线性回归

王敏杰

2022-11-06

线性模型是数据分析中最常用的一种分析方法。最基础的往往最深刻。

- Everything is regression (t tests, ANOVA, correlation, chi-square, non-parametric)
- The simplest linear regression model

1 从一个案例开始

这是一份 1994 年收集 1379 个对象关于收入、身高、教育水平等信息的数据集。首先，我们导入数据

```
library(tidyverse)
wages <- read_csv("./demo_data/wages.csv")
wages %>%
  head(5)
```

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

2 数据预处理

2.1 变量含义

通常，拿到一份数据，首先要了解数据每个变量的含义，

- `earn` 收入
- `height` 身高
- `sex` 性别
- `edu` 受教育程度
- `age` 年龄

2.2 缺失值检查

检查数据是否有缺失值，这点很重要。写代码的人都是偷懒的，希望写的简便一点，Tidyverse 函数总是很贴心、很周到。

```
wages %>%
  summarise(
    across(everything(), ~ sum(is.na(.x)))
  ) %>%
  pivot_longer(
    cols = everything()
  )
```

name	value
earn	0
height	0
sex	0
race	0
edu	0
age	0

2.3 初步统计

然后，探索下数据中每个变量的分布。比如男女数量分别是多少?

```
wages %>% count(sex)
```

sex	n
female	859
male	520

```
wages %>% count(race)
```

race	n
black	126
hispanic	77
other	29
white	1147

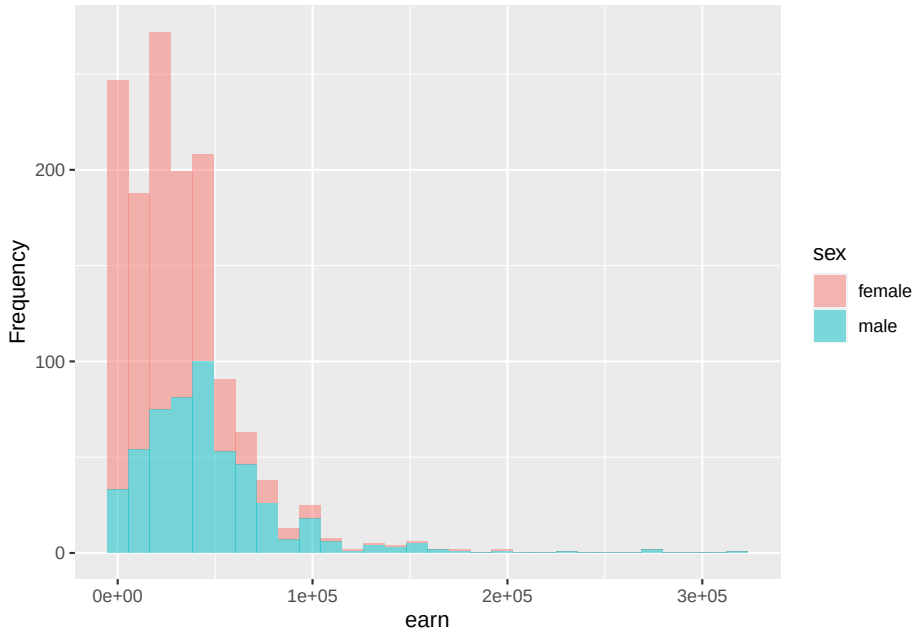
男女这两组的身高均值分别是多少？收入的均值分别是多少?

```r
wages %>%
  group_by(sex) %>%
  summarise(
    n = n(),
    across(c(height, earn), mean)
  )
```

sex	n	height	earn
female	859	64.50303	24245.65
male	520	70.04452	45993.13

也可以用可视化的方法，呈现男女收入的分布情况

```r
wages %>%
  ggplot(aes(x = earn, fill = sex)) +
  geom_histogram(alpha = 0.5) +
  labs(x = "earn", y = "Frequency" )
```



