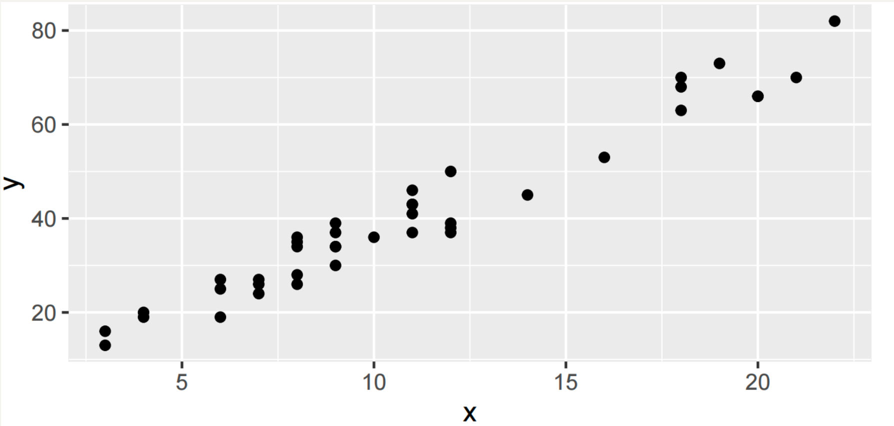
回归分析第一节课（复盘）

回归与建模

问题引入：

1. 外卖公司想要探究外卖送货时间与接单量之间的关系
2. 

   其中y坐标为外卖送达时间，x坐标为接单量。
3. 从上面的图中，我们可以发现接单量越大，送货时间就会越长，但是我们实际上无法使用一个精确的表达式进行表达，原因是外卖送达时间还有可能受到其他因素的影响，相当于外卖送达时间是受多变量控制的，而我们在上面的例子中控制的只是单一的变量，因此我们将会提炼出下面的表达式。
4. 一般回归模型
   - 第一种表达式（这种表达式中其实已经包含了ϵ的条件期望是0）

     $y = E(y|x) + \epsilon$
   - 第二种表达式

     $y = f(x) + \epsilon, \quad E(\epsilon|x) = 0$
   - 在表达式中$f(x) = E(y|x)$是完全由x确定的部分，称之为回归函数，ϵ表示的是随机误差
5. 在一般回归模型中的两个应该记住的点
   - $E(\epsilon|x) = 0$

     $Prove:$

$$\begin{aligned}E(\epsilon|x) &= E(y - f(x)|x) \\ &= E(y|x) - E(f(x)|x) \\ &= f(x) - f(x) \\ &= 0\end{aligned}$$

   - $Cov(\epsilon, x) = 0$

     $Prove:$

$$\because \quad E(\epsilon|x) = 0$$

$$\therefore \quad E(\epsilon) = \sum_{i=1}^{n} x_i E(\epsilon|x_i) = E(E(\epsilon|x_i)) = 0$$

$$\begin{aligned}
\therefore \quad Cov(\epsilon, x) &= E(\epsilon x) - E(\epsilon)E(x) \\
&= E(\epsilon x) \\
&= E(E(\epsilon x|x)) \\
&= E(xE(\epsilon|x)) \\
&= E(0) \\
&= 0
\end{aligned}$$

6. 简单线性模型

    - 在本例子中，我们可以观察到样本点是散布在某条直线附近的，有线性趋势，因此我们可以假设

    $$y = \beta_0 + \beta_1 x + \epsilon$$

    - 对于该模型来说，简单二字表示的是只有一个回归变量或者预测变量，对于该模型为一个线性模型，这里只是一个假设，或者一个经验模型，在实际问题中，需要在建模中对这一个假设进行适用性检验。

    $$E(y|x) = \mu_{y|x} = E(\beta_0 + \beta_1 x + \epsilon|x) = \beta_0 + \beta_1 x$$

