Variable Selection and Model Choice

EPI 204

Quantitative Epidemiology III

Statistical Models

Study Design

- Part of study design is deciding which variables to measure/collect.
- Given a data set, you may wish to use all of the variables or to select or reclassify. Often this is based on practical considerations such as sample size—you don't want 50 variables with 100 observations if you can help it.
- Often you will have prior ideas about which variables are important, and if their effects may differ among subjects.

Variables, transformations, interactions

- Numeric variables such as age, or cholesterol level may need some form of transformation.
- You can check for non-linearity by categorizing the variable and running an analysis and then graphing the coefficients. Are they increasing? Decreasing? Flat for some intervals?
- If the variable covers an order of magnitude or more (as some molecular measurements may), then maybe take logs.
- Use the numeric variable if the coefficients are roughly linear on the raw or log scale, otherwise consider categorization.

Variables, transformations, interactions

- Check for interaction terms. The R commands `drop1()` and `add1()` are handy for this.
- Interaction terms are very common, because the effects of an exposure or covariate can depend strongly on other factors.
- Three way interactions and higher are less common but can occur.
- We can decide to add an interaction based either on a significance test or something like the AIC.

Criteria for Model Choice

- This depends strongly on the purpose. The AIC and related measures try to maximize the predictive ability of the model.
- Significance tests try to choose relationships that may be directly or indirectly causal.
- Some parts of this can be automated, but it is always best to know what is likely to be important from subject matter knowledge.

AIC and BIC

- AIC = –2 log(likelihood) + k*p, where p is the number of parameters.
- AIC = Akaike Information Criterion
- AIC = deviance + k*p (omitting the log likelihood of the saturated model from all AICs for a given data set).
- The parameter k = 2 by default.
- For the BIC (Bayesian Information Criterion), k = log(n).

Model Comparison Statistics

```
drop1(object, scope, test = c("none", "Rao", "LRT",
"Chisq", "F"),k = 2, ...)
```

object returned from a glm() statement

scope is the set of things that can be dropped. Usually, this is omitted, meaning that any term in the model can be dropped, though hierarchy of interactions is respected automatically.

test is for logistic regression usually "Chisq" or "LRT", both of which produce the likelihood ratio test.

k is the AIC parameter.

Model Comparison Statistics

```
add1(object, scope, test = c("none", "Rao", "LRT",
"Chisq", "F"),k = 2, ...)
```

object returned from a glm() statement

scope is the formula of a larger object whose single terms may be added, with hierarchy of interactions respected.

test is for logistic regression usually "Chisq" or "LRT", both of which produce the likelihood ratio test.

k is the AIC parameter.

```
> summary(glm(CHD ~ CAT+CHL+SMK+HPT,family=binomial,data=evans))


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.975282   0.789031  -6.306 2.87e-10 ***
CAT          1.021916   0.310172   3.295 0.000985 ***
CHL          0.008963   0.003254   2.754 0.005885 **
SMK          0.714577   0.298261   2.396 0.016583 *
HPT          0.483481   0.287653   1.681 0.092805 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 438.56  on 608  degrees of freedom
Residual deviance: 406.52  on 604  degrees of freedom
AIC: 416.52

Number of Fisher Scoring iterations: 5
```

```
> drop1(glm(CHD~CAT+CHL+SMK+HPT,family=binomial,data=evans),test="Chisq")

       Df Deviance    AIC     LRT Pr(>Chi)
<none>      406.52 416.52
CAT     1   417.10 425.10 10.5822 0.001142 **
CHL     1   414.05 422.05  7.5383 0.006040 **
SMK     1   412.76 420.76  6.2472 0.012439 *
HPT     1   409.34 417.34  2.8273 0.092674 .

drop1(glm(CHD~CAT+CHL+SMK+HPT,binomial,data=evans),test="Chisq",k=log(609))

       Df Deviance    AIC     LRT Pr(>Chi)
<none>      406.52 438.57
CAT     1   417.10 442.74 10.5822 0.001142 **
CHL     1   414.05 439.70  7.5383 0.006040 **
SMK     1   412.76 438.41  6.2472 0.012439 *
HPT     1   409.34 434.99  2.8273 0.092674 .
```

