计量应用：最低工资水平与劳动力市场

By：仇昊晨 20140910204
Class of Econometrics

Dec. 2017

摘要

这篇文章是三年级计量课的最后作业，作者感兴趣的是美国一地区（Puerto Rico）的**最低工资标准**的提高是否降低其劳动力市场中**就业率**。分析的数据（1950-1987）来自 *Bureau of Labor Statistics* 和 Queens College 的网站。作者通过基本 OLS 回归，并辅以各类所学试图检验并解决模型中自相关，非平稳，异方差，变量选择等问题。最终结果表明，当其他变量不变时，最低工资水平每上升 1%，就业人口比重就下降 0.127%。

目录

一， 模型建立 ----------

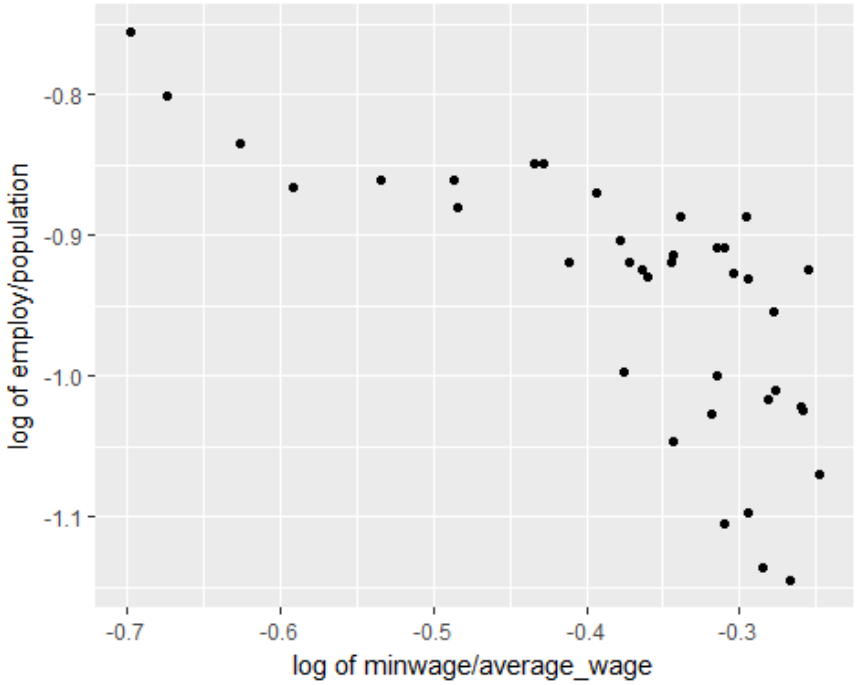
1 从图形出发：

I 数据先行处理：

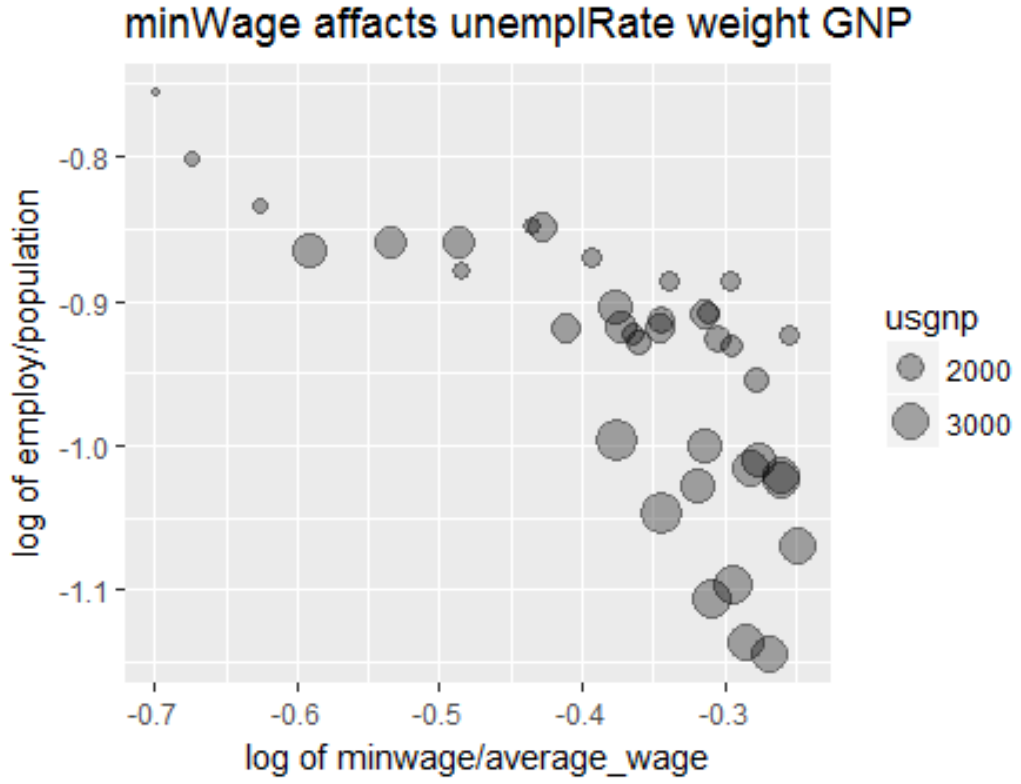
解释变量 X 是 Log（最低工资/制造业平均工资），描述了最低工资的水平，取对数便于解释，也有效压缩数据范围，减少异方差。

