Day43 딥러닝을 위한 빅데이터 기초-R(12)

☒ 통계학? 모집단을 알고 싶으나 전수조사가 불가능하므로 표본을 통하여 모집단을 추정하는 학문

자료유형

☒ 수치형 : boxplot, histogram

- 이산형 : 분리되는 숫자 자료
- 연속형 : 연속적인 값

☒ 범주형 : 도수분포표

- 명목형 : 순서가 없음
- 순위형 : 순서가 있음

☒ 입력변수(설명, 독립) → 반응변수(출력, 종속)

☒ 연속형

- → 범주형(로지스틱 회귀분석)
- → 연속형(회귀 분석)

☒ 범주형

- → 범주형(범주형 자료분석)
- → 연속형(분산분석)

자료요약

☒ 모집단 개체 수 : N

☒ 평균 $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$, 중앙값, 최빈값(mode, 가장 자주 등장한 값)

☒ 시각화 → 히스토그램

☒ 산포도(퍼진 정도)

- 분산(Var) : 중심으로 부터 멀리 떨어져 있는가
  - $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$
  - 특이값 영향 ⇒ 사분위수 범위
  - 왜도 : 비대칭 정도
  - 첨도 : 뾰족한 정도

☒ 모수 : μ, σ^2

☒ 통계량 : 표본의 관측값에 의해 계산된 값

- 표본 평균 : $\bar{x} = \frac{\sum x_i}{n} \Rightarrow$ 평균 모수 추정
- 표본 분산 : $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \Rightarrow$ 분산 모수 추정
- 추정량 : 모수를 추정하기 위한 용도의 통계량

확률

☒ 확률 실험 특징

- 실험 결과를 모음
- 실험 결과를 이미 인지(ex. 동전 앞/뒤)
- 실험 반복 가능(동전 여러개 던짐)
- 표본 공간(Ω) = 근원사건집합, 결과들의 모임(ex. 주사위 $s = 1, 2, 3, 4, 5, 6$)
- 근원사건 : 표본 공간 집합의 원소$(1, 2, 3, 4, 5, 6)$
- 사건 : 표본공간의 부분집합(홀수사건 = $1, 3, 5$)

☒ 확률? 사건이 일어날 가능성

- $P(A) = P(2, 4, 6)$
- $0 \leq P(A) \leq 1$
- $A_i, i = 1 \sim n$: 사건들이 서로 배반사건일 경우
  - 합사건의 확률은 각 사건이 발생할 확률의 합과 같다.

☒ 조건부 확률

- B가 주어졌을 때 사건 A가 일어날 확률
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- A와 B가 독립일 경우
  - $P(A|B) = P(A)$
  - $P(B|A) = P(B)$

☒ 확률변수? 근원사건들의 실수 값을 대응시킨 함수

- ex) A: 동전 앞면의 개수
  - $A(H, H) = 2$
  - $A(T, T) = 0$

☒ 확률분포? $f(x)$, 확률변수 -> 확률값

- ex. x : 동전 앞면 개수
  - $f(0)$: 동전 앞면개수가 0개일 확률 = $P(x = 0) = P(T, T) = 0.25$

☒ 확률변수의 기대값? 평균

- $E(X) = \sum_{i=1}^{n} x_i - f(x_i)$
- x 확률변수의 기대값

☒ 확률변수 분산

- $Var(X) = E(X - \mu)^2$

☒ 공분산

- $Cov(X, Y) = E(X - \mu_X)(Y - \mu_Y)$
- X, Y 관계, 변화하는 척도

☒ 상관계수

- $Corr(X, Y) = \dfrac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$
- $-1 \leq Corr(X, Y) \leq 1$

☒ 이산형 확률분포(0,1)

- 베르누이 시행 : 실험결과 2가지
  * X = 1(성공), 0(실패)
  * $\Rightarrow f(x) = p^x(1-p)^{1-x}$
  * 성공확률:p, 실패확률:$1-p$

☒ 이항 분포

- 베르누이 시행을 독립적으로 여러번(n) 수행했을 때, 성공한 횟수의 분포
- $f(x) = \dfrac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$
- x : 성공한 횟수

☒ 포아송 분포

- 평균적으로 몇 번 일어날지 알고 있을 때, 이번에는 몇번 발생할까?

