

Analysis On Factors Affecting Sales Of Commercial Districts In Seoul

Team EvE

2018년 4월 13일



Members : 전현수(leader), 강종일, 김도현, 장훈희

1. 분석목적 (Purpose of Analysis)

현재 국가의 창업 장려 정책에 따라, 청년창업자 혹은 창업지망자들이 날이 갈수록 늘어나고 있는 실정이다. 또한 창업자 대부분이 자본이 필요한 기술분야 혹은 제조업 분야가 아닌 소규모영업 분야로 나아가고 있다. 하지만 압구정로데오, 혹은 신촌 상권의 역사에서 볼 수 있듯이, 최근에는 떠오르는 상권 혹은 소멸해가는 상권의 변화가 매우 빠르게 일어나고 있다. 청년 창업가의 입장에서 본다면 당연하게도 활발한 상권에 입주하고 싶을 것이라 가정 하에, 어떤 요인이 활발한 상권을 만드는 지에 대한 분석을 해보기로 하였다.

활발한 상권을 정의하기는 상당히 어렵겠지만, 사용한 데이터는 넓은 범위의 상권이 아닌 골목 단위의 골목 상권의 데이터이므로 총 매출이 활발함의 정도를 파악하는 좋은 변수라 판단, 상권의 총 매출을 종속변수로 정하게 되었다. 사용한 데이터는 서울시 전체의 골목상권에 대해 집계한 데이터로서, 특이한 점이 있다면 상권 데이터뿐만 아니라 상권을 지탱하고 있는 상권 배후지에 대한 데이터도 포함하고 있었다. 상식적 추론에 따라서, 상권의 집객 시설 수, 유동인구 수, 직장 인구 수, 점포 수 등과 함께 배후지의 상주인구, 아파트 수, 인구의 소득/소비가 설명 변수에 해당할 것이라 예측했다. 하지만 혹 우리의 추론이 틀릴 경우를 가정, 데이터가 포함하고 있는 모든 변수를 설명 변수로 놓고, 변수별 유의성에 따라 설명 변수를 하나씩 소거하는 방법을 채택하기로 하였다.

한편, 최저임금과 임대료 등이 창업의 과정에서 중요한 영향을 미칠 것이라는 것은 예상 가능하나, 현재의 분석 목표는 상권 그 자체의 활발함을 계측해내는 것이 목적이기에 종속변수를 순이익이 아닌 총매출합으로 설정하였으므로 직접적인 상관관계를 보이진 않을 거라 생각하여 설명변수에서 제외하였다. 이번 프로젝트 데이터의 표본 집단은 서울특별시의 모든 상권으로서, 총 매출에 대한 표본으로서는 충분히 적합한 데이터라 판단하였다. 또한 전반적으로 모든 설명변수를 활용하기 위하여 변수에 결측치가 있는 데이터는 제외하고 분석을 진행하였다.

2. 모델 설정 (Set Analysis Model)

우리가 분석하고자 하는 모형에서의 종속변수는 각 골목상권별 월간 매출의 총합이며, 설명변수는 각 상권과 배후지 모두에 각기 8개씩 총 16개가 존재한다. 각각의 변수들은, N_Aparts(Number of Apartments)는 아파트의 수, N_Facs(Number of Gathering Facilities)는 집객시설의 수, N_Store(Number of Stores)는 점포의 수, Income은 그 지역 거주민의 수입, Spend는 그 지역 거주민의 소비금액, F_Pop(Floating Population)은 유동인구 수, S_Pop(Settled Population)는 거주인구 수, W_Pop(Working Population)는 직장인구를 의미한다. 그 중 배후지의 총 매출을 제외한 총 16개의 설명변수를 채택하였다.

일견 변수 간 다중공선성의 발생 확률이 존재할 것을 예상할 수 있다. 아파트의 수가 많을수록 관공서, 은행 등의 집객시설이 많을 수 있고, 수입이 많을수록 소비도 많을 것이 예측 가능하기 때문이다. 하지만 수입이 많다고 반드시 소비가 많은 것은 아니듯 특정 변수를 상식상의 예측을 통해서 빼는 것은 올바른 접근 방법이 아니라 판단하기에 일단은 모든 변수를 통해 분석하고 각 변수가 종속변수에 유의미한 영향을 미치는 후 확인한 뒤 소거를 통해서 유효한 모델을 선택하는 접근 방식을 택하기로 하였다. 앞서 설명했듯이, 상권은 트렌드에 민감하고 흥망의 변동이 굉장히 빠르기에 2015년부터 2018년 1월까지의 데이터 중 가장 최신 데이터인 2018년 1월 데이터를 활용하기로 결정하였다. 우리가 분석하고자 하는 1차적 모형은 다음과 같다.

1 차적으로 설정한 모델

$$SALES_i = \beta_0 + \beta_1 C_{NAparts} + \beta_2 C_{NFacs} + \beta_3 C_{NStore} + \beta_4 C_{Income} + \beta_5 C_{Spend} + \beta_6 C_{FPop} + \beta_7 C_{SPop} + \beta_8 C_{WPop} + \beta_9 H_{NAparts} + \beta_{10} H_{NFacs} + \beta_{11} H_{NStore} + \beta_{12} H_{Income} + \beta_{13} H_{Spend} + \beta_{14} H_{WPop} + \beta_{15} H_{SPop} + \beta_{16} H_{WPop} + \epsilon_i$$

다음으로 우리가 검정하고자 하는 가설은 다음과 같다

$$H_0 : \beta_1 = 0, H_0 : \beta_2 = 0, \dots, H_0 : \beta_{16} = 0$$

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{16} = 0$$

이 모델을 검정하기 위해 R을 이용하여 다중회귀분석을 실시하였다

3. 데이터 전처리 (Preprocess of Data)

필요한 패키지들의 라이브러리 불러오기 (Load necessary packages)

Directory 설정

```
setwd(dir = "C:/Users/eljuw/Desktop/Team EVE/")
```

i. 데이터 불러오기

참고 : 데이터는 서울시 공공데이터를 사용 서울시 열린데이터 광장 (<http://data.seoul.go.kr/dataList/datasetList.do>)

서울시열린데이터광장의 서울시 골목상권 프로파일 정보의 상권과 상권배후지의 데이터들을 엑셀형식으로 다운로드, 전체 데이터는 2015년부터 2018년 1월까지 존재, 서울시 열린데이터 광장에 상권관련한 자료는 총 20개가 존재, 하지만 실제로 사용할 데이터는 상권과 상권배후지의 데이터 15개이다.

RStudio에서 encoding 오류를 방지하기 위해 모든 파일명을 영어로 수정 상권의 헤더는 Commercial_로, 배후지 헤더는 Hinter_로 설정
아파트 = N_Aparts, 집객시설 = N_Facs, 점포수 = N_Store, 유동인구 = F_Pop, 상주인구 = S_Pop, 직장인구 = W_Pop, 추정매출 = Sales, 소득소비 = Income_Spend

상권 : Commercial District, 상권배후지 : Hinterland, 아파트 : Number of Apartments, 집객시설 : Number of Gathering Facilities, 유동인구 = Floating Population, 상주인구 : Settled Population, 직장인구 : Working Population, 추정매출 : Estimated Sales, 소득/소비 : Income and Spend

상권의 원데이터 불러오기

read_excel을 이용하여 원데이터 읽어오기

상권 배후지의 원데이터 불러오기

ii. 데이터 전처리

불러온 데이터들은 Raw Data로써 실제 사용할 데이터가 아닌 것들이 많기 때문에 전처리를 통해서 필요한 데이터들만 사용

필요한 열만 데이터로 저장하기(상권)

각각 데이터 프레임에서 사용할 데이터만 추출한다

사용할 데이터는 2018년 1월 데이터이므로 모든 데이터에서 2018년 1월 데이터를 추출하고

추출한 후 데이터를 보는 기준인 상권코드를 오름차순으로 정렬한다

모두 데이터프레임으로 다시 저장

필요한 열만 데이터로 저장하기(상권배후지)

모두 데이터프레임으로 다시 저장

iii. 불러온 데이터 모두 합치기

각 데이터들을 상권과 상권배후지로 각각 합쳐서 데이터를 두 개로 만든다

이번 프로젝트에서는 종속변수가 상권의 매출 C_Sales 이므로 이를 기준으로 잡는다

모든 변수들은 상권_코드로 분류가 되어 있기 때문에 이를 이용하여 데이터를 통합한다

C_Sales에는 1744개의 상권_코드가 존재한다 그러므로 1부터 1744까지의 데이터 프레임을 생성 후 이에 맞춰서 합친다

상권데이터 합치기

```
NaTotal_Commercial <- Total_Commercial
colnames(NaTotal_Commercial) <- c("Code", "Sales", "C_N_Aparts", "C_N_Facs",
                                    "C_Store", "C_Income", "C_Spend", "C_F_Pop",
                                    "C_S_Pop", "C_W_Pop")
```

상권배후지 데이터 합치기

iv. 결측치 제거 및 데이터 정리

두 파일 모두 C_Sales의 상권코드를 기준으로 합쳐졌지만 원데이터가 모든 변수들이 상권별로 조사된 것이 아니어서 결측치가 존재함이 결측치를 가지고 분석을 한다면 문제가 생길 여지가 있다. 이런 결측치가 있는 상권은 데이터 분석에서 제외하는 것이 나을 것으로 판단하여 제외하였다.