HPT is not significant. AIC would keep it, BIC would delete it.

```
> model1 <- glm(CHD ~ CAT+CHL+SMK+HPT,family=binomial,data=evans)
> drop1(model1,test="Chisq")

       Df Deviance    AIC     LRT Pr(>Chi)
<none>      406.52 416.52
CAT     1   417.10 425.10 10.5822 0.001142 **
CHL     1   414.05 422.05  7.5383 0.006040 **
SMK     1   412.76 420.76  6.2472 0.012439 *
HPT     1   409.34 417.34  2.8273 0.092674 .
> add1(model1,scope=~CAT*CHL*SMK*HPT,test="Chisq")

        Df Deviance    AIC    LRT  Pr(>Chi)
<none>       406.52 416.52
CAT:CHL  1   362.06 374.06 44.456 2.601e-11 ***
CAT:SMK  1   405.96 417.96  0.556  0.455760
CHL:SMK  1   405.10 417.10  1.414  0.234449
CAT:HPT  1   399.88 411.88  6.638  0.009982 **
CHL:HPT  1   406.33 418.33  0.183  0.668555
SMK:HPT  1   406.44 418.44  0.074  0.784899
```

Best AIC from adding CAT:CHL, next best from adding CAT:HPT.

HPT not significant, but AIC and significant interaction don't suggest dropping.

```
> model2 <- glm(CHD ~ CAT*CHL+SMK+HPT,family=binomial,data=evans)
> drop1(model2,test="Chisq")

        Df Deviance    AIC    LRT  Pr(>Chi)
<none>       362.06 374.06
SMK      1   367.46 377.46  5.405   0.02007 *
HPT      1   367.16 377.16  5.098   0.02395 *
CAT:CHL  1   406.52 416.52 44.456 2.601e-11 ***

> add1(model2,scope=~CAT*CHL*SMK*HPT,test="Chisq")

        Df Deviance    AIC    LRT Pr(>Chi)
<none>       362.06 374.06
CAT:SMK  1   361.61 375.61 0.4517 0.501538
CHL:SMK  1   360.50 374.50 1.5591 0.211797
CAT:HPT  1   352.92 366.92 9.1397 0.002501 **
CHL:HPT  1   357.01 371.01 5.0493 0.024636 *
SMK:HPT  1   361.94 375.94 0.1206 0.728367
```

Best AIC is from adding CAT:HPT, tests and AIC do not suggest dropping.

```
> model3 <- glm(CHD ~ CAT*CHL+SMK+CAT*HPT,family=binomial,data=evans)
> drop1(model3,test="Chisq")

        Df Deviance    AIC    LRT  Pr(>Chi)
<none>       352.92 366.92
SMK      1   357.93 369.93  5.011  0.025188 *
CAT:CHL  1   399.88 411.88 46.958 7.253e-12 ***
CAT:HPT  1   362.06 374.06  9.140  0.002501 **

> add1(model3,scope=~CAT*CHL*SMK*HPT,test="Chisq")

        Df Deviance    AIC    LRT Pr(>Chi)
<none>       352.92 366.92
CAT:SMK  1   352.29 368.29 0.6244   0.4294
CHL:SMK  1   351.43 367.43 1.4867   0.2227
CHL:HPT  1   348.80 364.80 4.1233   0.0423 *
SMK:HPT  1   352.66 368.66 0.2582   0.6114
```

CHL:HPT reduces AIC and is statistically significant

```
> model4 <- glm(CHD ~ CAT*CHL+SMK+CAT*HPT+CHL:HPT,binomial,data=evans)
> drop1(model4,test="Chisq")

        Df Deviance    AIC    LRT  Pr(>Chi)
<none>       348.80 364.80
SMK      1   353.94 367.94  5.143  0.023336 *
CAT:CHL  1   399.85 413.85 51.050 9.005e-13 ***
CAT:HPT  1   357.01 371.01  8.214  0.004158 **
CHL:HPT  1   352.92 366.92  4.123  0.042298 *

> add1(model4,scope=~CAT*CHL*SMK*HPT,test="Chisq")

            Df Deviance    AIC     LRT Pr(>Chi)
<none>           348.80 364.80
CAT:SMK      1   348.21 366.21 0.58432   0.4446
CHL:SMK      1   347.75 365.75 1.04800   0.3060
SMK:HPT      1   348.46 366.46 0.34086   0.5593
CAT:CHL:HPT  1   348.78 366.78 0.01918   0.8899
```

No addition is significant or lowers the AIC, no removal is suggested either.

Stepwise Logistic Regression

- Repeatedly adds/drops terms from the model, one at a time, working to minimize the AIC, until it cannot be further improved.
- Backward stepwise only drops terms.
- Forward stepwise only adds terms.
- Default is “both” which can either add or drop, whichever provides the most improvement.
- SAS uses significance tests for adding and dropping terms. This is all specified in the MODEL statement using SELECTION=FORWARD, BACKWARD, or STEPWISE.

