Multiple Linear Regression and Model Selection

Biostat 705

Hussein Al-Khalidi, PhD

Department of Biostatistics and Bioinformatics,
Duke University

January 21, 2024

Multiple linear regression model

- Multiple linear regression model:

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Here we are interested in linear associations between the response Y and predictors (independent variables) $X_1, X_2, \ldots, X_p$.
- Interpretation of the intercept β_0 is the expected value of Y when $X_1, X_2, \ldots, X_p$ are all equal zero
- We interpret the slopes (coefficient of predictors) β_j $(j = 1, \ldots, p)$ as the *average* effect on Y of a 1 unit change (increase or decrease) in X_j, *holding all other predictors constant*.

DukeHealth

Estimation of Multiple linear regression model

- Estimated multiple linear regression model:

  $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

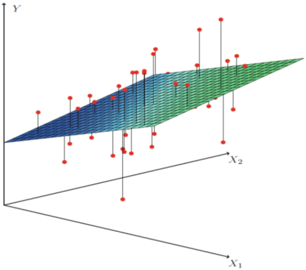
- We estimate regression parameters $\beta_0, \beta_1, \ldots, \beta_p$ using least-squares method, similar approach used in simple linear regression, that is by minimizing the sum squared errors:

$$\begin{aligned} \mathrm{SSE} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

  This is done using statistical software, such R or SAS. The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize SSE are multiple least-squares regression coefficient estimates.

DukeHealth

Example from ISLR book



A few important questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. Do all predictors help to explain Y, or only a subset of the predictors are useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

"Essentially, all model are wrong, but some are useful"
George Box

Multiple linear regression in a matrix form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

The regression model above, can be written in a matrix form as:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\underbrace{Y}_{(n\times 1)} = \underbrace{X}_{[n\times(p+1)]} \overbrace{\boldsymbol{\beta}}^{[(p+1)\times 1]} + \underbrace{\epsilon}_{(n\times 1)}$$

DukeHealth

Multiple linear regression in a matrix form

$$Y = X\beta + \epsilon$$

Y is a vector of observed responses which has a distribution (usually normal) with mean $X\beta$ and Variance $\sigma_\epsilon^2 \mathrm{I_n}$, X is called the design-matrix and measured without error (ie, fixed). Note, normality assumption is not required to estimate the model parameters $\beta' s$.
Thus, the estimated β's are given as:

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

Hence, variance of $\hat{\beta}$ given as: $\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathrm{X^{'}X})^{-1}$ and estimated varaince as: $\widehat{\mathrm{Var}(\hat{\beta})} = S_{y.x}^2(X^{'}X)^{-1} = \mathrm{MSE}(\mathrm{X^{'}X})^{-1}$

DukeHealth

Assumptions

Assumptions:

- Linearity, $E(\epsilon_i) = 0$, which implies that $EY_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$
- Homoscedasticity (constant variance), $\mathrm{var}(\epsilon_i) = \sigma_\epsilon^2$, which implies that $\mathrm{var}(Y_i | X_{i1}, \ldots, X_{ip}) = \sigma_\epsilon^2$
- Normality, $\epsilon_i \sim \mathsf{N}(0, \sigma_\epsilon^2)$, which implies that $Y_i | X_{i1}, \ldots, X_{ip} \sim \mathsf{N}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ik}, \sigma_\epsilon^2)$
- Independence, $\epsilon_i \epsilon_j$ are indpendent for $i \neq j$, ie Y values are independent from each other.

Partial F statistic

Testing overall (global) significance in multiple linear regression model, that is

- Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$
- Hypothesis: $\mathrm{H_0} : \beta_1 = \beta_2 = \ldots = \beta_\mathrm{p} = 0$
  $\mathrm{H_a}$: One or more of the β_j are nonzero
- Test statistic: $F = \frac{SS_{reg}/p}{SSE/(n-p-1)} \sim F_{(p,n-p-1)}$
  where $p+1$ is number of parameters in the model (including the intercept β_0).
- or equivalently, $F = \frac{R^2/p}{(1-R^2)/(n-p-1)} \sim F_{(p,n-p-1)}$
  where, $R^2 = \frac{SS_{reg}}{SST} = 1 - \frac{SSE}{SST}$

DukeHealth

```
weight height age
64 57  8
71 59 10
53 49  6
67 62 11
55 51  8
58 50  7
77 55 10
57 48  9
56 42 10
51 42  6
76 61 12
68 57  9
```

How to express the above example in a matrix form

$$\text{weight} = \beta_0 + \beta_1\,\text{hight} + \beta_2\,\text{age} + \epsilon$$

The regression model above, can be written in a matrix form as:

$$\begin{bmatrix} 64 \\ 71 \\ \vdots \\ 68 \end{bmatrix} = \begin{bmatrix} 1 & 57 & 8 \\ 1 & 59 & 10 \\ \vdots & \vdots & \\ 1 & 57 & 9 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{12} \end{bmatrix}$$

$$\underbrace{Y}_{(12\times 1)} = \underbrace{X}_{[12\times 3]} \overbrace{\boldsymbol{\beta}}^{[3\times 1]} + \underbrace{\epsilon}_{(12\times 1)}$$

```
library(MASS)
wha <- read.table("C:\\Users\\ha27\\Desktop\\wh.txt", header=T)
attach(wha)
m <- lm(weight ~ height + age)
#get summary of multiple resgression ANOVA
summary(m)
#################################################################################
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5530    10.9448   0.599   0.5641
height        0.7220     0.2608   2.768   0.0218 *
age           2.0501     0.9372   2.187   0.0565 .
---
Residual standard error: 4.66 on 9 degrees of freedom
Multiple R-squared:   0.78,     Adjusted R-squared:  0.7311
F-statistic: 15.95 on 2 and 9 DF,  p-value: 0.001099
##################################################################################
m1=lm(weight ~ height,data=wha)
summary(m1)
##################################################################################
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.1898    12.8487   0.482  0.64035
height        1.0722     0.2417   4.436  0.00126 **
---
Residual standard error: 5.471 on 10 degrees of freedom
Multiple R-squared:  0.663,     Adjusted R-squared:  0.6293
F-statistic: 19.67 on 1 and 10 DF,  p-value: 0.001263
```

