R 模型可视化

宋骁

模型可视化是应用统计学的重要内容。任何模型都离不开结果的可视化。所谓模型，不过是将一堆散点简化为一条线。结果的可视化需要预测值。Hadley Wickham 的 `modelr` 包提供了重要的用于预测的函数。预测的结果可以直接被 `ggplot2` 使用并画图。

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(modelr)
library(haven)
library(easyGgplot2)
`%>%` <- magrittr::`%>%`
```

1 基础回归

`hatdt` 为作者个人整理的中国家庭追踪调查(CFPS) 收入数据[1]。

```
hatdt <- hatdt %>%
  filter(type=='个人收入（元）') %>%
```

[1]可从Github下载

```
  drop_na(agem,inc,fswt_nat)


set.seed(20191001)
sample <- sample(1:nrow(hatdt),600,replace = F)
sampled <- hatdt[sample,]


plota <- ggplot(hatdt,aes(agem,inc,weight=fswt_nat)) +
  geom_jitter(data=sampled,height=550,width=5,
              size =1.5,alpha=1/3) +
  geom_smooth(span =10,size=1) +
  geom_smooth(method='lm',size=1,color='red') +
  ylim(0, 20000) +
  labs(x = "年龄",y = "人民币(元)") +
  theme_bw()


plotb <- ggplot() +
  geom_jitter(data=sampled,aes(agem,inc),
              height=550,width=5,size =1.5,alpha=1/3) +
  geom_quantile(data=hatdt,
  aes(agem,inc,weight=fswt_nat),
  size=1,color='red')+
  ylim(0, 20000) +
  labs(x = "年龄",y = "人民币(元)") +
  theme_bw()
```

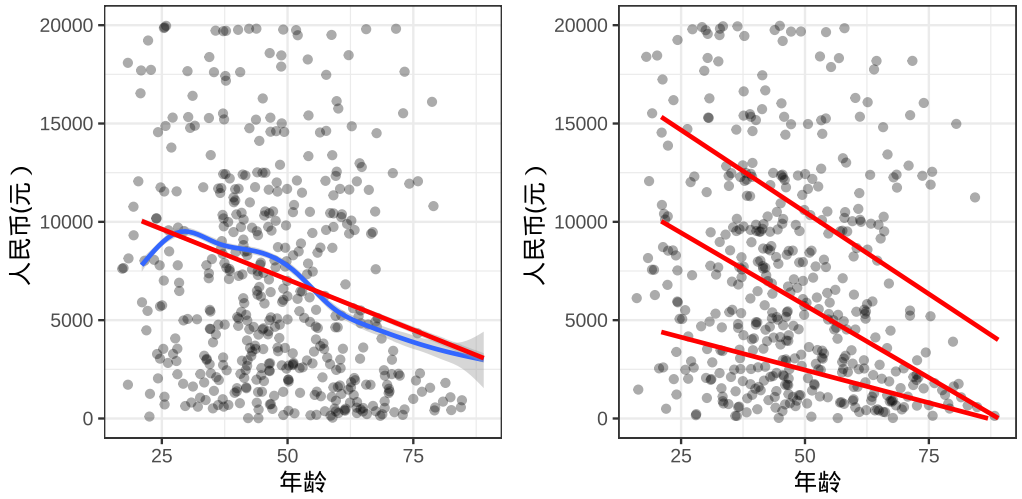

图 1: 个人收入与年龄。左图：红线为线性回归模型。蓝色曲线为非参数回归。右图：三条线分别是分位数回归。高收入者收入随年龄下降的速度快于低收入者。可将中位数回归与左图线性回归相比较，观测其中的差异。

