

Linear Regression and Its Applications in Finance

Lixia Wang



- General Form and Best Fitted Model (OLS Method)
- Statistical Inferences
- ANOVA tests
- Model Selection
- Regression Diagnostics
- Resampling Methods
- Extension to Random Predictors



1. General Form and Best Fitted Model (OLS Method)

General Form: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon = E(y) + \varepsilon$ (1)

Observations: $(y_i, x_{i1}, x_{i2}, \dots, x_{ik}), i = 1, 2, \dots, n$ $Y = X\beta + \varepsilon$ (1')

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$
$$X = (X_0, X_1, X_2, \dots, X_k) = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i = \mathbf{x}_i^T \beta + \varepsilon_i = \beta^T \mathbf{x}_i + \varepsilon_i \quad (1'')$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} \quad (2) \quad \text{Residual: } y_i - \hat{y}_i = \hat{\varepsilon}_i$$

$$\hat{Y} = X\hat{\beta} \quad (2')$$

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



Best Fitted Model - Ordinary Least Squares Method

Minimize the residual sum of squares

$$\begin{aligned} RSS &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik})^2 \\ &= \|Y - \hat{Y}\|^2 = (Y - \hat{Y})^T (Y - \hat{Y}) = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \end{aligned} \quad (3)$$

OLS estimates of the β_i 's: $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)^T = (X^T X)^{-1} X^T Y$ (4)

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY \quad (5)$$

Note. \hat{Y} can be considered as the projection of Y into $L(X_0, X_1, \dots, X_k)$, the linear space spanned by X_0, X_1, \dots, X_k , and $H = X(X^T X)^{-1} X^T$ is the projection matrix (hat matrix). Thus, \hat{Y} and $Y - \hat{Y}$ are orthogonal, or

$$\|Y\|^2 = \|\hat{Y}\|^2 + \|Y - \hat{Y}\|^2 \quad (6)$$



Statistical Properties of OLS estimates

Consider $Y = X\beta + \varepsilon$. Let $p = k + 1$, assume $n \geq p$. If

- (A) x_{ij} are nonrandom constants and X has full rank p ,
- (B) ε_i are unobserved random disturbances with $E(\varepsilon_i) = 0$.

Then $\hat{\beta} = (X^T X)^{-1} X^T Y$ is an unbiased estimate of β .

If additionally,

- (C) $\text{Var}(\varepsilon_i) = \sigma^2$ and $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$,

Then $\text{Cov}(\hat{\beta}) = \text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 (X^T X)^{-1}$. (7)

In this case, an unbiased estimate of σ^2 is

$$s^2 = \frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{\|Y - \hat{Y}\|^2}{n-p} = \frac{\text{RSS}}{n-p} \triangleq \text{MSE} \quad (8)$$

(C^{*}): ε_i are independent $N(0, \sigma^2)$



Statistical Properties of OLS estimates

Consider $Y = X\beta + \varepsilon$. Let $p = k + 1$, assume $n \geq p$. If

- (A) x_{ij} are nonrandom constants and X has full rank p ,
- (C*) ε_i are independent $N(0, \sigma^2)$

Then

$$\hat{\beta} \sim N\left(\beta, \sigma^2(X^T X)^{-1}\right) \quad (9)$$

$$\frac{(n-p)s^2}{\sigma^2} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / \sigma^2 = \chi^2_{n-p} \quad (10)$$

$$\hat{\beta} \text{ and } s^2 \text{ are independent.} \quad (11)$$

From (9), $\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii})$ (12)

c_{ii} : diagonal element of $C =$

$$(X^T X)^{-1}$$

$$\frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}} \sim t(n-p) \quad (13)$$

$$\frac{(\hat{\beta} - \beta)^T (X^T X) (\hat{\beta} - \beta) / p}{s^2} \sim F_{p, n-p} \quad (14)$$

Example. Multiple linear regression with interest rates

Dataset: weekly interest rates from 2/16/1977 to 12/31/1993.

There are 11 variables in the dataset:

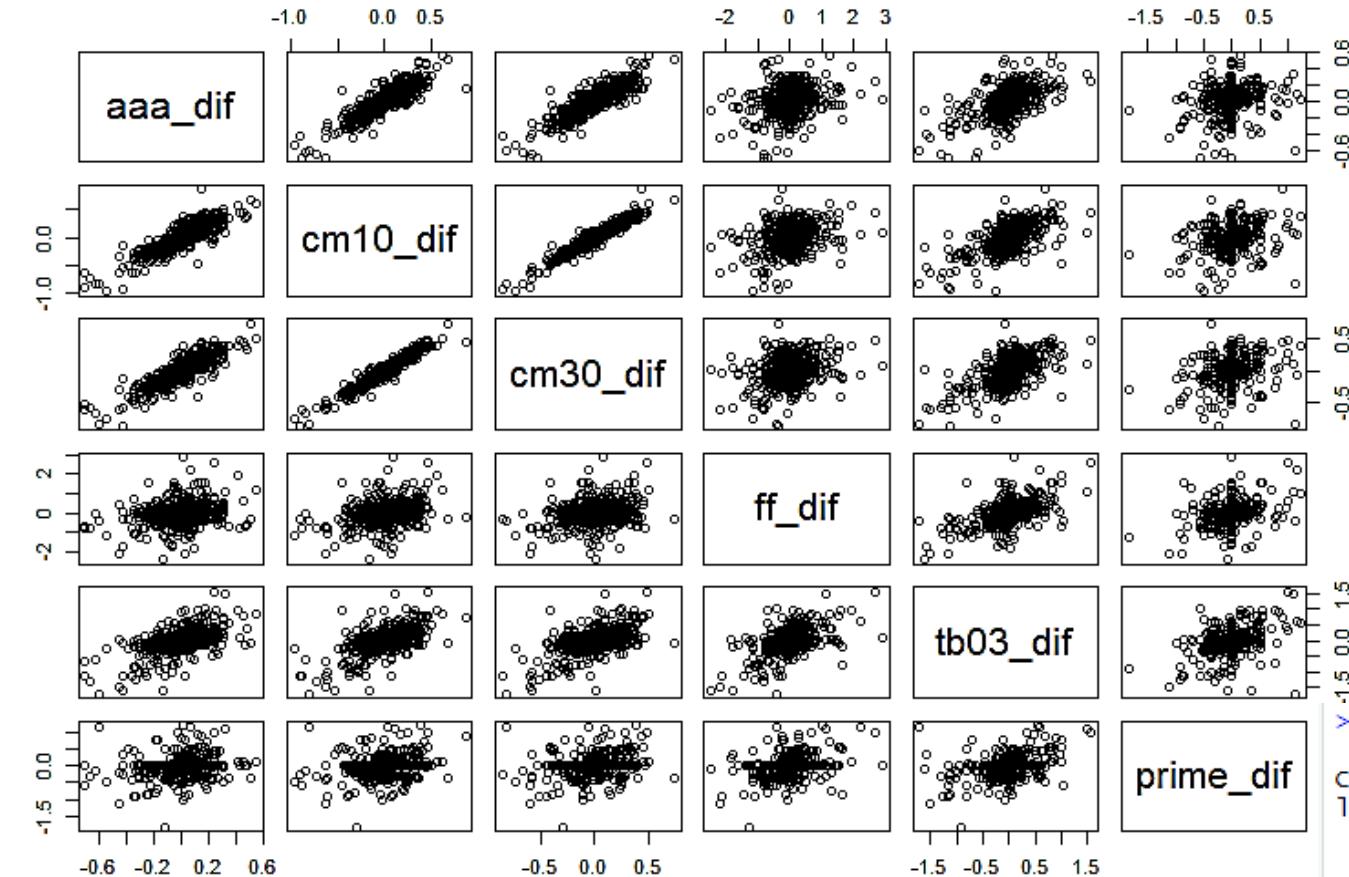
month, day, year, ff (Federal funds rate), tb03(3 month treasury-bill rate), cm10 (constant maturity 10-year Treasury bond rate), cm30 (constant maturity 30-year Treasury bond rate), discount, prime (prime rate), aaa (corporate AAA bond rate), and xxx.

Consider the six interest rates. We are interested in how the weekly changes in AAA bond rate are related to the weekly changes in 10-year Treasury bond rate (cm10), 30-year Treasury bond rate (cm30), prime rate (prime), Federal funds rate (ff) and 3 month treasury-bill rate (tb03)

	month	day	year	ff	tb03	cm10	cm30	discount	prime	aaa	xxx
1	2	16	77	4.70	4.62	7.36	7.69	5.25	6.25	8.04	105
2	2	23	77	4.74	4.67	7.39	7.75	5.25	6.25	8.08	105
3	3	2	77	4.68	4.70	7.47	7.81	5.25	6.25	8.10	105
4	3	9	77	4.63	4.64	7.49	7.82	5.25	6.25	8.12	105
5	3	16	77	4.62	4.59	7.45	7.80	5.25	6.25	8.09	105
6	3	23	77	4.77	4.57	7.44	7.77	5.25	6.25	8.00	105
7	3	30	77	4.74	4.59	7.46	7.79	5.25	6.25	8.10	105
8	4	6	77	4.60	4.56	7.44	7.78	5.25	6.25	8.10	105
9	4	13	77	4.65	4.58	7.39	7.77	5.25	6.25	8.05	105
10	4	20	77	4.71	4.51	7.27	7.66	5.25	6.25	7.99	105

	month	day	year	ff	tb03	cm10	cm30	discount	prime	aaa	xxx
872	10	27	93	2.97	3.06	5.42	5.98	3	6	6.73	94.59
873	11	3	93	3.04	3.07	5.54	6.03	3	6	6.87	95.12
874	11	10	93	2.96	3.09	5.70	6.20	3	6	6.92	95.08
875	11	17	93	3.03	3.10	5.67	6.17	3	6	6.94	95.22
876	11	24	93	2.98	3.12	5.82	6.31	3	6	6.99	95.81
877	12	1	93	3.09	3.12	5.80	6.27	3	6	6.95	96.21
878	12	8	93	2.92	3.10	5.75	6.21	3	6	6.86	95.73
879	12	15	93	2.94	3.05	5.77	6.23	3	6	6.97	95.62
880	12	22	93	2.99	3.05	5.81	6.29	3	6	6.95	95.65
881	12	29	93	2.99	3.05	5.73	6.24	3	6	6.94	95.33





$$\widehat{aaa}_{df} = 0.325cm10_{dif} + 0.301cm30_{dif} - 0.004ff_{dif} + 0.042tb03_{dif} - 0.017prime_{dif}$$

```

cm10_dif = diff( cm10 )
aaa_dif = diff( aaa )
cm30_dif = diff( cm30 )
ff_dif = diff( ff )
prime_dif = diff( prime )
tb03_dif = diff( tb03 )

> summary(lm(aaa_dif ~ cm10_dif + cm30_dif + ff_dif + tb03_dif + prime_dif))

call:
lm(formula = aaa_dif ~ cm10_dif + cm30_dif + ff_dif + tb03_dif +
    prime_dif)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.33563 -0.03129 -0.00137  0.03089  0.36729 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.159e-05  2.163e-03  -0.042 0.966235    
cm10_dif     3.248e-01   4.549e-02   7.140 1.96e-12 ***  
cm30_dif     3.006e-01   4.968e-02   6.050 2.15e-09 ***  
ff_dif       -4.322e-03   6.044e-03  -0.715 0.474801    
tb03_dif     4.192e-02   1.105e-02   3.794 0.000158 ***  
prime_dif   -1.703e-02   1.089e-02  -1.563 0.118342    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06416 on 874 degrees of freedom
Multiple R-squared:  0.7604,    Adjusted R-squared:  0.759 
F-statistic: 554.8 on 5 and 874 DF,  p-value: < 2.2e-16

```

2. Statistical Inference

2.1 Confidence Intervals

a) Confidence intervals for a regression coefficient

A $100(1 - \alpha)\%$ confidence interval for β_i is

$$\hat{\beta}_i - t_{\alpha/2} s \sqrt{c_{ii}} \leq \beta_i \leq \hat{\beta}_i + t_{\alpha/2} s \sqrt{c_{ii}} \quad (15)$$

$$\frac{\hat{\beta}_i - \beta_i}{s \sqrt{c_{ii}}} \sim t(n - p) \quad (13)$$

$$\frac{(\hat{\beta} - \beta)^T (X^T X) (\hat{\beta} - \beta) / p}{s^2} \sim F_{p, n-p} \quad (14)$$

b) Simultaneous confidence region for all regression coefficients

A $100(1 - \alpha)\%$ confidence region for β is

$$\left\{ \beta : \frac{(\hat{\beta} - \beta)^T (X^T X) (\hat{\beta} - \beta) / p}{s^2} \leq F_{p, n-p; 1-\alpha} \right\} \quad (16)$$

c) Confidence interval for the mean response

For an input \mathbf{x} , a $100(1 - \alpha)\%$ confidence region for $E(y) = \mathbf{x}^T \beta = \beta^T \mathbf{x}$ is

$$\hat{\beta}^T \mathbf{x} - t_{\alpha/2} s \sqrt{\mathbf{x}^T (X^T X)^{-1} \mathbf{x}} \leq E(y) \leq \hat{\beta}^T \mathbf{x} + t_{\alpha/2} s \sqrt{\mathbf{x}^T (X^T X)^{-1} \mathbf{x}} \quad (17)$$

d) Prediction interval for a new observation

For an input value \mathbf{x} , a $100(1 - \alpha)\%$ prediction interval for a future observation y is

$$\hat{\beta}^T \mathbf{x} - t_{\alpha/2} s \sqrt{1 + (X^T X)^{-1} \mathbf{x}} \leq y \leq \hat{\beta}^T \mathbf{x} + t_{\alpha/2} s \sqrt{1 + (X^T X)^{-1} \mathbf{x}} \quad (18)$$

3. ANOVA (analysis of variance) tests

F-tests of general linear hypotheses and ANOVA

Let $E(\varepsilon) \triangleq (E\varepsilon_1, E\varepsilon_2, \dots, E\varepsilon_n)^T$, then $E(\varepsilon) = 0$, $E(Y) \in L(X_0, X_1, \dots, X_k)^T$.

