



北京航空航天大学

— 经济管理学院 —

BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

第 3 讲：多元线性回归模型

康雁飞

2021-02-23

Outline

1 多元线性回归模型

2 模型估计

3 模型推断

4 模型诊断

5 变量选择

6 非线性回归

7 作业

Outline

1 多元线性回归模型

2 模型估计

3 模型推断

4 模型诊断

5 变量选择

6 非线性回归

7 作业

多元线性回归模型

- 因变量只受单一自变量影响的情况非常少见。
- 通常影响一个变量的变量有多个。
- 一元线性回归 \Rightarrow 多元线性回归。

两个自变量

我们先假设有两个自变量：

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \dots, n.$$

其中 $\epsilon_i \sim N(0, \sigma^2)$.

- 我们需要找到一个面，尽可能的拟合样本点。
- 问题：如何寻找？

两个自变量

我们可以最小化：

$$f(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}))^2,$$

然后根据优化理论：

$$\frac{\partial f}{\partial \beta_0} = 0$$

$$\frac{\partial f}{\partial \beta_1} = 0$$

$$\frac{\partial f}{\partial \beta_2} = 0$$

多元线性回归模型

对于多变量的情况：

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i(p-1)} + \epsilon_i, \quad i = 1, 2, \dots, n$$

其中 $\epsilon_i \sim N(0, \sigma^2)$.

- $p - 1$ 个自变量。
- $p + 1$ 个待估参数。

Outline

1 多元线性回归模型

2 模型估计

3 模型推断

4 模型诊断

5 变量选择

6 非线性回归

7 作业

矩阵形式

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n(p-1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

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

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n(p-1)} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad 9$$

矩阵形式

我们可以最小化：

$$f(\beta_0, \beta_1, \beta_2, \dots, \beta_{p-1}) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i(p-1)}))^2$$

然后得到方程：

$$X^T X \beta = X^T y.$$

然后我们可以得到参数估计：

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

R 中的多元线性回归

例：某商业银行 25 家分行 2002 年的主要业务数据

```
## 'data.frame':    25 obs. of  5 variables:
## $ 不良贷款      : num  0.9 1.1 4.8 3.2 7.8 2.7 1.6 12.5 1 2.6 ...
## $ 各项贷款余额  : num  67.3 111.3 173 80.8 199.7 ...
## $ 本年累计应收贷 : num  6.8 19.8 7.7 7.2 16.5 2.2 10.7 27.1 1.7 9.1 ...
## $ 贷款项目个数  : num  5 16 17 10 19 1 17 18 10 14 ...
## $ 本年固定资产投资额: num  51.9 90.9 73.7 14.5 63.2 2.2 20.2 43.8 55.9 64.3
```

相关性

```
par(family = 'SimHei')  
corrplot::corrplot(cor(loan))
```

相关系数矩阵

##	不良贷款	各项贷款余额	本年累计应收贷	贷款项目个数
## 不良贷款	1.0000000	0.8435714	0.7315050	0.7002815
## 各项贷款余额	0.8435714	1.0000000	0.6787718	0.8484164
## 本年累计应收贷	0.7315050	0.6787718	1.0000000	0.5858315
## 贷款项目个数	0.7002815	0.8484164	0.5858315	1.0000000
## 本年固定资产投资额	0.5185181	0.7797022	0.4724310	0.7466458
##	本年固定资产投资额			
## 不良贷款		0.5185181		
## 各项贷款余额		0.7797022		
## 本年累计应收贷		0.4724310		
## 贷款项目个数		0.7466458		
## 本年固定资产投资额		1.0000000		

回归模型

```
loan.model <- lm(不良贷款~各项贷款余额+  
                本年累计应收贷+  
                贷款项目个数+  
                本年固定资产投资额,  
                data = loan)  
loan.model.summary <- summary(loan.model)
```

