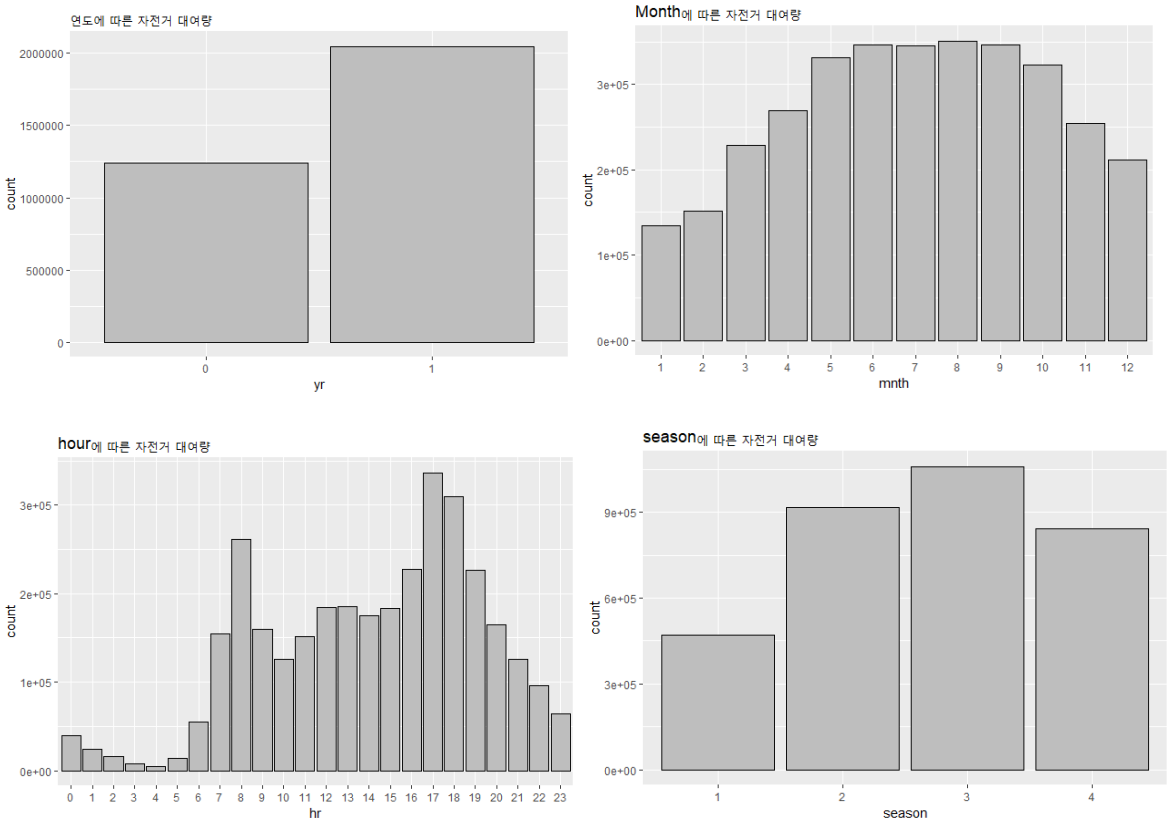
예측 애널리틱스 과제

2014170852 산업경영공학부 조영관

<데이터 파악>

먼저 Y와 몇 개의 독립변수들을 선택해 어떤 상관관계를 이루는지 살펴보았습니다.