Assume X has full rank, L_0 is a linear subspace of L with $\dim(L_0) = r < p$, \hat{Y}_0 is the projection of Y into L_0 . Then,

$$RSS(L_0) = \|Y - \hat{Y}_0\|^2, \quad (19)$$

$$RSS(L_0) - RSS = \|Y - \hat{Y}_0\|^2 - \|Y - \hat{Y}\|^2 = \|\hat{Y} - \hat{Y}_0\|^2, \quad (20)$$

$$\|\hat{Y}\|^2 = \|\hat{Y}_0\|^2 + \|\hat{Y} - \hat{Y}_0\|^2 \quad (21)$$

Now we want to test

$$H_0: E(Y) \in L_0 \quad (22)$$

Under H_0 , $\|\hat{Y} - \hat{Y}_0\|^2/(p - r)$ is an unbiased estimate of σ^2 and is independent of s^2 . Therefore, test statistic

$$F = \frac{\|\hat{Y} - \hat{Y}_0\|^2/(p - r)}{\|Y - \hat{Y}\|^2/(n - p)} = \frac{(RSS(L_0) - RSS)/(p - r)}{RSS/(n - p)} \sim F_{p-r, n-p} \quad (23)$$



Test of an individual parameter coefficient

$$H_0: \beta_i = 0$$

$$L_0(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p)$$

$$\text{t-Test : } t = \frac{\hat{\beta}_i}{s\sqrt{c_{ii}}} \sim t_{n-p} \quad (24)$$

$$\text{F-Test: } F = t^2 = \left(\frac{\hat{\beta}_i}{s\sqrt{c_{ii}}} \right)^2 = \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\|Y - \hat{Y}\|^2 / (n-p)} \quad (25)$$

Testing the global usefulness of the model

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_0 \leftrightarrow L_0 = L(\mathbf{1})$$

$$H_0 \leftrightarrow L_0 = L(\mathbf{1}), \hat{Y}_0 = \bar{y}\mathbf{1} = \bar{Y}, \text{ and } \dim(L_0) = 1, p - r = k + 1 - 1 = k$$

$$F = \frac{\|\hat{Y} - \bar{Y}\|^2 / k}{\|Y - \hat{Y}\|^2 / (n - (k + 1))} \sim F_{k, n-(k+1)} \quad (26)$$

$$RSS(L_0) = \sum_{i=1}^n (y_i - \bar{y})^2 = \|Y - \bar{Y}\|^2, \quad RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \|Y - \hat{Y}\|^2,$$

$$RSS(L_0) - RSS = \|Y - \bar{Y}\|^2 - \|Y - \hat{Y}\|^2 = \|\hat{Y} - \bar{Y}\|^2$$

Example (cont.) ANOVA Table for Linear Regression

$$\hat{y} = 0.325cm10_{dif} + 0.301cm30_{dif} - 0.004ff_{dif} + 0.042tb03_{dif} - 0.017prime_{dif}$$

```
> anova(lm(aaa_dif ~ cm10_dif + cm30_dif + ff_dif + tb03_dif + prime_dif))  
Analysis of Variance Table
```

Response: aaa_dif

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cm10_dif	1	11.2071	11.2071	2722.2529	< 2.2e-16 ***
cm30_dif	1	0.1481	0.1481	35.9859	2.908e-09 ***
ff_dif	1	0.0025	0.0025	0.6178	0.432069
tb03_dif	1	0.0515	0.0515	12.5017	0.000428 ***
prime_dif	1	0.0101	0.0101	2.4439	0.118342
Residuals	874	3.5981	0.0041		

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

```
> anova(lm(aaa_dif ~ ff_dif + tb03_dif + prime_dif + cm10_dif + cm30_dif ))  
Analysis of Variance Table
```

Response: aaa_dif

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ff_dif	1	0.9392	0.9392	228.1280	< 2.2e-16 ***
tb03_dif	1	3.8963	3.8963	946.4353	< 2.2e-16 ***
prime_dif	1	0.0143	0.0143	3.4699	0.06283 .
cm10_dif	1	6.4188	6.4188	1559.1668	< 2.2e-16 ***
cm30_dif	1	0.1507	0.1507	36.6022	2.146e-09 ***
Residuals	874	3.5981	0.0041		

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



Example (cont.) Partial F-Tests and the Corresponding ANOVA Tables

```
> fit1 = lm(aaa_dif ~ cm10_dif)
> fit2 = lm(aaa_dif~cm10_dif+cm30_dif)
> fit3 = lm(aaa_dif~cm10_dif+cm30_dif+ff_dif)
> fit4 = lm(aaa_dif~cm10_dif+cm30_dif+tb03_dif)

> anova(fit1, fit3)
Analysis of Variance Table

Model 1: aaa_dif ~ cm10_dif
Model 2: aaa_dif ~ cm10_dif + cm30_dif + ff_dif
  Res.Df  RSS Df Sum of Sq   F Pr(>F)
1     878 3.81
2     876 3.66  2      0.151 18 2.1e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(fit2, fit3)
Analysis of Variance Table

Model 1: aaa_dif ~ cm10_dif + cm30_dif
Model 2: aaa_dif ~ cm10_dif + cm30_dif + ff_dif
  Res.Df  RSS Df Sum of Sq   F Pr(>F)
1     877 3.66
2     876 3.66  1      0.00254 0.61   0.44

> anova(fit2, fit4)
Analysis of Variance Table

Model 1: aaa_dif ~ cm10_dif + cm30_dif
Model 2: aaa_dif ~ cm10_dif + cm30_dif + tb03_dif
  Res.Df  RSS Df Sum of Sq   F Pr(>F)
1     877 3.66
2     876 3.61  1      0.05 12.1 0.00052 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



4. Model Selection

4.1 Some Variable Selection Criteria

i) Partial F-statistics

ii) Multiple Coefficient of Determination R^2

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\|\hat{Y} - \bar{Y}\|^2}{\|Y - \bar{Y}\|^2} = \frac{\|Y - \bar{Y}\|^2 - \|Y - \hat{Y}\|^2}{\|Y - \bar{Y}\|^2} = 1 - \frac{RSS}{SS_{yy}}$$

$$= \frac{\text{variability of } Y \text{ explained by model}}{\text{Total variability of } Y} \quad (27)$$

Adjusted Multiple Coefficient of Determination R_a^2

Note: $R_a^2 \leq R^2$

$$R_a^2 = 1 - \left[\frac{n-1}{n-(k+1)} \right] \left(\frac{SSE}{SS_{yy}} \right) = 1 - \left[\frac{n-1}{n-(k+1)} \right] (1 - R^2) \quad (28)$$

One model selection procedure is to choose the model with the largest R_a^2 .



iii) Mallows' C_p

Let K be the largest possible number of predictors in the model.

Let RSS_p be the residual sum of squares when there are p predictors in the model.

$$s_p^2 = RSS_p / (n - p)$$

The Mallows' C_p -statistic is defined as

$$C_p = \frac{RSS_p}{s_K^2} + 2p - n \quad (29)$$

when a subset of p predictors is chosen from the full set $\{x_1, x_2, \dots, x_K\}$.

Mallows (1973) suggested to choose the subset that has the smallest C_p . Overfitting or underfitting tends to increase the value of C_p .



Let \hat{L} be the maximum value of the likelihood function for the model, and p be the number of estimated parameters in the model.

$$-2 \log(\hat{L}) + 2p \quad (30)$$

iv) Akaike's information criterion (AIC)

AIC of the model is defined as

$$AIC(p) = \log(\hat{\sigma}_p^2) + \frac{2p}{n} \quad (31)$$

The model selection procedure is to choose the model with the smallest $AIC(p)$.

v) Schwarz's Bayesian information criterion (BIC)

BIC of the model is defined as

$$BIC(p) = \log(\hat{\sigma}_p^2) + \frac{p(\log n)}{n} \quad (32)$$

The model selection procedure is to choose the model with the smallest $BIC(p)$.



Collinearity, Multicollinearity and VIF

Collinearity or multicollinearity occurs when two or more predictors are highly correlated with one another.

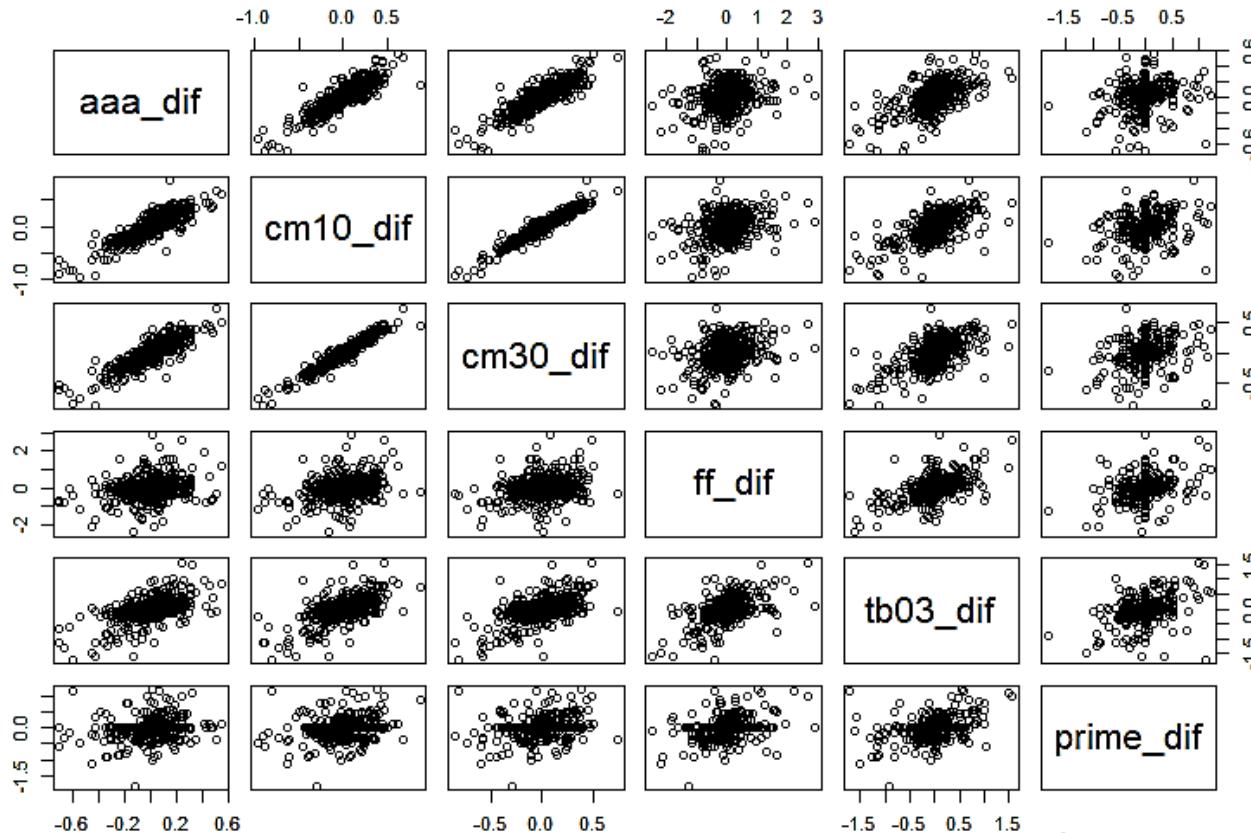
Suppose we have predictors x_1, x_2, \dots, x_p . Regressing x_j on other $p - 1$ predictors. Let R_j^2 be the R^2 -value of this regression, so R_j^2 measures how well x_j can be predicted by other predictors. Then the VIF (variance inflation factor) of x_j is defined as

$$VIF_j = \frac{1}{1 - R_j^2} \quad (33)$$

VIF_j tells nothing about the relationship between the response and x_j . When collinearity happens, the usual remedy is to reduce the number of predictors in the model by one of the model selection criteria.



Example (cont.) Collinearity & VIF



