从数据到论文—模型探索

关于小学生跑步的案例

王敏杰

2023-04-01

从一个案例开始

小学生50米短跑的成绩

```
library(tidyverse)

d <- read_csv("./data/data-50m.csv")
```

Y	Weight	Age
2.46	16.6	7
3.02	21.4	7
2.91	24.0	7
3.05	13.0	7
2.60	21.9	7
2.45	22.1	7
2.61	18.7	7
2.91	18.8	7

从一个案例开始

小学生50米短跑的成绩

```
library(tidyverse)

d <- read_csv("./data/data-50m.csv")
```

Y	Weight	Age
2.46	16.6	7
3.02	21.4	7
2.91	24.0	7
3.05	13.0	7
2.60	21.9	7
2.45	22.1	7
2.61	18.7	7
2.91	18.8	7

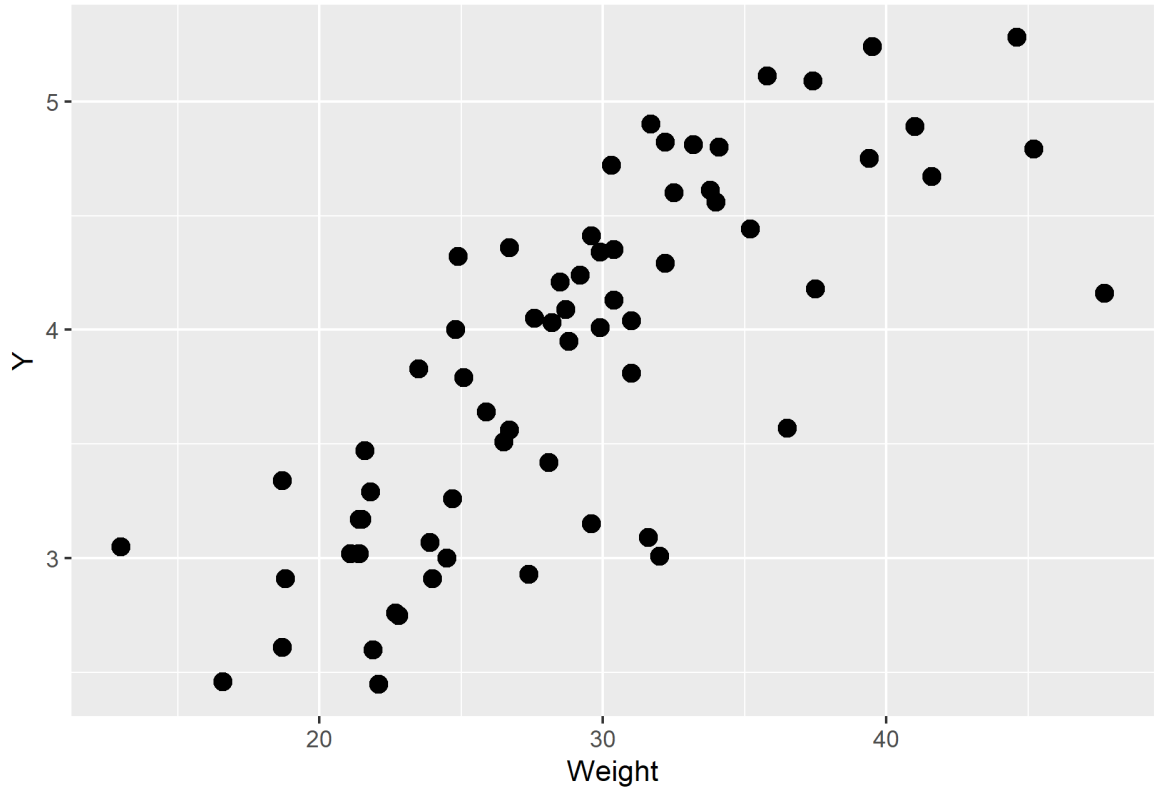
变量含义

变量	含义
Y	速度 (m/s)
Weight	体重 (kg)
Age	年龄 (year)

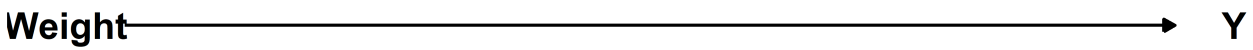
探索一：体重越大，跑步越慢?