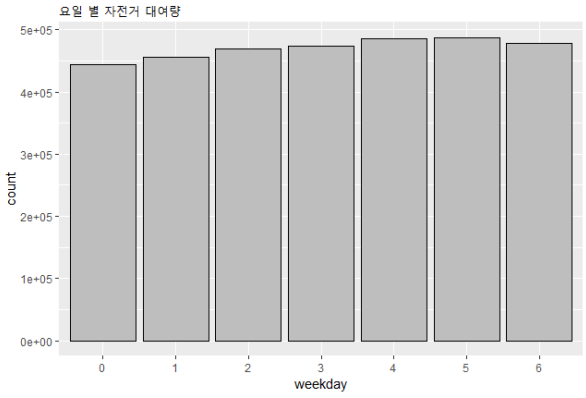


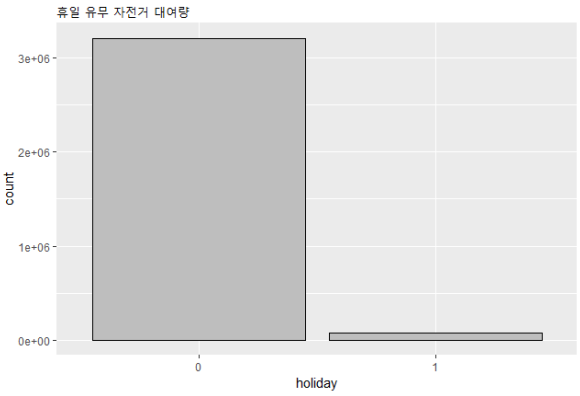
Bike data는 2011년~ 2012년에 걸친 2년간의 데이터입니다.

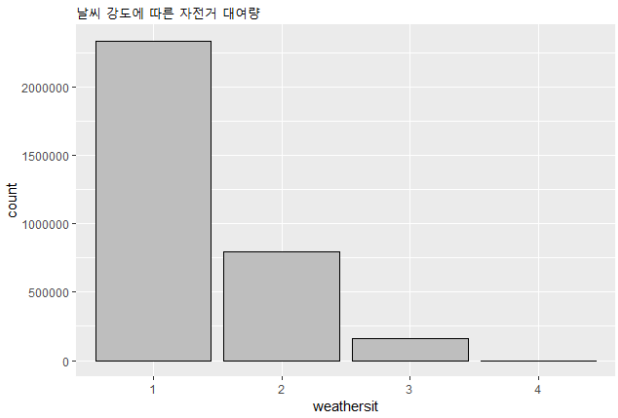
왼쪽 상단 그래프를 보면, 2011년 보다 2012년인 다음 해에 사람들이 bike를 더 자주 이용한 것을 확인할 수 있습니다.

오른쪽 하단 그래프를 보면, 분기(계절) 별 사람들의 bike 총 이용량을 확인할 수 있습니다. 1분기(1월~3월)을 제외하고는 뚜렷한 이용량의 차이가 보이지 않습니다. 좀 더 자세하게 월 별로 들어가 보면 오른쪽 상단 그래프를 통해 확인할 수 있습니다.

다음으로 시간에 따른 bike 이용량을 확인해보았습니다. 왼쪽 하단 그래프를 보면 6시까지는 비교적 bike 이용량이 적은 편입니다. 오전 8시, 그리고 17시, 18시에서 매우 높은 이용량을 보입니다. 추측하자면, 출근 시간대와 퇴근 시간대에 자전거 대여량이 많음을 파악할 수 있습니다.