DukeHealth

```
m2=lm(weight ~ age,data=wha)
summary(m2)
################################################################################
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.5714     8.6137   3.549  0.00528 **
age           3.6429     0.9551   3.814  0.00341 **
---
Residual standard error: 6.015 on 10 degrees of freedom
Multiple R-squared:  0.5926,    Adjusted R-squared:  0.5519
F-statistic: 14.55 on 1 and 10 DF,  p-value: 0.003407
##################################################################################
###No regression, ie model without predictors
m0=lm(weight ~ 1,data=wha)
summary(m0)
################################################################################
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.750      2.594   24.19 6.89e-11 ***
---
Residual standard error: 8.986 on 11 degrees of freedom
##################################################################################
```

```
###Regression model with interaction,
m3=lm(weight ~ height*age,data=wha)
summary(m3)
####################################################################################
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.06471   59.85950   0.018    0.986
height       0.83032    1.19112   0.697    0.505
age          2.63866    6.37515   0.414    0.690
height:age  -0.01146    0.12259  -0.093    0.928

Residual standard error: 4.94 on 8 degrees of freedom
Multiple R-squared:  0.7802,    Adjusted R-squared:  0.6978
F-statistic: 9.467 on 3 and 8 DF,  p-value: 0.005211
####################################################################################
Note: in situation where the interaction is significant, but the main effects are not significant,
it's a good practice to keep the main effects in the model, despite they're not significant, this
often called hierarchy principle.
```

```
data one;
input weight height age; int=height*age;datalines;
64 57 8
71 59 10
53 49 6
67 62 11
55 51 8
58 50 7
77 55 10
57 48 9
56 42 10
51 42 6
76 61 12
68 57 9
;
run;
title1 "*** Regression model without interaction***";
proc reg data=one;
model weight=height age;
run;
title1 "*** Regression model with interaction***";
proc reg data=one;
model weight=height age int;
run;
```

```
*** Regression model without interaction***
The REG Procedure
Model: MODEL1
Dependent Variable: weight

Number of Observations Read          12
Number of Observations Used          12

                          Analysis of Variance
                                   Sum of            Mean
Source                    DF      Squares          Square    F Value     Pr > F
Model                      2    692.82261       346.41130      15.95     0.0011
Error                      9    195.42739        21.71415
Corrected Total           11    888.25000

Root MSE              4.65984    R-Square     0.7800
Dependent Mean       62.75000    Adj R-Sq     0.7311
Coeff Var             7.42605
                     Parameter Estimates
                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

Intercept     1        6.55305       10.94483       0.60      0.5641
height        1        0.72204        0.26081       2.77      0.0218
age           1        2.05013        0.93723       2.19      0.0565
```

```
*** Regression model with interaction***

The REG Procedure
Model: MODEL1
Dependent Variable: weight

Number of Observations Read          12
Number of Observations Used          12

                          Analysis of Variance
                                   Sum of           Mean
Source                   DF       Squares         Square    F Value    Pr > F
Model                     3     693.03575      231.01192       9.47    0.0052
Error                     8     195.21425       24.40178
Corrected Total          11     888.25000

Root MSE               4.93982    R-Square     0.7802
Dependent Mean        62.75000    Adj R-Sq     0.6978
Coeff Var              7.87222

                     Parameter Estimates
                      Parameter      Standard
Variable     DF        Estimate         Error    t Value    Pr > |t|
Intercept     1         1.06471      59.85950       0.02      0.9862
height        1         0.83032       1.19112       0.70      0.5055
age           1         2.63866       6.37515       0.41      0.6898
int           1        -0.01146       0.12259      -0.09      0.9278
```

```
data one;
input weight height age; datalines;
64 57 8
71 59 10
53 49 6
67 62 11
55 51 8
58 50 7
77 55 10
57 48 9
56 42 10
51 42 6
76 61 12
68 57 9
;
run;
title1 "*** ANOVA model without interaction***";
proc glm data=one;
model weight=height age /ss3 ss1;
run;
title1 "*** ANOVA model with interaction***";
proc glm data=one;
model weight=height|age /ss3 ss1; ***this is same as weight=height age height*age;
run;
```

```
The GLM Procedure *** ANOVA model without interaction***

Dependent Variable: weight

                                          Sum of
Source                      DF           Squares     Mean Square    F Value    Pr > F
Model                        2       692.8226065     346.4113033      15.95    0.0011
Error                        9       195.4273935      21.7141548
Corrected Total             11       888.2500000

R-Square     Coeff Var      Root MSE    weight Mean
0.779986      7.426048      4.659845       62.75000

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
height                       1     588.9225232     588.9225232      27.12    0.0006
age                          1     103.9000834     103.9000834       4.78    0.0565

Source                      DF     Type III SS     Mean Square    F Value    Pr > F
height                       1     166.4297494     166.4297494       7.66    0.0218
age                          1     103.9000834     103.9000834       4.78    0.0565

                                    Standard
Parameter         Estimate             Error    t Value    Pr > |t|
Intercept      6.553048251       10.94482708       0.60      0.5641
height         0.722037958        0.26080506       2.77      0.0218
age            2.050126352        0.93722561       2.19      0.0565
```

```
The GLM Procedure *** ANOVA model with interaction***

Dependent Variable: weight

                                        Sum of
Source                    DF           Squares     Mean Square    F Value    Pr > F
Model                      3       693.0357479     231.0119160       9.47    0.0052
Error                      8       195.2142521      24.4017815
Corrected Total           11       888.2500000

R-Square     Coeff Var      Root MSE    weight Mean
0.780226      7.872217      4.939816       62.75000

Source                    DF         Type I SS     Mean Square    F Value    Pr > F
height                     1       588.9225232     588.9225232      24.13    0.0012
age                        1       103.9000834     103.9000834       4.26    0.0730
height*age                 1         0.2131413       0.2131413       0.01    0.9278

Source                    DF       Type III SS     Mean Square    F Value    Pr > F
height                     1       11.85765063     11.85765063       0.49    0.5055
age                        1        4.18031226      4.18031226       0.17    0.6898
height*age                 1        0.21314131      0.21314131       0.01    0.9278

                                    Standard
Parameter         Estimate             Error    t Value    Pr > |t|
Intercept      1.064709014       59.85950265       0.02      0.9862
height         0.830319291        1.19112289       0.70      0.5055
age            2.638664346        6.37515215       0.41      0.6898
height*age    -0.011457482        0.12259313      -0.09      0.9278
```

Multiple linear regression model with interaction

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 + \epsilon \end{aligned}$$

In our example:

$$\begin{aligned} \text{weight} &= \beta_0 + \beta_1 \text{height} + \beta_2 \text{age} + \beta_3 \text{height} * \text{age} + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \text{age}) \text{height} + \beta_2 \text{age} + \epsilon \\ &= \beta_0 + \beta_1 \text{height} + (\beta_2 + \beta_3 \text{height}) \text{age} + \epsilon \end{aligned}$$

$$\begin{aligned} \widehat{\text{weight}} &= 1.06 + 0.83 * \text{height} + 2.64 * \text{age} - 0.01 * \text{height} * \text{age} \\ &= 1.06 + (0.83 - 0.01 * \text{age}) * \text{height} + 2.64 * \text{age} \\ &= 1.06 + 0.83 * \text{height} + (2.64 - 0.01 * \text{height}) * \text{age} \end{aligned}$$

DukeHealth

Interpretation of interaction in linear regression model

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$

When the interaction term $(X_1 X_2)$ in the model above is significant, that means the relationship between the two predictors (X_1, X_2) and the response variable (Y) is not additive, but rather multiplicative. In other words, the effect of one predictor on Y depends on the level of the other predictor.

A significant interaction term does not necessarily mean that the main effects of (X_1, X_2) are not important, it only means that the effect of one predictor on the response variable is conditional on the level of the other predictor.

For example: Suppose a clinician wants to study the relationship between cancer treatment (X_1) and survival rate (Y), and whether the relationship between cancer treatment and survival rate is different for patients who are at different stages of cancer. Thus, the model should include an interaction term between cancer treatment (X_1) and cancer stage (X_2) in the regression model. This allows the clinician to examine whether the relationship between cancer treatment and survival rate is different for patients who are at different stages of cancer, and to determine whether different cancer treatments may have a different effect on patients with different stages of cancer.

