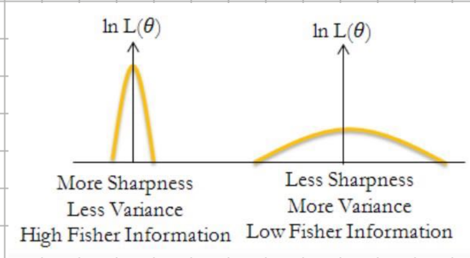
Fisher Information

$X \sim f_\theta$

$$\boxed{I(\theta) = E_\theta\left[\dot{\ell}_{x_1}(\theta)\,\dot{\ell}_{x_1}(\theta)^T\right] = -E_\theta\left[\ddot{\ell}_{x_1}(\theta)\right]}$$

where $\ell_{x_1}(\theta) = \log f_\theta(X_1)$



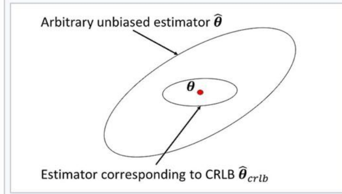


Illustration of the Cramer-Rao bound: there is no unbiased estimator which is able to estimate the (2-dimensional) parameter with less variance than the Cramer-Rao bound, illustrated as standard deviation ellipse.

Theorem: Cramér - Rao

Let $X_1, \dots, X_n \overset{iid}{\sim} f_{\theta_0}$ where

- Domain of f_θ $\mathcal{D}(f_\theta) = \{x \mid f_\theta(x) > 0\}$ does not depend on θ
- f_θ is 3 times continuously differentiable w.r.t θ

Let $\hat{\theta} = g(X_1, \dots, X_n)$ be an unbiased estimator of θ_0

$$\boxed{Var(\hat{\theta}) \geq [nI(\theta_0)]^{-1}}$$

where $I(\theta_0) = E_{\theta_0}\left[\dot{\ell}_{x_1}(\theta_0)\,\dot{\ell}_{x_1}(\theta_0)^T\right] = -E_{\theta_0}\left[\ddot{\ell}_{x_1}(\theta_0)\right]$ and $\ell_{x_1}(\theta) = \log f_\theta(X_1)$

Proof:

$$cor(\hat{\theta}(\vec{X}), \dot{\ell}_{\vec{x}}(\theta)) = \frac{Cov(\hat{\theta}, \dot{\ell}_{\vec{x}}(\theta))}{\sqrt{Var(\hat{\theta})\,Var(\dot{\ell}_{\vec{x}}(\theta))}} \leq 1$$

$$\iff \frac{Cov(\hat{\theta}, \dot{\ell}_{\vec{x}}(\theta))^2}{Var(\dot{\ell}_{\vec{x}}(\theta))} \leq Var(\hat{\theta})$$

$$\begin{aligned}
Var(\dot{\ell}_{\vec{x}}(\theta)) &= Var\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f_\theta(x_i)\right) \\
&= Var\left(\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f_\theta(x_i)\right) \\
&= \sum_{i=1}^n Var\left(\frac{\partial}{\partial\theta}\log f_\theta(x_i)\right) \\
&= n\cdot Var\left(\frac{\partial}{\partial\theta}\log f_\theta(x_i)\right) \\
&= n\cdot Var(\dot{\ell}_{x_1}(\theta)) \\
&= n\cdot\left(E[\dot{\ell}_{x_1}(\theta)^2] - \underbrace{E[\dot{\ell}_{x_1}(\theta)]^2}_{0}\right) \\
&= n\cdot I(\theta)
\end{aligned}$$

$$\begin{aligned}
E[\dot{\ell}_{x_1}(\theta)] &= E\left[\frac{\partial}{\partial\theta}\log f_\theta(x_1)\right] \\
&= \int \frac{\partial}{\partial\theta}(\log f_\theta(x_1))\cdot f_\theta(x_1)\,dx_1 \\
&= \int \frac{\frac{\partial}{\partial\theta}f_\theta(x_1)}{f_\theta(x_1)} f_\theta(x_1)\,dx_1 \\
&= \frac{\partial}{\partial\theta}\int f_\theta(x_1)\,dx_1 \\
&= \frac{\partial}{\partial\theta} 1 \\
&= 0
\end{aligned}$$

$$\begin{aligned}
Cov(\hat{\theta}, \dot{\ell}_{\vec{x}}(\theta)) &= E[\hat{\theta}\cdot\dot{\ell}_{\vec{x}}(\theta)] - E[\hat{\theta}]\cdot\overbrace{E[\dot{\ell}_{\vec{x}}(\theta)]}^{0} \\
&= \int\cdots\int \hat{\theta}\cdot\frac{\partial}{\partial\theta}(\log f_\theta(x_1,\dots,x_n))\cdot f_\theta(x_1,\dots,x_n)\,dx_1\cdots dx_n \\
&= \int\cdots\int \hat{\theta}\,\frac{\frac{\partial}{\partial\theta}f_\theta(x_1,\dots,x_n)}{f_\theta(x_1,\dots,x_n)}\cdot f_\theta(x_1,\dots,x_n)\,dx_1\cdots dx_n \quad \text{smoothness of density function} \\
&= \frac{\partial}{\partial\theta}\int\cdots\int \hat{\theta}\cdot f_\theta(x_1,\dots,x_n)\,dx_1\cdots dx_n \\
&= \frac{\partial}{\partial\theta}E[\hat{\theta}] \quad \hat{\theta}\text{ unbiased} \\
&= \frac{\partial}{\partial\theta}\theta \\
&= 1
\end{aligned}$$

$\square$

Stochastic Convergence

a sequence of R.V. $\{X_n\}$ converges in probability to a R.V. X, shown by $X_n \xrightarrow{P} X$,
if $\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$, for all $\varepsilon > 0$

$\Downarrow \quad \cancel{\Uparrow}$

a sequence of R.V. $\{X_n\}$ converges in distribution to a R.V. X, shown by $X_n \xrightarrow{D} X$,
if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$, for all x at which $F_X(x)$ is continuous

Consistency

an estimator $\hat{\theta}_n$ of θ_0 is consistent, if $\hat{\theta}_n \xrightarrow{P} \theta_0$, i.e. $\hat{\theta}_n$ "converges in probability" to the true value

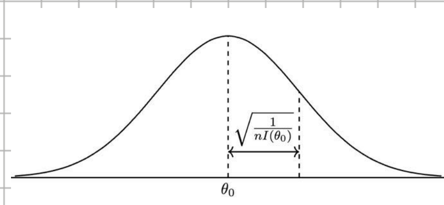
↳ e.g. an unbiased $\hat{\theta}_n$ with $Var(\hat{\theta}_n) \longrightarrow 0$

Efficiency

an estimator $\hat{\theta}_n$ of θ_0 is efficient, if

$$\boxed{\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \overbrace{\mathcal{N}}^{\text{asymptotically normal}}\big(\overbrace{0}^{\text{asymptotically unbiased}}, \overbrace{I(\theta_0)^{-1}}^{\text{asymptotically optimal in CR sense}}\big)}$$

Theorem: Asymptotic efficiency of MLE



Let $X_1, \ldots, X_n \overset{iid}{\sim} f_{\theta_0}$ where

- Domain of f_θ $\mathcal{D}(f_\theta) = \{x \mid f_\theta(x) > 0\}$ does not depend on θ
- f_θ is 3 times continuously differentiable w.r.t θ