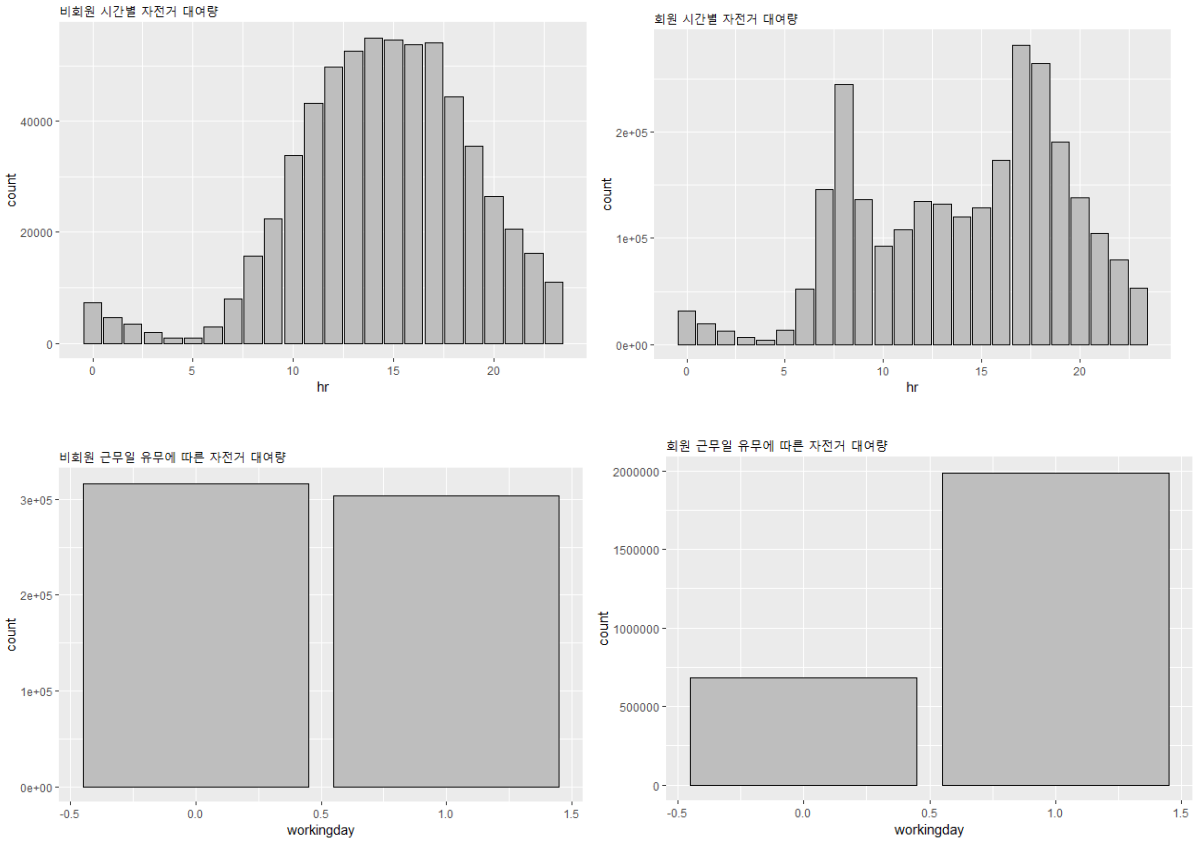




요일 별 자전거 이용량도 확인을 해보았습니다. 왼쪽 상단 그래프를 보시면, 요일에 따라서는 자전거 이용량의 차이가 거의 없는 것을 확인할 수 있습니다.

휴일 유무에 따라서 구분을 해보았습니다. 오른쪽 상단을 보시면 휴일이 아닐 때가 압도적으로 많은 자전거 이용량을 보입니다. 물론 휴일이 21일이고 아닌 날이 710일이라 차이가 많이 나는 것도 있지만, 일 수 비율을 조정하여 계산해보면 여전히 휴일이 아닌 날이 자전거 이용량이 훨씬 많습니다. 즉, 자전거는 휴일보다 휴일이 아닌 날에 이용하는 사람들이 더 많다는 것을 알 수 있습니다. 날씨의 좋은 정도(weathersit)도 마찬가지입니다. 날씨가 좋은 순으로 자전거 이용량이 많고 큰 차이가 보입니다. 비율을 조정해도 1, 2, 3, 4 순으로 자전거 이용량이 많습니다.

Y (자전거 대여량)을 구성하는 두 가지인 casual(비회원)과 registered(회원)의 변수에도 주목해보았습니다.



4개 그래프 중 왼쪽 상단 그래프는 casual(비회원)의 시간 별 자전거 이용량입니다. 그리고 오른쪽 상단 그래프는 registered(회원)의 시간 별 자전거 이용량입니다.

Casual(비회원)의 경우 아침6시부터 점차 증가하여 오후 시간대 (12~6)에 많이 자전거를 대여합니다. 반면에 registered(회원)의 경우 아침 그리고 저녁(8시, 17~18시)에 즉 출퇴근 시간대에 급격한 증가를 보였다가 그 이외의 시간에는 보통의 이용량을 보여줍니다. 이러한 특징을 통해 추측해보면, registered(회원)의 경우 직장이나 학교를 다니는 사람들이 대부분일 것이라 추정됩니다.

반면에 casual(비회원)의 경우 오후 시간대에 자전거가 많이 대여되는 것으로 보아, 정해진 근무시간이 없는 사람이거나, 근무하지 않는 날에 주로 자전거를 대여하는 사람들로 구성된 것으로 추정됩니다.

4개 그래프 중 왼쪽 하단 그래프는 casual(비회원)의 근무일 여부에 따른 자전거 대여량입니다. 오른쪽 그래프는 registered(회원)의 근무일 여부에 따른 자전거 대여량입니다.

