ST5202: Applied Regression Analysis

Department of Statistics and Applied Probability
National University of Singapore

19-March-2018
Lecture 8

Lecture 8

Model Selection and Validation (Ch. 9)

Announcement

- Assignment #4 will be released by tomorrow morning. Due on 26 March.

Main Idea

- We use multiple linear regression analysis to predict outcomes of the response variable, or examine associations between predictor variables and the response variables, e.g.
  a. predict survival of patients undergoing a liver operation (book Ch 9.2)
  b. life expectancy example (lecture 6)
- Often it is not clear which variable to include in the model
  - for $p-1$ candidate predictor variables, there are 2^{p-1} different candidate models.

    It can be even more once we include higher order terms (e.g., interactions)
- Why not include a convenient subset of variables, or all variables?

Overview of model building process

- Data collection and preparation
- Reduction of explanatory variables
- Model refinement and selection
- Model validation

Data collection - experimental studies

- Controlled experiments
  - levels of explanatory variables are set by the experimenter
  - The explanatory variables are *factors* or *control variables*
  - Example: investigate treatment effect in a random clinical trial. explanatory variable: treatment group or placebo group
- Controlled experiments with covariates
  - Covariates: extra *uncontrolled* variable, incorporated to the regression model to reduce the variance of the experimental error terms
  - Examples: age of patients in clinical trial.

Data collection - observational studies

- Confirmatory observational studies
  - variables known to affect response variable: control variables
  - new variables involved in hypotheses: primary variables
  - example: an observational study of effect of Vitamin E on occurrence of a certain type of cancer
    - control variables: age, gender, and race
    - primary variable: amount of Vitamin E taken
- Exploratory observational studies
  when controlled experiments can not be conducted and no adequate knowledge for confirmatory studies collect variables which might be related to response variable

Data preparation and preliminary model investigation

- Identify gross data errors and extreme outliers
  functional forms of explanatory variables in the model
  possible interaction terms
  the need for transformation
- Prior knowledge and relevant background expertise
  scatter plots, residual plots, lowess, etc

Reduction of explanatory variables

- controlled experiments-not important since explanatory variables are chosen for investigation
- controlled experiments with covariates-reduction may be possible, but the number of covariates is usually small
- confirmatory observational studies-no reduction needed
- explanatory observational studies
  investigators may have little idea of the driving predictors and so will cast a wide net in data collection, hoping that analysis will identify the important variables. We cannot expect all the predictors are useful. Want to Identify a few "good" subsets of variables for investigation

Why variable selection?

- What happens, if we simply use $\hat{Y} = 0$?
    - The variance of the prediction is 0 (the variance for any constant is 0).
    - The bias of the prediction may be fairly large.
    - So $E\{e^2\}$ may be fairly large, where $e = Y - \hat{Y}$
- What happens, if we use the full model?
    - The variance of the prediction > 0.
    - The bias of the prediction is 0, providing the true model is a sub-model of the full model.
    - Also $E\{e^2\}$ may be large

Why variable selection?

- $E\{e^2\} = (E\{e\})^2 + Var(e) = Bias(prediction)^2 + Var(prediction)$
- Variable selection is a trade off between the bias and variance.
- Including some important predictors, compared with $\hat{Y} = 0$, the variance of the prediction may increase a little bit, but the bias will decrease significantly.
- Or deleting some unimportant predictors, compared with the situation of full model, the variance of the prediction may decrease significantly and the bias will increase a little bit
- In the above two cases, we reach a smaller $E\{e^2\}$ for the prediction.

Too few predictor variables

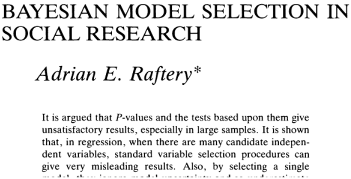
- Leaving out "key explanatory variables" can bias the estimates of the mean response
  - If income is very high region A, and very low in region B, leaving out region as a predictor leads to biased estimates in both regions
- Leaving out confounding variables can bias the estimates of the regression coefficients
  - If you want to examine the association between exercise and health, a possible confounder is age
    (there might be a negative associations between exercise and age, and age and health),
    not adjusting/controlling for age in your model can give a biased regression coefficients for exercise.

Too many predictor variables

- Then why not include all candidate predictor variables?
- Problems with (too) many predictor variables in the model:
  - Uncertainty in fitted regression line increases with p: $\sum_i \sigma^2\{\hat{Y}_i\} = \sum h_{ii}\sigma^2 = p\sigma^2$
  - Degrees of freedom $n - p$ decreases with p (thus spread of sampling distributions of b_k's, $\hat{Y}_h$ increases with p)
  - We can get collinearity among the predictors, which increases the standard errors of the coefficients of those predictors
  - We waste time/money to measure or collect unnecessary predictors

Conclusion

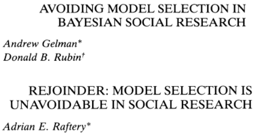
- We want to explain the data in the simplest way:
  the “best” model is the smallest model that fits the data
- No agreed upon best method, there is a lot of discussion/literature on this topic



BAYESIAN MODEL SELECTION IN SOCIAL RESEARCH

*Adrian E. Raftery**

It is argued that *P*-values and the tests based upon them give unsatisfactory results, especially in large samples. It is shown that, in regression, when there are many candidate independent variables, standard variable selection procedures can give very misleading results. Also, by selecting a single



