# Week 14: Choosing models

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019



Updated notes are here: https://clas.ucdenver.edu/marcelo-perraillon/teaching/health-services-research-methods-i-hsmp-7607

# Outline

- Modeling part II
- Information criteria: BIC and AIC
- Guidelines
- Variable selection procedures
- Oldies: forward, backward elimination
- Newish: lasso, ridge regression

# Brief review

- In the classes about "modeling" we covered transformation of variables (the outcome or covariates) for two reasons {Wait, before that: the song was Right Said Fred - I'm Too Sexy (find the video!)}:

  1) Changing the interpretation of the model (e.g. centering, logs, scales, etc)

  2) Making the assumptions of the linear model more plausible (e.g. taking the log of the outcome)
- We also covered **model specification**. Should we model age as linear or quadratic? Divide a continuous variable into categories?
- We saw that we could compare models using F tests (similar to comparing adjusted $R^2$) or with likelihood ratio tests (LRT). We could also use Wald tests (the "test" command in Stata)

# Overview

- Today, we will see other ways of choosing models/variables to include in regressions
- The most important part about today: BIC and AIC for non-nested models. Think of both essentially as an "adjusted $R^2$" using maximum likelihood or –more precisely– the estimate log-likelihood to **compare models**
- The other most important part: think theory/conceptual framewok to guide in **choosing variables**. **Never use automatic selection procedures** for choosing your variables (unless you are predicting; then **maybe**)
- Chapter 11 of Chaterjee covers selecting variables and choosing models

# Overview

- Often we worry about causal inference, confounders, etc. What you know about the research question will drive decisions about **which variables should be in a model**
- Say, if you are modeling cancer mortality it would be silly not to add other comorbid conditions–regardless of whether the bivariate relationship between the comorbid conditions/index is statistically significant or regardless of whether that variable is statistically significant in the final model. Conceptually, it has to be there
- The previous comment is one of those things that not all professors would be in agreement but I'd say that in HSR or health econ is the more accepted view
- On the other hand, if there is a variable that **conceptually** you are ambivalent about and it's not statistically significant, you would probably drop it
- Once you have the variables that you think should be in the model, you need to **worry about how they should be there**–that is, **model specification**

# Framing the problem the Chatterjee way

- We have a dataset with $j$ possible explanatory variables, $X_1, ..., X_j$
- How do we select a subset of them for inclusion in our regression model?
- Let's call the retained variables $X_1, ..., X_p$ and the deleted variables $X_{p+1}, X_{p+2}, ..., X_q$
- Two situations:
  1. In the **true** model that connects Y and X, all the parameters $\beta_0$ to $\beta_q$ are nonzero
  2. In the **true** model that connects Y and X, $\beta_0, ..., \beta_p$ are nonzero but $\beta_{p+1}, ..., \beta_q$ are zero
- In case 2 there is no problem if we leave the zero-coefficient variables out, but what is the problem with 1)?

# Consequences

- If we estimate the reduced model with $p$ variables, then the estimated coefficients will be **biased** if the omitted variables are correlated with some or all the $p$ variables
- The **variance** of the betas $var(\hat{\beta}_p)$ in the reduced model with $p$ variables is going to be at most the same or less because fewer parameters are estimated
- So, the full model might be less biased but the reduced model has smaller variance
- (This is an example of the **bias and variance** trade off we didn't have time to cover)
- In this view, **choosing the correct model is a question of bias versus variance trade-off**. **BUT...**

# Big picture

- Note how this way of framing the problem is conceptually different from what we have been doing before (and Wooldridge's)
- The textbook mixes two concepts. Whether some coefficients are truly zero (in the population) and whether the coefficients are correlated to some of the predictors left in the model
- If zero, they don't predict the outcome, so they can't be confounders either so not a problem leaving them out
- For Wooldridge –and HSR/health econ– the main concern is causality not bias/variance trade off
- In our field, we include the variables in the model that we know **conceptually** should be there, **regardless of their statistically significance or coefficient size in the *sample***
- Why? Because we know that the **true population model** should include those variables

# Maybe belaboring the point here but...

- Note one key consequences from both approaches. If you follow a conceptual framework, you don't run a single model to decide which variables will be in the model. You don't look at p-values, fit statistics–BIC, etc
- You may do so for the ones that you are not sure about
- **But for sure you should do all we are going to cover today to decide what is the best functional form for the model**
- As Will Manning used to say (paraphrasing): Theory doesn't tell you if age should be entered as a quadratic–it just tells you that age should be in the model
- For that, use today's class tools
- If you came from an epi or stats background, you might have been told to use today's tools to **choose variables**

