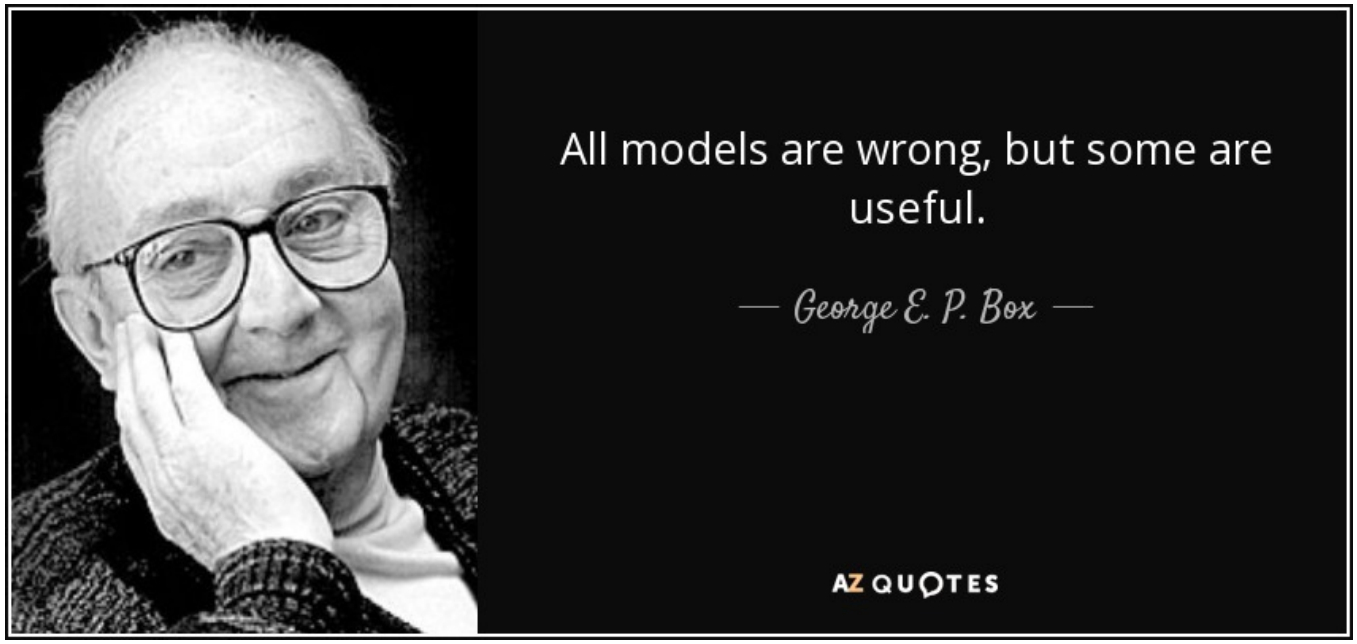
BST 210
Applied Regression Analysis



"*Science and Statistics",* JASA, 1976

Lecture 9
Plan for Today

1. A few course details
2. Lab this week: hands-on examples using subset of Framingham Heart Study data! (reading)
3. *Multivariable Model Selection!*
    a. Overview of Modeling/Selection
    b. Variable selection methods
        - Forward Selection/Backward Elimination
        - Stepwise; and Best subset; Bootstrap; Validation
    c. Criteria for model selection

What *is* a Model?

All Models are Wrong

Models are only useful *because* they are wrong

- Because the very nature of a model is a simplified and idealized representation of something, all models will be wrong in some sense.
- They must be wrong, in order to be generalizable!
- Models will never be “the truth” if truth means entirely representative of reality.
- *Goal typically is: model population mean (driven by underlying data process)*

What *good* is a Model?

Bias Variance Tradeoff

- Reduce dimensions -> Less accurate Model
- Increase Dimensions -> Less Useful Model
  (is just data again)

Objectives of a Model?