```
> cor(cm10_dif, cm30_dif)
[1] 0.96
```

```

> cm10_dif = diff( cm10 )
> aaa_dif = diff( aaa )
> cm30_dif = diff( cm30 )
> ff_dif = diff( ff )
> library(faraway)
> options(digits=2)
> vif(lm(aaa_dif~cm10_dif+cm30_dif+ff_dif+tb03_dif+prime_dif))
   cm10_dif   cm30_dif     ff_dif   tb03_dif prime_dif
       14.9       14.1      1.5       2.1      1.2
> detach(dat)

```



```

> summary(lm(aaa_dif ~ cm10_dif))

Call:
lm(formula = aaa_dif ~ cm10_dif)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.3894 -0.0330  0.0001  0.0293  0.4034 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.000109   0.002221  -0.05    0.96    
cm10_dif     0.615762   0.012117  50.82 <2e-16 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.066 on 878 degrees of freedom
Multiple R-squared:  0.746,    Adjusted R-squared:  0.746 
F-statistic: 2.58e+03 on 1 and 878 DF,  p-value: <2e-16

> summary(lm(aaa_dif ~ cm10_dif + cm30_dif))

Call:
lm(formula = aaa_dif ~ cm10_dif + cm30_dif)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.3345 -0.0314 -0.0005  0.0303  0.4075 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.38e-05  2.18e-03  -0.04    0.97    
cm10_dif     3.60e-01  4.45e-02   8.09  2.0e-15 ***  
cm30_dif     2.97e-01  4.98e-02   5.96  3.7e-09 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.065 on 877 degrees of freedom
Multiple R-squared:  0.756,    Adjusted R-squared:  0.756 
F-statistic: 1.36e+03 on 2 and 877 DF,  p-value: <2e-16

```

```

> summary(lm(aaa_dif ~ cm10_dif + ff_dif))

Call:
lm(formula = aaa_dif ~ cm10_dif + ff_dif)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.3883 -0.0327  0.0000  0.0294  0.4008 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.000108   0.002222  -0.05    0.96    
cm10_dif     0.614837   0.012647  48.62 <2e-16 ***  
ff_dif       0.001378   0.005367   0.26    0.80    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.066 on 877 degrees of freedom
Multiple R-squared:  0.746,    Adjusted R-squared:  0.746 
F-statistic: 1.29e+03 on 2 and 877 DF,  p-value: <2e-16

> summary(lm(aaa_dif ~ cm10_dif + cm30_dif + ff_dif))

Call:
lm(formula = aaa_dif ~ cm10_dif + cm30_dif + ff_dif)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.3348 -0.0311 -0.0006  0.0306  0.3999 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.07e-05  2.18e-03  -0.04    0.97    
cm10_dif     3.55e-01  4.51e-02   7.86  1.1e-14 ***  
cm30_dif     3.00e-01  5.00e-02   6.00  2.9e-09 ***  
ff_dif       4.12e-03  5.28e-03   0.78    0.44    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.065 on 876 degrees of freedom
Multiple R-squared:  0.756,    Adjusted R-squared:  0.755 
F-statistic: 906 on 3 and 876 DF,  p-value: <2e-16

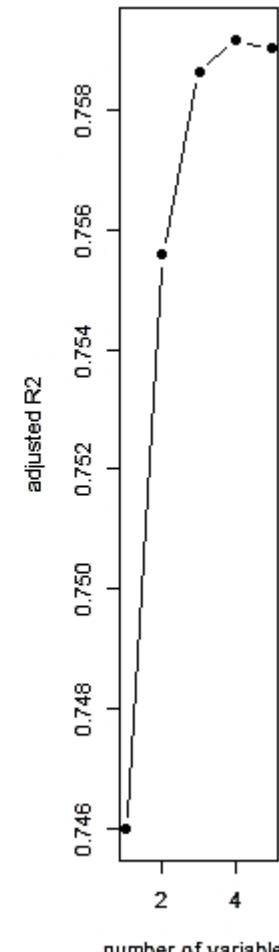
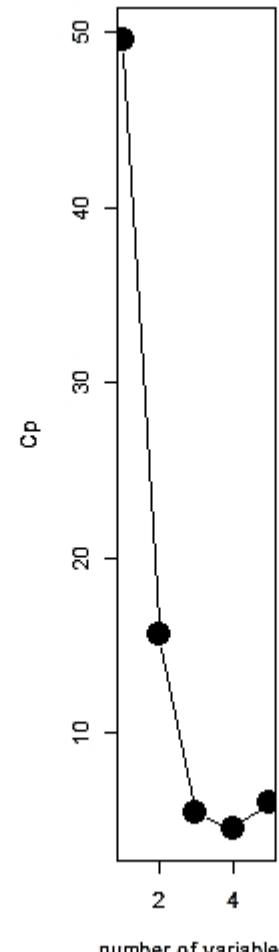
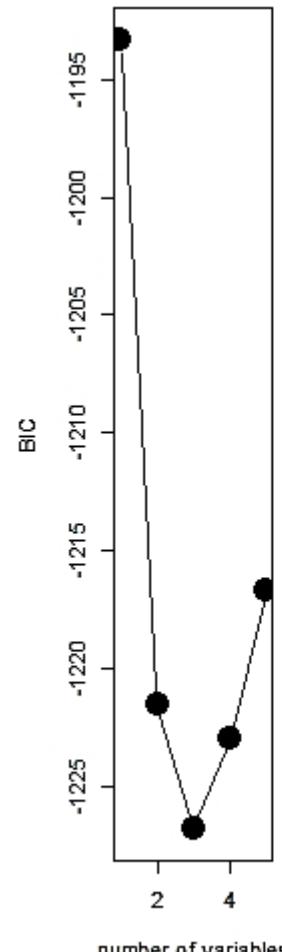
```



Example (cont.) Model Selection & VIF

```
Subset selection object
call: regsubsets.formula(ddd_dif ~ ., data = as.data.frame(cbind(cm10_dif,
  cm30_dif, ff_dif, tb03_dif, prime_dif)), nbest = 1)
5 Variables (and intercept)
  Forced in Forced out
cm10_dif      FALSE      FALSE
cm30_dif      FALSE      FALSE
ff_dif        FALSE      FALSE
tb03_dif      FALSE      FALSE
prime_dif     FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
  cm10_dif cm30_dif ff_dif tb03_dif prime_dif
1  ( 1 )    "*"      " "    " "    " "
2  ( 1 )    "*"      "*"    " "    " "
3  ( 1 )    "*"      "*"    " "    "*"    " "
4  ( 1 )    "*"      "*"    " "    "*"    "*" 
5  ( 1 )    "*"      "*"    "*"    "*"    "*" 

> cm10_dif = diff( cm10 )
> aaa_dif = diff( aaa )
> cm30_dif = diff( cm30 )
> ff_dif = diff( ff )
> library(faraway)
> options(digits=2)
> vif(lm(ddd_dif~cm10_dif+cm30_dif+ff_dif+tb03_dif+prime_dif))
  cm10_dif  cm30_dif   ff_dif  tb03_dif prime_dif
      14.9      14.1       1.5      2.1      1.2
> detach(dat)
```





4.2 Linear Regression Model Selection Methods

Subset, shrinkage, and dimension reduction

4.2.1 Subset Selection Methods

1. Best Subset Selection

1. Let \mathcal{M}_0 denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \dots, p$:
 - (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
 - (b) Pick the best among these $\binom{p}{k}$ models, and call it \mathcal{M}_k . Here *best* is defined as having the smallest RSS, or equivalently largest R^2 .
3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .

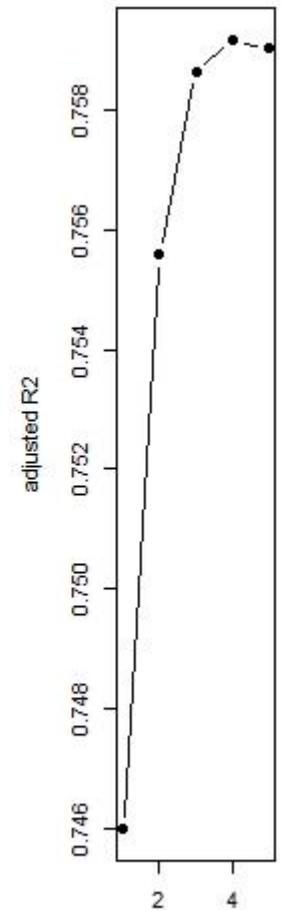
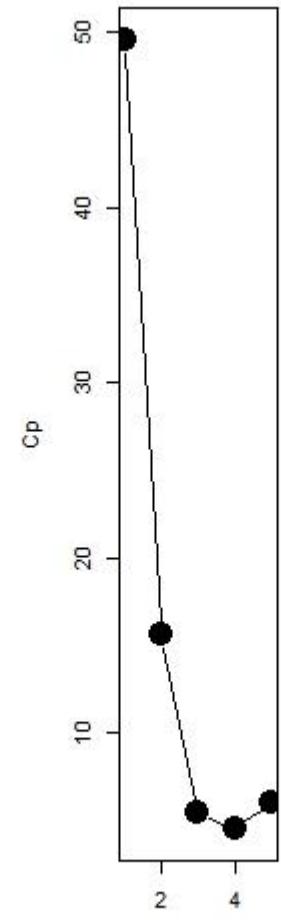
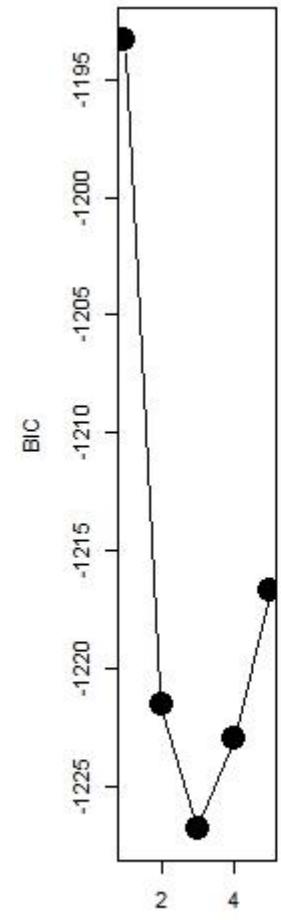
Problem:

Computationally infeasible (2^p models total) when p is very large.



Example (cont.) Best Subset Selection

