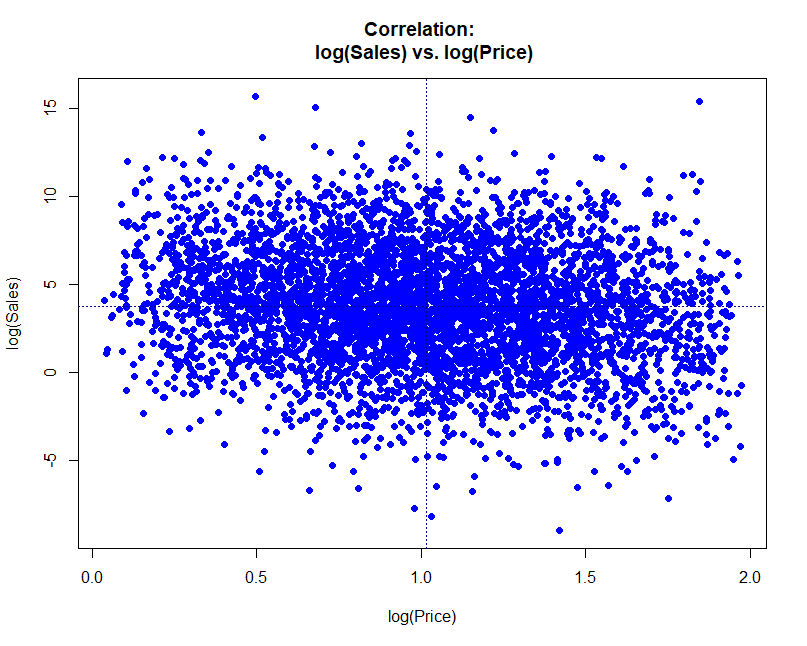
**Marketing Analytics**

**Homeworks**

**Homework #1**

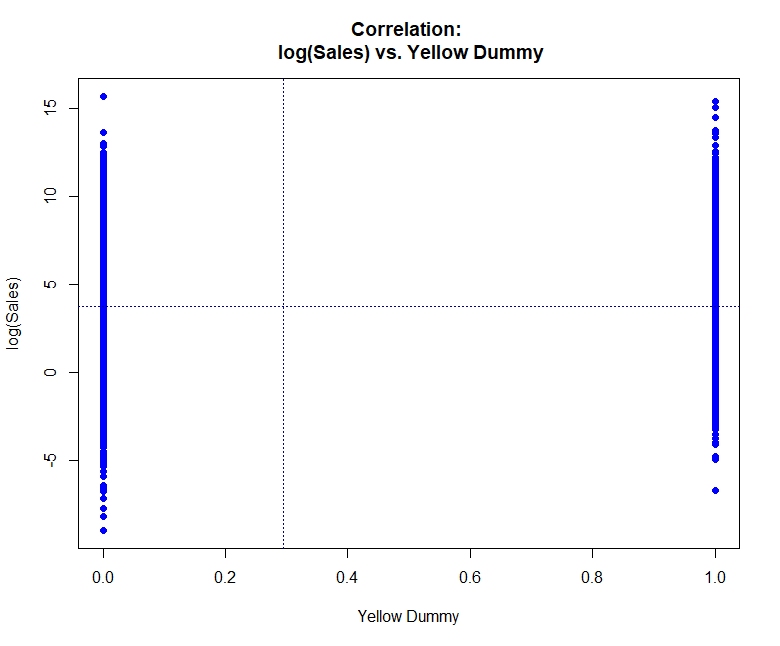
**이름: 허현**

1. 품질과 가격간 상관관계가 있는 모형
   1. 종속변수와 4개의 독립변수의 scatter plot
2. log(가격)과 log(판매)



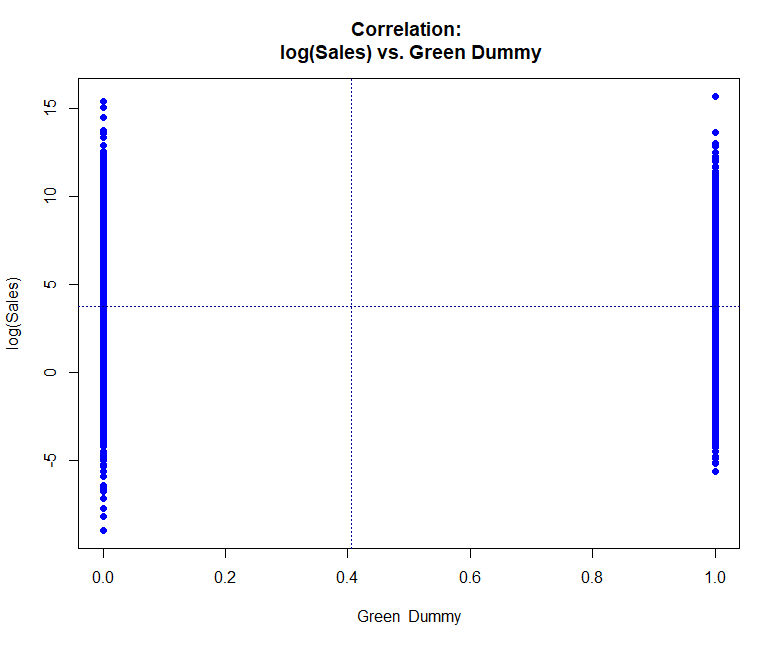
log(판매)가 넓게 퍼져 있지만 log(가격)과 약간의 음의 상관관계가 있음을 알 수 있다.

