R 模型可视化*

宋骁

English Version

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` 包提供用于预测的函数。预测的结果可以直接被 `ggplot2` 使用并画图。`modelr` 支持管道操作，是将数据分析流程化的利器。

`modelr` 包的主要函数有：

`data_grid`: 生成预测数据

`add_predictions`: 加入预测值

`crossv_kfold`、`crossv_mc`、`crossv_loo`: 交叉验证

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

*网页版本：https://xsong.ltd/zh/model

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）') %>%
  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),
```

[1] 可从Github下载

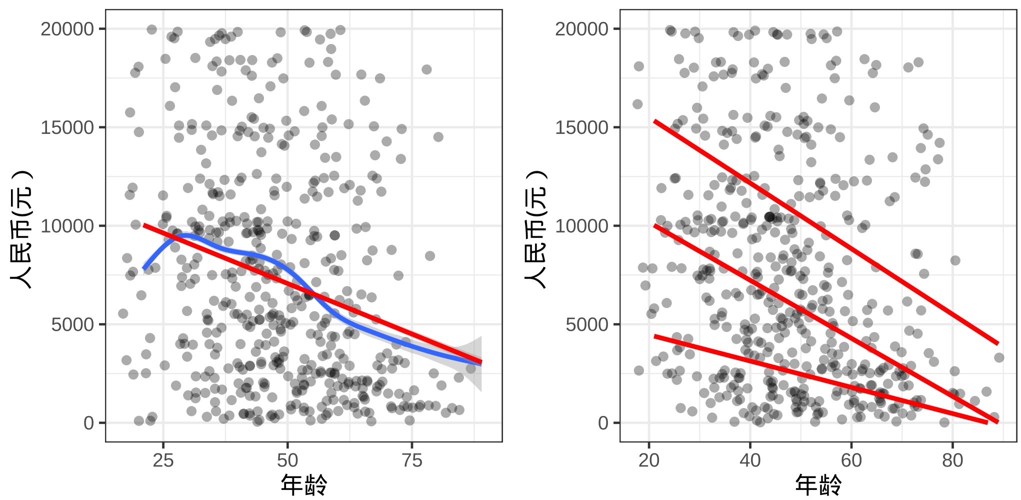


图 1: 个人收入与年龄。左图：红线为线性回归模型。蓝色曲线为非参数回归。右图：三条线分别是分位数回归。高收入者收入随年龄下降的速度快于低收入者。可将中位数回归与左图线性回归相比较，观测其中的差异。

```
  size=1,color='red')+
  ylim(0, 20000) +
  labs(x = "年龄",y = "人民币(元）") +
  theme_bw()
```

```
plot_grid(plota,plotb,ncol = 2)
```

2 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：

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

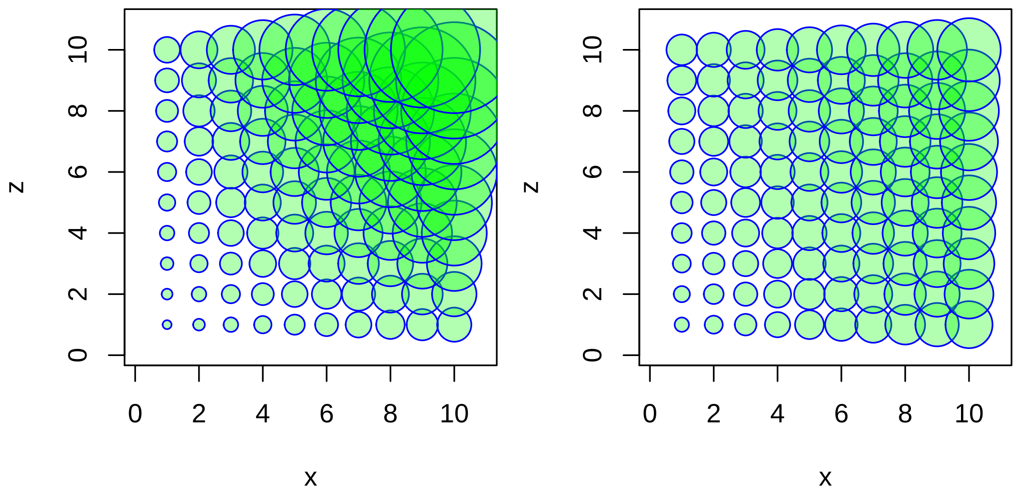


图 2: 谢益辉的交互效应表示方法。左图：$y = 2 + x + 0.5z + 0.5xz + \epsilon$。右图：$y = 2 + x + 0.5z + \epsilon$。圆圈面积表示因变量 y 的大小；坐标轴分别表示自变量 x 和 z。

