线性模型（Linear Regression）

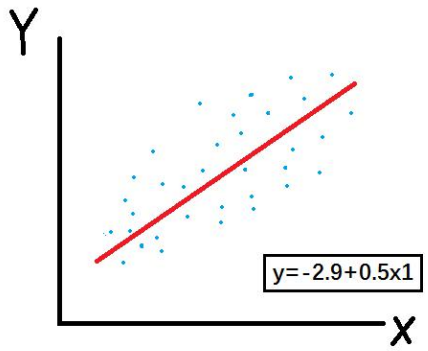
- 特点：目标变量必须满足数值变量/连续变量
- 根据自变量的多少，我们细分为：单变量模型/多变量模型
- 得到的结果是一个预测的数值
- 模型评价：预测值和真实值的接近程度
- 需要注意的问题：数据量、异常值、非线性关系、交互作用

单变量模型

- 目标变量 Y 为连续变量（价格、销售额）
- 自变量 X 只有一个
- 假设自变量和目标变量之间是线性关系

$$y = \beta_0 + \beta_1 x_1$$

(预测值) (截距) (系数) (自变量)



单变量模型原理：OLS 普通最小二乘法

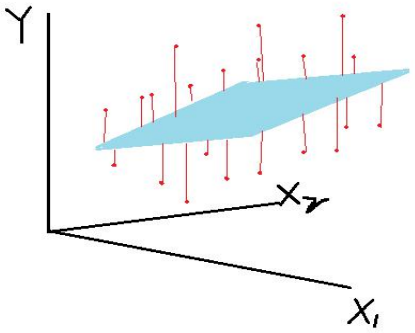
$$\min \sum_{t=1}^{T} (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t)^2$$

使残差平方和 RSS 最小，求偏导等于 0，达到对参数估计的目的。

多变量模型

多个自变量

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$



- 自变量影响大小：看系数 β 的大小，正为正相关，负为负相关
- 预测公式不代表因果关系，仅代表具有相关关系

模型评价：

一、拟合优度检验

①R-Square 属于[0,1]，越大越好

$$R^2=\frac{\sum(\hat{y}_t-\bar{y})^2}{\sum(y_t-\bar{y})^2}=\frac{ESS}{TSS}$$

②Ajusted R-Square $\overline{R}^2$ 放入多一个变量，则会增加变量的相关程度，由于自变量的增加产生的

拟合程度好的模型：R^2 大，加入一个变量后 R^2 更大，$\overline{R}^2 > R^2$，则该增加变量不是多余变量

②RMSE（Root Mean squared error）（均方根误差亦称标准误差）

$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi-\hat{y}i)^2}$ 越小模型越好，无上限，难以判断大小，与 $yi-\hat{y}i$ 比较？

二、假设检验

（1）t 检验，t>t α 显著，经验：若 t 值（如 0.4）位于[-1.96,1.96]则明显不显著

（2）F 检验（F>F α 拒绝原假设，回归模型有显著意义，意味着解释变量联合起来对 Y 有显著影响）

（3）置信区间法

（4）p 值检验（和 0.1/0.05/0.01 比较，p 值越小该变量越显著，影响因变量的可能性更高）

线性回归模型要注意的几个问题

- 数据量（sample size）：越多越好，数据条数是自变量的 10 倍以上

- 异常值（outlier）：线性回归对异常值很敏感（OLS），应当删除异常值
- （x 与 y 之间）非线性关系（non-linear relationship）：需要对自变量做出调整

—log,exp,square root

—x^2 ,x3,x4

—模型会变得很复杂，可能会产生过拟合现象

- （x 与 x 之间）交互作用（interaction effects）

比如相互促进作用，在模型中添加一个交互变量 X1 • X2，R^2 判断添加后的拟合效果

- 线性回归模型基本假设不满足时：横截面（面板）数据可能出现异方差（随机误差项的方差不为常数），影响模型效果，需要检验（散点图、残差图、GQ、BP、White）、修正异方差（WLS、对数变换）；时间序列数据可能出现自相关（随机项协方差不为零），检验（残差散点图、DW）、修正（广义差分法）

经验：分类数据要做独热编码，数值变量预处理归一化之后 RMSE、R^2 不变，系数大小改变

数据清洗与准备

1. 缺失数据忽略 or 补充

补充：全体均值法、临近值策略（KNN），基于专业知识补充

2. 异常值

Mean+/- 3std，boxplot

3. 数据预处理

数值变量：归一化数理（min-max、z-score）

分类变量：标签编码、独热编码

实现（R 语言）

2.使用函数lm()以mpg为因变量，horsepower为自变量，建立简单线性回归模型。使用函数summary()展示结果，并回答一下问题。

```
lm.fit<-lm(mpg~horsepower,data=Auto)

summary(lm.fit)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

2 多元线性回归模型

函数 `lm()` 也可以用来拟合一个多元线性回归模型，基本的用法是 `lm(y ~ x1 + x2 + x3)` ，表示以 `y` 为因变量，以 `x1, x2, x3` 为自变量建立回归模型。

```
> lm.fit <- lm(medv ~ lstat + age, data = Boston)
> lm.fit
```

```
##
## Call:
## lm(formula = medv ~ lstat + age, data = Boston)
##
## Coefficients:
## (Intercept)        lstat          age
##    33.22276     -1.03207      0.03454
```

7.使用函数lm()以mpg为因变量，除了变量"name"之外的所有变量为自变量，建立多元线性回归模型。使用函数summary()展示结果，并回答一下问题。

```
lm.fit2<-lm(mpg~.-name,data=Auto)
summary(lm.fit2)
```

可以看到 `ShelveLoc` 的数据形式是 `factor` 。现在，我们把除了 `Sales` 的变量都当成是自变量，

```
> lm.carseats <- lm(Sales ~ ., data = Carseats)
> summary(lm.carseats)
```

3 交互项

在R中，使用函数 `lm()` 可以很容易在线性模型中加入交互项，交互项用 `x1 : x2` 来表示，

```
> lm.fit <- lm(medv ~ lstat + age + lstat:age, data = Boston)
> summary(lm.fit)
```