In general, when interpreting a significant interaction term, it's important to examine the simple slopes of the predictors at specific levels of the other predictor. For example, you can look at the effect of one predictor on the response variable when the other predictor is at its lowest and highest values. If the slopes are different, it indicates that the effect of one predictor on the response variable changes depending on the level of the other predictor.

DukeHealth

Sequential and Partial SS

Definition of type I sum of sequares (SS) and type III sum of sequares in SAS output. Suppose we have a model with 3 independent variables (X_1, X_2, X_3).

Variables	Type I SS (sequential)	Type III SS (partial)
X_1	SS(X_1)	SS$(X_1 \mid X_2, X_3)$
X_2	SS$(X_2 \mid X_1)$	SS$(X_2 \mid X_1, X_3)$
X_3	SS$(X_3 \mid X_1, X_2)$	SS$(X_3 \mid X_1, X_2)$

- In Type I SS (sequential) order is important. Type I SS are statistically independent of each other, ie each associated with 1 df and they do add up to the SS regression, for example $\mathrm{SS(X_1) + SS(X_2/X_1) + SS(X_3/X_1, X_2) = SS_{reg}(X_1, X_2, X_3)}$. This type of SS is useful in polynomial regression modeling.
- There is also Type II SS (partial) which is similar to Type III and both produce same SS when the design is balanced. However, for unbalanced design we would use Type III SS. Unlike Type I SS, both Type II and Type III they don't add up to the $\mathrm{SS_{reg}}$. Also, both Type II and Type III SS are invariant to the ordering, ie, order is not important.

DukeHealth

Partial F test

In general, suppose we have the following model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \beta_{k+1} X_{k+1} + \ldots + \beta_p X_p + \epsilon$$

Thus, testing $\mathrm{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_{\mathrm{k}} = 0$
The test statistic (partial F) would be:

$$\begin{aligned} & F(X_1, X_2, \ldots X_k | X_{k+1}, X_{k+2}, \ldots, X_p) \\ &= \frac{[SS_{reg}(full) - SS_{reg}(reduced)]/k}{MSE(full)} \\ &= \frac{[SSE(reduced) - SSE(full)]/k}{MSE(full)} \end{aligned}$$

Note that the reduced model is nested within the full model meaning that all the terms remaining in the reduced model were in the full model as well.

DukeHealth

As an example, suppose

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ (Full-model)

Under $\mathrm{H}_0 : \beta_2 = 0 \rightarrow \mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \epsilon$ (Reduced-model)

- Test statistic:

$$F = \frac{[SS_{reg(full)} - SS_{reg(reduced)}]/1}{SSE_{(full)}/(n-3)}$$
$$= \frac{(SSE_{(reduced)} - SSE_{(full)})/1}{SSE_{(full)}/(n-3)} \sim F_{(1,n-3)}.$$

- numerator df $=\#$ of parameters in the full-model minus $\#$ of parameters in the reduced-model (or simply $\#$ of parameters tested in H_0)
- denominator df $= \#$ of observations (sample size n) minus $\#$ of parameters in the full-model.

DukeHealth

Partial F: in-class workout problem

Using the QUET example (quet.txt), in which systolic blood pressure (sbp) is the response and QUET (quetelet index, quet=100*(weight/hight^2), age and smoking history (smk).

- model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

  where Y = sbp, X_1 = quet,

  X_2 = age, X_3 = smk (1=smoker, 0=non-smoker)

- $SS(X_1)$ =sum squares regression explained using only X_1 predict Y.
- $SS(X_2|X_1)$ =sum squares regression explained using X_2 given X_1 in the model.
- $SS(X_3|X_1, X_2)$ =sum squares regression explained using X_3 given X_1 and X_2 in the model.

Example:

```
data one;
label sbp = 'Systolic blood pressure'
      quet= 'Quetelet index'
  age = 'Age'
  smk = 'Smoking histroy';
input sbp quet age smk @@;datalines;
135 2.876 45 0 122 3.251 41 0 130 3.100 49 0 148 3.768 52 0 146 2.979 54 1
129 2.790 47 1 162 3.668 60 1 160 3.612 48 1 144 2.368 44 1 180 4.637 64 1
166 3.877 59 1 138 4.032 51 1 152 4.116 64 0 138 3.673 56 0 140 3.562 54 1
134 2.998 50 1 145 3.360 49 1 142 3.024 46 1 135 3.171 57 0 142 3.401 56 0
150 3.628 56 1 144 3.751 58 0 137 3.296 53 0 132 3.210 50 0 149 3.301 54 1
132 3.017 48 1 120 2.789 43 0 126 2.956 43 1 161 3.800 63 0 170 4.132 63 1
152 3.962 62 0 164 4.010 65 0
;
run;
title1 "*** Full-model ***";
proc reg data=one;
  model sbp=quet age smk;
run;
title1 "*** ANOVA model ***";
proc glm data=one;
   class smk;
   model sbp=quet age smk/ss3 ss1;
run;
```

SAS outputs: PROC REG

```
The REG Procedure
                         Model: MODEL1
        Dependent Variable: sbp Systolic blood pressure

            Number of Observations Read          32
            Number of Observations Used          32


                      Analysis of Variance

                                  Sum of           Mean
Source                  DF       Squares         Square    F Value

Model                    3    4889.82570     1629.94190      29.71
Error                   28    1536.14305       54.86225
Corrected Total         31    6425.96875

                      Analysis of Variance

                   Source                Pr > F

                   Model                 <.0001

      Root MSE              7.40691    R-Square     0.7609
      Dependent Mean      144.53125    Adj R-Sq     0.7353
      Coeff Var             5.12478
```

DukeHealth

SAS outputs: PROC REG

```
The REG Procedure
                        Model: MODEL1
       Dependent Variable: sbp Systolic blood pressure

                      Parameter Estimates
                                       Parameter     Standard
Variable   Label                 DF     Estimate        Error

Intercept  Intercept              1     45.10319     10.76488
quet       Quetelet index         1      8.59245      4.49868
age        Age                    1      1.21271      0.32382
smk        Smoking histroy        1      9.94557      2.65606

                      Parameter Estimates
   Variable   Label                 DF  t Value   Pr > |t|

   Intercept  Intercept              1     4.19     0.0003
   quet       Quetelet index         1     1.91     0.0664
   age        Age                    1     3.75     0.0008
   smk        Smoking histroy        1     3.74     0.0008
```

SAS outputs: PROC GLM

```
The GLM Procedure
      Dependent Variable: sbp   Systolic blood pressure
                                        Sum of
Source                     DF          Squares     Mean Square
Model                       3      4889.825697     1629.941899
Error                      28      1536.143053       54.862252
Corrected Total            31      6425.968750

          Source                F Value    Pr > F
          Model                   29.71    <.0001

      R-Square     Coeff Var      Root MSE      sbp Mean
      0.760948      5.124778      7.406906      144.5313

Source                     DF        Type I SS     Mean Square
quet                        1      3537.945739     3537.945739
age                         1       582.646506      582.646506
smk                         1       769.233452      769.233452

          Source                F Value    Pr > F
          quet                    64.49    <.0001
          age                     10.62    0.0029
          smk                     14.02    0.0008

Source                     DF      Type III SS     Mean Square
quet                        1      200.1414685     200.1414685
age                         1      769.4592039     769.4592039
smk                         1      769.2334521     769.2334521

          Source                F Value    Pr > F
          quet                     3.65    0.0664
          age                     14.03    0.0008
          smk                     14.02    0.0008
```