Then $\hat{\theta}^{MLE}$ is efficient

Proof:

$$0 = \dot{\ell}_{X_1 \ldots X_n}(\hat{\theta}_n^{MLE}) = \dot{\ell}_{X_1 \ldots X_n}(\theta_0) + (\hat{\theta}_n^{MLE} - \theta_0)\, \ddot{\ell}_{X_1 \ldots X_n}(\theta_0) + \text{"small"}$$

$$\Leftrightarrow \sqrt{n}\,(\hat{\theta}_n^{MLE} - \theta_0) \approx - \frac{\dot{\ell}_{X_1 \ldots X_n}(\theta_0)\frac{1}{\sqrt{n}}}{\ddot{\ell}_{X_1 \ldots X_n}(\theta_0)\frac{1}{n}} \quad \begin{matrix} \text{B} \\ \text{A} \end{matrix}$$

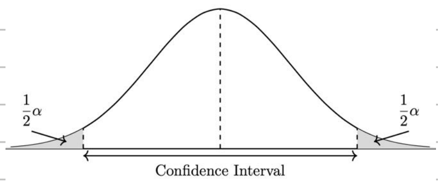
A $\quad -\frac{\ddot{\ell}_{X_1 \ldots X_n}(\theta_0)}{n} = -\frac{\sum_{i=1}^n \ddot{\ell}_{X_i}(\theta_0)}{n} \xrightarrow[LLN]{P} -E[\ddot{\ell}_{X_1}(\theta_0)] =: I(\theta_0)$

B $\quad \frac{\dot{\ell}_{X_1 \ldots X_n}(\theta_0)}{\sqrt{n}} = \sqrt{n}\,\frac{\sum_{i=1}^n \dot{\ell}_{X_i}(\theta_0)}{n} \xrightarrow[CLT]{D} \mathcal{N}(0, Var[\dot{\ell}_{X_1}(\theta_0)]) = \mathcal{N}(0, I(\theta_0))$

A + B: $\sqrt{n}\,(\hat{\theta}_n^{MLE} - \theta_0) \xrightarrow{D} \frac{\mathcal{N}(0, I(\theta_0))}{I(\theta_0)} = \mathcal{N}(0, I(\theta_0)^{-1})$

i.e. $\hat{\theta}_n^{MLE} \sim \mathcal{N}(\theta_0, (n I(\theta_0))^{-1})$ $\square$

Confidence Intervals



use efficiency of MLE to construct approximate CI:

$$\hat{\theta}_{MLE} \sim \mathcal{N}(\theta_0, (nI(\theta_0))^{-1})$$

$$\Rightarrow P\left(-z_{1-\frac{\alpha}{2}} < \sqrt{nI(\theta_0)}\,(\hat{\theta}-\theta_0) < z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

$$= P\left(-\hat{\theta} - \sqrt{\frac{1}{nI(\theta_0)}}\, z_{1-\frac{\alpha}{2}} < -\theta_0 < \hat{\theta} - \sqrt{\frac{1}{nI(\theta_0)}}\, z_{1-\frac{\alpha}{2}}\right)$$

$$= P\left(\hat{\theta} + \sqrt{\frac{1}{nI(\theta_0)}}\, z_{1-\frac{\alpha}{2}} > \theta_0 > \hat{\theta} - \sqrt{\frac{1}{nI(\theta_0)}}\, z_{1-\frac{\alpha}{2}}\right)$$

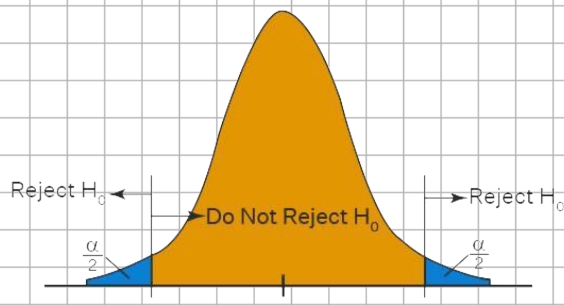
$$= P\left(\hat{\theta} - \sqrt{\frac{1}{nI(\theta_0)}}\, z_{1-\frac{\alpha}{2}} < \theta_0 < \hat{\theta} + \sqrt{\frac{1}{nI(\theta_0)}}\, z_{1-\frac{\alpha}{2}}\right)$$

$$\Rightarrow C(\vec{X}) = \left(\hat{\theta}_n^{MLE}(\vec{X}) - \sqrt{\frac{1}{nI(\theta_0)}}\, z_{1-\frac{\alpha}{2}},\ \hat{\theta}_n^{MLE}(\vec{X}) + \sqrt{\frac{1}{nI(\theta_0)}}\, z_{1-\frac{\alpha}{2}}\right)$$

$$P(\theta_0 \in C(\vec{X})) = 1-\alpha$$

replace θ_0 in $C(\vec{X})$ by $\hat{\theta}_n^{MLE}$ to get a practical CI

Hypothesis Testing



Let Θ_0 be accepted body of knowledge

We observe data $X_1 \dots X_n \overset{iid}{\sim} f_\theta$

$H_0: \ \theta = \theta_0$ status quo / null hypothesis

$H_1: \ \theta \neq \theta_0$ alternative hypothesis

Do the data give sufficient evidence to reject H_0?

Consider a test statistic T

(A) $T = \hat{\theta}^{MLE}(\vec{X})$

$t = \hat{\theta}(\vec{x})$ observed value of estimator

$$\text{p-value}(\vec{X}) = P\left(|\hat{\theta}(\vec{X}) - \theta_0| > |t - \theta_0| \;\middle|\; H_0 \text{ true}\right)$$

$$= 2 \cdot P\left(\hat{\theta}(\vec{X}) - \theta_0 > |t - \theta_0| \;\middle|\; H_0 \text{ true}\right)$$

asymptotic distribution of $\hat{\theta} - \theta_0$ is $N(0, (nI(\theta_0))^{-1})$

$$= 2 \cdot P\left(\frac{\hat{\theta}(\vec{X}) - \theta_0}{\sqrt{1/(nI(\theta_0))}} > \frac{|t - \theta_0|}{\sqrt{1/(nI(\theta_0))}} \;\middle|\; H_0 \text{ true}\right)$$

$$= 2 \cdot \left(1 - \Phi\left(\frac{|t - \theta_0|}{\sqrt{1/(nI(\theta_0))}}\right)\right)$$

$\Phi(\cdot)$ is cdf of $N(0,1)$

We reject H_0 if p-value $< \alpha$ where α = significance level
→ e.g. $\alpha = 0.05$ in science, $\alpha \leq 0.001$ in pharma

Interpretation p-value: "How likely is it to observe the data that we did observe (or more extreme) if the null hypothesis is true."

