

Regression Diagnostics

Kyubong Theodore Lee

적합으로 회귀분석은 끝나지 않는다.

- 모형: 현재의 모형이 반응변수와 설명변수 간 관계를 제대로 반영하는지 조사해야한다. 설명변수들 중에서도 어느 변수가 중요한지 또는 불필요한 것들이 없는지 살펴보아야 한다.
- 자료: 잔차분석을 통해 오차항의 등분산성, 독립성, 정규성이 조사되어야한다. 특이값, 이상치의 확인 또한 필요하다.

부분 F검정

- 두 개 이상의 회귀계수들에 대한 개별적인 t검정을 동시에 하는 것은, 가설검정의 개념에서 부적절. 이에 대한 대안으로 개개의 계수에 대한 검정 대신 두 개의 가능한 모형을 비교(full vs reduced) 이 경우 F 분포를 이용한 검정을 하고, 다중검정에서의 유의수준 문제가 발생하지 않는다.
- 설명변수가 (p-1)개 있는 회귀모형에 대한 분산분석표의 F 검정에서 귀무가설 $H_0: \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$ 을 기각하는 경우, 모든 β_j 가 0이 아니라는 것을 의미하지는 않는다. 일부는 0이 되어 해당 설명변수는 반응변수와 선형관계가 없음을 나타낼 수 있다.
- (p - 1)개 설명변수로 이루어진 회귀모형 vs (p - r)개의 설명변수로 이루어진 회귀모형 비교: 반응변수의 변동을 많이 설명해주는 설명변수를 모형에 포함시키면 SSE는 크게 감소, 기존의 모형에서 제거하면 크게 증가
- $H_0: \beta_r = \beta_{r+1} = \dots = \beta_{p-1} = 0$ 이 옳다면, 이 설명변수들은 반응변수와 선형관계가 없다는 것을 의미하므로, Full model(p-1개 설명변수 가진 모형)에서 이 변수들을 제거한다하더라도 SSE는 크게 증가하지 않는다. 역으로 이들 설명변수를 회귀모형에서 제거했을 때, SSE의 크기가 많이 변하지 않으면 귀무가설이 옳은 것이고, 많이 증가하면(SSE가 많이 증가했다는 뜻은 제외시켰던 변수들이 반응변수의 변동을 잘 설명해주는 유효한 설명변수라는 뜻) 귀무가설이 옳지 않다고 볼 수 있다. 그러므로 SSE 또는 SSR의 값의 변화 정도는 위의 가설에 대한 검정에서 중요한 역할을 한다.(SSR의 경우 상수항이 없는 모형에서는 접근 방식이 달라짐)
- 따라서 $SSE(Reduced) - SSE(Full)$ 의 크기를 조사하면 된다. 그리고 이를 $F_0 = \frac{SSE(R) - SSE(F)}{df_R - df_F} / \frac{SSE(F)}{df_F}$ 검정통계량 산출에 활용. 이 검정을 부분 F 검정(partial F test)
- '추가제곱합'은 완전모형에서 설명변수 X_j 가 제외될 때 발생하는 제곱합의 차이로 해석할 수 있으며, 이를 X_j 의 편제곱합(partial sum of squares)이라 부른다. 많은 경우 편제곱합의 크기로 설명변수의 중요도를 서열화하기도 한다.
- '순차제곱합(sequential sum of squares)'은 영모형부터 설명변수들을 하나씩 추가시켜 나갈 때 증가하는 회귀제곱합의 크기. 순차제곱합들은 서로 독립이고, 각각 1의 자유도를 갖는다. 순차제곱합은 설명변수의 개수가 많은 경우 모형선택에 유용하게 사용됨. 반응변수에 대한 설명력이 높은 순서대로 설명변수들을 하나씩 모형에 포함시켜 나가면서 매 단계에서 부분 F 검정을 실시해, 추가제곱합의 증가량이 통계적으로 유의할때까지만 포함시켜 적절한 회귀모형을 찾을 수 있다.(step, direction = 'forward')
- R에서 순차제곱합과 편제곱합은 lm 객체에 anova()와 drop1()을 적용해 구한다.

```
library(regbook)
```

```
## Loading required package: MASS
```

```
## Loading required package: leaps
```

```
lm.fit <- lm(price ~ year + mileage + cc + automatic, usedcars)
anova(lm.fit)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: price
```

```
##          Df Sum Sq Mean Sq F value    Pr(>F)
## year      1 2056608 2056608 201.2036 1.841e-13 ***
```

```
## mileage      1  135864  135864  13.2919 0.0012228 **
## cc           1   52409   52409   5.1273 0.0324794 *
## automatic    1  175828  175828  17.2018 0.0003389 ***
## Residuals    25  255538   10222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

• 순차제곱합 $SSR(F) = 2056608 + 135864 + 52409 + 175828 = 2420709$