# Information criterion - AIC

- A common way to compare models is by using the so-called **information criterion**
- It is a way to balance bias and variance or accuracy (fit) and simplicity (parsimony)
- One of them–the first one to be proposed–is the Akaike Information Criterion (AIC). In the **context of the linear model** (estimated using OLS), it is:

  $AIC_p = n * ln(\frac{SSE_p}{n}) + 2 * p$
- $p$ is the number of estimated parameters (including the constant), $n$ is the number of observations, and SSE is the residual sum of squares (recall SST = SSR + SSE)
- **The smaller the AIC the better**. When is a model going to be better? The larger the sample size, the lower the unexplained variance, the fewer parameters we use
- Usually, we compare models with the same sample size, n, or maybe a small difference in n (a covarite may have some missings)

# Information criterion - AIC

- That $2 * p$ is a **penalty** much like in adjusted $R^2$. The more parameters we estimate the better the model has to be
- We saw that there is formula linking SSE to the log-likehood. That's how Stata calculates the log-likelihood when you use the regress command since Stata doesn't use MLE in the regress command:

  $ll = -0.5 * n * [ln(2 * \pi) + ln(SEE/n] + 1$
- We could solve for SSE and plug in to get AIC as a function of the log-likelihood

# Some intuition

- AIC comes from information theory. Suppose that we know the population process that generated the data but we are using a model to represent that process
- Think of simulating some data and then using imperfect models to represent the data. There will be some information lost
- If we knew the true data generating process, we could find a way to estimate how much information we are losing in each model and thus choose the model that doesn't lose as much information
- **But we never know the true model/process**. Akaike, in 1974, showed that his information criterion, now called AIC, tells how much more information is lost comparing one model to the other using the estimated log-likelihood
- The above formula is a special case that applies to the linear model: $\epsilon_i \sim N(0, \sigma^2)$

# Some caveats

1. Valid **asymptotically**. You need much more observations than parameters
2. Balance between goodness of fit and parsimony (simplicity). The penalty is incurred to avoid overfitting models because models with more parameters will always fit the data better. If you use as many parameters as data points, model fit is perfect (essentially, the estimated model is a perfect representation of the data)
3. **Most important**: AIC is a **relative** measure that compares one model to another to choose the one that loses less information. It's NOT a measure of how good a model is. You may have two horrendous models (say, super low $R^2$); AIC helps you choose the best among terrible options–think of some elections...

# Information criterion - AIC

- Note that AIC is similar to the F-test we used to compare nested models. One way of writing the F-test was
  $F = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$
- Also, the F-test can be written in terms of the $R^2$:
  $F = \frac{(R_p^2 - R_q^2)/(p-q)}{(1-R^2 p)/(n-p-1)}$
- We are essentially comparing SSE (or $R^2$) in relationship to the number of estimated parameters
- So why is AIC useful? Because we can use it to **compare non-nested models**, although there is not statistical test like F test or MLE
- Again, **AIC is a relative measure**. Also, the absolute value doesn't tell us much; we use it **compare related models**
- Obviously, it wouldn't make much sense to compare models with different outcomes or different datasets

# Information criterion- AIC

- We also saw that we can use the likelihood ratio test to compare models and that the F-test and the LRT are asymptotically equivalent
- Stata's way to calculate AIC (except in GLM models) is $AIC = -2ll + 2k$
- $k$ is the number of parameters, $ll$ is the log-likelihood function. Again, $2k$ is the **penalty due to the number of parameters**; the more parameters, the higher AIC (we prefer models with lower AIC)
- **We are essentially comparing the log-likelihood**
- The advantage of using the log-likelihood rather than SSE is that we have an AIC for linear models, probit, logit, poisson, etc

# Information criterion - BIC

- The more common information criterion used is BIC–a modification of AIC:

  $BIC = -2ll + ln(n) * k$

- Note that we now add $ln(n)$ in the penalty term rather than 2
- If n = 8, ln(8) = 2.079. So in most application, the BIC penalty is always **higher than the AIC penalty** but it has a limit because the log function "flattens"
- ln(50) = 3.91, ln(100) = 4.60, ln(500) = 6.21, ln(1000) = 6.91, ln(2000) = 7.6
- The more observations we have, the lower the growth of the penalty $ln(n) * k$

# Example

- AIC and BIC for the low birth logit model

```
qui logit lw cigs faminc motheduc, nolog
estat ic

Akaike's information criterion and Bayesian information criterion

------------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df          AIC        BIC
-------------+----------------------------------------------------------------
           . |      1,387 -584.4731  -572.1589       4    1152.318   1173.257
------------------------------------------------------------------------------
               Note: N=Obs used in calculating BIC; see [R] BIC note.
```