```
par(mar = c(4,4,1,0.5), mfrow = c(1, 2), cex.main = 1)
sq = 1:10
x = rep(sq, 10)
z = rep(sq, each = 10)
y = c(outer(sq, sq, function(x, z) 2 + x + 0.5 *
z + 0.5 * x * z + runif(1)))
symbols(x, z, y, bg = rgb(0, 1, 0, 0.3), fg = "blue",
main = "",
inches = 0.4)
y = c(outer(sq, sq, function(x, z) 2 + x + 0.5 *
z + runif(1)))
symbols(x, z, y, bg = rgb(0, 1, 0, 0.3), fg = "blue",
main = "", inches = 0.2)
```

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。

2.1 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,
            aes(education,pred,color=language),size=1)+
  facet_wrap(~model)+
  xlim(0,25)+ ylim(0,40)+
  theme_bw()
```

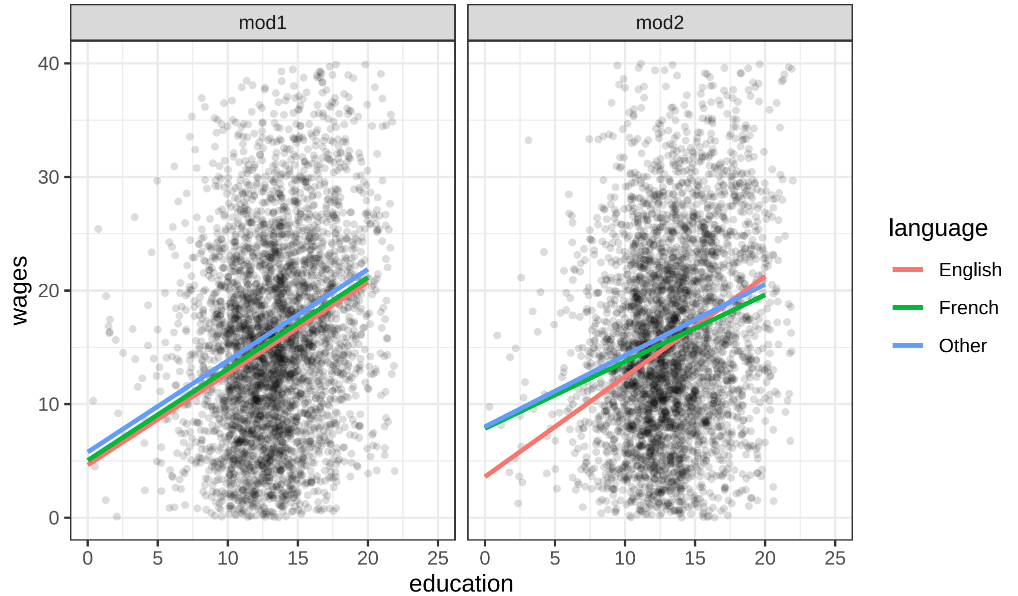


图 3: 左图：语言不与教育年限交互。不同语言使用者的斜率相同但截距不同。右图：交互模型，英语使用者的工资随教育回报率更高，假定其他条件不变。英语使用者在 15 年处超越了其他语言使用者。

2.2 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 %>%
data_grid(education,age = seq_range(age, 5)) %>%
gather_predictions(mod1,mod2)