```
drop1(lm.fit, test = 'F')
```

```
## Single term deletions
##
## Model:
## price ~ year + mileage + cc + automatic
##           Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                 255538 281.50
## year          1    398929 654467 307.71 39.0283 1.552e-06 ***
## mileage       1    100649 356187 289.46  9.8467 0.0043244 **
## cc            1     37794 293332 283.64  3.6975 0.0659577 .
## automatic     1    175828 431366 295.20 17.2018 0.0003389 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 위 결과는 편제곱합으로 설명변수 4개가 있는 full model에서 해당 설명변수를 제외할 때 감소하는 SSR의 크기 or 증가하는 SSE의 크기이다. 예를들어, cc를 제외한 경우 편제곱합은 37794
- 위의 경우 year의 편제곱합이 가장 크므로 price 설명에 제일 중요하다고 볼 수 있다.
- cc의 편제곱합 37794 검정통계량 $F_0 = 3.697$ 이 값은 F 분포표의 값 $F(0.05; 1, 25) = 4.242$ 보다 작으므로(p값이 0.05보다 크므로), 귀무가설 $H_0: \beta_3 = 0$ 을 기각하지 못한다. 각 검정 결과는 출력결과 오른쪽에 있다.
- 참고로 $H_0: \beta_3 = 0$ 에 대한 부분 F 검정통계량값 3.697은 완전모형에서의 β_3 에 대한 t값 1.9229의 제곱과 일치. 따라서 하나의 회귀계수에 대한 t 검정과 부분 F 검정이 일치함을 확인 가능

```
summary(lm.fit)
```

```
##
## Call:
## lm(formula = price ~ year + mileage + cc + automatic, data = usedcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -177.35  -63.91   -0.99   70.34  212.69
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.253e+02  3.998e+02   1.314 0.200823
## year        -5.800e+00  9.283e-01  -6.247 1.55e-06 ***
## mileage     -2.263e-03  7.211e-04  -3.138 0.004324 **
## cc           3.888e-01  2.022e-01   1.923 0.065958 .
## automatic    1.653e+02  3.986e+01   4.147 0.000339 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101.1 on 25 degrees of freedom
## Multiple R-squared:  0.9045, Adjusted R-squared:  0.8892
## F-statistic: 59.21 on 4 and 25 DF,  p-value: 2.184e-12
```

- anova()로도 가설 $H_0: \beta_3 = 0$ 검정 가능하다. 이는 full 모형과 이 두 변수가 빠진 모형을 비교하는 문제가

된다.

```
mod_1 <- lm(price ~ year + mileage, usedcars)
mod_2 <- lm(price ~ year + mileage + cc + automatic, usedcars)
anova(mod_1, mod_2)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ year + mileage
## Model 2: price ~ year + mileage + cc + automatic
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      27 483775
## 2      25 255538  2    228237 11.165 0.0003429 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- F 통계량 값이 11.165이며, p 값이 0.0003이므로 귀무가설을 기각한다.
- $SSE(R) - SSE(F) = 483775 - 255538 = 228237$, 검정통계량은 $(228237 / 2) / 10222 = 11.165$