MATH 341/641 LEC 15

So far, this class focused on the 3 goals of inference: estimation, testing, confidence set creation.

We began by assuming a DGP that explained where the data was realized from and what the params were and went from there.

How do you assume a DGP? Sometimes you really know e.g. coin flips are Binomial. Most often you don't. Wind speeds at JFK airport DGP? Summer times of lab rats? Unknown! We need to guess. Why not guess a few different DGP's and select the best one (model selection).

⇒ A DGP, a positive, is a "model" which is a useful approximation to reality. We then "fit" the model by estimating its params. Once we have this, we know properties of the model and can make predictions (BPZ machine).

Consider M candidate models $m = 1, 2, \ldots M$. We would like to pick the best one ("model selection").

Additionally, we would like to assign a probability weight score to each.

Model Selection is a fundamental problem in science!

$m=1$: $F = G\frac{m_1 m_2}{r^2}$ (Newton's Law)

$m=2$: $F = G_1\frac{m_1 m_2}{r^2} + G_2\frac{m_1 m_2}{r^3}$ (Newton's Extension)

$m=3$: $F = G_1\frac{m_1 m_2}{r^2}e^{-G_2 r}$ (Laplace's Extension)

Which one is "best"? We know mod #1 is a good approx but wrong as Einstein's relativity supercedes it.

DGP$_1$: $X_1, \dots X_n \overset{iid}{\sim} f_1(x_i; \theta_{1,1}, \dots \theta_{1,K_1}) = \mathcal{L}_1(x_i; \theta_{1,1}, \dots \theta_{1,K_1})$

$\vdots$

DGP$_m$: $X_1, \dots X_n \overset{iid}{\sim} f_m(x_i; \theta_{m,1}, \dots \theta_{m,K_m}) = \mathcal{L}_m(x_i; \theta_{m,1}, \dots, \theta_{m,K_m})$

The # of params could be different in each model $K_1, K_2, \dots, K_m$

Why not just take the model with the highest likelihood?

i.e. $m_* = \underset{m \in \{1,\dots m\}}{\operatorname{argmax}} \left\{ \mathcal{L}_m(\theta_{m,1}, \dots \theta_{m,K_m}; X_1, \dots X_n) \right\}$

$= \underset{m \in \{1,\dots m\}}{\operatorname{argmax}} \left\{ \ell_m(\theta_{m,1}, \dots \theta_{m,K_m}; X_1, \dots X_n) \right\}$

Can we do this? No! We don't know the values of the θ_m's for any m!

Why not use our best guess of the θ's i.e. $\hat{\theta}^{MLE}$?

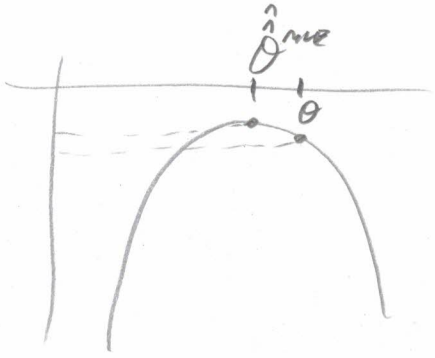
$$\Rightarrow m_* = \underset{m \in \{1 \ldots M\}}{\operatorname{argmax}} \left\{ \ell\left(\hat{\theta}^{MLE}_{m,1}, \ldots \hat{\theta}^{MLE}_{m,K_m}; X_1 \ldots X_n\right) \right\}$$

This works only if $K_1 = K_2 = \ldots = K_M$ i.e. the # params is the same in all models. If not, you're giving unfair advantage to models that have a higher # of params. Eventually as $K \to n$, the model will fit perfectly ("overfitting", a core concept in the 342 class).

> "With 4 parameters, I can fit an elephant; with 5 parameters I can make it wiggle its trunk" - John von Neumann famous mathematician, computer scientist

So we need a way to "penalize" the model by K_m, the # of parameters in the model. The penalty is to subtract the bias from $\ell\left(\hat{\theta}^{MLE}_{m,1} \ldots \hat{\theta}^{MLE}_{m,K_m}; X_1 \ldots X_n\right)$.

Why is there bias? If $K=1$



$\hat{\theta}^{MLE} \dot{\sim} N\left(\theta, \sqrt{\frac{I(\theta)^{-1}}{n}}^2\right)$

$\Rightarrow \hat{\theta}^{MLE} \neq \theta$

and by definition $\ell(\hat{\theta}^{MLE}; X_1 \ldots X_n) > \ell(\theta; X_1 \ldots X_n)$

$\Rightarrow$ over any datasets, this will always be the case $\Rightarrow E\left[\ell(\hat{\theta}^{MLE}; X)\right] > E\left[\ell(\theta; X)\right]$

It can be shown that this bias is K_m. Thus

$$\ell(\theta_{m,1} \dots \theta_{m,K_m}; X_1 \dots X_n) \approx \ell(\hat{\theta}^{MLE}_{m,1} \dots \hat{\theta}^{MLE}_{m,K_m}, X_1 \dots X_n) - K_m$$

very approximate! need large n to be true

Therefore, $m_* = \max_{m \in \{1 \dots M\}} \{ \ell_m - K_m \}$

For historical reasons we use $-2\times$ and take the min:

$$AIC_m := -2\,\ell(\hat{\theta}^{MLE}_{m,1} \dots \hat{\theta}_{m,K_m}; X_1 \dots X_n) + 2K_m$$