(B) $$T = \underbrace{-2 \log \underbrace{\frac{f_{\theta_0}(\vec{x})}{\sup_\theta f_\theta(\vec{x})}}_{\text{likelihood ratio (LR)}}}_{\in (0, \infty)}$$

likelihood ratio (LR):
→ close to 1: H_0 pretty good
→ close to 0: H_0 pretty bad

$\in (0, \infty)$:
→ close to 0: H_0 pretty good
→ close to ∞: H_0 pretty bad

Wilks Theorem

If $X_1 \dots X_n \overset{iid}{\sim} f_\theta$ with usual conditions

$$T = -2\log \frac{f_{\theta_0}(\vec{X})}{\sup_\theta f_\theta(\vec{X})}\Bigg|H_0 \xrightarrow{D} \chi^2_{df}$$

$df = \#$ parameters kept fixed in H_0

Proof:

$$T = -2\left(\ell_{\vec{X}}(\theta_0) - \ell_{\vec{X}}(\hat\theta)\right)$$

$$\ell_{\vec{X}}(\theta_0) = \ell_{\vec{X}}(\hat\theta) + \cancelto{0}{\dot\ell_{\vec{X}}(\hat\theta)}(\theta_0-\hat\theta) + \frac{1}{2}\ddot\ell_{\vec{X}}(\hat\theta)(\theta_0-\hat\theta)^2 + \text{"small"}$$

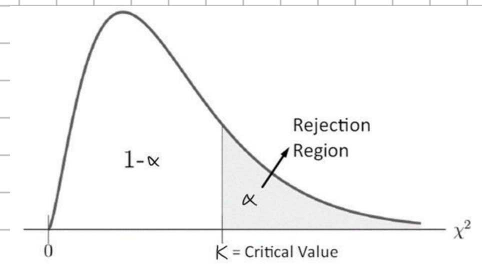
$$-2\left(\ell_{\vec{X}}(\theta_0)-\ell_{\vec{X}}(\hat\theta)\right) = T \overset{\sim}{=} -\ddot\ell_{\vec{X}}(\hat\theta)(\theta_0-\hat\theta)^2$$

$$= \frac{-\ddot\ell_{\vec{X}}(\hat\theta)}{nI(\theta_0)}(\theta_0-\hat\theta)^2\cdot nI(\theta_0)$$

$$= \underbrace{\frac{\sum_{i=1}^n(-\ddot\ell_{X_i}(\hat\theta))}{nI(\theta_0)}}_{\downarrow \text{LLN}\ \frac{E[-\ddot\ell_{X_1}(\hat\theta)]}{I(\theta_0)} = 1} \cdot \underbrace{\Big(\underbrace{\sqrt{nI(\theta_0)}\cdot(\theta_0-\hat\theta)}_{\mathcal{N}(0,1)}\Big)^2}_{\chi^2_1}$$

$\square$

Inverting the LRT



Rejection region: $R(\theta_0) = \{\vec{x} \mid T(\vec{x}) \geq K\}$

$\rightarrow$ We reject H_0 if we observe $T(\vec{x}) \geq K$

$P(\vec{X} \in R(\theta_0) \mid H_0 \text{ true}) = P(T(\vec{X}) \geq K \mid H_0 \text{ true}) = \alpha$

Using the asymptotic distr of $T(\vec{X}) \mid H_0$ true we can find K s.t. above equality holds:

$$K = \chi^2_{1,1-\alpha} = F^{-1}_{\chi^2_1}(1-\alpha)$$

$$P(T(\vec{X}) \geq \chi^2_{1,1-\alpha} \mid H_0 \text{ true}) = \alpha$$

$$\Rightarrow P(\underbrace{\vec{X} \notin R(\theta_0)}_{\text{"we accept } H_0\text{"}} \mid H_0 \text{ true}) = 1-\alpha$$

Hypothesis Testing $\longleftrightarrow$ Confidence Interval

$$P(\vec{X} \notin R(\theta_0) \mid H_0 \text{ true}) = 1-\alpha$$

$$\Downarrow$$

$$P(\underbrace{T(\vec{X}) \leq \chi^2_{1,1-\alpha}}) = 1-\alpha$$

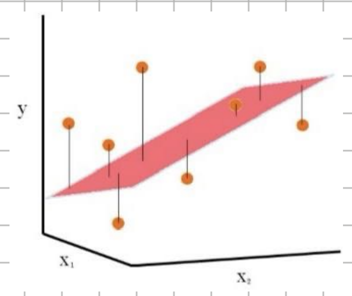
if we "invert" this by isolating θ_0 on one side

we can construct

$$C(\vec{X}) = \{\theta \mid T_\theta(\vec{X}) \leq \chi^2_{1,1-\alpha}\}$$

$$P(\theta_0 \in C(\vec{X})) = 1-\alpha$$

Linear Regression



Data: $\{(\vec{x}_i, Y_i)\}_{i=1}^{n}, \quad \vec{x}_i \in \mathbb{R}^p$

$Y_i = \vec{x}_i^T \vec{\beta} + \varepsilon_i, \quad \vec{\beta} \in \mathbb{R}^p, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

(Y_i: response, $\vec{x}_i$: covariates)

Estimation

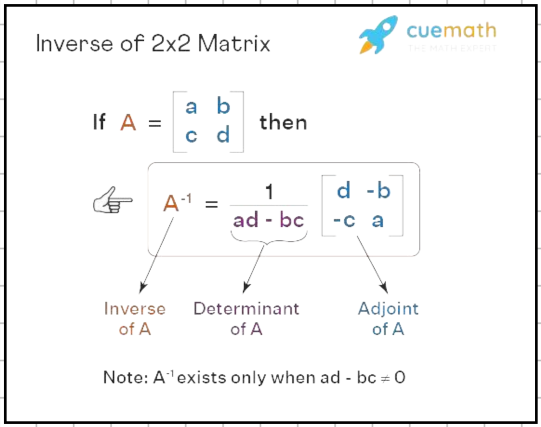
$$\begin{aligned} \ell_{\vec{Y}}(\vec{\beta}) &= \sum_{i=1}^{n} \log f(Y_i \mid \vec{\beta}, \vec{x}_i) = \log f(\vec{Y} \mid \vec{\beta}, \vec{x}_1, \dots, \vec{x}_n) \\ &= \log\left((2\pi)^{-\frac{n}{2}} \left|\sigma^2 \underline{I}\right|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\vec{Y} - \underline{X}\vec{\beta})^T (\sigma^2 \underline{I})^{-1} (\vec{Y} - \underline{X}\vec{\beta})\right)\right) \\ &= C - \tfrac{1}{2}\log \sigma^{2n} - \tfrac{1}{2\sigma^2}(\vec{Y} - \underline{X}\vec{\beta})^T(\vec{Y} - \underline{X}\vec{\beta}) \\ &= C - \tfrac{n}{2}\log\sigma^2 - \tfrac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \vec{x}_i^T\vec{\beta})^2 \end{aligned}$$

$$\underline{X} = \underbrace{\begin{bmatrix} - \vec{x}_1^T - \\ \vdots \\ - \vec{x}_n^T - \end{bmatrix}}_{p} \Big\} n$$