1. 노란색과 log(판매)



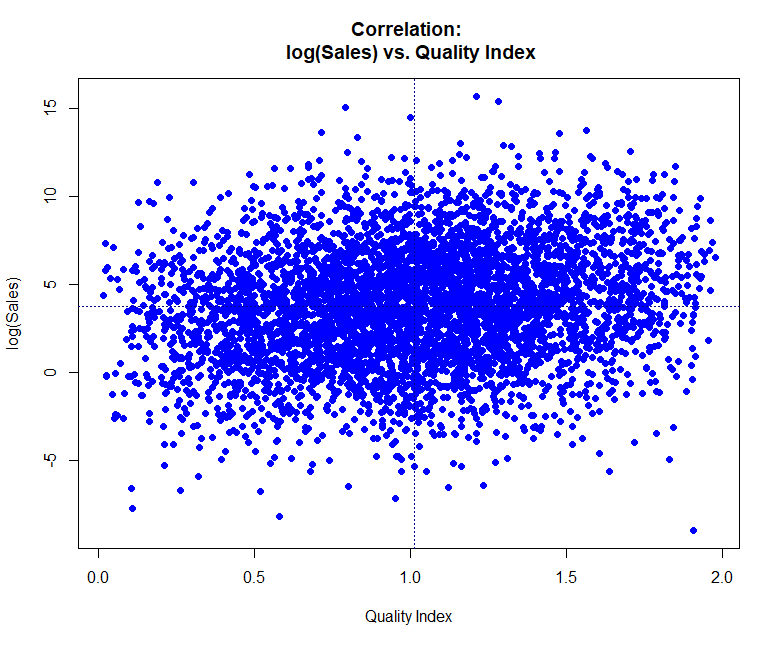
노란색이 맞다/아니다로 구분되는데 이 때, 노란색인 경우에 log(판매)가 전반적으로 높게 나타난다.

1. 초록색과 log(판매)



초록색이 맞다/아니다로 구분되며, 초록색인 경우에 log(판매)가 전반적으로 높아 보인다. 하지만 correlation으로 확인한 결과는 -0.0042로 아주 작게 음의 상관관계를 보인다.

1. 품질과 log(판매)



품질과 log(판매) 사이에 어느정도 양의 상관관계가 있음을 확인할 수 있다.

* 1. 품질(Quality)이 포함된 모형과 그렇지 않은 모형 간의 회귀 결과 비교

품질이 포함된 경우와 아닌 경우 모두 전체 변수가 유의미한 것으로 나타났다. 품질이 포함되지 않은 경우와 아닌 경우 모두 dummy1, dummy2의 추정값[포함: 1.4500, 0.6904 / 포함되지 않음: 1.4770, 0.6530]과 standard error[포함: 0.1108, 0.1029 / 포함되지 않음: 0.1170, 0.1087]가 비슷하게 나타났지만 log(가격)의 값, 특히 추정값이 다르게 나타났다. 따라서 log(가격)에 집중하자면, 품질이 포함된 경우의 log(가격)의 추정값은 -2.6872로 해석하자면 가격이 1% 변할 때 판매가 평균적으로 -2.6872% 변한다는 의미이다. 반면 품질이 포함되지 않은 경우에는 추정값이 -1.3116으로 가격이 1% 변할 때 판매가 평균적으로 -1.3116 변한다는 의미이다.

1. 품질과 가격간 상관관계가 없는 모형

1-2 문제 상황과 달리 품질과 가격간 상관관계가 없어졌기 때문에 품질 변수를 포함하는 경우와 아닌 경우 사이에 추정값과 standard error는 거의 차이나지 않는다.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.8862 0.1524 18.940 < 2e-16 \*\*\*

logPr -2.7498 0.1036 -26.541 < 2e-16 \*\*\*

dummy1 1.4494 0.1108 13.083 < 2e-16 \*\*\*

dummy2 0.6888 0.1029 6.693 2.43e-11 \*\*\*

quality 2.7887 0.1482 18.820 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.02 on 4995 degrees of freedom

Multiple R-squared: 0.1995, Adjusted R-squared: 0.1989

F-statistic: 311.2 on 4 and 4995 DF, p-value: < 2.2e-16

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.3560 0.1354 32.172 < 2e-16 \*\*\*

logPr -2.8159 0.1071 -26.282 < 2e-16 \*\*\*

dummy1 1.4559 0.1146 12.701 < 2e-16 \*\*\*

dummy2 0.6475 0.1065 6.082 1.28e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.125 on 4996 degrees of freedom

Multiple R-squared: 0.1427, Adjusted R-squared: 0.1422

F-statistic: 277.3 on 3 and 4996 DF, p-value: < 2.2e-16

**Code**

**문제1**

set.seed(30)

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

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))

regout\_short = lm(logQ ~ logPr+dummy1+dummy2)

print(summary(regout\_short))

**문제2**

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

regout\_full = lm(logQ ~ logPr+dummy1+dummy2+quality)

print(summary(regout\_full))

regout\_short = lm(logQ ~ logPr+dummy1+dummy2)

print(summary(regout\_short))