- So what do we do with that number? Nothing, really. We need to compare models **using the same data and outcome** –although there is some leeway if the sample sizes differ by a bit

# Example

- Comparing models

```
qui logit lw cigs faminc, nolog
est sto m1
qui logit lw cigs motheduc, nolog
est sto m2
qui logit lw cigs motheduc parity , nolog
est sto m3

est table m1 m2 m3, stats(r2_p aic bic)
```

- Note that we could do a LRT between model 2 and model 3

# Example

- The first model is better, adding the other covariates didn't improve things

```
est table m1 m2 m3, stats(r2_p aic bic)

----------------------------------------------------
    Variable |     m1           m2           m3
-------------+--------------------------------------
        cigs |  .04481613    .04666558    .04657893
      faminc | -.00785126
    motheduc |              -.02393827   -.02342169
      parity |                            .01208105
       _cons |  -1.644825   -1.5595502   -1.5857927
-------------+--------------------------------------
        r2_p |  .02104683    .01858425    .01860151
         aic |  1150.6598    1153.2221    1155.2019
         bic |  1166.3667    1168.9268    1176.1415
----------------------------------------------------

 lrtest m2 m3

Likelihood-ratio test                                 LR chi2(1)  =      0.02
(Assumption: m2 nested in m3)                         Prob > chi2 =    0.8871
```

- We do prefer m2 to m3 as well. Note that BIC/AIC yields similar results to pseudo-$R^2$

# Another way

- The "estimates" command to build table has an option "stat"

```
 est stat log pro

Akaike's information criterion and Bayesian information criterion

------------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+----------------------------------------------------------------
         log |      1,387 -584.4731  -572.1589       4    1152.318   1173.257
         pro |      1,387 -584.4731  -572.1118       4    1152.224   1173.163
------------------------------------------------------------------------------
               Note: N=Obs used in calculating BIC; see [R] BIC note.
```

- Stata has a note on how it calculates BIC

# Modeling suggestions

- Start from simple to complex (after an exploratory analysis, of course)
- Focus on the outcome and relationship of interest (say, low birth weight and smoking). In other words, simple linear regression
- Make sure you understand what the parameters mean. If logistic, make sure you go from log-odds to relative risk to differences in probability (marginal effects). Try the linear probability model
- If using transformations like $log(y)$, make sure you understand what the parameters mean in the scale you care about
- **WRITE DOWN THE MODEL**
- Write down the model write down the model write down the model write down the model write down the model write down the model write down the model write down the model write down the model write down the model write down the model write down the model write down the model write down the model write down the model write down the model

# Modeling suggestions

- Then add other variables one at a time or by logical groups, like demographics (sex, age, race)
- See how adding those variables changes the coefficient of the variable of interest
- Watch out for collinearity signs (large changes in SEs, parameters change a lot, signs reversed)
- Once you have a complete model, check for functional form. Age as quadratic? Dummy categories?
- Formally compare models. If nested using LRT/Wald tests. If not nested BIC
- Sometimes simplicity of presentation is preferred to a better fitting model, assuming, of course, the difference is not major
- If you know that conceptually a variable should be in the model, then the variable should be in the model (it doesn't matter if not statistically significant, although some people may argue against that)

# Modeling suggestions

- Do model diagnostics. Make sure assumptions are met. You may need to try other functional forms
- Check for interactions but be careful with complicated models because you can complicated yourself trying gazillion interactions (your conceptual model is your guide)
- Consider stratification if you have enough sample sizes
- If outcome is 1/0, start with the linear probability model to make sense of parameters. Then switch to logistic/probit models and make sure that you understand what the parameters mean

# Variable selection **procedures**

- We will briefly cover other ways of selecting variables because you may have heard of them
- The old-fashioned way: **forward selection** and **backward selection**
- The newer way: **ridge** and **lasso models** and there 100 variants
- Before, though, the **bottom line**: we don't use these methods often or at all in HSR and health economics
- In most situations we have a conceptual model and don't want our sample to dictate conceptual choices
- In all procedures, **you need to eliminate highly correlated (collinearity problem) variables**
- **The modern methods are ARE NOT for variable selection when you have correlated variables**. They are unstable–not a very well known issue yet

# Forward selection

- This follows the logic of starting with a simple model (one without predictors)
- Then the procedure adds the variable with the highest correlation with the outcome (let's call it $X_1$)
- If the Wald test for the added variable is significant, then keep it
- The next variable is the variable that has the highest correlation with Y once $X_1$ has been taken into account
- If significant, keep it and add another variable
- Repeat until all potential variables have been considered
- You can usually choose the significance level for the Wald test (sometimes 0.1 instead of 0.05)

