MarketingAnalyticsHomeworks1

Hyeonho Lee

2018년 9월 19일

DemandSimulation.r

"DemandSimulation.r.txt" 파일을 다운로드하여 내용을 확인한후, 관측치를 5,000개로 하는 동일한 시뮬레이션을 수행하되, 단 시뮬레이션 시드를 주어진 코드대로 하지 않고 본인의 학번 뒤의 5자리 숫자를 set.seed() 함수 안의 argument로 하여 시뮬레이션 시드를 조정할 것. (예. 학번이 2014-56789 인 경우 set.seed(56789)로 하여 시드를 조정)

seed 번호(26 고정)

```
set.seed(26)

n = 5000
trueB = c(3,-3, 1.5, 0.7, 3)
err = rnorm(n,sd=3)

u1 = runif(n)
u2 = runif(n)
u3 = runif(n)
u4 = runif(n)

logPr = u1 +  u2
quality = u3 + u2
dummy1 = (u4> 0.7)*1.0
dummy2 = ((u4<0.7)&(u4>0.3))*1.0

logQ= trueB[1] + trueB[2]*logPr + trueB[3]*dummy1 + trueB[4]*dummy2 + trueB[5]*quality + err
```

1. 품질과 가격간의 상관관계가 있는 현재의 모형에서

- 종속변수와 4개의 독립변수 각각과의 scatter plot

- 품질(Quality)이 포함된 모형과 그렇지 않은 모형간의 회귀 결과 비교 (특히 가격 계수를 중심으로)

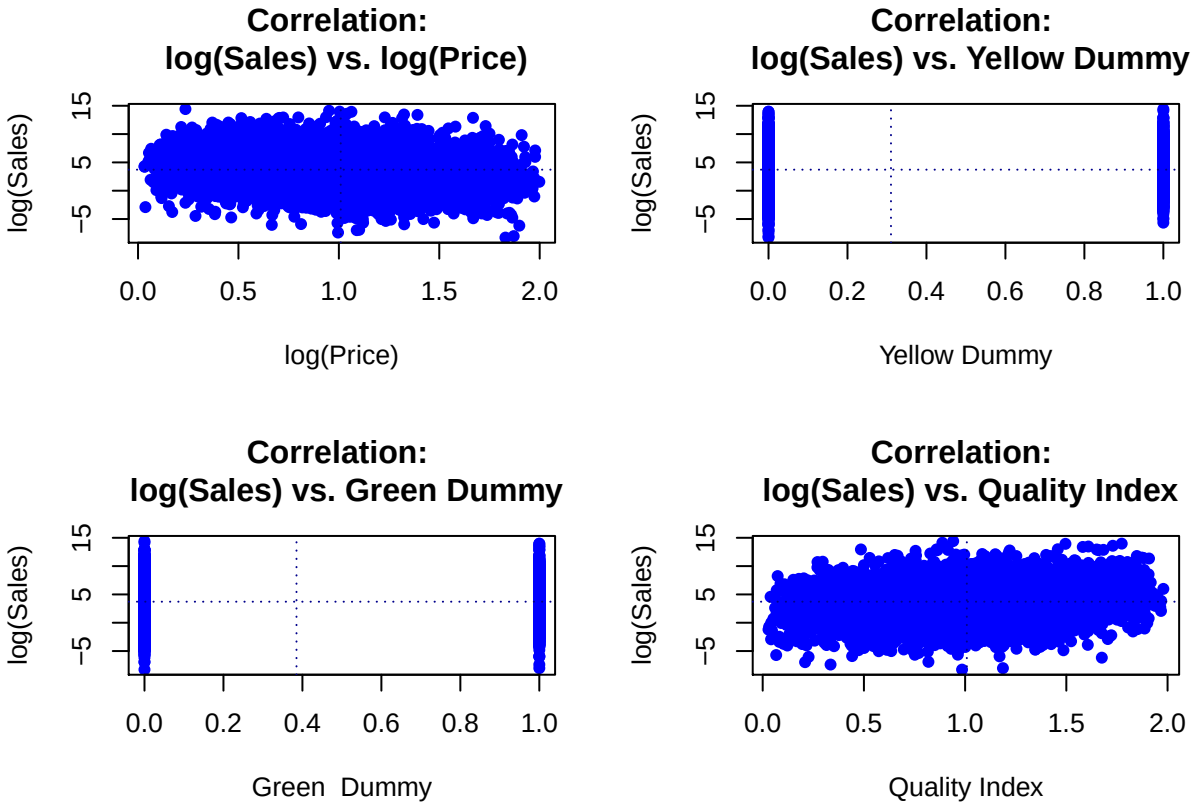
```
par(mfrow=c(2,2))
plot(x=logPr, y=logQ, col="blue", main="Correlation: \n log(Sales) vs. log(Price)",
     xlab="log(Price)", ylab = "log(Sales)", pch=16)
abline(h=mean(logQ),col="dark blue",lty="dotted")
abline(v=mean(logPr),col="dark blue",lty="dotted")

plot(x=dummy1, y=logQ, col="blue", main="Correlation: \n log(Sales) vs. Yellow Dummy",
     xlab="Yellow Dummy", ylab = "log(Sales)", pch=16)
abline(h=mean(logQ),col="dark blue",lty="dotted")
abline(v=mean(dummy1),col="dark blue",lty="dotted")

plot(x=dummy2, y=logQ, col="blue", main="Correlation: \n log(Sales) vs. Green Dummy",
```

```
     xlab="Green  Dummy", ylab = "log(Sales)", pch=16)
abline(h=mean(logQ),col="dark blue",lty="dotted")
abline(v=mean(dummy2),col="dark blue",lty="dotted")

plot(x=quality, y=logQ, col="blue", main="Correlation: \n log(Sales) vs. Quality Index",
     xlab="Quality Index", ylab = "log(Sales)", pch=16)
abline(h=mean(logQ),col="dark blue",lty="dotted")
abline(v=mean(quality),col="dark blue",lty="dotted")
```




```
regout_full = lm(logQ ~ logPr+dummy1+dummy2+quality)
print(summary(regout_full))
```

```
##
## Call:
## lm(formula = logQ ~ logPr + dummy1 + dummy2 + quality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6739 -1.9866  0.0493  1.9780 10.0474
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8900     0.1426  20.273  < 2e-16 ***
## logPr        -2.9612     0.1207 -24.541  < 2e-16 ***
## dummy1        1.4415     0.1078  13.367  < 2e-16 ***
## dummy2        0.6180     0.1025   6.027 1.79e-09 ***
```

```
## quality       3.1054     0.1203  25.815  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.99 on 4995 degrees of freedom
## Multiple R-squared:  0.1691, Adjusted R-squared:  0.1684
## F-statistic: 254.1 on 4 and 4995 DF,  p-value: < 2.2e-16
```

```
regout_short = lm(logQ ~ logPr+dummy1+dummy2)
print(summary(regout_short))
```

```
##
## Call:
## lm(formula = logQ ~ logPr + dummy1 + dummy2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0363  -2.0990   0.0117   2.1398  10.7248
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4447     0.1375  32.315  < 2e-16 ***
## logPr        -1.3775     0.1106 -12.453  < 2e-16 ***
## dummy1        1.4200     0.1148  12.369  < 2e-16 ***
## dummy2        0.5794     0.1091   5.309 1.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.183 on 4996 degrees of freedom
## Multiple R-squared:  0.05819,    Adjusted R-squared:  0.05763
## F-statistic: 102.9 on 3 and 4996 DF,  p-value: < 2.2e-16
```