결측치 수 확인

```
sum(is.na(Total_Commercial))
```

```
## [1] 16668
```

```
sum(is.na(Total_Hinter))
```

```
## [1] 14276
```

결측치가 있을경우엔 데이터 분석에 문제가 생길수 있음으로 모든 결측치를 제거한다

```
Total_Commercial <- Total_Commercial[complete.cases(Total_Commercial), ]
Total_Hinter <- Total_Hinter[complete.cases(Total_Hinter), ]
```

편리성과 통일성을 위하여 모든 colnames를 영어로 변경

기존 colnames확인

```
colnames(Total_Commercial)
```

```
## [1] "상권_코드"          "당월_매출_금액"      "아파트_단지_수"
## [4] "집객시설_수"        "점포_수"            "월_평균_소득_금액"
## [7] "지출_총금액"        "총_유동인구_수"      "총_상주인구_수"
## [10] "총_직장_인구_수"
```

```
colnames(Total_Hinter)
```

```
## [1] "상권_코드"          "아파트_단지_수"      "집객시설_수"      "점포_수"
## [5] "월평균소득_금액"    "지출_총금액"        "총_유동인구_수"    "총_상주인구_수"
## [9] "총_직장_인구_수"
```

colnames를 기준에 설정한 영어명으로 변경

```
colnames(Total_Commercial) <- c("Code", "Sales", "C_N_Aparts", "C_N_Facs", "C_Store", "C_Income",
                                 "C_Spend", "C_F_Pop", "C_S_Pop", "C_W_Pop")
colnames(Total_Hinter) <- c("Code", "H_N_Aparts", "H_N_Facs", "H_Store", "H_Income", "H_Spend",
                           "H_F_Pop", "H_S_Pop", "H_W_Pop")
```

엑셀로 저장

v. 상권과 상권 배후지 데이터 합치기

상권의 매출이 종속변수이기 때문에 상권의 매출을 기준으로 하여 모두 하나의 데이터로 합친다

합친 데이터에 결측치가 존재하는지 확인한다

```
sum(is.na(All_Total))
```

```
## [1] 5165
```

결측치가 존재하는 걸 확인하였으므로 결측치 제거

```
All_Total <- All_Total[complete.cases(All_Total), ]
```

이렇게 완성된 파일을 엑셀로 저장

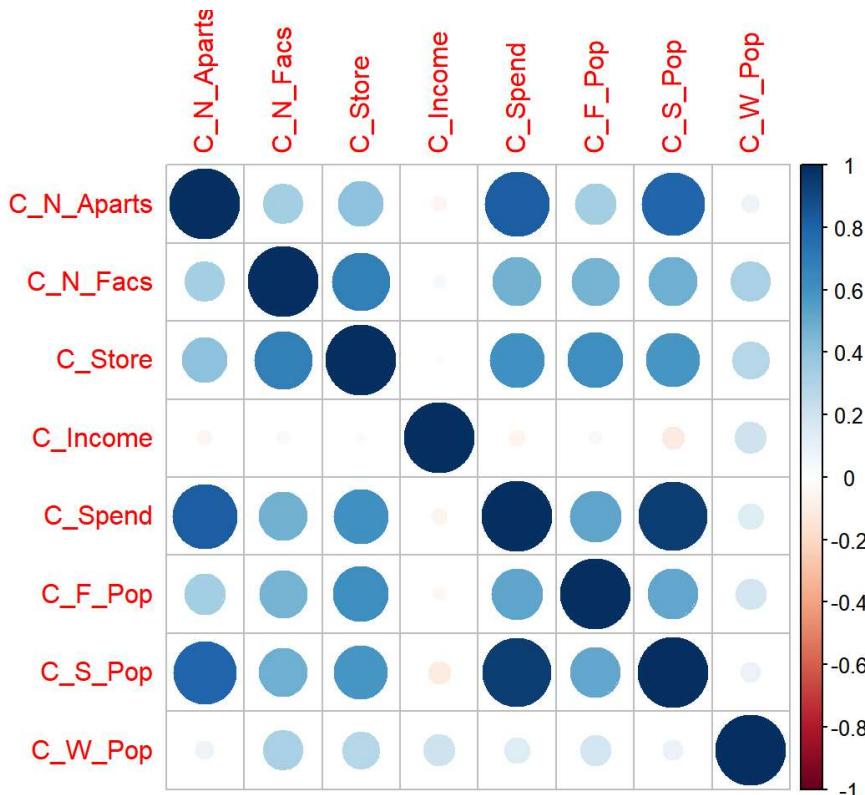
4. 회귀분석 전 탐색적 데이터 분석 (Data Briefing)

회귀분석 전 정리된 데이터를 분석해본다

```
##               mean      sd      min      max      se
## Sales       1.445418e+09 2.118692e+09 4027152 53357670022 5.931199e+07
## C_N_Aparts* 3.046473e+01 4.095313e+01     1      619 1.146468e+00
## C_N_Facs*   3.129232e+01 2.786159e+01     2      219 7.799748e-01
## C_Store     8.388166e+01 6.764471e+01     1      454 1.893688e+00
## C_Income*   2.940687e+06 7.567906e+05 1045878 8115507 2.118607e+04
## C_Spend*    9.397536e+08 8.426495e+08 13570299 10421880832 2.358966e+07
## C_F_Pop*   5.081332e+04 5.063067e+04     9      582739 1.417387e+03
## C_S_Pop*   1.936193e+03 1.648528e+03    31      19310 4.614992e+01
## C_W_Pop*   5.231066e+02 1.298205e+03     2      19502 3.634276e+01
```

변수들 간의 상관관계를 분석해보았다

```
Total_Commercial <- as.data.frame(Total_Commercial)
Total_Commercial <- dplyr::mutate_all(Total_Commercial, as.numeric)
M <- cor(Total_Commercial[, -c(1, 2)])
corrplot::corrplot(M, method = "circle")
```



5. 다중 회귀분석 (Multiple Regression)

Model_1 : 상권의 매출을 종속변수로 하여 상권과 상권배후지의 독립변수들 간의 다중회귀분석

i. 회귀분석과 변수선택 실시

```
lm.result.all.m1 <- lm(Sales ~ . - Code, data = All_Total) #Code는 독립변수가 아님으로 비교하지 않음
lm.result.forward.m1 <- step(lm.result.all.m1, direction = "forward")
lm.result.backward.m1 <- step(lm.result.all.m1, direction = "backward")
lm.result.stepwise.m1 <- step(lm.result.all.m1, direction = "both")
```