☒ 지수분포

☒ 정규분포

- 표준정규분포
  * 평균:0, 분산:1

---

p-vlaue

☒ ex. “이 모래가 매우 곱다” vs “모래는 모두 같다”
☒ 두 가설의 검정 결과를 판단하는 기준이 되는 값
☒ p-value가 0에 가까우면 ⇒ 두 가설에 차이가 확실히 존재하다.
☒ p-value가 1에 가까우면 ⇒ 우연하게 일어날 수 있는 흔한 차이
☒ 유의수준 : 5%

cf) ABtest

- 광고

광고	배너 A	배너 B	합
반응	35	25	60
무시	15	25	40
합	50	50	100

- 배너 A가 B보다 얼마만큼 효과가 있는지 확인하기 위한 test
- '2-표본 가설 검정'의 한 형태

유의수준

- 데이터의 차이가 유의미한 것인지 판단하는 기준
- 일반적으로 5%
- p값 < 유의수준 : 차이가 통계적으로 유의미하다.
- p값 > 유의수준 : 충분히 일어날 수 있는 차이

통계검정

- 귀무가설(H_0)
- 대립가설(H_1)
  - 귀무가설에 반대되는 가설
  - 입증해서 주장하고 싶은 가설
- 귀무가설이 아니라는 증거를 데이터로 부터 보여줌으로써 대립가설을 입증
- 귀무가설 기준 : 통계량 분포

t분포

- 표본평균이 0인지 아닌지 판단 사용
- ex. 맥주 4.2%(기준값)
  - 샘플 5잔 -> 알코올 도수
  - 관측치 : 4.15%, 4.19%, 4.21%, 4.23%, 4.27%
  - (관측치 - 기준값)으로 t분포 검정실시
  - $\bar{x} = \frac{-0.05 - 0.01 + 0.01 + 0.03 + 0.0.7}{5} = \frac{0.05}{5} = 0.01$
  - $s_x^2 = \frac{(-0.05 - 0.01)^2 + \cdots + (0.07 - 0.01)^2}{5 - 1} = 0.0004$
  - 표준편차 = $\sqrt{0.0004} = 0.02$
  - t-통계량 = $\frac{\quad}{\quad}$
  - t값? 데이터가 기준값으로 부터 얼마만큼 떨어져 있는지

다중공선성

- 어떤 독립(입력, x)변수가 다른 독립변수와 완벽한 선형 독립이 아닌 경우
- 회귀분석에 사용되는 일부변수가 다른변수와 상관정도가 큰 경우 => 분석 결과에 나쁜 영향을 줌
  - $\Rightarrow$ 회귀분석은 입력 변수기리 서로 독립이라고 가정하기 때문에
- 변수들 각각의 설명력이 낮아진다.

회귀분석

```
insurance = read.csv("dataset_for_ml/insurance.csv")
head(insurance)
```

```
##   age    sex  bmi children smoker    region expenses
## 1  19 female 27.9        0    yes southwest 16884.92
## 2  18   male 33.8        1     no southeast  1725.55
## 3  28   male 33.0        3     no southeast  4449.46
## 4  33   male 22.7        0     no northwest 21984.47
## 5  32   male 28.9        0     no northwest  3866.86
## 6  31 female 25.7        0     no southeast  3756.62
```

```
str(insurance)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ expenses: num  16885 1726 4449 21984 3867 ...
```

