线性回归 2

王敏杰

2023-03-28

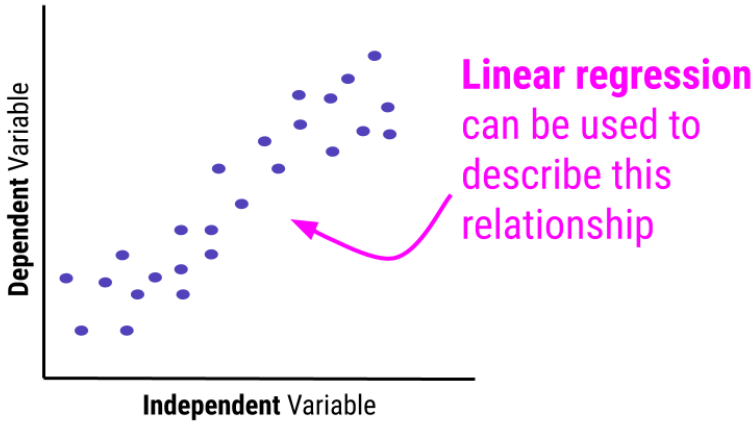
1 线性回归的前提假设

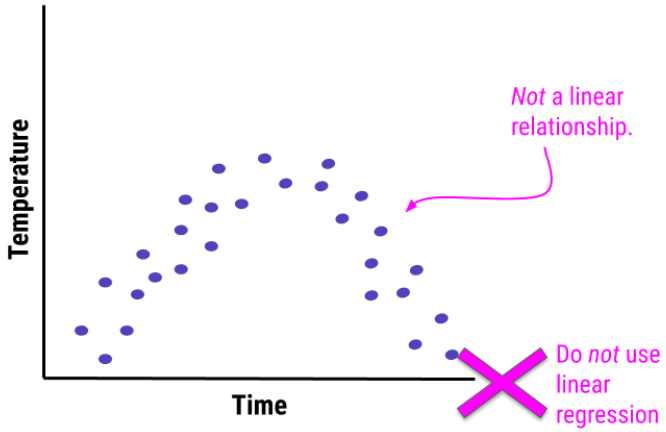
线性模型

$$y_n = \alpha + \beta x_n + \epsilon_n \quad \text{where} \quad \epsilon_n \sim \text{normal}(0, \sigma).$$