$$\begin{aligned} \dot{\ell}_{\vec{Y}}(\vec{\beta}) &= -\tfrac{1}{2\sigma^2}\sum_{i=1}^{n} 2(Y_i - \vec{x}_i^T\vec{\beta})\cdot(-\vec{x}_i^T) = \tfrac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \vec{x}_i^T\vec{\beta})\cdot\vec{x}_i^T = \tfrac{1}{\sigma^2}(\vec{Y} - \underline{X}\vec{\beta})^T\underline{X} = \tfrac{1}{\sigma^2}(\vec{Y}^T - \vec{\beta}^T\underline{X}^T)\underline{X} \\ &= \tfrac{1}{\sigma^2}(\vec{Y}^T\underline{X} - \vec{\beta}^T\underline{X}^T\underline{X}) \overset{!}{=} \vec{0}_p^T, \quad \vec{\beta}^T\underline{X}^T\underline{X} = \left[\sum_{i=1}^{p}\beta_i(\underline{X}^T\underline{X})_{i1} \ \cdots \ \sum_{i=1}^{p}\beta_i(\underline{X}^T\underline{X})_{ip}\right] \end{aligned}$$

$$\Rightarrow \vec{\beta}^T\underline{X}^T\underline{X} = \vec{Y}^T\underline{X} \iff \underline{X}^T\underline{X}\vec{\beta} = \underline{X}^T\vec{Y} \iff \vec{\beta} = (\underline{X}^T\underline{X})^{-1}\underline{X}^T\vec{Y}$$

directly normal! (not just asymptotically)



$$\ddot{\ell}_{\vec{Y}}(\vec{\beta}) = -\tfrac{1}{\sigma^2}\tfrac{d}{d\vec{\beta}}(\vec{\beta}^T\underline{X}^T\underline{X}) = -\tfrac{1}{\sigma^2}\underline{X}^T\underline{X} \overset{\text{Löwner}}{<} 0$$

$$\Rightarrow \hat{\vec{\beta}} = (\underline{X}^T\underline{X})^{-1}\underline{X}\vec{Y} \text{ is maximum} \Rightarrow MLE$$

Properties:

expected v:
$$\begin{aligned} E[\hat{\vec{\beta}}] &= E\left[(\underline{X}^T\underline{X})^{-1}\underline{X}^T\vec{Y}\right] \\ &= (\underline{X}^T\underline{X})^{-1}\underline{X}^T E[\vec{Y}] \\ &= (\underline{X}^T\underline{X})^{-1}\underline{X}^T\underline{X}\vec{\beta} \\ &= \vec{\beta} \quad \rightarrow \text{unbiased} \end{aligned}$$

variance:
$$\begin{aligned} Var[\hat{\vec{\beta}}] &= Var\left[(\underline{X}^T\underline{X})^{-1}\underline{X}^T\vec{Y}\right] \\ &= (\underline{X}^T\underline{X})^{-1}\underline{X}^T \, Var[\vec{Y}] \left((\underline{X}^T\underline{X})^{-1}\underline{X}^T\right)^T \\ &= (\underline{X}^T\underline{X})^{-1}\underline{X}^T\sigma^2\underline{I}\,\underline{X}(\underline{X}^T\underline{X})^{-1} \\ &= \sigma^2(\underline{X}^T\underline{X})^{-1} \end{aligned}$$

achieves CRLB → efficient

CRLB:
$$\begin{aligned} (I(\vec{\beta}))^{-1} &= \left(-E[\ddot{\ell}_{\vec{Y}}(\vec{\beta})]\right)^{-1} \\ &= \left(-E[-\tfrac{1}{\sigma^2}\underline{X}^T\underline{X}]\right)^{-1} \\ &= \sigma^2(\underline{X}^T\underline{X})^{-1} \end{aligned}$$

Confidence Interval

$\hat{\vec{\beta}} = (\underline{X}^T \underline{X})^{-1} \underline{X} \vec{Y}$ is a linear combination of normally distributed R.V.

$\Rightarrow$ so it is also normally distributed with mean and variance as calculated above (exact!)

$$\hat{\vec{\beta}} \sim N(\vec{\beta}, \sigma^2 (\underline{X}^T \underline{X})^{-1})$$

$\rightarrow$ we could create a joint confidence region for $\vec{\beta}$, but we stay modest and do univariate confidence for each of the β_j

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 ((\underline{X}^T \underline{X})^{-1})_{jj})$$

a $(1-\alpha)\cdot 100\%$ CI for β_j:

$$\boxed{\left(\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \sigma \sqrt{(\underline{X}^T \underline{X})^{-1}_{jj}}\right)}$$

Testing

particular covariate/predictor x_j significant or not, i.e. $\beta_j = 0$? ?

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$ Covariate j is interesting

but: if we estimate

Test-statistic:

$$\hat{\beta}_j \mid H_0 \sim \mathcal{N}(0, \sigma_j^2) \text{ where } \sigma_j^2 = \sigma^2 (\underline{X}^T \underline{X})^{-1}_{jj}$$

we estimate β_j as b_j

$$\text{P-value} = P(|\hat{\beta}_j| > |b_j| \mid H_0) = 2 \cdot P\Big(\underbrace{\frac{\hat{\beta}_j}{\sigma_j}}_{N(0,1)} > \frac{|b_j|}{\sigma_j} \,\Big|\, H_0\Big) = 2 \cdot \left(1 - \Phi\left(\frac{|b_j|}{\sigma_j}\right)\right)$$

reject H_0 if P-value $< \alpha$

Model Evaluation

does adding $x_{p_0+1}, \dots, x_{p_1}$ significantly improve our ability to predict y? ?

consider 2 competing models (nested, i.e. predictors in small model all contained in bigger model)

<u>Model 0:</u> $x_i^{(0)} = (x_{i1}, \dots, x_{ip_0}) \in \mathbb{R}^{p_0}$

<u>Model 1:</u> $x_i^{(1)} = (x_{i1}, \dots, x_{ip_1}) \in \mathbb{R}^{p_1}$

$p_0 < p_1$

H_0: model 0 is correct

H_1: model 1 is correct

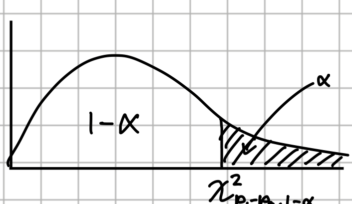
Test-statistic:

$$\Lambda = -2 \log \frac{L_0(\beta_0)}{L_1(\beta_1)} \quad \text{where} \quad \begin{aligned} L_0(\beta_0) &= \max_{\beta} f(\vec{Y} \mid \text{model } 0), \\ L_1(\beta_1) &= \max_{\beta} f(\vec{Y} \mid \text{model } 1) \end{aligned}$$

$$= -2\left(\ell_{\vec{Y}}(\hat{\beta}_0) - \ell_{\vec{Y}}(\hat{\beta}_1)\right)$$

$$\Lambda \mid H_0 \sim \chi^2_{p_1 - p_0} \quad \text{(exact!)}$$

reject H_0 if $\Lambda > \chi^2_{p_1 - p_0, 1-\alpha}$



this is where the field of data science was in the $70^s/80^s$

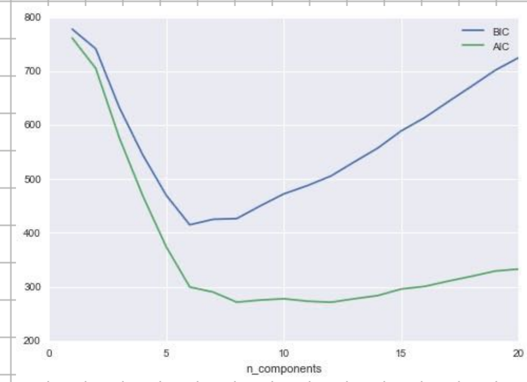
there was one big question: how to compare non-nested models?