回归模型

```
##
## Call:
## lm(formula = 不良贷款 ~ 各项贷款余额 + 本年累计应收贷 +
##      贷款项目个数 + 本年固定资产投资额, data = loan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.9198 -0.9507 -0.2880  1.0334  3.1037
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.02164    0.78237  -1.306  0.20643
## 各项贷款余额    0.04004    0.01043   3.837  0.00103 **
## 本年累计应收贷  0.14803    0.07879   1.879  0.07494 .
## 贷款项目个数   0.01453    0.08303   0.175  0.86285
## 本年固定资产投资额 -0.02919    0.01507  -1.937  0.06703 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.779 on 20 degrees of freedom
```

估计误差项的方差

误差项方差 σ^2 可以通过 s_e^2 来估计:

$$s_e^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p} = \frac{\mathbf{e}^\top \mathbf{e}}{n - p}$$

注意: $E(s_e^2) = \sigma^2$.

估计误差项的标准差

```
loan.model.summary$sigma
```

```
## [1] 1.778752
```

通过方差分解我们有： $SST = SSE + SSR$.

拟合优度 (Goodness of Fit)，又叫测定系数 (Coefficient of Determination)：

$$R^2 = \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

拟合优度越大模型就越好吗？拟合优度小模型就一定差吗？

R^2 是 p 的单调递增函数。

调整的拟合优度

$$R_{\text{adj}}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

调整的拟合优度

```
loan.model.summary$r.squared
```

```
## [1] 0.797604
```

```
loan.model.summary$adj.r.squared
```

```
## [1] 0.7571248
```

Outline

1 多元线性回归模型

2 模型估计

3 模型推断

4 模型诊断

5 变量选择

6 非线性回归

7 作业

Gauss-Markov 定理

如果基本假设成立，最小二乘估计量是总体参数的线性最小方差无偏估计量。

$\hat{\beta}$ 的分布

我们可以得到：

$$\hat{\beta} \sim N\left(\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right).$$

所以：

$$E[\hat{\beta}_j] = \beta_j, \text{Var}[\hat{\beta}_j] = \sigma^2 C_{jj},$$

其中 $\mathbf{C} = (\mathbf{X}^\top \mathbf{X})^{-1}$ 。

$\hat{\beta}$ 的标准误 (Standard Error)

$\hat{\beta}$ 的标准误为:

$$\text{SE}[\hat{\beta}] = s_e \sqrt{(X^T X)^{-1}}$$

对每一个 $\hat{\beta}_j$,

$$\text{SE}[\hat{\beta}_j] = s_e \sqrt{C_{jj}}.$$

对 $\hat{\beta}$ 的推断

因为：

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 C_{jj}).$$

我们得到：

$$\frac{\hat{\beta}_j - \beta_j}{s_e \sqrt{C_{jj}}} \sim t_{n-p}.$$

$\hat{\beta}$ 的区间估计

对每一个 $\hat{\beta}_j$ ，其区间估计为：

$$\hat{\beta}_j \pm t_{\alpha/2, n-p} \cdot s_e \sqrt{C_{jj}}$$

$\hat{\beta}$ 的区间估计

```
confint(loan.model, level = 0.99)
```

##		0.5 %	99.5 %
##	(Intercept)	-3.24775491	1.20447538
##	各项贷款余额	0.01035187	0.06972683
##	本年累计应收贷	-0.07616275	0.37223054
##	贷款项目个数	-0.22172819	0.25078690
##	本年固定资产投资额	-0.07208059	0.01369486

- 给定未知的 \mathbf{x}_0 , 点预测为:

$$\begin{aligned}\hat{y}(\mathbf{x}_0) &= \mathbf{x}_0^\top \hat{\beta} \\ &= \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_{01} + \hat{\beta}_2 \mathbf{x}_{02} + \cdots + \hat{\beta}_{p-1} \mathbf{x}_{0(p-1)}.\end{aligned}$$

- 这是一个无偏估计:

$$\begin{aligned}E[\hat{y}(\mathbf{x}_0)] &= \mathbf{x}_0^\top \beta \\ &= \beta_0 + \beta_1 \mathbf{x}_{01} + \beta_2 \mathbf{x}_{02} + \cdots + \beta_{p-1} \mathbf{x}_{0(p-1)}\end{aligned}$$