ggplot(SLID,aes(education,wages))+
  geom_jitter(size=1,width=2,
```

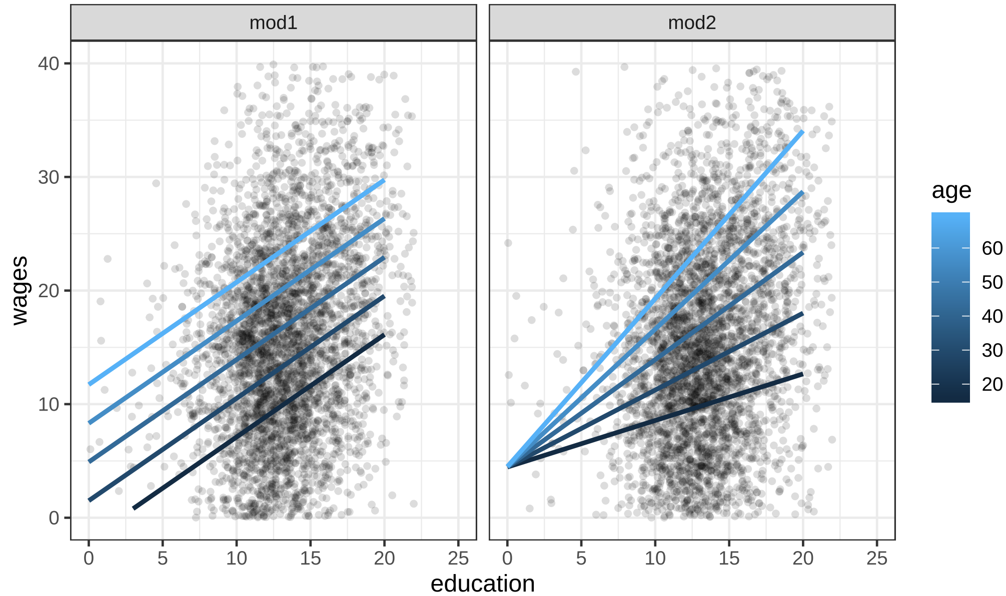


图 4: 无交互效应和有交互效应的区别：左图体现了不同年龄段者的教育回报率相同 (斜率相同)。右图体现了一个因素的大小随着另一个因素的变化而变化。随着年龄的升高教育回报率也在升高。

```
              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()
```

来个负相关的：

```
data(freeny)
partial <- lm(y~lag.quarterly.revenue+price.index+
                income.level+market.potential,freeny)


modela <- lm(y~price.index+market.potential,freeny)
```

```
modelb <- lm(y~price.index*market.potential,freeny)


stargazer(modela,modelb,partial,
          title='回归结果',
          dep.var.caption='',
          dep.var.labels='Quarterly Revenue',
          header=F,keep.stat=c('n','rsq'),
         no.space=T,type='latex')
```

表 1: 回归结果

	Quarterly Revenue		
	(1)	(2)	(3)
lag.quarterly.revenue			0.124
			(0.142)
price.index	-0.414^{*}	-39.796^{***}	-0.754^{***}
	(0.210)	(5.737)	(0.161)
income.level			0.767^{***}
			(0.134)
market.potential	4.030^{***}	-10.270^{***}	1.331^{**}
	(0.434)	(2.102)	(0.509)
price.index:market.potential		2.979^{***}	
		(0.434)	
Constant	-41.499^{***}	147.459^{***}	-10.473^{*}
	(6.602)	(27.863)	(6.022)
Observations	39	39	39
R^2	0.994	0.997	0.998
Note:	*p<0.1; **p<0.05; ***p<0.01		

```
gridt <- freeny %>%
  data_grid(price.index,
            market.potential=
```

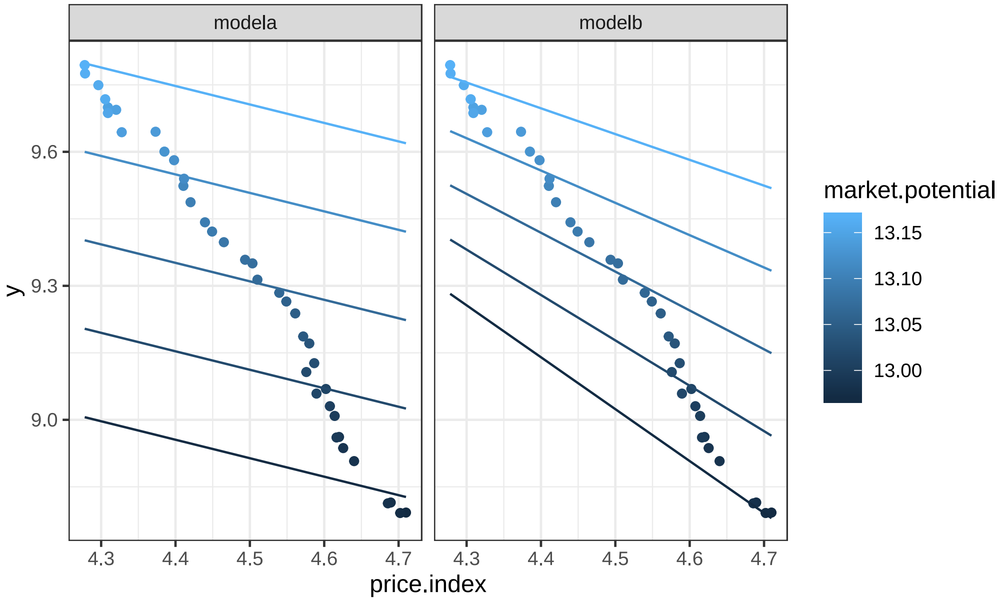


图 5: 左图无交互效应，可视为控制变量。右图为两个连续变量的交互效应