```
> subsets = regsubsets(ddd_dif~, data=as.data.frame(cbind(cm10_dif,cm30_dif,ff_dif,tb03_dif,prime_dif)),
+                      nbest=1)
> b = summary(subsets)
> b
Subset selection object
call: regsubsets.formula(ddd_dif ~ ., data = as.data.frame(cbind(cm10_dif,
  cm30_dif, ff_dif, tb03_dif, prime_dif)), nbest = 1)
5 variables (and intercept)
  Forced in  Forced out
cm10_dif      FALSE      FALSE
cm30_dif      FALSE      FALSE
ff_dif        FALSE      FALSE
tb03_dif      FALSE      FALSE
prime_dif     FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
      cm10_dif cm30_dif ff_dif tb03_dif prime_dif
1 ( 1 )    "*"      " "    " "    " "    " "
2 ( 1 )    "*"      "*"    " "    " "    " "
3 ( 1 )    "*"      "*"    " "    "*"    " "
4 ( 1 )    "*"      "*"    " "    "*"    "*" 
5 ( 1 )    "*"      "*"    "*"    "*"    "*" 
```



2. Forward Selection

1. Let \mathcal{M}_0 denote the *null* model, which contains no predictors.
 2. For $k = 0, \dots, p - 1$:
 - (a) Consider all $p - k$ models that augment the predictors in \mathcal{M}_k with one additional predictor.
 - (b) Choose the *best* among these $p - k$ models, and call it \mathcal{M}_{k+1} . Here *best* is defined as having smallest RSS or highest R^2 .
 3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .
-

- Searches through at most $1 + p(p+1)/2$ models.
- This method can be used even in the high-dimensional setting where $n < p$.
- It is not guaranteed to yield the *best* model containing a subset of the p predictors.



`lm(formula = aaa_dif ~ cm10_dif)`

Residuals:

Min	1Q	Median	3Q	Max
-0.38945	-0.03299	0.00011	0.02933	0.40336

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.0001094	0.0022208	-0.049	0.961
cm10_dif	0.6157616	0.0121172	50.817	<2e-16 ***

Residual standard error: 0.06588 on 878 degrees of freedom

Multiple R-squared: 0.7463, **Adjusted R-squared: 0.746**

F-statistic: 2582 on 1 and 878 DF, p-value: < 2.2e-16

`lm(formula = aaa_dif ~ prime_dif)`

Residuals:

Min	1Q	Median	3Q	Max
-0.72082	-0.04904	0.00122	0.05389	0.53976

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.001220	0.004338	-0.281	0.779
prime_dif	0.107052	0.019884	5.384	9.36e-08 ***

Residual standard error: 0.1287 on 878 degrees of freedom

Multiple R-squared: 0.03196, **Adjusted R-squared: 0.03086**

F-statistic: 28.99 on 1 and 878 DF, p-value: 9.363e-08

`lm(formula = aaa_dif ~ cm30_dif)`

Residuals:

Min	1Q	Median	3Q	Max
-0.34351	-0.03247	-0.00156	0.02961	0.40110

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.0001208	0.0022570	-0.054	0.957
cm30_dif	0.6853163	0.0137830	49.722	<2e-16 ***

Residual standard error: 0.06695 on 878 degrees of freedom

Multiple R-squared: 0.7379, **Adjusted R-squared: 0.7376**

F-statistic: 2472 on 1 and 878 DF, p-value: < 2.2e-16

`lm(formula = aaa_dif ~ tb03_dif)`

Residuals:

Min	1Q	Median	3Q	Max
-0.53179	-0.05148	-0.00133	0.04534	0.52397

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.0007905	0.0036404	-0.217	0.828
tb03_dif	0.2575684	0.0127245	20.242	<2e-16 ***

Residual standard error: 0.108 on 878 degrees of freedom

Multiple R-squared: 0.3182, **Adjusted R-squared: 0.3174**

F-statistic: 409.7 on 1 and 878 DF, p-value: < 2.2e-16

`lm(formula = aaa_dif ~ ff_dif)`

Residuals:

Min	1Q	Median	3Q	Max
-0.64670	-0.05224	0.00148	0.05827	0.52606

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.001103	0.004269	-0.258	0.796
ff_dif	0.075642	0.009884	7.653	5.16e-14 ***

Residual standard error: 0.1266 on 878 degrees of freedom

Multiple R-squared: 0.06254, **Adjusted R-squared: 0.06147**

F-statistic: 58.57 on 1 and 878 DF, p-value: 5.158e-14



`lm(formula = aaa_dif ~ cm10_dif + cm30_dif)`

Residuals:

Min	1Q	Median	3Q	Max
-0.33450	-0.03139	-0.00049	0.03032	0.40748

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-9.376e-05	2.178e-03	-0.043	0.966
cm10_dif	3.602e-01	4.452e-02	8.092	1.96e-15 ***
cm30_dif	2.968e-01	4.983e-02	5.956	3.73e-09 ***

Residual standard error: 0.06462 on 877 degrees of freedom

Multiple R-squared: 0.7561, **Adjusted R-squared: 0.7556**

F-statistic: 1360 on 2 and 877 DF, p-value: < 2.2e-16

`lm(formula = aaa_dif ~ cm10_dif + prime_dif)`

Residuals:

Min	1Q	Median	3Q	Max
-0.38616	-0.03308	-0.00001	0.02888	0.40659

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.0001085	0.0022218	-0.049	0.961
cm10_dif	0.6170039	0.0124153	49.697	<2e-16 ***
prime_dif	-0.0048351	0.0104302	-0.464	0.643

Residual standard error: 0.06591 on 877 degrees of freedom

Multiple R-squared: 0.7463, **Adjusted R-squared: 0.7458**

F-statistic: 1290 on 2 and 877 DF, p-value: < 2.2e-16

`lm(formula = aaa_dif ~ cm10_dif + ff_dif)`

Residuals:

Min	1Q	Median	3Q	Max
-0.38831	-0.03274	-0.00004	0.02942	0.40080

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.0001085	0.0022220	-0.049	0.961
cm10_dif	0.6148371	0.0126467	48.616	<2e-16 ***
ff_dif	0.0013783	0.0053666	0.257	0.797

Residual standard error: 0.06591 on 877 degrees of freedom

Multiple R-squared: 0.7463, **Adjusted R-squared: 0.7457**

F-statistic: 1290 on 2 and 877 DF, p-value: < 2.2e-16

`lm(formula = aaa_dif ~ cm10_dif + tb03_dif)`

Residuals:

Min	1Q	Median	3Q	Max
-0.38605	-0.03301	-0.00041	0.02931	0.35910

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.0001082	0.0022091	-0.049	0.96096
cm10_dif	0.5864905	0.0151056	38.826	< 2e-16 ***
tb03_dif	0.0311095	0.0096768	3.215	0.00135 **

Residual standard error: 0.06553 on 877 degrees of freedom

Multiple R-squared: 0.7492, **Adjusted R-squared: 0.7487**

F-statistic: 1310 on 2 and 877 DF, p-value: < 2.2e-16



```
lm(formula = aaa_dif ~ cm10_dif + cm30_dif + ff_dif)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.33484	-0.03115	-0.00059	0.03062	0.39986

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-9.069e-05	2.179e-03	-0.042	0.967
cm10_dif	3.545e-01	4.512e-02	7.858	1.14e-14 ***
cm30_dif	3.002e-01	5.003e-02	6.000	2.88e-09 ***
ff_dif	4.122e-03	5.282e-03	0.780	0.435

Residual standard error: 0.06463 on 876 degrees of freedom
Multiple R-squared: 0.7563, **Adjusted R-squared: 0.7555**
F-statistic: 906.2 on 3 and 876 DF, p-value: < 2.2e-16

```
lm(formula = aaa_dif ~ cm10_dif + cm30_dif + tb03_dif)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.33612	-0.03134	-0.00153	0.03054	0.36055

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-9.209e-05	2.165e-03	-0.043	0.966078
cm10_dif	3.242e-01	4.543e-02	7.136	2.01e-12 ***
cm30_dif	3.025e-01	4.954e-02	6.106	1.53e-09 ***
tb03_dif	3.303e-02	9.488e-03	3.482	0.000523 ***

Residual standard error: 0.06421 on 876 degrees of freedom
Multiple R-squared: 0.7595, **Adjusted R-squared: 0.7586**
F-statistic: 922 on 3 and 876 DF, p-value: < 2.2e-16

```
lm(formula = aaa_dif ~ cm10_dif + cm30_dif + prime_dif)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.33433	-0.03164	-0.00067	0.03038	0.41073

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-9.282e-05	2.179e-03	-0.043	0.966
cm10_dif	3.615e-01	4.461e-02	8.102	1.81e-15 ***
cm30_dif	2.968e-01	4.985e-02	5.954	3.79e-09 ***
prime_dif	-4.868e-03	1.023e-02	-0.476	0.634

Residual standard error: 0.06465 on 876 degrees of freedom
Multiple R-squared: 0.7562, **Adjusted R-squared: 0.7554**
F-statistic: 905.7 on 3 and 876 DF, p-value: < 2.2e-16



```
lm(formula = aaa_dif ~ cm10_dif + cm30_dif + tb03_dif + prime_dif)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.33579	-0.03130	-0.00121	0.03108	0.36489

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-8.829e-05	2.162e-03	-0.041	0.967440
cm10_dif	3.228e-01	4.539e-02	7.112	2.38e-12 ***
cm30_dif	3.035e-01	4.949e-02	6.133	1.30e-09 ***
tb03_dif	3.860e-02	1.002e-02	3.851	0.000126 ***
prime_dif	-1.832e-02	1.074e-02	-1.707	0.088265 .

Residual standard error: 0.06414 on 875 degrees of freedom

Multiple R-squared: 0.7603, **Adjusted R-squared: 0.7592**

F-statistic: 693.7 on 4 and 875 DF, p-value: < 2.2e-16

```
lm(formula = aaa_dif ~ cm10_dif + cm30_dif + tb03_dif + prime_dif + ff_dif)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.3356	-0.0313	-0.0014	0.0309	0.3673

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-9.16e-05	2.16e-03	-0.04	0.96623
cm10_dif	3.25e-01	4.55e-02	7.14	2.0e-12 ***
cm30_dif	3.01e-01	4.97e-02	6.05	2.1e-09 ***
tb03_dif	4.19e-02	1.10e-02	3.79	0.00016 ***
prime_dif	-1.70e-02	1.09e-02	-1.56	0.11834
ff_dif	-4.32e-03	6.04e-03	-0.71	0.47480

Residual standard error: 0.0642 on 874 degrees of freedom

Multiple R-squared: 0.76, **Adjusted R-squared: 0.759**

F-statistic: 555 on 5 and 874 DF, p-value: <2e-16

```
lm(formula = aaa_dif ~ cm10_dif + cm30_dif + tb03_dif + ff_dif)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.33588	-0.03128	-0.00152	0.03079	0.36425

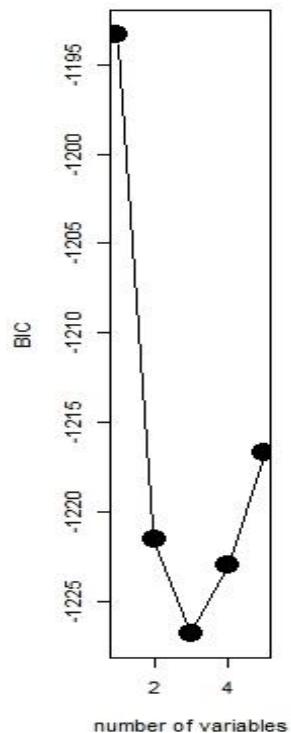
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-9.622e-05	2.165e-03	-0.044	0.964558
cm10_dif	3.268e-01	4.551e-02	7.181	1.48e-12 ***
cm30_dif	2.986e-01	4.970e-02	6.007	2.77e-09 ***
tb03_dif	3.810e-02	1.078e-02	3.533	0.000433 ***
ff_dif	-5.894e-03	5.965e-03	-0.988	0.323391

Residual standard error: 0.06422 on 875 degrees of freedom

Multiple R-squared: 0.7597, **Adjusted R-squared: 0.7586**

F-statistic: 691.7 on 4 and 875 DF, p-value: < 2.2e-16



3. Backward Elimination

1. Let \mathcal{M}_p denote the *full* model, which contains all p predictors.
 2. For $k = p, p - 1, \dots, 1$:
 - (a) Consider all k models that contain all but one of the predictors in \mathcal{M}_k , for a total of $k - 1$ predictors.
 - (b) Choose the *best* among these k models, and call it \mathcal{M}_{k-1} . Here *best* is defined as having smallest RSS or highest R^2 .
 3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .
-

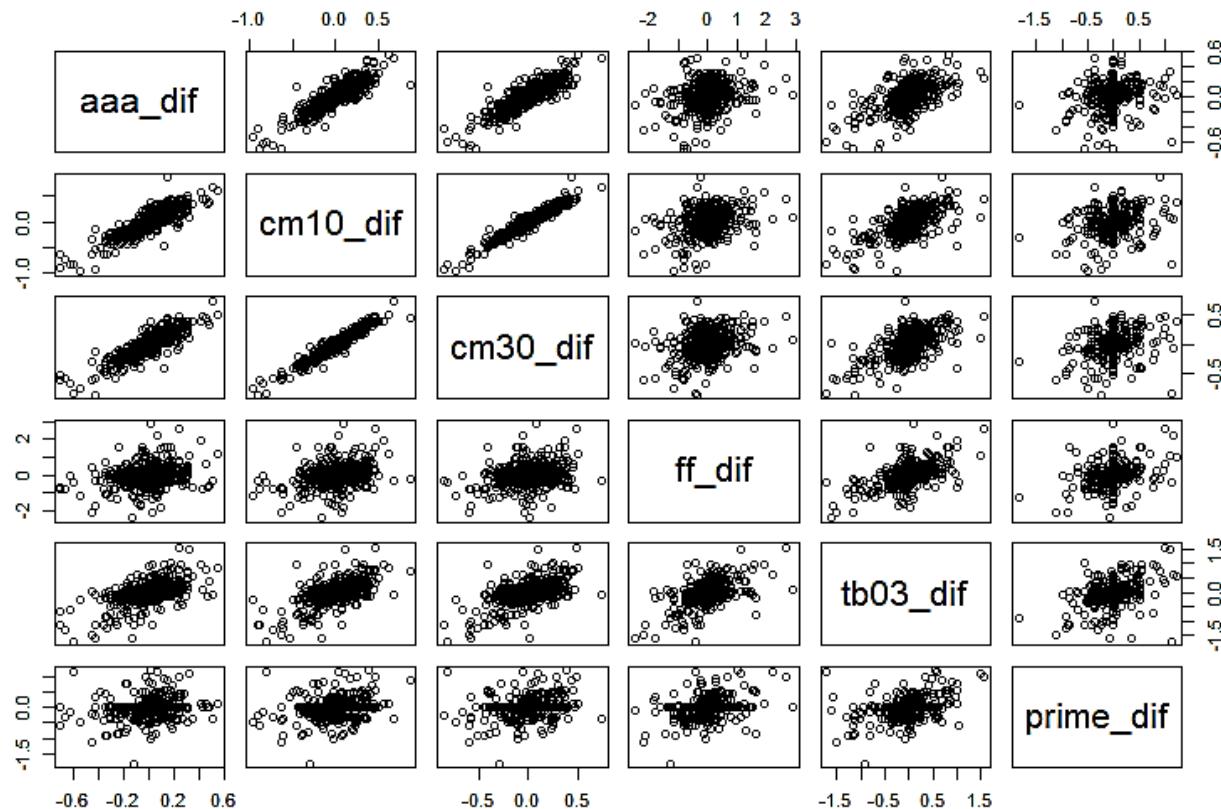
Searches through only $1 + p(p+1)/2$ models. Remove the least significant variable from the model, one-at-a-time.

It requires n is larger than p (so that the full model can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.

It is not guaranteed to yield the *best* model containing a subset of the p predictors.



Example (cont.) Backward Elimination