同样的，被解释变量 Y 是 Log（就业人口/总人口）。

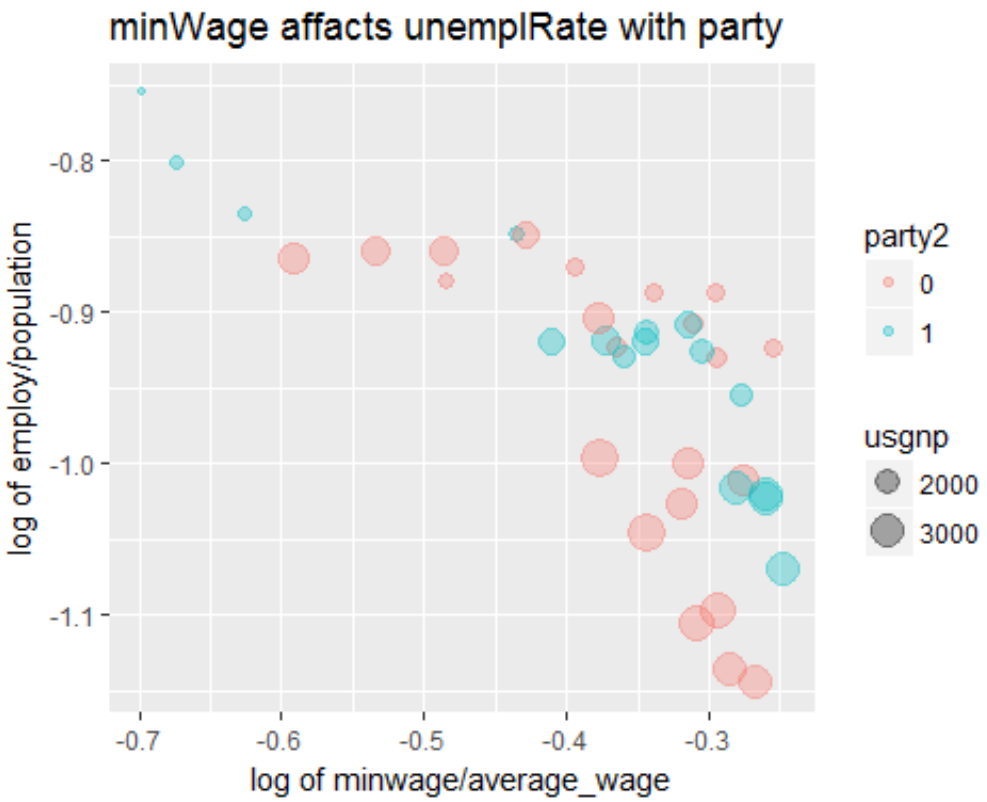
I 绘制 X，Y 的散点图：



- 加入美国人均 GNP 作为权重后，绘制 X， Y 的散点图：



- 加入美国执政党派作为虚拟变量：民主党为 0，共和党为 1：



- 猜测：不难发现，X 与 Y 有负相关，即更高的最低工资标准导致就业率的下降，并且在人均 GNP 大的时间里似乎更明显，而两党的影响差异尚未知道。以下通过 OLS 来验证。

2 OLS

- 对 $Y_t = \mathrm{a} + \mathrm{b}X_t + u_t$ 进行回归：

```
##
## Call:
## lm(formula = log(prepop) ~ log(avgmin/avgwage), data = d)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.142313 -0.034802  0.002593  0.054145  0.101104
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.15770    0.03517 -32.914  < 2e-16 ***
## log(avgmin/avgwage) -0.57380    0.09018  -6.363 2.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06467 on 36 degrees of freedom
```

```
## Multiple R-squared:  0.5293, Adjusted R-squared:  0.5163
## F-statistic: 40.49 on 1 and 36 DF,  p-value: 2.28e-07
```

❙ 加入 GNP 变量后，对 $Y_t = a + bX_t + cG_t + v_t$ 进行回归：G_t 为 log(usgnp)

```
##
## Call:
## lm(formula = log(prepop) ~ log(avgmin/avgwage) + log(usgnp),
##     data = d)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.109740 -0.020316  0.005102  0.027187  0.085724
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.10787    0.20053   0.538    0.594
## log(avgmin/avgwage) -0.37155    0.06995  -5.312 6.25e-06 ***
## log(usgnp)          -0.15424    0.02426  -6.358 2.61e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04468 on 35 degrees of freedom
## Multiple R-squared:  0.7816, Adjusted R-squared:  0.7691
## F-statistic: 62.63 on 2 and 35 DF,  p-value: 2.736e-12
```

❙ 加入虚拟变量后回归，发现虚拟变量不显著：

```
##
## Call:
## lm(formula = log(prepop) ~ log(avgmin/avgwage) + log(usgnp) +
##     party, data = d2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.107218 -0.017309  0.007152  0.028691  0.088985
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.089764   0.205569   0.437    0.665
## log(avgmin/avgwage) -0.370282   0.070729  -5.235 8.51e-06 ***
## log(usgnp)          -0.152260   0.024802  -6.139 5.69e-07 ***
## party                0.007916   0.015108   0.524    0.604
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04515 on 34 degrees of freedom
## Multiple R-squared:  0.7833, Adjusted R-squared:  0.7642
## F-statistic: 40.98 on 3 and 34 DF,  p-value: 2.165e-11
```

- 初步结论：经过回归后发现**最低工资与平均工资比率**每上升 1%，**就业率**下降 0.574%，t 值为 – 6.363，说明其具有统计意义的重要性。 加入**人均 GNP 变量**后影响略有减少（0.372%），t 值为 -5.312。所以基本验证了图形的猜测，**虚拟变量**的 p 值为 0.6，故舍去。

