12

데이터분석방법론(1)

Selection of Variables in Regression Equation

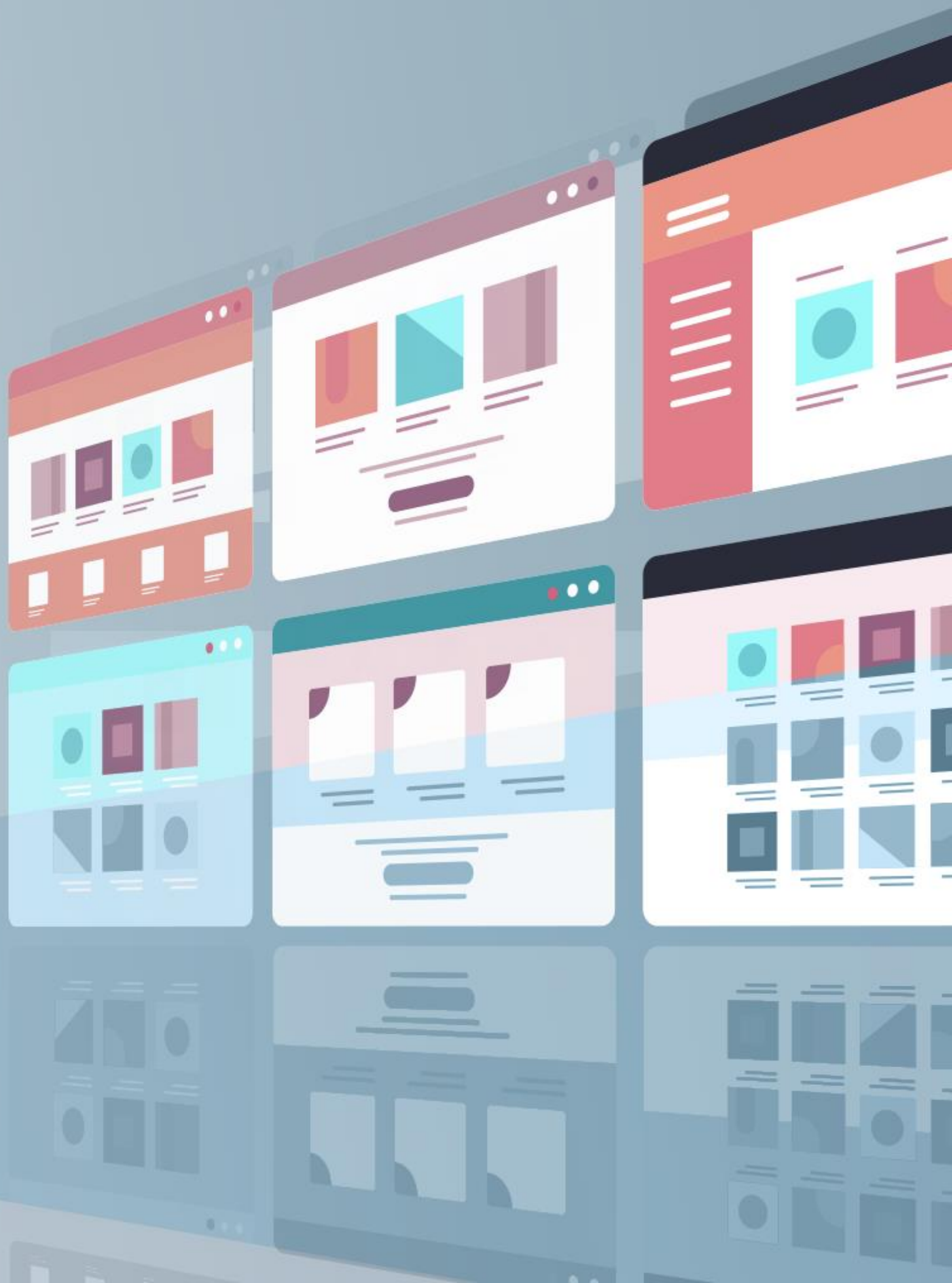
통계데이터과학과 장영재 교수

학습목차

1. Variable Selection
2. Biased Estimation
3. Prior Information
4. Model Selection

01

Variable Selection



1. Model Re-Specification

- Although collinearity is a data problem, not (necessarily) a deficiency of the model, one approach is to re-specify the model.
  - Perhaps several regressors in the model can be conceptualized as alternative indicators of the same construct.
    - Then these measures can be combined or one can be chosen to represent the others.
    - High correlations among the X's indicate high reliability.
  - Alternatively, we can reconsider whether we really need to control for X_2 (for example) in examining the relationship of Y to X_1.
    - Re-specification of this variety is possible only where the original model was poorly thought out, or where the researcher is willing to abandon (some of) the goals of the research.