ANOVA Table

- Global null hypothesis: $\mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$
- ANOVA Table

Source	df	SS	MS	F
Regression	3	4889.83	1629.9=$\frac{4889.83}{3}$	$\frac{1629.9}{54.86} \sim F_{3,28}$
X_1	1	3537.95		
$X_2 \| X_1$	1	582.65		
$X_3 \| X_1, X_2$	1	769.23		
Error	28	1536.14	54.86=$\frac{1536.14}{28}$	
Total	31	6425.97		

- Calculated $F = 29.71 \sim F_{3,28}$. $F_{.95,3,28} = 2.95$, thus, $F = 29.71 > 2.95 \Rightarrow$ reject the global null hypothesis H_0

DukeHealth

ANOVA Table

In general,

$$SS(X^*|X_1, X_2, \ldots, X_p) = SS_{reg}(X_1, X_2, \ldots, X_p, X^*) - SS_{reg}(X_1, X_2, \ldots, X_p)$$
$$= SSE(X_1, X_2, \ldots, X_p) - SSE(X_1, X_2, \ldots, X_p, X^*).$$

Definition

$$F(X^*|X_1, X_2, \ldots, X_p) = \frac{\text{extra sum of squares adding } X^*\text{, given} X_1,\ldots,X_p}{\text{mean square residual for all variables } (X_1,X_2,\ldots,X_p,X^*) \text{ in the model}}.$$

ANOVA Table

Definition

$$F(X^*|X_1, X_2, \ldots, X_p) = \frac{SS_{reg}(X^*|X_1,\ldots,X_p)}{MSE(X_1,X_2,\ldots,X_p,X^*)}$$

Definition

$$F(X^*|X_1, X_2, \ldots, X_p)$$
$$= \frac{(SS_{reg}(X_1,X_2,\ldots,X_p,X^*) - SS_{reg}(X_1,X_2,\ldots,X_p))/1}{MSE(X_1,X_2,\ldots,X_p,X^*)}$$

Definition

$$F(X^*|X_1, X_2, \ldots, X_p) = \frac{(SSE(X_1,X_2,\ldots,X_p) - SSE(X_1,X_2,\ldots,X_p,X^*))/1}{MSE(X_1,X_2,\ldots,X_p,X^*)}$$

DukeHealth

Sum Squares Decomposition

Some notes on decomposition of SSR and SSE:

$$
\begin{aligned}
SSR(X_1, X_2, X_3) &= SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) \\
&= SSR(X_2) + SSR(X_3|X_2) + SSR(X_1|X_2, X_3) \\
&= SSR(X_1) + SSR(X_2, X_3|X_1) \\
SSR(X_3|X_1, X_2) &= SSE(X_1, X_2) - SSE(X_1, X_2, X_3) \\
SSR(X_3|X_1, X_2) &= SSR(X_1, X_2, X_3) - SSR(X_1, X_2) \\
SSR(X_2, X_3|X_1) &= SSE(X_1) - SSE(X_1, X_2, X_3) \\
&= SSR(X_1, X_2, X_3) - SSR(X_1)
\end{aligned}
$$

Model Selection

There are several ways (criteria) can be used to find the 'best' subset model. In general, there are 2^p models that involve subsets of p predictors.

For example, in a case with 3 predictors, there are $2^3 - 1 = 7$ different possible subsets that can be formed from a pool of 3 independent variables. For p=10, leads to 1,023 to subsets, if p=20, there are 1,048,575 subsets. In practice, having 20 predictors is not unusual, especially in medical dataset.

DukeHealth

"Best" Subset Selection

- There are primarily two types of variable selection:
  i) Penalty based criterion approaches that attempt to find the model that optimizes some measure of goodness.
  ii) A stepwise approach that compares one model to another, assessing the change in fit at each step.

DukeHealth

"Best" Subset Selection

While there are many Penalty based criteria for comparing regression models, but the most commonly used are Mallow's C_p, mean-square error (MSE_p), R_p^2, adjusted-R_p^2 ($R_{a,p}^2$), Akaike's information criterion (AIC_p) and Schwarz' Bayesian criterion (SBC_p, also called Bayesian information criterion [BIC_p]). In practice, it is always good to weigh-in several criteria in selecting the 'best' model.

DukeHealth

Model Selection: C_p

The Mallow's C_p is defined as:

Definition

$C_p = \frac{SSE_p}{MSE_{full}} - (n - 2p)$, where SSE_p is the error sum squares of the reduced model.

Plot C_p values against p, those models with small bias will tend to be near the line $C_p = p$.

Note: the expected value of C_p is approximately p (ie, $E(C_p) \approx p$).

Note: For the full model (ie, model with all variables) has $C_P = P$ exactly.

Model Selection: MSE_p and R_p^2

The mean-square error is defined as:

Definition

$MSE_p = \frac{SSE_p}{n-p}$, where SSE_p is the sum-squares error for the reduced model.
Select models with small MSE_p.

Adjusted-R_p^2 is defined as:

Definition

$R_{a,p}^2 = 1 - (1 - R_p^2)\frac{n-1}{n-p}$, where $R_p^2 = \frac{SS_{reg(p)}}{SST}$.

Similar, to C_p, it is useful to plot $R_{a,p}^2$ (or R_p^2) values against p.

Note: That $R_{a,p}^2$ and MSE_p provide equivalent information.

DukeHealth

Model Selection: AIC

Akaike's information criterion (AIC_p) is defined below. Assuming that all candidate regression models use the same number of observations (n)

Definition

$AIC_p = n \ln \frac{SSE_p}{n} + 2p$, where $\ln$ is the natural log. p includes the intercept parameter. Model with the smallest AIC_p is the best-fitting model from among the candidates.

- Akaike Information Criterion (AIC) is a measure of the divergence between the true distribution (model) and a candidate, measured in terms of the Kullback-Leibler distance.
- AIC is defined slightly differently in different software packages, but was originally defined as AIC $= -2l_{max} + 2p$, where l_{max} is the log-likelihood maximum and p is the number of unknown parameters.

DukeHealth

Model Selection: AIC

- Note that the $+2p$ term is a term that increases the value of AIC_p when the number of predictors (p) is larger.
- Since we are looking for small values of AIC_p, this $+2p$ term is a penalty term - penalizing for more variables - which encourages the researcher to use as few predictors as needed.
- Several researchers have suggested other types of information criteria - usually altering the penalty term, often as a function of sample size.

Model Selection: BIC

Another common information criterion measure is Schwartzs Bayesian information criterion (BIC_p) defined as:

Definition

$BIC_p = n \ln \frac{SSE_p}{n} + [\ln n]p.$
Again, a better model is one with a smaller BIC_p.

Models with small SSE_p will do well under both criteria (AIC_p and BIC_p) as long as the penalties ($2p$ or $[\ln n]p$) are not too large. Only when $n \geq 8$ the penalty for BIC_p is larger.

