Model selection
560 Hierarchical modeling

Peter Hoff

Statistics, University of Washington
Modeling choices

**Model:** A *statistical model* is a set of probability distributions for your data.

- In HLM, the model is a specification of fixed effects and random effects.
- Once we select a model, we can estimate the parameters in the model and make further inference.

```r
nels[1:5,]
```

<table>
<thead>
<tr>
<th>school</th>
<th>enroll</th>
<th>flp</th>
<th>public</th>
<th>urbanicity</th>
<th>hwh</th>
<th>ses</th>
<th>mscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>2</td>
<td>-0.23</td>
</tr>
<tr>
<td>2</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>0</td>
<td>0.69</td>
</tr>
<tr>
<td>3</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>4</td>
<td>-0.68</td>
</tr>
<tr>
<td>4</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>5</td>
<td>-0.89</td>
</tr>
<tr>
<td>5</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>3</td>
<td>-1.28</td>
</tr>
</tbody>
</table>

**What kinds of effects could we include?**

- fixed effects: `enroll, flp, public, urbanicity, hwh, ses`
- random effects: `hwh, ses`
- fixed effect interactions: `enroll*flp, public*flp, ...`
- random effect interactions: `hwh*ses`
- higher order terms: `ses^2, ...`
Model selection

We would like a procedure that can identify the “best” model from the data.

- “best=true” if the truth is one of the potential models.
- “best” means giving the best prediction or description otherwise.

Setup: Let $M_1, M_2, \ldots, M_K$ be candidate models. For example, maybe

- $M_1$: $y \sim flp$
- $M_2$: $y \sim flp + ses$
- $M_3$: $y \sim flp + ses + (ses|school)$

Model selection procedure: A procedure that takes data $(y, X)$ as input and outputs a model.

$$msel(y, X) \in \{M_1, \ldots, M_K\}$$
Consistent model selection

As our data are subject to sampling variability, we can’t expect a model selection procedure to select the best model with probability 1. However, we do expect that

\[ \Pr(\text{msel}(y, X) = M_k) \] is large if \( M_k \) is correct.

As more data comes in, a good procedure should have an increasingly large chance of selecting the right model. Such a procedure is consistent.

**Consistency:** \( \text{msel}(y, X) \) is consistent if

when \( M_k \) is true, then \( \Pr(\text{msel}(y, X) = M_k) \to 1 \) as \( n, m \to \infty \).

Unfortunately, model selection based on \( p \)-values is *not consistent*. 


**Diabetes example:**

- 442 subjects
- $y_i = \text{diabetes progression}$
- $x_i = \text{explanatory variables}$.

**Each $x_i$ includes**

- 13 subject specific measurements ($x_{\text{age}}, x_{\text{sex}}, \ldots$);
- $78 = \binom{13}{2}$ interaction terms ($x_{\text{age}} \cdot x_{\text{sex}}, \ldots$);
- 9 quadratic terms ($x_{\text{sex}}$ and three genetic variables are binary)

100 explanatory variables total!
Backwards elimination

1. Obtain the estimator $\hat{\beta}_{ols} = (X^T X)^{-1} X^T y$ and its $t$-statistics.
2. If there are any regressors $j$ such that $|t_j| < t_{cutoff}$,
   2.1 find the regressor $j_{min}$ having the smallest value of $|t_j|$;
   2.2 remove column $j_{min}$ from $X$;
   2.3 return to step 1.
3. If $|t_j| > t_{cutoff}$ for all variables $j$ remaining in the model, then stop.
### Backwards elimination

```r
ZSTATS <- NULL; zmin <- 0; zcut <- qnorm(.975)
while (zmin < zcut)
{
    fit <- lm(y ~ -1 + XS)
    zscore <- summary(fit)$coef[,3]

    zmin <- min(abs(zscore))
    if (zmin < zcut)
    {
        jmin <- which.min(abs(zscore))
        XS <- XS[, -jmin]
    }

    zs <- rep(0, ncol(X))
    zs[ match(substr(names(zscore), 3, 9), colnames(X)) ] <- zscore
    ZSTATS <- rbind(ZSTATS, zs)
}
###
```
Backwards elimination

Initial $z$-scores:

![Graph showing initial z-scores with predictor on the x-axis and zstat on the y-axis. The graph displays a series of vertical bars representing the magnitude of the z-scores for different predictor values.]
Backwards elimination

After ten iterations:
Backwards elimination

After twenty iterations:

![Graph showing z-stat values for predictor variables after 20 iterations. The graph has z-stat values on the y-axis and predictor values on the x-axis. The graph shows a range from predictor 0 to 100 on the x-axis and values from -4 to 8 on the y-axis. The horizontal dashed lines indicate critical z-stat values.](image-url)
Backwards elimination