$\rightarrow$ next page

Residual Sum of Squares

$$RSS = \hat{\vec{\varepsilon}}^T \hat{\vec{\varepsilon}} = (\vec{y} - \underset{\sim}{X}\hat{\vec{\beta}})^T(\vec{y} - \underset{\sim}{X}\hat{\vec{\beta}}) = \sum_{i=1}^{n}(y_i - \vec{x}_i^T\hat{\vec{\beta}})^2$$

Akaike Information Criterion



$$KL(M_0 | M) = \int_{-\infty}^{\infty} f_{M_0}(y) \log\left(\frac{f_{M_0}(y)}{f_M(y)}\right) dy \quad \rightarrow \text{minimize}$$

$$= E_{Y \sim M_0}\left[\ell_Y(M_0) - \ell_Y(M)\right]$$

$$= C - E_{Y \sim M_0} \ell_Y(M)$$

$$\approx \cdots \approx C - \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_{Y_i}(\hat{\beta})}_{\ell_Y(\hat{\beta})} + \underbrace{\frac{p}{n}}_{\text{bias}} \qquad / \times 2n$$

$$AIC(M) = \boxed{-2\,\ell_Y(\hat{\vec{\beta}}) + 2p} \quad \rightarrow \text{prediction}$$

$$BIC(M) = -2\,\ell_Y(\hat{\vec{\beta}}) + p\log(n) \quad \rightarrow \text{"truth"}$$

ANOVA

Mixed Effects Models

Consider $\underline{\boldsymbol{X}}$ to be the covariates describing the quantities of interest, $\underline{\boldsymbol{Z}}$ to be the covariates describing the nuisance effects

Random effects model:

$$\boxed{\vec{\boldsymbol{Y}} = \underbrace{\underline{\boldsymbol{X}}\vec{\boldsymbol{\beta}}}_{\text{fixed effects}} + \underbrace{\underline{\boldsymbol{Z}}\vec{\boldsymbol{\gamma}}}_{\substack{\text{random effects} \\ \gamma_j \sim N(0,\sigma_\gamma^2)}} + \underbrace{\vec{\boldsymbol{\epsilon}}}_{\substack{\text{residuals} \\ \sim N(0,\sigma^2 \underline{\boldsymbol{I}})}}}$$

Why is γ random?

- Philosophical: γ is a random draw out of the population of classes
- Pragmatic: if γ is random, then γ is **not** a paramater, σ_γ^2 is!

Estimation: REML (Restricted Maximum Likelihood)

Way to estimate the variance parameters σ_γ^2 and σ^2

projection matrix into the residual space:

$$\underline{\boldsymbol{P}} = \underline{\boldsymbol{I}} - \underline{\boldsymbol{X}}(\underline{\boldsymbol{X}}^T\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}^T$$

note that $\underline{\boldsymbol{P}}\vec{\boldsymbol{Y}} = \vec{\boldsymbol{Y}} - \underline{\boldsymbol{X}}\underbrace{(\underline{\boldsymbol{X}}^T\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}^T\vec{\boldsymbol{Y}}}_{\hat{\vec{\boldsymbol{\beta}}}} = \hat{\vec{\boldsymbol{\epsilon}}}$

$$\mathbb{E}(\underline{\boldsymbol{P}}\vec{\boldsymbol{Y}}) = \vec{\boldsymbol{0}}$$

$$\begin{aligned} \mathbb{V}(\underline{\boldsymbol{P}}\vec{\boldsymbol{Y}}) &= \underline{\boldsymbol{P}}\mathbb{V}(\vec{\boldsymbol{Y}})\underline{\boldsymbol{P}}^T \\ &= \underline{\boldsymbol{P}}\left(\mathbb{V}(\underline{\boldsymbol{X}}\vec{\boldsymbol{\beta}} + \underline{\boldsymbol{Z}}\vec{\gamma} + \vec{\epsilon})\right)\underline{\boldsymbol{P}}^T \\ &= \underbrace{\underline{\boldsymbol{P}}\Big(\underline{\boldsymbol{Z}}\underbrace{\mathbb{V}(\vec{\gamma})}_{\text{function of } \sigma_\gamma^2}\underline{\boldsymbol{Z}}^T + \sigma^2\underline{\boldsymbol{I}}\Big)\underline{\boldsymbol{P}}^T}_{\text{known function of } \sigma_\gamma^2 \text{ and } \sigma^2} \end{aligned}$$

If I maximize the likelihood for $\underline{\boldsymbol{P}}\vec{\boldsymbol{Y}}$ with respect to σ_γ^2 and σ^2, then estimates are unbiased.

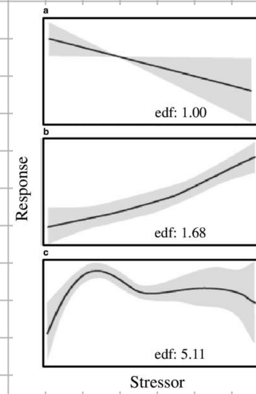
$\Longrightarrow$ use those estimates to estimate $\vec{\boldsymbol{\beta}}$

REML

problem with ML:
doesn't take into account the degrees of freedom that we lose when we estimate the mean

REML basically corrects for bias in variance components of the ML estimate

Additive Models



non-linearity: $Y = f(X) + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$

how to find f?

<u>Option 1</u>: we have a known form of functional space; $f(X) = f(X, \theta)$

find f by estimating θ; $\hat{\theta} = \operatorname{argmax} -\frac{1}{2}\sum_{i=1}^{n}(Y_i - f(X_i, \theta))^2 \longrightarrow \hat{f} = f(\cdot, \hat{\theta})$

<u>Option 2</u>: data driven non-linear functions

find the true function that belongs to general function class, say $C = \{f \mid f: \mathbb{R} \to \mathbb{R} \text{ continuous}\}$

construct a sequence of functional spaces $S_1 \subset S_2 \subset S_3 \subset \cdots$, such that $\bigcup_{i=1}^{\infty} S_i = C$ $\Rightarrow$ SIEVE

IDEA: estimate f inside $S_{m(n)}$ where n sample size and $m(n)$ increasing sequence with $\lim_{n\to\infty} m(n) = \infty$

How to choose S_m?
finite dimensional space with basis $\{b_1, \dots, b_m\}$ where $b_j: \mathbb{R} \to \mathbb{R}$

Options

(A) $b_j(x) = x^j$, note $S_m = \{\text{all polynomials of order } m\}$ and $\overline{\bigcup_{i=1}^{\infty} S_i} = C$

disadvantage: tends to be quite "unstable" in practice
→ any choice made to fit data at one point affects the fit everywhere else (NON-LOCAL)
→ drawback

(B) LOCAL BASIS, e.g. cubic spline

3 constraints at knot points ξ: continuous, continuous 1st & 2nd derivative

\# independent parameters: $\underbrace{(n_\xi + 1)\cdot 4}_{\text{intervals}} - \underbrace{n_\xi \cdot 3}_{\text{constraints at knots}} = n_\xi + 4$

n_ξ knots $\Rightarrow S_{n_\xi + 4}$

by increasing n_ξ, we obtain a "covering" of C

Estimation

given some space S_m with basis $\mathcal{B}_m = \{b_1, \dots, b_m\}$ and data $(x_1, Y_1), \dots, (x_n, Y_n)$ MLE of $f \in S_m$