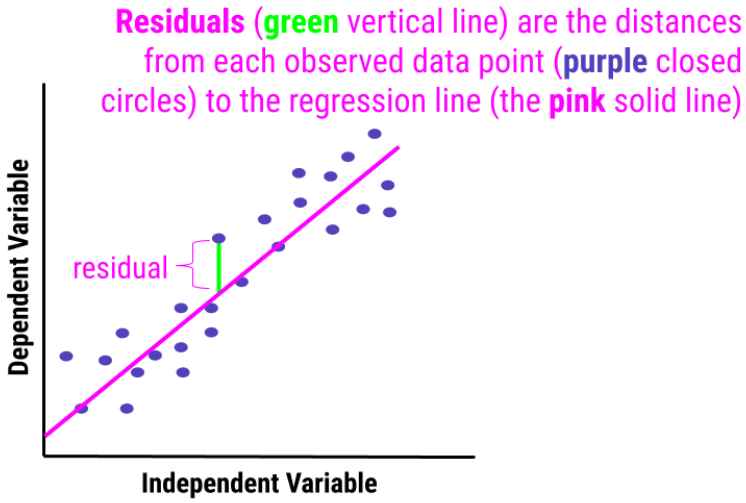
线性回归需要满足四个前提假设：

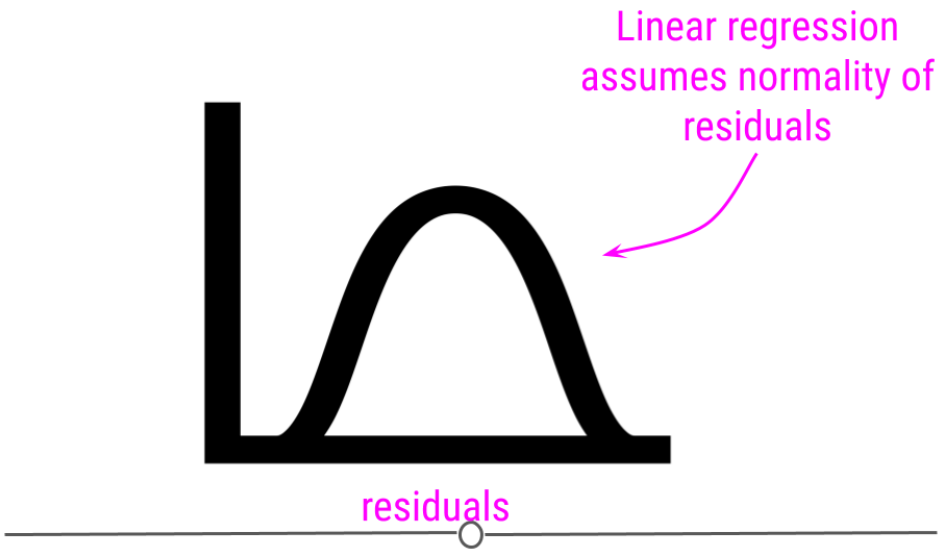
1. **Linearity**
   - 因变量和每个自变量都是线性关系



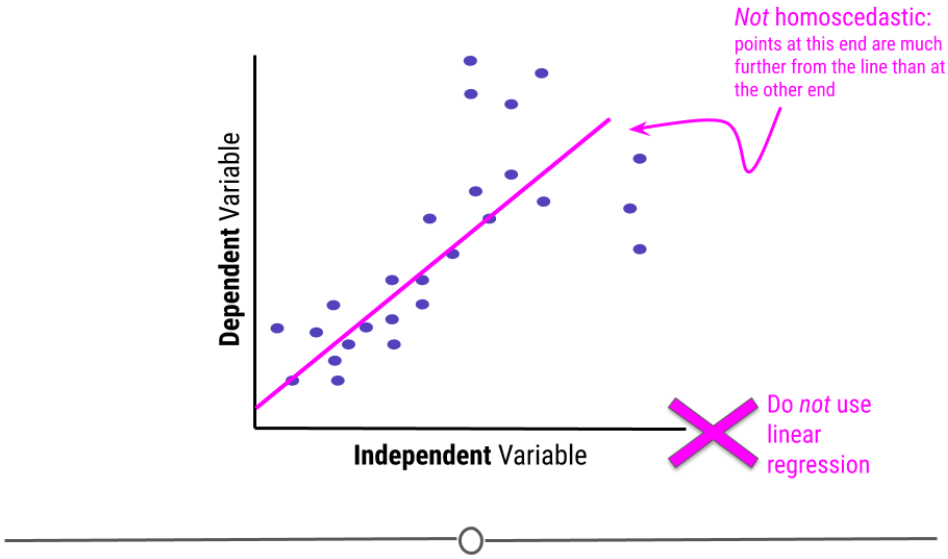


2. **Indpendence**
   - 对于所有的观测值，它们的误差项相互之间是独立的
3. **Normality**
   - 误差项服从正态分布



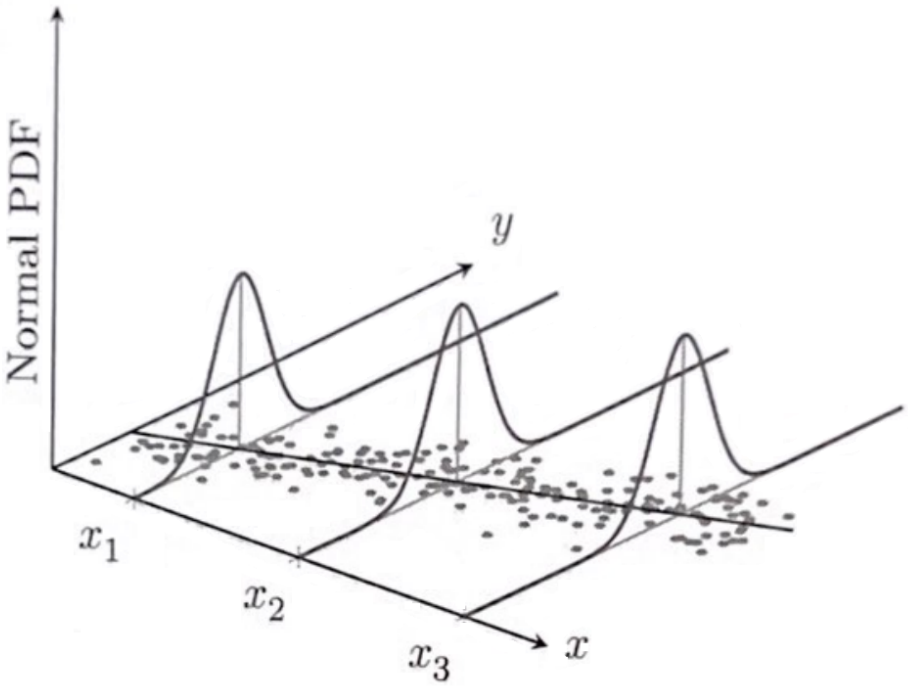


4. **Equal-variance**
    - 所有的误差项具有同样方差



这四个假设的首字母，合起来就是 **LINE**，这样很好记

把这**四个前提**画在一张图中



课堂练习：以下是否违背了 LINE 假设

1. 努力学习与是否通过 R 语言考试?
   - 响应变量 是否通过考试 (Pass or Fail)
   - 解释变量: 课后练习时间 (in hours)
2. 汽车音乐音量大小与司机刹车的反应时
   - 响应变量 反应时
   - 解释变量: 音量大小

2 回到案例

```
library(tidyverse)
wages <- read_csv("./demo_data/wages.csv")

wages %>% head()
```

earn	height	sex	race	edu	age
79571.30	73.89	male	white	16	49
96396.99	66.23	female	white	16	62
48710.67	63.77	female	white	16	33
80478.10	63.22	female	other	16	95
82089.35	63.08	female	white	17	43

earn	height	sex	race	edu	age
15313.35	64.53	female	white	15	30

$$y_i = \beta_0 + \beta_1 \text{height}_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2)$$

```
mod1 <- lm(
  formula = earn ~ 1 + height,
  data = wages
)

summary(mod1)
```

```
##
## Call:
## lm(formula = earn ~ 1 + height, data = wages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47903 -19744  -5184  11642 276796
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -126523      14076  -8.989   <2e-16 ***