- **经济意义检验**：两者的数字关系符合劳动力市场中最低工资设立让企业裁员，导致失业率高的经济理论。

然而，由于宏观经济的时间序列很容易产生自相关问题，且平稳性尚未知道，故以下检验的是 Y_t = a + bX_t + cG_t + v_t 的相关问题。

二， 模型检验与修正 ----------

3 平稳性检验（单位根检验（ADF））

- 对 Y_t 进行单位根检验：

```
##
##  Augmented Dickey-Fuller Test
##
## data:  log(prepop)
## Dickey-Fuller = -3.3797, Lag order = 1, p-value = 0.0932
## alternative hypothesis: stationary
```

由上，p 值为 0.093，因此在 10%的置信水平下， Y_t 为弱平稳。

- 同样，对 X_t 进行单位根检验：

```
##
##  Augmented Dickey-Fuller Test
##
## data:  log(avgmin/avgwage)
## Dickey-Fuller = -3.9414, Lag order = 1, p-value = 0.02282
## alternative hypothesis: stationary
```

由上，p 值为 0.023，因此在 5%的置信水平下， X_t 为弱平稳。

然而在1%的置信水平下，两者都是非平稳的，下讨论在1%置信水平下，查看其是否可用误差修正模型。

- 对 Y_t 进行一次差分，再进行ADF检验，发现其在1%置信水平下弱平稳：

```
##  Augmented Dickey-Fuller Test
##
## data:  diff(diff(log(prepop)))
## Dickey-Fuller = -6.1464, Lag order = 1, p-value = 0.01
## alternative hypothesis: stationary
```

- 对 X_t 进行一次差分，再进行ADF检验，发现其在1%置信水平下弱平稳：

```
##  Augmented Dickey-Fuller Test
##
## data:  diff(diff(log(avgmin/avgwage)))
## Dickey-Fuller = -7.6832, Lag order = 1, p-value = 0.01
## alternative hypothesis: stationary
```

说明：R语言中此类ADF检验p值小于0.01会抛出提醒，而只显示0.01.

对 $Y_t = a + bX_t + u_t$ 的随机误差项进行ADF检验，结果显示，在5%的置信水平下，两者存在协整关系：

```
##  Augmented Dickey-Fuller Test
##
## data:  model1$residuals
## Dickey-Fuller = -3.799, Lag order = 1, p-value = 0.0429
## alternative hypothesis: stationary
```

构造误差修正模型 $\triangle Y_t = a + b \triangle X_t + c\gamma_{t-1} + \mu_t$ 进行回归：

```
## Call:
## lm(formula = diff(log(prepop)) ~ diff(log(avgmin/avgwage)) +
##     model1$residuals[1:37])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.082491 -0.009881 -0.001408  0.020651  0.068187
##
## Coefficients:
```

```
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.004722   0.005038  -0.937   0.3552
## diff(log(avgmin/avgwage)) -0.183798   0.073624  -2.496   0.0176 *
## model1$residuals[1:37]     -0.145279   0.079901  -1.818   0.0778 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03038 on 34 degrees of freedom
## Multiple R-squared:  0.1987, Adjusted R-squared:  0.1516
## F-statistic: 4.216 on 2 and 34 DF,  p-value: 0.02315
```

- **误差修正模型结果：**$\triangle \hat{Y}_t = -0.0047 - 0.1838 \triangle X_t - 0.1453\gamma$ ，因此 log（就业人口比重）的变化，不仅取决于最低工资水平的变化，还取决于上期 log（就业人口比重）对均衡水平的偏离，γ 的系数 -0.1453 体现了对偏离的修正。

- 以下以 10%的置信水平接受 Y_t，X_t 的弱平稳（上文的误差修正模型基于 1%的置信水平接受原假设），并进行进一步检验：

4 多重共线检验（方差膨胀因子测量，试图加入变量）

- 方差膨胀因子测量：

```
## log(avgmin/avgwage)          log(usgnp)
##           1.260707           1.260707
```

- 试图加入变量：

```
##
## Call:
## lm(formula = log(prepop) ~ log(avgmin/avgwage) + log(usgnp) +
##     prdef, data = d)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.065632 -0.018763 -0.002238  0.020351  0.084452
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.36493    0.27466  -4.969 1.88e-05 ***
## log(avgmin/avgwage) -0.37860    0.04855  -7.797 4.49e-09 ***
## log(usgnp)           0.05700    0.03791   1.504    0.142
## prdef               -0.08686    0.01397  -6.220 4.47e-07 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.031 on 34 degrees of freedom
## Multiple R-squared:  0.8978, Adjusted R-squared:  0.8888
## F-statistic:  99.6 on 3 and 34 DF,  p-value: < 2.2e-16
```

此时方差膨胀因子：

```
## log(avgmin/avgwage)          log(usgnp)              prdef
##            1.261395            6.392358            6.078295
```

- 综上分析：原始模型方差膨胀因子为 1.26，满足要求，这里试图加入**价格平减指数**变量，发现原有解释变量变得不显著了。而且，除了 log(avgmin/avgwage)，方差膨胀因子都大于 6，故舍去新变量，保持原有模型。