AVOIDING MODEL SELECTION IN BAYESIAN SOCIAL RESEARCH

*Andrew Gelman**
Donald B. Rubin[†]

REJOINDER: MODEL SELECTION IS UNAVOIDABLE IN SOCIAL RESEARCH

*Adrian E. Raftery**

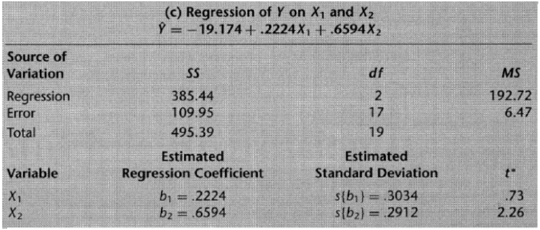
Methods for variable selection

- Stepwise methods:
  Include or exclude a predictor based on its p-value
  - Backward elimination
  - Forward selection
  - Both
- Consider all possible models, and compare them using some criterion

Why can't a simple F test tell?

- Questions: why not just check the p value for each predictor under a "FULL" model for model selection?
    - or through a t test
- This is equivalent to deleting/adding several predictors simultaneously from/into the model

Body fat example re-vist



(c) Regression of Y on X_1 and X_2

$\hat{Y} = -19.174 + .2224X_1 + .6594X_2$

Source of Variation	SS	df	MS
Regression	385.44	2	192.72
Error	109.95	17	6.47
Total	495.39	19	
Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = .2224$	$s\{b_1\} = .3034$	.73
X_2	$b_2 = .6594$	$s\{b_2\} = .2912$	2.26

- Consider model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$
- Test $\beta_1 = 0$: $\frac{SSR(X_1|X_2)/1}{SSE(X_1,X_2)/17} = 0.537 < F(0.99, 1, 17) = 8.400$
- Test $\beta_2 = 0$: $\frac{SSR(X_2|X_1)/1}{SSE(X_1,X_2)/17} = 5.129 < F(0.99, 1, 17) = 8.400$
- Test $\beta_1 = \beta_2 = 0$: $\frac{SSR(X_1,X_2)/2}{SSE(X_1,X_2)/17} = 29.798 > F(0.99, 2, 17) = 6.112$

More issues with a simple F test

- Answer to the previous question:
  - In some cases, the number of predictors is even larger than the number of observations (e.g., gene expression data). It is even impossible to fit a linear model with all predictors. Therefore, no way to perform the single F (or t) test for all the predictors.
  - Even though the data can support us to do in this way, it is a lousy method. Refer to the body fat example (correlated predictors) above.

Stepwise regression

- Backward elimination:
  - Start with all predictors in the model
  - Remove the predictor with the highest p-value greater than α_{drop}
  - Continue until all p-values are smaller than α_{drop}
  - Note that α_{drop} does not need to be 0.05. If prediction is the main goal, a higher cut-off may work better
- Forward selection:
  - Start with no predictors in the model
  - For all predictors not in the model, compute their p-values for adding them to the model. Choose the one with the lowest p-value, lower than α_{add}
  - Continue until no new predictors can be added
- Both directions: add or remove a variable at each step
- Stepwise methods in R: use "update"

Backward elimination: main procedure

- We start with all the predictors in the model.
- Delete the least significant predictor.
  - Fit the model containing all the p predictors

    $$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \epsilon$$

    and for each predictor calculate the $p-$value of the single $t-$test.
  - Check whether the $p-$values for all the p predictors are smaller than α.
  - If yes, stop the algorithm and all the $p-1$ predictors should be retained.
  - If not, delete the least significant variable, i.e., the variable with the largest $p-$value and go to the next step. Suppose X_j is deleted.

Backward elimination: main procedure

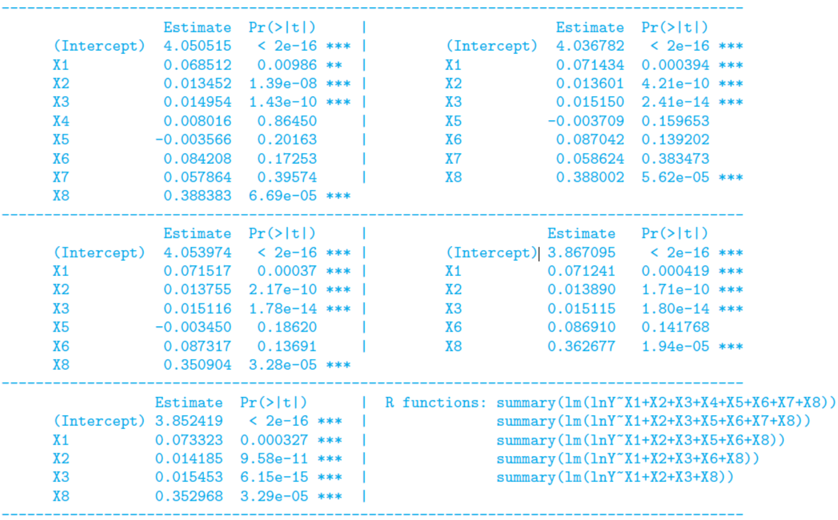
- Delete the second least significant predictor.
  - Refit the model containing the remaining $p-2$ predictors $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{j-1}X_{j-1} + \beta_{j+1}X_{j+1} + \cdots + \beta_{p-1}X_{p-1} + \epsilon$ and for each predictor calculate the $p-$value of single $t-$test.
  - Check whether the $p-$values for all the remaining $p-2$ predictors are smaller than the level α
  - If yes, stop the algorithm and the $p-2$ predictors are retained
  - If not, delete the least significant variable, i.e., the predictors with the largest p-value.
- Continue the above process until all the p-values of the single $t-$test for the predictors in the model are less than α.
- With this algorithm, once a predictor is removed from the model it does not reenter. The present "α" is called "alpha to drop"