```
ggplot2.multiplot(plota,plotb,cols=2)
```

2 多项式回归

```
set.seed(2019)

x <- seq(0,4,length=100)

y <- -x^2 + 3*x + jitter(rep(5:9,each =20),2) +3

df <- data.frame(x,y)


reg <- lm(y ~ x + I(x^2),df)


grid <- df %>%

data_grid(x) %>%

gather_predictions(reg)
```

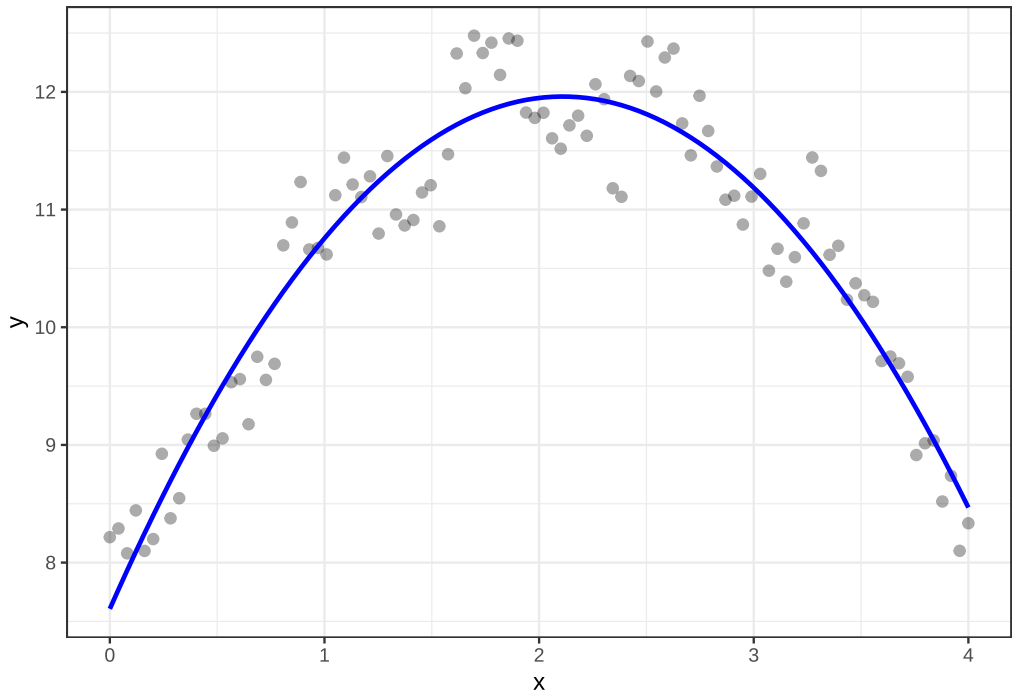

图 2: 对一个模拟数据进行二次项回归。

```
ggplot(df,aes(x,y))+
  geom_point(size =2,alpha=1/3)+
  geom_line(data=grid,aes(x,pred),size=1,color='blue')+
  theme_bw()
```

下面使用多项式回归拟合 CFPS 数据:

$$y = \alpha_0 + \alpha_1 x_1 + x_1^2$$

$$y = \alpha_0 + \alpha_1 x_1 + x_1^2 + x_1^3$$

```
mtrga <- lm(inc~agem+I(agem^2),hatdt)
mtrgb <- lm(inc~agem+I(agem^2)+I(agem^3),hatdt)
```

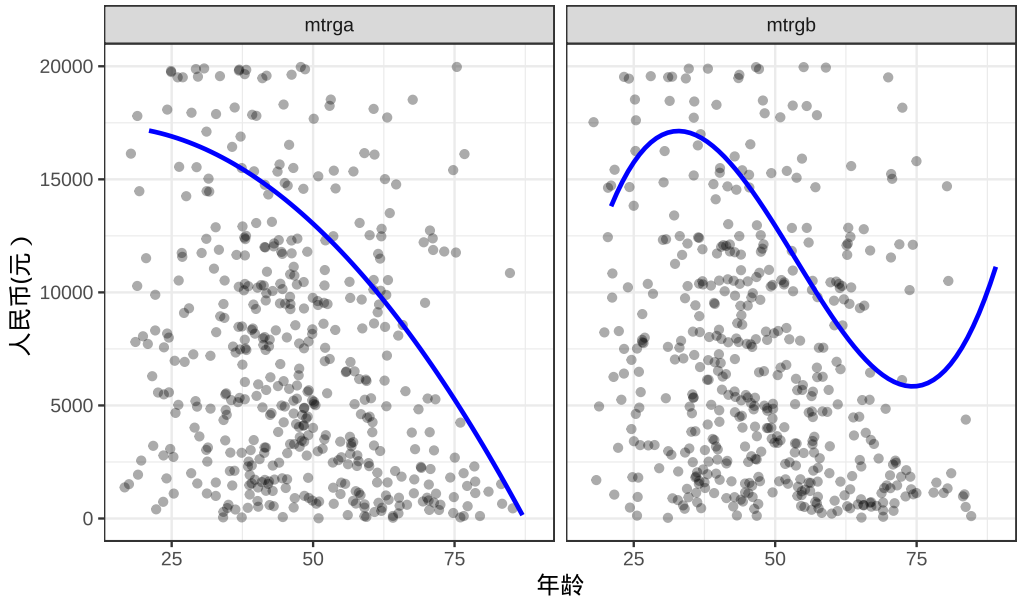

图 3: 分别对 CFPS 数据进行二次项和三次项回归。三次项导致了过拟合。