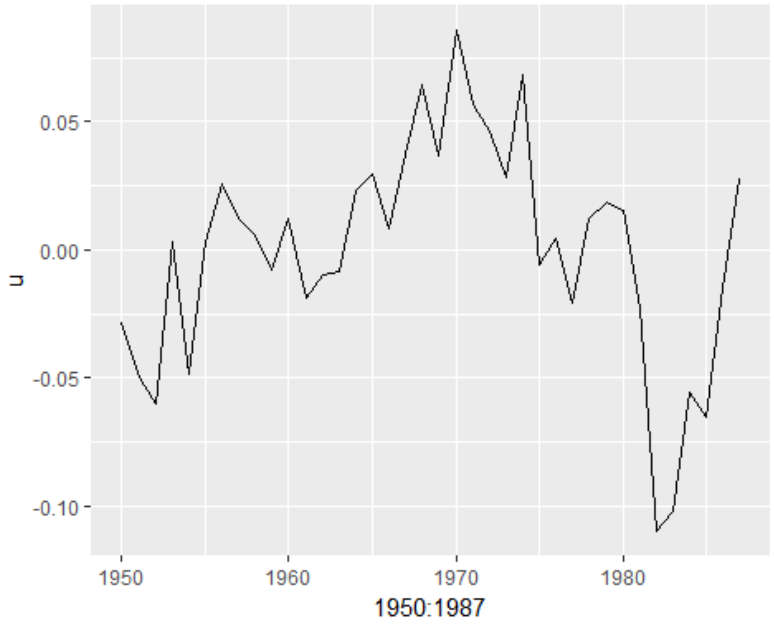
5 自相关检验（DW 检验， 图示法， BG 检验法）

DW 检验：

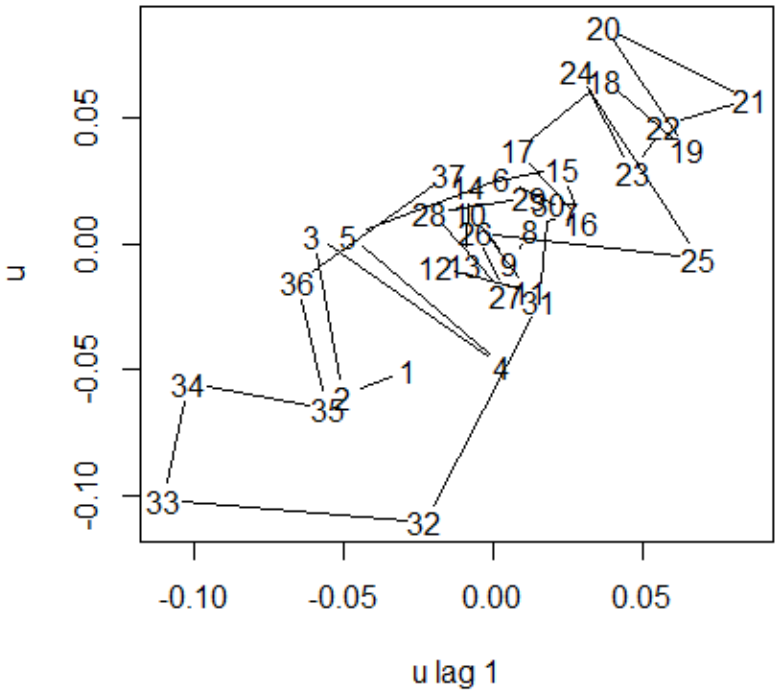
```
##
##  Durbin-Watson test
##
## data:  model2
## DW = 0.63704, p-value = 6.975e-08
## alternative hypothesis: true autocorrelation is greater than 0
```

图示法：

E(t) 随时间的波动：



e(t-1) 和 e(t) 的关系：



对上图作回归：

```
##
## Time series regression with "ts" data:
## Start = 2, End = 38
##
## Call:
```

```
## dynlm(formula = u ~ L(u), data = d)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.095590 -0.025175  0.001899  0.025121  0.059356
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.001270   0.005365   0.237    0.814
## L(u)        0.678195   0.124151   5.463 3.95e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03263 on 35 degrees of freedom
## Multiple R-squared:  0.4602, Adjusted R-squared:  0.4448
## F-statistic: 29.84 on 1 and 35 DF,  p-value: 3.947e-06
```