## height          2387        211  11.312   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29910 on 1377 degrees of freedom
## Multiple R-squared:  0.08503,    Adjusted R-squared:  0.08437
## F-statistic:   128 on 1 and 1377 DF,  p-value: < 2.2e-16
```

3 更多模型

模型只是一种探测手段，通过捕获数据特征，探测数据的产生机制。所以说，模型只是人的一种假设，或者人的期望，试图去解释现象，但真实的情况，可能不是这样，任何一种模型都不正确。

所以，遇到不符合预期的情况，是正常的。我们可以带入更多的解释变量，建立若干个模型，逐一尝试，然后从中选择一个最好模型。

但要记住，所有模型都是人的”意淫”，要正确看待模型结果。

3.1 增加解释变量

之前是用单个变量 `height` 预测 `earn`，我们可以增加一个解释变量 `edu`，稍微扩展一下我们的一元线性模型，就是多元回归模型

$$\text{earn} = \beta_0 + \beta_1 \text{height} + \beta_2 \text{edu} + \epsilon$$

R 语言代码实现也很简单，只需要把变量 `edu` 增加在公式的右边

```
mod2 <- lm(earn ~ 1 + height + edu, data = wages)
```

同样，我们打印 `mod2` 看看

```
mod2
```

```
##
## Call:
## lm(formula = earn ~ 1 + height + edu, data = wages)
##
## Coefficients:
## (Intercept)       height          edu
##     -161541         2087         4118
```

3.2 分类变量

我们将**身高**和**性别**同时考虑进模型

$$Y_i = \beta_0 + \beta_1 \text{height}_i + \beta_2 \text{sex}_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2)$$

此时预测变量是一个分类变量和一个连续变量

```
mod5 <- lm(earn ~ 1 + height + sex, data = wages)
coef(mod5)
```

```
## (Intercept)      height     sexmale
##  -32479.861     879.424   16874.158
```

- `height = 879.424` 当 sex 保持不变时，height 变化一个单位引起的 earn 变化
- `sexmale = 16874.158` 当 height 保持不变时，sex 变化引起的 earn 变化 (male 与 female 的差值)