- **To identify predictors of outcome of interest**

  Ex: Modeling BMI, total run time, fluid intake as risk factors for predicting lower sodium levels during Boston Marathon

- **To build prediction model for outcome of interest**

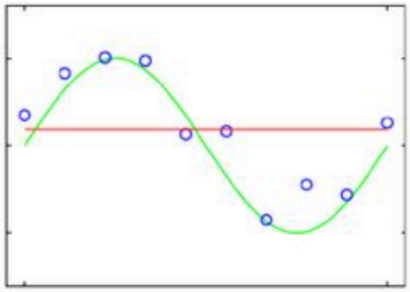
  Ex: Prognostic tool around water-intake limits, and potential salt intake guidelines for along route

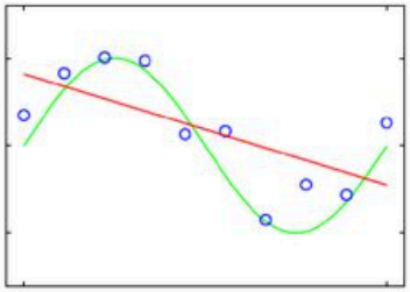
- **To quantify association between outcome and exposures**

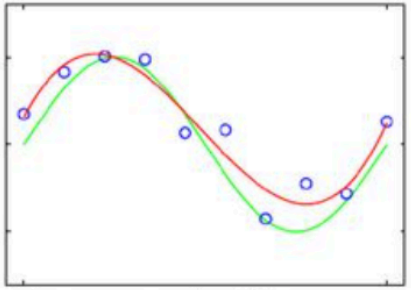
  Ex: Modeling association between BMI, total run time, fluid intake, gender, age and lower sodium levels

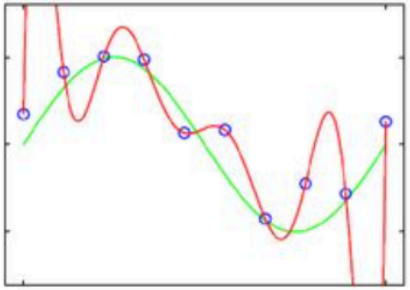
Objectives of Model Selection?

- Must decide what factors in world best model outcome
- Model selection is *not* Model fitting
- Goal: Evaluate balance between goodness of fit and generalizability


Intercept only


Intercept + height


Intercept + height + age


Intercept + height + age + ...

Why Model Selection?

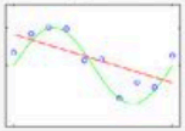
- Clinical trials (RCT) vs Observational Studies
- Numerous potential explanatory (predictor) variables
- Collinearity!

Main Points of Model Selection?

(one demonstration of process)

1. Define models

Simple Model = Dog preference is predicted well by height.



$$\text{Model (S)} = m \cdot x + c$$

Augmented Model = Dog preference is predicted better by height and age

$$\text{Model (A)} = m \cdot x + n \cdot x^2 + c$$

2. Comparing Error reductions

$$F = \frac{(SSE - SSE)/(P - P)}{SSE / n - P}$$

SSE: the sum of squared error;
P: the number of parameters
n: the number of observations

3. Significance test

Criteria for Model Selection?

- Adjusted R^2
- F-test
- t-test
- AIC
- BIC
- Root MSE
- P-values

Principles of Model Selection?

- **Parsimony**
  - find simplest model that meets objective
- **Hierarchy**
  - respect natural hierarchy amongst variables while modeling
- **Simplicity, Interpretability, Feasibility**

Multivariable Model Building

- In most practical situations, we have a set of potential explanatory variables and we must decide which ones to include in the regression model and which ones to leave out
- These decisions must be based on both statistical and non-statistical considerations (never just one alone)

Model Building

- One also needs to consider the goals of the model building
- Is it to develop a model for interpretation of effects? Perhaps focus on statistical significance, confounding, and effect modification.
- In this case, it might be best to model covariates in a way that allows easy interpretation of the β coefficients

Model Building

- Is it the development of a good prediction model for future subjects? Perhaps focus on adjusted R^2 or other criteria, less on statistical significance.
- In this case, it might be appropriate to consider the use of splines or generalized additive models for covariate adjustment

Model Building

- Ideally, we would have some prior knowledge as to which variables might be relevant and interesting to consider from a medical/epidemiologic perspective

- Similarly, it may be well-known that it is important to adjust for certain factors in the model (e.g., age and gender in many epidemiologic studies)

- Ask collaborators or field experts

Model Building

- Comparing models statistically (by looking at *p*-values) should help focus the looking, but a *p*-value is not the only reason to keep a variable in the model
- However, if a covariate is not statistically predictive of the outcome, we would often drop it from the model (unless it is the primary exposure of interest or is an appreciable confounding variable)