DukeHealth

Drawbacks: AIC and BIC

- These information criteria measures have no distributional properties that would help us determine if the differences seen between models is "large" or "small".
- These measures work poorly in the presence of multicollinearity.
  Consider adding (separately) two variables u and v - where u is highly collinear with variables already in the model, but v is not. Suppose neither change the term $n \ln \frac{SSE_p}{n}$ much (and both increase p by 1), so the AIC_p (or BIC_p) values are about the same for the two choices. However, adding u will adversely impact the t-statistics of the variables with which it is correlated, making them appear less important, which v does not. Clearly, v would be the better choice, but AIC_p (or BIC_p) can not distinguish between the two.

DukeHealth

Example:

```
data one;
label sbp = 'Systolic blood pressure'
      quet= 'Quetelet index'
  age = 'Age'
  smk = 'Smoking histroy';
input sbp quet age smk @@;datalines;
135 2.876 45 0 122 3.251 41 0 130 3.100 49 0 148 3.768 52 0 146 2.979 54 1
129 2.790 47 1 162 3.668 60 1 160 3.612 48 1 144 2.368 44 1 180 4.637 64 1
166 3.877 59 1 138 4.032 51 1 152 4.116 64 0 138 3.673 56 0 140 3.562 54 1
134 2.998 50 1 145 3.360 49 1 142 3.024 46 1 135 3.171 57 0 142 3.401 56 0
150 3.628 56 1 144 3.751 58 0 137 3.296 53 0 132 3.210 50 0 149 3.301 54 1
132 3.017 48 1 120 2.789 43 0 126 2.956 43 1 161 3.800 63 0 170 4.132 63 1
152 3.962 62 0 164 4.010 65 0
;
run;

title1 "*** R-square selection ***";
proc rsquare data=one;
  model sbp=quet age smk/adjrsq sse bic aic sbc cp mse rmse;
run;
```

Model Selection:

```
*** R-square selection ***

The RSQUARE Procedure
Model: MODEL1
Dependent Variable: sbp

R-Square Selection Method

Number of Observations Read          32
Number of Observations Used          32
```

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	BIC	MSE	Root MSE
1	0.6009	0.5876	18.7414	144.2790	144.8186	85.47795	9.24543
1	0.5506	0.5356	24.6414	148.0829	148.2070	96.26743	9.81160
1	0.0612	0.0299	81.9640	171.6558	169.8144	201.09569	14.18082
2	0.7298	0.7112	5.6481	133.8005	135.8670	59.87188	7.73769
2	0.6412	0.6165	16.0212	142.8724	143.3278	79.49574	8.91604
2	0.6412	0.6165	16.0253	142.8756	143.3304	79.50353	8.91647
3	0.7609	0.7353	4.0000	131.8814	134.9835	54.86225	7.40691

Number in Model	R-Square	SSE	Variables in Model
1	0.6009	2564.33838	age
1	0.5506	2888.02301	quet
1	0.0612	6032.87059	smk
2	0.7298	1736.28452	age smk
2	0.6412	2305.37650	quet age
2	0.6412	2305.60226	quet smk
3	0.7609	1536.14305	quet age smk

DukeHealth

Model Selection: Testing-Based Procedures

Backward Elimination

1) The "backward elimination" procedure for model selection means that you start with all possible predictors in the model, and then remove the predictor that meets some criterion for "least important".
2) When implementing a backward elimination procedure using statistical testing, the criterion for "least important" is the variable with the highest p-value (from partial $F-$statistic) greater than some cutoff (e.g., α_{crit}). That is, Determine the partial F statistic for every variable in the model as if were the last variable to enter to the model,
3) The α_{crit} value; is often called the "p to remove" value, and does not need to be (nor probably should be) as small as 0.05. It is often chosen to be larger (e.g., 0.10 or 0.15), to allow variables to remain in the model that are correlated with other predictors.

DukeHealth

4) Once the "least important" is removed, the model is re-fit without it and the procedure repeated. It stops when all the variables remaining are significant at the α_{crit} level.

- The drawback to this type of stepwise procedure is that it cannot guarantee that the final model is optimal in any way (e.g., not largest R^2 ,R^2_{adj} or smallest MSE of any set of possible predictors).

DukeHealth

Example:

```
data one;
label sbp = 'Systolic blood pressure'
      quet= 'Quetelet index'
  age = 'Age'
  smk = 'Smoking histroy';
input sbp quet age smk @@;datalines;
135 2.876 45 0 122 3.251 41 0 130 3.100 49 0 148 3.768 52 0 146 2.979 54 1
129 2.790 47 1 162 3.668 60 1 160 3.612 48 1 144 2.368 44 1 180 4.637 64 1
166 3.877 59 1 138 4.032 51 1 152 4.116 64 0 138 3.673 56 0 140 3.562 54 1
134 2.998 50 1 145 3.360 49 1 142 3.024 46 1 135 3.171 57 0 142 3.401 56 0
150 3.628 56 1 144 3.751 58 0 137 3.296 53 0 132 3.210 50 0 149 3.301 54 1
132 3.017 48 1 120 2.789 43 0 126 2.956 43 1 161 3.800 63 0 170 4.132 63 1
152 3.962 62 0 164 4.010 65 0
;
run;


title1 "*** Variable slection in regression: backward ***";
proc reg data=one;
  model sbp=quet age smk/selection=backward slentry=.05 slstay=.05;
run;
```

Backward Elimination:

```
      *** Variable slection in regression: backward ***

                 Backward Elimination: Step 0
   All Variables Entered: R-Square = 0.7609 and C(p) = 4.0000
                      Analysis of Variance
                                  Sum of          Mean
Source                  DF       Squares        Square   F Value

Model                    3    4889.82570    1629.94190     29.71
Error                   28    1536.14305      54.86225
Corrected Total         31    6425.96875
            Parameter    Standard
  Variable     Estimate       Error  Type II SS F Value Pr > F

  Intercept    45.10319    10.76488   963.09739   17.55 0.0003
  quet          8.59245     4.49868   200.14147    3.65 0.0664
  age           1.21271     0.32382   769.45920   14.03 0.0008
  smk           9.94557     2.65606   769.23345   14.02 0.0008

                  Backward Elimination: Step 1
   Variable quet Removed: R-Square = 0.7298 and C(p) = 5.6481
```

Backward Elimination:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	2	4689.68423	2344.84211	39.16
Error	29	1736.28452	59.87188	
Corrected Total	31	6425.96875		

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	48.04960	11.12956	1115.95464	18.64	0.0002
age	1.70916	0.20176	4296.58607	71.76	<.0001
smk	10.29439	2.76811	828.05385	13.83	0.0009

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square
1	quet	Quetelet index	2	0.0311	0.7298

Summary of Backward Elimination

Step	C(p)	F Value	Pr > F
1	5.6481	3.65	0.0664

Forward Selection

1) The forward selection method is just the reverse of backward selection. Find the one predictor that has the lowest $p-$value smaller than α_{crit}.
2) Given that this first variable is in the model, find the next (single) variable that again has the lowest $p-$value smaller than α_{crit}. that is, Determine the partial $F-$statistic for the remaining variables giving the variable initially selected,
3) Continue in this fashion until no remaining variables have a $p-$value smaller than α_{crit}.
4) As with backward selection, this method is not guaranteed to result in the selection of a best model under any criterion.