Surgical unit example

- Y : survival time of patients after liver operation
  X_1 : blood clotting score
  X_2 : prognostic index
  X_3 : enzyme function test score
  X_4 : liver function test score
  X_5 : age, in years
  X_6 : indicator variable for gender (0=male, 1=female)
  X_7 and X_8: indicator variables for history of alcohol use

Alcohol Use	X_7	X_8
None	0	0
Moderate	1	0
Severe	0	1

Backward elimination: surgical unit example



```
------------------------------------------------------------------------------
             Estimate  Pr(>|t|)     |             Estimate  Pr(>|t|)
(Intercept)  4.050515  < 2e-16 *** |  (Intercept)  4.036782  < 2e-16 ***
X1           0.068512  0.00986 **  |  X1           0.071434  0.000394 ***
X2           0.013452  1.39e-08 *** | X2           0.013601  4.21e-10 ***
X3           0.014954  1.43e-10 *** | X3           0.015150  2.41e-14 ***
X4           0.008016  0.86450      | X5          -0.003709  0.159653
X5          -0.003566  0.20163      | X6           0.087042  0.139202
X6           0.084208  0.17253      | X7           0.058624  0.383473
X7           0.057864  0.39574      | X8           0.388002  5.62e-05 ***
X8           0.388383  6.69e-05 ***
------------------------------------------------------------------------------
             Estimate  Pr(>|t|)     |             Estimate  Pr(>|t|)
(Intercept)  4.053974  < 2e-16 *** |  (Intercept)| 3.867095  < 2e-16 ***
X1           0.071517  0.00037 *** |  X1           0.071241  0.000419 ***
X2           0.013755  2.17e-10 *** | X2           0.013890  1.71e-10 ***
X3           0.015116  1.78e-14 *** | X3           0.015115  1.80e-14 ***
X5          -0.003450  0.18620      | X6           0.086910  0.141768
X6           0.087317  0.13691      | X8           0.362677  1.94e-05 ***
X8           0.350904  3.28e-05 ***
------------------------------------------------------------------------------
             Estimate  Pr(>|t|)     | R functions: summary(lm(lnY~X1+X2+X3+X4+X5+X6+X7+X8))
(Intercept) 3.852419  < 2e-16 ***   |              summary(lm(lnY~X1+X2+X3+X5+X6+X7+X8))
X1          0.073323  0.000327 ***  |              summary(lm(lnY~X1+X2+X3+X5+X6+X8))
X2          0.014185  9.58e-11 ***  |              summary(lm(lnY~X1+X2+X3+X6+X8))
X3          0.015453  6.15e-15 ***  |              summary(lm(lnY~X1+X2+X3+X8))
X8          0.352968  3.29e-05 ***  |
------------------------------------------------------------------------------
```

Forward selection: main idea

- When the number of predictors is larger than the number of observations, backward elimination does not work.
- Forward selection, which proceeds in the opposite way, can work in this situation.
- We start with no predictor in the model and sequentially add more significant variables.
- In principle, the algorithm is as follows.

Forward selection: main procedure

- We start with no predictor in the model.
- Pick the most significant predictor.
  - Fit p simple linear regression models

    $$Y = \beta_0 + \beta_1 X_j + \epsilon, \, j = 1, \cdots, p-1.$$

    For each predictor, we calculate the p-value of the single t-test for the hypothesis $H_0 : \beta_1 = 0$.
  - Choose the most significant predictor (say $X_{(1)}$) such that the p-value of the t-test statistic for the hypothesis $H_0 : \beta_1 = 0$ is smallest.
  - If the p-value for the most significant predictor is larger than α, we stop the algorithm and no predictor is needed.
  - If not, the most significant predictor is added to the model and we go to the next step.

- Pick the second most significant predictor.
    - Fit all the models containing $X_{(1)}$ and another predictor

      $$Y = \beta_0 + \beta_1 X_{(1)} + \beta_2 X_j + \epsilon, , j = 1, \cdots , p - 1, \; X_j \neq X_{(1)}$$

      For each predictor ($j = 1, \cdots , p - 1, \; X_j \neq X_{(1)}$), we calculate the p-value of the single t-test for the hypothesis $H_0 : \beta_2 = 0$ (No interest for $X_{(1)}$).
    - If all the p-values for X_j's are larger than α, stop the algorithm.
    - If one of the p-values is smaller than α, add the predictor with the smallest p-value in the model.
- Continue the above process until no remaining predictors (outside the model) generate a p-value smaller than α.
- With this algorithm, once a predictor enters the model it remains in the model. The present "α" is called "alpha to enter"

Forward selection: surgical unit example

- Select the first variable to enter the model:

