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 $H_0 : {}_1 = {}_2 = ... = {}_{p-1} = 0$ 을 기각하는 경우, 모든 beta j가 0이 아니라는 것을 의지하지는 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
- $H_0 : {}_r = {}_{r+1} = ... = {}_{p-1} = 0$ 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 변화정도는 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 $F_0 = SSE(R) - SSE(F)/df_R - df_F / SSE(F)/df_F$ 검정통계량 산출에 활용. 이 검정을 **부분 F 검정(partial F test)**
- '추가제곱합'은 완전모형에서 설명변수 X_j가 제외될 때 발생하는 제곱합의 차이로 해석할 수 있으며, 이를 X_j의 편제곱합(partial sum of squares)이라 부른다. 많은 경우 편제곱합의 크기로 설명변수의 중요도를 서열화하기도 한다.
- '순차제곱합(sequential sum of sqaures)는 영모델부터 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 $H_0 : _3 = 0$을 기각하지 못한다. 각 검정 결과는 출력결과 오른편에 있다.
- 참고로 $H_0 : _3 = 0$에 대한 부분 F 검정통계량값 3.697은 완전모형에서의 beta3 hat에 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 $H_0 : _3 = _4 = 0$ 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 ***** ### 편회귀, 편상관계수
- 중회귀모형에서 회귀계수 $_j$는 다른 설명변수들의 값이 고정되었을 때 설명변수 X_j의 값이 한 단위 증가함에 따른 반응변수의 '평균변화량'으로 해석. 그리고 최소제곱추정값 $\hat{}_j$은 그 평균변화량의 추정값이다.
- 그렇지만 일반적으로 중회귀모형에서는 반응변수 Y와 특정한 설명변수 X_j와의 관계가, 다른 설명변수들에 의해 영향을 받기 때문에 해석에 조심할 필요가 있음.
- 중고차 가격이 '연식(X1)'과 '주행거리(X2)'로 산정된다고 가정.
- 1차적으로 X1만으로 모형을 구축하면, $Y = \hat{Y}(X) + e(Y|X_1)$. $e(Y|X_1)$은 'X1에 의해 설명되지 않은 부분'이므로, X2가 추가된 풀 모형은 이 $e(Y|X_1)$을 추가적으로 설명하는 것.
- X1과 X2가 어느 정도의 선형관계를 가지므로, X1에 대한 정보로 X2의 값을 어느 정도 예측할 수 있는데, 이는 X1만 포함된 회귀모형도 X2가 갖는 정보를 이미 어느 정도 포함하고 있다는 것을 의미. 그러므로 X2를 추가하는 경우, 모형의 설명에는 X1에 의해 알려지지 않는 X2의 정보만 추가로 사용된다고 할 수 있음. 즉, $X_2 = \hat{X}_2(X_1) + e(X_2|X_1)$. 'X1에 의해 이미 주어진 X2의 정보' + 'X1에 의해 주어지지 않은 X2의 정보' 즉, X_1이 이미 포함된 모형에 X_2를 추가할 때는, X1에 의해 설명되지 않은 부분인 $e(Y|X_1)$를 X2가 추가적으로 제공하는 정보인 $e(X_2|X_1)$으로 설명을 하게 된다. 즉, Y와 X2 각각에서 X1의 영향력을 제거한 나머지 부분들로 분석이 진행. 잔차 $e(Y|X_1)$와 $e(X_2|X_1)$를 이용하면 X1의 영향력이 제거된 순수한 Y와 X2만의 관계를 파악할 수 있는데, 이 두 잔차들의 상관계수값을 Y와 X2의 $sample partial correlation coefficient$라고 한다. 회귀분석에서는 $e(Y|X_1)$와 $e(X_2|X_1)$ 값들의 산점도를 X2의 편회귀그림 partial regression plot
- 편회귀 그림의 극단적 경우로 기울기가 수직에 가까우면 다중공선성을 의심해야 한다. ### 변수변환
- 반응변수와 설명변수드르이 관계가 이론적 근거나 산점도에 의해 비선형함수 관계를 가져도, 변수변환을 하면 선형함수 관계가 성립해 선형모형을 가정할 수 있다. 또한 분포의 정규성이 만족되지 않는 경우도 변수변환을 이용해 문제점을 해결할 수 있다. ### 표준화된 회귀계수
- 회귀계수의 의미는 항상 x와 y의 단위와 함께 해석되어야한다. Y = X1 + 100 * X2에서, X2가 100배 영향력이 크다고 할 수 없다. (x1이 m, x2가 cm 단위라면?)
- R의 lm은 비표준화계수를 보여주고 있다. 절대로 이 출력값들이 상대적중요도를 나타내는 것이 아니다. ### 다중공선성의 탐색
- n x p 디자인행렬 X의 rank는 p라고 가정하고 있다. 행렬 X의 계수가 열이나 행의 개수와 같을 때 행렬 X를 Full rank를 갖는다고 하는데, 이 경우에만 $X'X$도 최대계수를 가지고 따라서 역행렬이 존재한다.
- 허나 설명변수들 사이에 선형종속 관계가 존재하면 행렬 X의 계수가 p보다 작게된다. 이 경우 $X'X$의 역행렬이 존재하지 않아 정규방정식의 유일해를 구할 수 없게 된다.(무수히 많은 해) 선형종속이 아니라 상관관계가 높은 경우에도 회귀분석의 추론과정에 문제가 생긴다. 두개의 설명변수가 있는 모형에서 각 변수가 표준화되었을 때 디자인행렬을 W라 하면, 이 때 행렬식의 값이 0에 가까워져 회귀계수 추정량의 분산이 매우 커지게되어 추정의 정확도가 낮아진다.
- 보통 상관계수가 0.9 이상이거나, VIF(1 / 1 - R_k^2(= k번째 설명변수를 반응변수로 하여 나머지 설명변수들에 대해 회귀분석을 했을 때 얻어지는 R-square, 선형관계가 전혀 없으면 0이되어 VIF는 1에 가까워진다. 허나 밀접한 관계에 있어 R^2값이 올라가면 VIF는 커진다))가 10 이상일 경우 다중공선성 가능성을 판단. Variance inflation factor는 다중공선성의 존재 유무만 알려줄 뿐, 설명변수 간 존재하는 선형종속관계의 개수나 형태에 대해서는 알려주지 않는다. VIF값은 표본상관계수행렬의 역행렬의 대각원소
- 다중공선성으로 특정 변수를 제거 후 lm을 돌리면 설명력은 낮아지지 않으나, 계수들의 추정값들이 크게 바뀐다.

표준오차는 작아져 유의도는 높아진다. 전문가와의 상의를 통한 제거 + Lasso or Elastic Lasso