2. Variable Selection

- A common, but usually misguided, approach to collinearity is variable selection, where some procedure is employed to reduce the regressors in the model to a less highly correlated set.
- Variable Selection Methods:
  - Forward selection methods add explanatory variables to the model one at a time. At each step, the variable that yields the largest increment in R^2 is selected. The procedure stops when the increment is smaller than a preset criterion.
  - Backward elimination methods are similar, except that the procedure starts with the full model and deletes variables one at a time.
  - Stepwise(Forward/backward) methods combine the two approaches.

2. Variable Selection

- Stepwise methods frequently are abused by researchers who interpret the order of entry of X's as an index of their 'importance.'
  - Suppose that there are two highly correlated X's that have nearly identical large correlations with Y; only one X will enter the regression equation.
  - A small modification to the data, or a new sample, could easily reverse the result.

◆ Subset Methods:

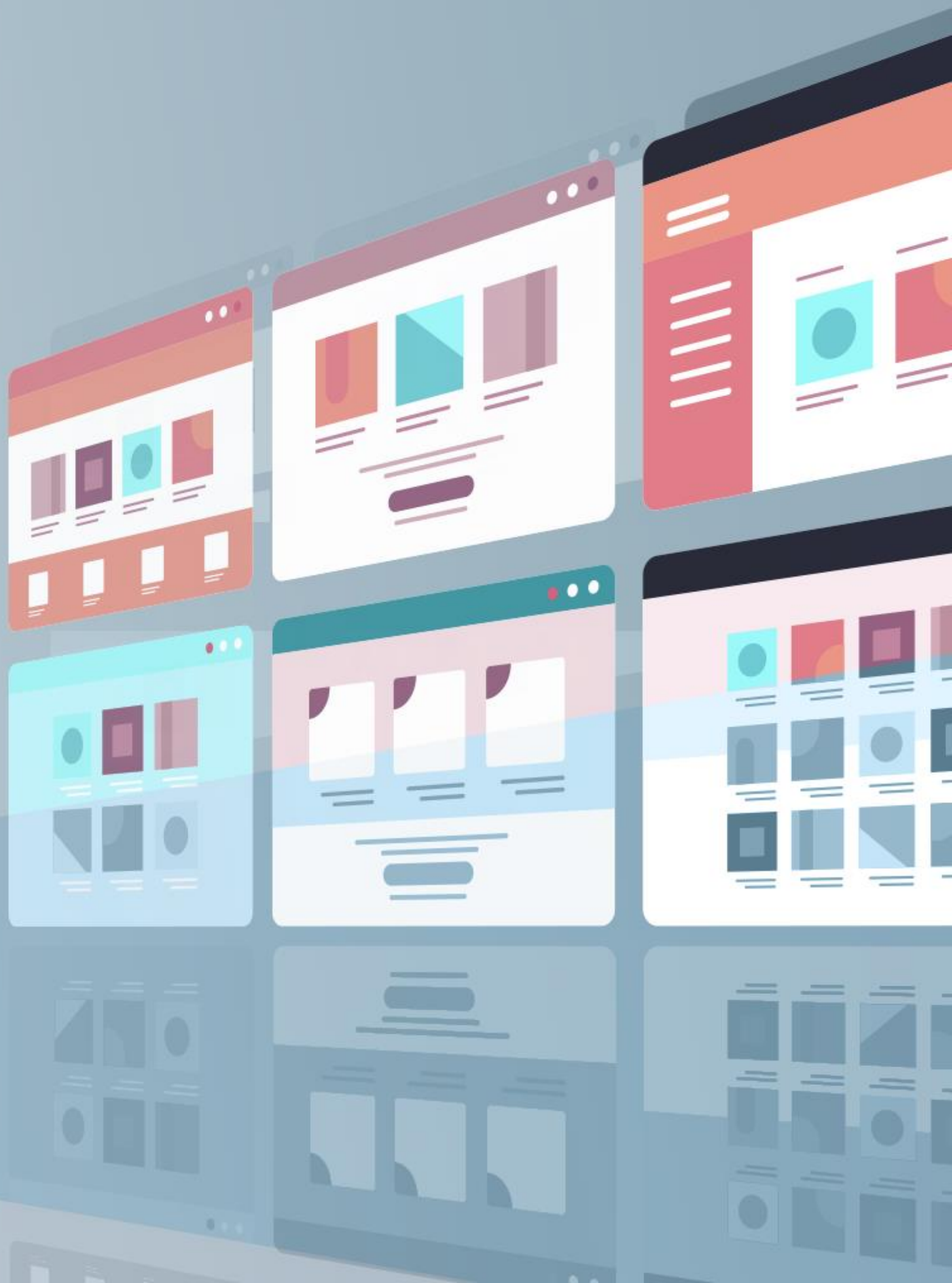
- Stepwise methods can fail to turn up the optimal subset of regressors of a given size.
- It is feasible to examine all subsets of regressors even when k is large.
- Subset techniques also have the advantage of revealing alternative, nearly equivalent models, and thus avoid the appearance of a uniquely 'correct' result.