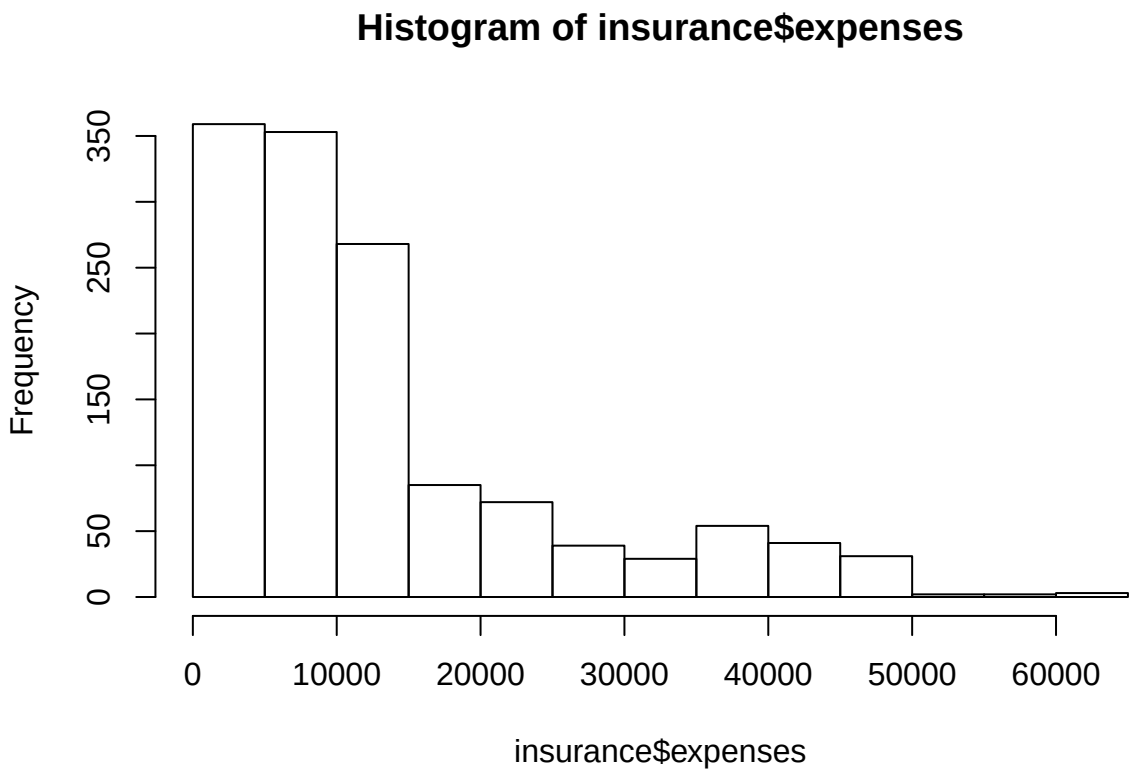
- 다른 변수들을 이용하여 expenses(보험료)를 예측

- 회귀모델을 구축하기 전에 정규성 확인
- 종속변수가 정규분포를 따르는 경우 모델이 잘 만들어짐

```
summary(insurance$expenses)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1122    4740    9382   13270   16640   63770
```

```r
hist(insurance$expenses)
```




```r
table(insurance$region)
```

