## Model selection

$X_1$, $X_2$, ..., $X_k$: which ones should be included?

Best R-squared? Include them all (but forget about prediction)

Highest likelihood? Remember, that is just like R-squared.

"Stepwise" regression: various schemes

> Build up: find best one-variable model (according to hypothesis test), then best two-variable model with that first variable included, etc., until no additional variables are "significant"
>
> Break down: start with all variables, remove the one that is least significant, then the next, etc., until all remaining variables are significant

Stepwise schemes do not necessarily find the best model among all the possible variable combinations

**Professor Hirotugu Akaike's approach**

Akaike (1973) and subsequent papers (see also recent book by K. P. Burnham and D. R. Anderson: *Model selection and inference: a practical information-theoretic approach*. Springer)

$f(y)$: true model (unknown) giving rise to data $y$ ($y$ is a vector of data)

$g(y; \theta)$: candidate model (parameter vector $\theta$)

Want to find a model $g(y; \theta)$ that is "close to" $f(y)$

Kullback-Leibler discrepancy:

$$K(f, g) = \mathrm{E}_f \left[ \log \left( \frac{f(Y)}{g(Y; \theta)} \right) \right]$$

This is a measure of how "far" model $g$ is from model $f$ (with reference to model $f$). Properties:

$$K(f, g) \geq 0$$

$$K(f, g) = 0 \quad \Leftrightarrow \quad f = g$$

Of course, we can never know how far our model $g$ is from $f$. But Akaike showed that we might be able to estimate something almost as good.

Suppose we have two models under consideration: $g(y, \theta)$ and $h(y, \phi)$. Akaike showed that we can *estimate*

$$K(f, g) - K(f, h)$$

(!!!). It turns out that the difference of maximized log-likelihoods, corrected for a bias, estimates the difference of KL distances. The maximized likelihoods are

$$\widehat{L}_g = g\left(y, \widehat{\theta}\right)$$

$$\widehat{L}_h = h\left(y, \widehat{\phi}\right)$$

where $\widehat{\theta}$ and $\widehat{\phi}$ are the ML estimates of the parameters.

Akaike's result:

$$\left[\log\left(\widehat{L}_g\right) - q\right] - \left[\log\left(\widehat{L}_h\right) - r\right]$$

is an asymptotically unbiased estimate (i.e. bias approaches zero as sample size increases) of $K(f, g) - K(f, h)$. Here $q$ is the number of parameters estimated in $\theta$ (model $g$) and $r$ is the number of parameters estimated in $\phi$ (model $h$).

The price of parameters: the likelihoods in the above expression are *penalized* by the number of parameters.

The Akaike Information Criterion (AIC) for model $g$:

$$\text{AIC} = -2\log\left(\widehat{L}_g\right) + 2q$$

Model selection procedure:

- Decide, on scientific grounds, upon a small suite of models to be compared (not a fishing expedition), that is, narrow down the number of variables to be considered for inclusion in a multiple regression

- Fit all possible models in the suite (ML estimates of parameters in all the models under consideration)

- For each model, calculate its AIC (remember that the number of parameters in a multiple regression is $k + 2$)

- Pick the model with the smallest AIC. That is the model in the suite with the best overall statistical properties and parameter balance

Akaike's rule of thumb: two models are essentially indistinguishable if the difference of their AICs is less than 2.

**AIC/multiple regression in SAS**

PROC RSQUARE fits all possible regression models!!!

```
DATA;
  INPUT Y X1 X2 X3 X4 X5;
  CARDS;
      ⋮
  ;
PROC RSQUARE;
  MODEL Y=X1 X2 X3 X4 X5 / AIC;
```

**Remarks on model selection with AIC**

1. Too many models (or variables) will defeat AIC properties

2. AIC puts all models on a "level playing field"

3. Picking by AIC is asymptotically equivalent to picking by "cross-validation"

4. One has to know model form, know what likelihood is, know how to calculate ML estimates

5. Normal linear models:

$$-2\log\left(\widehat{L}\right) = n\left[\log\left(\frac{2\pi \text{SS(residual)}}{n}\right) + 1\right]$$

6. AIC approach is valid for any families of statistical models: AOV (can use it for multiple comparisons), time series models, categorical data models, multivariate models, mark-recapture, ...

**Occam's razor**

> *"Quia frustra fit per plura quod potest fieri per pauciora"*
>
> (Because it is vain to do with more what can be done with less)
>
> — William of Occam

**Brian's bludgeon**

> "No amount of computing is too much, if it gets the job done"
>
> — Brian of Idaho