```
grid <- hatdt %>%
data_grid(agem) %>%
gather_predictions(mtrga,mtrgb)


ggplot() +
  geom_jitter(data=sampled,aes(agem,inc),
              height=550,width=5,size =1.5,alpha=1/3) +
  geom_line(data=grid,aes(agem,pred),
            size=1,color='blue')+
  facet_wrap(~model) +
  ylim(0, 20000) +
  labs(x = "年龄",y = "人民币(元)") +
  theme_bw()
```

3 交互项

交互项是计量经济学和应用统计学常用的机制分析技术。公式如下：

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2$$

下面使用 R 自带数据，1994 年加拿大劳动与收入动态调查 (SLID)。详细信息请在 R 中输入?carData::SLID 查看。

3.1 分类变量与连续变量交互

因变量为收入。自变量为教育年限 (年) 和使用的语言 (英语、法语、其他)。下面分别展示了没有交互项和有交互项的模型。

```
#?carData::SLID
data(SLID,package = 'carData')
SLID <- SLID %>% drop_na()

mod1 <- lm(wages ~ education + language,SLID)
mod2 <- lm(wages ~ education * language,SLID)

grid <- SLID %>%
data_grid(education,language) %>%
gather_predictions(mod1,mod2)

ggplot(SLID,aes(education,wages))+
  geom_jitter(size=1,width=2,height=10,alpha=1/7)+
  geom_line(data=grid,
```

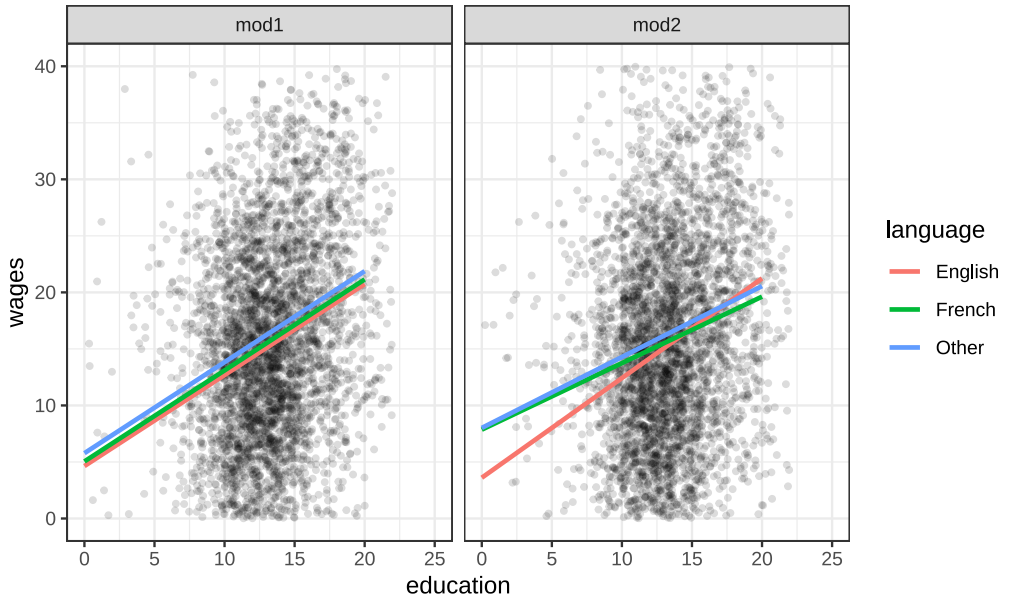

图 4: 左图：语言不与教育年限交互。不同语言使用者的斜率相同但截距不同。右图：交互模型，英语使用者的工资随教育回报率更高，假定其他条件不变。英语使用者在 15 年处超越了其他语言使用者。