BG 检验法：

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  model2
## LM test = 17.5, df = 1, p-value = 2.873e-05
```

- **自相关检验结果：以上三种检验都表明该模型有一阶的正自相关问题，拟用 co 迭代进行修正：**

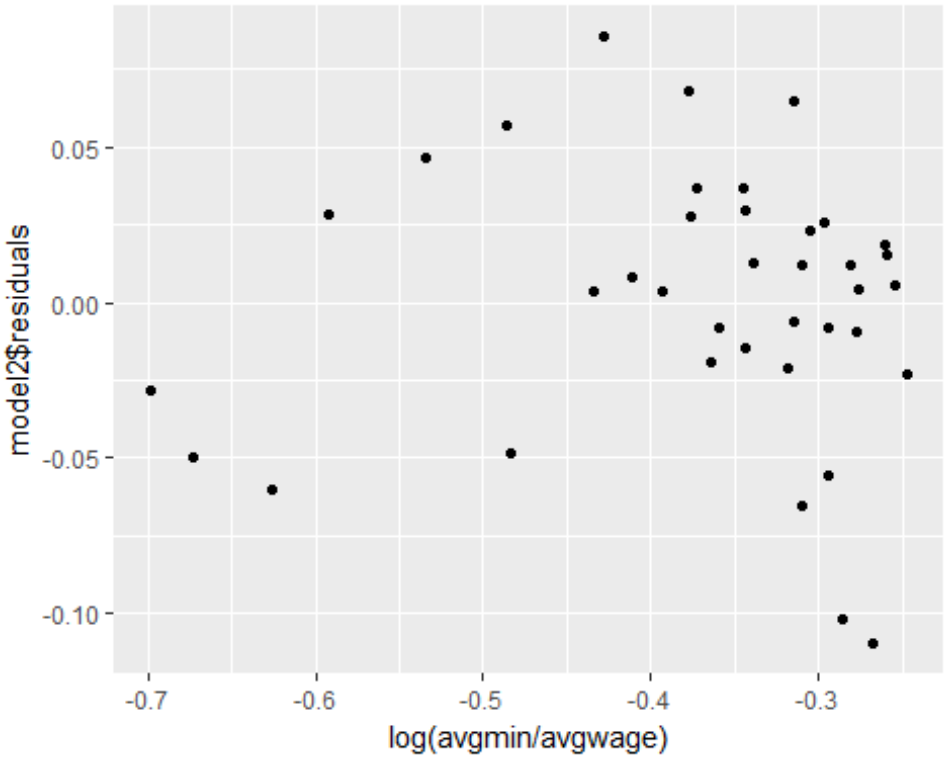
```
## Call:
## lm(formula = log(prepop) ~ log(avgmin/avgwage) + log(usgnp),
##     data = d)
##
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -4.072146   1.598657  -2.547  0.01555 *
## log(avgmin/avgwage) -0.127312   0.074871  -1.700  0.09818 .
## log(usgnp)           0.350477   0.189412   1.850  0.07297 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0296 on 34 degrees of freedom
## Multiple R-squared:  0.1889 ,  Adjusted R-squared:  0.1412
## F-statistic: 4 on 2 and 34 DF,  p-value: < 2.847e-02
##
```

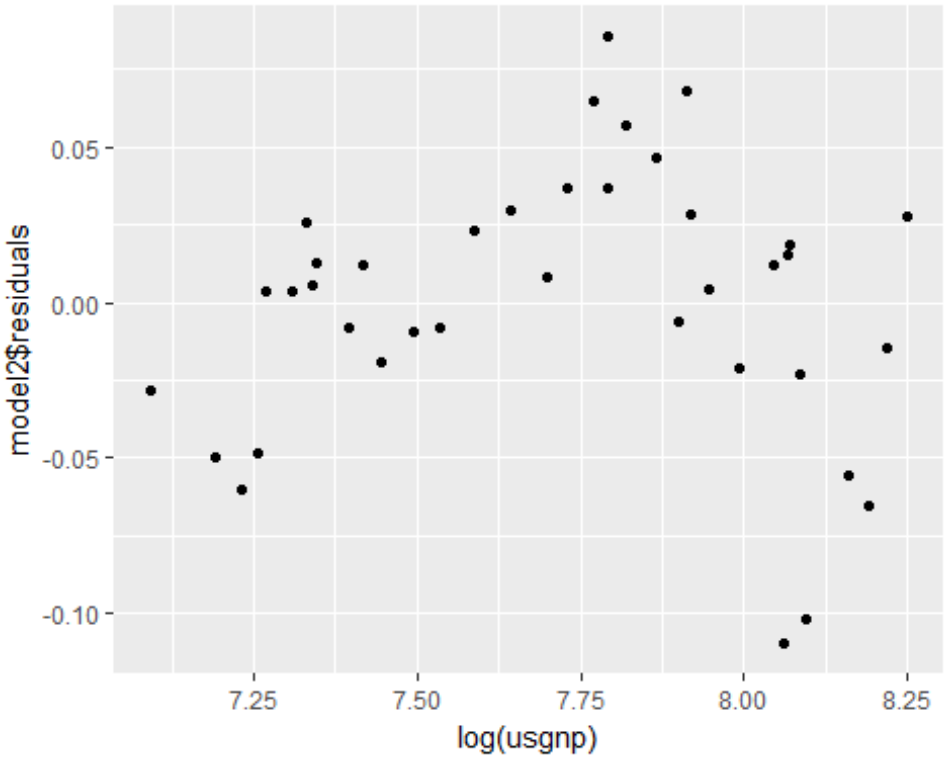
```
## Durbin-Watson statistic
## (original):    0.63704 , p-value: 6.975e-08
## (transformed): 1.73221 , p-value: 1.987e-01
```

可以发现，DW 值从 0.637 变为 1.732，转换后 p 值是 0.1987。

6　异方差检验（图示法， BP 检验， white 检验）

图示法：





BP 检验：

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 3.9374, df = 2, p-value = 0.1396
```

P 值为 0.14，没有强烈异方差。

White 检验：

```
## Call:
## lm(formula = (model2$residuals)^2 ~ log(avgmin/avgwage) + I((log(a
vgmin/avgwage))^2) +
##     log(usgnp) + I((log(usgnp))^2) + I(log(avgmin/avgwage) *
##     log(usgnp)), data = d)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.0026566 -0.0015578 -0.0006391  0.0010189  0.0092123
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>
|t|)
## (Intercept)                         -0.372899   0.409992  -0.910
```

```
0.370
## log(avgmin/avgwage)                -0.187167   0.153433  -1.220
0.231
## I((log(avgmin/avgwage))^2)          -0.020117   0.042435  -0.474
 0.639
## log(usgnp)                          0.085228   0.103005   0.827
0.414
## I((log(usgnp))^2)                   -0.004814   0.006472  -0.744
0.462
## I(log(avgmin/avgwage) * log(usgnp))  0.021831   0.017065   1.279
  0.210
##
## Residual standard error: 0.002785 on 32 degrees of freedom
## Multiple R-squared:  0.1496, Adjusted R-squared:  0.01674
## F-statistic: 1.126 on 5 and 32 DF,  p-value: 0.3666
```