Model Building

- In addition to statistical tests, subject matter knowledge and common sense should help guide which variables are important in predicting outcome
- Be sure to be thinking about possible confounding or effect modification

Steps that May Help

- Begin with a careful univariable screening of each covariate
- Run a multivariable model including all variables thought (possibly) to influence outcome from the univariable analysis
- Perhaps use $p < 0.25$ as a screening criteria (using $p < 0.05$ may fail to identify important factors that may become more statistically significant after adjustment for other factors)

Steps that May Help

- Throughout you are considering outliers and high influence observations
  - Scatterplots of *Y* vs. any covariate *x* (or histograms, for categorical *x*) can be helpful before beginning any analysis
  - Note: an observation being an outlier or high influence point can depend on what other variables are included in your model
  - Maybe preliminary work is needed at the start, and then a more comprehensive look near the end of model building

Steps that May Help

- Include variables known or thought to be important from subject matter considerations
  - Treatment or exposure variables
  - Possible or known confounding factors
  - Plausible effect modifiers
  - Variables “everybody” includes
- Be careful about covariates with missing values (using them leads to smaller sample sizes and possible bias)

Steps that May Help

- Considering interactions can be challenging
  - In addition, one would always include the main effects of two covariates when their interaction is included (hierarchy principle)
  - Assessment of interactions can be particularly challenging with computerized procedures, and thus may require human involvement

Steps that May Help

- Considering interactions can be challenging
  - Which interactions are medically plausible?
  - With 10 covariates, you have $10 \times 9 / 2 = 45$ possible interactions; with 20 covariates you have $20 \times 19 / 2 = 190$ possible interactions
  - Multiple testing of so many interactions is clearly an issue
  - May be better to focus on a limited number of possible interactions thought interesting to assess in advance

Steps that May Help

- Begin to fit smaller, reduced models if some of the estimated coefficients in the multivariable model do not reach statistical significance
- Usually consider dropping (or adding) one variable at a time, rather than dropping (or adding) multiple variables at one time (except when you add, say, a number of indicator variables to model age groups)

Steps that May Help

- Once you feel close to a final model, also look more carefully for
  - Contributions of continuous covariates (possible nonlinear, spline, or GAM terms to assess linearity)
  - Comparisons of continuous vs. categorical coding of a covariate (e.g., continuous age vs. age categories)
  - Possible interactions of interest

Model Building

- Decisions made about inclusion of a variable, or linearity of a variable, etc. can change as other covariates are added or dropped
- Use logic, good sense, caution, and biologic plausibility when using model building techniques
- Have fun and be creative
- There is as much an “art” to model building as “science”

Model Building

- There may be more than one “final model”
- In complex data sets, we may present the results of several related models, or choose between models on the basis of subject matter considerations
- Often many models will have acceptable fit (similar MSE or adjusted R^2, say)
- Different, reasonable people may come up with different models

Variable Selection Methods

There are several variable selection methods that are commonly used:

1. Forward selection
2. Backward elimination
3. Stepwise selection (forward or backward)
4. Best subsets
5. Criterion methods

1. Forward Selection

Start with a model including only the intercept term

Perform a univariable analysis of each risk factor separately

Multi-category exposures represented by several indicator variables count as a single variable

1. Forward Selection

Identify the variable (V_1) with the lowest *p*-value such that $p < \alpha_1$ (often $\alpha_1 = 0.05$). If $p < \alpha_1$, then keep V_1 in the model, else don't and stop.

Fit separate two variable models with V_1 plus one additional variable

Find the variable V_2 with the lowest $p < \alpha_1$ after including V_1

1. Forward Selection

If there is no such variable, then stop and declare the model with V_1 as the final model

Otherwise, consider adding a third variable to V_1 and V_2 and find the variable V_3 with the lowest *p*-value, …, etc.