regout_full모형과 regout_short모형을 비교해 보았을 때, 가장 먼저 F-statistic 을 비교했습니다. F-statistic이 short모델이 감소한 것을 볼 수 있었고, 모델의 유의함이 감소하였습니다. 그러나 p-value는 여전히 상당히 낮으므로 F-statistic아 낮은 것이 문제가 되지 않습니다. 그 다음에 Adjusted R-squared을 확인하였는데, 설명력이 매우 낮아 진 것을 확인 할 수 있습니다. 16.84%에서 5.763%, 약 10%가 감소하였습니다. 그러나 우리가 관심있는 것은 각 변수가 y에 어떤 영향을 미치는 가 이기 때문에 낮아지는 것을 크게 신경 쓰지 않는다. 다음으로 Coefficients의 p-value를 보는데, full, short모형 둘다 모든 변수가 유의하므로 Estimate를 봅니다. 그 중 우리가 관심있는 X는 가격계수인 logPr이므로 logPr을 집중적으로 볼 것이다. quality 를 뺐을 때 logPr이 계수가 약 1.6정도 증가하는 것을 볼 수 있다. 이 말은 즉슨 logPr변수가 Y에 대한 영향력이 감소한다는 이야기 이다. p-value는 두 모델이 모두 유의하다는 결과로 나오기 때문에 두 경우 모두 유의하다고 할 수 있다. Residual standard error은 증가하였다. RMSE가 증가했다고 판단을 내린다.

OVB(Omitted Variable Bias)관점으로 본다면, 상관성을 먼저 볼 필요성이 있다.

```
round(cor(cbind(logQ,logPr, dummy1, dummy2,quality)),2)
```

```
##          logQ logPr dummy1 dummy2 quality
## logQ     1.00 -0.17   0.15  -0.02    0.20
## logPr   -0.17  1.00   0.00   0.01    0.51
## dummy1   0.15  0.00   1.00  -0.53    0.00
## dummy2  -0.02  0.01  -0.53   1.00    0.00
## quality  0.20  0.51   0.00   0.00    1.00
```

logPr과 quality의 상관성은 매우 강하게 나타나고 있다. 또한 위에서 모든 변수를 모델에 적합하였을 때, 모델의 유의함을 F통계량과 p-value로 보여주고 Coefficients, 각 계수들은 참값인 3, -3, 1.5, 0.7, 3과 비슷하게 나타난다. 이는 모델이 상당히 정확하다고 판단하는 근거가 된다. 또한 회귀추정계수가 positive를 나타내고 있다.