Final solution:

```
  predictor  zstat
   0    20   40   60   80  100
  -4   -2    0    2    4    6    8
```

11/41
summary(fit)

## Call:
## lm(formula = y ~ -1 + XS)
##
## Residuals:
##    Min 1Q Median 3Q Max
##-2.02487 -0.50086 -0.02309 0.39881 1.82902
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## XSsex            -0.14960   0.03613  -4.141   4.16e-05 ***
## XSbmi             0.30871   0.03962   7.792   5.01e-14 ***
## XSmapi           -0.02309    0.02309 -1.000  0.316593    *
## XStc            -3.46601   0.93093  -3.723  0.000223 ***
## XSld1            2.91438   0.81340   3.583  0.000379 ***
## XShdl             1.11968   0.35008   3.198  0.001484 **
## XSltg            1.56747   0.33733   4.647  4.49e-06 ***
## XSglu2           -0.17921   0.03421  -5.239  2.54e-07 ***
## XSsex.age         0.12978   0.03297   3.936  9.66e-05 ***
## XSmapi.bmi       -0.02309    0.02309 -1.000  0.316593    *
## XSltg2            0.94347   0.13505   7.000  2.94e-11 ***
## XShdl2            0.08099   0.03336   2.428  0.015609 *
##
## Signif. codes:  < 0.001 '***' 0.001 '**' 0.01 '*' 0.05 '.' 1
##
## Residual standard error: 0.6752 on 430 degrees of freedom
## Multiple R-squared:  0.5554, Adjusted R-squared:  0.543
## F-statistic: 44.77 on 12 and 430 DF,  p-value: < 2.2e-16

How would you interpret the p-values, standard errors, CIs?
A problem with backwards selection

Let $y_{\pi}$ be a permutation of $y$, eg.

$$y = (2.2, -1.2, 0.5, \ldots, -0.7)$$
$$y_{\pi} = (0.5, -0.7, 2.2, \ldots, -1.2)$$

**Question:** What is the relationship between $y_{\pi}$ and $X$?

**Question:** What would happen if we did backwards elimination on $y_{\pi} \sim X$?
yp<-sample(y)
XS<-X

### backwards elimination
ZSTATS<-NULL ; zmin<-0 ; zcut<-qnorm(.975)
while(zmin< zcut)
{
  fit<-lm(yp~ -1+XS)
  zscore<-summary(fit)$coef[,3]
  zmin<-min(abs(zscore))
  if(zmin<zcut)
  {
    jmin<-which.min(abs(zscore))
    XS<-XS[,,-jmin]
  }
  zs<-rep(0,ncol(X))
  zs[ match(substr(names(zscore),3,9),colnames(X)) ] <-zscore
  ZSTATS<-rbind(ZSTATS,zs)
}

###
Backwards elimination

Initial $z$-scores:

```
<table>
<thead>
<tr>
<th>predictor</th>
<th>zstat</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-4</td>
</tr>
<tr>
<td>20</td>
<td>-2</td>
</tr>
<tr>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>80</td>
<td>4</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
</tr>
</tbody>
</table>
```

15/41
Backwards elimination

After 10 iterations:
Backwards elimination

After twenty iterations:
Backwards elimination

Final solution:

```
<table>
<thead>
<tr>
<th>Predictor</th>
<th>Zstat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>
```

18/41
**Final solution**

```r
summary(fit)

## Call:
## lm(formula = yp ~ -1 + XS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##-1.8359 -0.7679 -0.1648  0.6944  2.5759
##
## Coefficients:
##            Estimate Std. Error t value  Pr(>|t|) 
## XSsex    -0.10864   0.05148  -2.110  0.035414 *
## XStc     0.36861    0.12127   3.040  0.002517 **
## XShdl   -0.47080    0.15785  -2.982  0.003026 **
## XStch   -0.54173    0.20150  -2.688  0.007464 **
## XStc.age -0.49277    0.15995  -3.081  0.002201 **
## XStc.map  0.53362    0.15753   3.387  0.000772 ***
## XSldl.age  0.55720    0.16124   3.456  0.000716 ***
## XSldl.map -0.45468    0.15914  -2.857  0.004485 *
## XSldl.tc -0.58580    0.27615  -2.121  0.034485 *
## XShdl.age  0.17613    0.06262   2.813  0.005142 **
## XShdl.map -0.18940    0.07789  -2.429  0.015547 *
## XShdl.tch  0.22954    0.10399   2.207  0.027830 *
## XStch.bmi  0.14459    0.05421   2.667  0.007945 **
## XStc2     0.62743    0.28485   2.203  0.028161 *
## XStch2   0.25008    0.08824   2.834  0.004818 **
##
## Residual standard error: 0.9629 on 420 degrees of freedom
## Multiple R-squared:  0.1169,  Adjusted R-squared:  0.07064
## F-statistic: 2.527 on 22 and 420 DF,  p-value: 0.0001975
```

Would you interpret the p-values, standard errors, CIs?
Inconsistency of backwards elimination

Backwards elimination (and forwards selection) generally rely on a comparison of models based on a $p$-value.