置信区间

- 标准误为: $SE[\hat{y}(x_0)] = s_e \sqrt{x_0^\top (X^\top X)^{-1} x_0}$.
- 置信区间:
$$\hat{y}(x_0) \pm t_{\alpha/2, n-p} \cdot s_e \sqrt{x_0^\top (X^\top X)^{-1} x_0}.$$

置信区间

```
new.loan = data.frame(各项贷款余额 = c(100, 150, 200), 本年累计应收贷 = c(7, 20, 13),  
                      贷款项目个数 = c(10, 6, 5), 本年固定资产投资额 = c(40, 76,  
predict(loan.model, newdata = new.loan, interval = "confidence", level = 0.99)
```

```
##          fit          lwr          upr  
## 1 2.996112 1.631683 4.360541  
## 2 5.813459 2.089203 9.537715  
## 3 8.837354 3.896133 13.778574
```

- 计算预测区间，我们需要加一个随机误差项，因此标准误为：

$$SE[\hat{y}(x_0) + \epsilon] = s_e \sqrt{\mathbf{1} + x_0^\top (X^\top X)^{-1} x_0}.$$

- 预测区间为：

$$\hat{y}(x_0) \pm t_{\alpha/2, n-p} \cdot s_e \sqrt{\mathbf{1} + x_0^\top (X^\top X)^{-1} x_0}.$$

预测区间

```
predict(loan.model, newdata = new.loan, interval = "prediction", level = 0.99)
```

```
##           fit           lwr           upr  
## 1 2.996112 -2.2457345  8.237958  
## 2 5.813459 -0.4702792 12.097198  
## 3 8.837354  1.7640983 15.910609
```


单参数检验

我们要检验：

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

统计量

$$t = \frac{\hat{\beta}_j - \beta_j}{\text{SE}[\hat{\beta}_j]} = \frac{\hat{\beta}_j - 0}{s_e \sqrt{C_{jj}}}$$

在零假设成立的情况下，服从 t_{n-p} 分布。

单参数检验

```
loan.model.summary$coefficients
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	-1.02163976	0.78237236	-1.3058229	0.206433969
## 各项贷款余额	0.04003935	0.01043372	3.8374953	0.001028464
## 本年累计应收贷	0.14803389	0.07879433	1.8787378	0.074935421
## 贷款项目个数	0.01452935	0.08303316	0.1749825	0.862852686
## 本年固定资产投资额	-0.02919287	0.01507297	-1.9367689	0.067030078

回归模型的显著性检验

通过方差分解我们有： $SST = SSE + SSR$.

多元回归中，回归模型的显著性检验的零假设为：

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0.$$

也就是说零假设为：

$$Y_i = \beta_0 + \epsilon_i.$$

备择假设为：

$$H_1 : \text{至少存在一个 } \beta_j \neq 0, j = 1, 2, \cdots, (p - 1)$$

回归模型的 F 检验

我们有方差分析表：

来源	平方和	自由度	均方	F
回归	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$p - 1$	$SSR/(p-1)$	MSR/MSE
误差	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - p$	$SSE/(n-p)$	
总	$\sum_{i=1}^n (y_i - \bar{y})^2$	$n - 1$		

回归模型的 F 检验

F 统计量为:

$$F = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / (p - 1)}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - p)},$$

p -值为:

$$p = P(F_{p-1, n-p} > F)$$

回归模型的 F 检验