그러나 품질인 quality를 제거했을 때 변화한 logPr은 logPr과 quality의 상관성과 모델의 계수추정치가 positive로 인하여 Upward bias의 가능성이 있다고 판단할 수 있다.

$$(X_1'X)^{-1}X_1'X_2 = Corr(X_1, X_2)$$

결과적으로 회귀계수는 과대추정 되었음을 알 수 있고, 가격의 효과에 대해서는 계수가 음수이므로 과소추정이 되었다고 결론을 내릴 수 있다. 이는 가격이 한 단위 증가할 때, 판매량이 적게 감소하는 것을 의미한다. 그러므로 가격 탄력성에 대해서도 과소추정했다는 결론을 내릴 수 있다.

2. 품질과 가격간의 상관관계가 없는 모형

- 주어진 코드에서 다음 내용을 수정하여 품질과 가격간의 상관관계가 "없는" 자료를 생성 (quality 생성할 때 u2를 포함하지 않으면 됨) 즉,

☒ 현재 코드: quality = u3 + u2s

☒ 수정된 코드: quality = u3

- 품질(Quality)이 포함된 모형과 그렇지 않은 모형간의 회귀 결과 비교

```
set.seed(26)

n = 5000
trueB = c(3,-3, 1.5, 0.7, 3)
err = rnorm(n,sd=3)

u1 = runif(n)
u2 = runif(n)
u3 = runif(n)
u4 = runif(n)

logPr = u1 +  u2
quality = u3
dummy1 = (u4> 0.7)*1.0
dummy2 = ((u4<0.7)&(u4>0.3))*1.0

logQ= trueB[1] + trueB[2]*logPr + trueB[3]*dummy1 + trueB[4]*dummy2 + trueB[5]*quality + err
```

```
round(cor(cbind(logQ,logPr, dummy1, dummy2,quality)),2)
```

```
##          logQ logPr dummy1 dummy2 quality
## logQ     1.00 -0.35   0.15  -0.02    0.26
## logPr   -0.35  1.00   0.00   0.01    0.01
## dummy1   0.15  0.00   1.00  -0.53   -0.01
## dummy2  -0.02  0.01  -0.53   1.00   -0.01
## quality  0.26  0.01  -0.01  -0.01    1.00
```

logPr과 quality의 상관관계를 줄이기 위해 quality를 새롭게 정의하였다. 두 변수의 상관관계는 0.01로써 거의 없어졌다고 볼 수 있다. 단 logQ와의 상관관계는 증가한 것을 확인 할 수 있다.

```
regout_full = lm(logQ ~ logPr+dummy1+dummy2+quality)
print(summary(regout_full))
```

```
##
## Call:
## lm(formula = logQ ~ logPr + dummy1 + dummy2 + quality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6627 -1.9967  0.0534  1.9850 10.0897
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8456     0.1487  19.141  < 2e-16 ***
## logPr        -2.9092     0.1039 -27.999  < 2e-16 ***
## dummy1        1.4430     0.1078  13.381  < 2e-16 ***
## dummy2        0.6185     0.1025   6.032 1.73e-09 ***
## quality       3.1944     0.1472  21.708  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.99 on 4995 degrees of freedom
## Multiple R-squared:  0.2199, Adjusted R-squared:  0.2193
## F-statistic:   352 on 4 and 4995 DF,  p-value: < 2.2e-16
```

```
regout_short = lm(logQ ~ logPr+dummy1+dummy2)
print(summary(regout_short))
```

```
##
## Call:
## lm(formula = logQ ~ logPr + dummy1 + dummy2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2920  -2.0719  -0.0026   2.0894  10.4702
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4425     0.1351  32.874  < 2e-16 ***
## logPr        -2.8800     0.1087 -26.501  < 2e-16 ***
## dummy1        1.4063     0.1128  12.468  < 2e-16 ***
## dummy2        0.5891     0.1072   5.493 4.14e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.127 on 4996 degrees of freedom
## Multiple R-squared:  0.1463, Adjusted R-squared:  0.1458
## F-statistic: 285.4 on 3 and 4996 DF,  p-value: < 2.2e-16
```

두 모델 모두 F-statistic이 높고 p-value가 모델의 유의함을 나타내고 있다. 또한 위의 모델은 계수들이 참값들과 매우 근접함을 보이므로 좋은 모델이라고 할 수 있다.

또한 두번째 모델의 logPr을 보면 1번의 예시와는 다르게 회귀계수의 변화가 거의 없음을 알 수 있다. 3개의 변수는 모두 참값과 유사하며 intercept 값이 변했음을 알 수 있다. 이를 통해 X변수의 상관성이 적을 경우, quality와 상관성이 높은 다른 X의 변수가 없을 때, 변수를 제거 하면 다른변수에 큰 영향이 없다고 할 수 있으며(공분산이 낮으므로) 상수항인 intercept 에 풀링되었음을 알 수 있다.