Finance Analytics
Chapter 6. Multicolinearity, Variable selection and Linear transformation

권태연

한국외대 국제금융학과

Multicolinearity (다중공선성)

선형회귀모형의 가정

- 가정 1. 변수 Y와 X의 관계는 선형(Linear)이다. ✓
- 가정 2. X는 확률변수가 아닌 주어진 상수값이다. ✓
- 가정 3. X값이 주어져 있을 때, 오차항의 평균은 0이다. ✓
- 가정 4. X값이 주어져 있을 때, 오차항의 분산은 σ^2로 모든 개체 i에 대해 동일하다 ✓
- (가정 5. 서로 다른 개체간 오차항들은 상관되어 있지 않다. = 자기상관(autocorrelation)이 없다. → 6장, 시계열모형)
- 가정 6. X변수들이 여러개 있을때, X변수들 사이에는 선형관계가 없다. = 다중공선성(Multicollinearity) 문제가 없음을 가정
- 가정 7. 모형 설정 오류가 없음
- 가정 8. 오차항은 정규분포를 따름을 가정 ✓

다중공선성 문제

- 설명변수들 간에 선형,상관관계 존재 할 경우, 다중공선성 (multicollinearity) 문제 발생
- 실제 경제, 재무 현상을 분석할 경우 완전한 다중공선성 (perfect-collinearity)은 발생하기 어렵지만 독립변수들 간에는 어느정도의 상관관계가 존재할 수 밖에 없기 때문에 다중공선성은 어느 정도 존재 (imperfect-collinearity, near-collinearity)하게 된다.
- 즉, 다중공선성의 문제는 있느냐 없느냐의 문제가 아닌, 어느정도 심각하게 존재하느냐의 문제임

완전한 다중공선성

- 설명변수들간의 완전한 다중공선성 (perfect-colllinearity), 즉 $cor(X_1, X_2) = 1$ or -1

```
> x1<-rnorm(100,0,1)
> x2<-2*x1
> error<-rnorm(100,0,0.1)
> y<-0.5*x1+0.7*x2 + error
> cor(x1,x2)
[1] 1
>
> model<-lm(y~x1+x2)
> summary(model)
```

완전한 다중공선성

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
Min        1Q    Median        3Q       Max
-0.27421 -0.06032 -0.00108  0.05478  0.21094

Coefficients: (1 not defined because of singularities)
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.011023   0.009855  -1.119    0.266
x1           1.902176   0.009949 191.191   <2e-16 ***
x2                 NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09849 on 98 degrees of freedom
Multiple R-squared:  0.9973,    Adjusted R-squared:  0.9973
F-statistic: 3.655e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

완전한 다중공선성

- 위와 같이 추정 자체가 불가능
- 실제에서 발생할 가능성 거의 없다.
- 단, 연구자가 임의로 만든 변수를 독립변수로 사용할 때는 발생할 수 있으므로 주의하여야 한다.
- (ex) 더미변수, 선형변환된 변수의 추가적 사용
- (ex) 임금예제에서 이와 같은 상황 발생함. (exper, edu, age)

가상 데이터: 그릴리치 자료

그릴리치 (Griliches et al. 1962) 자료의 일부를 이용하여 유동성 자산을 포함한 소비함수 모형을 최소제곱방법을 이용하여 추정하여 보고 다중공선성 여부를 살펴보자. Table4_2 자료는 가상의 자료로 소비지출, 가처분소득, 그리고 유동성 자산에 대한 자료이다.

$lm(expend\ \ income + wealth, data = data2)$

다중공선성의 탐지 방법 1.

- 독립변수의 갯수가 많을 때 : 의심해봐야 함!
- 높은 독립변수 간 상관계수
- 높은 결정계수 (R^2) 그러나 대부분의 유의적이지 않은 t 값

✓ 위의 회귀모형에서 종속변수와 개별 변수간의 상관계수값들과 비교하였을 때, 높은 결정계수가 관찰되었으며

✓ 독립변수간 상관계수가 높은 변수 쌍이 있음을 확인하시오.

다중공선성의 탐지 방법 1.

```
> cor(data2)
           expend    income    wealth
expend  1.0000000 0.9808474 0.9780997
income  0.9808474 1.0000000 0.9989624
wealth  0.9780997 0.9989624 1.0000000
```

다중공선성의 탐지 방법 2.

- 분산 팽창인자 (VIF: Variance Inflation Factor)

$$VIF(b_k) = \frac{1}{1 - R_k^2}$$

  이때, R_k^2는 k번째 독립변수를 종속변수로 나머지 독립변수를 독립변수로 하는 회귀모형의 결정계수이다.

- VIF값이 크면? 다중공선성이 존재할 가능성 매우 높음 !!
- 일반적으로 기준 10.

다중공선성의 탐지 방법 2.

```
install.packages("HH")
> library(HH)
> vif(data2[,2:3])
income   wealth
482.1275 482.1275
```

다중 공선성의 해결방안 - 1. 쉬운 방법

1. 가장 쉬운 방법은 상관계수가 높은 독립변수 중 하나 혹은 일부를 회귀모형에서 제거
   - 제거 할 때 모형 설정 오류에 유의
   - 즉 중요한 변수 (경제이론, 즉 해석상)는 제거하지 않도록 한다.
   - R^2는 감소 함
2. 때로는 로그변환 차분과 같은 변수 변환을 통해 다중공선성을 완화시킬 수 있다.
3. 다중 공선성이 높은 변수 대신 유사한 의미를 갖는 다른 변수로 대체할 수 있다.

