Lecture 9
Variable selection

BIOST 515

February 3, 2004

Model Building

Previously, we have assumed that the regressors included in the model are known to be important. We focused on

- Correct functional form
- Verification of underlying assumptions
- Inclusion of known confounders

Sometimes, we will have a database of candidate regressors that should include all influential variables, but the actual subset to be included has yet to be determined. Finding the appropriate subset of regressors is a **variable selection** problem.

Variable selection

- In controlled settings (such as clinical trials), variable subset selection is usually not necessary.

- However, in observational studies, we may collect a large number of potentially predictive variables without adequate scientific or historical information to tell us which variables are important to include in regression.

- In prediction settings, we may also be unsure about which variables will be predictive of the outcome. In that case, we may want to determine the "best" set of predictors.

Possible objectives in model building and variable selection

- Include as many regressors as possible to have the most information when predicting values of the outcome.
- We want the model to be as parsimonious as possible because
  - Interpretation is difficult with a large number of predictors
  - Excess variables may increase the variability of $\hat{\beta}$
  - Our interest may be in which variables are predictive and not in prediction itself.
  - Variance of the prediction of $\hat{y}$ increases as the number of regressors increases
  - We may want the smallest subset for prediction for practical reasons - cost, time, availability, etc.
  - Estimation problems may occur with too many variables (multi-collinearity)

Selection the "best" regression model

There is no "best" regression model.

- What is "best"?
- There are a number of ways we can choose the "best" - they will not all yield the same results.
- What about the other potential problems with the model that might have been ignored while selecting the "best" model?

Consequences of model misspecification

Suppose there are K candidate regressors, $x_1, x_2, \ldots, x_k$ and $n \geq K + 1$ observations. The full model is

$$y_i = \beta_0 + \sum_{j=1}^{K} \beta_j x_{ij} + \epsilon_i, \ i = 1, \ldots, n.$$

We will assume

- The list of condidate regressors includes all the important variables (no unmeasured confounders).
- The intercept, β_0, will always be in the model.

Suppose we delete r regressors from the model and retain $p = K - r$ ($p + 1$ total regressors including the intercept).

The model may be rewritten as

$$y = X_{p+1}\beta_{p+1} + X_r\beta_r + \epsilon,$$

where the X matrix has been partitioned into X_{p+1} and X_r and β has been partitioned into β_{p+1} and β_r.

For the full model, the least squares estimate of β is

$$\hat{\beta}^* = (X'X)^{-1}X'y$$

and an estimate of the variance (σ^2) is

$$\hat{\sigma}^{2*} = \frac{y'(I - X(X'X)^{-1}X')y}{n - K - 1}.$$

The components of $\hat{\beta}^*$ are denoted by $\hat{\beta}^*_{p+1}$ and $\hat{\beta}^*_r$, and $\hat{y}^*_i$ denotes the fitted values.

The subset model is denoted by

$$y = X_{p+1}\beta_{p+1} + \epsilon,$$

where the least squares estimate of β_{p+1} is

$$\hat{\beta}_{p+1} = (X'_{p+1}X_{p+1})^{-1}X'_{p+1}y$$

and the estimate of the residual variance is

$$\hat{\sigma}^2 = \frac{y'(I - X_{p+1}(X'_{p+1}X_{p+1})^{-1}X'_{p+1})y}{n - p - 1}$$

and the fitted values are $\hat{y}_i$.

Some properties of the estimates $\hat{\beta}_{p+1}$ and $\hat{\sigma}^2$

1. $E[\hat{\beta}_{p+1}] = \beta_p + (X'_{p+1}X_{p+1})^{-1}X'_{p+1}X_r\beta_r$
   $\hat{\beta}_{p+1}$ is biased except . . . ?
2. $\text{var}(\hat{\beta}_{p+1}) = \sigma^2(X'_{p+1}X_{p+1})^{-1}$ and $\text{var}(\hat{\beta}^*) = \sigma^2(X'X)^{-1}$. Also, $\text{var}(\hat{\beta}_{p+1}) - \text{var}(\hat{\beta}^*_{p+1})$ is positive semidefinite; therefore, elementwise, $\text{var}(\hat{\beta}_{p+1}) \leq \text{var}(\hat{\beta}^*_{p+1})$. Deleting variables never increases the variances of the estimates of the remaining parameters.
3. The estimate $\hat{\sigma}^{2*}$ from the full model is an unbiased estimate of σ^2. For the subset model

$$E[\hat{\sigma}^2] = \sigma^2 + \frac{\beta'_r X'_r(I - X_{p+1}(X'_{p+1}X_{p+1})^{-1}X'_{p+1})X_r\beta_r}{n-p-1}$$

Criteria for Evaluation

How can we evaluate and compare different candidate models?

