**线性模型（Linear Regression）**

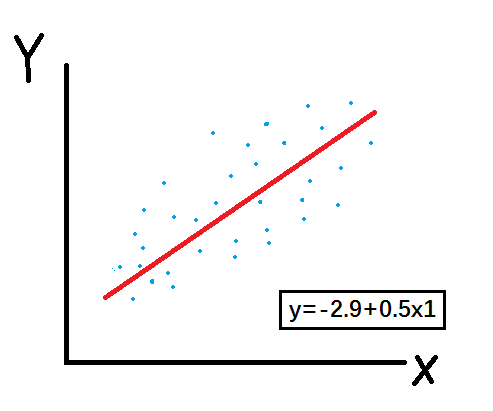
* 特点：目标变量必须满足数值变量/连续变量
* 根据自变量的多少，我们细分为：单变量模型/多变量模型
* 得到的结果是一个预测的数值
* 模型评价：预测值和真实值的接近程度
* 需要注意的问题：数据量、异常值、非线性关系、交互作用

**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



(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

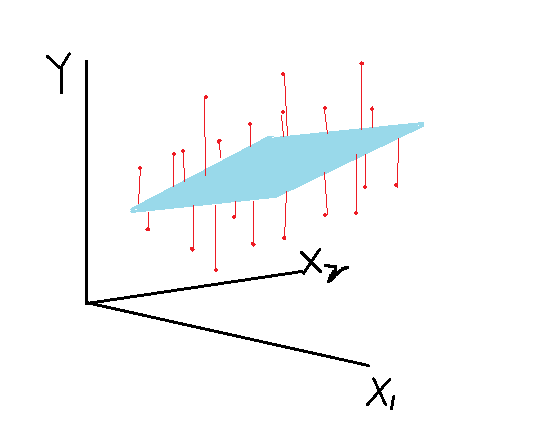


使残差平方和RSS最小，求偏导等于0，达到对参数估计的目的。

**多变量模型**

多个自变量





* 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



②Ajusted R-Square 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：大，加入一个变量后更大，>，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）

越小模型越好，无上限，难以判断大小，与比较？

1. 假设检验
2. t检验，t>tα显著，经验：若t值（如0.4）位于[-1.96,1.96]则明显不显著
3. F检验（F>Fα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）
4. 置信区间法
5. 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

—x²,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，R²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、R²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

1. 异常值

Mean+/- 3std，boxplot

1. 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）**

