# 回归应用：葡萄酒价格的解释

By：仇昊晨 20140910204
Class of Econometrics

Oct. 2017

# 1 Introduction

这篇文章是三年级计量课的作业，作者探究葡萄酒价格的影响因素，使用 MIT 提供的数据，该数据显示了葡萄酒(Chateau Latour) 在 1986 年伦敦拍卖会上的价格。这些酒的封装年份从 1952 年到 1983 年不等。同时数据包括在该酒在生产地法国 Chateau Latour 每年的的气温与降水。此外基本的回归模型和 R 语言的应用构成的探究的主要手段。
尽管有很多因素，葡萄酒的酿造期是决定酒的质量的关键。这意味着酒保存的时间越长，酒的价格越高。此外，用以酿造的葡萄质量不可忽视，葡萄的质量取决于当地的气候，主要表现在降水与气温上，一般来说：上等的葡萄来自干燥且温热的环境。

# 2 Outline of model

2.1 我们先通过标准的一元回归模型来介入问题，即探究葡萄酒价格和酒的珍藏时间的关系。
令 t 为酒封装的时间（年），令 T 为酒卖出的时间（年）。x = T − t,x 则是酒的珍藏时间（年）
令 y 是酒卖出的价格
构建总体回归模型：

$$\log(y) = \beta_0 + \beta_1 x + \epsilon$$

2.2 我们试图加入降水变量
此时总体回归模型：

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 \log(x_2) + v$$

2.3 再加入气温变量时，总体回归模型：

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 \log(x_2) + \beta_3 \log(x_3) + u$$

# 3 Run the Regression

3.1 总体回归模型 1：

$$\log(y) = \beta_0 + \beta_1 x + \epsilon$$

运行结果：hat(log(y)) = 5.49129 + 0.02194x

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.49129    0.23705  23.165   <2e-16 ***
TIME         0.02194    0.01147   1.913   0.0657 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5984 on 29 degrees of freedom
Multiple R-squared:  0.112,  Adjusted R-squared:  0.0814
F-statistic: 3.658 on 1 and 29 DF,  p-value: 0.06571
```

3.2 总体回归模型 2：

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 \log(x_2) + v$$

运行结果：hat(log(y)) = 7.98122 + 0.02411 x1 + -0.61067 log(x2)

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 7.98122 | 0.82188 | 9.711 | 1.83e-10 | *** |
| TIME | 0.02411 | 0.01007 | 2.394 | 0.02359 | * |
| log(RAIN) | -0.61067 | 0.19503 | -3.131 | 0.00405 | ** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5241 on 28 degrees of freedom

Multiple R-squared: 0.3423, Adjusted R-squared: 0.2953

F-statistic: 7.286 on 2 and 28 DF, p-value: 0.002834

### 3.3 总体回归模型 3：

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 \log(x_2) + \beta_3 \log(x_3) + u$$

运行结果：hat(log(y)) = -8.495372 + 0.034924x1 + -0.704211log(x2) + 5.618405(x3)

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -8.495372 | 5.113723 | -1.661 | 0.108225 | |
| TIME | 0.034924 | 0.009306 | 3.753 | 0.000848 | *** |
| log(RAIN) | -0.704211 | 0.170775 | -4.124 | 0.000319 | *** |
| log(TEMP) | 5.618405 | 1.726885 | 3.253 | 0.003060 | ** |

---

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4524 on 27 degrees of freedom

Multiple R-squared: 0.5275, Adjusted R-squared: 0.475

F-statistic: 10.05 on 3 and 27 DF, p-value: 0.0001281

## 4 Test

### 4.1 经济意义检验

通过 的知识，我们判断出各系数符号大小与实际相符，因此通过经济意义检验。

### 4.2 计量检验

#### 4.2.1 异方差性

通过绘制残差的分布图，发现并没有显著异方差性。详见 7.1 图 1.3。

#### 4.2.2 多重共线性

4.2.2.1　X2,x3 对 x1 回归：

运行结果：

```
Call:
lm(formula = TIME ~ log(RAIN) + log(TEMP), data = m5)

Residuals:
     Min       1Q   Median       3Q      Max
-13.352  -7.303  -0.463   6.593  16.999

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    205.57      96.31    2.13    0.042 *
log(RAIN)        2.27       3.44    0.66    0.515
log(TEMP)      -66.26      32.76   -2.02    0.053 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.19 on 28 degrees of freedom
Multiple R-squared:  0.132,  Adjusted R-squared:  0.0696
F-statistic: 2.12 on 2 and 28 DF,  p-value: 0.139
```

VIF1= 1.15 < 10

4.2.2.2　X1,x3 对 x2 回归：

运行结果：

