STAT 538 LEC. # 7

Model Choice

Say measure $(Y, X_1, \ldots, X_m)$ for n subjects.

Data aquisition may have been "fishing expedition" - don't necessarily believe all m predictors are relevant.

May seek a final model including only m^* of the predictors (or equivalently, set $m - m^*$ of the regression coefficients to zero).

First, a more focussed question. How to compare *two* models.

but also consider interactions, possibly quadratic effects - end up with p coefficients/parameters comparing 2 models

Likelihood Ratio Test (LRT)

Data D, Model M_0 (p_0 params) nested within M_1 (p_1 params).

$> p_0$

If M_0 true,

$$2\left\{l_1(\hat{\theta}_1; D) - l_0(\hat{\theta}_0; D)\right\} \overset{approx}{\sim} \chi^2_{p_1 - p_0}$$

$= D_0 - D_1$

saturated model terms cancel

Usual hypothesis testing implementation and interpretation.

can be more reliable than

comparing $\dfrac{\hat{\beta}_j - 0}{\widehat{SE}(\hat{\beta}_j)}$ to $N(0,1)$

Wald test

Akaike

$\boxed{\text{?IC}}$

An Information Criterion (AIC)

No requirement that competing models be nested.

Choose the one maximizing

maximized log-likelihood under ith model

$$2\, l_i(\hat{\theta}_i; D) - 2\, p_i,$$

again # parameters in ith model

i.e., notion of *complexity penalty*.

Motivated as a measure of *predictive performance*.

Bayesian Information Criterion (BIC)

Again no nesting requirement.

Choose the model maximizing

was 1 with AIC

$$l_i(\hat{\theta}_i; D) - \boxed{\{(1/2)\log n\}}p_i,$$

i.e., bigger complexity penalty than AIC, especially for large samples.

Rationale: Bayesian - somewhat crude approximation to choosing the model for which Pr(Model i is true|Data=D) is largest.

Practical Difference - YES

Xj should be in the model no/yes

Say comparing M_0 and M_1, nested, with $p_1 = p_0 + 1$. Choose M_1 if $2\{l_1(\hat{\theta}_1; D) - l_0(\hat{\theta}_0; D)\} > c$.

LRT: $c = \chi^2_1$ quantile, i.e.,

$$c = \begin{cases} 2.71 & 10\% \text{ sig.}, \\ 3.84 & \underline{5\% \text{ sig.}}, \\ 6.63 & 1\% \text{ sig.} \end{cases}$$

liberal

AIC: $c = 2$.

BIC: $c = \log n$, i.e.,

conservative

$$c = \begin{cases} \underline{4.6} & \text{if } n = 100, \\ 6.2 & \text{if } n = 500, \\ 6.9 & \text{if } n = 1000. \end{cases}$$

Comparing all possible models

?IC can compare any collection of models.

There are 2^m subsets of m predictor variables.

Fitting 2^{10} models may be tolerable. ← ≈ 1000

Fitting 2^{20} models may not be. ← 1000000

Situation worse if want to consider possibilities of 'curved' effects and/or interactions.

e.g. interactions: m physical variables, but $m + m(m-1)/2$ possible predictors for inclusion/exclusion.

Motivates **stepwise** procedures. Search for models with high values of criterion function without evaluating all possible models.

one of many stepwise implementations - details differ

stepAIC() in R (part of MASS library)

Iterative scheme.

From current model, consider all possible ‘one-term deletions‘ (backward) AND/OR ‘one-term additions (forward).’

Of these, the new model is the one with the best improvement in AIC (or BIC).

Iterate this scheme until no such changes improve AIC.

Practical, but no guarantee of global max.

$y = \begin{cases} 1 & \text{'low'} \\ 0 & \text{'normal'} \end{cases}$ birthweight

n=189

8 predictors of various types

Stepwise Example (low birthweight dataset)

```
fit0 <- glm(low ~ ., family=binomial, data=bwt)
```

starting model

```
fit1 <- stepAIC(fit0, ~.)
```

scope

```
fit2 <- stepAIC(fit0, ~.^2 + I(scale(age)^2) +
                             I(scale(lwt)^2)   )
```

changing initial model

Same?

```
fit3 <- stepAIC(fit1, ~.^2 + I(scale(age)^2) +
                             I(scale(lwt)^2)  )
```

```
> summary(fit0)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.82302    1.24471   0.661  0.50848
age          -0.03723    0.03870  -0.962  0.33602
lwt          -0.01565    0.00708  -2.211  0.02705 *
raceblack     1.19241    0.53597   2.225  0.02609 *
raceother     0.74069    0.46174   1.604  0.10869
smokeTRUE     0.75553    0.42502   1.778  0.07546 .
ptdTRUE       1.34376    0.48062   2.796  0.00518 **
htTRUE        1.91317    0.72074   2.654  0.00794 **
uiTRUE        0.68019    0.46434   1.465  0.14296
ftv1         -0.43638    0.47939  -0.910  0.36268
ftv2+         0.17901    0.45638   0.392  0.69488
---
Residual deviance: 195.48  on 178  degrees of freedom
AIC: 217.48
```

mother

prev. premature labour

high BP

1st trimester visits to physician

$n_i = 1$

can't trust deviance GoF test

```
> fit1$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
low ~ age + lwt + race + smoke + ptd + ht + ui + ftv

Final Model:
low ~ lwt + race + smoke + ptd + ht + ui


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                       178   195.4755 217.4755
2 - ftv  2 1.358185     180   196.8337 214.8337
3 - age  1 1.017866     181   197.8516 213.8516
```

