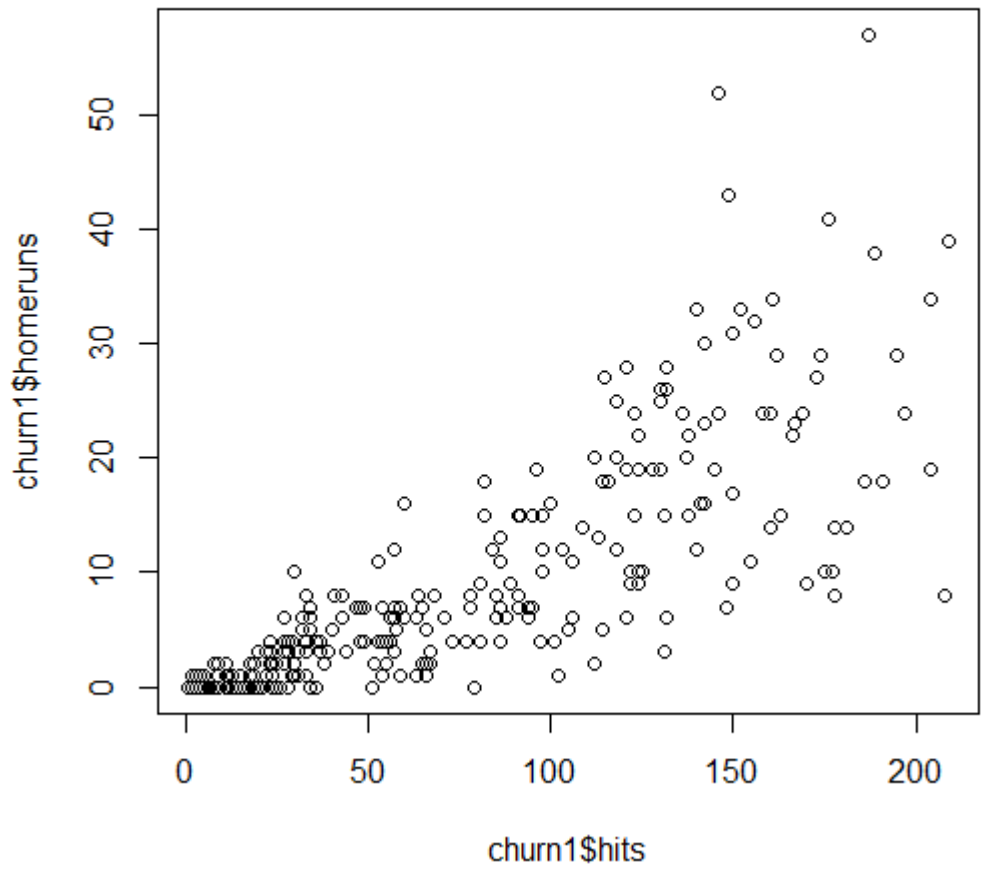
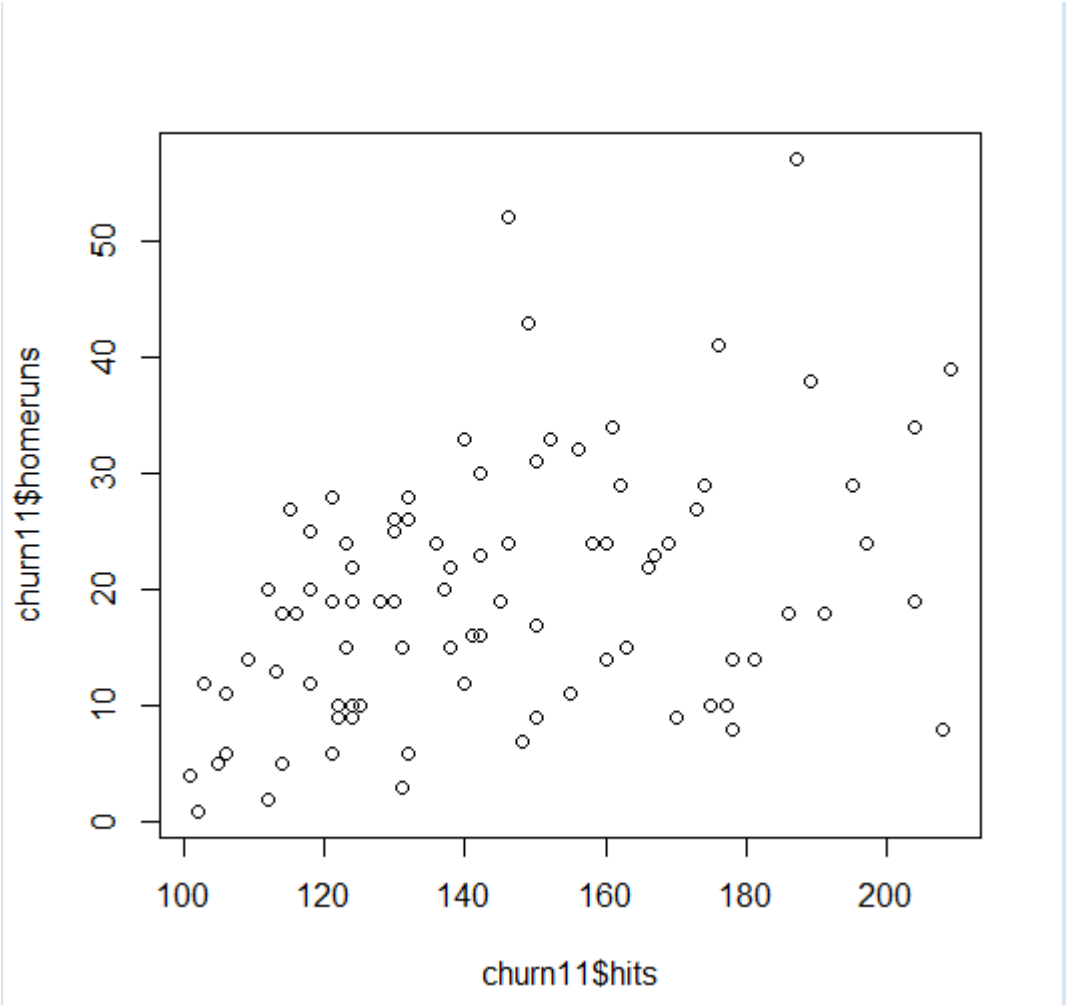
第 8 章