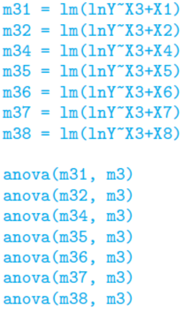
```
m0 = lm(lnY~1)
m1 = lm(lnY~X1)
m2 = lm(lnY~X2)
m3 = lm(lnY~X3)
m4 = lm(lnY~X4)
m5 = lm(lnY~X5)
m6 = lm(lnY~X6)
m7 = lm(lnY~X7)
m8 = lm(lnY~X8)

anova(m0,m1)
anova(m0,m2)
anova(m0,m3)
anova(m0,m4)
anova(m0,m5)
anova(m0,m6)
anova(m0,m7)
anova(m0,m8)
```

- From the R output $m3$ produced the minimum p-value: $8.26e - 08$. Therefore X_3, as the first variable, enters our model.
- Model $\log(Y) \sim X_3$

Forward selection: surgical unit example

- Select the second variable to enter the model:



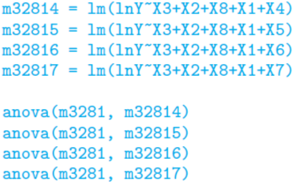
```
m31 = lm(lnY~X3+X1)
m32 = lm(lnY~X3+X2)
m34 = lm(lnY~X3+X4)
m35 = lm(lnY~X3+X5)
m36 = lm(lnY~X3+X6)
m37 = lm(lnY~X3+X7)
m38 = lm(lnY~X3+X8)

anova(m31, m3)
anova(m32, m3)
anova(m34, m3)
anova(m35, m3)
anova(m36, m3)
anova(m37, m3)
anova(m38, m3)
```

- From the R output $m32$ produced the minimum p-value: $2.242e-07$. Therefore X_2, as the second variable, enters our model.
- Model $\log(Y) \sim X_2 + X_3$

Forward selection: surgical unit example

- Continue the above procedure, the added variable, in order, are X_3, X_2, X_8, and X_1.
- The selected model is $\log(Y) \sim X_1 + X_2 + X_3 + X_8$
- In the last step:



```
m32814 = lm(lnY~X3+X2+X8+X1+X4)
m32815 = lm(lnY~X3+X2+X8+X1+X5)
m32816 = lm(lnY~X3+X2+X8+X1+X6)
m32817 = lm(lnY~X3+X2+X8+X1+X7)

anova(m3281, m32814)
anova(m3281, m32815)
anova(m3281, m32816)
anova(m3281, m32817)
```

- The minimum p-value is 0.1418, and therefore, no variable should be added to the model further.

Backward + Forward: main idea

- As we discussed before, a disadvantage of backward elimination is that once a predictor is removed, the algorithm does not allow it to be reconsidered.
- Similarly, with forward selection once a predictor is in the model, its usefulness is not re-assessed at later steps.
- A hybrid of the backward elimination and the forward selection, allows the predictors enter and leave the model several times.
- In principle, the algorithm is as follows.

Backward + Forward: main procedure

- Forward stage: we start as the forward selection using the specified α to enter.
  - If no predictor enters the model, stop the algorithm.
  - If the most significant predictor enters the model, come to the next step.
- Backward stage: after a predictor enters the model, refit the model and check the p-values for the single t-test for all predictors in the model.
  - If all the p-values are smaller than the specified α to drop, go to the next step.
  - If not, conduct the backward elimination until all the p-values for the predictors in the model are smaller than the specified α to drop.
- Continue until no predictor can be added and no predictor can be removed according to the specified α to enter and α to drop.

Backward + Forward: surgical unit example

- Select the first variable to enter the model (forward stage):

```
m0 = lm(lnY~1)
m1 = lm(lnY~X1)
m2 = lm(lnY~X2)
m3 = lm(lnY~X3)
m4 = lm(lnY~X4)
m5 = lm(lnY~X5)
m6 = lm(lnY~X6)
m7 = lm(lnY~X7)
m8 = lm(lnY~X8)

anova(m0,m1)
anova(m0,m2)
anova(m0,m3)
anova(m0,m4)
anova(m0,m5)
anova(m0,m6)
anova(m0,m7)
anova(m0,m8)
```

- From the R output $m3$ produced the minimum p-value: $8.26e-08$. Therefore X_3, as the first variable, enters our model.
- Model $\log(Y) \sim X_3$

Backward + Forward: surgical unit example

- Select the 2nd variable to enter the model (forward stage):

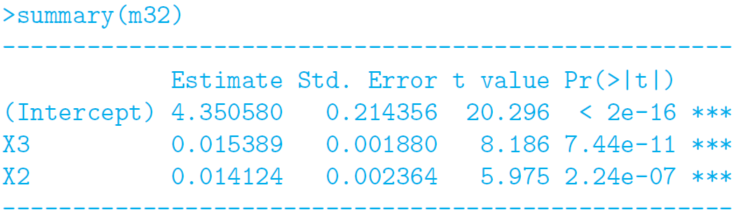
```r
m31 = lm(lnY~X3+X1)
m32 = lm(lnY~X3+X2)
m34 = lm(lnY~X3+X4)
m35 = lm(lnY~X3+X5)
m36 = lm(lnY~X3+X6)
m37 = lm(lnY~X3+X7)
m38 = lm(lnY~X3+X8)

anova(m31, m3)
anova(m32, m3)
anova(m34, m3)
anova(m35, m3)
anova(m36, m3)
anova(m37, m3)
anova(m38, m3)
```

- From the R output $m32$ produced the minimum p-value: $2.242e-07$. Therefore X_2, as the second variable, enters our model.
- Model: $\log(Y) \sim X_2 + X_3$
- **Question:** why don't we use backward stage, insated, directly go another forward stage.