回归的整体显著性：p 值为 0.366。可以认定为没有强烈异方差。

- **综上，在 10%和更小的置信水平下，模型可以接受，作者决定不再加权调整。**

- 最终作者使用 co 迭代法调整的结果作为最终估计并说明其统计意义与预测：

$$\hat{Y}_t = -4.072 - 0.127X_t + 0.350G_t$$

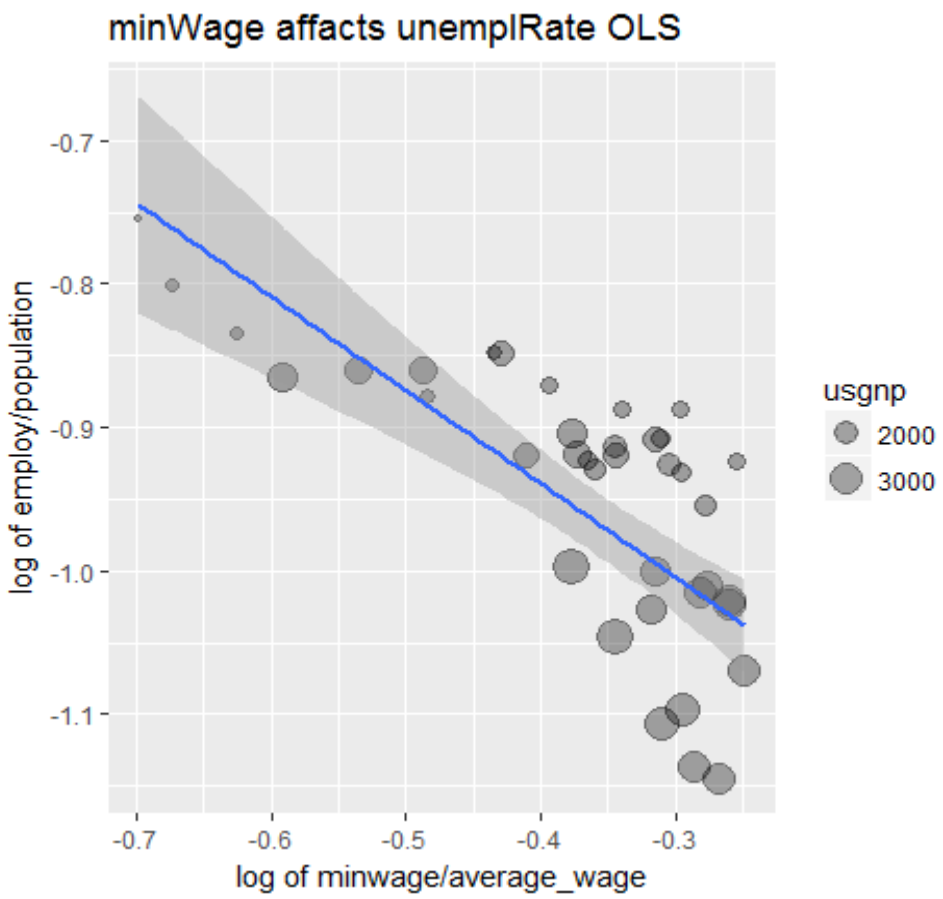
三， 模型统计意义 ----------

- **拟合优度**：由 co 迭代结果得到调整的 R2 为 0.1412，说明上述两个解释变量即 Log（最低工资/制造业平均工资）与 Log（人均 GNP）解释了 14%的被解释变量 Log（就业人口/总人口）的变化。

- **F 值**：（由 co 迭代结果）该模型 F 值为 4 ，p 值为 0.028 ，因此在 5% 的置信水平下该模型总体显著。

- **T 值**：log(avgmin/avgwage)的 t 值为 – 1.7，p 值为 0.098 ，在 10% 的置信水平下该变量显著，log(usgnp) 的 t 值为 1.85 ，p 值为 0.073 ，在 10% 的置信水平下该变量显著。

四， 模型预测 ----------



注：上图是一元回归的结果。

- 预测 1990 年 log（就业人口/总人口），现已知 1990 年每小时最低工资 3.472$，每小时平均工资 5.107$，人均 GNP4126.0$.

```
min90 <- 3.472
wage90 <- 5.107
usgnp90 <- 4126.0
yhat <- -4.072146 + (-0.127312) * log(min90/wage90) + (0.350477) * lo
g(usgnp90)
yhat
```

```
## [1] -1.105275
```

```
sd <- 0.01684
yhat+1.96*sd
```

```
## [1] -1.072269
```

```
yhat-1.96*sd
```

```
## [1] -1.138282
```

```
to <- (coch$residuals)^2
to_hat <- sum(to)/(38-3)
```

```
fuc <- sqrt(sd^2 + to_hat^2)

yhat+1.96*fuc

## [1] -0.7271304

yhat-1.96*fuc

## [1] -1.48342
```

经过计算，1990 年的 log(就业人口/总人口)的预测值为 yhat = - 1.105.即 1990 年预测就业人口占总人口 33.12% 。
在 95%的置信水平下，yhat 均值区间为[-1.138, -1.072] ，
个体值区间为[-1.483, -0.727] 。

❙ 相对误差分析：1990 年实际的就业人口/总人口比值为 34.7%，经计算：

```
delta <- abs(log(0.347)-yhat)
reerror <- delta/log(0.347)
reerror

## [1] -0.04425876
```

