Ch.7 Alternative modeling of binary response data

Jaehyoung Hong

What we will study?

- Probit model
- Bayesian approach
- Conditional likelihood
- Kernel smoothing

Probit models and related latent variable model

- Alternative to logistic models for binary response
  : $g[\pi(x)] = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$ with logit g

- Probit model $\quad \Phi^{-1}[\pi(x)] = \alpha + \beta x$ with standard normal cdf Φ

- Threshold model

$$\pi(x) = P(Y = 1) = P(Y^* > \tau) = P(\alpha + \beta x + \epsilon > \tau)$$
$$= P(-\epsilon < \alpha + \beta x - \tau) = \Phi[(\alpha + \beta x - \tau)/\sigma]$$

With $\{\epsilon_i\}$ are independent from $N(0, \sigma^2)$

- Utility model

$$\pi(x) = P(Y = 1) = P(\alpha_1 + \beta_1 x_1 + \epsilon_1 > \alpha_0 + \beta_0 x_0 + \epsilon_0)$$
$$= P\left((\epsilon_0 - \epsilon_1)/\sqrt{2} < [(\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)x]/\sqrt{2}\right) = \Phi[(\alpha^* + \beta^* x]$$

With ϵ_0, ϵ_1 are independent from $N(0,1)$

Probit model fitting

- Let y_i : #(success) out of n_i trials at setting $\boldsymbol{x_i}$ for $i = 1, \cdots, N$ with predictor x_{ij}, probit model $\Phi^{-1}[\pi(\boldsymbol{x_i})] = \sum_j \beta_j x_{ij}$

$$L(\beta) = \log\{\prod_{i=1}^{N}\left[\Phi\left(\sum_j \beta_j x_{ij}\right)\right]^{y_i}\left[1-\Phi\left(\sum_j \beta_j x_{ij}\right)\right]^{n_i-y_i}\}$$

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_i \frac{n_i\left[y_i - \Phi\left(\sum_j \beta_j x_{ij}\right)\right]x_{ij}}{\Phi\left(\sum_j \beta_j x_{ij}\right)\left[1-\Phi\left(\sum_j \beta_j x_{ij}\right)\right]}\phi\left(\sum_j \beta_j x_{ij}\right) = 0$$

Fisher scoring

$$\widehat{cov}\left(\hat{\beta}\right) = \left(X^T \widehat{W} X\right)^{-1}$$

$$with\ \widehat{w}_i = n_i\left[\phi\left(\sum_j \hat{\beta}_j x_{ij}\right)\right]^2 \Big/ \{\Phi\left(\sum_j \hat{\beta}_j x_{ij}\right)[1-\Phi\left(\sum_j \hat{\beta}_j x_{ij}\right)]\}$$

Complementary Log-Log link models

- Logit and probit links are symmetric about 0.5

$$logit[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = -\log\left[\frac{1-\pi(x)}{\pi(x)}\right] = -\text{logit}[1-\pi(x)]$$