여기서 근무일은 주말, 공휴일을 뺀 날을 말합니다.

Casual(비회원)의 경우 근무일이나 근무일이 아닌 날이나 자전거 대여량에 큰 차이를 보이지 않습니다. 즉, 근무일이든 아니든 자전거를 대여하는 횟수는 비슷합니다.

반면에 registered(회원)의 경우 근무일이 아닌 날(주말, 공휴일)은 근무일보다 훨씬 적은 자전거 대여량을 보여줍니다. 위의 시간대 별 회원/비회원의 자전거 대여량과 같이 고려 해보았을 때 이를 통해 알 수 있는 점은 이렇습니다.

회원의 경우 근무시간이 정해진 직장인 혹은 등/하교하는 학생의 비율이 높을 것이며 주로 출퇴근 시에 자전거를 이용한다.

비회원의 경우 근무시간이 정해져 있지 않은 사람의 비율이 높을 것이며, 주로 오후 시간대(12~6)에 자전거를 이용한다.

<다중회귀모델 구축 및 해석>

Season, yr, mnth, hr, holiday, weekday, weathersit은 범주형 데이터이므로 factor형 변수로 설정하여 회귀분석을 진행하였습니다. 그리고 workingday는 holiday와 weekday 두 변수로 설명이 가능하여 제외하였습니다. 또한 instant, dteday도 회귀분석에 불필요한 변수이므로 제거하였습니다.

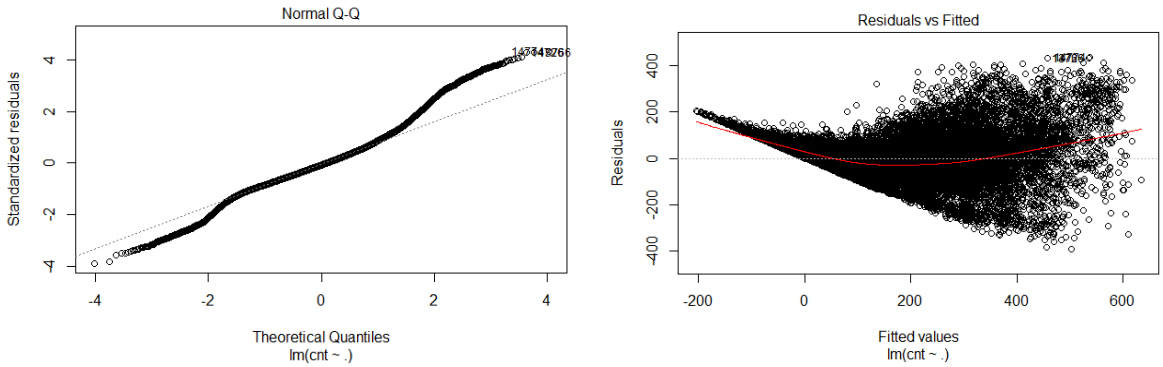
그 결과, 다음과 같이 나왔습니다.

R2 값은 0.6864로, 약 68.64%의 설명력을 가지는 모델임을 알 수 있습니다.

```
Residual standard error: 101.7 on 17326 degrees of freedom
Multiple R-squared:  0.6864,    Adjusted R-squared:  0.6854
F-statistic: 729.1 on 52 and 17326 DF,  p-value: < 2.2e-16
```

문제는, 이 모델이 3가지 가정을 충족하는지의 여부를 파악해야 합니다.

정규성, 등분산성, 독립성이 그 3가지입니다.



위 그림을 보면, 왼쪽의 Normal Q-Q Plot을 통해 정규성을 띄는 모습을 보일 수 있습니다.

하지만 오른쪽 residual plot을 보면 점의 분포가 패턴을 보이는 것을 확인할 수 있습니다.

점이 왼쪽 위 한군데에 집중해 몰려 있고 점점 오른쪽으로 갈수록 퍼지는 모습을 확인할 수 있습니다. 즉 등분산성을 만족하지 않습니다. 따라서 이 모델은 가정을 만족하지 않습니다.