D + 2 # params

bigger scope

```
> fit2$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
low ~ age + lwt + race + smoke + ptd + ht + ui + ftv

Final Model:
low ~ age + lwt + smoke + ptd + ht + ui + ftv + age:ftv +
                                                 smoke:ui


        Step Df  Deviance Resid. Df Resid. Dev      AIC
1                               178   195.4755 217.4755
2   + age:ftv  2 12.474896       176   183.0006 209.0006
3 + smoke:ui  1  3.056805       175   179.9438 207.9438
4      - race  2  3.129586       177   183.0734 207.0734
```

```
> fit3$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
low ~ lwt + race + smoke + ptd + ht + ui

Final Model:
low ~ lwt + race + smoke + ptd + ht + ui

  Step Df Deviance Resid. Df Resid. Dev      AIC
1                        181   197.8516 213.8516
```

6 predictors

final model can depend on the initial model - can't expect to find the global minimum

```
> summary(fit2)
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.582374   1.421613  -0.410 0.682057
age                 0.075539   0.053967   1.400 0.161599
lwt                -0.020373   0.007497  -2.717 0.006580 **
smokeTRUE           0.780044   0.420385   1.856 0.063518 .
ptdTRUE             1.560317   0.497001   3.139 0.001693 **
htTRUE              2.065696   0.748743   2.759 0.005800 **
uiTRUE              1.818530   0.667555   2.724 0.006446 **
ftv1                2.921088   2.285774   1.278 0.201270
ftv2+               9.244907   2.661497   3.474 0.000514 ***
age:ftv1           -0.161824   0.096819  -1.671 0.094642 .
age:ftv2+          -0.411033   0.119144  -3.450 0.000561 ***
smokeTRUE:uiTRUE -1.916675   0.973097  -1.970 0.048877 *
Residual deviance: 183.07  on 177  degrees of freedom
AIC: 207.07
```

Now - try the same stepwise procedures using BIC.

```
> fit1a <- stepAIC(fit0, ~., k=log(nrow(bwt)))
> fit1a$anova
Initial Model:
low ~ age + lwt + race + smoke + ptd + ht + ui + ftv
Final Model:
low ~ lwt + ptd + ht
```

sample size

scope

main effects only

replacing default k=2

BIC being more conservative

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1			178	195.4755	253.1347
2 - ftv	2	1.358185	180	196.8337	244.0094
3 - age	1	1.017866	181	197.8516	239.7855
4 - race	2	7.614209	183	205.4658	236.9163
5 - smoke	1	2.046576	184	207.5124	233.7211
6 - ui	1	2.611024	185	210.1234	231.0904

```
> fit2a <- stepAIC(fit0, ~.^2 +  I(scale(age)^2) +
                         I(scale(lwt)^2), k=log(nrow(bwt)))
> fit2a$anova
Initial Model:
low ~ age + lwt + race + smoke + ptd + ht + ui + ftv
Final Model:
low ~ lwt + ptd + ht

                                                        BIC
     Step Df Deviance Resid. Df Resid. Dev         AIC
1                           178   195.4755 253.1347
2    - ftv  2 1.358185       180   196.8337 244.0094
3    - age  1 1.017866       181   197.8516 239.7855
4   - race  2 7.614209       183   205.4658 236.9163
5 - smoke  1 2.046576       184   207.5124 233.7211
6     - ui  1 2.611024       185   210.1234 231.0904
```

More purely empirical model comparison?

CROSS-VALIDATION

- Randomly split data intro *training* (T) and *validation* (V) cases.
- Fit model to (X_T, Y_T) data.
- Use the fitted model to generate predictions Y_V^* given X_V.

How close is Y_V^* to the actual Y_V?

One formalization - pick the model M for which

$$\log f_M(y_V | x_V, \hat{\theta}_T)$$

is largest.

Biggest model doesn't necessarily win.

Sensitivity to random split?

k-fold cross-validation

Randomly split cases into k blocks: (Y_j, X_j), $j = 1, \dots, k$.

Let $(Y_{(j)}, X_{(j)})$ denote all data except (Y_j, X_j).

Do cross-validation k times, each time with $k-1$ blocks as training data, one block as validation data.

