36-401 Variable (Model) Selection (part 1)

- *Goals*: Learning relationships and/or coefficients vs. predicting future cases
- *Constraints*: Costs of measurement, parsimony

```
> wages _ read.table("wage.dat",header=T)
> wages _ cbind(wages[,names(wages)!="Wage"],LogWage=log(wage$Wage))
> cor(wages[names(wages)="LogWage",]
       Educ      South        Sex     Exper     Union       Age       Race
  0.3803983 -0.1715523 -0.2185518 0.1076815 0.2080167 0.198224 0.09902679
      Occup     Sector       Marr LogWage
-0.01110503 0.07721102 0.1357926       1
```

- *Mean squared error* = variance + bias^2. More variables increases variance and may reduce bias.
- *Risks:*
  - ◇ Inclusion of irrelevant variables
    (Any X_k for which true $\beta_k \approx 0$. Drives up all $Var(\hat{\beta}_j)$'s, $Var(\hat{y}_i)$'s.)
  - ◇ Exclusion of relevant variables
    (Any X_k for which true $\beta_k \neq 0$. If X_k is correlated with other X_j's, excluding X_k drives up bias in $\hat{\beta}_j$'s $\hat{y}_i$'s. *Meaning and sign of $\hat{\beta}_j$'s suspect.*)
- *Correlation and multicollinearity of covariates (X's)*:
  - ◇ More uncertainty (higher Std Err's) in estimation of individual coefficients;
  - ◇ Interpretations based on changing one covariate at a time less meaningful;
  - ◇ Model selection more complex; may result in many similarly good models.

```
> summary(lm(LogWage ~ . , data=wages))  # ``.'' means ``include everything''!

              Value Std. Error t value Pr(>|t|)
(Intercept)  1.0786  0.6875     1.5688  0.1173
       Educ  0.1794  0.1108     1.6195  0.1059
      South -0.1024  0.0428    -2.3903  0.0172
        Sex -0.2220  0.0399    -5.5628  0.0000
      Exper  0.0958  0.1108     0.8648  0.3875
      Union  0.2005  0.0525     3.8206  0.0001
        Age -0.0854  0.1107    -0.7716  0.4407
       Race  0.0504  0.0285     1.7667  0.0779
      Occup -0.0074  0.0131    -0.5658  0.5718
     Sector  0.0915  0.0387     2.3611  0.0186
       Marr  0.0766  0.0419     1.8271  0.0683

[...]
```

- Horrid multicollinearity: "Error in lm.fit.qr(x, y): computed fit is singular, rank K"
- Bad multicollinearity: Correlation of *coefficients* greater than e.g. 0.9 in absolute value

```
> summary(lm(LogWage ~ . , data=wages))  # ``.'' means ``include everything''!

[...]

Correlation of Coefficients:
       (Intercept)    Educ   South     Sex   Exper   Union     Age    Race
  Educ  0.9566
 South -0.0342      0.0318
   Sex -0.0921     -0.0515  0.0302
 Exper  0.9708      0.9976  0.0223 -0.0550
 Union  0.0161      0.0075  0.0975  0.1201  0.0102
   Age -0.9725     -0.9973 -0.0215  0.0537 -0.9999 -0.0122
  Race -0.1048     -0.0172  0.1112 -0.0200 -0.0109  0.1077  0.0108
 Occup -0.1486     -0.0294 -0.0084  0.1428 -0.0421 -0.2130  0.0441 -0.0575
Sector  0.0117      0.0213  0.0215  0.1121  0.0133  0.0135 -0.0146 -0.0064
  Marr  0.0322      0.0403 -0.0304 -0.0042  0.0410 -0.0689 -0.0451 -0.0556

         Occup  Sector
Sector -0.3147
  Marr  0.0186 -0.0365
```

- A check for multicollinearity: collin(my.x.vars) from collin.q. Gives R^2 from regressing each X on all other X's.

```
> source("collin.q")
> args(collin)
function(X, cutoff = 0.7)
> collin(wages[,names(wages)!="LogWage"])
Educ 0.996
Exper 1
Age 1
> summary(lm(Age ~ Educ + Exper,data=wages))

                Value Std. Error   t value  Pr(>|t|)
(Intercept)    6.0067    0.0459   131.0040    0.0000
       Educ    0.9979    0.0031   325.8991    0.0000
      Exper    1.0007    0.0006  1546.9369    0.0000

Residual standard error: 0.173 on 531 degrees of freedom
Multiple R-Squared: 0.9998
>
> # So:     Age  =  6  +  Educ  +  Exper   ; This makes sense!!
>
```

- A common, but not ideal, model selection method is stepwise regression. It is a "greedy" algorithm that guarantees nothing, but often gives a decent model with minimal effort. It is based on an arbitrary cutoff for the F statistic comparing the RSS for two nested models chosen by looking at all possible single variable additions (forward stepwise) or deletions (backward stepwise). The S-plus function is stepwise().