Stepwise Logistic Regression

```
step(object, scope, direction = c("both", "backward",
"forward"), trace = 1, keep = NULL, steps = 1000, k = 2)
```

object an lm or glm object. This is used as the initial model in the stepwise search.

scope defines the range of models examined in the stepwise search. This should be either a single formula, or a list containing components upper and lower, both formulae.

direction the mode of stepwise search, can be one of "both", "backward", or "forward", with a default of "both". If the scope argument is missing the default for direction is "backward".

Stepwise Logistic Regression

```
step(object, scope, direction = c("both", "backward",
"forward"), trace = 1, keep = NULL, steps = 1000, k = 2)
```

trace if positive, information is printed during the running of step. Larger values may give more detailed information.

steps the maximum number of steps to be considered. The default is 1000 (essentially as many as required). It is typically used to stop the process early.

k the multiple of the number of degrees of freedom used for the penalty. Only k = 2 gives the genuine AIC: k = log(n) is sometimes referred to as BIC or SBC.

```
> step(model1,scope=~CAT*CHL*SMK*HPT)
......
Step:  AIC=364.8
CHD ~ CAT + CHL + SMK + HPT + CAT:CHL + CAT:HPT + CHL:HPT

              Df Deviance    AIC
<none>             348.80 364.80
+ CHL:SMK      1   347.75 365.75
+ CAT:SMK      1   348.21 366.21
+ SMK:HPT      1   348.46 366.46
+ CAT:CHL:HPT  1   348.78 366.78
- CHL:HPT      1   352.92 366.92
- SMK          1   353.94 367.94
- CAT:HPT      1   357.01 371.01
- CAT:CHL      1   399.85 413.85


Coefficients:
(Intercept)          CAT          CHL          SMK          HPT      CAT:CHL
  -3.981678   -13.723211     0.003506     0.712280     4.603360     0.075636

CAT:HPT      CHL:HPT
-2.158014    -0.016542

Degrees of Freedom: 608 Total (i.e. Null);  601 Residual
Null Deviance:      438.6
Residual Deviance: 348.8        AIC: 364.8
```

Hazards of Variable Selection

- Variable selection can capitalize on chance, meaning that it produces relationships that randomly look good and will occur if we can select from many variables.
- Even if there is no association at all between any variable and the disease, often a few variables will produce apparently excellent results, that will not generalize at all
- This can sometimes be accounted for in cross validation, though this is rarely done in epidemiological studies.
- It is worse if there are many possible variables.
- It is worse if there are other modeling choices that can be made.
- "Garden of Forking Paths." Jorge Luis Borges short story and metaphor by Andrew Gelman.

Generate 100 obs with 10 unrelated variables

```
p1 <- 10
y <- rep(0:1,each=50)
x <- matrix(rnorm(p1*100),nrow=p1)
```

Generate 10 named variables x1, x2, ...

```
for (i in 1:p1)
{
  assign(paste("x",i,sep=""),x[i,])
}
------
> paste("z",12,sep="")
[1] "z12"

> assign("z12",1:5)

> z12
[1] 1 2 3 4 5
```

Generate full model formula+interactions

```
for (i in 1:p1)
{
  assign(paste("x",i,sep=""),x[i,])
}

fchar1 <- "y~x1"
fchar2 <- fchar1
for (i in 2:p1)
{
  fchar1 <- paste(fchar1,"+x",i,sep="")
  fchar2 <- paste(fchar2,"*x",i,sep="")
}
form1 <- as.formula(fchar1)
form2 <- as.formula(fchar2)
------

> fchar1
[1] "y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10”
> form1
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10
```

```
step1.res <- step(glm(form1,binomial),trace=0)
step2.res <- step(glm(form1,binomial),
                  list(lower=form1,upper=form2),trace=0)
print(summary(step1.res))
print(summary(step2.res))
-----------------
```

With 10 variables, 4 were selected, though none was statistically significant.

With the interactions possible, all 10 variables were kept with 10 two-way interactions out of 45.

One variable and 6 interactions were statistically significant.

This problem will clearly get worse as the number of variables increases.

Overfitting

- When we fit a statistical model to data, we adjust the parameters so that the fit is as good as possible and the errors are as small as possible
- Once we have done so, the model may fit well, but we don't have an unbiased estimate of how well it fits if we use the same data to assess as to fit.
- This is worse if we select variables/interactions.
- After substantial model selection, p-values are no longer meaningful.

Homework (Due 4/15/2021)

- Possibly using all the variables in the Evans County data set, make the best model you can for CHD. Use explicit interactions, not CH and CC.
- Assume that catecholamine levels are the primary exposure variable of interest.
- Consider interactions, especially if the effect is large.
- But try not to overfit.
- Then interpret the resulting model, especially recognizing confounders, effect modifiers, and control variables.