growth hackers quest

정유진

2018 년 3 월 8 일

(1) [Coding] 다이아몬드 가격의 평균, 중앙값, 최댓값, 최솟값을 출력해주세요.

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
-
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
diamond <- read.csv("c:/temp/diamonds_data.csv")
summary(diamond$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326     950    2401    3933    5324   18823
```

(2) [Coding] carat 값을 기준으로 다이아몬드의 가격을 재분류해주세요. [Hint: Dictionary]

처음 6 개와 끝 6 개자료

```
diamond %>%  arrange(carat) %>% select(carat,price) %>% head()
```

```
##   carat price
## 1   0.2   345
## 2   0.2   367
## 3   0.2   367
## 4   0.2   367
## 5   0.2   367
## 6   0.2   367
```

```
diamond %>%  arrange(carat) %>% select(carat,price) %>% tail()
```

```
##       carat price
## 53934  4.00 15984
```

```
## 53935  4.01 15223
## 53936  4.01 15223
## 53937  4.13 17329
## 53938  4.50 18531
## 53939  5.01 18018
```

(3) 주어진 데이터를 바탕으로 가격 예측 모델을 결정하고 해당 모델을 선택한 이유를 설명해주세요.

변수 y, z 를 제외하고 carat, x, depth, table 변수가 있는 중회귀모형이 가장 가격을 잘 예측하는 모델이라고 생각했다.

price=20765.594 + 10692.393carat -1226.722x-201.229depth-101.831table

독립변수에 price 를 제외한 모든 변수를 넣은 모형에서 변수의 P-값을 확인하여 보았을 때 변수 z 가 0.05 를 넘는 다는 것을 알 수 있다. 이는 추가설명력이 없다고 판단해 변수 z 를 적합모형에서 제외시켰다. 또한 변수 y 를 제외할지 말지에 대해 알아보기 위해 y 를 제외 후와 제외하지 않은 후 adjusted R-square 을 살펴보았다. 둘은 같은 0.8592 였지만 변수 y 를 포함하지 않은 모형의 F 통계량 값이 더 커서 변수 y 또한 제외시켰다.

(4) [Coding] 문제(3)에서 제시한 모델을 실행하는 코드를 작성해주세요.

```
model2 <-lm(price ~ carat + x + depth + table, data=diamond)
summary (model2)
```

```
##
## Call:
## lm(formula = price ~ carat + x + depth + table, data = diamond)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23894.0   -615.0    -50.6    346.9  12760.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20765.594    418.987   49.56   <2e-16 ***
## carat       10692.393     63.169  169.27   <2e-16 ***
## x           -1226.722     26.678  -45.98   <2e-16 ***
## depth         -201.229      4.852  -41.48   <2e-16 ***
## table         -102.831      3.082  -33.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1497 on 53934 degrees of freedom
```

```
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.8592
## F-statistic: 8.228e+04 on 4 and 53934 DF,  p-value: < 2.2e-16
```

(5) 문제(4)의 결과로 얻은 통계적 관측치들을 제시하고 이를 구체적으로 설명해주세요.

잔차통계량을 residual 을 이용해서 보았을 때 최대값과 최소값이 너무 크고 너무 작은 것을 확인해 이상치가 있다는 것을 알 수 있다. 계수들의 모든 p-value 가 0.05 보다 낮은 것을 확인할 수 있다.

결정계수보다 adjusted R-squared 를 확인하는 것이 중회귀모형에서 좀 더 도움이 되기 위해 이를 살펴보면 0.8592 로 높은 수치를 기록한다는 것을 알 수 있다. 이는 설명력이 꽤 높다는 것을 말한다.

F 통계량과 p-value 는 높을수록 좋은데 각각 8.228e+04 와 2.2e-16 을 기록했다는 것을 볼 수 있다.

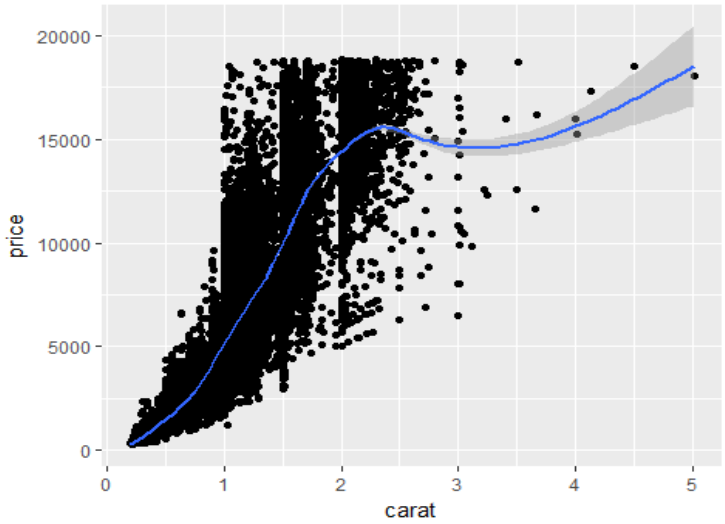
잔차의 표준오차는 작을수록 좋은데 높은 수치를 기록하고 있다는 것을 알 수 있다.

(6) 다이아몬드 가격 예측의 정확도를 높이기 위해서는 무엇을 고려해서 모델을 보완해야 할지 아이디어를 서술해주세요.

carat 과 price 의 그림을 그려보았을 때 밑과 같다.

