n (Y_i - \beta^T \mathbf{X}_i)^2$ with

$$\ell(\widehat{\beta}, \sigma) = \text{const} - n \log \sigma - \frac{\text{RSS}}{2\sigma^2}$$

Taking derivative w.r.t. σ and setting it to zero, we get

$$\widehat{\sigma}^2_{MLE} = n^{-1} \text{RSS}, \qquad \text{—a diff normalization constant}$$

Variance of MLE: Hold exactly as inverse of the negative Hessian matrix

$$\text{var}(\widehat{\beta}) = [-\ell''(\beta)]^{-1} = \sigma^2 [\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T]^{-1} = \sigma^2 [\mathbf{X}^T \mathbf{X}]^{-1}$$

Bernoulli MLE

★Conditional dist: $(Y_i|X_i) \sim \text{Bernoulli}(p(\mathbf{X}_i))$ with "density" $p(X_i)^{Y_i}(1-p(X_i))^{1-Y_i}$

★Conditional joint density: $\prod_{i=1}^n p(X_i)^{Y_i}(1-p(X_i))^{1-Y_i}$ with log-likelihood

$$\ell(\beta) = \sum_{i=1}^n Y_i \log p(X_i) + (1-Y_i)\log(1-p(X_i)), \qquad p(X_i) = F(\mathbf{X}_i^T\beta)$$

Logist regression: For logit link $p(\mathbf{x}) = \frac{\exp(\beta^T\mathbf{x})}{1+\exp(\beta^T\mathbf{x})}$, we have

$$\ell(\beta) = \sum_{i=1}^n Y_i\mathbf{X}_i^T\beta - \log(1+\exp(\mathbf{X}_i^T\beta))$$

—MLE $\widehat{\beta}$ has no explicit solution, but it is a convex optimization

—Uncertainty quantification of estimate is given by

$$\boxed{\ell''(\beta) = -\textstyle\sum_{i=1}^n \widehat{p}(\mathbf{X}_i)(1-\widehat{p}(\mathbf{X}_i))\mathbf{X}_i\mathbf{X}_i^T}, \qquad \widehat{p}(\mathbf{x}) = \frac{\exp(\widehat{\beta}^T\mathbf{x})}{1+\exp(\widehat{\beta}^T\mathbf{x})}$$

Example: (**Gene expression and autism**) Over 60K gene expression profiles (Next Generation Sequence) are measured among 104 samples: 47 autisms and 57 healthy controls, along with gender, brain region, age, and sites. Of interest is to find the genes that are associated with autism. We select top 5 differently expressed by using two-sample *t*-test and fit logistic regression along with other variables.

Data: autism.csv

```
> autism  = read.csv("autism.csv")    #reading the data
> aut.glm = glm(Autism ~ . , family=binomial, data=autism)
>    #fitting the model
> summary(aut.glm)   #summarize the fit

Call:
glm(formula = Autism ~ ., family = binomial, data = autism)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.4105  -0.5834  -0.1647   0.4863   2.5613

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.33425    2.56463  -0.520 0.602889
GenderM      0.14585    0.73279   0.199 0.842233
Age         -0.05945    0.02871  -2.071 0.038365 *
SiteM       -3.43602    0.95416  -3.601 0.000317 ***
Reg          1.17445    0.57933   2.027 0.042636 *
Gene1       -0.10237    0.14148  -0.724 0.469332
```

```
Gene2         0.43250    0.32752   1.321 0.186658
Gene3         0.78675    0.26275   2.994 0.002751 **
Gene4        -0.66137    0.30426  -2.174 0.029729 *
Gene5         0.08676    0.26373   0.329 0.742165
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 143.212  on 103  degrees of freedom
Residual deviance:  74.617  on  94  degrees of freedom
AIC: 94.617
```

We now select model by using stepwise procedure `step(aut.glm)`. It selects the model:

