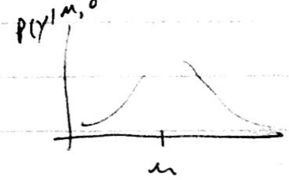
1

Maximum Likelihood

Suppose $y \sim N(\mu, \sigma^2)$

$$p(y \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$



Suppose $\underset{\sim}{y} \sim N(\underset{\sim}{\mu}, \sigma^2 I)$

$$p(\underset{\sim}{y} \mid \underset{\sim}{\mu}, \sigma^2) = p(y_1 \mid \mu_1, \sigma^2) \times \cdots \times p(y_n \mid \mu_n, \sigma^2)$$

$$= \prod_{i=1}^{n} p(y_i \mid \mu_i, \sigma^2)$$

$$= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(y_i-\mu_i)^2}$$

$$= \underline{(2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum (y_i - \mu_i)^2}}$$

Suppose $\underset{\sim}{\mu} = \underset{n\times p}{X}\underset{p\times 1}{\beta}$ for some unknown β. $(\mu_i = x_i^T \beta)$

$$\underline{p(\underset{\sim}{y} \mid \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum (y_i - x_i^T\beta)^2}}$$

Now observe $\underset{\sim}{y}$. What value of $\underset{\sim}{\beta}$ maximizes the probability of the observed data?

$$\max_{\beta} p(\underset{\sim}{y} \mid \beta, \sigma^2) = \max_{\beta} \log p(y \mid \beta, \sigma^2)$$

$$= \max_{\beta} -\frac{1}{2\sigma^2}\sum (y_i - x_i^T\beta)^2$$

$$= \min_{\beta} \sum (y_i - x_i^T\beta)^2$$

Terminology:

* the pdf evaluated at the obs. data, as a function of (β, σ^2) is called the likelihood function for β, σ^2

$$L(\beta, \sigma^2 : y) = p(y \mid \beta, \sigma^2)$$

$$\ell(\beta, \sigma^2 : y) = \log p(y \mid \beta, \sigma^2) \quad \text{(log likelihood function)}$$

* The maximizer of the likelihood (or log likelihood) is called the max. lik. estimator (MLE).

* for linear regression, $\hat{\beta}_{MLE} = \hat{\beta}_{OLS}$

* for linear regression, $\hat{\sigma}^2_{MLE} = \sum (y_i - x_i^T \hat{\beta}_{MLE})^2 / n$

$$= \frac{n-1}{n} \hat{\sigma}^2_{unbiased} < \hat{\sigma}^2_{unbiased}$$

Maximum Likelihood for Logistic Regression

Let $Y \sim \text{binary}(\mu) \quad (E[Y \mid \mu] = \mu)$

$$\left.\begin{array}{l} \Pr(Y=0) = 1-\mu \\ \Pr(Y=1) = \mu \end{array}\right\} \Rightarrow p(y \mid \mu) = \mu^y (1-\mu)^{1-y}$$

Now suppose $\mu = \dfrac{e^{x^T\beta}}{1+e^{x^T\beta}}$: $p(y \mid \beta) = \left(\dfrac{e^{x^T\beta}}{1+e^{x^T\beta}}\right)^y \left(\dfrac{1}{1+e^{x^T\beta}}\right)^{1-y}$

$$= \frac{e^{x^T\beta y}}{1+e^{x^T\beta}}$$

Now suppose $y_i \sim \text{bin}\left(\dfrac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)$, $y_1 \ldots y_n$ indep.

$$p(y \mid \beta) = \prod_{i=1}^{n} p(y_i \mid \beta) = \prod_{i=1}^{n} \frac{e^{x_i^T\beta y_i}}{1+e^{x_i^T\beta}} = e^{\sum y_i (x_i^T\beta)} \left(\prod (1+e^{x_i^T\beta})\right)^{-1}$$

$$\log p(\underset{\sim}{y} \mid \beta) = \ell(\beta : y) = \underline{\underline{\sum y_i (x_i^T \beta) - \sum \log (1 + e^{x_i^T \beta})}}$$

what value maximizes $\ell(\beta : y)$? What is the MLE?

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum y_i \underset{\sim}{x_i} - \sum \frac{\underset{\sim}{x_i} e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$$

$$= \sum_{i=1}^{n} \underset{\sim}{x_i} \Big(y_i - \underbrace{\frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}}_{E(y_i \mid \beta)!} \Big)$$

MLE satisfies

$$\underbrace{\sum \underset{\sim}{x_i} y_i}_{\text{observed}} = \underbrace{\sum \underset{\sim}{x_i} \frac{e^{x_i^T \hat\beta}}{1 + e^{x_i^T \hat\beta}}}_{\text{fitted}}$$

Compare to linear regression

$$\frac{\partial}{\partial \beta} \sum (y_i - x_i^T \beta)^2 = -2 \sum \underset{\sim}{x_i} (y_i - x_i^T \beta)$$