```
Call:
lm(formula = log(RAIN) ~ TIME + log(TEMP), data = m5)

Residuals:
     Min       1Q   Median       3Q      Max
-1.2013 -0.2417 -0.0673  0.3096  0.9023

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.03056    5.65558   -0.18     0.86
TIME         0.00673    0.01022    0.66     0.52
log(TEMP)    1.70236    1.88372    0.90     0.37

Residual standard error: 0.501 on 28 degrees of freedom
Multiple R-squared:  0.033,  Adjusted R-squared:  -0.0361
F-statistic: 0.477 on 2 and 28 DF,  p-value: 0.626
```

VIF2= 1.03 < 10

4.2.2.3　X1,x2 对 x3 回归：

运行结果：

```
Call:
lm(formula = log(TEMP) ~ TIME + log(RAIN), data = m5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.08948 -0.04150  0.00265  0.03838  0.08283

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.932611   0.077631   37.78   <2e-16 ***
TIME        -0.001924   0.000951   -2.02    0.053 .
log(RAIN)    0.016648   0.018422    0.90    0.374
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0495 on 28 degrees of freedom
Multiple R-squared:  0.143,  Adjusted R-squared:  0.082
F-statistic: 2.34 on 2 and 28 DF,  p-value: 0.115
```

VIF3= 1.17 < 10

因此各变量没有强烈的多重共线性。

## 4.3　统计检验

### 4.3.1　$R^2$

比较三次回归的 adjusted$R^2$，0.0814 < 0.2953 < 0.475, 不难发现，模型三能够解释 47.5%被解释变量 log（y）的差异，比前两种模型解释得更多。

### 4.3.2　t 检验

在模型三中，t 值分别为 3.753，-4.124，3.253，对应的 p 值 0.000848，0.000319，0.003060，在 99%的置信水平下拒绝原假设（系数为 0），它们都具有统计意义的重要性。

### 4.3.3　F 检验

此处检验降水和气温是否共同具有显著影响。

运行结果：