```
> summary(lm(aaa_dif ~ cm10_dif + cm30_dif + ff_dif + tb03_dif + prime_dif))

Call:
lm(formula = aaa_dif ~ cm10_dif + cm30_dif + ff_dif + tb03_dif +
prime_dif)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.33563 -0.03129 -0.00137  0.03089  0.36729 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.159e-05 2.163e-03 -0.042 0.966235  
cm10_dif     3.248e-01 4.549e-02  7.140 1.96e-12 ***
cm30_dif     3.006e-01 4.968e-02  6.050 2.15e-09 ***
ff_dif       -4.322e-03 6.044e-03 -0.715 0.474801  
tb03_dif     4.192e-02 1.105e-02  3.794 0.000158 *** 
prime_dif   -1.703e-02 1.089e-02 -1.563 0.118342  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06416 on 874 degrees of freedom
Multiple R-squared:  0.7604,    Adjusted R-squared:  0.759 
F-statistic: 554.8 on 5 and 874 DF,  p-value: < 2.2e-16
```

So we begin with

$$\widehat{\text{aaa}_{\text{df}}} = -0.00009159 + 0.3248 \text{cm10}_{\text{dif}} + 0.3006 \text{cm30}_{\text{dif}} - 0.004322 \text{ff}_{\text{dif}} + 0.03303 \text{tb03}_{\text{dif}} - 0.01703 \text{prime}_{\text{dif}}$$

Then remove: ff_{dif}

```

## remove ff_dif from the full model##
aaa_dif = 0.323cm10_dif + 0.304cm30_dif + 0.0386tb03_dif + p0.0183rime_dif
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.83e-05 2.16e-03 -0.04 0.96744
cm10_dif 3.23e-01 4.54e-02 7.11 2.4e-12 ***
cm30_dif 3.04e-01 4.95e-02 6.13 1.3e-09 ***
tb03_dif 3.86e-02 1.00e-02 3.85 0.00013 ***
prime_dif -1.83e-02 1.07e-02 -1.71 0.08826 .
---
Residual standard error: 0.0641 on 875 degrees of freedom
Multiple R-squared: 0.76, Adjusted R-squared: 0.759
F-statistic: 694 on 4 and 875 DF, p-value: <2e-16

```

```

> ##remove prime_dif from the full model#
> summary(lm(aaa_dif ~ cm10_dif + cm30_dif + tb03_dif + ff_dif))

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.62e-05 2.16e-03 -0.04 0.96456
cm10_dif 3.27e-01 4.55e-02 7.18 1.5e-12 ***
cm30_dif 2.99e-01 4.97e-02 6.01 2.8e-09 ***
tb03_dif 3.81e-02 1.08e-02 3.53 0.00043 ***
ff_dif -5.89e-03 5.96e-03 -0.99 0.32339
---
Residual standard error: 0.0642 on 875 degrees of freedom
Multiple R-squared: 0.76, Adjusted R-squared: 0.759
F-statistic: 692 on 4 and 875 DF, p-value: <2e-16

```

Since ff_dif is least significant in the full model, remove it.

aaa_dif = 0.323cm10_dif +
 0.304cm30_dif + 0.0386tb03_dif +
 p0.0183rime_dif

```

> ##remove tb03_dif from the full model#
> summary(lm(aaa_dif ~ cm10_dif + cm30_dif + prime_dif + ff_dif))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.83e-05 2.18e-03 -0.04 0.97
cm10_dif 3.55e-01 4.51e-02 7.86 1.1e-14 ***
cm30_dif 3.01e-01 5.01e-02 6.02 2.6e-09 ***
prime_dif -7.88e-03 1.07e-02 -0.74 0.46
ff_dif 5.32e-03 5.53e-03 0.96 0.34
---
Residual standard error: 0.0647 on 875 degrees of freedom
Multiple R-squared: 0.756, Adjusted R-squared: 0.755
F-statistic: 679 on 4 and 875 DF, p-value: <2e-16

> ##remove cm30_dif from the full model#
> summary(lm(aaa_dif ~ cm10_dif + tb03_dif + ff_dif+prime_dif))

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.00011 0.00221 -0.05 0.9603
cm10_dif 0.58501 0.01511 38.71 <2e-16 ***
tb03_dif 0.04213 0.01127 3.74 0.0002 ***
ff_dif -0.00738 0.00614 -1.20 0.2301
prime_dif -0.01533 0.01111 -1.38 0.1678
---
Residual standard error: 0.0655 on 875 degrees of freedom
Multiple R-squared: 0.75, Adjusted R-squared: 0.749
F-statistic: 658 on 4 and 875 DF, p-value: <2e-16

```

```

> ##remove cm10_dif from the full model#
> summary(lm(aaa_dif ~ cm30_dif + cm30_dif + tb03_dif + ff_dif))

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.000117 0.002226 -0.05 0.96
cm30_dif 0.636003 0.016654 38.19 < 2e-16 ***
tb03_dif 0.051462 0.010924 4.71 2.9e-06 ***
ff_dif -0.003424 0.006125 -0.56 0.58
---
Residual standard error: 0.066 on 876 degrees of freedom
Multiple R-squared: 0.746, Adjusted R-squared: 0.745
F-statistic: 856 on 3 and 876 DF, p-value: <2e-16

```



```

##remove cm10_dif from the full model##
> summary(lm(ddd_dif ~ cm30_dif + tb03_dif + prime_dif))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.00011   0.00222   -0.05    0.960
cm30_dif     0.63628   0.01658   38.39 < 2e-16 ***
tb03_dif     0.05436   0.01005    5.41 8.1e-08 ***
prime_dif   -0.01970   0.01103   -1.79    0.074 .
---
Residual standard error: 0.0659 on 876 degrees of freedom
Multiple R-squared:  0.746,  Adjusted R-squared:  0.746
F-statistic: 859 on 3 and 876 DF,  p-value: <2e-16

>
> ##remove cm30_dif from the full model##
> summary(lm(ddd_dif ~ cm10_dif + tb03_dif + prime_dif))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.000105  0.002207  -0.05  0.96222
cm10_dif     0.585989  0.015095   38.82 < 2e-16 ***
tb03_dif     0.036429  0.010224    3.56  0.00039 ***
prime_dif   -0.017531  0.010957   -1.60  0.10995
---
Residual standard error: 0.0655 on 876 degrees of freedom
Multiple R-squared:  0.75,  Adjusted R-squared:  0.749
F-statistic: 876 on 3 and 876 DF,  p-value: <2e-16

```

Remove prime_dif

ddd_dif = 0.0000921+0.324cm10_dif + 0.303cm30_dif + 0.033tb03_dif

```

> ##remove tb03_dif from the full model##
> summary(lm(ddd_dif ~ cm10_dif + cm30_dif + prime_dif))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.28e-05  2.18e-03  -0.04    0.97
cm10_dif     3.61e-01  4.46e-02   8.10 1.8e-15 ***
cm30_dif     2.97e-01  4.98e-02   5.95 3.8e-09 ***
prime_dif   -4.87e-03  1.02e-02  -0.48    0.63
---
Residual standard error: 0.0646 on 876 degrees of freedom
Multiple R-squared:  0.756,  Adjusted R-squared:  0.755
F-statistic: 906 on 3 and 876 DF,  p-value: <2e-16

>
> ##remove prime_dif from the full model##
> summary(lm(ddd_dif ~ cm10_dif + cm30_dif + tb03_dif))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.21e-05  2.16e-03  -0.04  0.96608
cm10_dif     3.24e-01  4.54e-02   7.14 2.0e-12 ***
cm30_dif     3.03e-01  4.95e-02   6.11 1.5e-09 ***
tb03_dif     3.30e-02  9.49e-03   3.48  0.00052 ***
---
Residual standard error: 0.0642 on 876 degrees of freedom
Multiple R-squared:  0.759,  Adjusted R-squared:  0.759
F-statistic: 922 on 3 and 876 DF,  p-value: <2e-16

```



```

> ##remove cm10_dif from the full model##
> summary(lm(aaa_dif ~ cm30_dif + tb03_dif))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.000114   0.002226   -0.05    0.96
cm30_dif     0.636747   0.016594   38.37 < 2e-16 ***
tb03_dif     0.048448   0.009498    5.10  4.1e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.066 on 877 degrees of freedom
Multiple R-squared:  0.745,  Adjusted R-squared:  0.745
F-statistic: 1.28e+03 on 2 and 877 DF,  p-value: <2e-16

> ##remove cm30_dif from the full model##
> summary(lm(aaa_dif ~ cm10_dif + tb03_dif))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.000108   0.002209   -0.05    0.9610
cm10_dif     0.586491   0.015106   38.83 < 2e-16 ***
tb03_dif     0.031110   0.009677    3.21  0.0014 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0655 on 877 degrees of freedom
Multiple R-squared:  0.749,  Adjusted R-squared:  0.749
F-statistic: 1.31e+03 on 2 and 877 DF,  p-value: <2e-16

```

Remove tb03_dif
aaa_dif =0.0000938+0.36cm10_dif + 0.297cm30_dif

```

> ##remove tb03_dif from the full model##
> summary(lm(aaa_dif ~ cm10_dif + cm30_dif))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.38e-05   2.18e-03   -0.04    0.97
cm10_dif     3.60e-01   4.45e-02    8.09  2.0e-15 ***
cm30_dif     2.97e-01   4.98e-02    5.96  3.7e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0646 on 877 degrees of freedom
Multiple R-squared:  0.756,  Adjusted R-squared:  0.756
F-statistic: 1.36e+03 on 2 and 877 DF,  p-value: <2e-16

```



```

> ## remove cm30_dif from the full model ##
> summary(lm(aaa_dif ~ cm30_dif))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.000121   0.002257   -0.05    0.96
cm30_dif     0.685316   0.013783   49.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.067 on 878 degrees of freedom
Multiple R-squared:  0.738,  Adjusted R-squared:  0.738
F-statistic: 2.47e+03 on 1 and 878 DF,  p-value: <2e-16

>
> ## remove cm30_dif from the full model ##
> summary(lm(aaa_dif ~ cm10_dif))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.000109   0.002221   -0.05    0.96
cm10_dif     0.615762   0.012117   50.82   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0659 on 878 degrees of freedom
Multiple R-squared:  0.746,  Adjusted R-squared:  0.746
F-statistic: 2.58e+03 on 1 and 878 DF,  p-value: <2e-16

```

**Remove cm30_dif
 aaa_dif =-0.000109+0.615762cm10_dif**



$$\widehat{\text{aaa}_{\text{df}}} = -0.00009159 + 0.3248 \text{cm10}_{\text{dif}} + 0.3006 \text{cm30}_{\text{dif}} - 0.004322 \text{ff}_{\text{dif}} + 0.03303 \text{tb03}_{\text{dif}} - 0.01703 \text{prime}_{\text{dif}}$$

$$\text{aaa_dif} = 0.323\text{cm10_dif} + 0.304\text{cm30_dif} + 0.0386\text{tb03_dif} + p0.0183\text{rime_dif}$$

$$\text{aaa_dif} = 0.0000921 + 0.324\text{cm10_dif} + 0.303\text{cm30_dif} + 0.033\text{tb03_dif}$$

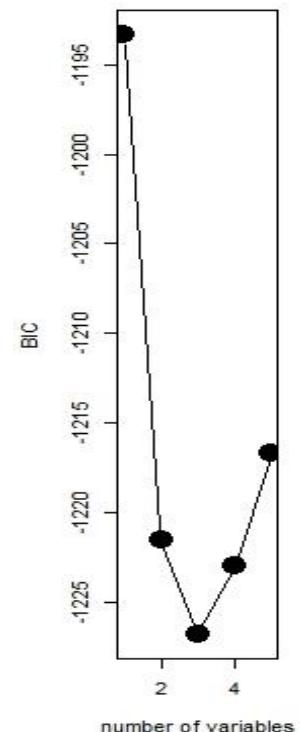
$$\text{aaa_dif} = 0.0000938 + 0.36\text{cm10_dif} + 0.297\text{cm30_dif}$$

$$\text{aaa_dif} = -0.000109 + 0.615762\text{cm10_dif}$$