Backward + Forward: surgical unit example

- Try to remove one variable from the model above:



```
>summary(m32)
-----------------------------------------------------
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.350580   0.214356  20.296  < 2e-16 ***
X3          0.015389   0.001880   8.186 7.44e-11 ***
X2          0.014124   0.002364   5.975 2.24e-07 ***
-----------------------------------------------------
```

- All variables are significant. No variable should be deleted. Return back to forward stage.
- Repeat the above steps, finally we obtain the model $\log(Y) \sim X_1 + X_2 + X_3 + X_8$.

Disadvantages of stepwise regression

- Trouble with stepwise regression fans:
  - Motivation: not enough data to estimate the coefficients in any meaningful way
  - At any given step, the model is fit using unconstrained least squares
- Disadvantages
  - Because the "one-at-a-time" adding/dropping of variables, it's possible to miss the "optimal" model
  - Multiple testing issues and how to choose α_{drop} and α_{add}
  - Ad-hoc method: the selected model does not need to optimize any reasonable criterion.
  - Freedman (1983)'s experiment:
    - He sampled 100 "observations" $(X_{i,1}, X_{i,2}, \cdots, X_{i,50}, Y_i)$, all independent random draws from $N(0,1)$
    - Stepwise regression gave a model with 4 predictor variables that were significant at $\alpha = 0.01$, and $R^2 = 0.18$

Modern methods

- If you have to use stepwise regression, I discourage you from using stepwise method based on $p-$values. With the computational power available today, we can do much better.
- A better way is to examine criteria for model selection:
  - These criteria are all based on some measure of model fit, while taking into account the number of parameters in the model
  - Examples: $R^2_{a,p}, C_p, AIC_p, BIC_p = SBC_p, PRESS_p$
- We can search through all possible models, and consider the best model(S) according to the criterion

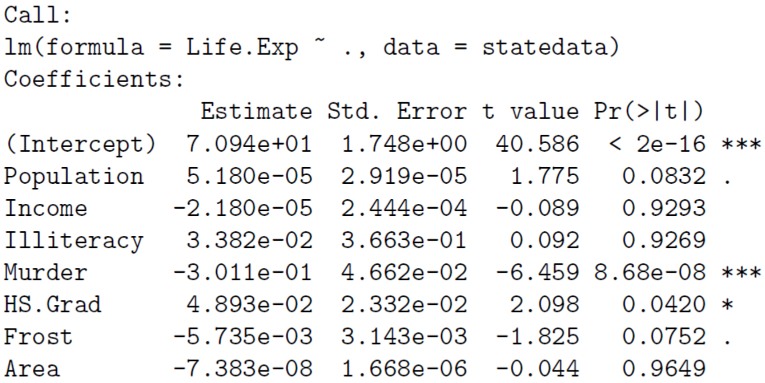
Variable selection Criteria

- Denote number of potential predictor variables with $p - 1$
- We'll include $p - 1$ variables (thus p parameters) and calculate the criterion
- Criteria:
    - Adjusted coefficient of multiple determination
    - Mallow's C_p
    - AIC and BIC
    - PRESS

Example data set

- Data set with for each US state in 1970's:
  income, illiteracy, life expectancy, murder, high-school graduates (%), mean number of days with minimum temperature ¡ 32 degrees (in 1931-1960), land area
- Goal: predict life expectancy (and examine associations between life expectancy and the other variables)

State data: model fit with all predictors



```
Call:
lm(formula = Life.Exp ~ ., data = statedata)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
Population   5.180e-05  2.919e-05   1.775   0.0832 .
Income      -2.180e-05  2.444e-04  -0.089   0.9293
Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *
Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
Area        -7.383e-08  1.668e-06  -0.044   0.9649
```

- Data set with for each US state in 1970's:
  income, illiteracy, life expectancy, murder, high-school graduates (%), mean number of days with minimum temperature ¡ 32 degrees (in 1931-1960), land area
- Goal: predict life expectancy (and examine associations between life expectancy and the other variables)

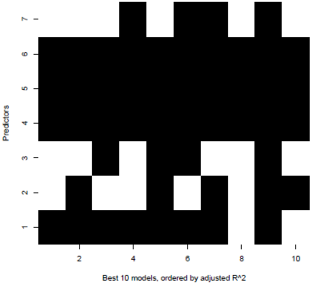
Adjusted coefficients of multiple determination

- Adjusted coefficient of multiple determination:

$$R^2_{a,p} = 1 - \frac{MSE_p}{SSTO/(n-1)}$$

  - $R^2_{a,p}$ decreases if extra predictor(s) increases the MSE
  - We prefer models with large $R^2_{a,p}$, and the smallest number of predictors
  - “leaps” can be used in R to get $R^2_{a,p}$(and C_p) for all models

Results for $R^2_{a,p}$ for state life expectancy example



- How to read the plot:
  The k^{th} column represents the k^{th} best model with respect to $R^2_{a,p}$(i.e., model with the k^{th} highest $R^2_{a,p}$). The j^{th} row represents the j^{th} variable, and it being marked implies that the variable is in the model
  (e.g., For the 8^{th} column, 4^{th}, 5^{th}, and 6^{th} rows are marked. It implies that the model with the 8^{th} highest $R^2_{a,p}$ value (among all) is the model with X_4, X_5, and X_6)
- Model 1456 is a candidate model because it has the largest $R^2_{a,p}$ (1456=Population, Murder, HS.grad, and Frost)
- Model 456 can be considered too, as $R_{a,p}$ doesn't differ that much and it has less parameters.