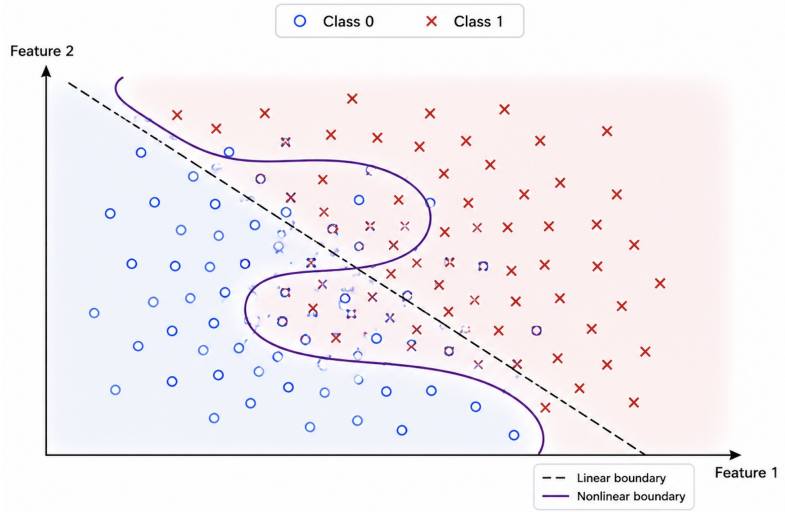
```
> aut.glm1 = glm(Autism ~ Age + Site + Reg + Gene3 + Gene4,
family=binomial, data=autism)
> summary(aut.glm1)     #summarize the fit

Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.01125    2.12388   0.005 0.995773
Age         -0.06377    0.02804  -2.275 0.022928 *
SiteM       -3.31923    0.85777  -3.870 0.000109 ***
Reg          1.05110    0.52212   2.013 0.044099 *
Gene3        0.89643    0.22623   3.962 7.42e-05 ***
Gene4       -0.51391    0.18172  -2.828 0.004684 **
Residual deviance:  137.226  on  98  degrees of freedom
```

We now predict (in-sample) and compute the misclassification rate. For each given $\mathbf{x}$, we compute $p(\mathbf{x}) = \frac{\exp(\widehat{\beta}^T \mathbf{x})}{1+\exp(\widehat{\beta}^T \mathbf{x})}$, which is the estimated probability $P(Y|\mathbf{X} = x)$. Classify it as 1 if $p(\mathbf{x}) > 0.5$. The in-sample misclassification rate is 13.46%

```
> logit = predict(aut.glm1)              #fitted log(odd-ratios)
> prob = exp(logit)/(1+exp(logit))      #fitted probability
> classification = (prob > 0.5)         #classification
### equivalent to directly using    (logit > 0)
> mean(autism[,1] != classification)    #compute misclassification rate
[1] 0.1346154
```

★Logist reg. results are organized in the same way as OLS

★Logistic regression can be used for classification



Can the two cases be further generalized?

Likelihood ratio test and ANAOVA

Hypothesis: $H_0 : \theta_0 \in \mathbf{\Theta}_0$ vs. $H_1 : \theta_0 \in \mathbf{\Theta} - \mathbf{\Theta}_0$

Example: Is a subset of variable $\mathbf{X}_S$ adequate? $\mathbf{\Theta}_0 = \{\beta_{S^c} = 0\}$

MLR test: Reject H_0 when the MLR is large:

$$\Lambda = \frac{\max_{\theta \in \mathbf{\Theta}} L(\theta)}{\max_{\theta \in \mathbf{\Theta}_0} L(\theta)} \geq c' \Longleftrightarrow 2\log\Lambda = 2\Big(\ell(\widehat{\theta}_{MLE}) - \ell(\widehat{\theta}_{MLE,0})\Big) \geq c.$$

Wilks Theorem: $2\log\Lambda \overset{a}{\sim}_{H_0} \chi^2_{dim(\Theta)-dim(\Theta_0)}$, *(No. of restrictions)*.

Example: For Gaussian MLE, $\max_{\theta \in \mathbf{\Theta}} L(\theta) = \text{const} - n\log\widehat{\sigma}_{MLE}$ and

$$2\log\Lambda = n\log\frac{\widehat{\sigma}^2_{MLE,0}}{\widehat{\sigma}^2_{MLE}} \approx n\frac{\text{RSS}_0 - \text{RSS}}{\text{RSS}} \overset{a}{\sim}_{H_0} \chi^2_{p-s}, \qquad s = |S| \qquad \text{(approx F-test)}$$

Are genes and sites related to Audisim: Full model $2\ell(\widehat{\theta}_{MLE,0}) = 74.617$ (df =94);

Null: $2\ell(\widehat{\theta}_{MLE,0}) = 143.212$ (df =103), diff = 68.595 (df = 9); Pvalue ≈ 0.