39、

```
> churnl1<-churnl[churnl$hits>100,]
> head(churnl1)
  firstname lastname age team games at_bats runs hits doubles triples
1   Alfonso  Soriano  24  NYY   156     696  128  209      51       2
2    Miguel   Tejada  26  OAK   162     662  108  204      30       0
3   *Ichiro   Suzuki  28  SEA   157     647  111  208      27       8
4     Derek    Jeter  28  NYY   157     644  124  191      26       0
5   *Garret Anderson  30  ANA   158     638   93  195      56       3
6   #Carlos  Beltran  25  KCR   162     637  114  174      44       7
  homeruns RBIs walks strikeouts bat_ave on_base_pct slugging_pct
1       39  102    23        157   0.300       0.332        0.547
2       34  131    38         84   0.308       0.354        0.508
3        8   51    68         62   0.321       0.388        0.425
4       18   75    73        114   0.297       0.373        0.421
5       29  123    30         80   0.306       0.332        0.539
6       29  105    71        135   0.273       0.346        0.501
  stolen_bases caught_stealing
1           41              13
2            7               2
3           31              15
4           32               3
5            6               4
6           35               7
> plot(churnl$hits,churnl$homeruns)
```



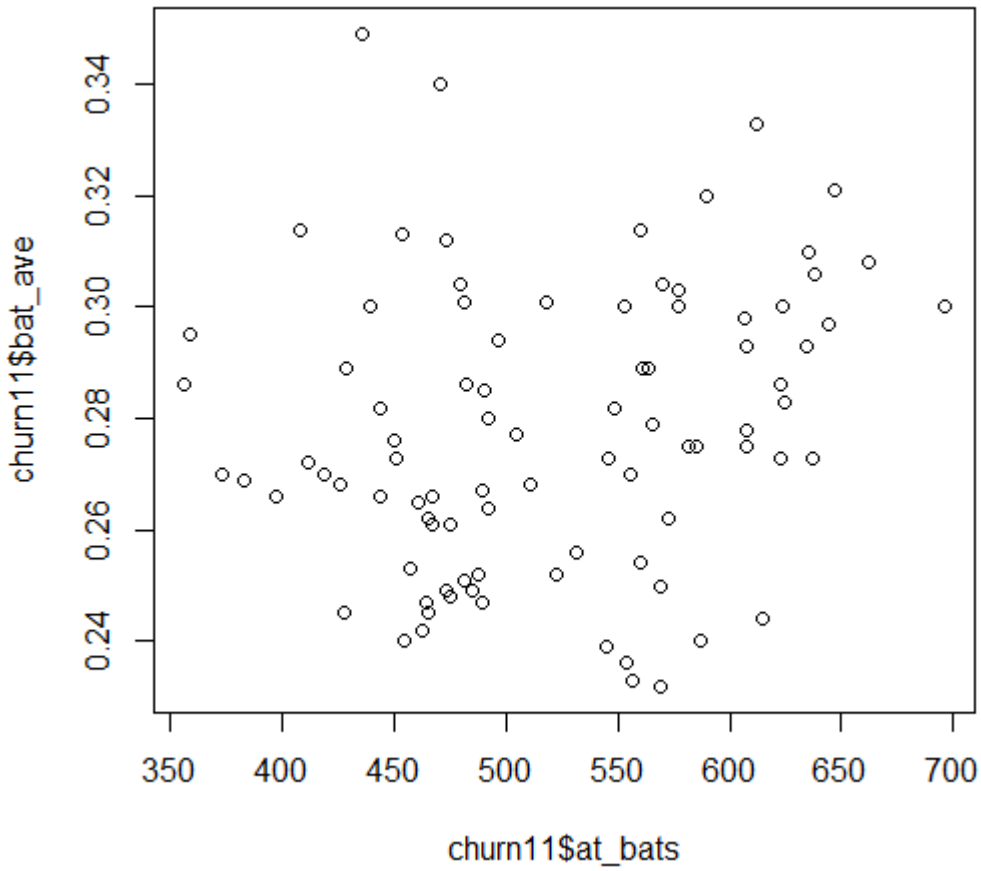

```
> plot(churnll$hits,churnll$homeruns)
```