```
              seq_range(market.potential,5)) %>%
  gather_predictions(modela,modelb)



ggplot(freeny,aes(price.index,y,
                  color=market.potential))+
  geom_point()+
  geom_line(data=gridt,aes(price.index,pred,
color=market.potential,
group=market.potential))+
  facet_wrap(~model)+
  theme_bw()
```

3 多项式回归

- 多项式回归是平滑方法的基础。

```
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)

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:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_1^2$$

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_1^2 + \alpha_3 x_1^3$$

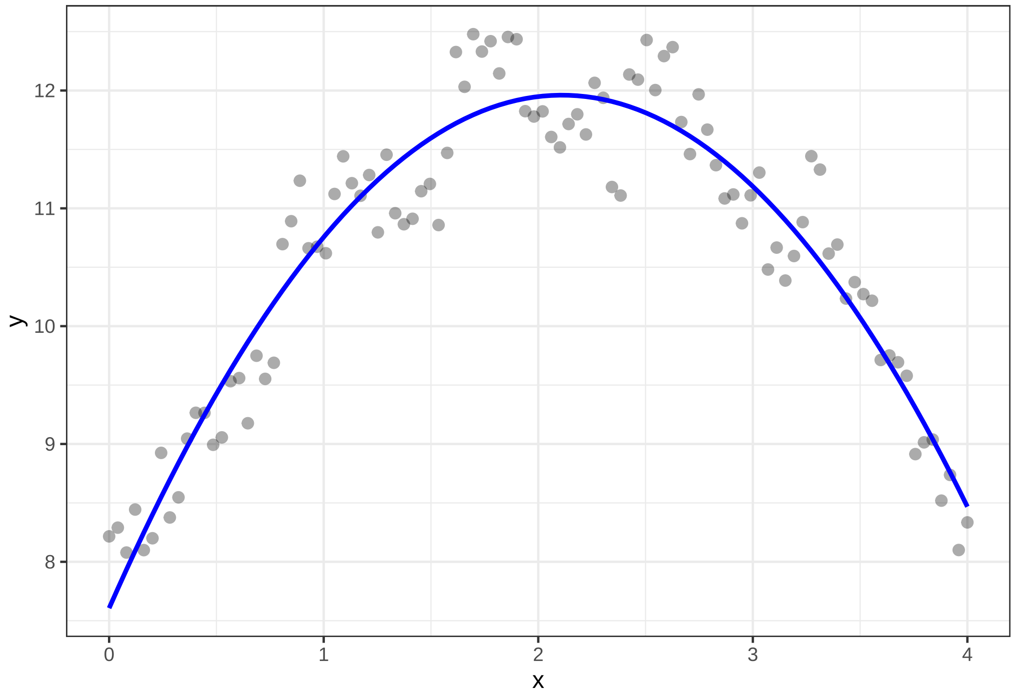


图 6: 对一个模拟数据进行二次项回归。

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

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) +
```

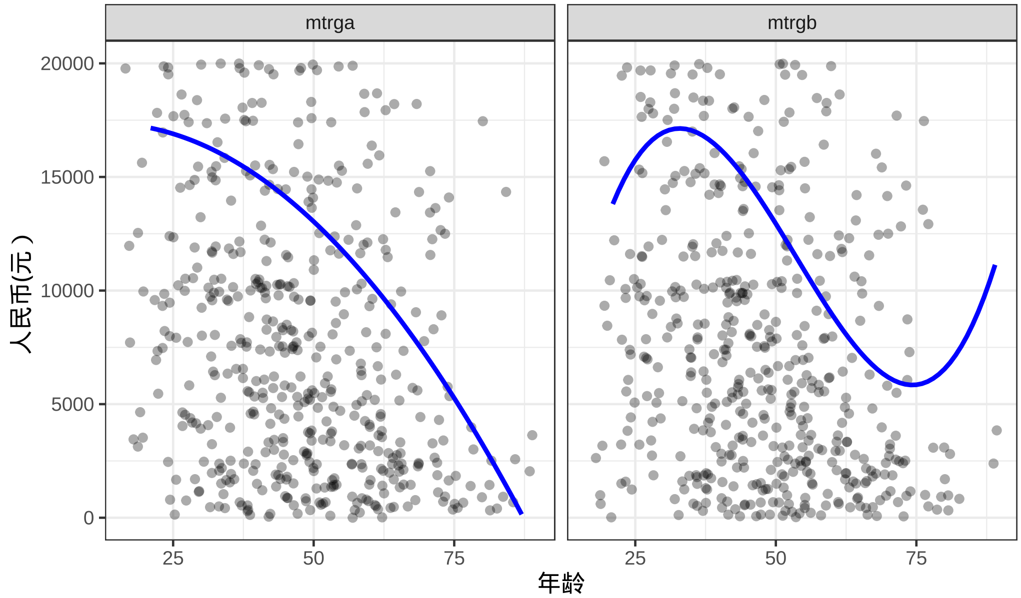


图 7: 分别对 CFPS 数据进行二次项和三次项回归。三次项导致了过拟合。