  - Such symmetric is violated, logit and probit model fitting is poor

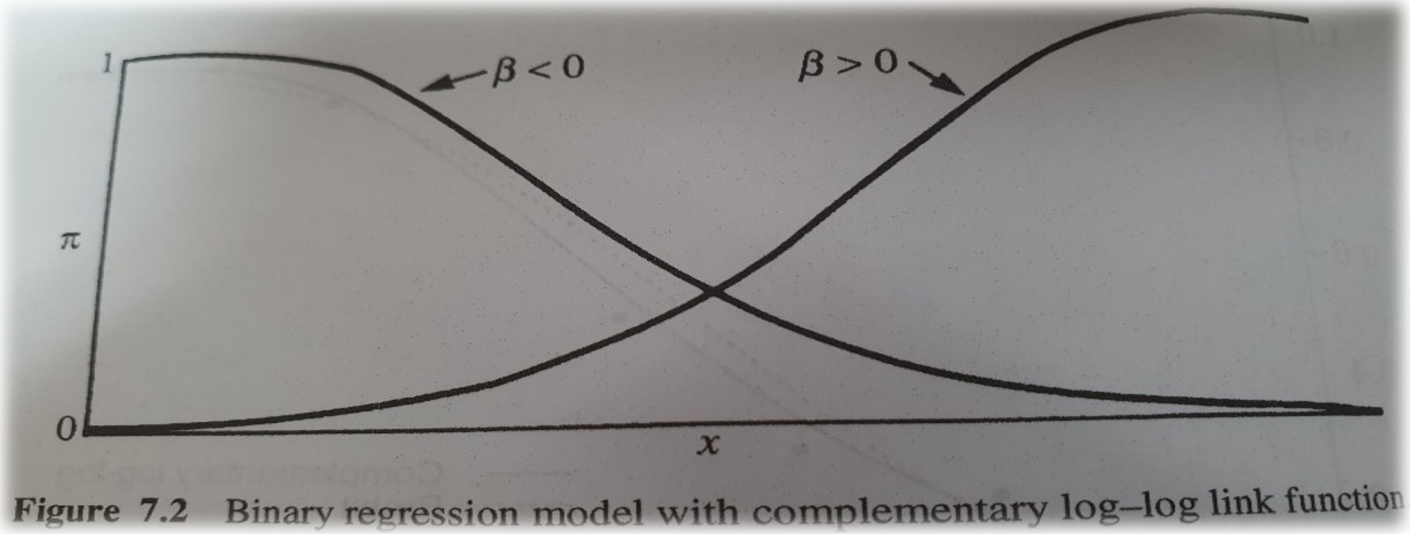


Figure 7.2 Binary regression model with complementary log–log link function

asymmetric

Complementary Log-Log link models

- Complementary log-log link

$$\log[-\log(1-\pi(x))] = \alpha + \beta x$$

- Interpretation of Complementary log-log link

$$\log[-\log(1-\pi(x_2))] - \log[-\log(1-\pi(x_1))] = \beta(x_2 - x_1)$$

$$\frac{\log[1-\pi(x_2)]}{\log[1-\pi(x_1)]} = \exp[\beta(x_2 - x_1)]$$

$$1-\pi(x_2) = [1-\pi(x_1)]^{\exp[\beta(x_2-x_1)]}$$

Fitting result with probit and comp. Log-Log

Table 7.1 Beetles Killed After Exposure to Carbon Disulfide

Log Dose	Number of Beetles	Number Killed	Fitted Values: Comp. Log–Log	Fitted Values: Probit	Fitted Values: Logit
1.6907	59	6	5.6	3.4	3.5
1.7242	60	13	11.3	10.7	9.8
1.7552	62	18	21.0	23.5	22.5
1.7842	56	28	30.4	33.8	33.9
1.8113	63	52	47.8	49.6	50.1
1.8369	59	53	54.1	53.3	53.3
1.8610	62	61	61.1	59.7	59.2
1.8839	60	60	59.9	59.2	58.7

Source: Data reprinted with permission from Bliss (1935).