    - 简单线性回归模型的两种假设
        - 第一种假设

        $$y = \beta_0 + \beta_1 x + \epsilon, \quad E(\epsilon|x) = 0$$

        - 第二种假设

        $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad, i = 1, \cdots, n$$

7. 多元线性回归模型

    - 在实际问题里面，可以考虑响应变量y与k个变量$x_1, x_2, \cdots, x_k$之间的关系，如果假设

    $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

    则称之为多元线性回归模型，注意这里说的线性表示的是回归函数是关于$\beta_0, \cdots, \beta_k$是线性的，而非y是关于x的线性函数，因为显然地，下列的多项式回归模型也可以表示为线性回归的范畴

    $$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

    - 多元线性回归模型的假设

    $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \quad E(\epsilon|x_1, \cdots, x_k) = 0$$

数据的搜集

- 回溯研究——表示的是通过观察过去已有的数据，这里通常会遇到一个缺失值的问题
- 观察研究——进行抽样调查之类的
- 实验设计——比如观察产品浓度和回流率之间的关系

回归分析的目标

- 数据描述
- 参数估计，推断与解释
- 预测和估计
- 控制

回归模型流程

- 具体流程如下：
    - 首先先结合研究问题的背景以及可获得的数据，设定一个初始模型，这里对数据的探索性分析如散点图或者散点图矩阵，有助于设定一个合适的模型。
    - 根据数据对模型参数进行估计，主要方法是最小二乘法或者极大似然估计法。
    - 估计完模型之后，我们需要对模型的充分性进行检验，例如模型形式，变量选择，异常点检测，模型假设是否成立
    - 最后是模型验证和模型部署应用
- 流程图如下：

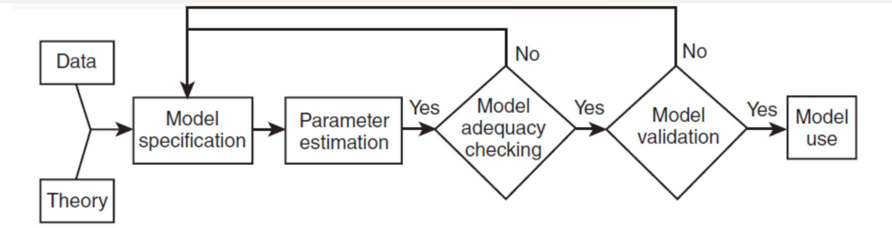


Figure 1.8 Regression model-building process.

简单线性回归

简单线性回归模型

- 定义：所谓的简单线性回归模型，是只有一个自变量x（也称为解释变量，预测变量）对应的变量y称之为因变量（或者被称为解释变量，响应变量），y与x之间的关系的假设为

$$y = \beta_0 + \beta_1 x + \epsilon$$

这里$E(y|x) = \beta_0 + \beta_1 x$称为总体回归线（总体回归函数），其中$\beta_0, \beta_1$为未知常数，称为回归系数，$\beta_0$为截距，$\beta_1$为斜率，$\epsilon$为随机误差项，满足$E(\epsilon|x) = 0$，假如我们再假设

$$Var(\epsilon|x) = \sigma^2$$

则$Var(y|x) = \sigma^2$，称为同方差假定，此时模型称为经典线性模型（经典线性模型就是简单线性模型加上同方差假定）$F_{y|x}$的期望和x有关，但是方差不随着x变换

参数的最小二乘估计

- 假设$(y_i,x_i),i=1,\cdots,n$为随机样本，下面将会研究如何使用这些数据估计模型中的未知参数$\hat{\beta}_0$,$\hat{\beta}_1$,对应的$\hat{\beta}_0+\hat{\beta}_1 x$称为样本回归线，回归线的确定的考虑为:这条线应该尽可能靠近这些样本点，本来应该考虑的是点(x_i,y_i)到这条线的距离是$|y_i-\hat{\beta}_0-\hat{\beta}_1 x_i|$，但是由于考虑到绝对值在计算上的困难程度，我们将引入残差平方和