- Model comparison for the same Y and different covariate subsets *of the same size* (same number of covariates) can be based on big $R^2 = \frac{ESS}{TSS_y} = 1 - \frac{RSS}{TSS_y}$ or small $RMS = \frac{RSS}{n-K}$.

- Model comparison between models of different sizes (different numbers of covariates) requires other criteria (bigger is better):

  ◇ adjusted $R^2 = 1 - \frac{RSS/(n-K)}{TSS_y/(n-1)} = R^2 - \frac{K-1}{n-K}(1 - R^2)$

  ◇ (Akaike Information Criterion) AIC = log(likelihood) - K

  ◇ (Bayesian Information Criterion) BIC = log(likelihood) - $\frac{K}{2}\log(n)$

  Important: unadjusted R^2 and RSE *always* increase when variables are added, regardless of their importance.

- Exhaustive model selection is best, if possible. Use sum.step() from sum.step.q. For larger datasets (many cases and/or variables), stepwise selection may be needed.

```
> source("sum.step.q")
> args(sum.step)
function(step.rslt, data, y, sortcol = 1, sigfig = c(IC = 1, RSE = 1, R = 3),
        includeBIC = T, includeAIC = F, includeRSE = F, includeR2 = F,
        includeR2a = T, top = NULL, compact = F)
> args(stepwise)
function(x, y, wt, intercept = T, tolerance = 1e-07, method = "efroymson",
        size.max = ncol(x), nbest = 3, f.crit = 2, xinclude, plot.rss = dev.cur(
        ) != 1, time = 0.02)
> sw.result _ stepwise(wages[,names(wages)!="LogWage"],wages$LogWage,
+ method="exhaustive",nbest=2)
> sum.step(sw.result,wages,"LogWage", includeAIC=T, includeR2=T, sort=0)
           BIC    AIC    R2   R2a Educ South Sex Exper Union Age Race Occup Sector Marr
 1(#1) -380.5 -376.2 0.145 0.143    1     0   0     0     0   0    0     0      0    0
 1(#2) -409.1 -404.8 0.048 0.046    0     0   1     0     0   0    0     0      0    0
 2(#1) -361.9 -355.5 0.211 0.209    1     0   0     1     0   0    0     0      0    0
 2(#2) -361.9 -355.5 0.211 0.208    1     0   0     0     0   1    0     0      0    0
 3(#1) -344.6 -336.0 0.270 0.265    1     0   1     1     0   0    0     0      0    0
 3(#2) -344.7 -336.1 0.269 0.265    1     0   1     0     0   1    0     0      0    0
 4(#1) -339.4 -328.7 0.292 0.287    1     0   1     1     1   0    0     0      0    0
 4(#2) -339.5 -328.8 0.292 0.286    1     0   1     0     1   1    0     0      0    0
 5(#1) -339.1 -326.3 0.301 0.294    1     1   1     1     1   0    0     0      0    0
 5(#2) -339.2 -326.4 0.301 0.294    1     1   1     0     1   1    0     0      0    0
 6(#1) -339.4 -324.4 0.309 0.301    1     1   1     1     1   0    0     0      1    0
 6(#2) -339.4 -324.5 0.308 0.301    1     1   1     0     1   1    0     0      1    0
 7(#1) -340.7 -323.6 0.313 0.304    1     1   1     1     1   0    0     0      1    1
 7(#2) -340.8 -323.6 0.313 0.304    1     1   1     0     1   1    0     0      1    1
 8(#1) -342.3 -323.0 0.317 0.307    1     1   1     1     1   0    1     0      1    1
 8(#2) -342.4 -323.1 0.317 0.307    1     1   1     0     1   1    1     0      1    1
 9(#1) -345.1 -323.7 0.318 0.306    1     1   1     1     1   1    1     0      1    1
 9(#2) -345.3 -323.9 0.318 0.306    1     1   1     1     1   0    1     1      1    1
10(#1) -348.1 -324.6 0.318 0.305    1     1   1     1     1   1    1     1      1    1
 0(#2) -419.0 -416.9 0.000 0.000    0     0   0     0     0   0    0     0      0    0
```