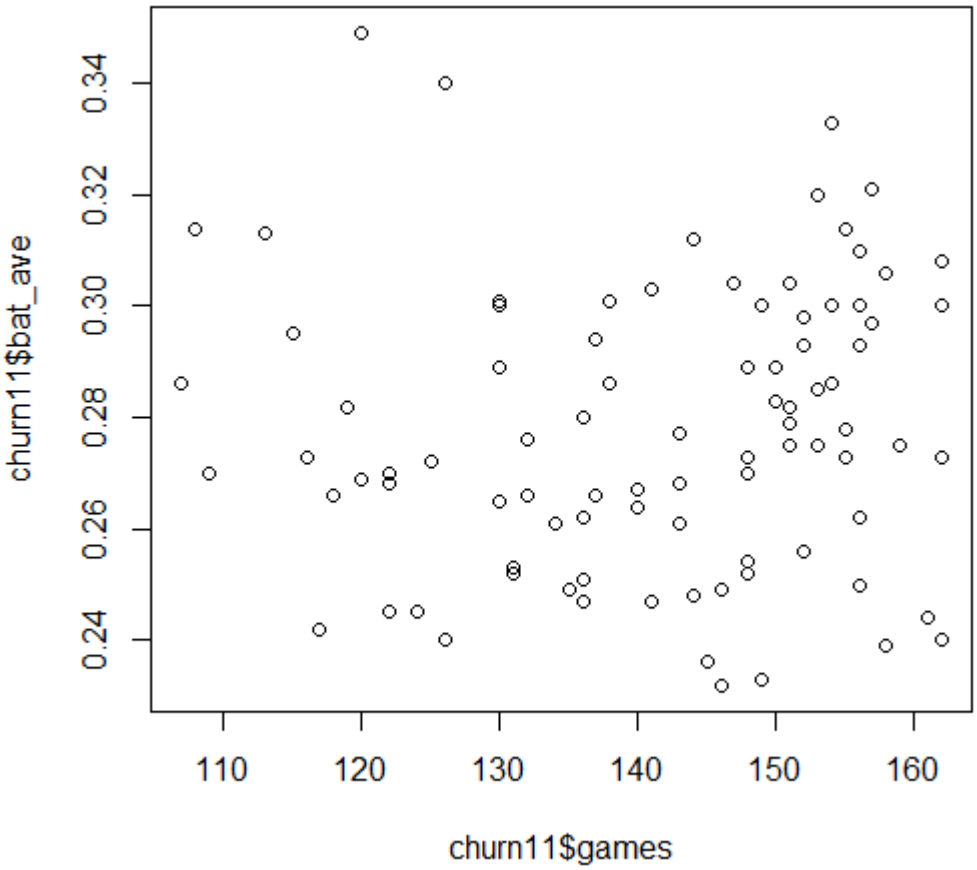


分析：首先去除击球数小于 100 的数据，分别画出去除前和去除后的数据再观察。

40、

```
> plot(churn11$at_bats,churn11$bat_ave)
> plot(churn11$games,churn11$bat_ave)
```

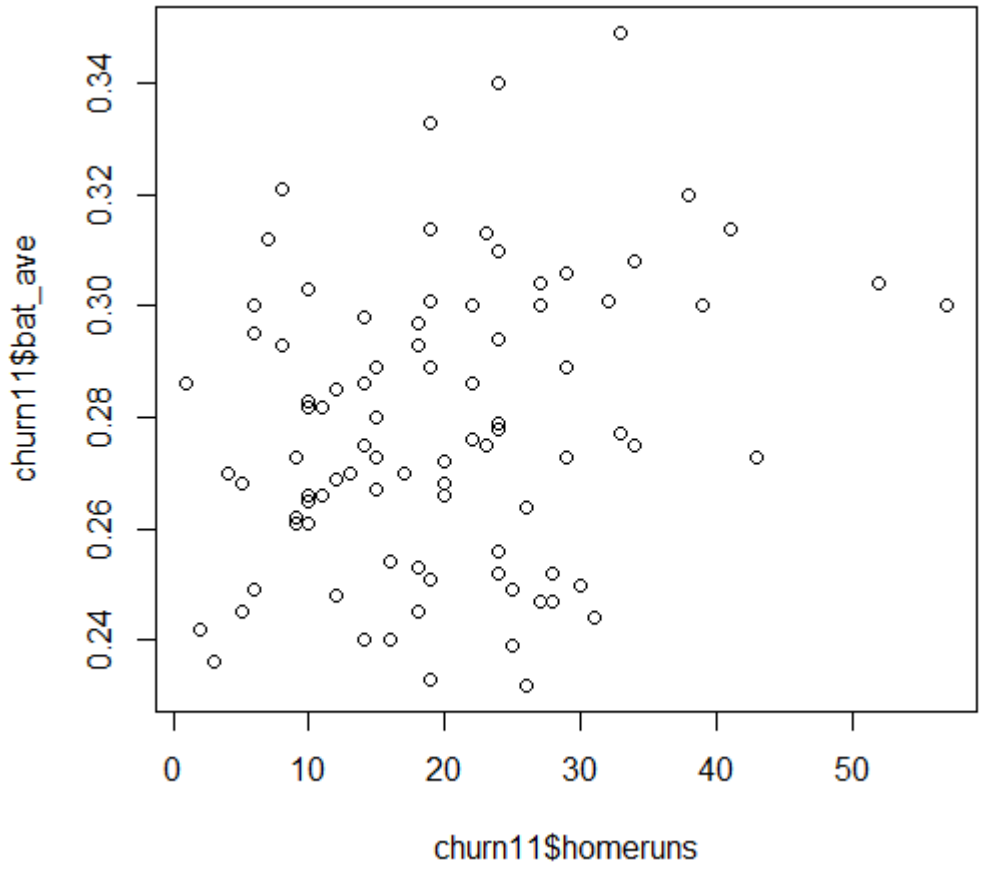

churn11$bat_ave
0.24
0.26
0.28
0.30
0.32
0.34
350
400
450
500
550
600
650
700
churn11$at_bats