Mallows' C_p criterion

- This criterion is based on the mean squared error of all $\hat{Y}$'s
- The mean squared error of an estimator quantifies how much an estimator differs from its true values:
  the mean squared error for the fitted $\hat{Y}_i$'s defined as:

$$E\{(\hat{Y}_i - \mu_i)^2\},$$

  with $\mu_i = E\{Y_i\}$, the true (unknown) mean of Y_i
- Notes:
  - This is not the MSE as we defined earlier
  - $E\{\hat{Y}_i\} \neq \mu_i$ if the fitted model is not correct (if $\hat{Y}_i$ is biased)
- Each mean squared error = squared bias + sampling variability:

$$E\{(\hat{Y}_i - \mu_i)^2\} = \left(E\{\hat{Y}_i\} - \mu_i\right)^2 + \sigma^2\{\hat{Y}_i\}$$

(using $(\hat{Y}_i - \mu_i)^2 = [(E\{\hat{Y}_i\} - \mu_i) + (\hat{Y}_i - E\{\hat{Y}_i\})]^2$)

Mallows' C_p criterion: some deviation

- The criterion C_p estimates Γ_p, which is the total mean squared error for n fitted values, over σ^2:

$$\begin{aligned}\Gamma_p &= \frac{1}{\sigma^2}\sum_i E\left[\left(\hat{Y}_i - \mu_i\right)^2\right] \\ &= \frac{\sum_i \text{bias}\{\hat{Y}_i\}^2}{\sigma^2} + \frac{\sum_i \sigma^2\{\hat{Y}_i\}}{\sigma^2} \\ &= \frac{E[SSE_p] - (n-p)\sigma^2}{\sigma^2} + p.\end{aligned}$$

  Note small p and big P in the notation

- Estimate Γ_p by:

$$\begin{aligned}C_p &= \frac{SSE_p - (n-p)MSE(X_1,\ldots,X_{P-1})}{MSE(X_1,\ldots,X_{P-1})} + p, \\ &= \frac{SSE_p}{MSE(X_1,\ldots,X_{P-1})} - (n-2p),\end{aligned}$$

  with $p = P$
  - $C_P = P$

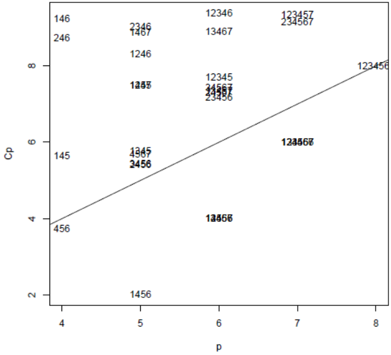
Mallows' C_p criterion: summary

- The criterion C_p estimates the total mean squared error for n fitted values (standardized by σ^2):

$$
\begin{aligned}
C_p &= \frac{SSE_p - (n-p)MSE(X_1,\ldots,X_{P-1})}{MSE(X_1,\ldots,X_{P-1})} + p, \\
&= \frac{SSE_p}{MSE(X_1,\ldots,X_{P-1})} - (n-2p),
\end{aligned}
$$

- We prefer a model with a small C_p, but we have to keep in mind the following:
  If C_p is larger than p, the $\hat{Y}_i$'s could be biased, so we prefer a model with C_p below or at least close to p

Results for C_p for state life expectancy example



- 456 and 1456 both candidate models

AIC_p and BIC_p

- Akaike's information criterion (AIP_p) and Schwarz' Bayesian criterion (SBC_p or BIC_p) are combination of
  1. (negative) measure of model fit
  2. penalty for extra parameters
- Good candidate models have smaller AIC/BIC (max the likelihood and min the number of parameters)
- BIC has a bigger penalty (when $n \geq 8$) thus favors smaller models:

$$\begin{aligned} AIC_p &= n\log(SSE/n) + 2p, \\ BIC_p &= n\log(SSE/n) + p\log(n) \end{aligned}$$

AIC_p and BIC_p

- To get AICs and BICs in R:
  - Use "step(k=$\cdots$)", which drops one predictor at the time (as in backward selection with p-values) if dropping the predictor lowers the AIC/BIC. k is the multiplier of p in the penalty term, so $k = 2$ for AIC and $k = \log(n)$ for BIC
  - Predictors are only dropped if it doesn't violate the hierarchy (e.g., X_1 is not dropped if there is an interaction term $X_1 \cdot X_2$ in the model)
  - To get the AIC/BIC for models with other subsets of interest (e.g., leaving out a variable and its interaction at once), use "AIC($lm(Y \sim \cdots)$)" and "BIC($lm(Y \sim \cdots)$)"

Prediction sum of squares $PRESS_p$

- Get $Y_i - \hat{Y}_{i(i)}$ based on the model under consideration, leaving out observation Y_i when estimating the β_k's and σ^2
- Prediction sum of squares

$$\begin{aligned} PRESS_p &= \sum_i \left(Y_i - \hat{Y}_{i(i)} \right)^2 \\ &= \sum_i \left(\frac{e_i}{1 - h_{ii}} \right)^2 \end{aligned}$$

  where h_{ii} is the $ii-$th element of the hat matrix $\mathbf{H}$
- Good candidate models: smaller $PRESS_p$