Example:

```
data one;
label sbp = 'Systolic blood pressure'
      quet= 'Quetelet index'
  age = 'Age'
  smk = 'Smoking histroy';
input sbp quet age smk @@;datalines;
135 2.876 45 0 122 3.251 41 0 130 3.100 49 0 148 3.768 52 0 146 2.979 54 1
129 2.790 47 1 162 3.668 60 1 160 3.612 48 1 144 2.368 44 1 180 4.637 64 1
166 3.877 59 1 138 4.032 51 1 152 4.116 64 0 138 3.673 56 0 140 3.562 54 1
134 2.998 50 1 145 3.360 49 1 142 3.024 46 1 135 3.171 57 0 142 3.401 56 0
150 3.628 56 1 144 3.751 58 0 137 3.296 53 0 132 3.210 50 0 149 3.301 54 1
132 3.017 48 1 120 2.789 43 0 126 2.956 43 1 161 3.800 63 0 170 4.132 63 1
152 3.962 62 0 164 4.010 65 0
;
run;


title1 "*** Variable slection in regression: forward ***";
proc reg data=one;
  model sbp=quet age smk/selection=forward slentry=.05 slstay=.05;
run;
```

Forward Procedure:

```
*** Variable slection in regression: forward ***
            Forward Selection: Step 1
               Statistics for Entry
                    DF = 1,30

                                 Model
Variable        Tolerance     R-Square    F Value    Pr > F

quet             1.000000       0.5506      36.75    <.0001
age              1.000000       0.6009      45.18    <.0001
smk              1.000000       0.0612       1.95    0.1723

 Variable age Entered: R-Square = 0.6009 and C(p) = 18.7414
        Analysis of Variance

                             Sum of          Mean
Source               DF     Squares        Square   F Value

Model                 1  3861.63038    3861.63038     45.18
Error                30  2564.33838      85.47795
Corrected Total      31  6425.96875
```

Forward Procedure:

```
      Forward Selection: Step 2
                   Statistics for Entry
                        DF = 1,29
                                   Model
  Variable        Tolerance      R-Square    F Value    Pr > F
  quet             0.355591        0.6412       3.26    0.0815
  smk              0.980544        0.7298      13.83    0.0009

 Variable smk Entered: R-Square = 0.7298 and C(p) = 5.6481
                   Analysis of Variance

                                 Sum of          Mean
Source                  DF      Squares        Square   F Value
Model                    2   4689.68423    2344.84211     39.16
Error                   29   1736.28452      59.87188
Corrected Total         31   6425.96875

             Parameter    Standard
  Variable     Estimate       Error  Type II SS F Value Pr > F
  Intercept    48.04960    11.12956  1115.95464    18.64 0.0002
  age           1.70916     0.20176  4296.58607    71.76 <.0001
  smk          10.29439     2.76811   828.05385    13.83 0.0009
----------------------------------------------------------------
               Forward Selection: Step 3
                   Statistics for Entry
                        DF = 1,28
                                   Model
  Variable        Tolerance      R-Square    F Value    Pr > F
  quet             0.353910        0.7609       3.65    0.0664

 No other variable met the 0.0500 significance level for entry
                       into the model.
```

Stepwise Selection

1) Stepwise selection is a combination of forward and backward selection. At each step, a variable can be removed or entered, based on a criterion, such as comparing its $p-$value to α_{crit}.
2) Again, it is not guaranteed to result in the selection of a "best" set of predictors.

Example:

```
data one;
label sbp = 'Systolic blood pressure'
      quet= 'Quetelet index'
  age = 'Age'
  smk = 'Smoking histroy';
input sbp quet age smk @@;datalines;
135 2.876 45 0 122 3.251 41 0 130 3.100 49 0 148 3.768 52 0 146 2.979 54 1
129 2.790 47 1 162 3.668 60 1 160 3.612 48 1 144 2.368 44 1 180 4.637 64 1
166 3.877 59 1 138 4.032 51 1 152 4.116 64 0 138 3.673 56 0 140 3.562 54 1
134 2.998 50 1 145 3.360 49 1 142 3.024 46 1 135 3.171 57 0 142 3.401 56 0
150 3.628 56 1 144 3.751 58 0 137 3.296 53 0 132 3.210 50 0 149 3.301 54 1
132 3.017 48 1 120 2.789 43 0 126 2.956 43 1 161 3.800 63 0 170 4.132 63 1
152 3.962 62 0 164 4.010 65 0
;
run;


title1 "*** Variable slection in regression: stepwise ***";
proc reg data=one;
  model sbp=quet age smk/selection=stepwise slentry=.05 slstay=.05;
run;
```

Stepwise Selection: Step 1

Variable age Entered: R-Square = 0.6009 and C(p) = 18.7414

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	3861.63038	3861.63038	45.18
Error	30	2564.33838	85.47795	
Corrected Total	31	6425.96875		

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	59.09163	12.81626	1817.11840	21.26	<.0001
age	1.60450	0.23872	3861.63038	45.18	<.0001

Stepwise Selection: Step 2

Variable smk Entered: R-Square = 0.7298 and C(p) = 5.6481

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	2	4689.68423	2344.84211	39.16
Error	29	1736.28452	59.87188	
Corrected Total	31	6425.96875		

DukeHealth

Stepwise Procedure:

Dependent Variable: sbp Systolic blood pressure

Stepwise Selection: Step 2

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	48.04960	11.12956	1115.95464	18.64	0.0002
age	1.70916	0.20176	4296.58607	71.76	<.0001
smk	10.29439	2.76811	828.05385	13.83	0.0009

All variables left in the model are significant at the 0.0500 level.
No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In
1	age		Age	1
2	smk		Smoking histroy	2

Summary of Stepwise Selection

Step	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	0.6009	0.6009	18.7414	45.18	<.0001
2	0.1289	0.7298	5.6481	13.83	0.0009

Variable selection R-code for quet example:

```
library(MASS)
#
quet <- read.table("FILE PATH\\Lecture 2\\QUET.txt", header=T)
quet
###################################
fit=lm(sbp~QUET+age+smk, data=quet)
summary(fit)
####################################
library(olsrr)
####################################
##all possible models
ols_step_all_possible(fit)
####################################
##best subset --this is similar to proc rsquare in SAS;
ols_step_best_subset(fit)
####################################
##backward elimination
ols_step_backward_p(fit,prem = 0.05)
#####################################
##forward selection
ols_step_forward_p(fit,penter = 0.05)
####################################
##stepwise selection
ols_step_both_p(fit,pent = 0.05,prem = 0.05)
```

Model Selection: more about Testing-Based Methods

- Drawbacks:

1) Because of the one-at-a-time process, the optimal model (based on some defined criterion) may be missed.
2) The multiple testing issue is present many significance tests means that $p-$values cannot be interpreted literally.
3) The removal of less significant predictors tends to increase the significance of the remaining predictors - which could overstate their importance.
4) The procedure is not linked to the objectives of the study, and may not result in a model that can address key study objectives (e.g., assessing the impact of a key variable).