ii. 결과확인

```
summary(lm.result.all.m1)
```

```
##  
## Call:  
## lm(formula = Sales ~ . - Code, data = All_Total)  
##  
## Residuals:  
##      Min       1Q   Median     3Q    Max  
## -1.236e+10 -3.346e+08 -3.771e+07  2.380e+08  3.505e+10  
##  
## Coefficients: (1 not defined because of singularities)  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -6.532e+08 3.346e+08 -1.952 0.0512 .  
## C_N_Aparts -2.467e+06 2.692e+06 -0.916 0.3596  
## C_N_Facs   1.437e+07 2.810e+06  5.114 3.75e-07 ***  
## C_Store    1.052e+07 1.291e+06  8.149 1.05e-15 ***  
## C_Income   1.102e+02 8.316e+01  1.325 0.1854  
## C_Spend    2.840e-01 2.157e-01  1.317 0.1881  
## C_F_Pop   -3.196e+02 1.320e+03 -0.242 0.8087  
## C_S_Pop   -2.666e+05 1.037e+05 -2.572 0.0103 *  
## C_W_Pop    7.175e+05 4.398e+04 16.313 < 2e-16 ***  
## H_N_Aparts 3.504e+05 3.814e+05  0.919 0.3586  
## H_N_Facs   1.044e+06 1.187e+06  0.879 0.3794  
## H_Store    -2.190e+05 1.785e+05 -1.226 0.2203  
## H_Income   6.059e+00 1.048e+02  0.058 0.9539  
## H_Spend    2.287e-02 3.835e-02  0.596 0.5512  
## H_F_Pop   3.018e+04 3.319e+04  0.910 0.3633  
## H_S_Pop   -5.285e+03 1.565e+04 -0.338 0.7356  
## H_W_Pop      NA       NA       NA       NA  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 1.552e+09 on 1034 degrees of freedom  
## Multiple R-squared:  0.4968, Adjusted R-squared:  0.4895  
## F-statistic: 68.04 on 15 and 1034 DF, p-value: < 2.2e-16
```

```
summary(lm.result.forward.m1)
```

```

## 
## Call:
## lm(formula = Sales ~ (Code + C_N_Aparts + C_N_Facs + C_Store +
##   C_Income + C_Spend + C_F_Pop + C_S_Pop + C_W_Pop + H_N_Aparts +
##   H_N_Facs + H_Store + H_Income + H_Spend + H_F_Pop + H_S_Pop +
##   H_W_Pop) - Code, data = All_Total)
## 
## Residuals:
##      Min       1Q     Median      3Q      Max 
## -1.236e+10 -3.346e+08 -3.771e+07  2.380e+08  3.505e+10 
## 
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -6.532e+08 3.346e+08 -1.952 0.0512 .  
## C_N_Aparts -2.467e+06 2.692e+06 -0.916 0.3596    
## C_N_Facs   1.437e+07 2.810e+06  5.114 3.75e-07 *** 
## C_Store    1.052e+07 1.291e+06  8.149 1.05e-15 *** 
## C_Income   1.102e+02 8.316e+01  1.325 0.1854    
## C_Spend    2.840e-01 2.157e-01  1.317 0.1881    
## C_F_Pop   -3.196e+02 1.320e+03 -0.242 0.8087    
## C_S_Pop   -2.666e+05 1.037e+05 -2.572 0.0103 *  
## C_W_Pop    7.175e+05 4.398e+04 16.313 < 2e-16 *** 
## H_N_Aparts 3.504e+05 3.814e+05  0.919 0.3586    
## H_N_Facs   1.044e+06 1.187e+06  0.879 0.3794    
## H_Store   -2.190e+05 1.785e+05 -1.226 0.2203    
## H_Income   6.059e+00 1.048e+02  0.058 0.9539    
## H_Spend    2.287e-02 3.835e-02  0.596 0.5512    
## H_F_Pop    3.018e+04 3.319e+04  0.910 0.3633    
## H_S_Pop   -5.285e+03 1.565e+04 -0.338 0.7356    
## H_W_Pop      NA       NA       NA       NA      
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 1.552e+09 on 1034 degrees of freedom 
## Multiple R-squared:  0.4968, Adjusted R-squared:  0.4895 
## F-statistic: 68.04 on 15 and 1034 DF, p-value: < 2.2e-16

```

```
summary(lm.result.backward.m1)
```

```

## 
## Call:
## lm(formula = Sales ~ C_N_Facs + C_Store + C_Income + C_S_Pop +
##   C_W_Pop + H_N_Aparts, data = All_Total)
## 
## Residuals:
##      Min       1Q     Median      3Q      Max 
## -1.259e+10 -3.289e+08 -4.645e+07  2.422e+08  3.517e+10 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -4.765e+08 2.099e+08 -2.270 0.0234 *  
## C_N_Facs   1.508e+07 2.533e+06  5.955 3.56e-09 *** 
## C_Store    1.065e+07 1.107e+06  9.616 < 2e-16 *** 
## C_Income   1.327e+02 6.380e+01  2.079 0.0378 *  
## C_S_Pop   -1.750e+05 4.307e+04 -4.063 5.21e-05 *** 
## C_W_Pop    7.260e+05 4.227e+04 17.175 < 2e-16 *** 
## H_N_Aparts 3.820e+05 2.124e+05  1.798 0.0725 .  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 1.55e+09 on 1043 degrees of freedom 
## Multiple R-squared:  0.4934, Adjusted R-squared:  0.4905 
## F-statistic: 169.3 on 6 and 1043 DF, p-value: < 2.2e-16

```

```
summary(lm.result.stepwise.m1)
```

```

## 
## Call:
## lm(formula = Sales ~ C_N_Facs + C_Store + C_Income + C_S_Pop +
##      C_W_Pop + H_N_Aparts, data = All_Total)
## 
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -1.259e+10 -3.289e+08 -4.645e+07  2.422e+08  3.517e+10 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -4.765e+08  2.099e+08 -2.270  0.0234 *  
## C_N_Facs     1.508e+07  2.533e+06  5.955 3.56e-09 *** 
## C_Store      1.065e+07  1.107e+06  9.616 < 2e-16 *** 
## C_Income     1.327e+02  6.380e+01  2.079  0.0378 *  
## C_S_Pop      -1.750e+05 4.307e+04 -4.063 5.21e-05 *** 
## C_W_Pop      7.260e+05  4.227e+04 17.175 < 2e-16 *** 
## H_N_Aparts   3.820e+05  2.124e+05  1.798  0.0725 .  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 1.55e+09 on 1043 degrees of freedom 
## Multiple R-squared:  0.4934, Adjusted R-squared:  0.4905 
## F-statistic: 169.3 on 6 and 1043 DF,  p-value: < 2.2e-16

```

어떤 방식을 사용할지를 정하기 위해 Akaike Information Criterion(AIC)으로 확인한다

```
AIC(lm.result.all.m1)
```

```
## [1] 47439.57
```

```
AIC(lm.result.forward.m1)
```

```
## [1] 47439.57
```

```
AIC(lm.result.backward.m1)
```

```
## [1] 47428.46
```

```
AIC(lm.result.stepwise.m1)
```

```
## [1] 47428.46
```

stepwise와 **backward** 방식의 AIC가 42428.46으로 **all**과 **forward**의 47439.57 보다 낮다

summary로 확인한 **backward**와 **stepwise**모델이 동일한 형태임을 확인하였으므로 **stepwise** 방식을 선택

iii. 모델해석

```

## 
## Call:
## lm(formula = Sales ~ C_N_Facs + C_Store + C_Income + C_S_Pop +
##      C_W_Pop + H_N_Aparts, data = All_Total)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -1.259e+10 -3.289e+08 -4.645e+07  2.422e+08  3.517e+10 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -4.765e+08  2.099e+08 -2.270  0.0234 *  
## C_N_Facs     1.508e+07  2.533e+06  5.955 3.56e-09 *** 
## C_Store      1.065e+07  1.107e+06  9.616 < 2e-16 *** 
## C_Income     1.327e+02  6.380e+01  2.079  0.0378 *  
## C_S_Pop      -1.750e+05 4.307e+04 -4.063 5.21e-05 *** 
## C_W_Pop      7.260e+05  4.227e+04 17.175 < 2e-16 *** 
## H_N_Aparts   3.820e+05  2.124e+05  1.798  0.0725 .  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 1.55e+09 on 1043 degrees of freedom 
## Multiple R-squared:  0.4934, Adjusted R-squared:  0.4905 
## F-statistic: 169.3 on 6 and 1043 DF, p-value: < 2.2e-16

```

1단계 : 회귀모형은 타당한가?

귀무가설 : 회귀모형은 타당하지 않다

대립가설 : 회귀모형은 타당하다

F-statistic: 169.3 on 6 and 1043 DF, p-value: < 0.0000000000000022

결론 : 유의확률이 0.000 이므로 유의수준 0.05보다 낮아 대립가설을 지지한다. 그러므로 회귀모형은 통계적으로 타당하다

2단계 : 독립변수들은 종속변수에게 영향을 주는가?