Continue until there are no other variables that satisfy the α_1 *p*-value criterion

1. Forward Selection

Once a variable is in the model, it cannot be removed

Occasionally a criterion other than a *p*-value is used, such as adjusted R^2

In that case, variables would be chosen to maximize the adjusted R^2 (or increase it enough to be worth including the covariate in your model)

1. Forward Selection

- We started with a model with no covariates (the intercept only model)
- May make more sense to start with a simple model that includes factors you know you want to include in your final model (e.g., treatment or exposure variables, likely confounding variables)
- Due to risk of finding local min instead of global min

2. Stepwise Selection (forward)

The forward stepwise method is different in one respect

After each step, re-evaluate the contribution of each variable currently in the model and delete the variable with the highest *p*-value if $p > \alpha_2$

Basically, after you add a variable in, you could go back and delete a previously added variable that is not as significant anymore

2. Stepwise Selection (forward)

Note that α_1 and α_2 do not have to be the same

Usually we have $\alpha_2 > \alpha_1$ (perhaps $\alpha_1 = 0.05$ to enter, $\alpha_2 = 0.10$ to exit), because otherwise a covariate you enter could leave right away

Stepwise selection ends if no further variables can be entered or removed

Recall:

The New England
Journal of Medicine

Volume 329 OCTOBER 7, 1993 Number 15

A COMPARISON OF THE PERIOPERATIVE NEUROLOGIC EFFECTS OF HYPOTHERMIC CIRCULATORY ARREST VERSUS LOW-FLOW CARDIOPULMONARY BYPASS IN INFANT HEART SURGERY

JANE W. NEWBURGER, M.D., M.P.H., RICHARD A. JONAS, M.D., GIL WERNOVSKY, M.D., DAVID WYPIJ, PH.D., PAUL R. HICKEY, M.D., KARL C.K. KUBAN, M.D., S.M., DAVID M. FARRELL, M.A., C.C.P., GREGORY L. HOLMES, M.D., SANDRA L. HELMERS, M.D., JULES CONSTANTINOU, F.R.A.C.P., ENRIQUE CARRAZANA, M.D., JOHN K. BARLOW, M.D.,* AMY Z. WALSH, R.N., B.S.N., KRISTIN C. LUCIUS, R.N., M.S., JANE C. SHARE, M.D., DAVID L. WESSEL, M.D., FRANK L. HANLEY, M.D., JOHN E. MAYER, JR., M.D., ALDO R. CASTANEDA, M.D., AND JAMES H. WARE, PH.D.

(At the time of this study, hypothermic circulatory arrest (HCA) and low-flow cardiopulmonary bypass (CPB) were accepted techniques for the operative management of complex cardiovascular pathology, with the potential for neurologic sequelae being a concern.)

Example: Circulatory Arrest Study

- Bellinger *et al*. (NEJM, 1995) report on a clinical trial comparing CA (Circulatory Arrest) vs. LFB (Low Flow Bypass) for repair of transposition of the great arteries

- Primary outcome at one year was PDI (Psychomotor Development Index), a continuous measure of motor skills (scaled similar to IQ)

Example: Circulatory Arrest Study

- Predictor variables include:
  - treatment group ‘dhca’ (CA vs. LFB) - nominal
  - duration of circulatory arrest ‘minutes’ - continuous
  - diagnosis group ‘vsd’ (IVS, Intact Ventricular Septum, vs. VSD, Ventricular Septal Defect) - nominal
  - age at surgery, birth weight, etc.

USING THE CIRCULATORY ARREST DATA SET FOR SOME MODEL BUILDING (with Stata)

```
. use "P:\BST 210 \circarrest.dta"

. stepwise, pe(.10): regress pdi dhca vsd minutes birthwt age
                     begin with empty model
p = 0.0086 <  0.1000  adding  minutes
p = 0.0053 <  0.1000  adding  birthwt

      Source |       SS             df       MS      Number of obs   =       142
-------------+------------------------------------   F(2, 139)       =      7.75
       Model |  3377.28384           2  1688.64192   Prob > F        =    0.0006
    Residual |  30281.3359         139  217.851337   R-squared       =    0.1003
-------------+------------------------------------   Adj R-squared   =    0.0874
       Total |  33658.6197         141  238.713615   Root MSE        =     14.76

------------------------------------------------------------------------------
         pdi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     minutes |   -.164923   .0566858    -2.91   0.004    -.277001    -.052845
     birthwt |   .0084129   .0029668     2.84   0.005     .0025471    .0142788
       _cons |   71.20498   10.62018     6.70   0.000     50.20699    92.20297
------------------------------------------------------------------------------
```

3. Backward Elimination

Here we start with a model containing all potential predictor variables

At the next step, delete the variable with the largest *p*-value such that $p > \alpha_2$, and continue until all variables retain statistical significance at the α_2 level