Aggregate results. For instance choose model for which

$$\sum_{j=1}^{k} \log f_M(y_j | x_j, \hat{\theta}_{(j)})$$

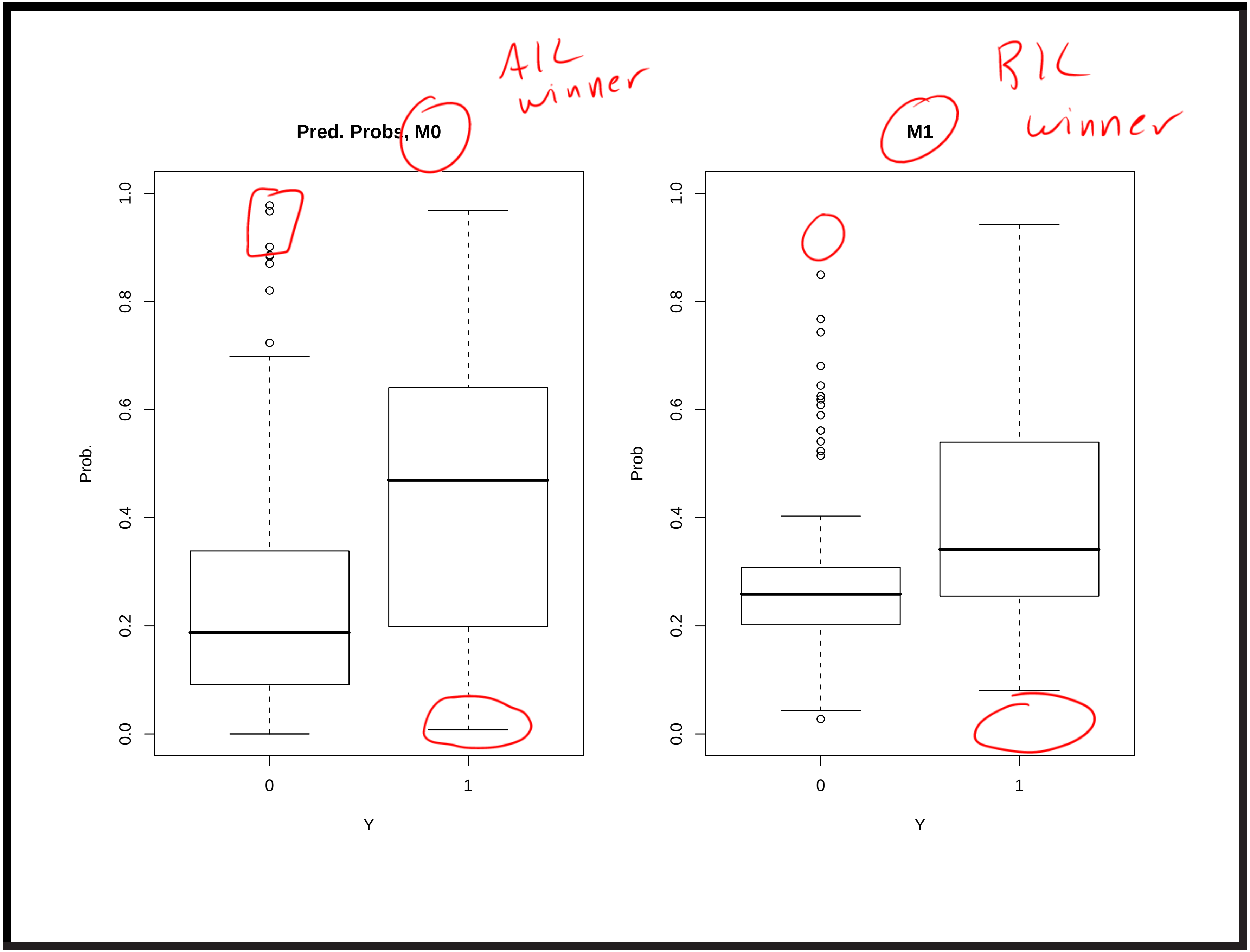
is largest.

Ex.: Compare our AIC and BIC champions.

```
### randomly assign 189 subjects to five blocks
ind <- sample( c(rep(1,38), rep(2,38), rep(3,38),
                 rep(4,38), rep(5,37)) )

for (i in 1:5) {
### fit models to all but i-th block
                  AIC
m0 <- glm(low~age+lwt+smoke+ptd+ht+ui+ftv+age:ftv+smoke:ui,
          family=binomial, data=bwt, subset=(ind!=i) )
m1 <- glm(low~lwt+ptd+ht,  BIC
          family=binomial, data=bwt, subset=(ind!=i) )

### predicted prob(Y=1) for i-th block
ftpr0[ind==i] <- predict(m0, newdata=bwt,
                         type="response")[ind==i]
ftpr1[ind==i] <- predict(m1, ...)
```

```
### predictive log-likelihoods and magnitude of diff.
> predll0 <- sum(as.numeric(bwt$low)*log(ftpr0) +
                (1-as.numeric(bwt$low))*log(1-ftpr0))
> predll1 <- sum(as.numeric(bwt$low)*log(ftpr1) +
                (1-as.numeric(bwt$low))*log(1-ftpr1))

> c(predll0, predll1)
 -341.5994 -274.8250

> exp((predll0-predll1)/189)
 0.7023638
```

HUGE preference for second (smaller, BIC-champ.) model. Why?


AIC winner
BIC winner
Pred. Probs, M0
M1
Prob.
Prob
0.0
0.2
0.4
0.6
0.8
1.0
0
1
Y