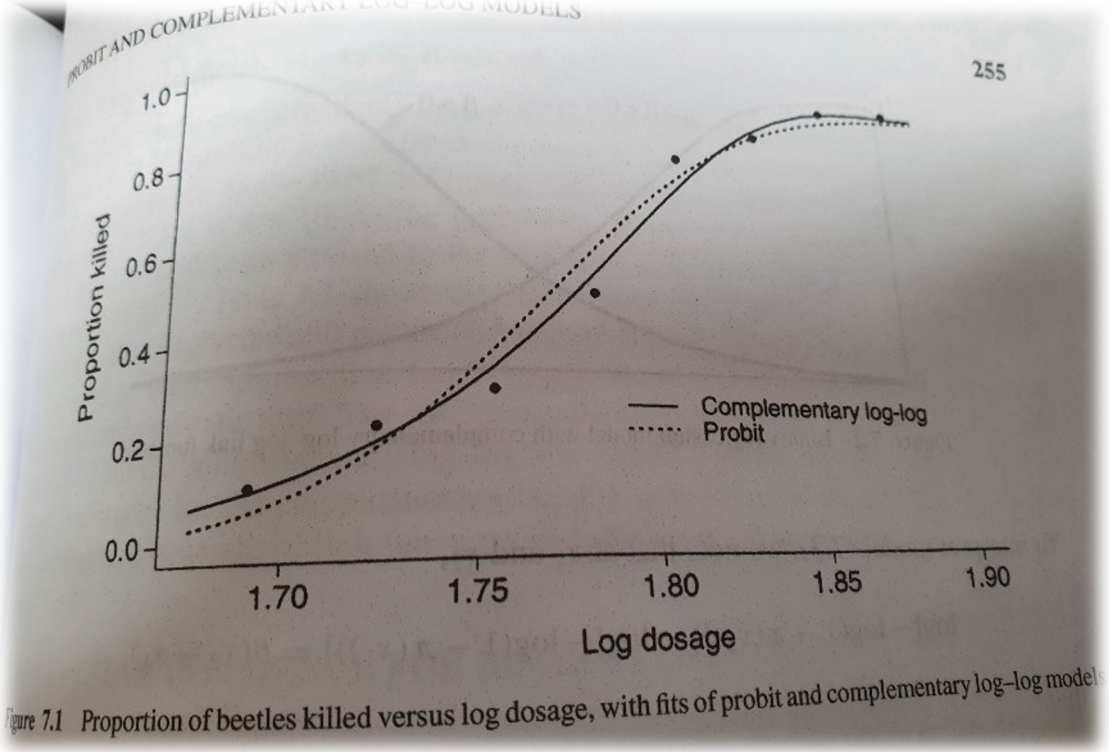


Figure 7.1 Proportion of beetles killed versus log dosage, with fits of probit and complementary log–log models

Prior specifications for binary regression models

- Improper prior : not integrating out to 1

→ Posterior can also be improper so that MCMC cannot work

- Simple prior : $N(0, \sigma^2)$ prior for each parameter

→ Using standardization we can compare the effect (good interpretation)

- To choose hyper parameter of prior

1. Construct prior distribution on the probability scale rather than a link function scale

2. Jeffreys prior : pdf proportional to $|\mathcal{J}|^{0.5}$
   : Invariance to the parametrization + good properties (proper / symmetric and unimodal at 0 / posterior have thinner tails than t-distribution / can used for comp.log-log)

Potential impact of the choice of prior

Table 7.2 Part of Endometrial Cancer Data Set[a]

HG	NV	PI	EH	HG	NV	PI	EH	HG	NV	PI	EH
0	0	13	1.64	0	0	16	2.26	0	0	8	3.14
...											
1	1	21	0.98	1	0	5	0.35	1	1	19	1.02

[a] HG = histology grade, NV = neovasculation, PI = pulsatility index, EH = endometrium height.
Source: Data courtesy of Michael Schemper and Georg Heinze. Complete data ($n = 79$) at `www.stat.ufl.edu/~aa/cda/cda.html`.

- 79 case

$$logit[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- For all $x_1 = 1 \rightarrow y = 1$ (Very important)

Table 7.3 Results of Fitting Models to Cancer Data Set of Table 7.2[a]

Analysis	$\hat{\beta}_1$	SD	Interval	$\hat{\beta}_2$	SD	Interval	$\hat{\beta}_3$	SD	Interval
ML	∞	—	$(1.3, \infty)$	−0.42	0.44	(−1.4, 0.4)	−1.92	0.56	(−3.2, −1.0)
Bayes, $\sigma = 10$	9.12	5.10	(2.1, 21.3)	−0.47	0.45	(−1.4, 0.4)	−2.14	0.59	(−3.4, −1.1
Bayes, $\sigma = 1$	1.65	0.69	(0.3, 3.0)	−0.22	0.33	(−0.9, 0.4)	−1.77	0.43	(−2.7, −1.(

[a] Interval is profile likelihood interval for ML and equal-tail posterior interval for Bayes.

