DemandSimulation

Dongkeon Oh

2018년 9월 18일

#### Introduction.

Simulate a data set for regression illustratoin true model logQ is funciton of logP, dummy1, dummy2, a quality

### 1. 품질과 상관관계가 있는 모형.

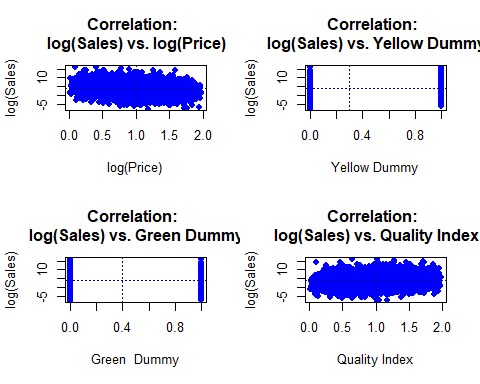
#### data generation

set.seed(20)  
n = 5000 # number of observation  
trueB = c(3,-3, 1.5, 0.7, 3)  
err = rnorm(n,sd=3) # error term simulation  
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  
summary(cbind(logQ,logPr, dummy1, dummy2,quality))

## logQ logPr dummy1 dummy2   
## Min. :-7.859 Min. :0.01632 Min. :0.0000 Min. :0.000   
## 1st Qu.: 1.517 1st Qu.:0.70927 1st Qu.:0.0000 1st Qu.:0.000   
## Median : 3.655 Median :1.00333 Median :0.0000 Median :0.000   
## Mean : 3.694 Mean :1.00466 Mean :0.3008 Mean :0.396   
## 3rd Qu.: 5.939 3rd Qu.:1.30106 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :16.296 Max. :1.97327 Max. :1.0000 Max. :1.000   
## quality   
## Min. :0.01714   
## 1st Qu.:0.71201   
## Median :1.00130   
## Mean :1.00040   
## 3rd Qu.:1.28754   
## Max. :1.97666

#### plotting

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")



산점도를 보면 품질과 가격은 y값(판매량)과 관련이 있어 보인다. 가격은 판매량과 음의 상관관계를 갖고, 품질은 양의 상관관계를 갖는다. 정확한 수치는 상관 행렬을 통해 확인할 수 있다.

#### correlation matrix

library(knitr)  
corrout = cor(cbind(logQ,logPr, dummy1, dummy2,quality))  
print(kable(round(corrout,digits=4)))

##   
##   
## logQ logPr dummy1 dummy2 quality  
## -------- -------- -------- -------- -------- --------  
## logQ 1.0000 -0.2035 0.1704 -0.0138 0.1521  
## logPr -0.2035 1.0000 -0.0012 0.0049 0.5028  
## dummy1 0.1704 -0.0012 1.0000 -0.5311 0.0232  
## dummy2 -0.0138 0.0049 -0.5311 1.0000 -0.0107  
## quality 0.1521 0.5028 0.0232 -0.0107 1.0000

상관행렬에서 dummy1는 판매량과 상관계수는 양수이고, dummy2는 음수이다.  
또한 가격과 품질 사이의 상관계수가 높은 것을 통해, 두 설명 변수 사이의 공선성을 의심할 수 있다.

#### regression output

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   
## -10.3106 -2.0575 -0.0191 2.0413 10.7273   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.2113 0.1418 22.640 < 2e-16 \*\*\*  
## logPr -2.9625 0.1192 -24.863 < 2e-16 \*\*\*  
## dummy1 1.5646 0.1088 14.375 < 2e-16 \*\*\*  
## dummy2 0.7230 0.1020 7.086 1.58e-12 \*\*\*  
## quality 2.7008 0.1206 22.388 < 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: 0.1628, Adjusted R-squared: 0.1621   
## F-statistic: 242.8 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   
## -11.1777 -2.1175 -0.0182 2.1248 11.6133   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.5493 0.1349 33.72 < 2e-16 \*\*\*  
## logPr -1.6209 0.1080 -15.01 < 2e-16 \*\*\*  
## dummy1 1.6204 0.1141 14.20 < 2e-16 \*\*\*  
## dummy2 0.7213 0.1070 6.74 1.77e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.136 on 4996 degrees of freedom  
## Multiple R-squared: 0.07875, Adjusted R-squared: 0.07819   
## F-statistic: 142.3 on 3 and 4996 DF, p-value: < 2.2e-16

회귀 결과를 비교하기 전에 완전 모델을 기준으로 회귀계수를 해석해보자. 로그 가격(이하 가격)의 회귀계수는 -2.96이고 이는 가격탄력성과 비슷하게 해석된다. 즉 다른 변수들이 고정되어 있을 때, 가격이 1% 증가하면 판매량은 평균적으로 2.96% 감소한다. dummy1은 노란색 상품일 때를 의미하고 회귀계수는 1.56이다. 따라서 기준이 되는 색일 때에 비해, 판매량이 평균적으로 376% 증가한다. dummy2은 초록색 상품일 때를 의미하고 회귀계수는 0.72이다. 즉 초록색 상품일 때는 평균적으로 판매량이 105% 증가한다. 앞의 두 경우에도 다른 조건은 고정이다. 끝으로 품질의 회귀계수는 2.7이기 때문에 품질이 1단위 증가하면 판매량은 평균적으로 270% 증가한다.