Variables	t-value	p-value
C_N_Facs	(t = 5.952,	p < 0.001)
C_Store	(t = 9.565,	p < 0.001)
C_S_Pop	(t = -3.667,	p < 0.001)
C_W_Pop	(t = 17.070,	p < 0.001)
C_Income	(t = 1.994,	p < 0.05)
H_N_Aparts	(t = 1.798,	p > 0.05)

결론 : H_N_Aparts 는 유의수준 0.05에서 종속변수에게 통계적으로 유의한 영향이 없음

그럼으로 통계적으로 무의미한 변수 H_N_Aparts 를 제외하고 모델2를 새로 구성함

Model_2 : Model_1의 stepwise방식에 통계적으로 무의미한 변수 H_N_Aparts를 제거한 모델

```

lm.result.final.m2 <- lm(Sales ~ C_N_Facs + C_Store + C_S_Pop + C_W_Pop + C_Income, data = All_Total)
summary(lm.result.final.m2)

```

```

## 
## Call:
## lm(formula = Sales ~ C_N_Facs + C_Store + C_S_Pop + C_W_Pop +
##      C_Income, data = All_Total)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.262e+10 -3.149e+08 -4.806e+07  2.308e+08  3.533e+10 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -3.968e+08  2.054e+08 -1.932 0.053619 .  
## C_N_Facs     1.509e+07  2.536e+06  5.952 3.61e-09 *** 
## C_Store      1.060e+07  1.108e+06  9.565 < 2e-16 *** 
## C_S_Pop     -1.334e+05  3.639e+04 -3.667 0.000258 *** 
## C_W_Pop      7.202e+05  4.219e+04 17.070 < 2e-16 *** 
## C_Income     1.272e+02  6.380e+01  1.994 0.046460 *  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 1.552e+09 on 1044 degrees of freedom 
## Multiple R-squared:  0.4919, Adjusted R-squared:  0.4894 
## F-statistic: 202.1 on 5 and 1044 DF, p-value: < 2.2e-16

```

1단계 : 회귀모형은 타당한가?

F-statistic: 202.1 on 5 and 1044 DF, p-value: < 0.00000000000000022

결론 : 유의확률이 0.000 이므로 유의수준 0.05보다 낮아 대립가설을 지지한다 그럼으로 회귀모형은 통계적으로 타당하다

2단계 : 독립변수들이 종속변수에게 영향을 주는가?

Variables	t-value	p-value
C_N_Facs	(t = 6.657,	p < 0.001)
C_Store	(t = 12.109,	p < 0.001)
C_S_Pop	(t = -4.159,	p < 0.001)
C_W_Pop	(t = 21.881,	p < 0.001)
C_Income	(t = 2.192,	p < 0.05)

독립변수 5개 모두가 종속변수에게 유의수준 0.05에서 통계적으로 유의한 영향을 주는 것으로 나타났다

3단계 : 독립변수들은 어떤 영향을 주는가?

Variables	
C_N_Facs	15093128.3
C_Store	10599641.0
C_S_Pop	-133433.1
C_W_Pop	720224.5
C_Income	127.2

C_N_Facs는 다른 네 개의 독립변수 (C_Store, C_S_Pop, C_W_Pop, C_Income)가 고정되어있을 때에, C_N_Facs의 기본단위가 1 증가하면, 종속변수, Sales는 약 15093128.3 정도 증가된다

C_Store는 다른 네 개의 독립변수 (C_N_Facs, C_S_Pop, C_W_Pop, C_Income)가 고정되어있을 때에, C_Store의 기본단위가 1 증가하면, 종속변수, Sales는 약 10599641.0 정도 증가된다

C_S_Pop는 다른 네 개의 독립변수 (C_N_Facs, C_Store, C_W_Pop, C_Income)가 고정되어있을 때에, C_S_Pop의 기본단위가 1 증가하면, 종속변수, Sales는 약 133433.1 정도 감소된다

C_W_Pop는 다른 네 개의 독립변수 (C_N_Facs, C_Store, C_S_Pop, C_Income)가 고정되어있을 때에, C_W_Pop의 기본단위가 1 증가하면, 종속변수, Sales는 약 720224.5 정도 증가된다

C_Income는 다른 네 개의 독립변수 (C_N_Facs, C_Store, C_S_Pop, C_W_Pop)가 고정되어있을 때에, C_Income의 기본단위가 1 증가하면, 종속변수, Sales는 약 127.2 정도 증가된다

4단계 : 독립변수들의 설명력

Adjusted R-squared: 0.4894

네 개의 독립변수가 종속변수를 약 48.94% 설명

iv. 모델 검정

다중공선성 점검

```
## C_N_Facs C_Store C_S_Pop C_W_Pop C_Income
## 2.163874 2.443343 1.669860 1.212911 1.073630
```

모든 변수의 vif가 4 미만이므로 다중공선성은 존재하지 않는 것으로 보임

잔차분석

```
##
## Call:
## lm(formula = Sales ~ C_N_Facs + C_Store + C_S_Pop + C_W_Pop +
##     C_Income, data = All_Total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.262e+10 -3.149e+08 -4.806e+07  2.308e+08  3.533e+10 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -3.968e+08  2.054e+08 -1.932 0.053619 .  
## C_N_Facs     1.509e+07  2.536e+06  5.952 3.61e-09 *** 
## C_Store      1.060e+07  1.108e+06  9.565 < 2e-16 *** 
## C_S_Pop     -1.334e+05  3.639e+04 -3.667 0.000258 *** 
## C_W_Pop      7.202e+05  4.219e+04 17.070 < 2e-16 *** 
## C_Income     1.272e+02  6.380e+01  1.994 0.046460 *  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 1.552e+09 on 1044 degrees of freedom
## Multiple R-squared:  0.4919, Adjusted R-squared:  0.4894 
## F-statistic: 202.1 on 5 and 1044 DF, p-value: < 2.2e-16
## 
## 
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance = 0.05
## 
## Call:
## gvlma::gvlma(x = lm.result.final.m2)
## 
##          Value p-value      Decision
## Global Stat 3.100e+06      0 Assumptions NOT satisfied!
## Skewness    2.073e+04      0 Assumptions NOT satisfied!
## Kurtosis    3.079e+06      0 Assumptions NOT satisfied!
## Link Function 1.574e+02      0 Assumptions NOT satisfied!
## Heteroscedasticity 7.264e+01      0 Assumptions NOT satisfied!
```

정규성, 선형성, 등분산성 모두 불만족

더빈 왓슨 테스트로 독립성 점검

```
##
## Durbin-Watson test
##
## data: lm.result.final.m2
## DW = 1.9804, p-value = 0.3639
## alternative hypothesis: true autocorrelation is greater than 0
```

DW test 결과상 2에 인접함으로 독립성은 만족

Model_3: 상권의 매출과 상권의 독립변수들 간의 다중회귀분석 (상권배후지 변수 소거)

위에서 Model_1 과 2를 확인했을 때 상권배후지(Hinterland)의 데이터는 통계적으로 전혀 유의미한 영향을 주지 않는 것으로 나타난다
그러므로 Model_3에서는 상권(Commercial)의 데이터만 사용해서 분석한다