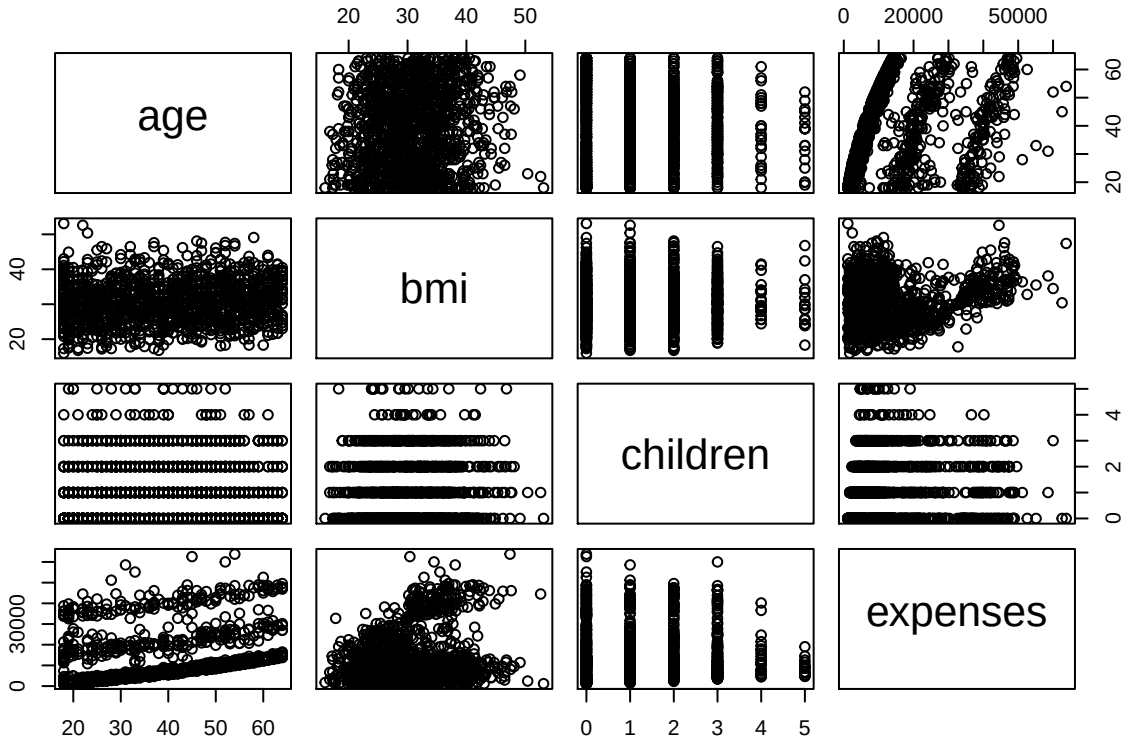
```
##
## northeast northwest southeast southwest
##       324       325       364       325
```

변수 상관 관계(상관행렬)

```r
cor(insurance[c("age", "bmi", "children", "expenses")])
```

```
##                age        bmi   children   expenses
## age      1.0000000 0.10934101 0.04246900 0.29900819
## bmi      0.1093410 1.00000000 0.01264471 0.19857626
## children 0.0424690 0.01264471 1.00000000 0.06799823
## expenses 0.2990082 0.19857626 0.06799823 1.00000000
```

```r
pairs(insurance[c("age", "bmi", "children", "expenses")])
```



pairs.panels()

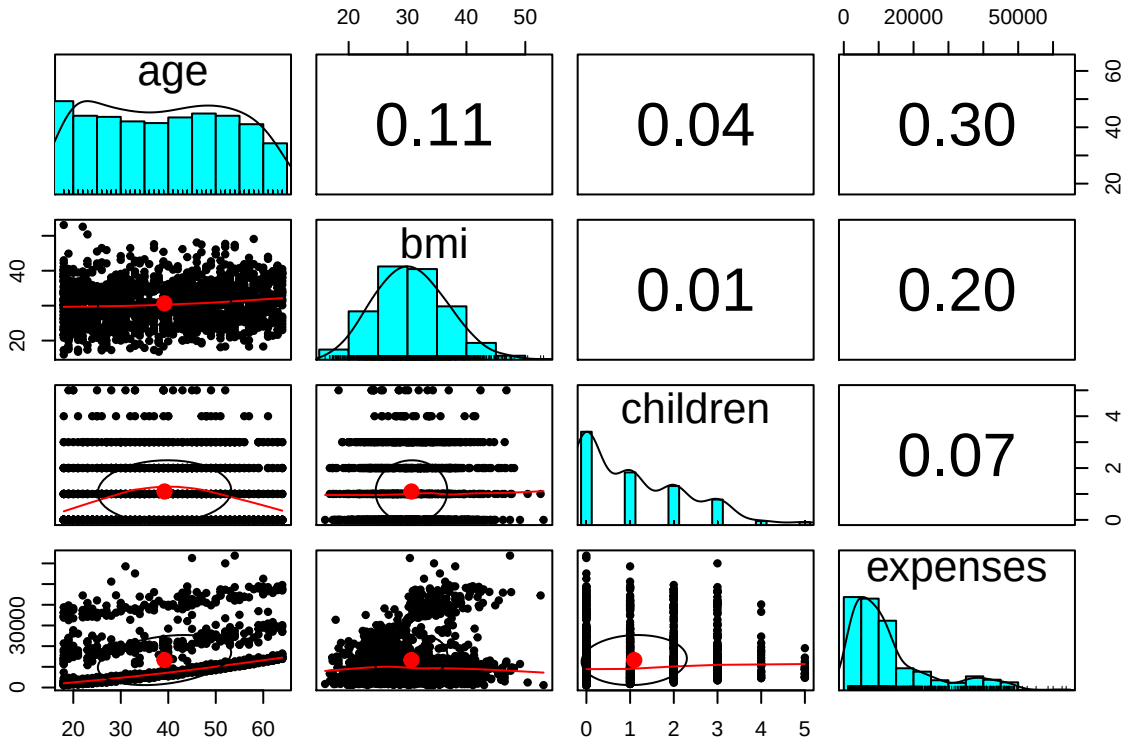
```
install.packages("psych")
```

```
## Installing package into 'C:/Users/student/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
```

```
library(psych)
```

```
pairs.panels(insurance[c("age", "bmi", "children", "expenses")])
```