```
  ylim(0, 20000) +
  labs(x = "年龄",y = "人民币(元)") +
  theme_bw()
```

4 局部加权回归散点平滑

- Locally Weighted Scatterplot Smoother，LOWESS

$$y_i = g(x_i) + \varepsilon_i$$

g 是在 x 带宽 α 范围内进行的多项式回归。

```
data(PlantCounts,package = 'MSG')
par(mar = c(4,4,1,0.5), mfrow = c(1, 2), pch = 20)
with(PlantCounts, {
```

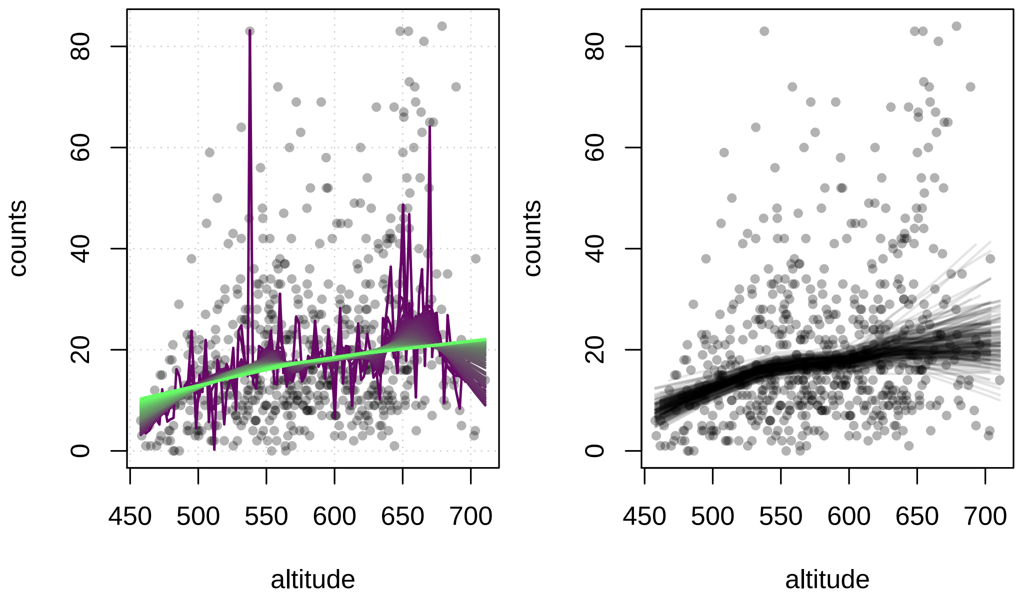


图 8: 使用 R 自带作图工具绘图。左图：设置不同带宽进行 LOWESS 回归。右图：Bootstrap 重抽样 200 次的结果。

```
plot(altitude, counts, col = rgb(0, 0, 0, 0.3),
panel.first = grid())
for (i in seq(0.01, 1, length = 70)) {
lines(lowess(altitude, counts, f = i),
      col = rgb(0.4,i, 0.4), lwd = 1.5)
 }
plot(altitude, counts, col = rgb(0, 0, 0, 0.3))
for (i in 1:200) {
idx = sample(nrow(PlantCounts), 300, T)
lines(lowess(altitude[idx], counts[idx]),
col = rgb(0, 0, 0, 0.1), lwd = 1.5)
}
})
```

- ggplot2 版本