- Strong belief that the effects are not extremely strong

Potential impact of the choice of prior

Table 7.4 Bayesian and ML Fit of Logistic Regression Model for Trauma Data

	Bayesian, Beta Priors		Bayesian, Normal Priors		Frequentist ML	
Variable	Mean	Std. dev.	Mean	Std. dev.	Estimate	SE
Intercept	−1.79	1.10	−2.02	1.57	−2.061	1.526
Injury score	0.07	0.02	0.09	0.03	0.083	0.028
Trauma score	−0.60	0.14	−0.60	0.18	−0.553	0.171
Age	0.05	0.01	0.05	0.02	0.051	0.017
Injury type	1.10	1.06	1.44	1.41	1.338	1.334
Age × Injury type	−0.02	0.03	−0.01	0.03	−0.005	0.032

Source: Results with beta priors based on Table 2 in Bedrick et al. (1997).

$$logit[P(Y=1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 (x_3 x_4)$$

Using six prior distributions for different setting (different surgeon)

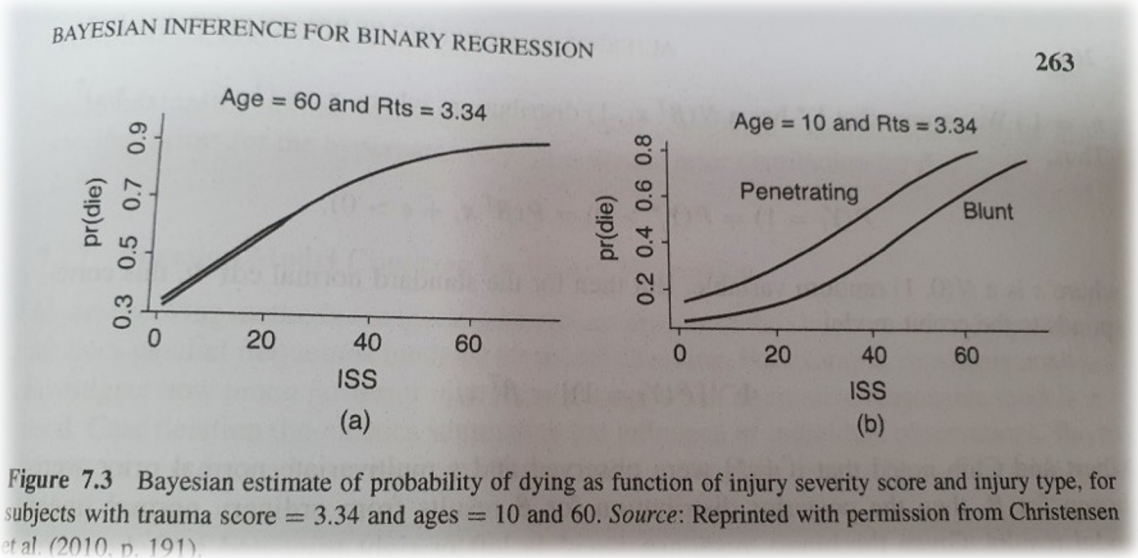


Figure 7.3 Bayesian estimate of probability of dying as function of injury severity score and injury type, for subjects with trauma score = 3.34 and ages = 10 and 60. *Source*: Reprinted with permission from Christensen et al. (2010, p. 191).

- Bayes factor (BF) = $p(y|M_1)/p(y|M_2)$

$where, p(y|M)$ is obtained by integrating the likelihood function for that model with respect to the prior on β.