Note: if all models have same # of params, the highest $\ell(\hat{\theta}^{MLE}_m)$ is selected still

the model m with lowest AIC_m is "selected", AIC_{m_*}

"Akaike's Information Criterion" (Akaike, 1973)

Further $w_m := e^{-(AIC_m - AIC_{(*)})/2} = e^{-\left((-2\ell_m + 2K_m) - (-2\ell_{m_*} + 2K_{m_*})\right)/2}$

$= e^{(\ell_m - \ell_{m_*}) + (K_m - K_{m_*})} = \frac{L_m}{L_{m_*}} \frac{e^{K_m}}{e^{K_{m_*}}}$, i.e. prop of likelihood

is called the "weight".

Thm: if one of the M models is correct, the weights are the prob's of the model being true.

For small n, the AIC ~~is~~ not accurate. There is a small-n correction:

$$AICC_m := -2\ell_m + 2K_m \left(\frac{n}{n - K_m - 1} \right)$$

If all models have same # of params K, nothing changes

↑ inflates penalty for small n.

Troubles..... Paradoxes...

① During pt. estimation, we sometimes get pt. estimates we know are silly.

Eg. DGP: $X_1, \dots X_n \overset{iid}{\sim} \text{Bern}(\theta)$

$X_1=0, X_2=0, X_3=0 \quad \hat{\theta} = \bar{X} = \frac{0+0+0}{3} = 0.$

This seems overly harsh — saying the Head on the coin is impossible.

② Consider $X_1 - X_n \overset{iid}{\sim} N(\theta, \sigma^2)$

$\hat{CI}_{\theta, 1-\alpha} = \left[\bar{X} \pm t_{n-1, 1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$ which is exact...

Does this CI mean anything? You could say $P(\theta \in \hat{CI}_{\theta, 1-\alpha} \mid \vec{X}) = 1-\alpha$!

③ Consider $X_1 - X_n \overset{iid}{\sim} \text{Bern}(\theta)$

$X_1=0, X_2=1, X_3=0$

$\hat{CI}_{\theta, 95\%} = \left[\frac{1}{3} \pm 1.96\sqrt{\frac{\frac{1}{3}\frac{2}{3}}{3}}\right] = [-0.20, 0.87] \notin \Theta$ (H) which is absurd

This happened due to the approximate nature of the CI.

These approximations get us into trouble

④ Run a test. Get $p_{val} = 2\%$ ⇒ Reject H_0.

$p_{val} = P(H_0 \mid \vec{X})$? Does it mean any prob! No, it's just the smallest α that Returns H_0.

⑤ $X_1 \dots X_n \overset{iid}{\sim} \text{Bern}(\theta)$ But you know $\Theta = [0.3, 0.7] \neq [0,1]$ due to some physical limitation. No way to account for this!!!

MANY MORE...

All these problems could be solved, and more, if you allow for $p(\theta|\vec{x})$ to be a fully fledged legitimate probabilistic entity. Why the fuss? Let θ_R be the real value of θ.

Assume $X_1 \dots X_n$ discrete

$$\underbrace{p(\theta|\vec{x})}_{\text{posterior distr.}} = \frac{\overbrace{p(\vec{x}|\theta)}^{\text{joint/just been called likelihood as } = \mathcal{L}(\theta|\vec{x})}\ \overbrace{p(\theta)}^{\text{prior distr}}}{\underbrace{p(\vec{x})}_{\text{marginal likelihood distr}}} = \frac{p(\vec{x}|\theta)\, p(\theta)}{\sum_{\theta \in \Theta} p(\vec{x}|\theta)\, p(\theta)}$$

If $X_1, X_2 \dots$ cont..

$$p(\theta|\vec{x}) = \frac{f(\vec{x}|\theta)\, p(\theta)}{f(\vec{x})} = \frac{f(\vec{x}|\theta)\, p(\theta)}{\sum_{\theta \in \Theta} f(\vec{x}, \theta)} = \frac{f(\vec{x}|\theta)\, p(\theta)}{\sum_{\theta \in \Theta} f(\vec{x}|\theta)\, p(\theta)}$$

$$p(\theta) = \text{Deg}(\theta_R) = \begin{cases} 1 & \text{if } \theta = \theta_R \\ 0 & \text{if } \theta \neq \theta_R \end{cases}$$

$$\Rightarrow p(\theta|\vec{x}) = \begin{cases} \dfrac{p(\vec{x}|\theta=\theta_R)}{p(\vec{x}|\theta=\theta_R)} = 1 & \text{if } \theta = \theta_R \\ 0 & \text{if } \theta \neq \theta_R \end{cases}$$

This is the source of the intransigence. If this is accepted as non-degenerate $\Rightarrow p(\theta|\vec{x})$ becomes non-degenerate.

This is the big leap, the big pill to swallow! Can you call this a real r.v.! It represents uncertainty in θ before data is seen - you can encode any ideas into this.

Frequentists scream: this is subjective! You can allow your own ideas to influence science? Absurd! We will address this concern later.

Can θ still be considered fixed? Yes, $p(\theta|x)$ just codes the uncertainty in it. Or... you can go radical and consider θ in a quantum state ⟼ ! ∫ until measured classically...