分析：画出比赛次数和总击球数与平均击球率之间的关系的散点图，可以看出虽然变化不太明显，但是击球率是随着这两个变量在渐渐上升的，说明经验和技术的锻炼会使击球率变高。

41、

```
> plot(churn11$homeruns,churn11$bat_ave)
```



分析：可以从上面的图中观察到上升的趋势。

42、

多元线性回归假设。偏回归图：在简单线性回归（一个 X 一个 Y）中，我们画出自变量和因变量的散点图大致可以判断是否为线性关系。但是在多元线性回归中，我们不能再用这种一个自变量和一个因变量的 bivariate plot，因为它没有控制其他自变量的影响，而是应该用偏回归图，因为多组变量之间的关系可能并不是线性的。

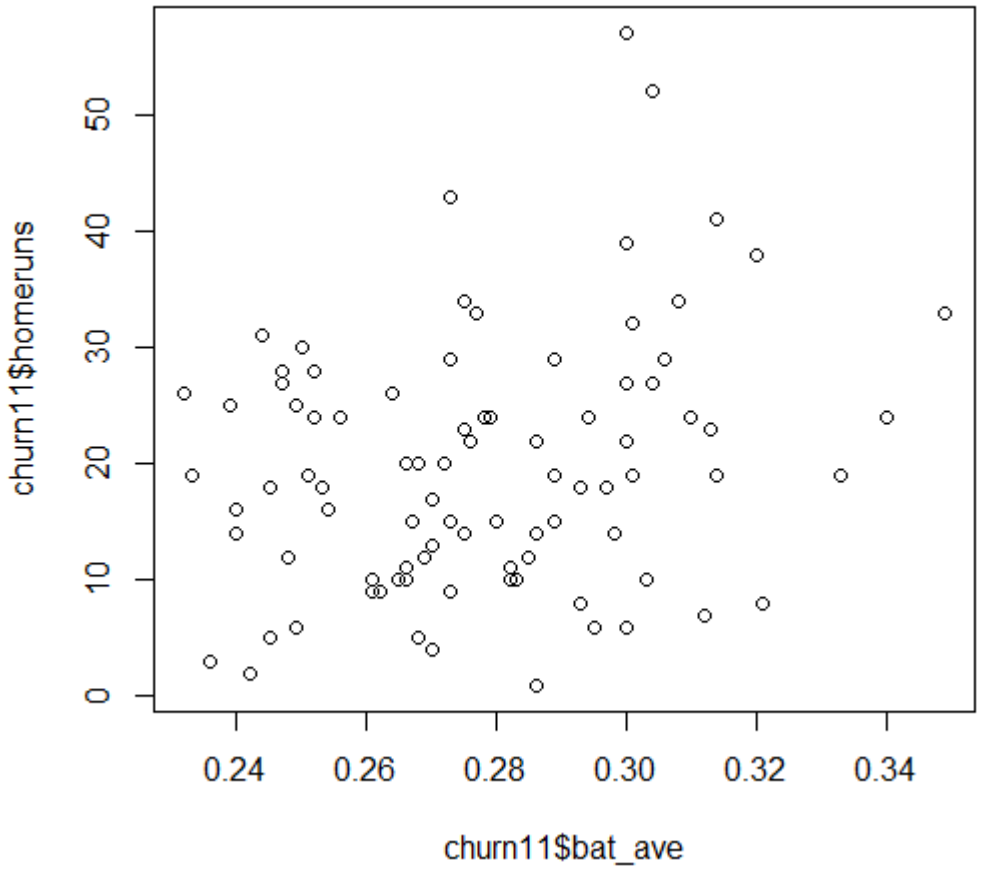
1. 既然是线性模型，那关系必然是**线性**的。
2. 误差与自变量不相关
3. 方差齐性 homoscedasticity (equal variance of ui)