简单探索

我们先画出体重与速度的散点图



建立线性模型



建立线性模型

```
mod1 <- lm(Y ~ Weight, data = d)

mod1 %>% summary()
```

```
##
## Call:
## lm(formula = Y ~ Weight, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2646 -0.3679  0.0473  0.3838  0.8113
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.44204    0.26294    5.48  7.5e-07 ***
## Weight       0.08349    0.00882    9.47  8.6e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.504 on 64 degrees of freedom
## Multiple R-squared:  0.584,    Adjusted R-squared:  0.577
## F-statistic: 89.7 on 1 and 64 DF,  p-value: 8.65e-14
```

建立线性模型

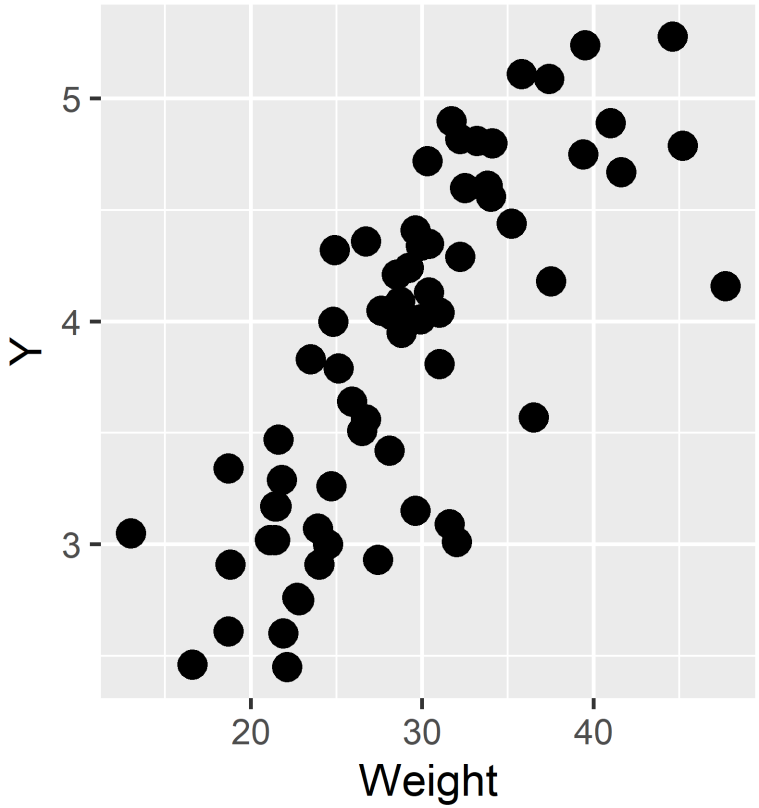
```
mod1 <- lm(Y ~ Weight, data = d)

mod1 %>% summary()
```

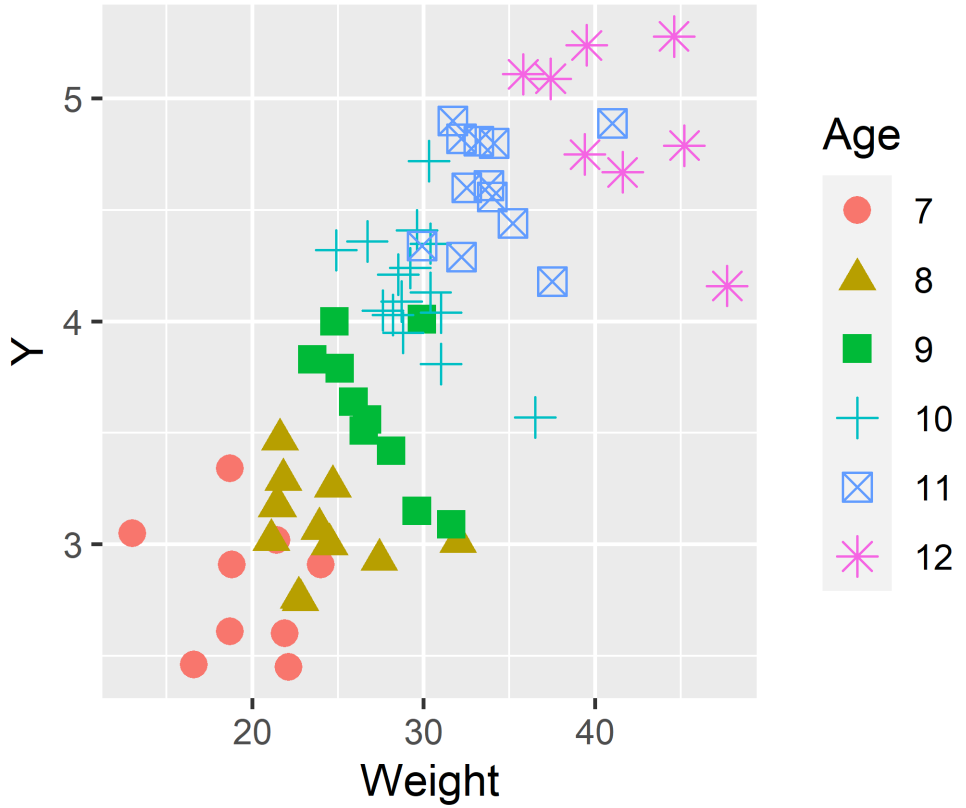
```
##
## Call:
## lm(formula = Y ~ Weight, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2646 -0.3679  0.0473  0.3838  0.8113
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.44204    0.26294    5.48  7.5e-07 ***
## Weight       0.08349    0.00882    9.47  8.6e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.504 on 64 degrees of freedom
## Multiple R-squared:  0.584,    Adjusted R-squared:  0.577
## F-statistic: 89.7 on 1 and 64 DF,  p-value: 8.65e-14
```