```
loan.model.summary$fstatistic
```

```
##      value      numdf      dendf  
## 19.70404    4.00000   20.00000
```

Outline

1 多元线性回归模型

2 模型估计

3 模型推断

4 模型诊断

5 变量选择

6 非线性回归

7 作业

是否符合模型假设？

作业

不正常点

- 除了检查模型假设之外，还应该注意“不正常点” (unusual observations)。
- 因为有时少量的不正常点对回归的影响是非常大的。

常见的不正常点：

- 1 异常点 (Outliers)
- 2 高杠杆点 (Points with high leverage)
- 3 有影响点 (Influential points)

异常点是没有被模型很好的拟合的点，通常是有很大的标准化残差（standardized residual）的观测值。

判断标准如：标准化残差的绝对值大于 3。

高杠杆点，即杠杆值很大点。杠杆值只受自变量取值的影响。

杠杆值（一元线性回归）

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

多高的杠杆值应该引起注意呢？

通常如果某个观测值的杠杆值大于二倍的平均杠杆值，被认为是高杠杆值。

杠杆值

```
hatvalues(loan.model)
```

```
##           1           2           3           4           5           6           7
## 0.14699817 0.37700617 0.11985647 0.10804892 0.14506018 0.16892163 0.16380175
##           8           9          10          11          12          13          14
## 0.51247764 0.14807691 0.11234041 0.10369810 0.13773405 0.13179714 0.18982442
##          15          16          17          18          19          20          21
## 0.28661792 0.10235522 0.21113050 0.07722630 0.11290567 0.35554736 0.57824701
##          22          23          24          25
## 0.10095553 0.05756815 0.26663244 0.28517194
```

杠杆值

```
hatvalues(loan.model) > 2 * mean(hatvalues(loan.model))
```

```
##      1      2      3      4      5      6      7      8      9     10     11     12     13
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE
##     14     15     16     17     18     19     20     21     22     23     24     25
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE
```

有影响点

有影响点是对回归模型有显著影响的观测值，一般同时具有大残差和高杠杆值的特点。可以通过库克距离（Cook's Distance） D_i 来计算。一般当 $D_i > \frac{4}{n}$ 时，判断为有影响点。

库克距离（Cook's Distance）

$$D_i = \frac{1}{p} r_i^2 \frac{h_i}{1 - h_i},$$

其中 r_i 表示标准化残差。

有影响点

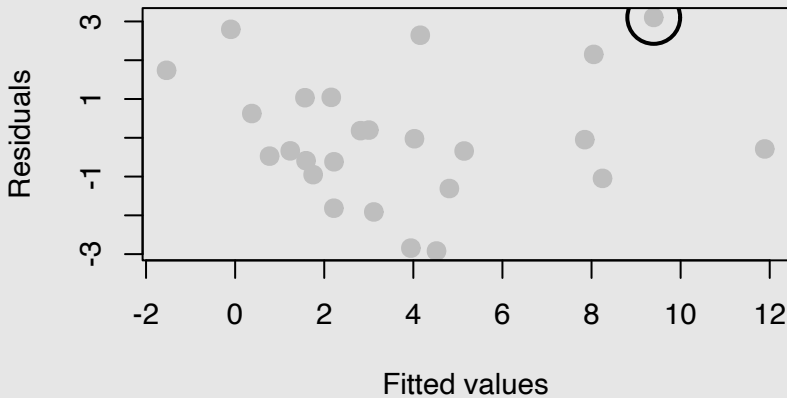
```
cooks.distance(loan.model)[8] > 4/nrow(loan)
```

```
##      8  
## TRUE
```

```
cooks.distance(loan.model)[21] > 4/nrow(loan)
```

```
##     21  
## FALSE
```

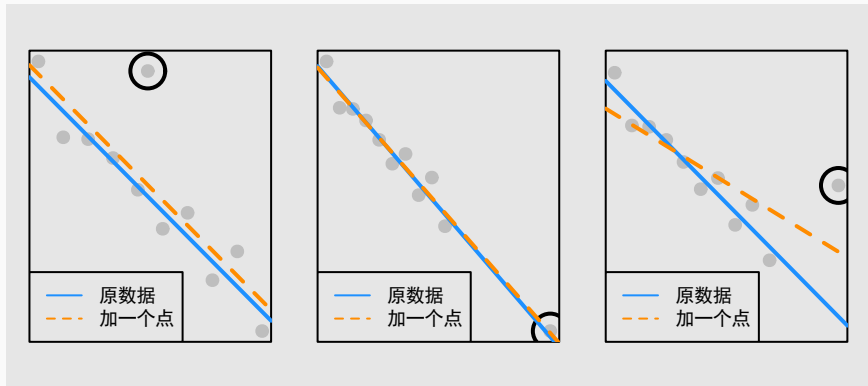

有影响点



去除有影响点?