DukeHealth

Model Selection: more about Testing-Based Methods

5) Stepwise procedures tend to result in smaller models than you may want for prediction purposes. That is, it may eliminate variables that are not statistically significant enough, but still useful for prediction.

6) There is a good chance that you will overfit the model to the particular dataset, resulting in a model that is very good for this dataset, but may not work well for a replicated study (and may not even make a lot of sense in terms of the problem at hand).

7) Automated procedures do not take into consideration residual analyses, influential points, etc.

8) Unless modified, the process would not keep all dummy variables for a categorical variable together (either in or out of the model). Unless modified, the process would not keep all dummy variables for a categorical variable together (either in or out of the model).

DukeHealth

As an alternative to model selection, we can fit a model which includes all p predictors using a technique that shrinks the coefficient estimates towards zero. The two commonly used techniques for shrinking the regression coefficients towards zero are ridge regression and the LASSO (Least Absolute Shrinkage and Selection Operator).

- Ridge Regression:
  Recall, in the least-squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ by minimizing

$$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

Ridge Regression

Ridge regression is very similar to least-squares fitting, except that the coefficients are estimated by minimizing a slightly different quantity. That is, ridge regression coefficients $\hat{\beta}$ are estimated by minimizing

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

where $\lambda \geq 0$ is referred to as a tuning parameter.

Ridge Regression

- Similar to least-squares, ridge regression seeks coefficient estimates that fit the data well, by making the SSE small. However, the second term $\lambda \sum_{j=1}^{p} \beta_j^2$, called a shrinkage penalty, is small when $\beta_1, \ldots, \beta_p$ are close to zero. The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.
- When $\lambda = 0$, then ridge regression will produce the least squares estimates.
- However, as $\lambda \longrightarrow \infty$, the impact of shrinkage penalty increases, and the ridge regression coefficient estimates will approach zero.

DukeHealth

Ridge Regression

- Unlike leas-squares fitting, which generates only one set of coefficient estimates $\hat{\beta}$, ridge regression will produce a different set of coefficient estimates $\hat{\beta}_\lambda$, for each value of λ, thus selecting a good value for λ is critical.
- Note, the shrinkage penalty is applied to $\beta_1, \ldots, \beta_p$, but not to the intercept β_0.
- Ridge regressions advantage over least squares is based on the bias-variance trade-off. As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.
- generally, cross-validation method has been used to select a good value for λ.

DukeHealth

LASSO

- Ridge regression does have one obvious disadvantage. Unlike best subset, forward stepwise, and backward stepwise selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model.
- The penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$).
- This may not be a problem for prediction accuracy, but it can create a challenge in model interpretation in settings in which the number of predictors p is quite large.
- The lasso is an alternative to ridge regression that overcomes the disadvantage above.

DukeHealth

LASSO

The LASSO has a very similar formulations to ridge regression, which differ in their expression of the shrinkage penalty. That is, LASSO coefficients $\hat{\beta}$ are estimated by minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{SSE} + \lambda\sum_{j=1}^{p}|\beta_j|$$

where $\lambda \geq 0$ is a tuning parameter, similar to the rideg regression.

- The only difference is that β_j^2 term in the Ridge regression penalty has been replaced with $|\beta_j|$ in the lasso penalty. In statistical term, the lasso uses l_1 penalty instead of an l_2 penalty.
- Similar to ridge regression, the lasso shrinks the model coefficient estimates towards zero. However, in the case of the lasso, the l_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.

DukeHealth

LASSO

- Hence, much like best subset selection, the lasso performs variable selection. As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression.
- As in ridge regression, selecting a good value of λ for the lasso is critical; via cross-validation methods, in the sense select λ which produces smallest cross-valiation error.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

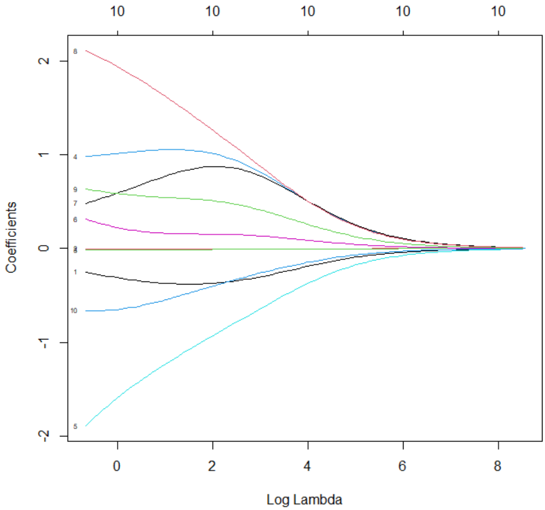
DukeHealth

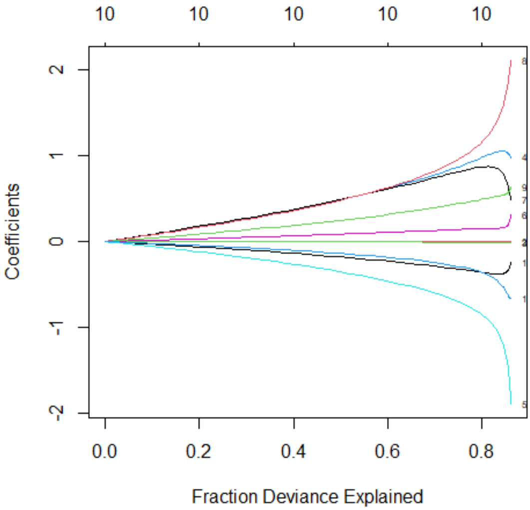
help(mtcars) in R data:

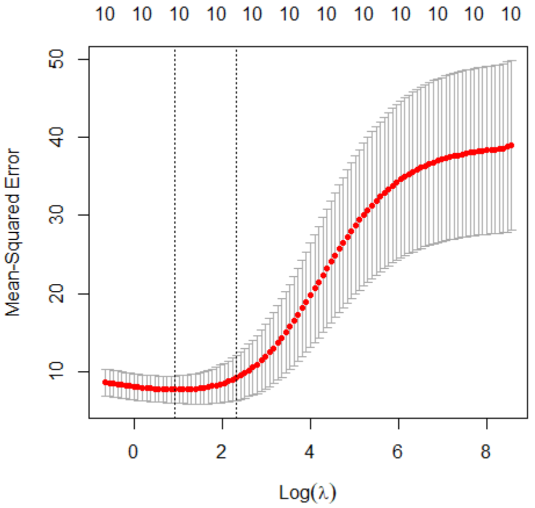
```
*  mtcars dataset: mpg is the response and there're 10 predictors
   in R if you invoke help(mtcars) this will give you the description of each predictors;

Description

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption
and 10 aspects of automobile design and performance for 32 automobiles (1973/74 models).

Usage
mtcars

Format
A data frame with 32 observations on 11 (numeric) variables.

[, 1]  mpg  Miles/(US) gallon
[, 2]  cyl  Number of cylinders
[, 3]  disp  Displacement (cu.in.)
[, 4]  hp  Gross horsepower
[, 5]  drat  Rear axle ratio
[, 6]  wt  Weight (1000 lbs)
[, 7]  qsec  1/4 mile time
[, 8]  vs  Engine (0 = V-shaped, 1 = straight)
[, 9]  am  Transmission (0 = automatic, 1 = manual)
[,10]  gear  Number of forward gears
[,11]  carb  Number of carburetors
```

```
library(glmnet)
## There is no model statement in glmnet, thus we need to create the x matrix and the response y
x=model.matrix(mpg~.-1,data=mtcars)
y=mtcars$mpg
########################
### Ridge regression ###
########################

### glmnet has a parameter alpha=0 (ridge regression)
fit.ridge=glmnet(x,y,alpha=0)
### plot log(lambda) vs. the regression coefficient
plot(fit.ridge,xvar="lambda",label=T)
### plot of fraction of deviance explained, similar to R-squares
plot(fit.ridge,xvar="dev",label=T)
### glmnet also does the cross-validation (cv)
cv.ridge=cv.glmnet(x,y,alpha=0)
plot(cv.ridge)
coef(cv.ridge)

########################
### LASSO regression ###
########################
### alpha=1 (LASSO regression) is the default in glmnet
fit.lasso=glmnet(x,y,alpha=1) ###or fit.lasso=glmnet(x,y)
### plot log(lambda) vs. the regression coefficient
plot(fit.lasso,xvar="lambda",label=T)
### plot of fraction of deviance explained, similar to R-squared
plot(fit.lasso,xvar="dev",label=T)
### glmnet also does the cross-validation (cv)
cv.lasso=cv.glmnet(x,y,alpha=1)
plot(cv.lasso)
coef(cv.lasso)
################################################################
### Cross-validation                                           #
### use train/validation division to select "lambda" for lasso #
################################################################
set.seed(1)
train=sample(seq(32),8,replace=FALSE)
train
lasso.tr=glmnet(x[train,],y[train])
pred=predict(lasso.tr,x[-train,])
rmse=sqrt(apply((y[-train]-pred)^2,2,mean))
plot(log(lasso.tr$lambda),rmse,type="b",xlab="log(lambda)")
lam.best=lasso.tr$lambda[order(rmse)[1]]
lam.best
coef(lasso.tr,s=lam.best)
```

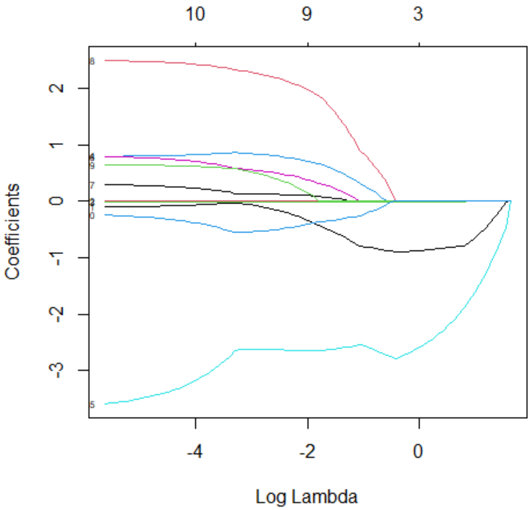
Ridge regression fit plot: plot(fit.ridge,xvar="lambda",label=T)



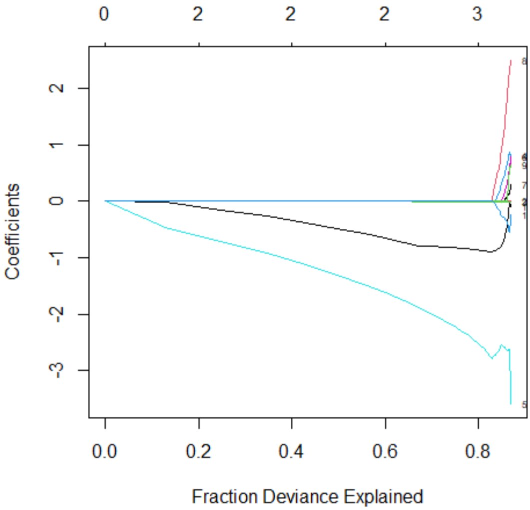

10
10
10
10
10
2
1
0
-1
-2
Coefficients
0.0
0.2
0.4
0.6
0.8
Fraction Deviance Explained

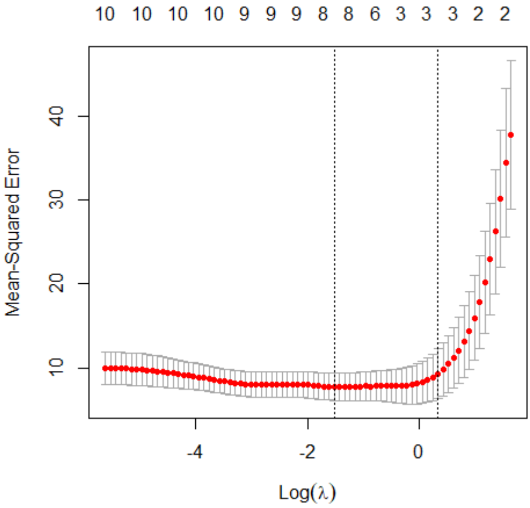

10 10 10 10 10 10 10 10 10 10 10 10
Mean-Squared Error
10
20
30
40
50
0
2
4
6
8
Log(λ)

```
> coef(cv.ridge)
11 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) 19.772087586
cyl         -0.356508690
disp        -0.005166736
hp          -0.009405988
drat         0.968566691
wt          -0.842240477
qsec         0.150254825
vs           0.867150289
am           1.143733962
gear         0.487695611
carb        -0.356201375
```


10
9
3
Coefficients
Log Lambda
-4
-2
0
2
1
0
-1
-2
-3

LASSO regression % deviance plot: plot(fit.lasso,xvar="dev",label=T)

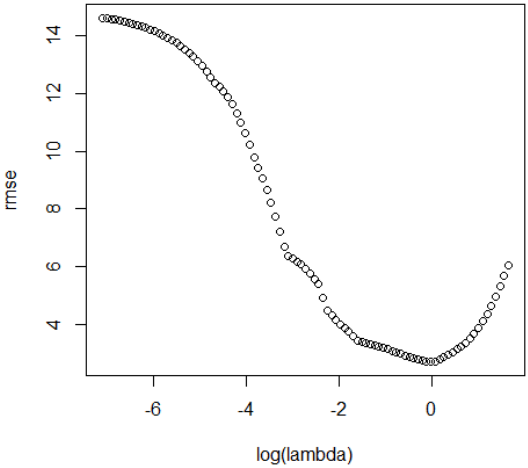



10 10 10 10 9 9 9 8 8 6 3 3 3 2 2
Mean-Squared Error
10
20
30
40
-4
-2
0
Log(λ)

LASSO regression coefficients:

```
> coef(cv.lasso)
11 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) 33.940487806
cyl         -0.843038418
disp         .
hp          -0.006965929
drat         .
wt          -2.365917424
qsec         .
vs           .
am           .
gear         .
carb         .
```

Cross-validation rmse plot:



```
> lam.best=lasso.tr$lambda[order(rmse)[1]]
> lam.best
[1] 1.070826
> coef(lasso.tr,s=lam.best)
11 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) 32.43505139
cyl         -0.22313181
disp         .
hp          -0.02435341
drat         0.53761712
wt          -2.42092591
qsec         .
vs           .
am           .
gear         .
carb        -0.29086140
```