LESSON 13: UNDERSTANDING REGRESSION

CHRIS QI

拟合的过程：

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon\,, \qquad \epsilon \sim N(0, \sigma_\epsilon)$$

$$e = Y - \hat{Y}$$

- 有n个样本量，也就是n个数据对 （x_i,y_i）
- 找到 截距和斜率，使得残差平方和最小

$$\sum_{i=1}^{n} e_i^2$$

$$Y_i = \beta_0 + \beta_1 X + \epsilon$$ 最小二乘法求线性回归系数

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

$$= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$f = \sum e_i^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial f}{\partial \hat{\beta}_0} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$\frac{\partial f}{\partial \hat{\beta}_1} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i$$

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

$$Y_i = \beta_0 + \beta_1 X + \epsilon$$

$$\beta_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

评估模型好坏的标准

STATISTIC	CRITERION
R-Squared	Higher the better *(> 0.70)*
Adj R-Squared	Higher the better
F-Statistic	Higher the better
Std. Error	Closer to zero the better
t-statistic	Should be greater 1.96 for p-value to be less than 0.05
AIC	Lower the better
BIC	Lower the better

R 方

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{Var(e)}{Var(y)}$$

调整后的R方

$$R^2_{adj} = 1 - \frac{MSE}{MST} \qquad R^2_{adj} = 1 - \left(\frac{\left(1 - R^2\right)(n-1)}{n-q} \right)$$

$$MSE = \frac{SSE}{(n-q)}$$

$$MST = \frac{SST}{(n-1)}$$

模型标准差和F统计量

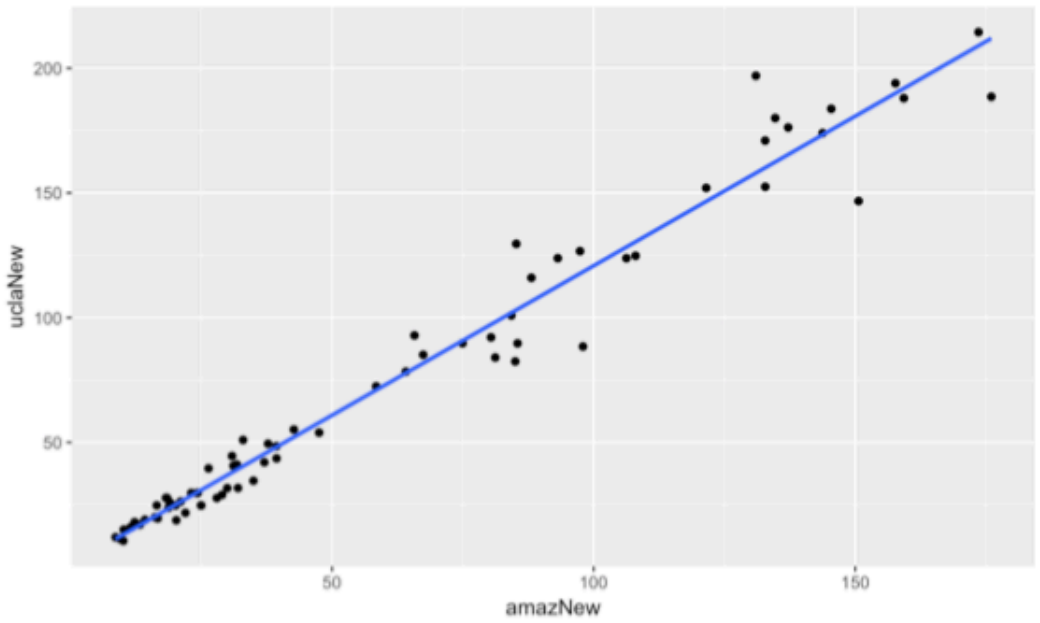
$$Std.\,Error = \sqrt{MSE} = \sqrt{\frac{SSE}{n-q}}$$

$$F - statistic = \frac{MSR}{MSE}$$

$$MSR = \frac{SST - SSE}{q-1}$$

回归系数的理解

```
ggplot(data = textbooks, aes(x = amazNew, y = uclaNew)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE)
```



```
> lm(uclaNew ~ amazNew, data = textbooks)

Call:
lm(formula = uclaNew ~ amazNew, data = textbooks)

Coefficients:
(Intercept)      amazNew
      0.929        1.199
```

$$\widehat{uclaNew} = 0.929 + 1.199 \cdot amazNew$$

```
> summary(mod)

Call:
lm(formula = uclaNew ~ amazNew, data = textbooks)

Residuals:
   Min     1Q Median     3Q    Max
-34.78  -4.57   0.58   4.01  39.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9290     1.9354    0.48     0.63
amazNew       1.1990     0.0252   47.60   <2e-16

Residual standard error: 10.5 on 71 degrees of freedom
Multiple R-squared:  0.97,   Adjusted R-squared:  0.969
F-statistic: 2.27e+03 on 1 and 71 DF,  p-value: <2e-16
```

```
> fitted.values(mod)
    1      2      3      4      5      6      7      8      9     10
34.44  38.27  39.30  14.74  17.97  13.12  24.98  20.90 128.32  16.83
   11     12     13     14     15     16     17     18     19     20
36.84 106.55  23.05  20.68 117.69  57.89  90.77 160.12 146.61 130.42
   21     22     23     24     25     26     27     28     29     30
14.92  23.64  15.60  27.25  38.27  35.64  20.29  46.19  39.03  40.46
   31     32     33     34     35     36     37     38     39     40
37.94 102.84  42.83 118.37  98.26  12.32  13.16 162.42 173.29 211.95
   41     42     43     44     45     46     47     48     49     50
181.53 175.26 209.03 158.00 189.99 165.40  30.84 191.91  28.59  26.16
   51     52     53     54     55     56     57     58     59     60
52.10  48.13 103.08 112.59  81.74 160.14  30.08  30.84 103.38  13.01
   61     62     63     64     65     66     67     68     69     70
79.74 101.96  11.24  70.97  97.29  77.77  45.34  25.16  48.10  32.55
   71     72     73
29.93  23.37  22.77
```

```
> residuals(mod)
       1        2        3        4        5        6        7
-6.77105  2.32413 -7.61701  1.25854  0.98322  1.82719 -0.28093
       8        9       10       11       12       13       14
-1.40433 -4.48287  0.17228 -5.20906  9.45100  4.61946  4.02348
      15       16       17       18       19       20       21
 8.98228 -3.99352 -1.04014 10.87962  5.39236 -5.62112  1.07869
      22       23       24       25       26       27       28
 2.31195  2.39526 -5.51705  2.32413 -6.69006 -0.34284  3.25873
      29       30       31       32       33       34       35
 2.05677 10.48996  6.55786 -20.39409 -8.23406 -29.95115 -14.26390
      36       37       38       39       40       41       42
-1.06948  1.84122 17.60753  0.71458 -23.37321 -34.78455  8.48623
      43       44       45       46       47       48       49
 5.47235 39.00185  4.01249 10.85401 -6.14405 -3.90591  1.11007
      50       51       52       53       54       55       56
 0.08405  3.02765 -4.57365 26.51611 11.24803  3.37834 -7.66436
```

用估计出来的模型做预测

- predict(lm, data)
- 得到每一个新数据的预测值
- 机器学习的基础