```

coef(regfit.fwd, 3)
(Intercept) cm10_dif cm30_dif tb03_dif
-9.21e-05 3.24e-01 3.03e-01 3.30e-02
> coef(regfit.bwd, 3)
(Intercept) cm10_dif cm30_dif tb03_dif
-9.21e-05 3.24e-01 3.03e-01 3.30e-02
> coef(subsets, 3)
(Intercept) cm10_dif cm30_dif tb03_dif
-9.21e-05 3.24e-01 3.03e-01 3.30e-02

```



4. Hybrid Approaches

Hybrid versions of forward and backward stepwise selection.

4.2.2 Shrinkage Methods

Idea: Fit a model containing all p predictors using a technique that shrinks the coefficient estimates towards zero thus reduce the variance of the coefficient estimates.

1. Ridge Regression

The OLS method estimates $\hat{\beta}_i$ minimize

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2$$

Ridge regression coefficient estimates $\hat{\beta}_i^R$ minimize

$$RSS + \lambda \sum_{j=1}^k \beta_j^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^k \beta_j^2 \quad (34)$$

$\lambda \sum_{j=1}^k \beta_j^2$ is a shrinkage penalty. $\lambda \geq 0$ is a tuning parameter. It controls the bias-variance trade-off. Ridge regression is particularly useful to mitigate the problem of multicollinearity in linear regression, which commonly occurs in models with large numbers of parameters.



2. The LASSO (least absolute shrinkage and selection operator)

The LASSO *Ridge regression* coefficient estimates $\hat{\beta}^L$ minimize

$$RSS + \lambda \sum_{j=1}^k |\beta_j| = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^k |\beta_j| \quad (35)$$

- The lasso shrinks the coefficient estimates towards zero.
- The l_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when λ is sufficiently large.
- The lasso performs *variable selection*.



3. The Elastic Net

The Elastic Net *regression* coefficient estimates $\hat{\beta}^{EN}$ minimize

$$\begin{aligned} & RSS + \lambda_2 \sum_{j=1}^k \beta_j^2 + \lambda_1 \sum_{j=1}^k |\beta_j| \\ &= \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij} \right)^2 + \lambda_2 \sum_{j=1}^k \beta_j^2 + \lambda_1 \sum_{j=1}^k |\beta_j| \quad (36) \end{aligned}$$

- The Elastic Net overcomes some limitations of the LASSO.
- The elastic net method includes the LASSO and ridge regression.
- Double shrinkage leads to increased bias and poor predictions. (To improve the prediction performance, the inventor suggested multiplying the estimated coefficients by $(1 + \lambda_2)$).



4.2.3 The Dimension Reduction Methods

Let Z_1, Z_2, \dots, Z_M represent $M < k$ linear combinations of our original p predictors:

$$Z_m = \sum_{j=1}^k \phi_{jm} X_j, \quad m = 1, 2, \dots, M \quad (37)$$

Fitting (1'') $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon$

is converted to fitting

$$y_i = \theta_0 + \theta_1 Z_{i1} + \theta_2 Z_{i2} + \dots + \theta_M Z_{iM} + \varepsilon \quad (38)$$

If $\phi_{1m}, \phi_{2m}, \dots, \phi_{km}$ are chosen wisely, then fitting (38) using least squares can lead to better results than fitting (1'') using least squares.

From (37),

$$\sum_{m=1}^M \theta_m Z_{im} = \sum_{m=1}^M \theta_m \sum_{j=1}^k \phi_{jm} x_{ij} = \sum_{j=1}^k \sum_{m=1}^M \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^k \beta_j x_{ij}$$

Where

$$\beta_j = \sum_{m=1}^M \theta_m \phi_{jm} \quad (39)$$



1. Principal Components Regression

Principal Components Analysis (PCA)

An *unsupervised learning* approach for deriving a low-dimensional set of features.

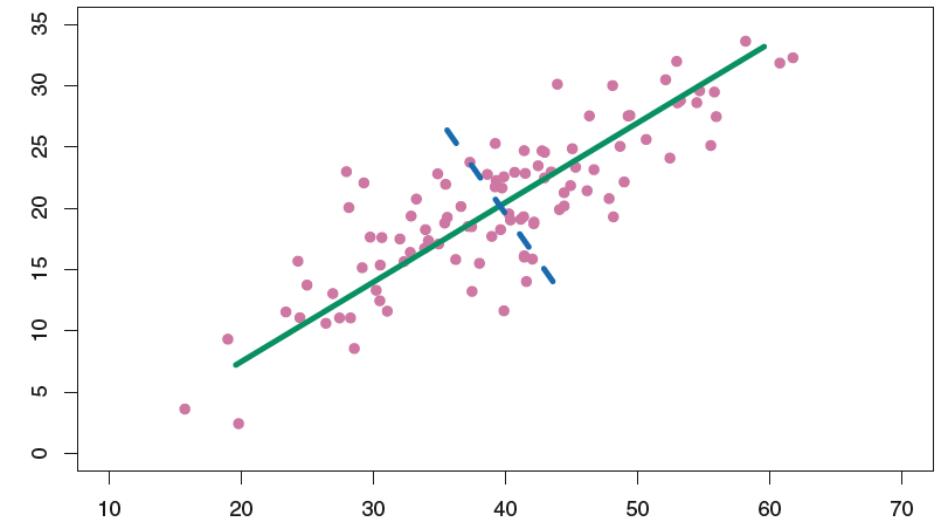
The *first principal component* (Z_1) direction of the data is that along which the observations *vary the most*.

The *second principal component* Z_2 is a linear combination of the variables that is uncorrelated with Z_1 , and has largest variance subject to this constraint...

Principal Components Regression Approach

The *principal components regression* (PCR) approach involves

- Constructing the first M principal components, Z_1, \dots, Z_M .
- Using these components as the predictors in a linear regression model that is fit using least squares.



2. Partial Least Squares (PLS): A supervised alternative to PCR.

Attempts to find directions that help explain both the response and the predictors.

- PLS computes the first direction Z_1 by setting each ϕ_{jm} equal to the coefficient from the simple linear regression of Y onto X_j . (One can show that this coefficient is proportional to the correlation between Y and X_j .)

Hence, in computing $Z_1 = \sum_{j=1}^k \phi_{j1} X_j$, PLS places the highest weight on the variables that are most strongly related to the response.

- To identify the second PLS direction we first *adjust* each of the variables for Z_1 , by regressing each variable on Z_1 and taking *residuals*. Using this *residual* to compute Z_2 . Z_2 is orthogonal to Z_1 .
- Repeat this procedure to identify multiple PLS components Z_1, \dots, Z_M .
- Then use least squares to fit a linear model to predict Y using Z_1, \dots, Z_M .

In practice PLS often performs no better than ridge regression or PCR.