# Backward selection

- Start with full model and then remove one by one
- The variables are dropped in relation to their contribution to the error sum of squares (SSE)
- First drop the variable with the smallest contribution (so the variable that doesn't "matter" as much)

# Example

- Beauty example, forward selection

```
 stepwise, pe(0.1): reg lwage abvavg exper expersq union goodhlth black female married south
                    begin with empty model

p = 0.0000 <  0.1000  adding  female
p = 0.0000 <  0.1000  adding  exper
p = 0.0000 <  0.1000  adding  expersq
p = 0.0000 <  0.1000  adding  union
p = 0.0041 <  0.1000  adding  south
p = 0.0176 <  0.1000  adding  black
p = 0.0255 <  0.1000  adding  abvavg
p = 0.0525 <  0.1000  adding  goodhlth

(omitted)
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.4423845   .0310588   -14.24   0.000    -.5033175   -.3814515
       exper |   .0441188   .0046224     9.54   0.000     .0350503    .0531873
     expersq |  -.0007647   .0001029    -7.43   0.000    -.0009665   -.0005628
       union |   .1607766   .0321088     5.01   0.000     .0977836    .2237695
       south |   .1027258   .0372794     2.76   0.006     .0295887    .1758629
       black |  -.1283028   .0546851    -2.35   0.019    -.2355874   -.0210182
      abvavg |   .0671251   .0312239     2.15   0.032     .0058681    .1283821
    goodhlth |    .111338    .057356     1.94   0.052    -.0011866    .2238627
       _cons |   1.194908   .0738119    16.19   0.000     1.050099    1.339717
------------------------------------------------------------------------------
```

# Example

- Backwards

```
 . stepwise, pr(0.1): reg lwage abvavg exper expersq union goodhlth black female married south
                      begin with full model
p = 0.8518 >= 0.1000  removing married

      Source |       SS           df       MS      Number of obs   =     1,260
-------------+----------------------------------   F(8, 1251)      =     65.21
       Model |   130.94827         8  16.3685338   Prob > F        =    0.0000
    Residual |  314.031702     1,251  .251024542   R-squared       =    0.2943
-------------+----------------------------------   Adj R-squared   =    0.2898
       Total |  444.979972     1,259  .353439215   Root MSE        =    .50102

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      abvavg |   .0671251   .0312239     2.15   0.032     .0058681    .1283821
       exper |   .0441188   .0046224     9.54   0.000     .0350503    .0531873
     expersq |  -.0007647   .0001029    -7.43   0.000    -.0009665   -.0005628
       union |   .1607766   .0321088     5.01   0.000     .0977836    .2237695
    goodhlth |    .111338    .057356     1.94   0.052    -.0011866    .2238627
       black |  -.1283028   .0546851    -2.35   0.019    -.2355874   -.0210182
      female |  -.4423845   .0310588   -14.24   0.000    -.5033175   -.3814515
       south |   .1027258   .0372794     2.76   0.006     .0295887    .1758629
       _cons |   1.194908   .0738119    16.19   0.000     1.050099    1.339717
------------------------------------------------------------------------------
```

# Problems

- Oh... where to start?
- The order sometimes matter and can give different results
- The method doesn't work with highly correlated variables (remember that collinearity makes p-values unstable)
- The usual advice is to do backwards elimination because you start with the full model
- It gets the modeling problem conceptually wrong. It uses the data to dictate what should be in a model that describe a true and unknown model – where is the theory?
- Not once in my life I have used these methods. I can't remember ever reading a paper that used this methods in HSR or health econ

# The modern methods

- Same conceptual issues in terms of selection of variables
- The idea is that we change the objective function in linear regression. Besides minimizing the sum or squares, we try to shrink the coefficients to zero
- Our usual OLS: min $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} X_j\beta_j)^2$
- Ridge **penalized** sum of squares: min $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} X_j\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$

# The modern methods

- Some coefficients may shrink to zero; the idea is that we could drop the ones that are closer to zero
- Ridge regression is better at **prediction** rather than variable selection
- It introduces bias by forcing coefficients to shrink but reduces variance (so better RMSE)
- If some variables are really zero and thus not predictive of Y, ridge regression performs well
- The other alternative is **lasso regression**; it imposes a different penalty
- There are new methods for binary outcomes and different types...

# Summary

- Use BIC/AIC to select non-nested models
- Don't be seduced by variable selection procedures **unless you are building models for prediction**; if so, learn about ridge and lasso regression and their many new variants
- But be careful with the **newest-is-best** effect
- Take into account that you need to eliminate highly correlated predictors first (use VIF) in all these methods