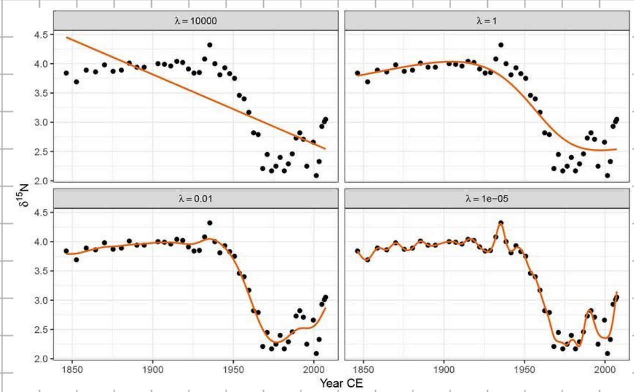
$$\hat{f} = \underset{f \in S_m}{\operatorname{argmax}} -\frac{1}{2}\sum_{i=1}^{n}(Y_i - f(x_i))^2$$

$$\to \hat{\vec{\beta}}_{MLE} = \underset{\beta \in \mathbb{R}^m}{\operatorname{argmax}} -\frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{m}\beta_j b_j(x_i)\right)^2$$ → linear model

$$= (\underline{X}^T\underline{X})^{-1}\underline{X}^T\vec{Y}, \quad \vec{Y} = [Y_1, \dots, Y_n]^T, \quad \underline{X} = \begin{bmatrix} b_1(x_1) & \cdots & b_m(x_1) \\ \vdots & & \vdots \\ b_1(x_n) & \cdots & b_m(x_n) \end{bmatrix}$$

this may be a problematic solution, e.g. if m is too close to n → $\hat{f}$ becomes too "wiggly"
→ consequence of overfitting

solution: regularization / penalty on wiggliness W



$$W(f) = \int_L^U (f''(x))^2 dx$$

$$\overset{f \in S_m}{=} \int_L^U \left(\sum_{j=1}^m \beta_j b_j''(x) \right)^2 dx$$

$$= \int_L^U \sum_{j=1}^m \sum_{k=1}^m \beta_j \beta_k b_j''(x) b_k'' dx$$

$$= \sum_{j=1}^m \sum_{k=1}^m \beta_j \beta_k \int_L^U b_j''(x) b_k''(x) dx$$

$$= \sum_{j=1}^m \sum_{k=1}^m \beta_j \beta_k W_{jk}^* \qquad \rightarrow W_{jk}^* \text{ is known}$$

$$= \vec{\beta}^T \underline{W}^* \vec{\beta}$$

penalized maximum likelihood:

$$\hat{\beta} = \underset{\vec{\beta}}{\operatorname{argmax}} \left(-\frac{1}{2} (\vec{Y} - \underline{X}\vec{\beta})^T (\vec{Y} - \underline{X}\vec{\beta}) - \frac{\lambda}{2} \vec{\beta}^T \underline{W}^* \vec{\beta} \right)$$

if λ is $\begin{cases} = 0: & \text{wiggly solution} \\ \rightarrow \infty: & \text{linear solution} \end{cases}$

$$\frac{\partial \ell^P}{\partial \vec{\beta}} = \underline{X}^T \vec{Y} - \underline{X}^T \underline{X} \vec{\beta} - \lambda \underline{W}^* \vec{\beta} = 0$$

$$\Rightarrow (\underline{X}^T \underline{X} + \lambda \underline{W}^*) \vec{\beta} = \underline{X}^T \vec{Y}$$

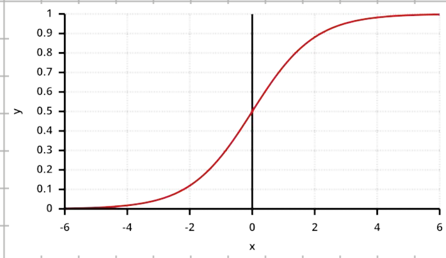
$$\rightarrow \hat{\beta}_{PMLE} = (\underline{X}^T \underline{X} + \lambda \underline{W}^*)^{-1} \underline{X}^T \vec{Y}$$

how to choose λ: cross-validation

note on wiggliness:

among all functions $\in C$ that fit the data "in the same way", cubic spline is the least wiggly

Logistic Regression



consider data $(Y, \vec{x})$ where $\begin{cases} Y \sim \text{Bern}(\pi(\vec{x})) \\ \vec{x} \in \mathbb{R}^p \end{cases}$

How to model $\pi(\vec{x})$?

$$\begin{cases} \eta = \vec{x}^T \vec{\beta} & \text{Linear predictor} \\ \pi = g^{-1}(\eta) & \text{for some suitable function } g^{-1}: \mathbb{R} \to [0,1] \text{ that fulfills} \end{cases}$$

→ inverse link function

* one-to-one
* monotone increasing

(i) logistic function $g^{-1}(\eta) = \dfrac{e^{\eta}}{1+e^{\eta}}$ ✓

(ii) cdf of $N(0,1)$ $g^{-1}(\eta) = \Phi_{0,1}(\eta)$

Logistic Regression

$$\begin{cases} Y \sim \text{Bern}(\pi) \\ \pi = \dfrac{e^{\eta}}{1+e^{\eta}} \\ \eta = \vec{x}^T \vec{\beta} \end{cases} \Longrightarrow \eta = \overbrace{\text{Log}\Big(\underbrace{\frac{\pi}{1-\pi}}_{\text{odds} \in \mathbb{R}^+}\Big)}^{\text{log-odds} \in \mathbb{R}}$$

Logit function

$\vec{x}^T \vec{\beta}$ has interpretation in terms of log-odds:

β_j = increase in log-odds if x_j increases by 1 and rest stayed the same

e^{β_j} = multiplicative increase in odds if x_j increases by 1 and rest stayed the same

"chances of Y are increased by a factor of e^{β_j} if $x_j \to x_j + 1$"