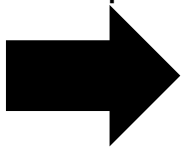
Conditional likelihood eliminate nuisance parameters by conditioning on their sufficient statistics

- When $\#(sample) = n$ is small or $\#(parameters)$ grows as n does

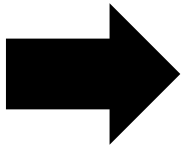
- Model for binary response y_i with subject i and predictor x_{ij} :

$$P(Y_i = y_i) = \frac{\exp[y_i(\alpha + \sum_{j=1}^{p} \beta_j x_{ij})]}{1 + \exp(\alpha + \sum_{j=1}^{p} \beta_j x_{ij})}$$

N-indep obs



$$P(Y_1 = y_1, \cdots, Y_N = y_N) = \frac{\exp[(\sum_i y_i)\alpha + \sum_{j=1}^{p}(\sum_i y_i x_{ij})\beta_j]}{\prod_i [1 + \exp(\alpha + \sum_{j=1}^{p} \beta_j x_{ij})]}$$



Sufficient statistics for α is $\sum_i y_i$, the sufficient statistics for β_j is $\sum_i y_i x_{ij}$

Conditional likelihood eliminate nuisance parameters by conditioning on their sufficient statistics

- We will eliminate the effect of parameter α

- Sufficient statistics for α is $\sum_i y_i$, the sufficient statistics for β_j is $\sum_i y_i x_{ij}$

Denote $S(t) = \{(y_1^*, \cdots, y_N^*) : \sum_i y_i^* = t\}$

Conditional likelihood

$$P(Y_1 = y_1, \cdots, Y_N = y_N | \sum_i y_i = t) = \frac{P(Y_1 = y_1, \cdots, Y_N = y_N)}{\sum_{S(t)} P(Y_1 = y_1^*, \cdots, Y_N = y_N^*)}$$

$$= \frac{\exp[t\alpha + \sum_{j=1}^{p} (\sum_i y_i x_{ij}) \beta_j] / \prod_i [1 + \exp(\alpha + \sum_{j=1}^{p} \beta_j x_{ij})]}{\sum_{S(t)} \exp[t\alpha + \sum_{j=1}^{p} (\sum_i y_i^* x_{ij}) \beta_j] / \prod_i [1 + \exp(\alpha + \sum_{j=1}^{p} \beta_j x_{ij})]}$$

$$= \frac{\exp[\sum_{j=1}^{p} (\sum_i y_i x_{ij}) \beta_j]}{\sum_{S(t)} \exp[\sum_{j=1}^{p} (\sum_i y_i^* x_{ij}) \beta_j]}$$

Not depends on α

We can perform “exact” inference using conditional likelihood

- To infer parameter β_p, We will eliminate the effect of parameter $\beta_1, \cdots, \beta_{p-1}$

- The sufficient statistics for β_j is $\sum_i y_i x_{ij}$

Denote $T_j = \sum_i y_i x_{ij}\,, j = 0, \cdots, p-1\ (x_{i0} = 1)$

$$S(t_0, \cdots, t_{p-1}) = \{(y_1^*, \cdots, y_N^*) \colon \sum_i y_i^* x_{ij} = t_j, j = 0, \cdots, p-1\}$$

$$P(Y_1 = y_1, \cdots, Y_N = y_N | T_j = t_j, j = 0, \cdots, p-1)$$

$$= \frac{\exp[(\sum_i y_i x_{ip})\beta_p]}{\sum_{S(t_0,\cdots,t_{p-1})} \exp[(\sum_i y_i^* x_{ip})\beta_p]} = \frac{\exp(t_p \beta_p)}{\sum_{S(t_0,\cdots,t_{p-1})} \exp(t_p^* \beta_p)}$$

Depends only on β_p

We can perform "exact" inference using conditional likelihood

- Inference for β_p uses the conditional distribution of its sufficient statistic, $T_p = \sum_i y_i x_{ip}$, given others.