```
g <- ggplot(PlantCounts,
            aes(altitude,counts)) +
  geom_point(size=1.5,alpha=1/3) +
  ylim(0,80)+
  theme_bw()

for (i in seq(1,1000,10)){
  col = rgb(0.4,i/1000,0.4)
  g <- g + stat_smooth(geom='line',
                       span=i/1000,
                       size=0.5,
                       se=F,color=col)
}

f <- ggplot(PlantCounts,
            aes(altitude, counts)) +
  geom_point(size=1.5,alpha=1/3) +
  ylim(0,80)+
  theme_bw()

for (i in 1:200) {
idx <- sample(nrow(PlantCounts),300,T)
df <- PlantCounts[idx,]
f <- f + stat_smooth(geom='line',
```

```
                      data=df,
                      aes(altitude,counts),
                      span=1,size=0.5,
                      se=F,alpha=1/10)
}

e <- ggplot(PlantCounts,
             aes(altitude, counts)) +
  geom_point(size=1.5,alpha=1/3) +
  geom_smooth(span=1,size=1)+
  ylim(0,80)+
  theme_bw()

plot_grid(g,f,e,ncol = 2)
```

5 样条

- Splines
- 结点为 a, b, c 的样条回归函数为：

$$y = \alpha + \beta_1 x + \beta_2 (x-a)_+ + \beta_3 (x-b)_+ + \beta_4 (x-c)_+$$

$(\mu)_+ = \mu$ 当 $\mu > 0$，否则 $(\mu)_+ = 0$。

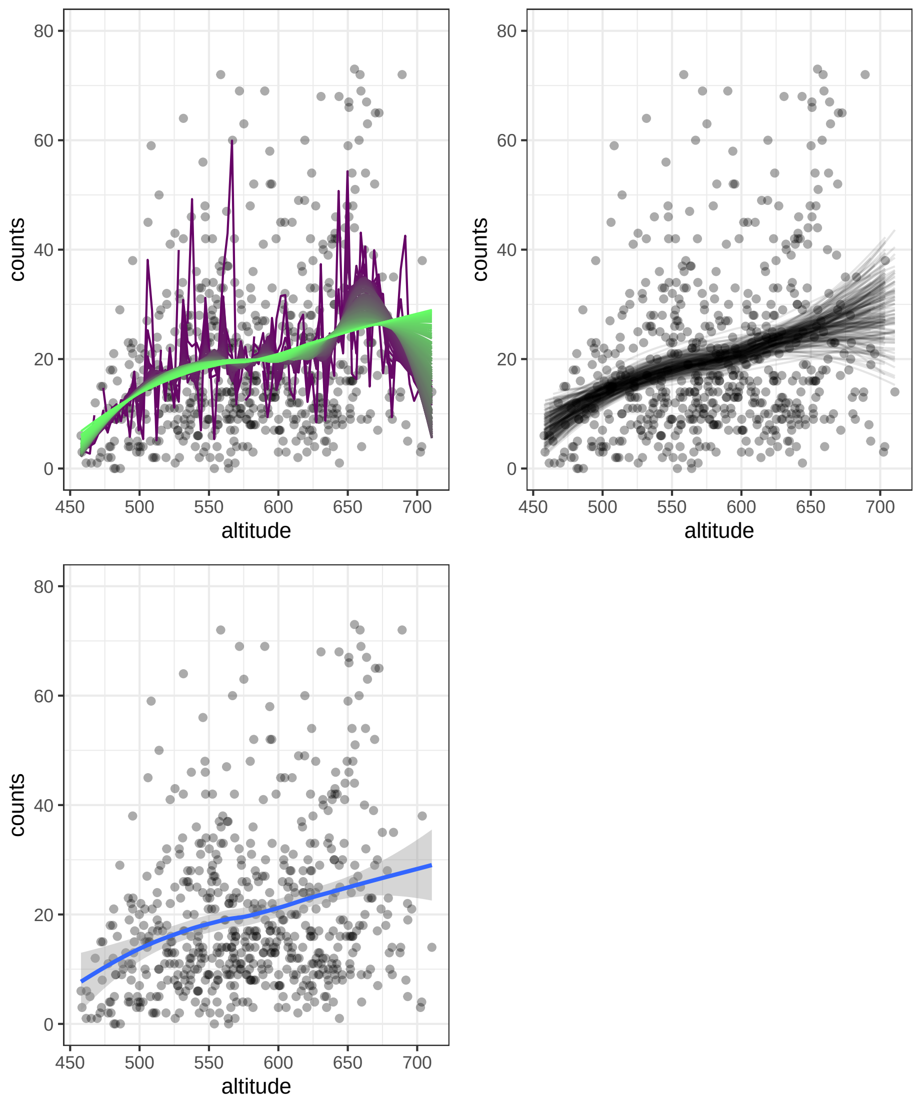


图 9: 使用 'ggplot2'和 'for'循环绘图