课堂练习：

- 用分面的方法，呈现男女收入的分布情况
- 用分面的方法，呈现男女身高的分布情况

3 问题

现在提出几个问题，希望大家带着这些问题去思考：

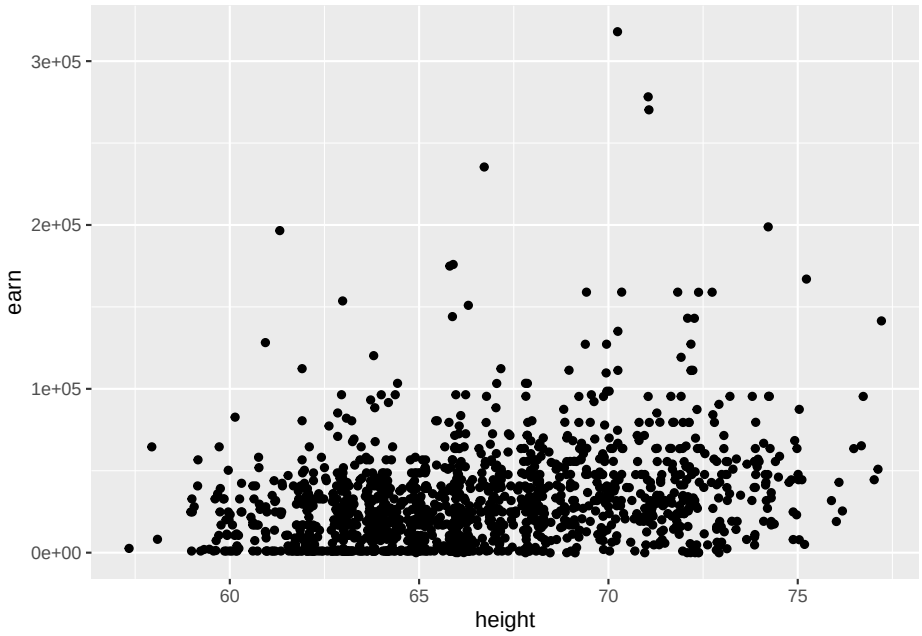
- **长的越高的人挣钱越多？**
- 男性是否比女性挣的多?
- 受教育程度越高，挣钱越多?
- 如果知道年龄、性别和受教育程度，我们能预测他的收入情况吗？预测在多大范围是可信的?

从统计学的角度回答以上问题，需要引入统计模型。

4 长的越高的人挣钱越多？

回答这个问题，可以先可视化收入和身高的关系

```
wages %>%
  ggplot(aes(x = height, y = earn)) +
  geom_point()
```



图形给出的趋势，非常值得我们通过建立模型加以验证。

假定 y_i 是每个样本 i 的收入，x_i 是每个样本 i 的身高，如果将**收入**视为**身高**的线性函数，可以表示为

$$y_i = \alpha + \beta x_i + \epsilon_i$$
$$\epsilon_i \in \text{Normal}(0, \sigma^2)$$

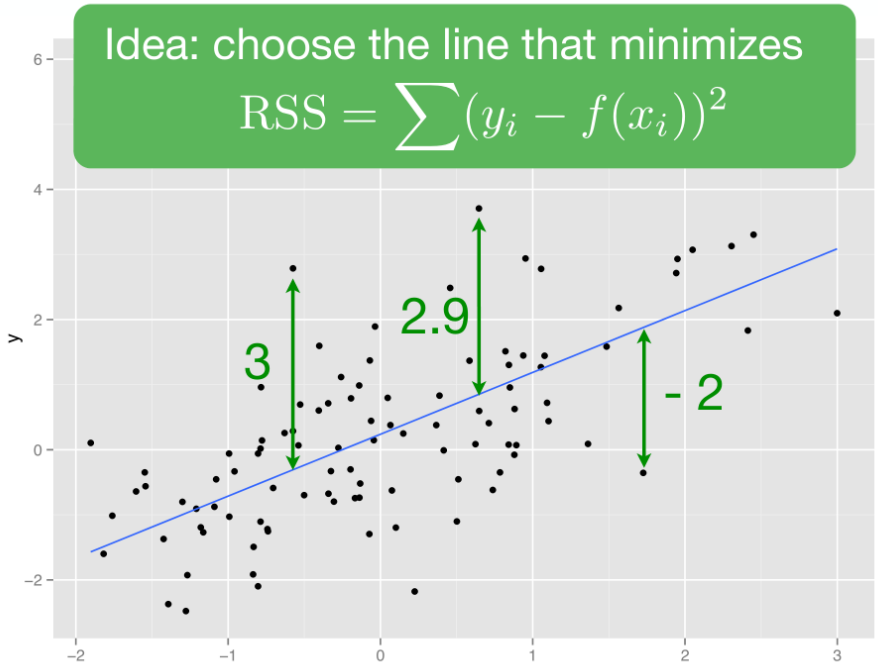
先解释一些概念：

- 观测数据 y_i 分成两个部分，即 `data = model + error`
- 模型部分，系数与预测变量的线性组合，$\alpha + \beta x_i$
- 残差部分 (误差项部分)，ϵ_i，代表着**真实收入**与模型给出的**期望收入**之间的**偏差**，它与 x 无关，服从正态分布。

