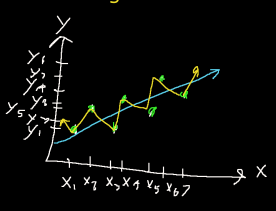
We can choose a model by computing:

$$m_* = \underset{m \in \{1,\ldots,M\}}{\arg\max} \left\{ \ell\left(\hat{\hat{\theta}}_1^{MLE}, \ldots, \hat{\hat{\theta}}_{K_m}^{MLE} ; x_i\right) \right\}$$

Unfortunately this won't generally yield the best model of the M candidates. Because the more parameters K_m in the model, the better the fit. In MATH 342W we call this "overfitting".

"With four parameters I can fit an elephant and with five I can make it wiggle its trunk!" - John von Neumann (he is famous)

You only have n observations. It K_m gets close to n, then the model begins to fit more and more perfectly.



$$y = \underbrace{\theta_1 + \theta_2 x}_{f(x;\,\theta_1,\theta_2)} + \varepsilon$$

2 params

$K_1 = 2$

$$y = \begin{cases} \theta_1 & \text{if } x = x_1 \\ \theta_2 & \text{if } x = x_2 \\ \theta_3 & \\ \theta_4 & \vdots \\ \theta_5 & \vdots \\ \theta_6 & \vdots \\ \theta_7 & \text{if } x = x_7 \end{cases} = f_2(x, \ldots)$$

$K_2 = 7$

which model do you like better? The first since it's simple and we like simple (Occam's Razor). It's also more generalizable. It can predict for each x while model 2 which fits perfectly cannot predict for each x.

Due to this overfitting phenomenon, we need to penalize candidate models by their number of parameters, K_m. How do we create this penalty? Hirotugo Akaike, a Japanese statistician who showed in 1974 the following:

$$E\left[\ell\left(\hat{\hat{\theta}}_1^{MLE}, \ldots, \hat{\hat{\theta}}_{K_m}^{MLE} ; x_1, \ldots, x_n\right)\right] > \ell\left(\theta_1, \ldots, \theta_{K_m} ; x_1, \ldots, x_n\right)$$

This means substituting the MLE's as estimates for the true parameters will bias your log-likelihood upwards, i.e. your candidate seems to fit better than it actually does (inflation).

With many assumptions, you can prove that the asymptotic bias as $n \to \infty$ is:

$$\lim_{n\to\infty} \text{Bias}\left[\ell\left(\hat{\hat{\theta}}_1^{MLE}, \ldots, \hat{\hat{\theta}}_{K_m}^{MLE} ; x_1, \ldots, x_n\right)\right] = K_m$$

It makes sense since the higher the number of parameters, the more your model seems to fit better (overfitting). Thus,

$$\ell\left(\theta_1, \ldots, \theta_{K_m} ; x_1, \ldots, x_n\right) \approx \ell\left(\hat{\hat{\theta}}_1^{MLE}, \ldots, \hat{\hat{\theta}}_{K_m}^{MLE} ; x_1, \ldots, x_n\right) - K_m$$

This bias-corrected log-likelihood can be used to select models.

For historical reasons, the rhs of the above is multiplied by -2 to yield a very famous metric called Akaike's Information Criterion (AIC),

$$AIC_m := -2\,\ell\left(\hat{\hat{\theta}}_{m,1}^{MLE}, \ldots, \hat{\hat{\theta}}_{m,K_m}^{MLE} ; x_1, \ldots, x_n\right) + 2K_m$$

Since largest log-likelihoods (i.e. those least negative or closest to zero) are the better models, the smallest AIC are the better models.

The AIC also allows us to do goal (b) which is to score each of the candidate models. Once we compute

$$AIC_1, AIC_2, \ldots, AIC_M$$

we can then select the best AIC and thus the best model:

$$AIC_* := \min\{AIC_1, \ldots, AIC_M\}, \quad m_* := \arg\min\{AIC_1, \ldots, AIC_M\}$$

we can then compute "Akaike weights" for each model:

$$w_m := \frac{e^{-(AIC_m - AIC_*)/2}}{\sum_{j=1}^{M} e^{-(AIC_j - AIC_*)/2}}$$

He showed that if the candidate model list contained the true model, then w_m is the probability that model m is the true model.