A comparison of 2^{p-1} models-surgical unit example

- possibly incldue X_2, X_3, X_4, X_8, $2^4 = 16$ all possible models

variables	p		
None	1	$X3, X4$	3
$X2$	2	$X3, X8$	3
$X3$	2	$X4, X8$	3
$X4$	2	$X2, X3, X4$	4
$X8$	2	$X2, X3, X8$	4
$X2, X3$	3	$X2, X4, X8$	4
$X2, X4$	3	$X3, X4, X8$	4
$X2, X8$	3	$X2, X3, X4, X8$	5

R implementation

- Preparation

```
Data_example = read.table(“CH09TA01.txt”, header = F)
lnY=Data_example$V10[1:54]
X1=Data_example$V1[1:54]
X2=Data_example$V2[1:54]
X3=Data_example$V3[1:54]
X4=Data_example$V4[1:54]
X5=Data_example$V5[1:54]
X6=Data_example$V6[1:54]
X7=Data_example$V7[1:54]
X8=Data_example$V8[1:54]
X=cbind(X2,X3,X4,X8)
```

R implementation

- R code for C_p, R^2, R_a^2

```
leaps(X,lnY,method=“Cp”)

leaps(X,lnY,method=“r2”)

leaps(X,lnY,method=“adjr2”)
```

- Note: “leaps” function is contained in “leaps” library, so install and load the library [>library(leaps)] to R environment before using “leaps” function

R implementation

- R code for AIC

$$\text{step}(\text{lm}(lnY \sim X2 + X3 + X4 + X8))$$
$$\text{step}(\text{lm}(lnY \sim X2 + X3 + X4))$$
$$\text{step}(\text{lm}(lnY \sim X2 + X3 + X8))$$
$$\text{step}(\text{lm}(lnY \sim X3 + X4 + X8))$$
$$\text{step}(\text{lm}(lnY \sim X2 + X4 + X8))$$
$$\text{step}(\text{lm}(lnY \sim X2 + X3))$$
$$\text{step}(\text{lm}(lnY \sim X4 + X8))$$

C_p and R^2 results

variables	p	Cp	R^2
None	1		
$X2$	2	155.7	0.22
$X3$	2	101.2	0.43
$X4$	2	102.8	0.42
$X8$	2	177.3	0.14
$X2, X3$	3	40.9	0.66
$X2, X4$	3	88.5	0.48
$X2, X8$	3	112.6	0.39
$X3, X4$	3	57.8	0.60
$X3, X8$	3	79.7	0.52
$X4, X8$	3	82.0	0.51
$X2, X3, X4$	4	28.5	0.72
$X2, X3, X8$	4	12.6	0.78
$X2, X4, X8$	4	61.4	0.59
$X3, X4, X8$	4	41.4	0.67
$X2, X3, X4, X8$	5	5.00	0.81

AIC and R_a^2 results

variables	p	AIC	R_a^2
None	1		
$X2$	2	-87	0.21
$X3$	2	-104	0.42
$X4$	2	-103	0.41
$X8$	2	-82	0.12
$X2, X3$	3	-130	0.65
$X2, X4$	3	-107	0.46
$X2, X8$	3	-99	0.37
$X3, X4$	3	-121	0.58
$X3, X8$	3	-110	0.50
$X4, X8$	3	-109	0.49
$X2, X3, X4$	4	-138	0.70
$X2, X3, X8$	4	-151	0.76
$X2, X4, X8$	4	-118	0.57
$X3, X4, X8$	4	-129	0.65
$X2, X3, X4, X8$	5	-158	0.80

‘Best’ subset selection

- Choose the subset model with the largest R_a^2 or
- the smallest AIC_p or
- the smallest Mallows’ C_p
- the smallest PRESS
- different criteria may result in different “best” model

Surgical unit example: Mallow's C_p criterion

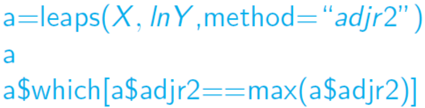
- Consider $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$
- R-code:

```
library(leaps)
X=cbind(X1,X2,X3,X4,X5,X6,X7,X8)
a=leaps(X,lnY,method=“Cp”)
a
a$which[a$Cp==min(a$Cp)]
```

- Selected model:
  $\log(Y) \sim X_1 + X_2 + X_3 + X_6 + X_8$

Surgical unit example: R_a^2 criterion

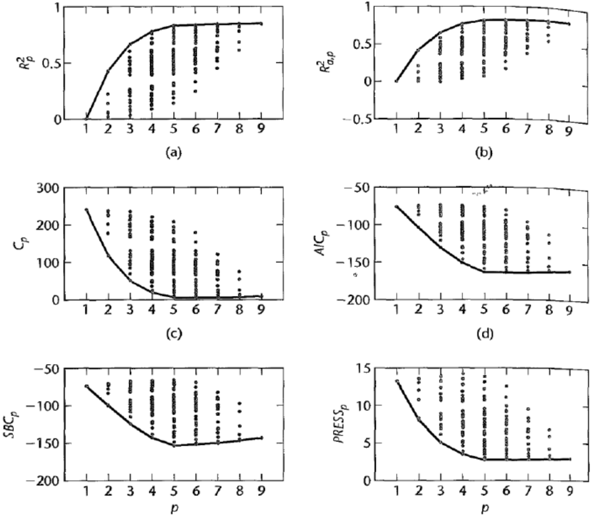
- Consider $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$
- R-code:



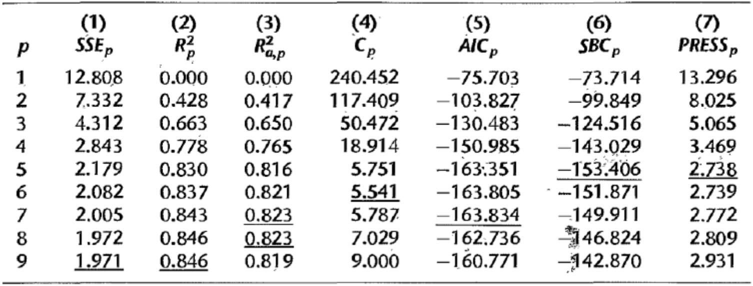
```
a=leaps(X, lnY,method=“adjr2”)
a
a$which[a$adjr2==max(a$adjr2)]
```

- Selected model:
  $\log(Y) \sim X_1 + X_2 + X_3 + X_5 + X_6 + X_8$

Surgical unit example: all criterion



Surgical unit example: all criterion



p	(1) SSE_p	(2) R^2_p	(3) $R^2_{a,p}$	(4) C_p	(5) AIC_p	(6) SBC_p	(7) $PRESS_p$
1	12.808	0.000	0.000	240.452	−75.703	−73.714	13.296
2	7.332	0.428	0.417	117.409	−103.827	−99.849	8.025
3	4.312	0.663	0.650	50.472	−130.483	−124.516	5.065
4	2.843	0.778	0.765	18.914	−150.985	−143.029	3.469
5	2.179	0.830	0.816	5.751	−163.351	−153.406	2.738
6	2.082	0.837	0.821	5.541	−163.805	−151.871	2.739
7	2.005	0.843	0.823	5.787	−163.834	−149.911	2.772
8	1.972	0.846	0.823	7.029	−162.736	−146.824	2.809
9	1.971	0.846	0.819	9.000	−160.771	−142.870	2.931

Surgical unit example: all criterion

- t/F-test is not the unique criteria that can be used in stepwise regression.
- Recall "Life expectancy example", R_a^2 was used in a backward elimination (lecture 6)
- Similar steps (backward elimination, forward selection) can be similarly established by replacing t/F-test criteria with AIC, SBC, C_p criteria. The details are omitted.
- These criteria for stepwise regression, in R, are implemented by function "step()". Refer to the following slides.

Forward selection: *AIC* criterion

- Consider $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$
- R-code:

```
m0=lm(lnY ~ 1)
step(m0,scope=list(lower=~ 1, upper=~
X1+X2+X3+X4+X5+X6+X7+X8),direction=''forward'')
```

- Selected model:
  $\log(Y) \sim X_1 + X_2 + X_3 + X_5 + X_6 + X_8$

Backward elimination: it AIC criterion

- Consider $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$
- R-code:

  m1=lm($lnY \sim X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8$)
  step(m1,scope=list(lower=$\sim$ 1, upper=$\sim| X1 + X2 + X3 +$
  $X4 + X5 + X6 + X7 + X8$),direction="backward")

- Selected model:
  $\log(Y) \sim X_1 + X_2 + X_3 + X_5 + X_6 + X_8$

Forward selection: BIC criterion

- Consider $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$
- R-code:

```
m0=lm(lnY ~ 1)
step(m0,scope=list(lower=~ 1,
upper=~ X1 + X2 + X3 + X4 + X5 + X6 + X7 +
X8),direction=”forward”,k=log(length(lnY)))
```

- Selected model:
  $\log(Y) \sim X_1 + X_2 + X_3 + X_8$

Backward elimination: BIC criterion

- Consider $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$
- R-code:

```
m1=lm(lnY ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8)
step(m1,scope=list(lower=~ 1,
upper=~ X1 + X2 + X3 + X4 + X5 + X6 + X7 +
X8),direction=”backward”,k=log(length(lnY)))
```

- Selected model:
  $\log(Y) \sim X_1 + X_2 + X_3 + X_8$

Forward selection: C_p criterion

- Consider $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$
- R-code:

```
m0=lm(lnY ~ 1)
step(m0,scope=list(lower=~ 1,
upper=~ X1 + X2 + X3 + X4 + X5 + X6 + X7 +
X8),direction=”forward”,scale=(summary(m0)$sigma)∧2)
```

- Selected model:
  $\log(Y) \sim X_1 + X_2 + X_3 + X_6 + X_8$

Backward elimination: C_p criterion

- Consider $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$
- R-code:

```
m1=lm(lnY ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8)
step(m1,scope=list(lower=~ 1,
upper=~ X1 + X2 + X3 + X4 + X5 + X6 + X7 +
X8),direction="backward",scale=(summary(m1)$sigma)∧2)
```

- Selected model:
  $\log(Y) \sim X_1 + X_2 + X_3 + X_6 + X_8$

Summary

- Model selection methods should be used as a guide only
- Look at the difference between the selected models. If they are all very different, there is a lot of uncertainty about which model to use.
- For model validation:
  - Use knowledge of the study area, including signs and magnitude of coefficients
  - Collect new data and check predictive ability
  - Use a holdout sample to check the model and its predictive ability (cross validation)
- Other things to consider:
  - Do they make similar predictions?
  - What is the cost of measuring the predictors?
  - Which has the best diagnostics?