$M_1$: $y \sim x_1 + x_2 + x_3$

$M_0$: $y \sim x_1 + x_2$

Variable $x_3$ is eliminated if

- its $z$-score is $< 1.96$ in absolute value
- (more or less) equivalently, if the $p$-value from the LRT is $> 0.05$. 
Inconsistency of backwards elimination

Now suppose $M_0$ is true. What is the probability of selecting $M_1$?

$$\Pr(\text{bsel}(y, X) = M_1|M_0) = \Pr(\text{reject } M_0|M_0) = \text{type I error rate} = \Pr(p - \text{value} > 0.05|M_0) = 0.05$$

This does not change as $m, n \to \infty$.

(Actually, for the LRT the probability gets closer to 0.05 as $m, n \to \infty$).
Problems with backwards elimination

There are other problems with backwards elimination (and forwards selection):

**Problem 1:** The method doesn’t search over all possible models.

**Problem 2:** The resulting $p$-values and standard errors may be misleading.

**Problem 3:** The model selection procedure is *not consistent*

Problems 1-2 are issues for any model selection procedure.

However, some model selection procedures do not have problem 3.
Building a better model selection procedure

Suppose only two models are under consideration, $M_0$ and $M_1$.
Maximize the likelihoods under each model:

$$l_1 = \log p(y|\hat{\theta}_1)$$
$$l_0 = \log p(y|\hat{\theta}_0)$$

If $l_1$ is much bigger than $l_0$, then it makes sense to prefer $M_1$ to $M_0$.

However, recall that if

- $M_0$ is nested in $M_1$, or
- $M_0$ has many fewer parameters than $M_1$,

then $l_1$ with always/typically be larger than $l_0$. 
Building a better model selection procedure

**Idea:** Prefer $M_1$ to $M_0$ if

- $l_1$ is bigger than $l_0$ by an amount that depends on $p_0, p_1$.
- $l_1 - l_0 > c_{p_0, p_1}$

This should remind you of the LRT, where we prefer $M_1$ to $M_0$ if

$$\lambda = 2 \times (l_1 - l_0) > q_{p_0, p_1},$$

where $q_{p_0, p_1}$ is a quantile of the appropriate null distribution.

**Exercise:** Show that the LRT procedure has the above form.
LRT as a model selection procedure

**LRT:** Reject $M_0$, favor $M_1$ if

$$
\lambda = 2 \times (l_1 - l_0) > \chi^2_{\rho_1 - \rho_0, 0.95}
$$

$$
l_1 - l_0 > \frac{1}{2} \chi^2_{\rho_1 - \rho_0, 0.95} = c_{\rho_1, \rho_0}
$$

**Problem:** If $M_0$ is true, probability of selecting $M_1$ is $\approx 0.05$, regardless of $m, n$.

Model selection via hypotheses test is *not consistent*. 
Consider *any* procedure that prefers $M_1$ to $M_0$ if

$$l_1 - l_0 > c_{p_0,p_1},$$

where $c_{p_0,p_1}$ is constant in $m,n$.

Any such procedure corresponds to a LRT for some particular type I error rate, and hence will not be consistent.

**Solution:** Have the cutoff $c$ depend on $m,n$ - favor $M_1$ over $M_0$ if

$$l_1 - l_0 > c_{p_0,p_1,m,n}$$
Modified selection criteria

Question: How should $c$ change with $N = m \times n$? Go up, or go down?

Answer:

- The inconsistency comes from rejecting $M_0$ too often.
- The threshold for favoring $M_1$ over $M_0$ should go up.
- We will still be able to select $M_1$ correctly if $M_1$ is true - as $N$ increases our ability to distinguish $M_1$ from $M_0$ increases as well.

Selection criteria: Favor $M_1$ over $M_0$ if

$$l_1 - l_0 > c_{p_0, p_1, m, n},$$

where $c_{p_0, p_1, m, n}$ is increasing in $m, n$. 
BIC - Bayes information criteria

\[ b_0 = l_0 - \frac{1}{2} p_0 \log N \]
\[ b_1 = l_1 - \frac{1}{2} p_1 \log N \]

Model selection via BIC: Favor \( M_1 \) over \( M_0 \) if \( b_1 > b_0 \).