Linear Model test를 하기 위해 모든 변수들을 Numeric으로 변경

i. 회귀분석과 변수선택 실시

ii. 결과확인

```
##  
## Call:  
## lm(formula = Sales ~ ., data = Total_Commercial)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -1.275e+10 -3.061e+08 -4.184e+07  2.287e+08  3.525e+10  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -3.390e+08  1.868e+08 -1.815  0.06982 .  
## Code        -2.775e+04  8.233e+04 -0.337  0.73616  
## C_N_Aparts -4.649e+05  1.826e+06 -0.255  0.79908  
## C_N_Facs    1.379e+07  2.052e+06  6.719 2.77e-11 ***  
## C_Store     1.054e+07  9.750e+05 10.812 < 2e-16 ***  
## C_Income    1.139e+02  5.534e+01  2.057  0.03987 *  
## C_Spend     2.520e-01  1.626e-01  1.550  0.12142  
## C_F_Pop     2.690e+02  1.063e+03  0.253  0.80028  
## C_S_Pop     -2.385e+05  7.526e+04 -3.169  0.00157 **  
## C_W_Pop     7.333e+05  3.409e+04 21.513 < 2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 1.44e+09 on 1266 degrees of freedom  
## Multiple R-squared:  0.5414, Adjusted R-squared:  0.5382  
## F-statistic: 166.1 on 9 and 1266 DF,  p-value: < 2.2e-16
```

```

## 
## Call:
## lm(formula = Sales ~ Code + C_N_Aparts + C_N_Facs + C_Store +
##     C_Income + C_Spend + C_F_Pop + C_S_Pop + C_W_Pop, data = Total_Commercial)
## 
## Residuals:
##      Min       1Q    Median       3Q      Max 
## -1.275e+10 -3.061e+08 -4.184e+07  2.287e+08  3.525e+10 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -3.390e+08  1.868e+08 -1.815  0.06982 .  
## Code        -2.775e+04  8.233e+04 -0.337  0.73616  
## C_N_Aparts -4.649e+05  1.826e+06 -0.255  0.79908  
## C_N_Facs   1.379e+07  2.052e+06  6.719  2.77e-11 *** 
## C_Store    1.054e+07  9.750e+05 10.812 < 2e-16 *** 
## C_Income   1.139e+02  5.534e+01  2.057  0.03987 *  
## C_Spend    2.520e-01  1.626e-01  1.550  0.12142  
## C_F_Pop    2.690e+02  1.063e+03  0.253  0.80028  
## C_S_Pop    -2.385e+05 7.526e+04 -3.169  0.00157 ** 
## C_W_Pop    7.333e+05  3.409e+04 21.513 < 2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 1.44e+09 on 1266 degrees of freedom 
## Multiple R-squared:  0.5414, Adjusted R-squared:  0.5382 
## F-statistic: 166.1 on 9 and 1266 DF,  p-value: < 2.2e-16

```

```

## 
## Call:
## lm(formula = Sales ~ C_N_Facs + C_Store + C_Income + C_Spend +
##     C_S_Pop + C_W_Pop, data = Total_Commercial)
## 
## Residuals:
##      Min       1Q    Median       3Q      Max 
## -1.277e+10 -3.101e+08 -4.475e+07  2.312e+08  3.524e+10 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -3.513e+08  1.775e+08 -1.979  0.04799 *  
## C_N_Facs   1.377e+07  2.042e+06  6.745  2.32e-11 *** 
## C_Store    1.066e+07  9.095e+05 11.725 < 2e-16 *** 
## C_Income   1.114e+02  5.503e+01  2.024  0.04313 *  
## C_Spend    2.362e-01  1.462e-01  1.615  0.10646  
## C_S_Pop    -2.387e+05 7.468e+04 -3.197  0.00142 ** 
## C_W_Pop    7.340e+05  3.401e+04 21.584 < 2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 1.438e+09 on 1269 degrees of freedom 
## Multiple R-squared:  0.5413, Adjusted R-squared:  0.5392 
## F-statistic: 249.6 on 6 and 1269 DF,  p-value: < 2.2e-16

```

```

## 
## Call:
## lm(formula = Sales ~ C_N_Facs + C_Store + C_Income + C_Spend +
##      C_S_Pop + C_W_Pop, data = Total_Commercial)
## 
## Residuals:
##       Min        1Q     Median        3Q       Max
## -1.277e+10 -3.101e+08 -4.475e+07  2.312e+08  3.524e+10
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -3.513e+08  1.775e+08 -1.979  0.04799 *  
## C_N_Facs     1.377e+07  2.042e+06  6.745  2.32e-11 *** 
## C_Store      1.066e+07  9.095e+05 11.725 < 2e-16 *** 
## C_Income     1.114e+02  5.503e+01  2.024  0.04313 *  
## C_Spend      2.362e-01  1.462e-01  1.615  0.10646    
## C_S_Pop     -2.387e+05  7.468e+04 -3.197  0.00142 **  
## C_W_Pop      7.340e+05  3.401e+04 21.584 < 2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.438e+09 on 1269 degrees of freedom
## Multiple R-squared:  0.5413, Adjusted R-squared:  0.5392 
## F-statistic: 249.6 on 6 and 1269 DF,  p-value: < 2.2e-16

```

어떤 방식을 사용할지를 정하기 위해 Akaike Information Criterion(AIC)으로 확인한다

```
AIC(lm.result.all.m3)
```

```
## [1] 57449.1
```

```
AIC(lm.result.forward.m3)
```

```
## [1] 57449.1
```

```
AIC(lm.result.backward.m3)
```

```
## [1] 57443.34
```

```
AIC(lm.result.stepwise.m3)
```

```
## [1] 57443.34
```

stepwise와 **backward** 방식의 AIC가 57443.34으로 **all**과 **forward**의 57449.1 보다 낮다

summary로 확인한 **backward**와 **stepwise** 모델이 동일한 형태임을 확인하였으므로 **stepwise** 방식을 선택

iii. 모델해석

```

## 
## Call:
## lm(formula = Sales ~ C_N_Facs + C_Store + C_Income + C_Spend +
##      C_S_Pop + C_W_Pop, data = Total_Commercial)
## 
## Residuals:
##       Min        1Q     Median        3Q       Max
## -1.277e+10 -3.101e+08 -4.475e+07  2.312e+08  3.524e+10
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -3.513e+08  1.775e+08 -1.979  0.04799 *  
## C_N_Facs     1.377e+07  2.042e+06  6.745  2.32e-11 *** 
## C_Store      1.066e+07  9.095e+05 11.725 < 2e-16 *** 
## C_Income     1.114e+02  5.503e+01  2.024  0.04313 *  
## C_Spend      2.362e-01  1.462e-01  1.615  0.10646    
## C_S_Pop     -2.387e+05  7.468e+04 -3.197  0.00142 **  
## C_W_Pop      7.340e+05  3.401e+04 21.584 < 2e-16 *** 
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.438e+09 on 1269 degrees of freedom
## Multiple R-squared:  0.5413, Adjusted R-squared:  0.5392 
## F-statistic: 249.6 on 6 and 1269 DF,  p-value: < 2.2e-16

```

C_Spend의 t 검정 결과가 0.1 으로 유의수준 0.05 이상으로 통계적으로 유의하지 않음. C_Spend변수를 제거한 모델을 재구성.

Model_4: Model_3 에서 C_Spend변수를 제거한 모델

```

lm.result.m4 <- lm(Sales ~ C_N_Facs + C_Store + C_Income + C_S_Pop + C_W_Pop, data = Total_Commercial)
summary(lm.result.m4)

```

```

## 
## Call:
## lm(formula = Sales ~ C_N_Facs + C_Store + C_Income + C_S_Pop +
##      C_W_Pop, data = Total_Commercial)
## 
## Residuals:
##       Min        1Q     Median        3Q       Max
## -1.279e+10 -3.102e+08 -4.109e+07  2.326e+08  3.523e+10
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -3.836e+08  1.764e+08 -2.174  0.0299 *  
## C_N_Facs     1.358e+07  2.039e+06  6.657  4.13e-11 *** 
## C_Store      1.089e+07  8.993e+05 12.109 < 2e-16 *** 
## C_Income     1.201e+02  5.480e+01  2.192  0.0285 *  
## C_S_Pop     -1.290e+05  3.100e+04 -4.159  3.41e-05 *** 
## C_W_Pop      7.401e+05  3.382e+04 21.881 < 2e-16 *** 
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.439e+09 on 1270 degrees of freedom
## Multiple R-squared:  0.5404, Adjusted R-squared:  0.5386 
## F-statistic: 298.7 on 5 and 1270 DF,  p-value: < 2.2e-16

```

1단계 : 회귀모형은 타당한가?

F-statistic: 298.7 on 5 and 1270 DF, p-value: < 0.0000000000000022

결론 : 유의확률이 0.000 이므로 유의수준 0.05에서 회귀모형은 통계적으로 타당하다.

2단계 : 독립변수들이 종속변수에게 영향을 주는가?

독립변수 5개 모두가 종속변수에게 유의수준 0.05에서 통계적으로 유의한 영향을 주는 것으로 나타났다.

3단계 : 독립변수들은 어떤 영향을 주는가?

4단계 : 독립변수들의 설명력

Adjusted R-squared: 0.5386

네개의 독립변수가 종속변수를 약 53.86% 설명.