```r
library(ISLR)
library(splines)
data(wage,package = 'ISLR')
fita <- lm(wage ~ bs(age,degree=1,knots = c(25,40,60)),Wage)
fitb <- lm(wage ~ bs(age,knots = c(25,40,60)),Wage)
summary(fita)
```

```
Call:
lm(formula = wage ~ bs(age, degree = 1, knots = c(25, 40, 60)),
    data = Wage)

Residuals:
    Min      1Q  Median      3Q     Max
-99.795 -24.686  -4.856  15.344 204.671

Coefficients:
                                            Estimate Std. Error t value
(Intercept)                                   54.333      5.957   9.120
bs(age, degree = 1, knots = c(25, 40, 60))1   37.645      6.817   5.522
bs(age, degree = 1, knots = c(25, 40, 60))2   65.847      6.019  10.940
bs(age, degree = 1, knots = c(25, 40, 60))3   63.850      6.319  10.104
bs(age, degree = 1, knots = c(25, 40, 60))4   33.772     10.580   3.192
                                            Pr(>|t|)
(Intercept)                                  < 2e-16 ***
bs(age, degree = 1, knots = c(25, 40, 60))1 3.64e-08 ***
```

```
bs(age, degree = 1, knots = c(25, 40, 60))2  < 2e-16 ***
bs(age, degree = 1, knots = c(25, 40, 60))3  < 2e-16 ***
bs(age, degree = 1, knots = c(25, 40, 60))4  0.00143 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.91 on 2995 degrees of freedom
Multiple R-squared:  0.08665,   Adjusted R-squared:  0.08543
F-statistic: 71.03 on 4 and 2995 DF,  p-value: < 2.2e-16
```

```
summary(fitb)
```

```


Call:
lm(formula = wage ~ bs(age, knots = c(25, 40, 60)), data = Wage)

Residuals:
    Min      1Q  Median      3Q     Max 
-98.832 -24.537  -5.049  15.209 203.207 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       60.494      9.460   6.394 1.86e-10 ***
bs(age, knots = c(25, 40, 60))1    3.980     12.538   0.317 0.750899    
bs(age, knots = c(25, 40, 60))2   44.631      9.626   4.636 3.70e-06 ***
bs(age, knots = c(25, 40, 60))3   62.839     10.755   5.843 5.69e-09 ***
bs(age, knots = c(25, 40, 60))4   55.991     10.706   5.230 1.81e-07 ***
```

```
bs(age, knots = c(25, 40, 60))5   50.688     14.402   3.520 0.000439 ***
bs(age, knots = c(25, 40, 60))6   16.606     19.126   0.868 0.385338
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.92 on 2993 degrees of freedom
Multiple R-squared:  0.08642,   Adjusted R-squared:  0.08459
F-statistic: 47.19 on 6 and 2993 DF,  p-value: < 2.2e-16
```

```
grid <- Wage %>%
  data_grid(age) %>%
  gather_predictions(fita,fitb)