Estimation

$$\begin{aligned}
\ell(\vec{\beta}) &= \text{Log} \prod_{i=1}^{n} \pi_i^{Y_i} (1-\pi_i)^{(1-Y_i)} \\
&= \sum_{i=1}^{n} \Big(Y_i \,\text{Log}\Big(\frac{e^{\eta_i}}{1+e^{\eta_i}}\Big) + (1-Y_i)\,\text{Log}\Big(\overbrace{1 - \frac{e^{\eta_i}}{1+e^{\eta_i}}}^{\frac{1}{1+e^{\eta_i}}}\Big)\Big) \\
&= \sum_{i=1}^{n} \Big(Y_i \big(\vec{x}_i^T \vec{\beta} - \text{Log}(1+e^{\eta_i})\big) + (1-Y_i)\big(-\text{Log}(1+e^{\eta_i})\big)\Big) \\
&= \sum_{i=1}^{n} \Big(Y_i \vec{x}_i^T \vec{\beta} - \cancel{Y_i \,\text{Log}(1+e^{\eta_i})} + \cancel{Y_i \,\text{Log}(1+e^{\eta_i})} - \text{Log}(1+e^{\eta_i})\Big) \\
&= \sum_{i=1}^{n} \Big(Y_i \vec{x}_i^T \vec{\beta} - \text{Log}(1+e^{\vec{x}_i^T \vec{\beta}})\Big)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta_j} \ell(\vec{\beta}) &= \sum_{i=1}^{n} \Big(Y_i x_{ij} - \frac{1}{1+e^{\vec{x}_i^T \vec{\beta}}} \cdot e^{\vec{x}_i^T \vec{\beta}} \cdot x_{ij} \Big) \\
&= \sum_{i=1}^{n} (Y_i - \pi_i) x_{ij}
\end{aligned}$$

$$\begin{cases} \sum_{i=1}^{n} (Y_i - \pi_i) x_{i1} \overset{!}{=} 0 \\ \vdots \\ \sum_{i=1}^{n} (Y_i - \pi_i) x_{ip} \overset{!}{=} 0 \end{cases}$$

find $\hat{\vec{\beta}}$ numerically using

(A) Newton-Raphson

(B) Fisher Scoring

Properties of $\hat{\vec{\beta}}$

asymptotically (under certain constraints on x):

$$\lim_{n\to\infty} \sqrt{n}(\hat{\vec{\beta}} - \vec{\beta}_0) \sim N(0, \underline{I}^{-1}), \quad \underline{I} = \lim_{n\to\infty} -E\left[\frac{\partial^2 \ell}{\partial \vec{\beta}^2}\right] / n$$

so approximately:

$$\hat{\vec{\beta}} \sim N(\vec{\beta}_0, \tfrac{1}{n}\underline{I}^{-1})$$

→ we can now construct CI's and test hypotheses $H_0: \beta_j = 0$

other things:

Ⓐ testing of nested models $\begin{cases} H_0: & M = M_0 \\ H_1: & M = M_1 \end{cases}$ where $M_0 \subset M_1$

likelihood ratio test: $\Lambda = -2(\ell(\hat{\vec{\beta}}_{M_0}) - \ell(\hat{\vec{\beta}}_{M_1}))$ where $\Lambda | H_0 \sim \chi^2_{df(M_1)-df(M_0)}$

Ⓑ $AIC(M) = -2\ell(\hat{\vec{\beta}}_M) + 2df(M)$

Ⓒ goodness-of-fit test $\begin{cases} H_0: & M \text{ is true/correct} \\ H_1: & M \text{ is not true} \end{cases}$

test-statistic: $T = -2\ell(\hat{\vec{\beta}}_M)$ deviance where $T | H_0 \sim \chi^2_{n-df(M)}$

reject H_0 if $T > \chi^2_{n-df(M), 1-\alpha}$

Generalized Linear Models

Let $Y \sim$ Exponential Family, i.e.

$$f_Y(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\right)$$

for parameters

θ : location parameter
φ : dispersion parameter

for some functions a, b, c

expected value and variance of Y:

$$0 = E[\dot{\ell}_Y(\theta)] \quad \text{trick...}$$
$$= E\left[\frac{\partial}{\partial\theta}\left(\frac{Y\theta - b(\theta)}{a(\varphi)} + c(Y,\varphi)\right)\right]$$
$$= E\left[\frac{Y - \frac{\partial}{\partial\theta}b(\theta)}{a(\varphi)}\right]$$
$$= \frac{E[Y] - \dot{b}(\theta)}{a(\varphi)}$$

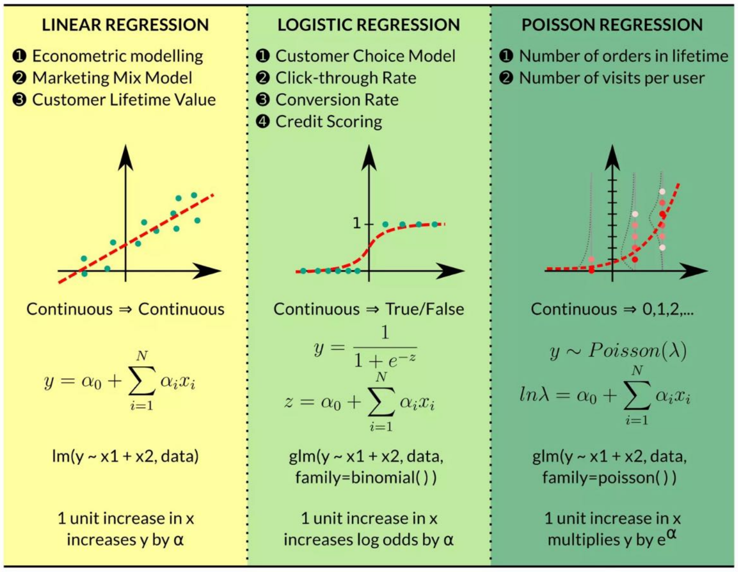
$$\rightarrow E[Y] = \dot{b}(\theta) \implies \theta = \dot{b}^{-1}(E[Y]) = \dot{b}^{-1}(\mu)$$

$$0 = E[\dot{\ell}_Y(\theta)^2] + E[\ddot{\ell}_Y(\theta)]$$
$$= E\left[\left(\frac{Y - \dot{b}(\theta)}{a(\varphi)}\right)^2\right] - \ddot{b}(\theta)/a(\varphi)$$
$$= E[(Y - E[Y])^2]/a(\varphi)^2 - \ddot{b}(\theta)/a(\varphi)$$

$$\rightarrow Var[Y] = \ddot{b}(\theta)\, a(\varphi)$$

Generalized Linear Models (GLM) extend the ordinary linear regression and allow the response variable y to have an error distribution other than the normal distribution.

GLMs are:
a) Easy to understand
b) Simple to fit and interpret in any statistical package
c) Sufficient in a lot of practical applications



Linear Model and Link Function

consider $(\vec{x}, Y)$ where $Y \sim$ Exponential Family and $\vec{x} \in \mathbb{R}^p$

linear predictor for Y:

$$\eta = \vec{x}^T\vec{\beta}, \quad \vec{\beta} \in \mathbb{R}^p$$

the aim of the link function $g: M \rightarrow \mathbb{R}$ where $EY \in M \subset \mathbb{R}$ is to map EY into the space of linear predictors $\mathbb{R} \ni \vec{x}^T\vec{\beta}$:

$$\boxed{E[Y] = g^{-1}(\vec{x}^T\vec{\beta})} \Leftrightarrow g(\mu) = \eta$$

constraints on g:
* 1-to-1
* continuous
* strictly increasing

→ computationally more efficient but not always best choice from interpretation point of view

canonical link function: $\boxed{g = \dot{b}^{-1}}$ then $\eta = \vec{x}^T\vec{\beta} = g(E[Y]) = \theta$