Model_3 보다 Model_4 의 설명력이 높으므로, 현재까지는 Model_4가 가장 유 효해보임.

iii. 모델 검정

다중공선성 점검

```
## C_N_Facs C_Store C_Income C_S_Pop C_W_Pop
## 1.987406 2.277916 1.058892 1.608169 1.186786
```

모든 변수의 vif가 4 미만이므로 다중공선성은 존재하지 않는 것으로 보임

잔차분석

```
##
## Call:
## lm(formula = Sales ~ C_N_Facs + C_Store + C_Income + C_S_Pop +
##     C_W_Pop, data = Total_Commercial)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.279e+10 -3.102e+08 -4.109e+07  2.326e+08  3.523e+10 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -3.836e+08  1.764e+08  -2.174  0.0299 *  
## C_N_Facs     1.358e+07  2.039e+06   6.657 4.13e-11 *** 
## C_Store       1.089e+07  8.993e+05  12.109 < 2e-16 *** 
## C_Income      1.201e+02  5.480e+01   2.192  0.0285 *  
## C_S_Pop      -1.290e+05  3.100e+04  -4.159 3.41e-05 *** 
## C_W_Pop       7.401e+05  3.382e+04  21.881 < 2e-16 *** 
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 1.439e+09 on 1270 degrees of freedom
## Multiple R-squared:  0.5404, Adjusted R-squared:  0.5386 
## F-statistic: 298.7 on 5 and 1270 DF,  p-value: < 2.2e-16 
## 
## 
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance = 0.05
## 
## Call:
## gvlma::gvlma(x = lm.result.m4)
## 
##             Value p-value          Decision
## Global Stat 4.570e+06 0.000e+00 Assumptions NOT satisfied!
## Skewness    2.613e+04 0.000e+00 Assumptions NOT satisfied!
## Kurtosis    4.544e+06 0.000e+00 Assumptions NOT satisfied!
## Link Function 1.514e+02 0.000e+00 Assumptions NOT satisfied!
## Heteroscedasticity 5.795e+01 2.698e-14 Assumptions NOT satisfied!
```

정규성, 선형성, 등분산성 모두 불만족

더빈 왓슨 테스트로 독립성 확인

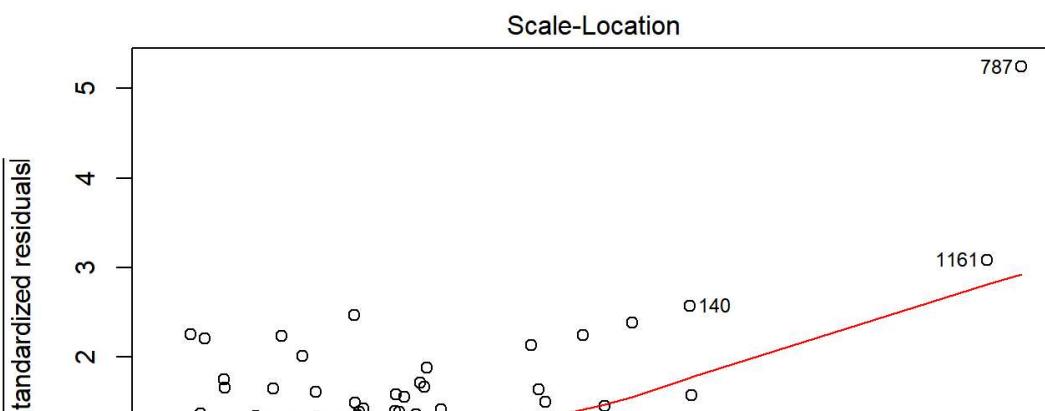
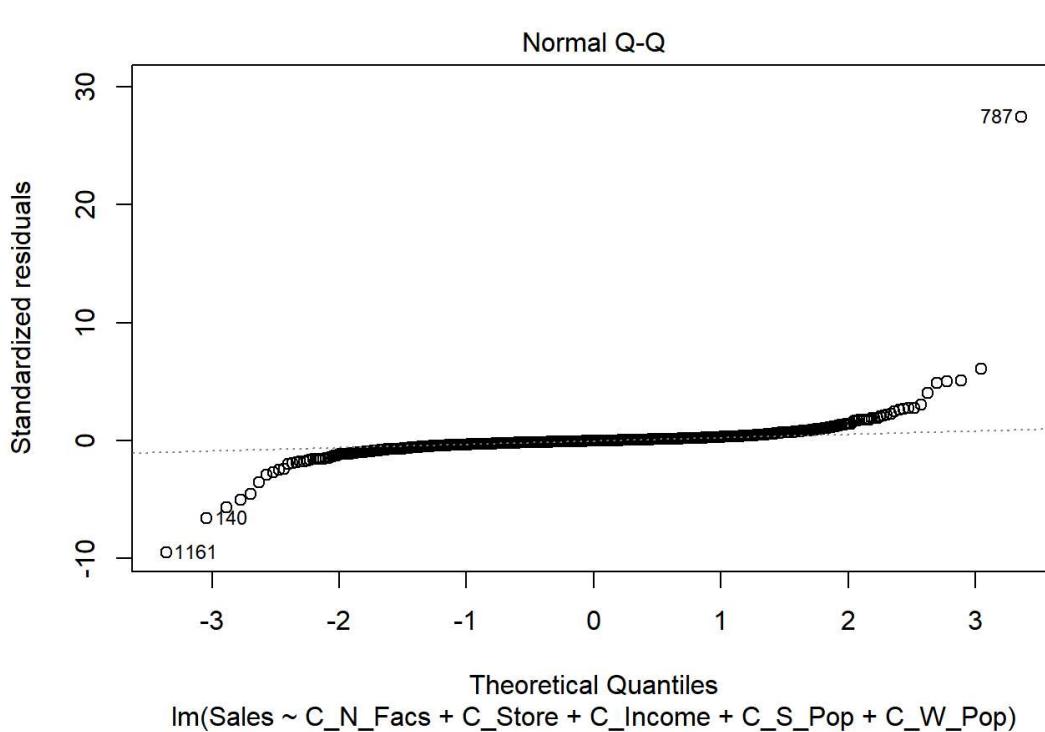
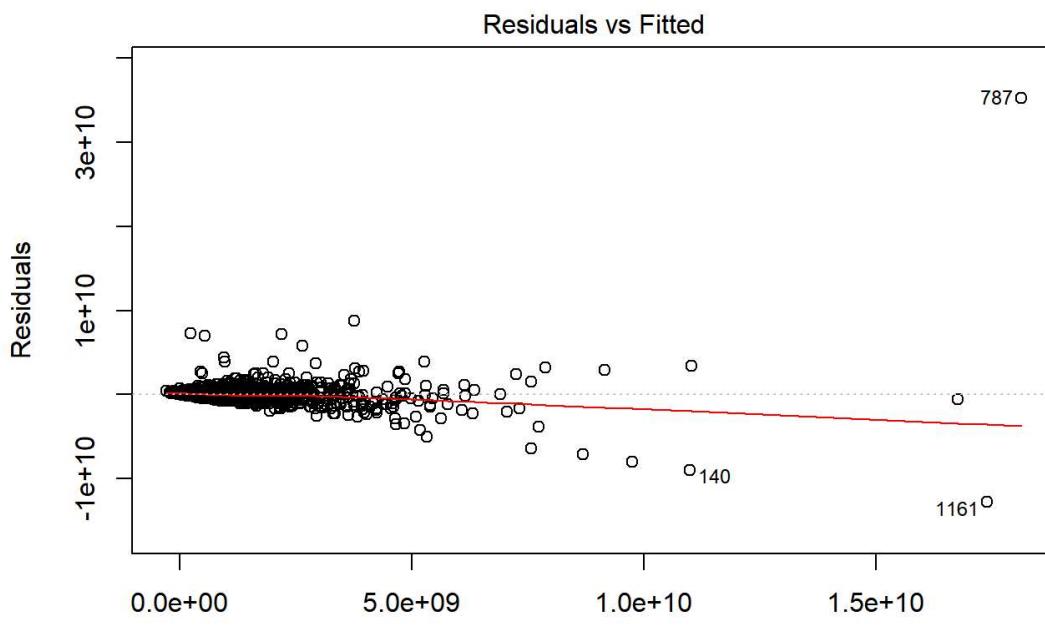
```
##  
## Durbin-Watson test  
##  
## data: lm.result.m4  
## DW = 1.9796, p-value = 0.3473  
## alternative hypothesis: true autocorrelation is greater than 0
```