- Coefficient of multiple determination
- Adjusted R^2
- Residual mean square
- AIC

Note: We should always balance these criteria-based model selection techniques with scientific knowledge.

Criteria for Evaluation – Coefficient of multiple determination

Let $R^2_{(subset,p)}$ denote the R^2 for the regression model with p regressors ($p+1$ terms inc. the intercept).

$$R^2_{(subset,p)} = \frac{SSR_{(subset,p)}}{SSTO} = 1 - \frac{MSE_{(subset,p)}}{SSTO},$$

where $SSR_{(subset,p)}$ and $MSE_{(subset,p)}$ denote the regression sum of squares and residual sum of squares for a $p+1$ term subset model. Choose the "best" model by comparing R^2 for different models. Unfortunately, $R^2_{(subset,p)}$ increases with p with a maximum when $p = K$.

Criteria for Evaluation - Adjusted R^2

Instead of R^2, we may use the adjusted R^2 as it may be more interpretable. The ajusted R^2 for a $p+1$-term model is

$$R^2_{adj(subset,p)} = 1 - \frac{n-1}{n-p-1}(1 - R^2_{(subset,p)}.$$

We can then choose a model based on the largest $R^2_{adj(subset,p)}$.

Criteria for Evaluation - Residual mean square

The residual mean square for a subset regression model with p regressors,

$$MSE_{(subset,p)} = \frac{MSE_{(subset,p)}}{n-p-1},$$

may be used to evaluate regression models.

- $MSE_{(subset,p)}$ always decreases as the number of regressors in the subset increases

- $MSE_{(subset,p)}$ initially decreases, then stabilizes, then may increase.

To choose a model based on MSE, choose the model with

- The minimum $MSE_{(subset,p)}$, or
- The subset of regressors where $MSE_{(subset,p)}$ is approximately equal to MSE from the full model.

The subset regression that minimizes $MSE_{(subset,p)}$ will also maximize $R^2_{adj,subset}$.

Criteria for Evaluation - change in AIC

$AIC = -2 \times$ maximized log-likelihood $+ 2 \times$ no. of parameters

For a regression model with n observations, $p+1$ regressors and $normally\text{-}distributed\ errors$, the log-likelihood is

$$l(\beta, \sigma^2; y) = c + \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \parallel y - X\beta \parallel^2 .$$

Maximizing over β yields

$$l(\hat{\beta}, \sigma^2; y) = c + \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} SSE.$$

If σ^2 is known,

$$AIC = \frac{SSE}{\sigma^2} + 2p + c.$$

If σ^2 is unknown,

$$l(\hat{\beta}, \hat{\sigma}^2; y) = c + \frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2}$$

and

$$AIC = n\log(SSE/n) + 2p + c.$$

We then choose the model that leads to the largest reduction in AIC.

Techniques for variable selection

Now that we have some criteria for comparing models, what are some approaches we can take to compare models? Ideally, we'd have some scientifically meaningful and specified models to compare that were determined before seeing the data. This might not always be the case, though.

- All possible regressions
- Stepwise regression methods
  - Forward selection
  - Backward elimination
  - Stepwise regression