相对误差为 4.426%，是较为满意的结果。

五， 模型延伸 ----------

六， 附数据，注释及出处 ----------

	year	avgmin	avgwage	prdef	prepop	usgnp	party2
1	1950	0.198	0.398	0.859	0.470	1203.7	1
2	1951	0.209	0.410	0.881	0.449	1328.2	1
3	1952	0.225	0.421	0.953	0.434	1380.0	1
4	1953	0.311	0.480	0.970	0.428	1435.3	1
5	1954	0.313	0.508	1.000	0.415	1416.2	0
6	1955	0.369	0.547	1.003	0.419	1494.9	0
7	1956	0.447	0.601	1.011	0.412	1525.6	0
8	1957	0.488	0.685	1.035	0.412	1551.1	0
9	1958	0.555	0.716	1.089	0.397	1539.2	0
10	1959	0.588	0.789	1.110	0.394	1629.1	0
11	1960	0.616	0.840	1.138	0.403	1665.3	0

12	1961	0.608	0.875	1.173	0.397	1708.7	0
13	1962	0.707	0.933	1.216	0.385	1799.4	1
14	1963	0.723	1.036	1.247	0.395	1873.3	1
15	1964	0.809	1.097	1.298	0.396	1973.3	1
16	1965	0.834	1.176	1.327	0.401	2087.6	1
17	1966	0.854	1.288	1.358	0.399	2208.3	1
18	1967	0.971	1.371	1.421	0.399	2271.4	1
19	1968	1.104	1.512	1.500	0.403	2365.6	1
20	1969	1.149	1.667	1.552	0.399	2423.3	1
21	1970	1.209	1.856	1.616	0.428	2416.2	0
22	1971	1.224	1.990	1.708	0.423	2484.8	0
23	1972	1.257	2.144	1.780	0.423	2608.5	0
24	1973	1.262	2.281	1.817	0.421	2744.1	0
25	1974	1.681	2.452	1.946	0.405	2729.3	0
26	1975	1.871	2.562	2.082	0.368	2695.0	0
27	1976	2.034	2.681	2.174	0.364	2826.7	0
28	1977	2.198	3.023	2.240	0.358	2958.6	0
29	1978	2.509	3.323	2.340	0.362	3115.2	1
30	1979	2.768	3.589	2.483	0.360	3192.4	1
31	1980	2.997	3.883	2.716	0.359	3187.1	1
32	1981	3.264	4.181	2.954	0.343	3248.8	1
33	1982	3.305	4.318	3.175	0.318	3166.0	0
34	1983	3.350	4.456	3.321	0.321	3279.1	0
35	1984	3.350	4.498	3.461	0.334	3501.4	0
36	1985	3.350	4.565	3.548	0.331	3607.5	0
37	1986	3.350	4.725	3.697	0.351	3713.3	0
38	1987	3.350	4.879	3.787	0.369	3819.6	0

注释：

1 avgmin avgwage 均为每小时工资（$）

2 prdef 为价格平减指数，以 1954 为基数

3 prepop 为工作人口除以总人口

4 usgnp 为美国人均国内生产总值，单位是美元

5 party2 为虚拟变量，0 为民主党，1 为共和党

6 此数据中 usgnp 来自 https://www.bls.gov ，party2 为作者搜集整理，其余数据来自 http://qcpages.qc.cuny.edu/~rvesselinov/statafiles.html 。

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 adjustedR^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 ± 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