2. Variable Selection

- Some additional cautions about variable selection:
  - Variable selection results in a re-specified model that usually does not address the original research questions.
    - If the original model is correctly specified, then coefficient estimates following variable selection are biased.
    - High correlations among the X's indicate high reliability.
  - When regressors occur in sets (e.g., of dummy variables), then these sets should be kept together during selection.
    - Likewise, when there are hierarchical relations among regressors, these relations should be respected — for example, don't remove a main effect when an interaction to which it is marginal is included in the model.
  - Coefficient standard errors calculated following explanatory-variable selection overstate the precision of results.
- Variable selection has applications to statistical modeling even when collinearity is not an issue, particularly in prediction.

02

Biased Estimation



1. Bias and Variance Trade-off

- The essential idea here is to trade a small amount of bias in the coefficient estimates for a substantial reduction in coefficient sampling variance, producing a smaller mean-squared error of estimation of the β's.
- By far the most common biased estimation method is ridge regression (due to Hoerl and Kennard).
- Like variable selection, biased estimation is not a panacea for collinearity.
  - Ridge regression involves the arbitrary selection of a 'ridge constant,' which controls the extent to which ridge estimates differ from the least-squares estimates.
  - The larger the ridge constant, the greater the bias and the smaller the variance of the ridge estimator .
  - To pick an optimal ridge constant — or even a good one — generally requires knowledge about the unknown β's.

2. Ridge Regression

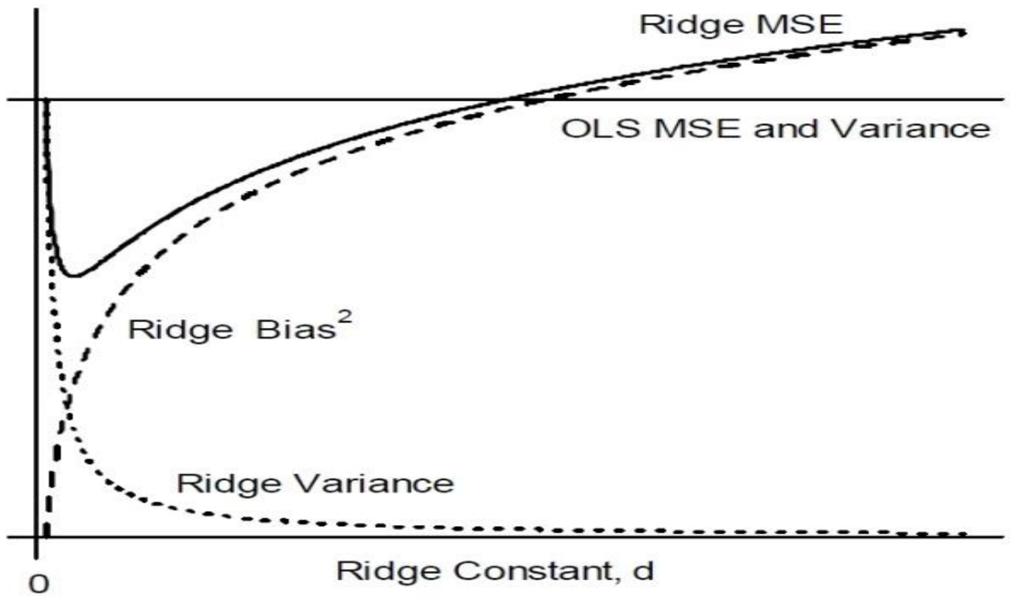
◆ Shrinkage Method

- Ridge regression : Set of estimates according to parameter λ

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

  - In case of $p > n$, handle least squares problem : $X'X \rightarrow X'X + \lambda I$
  - Adjusted least squares estimates (ridge estimates) $\hat{\beta} = (X'X + \lambda I)^{-1}X'Y$
  - A least squares estimate can be shrunk using a ridge estimator to improve the estimate.