ggplot(Wage,aes(age,wage))+
  geom_point(size=1,alpha=1/7)+
  geom_line(data=grid,aes(age,pred),
            size=1,color='purple')+
  facet_wrap(~model)+
  theme_bw()
```

6 Box-Cox 变换

为保证变量的正态性进行的统计学转换。

$$y(\lambda)=\begin{cases}\frac{y^\lambda-1}{\lambda}, & \text{if} \quad \lambda\neq 0\\ \log y, & \text{if} \quad \lambda=0\end{cases}$$

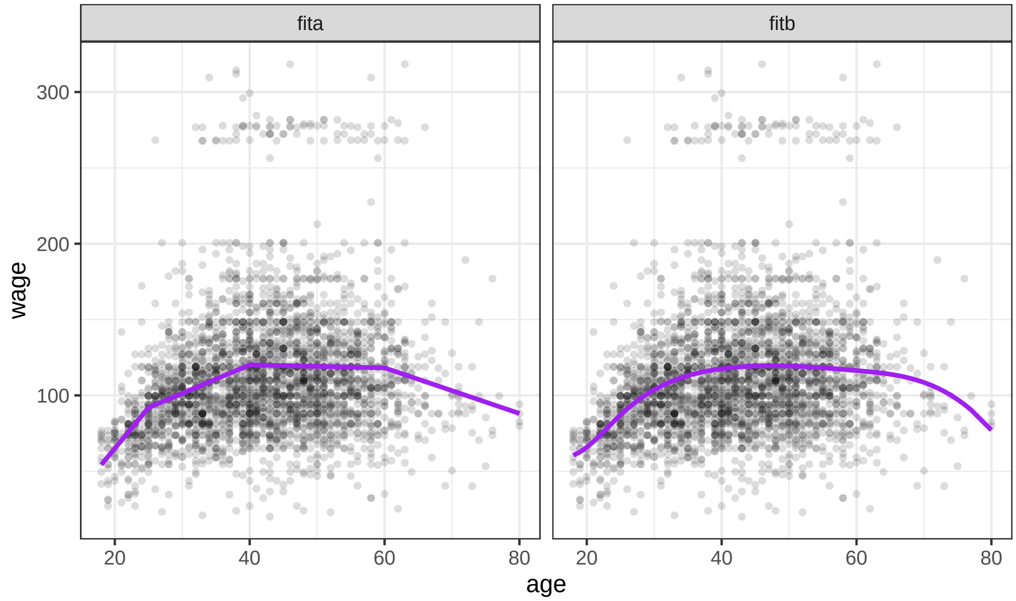
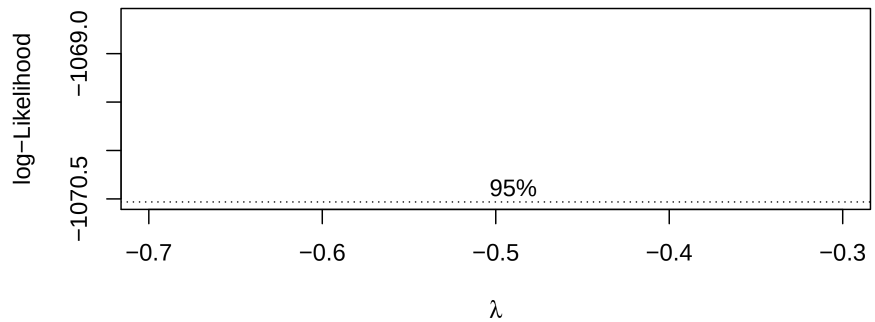


图 10: 左图：一次项样条。右图：三次项样条

$$y(\lambda)=\begin{cases}\frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \text{if} \quad \lambda_1\neq 0\\ \log(y+\lambda_2), & \text{if} \quad \lambda_1=0\end{cases}$$

```
library(MASS)
x = rf(500,30,30)
boxa <- ggplot(as.data.frame(x),aes(x))+
  geom_histogram(color='black',
                 fill='white')


result = boxcox(x~1,lambda = seq(-0.5,0,5))
```




```r
mylambda = result$x[which.max(result$y)]
mylambda
```

```
[1] -0.5
```

```r
x2 <- (x^mylambda-1)/mylambda
x2 = as.data.frame(x2)

boxb <- ggplot(x2,aes(x2)) +
  geom_histogram(color='black',
                 fill='white')
```

```r
plot_grid(boxa,boxb,ncol = 2)
```

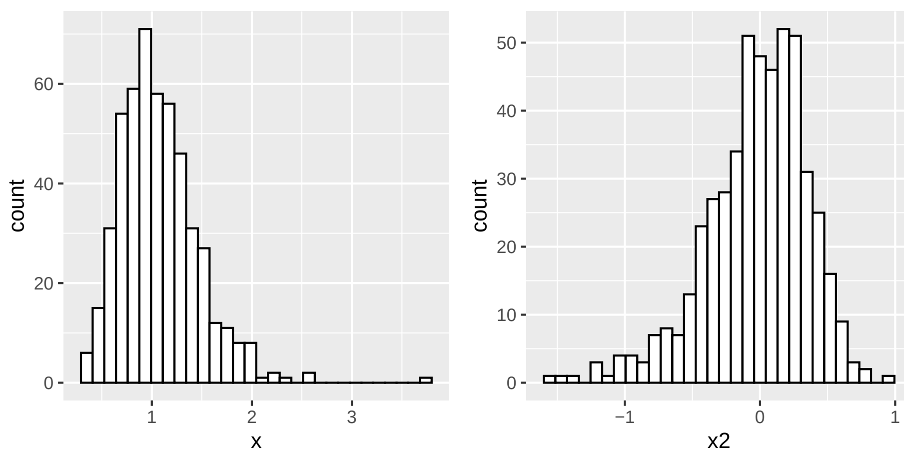


图 11: 左图：原分布。右图：变换后的分布