# 36-401 Class Notes and Exercise for 9/26/00

**Howard Seltman**

## Variable (Model) Selection (part 1)

- Goals: Learning relationships and/or coefficients vs. predicting future cases
- Constraints: Costs of measurement, parsimony
- Mean squared error of coefficient estimates = variance + bias$^2$.
  More variables increases variance and may reduce bias.
- Inclusion of irrelevant variables
- Exclusion of relevant variables
  RwG defines covariate $X_k$ as relevant if $\beta_k \neq 0$ and $X_k$ is correlated with some other X variables. Under this definition, exclusion of relevant variables makes the model un-trustworthy, particularly for meaning (and even sign) of the coefficients.
- Correlation and multicollinearity of covariates: prevents precise estimation of individual coefficients; makes interpretations based on changing one covariate at a time less meaningful; makes model selection more complex and may result in many similarly good models.

```
> jnk_data.frame(a=c(3,5,2,7,1,4), b=c(2,3,5,4,2,3), d=c(7,11,12,15,5,10))
> round(cor(jnk),3)
      a     b     d
a 1.000 0.264 0.776
b 0.264 1.000 0.813
d 0.776 0.813 1.000
> lm(d~a+b, jnk)
Coefficients:
   (Intercept) a b
 -2.175584e-15 1 2
Residual standard error: 3.455541e-16
Multiple R-squared: 1  # from summary()
```

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

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

- 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 $R^2$ or small RSE.

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

  - ◇ adjusted $R^2 = R^2 - \frac{K-1}{n-K}(1 - R^2)$
  - ◇ (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 step.sum() from step.sum.q. For larger datasets (many cases and/or variables), stepwise selection may be needed.

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

```
rslt(lm(y~x1+x2+x3, original))
pred_predict(rslt, moredata)
MSEPred_sum( (pred$fit - moredata$y)^2 ) / (nrow(moredata)-4)
```

# Breakout Session for 9/26/00

```
# 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 $b_3$?