2. Ridge Regression

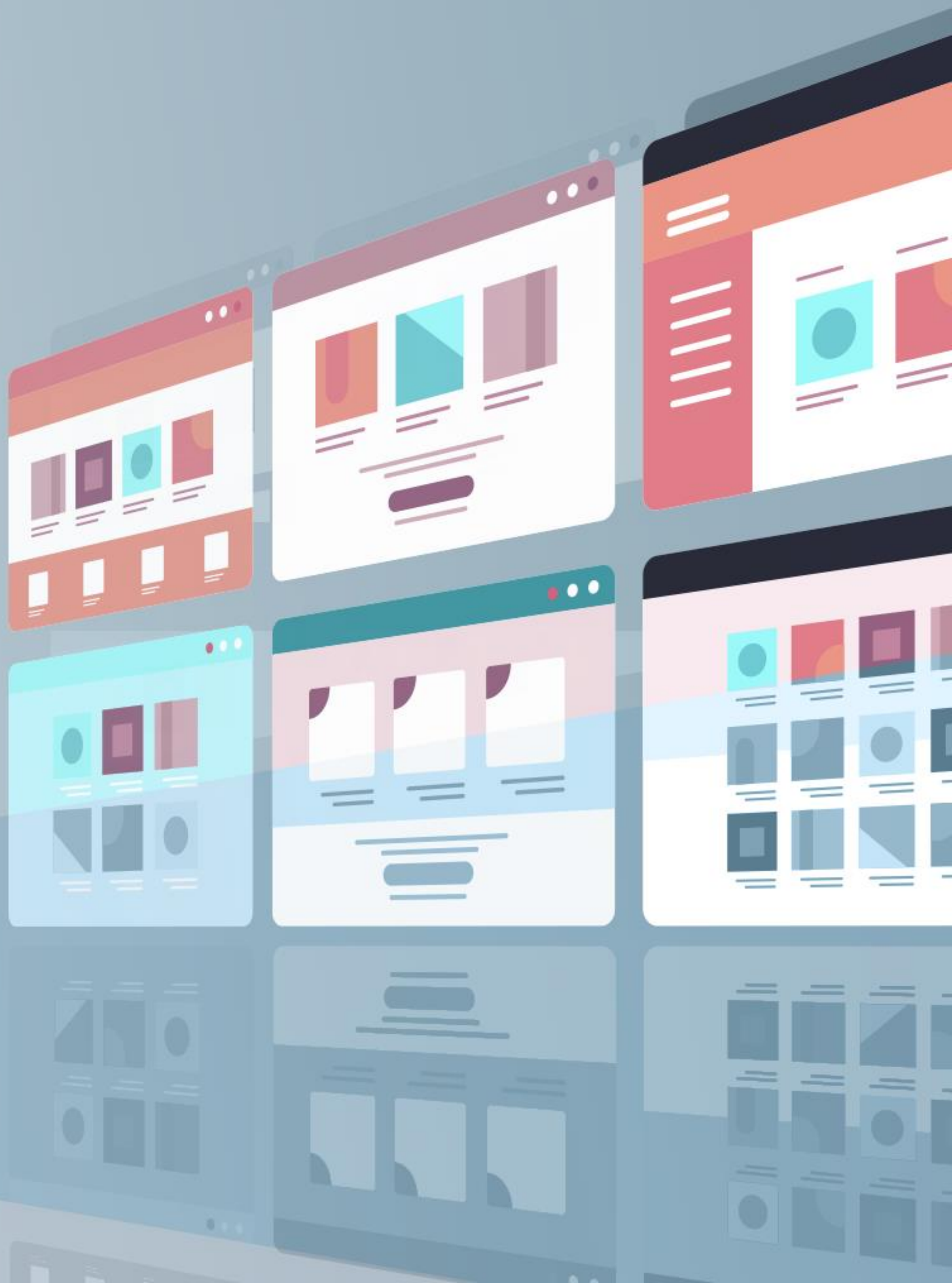


- Figure 3. The squared bias, variance, and MSE of the ridge estimator as a function of the ridge constant, compared to the MSE of the OLS estimator.

$$\text{ridge}MSE=\text{ridge}\mathrm{BIAS}^2+\text{ridgeVariance}$$

03

Prior Information



1. Prior Information About the Regression Coefficients

- A final approach is to introduce additional prior information that reduces the ambiguity produced by collinearity.
- Here is a particularly simple case:
  - We wish to estimate the model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where Y is savings, X_1 is income from wages and salaries, X_2 is dividend income from stocks, and X_3 is interest income.

  - We have trouble estimating β_2 and β_3 because X_2 and X_3 are highly correlated.
  - We have reason to believe that $\beta_2 = \beta_3$, and denote the common quantity β_*.
  - If X_2 and X_3 were not so highly correlated, then we could test this belief as a hypothesis.

1. Prior Information About the Regression Coefficients

- In the current situation, we can fit the model

$$Y = \alpha + \beta_1 X_1 + \beta_*(X_2 + X_3) + \varepsilon$$

incorporating our belief in the equality of X_2 and X_3 in the specification of the model, and thus eliminating the collinearity problem (along with the possibility of testing the belief).