```
Model 1: log(PRICE) ~ TIME
Model 2: log(PRICE) ~ TIME + log(RAIN) + log(TEMP)
  Res.Df     RSS Df Sum of Sq      F     Pr(>F)
1     29 10.3851
2     27  5.5256  2    4.8595 11.873 0.0001998 ***
---
```

Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
运行结果显示：f 值 11.873，p 值 0.0001998，因此在 99%的置信水平下拒绝原假设（系数同时 0），它们都具有统计意义的重要性。

并且模型三在所有系数=0 的 F 检验中，p 值为 0.0001281，故拒绝原假设，它们总体具有统计意义重要性。

# 5 Conclusion

通过以上比较与检验，不难发现，模型三有较好的适用性。我们因此以模型三

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

进行预测。

# 6 Prediction

## 6.1 均值预测

由于 1984 年的价格是缺失值，我们预测该年卖出的葡萄酒的平均价格：
代入 x1= 2, x2 = 72.5 x3 = 20.15
得出 hat(logy) =5.43

下计算 hat(logy) 区间：
令 Hat(logy) = w
1984 的预测式子

$$w = \widehat{\beta_0} + \widehat{\beta_1} c_1 + \widehat{\beta_2} c_2 + \widehat{\beta_3} c_3$$

原模型

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

变形得

$$\log(y) = w + \beta_1 (x_1 - c_1) + \beta_2 (x_2 - c_2) + \beta_3 (x_3 - c_3) + \varepsilon$$

回归后得 se(w) = 0.16471
Hat(logy)5.43 $\pm$ 1.96 * 0.16471
= 5.11 ~ 5.75
Y : 116 ~ 314
故在 95%置信水平下,1984 年葡萄酒平均价格在 116~314（£）之间。

## 6.2 个体预测

我们预测一瓶 1984 年卖出的酒的价格区间：
此时

$$se(w_2)^2 = se(w)^2 + \sigma^2$$

又

$$\widehat{\sigma^2} = SSR/(n - k - 1)$$
$$= 5.5256/27 = 0.02465$$

因此

$$\mathrm{se}(w_2) = 0.4814$$

所以：

$w_2$= 5.43 ± 1.96 * 0.4814

= 4.4865~6.3735

Y : 88.8062 ~586.1314

在 95%的置信水平下一瓶在 1984 年卖出的酒的价格区间是：88.81 ~586.13（£）之间。

# 7 Appendix

## 7.1 回归可视化

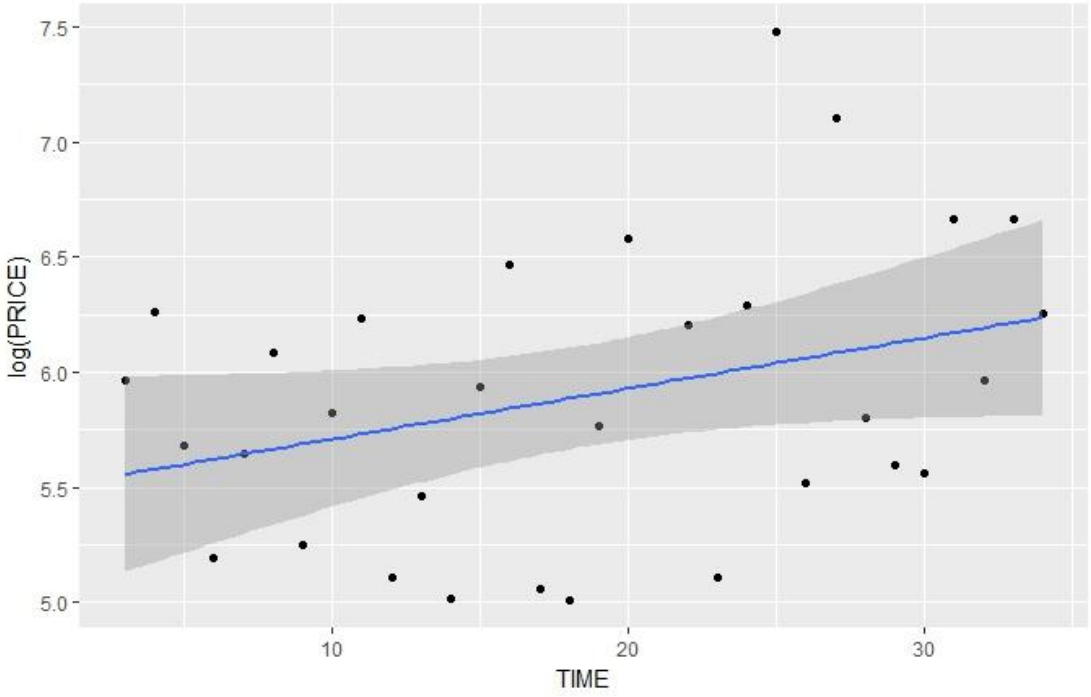


图 1.1 一元回归，时间对价格的影响

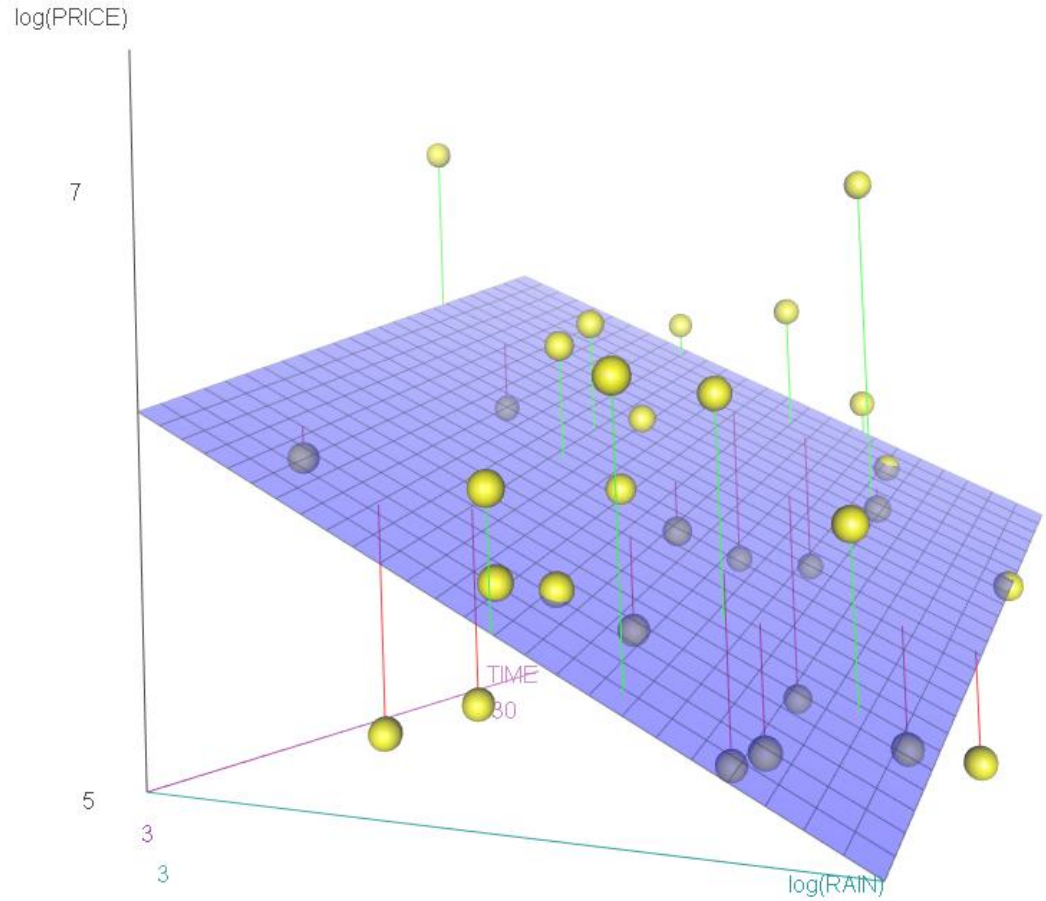


图 1.2 二元回归：时间，降水对价格的影响

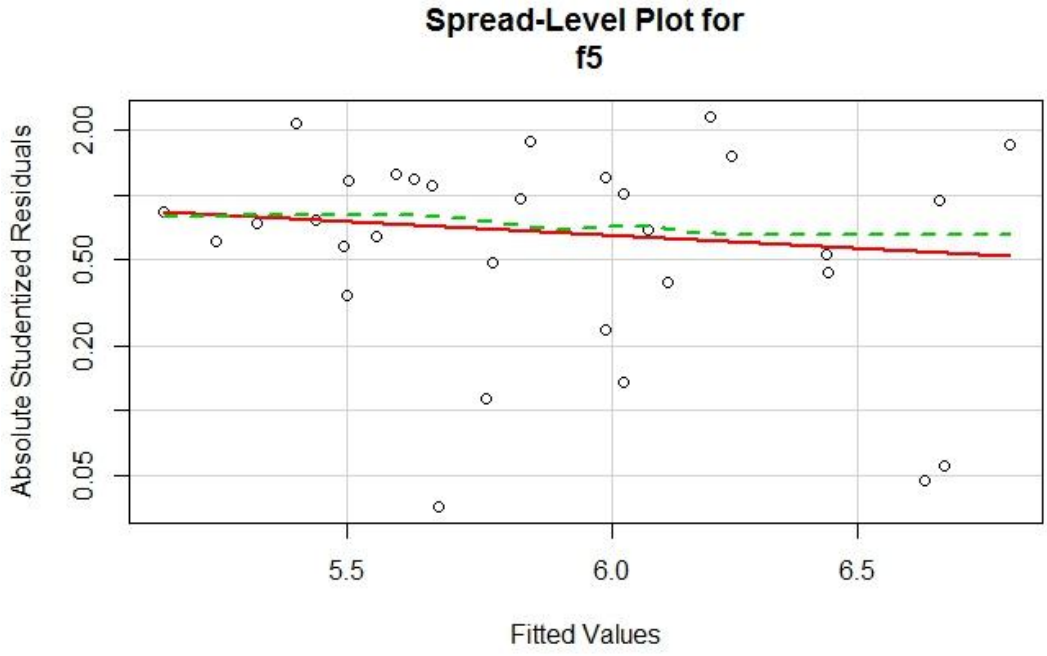


图 1.3 异方差性检查

## 7.2 主要 R 语言代码