$$e^{-(AIC_m - AIC_*)/2} = e^{-\left(\left(-\cancel{2}\ell_m + \cancel{2}K_m\right) - \left(-\cancel{2}\ell_* + \cancel{2}K_*\right)\right)/\cancel{2}}$$
$$= e^{(\ell_m - \ell_*) + (K_* - K_m)}$$
$$= \frac{\mathcal{L}_m}{\mathcal{L}_*} \frac{e^{K_*}}{e^{K_m}} \quad \text{a likelihood ratio} \leq 1$$

Note: not in this course. Some people "model average" and isntead of picking one model, they use a mixture distribution based on their M models weighted by the Akaike weights i.e.

$$\bar{f}(x; \ldots) = \sum_{m=1}^{M} w_m f_m\left(x; \hat{\hat{\theta}}_{m,1}^{MLE}, \ldots, \hat{\hat{\theta}}_{m,K_m}^{MLE}\right)$$

He proved that the asymptotic bias of the log likelihood expectation is K_m. However, it is really approximate. There is a finite sample approximation to this bias that performs better:

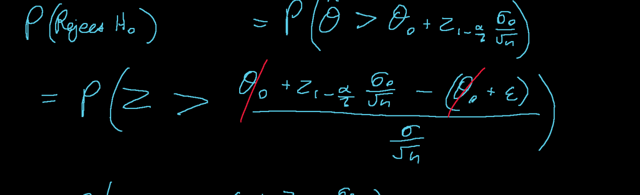
$$AICC_m := -2\,\ell\left(\hat{\hat{\theta}}_{m,1}^{MLE}, \ldots, \hat{\hat{\theta}}_{m,K_m}^{MLE} ; x_1, \ldots, x_n\right) + 2K_m\left(\frac{n}{n - K_m - 1}\right)$$

"Correction"

Akaike's Information Criterion Corrected. This is the recommended metric for this class for model selection.

New core concept: clinical / practical significance of a result.

$H_a: \theta \neq \theta_0$ vs. $H_0: \theta = \theta_0$, $\hat{\theta}$ is asymptotically normal



Let's assume that the real parameter value is $\theta = \theta_0 + \varepsilon$ where ε is a small positive value

$$P(\text{Reject } H_0) = P\left(\hat{\theta} > \theta_0 + z_{1-\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}\right)$$

$$= P\left(Z > \frac{\cancel{\theta_0} + z_{1-\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}} - (\cancel{\theta_0} + \varepsilon)}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= P\left(Z > \frac{-\varepsilon + z_{1-\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z > -\underbrace{\frac{\varepsilon}{\sigma}}_{a>0}\sqrt{n} + \underbrace{z_{1-\frac{\alpha}{2}} \frac{\sigma_0}{\sigma}}_{b>0}\right)$$

$$= P\left(Z > -a\sqrt{n} + b\right). \quad \text{Consider } \lim_{n\to\infty} P(\text{Reject } H_0)$$
$$= P(Z > -\infty) = 1$$

This calculation is valid for all ε, θ_0 values. The implication is with enough n (with sample size large enough), any hypothesis test will be rejected. This means all tests are "statistically significant" if n is large enough.

However, not all findings (estimated differences from the null) i.e. $\hat{\hat{\theta}} - \theta_0$ are "clinically significant" or "practically significant". The latter means that "for all practical purposes, the null is true and the real difference can be ignored". The former means that "in a clinical setting, the finding can be ignored" which is the terminology in medical literature.

You define in advance what "practical" / "clinical" significance means as a deviation from the null hypothesis. And then if the effect you find is less than that threshold, you don't care if it's statisticall significant.

"Multiple hypothesis testing problem" or "multiple comparisons problem".

Recall one hypothesis test and every possible outcome of its decision:

Decision

Truth	Retain	Reject
H_0	✓	Type I error
H_a	Type II error	✓

$P(\text{Type I error}) = \alpha$

this is set beforehand by you

What if you were doing many hypothesis tests? Let's say m tests where each P(Type I error) was the same at your setting α. These m tests are called a "family of tests". Among these tests, you reject R of them (a rv) and retain F := m - R of them (a rv). But you also make Type I and Type II errors (unobserved rv's). Denote the number of Type I errors by V. It's also called the # of "false negatives" or "false discoveries". Here is the table:

Decision

Truth	Retain	Reject	
H_0	U, u	V, v	m_0
H_a	T, t	S, s	m_a
	F, f	R, r	m

rv's are capital letters and realization are lowercase letters and constants are lowercase as well.

n