04

Model Selection

1. Some Comparisons

- The several approaches to collinear data have much in common:
  - Model re-specification can involve variable selection, and variable selection re-specifies the model.
  - Variable selection constrains the coefficients of deleted regressors to zero.
  - Variable selection produces biased coefficient estimates.
    - We hope that the trade off of bias against variance is favorable, but because the bias depends on the unknown regression coefficients, we have no assurance
      that this will be the case.
  - Certain types of prior information result in a re-specified model.
  - Biased-estimation methods like ridge regression place prior constraints on the values of the β's.

1. Some Comparisons

- Conclusion: Mechanical model-selection and modification procedures disguise the substantive implications of modeling decisions.
  - These methods generally cannot compensate for weaknesses in the data and are no substitute for judgment and thought.

2. Variable Selection and Model Selection

- We have touched on issues of model selection at several points, often simplifying a model after preliminary statistical hypothesis tests
- Issues of model search extend beyond the selection of explanatory variables or terms to include in a regression model to questions such as the removal of outliers and variable transformations.
- The strategy of basing model selection on hypothesis tests is problematic for a number of reasons:
  - Simultaneous inference.
  - The fallacy of affirming the consequent.
  - The impact of large samples on hypothesis tests.
  - Exaggerated precision following model selection.

3. Model Selection Criteria

- Model selection is conceptually simplest when our goal is prediction— that is, the development of a regression model that will predict new data as accurately as possible.
- Assume that we have n observations on a response variable Y, and a set of m contending statistical models M= $\{M_1, M_1, \ldots, M_m\}$ for Y; model M_j has s_j regression coefficients.
- Adjusted R^2 for p predictor model:

$$R_a^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$$
$$= 1 - \left(\frac{n-1}{n-p}\right) \times (1 - R^2)$$

3. Model Selection Criteria

- Mallow's C_p statistic, which estimates the mean-squared error of prediction under the model:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n$$

where $\hat{\sigma}^2$ is from the model with all predictors and RSS_p indicates the RSS from a model with p parameters.

- It is usual to plot C_p against p.
- We desire model with small p and C_p around or less than p.

3. Model Selection Criteria

- The cross-validation criterion:

$$CV_j \equiv \frac{\sum_{i=1}^{n}(\hat{Y}_{-i}^{(j)} - Y_i)^2}{n}$$

where $\hat{Y}_{-i}^{(j)}$ is the predicted value for the ith observation obtained from a model fit without this observation. We prefer the model with the smallest value of CV_j.

- The generalized cross-validation criterion:

$$\text{AIC}_j \equiv n log_e \hat{\sigma}_{\varepsilon}^{(j)2} + 2s_j$$

where $\hat{\sigma}_{\varepsilon}^{(j)2} = \sum(E_i^{(j)2})\,/n$ is the MLE of the error variance for model j.

- Schwartz's Bayesian information criterion:

$$\text{BIC}_j \equiv n log_e \hat{\sigma}_{\varepsilon}^{(j)2} + s_j\, log_e\, n$$

Note that the BIC penalizes lack of parsimony more than the AIC does.

- We want to minimize AIC or BIC.

4. Model Selection Criteria application

- Regardless of the criterion applied, automatic model-selection methods attend to the predictive adequacy of regression models and are blind to their substantive interpretability.

- Model-selection criteria such as the AIC and BIC are not limited to comparing models selected by automatic methods.
  - One of the currently popular applications of the BIC is to justify the removal of small, but "statistically significant," terms in regression models fit to large samples of data.
  - But researchers should feel free to remove "statistically significant" terms from a statistical model based on the substantive judgment that these terms are too small to be of interest regardless of the value of the BIC.

5. Model Selection Example

- R Example of the variable selection methods:
  - Data : Data were collected from U.S. Bureau of the Census on the 50 states from the 1970s. We will take life expectancy as the response and the remaining variables as predictors.

```
> library(faraway)
> data(state)
> statedata = data.frame(state.x77, row.names=state.abb)
> statedata[c(1:5),]
```

```
> statedata[c(1:5),]
   Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
AL       3615   3624        2.1    69.05   15.1    41.3    20  50708
AK        365   6315        1.5    69.31   11.3    66.7   152 566432
AZ       2212   4530        1.8    70.55    7.8    58.1    15 113417
AR       2110   3378        1.9    70.66   10.1    39.9    65  51945
CA      21198   5114        1.1    71.71   10.3    62.6    20 156361
```

5. Model Selection Example

◆ Implementation of backward elimination methods

```
> g=lm(Life.Exp ~ . , data=statedata)
> summary(g)
```

```
> g=lm(Life.Exp ~ . , data=statedata)
> summary(g)

Call:
lm(formula = Life.Exp ~ ., data = statedata)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.48895 -0.51232 -0.02747  0.57002  1.49447 

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    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared: 0.7362,     Adjusted R-squared: 0.6922 
F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

5. Model Selection Example

1) Remove the predictor with the largest p-value over 0.05:

```
> g = update(g,  . ~ . – Area)
> summary(g)
```

```
> g = update(g, . ~ . - Area)
> summary(g)

Call:
lm(formula = Life.Exp ~ Population + Income + Illiteracy + Murder +
    HS.Grad + Frost, data = statedata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.49047 -0.52533 -0.02546  0.57160  1.50374

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.099e+01  1.387e+00  51.165  < 2e-16 ***
Population   5.188e-05  2.879e-05   1.802   0.0785 .
Income      -2.444e-05  2.343e-04  -0.104   0.9174
Illiteracy   2.846e-02  3.416e-01   0.083   0.9340
Murder      -3.018e-01  4.334e-02  -6.963 1.45e-08 ***
HS.Grad      4.847e-02  2.067e-02   2.345   0.0237 *
Frost       -5.776e-03  2.970e-03  -1.945   0.0584 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7361 on 43 degrees of freedom
Multiple R-squared: 0.7361,     Adjusted R-squared: 0.6993
F-statistic: 19.99 on 6 and 43 DF,  p-value: 5.362e-11
```

5. Model Selection Example

> g = update(g, . ~ . – Illiteracy)

> summary(g)

```
> g = update(g, . ~ . - Illiteracy)
> summary(g)

Call:
lm(formula = Life.Exp ~ Population + Income + Murder + HS.Grad +
    Frost, data = statedata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4892 -0.5122 -0.0329  0.5645  1.5166

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.107e+01  1.029e+00  69.067  < 2e-16 ***
Population   5.115e-05  2.709e-05   1.888   0.0657 .
Income      -2.477e-05  2.316e-04  -0.107   0.9153
Murder      -3.000e-01  3.704e-02  -8.099 2.91e-10 ***
HS.Grad      4.776e-02  1.859e-02   2.569   0.0137 *
Frost       -5.910e-03  2.468e-03  -2.395   0.0210 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7277 on 44 degrees of freedom
Multiple R-squared: 0.7361,     Adjusted R-squared: 0.7061
F-statistic: 24.55 on 5 and 44 DF,  p-value: 1.019e-11
```

5. Model Selection Example

> g = update(g, . ~ . – Income)
> summary(g)

```
> g = update(g, . ~ . - Income)
> summary(g)

Call:
lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
    data = statedata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47095 -0.53464 -0.03701  0.57621  1.50683

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
Population   5.014e-05  2.512e-05   1.996  0.05201 .
Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared: 0.736,      Adjusted R-squared: 0.7126
F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

5. Model Selection Example

> g = update(g, . ~ . – Population)

> summary(g)

```
> g = update(g, . ~ . - Population)
> summary(g)

Call:
lm(formula = Life.Exp ~ Murder + HS.Grad + Frost, data = statedata)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5015 -0.5391  0.1014  0.5921  1.2268 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
Murder      -0.283065   0.036731  -7.706 8.04e-10 ***
HS.Grad      0.049949   0.015201   3.286  0.00195 ** 
Frost       -0.006912   0.002447  -2.824  0.00699 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7427 on 46 degrees of freedom
Multiple R-squared: 0.7127,     Adjusted R-squared: 0.6939 
F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

- Notice that the R^2 for the full model of 0.736 is reduced only slightly to 0.713 in the final model.

5. Model Selection Example

◆ Stepwise selection methods

```
> g=lm(Life.Exp ~ . , data=statedata)
> step(g)
```

```
> g=lm(Life.Exp ~ . , data=statedata)
> step(g)
Start:  AIC=-22.18
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
    Frost + Area

             Df Sum of Sq    RSS     AIC
- Area        1    0.0011 23.298 -24.182
- Income      1    0.0044 23.302 -24.175
- Illiteracy  1    0.0047 23.302 -24.174
<none>                    23.297 -22.185
- Population  1    1.7472 25.044 -20.569
- Frost       1    1.8466 25.144 -20.371
- HS.Grad     1    2.4413 25.738 -19.202
- Murder      1   23.1411 46.438  10.305

Step:  AIC=-24.18
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
    Frost
```

```
Step:  AIC=-28.16
Life.Exp ~ Population + Murder + HS.Grad + Frost

             Df Sum of Sq    RSS     AIC
<none>                    23.308 -28.161
- Population  1     2.064 25.372 -25.920
- Frost       1     3.122 26.430 -23.877
- HS.Grad     1     5.112 28.420 -20.246
- Murder      1    34.816 58.124  15.528

Call:
lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
    data = statedata)

Coefficients:
(Intercept)   Population       Murder      HS.Grad        Frost
  7.103e+01    5.014e-05   -3.001e-01    4.658e-02   -5.943e-03
```

5. Model Selection Example

- Try the C_p and ${R_a}^2$ for the selection of variables