```
##  
## Call:  
## lm(formula = 不良贷款 ~ 各项贷款余额 + 本年累计应收贷 +  
##      贷款项目个数 + 本年固定资产投资额, data = loan.new)  
##  
## Residuals:  
##      Min      1Q  Median      3Q      Max  
## -2.09473 -1.24993 -0.09849  0.98472  2.77202  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept)    -0.674348   0.676097  -0.997  0.331098  
## 各项贷款余额     0.039071   0.008884   4.398  0.000309 ***  
## 本年累计应收贷  -0.014968   0.087032  -0.172  0.865271  
## 贷款项目个数     0.039881   0.071174   0.560  0.581803  
## 本年固定资产投资额 -0.017078   0.013472  -1.268  0.220223  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 1.513 on 19 degrees of freedom
```

例子



Outline

1 多元线性回归模型

2 模型估计

3 模型推断

4 模型诊断

5 变量选择

6 非线性回归

7 作业

变量选择

变量选择通常是一个迭代过程，在每一步中，以某种预先定好的标准来决定是否加入或提出某个自变量。这个标准可以是：

- 假设检验，如 F 检验或 t 检验
- R_{adj}^2
- Akaike information criterion (AIC) 或 Bayesian information criterion (BIC)
- Mallows's C_p
- 其他标准

向后筛选法 (Backward Elimination)

在一开始假设模型中包含所有自变量，然后依据某种标准逐渐剔除不显著的变量，重复直到现存变量均不符合剔除条件。以 t 检验为例，向后筛选法为：

- 1 所有自变量 X_1, X_2, \dots, X_{p-1} 均包含在模型中；
 - ▶ 如果 t 检验都显著，则 X_1, X_2, \dots, X_{p-1} 均包含在模型中；
 - ▶ 若存在若干 t 检验不通过的参数，则先把 p -值最大的变量删除；
- 2 对剩余的 $p - 2$ 个变量做回归方程，删除 t 检验不通过中 p -值最大的变量；
- 3 重复以上步骤。直到模型中所有变量均通过 t 检验。

向后筛选法

```
step(loan.model, direction = "backward")
```

```
## Start: AIC=33.22
```

```
## 不良贷款 ~ 各项贷款余额 + 本年累计应收贷 + 贷款项目个数 +  
## 本年固定资产投资额
```

```
##
```

##	Df	Sum of Sq	RSS	AIC
## - 贷款项目个数	1	0.097	63.376	31.255
## <none>			63.279	33.217
## - 本年累计应收贷	1	11.168	74.447	35.280
## - 本年固定资产投资额	1	11.868	75.147	35.514
## - 各项贷款余额	1	46.594	109.873	45.011

```
##
```

```
## Step: AIC=31.26
```

```
## 不良贷款 ~ 各项贷款余额 + 本年累计应收贷 + 本年固定资产投资额
```

```
##
```

##	Df	Sum of Sq	RSS	AIC
## <none>			63.376	31.255
## - 本年累计应收贷	1	11.333	74.709	33.368
## - 本年固定资产投资额	1	12.147	75.523	33.639

向前选择法 (Forward Selection)

在一开始假设模型中没有变量，计算 Y 和每一个 X_i 的一元线性回归模型，选择 p -值最小的变量，然后依据某种标准逐渐加入一个变量，重复直到剩下的变量均不符合加入条件。

向前选择法

```
step(lm(不良贷款 ~ 1, data = loan),  
      scope = 不良贷款 ~ 各项贷款余额 + 本年累计应收贷 +  
              贷款项目个数 + 本年固定资产投资额,  
      direction = "forward")
```