5. Regression Diagnostic

5.1 Analysis of Residuals

Recall

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}^T x_i$$
$$\hat{\varepsilon} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \dots, \hat{\varepsilon}_n)^T = Y - HY = (I - H)Y, Cov(\hat{\varepsilon}) = \sigma^2(I - H)$$
$$Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}), Cov(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = -\sigma^2 h_{ij} \text{ for } i \neq j$$

$H = X(X^T X)^{-1}X^T$ is the *projection matrix (hat matrix)*.

Standardized Residuals

$$\hat{\varepsilon}'_i = \frac{\hat{\varepsilon}_i}{s\sqrt{1 - h_{ii}}} \quad (40)$$

$s^2 = \frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ is an unbiased estimate for σ^2 . $s = \sqrt{s^2}$.

The standardized residuals $\hat{\varepsilon}'_i$ have zero means and approximately unit variance if the regression model indeed holds.



Jackknife (Studentized) Residuals

If one ε_i is much larger than others, it can deflate all $\hat{\varepsilon}_j'$ by increasing s^2 . In this scenario, we omit the i -th observation (x_i, y_i) then refit the model, use this new model and denote the predicted response at x_i as $\hat{y}_{(-i)}$

$$\hat{\varepsilon}_i^* = \frac{y_i - \hat{y}_{(-i)}}{\sqrt{\widehat{Var}(y_i - \hat{y}_{(-i)})}}, \quad i = 1, 2, \dots, n \quad (41)$$

$\widehat{Var}(y_i - \hat{y}_{(-i)}) = s_{(-i)}^2 \left\{ 1 + \mathbf{x}_i^T \left(X_{(-i)}^T X_{(-i)} \right)^{-1} \mathbf{x}_i \right\}$ is an unbiased estimate for

$$Var(y_i - \hat{y}_{(-i)}) = Var(y_i) + Var(\hat{y}_{(-i)}) = \sigma_{(-i)}^2 \left\{ 1 + \mathbf{x}_i^T \left(X_{(-i)}^T X_{(-i)} \right)^{-1} \mathbf{x}_i \right\}$$

Note. It is not necessary to refit the model and compute, because

$$\hat{\varepsilon}_i^* = \hat{\varepsilon}'_i / \sqrt{\frac{n - p - (\hat{\varepsilon}'_i)^2}{n - p - 1}} \quad (42)$$



(a) Normal Probability Plot

A simple method of checking the normality assumption of ε is to contrast a Q-Q plot of the quantiles of studentized residuals versus standard normal quantiles.

Definition. The quantile-quantile (Q-Q) plot of a distribution G versus another distribution F plots the p th quantile y_p of G against the p th quantile x_p of F for $0 < p < 1$.

(b) Plot of Residuals against the Fitted Values

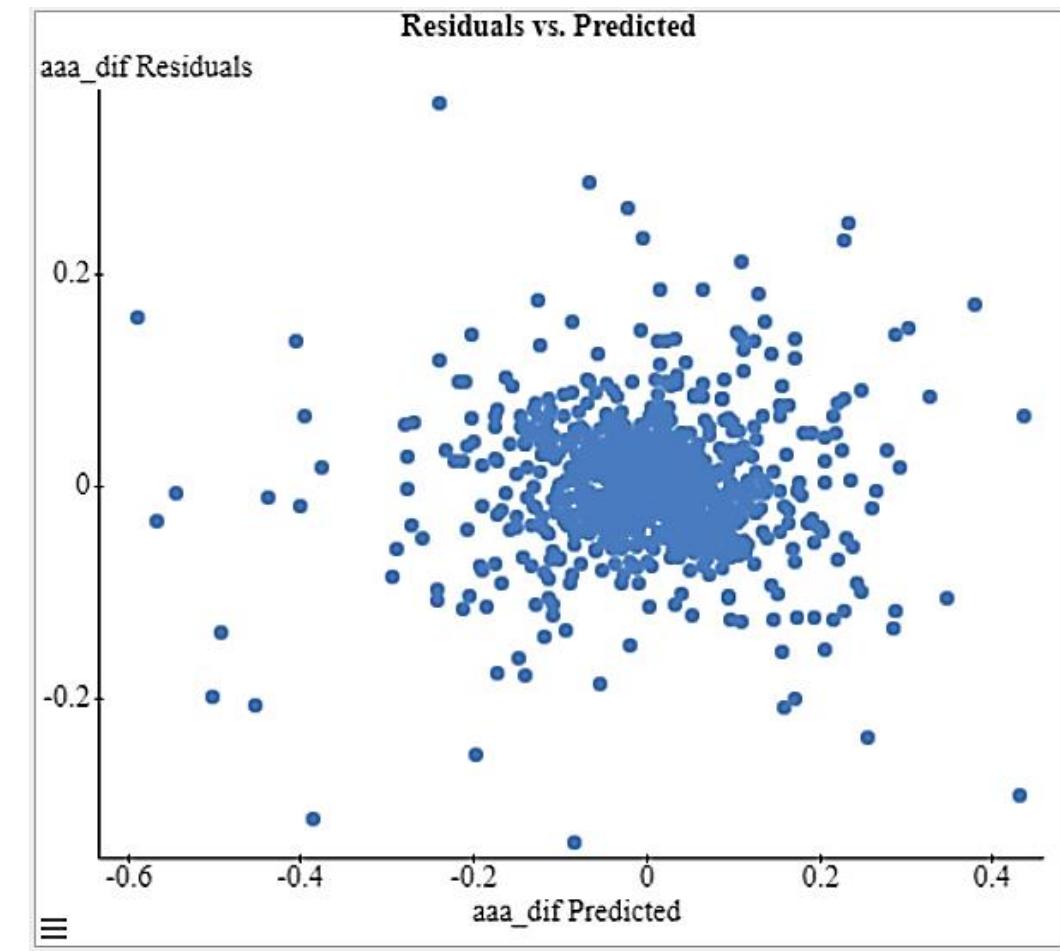
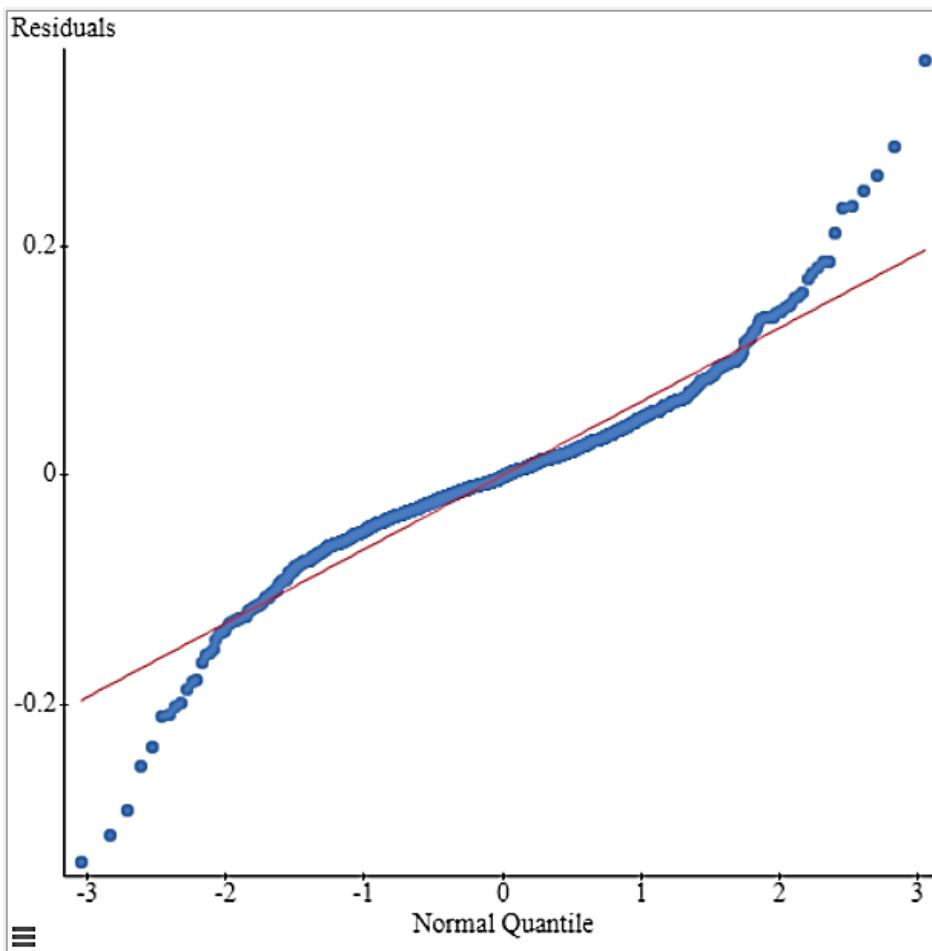
It is used to detect if the model is adequate. In general, if the plot of residuals spreads uniformly in a rectangular region around the horizontal axis, then there are no obvious model defects.

However, if the plot of residuals exhibits other patterns, one can use the patterns to determine how the assumed model can be amended.

(c) Plot of Residuals against the Regressors

In multiple regression, it is also helpful to plot the residuals against each regressor. The patterns displayed in these plots can be used to assess if the assumed relationship between the response and the regressor is adequate.



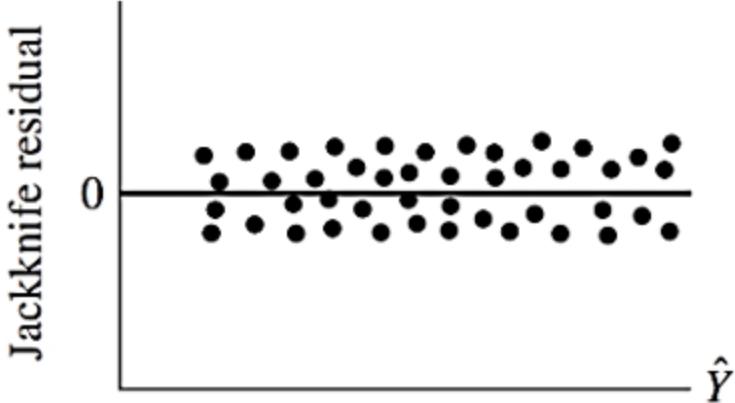


Multiple linear regression results:

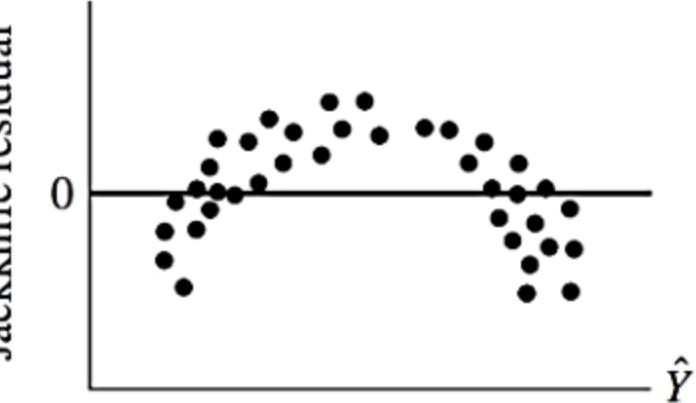
Dependent Variable: aaa_dif

Independent Variable(s): cm10_df, cm30_df, tb_df

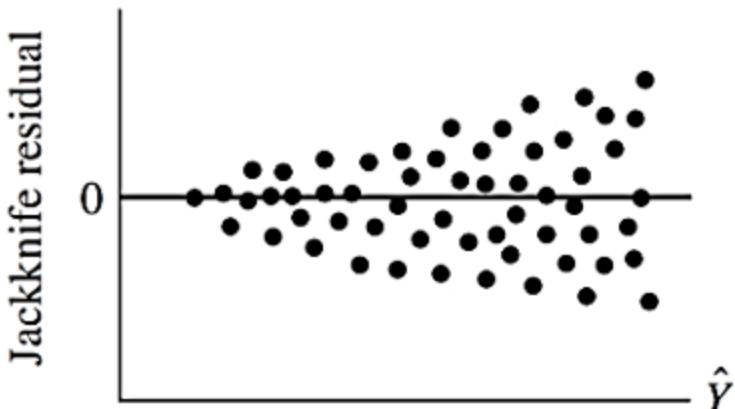
$$\text{aaa_dif} = -0.000092090414 + 0.32420932 \text{ cm10_df} + 0.30250823 \text{ cm30_df} + 0.033033319 \text{ tb_df}$$



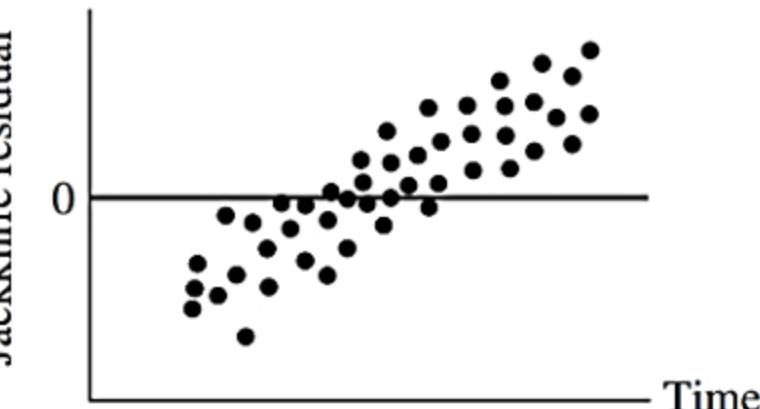
(a) Data satisfy all multiple regression assumptions



(b) Data depart from linearity



(c) Error variance increases with \hat{Y}



(d) Residuals plotted against time

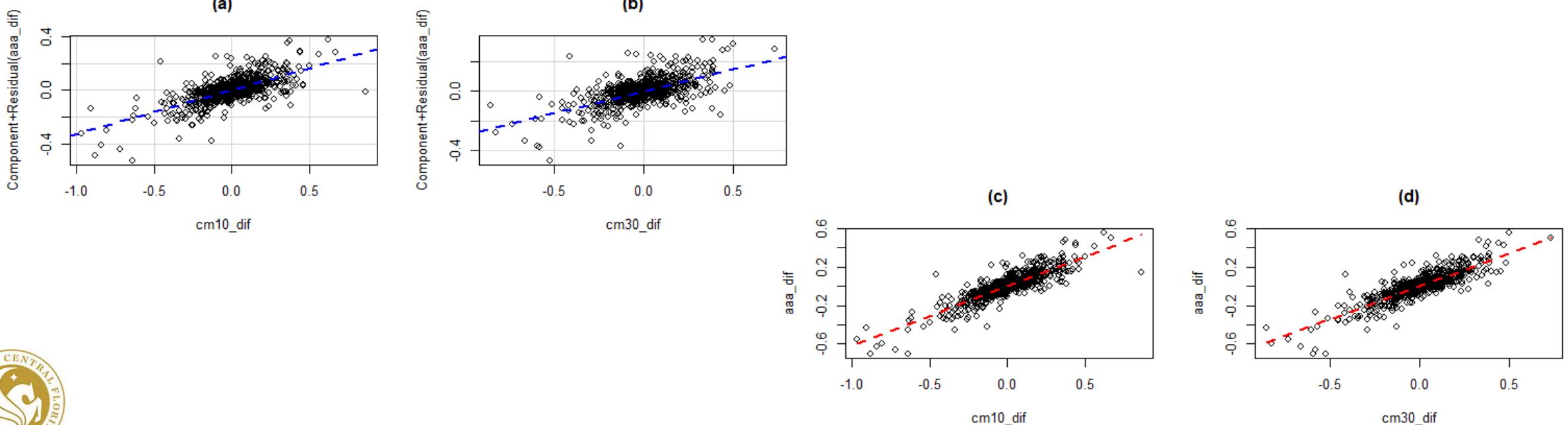
© Cengage Learning

FIGURE 14.6 Typical jackknife residual plots as a function of predicted value, \hat{Y} , or time of data collection for hypothetical data

(d) Partial Residual Plot

A partial residual plot is used to visualize the effect of a predictor on the response while removing the effects of the other predictors. The partial residual for the j th predictor variable is

$$Y_i - \left(\hat{\beta}_0 + \sum_{j' \neq j} X_{i,j'} \hat{\beta}_{j'} \right) = \hat{Y}_i + \hat{\epsilon}_i - \left(\hat{\beta}_0 + \sum_{j' \neq j} X_{i,j'} \hat{\beta}_{j'} \right) = X_{i,j} \hat{\beta}_j + \hat{\epsilon}_i,$$



5.2 Influence Diagnostics

“influential” observations

The hat matrix $H = X(X^T X)^{-1}X^T$ plays an important role in finding influential observations. The elements of H are $h_{ij} = x_i^T (X^T X)^{-1} x_j$.

Since $\hat{Y} = HY$, $\hat{y}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j$, h_{ii} is the *leverage* of the i th observation.

Cook's distance

$$D_i = \frac{(\hat{\varepsilon}'_i)^2}{p} \frac{Var(\hat{y}_i)}{Var(\hat{\varepsilon}_i)} = \frac{(\hat{\varepsilon}'_i)^2}{p} \frac{h_{ii}}{1 - h_{ii}} \quad (43)$$

where $\hat{\varepsilon}'_i$ is the standardized residual, measures the influence of y_i . Observations with large D_i (e.g., $D_i \geq 4/n$) are considered to be influential on the OLS estimate $\hat{\beta}$.