4. 误差之间不相关

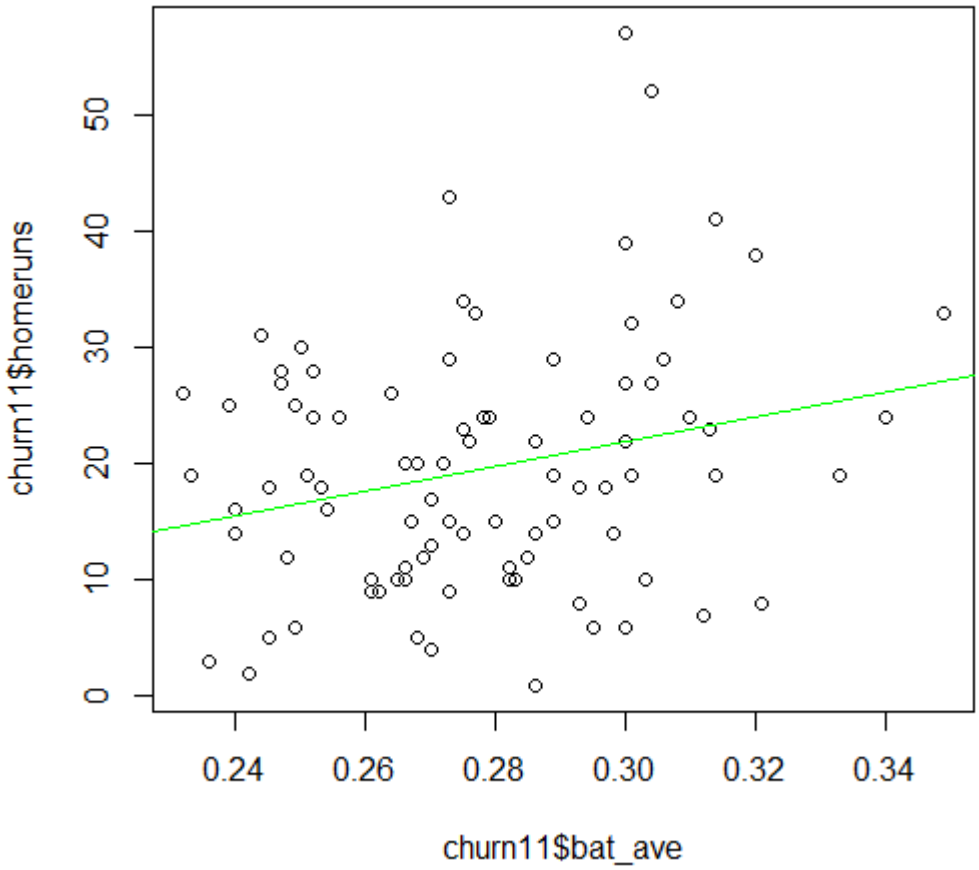
5. 误差正态分布 normality disturbance

43、

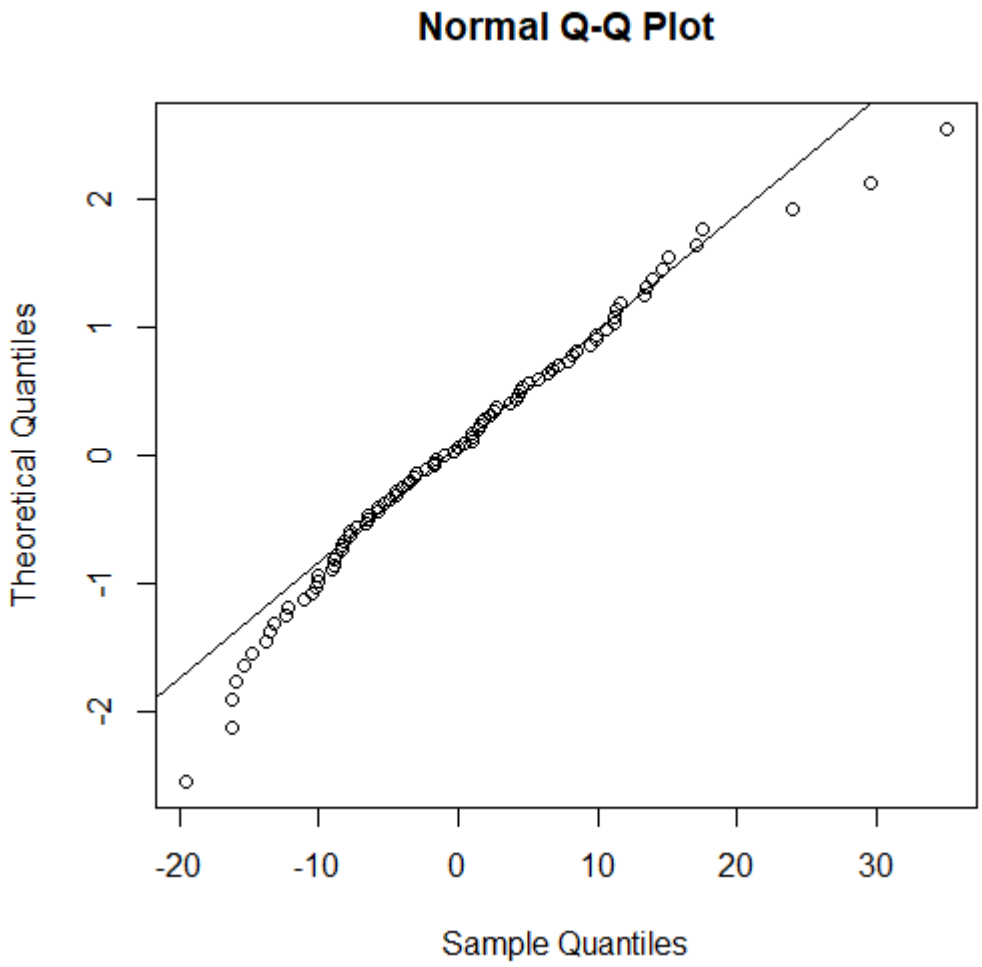
```
> bian=lm(homeruns~bat_ave,data=churn11)
> summary(bian)

Call:
lm(formula = homeruns ~ bat_ave, data = churn11)

Residuals:
    Min      1Q  Median      3Q     Max
-19.455  -8.215  -1.042   6.745  35.052

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -10.05      12.30  -0.817   0.4162
bat_ave       106.66      44.11   2.418   0.0177 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.6 on 87 degrees of freedom
Multiple R-squared:  0.06298,   Adjusted R-squared:  0.05221
F-statistic: 5.847 on 1 and 87 DF,  p-value: 0.01769

> plot(churn11$bat_ave,churn11$homeruns)
```




```
> abline(bian,col='green')
```