```
> library(leaps)
> b = regsubsets(Life.Exp ~. , data=statedata)
> (rs = summary(b) )
```

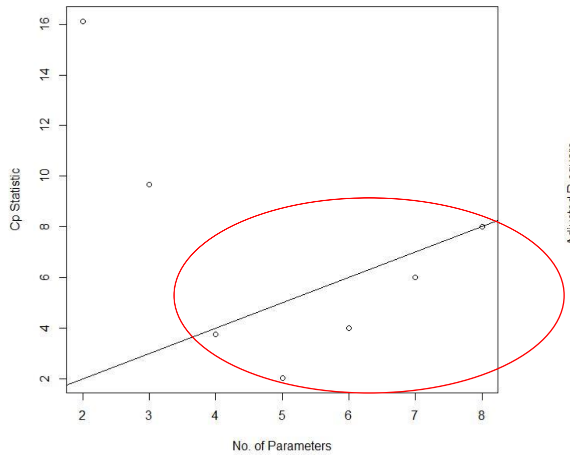
Best models for each size

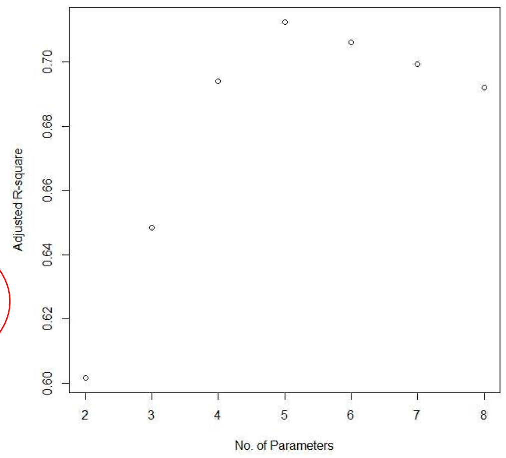
```
> (rs = summary(b) )
Subset selection object
Call: regsubsets.formula(Life.Exp ~ ., data = statedata)
7 Variables  (and intercept)
           Forced in Forced out
Population     FALSE      FALSE
Income         FALSE      FALSE
Illiteracy     FALSE      FALSE
Murder         FALSE      FALSE
HS.Grad        FALSE      FALSE
Frost          FALSE      FALSE
Area           FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         Population Income Illiteracy Murder HS.Grad Frost Area
1  ( 1 ) " "        " "    " "        "*"    " "     " "   " "
2  ( 1 ) " "        " "    " "        "*"    "*"     " "   " "
3  ( 1 ) " "        " "    " "        "*"    "*"     "*"   " "
4  ( 1 ) "*"        " "    " "        "*"    "*"     "*"   " "
5  ( 1 ) "*"        "*"    " "        "*"    "*"     "*"   " "
6  ( 1 ) "*"        "*"    "*"        "*"    "*"     "*"   " "
7  ( 1 ) "*"        "*"    "*"        "*"    "*"     "*"   "*"
> rs$cp
[1] 16.126760  9.669894  3.739878  2.019659  4.008737  6.001959  8.000000
> rs$adjr2
[1] 0.6015893 0.6484991 0.6939230 0.7125690 0.7061129 0.6993268 0.6921823
```

5. Model Selection Example

◆ Cp and Adjusted R-square plot

```
> plot(2:8, rs$cp, xxlab="No. of Parameters", ylab="Cp statistic")
> abline(0,1)
> plot(2:8, rs$adjr2, xxlab="No. of Parameters", ylab="Adjusted R-square")
```





These models are on and below the Cp = p, indicating good fits.
Adjusted R-square is largest for p=5.
So, we choose "Population, Frost, HS.grad, and Murder".

5. Model Selection Example

◆ Variable selection methods are sensitive to outliers and influential points.

- Check for high leverage points.

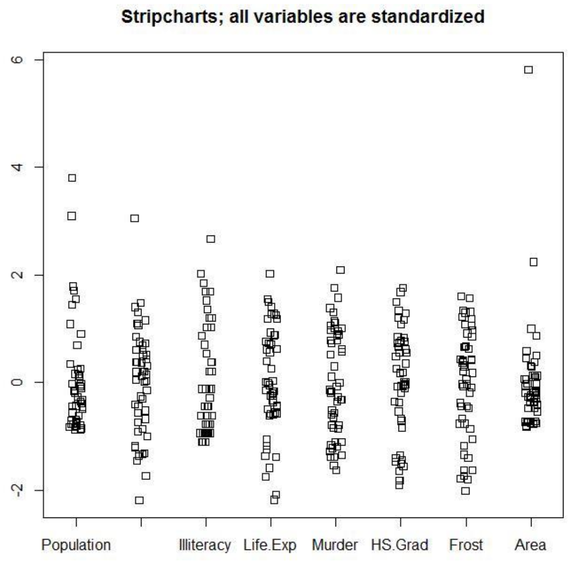
```
> h = lm.influence(g)$hat
> names(h) = state.abb
> round(rev(sort(h)),3)
   AK    CA    HI    NV    NM    TX    NY    WA    OR    ND    LA    CT
0.810 0.409 0.379 0.365 0.325 0.284 0.257 0.223 0.222 0.220 0.195 0.194
> b = regsubsets(Life.Exp ~ . , data=statedata,
      subset=(state.abb!="AK"))
> (rs= summary(b))
> rs$adjr2
```