Estimation

$$\ell_{\vec Y}(\vec\beta) = \sum_{i=1}^{n} \ell_{Y_i}(\vec\beta)$$
$$= \sum_{i=1}^{n}\left(\frac{Y_i\theta_i - b(\theta_i)}{a_i(\varphi)} + c(Y_i,\varphi)\right)$$

$$\frac{d}{d\beta_k}\ell_{\vec Y}(\vec\beta) = \sum_{i=1}^{n}\frac{d}{d\beta_k}\ell_{Y_i}(\vec\beta)$$
$$= \sum_{i=1}^{n}\frac{d\ell_{Y_i}}{d\theta_i}\,\frac{d\theta_i}{d\mu_i}\,\frac{d\mu_i}{d\eta_i}\,\frac{d\eta_i}{d\beta_k}$$
$$= \sum_{i=1}^{n}\frac{Y_i-\dot b(\theta_i)}{a_i(\varphi)}\,\frac{1}{\ddot b(\theta_i)}\,\frac{d\mu_i}{d\eta_i}\,\frac{d\eta_i}{d\beta_k}$$
$$= \sum_{i=1}^{n}\frac{(Y_i-\mu_i)\cdot {}^{d\eta_i}/_{d\mu_i}}{\left({}^{d\eta_i}/_{d\mu_i}\right)^2\ddot b(\theta_i)a_i(\varphi)}\,\frac{d\eta_i}{d\beta_k}$$

Iterative Re-weighted Least Squares. Aiming to maximize the log, we set the derivative of the log-likelihood to zero. In class we used the usual Taylor method. Here we consider another technique, called *iterative reweighted least squares.*

i. The problem of the above equation for the derivative of the log-likelihood is that it is not linear in β. Write $\mu_i = \mu_i(\beta_0) + (\mu_i(\beta) - \mu_i(\beta_0))$ to derive the linear approximation of the numerator using the Taylor expansion of η_i around $\mu_i(\beta_0)$.

ii. Write now the full derivative of the likelihood in the format
$$\frac{d\ell_Y}{d\beta} \approx \sum_{i,j}\frac{(Y_i^* - x_i^t\beta)}{\left(\frac{d\eta_i}{d\mu_i}\right)^2 V(\mu_i)}\frac{d\eta_i}{d\beta}.$$
What is the *adjusted dependent variable* Y^*.

iii. Show that we can further approximate the denominator of the above expression to get,
$$\frac{d\ell_Y}{d\beta} \approx X^tW(Y^* - X\beta),$$
where X is the matrix of dyadic covariates $\{x_i\}_{i,j}$ and W is the diagonal matrix with $1/w_i$ on the diagonal, where $w_i = \left(\frac{d\eta_i}{d\mu_i}(\beta_0)\right)^2 V(\mu_i(\beta_0))$. The resulting expression is now linear in β.

iv. What is the expression of the term β that maximizes the approximation of the likelihood?

aim: write this as a linear function in β.

(i) write μ_i as approx linear in η_i:

$$\mu(\eta) \cong \mu(\eta_0) + (\eta-\eta_0)\frac{\partial\mu}{\partial\eta}(\eta_0)$$

(ii)
$$\frac{\partial\ell_Y}{\partial\beta_k} \cong \sum_{i=1}^{n}\frac{\left(y_i - \left[\overbrace{\mu(\eta_{i0})}^{\mu_{i0}} + (\eta_i-\eta_{i0})\frac{\partial\mu_i}{\partial\eta_i}(\eta_0)\right]\right)\frac{\partial\mu_i}{\partial\eta_i}(\mu_{i0})}{\left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2\ddot b(\theta)\,a(\varphi)}\underbrace{\frac{\partial\eta_i}{\partial\beta_k}}$$

$$= \frac{\partial x_i^t\beta}{\partial\beta_k} = x_{ik}$$

$$= \sum_{i=1}^{n}\frac{\overbrace{\boxed{(y_i-\mu_{i0})\frac{\partial\eta_i}{\partial\mu_i}(\mu_{i0})+\eta_{i0}}}^{y_i^*} - \eta_i}{\left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2\ddot b(\theta)\,a(\varphi)}\,x_{ik}$$

$$= \sum_{i=1}^{n}\frac{y_i^* - x_i^t\beta}{\left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2\ddot b(\theta)\,a(\varphi)}\cdot x_{ik}$$

almost linear

but still depend on β

$$\propto \sum_{i=1}^{n}\frac{y_i^*-x_i^t\beta}{\boxed{\left(\frac{\partial\eta_i}{\partial\mu_i}[\mu_{i0}]\right)^2\ddot b(\theta_i(\mu_{i0}))\,a(\varphi)}}\cdot x_{ik}$$

$\to w_i(\mu_{i0})$

$$= \sum_{i=1}^{n}\frac{y_i^*-x_i^t\beta}{w_i(\mu_{i0})}\cdot x_{ik}$$

$$\frac{\partial\ell}{\partial\beta} \cong \underset{p\times n}{X^t}\,\underset{n\times n}{W}\,\underset{n\times 1}{(y^*-X\beta)} = 0$$

W ($n\times n$) $\to$ diagonal matrix with $\frac{1}{w_i}$ on diagonal

$$x^tWy^* - x^tWX\beta = 0$$

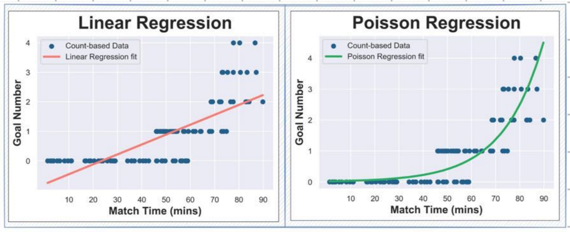
$$\boxed{\beta = (x^tWx)^{-1}x^tWy^*}$$

Poisson Regression

Y_i = #particles that decayed in time interval i

$$f_Y(y) = e^{-\lambda}\frac{\lambda^y}{y!} \quad \text{where } y \in \mathbb{N}_0$$
$$= \exp\left(\log\left(e^{-\lambda}\frac{\lambda^y}{y!}\right)\right)$$
$$= \exp\left(y\log(\lambda) - \log(y!) - \lambda\right)$$
$$= \exp\left(y\log(\lambda) - \lambda - \log(y!)\right)$$
$$\overset{!}{=} \exp\left(\frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\right)$$

$\rightarrow \theta = \log(\lambda),\ b(\theta) = \exp(\theta),\ a(\varphi) = 1,\ c(y,\varphi) = -\log(y!)$



❌ Linear Regression	✅ Poisson Regression
Unsuitable for modeling count data	Specifically designed to model count data
May return negative output for some inputs	All outputs are non-negative
Errors must follow a symmetric distribution	Works well even if errors are asymmetrically distributed

link

$$\theta \xrightarrow{e^\theta} \lambda \xrightarrow{g} \eta = \vec{x}^T\vec{\beta}$$

$\mathcal{M} = \{\lambda = EY\} = (0,\infty)$ = set of all Poisson rates

$g: (0,\infty) \rightarrow \mathbb{R}$

canonical link:

$$g(\lambda) = \dot{b}^{-1}(\lambda) = \exp^{-1}(\lambda) = \log(\lambda) = \eta = \vec{x}^T\vec{\beta}$$

offset

#particles that decay depends on length of time interval:

$$Y \sim \text{Poisson}(\text{length} \cdot e^{\vec{x}^T\vec{\beta}})$$
$$= \text{Poisson}(e^{\vec{x}^T\vec{\beta} + \log(\text{length})})$$

no additional covariate (bad idea!), we do not estimate the effect of length

in Poisson Regression offset are quite common and need to be included "manually"