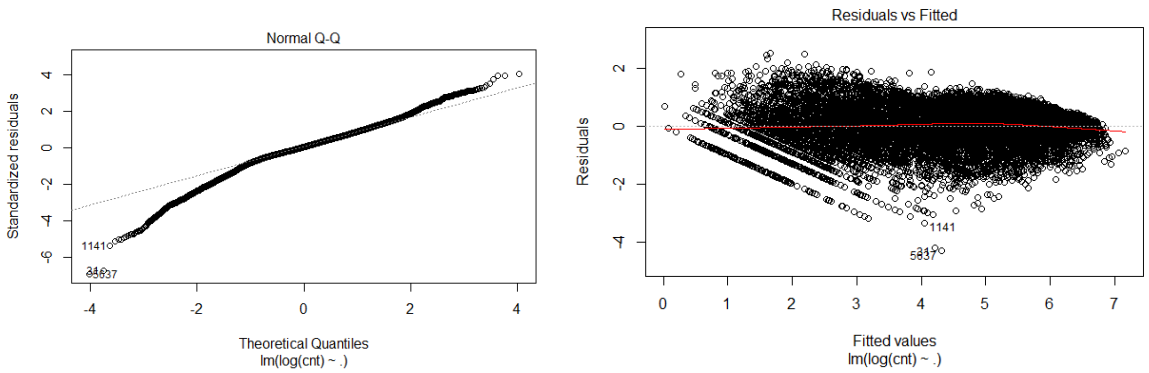
<다중회귀모델 변환 및 가정 충족 여부 확인>

그래서 Y값에 log를 씌워 다시 모델을 구축해보았습니다.

그 결과 다음과 같습니다.

```
Call:
lm(formula = log(cnt) ~ ., data = bikedata)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3185 -0.3002  0.0314  0.3771  2.5365

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.616548   0.040706  64.279  < 2e-16 ***
season2      0.315764   0.029800  10.596  < 2e-16 ***
season3      0.375101   0.035282  10.632  < 2e-16 ***
season4      0.611160   0.029959  20.400  < 2e-16 ***
yr1          0.473403   0.009592  49.355  < 2e-16 ***
mnth2        0.109315   0.024059   4.544 5.56e-06 ***
mnth3        0.127843   0.027045   4.727 2.30e-06 ***
mnth4        0.084463   0.040185   2.102 0.035579 *
mnth5        0.224996   0.042998   5.233 1.69e-07 ***
mnth6        0.121234   0.044213   2.742 0.006112 **
mnth7       -0.028485   0.049596  -0.574 0.565742
mnth8        0.046301   0.048351   0.958 0.338279
mnth9        0.129880   0.042964   3.023 0.002506 **
mnth10       0.042157   0.039787   1.060 0.289353
mnth11      -0.048810   0.038282  -1.275 0.202329
mnth12      -0.040396   0.030402  -1.329 0.183957
hr1         -0.626723   0.032801 -19.107  < 2e-16 ***
hr2         -1.165061   0.032916 -35.395  < 2e-16 ***
hr3         -1.749537   0.033154 -52.770  < 2e-16 ***
hr4         -2.044652   0.033187 -61.610  < 2e-16 ***
hr5         -0.953466   0.032972 -28.917  < 2e-16 ***
hr6          0.264713   0.032885   8.050 8.84e-16 ***
hr7          1.248343   0.032820  38.036  < 2e-16 ***
hr8          1.882225   0.032781  57.418  < 2e-16 ***
hr9          1.565427   0.032815  47.704  < 2e-16 ***
hr10         1.239072   0.032951  37.603  < 2e-16 ***
hr11         1.353384   0.033196  40.769  < 2e-16 ***
hr12         1.535804   0.033480  45.872  < 2e-16 ***
hr13         1.509262   0.033712  44.769  < 2e-16 ***
hr14         1.430093   0.033903  42.182  < 2e-16 ***
hr15         1.482693   0.033969  43.649  < 2e-16 ***
hr16         1.738501   0.033897  51.287  < 2e-16 ***
hr17         2.133607   0.033699  63.314  < 2e-16 ***
hr18         2.056809   0.033477  61.440  < 2e-16 ***
hr19         1.771293   0.033163  53.412  < 2e-16 ***
hr20         1.482792   0.032982  44.957  < 2e-16 ***
hr21         1.229074   0.032848  37.417  < 2e-16 ***
hr22         0.978822   0.032786  29.855  < 2e-16 ***
hr23         0.584304   0.032759  17.836  < 2e-16 ***
holiday1    -0.137293   0.029951  -4.584 4.60e-06 ***
weekday1    -0.042034   0.018243  -2.304 0.021230 *
weekday2    -0.055943   0.017818  -3.140 0.001694 **
weekday3    -0.040253   0.017796  -2.262 0.023719 *
weekday4     0.008926   0.017803   0.501 0.616096
weekday5     0.113181   0.017747   6.377 1.85e-10 ***
weekday6     0.101132   0.017660   5.727 1.04e-08 ***
weathersit2 -0.043768   0.011783  -3.715 0.000204 ***
weathersit3 -0.587242   0.019858 -29.571  < 2e-16 ***
weathersit4 -0.064253   0.361405  -0.178 0.858893
temp         0.499171   0.181114   2.756 0.005855 **
atemp        1.050997   0.187928   5.593 2.27e-08 ***
hum         -0.273329   0.034082  -8.020 1.13e-15 ***
windspeed   -0.190773   0.043276  -4.408 1.05e-05 ***

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

Residual standard error: 0.6243 on 17326 degrees of freedom
Multiple R-squared:  0.824,     Adjusted R-squared:  0.8235
F-statistic:  1560 on 52 and 17326 DF,  p-value: < 2.2e-16
```