Exercise: Rewrite this procedure to have the form used previously.

\[ b_1 > b_0 \Leftrightarrow l_1 - l_0 > \frac{1}{2} ((p_1 - p_0) \times \log N) \]

Notice: The cutoff
- is increasing in \( p_1 - p_0 \),
- is increasing in \( N = m \times n \).
Model selection via BIC: Favor $M_1$ over $M_0$ if $BIC_1 < BIC_0$. This is the same as favoring $M_1$ over $M_0$ if $b_1 < b_0$:  

\[
BIC_0 = -2 \times b_0 \\
BIC_1 = -2 \times b_1
\]
Do we trust BIC?

\[ y_{i,j} = \beta_1 + \beta_2 x_{i,j} + b_{1,j} + \epsilon_{i,j} \]

\[ b_{1,j} \sim N(0, \tau^2) \]

Consider selecting from among the following four models:

- **M\(_{00}\)**: \( \beta_2 = 0, \tau^2 = 0 \)
- **M\(_{10}\)**: \( \beta_2 \neq 0, \tau^2 = 0 \)
- **M\(_{01}\)**: \( \beta_2 = 0, \tau^2 \neq 0 \)
- **M\(_{11}\)**: \( \beta_2 \neq 0, \tau^2 \neq 0 \)

**Question:** What are the number of parameters in each model?

- **M\(_{11}\)** \( p = 4 \)
- **M\(_{01}\)** \( p = 3 \)
- **M\(_{10}\)** \( p = 3 \)
- **M\(_{00}\)** \( p = 2 \)

**Comment:** Which models could be compared with LRT?
Simulation study

m<-50 ; n<-5 ; g<-rep(1:m,times=rep(n,m))

BIC.RES<-NULL

for(t2 in c(0,1)) {
  for(beta2 in c(0,1)) {

    BIC.SIM<-NULL
    for(s in 1:100) {
      b<-rnorm(m,0,sqrt(t2) )
      x<-rnorm(m*n)

      y<- 1 + beta2*x + b[g] + rnorm(m*n)

      fit.00<-lm(y~1)
      fit.01<-lm(y~x)

      fit.10<-lmer(y ~ 1 + (1|g), REML=FALSE )
      fit.11<-lmer(y ~ x + (1|g), REML=FALSE )

      BIC.SIM<-rbind(BIC.SIM,c(BIC(fit.00),BIC(fit.01),BIC(fit.10),BIC(fit.11)))
    }

    BIC.RES<-rbind(BIC.RES,(table( c(1:4,apply(BIC.SIM,1,which.min)) ) -1))
  }
}
# Simulation study

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>95</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>97</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>
A harder simulation study

m<-10 ; n<-5 ; g<-rep(1:m,times=rep(n,m))

BIC.RES<-NULL

for(t2 in c(0,.5)){
    for(beta2 in c(0,.5)) {

        BIC.SIM<-NULL
        for(s in 1:100){
            b<-rnorm(m,0,sqrt(t2) )
            x<-rnorm(m*n)

            y<- 1 + beta2*x + b[g] + rnorm(m*n)

            fit.00<-lm(y~1)
            fit.01<-lm(y~x)

            fit.10<-lmer(y ~ 1 + (1|g), REML=FALSE )
            fit.11<-lmer(y ~ x + (1|g), REML=FALSE )

            BIC.SIM<-rbind(BIC.SIM,c(BIC(fit.00),BIC(fit.01),BIC(fit.10),BIC(fit.11)))
        }

        BIC.RES<-rbind(BIC.RES,(table( c(1:4,apply(BIC.SIM,1,which.min)) ) -1))
    }
}}
## Simulation study

```
BIC.RES

##
## [1,]  90  7  2  1
## [2,]  4  94  1  1
## [3,]  45  3  47  5
## [4,]  6  26  9  59
```
Model selection for NELS data

```r
fit.full<-lmer( mscore ~
    as.factor(flp) + as.factor(urbanicity) + public +
    ses + ses:public + (ses|school) , data=nels,REML=FALSE)

summary(fit.full)$coef
```