$$
\begin{aligned}
S(\hat{\beta}_0,\hat{\beta}_1) &= \sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2 \\
\left.\frac{\partial S}{\partial \hat{\beta}_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} &= -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)\Big|_{\hat{\beta}_0,\hat{\beta}_1} \\
\left.\frac{\partial S}{\partial \hat{\beta}_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} &= 2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)(-x_i)\Big|_{\hat{\beta}_0,\hat{\beta}_1} \\
(\hat{\beta}_0,\hat{\beta}_1) &= argminS(\hat{\beta}_0,\hat{\beta}_1)
\end{aligned}
$$

$$
\begin{cases}
\left.\frac{\partial S}{\partial \hat{\beta}_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 0 \\
\left.\frac{\partial S}{\partial \hat{\beta}_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 0
\end{cases}
$$

通过将上述的方程组进行展开并且化简我们可以得到

$$
\begin{cases}
\left.\frac{\partial S}{\partial \hat{\beta}_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 0 \\
\left.\frac{\partial S}{\partial \hat{\beta}_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 0
\end{cases}
\Longrightarrow
\begin{cases}
\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)\Big|_{\hat{\beta}_0,\hat{\beta}_1} = 0 \\
\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)(x_i)\Big|_{\hat{\beta}_0,\hat{\beta}_1} = 0
\end{cases}
\Longrightarrow
\begin{cases}
n\hat{\beta}_0+(\sum_{i=1}^{n}x_i)\hat{\beta}_1 \\
(\sum_{i=1}^{n}x_i)\hat{\beta}_0+(\sum_{i=1}^{n}x_i^2)\hat{\beta}_1
\end{cases}
$$

化简上面的表达式我们能够得到下面的解

$$
\begin{cases}
\hat{\beta}_0 = \bar{y}-\hat{\beta}_1\bar{x} \\
\hat{\beta}_1 = \frac{\sum_{i=1}^{n}y_i x_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{S_{xy}}{S_{xx}}
\end{cases}
$$

计算残差

$$
e_i = y_i-\hat{y}_i = y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i), i=1,2,\cdots,n
$$

- 下面为课上的一些代码（加以注释）

```r
# 首先先进行加载包
library("tidyverse")
library("readxl")
# 首先先进行读取对应的数据
da2_1 = readxl::read_xls("DataSets4e/Chapter 2/Examples/data-ex-2-1
(Rocket Prop).xls")
# 首先对数据框的列进行取名字
colnames(da2_1) = c("ID","y","x")

# 下面开始进行绘制图像
# 首先先绘制出对应的散点图
p = ggplot(da2_1) + geom_point(aes(x = x,y=y))
```

```r
# 下面开始根据公式进行计算简单线性回归模型的总体回归线的参数
x = da2_1$x
y = da2_1$y
# 首先先进行计算参数beta_1
S_xy = sum((x-mean(x)*(y-mean(y))))
S_xx = sum((x-mean(x))^2)
beta_1 = S_xy/S_xx
beta_1

# 接下来进行计算beta_0
beta_0 = mean(y) - beta_1*mean(x)
# 添加辅助线
p + geom_abline(slope = beta_1, intercept = beta_0, colour = "red")

# 接下来计算出y_hat拟合值以及res残差
y_hat = beta_0 + beta_1*x
res = y - y_hat
res

# 为数据框添加上新的三列
da2_1 = dplyr::mutate(da2_1,residuals = res)
# 将点加到图上去
# ggplot(da2_1) + aes(x = x,y = residuals) + geom_point()
# 这里也能够使用另外的函数进行代替
ggplot(da2_1) + geom_point(aes(x = x,y = residuals))
```

```r
# 我们也可以直接调用函数实现OLS估计
# OLS表示的是ordinary least squares 普通最小二乘法
fit = lm(y~x,data = da2_1)
fit # 这个直接能得到截距和斜率
```