R2 값이 0.824로 향상된 것을 확인할 수 있습니다. 즉 Y에 log를 씌운 모델은 82.4%의 설명력을 가집니다. 가정을 만족하는지 확인해보겠습니다.



왼쪽 Normal Q-Q Plot 그래프를 보면 정규성을 가짐을 확인할 수 있습니다. 그리고 오른쪽

Residual plot을 보면 점들이 고르게 분포되어 있음을 파악할 수 있습니다. 즉 등분산성, 독립성을 만족합니다. 따라서, 3가지 가정을 만족함을 알 수 있습니다.

<각 회귀계수에 대한 95% 신뢰구간 추정>

각 회귀계수에 대한 95%의 신뢰구간 추정은 다음과 같습니다. (양측 추정)

	2.5 %	97.5 %
(Intercept)	2.536760078	2.696336082
season2	0.257353659	0.374174504
season3	0.305945489	0.444257080
season4	0.552437284	0.669882734
yr1	0.454602275	0.492203875
mnth2	0.062158019	0.156472325
mnth3	0.074830875	0.180854209
mnth4	0.005697056	0.163228973
mnth5	0.140714715	0.309276391
mnth6	0.034571888	0.207895982
mnth7	-0.125697656	0.068727572
mnth8	-0.048472204	0.141074325
mnth9	0.045666265	0.214093329
mnth10	-0.035829360	0.120144041
mnth11	-0.123846901	0.026227570
mnth12	-0.099986266	0.019195064
hr1	-0.691016948	-0.562429670
hr2	-1.229579619	-1.100541924
hr3	-1.814521882	-1.684552398
hr4	-2.109702316	-1.979601788
hr5	-1.018094920	-0.888838032
hr6	0.200254689	0.329171699
hr7	1.184012110	1.312674381
hr8	1.817971357	1.946479541
hr9	1.501105617	1.629747687
hr10	1.174484407	1.303659900
hr11	1.288315952	1.418451169
hr12	1.470179042	1.601428912
hr13	1.443182378	1.575341622
hr14	1.363640474	1.496546025
hr15	1.416110169	1.549274900
hr16	1.672058511	1.804943506
hr17	2.067553502	2.199660042
hr18	1.991191395	2.122426911
hr19	1.706290715	1.836296186
hr20	1.418144198	1.547440727
hr21	1.164688454	1.293458674
hr22	0.914557679	1.043086202
hr23	0.520092425	0.648515711
holiday1	-0.196000605	-0.078584904
weekday1	-0.077791707	-0.006275338
weekday2	-0.090869055	-0.021017929
weekday3	-0.075135566	-0.005369999
weekday4	-0.025968983	0.043821657
weekday5	0.078394308	0.147967316
weekday6	0.066516567	0.135747250
weathersit2	-0.066863468	-0.020672462
weathersit3	-0.626166663	-0.548317769
weathersit4	-0.772643453	0.644137515
temp	0.144169814	0.854172075
atemp	0.682638539	1.419355997
hum	-0.340132301	-0.206525081
windspeed	-0.275597826	-0.105948650

<각 회귀계수에 대한 기울기=0 여부 검정>

기울기=0인지의 여부에 대한 검정은 위의 다중회귀분석 모델 결과에서 확인할 수 있습니다.

p-value <0.05인 회귀계수는 기울기=0에 대한 귀무 가설을 기각합니다. 즉, 기울기가 0이 아닙니다. Factor로 범주화 한 특정 몇 개를 제외한 대부분의 회귀계수는 기울기가 0이 아닙니다.