我们的任务是估计**系数**

- α 代表截距
- β 代表斜率

模型中的参数 (α, β, 和 σ^2) 通过最小二乘法计算得出，基本思路是**使得残差的平方和最小（偏差的总量最小）**



当然，数据量很大，手算是不现实的，我们借助 R 语言代码吧。

4.1 使用 `lm()` 函数

用 R 语言代码 `lm(y ~ 1 + x, data)` 是最常用的线性模型函数，`lm()` 是 linear model 的缩写

`lm` 参数很多，但很多我们都用不上，所以我们只关注其中重要的两个参数

```
lm(formula = y ~ 1 + x, data = __)
```

解释说明：

- formula：指定回归模型的公式，对于简单的线性回归模型写为 `y ~ 1 + x`.
- ~ 符号：代表 “预测”，可以读做 “y 由 x 预测”。有些学科不同的表述，比如下面都是可以的

- response ~ explanatory
- dependent ~ independent
- outcome ~ predictors

- data：代表数据框，数据框包含了响应变量和独立变量

等不及了，马上运行代码吧

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

公式中为什么有个 1，我们可以这样理解

表达式	符号						
数学表达	y	=	a	*	1	+	b * x
R 表达式	y	~			1	+	x

注意，以下两者是等价的

```
lm(formula = earn ~ 1 + height,  data = wages)
lm(formula = earn ~     height,  data = wages)   # 偷懒版本
```

lm() 返回赋值给 mod1，mod1 是一个叫 lm object 或者叫类的东西，我们打印看看，

```
print(mod1)
```

```
##
## Call:
## lm(formula = earn ~ 1 + height, data = wages)
##
## Coefficients:
## (Intercept)       height
##     -126523         2387
```

它告诉我们，建立的线性回归模型是

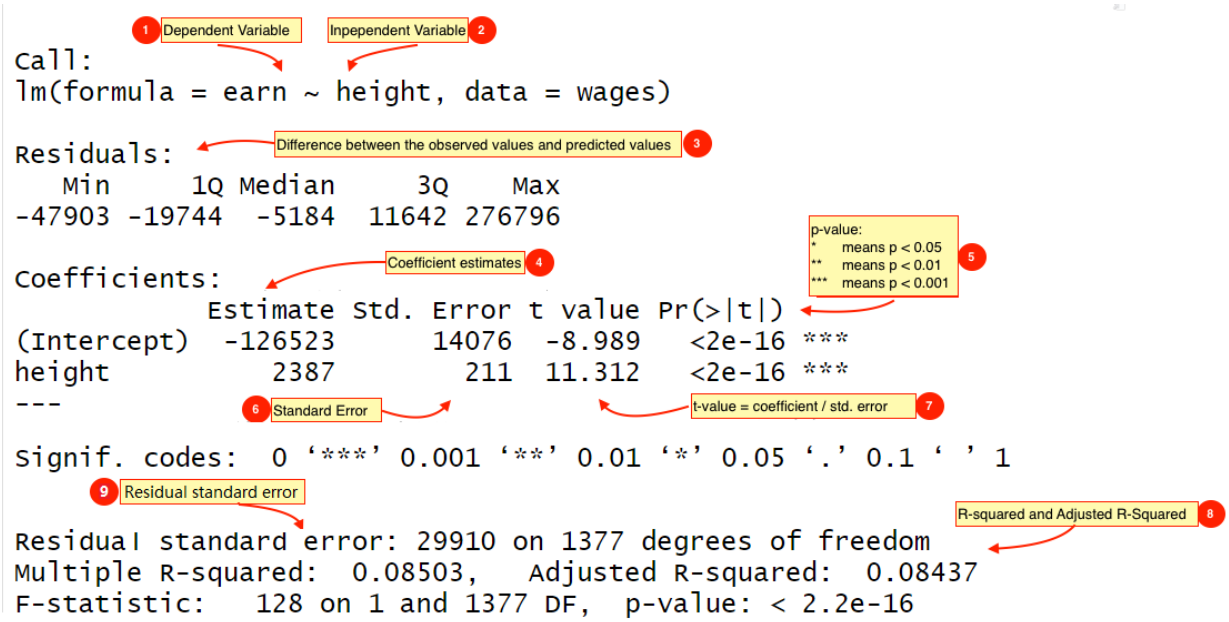
$$\hat{\text{earn}} = -126532 + 2387 \text{ height}$$