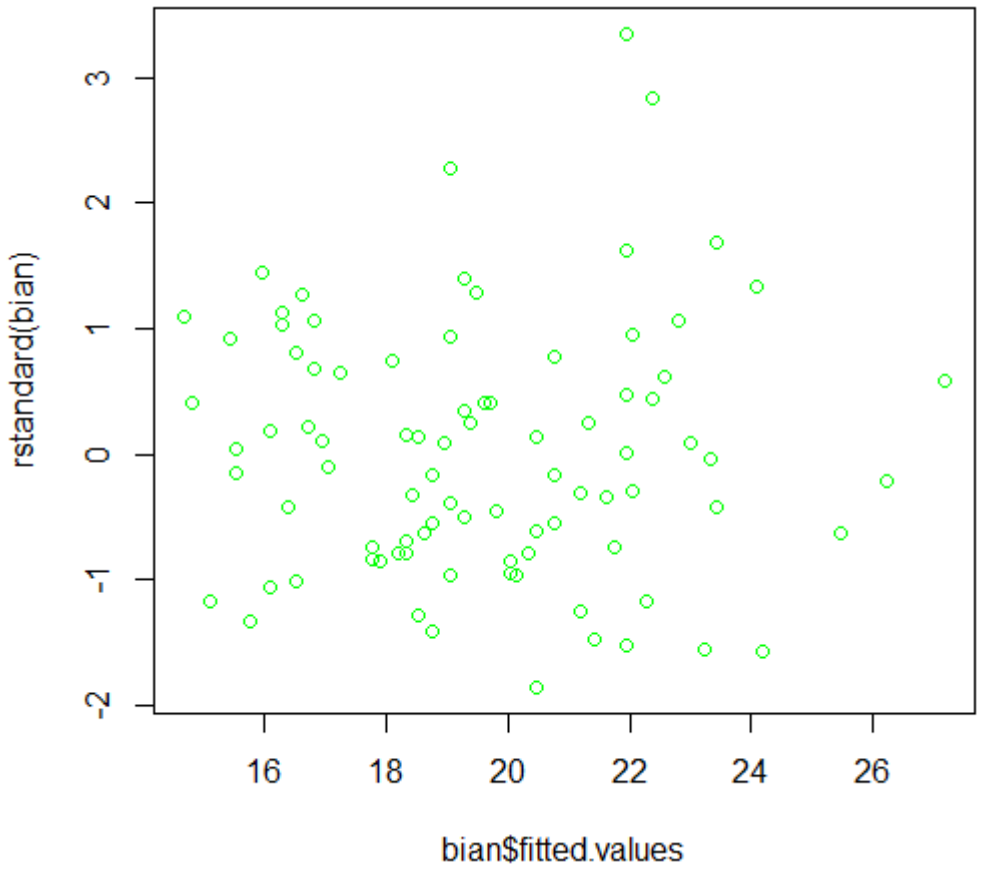

```
> qqnorm(bian$residuals,datax = TRUE)
> qqline(bian$residuals,datax = TRUE)
```

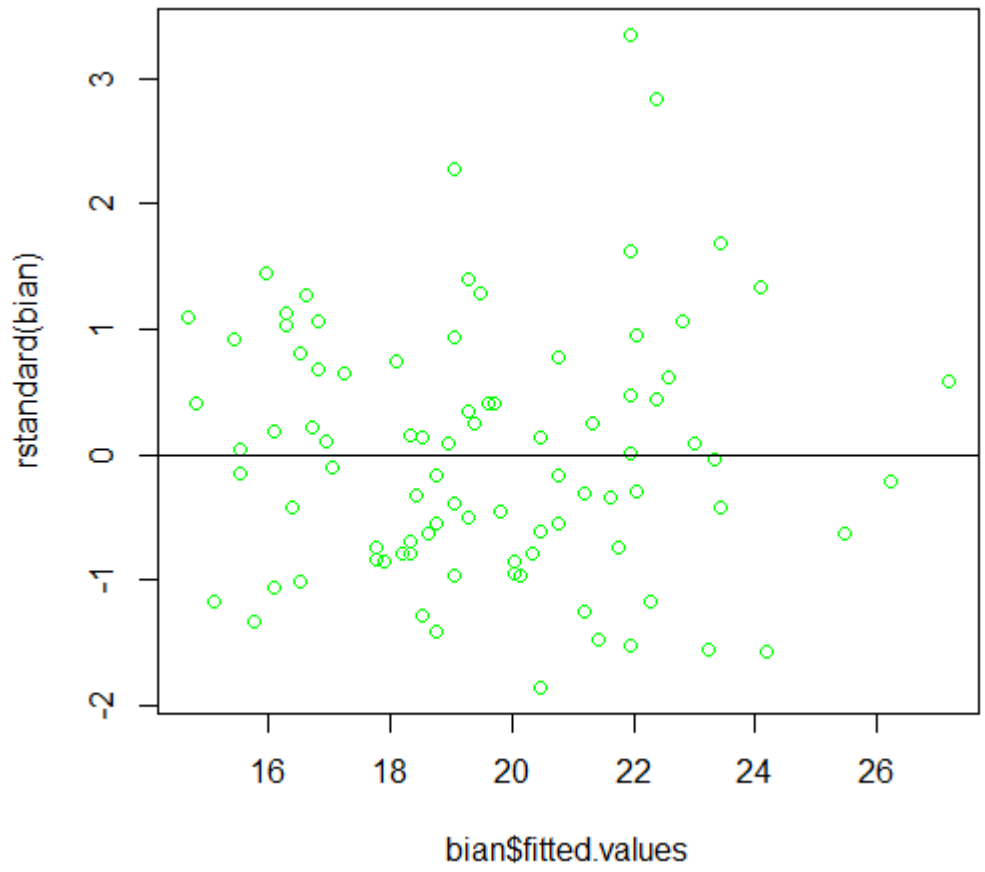


分析:首先执行本垒打与平均击球次数的回归，再画出本垒打和平均击球吃书的散点图再拟合。之后获取残差分布的正态分布概率分布图。

44、

```
> plot(bian$fitted.values,rstandard(bian),col = 'green')
> abline(0,0)
```