<training set과 test set으로 나누고 test 데이터에 대한 예측 정확도 계산하기>

Training set, test set을 8대 2로 나누었습니다.

그리고 training set을 이용해 model을 구축하였습니다.

그 후 test set을 구축한 model에 넣어 predict를 하였습니다.

```
#train, test set 쪼개기
sample_num <- sample(1:nrow(bikedata), size=round(0.2*nrow(bikedata)))
test_bikedata <- bikedata[sample_num,]
train_bikedata <- bikedata[-sample_num,]
```

```
#train set의 다중회귀분석모델 그리고 test set을 통한 검증
train_logmodel <- lm(log(cnt)~. , data=train_bikedata)
pred <- predict(train_logmodel, test_bikedata)
cor(pred, test_bikedata$cnt)
```

```
> cor(pred, test_bikedata$cnt)
[1] 0.7505517
```

그 결과 0.7506, 다시 말해 75%의 예측 정확도를 가지는 모델임을 결론 내릴 수 있습니다.

<ANOVA 검정 실시 후 결과>

구축한 다중회귀분석 모델에 대해 ANOVA 검정을 실시하였습니다.

```
Analysis of Variance Table

Response: log(cnt)
              Df  Sum Sq Mean Sq  F value    Pr(>F)
season         3  2043.9  681.31 1747.944 < 2.2e-16 ***
yr             1  1084.4 1084.40 2782.103 < 2.2e-16 ***
mnth          11   333.3   30.30   77.731 < 2.2e-16 ***
hr            23 27199.0 1182.56 3033.939 < 2.2e-16 ***
holiday        1     6.1    6.14   15.759 7.222e-05 ***
weekday        6    73.2   12.19   31.280 < 2.2e-16 ***
weathersit     3   601.2  200.41  514.161 < 2.2e-16 ***
temp           1   244.2  244.22  626.550 < 2.2e-16 ***
atemp          1    14.3   14.27   36.608 1.474e-09 ***
hum            1    20.7   20.68   53.059 3.377e-13 ***
windspeed      1     7.6    7.57   19.433 1.048e-05 ***
Residuals  17326  6753.3    0.39
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

결과를 보면, 모든 독립변수들의 P-VALUE 값이 0.05보다 작습니다.

모든 독립변수들이 Y값(자전거 대여량)에 유의미한 영향을 미침을 확인할 수 있습니다.

<분석 시 어려웠던 점>

#처음에 데이터를 파악하는 작업이 어려웠습니다. 다양한 변수들을 활용해 상관관계를 살펴보거나, 유의미한 정보를 도출해내는 것이 많은 생각과 시간을 요구하는 작업이었습니다.

#앞으로 새로운 데이터를 접할 때 변수가 점점 더 늘어날수록 고려해야 할 제약조건이 늘어나고 분석이 복잡해질 것이라는 생각이 들었습니다.

Train set과 test set을 나누고 r을 이용해 모델의 예측 정확도를 검정하는 방법을 찾는 과정이 오래 걸렸습니다. 예측 값들과 test set의 값들을 비교해보며 비슷한 분포를 띄는 것을 파악할 수는 있었지만, 어느 정도의 높은 확률로 예측할 수 있는지 계산하는 것이 쉽지 않았습니다.

<수업 시간에 배웠던 내용 외에 필요했던 지식>

#R 명령어를 기본적으로 다룰 줄 알아야 합니다.

데이터를 정제, 변형, 가공하기 위해서는 R 명령어를 통해 자유자재로 다루는 것이 필수적이기 때문입니다.

#구축한 모델이 잘 예측할 수 있는 모델인지를 평가하기 위해서는 교차 검증, 혹은 7 대 3 train, test set 나누기를 통해 평가해야 함을 공부하였습니다.