All possible regressions

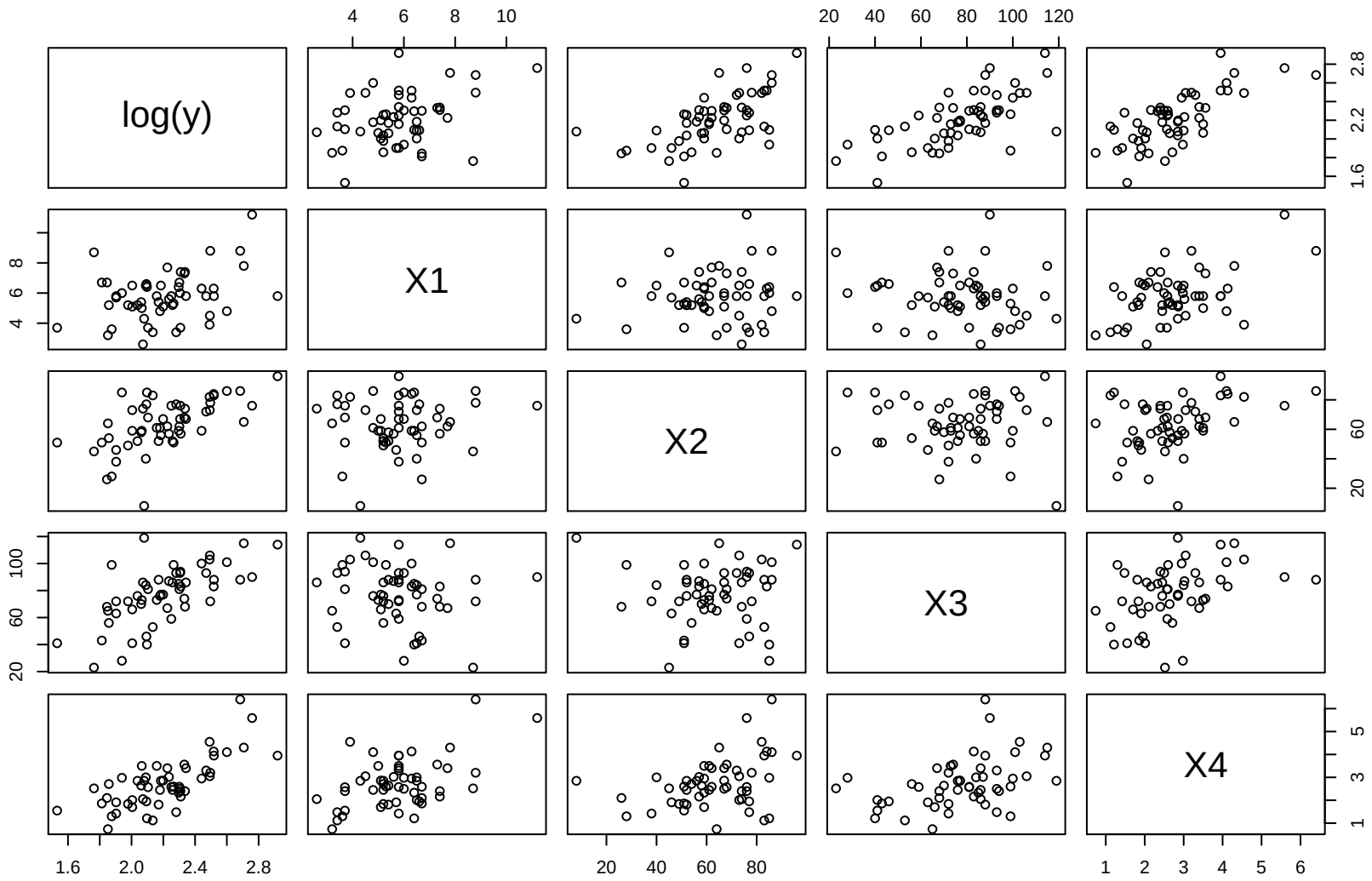
- For all candidate regressors, fit all possible models.
- If there are K candidate regressors, fit and examine 2^K models.

Example

A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 54 patients was available for analysis. The potential predictors are

X_1	blood clotting score
X_2	prognostic index, includes age of patient
X_3	enzyme function test score
X_4	liver function test score

The number of candidate regressors is small. We can easily eplore the different models.


log(y)
X1
X2
X3
X4
4
6
8
10
20
40
60
80
100
120
1.6
2.2
2.8
4
6
8
60
20
100
60
20
1.6
2.0
2.4
2.8
20
40
60
80
1
2
3
4
5
6
1
3
5

Model	AIC	mse	R^2	R^2_{adj}
X_1, X_2, X_3, X_4	-324.71	0.00224	0.972^*	0.970
X_1, X_2, X_3	-326.67^*	0.00220^*	0.972	0.971^*
X_1, X_2, X_4	-189.60	0.0278	0.65	0.63
X_1, X_2	-166.04	0.0438	0.44	0.42
X_1, X_3, X_4	-201.50	0.0223	0.72	0.70
X_1, X_3	-190.96	0.0276	0.65	0.63
X_1, X_4	-175.44	0.0368	0.53	0.51
X_1	-143.82	0.0672	0.12	0.10
X_2, X_3, X_4	-248.73	0.0093	0.88	0.88
X_2, X_3	-225.45	0.046	0.81	0.81
X_2, X_4	-191.54	0.0273	0.65	0.64
X_2	-160.30	0.0495	0.35	0.34
X_3, X_4	-197.56	0.0244	0.69	0.67
X_3	-168.45	0.0426	0.44	0.43
X_4	-177.38	0.0361	0.53	0.52

