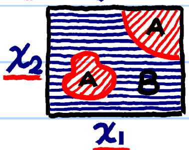
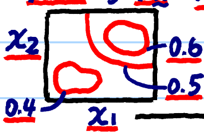
- Prediction of future binary observation $z_{\mathbf{x}_0}$ and classification

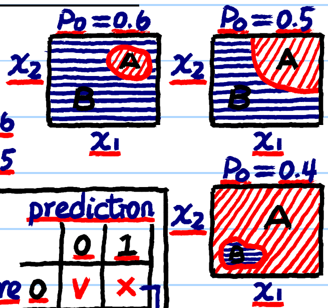
2 approaches ① z: response, $\mathbf{x}$: covariates ② $\mathbf{x}$: response, z: covariate (e.g., Fisher's LDA)

Q: What is a classification problem for binary data? → 2 categories: A, B → 1, 0 ← Bernoulli



contour plot of $\hat{p}_{\mathbf{x}}$





➢ The prediction of $z_{\mathbf{x}_0}$ is based on whether under the fit $\hat{p}_{\mathbf{x}_0}$ at $\mathbf{x}_0$ is greater (then, predict $z_{\mathbf{x}_0}$=1) or less (then, predict $z_{\mathbf{x}_0}$=0) than a given p_0

future z \ prediction	0	1
0	V	× (loss1)
1	× (loss2)	V

- p_0 often chosen to equal 0.5 ← symmetric loss
  - i.e., choose the category with larger predicted probability
  - ▪ for cases where the losses due to misclassification are not symmetrical, p_0 other than 0.5 should be used, such as disease diagnosis, credit scoring

- loss1 = loss2 ⇒ $p_0=0.5$
- loss1 > loss2 ⇒ tend to predict $z_{\mathbf{x}_0}=0$ ⇒ $p_0>0.5$
- loss1 < loss2 ⇒ tend to predict $z_{\mathbf{x}_0}=1$ ⇒ $p_0<0.5$

p_0 determined by decision Thm

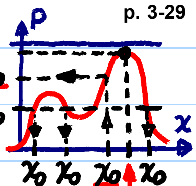
Q: Why no such issue in LM? Ans. LM uses function of $|y_i-\hat{y}_i|$ as loss. symmetric | asymmetric case: larger loss if $\hat{y}_i>y_i$, smaller o.w.

➢ In contrast to the linear model situation, there is no distinction possible between prediction interval (P.I.) for a future observation z and confidence interval (C.I.) for the mean response (check LN p.1-18)

possible P.I. for future z: {0}, {1}, {0,1} ← useless | need to consider 2 sources of variation: (a) in $\hat{p}_{\mathbf{x}_0}$ (obs. data) (b) in $B(n, p_{\mathbf{x}_0})$ (future $y_{\mathbf{x}_0}$) and (c) non-constant variance (d) discrete nature in $y_{\mathbf{x}_0}$

Q: How about P.I. for a future binomial observation $y_{\mathbf{x}_0}$? → complicate

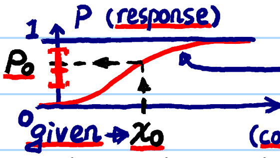
- Prediction of effective dose (an inverse problem): unique



multiple solutions

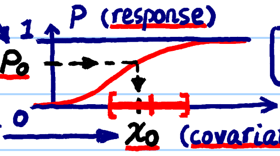
➢ **Q**: What is an inverse problem? Estimate/predict the set of covariate values $\mathbf{x}$ that meet certain condition. (e.g., settings that produce good product)

prediction ↔ parameter ↔ confidence interval



containing parameters β ← estimation

given confidence interval → prediction parameter



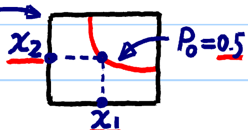
Find optimal setting

➢ When there is a single covariate or when other covariates are held fixed, we sometimes wish to estimate the value of $\mathbf{x}_{p_0}$ corresponding to a given p_0, i.e., find $\mathbf{x}_{p_0}$ such that $\eta_{\mathbf{x}_{p_0}}=g(p_0)$

$\{\mathbf{x}: p_0 = g^{-1}(\eta_{\mathbf{x}}) = g^{-1}(\mathbf{x}^T\beta)\}$

➢ For example, determine which dose $\mathbf{x}_{p_0}$ will lead to a probability of success p_0

How is it related to and different from classification?



$\text{logit}(p_x) = \eta_x = \beta_0+\beta_1 x$

- ▪ ED50 (effective dose, e.g., for drug) or LD50 (lethal dose, e.g., for toxin): p_0=1/2
- For a logit link with $\eta_x=\beta_0+\beta_1 x$, $\hat{x}_{p_0} = (\text{logit}(p_0) - \hat{\beta}_0)/\hat{\beta}_1$ ← MLE
  - a parameter $x_{p_0} = \dfrac{\text{logit}(p_0)-\beta_0}{\beta_1}$
- ▪ Set p_0=1/2 (logit(p_0)=0) and solve for x_{p_0} ⇒ $\hat{x}_{p_0} = -\hat{\beta}_0/\hat{\beta}_1$
- ▪ To determine its standard error, we can use δ-method: $\gamma(\beta_1, \beta_2, p_0)$

$$var(\gamma(\hat{\beta})) \approx \gamma'(\hat{\beta})^T \; var(\hat{\beta}) \; \gamma'(\hat{\beta})$$