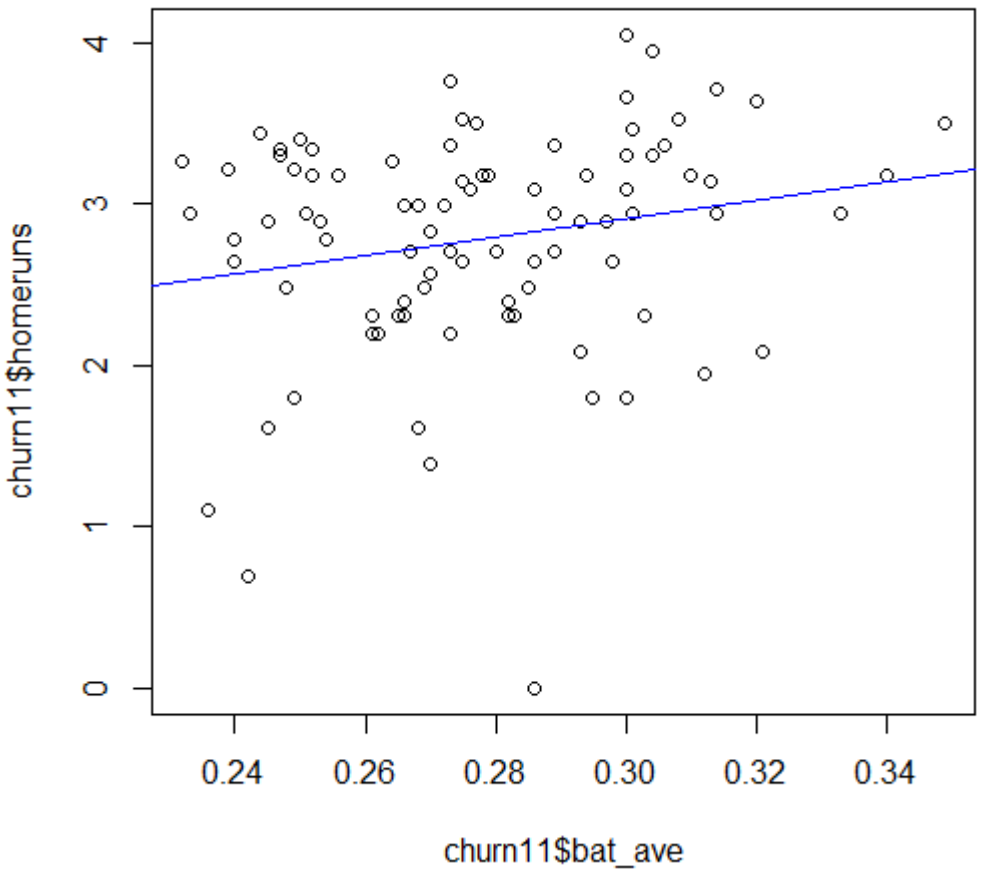

rstandard(bian)
3
2
1
0
-1
-2
16
18
20
22
24
26
bian$fitted.values


rstandard(bian)
3
2
1
0
-1
-2
16
18
20
22
24
26
bian$fitted.values

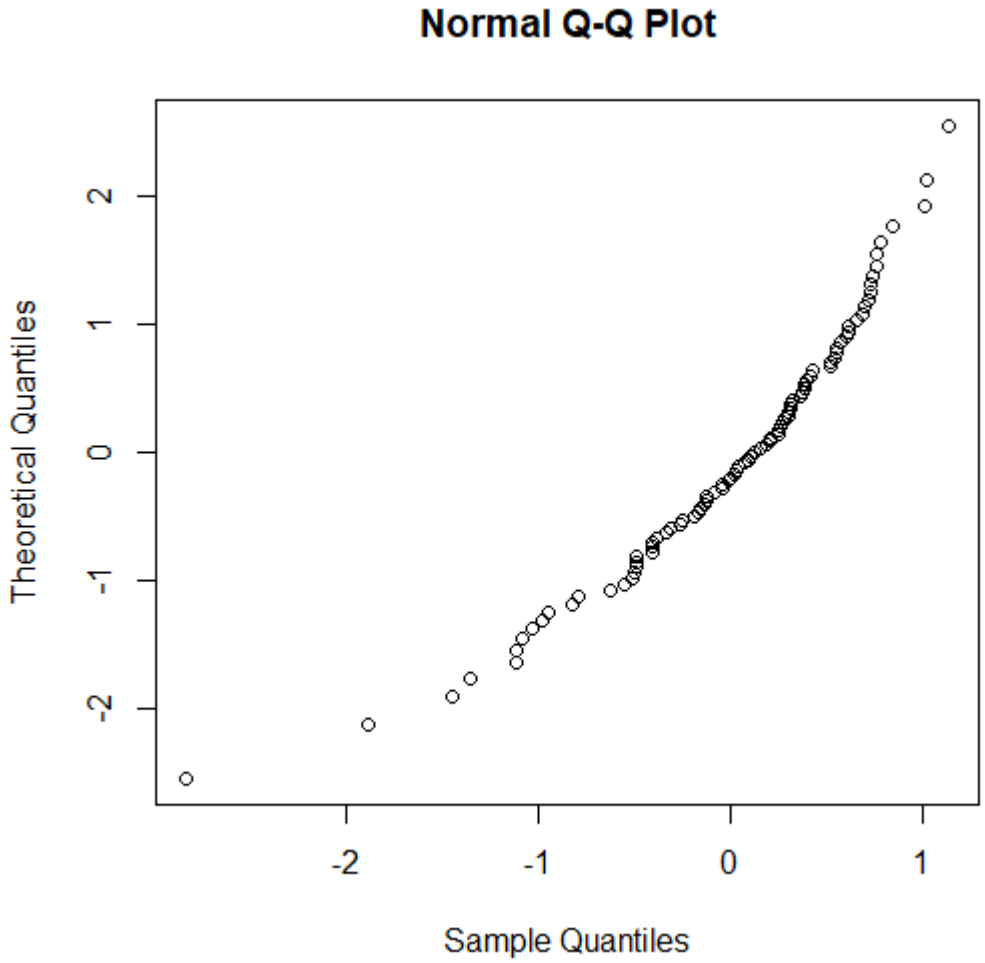
分析：它违背了方差是常数的假设。残差随 x 值的变化而变化。X 值越小，残差越小，X 值越大。这违背了方差是常数的假设。