收入的期望值是 -126532 乘以 1，加上 2387 乘以 height

4.2 模型输出 (害羞的小男孩)

4.2.1 查看详细信息

```
summary(mod1)
```



4.2.2 系数

```
coef(mod1)
```

```
## (Intercept)      height
## -126523.359    2387.196
```

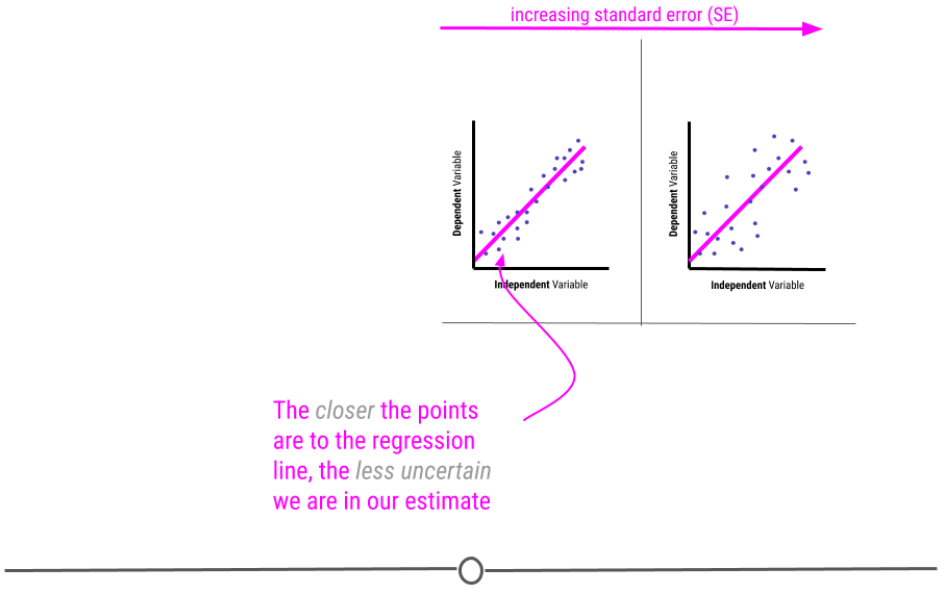
4.2.3 系数的置信区间

```
confint(mod1, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) -154135.798 -98910.920
## height         1973.228   2801.163
```

4.2.4 标准误 (Std. Error)

标准误反映的是系数估计的不确定程度。标准误越大，系数估计的不确定越大。



4.2.5 残差

```
resid(mod1) %>%
  quantile()
```

```
##          0%         25%         50%         75%        100%
## -47903.128 -19743.886  -5183.842  11642.160 276795.873
```

4.2.6 残差标准误

```
summary(mod1)$sigma
```

```
## [1] 29909.5
```

我们也手动计算

$$\hat{\sigma}^2 = \sum \hat{\epsilon}_i^2/(n-2)$$

注意公式左边是平方，所以，我们最后需要开方

```
sum( resid(mod1)^2 / 1377) %>% sqrt()
```

```
## [1] 29909.5
```

4.2.7 拟合优度

R^2 代表着拟合优度（Goodness of Fit），它指回归直线对观测值的拟合程度。

```
summary(mod1)$r.squared
```

```
## [1] 0.08503071
```

4.3 模型的解释

建立一个 `lm` 模型是简单的，然而最重要的是，我们能解释这个模型。

$$\hat{\text{earn}} = -126523 + 2387(\text{height})$$

- 对于斜率 $\beta = 2387$ 意味着，当一个人的身高是 68 英寸时，他的预期收入 $earn = -126532 + 2387 \times 68 = 35793$ 美元，如果是 69 英寸时，他的预期收入 $earn = -126532 + 2387 \times 69 = 38180$ 美元，换个方式说，身高 $height$ 每增加一个 1 英寸，收入 $earn$ 会增加 2387 美元。
- 对于截距 $\alpha = -126532$，即当身高为 0 时，期望的收入值-126532。呵呵，人的身高不可能为 0，所以这是一种极端的理论情况，现实不可能发生。
- $R^2 = 0.08503$ 意味着**该模型**只能解释 8% 的**收入随身高变化的特征**。

```
wages %>%
  ggplot(aes(x = height, y = earn)) +
  geom_point(alpha = 0.25) +
  geom_smooth(method = "lm", se = FALSE)
```