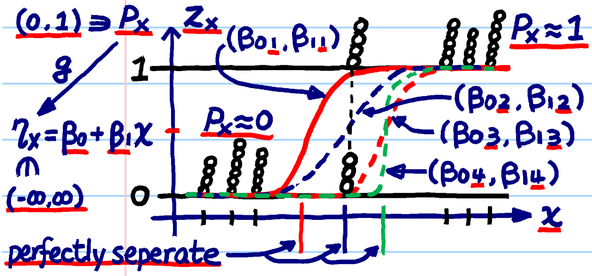
$var(\hat{\beta}) = (X^T\hat{W}X)^{-1}$

confidence interval of x_{p_0}: $\{x;\ p_0 \in [\underline{p}_x, \bar{p}_x]\}$, where $[\underline{p}_x, \bar{p}_x]$ is the C.I. of p at x

❖ **Reading**: Faraway (2006, 1[st] ed.), 2.10

Estimation Problems → check Hauck-Donner effect (LNp 3-12)

- **Q**: If we observe a dataset that has all 0 at lower doses and all 1 at high doses, (1) what does it imply? (2) is it a good thing?



- For (1), clearly, the effect of x (dose) on P_x is very significant (can be identified even without analysis)
- For (2), but cannot determine a curve (i.e., a fit) for P_x (∵ upper & lower bounds in P_x) ⇒ $\hat{\beta}$ very unstable (large $se(\hat{\beta})$) (cf., unidentifiable & strong collinear problems in LM)

Recall. deviance: $y_i \overset{cf.}{\longleftrightarrow} \hat{y}_i$, all the curves perform perfect in deviance.

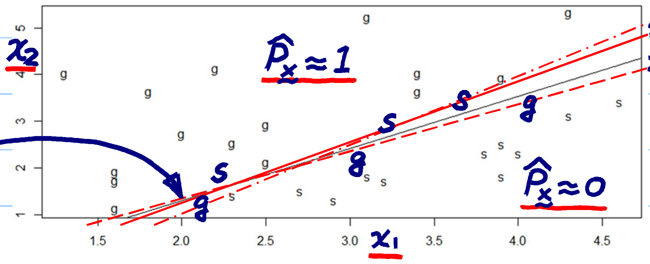
- The IRWLS algorithm for MLE (future lecture) usually converge (approximate) to the MLE fast — usually within 10 iterations — e.g., log-likelihood $\ell(\beta)$ is very flat around MLE

$p=e^{\eta}$ e.g., non-standard links

- However, difficulties can sometimes arise ⇒ convergence fail

⊙ (from Lab) The following is a condition that cause the failure ⇒ the two groups are *linearly separable* ← 2 groups of data can be perfectly seperated by a linear function of x ($\eta=\beta_0+\beta_1x_1+\beta_2x_2$)

$z_x = \begin{cases} g, & \text{prob. } P_x \\ s, & \text{prob. } 1-P_x \end{cases}$ ← 2 categories

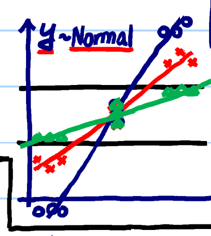


Prediction of an inverse problem based on a fit

each lines corresponds to an estimate of β → infinitely many lines (i.e., $\hat{\beta}$) can perfectly seperate the 2 groups

- ➢ Deviance would be small ← $\hat{P}_i \approx y_i/n_i$ ⇒ $\hat{y}_i \approx y_i$ ⇒ very good fitting → goodness-of-fit test not rejected
- ➢ Estimation of β is very unstable ⇒ large standard error ← ∵ close to unidentifiable ⇒ many good solutions of β
  ⇒ Wald test insignificant ← ∵ $\hat{\beta}_i/se(\hat{\beta}_i)$ small (But, deviance-based test for $\beta_i=0$ still valid) Why? Consider the example in LNp 3-30 Ω: $\eta=\beta_0+\beta_1x_1+\beta_2x_2$. ω_1: $\eta=\beta_0+\beta_2x_2$ ← horizontal cut; ω_2: $\eta=\beta_0+\beta_1x_1$ ← vertical cut — Both not good fits ⇒ large deviance
- ➢ This is an "embarrassment of riches"
  ⇒ Perfect fit is possible, ($y_i \approx \hat{y}_i$ on observed data)
  but estimation is a problem & Wald test
- ➢ Lesson for data collection ⇒ should get y_x on some x's where $y_x/n_x \neq 0$ or 1. (**Q**: why there is no such issue in Normal y?)

might be impossible if have no idea about the value of true β



or 0.1 mixed around the boundary. In classification, data closer to the boundary ⇒ more information.

- Alternative fitting approaches — use conditional likelihood, conditioned on the suff. stat. of nuisance parameters, e.g., other β_j's, $j\neq i$ (check Table 3.9, LNp 2-11)
  - ➢ exact logistic regression (Cox, 1970; Mehta and Patel, 1995)
  - ➢ Bias reduction (BR) method of Firth (1993): remove the $O(n^{-1})$ term from the asymptotic bias of $\hat{\beta}_{\text{MLE}}$ (check Lab; cf. asymptotic dist. in test & C.I.)

Although $\lim E(\hat{\beta}_{MLE})=\beta$, $E(\hat{\beta}_{MLE})\neq\beta$ in finite sample. bias $=E(\hat{\beta}_{MLE}-\beta)\sim O(n^{-1})$, $E(\hat{\beta}_{BR}-\beta)\sim O(n^{-2})$

- Instability in parameter estimation will also occur in datasets that approach linear separability

Recall. unidentifiability $\overset{cf.}{\longleftrightarrow}$ strong collinearity

❖ **Reading**: Faraway (1st ed.), 2.8