Selection using AIC, MSE and adjusted R^2 all yield the same

model,

$$E(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

Stepwise regression methods

- Evaluating all possible regressions can be burdensome computationally and for the analyst.
- An alternative might be to compare a scientifically meaningful subset of models.
- When that's not possible or for some reason, not desired, procedures have been developed to evaluate only a small subset of regression models by either adding or deleting regressors one at a time. These fit into 3 categories
  - forward selection
  - backward elimination
  - stepwise regression

Forward selection

1. Begin with the assumption that there are no regressors in the model.

2. Check models with all possible regressors added individually.

3. Add the regressor that most changes your criterion in the correct direction. Go back to 2.

4. If none of the regressors have a positive effect on your criterion, stop with the regressors you have. This is your final model.

Example using $add1()$ in R

```
Forward selection
Single term additions

Model:
surg$logY ~ -1 + 1
        Df Sum of Sq      RSS      AIC
<none>                  3.973 -138.914
surg$X1  1     0.477    3.496 -143.817
surg$X2  1     1.396    2.576 -160.303
surg$X3  1     1.758    2.215 -168.455
surg$X4  1     2.095    1.878 -177.385
Single term additions

Model:
surg$logY ~ surg$X4
        Df Sum of Sq      RSS      AIC
<none>                  1.878 -177.385
surg$X1  1     0.002    1.876 -175.437
surg$X2  1     0.485    1.392 -191.539
```

```
surg$X3  1     0.632    1.245 -197.558
Single term additions

Model:
surg$logY ~ surg$X4 + surg$X3
        Df Sum of Sq      RSS      AIC
<none>                   1.245 -197.558
surg$X1  1     0.130    1.116 -201.498
surg$X2  1     0.780    0.465 -248.730
Single term additions

Model:
surg$logY ~ surg$X4 + surg$X3 + surg$X2
        Df Sum of Sq     RSS     AIC
<none>                  0.47 -248.73
surg$X1  1      0.36    0.11 -324.71
```

Note: RSS=residual sums of squares (SSE)

Model selected by forward selection using AIC:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4888	0.0502	9.73	0.0000
X4	0.0019	0.0097	0.20	0.8436
X3	0.0095	0.0004	23.91	0.0000
X2	0.0093	0.0004	21.19	0.0000
X1	0.0685	0.0054	12.60	0.0000

Backward elimination

1. Start with all p candidate regressors in the model

2. Drop the predictor that improves your selection criterion the least

3. Continue until there is no predictor that can be dropped and result in an improvement of your selection criterion, then all the remaining predictors define your final model.

Example using $drop1()$ in R

```
Backward elimination
Single term deletions

Model:
surg$logY ~ surg$X4 + surg$X3 + surg$X2 + surg$X1
        Df Sum of Sq     RSS     AIC
<none>                  0.11 -324.71
surg$X1  1      0.36    0.47 -248.73
surg$X2  1      1.01    1.12 -201.50
surg$X3  1      1.28    1.39 -189.60
surg$X4  1  8.81e-05    0.11 -326.67

Model:
surg$logY ~ surg$X3 + surg$X2 + surg$X1
        Df Sum of Sq     RSS     AIC
<none>                  0.11 -326.67
surg$X1  1      0.63    0.74 -225.45
surg$X2  1      1.30    1.41 -190.96
surg$X3  1      2.12    2.23 -166.04
```

Model selected by backward elimination

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4836	0.0426	11.34	0.0000
X3	0.0095	0.0003	31.08	0.0000
X2	0.0093	0.0004	24.30	0.0000
X1	0.0692	0.0041	16.98	0.0000

Stepwise regression

General stepwise regression techniques are usually a combination of backward elimination and forward selection, alternating between the two techniques at different steps.