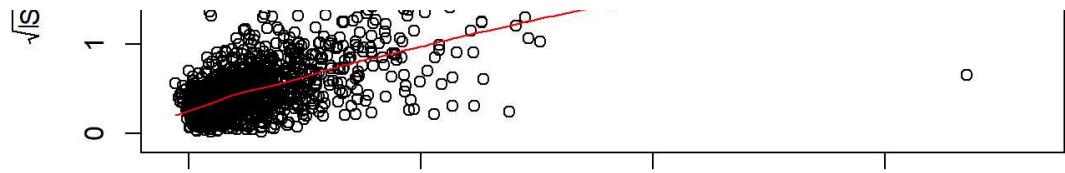
DW test 결과상 2에 인접함으로 독립성은 만족

6. 최종모델의 모델 신뢰성 문제 개선

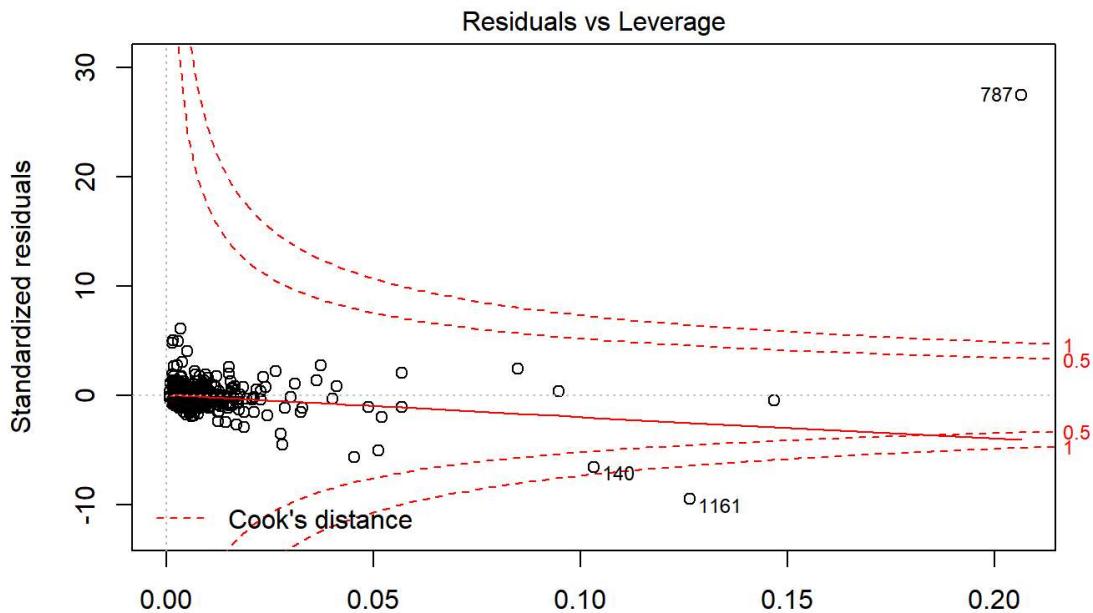
Model_4가 잠정적으로 최종모델로 결론났으나, 모델 추정치의 신뢰도 개선을 위해 각종 시도를 해보기로 하였음.

그림을 그려본다

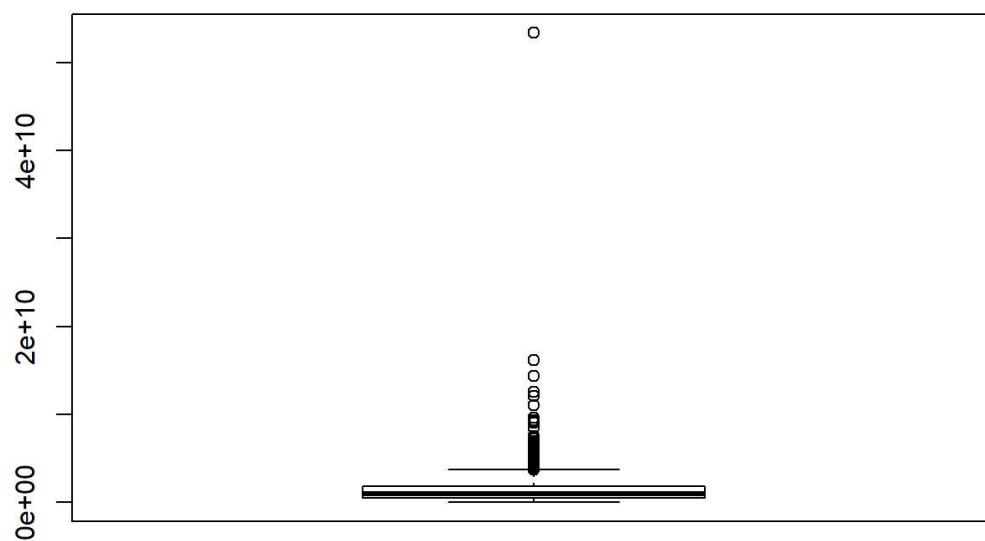




Fitted values
 $\text{lm}(\text{Sales} \sim \text{C_N_Facs} + \text{C_Store} + \text{C_Income} + \text{C_S_Pop} + \text{C_W_Pop})$



Residuals vs Leverage
 $\text{lm}(\text{Sales} \sim \text{C_N_Facs} + \text{C_Store} + \text{C_Income} + \text{C_S_Pop} + \text{C_W_Pop})$



Model_5 이상치 제거

boxplot을 확인하니 많은 양의 이상치 (Outlier)가 Q3이후에 존재함을 알 수 있다

추정치 신뢰성을 개선하기 위한 이상치 제거

```
IQR(Total_Commercial$Sales)
```

```
## [1] 1318641589
```

```
summary(Total_Commercial$Sales)
```

```
##      Min.   1st Qu.    Median     Mean   3rd Qu.   Max. 
## 4.027e+06 4.593e+08 9.701e+08 1.445e+09 1.778e+09 5.336e+10
```

```
# Q3 + 1.5*IQR
1777896903 + (1.5*IQR(Total_Commercial$Sales))
```

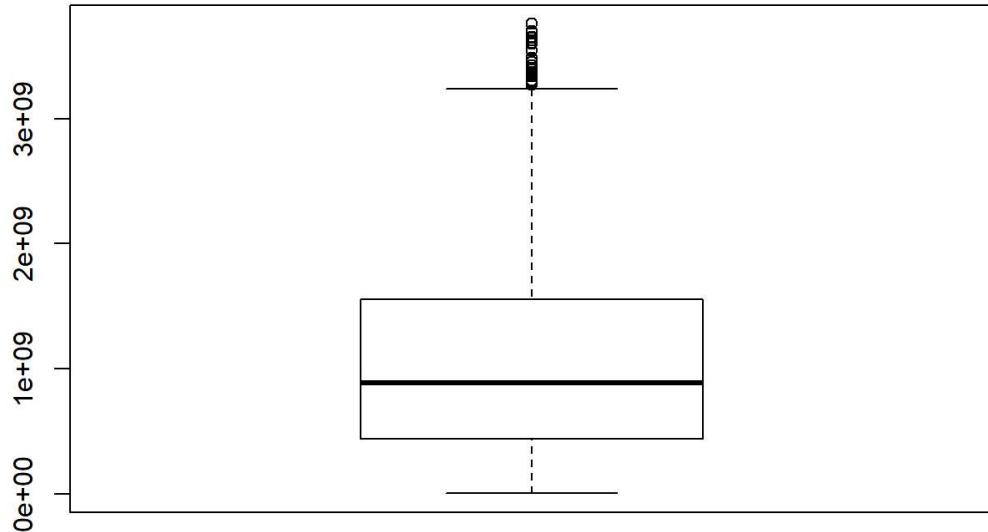
```
## [1] 3755859287
```

```
## [1] 1188
```

```
## [1] 1276
```

```
## [1] 88
```

이상치가 빠진 데이터로 boxplot을 다시 그려본다



이상치가 빠진 데이터로 회귀분석

```

## 
## Call:
## lm(formula = Sales ~ C_N_Facs + C_Store + C_Income + C_S_Pop +
##      C_W_Pop, data = Total_Commercial_N)
## 
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -1.415e+09 -2.597e+08 -9.098e+07  1.841e+08  2.627e+09 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2.278e+08  6.212e+07 -3.667 0.000256 ***
## C_N_Facs     4.233e+06  7.659e+05  5.526 4.02e-08 ***
## C_Store      1.262e+07  3.627e+05 34.800 < 2e-16 ***
## C_Income     1.240e+02  1.928e+01  6.434 1.80e-10 ***
## C_S_Pop      -6.934e+04 1.120e+04 -6.191 8.24e-10 ***
## C_W_Pop      8.559e+04  1.709e+04  5.007 6.38e-07 ***
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 467900000 on 1182 degrees of freedom
## Multiple R-squared:  0.6951, Adjusted R-squared:  0.6938 
## F-statistic: 538.9 on 5 and 1182 DF,  p-value: < 2.2e-16

```

설명력이 69.38% 까지 증가한 것을 알 수 있다. 전반적으로 **Adjusted R-Square**가 상승한 것으로 보아, 이상치 제거가 적합한 선택이었음을 알 수 있다

모델 검정

다중공선성 점검

```

## C_N_Facs  C_Store  C_Income  C_S_Pop  C_W_Pop
## 1.745246 1.994091 1.076511 1.491482 1.089712

```

모든 변수의 vif가 4 미만이므로 다중공선성은 존재하지 않는 것으로 보임.