Different models may be selected using the different variable selection methods

3. Backward Elimination

One rule of thumb states that to avoid over-fitting (getting a model that is not generalizable), there should be at least 10 observations for each variable in the regression model

But even that does not guarantee that a model with many covariates can be fit, or that the covariates will reach statistical significance

With smaller data sets or large numbers of covariates, therefore, backward elimination may be inappropriate to use

4. Stepwise Selection (backward)

The backward stepwise method is different in one respect

After each step, re-evaluate the contribution of each variable not currently in the model and add the variable with the lowest *p*-value back in if $p < \alpha_1$

Basically, after you drop a variable, you could go back and add it back in if it later looks more significant

4. Stepwise Selection (backward)

Note that α_1 and α_2 do not have to be the same

Again, $\alpha_2 > \alpha_1$ (perhaps $\alpha_2 = 0.10$ to exit, $\alpha_1 = 0.05$ to enter)

Backward stepwise selection ends if no further variables can be removed or entered

```
. stepwise, pr(.10): regress pdi dhca vsd minutes birthwt age
                     begin with full model
p = 0.6005 >= 0.1000  removing minutes
p = 0.1258 >= 0.1000  removing age

      Source |       SS       df       MS      Number of obs   =       142
-------------+------------------------------   F(3, 138)       =      6.54
       Model |  4190.84279     3  1396.9476   Prob > F        =    0.0004
    Residual |  29467.7769   138  213.534615   R-squared       =    0.1245
-------------+------------------------------   Adj R-squared   =    0.1055
       Total |  33658.6197   141  238.713615   Root MSE        =    14.613

------------------------------------------------------------------------------
         pdi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dhca |   -6.952038   2.458269    -2.83   0.005    -11.81278   -2.091293
         vsd |   -6.224385   2.972168    -2.09   0.038    -12.10126   -.3475067
     birthwt |    .0080636   .0029403     2.74   0.007     .0022497    .0138775
       _cons |    71.53173   10.51437     6.80   0.000     50.74162    92.32183
------------------------------------------------------------------------------
```

```
. stepwise, pe(.05) pr(.10): regress pdi dhca vsd minutes birthwt age
                      begin with full model
p = 0.6005 >= 0.1000  removing minutes
p = 0.1258 >= 0.1000  removing age

      Source |       SS       df       MS      Number of obs   =       142
-------------+------------------------------   F(3, 138)       =      6.54
       Model |  4190.84279     3  1396.9476   Prob > F        =    0.0004
    Residual |  29467.7769   138  213.534615   R-squared       =    0.1245
-------------+------------------------------   Adj R-squared   =    0.1055
       Total |  33658.6197   141  238.713615   Root MSE        =    14.613

------------------------------------------------------------------------------
         pdi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dhca |   -6.952038   2.458269    -2.83   0.005    -11.81278   -2.091293
         vsd |   -6.224385   2.972168    -2.09   0.038    -12.10126   -.3475067
     birthwt |    .0080636   .0029403     2.74   0.007     .0022497    .0138775
       _cons |    71.53173   10.51437     6.80   0.000     50.74162    92.32183
------------------------------------------------------------------------------
```

5. Best Subsets

- Perhaps find the five best one covariate models, the five best two covariate models, the five best three covariate models, etc., based on some criterion you like (e.g., adjusted R^2)
- Choose between these based on some numeric criterion or logic and common sense
- There are different numeric criteria to choose from (more next time)

Variable Selection Methods

- Sometimes it is desirable to force one or more variables into the model (e.g., your primary exposure variable, or important confounders like age or sex, lower order terms, or levels of a categorical variable), and include them in all models at each step – even if they are not statistically significant

- Keep in mind that sets of variables sometimes function as a group (e.g., indicator variables for a multi-level categorical variable)

- Automated procedures could allow in quadratic or cubic terms without having a linear term included, which would be odd (hierarchy principle)

Variable Selection Methods

- Variable selection methods don't automatically give any consideration to interaction terms, categorical vs. continuous covariates, linear vs. quadratic covariates, etc.