```
[Workspace loaded from ~/.RData]

> #simple regression
> setwd("G:/R/jl1")
> m4 <- read.csv("wine3.csv")
> m5 <- m4[,c(15,16,17,18)]
> attach(m5)
> f3<- lm(log(PRICE) ~ TIME, data = m5)
> library(ggplot2)
> qplot(TIME, log(PRICE), data = m5, geom = c("point", "smooth"),
+           method = "lm")
Warning: Ignoring unknown parameters: method
> summary(f3)

Call:
lm(formula = log(PRICE) ~ TIME, data = m5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.88996 -0.48280  0.01658  0.43467  1.43725

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.49129    0.23705  23.165   <2e-16 ***
TIME         0.02194    0.01147   1.913   0.0657 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5984 on 29 degrees of freedom
Multiple R-squared:  0.112,  Adjusted R-squared:  0.0814
F-statistic: 3.658 on 1 and 29 DF,  p-value: 0.06571

> detach(m5)
> #second regression
> attach(m5)
> f4<- lm(log(PRICE) ~ TIME + log(RAIN), data = m5)
> #plot(TIME, log(PRICE))
> library("Rcmdr")
载入需要的程辑包：splines
载入需要的程辑包：RcmdrMisc
载入需要的程辑包：car
载入需要的程辑包：sandwich
```