다중 공선성의 해결방안 - 어려운 방법

1. 주성분 분석 (Principal component analysis) : 서로 연관되어 있는 독립변수들을 그룹으로 만드러 소그룹으로 만든다. 이 소그룹들은 서로 상관되지 않은 가상 변수가 된다.

- 독립변수, 회귀계수의 의미부여가 어려울 수 있다.

2. 능형회귀 (Ridge Regression) - by Hoerl and Jennard (1970)
   - OLS가 아닌 다른 추정방법을 사용 (*lm.ridge*() 함수)
   - 그러나 자체로 다중공선성을 해결은 못함
   - R^2는 매우 높지만 모든회귀계수가 유의미하지 않은 문제를 회귀계수의 기여도 조정을 통해 해결해줌
   - 변수선택의 가이드를 줄 수는 있음.

생각해 보기: 다중 공선성은 언제나 문제가 되는가?

Model Selection and Regulation

모형의 적합도와 변수 선택(variable selection)

- 변수 선택 방법 : Forward, Backward, Stepwise selection
  - → 모형의 적합도에 근거하여 설명변수들의 부분집합 선택
- 모형의 적합도 평가 기준:
  - R^2
  - adj-R^2
  - AIC (Akaike Information Criterion)
  - BIC or SBS (Bayesian Information Criterion)
  - SIC (Schwartz information criterion)

  ✓ R^2, adj-R^2는 클수록, AIC, BIC, SIC는 작을 수록 좋음.

  ✓ R^2, adj-R^2는 선형 회귀 모형 에서만

  ✓ AIC, BIC, SIC는 다른 종류의 모형에서도 모두 사용가능

모형조정 (몰라도 됩니다.)

- ML technique (회귀계수 수축(shrink) 방법): Lasso regression, Elastic-Net(Lasso와 Ridge의 조합)
  - Lasso : 변수선택 (회귀계수들을 0으로 보냄)
  - Ridge: 변수선택은 하지 못함. 변수들의 기여도를 조정하여 다중공선성 문제 (일부- 이상한 회귀계수 부호, 유의미하지 않은데 R-square는 큼.) 해결.
  - EN : Lasso와 Ridge의 조합.
  - (주의) Ridge, Lasso, Elastic-Net, LARS regression을 위해서는 독립변수들 표준화 필요함.
- LARS (Least-angle regression) Regression

Ridge, Lasso regression, Elastic-Net (몰라도 됩니다.)

$$
\begin{aligned}
\hat{\beta}^{ols} &= argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \\
\hat{\beta}^{lasso} &= argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 - \lambda \sum_{j=1}^{p} |\beta_j| \\
\hat{\beta}^{ridge} &= argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 - \lambda \sum_{j=1}^{p} \beta_j^2 \\
\hat{\beta}^{elastic-net} &= argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 - \lambda \sum_{j=1}^{p} [\alpha |\beta_j| + (1-\alpha)\beta_j^2]
\end{aligned}
$$

변수의 선형변환

단위 변환과 회귀계수

로그-로그 변환 없는 다음의 생산함수를 고려하자.

$$Q_i = B_1 + B_2 L_i + B_3 K_i + u_i$$

- 자료 Table2_1.csv에서 생산량 (Q)의 단위는 1000달러, 노동시간 (L)의 단위는 1000시간, 자본투입량 (K)의 단위는 1000달러이다.
- 측정 단위가 위와 같을때, 회귀모형 적합후 추정된 b_1, b_2, b_3와
- 생산량의 측정 단위가 10000달러로 변경된 후, 회귀모형 적합 후 추정된 b_1^*, b_2^*, b_3^*의
  1. 회귀계수 추정치의 차이?
  2. 각 회귀계수의 유의성 차이?
  3. 추정된 회귀계수의 해석의 차이?

표준변수와 회귀계수

표준변수(표준화된 변수) 만들기

$$Y^* = \frac{Y - \bar{Y}}{S_Y}$$
$$X^* = \frac{X - \bar{X}}{S_X}$$

$$Y_i^* = B_1^* + B_2^* X_i^* + u_i^*$$

- 표준변수(standardized variable): 단위의 개념 없음, 평균이 0이고 분산이 1
- B_1^*, B_2^*를 표준계수라 한다.
- B_2^*의 해석:
- 표준계수를 고려해야 하는 경우:

표준변수와 회귀계수

로그-로그 변환 없는 다음의 생산함수를 고려하자.

$$Q_i = B_1 + B_2 L_i + B_3 K_i + u_i$$

- 모두 변수를 표준변수로 변환한 후 회귀모형을 적합하여 보자.
  1. 회귀계수 추정치의 차이?
  2. 각 회귀계수의 유의성 차이?
  3. 추정된 회귀계수의 해석의 차이?