- I sometimes use these variable selection methods as a starting point towards final model selection, but then use logic, common sense, and medical/epidemiological knowledge and advice to finalize a model

Modeling Rules of Thumb

- Forward selection, backward elimination, and stepwise procedures can be helpful to focus your modeling

- Always go back and consider a purposeful selection of covariates (including confounders and effect modifiers) and check model assumptions (LINE), outliers, and high influence observations

- There may be more than one “final model”

Parsimony

- “Economy or simplicity of assumptions in logical formulation”
- What is the smallest/simplest model that adequately and appropriately fits your data?
- Sometimes a simple model is more helpful than a complex model, because of ease of interpretation

Hierarchical Modelling

- Including lower order terms when higher order terms are included in the model (e.g., linear when quadratic is included, main effects when interaction is included)
- Can sometimes refer to hierarchies of covariates used in modelling (e.g., first demographic factors only, then demographic factors plus exposures, then all covariates)

```
. gen int1 = dhca * vsd

. gen int2 = minutes * vsd

. gen minsq = minutes * minutes

. gen bwtsq = birthwt * birthwt

. gen agesq = age * age

. stepwise, pe(.10): regress pdi dhca vsd minutes birthwt age int1 int2 minsq b
> wtsq agesq
                      begin with empty model
p = 0.0024 <  0.1000  adding  int1
p = 0.0164 <  0.1000  adding  birthwt
p = 0.0156 <  0.1000  adding  minsq
p = 0.0274 <  0.1000  adding  agesq

      Source |       SS             df       MS      Number of obs   =       142
-------------+------------------------------------   F(4, 137)       =      6.99
       Model |  5702.54299           4  1425.63575   Prob > F        =    0.0000
    Residual |  27956.0767         137  204.058954   R-squared       =    0.1694
-------------+------------------------------------   Adj R-squared   =    0.1452
       Total |  33658.6197         141  238.713615   Root MSE        =    14.285

------------------------------------------------------------------------------
         pdi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        int1 |  -12.33457    4.27716    -2.88   0.005    -20.79236   -3.876778
     birthwt |    .008021   .0028849     2.78   0.006     .0023162    .0137258
       minsq |  -.0020328   .0007693    -2.64   0.009    -.0035541   -.0005115
       agesq |    .004682   .0020995     2.23   0.027     .0005304    .0088336
       _cons |   70.67791   10.28458     6.87   0.000     50.34087    91.01496
------------------------------------------------------------------------------
```

```
. stepwise, pr(.10): regress pdi dhca vsd minutes birthwt age int1 int2 minsq b
> wtsq agesq
                      begin with full model
p = 0.9104 >= 0.1000  removing bwtsq
p = 0.8081 >= 0.1000  removing int2
p = 0.6449 >= 0.1000  removing age
p = 0.3043 >= 0.1000  removing dhca
p = 0.3940 >= 0.1000  removing vsd
p = 0.4614 >= 0.1000  removing minutes

      Source |       SS             df       MS      Number of obs   =       142
-------------+------------------------------------   F(4, 137)       =      6.99
       Model |  5702.54299           4  1425.63575   Prob > F        =    0.0000
    Residual |  27956.0767         137  204.058954   R-squared       =    0.1694
-------------+------------------------------------   Adj R-squared   =    0.1452
       Total |  33658.6197         141  238.713615   Root MSE        =    14.285

------------------------------------------------------------------------------
         pdi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       agesq |    .004682   .0020995     2.23   0.027      .0005304    .0088336
        int1 |  -12.33457    4.27716    -2.88   0.005     -20.79236   -3.876778
       minsq |  -.0020328   .0007693    -2.64   0.009     -.0035541   -.0005115
     birthwt |    .008021   .0028849     2.78   0.006      .0023162    .0137258
       _cons |   70.67791   10.28458     6.87   0.000      50.34087    91.01496
------------------------------------------------------------------------------
```

```
. stepwise, pe(.10) lockterm1: regress pdi vsd dhca minutes birthwt age
                      begin with term 1 model
p = 0.0103 <  0.1000  adding  dhca
p = 0.0069 <  0.1000  adding  birthwt

      Source |       SS             df       MS      Number of obs   =       142
-------------+------------------------------------   F(3, 138)       =      6.54
       Model |  4190.84279           3   1396.9476   Prob > F        =    0.0004
    Residual |  29467.7769         138   213.534615   R-squared       =    0.1245
-------------+------------------------------------   Adj R-squared   =    0.1055
       Total |  33658.6197         141   238.713615   Root MSE        =    14.613

------------------------------------------------------------------------------
         pdi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         vsd |  -6.224385   2.972168    -2.09   0.038    -12.10126   -.3475067
        dhca |  -6.952038   2.458269    -2.83   0.005    -11.81278   -2.091293
     birthwt |   .0080636   .0029403     2.74   0.007     .0022497    .0138775
       _cons |   71.53173   10.51437     6.80   0.000     50.74162    92.32183
------------------------------------------------------------------------------
```