잔차분석

```

## 
## Call:
## lm(formula = Sales ~ C_N_Facs + C_Store + C_Income + C_S_Pop +
##      C_W_Pop, data = Total_Commercial_N)
## 
## Residuals:
##       Min        1Q     Median        3Q       Max
## -1.415e+09 -2.597e+08 -9.098e+07  1.841e+08  2.627e+09
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2.278e+08  6.212e+07 -3.667 0.000256 ***
## C_N_Facs     4.233e+06  7.659e+05  5.526 4.02e-08 ***
## C_Store      1.262e+07  3.627e+05 34.800 < 2e-16 ***
## C_Income     1.240e+02  1.928e+01  6.434 1.80e-10 ***
## C_S_Pop      -6.934e+04 1.120e+04 -6.191 8.24e-10 ***
## C_W_Pop      8.559e+04  1.709e+04  5.007 6.38e-07 ***
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 467900000 on 1182 degrees of freedom
## Multiple R-squared:  0.6951, Adjusted R-squared:  0.6938
## F-statistic: 538.9 on 5 and 1182 DF, p-value: < 2.2e-16
## 
## 
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance = 0.05
## 
## Call:
## gvlma::gvlma(x = lm.outlier)
## 
##             Value    p-value          Decision
## Global Stat 930.53 0.000e+00 Assumptions NOT satisfied!
## Skewness     260.97 0.000e+00 Assumptions NOT satisfied!
## Kurtosis     600.90 0.000e+00 Assumptions NOT satisfied!
## Link Function 53.34 2.801e-13 Assumptions NOT satisfied!
## Heteroscedasticity 15.31 9.110e-05 Assumptions NOT satisfied!

```

정규성, 선형성, 등분산성 모두 불만족

더 빈 왓슨 테스트로 독립성 확인

```

## 
## Durbin-Watson test
## 
## data: lm.outlier
## DW = 1.9622, p-value = 0.2489
## alternative hypothesis: true autocorrelation is greater than 0

```

DW test 결과상 2에 인접함으로 독립성은 만족

Model_6 로그 변환

잔차분석을 해본 결과, 선형회귀분석의 기본가정이 위배된 것으로 확인, 개선시도를 해보았음

종속변수인 매출을 로그변환한다

```

logSales <- log10(Total_Commercial_N$Sales)
Total_Commercial_N <- data.frame(Total_Commercial_N, logSales)
lm.log <- lm(logSales ~ C_N_Facs + C_Store + C_Income + C_S_Pop + C_W_Pop, data = Total_Commercial_N)
gvlma(lm.log)

```

```

## 
## Call:
## lm(formula = logSales ~ C_N_Facs + C_Store + C_Income + C_S_Pop +
##      C_W_Pop, data = Total_Commercial_N)
## 
## Coefficients:
## (Intercept)    C_N_Facs     C_Store     C_Income     C_S_Pop
## 8.240e+00    1.707e-03   5.829e-03   6.568e-08  -2.840e-05
## C_W_Pop
## 4.539e-05
## 
## 
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance = 0.05
## 
## Call:
## gvlma(x = lm.log)
## 
##          Value p-value      Decision
## Global Stat 1532.738 0.000000 Assumptions NOT satisfied!
## Skewness     336.881 0.000000 Assumptions NOT satisfied!
## Kurtosis     829.714 0.000000 Assumptions NOT satisfied!
## Link Function 358.041 0.000000 Assumptions NOT satisfied!
## Heteroscedasticity 8.102 0.004423 Assumptions NOT satisfied!

```

잔차분석결과가 개선되지 않았으므로 모델을 채택하지 않는다.

최종모델은 Model_5로 정한다

7. 예측

결측치 제거 전 데이터 가지고 예측해보기

우리는 최종적으로 상권의 5개 변수만 가지고 회귀분석을 실시하였다

이 모형의 예측력을 확인해보기 위해 Raw data에서 해당 5개의 변수는 모두 살아있고 다른 변수들에 결측치가 있어서 탈락된 행들을 추출하여 이 행들로 구성된 Dataset에 대하여 Sales를 예측하고, 실값과 대조해 보았다

예측은 최종적으로 선택된 이상치가 제거된 Model_5로 진행한다

예측 모델

$$\begin{aligned} Sales = & -227787016.77 + 4232766.08 * C_{NFacs} + 12620429.96 * C_{Store} + 124.03 * C_{Income} \\ & - 69337.61 * C_{SPop} + 85588.32 * C_{WPop} \end{aligned}$$

predict를 사용하여 예측

```
lm.predict <- predict(lm.outlier, newdata= data.frame(omitTotal_commercial), interval = 'predict')
```

interval = 'predict' : 회귀모형 + 잔차를 고려하여 예측한 구간의 상한 + 하한값 표시

예측률 생성 : 실값이 predict의 상한값 이하 및 하한값 이상이면 예측 성공, 아니면 실패

예측률 출력

```
## 예측률 : 88.888888888889
```

8. 결론

우리는 상권의 활성화 정도의 척도를 상권의 총 매출로 상정하고, 상권의 총 매출에 영향을 주는 설명 변수는 어떤 것이 있는지 분석해 보았다. 상권 및 배후지의 총 15개의 설명 변수를 채택하여 다중선행회귀분석 모형을 구축하였다. 그 중 종속변수에 유의한 영향을 주지 않는 10개의 설명 변수를 소거하고, 나머지 5개의 설명 변수를 선택하여 모델을 재구성하였다.

설명 변수 제거과정에서 우리는 배후지의 각종 변수들이 상권의 총 매출에 유의한 영향을 미치지 않는다는 점을 알게 되었다. 5개의 설명 변수는 상권의 점포 수, 상권 인구의 소득, 상권의 집객 시설, 상권의 주거 및 직업 인구로 직관적으로 봤을 때 그 결과는 매우 타당해 보였다. 다만, 배후지의 데이터에서도 특정 변수가 분명히 유의한 결과를 나타낼 것으로 생각한 초기의 생각은 통계적으로 정확하지 않았음을 알게 되었다.

분석 결과 중 안타까운 부분은 선형회귀분석의 가정이 대부분 깨졌다는 것이다. 이는 모형 추정치의 신뢰성이 확보되지 않을 수 있음을 의미한다. 현재 우리가 시행한 변수의 log변환, 이상치 제거 말고도 설명변수의 다항식 포함, WLS(Weighted Least Squares) 혹은 FGLS(Feasible Generalized Least Squares)등의 방법을 활용할 수 있지만, 현 프로젝트의 시간 및 이해도 관계상 유의미한 회귀계수를 얻는 것에 만족하였다.

결론적으로, 우리의 분석결과에 따르면 서울특별시의 골목상권의 매출은 상권의 점포 수, 상권 인구의 소득, 상권의 집객 시설, 상권의 주거 및 직업 인구에 의해 약 69.4% 정도 설명되는 것을 알 수 있었다. 그 중 상권의 상주인구는 유일하게 음의 상관관계를 보이는 설명 변수 였다. 회귀 계수를 표준화 하였을 때 가장 큰 영향력을 가지는 설명 변수는 예상대로 점포의 수였으나, 인구의 소득 혹은 직장 인구 보다도 집객 시설의 개수가 더 영향력이 크다는 점은 새로운 발견이었다. 마지막으로, predict 함수를 통한 예측모형은 오차범위 내 88.89% 의 적중률을 보였다.