모델 생성

```
ins_model = lm(expenses~age+bmi+children+sex+smoker+region, data = insurance)
```

```
ins_model
```

```
##
## Call:
## lm(formula = expenses ~ age + bmi + children + sex + smoker +
##     region, data = insurance)
##
## Coefficients:
##     (Intercept)              age              bmi         children
##       -11941.6            256.8            339.3            475.7
##         sexmale        smokeryes  regionnorthwest  regionsoutheast
##          -131.4          23847.5           -352.8          -1035.6
## regionsouthwest
##          -959.3
```

☒ 모델 요약

```r
summary(ins_model)
```

```
##
## Call:
## lm(formula = expenses ~ age + bmi + children + sex + smoker +
##     region, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11302.7  -2850.9   -979.6   1383.9  29981.7
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11941.6      987.8 -12.089  < 2e-16 ***
## age                256.8       11.9  21.586  < 2e-16 ***
## bmi                339.3       28.6  11.864  < 2e-16 ***
## children           475.7      137.8   3.452 0.000574 ***
## sexmale           -131.3      332.9  -0.395 0.693255
## smokeryes        23847.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -352.8      476.3  -0.741 0.458976
## regionsoutheast  -1035.6      478.7  -2.163 0.030685 *
## regionsouthwest   -959.3      477.9  -2.007 0.044921 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

- R-squared(결정계수, r제곱값)
- 모델이 종속변수 값을 얼마나 잘 설명하는가?

- 독립/종속 변수 : 선형 가정
- 비선형 관계 : 높은 차수 항을 모델에 추가

```r
insurance$age2 = insurance$age^2
# lm(expenses~age+age2, data = insurance)
```

- bmi지수에 따라 구간화

```r
insurance$bmi30 = ifelse(insurance$bmi >= 30, 1, 0)
```

```r
ins_model2 = lm(expenses ~ age+age2+children+bmi+sex+bmi30*smoker+region, data = insurance)
summary(ins_model2)
```

```
##
## Call:
## lm(formula = expenses ~ age + age2 + children + bmi + sex + bmi30 *
##     smoker + region, data = insurance)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -17297.1  -1656.0  -1262.7   -727.8  24161.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       139.0053  1363.1359   0.102 0.918792
## age               -32.6181    59.8250  -0.545 0.585690
## age2                3.7307     0.7463   4.999 6.54e-07 ***
## children          678.6017   105.8855   6.409 2.03e-10 ***
## bmi               119.7715    34.2796   3.494 0.000492 ***
## sexmale          -496.7690   244.3713  -2.033 0.042267 *
## bmi30            -997.9355   422.9607  -2.359 0.018449 *
## smokeryes       13404.5952   439.9591  30.468  < 2e-16 ***
## regionnorthwest  -279.1661   349.2826  -0.799 0.424285
## regionsoutheast  -828.0345   351.6484  -2.355 0.018682 *
## regionsouthwest -1222.1619   350.5314  -3.487 0.000505 ***
## bmi30:smokeryes 19810.1534   604.6769  32.762  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4445 on 1326 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8653
## F-statistic: 781.7 on 11 and 1326 DF,  p-value: < 2.2e-16
```

연습문제

air quailty data