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>53.7270</td>
<td>0.4672</td>
<td>114.98</td>
</tr>
<tr>
<td>as.factor(flp)2</td>
<td>-1.7354</td>
<td>0.4026</td>
<td>-4.31</td>
</tr>
<tr>
<td>as.factor(flp)3</td>
<td>-4.4500</td>
<td>0.4379</td>
<td>-10.16</td>
</tr>
<tr>
<td>as.factor(urbanicity)</td>
<td>-0.0207</td>
<td>0.3834</td>
<td>-0.05</td>
</tr>
<tr>
<td>as.factor(urbanicity)</td>
<td>-0.9465</td>
<td>0.4193</td>
<td>-2.26</td>
</tr>
<tr>
<td>public</td>
<td>-0.8437</td>
<td>0.4425</td>
<td>-1.91</td>
</tr>
<tr>
<td>ses</td>
<td>3.4174</td>
<td>0.2586</td>
<td>13.21</td>
</tr>
<tr>
<td>public:ses</td>
<td>0.9086</td>
<td>0.2946</td>
<td>3.08</td>
</tr>
</tbody>
</table>
Model selection for NELS data

\[ \text{BIC}(\text{fit.full}) \]

## [1] 92472.76

\[
\text{fit.r1} \leftarrow \text{lmer}( \text{mscore} \sim \text{as.factor(flp)} + \text{as.factor(urbanicity)} + \text{public} + \text{ses} + (\text{ses|school}), \text{data=nels}, \text{REML=FALSE})
\]

\[ \text{BIC}(\text{fit.r1}) \]

## [1] 92472.71

\[
\text{fit.r2} \leftarrow \text{lmer}( \text{mscore} \sim \text{as.factor(flp)} + \text{as.factor(urbanicity)} + \text{ses} + (\text{ses|school}), \text{data=nels}, \text{REML=FALSE})
\]

\[ \text{BIC}(\text{fit.r2}) \]

## [1] 92464.98
Futher reductions

```r
fit.r3 <- lmer(mscore ~ as.factor(flp) + ses + (ses | school), data=nels, REML=FALSE)
BIC(fit.r3)
```

```
## [1] 92454.31
```
Futher reductions

```r
fit.r4a <- lm( mscore ~ as.factor(flp) + ses , data = nels)
BIC(fit.r4a)
## [1] 93151.9

fit.r4b <- lmer( mscore ~ ses + (ses | school) , data = nels, REML = FALSE)
BIC(fit.r4b)
## [1] 92597.89

fit.r4c <- lmer( mscore ~ (ses | school) , data = nels, REML = FALSE)
BIC(fit.r4c)
## [1] 93267.56
```
Where does BIC come from?

Suppose there are only two models $M_0$ and $M_1$. 

In a Bayesian analysis, one would be able to compute

$$
\Pr(M_1 | y) = \frac{\Pr(M_1) p(y | M_1)}{\Pr(M_1) p(y | M_1) + \Pr(M_0) p(y | M_0)}
$$

Alternatively, the odds that $M_1$ is true are

$$
\frac{\Pr(M_1 | y, X)}{\Pr(M_0 | y, X)} = \frac{\Pr(M_1)}{\Pr(M_0)} \times \frac{p(y | M_1)}{p(y | M_0)}
$$

If $Pr(M_1) = Pr(M_0)$, then

$$
\frac{\Pr(M_1 | y, X)}{\Pr(M_0 | y, X)} = \frac{p(y | M_1)}{p(y | M_0)}
$$
Where does BIC come from?

We would select $M_1$ if $\frac{p(y|M_1)}{p(y|M_0)} > 1$, or equivalently

$$\log \frac{p(y|M_1)}{p(y|M_0)} = \log p(y|M_1) > p(y|M_0).$$

It can be shown that in many cases for large $N$,

$$\log p(y|M_1) \approx \log p(y|\hat{\theta}_1) - \frac{1}{2} p_1 \log N$$
$$\log p(y|M_0) \approx \log p(y|\hat{\theta}_0) - \frac{1}{2} p_0 \log N$$

and so we prefer $M_1$ ot $M_0$ if

$$\log p(y|\hat{\theta}_1) - \frac{1}{2} p_1 \log N > \log p(y|\hat{\theta}_0) - \frac{1}{2} p_0 \log N$$
$$- 2 \log p(y|\hat{\theta}_1) + p_1 \log N < - 2 \log p(y|\hat{\theta}_0) + p_0 \log N$$

$$BIC(M_1) < BIC(M_0)$$
Other information criteria: AIC, TIC, GIC.
See Müller, Sealy and Welsh (2013) for a review.

Don’t do the following:

- \( BIC(M_1) = 100 \), but has many parameters;
- \( BIC(M_0) = 101 \), but has few parameters.

“Since the BICs are close, and \( M_1 \) has more parameters, I’ll go with \( M_0 \).”

\( M_1 \) has already been penalized for its number of parameters. The BIC selection rule would be to select \( M_1 \).