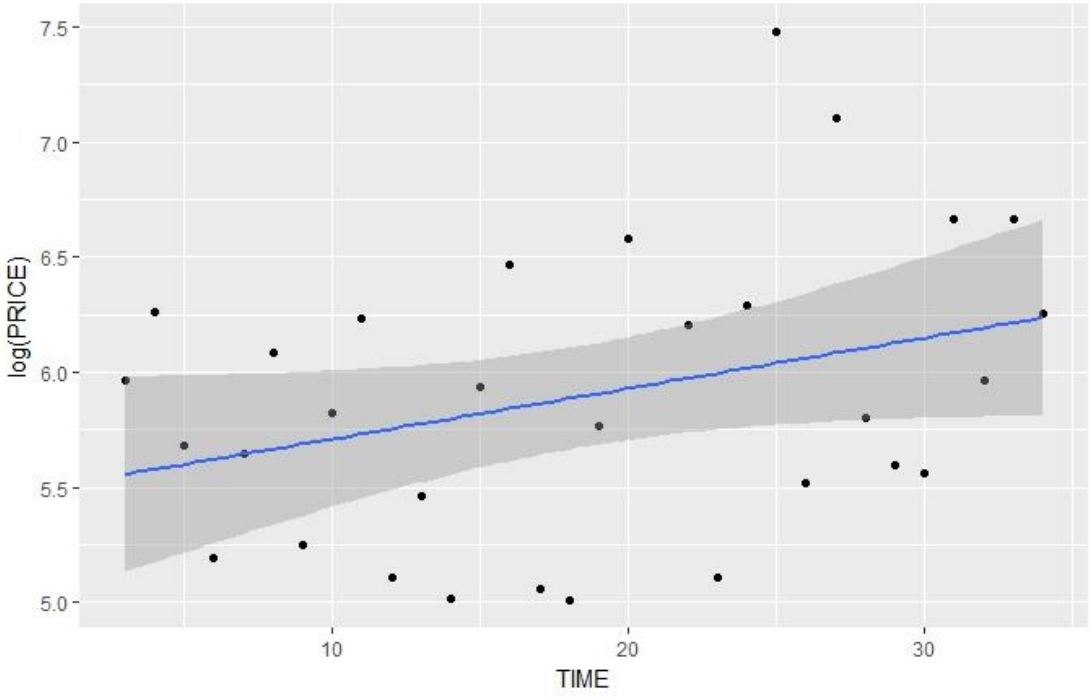


图 1.1 一元回归，时间对价格的影响

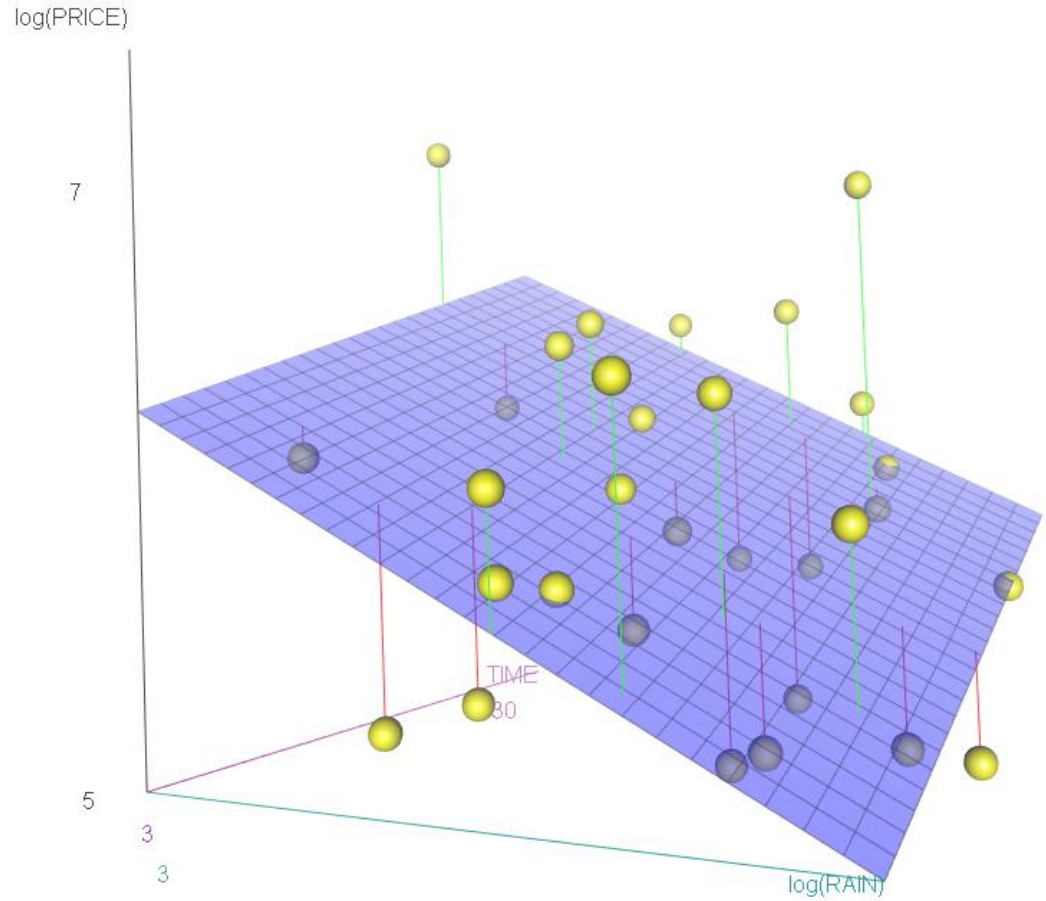


图 1.2 二元回归：时间，降水对价格的影响

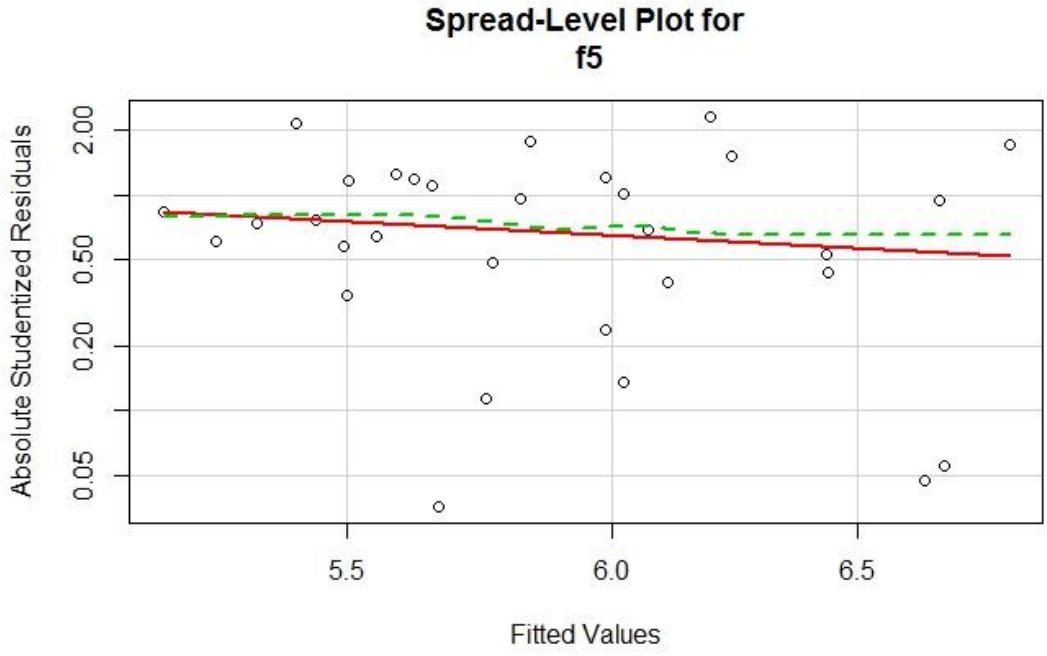


图 1.3 异方差性检查

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