Example using $stepAIC()$ in the MASS library in R

```
Start with no regressors
Start:  AIC= -138.91
 surg$logY ~ -1 + 1

          Df Sum of Sq      RSS      AIC
+ surg$X4  1     2.095    1.878 -177.385
+ surg$X3  1     1.758    2.215 -168.455
+ surg$X2  1     1.396    2.576 -160.303
+ surg$X1  1     0.477    3.496 -143.817
<none>                    3.973 -138.914

Step:  AIC= -177.38
 surg$logY ~ surg$X4

          Df Sum of Sq      RSS      AIC
+ surg$X3  1     0.632    1.245 -197.558
+ surg$X2  1     0.485    1.392 -191.539
```

```
<none>                    1.878 -177.385
+ surg$X1  1     0.002    1.876 -175.437
- surg$X4  1     2.095    3.973 -138.914

Step:  AIC= -197.56
 surg$logY ~ surg$X4 + surg$X3

          Df Sum of Sq      RSS      AIC
+ surg$X2  1     0.780    0.465 -248.730
+ surg$X1  1     0.130    1.116 -201.498
<none>                    1.245 -197.558
- surg$X3  1     0.632    1.878 -177.385
- surg$X4  1     0.970    2.215 -168.455

Step:  AIC= -248.73
 surg$logY ~ surg$X4 + surg$X3 + surg$X2

          Df Sum of Sq     RSS     AIC
+ surg$X1  1      0.36    0.11 -324.71
<none>                    0.47 -248.73
- surg$X4  1      0.28    0.74 -225.45
```

```
- surg$X2  1      0.78    1.25 -197.56
- surg$X3  1      0.93    1.39 -191.54

Step:  AIC= -324.71
 surg$logY ~ surg$X4 + surg$X3 + surg$X2 + surg$X1

          Df Sum of Sq     RSS     AIC
- surg$X4  1  8.81e-05    0.11 -326.67
<none>                    0.11 -324.71
- surg$X1  1      0.36    0.47 -248.73
- surg$X2  1      1.01    1.12 -201.50
- surg$X3  1      1.28    1.39 -189.60

Step:  AIC= -326.67
 surg$logY ~ surg$X3 + surg$X2 + surg$X1

          Df Sum of Sq     RSS     AIC
<none>                    0.11 -326.67
+ surg$X4  1  8.81e-05    0.11 -324.71
- surg$X1  1      0.63    0.74 -225.45
- surg$X2  1      1.30    1.41 -190.96
```

```
- surg$X3  1      2.12    2.23 -166.04

Call:
lm(formula = surg$logY ~ surg$X3 + surg$X2 + surg$X1)

Coefficients:
(Intercept)      surg$X3      surg$X2      surg$X1
   0.483621     0.009524     0.009295     0.069225

Start with full model
Start:  AIC= -324.71
 surg$logY ~ surg$X4 + surg$X3 + surg$X2 + surg$X1

          Df Sum of Sq     RSS     AIC
- surg$X4  1  8.81e-05    0.11 -326.67
<none>                    0.11 -324.71
- surg$X1  1      0.36    0.47 -248.73
- surg$X2  1      1.01    1.12 -201.50
- surg$X3  1      1.28    1.39 -189.60

Step:  AIC= -326.67
```

```
 surg$logY ~ surg$X3 + surg$X2 + surg$X1

          Df Sum of Sq     RSS     AIC
<none>                    0.11 -326.67
- surg$X1  1      0.63    0.74 -225.45
- surg$X2  1      1.30    1.41 -190.96
- surg$X3  1      2.12    2.23 -166.04

Call:
lm(formula = surg$logY ~ surg$X3 + surg$X2 + surg$X1)

Coefficients:
(Intercept)      surg$X3      surg$X2      surg$X1
   0.483621     0.009524     0.009295     0.069225
```

Comments on stepwise regression techniques

- The techniques we've discussed in this lecture are all quantitative/computational and therefore have many scientific drawbacks.
- They should NEVER replace careful scientific thought and consideration in model building.
- How do these techniques fit in with our scientific examination of precision variables and confounders?
- How should we treat influential observations in a stepwise selection setting?

Frank Harrell's summary of problems with stepwise variable selection

1. It yields R^2 values that are biased high.

2. The ordinary F and χ^2 statistics do not have the claimed distribution. Variable selection is based on methods that were intended to be used to test only prespecified hypotheses.

3. The method yields standard errors of regression coefficient estimates that are biased low and confidence intervals for effects and predicted values that are falsely narrow.

4. It yields p-values that are too small (i.e.- there are severe multiple testing problems) and that do not have the proper

meaning, and the proper correction for them is a difficult problem.

5. It provides regression coefficients that are biased high in absolute value and need shrinkage.

6. Rather than solving problems caused by collinearity, variable selection is made arbitrary by collinearity.

7. It allows us not to think about the problem.