exp(1.56)-1

## [1] 3.758821

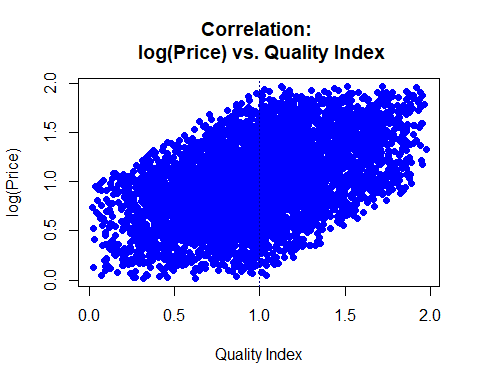
exp(0.72)-1

## [1] 1.054433

두 회귀 결과를 비교해보자. 둘 다 품질이 포함된 모델의 가격의 회귀계수는 -2.96인 반면, 품질이 제거된 모델의 가격의 회귀계수는 -1.62이다. 가격의 회귀계수는 가격탄력성과 같은 의미인데 가격탄력성의 크기는 절대값으로 비교하기 때문에, 제거된 모델에서 가격탄력성이 과소평가 되었다고 할 수 있다. 반면 회귀계수의 측면에서 보면 품질이 제거된 모델은 모회귀계수(-3)를 상대적으로 과대추정 하는데, 이는 제거된 변수로 인해 회귀계수 추정에 편향이 발생했기 때문이다. 품질을 제외해서 편향이 생긴 이유는 품질과 가격이 공선성을 가지고 있기 때문인데, 품질과 가격의 상관계수가 0이 아니면 두 모델의 가격의 회귀계수가 달라진다. 제거된 모델의 가격의 회귀계수는 아래 식으로 구할 수 있다.

수식에 따르면, X1와 X2의 공분산에 따라 제거된 모델의 회귀계수가 영향을 받는다. 두 변수 사이의 상관계수가 커지면 편향이 더욱 커지는 경향이 있으며, beta2의 부호와 상관계수의 부호에 따라 모회귀계수를 과대추정할 수 도 과소추정할 수도 있다. 문제에서 가격과의 회귀계수가 양수이고 가격과 품질의 상관계수가 양수이기 때문에 회귀계수를 과대추정하는 것이다. 아래 그래프를 통해, 품질과 가격의 양의 상관관계를 확인할 수 있다.

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



cor(logPr,quality)

## [1] 0.5027613

#### 질문1 vif 값은 낮게 나오는 이유는?

library(car)  
vif(regout\_full)

## logPr dummy1 dummy2 quality   
## 1.338598 1.393599 1.392902 1.339285

#### 질문2 완전한 모델에서 공선성이 있는 변수들을 포함해서 돌렸는데도, 회귀계수 값이 잘 추정된 이유는? n이 커서? u1,u2,u3가 같은 분포에서 나와서?

#### 질문3 X1과 X2에 상관관계가 있을 때, X2를 제거하고 회귀 모델을 만들면 오차항에 제거된 변수가 포함돼서, X1과 오차항이 상관성을 갖게돼서 OLS 추정량이 편향되는지? 데이터를 보면 아닌 것 같은데, 왜 아닌지를 설명하지 못하겠다.. Some Math 가 나온 맥락?

cor(logPr,regout\_short$residuals)

## [1] 1.221369e-16

cor(logPr,regout\_full$residuals)

## [1] 9.557367e-17

#### 질문4 더미변수의 log-level 해석.

### 2. 품질과 상관관계가 있는 모형.

#### data generation

set.seed(20)  
n = 5000 # number of observation  
trueB = c(3,-3, 1.5, 0.7, 3)  
err = rnorm(n,sd=3) # error term simulation  
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  
summary(cbind(logQ,logPr, dummy1, dummy2,quality))

## logQ logPr dummy1 dummy2   
## Min. :-10.80081 Min. :0.01632 Min. :0.0000 Min. :0.000   
## 1st Qu.: -0.08362 1st Qu.:0.70927 1st Qu.:0.0000 1st Qu.:0.000   
## Median : 2.16629 Median :1.00333 Median :0.0000 Median :0.000   
## Mean : 2.18826 Mean :1.00466 Mean :0.3008 Mean :0.396   
## 3rd Qu.: 4.50929 3rd Qu.:1.30106 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. : 16.12355 Max. :1.97327 Max. :1.0000 Max. :1.000   
## quality   
## Min. :0.0001372   
## 1st Qu.:0.2507805   
## Median :0.4947010   
## Mean :0.4985408   
## 3rd Qu.:0.7473739   
## Max. :0.9996378

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   
## -10.4596 -2.0685 -0.0182 2.0429 10.7117   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.2485 0.1484 21.891 < 2e-16 \*\*\*  
## logPr -3.1135 0.1030 -30.231 < 2e-16 \*\*\*  
## dummy1 1.5634 0.1088 14.365 < 2e-16 \*\*\*  
## dummy2 0.7226 0.1020 7.081 1.63e-12 \*\*\*  
## quality 2.6305 0.1475 17.839 < 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: 0.2259, Adjusted R-squared: 0.2253   
## F-statistic: 364.4 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   
## -11.6583 -2.1395 0.0011 2.1008 11.2996   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.5682 0.1327 34.436 < 2e-16 \*\*\*  
## logPr -3.1306 0.1062 -29.477 < 2e-16 \*\*\*  
## dummy1 1.5986 0.1122 14.245 < 2e-16 \*\*\*  
## dummy2 0.7180 0.1052 6.823 9.96e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.083 on 4996 degrees of freedom  
## Multiple R-squared: 0.1766, Adjusted R-squared: 0.1761   
## F-statistic: 357.1 on 3 and 4996 DF, p-value: < 2.2e-16

품질과 가격 간의 상관관계가 없는 모형의 경우, 두 모형 모두 모회귀계수를 잘 추정한다. 앞서 설명한 바와 같이 두 변수 사이의 공선성이 사라졌기 때문이다.