- Let $c(t_0, \cdots, t_{p-1}, t)$: #(data vectors in $S(t_0, \cdots, t_{p-1})$ for which $T_p = t$

$$\Rightarrow \quad P(T_p = t | T_j = t_j, j = 0, \cdots, p-1) = \frac{c(t_0, \cdots, t_{p-1}, t)\exp(t\beta_p)}{\sum_u c(t_0, \cdots, t_{p-1}, u)\exp(u\beta_p)}$$

- For $H_a: \beta_p > 0$ and observed $T_p = t_{obs}$, the exact conditional P-value is

$$\Rightarrow \quad \sum_{t \geq t_{obs}} P(T_p = t | T_j = t_j, j = 0, \cdots, p-1) = \frac{\sum_{t \geq t\_obs} c(t_0, \cdots, t_{p-1}, t)}{\sum_u c(t_0, \cdots, t_{p-1}, u)}$$

This is the proportion of data configurations in the conditional set for which the sufficient static for β_p is at least as large as observed

Small-sample conditional inference for 2×2 tables

- Model : $logit[P(Y_i = 1)] = \alpha + \beta x_i, i = 1, \cdots, N$

	$y_i = 1$	$y_i = 0$	
$x_i = 1$	t		n_1
$x_i = 0$	s		n_2

- To eliminate α, we condition on $\sum_i y_i = t + s$

$$f(t|t+s; n_1, n_2, \beta) = \frac{\binom{n_1}{t}\binom{n_2}{s}e^{\beta s_1}}{\sum_{u=m_-}^{m_+}\binom{n_1}{u}\binom{n_2}{s+t-u}e^{\beta u}}$$

$$where\ m_- \le t \le m_+\ with\ m_- = \max(0, n_{1+} + n_{+1} - n)\ \&\ m_+ = \min(n_{1+}, n_{+1})$$

$c(t_0, t) = \binom{n_1}{t}\binom{N - n_1}{t_0 - t}$ and $t_0 = t + s$

/ The resulting exact conditional test that $\beta = 0$ is Fisher's exact test for 2×2 tables

Small-sample conditional inference for $\mathbf{I}\times\mathbf{2}$ tables

- Model : $logit[P(Y_i = 1)] = \alpha + \beta x_i, i = 1, \cdots, N$

The data $\{y_i\}_{i=1}^{I} \sim \{bin(n_i, \pi_i)\}$ with fixed row counts $\{n_i\}$

- To infer β, use $T = \sum_i x_i y_i$ eliminate α by conditioning on $\sum_i y_i = t + s$

$$P(T = t | \sum_i y_i = t_0; \beta) = \frac{c(t_0, t)e^{\beta t}}{\sum_u c(t_0, u)e^{\beta u}}$$

where $c(t_0, t)$ =the sum of $[\prod_i \binom{n_i}{y_i}]$ for all tables with the given marginal totals that $T = u$

When $\beta = 0$, the cell counts have distribution that is a special case of multivariate hypergeometric distribution

To test $H_0: \beta = 0$, Cochran-Armitage statistic

Ways of smoothing categorical data, mainly in the context of analyzing a binary response variable

- **Smoothing method**
  - Non-parametric fashion → Model 설정 miss로 인한 오류↓
  - Overfitting↑

- **Model based method vs other method**
  - Model : Bias↑ & Variance↓
  - Others : Bias↓ & Variance↑

Variance / Bias trade-off

Kernel estimation for binary response data

- Kernel estimation : Completely non-model based
  + to estimate mean at particular point, also use other points

- Estimating $\boldsymbol{\pi}$ joint cell probability in a multiway contingency table by smoothing the sample cell proportions $\boldsymbol{p}$

Kernel estimates of $\tilde{\boldsymbol{\pi}} = \boldsymbol{K}\boldsymbol{p}$.