```

> cooks.distance(lm(aaa_dif ~ cm10_dif+cm30_dif+tb03_dif))
   1      2      3      4      5      6      7      8      9      10
1.279278e-05 5.531390e-05 1.269534e-05 6.808850e-06 5.086865e-04 5.310717e-04 8.053165e-06 1.184496e-04 2.164819e-05 1.265600e-04
   11     12     13     14     15     16     17     18     19     20
2.198915e-05 2.528427e-05 2.280647e-05 2.029455e-05 6.664379e-06 1.192433e-07 2.078078e-05 8.053165e-06 5.647577e-09 2.832737e-04
   151    152    153    154    155    156    157    158    159    160
1.130780e-03 5.450711e-04 3.473370e-03 2.379836e-03 4.313699e-02 1.839713e-05 2.940187e-02 5.183132e-01 4.599241e-01 1.292175e-05
   161    162    163    164    165    166    167    168    169    170
2.246808e-04 1.014359e-02 4.643555e-03 4.848428e-03 9.555724e-02 4.694924e-04 2.537676e-03 2.873925e-04 9.408186e-03 4.128451e-05

```

```

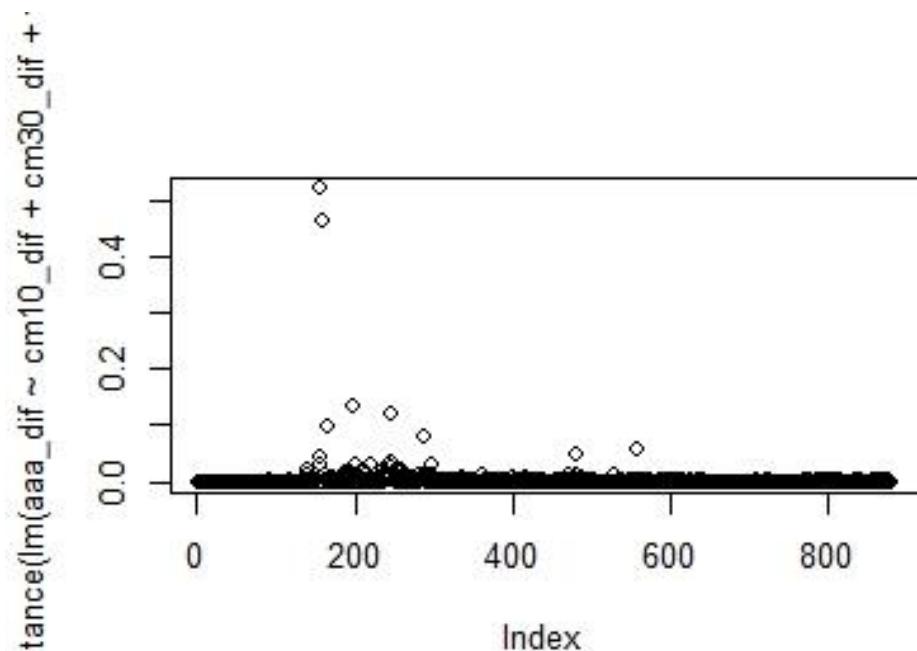
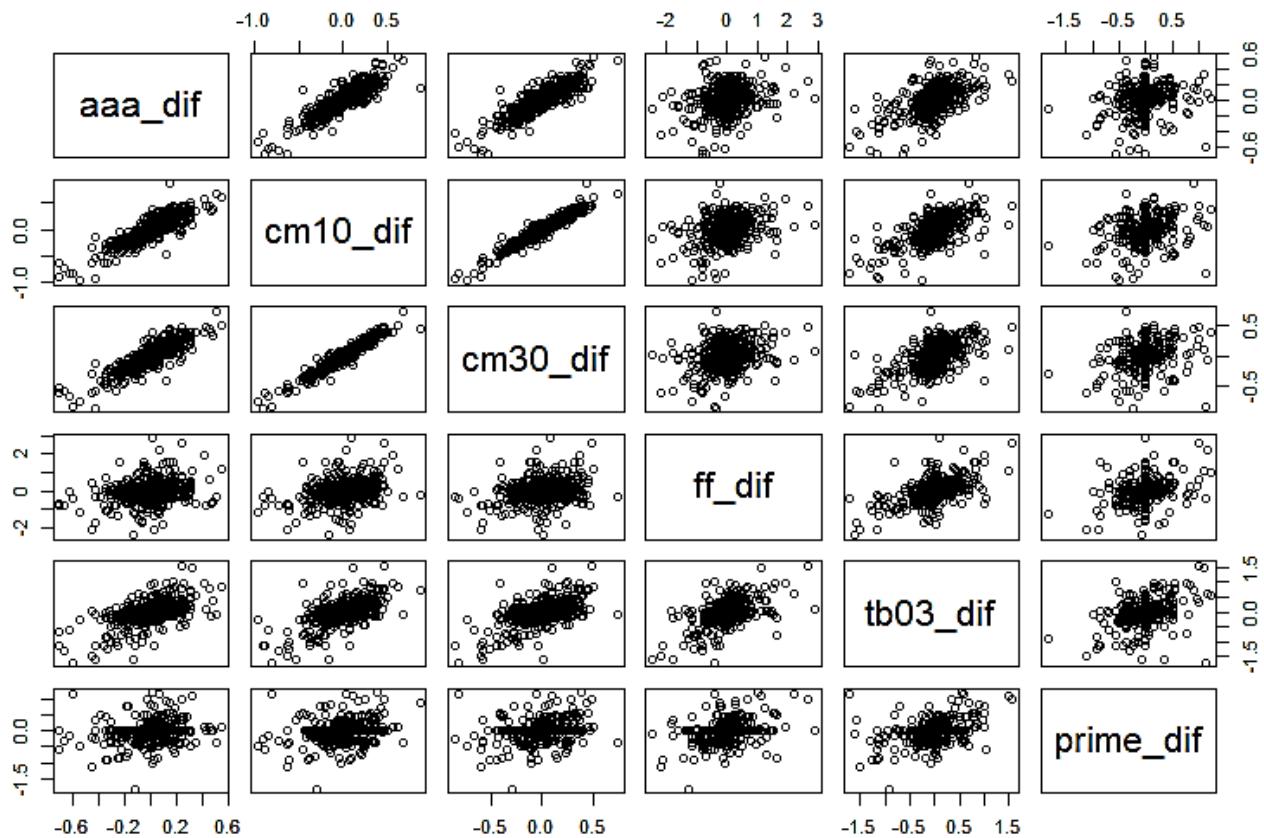
> order(-cooks.distance(lm(aaa_dif ~ cm10_dif+cm30_dif+tb03_dif)))

```

```

[1] 158 159 199 247 165 287 558 482 155 246 299 220 200 242 157 248 141 256 260 210 252 236 193 190 195 259 228 471 214 361 531
[32] 289 237 294 186 241 265 483 283 140 162 298 263 216 169 276 196 245 240 171 418 400 203 286 222 181 288 255 338 208 479 487
[63] 202 339 172 164 480 226 163 433 175 326 404 493 250 173 147 430 219 201 485 389 391 212 189 153 383 285 858 472 207 559 556

```



6 Resampling Methods

6.1 Cross-Validation

6.1.1 The Validation Set Approach

Randomly split the data set into two parts: training and validation

```
> ## 50% of the sample size ####  
>  
> smp_size <- floor(0.5*nrow(dat))  
>  
> ## set the seed to make your partition reproducible  
> set.seed(123)  
> train_ind <- sample(seq_len(nrow(dat)), size = smp_size)  
>  
> train <- dat[train_ind, ]  
> validation <- dat[-train_ind, ]  
>  
> train  
  month day year   ff tb03 cm10 cm30 discount prime aaa xxx  
415    1  23   85 8.19 7.71 11.34 11.45    8.00 10.50 11.97 153.31 9  
463   12  25   85 8.02 7.07  9.07  9.34    7.50  9.50  9.97 126.05 12  
179    7  16   80 8.98 8.02 10.21 10.22   11.00 11.50 11.09  84.36 15  
526    3  11   87 6.12 5.63  7.22  7.51    5.50  7.50  8.36  99.86 17  
195   11    5   80 13.99 12.96 12.50 12.27   11.00 14.50 12.96  88.66 18  
818   10   14   92 3.20 2.85  6.49  7.50    3.00  6.00  7.96  84.36 21  
118     5  16   79 10.25 9.58  9.33  9.24    9.50 11.75  9.51  90.00 22  
299   11    3   82 9.43 7.85 10.63 10.92    9.50 12.00 11.68 124.60 25
```

```
> validation  
  month day year   ff tb03 cm10 cm30 discount prime aaa xxx  
1      2  16   77 4.70 4.62 7.36 7.69    5.25 6.25 8.04 105.30  
2      2  23   77 4.74 4.67 7.39 7.75    5.25 6.25 8.08 105.24  
3      3  2   77 4.68 4.70 7.47 7.81    5.25 6.25 8.10 105.30  
4      3  9   77 4.63 4.64 7.49 7.82    5.25 6.25 8.12 105.26  
6      3  23   77 4.77 4.57 7.44 7.77    5.25 6.25 8.00 105.14  
7      3  30   77 4.74 4.59 7.46 7.79    5.25 6.25 8.10 105.00  
9      4  13   77 4.65 4.58 7.39 7.77    5.25 6.25 8.05 104.69  
12     5  4   77 5.15 4.65 7.44 7.79    5.25 6.25 8.04 104.40  
15     5  25   77 5.45 5.11 7.45 7.78    5.25 6.50 8.04 104.51  
17     6  8   77 5.31 5.04 7.37 7.70    5.25 6.75 7.98 104.37  
18     6  15   77 5.37 5.04 7.27 7.63    5.25 6.75 7.94 104.35  
21     7  6   77 5.35 5.06 7.32 7.61    5.25 6.75 7.93 103.50  
22     7  13   77 5.33 5.14 7.31 7.64    5.25 6.75 7.94 102.79  
25     8  3   77 5.80 5.37 7.42 7.71    5.25 6.75 8.00 102.71
```

Fit models on the training dataset, then evaluate their performance on the validation dataset (*model assessment*) by comparing their MSE.



6.1.2 *Leave-One-Out Cross-Validation* (LOOCV)

Take the i -th observation ($i = 1, 2, 3, \dots, n$) as the validation dataset and keep all the other $n-1$ observations as the training dataset.

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n MSE_i$$

For least squares linear or polynomial regression,

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2$$

Where \hat{y}_i is the i th fitted value from the original least squares fit, h_{ii} is the i -th diagonal element for the hat matrix $H = X(X^T X)^{-1} X^T$.



6.1.3 *k*-Fold Cross-Validation

Randomly dividing the dataset into k groups (folds), of approximately equal size. Treat the i -th ($i = 1, 2, 3, \dots, n$) fold as the validation set, fit the model on the other $k-1$ folds.

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k MSE_i$$

6.2 The Bootstrap

Obtaining data sets by repeatedly sampling observations *from the original data set.*



7. Extension to Stochastic Regressors

7.1 Minimum-Variance Linear Predictors

Assume Y and X_1, X_2, \dots, X_k are all random variables. Consider the problem of predicting Y by

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (44)$$

The coefficients $\beta_0, \beta_1, \dots, \beta_k$ can be determined by minimizing

$$S(\beta_0, \beta_1, \dots, \beta_k) = E\{[Y - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)]^2\}$$

From data $(y_t, x_{t1}, x_{t2}, \dots, x_{tk})$, the minimum-variance estimates for the coefficients are

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \left(\sum_{t=1}^n (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j) \right)^{-1} \begin{pmatrix} \sum_{t=1}^n (x_{t1} - \bar{x}_1)(y_t - \bar{y}) \\ \vdots \\ \sum_{t=1}^n (x_{tk} - \bar{x}_k)(y_t - \bar{y}) \end{pmatrix} \quad (45)$$

$$\hat{\beta}_0 = \bar{y} - (\hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \cdots + \hat{\beta}_k \bar{x}_k) \quad (46)$$



7.2 Inference in the Case of Stochastic Regressors

The estimates are consistent and asymptotically normal under certain regularity conditions. Crucial: $\mathbf{x}_t = (x_{t1}, x_{t2}, \dots, x_{tk})^T$ is uncorrelated with ε_t . (insufficient)
Stronger: $\{\mathbf{x}_t\}$ and $\{\varepsilon_t\}$ are independent. (too strong for financial time series)

Martingale difference assumption

$$E(\varepsilon_t | x_t, \varepsilon_{t-1}, x_{t-1}, \dots, \varepsilon_1, x_1) = 0 \quad \text{for all } t, \quad (47)$$

$$\lim_{t \rightarrow \infty} E(\varepsilon_t^2 | x_t, \varepsilon_{t-1}, x_{t-1}, \dots, \varepsilon_1, x_1) = \sigma^2 \text{ with probability 1.} \quad (48)$$

If we assume that for some nonrandom constants c_n such that $\lim_{n \rightarrow \infty} c_n = \infty$,

$$(X^T X)/c_n \text{ converges in probability to a nonrandom matrix } \neq 0 \quad (49)$$

then the distributional properties of the OLS estimates still hold asymptotically under some additional regularity conditions.

Rule of thumb:

Even when the \mathbf{x}_t are random, we can treat them as if they were nonrandom in constructing tests and confidence intervals for the regression parameters, by appealing to symptotic approximations under certain regularity conditions.



References

- D. Ruppert, D. S. Matteson, *Statistical and Data Analysis for Financial Engineering with R Examples* (2nd ed), Springer, 2015
- T. Z., Lai, H. Xing, *Statistical Models and Methods for Financial Markets*, Springer, 2008
- G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013
- Kleinbaum, Kupper, Nizam, Rosenberg, *Applied Regression Analysis and Other Multivariable Methods*, Cengage, 2014



Thank you!