```
            aes(education,pred,color=language),size=1)+
  facet_wrap(~model)+
  xlim(0,25)+ ylim(0,40)+
  theme_bw()
```

3.2 两个连续变量交互

对两个连续交互变量的可视化是一个难题。较好的解决办法是分箱。使用 `modelr` 的 `seq_range` 函数对其中一个连续变量进行分箱。

```
mod1 <- lm(wages ~ education + age,SLID)
mod2 <- lm(wages ~ education * age,SLID)


grid <- SLID %>%
```

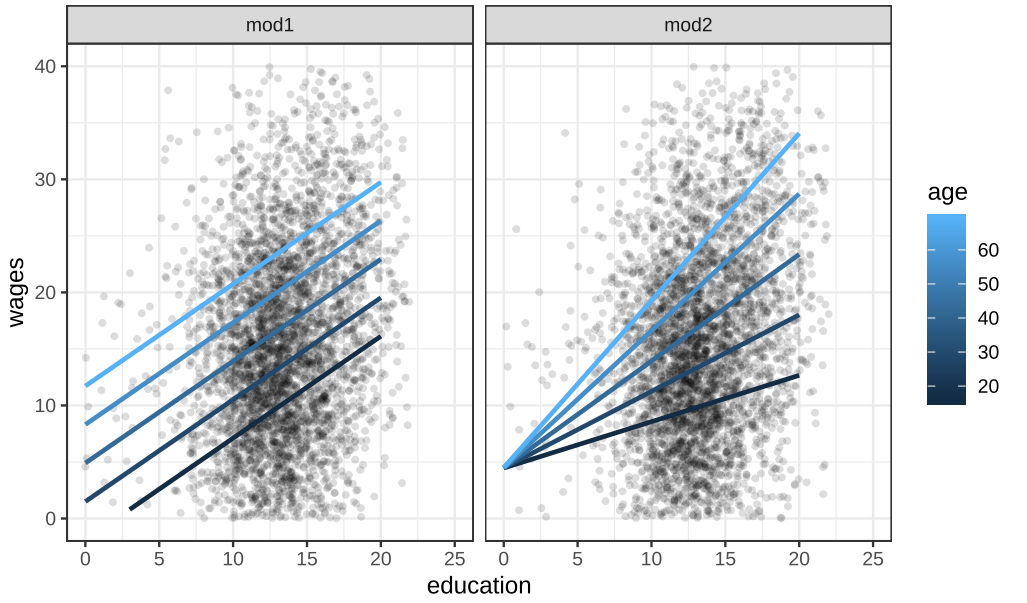

图 5: 无交互效应和有交互效应的区别：左图体现了同一年龄段者的教育回报率相同 (斜率相同)。右图体现了一个因素的大小随着另一个因素的变化而变化。随着年龄的升高教育回报率也在升高。

```
data_grid(education,age = seq_range(age, 5)) %>%
gather_predictions(mod1,mod2)


ggplot(SLID,aes(education,wages))+
  geom_jitter(size=1,width=2,height=10,alpha=1/7)+
  geom_line(data=grid,aes(education,pred,
                          color=age,group=age),size=1)+
  facet_wrap(~model)+
  xlim(0,25)+ ylim(0,40)+
  theme_bw()
```