```
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         Population Income Illiteracy Murder HS.Grad Frost Area
1  ( 1 ) " "        " "    " "        "*"    " "     " "   " "
2  ( 1 ) " "        " "    " "        "*"    "*"     " "   " "
3  ( 1 ) " "        " "    " "        "*"    "*"     "*"   " "
4  ( 1 ) "*"        " "    " "        "*"    "*"     "*"   " "
5  ( 1 ) "*"        " "    " "        "*"    "*"     "*"   "*"
6  ( 1 ) "*"        "*"    " "        "*"    "*"     "*"   "*"
7  ( 1 ) "*"        "*"    "*"        "*"    "*"     "*"   "*"
> rs$adjr2
[1] 0.5923260 0.6603281 0.6948855 0.7086703 0.7104405 0.7073027 0.7008899
```

5. Model Selection Example

◆ Transforming the predictors

```
> stripchart(data.frame(scale(statedata)), vertical=TRUE, method="jitter")
> title("StripCharts; all variables are standardized")
> b = regsubsets(Life.Exp ~ log(Population)+Income+Illiteracy+Murder+
      , +HS.Grad+Frost+log(Area), data=state.data)
> (rs= summary(b))
> rs$adjr2
```



- Jittering adds a small amount of noise to avoid the overprint each other.
- Population and Area are skewed, so try transformation.

5. Model Selection Example

◆ Transforming the predictors

```
> rs$which[which.max(rs$adjr2),]
> b2 = lm(Life.Exp ~ log(Population)+Murder+ HS.Grad+Frost+log(Area),
  data=state.data)
> summary(b2)
```

```
> rs$which[which.max(rs$adjr2),]
    (Intercept) log(Population)          Income      Illiteracy
           TRUE            TRUE           FALSE           FALSE
         Murder         HS.Grad           Frost       log(Area)
           TRUE            TRUE            TRUE           FALSE
> b2 = lm(Life.Exp ~ log(Population)+Murder+HS.Grad+Frost,statedata)
> summary(b2)

Call:
lm(formula = Life.Exp ~ log(Population) + Murder + HS.Grad +
    Frost, data = statedata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.41760 -0.43880  0.02539  0.52066  1.63048

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     68.720810   1.416828  48.503  < 2e-16 ***
log(Population)  0.246836   0.112539   2.193 0.033491 *
Murder          -0.290016   0.035440  -8.183 1.87e-10 ***
HS.Grad          0.054550   0.014758   3.696 0.000591 ***
Frost           -0.005174   0.002482  -2.085 0.042779 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7137 on 45 degrees of freedom
Multiple R-squared: 0.7404,     Adjusted R-squared: 0.7173
F-statistic: 32.09 on 4 and 45 DF,  p-value: 1.17e-12
```

Best Model : log(Population), Murder, HS.Grad, frost

The adjusted R-square of 74% is the highest models we have seen so far.

6. Model Validation

- In model validation, part of the data (called the "training" or "exploratory" sub-sample) is used to specify a statistical model, which is then evaluated using the other part of the data (the "validation" or "confirmatory" sub-sample).
- The process of data exploration, model fitting, model criticism, and model re-specification is typically iterative, requiring several failed attempts before an adequate description of the data is achieved.
  - In the process, variables may be dropped from the model; terms such as interactions may be incorporated or deleted; variables may be transformed; and unusual data may be corrected, removed, or otherwise accommodated.
  - The risk of iterative modeling is that we will capitalize on chance—over-fitting the data and overstating the strength of our results.

6. Model Validation

- An ideal solution would be to collect new data with which to validate a model, but this solution is often impractical.
- Model validation simulates the collection of new data by randomly dividing the data that we have in hand into two parts—the first part to be used for exploration and model formulation, the second for checking the model, formal estimation, and testing.
- Barnard: "The simple idea of splitting a sample into two and then developing the hypothesis on the basis of one part and testing it on the remainder may perhaps be said to be one of the most seriously neglected ideas in statistics, if we measure the degree of neglect by the ratio of the number of cases where a method could give help to the number where it is actually used."

다음시간 안내

13

Logistic Regression