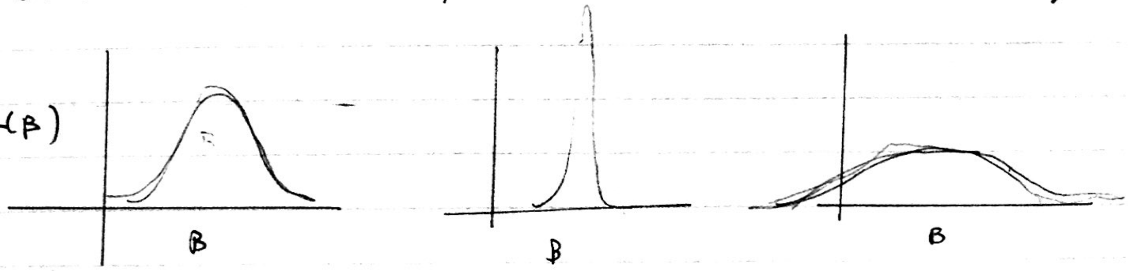
MLE/OLS $\hat\beta$ satisfies $\sum \underset{\sim}{x_i} (y_i - x_i^T \beta) = 0$

$$\underbrace{\sum \underset{\sim}{x_i} y_i}_{\text{observed}} = \underbrace{\sum \underset{\sim}{x_i} (x_i^T \beta)}_{\text{fitted.}}$$

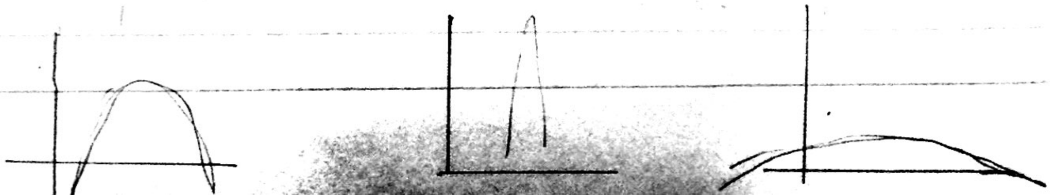
Variances and Standard Errors

$\text{Var}(\hat\beta) \approx$ uncertainty in $\hat\beta$ as estimate of β

$L(\beta)$



$\ell(\beta)$



Certainty in $\hat{\beta} \approx$ "$- \frac{\partial^2}{\partial \beta^2} \ell(\beta)$"

Uncertainty in $\hat{\beta} \approx \left(+ \frac{\partial^2}{\partial \beta^2} \ell(\beta) \right)^{-1}$

Linear Regression:

$$\ell(\beta) = -\frac{n}{2} \log(2\pi\bar{\sigma}^2) - \frac{1}{2\sigma^2} \Sigma (y_i - x_i^T \beta)^2$$

$$\frac{\partial \ell(\beta)}{\partial \beta} = + \frac{1}{2\sigma^2} \left(-2 \Sigma \underset{\sim}{x}_i (y_i - x_i^T \beta) \right)$$

$$= \frac{1}{\sigma^2} \Sigma \underset{\sim}{x}_i (y_i - \underset{\sim}{x}_i^T \beta) = \frac{1}{\sigma^2} \left[\Sigma \underset{\sim}{x}_i y_i - \underset{\sim}{x}_i \underset{\sim}{x}_i^T \beta \right]$$

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = -\frac{1}{\sigma^2} \Sigma \underset{\sim}{x}_i \underset{\sim}{x}_i^T$$

$$-\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = \frac{1}{\sigma^2} X^T X$$

$$\left(-\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \right) = \underline{\underline{(X^T X)^{-1} \sigma^2}} = Var(\hat{\beta}_{OLS})$$

$$Var(\hat{\beta}_{OLS}) = \sigma^2 (X^T X)^{-1}$$
$$\underline{\underline{\hat{Var}(\hat{\beta}_{OLS})}} = \underline{\hat{\sigma}^2 (X^T X)^{-1}}$$

Logistic Regression:

$$\frac{\partial}{\partial \beta} \ell(\beta) : \quad \Sigma \underset{\sim}{x}_i \left(y_i - \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right)$$

$$\frac{\partial^2}{\partial \beta \partial \beta^T} \ell(\beta) : \quad -\Sigma \underset{\sim}{x}_i \left[\frac{(1+e^{\theta_i}) e^{\theta_i} x_i^T - (e^{\theta_i})^2 x_i^T}{(1+e^{\theta_i})^2} \right]$$

$$= -\Sigma \underset{\sim}{x}_i \underset{\sim}{x}_i^T \left[\frac{e^{\theta_i}}{1+e^{\theta_i}} - \frac{(e^{\theta_i})^2}{(1+e^{\theta_i})^2} \right]$$

$$= \underline{-\Sigma \underset{\sim}{x}_i \underset{\sim}{x}_i^T \mu_i (1-\mu_i)} \Rightarrow Var(\hat{\beta}) \approx (X^T D X)^{-1}$$

$$D = diag\left(\left(\frac{e^{\hat{\beta}^T x_i}}{1+e^{\hat{\beta}^T x_i}} \right) \left(\frac{1}{1+e^{\hat{\beta}^T x_i}} \right) \right)$$
$$= diag\left(\hat{\mu}_i (1-\hat{\mu}_i) \right)$$