Criterion Methods

- Some model criteria are not very useful in model selection because they can always be improved by the addition of additional covariates (e.g., R^2, $\hat{\sigma}^2$ = SSE / n [rather than $\hat{\sigma}^2$ = SSE / (n – (p + 1))])
- However, other criteria exist that could be helpful in selecting between models

Adjusted R^2

- Adjusted R^2 can be used as a possible criterion for selecting between models—with drawback
- It roughly has the interpretation as being the proportion of variability in the outcomes explained by your model, adjusting for the number of covariates in the model

Adjusted R^2

- Although adding a new covariate will increase R^2, it may or may not increase adjusted R^2
- Given a set of candidate models, the preferred model could be the one with the *maximum* R^2 value
- A covariate that increases adjusted R^2 may not necessarily be statistically significant

Akaike's Information Criteria (AIC)

- Pronounced 'ah-kai-ee-kay'

$$AIC = \{2x(\#\text{ of parameters}) - 2\log(L)\}$$

- where L is the maximized likelihood from the data

- Given a set of candidate models, the preferred model could be the one with the *minimum* AIC value
- Models don't need to be nested; don't need to have same covariates, but must have same outcome
- Dataset and number of observations must be the same
- (Same holds for adjusted R-squared)

AIC: Linear Regression Likelihood

- Here $Y_i \sim N(\beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \ldots + \beta_p \cdot x_{ip}, \sigma^2)$

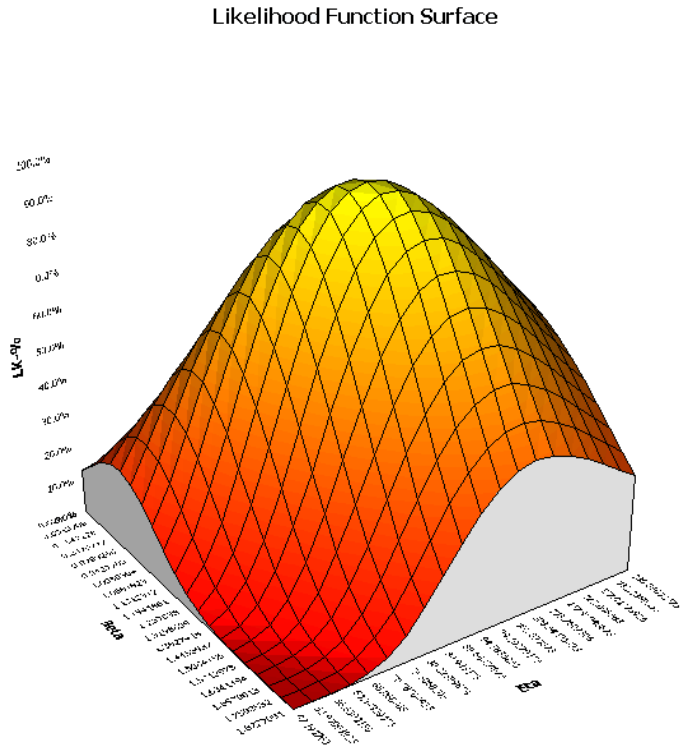
  or $Y_i \sim N(\mu_i, \sigma^2)$

- The likelihood function is given by

$$L(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2) = \prod_{i=1}^{n} \exp(-(y_i - \mu_i)^2/2\sigma^2)/\sqrt{2\pi\sigma^2}$$

- The β vector (MLE) that maximizes this likelihood is also the least squares estimator (LSE) of β

Akaike's Information Criteria (AIC)



AIC = {2x(# of parameters) – 2 log(*L*)} where *L* is the maximized likelihood (MLE) from the data

Akaike's Information Criteria (AIC)

- AIC is based on information theory, dealing with the trade-off between the goodness of fit of the model and the complexity of the model, but tells us nothing about the quality of the model in an absolute sense

Akaike's Information Criteria (AIC)

- AIC is more useful in predictions of fitted values for future observations rather then estimation of β coefficients
- Generally more generous in allowing covariates than hypothesis testing (which recall is based on *p*-values)