➡ $where\ \boldsymbol{K}$ square matrix with nonnegative elements having column totals 1

$$\tilde{\pi}_i = (1-\lambda)p_i + \lambda(smoother_i)$$

➡ $where\ \lambda : constant\ of\ the\ degree\ of\ smoothing$ (if $\lambda \uparrow$, more smoothing)

Kernel estimation for binary response data

- For binary response data, let $\phi(\cdot)$: symmetric unimodal $kernel\ function$, kernel smoothed estimate of $P(Y=1|X=x)$ is

$$\tilde{\pi}(x) = \frac{\sum_i y_i \phi[\frac{x-x_i}{\lambda}]}{\sum_i \phi[\frac{x-x_i}{\lambda}]}$$

$where\ \lambda > 0$ is a smoothing parameter

- Effect of ϕ

If $\phi(\mathrm{u}) = 1\ (\text{when } \mathrm{u} = 0), \phi(u) = 0\ (o.w.)$

$\tilde{\pi}(x_k)$: sample proportion at $x = x_k$ (no smoothing)

- Effect of λ

The choice of λ is more important than the choice of ϕ.

If $\lambda \to 0$: no smoothing (local)

If $\lambda \to \infty$: overall sample proportion (global)

Effect of λ

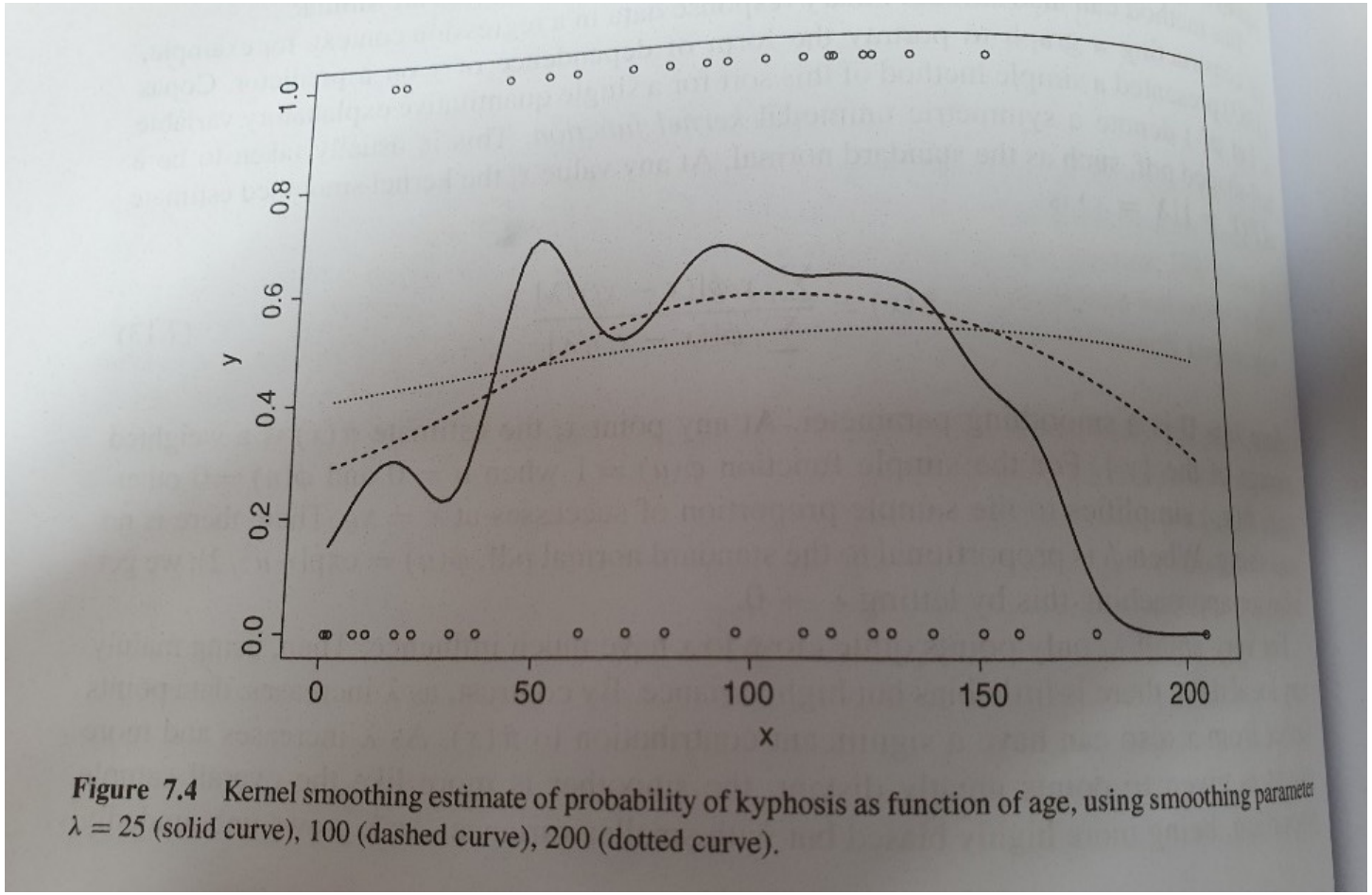


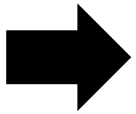
Figure 7.4 Kernel smoothing estimate of probability of kyphosis as function of age, using smoothing parameter $\lambda = 25$ (solid curve), 100 (dashed curve), 200 (dotted curve).

- Binary outcome y
  (1=kyphosis present , 0=absent)
- x : Age in months

Nearest Neighbors Smoothing

- Nearest neighbors smoothing : Weighted average of observations for k subjects

- Estimate the probability $\pi_i = P(Y_i = 1)$ for subject i, the similarity measure s_{ij}



$$\widehat{\pi_i} = \frac{\sum_{j \in N(i)} s_{ij} y_j}{\sum_{j \in N(i)} s_{ij}}$$

$where\ N(i)$: set of k subjects who are the nearest neighbors

- k : usually determined by cross-validation
- Simple to implement & Decision boundary is quite simple for standard binary regression models
- When the number of explanatory variable is large or highly correlated with each other or some of them is qualitative, decision of measure is not trivial