Adequacy of GLIM 1: $2\ell(\widehat{\theta}^{GLIM1}_{MLE}) = 137.226$ (df = 98)

diff = 5.986 (df = 5), Pvalue = 0.3076.

2.3. Generalized Linear Models

put normal, binary regression, Poisson, Gamma, etc. in one framework with similar concepts (RSS, residuals)

Members of the exponential family

Bernoulli: $P(Y=y|\mathbf{X}=\mathbf{x}) = p(\mathbf{x})^y(1-p(\mathbf{x}))^{1-y}$

$$= \exp\left\{ y \underbrace{\log\frac{p(\mathbf{x})}{1-p(\mathbf{x})}}_{\theta(\mathbf{x})=\text{canonical parameter, interact with data } y} + \underbrace{\log(1-p(\mathbf{x}))}_{-b(\theta)} + \underbrace{0}_{c(y)} \right\}.$$

with $b(\theta) = \log(1+\exp(\theta))$.

Normal distribution: $f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\mu(\mathbf{x}))^2}{2\sigma^2}\right)$

$$= \exp\left(\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log\sqrt{2\pi}\sigma\right).$$

Here $\theta=\mu, \phi=\sigma^2$, $b(\theta)=\theta^2/2$ and $c(y,\phi) = -\frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)$.

- The canonical link function is the identity link $g(t)=t$.

Generalized linear models

Purpose: To accommodate various types of responses (binary, categorical, counts, continuous)

GLIM: $f(y|\mathbf{X}=\mathbf{x};\theta) = \exp\left\{\frac{y\theta(\mathbf{x})-b(\theta(\mathbf{x}))}{\phi} + c(y,\phi)\right\}$ with $g(\mu(\mathbf{x})) = \mathbf{x}^T\beta$

can. param. ↑ θ ; disp. para ↑ ϕ

Regression: $\mu(\mathbf{x}) \equiv E(Y|\mathbf{x}) = b'(\theta(\mathbf{x}))$ (fact)

Canonial link: Takes $g(\mu) = b'^{-1}(\mu)$. $\iff$ to assume $\theta(\mathbf{x}) = \mathbf{x}^T\beta$

★Normal: $g(\mu) = \mu$ ★Bernoulli: $g(p) = \log\frac{p}{1-p}$ = logit link

Poisson Distribution and MLE

Poisson Dist.: $P(Y = y|\mathbf{X} = \mathbf{x}) = \frac{\lambda(\mathbf{x})^y \exp(-\lambda(\mathbf{x}))}{y!}$

$$= \exp(y \underbrace{\log \lambda(\mathbf{x})}_{\theta(\mathbf{x})} - \underbrace{\lambda(\mathbf{x})}_{b(\theta(\mathbf{x}))} \underbrace{- \log y!}_{c(y,\phi)}).$$

with $b(\theta) = \lambda = \exp(\theta)$, $c(y, \phi) = -\log y!$, with $\phi = 1$.

Likelihood: $\ell_n(\beta) = \sum_{i=1}^n \log f(y_i|\mathbf{x}_i) \propto \sum_{i=1}^n [y_i\theta_i - b(\theta_i)]$, $\theta_i = \mathbf{x}_i^T\beta$.

Estimated Variance: $\widehat{\text{var}}(\widehat{\beta}) = -[\ell_n''(\widehat{\beta})]^{-1} = \phi[\sum_{i=1}^n b''(\theta_i)\mathbf{x}_i\mathbf{x}_i^T]^{-1}$, $\widehat{\theta}_i = \mathbf{x}_i^T\widehat{\beta}$

★Extend RSS and residuals to deviance and deviance residuals (omit here).

★Model building arts (nonlinearity, interactions, kernels) continue to apply