```
## Start: AIC=65.16
```

```
## 不良贷款 ~ 1
```

```
##
```

##	Df	Sum of Sq	RSS	AIC
## + 各项贷款余额	1	222.49	90.164	36.069
## + 本年累计应收贷	1	167.30	145.351	48.007
## + 贷款项目个数	1	153.32	159.328	50.302
## + 本年固定资产投资额	1	84.06	228.591	59.326
## <none>			312.650	65.155

```
##
```

```
## Step: AIC=36.07
```

```
## 不良贷款 ~ 各项贷款余额
```

```
##
```

##	Df	Sum of Sq	RSS	AIC
## + 本年固定资产投资额	1	15.4555	74.709	33.368

逐步回归法 (Stepwise Regression)

- 前进法的问题：一旦某自变量进入模型后，它就永远留在模型中。然而，随着其他自变量的引入，一些先进入模型的变量的作用会变得不再显著。
- 向后法的问题：一旦某自变量被删除后，就永远不再进入模型。然而，随着其他自变量被删除，它的作用有可能会显著起来。

逐步回归法 (Stepwise Regression)

- 对于模型外部的变量，只要还能提供显著的解释作用，则可以再次进入模型。而在模型内部的变量，只要它的 t 检验不再显著，则可以从模型中删除。
- 方法：边进边退
- 起始：同前进法
- 结束：模型外所有变量均不能通过 t 检验

逐步回归法

```
step(lm(不良贷款 ~ 1, data = loan),  
      scope = 不良贷款 ~ 各项贷款余额 + 本年累计应收贷 +  
      贷款项目个数 + 本年固定资产投资额,  
      direction = "both")
```

```
## Start: AIC=65.16
```

```
## 不良贷款 ~ 1
```

```
##
```

##	Df	Sum of Sq	RSS	AIC
## + 各项贷款余额	1	222.49	90.164	36.069
## + 本年累计应收贷	1	167.30	145.351	48.007
## + 贷款项目个数	1	153.32	159.328	50.302
## + 本年固定资产投资额	1	84.06	228.591	59.326
## <none>			312.650	65.155

```
##
```

```
## Step: AIC=36.07
```

```
## 不良贷款 ~ 各项贷款余额
```

```
##
```

##	Df	Sum of Sq	RSS	AIC
## + 本年固定资产投资额	1	15.455	74.709	33.368

Outline

1 多元线性回归模型

2 模型估计

3 模型推断

4 模型诊断

5 变量选择

6 非线性回归

7 作业

为什么非线性回归?

- 有些时候自变量和因变量之间的关系是非线性的。
- 为了捕捉到这种非线性关系，常用的方法有：
 - 1 非线性关系线性化，再进行线性回归估计（如多项式模型，对数模型，指数模型，幂指数模型等）
 - 2 样条回归（Spline Regression）
 - 3 广义可加模型（Generalized Additive Models）

线性回归中拟合效果好，但外推预测效果不好。

- 样本噪音干扰过大，将部分噪音认为是特征，从而扰乱了建模规则。
- 参数太多，模型复杂度过高。

如何避免过拟合？交叉验证

- 1 交叉验证：比如将数据集分成两部分，一部分作为训练集；另一部分作为测试集。用训练集建立模型，然后将测试集带入模型，验证模型的表现。
- 2 模型的预测精度：可以用验证集的均方误差来测量。
- 3 由于诸多实际情况下并不存在足够的数据留作验证数据，这个时候通常用留一交叉验证法 (Leave-one-out)。

Outline

1 多元线性回归模型

2 模型估计

3 模型推断

4 模型诊断

5 变量选择

6 非线性回归

7 作业

1 对课上的不良贷款例子进行模型诊断（主要针对回归假设）

2 运用“污染数据”开展分析

1 计算和分析变量之间的相关关系

2 建立回归模型，并完成多元线性回归的建模、推断、诊断及预测过程。