```
载入需要的程辑包：effects
载入需要的程辑包：carData

载入程辑包：‘carData’

The following objects are masked from ‘package:car’:

    Guyer, UN, Vocab

lattice theme set by effectsTheme()
See ?effectsTheme for details.
RcmdrMsg: [1] 注意: R Commander 版本 2.4-0: Thu Oct 19 15:20:04 2017

Rcmdr 版本 2.4-0

> attach(m5)
The following objects are masked from m5 (pos = 10):

    PRICE, RAIN, TEMP, TIME

> scatter3d(TIME,log(PRICE), log(RAIN))
Loading required namespace: rgl
> summary(f4)

Call:
lm(formula = log(PRICE) ~ TIME + log(RAIN), data = m5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.77328 -0.38124 -0.03878  0.41021  1.24390

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.98122    0.82188   9.711 1.83e-10 ***
TIME         0.02411    0.01007   2.394  0.02359 *
log(RAIN)   -0.61067    0.19503  -3.131  0.00405 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5241 on 28 degrees of freedom
Multiple R-squared:  0.3423,    Adjusted R-squared:  0.2953
F-statistic: 7.286 on 2 and 28 DF,  p-value: 0.002834

> detach(m5)
```

```
> #last regression
> attach(m5)
The following objects are masked from m5 (pos = 10):

        PRICE, RAIN, TEMP, TIME

> f5<- lm(log(PRICE) ~ TIME + log(RAIN) + log(TEMP), data = m5)
> summary(f5)

Call:
lm(formula = log(PRICE) ~ TIME + log(RAIN) + log(TEMP), data = m5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.64004 -0.31454 -0.05545  0.30397  0.91058

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.495372   5.113723  -1.661 0.108225
TIME         0.034924   0.009306   3.753 0.000848 ***
log(RAIN)   -0.704211   0.170775  -4.124 0.000319 ***
log(TEMP)    5.618405   1.726885   3.253 0.003060 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4524 on 27 degrees of freedom
Multiple R-squared:  0.5275,    Adjusted R-squared:  0.475
F-statistic: 10.05 on 3 and 27 DF,  p-value: 0.0001281

> detach(m5)
> #F test
> anova(f3, f5)
Analysis of Variance Table

Model 1: log(PRICE) ~ TIME
Model 2: log(PRICE) ~ TIME + log(RAIN) + log(TEMP)
  Res.Df     RSS Df Sum of Sq      F        Pr(>F)
1     29 10.3851
2     27  5.5256  2    4.8595 11.873 0.0001998 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7.3 使用数据

| YEAR | PRICE | RAIN | TEMP | TIME |
|---|---|---|---|---|
| 1983 | 390 | 59.5 | 21.7 | 3 |
| 1982 | 525 | 81 | 20.1 | 4 |
| 1981 | 294 | 55.5 | 19.95 | 5 |
| 1980 | 180 | 37 | 18.8 | 6 |
| 1979 | 284 | 61 | 19.35 | 7 |
| 1978 | 439 | 25.5 | 18.8 | 8 |
| 1977 | 191 | 43.5 | 18.3 | 9 |
| 1976 | 338 | 123.5 | 20.95 | 10 |
| 1975 | 508 | 85.5 | 20.85 | 11 |
| 1974 | 165 | 92 | 19.55 | 12 |
| 1973 | 236 | 61.5 | 20.35 | 13 |
| 1972 | 151 | 79 | 18.6 | 14 |
| 1971 | 378 | 56 | 20.3 | 15 |
| 1970 | 644 | 44.5 | 19.3 | 16 |
| 1969 | 157 | 122 | 20.15 | 17 |
| 1968 | 150 | 146 | 19.35 | 18 |
| 1967 | 319 | 59 | 19.8 | 19 |
| 1966 | 720 | 43 | 18.15 | 20 |
| 1964 | 494 | 48 | 20.2 | 22 |
| 1963 | 165 | 77.5 | 18.45 | 23 |
| 1962 | 538 | 26 | 19.3 | 24 |
| 1961 | 1767 | 19 | 18.95 | 25 |
| 1960 | 249 | 145 | 18.1 | 26 |
| 1959 | 1218 | 93.5 | 20.4 | 27 |
| 1958 | 332 | 93.5 | 18.75 | 28 |
| 1957 | 270 | 55 | 19 | 29 |
| 1956 | 260 | 70 | 17.9 | 30 |
| 1955 | 784 | 65 | 20.6 | 31 |
| 1954 | 390 | 90 | 17.4 | 32 |
| 1953 | 786 | 40 | 19.3 | 33 |
| 1952 | 520 | 80 | 20 | 34 |