体重越大，速度越快？似乎与我们生活常识不太相符啊。

探索二：考虑年龄因素

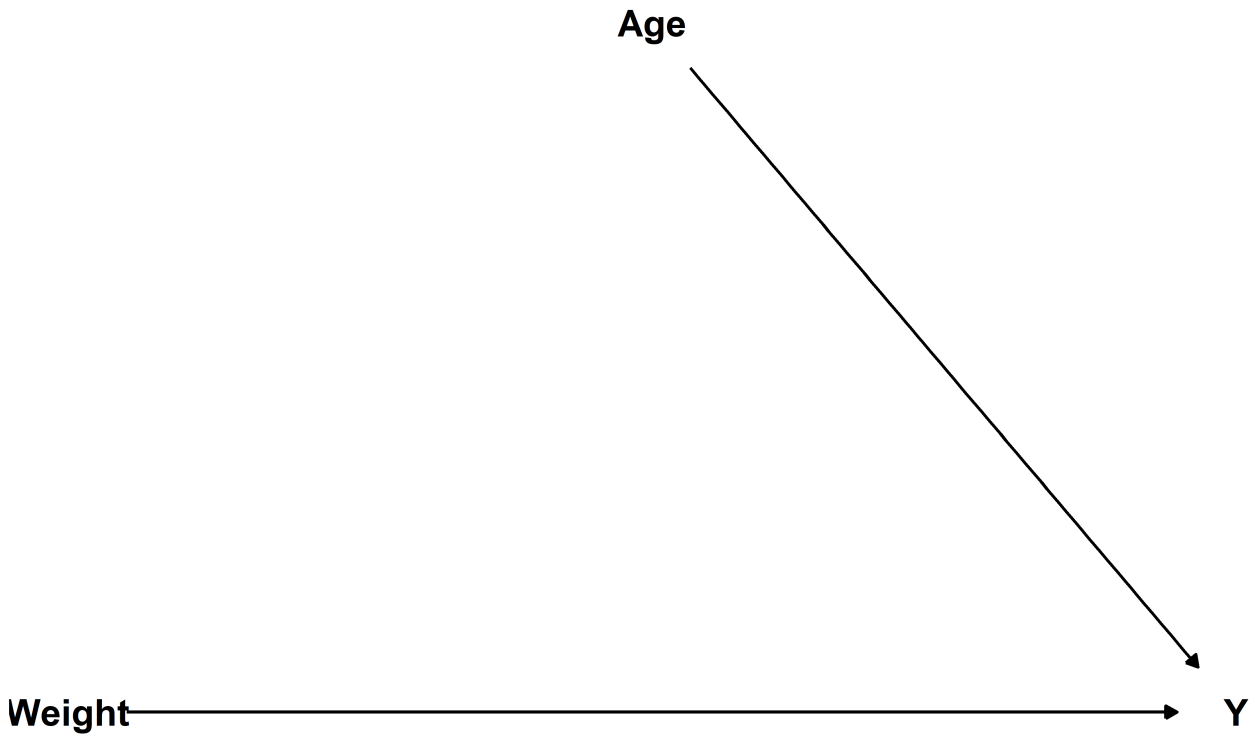
考虑年龄因素





我们在建模中不能忽略小学生年龄这个重要的解释变量。同时也看到。在同龄的小学生中，胖子要跑的慢些。

多元回归



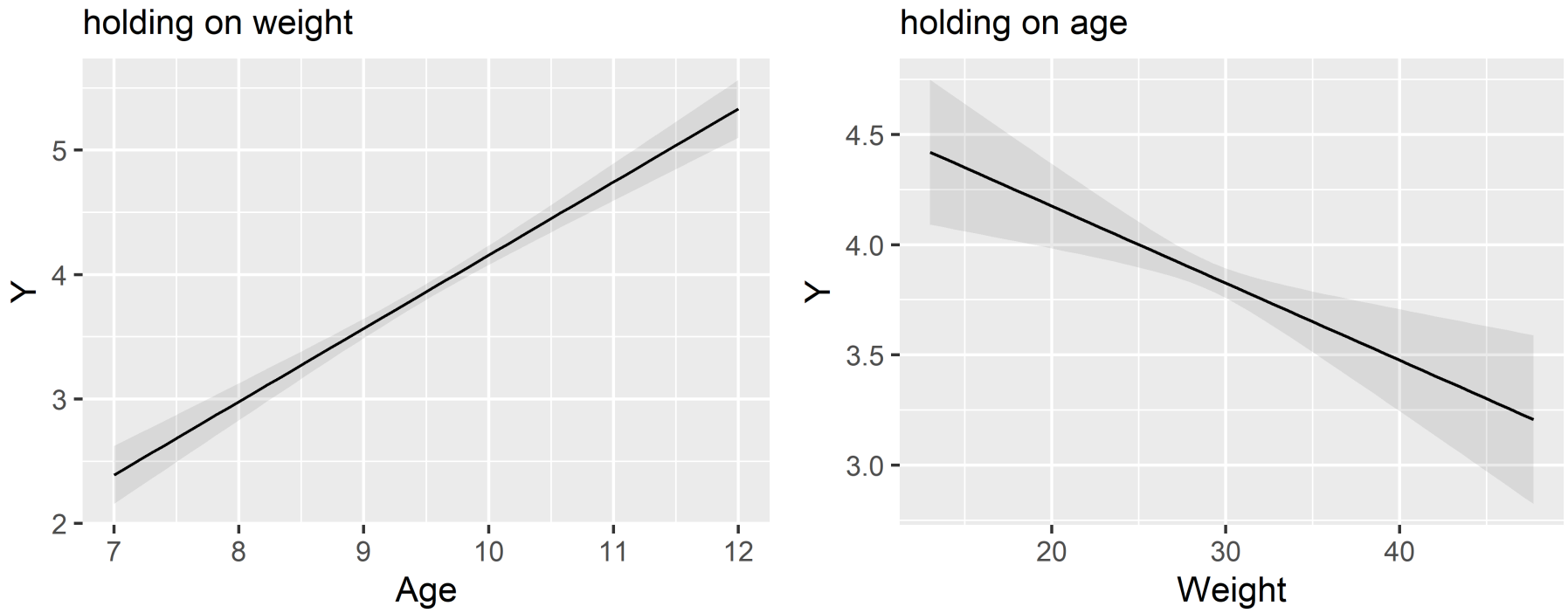
多元回归

```
mod2 <- lm(Y ~ Weight + Age, data = d)

mod2 %>% summary()
```

```
##
## Call:
## lm(formula = Y ~ Weight + Age, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5186 -0.1720 -0.0151  0.1573  0.6104
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.7169     0.2171    -3.3   0.0016 **
## Weight       -0.0349     0.0103    -3.4   0.0012 **
## Age           0.5885     0.0455    12.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.266 on 63 degrees of freedom
## Multiple R-squared:  0.886,    Adjusted R-squared:  0.882
## F-statistic:  245 on 2 and 63 DF,  p-value: <2e-16
```

多元回归

对于小学生，随着年龄的增长，肌肉越发达，所以跑步速度会越快。 但如果学生超重，那么也会影响速度。



上图是在体重不变的前提下，跑步速度随年龄的变化；以及在年龄不变的前提下，跑步速度随体重的变化。

未完待续!