```
ggplot(data=diamond,aes(carat,price))+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'gam'
```



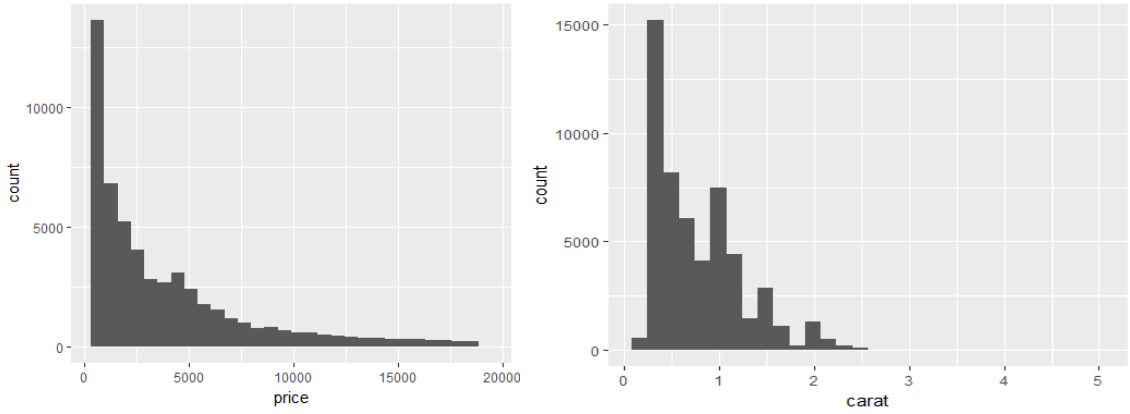
그림을 보았을 때 일정 수준에서 carat 이 몰려있다는 것을 알 수 있다. 이는 아마 carat 의 단위를 측정할 때 일정한 수준이 넘으면 바로 단위가 올라갔다는 것을 예상할 수있어 carat 에 단위를 쪼개어 보거나 좀 더 정확한 carat 이 나올 수 있는 방법이 있는지 찾아보는 작업이 필요할 것 같다.

```
ggplot(diamond,aes(price))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
ggplot(diamond,aes(carat))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



그리고 위의 price 와 carat 의 count 를 살펴보았을 때 상당히 많은 양이 왼쪽에 치우쳐진 결과가 나타나는 것을 확인할 있다. 이는 두 변수 모두 log 를 씌웠을 때 좀 더 보기 나은 그래프가 나온다는 것을 알 수 있어 이러한 것이 나중 가격예측에 더 도움이 되는지도 알아봐야 할 것 같다.

또한 연속형 범주만 주어졌는데 price 를 결정하는 요인에는 범주형 변수 또한 요구될 것으로 보여진다. 색깔이나 다양한 자료들이 가격에 변동을 줄 수 있을 것이라고 예상한다.

마지막으로 outlier 에 대한 제거 작업이 진행되지 않았다. 회귀모형을 설정하기 이전 outlier 에 대한 제거나 처리방법을 취했어야 했는데 거기까지는 진행하지 못하고 모형을 설정하게 되었다.