3.2.1 换种方式理解

$$Y_i = \beta_0 + \beta_1 \text{height}_i + \beta_2 \text{sex}_i + \epsilon_i$$

事实上，分类变量 `sex` 在 R 语言代码里，会转换成 0 和 1 这种**虚拟变量**，然后再计算。

我们这里显式地构建一个新变量 `gender`，将 `sex` 中 `(male, female)` 替换成 `(1, 0)`

```
wages_gender <- wages %>%
  mutate(gender = if_else(sex == "male", 1, 0))

wages_gender %>% head()
```

earn	height	sex	race	edu	age	gender
79571.30	73.89	male	white	16	49	1
96396.99	66.23	female	white	16	62	0
48710.67	63.77	female	white	16	33	0
80478.10	63.22	female	other	16	95	0
82089.35	63.08	female	white	17	43	0
15313.35	64.53	female	white	15	30	0

这样，我们获得一个等价的模型

$$y_i = \beta_0 + \beta_1 \text{height}_i + \beta_2 \text{gender}_i + \epsilon_i$$

然后放入 `lm()`

```
mod5a <- lm(earn ~ 1 + height + gender, data = wages_gender)
coef(mod5a)
```

```
## (Intercept)      height      gender
##  -32479.861     879.424   16874.158
```

我们发现系数没有发生变化，但更容易理解。

在固定的身高上，比较不同性别的收入**差异**:

- 当 `gender = 0` 情形

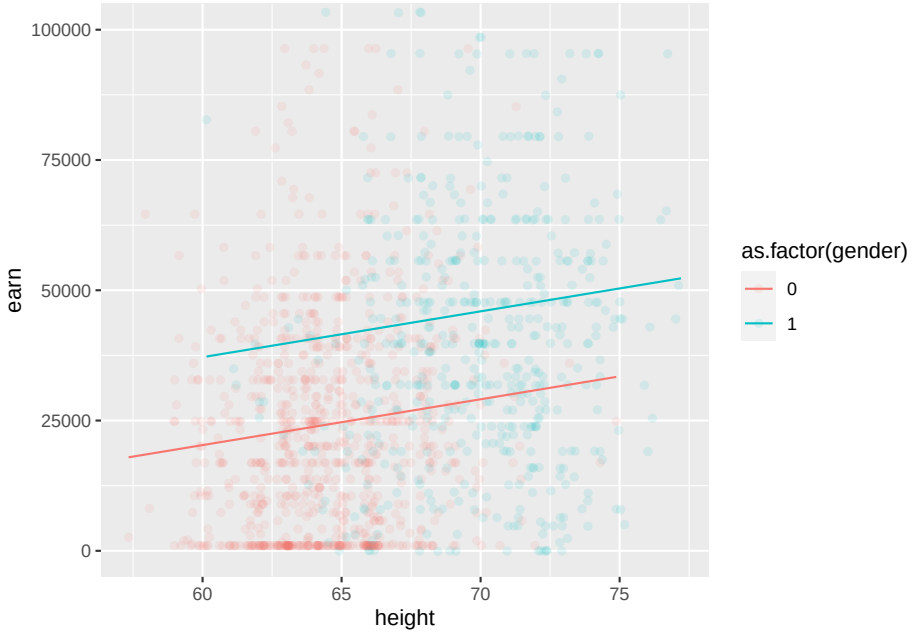
$$\begin{aligned} y_i &= \beta_0 + \beta_1 \text{height}_i + \beta_2 \text{gender}_i + \epsilon_i \\ &= \beta_0 + \beta_1 \text{height}_i + \epsilon_i \end{aligned}$$

- 当 `gender = 1` 情形

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \text{height}_i + \beta_2 \text{gender}_i + \epsilon_i \\ &= \beta_0 + \beta_1 \text{height}_i + \beta_2 + \epsilon_i \\ &= (\beta_0 + \beta_2) + \beta_1 \text{height}_i + \epsilon_i \end{aligned}$$

男性和女性的两条拟合直线，斜率相同、截距不同。

```
wages_gender %>%
  ggplot(aes(x = height, y = earn, color = as.factor(gender))) +
  geom_point(alpha = 0.1) +
  geom_line(aes(y = predict(mod5a))) +
  coord_cartesian(ylim = c(0, 100000))
```



3.3 交互项

然而，模型 5 的一个局限性，因为模型的结论从图形上看，是两条**平行的直线**:

1. 性别的影响对不同身高的人是相同的，
2. 或者，不管男性女性，收入随身高的增长是相同的。

虽然我们分组考虑不同性别的影响，但模型结论不一定符合现实情况。

为了扩展模型能力，允许预测因子之间相互影响，即需要考虑**交互项**。

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 \text{height}_i + \beta_2 \text{gender}_i + \\ & \beta_3 \text{height}_i \times \text{gender}_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2) \end{aligned}$$

```
mod6 <- lm(earn ~ 1 + height + gender + height:gender, data = wages_gender)
```

```
summary(mod6)
```

```
##
## Call:
## lm(formula = earn ~ 1 + height + gender + height:gender, data = wages_gender)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49699 -20090  -5034  11553 271709
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -12167.0    25340.4  -0.480    0.631
## height           564.5      392.5   1.438    0.151
## gender        -30510.4    39644.1  -0.770    0.442
## height:gender    701.4      585.8   1.197    0.231
##
## Residual standard error: 29340 on 1375 degrees of freedom
## Multiple R-squared:  0.1205, Adjusted R-squared:  0.1186
## F-statistic: 62.82 on 3 and 1375 DF,  p-value: < 2.2e-16
```

3.3.1 解释

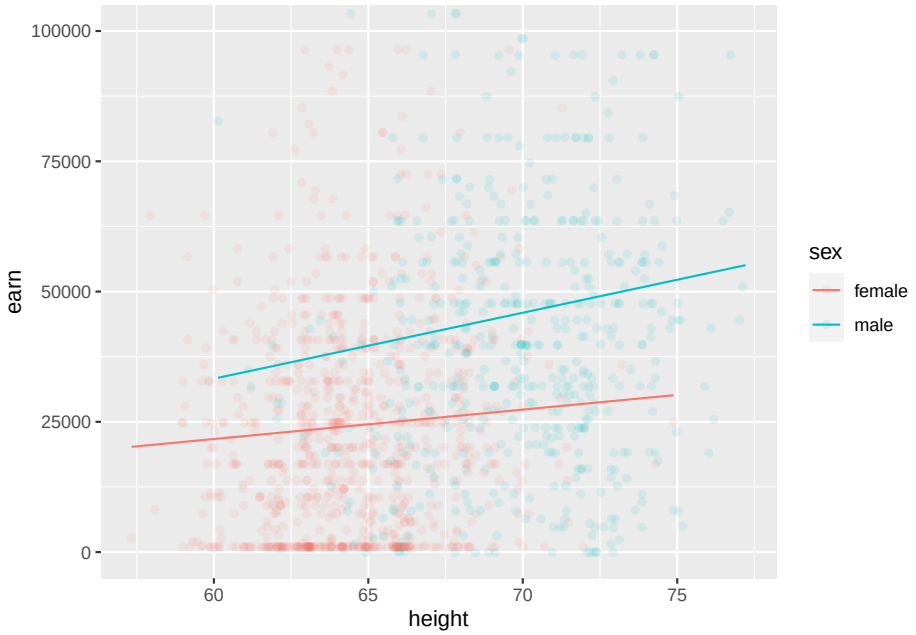
为了方便理解，我们仍然分开来看

$$\begin{aligned}
&\text{gender} = 0: \\
&\qquad \hat{y}_i = -12166.97 + 564.51\text{height}_i \\
&\text{gender} = 1: \\
&\qquad \hat{Y}_i = (-12166.97 - 30510.43) + (564.51 + 701.41)\text{height}_i
\end{aligned}$$

- 对于女性，height 增长 1 个单位，引起 earn 的增长 `564.5102`
- 对于男性，height 增长 1 个单位，引起 earn 的增长 `564.5102 + 701.4065 = 1265.92`

两条拟合直线，不同的截距和不同的斜率。

```
wages %>%
  ggplot(aes(x = height, y = earn, color = sex)) +
  geom_point(alpha = 0.1) +
  geom_line(aes(y = predict(mod6))) +
  coord_cartesian(ylim = c(0, 100000))
```



对于男性和女性，截距和系数都不同，因此这种情形**等价于**，按照 sex 分成两组，男性算男性的斜率，女性算女性的斜率（是不是似曾相识?）

```
wages %>%
  ggplot(aes(x = height, y = earn, color = sex)) +
  geom_point(alpha = 0.1) +
  geom_smooth(method = "lm", se = FALSE) +
  coord_cartesian(ylim = c(0, 100000))
```