```
library(readxl)
air = read_xlsx("air_0.2_.xlsx")
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting date in A99979 / R99979C1: got ' '
```

```
head(air)
```

```
## # A tibble: 6 x 16
##   Date                Locate  pm10  pm24 pm2.5    o3   no2    co   so2
##   <dttm>              <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 549776392-07-23 00:00:02       37    42    18 0.016 0.014   0.5 0.006
## 2 549776392-10-30 23:59:58       34    41    24 0.016 0.012   0.4 0.007
## 3 549776393-02-08 00:00:00       31    39    18 0.016 0.011   0.4 0.007
## 4 549776393-05-19 00:00:00       41    38    10 0.017 0.01    0.4 0.006
## 5 549776393-08-27 00:00:02       26    36    11 0.017 0.01    0.4 0.006
## 6 549776393-12-04 23:59:58       22    34    17 0.016 0.012   0.4 0.006
## # ... with 7 more variables: `PM10(aqi)` <dbl>, `PM2.5(aqi)` <dbl>,
## #   `o3(aqi)` <dbl>, `no2(aqi)` <dbl>, `co(aqi)` <dbl>, `so2(aqi)` <dbl>,
## #   humidity <dbl>
```

```r
air_sub = air[c("pm2.5", "o3", "no2", "co", "so2")]
head(air_sub)
```

```
## # A tibble: 6 x 5
##   pm2.5    o3   no2    co   so2
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1    18 0.016 0.014   0.5 0.006
## 2    24 0.016 0.012   0.4 0.007
## 3    18 0.016 0.011   0.4 0.007
## 4    10 0.017 0.01    0.4 0.006
## 5    11 0.017 0.01    0.4 0.006
## 6    17 0.016 0.012   0.4 0.006
```

```r
air_model = lm(pm2.5 ~ o3+co+no2+so2, data = air_sub)
summary(air_model)
```

```
##
## Call:
## lm(formula = pm2.5 ~ o3 + co + no2 + so2, data = air_sub)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -99.33  -7.31  -1.86   5.13 451.54
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.7231     0.1413  -47.57   <2e-16 ***
## o3           224.1681     2.4961   89.81   <2e-16 ***
## co            29.9557     0.2465  121.53   <2e-16 ***
## no2          161.4466     3.8465   41.97   <2e-16 ***
## so2         1032.2903    21.3164   48.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.99 on 99972 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4164, Adjusted R-squared:  0.4164
## F-statistic: 1.783e+04 on 4 and 99972 DF,  p-value: < 2.2e-16
```

```r
air_sub$o3_2 = air_sub$o3^2
air_sub$no2_2 = air_sub$no2^2
air_sub$co_2 = air_sub$co^2
air_sub$so2_2 = air_sub$so2^2
```

```r
air_model2 = lm(pm2.5 ~ o3+co+no2 +o3_2+no2_2+co_2+so2_2+ o3*co , data = air_sub)
summary(air_model2)
```

```
##
## Call:
## lm(formula = pm2.5 ~ o3 + co + no2 + o3_2 + no2_2 + co_2 + so2_2 +
##     o3 * co, data = air_sub)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -70.90  -7.12  -1.47   5.13 452.16
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.7659     0.2725  -6.481 9.13e-11 ***
## o3            -96.8098     8.8871 -10.893  < 2e-16 ***
## co             31.1451     0.6287  49.540  < 2e-16 ***
## no2           122.1678     9.6874  12.611  < 2e-16 ***
## o3_2         -876.0150    73.5709 -11.907  < 2e-16 ***
## no2_2         258.9185   119.0331   2.175   0.0296 *
## co_2           -4.9216     0.3139 -15.681  < 2e-16 ***
## so2_2       63185.3119  1246.6721  50.683  < 2e-16 ***
## o3:co         932.5702    13.5728  68.709  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.45 on 99968 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4642, Adjusted R-squared:  0.4642
## F-statistic: 1.083e+04 on 8 and 99968 DF,  p-value: < 2.2e-16
```