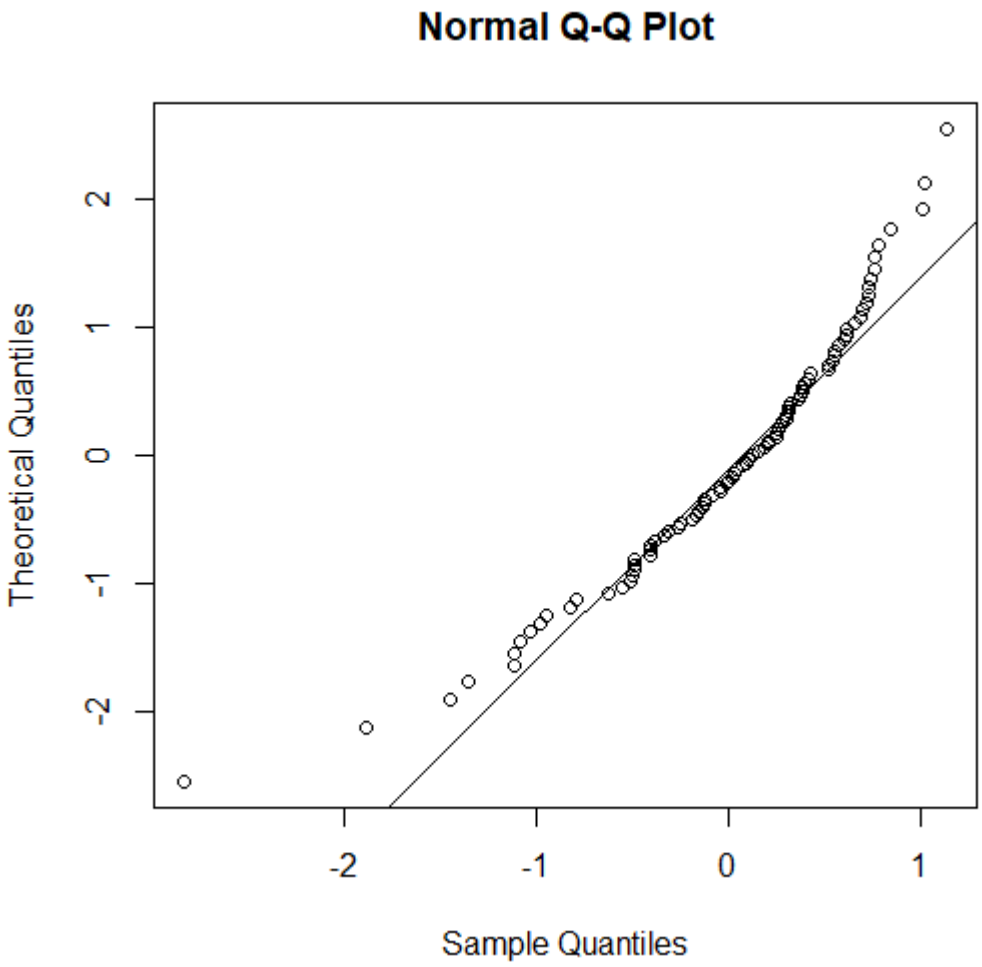
45、

```
> churn11$homeruns<-log(churn11$homeruns)
> churn11$homeruns[is.infinite(churn11$homeruns)]<-0
> head(churn11)

  homeruns |   homeruns
1       39   1 3.663562
2       34   2 3.526361
3        8   3 2.079442
4       18   4 2.890372
5       29   5 3.367296
6       29   6 3.367296

> plot(churn11$bat_ave,churn11$homeruns)
> abline(bianl,col='blue')
```




```
> qqnorm(bianl$residuals,datax = TRUE)
> qqline(bianl$residuals,datax = TRUE)
```


Normal Q-Q Plot
Theoretical Quantiles
Sample Quantiles
-2
-1
0
1
2
-2
-1
0
1



分析：取自然对数，将计算结果为-inf 的所有数据替换为 0。然后绘制线性拟合后的散点图，并在其上绘制直线图，最后绘制图。观察到预期正态分布。