- ◇ Best BIC model:

```
LogWage ˜  Educ + South + Sex + Exper + Union +                      Sector
LogWage ˜  Educ + South + Sex +         Union + Age +                Sector
```

- ◇ Best AIC model:

```
LogWage ˜  Educ + South + Sex + Exper + Union +       Race +         Sector + Marr
```

- ◇ Best R^2_{Adj} model:

```
LogWage ˜  Educ + South + Sex + Exper + Union +       Race +         Sector + Marr
LogWage ˜  Educ + South + Sex +         Union + Age + Race +         Sector + Marr
```

- ◇ Consistently missing: `Occupation`

- If we can collect additional data, model selection can be based on minimizing prediction MSE.
  Estimate of prediction MSE= $\frac{\sum(\hat{Y}_i - Y_i)^2}{n-K}$
  We will see this in the breakout exercise.

Breakout Session – Variable Selection I

```
# generate random covariates
hours_runif(30,5,20)
temp_runif(30,150,200)
f1_rnorm(30); f2_rnorm(30)
f3_rnorm(30); f4_rnorm(30)
f5_rnorm(30); f6_rnorm(30)
f7_rnorm(30); f8_rnorm(30)
f9_rnorm(30); f10_rnorm(30)
# generate dependent variable
yield_100+5.0*hours+0.5*temp + rnorm(30,0,5)
dtf_data.frame(yield,hours,temp,f1,f2,f3,f4,f5,f6,f7,f8,f9,f10)

# Give variable selection info on all possible models
sw_stepwise(dtf[,2:13], dtf$yield, method="exhaustive", nbest=2, plot=F)
sum.step(sw, dtf, "yield", includeAIC=T, includeR2=T, sort=0)
#           BIC    AIC    R2   R2a hours temp f1 f2 f3 f4 f5 f6 f7 f8 f9 f10
#  1(#1) -110.9 -109.5 0.876 0.872     1    0  0  0  0  0  0  0  0  0  0   0
#  1(#2) -140.4 -139.0 0.115 0.084     0    0  0  1  0  0  0  0  0  0  0   0
#  2(#1)  -92.4  -90.3 0.968 0.965     1    1  0  0  0  0  0  0  0  0  0   0
#  2(#2) -111.5 -109.4 0.885 0.876     1    0  0  0  0  0  0  0  1  0  0   0
#  3(#1)  -92.8  -90.1 0.971 0.967     1    1  0  0  0  0  0  1  0  0  0   0
#  3(#2)  -93.1  -90.3 0.970 0.967     1    1  0  1  0  0  0  0  0  0  0   0
#  4(#1)  -93.5  -90.0 0.973 0.968     1    1  0  1  0  0  0  1  0  0  0   0
#  4(#2)  -93.9  -90.4 0.972 0.967     1    1  0  0  0  0  0  1  0  1  0   0
...
# 11(#1) -104.6  -96.2 0.977 0.962     1    1  1  1  1  1  1  1  0  1  1   1
# 11(#2) -104.7  -96.3 0.976 0.962     1    1  1  1  1  1  1  1  1  1  0   1
# 12(#1) -106.6  -97.5 0.977 0.960     1    1  1  1  1  1  1  1  1  1  1   1

# New data (ndtf) is generated from the same model (n=30)
...

# Check prediction error on best AIC model
bstAIC_lm(yield~hours+temp+f2+f6, dtf)
sum((predict(bstAIC,ndtf) - ndtf$yield)^2) / (nrow(ndtf)-5)
# 57.40387

# Best BIC model
bstBIC_lm(yield~ ????????, dtf)
sum((predict(bstBIC,ndtf) - ndtf$yield)^2) / (nrow(ndtf)-3)
# [1] 44.62793

# Use all covariates
allvar_lm(yield~hours+temp+f1+f2+f3+f4+f5+f6+f7+f8+f9+f10, dtf)
sum((predict(allvar,ndtf) - ndtf$yield)^2) / (nrow(ndtf)-13)
# [1] 100.7593
```

1. What are the ”true” values of the intercept and slopes for the simulated data?

2. In the stepwise procedure, method="exhaustive" means that all models are tried. Best=2 means that only the two best models of each size are reported to you. How do you interpret the line labeled "1(#2)"?

3. What is the best BIC model? What exactly is being calculated by the "sum((predict(bstBIC,ndtf) - ndtf$yield)^2)..." line? Which model predicts best?

4. What is the best model using adjusted R^2? In what sense is BIC "more conservative" than AIC and adjusted R^2?

5. Consider the case where one meaningless variables is added to a model along with 2 relevant variables. Over many datasets what, on average, will be the value of b_3?