Smoothing using penalized likelihood estimation

- Penalized likelihood estimation (Ridge / Lasso / Elastic net)
  : Can consider probability distribution of Y or dependence among variables

- Model with generic parameter $\boldsymbol{\beta}$ and log-likelihood function $L(\boldsymbol{\beta})$. The penalized likelihood estimator of $\boldsymbol{\beta}$ maximizes

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) - \lambda(\boldsymbol{\beta})$$

$where, \lambda(\cdot)$: roughness penalty (As λ increase β become more smoother)

- λ : Cross-validation
- $\lambda(\beta) = \lambda \sum_j \beta_j^2$: Ridge / $\lambda(\beta) = \lambda \sum_j |\beta_j|$: Lasso / $\lambda(\beta) = \lambda \sum_j |\beta_j| + (1-\lambda) \sum_j \beta_j^2$: Elastic net
- Advantage1 : If a large number of explanatory variable has no effect, $\widehat{\beta_j}$ is larger than true value without shrinkage
- Advantage2 : Forward / Backward variable selection method are discrete, shrinkage method is continuous

Generalized additive models : Approach to generalized GLM

- GAM : Replace the linear predictor of GLM by additive smooth functions of the predictors

- The GLM structure of $g(\mu_i) = \sum_j \beta_j x_{ij}$ then generalizes to

$$g(\mu_i) = \sum_j s_j(x_{ij})$$

$where, s(\cdot)$: smooth function of predictor j

- $s(\cdot)$: Such as *cubic spline*
- $\widehat{s(\cdot)}$: Given by *backfitting algorithm* which is generalization of the Newton-Raphson method
- Choosing df of each s_j determines how smooth the resulting GAM fit looks.
- Cons : Loss of interpretability for describing effect

Advantage and disadvantage of smoothing method

- Pros 1. GAM & Penalized likelihood : Ordinary GLM is a special case

- Cons 1. All smoothing method need to choose the degree of smoothing (extra budget)
  : Unlike Bayesian approach such choosing has little theoretical basis