

고급회귀분석론

Ch1. Introduction

양성준

회귀분석 (Regression analysis)

- ▶ 변수들간의 관계를 모형화하고 조사하는 통계적 기법
- ▶ 가장 널리 활용되는 통계 기법 중 하나
- ▶ 매우 넓은 영역에 활용됨
- ▶ ‘회귀’라는 단어는 키가 큰 아버지의 아들 키가 세대를 거치면서 모집단의 평균키로 회귀한다는 발견에서 유래됨

함수적 관계 vs 통계적 관계

- ▶ 함수적 관계 : 두 변수의 관측치가 수식으로 표현되는 그래프 위에 정확히 나타남

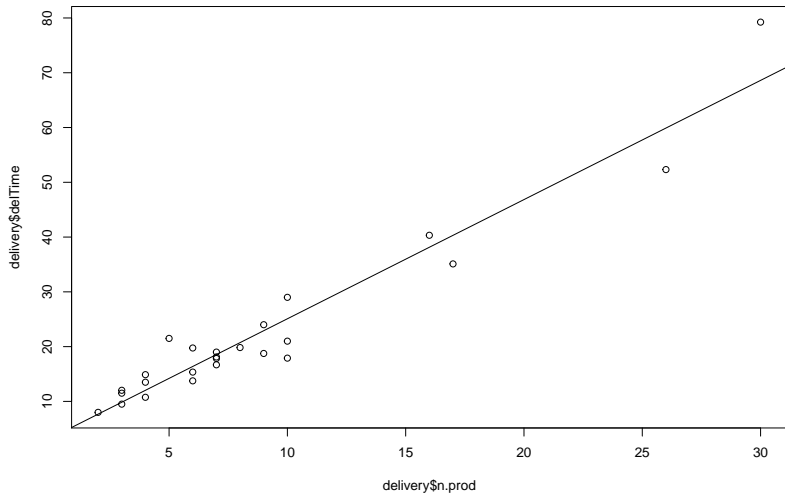
$$(\text{지출}) = (\text{수입}) * 0.9$$

- ▶ 통계적 관계 : 두 변수의 관측치가 수식으로 표현되는 그래프 위에 정확히 나타나지는 않음

$$(\text{지출}) = (\text{수입}) * 0.9 + (\text{오차})$$

- ▶ 오차의 존재 하에서 변수 사이의 관계를 나타내는 수식을 찾아내는 것이 회귀분석의 일차적인 목표 중 하나

Example



선형회귀모형

$$y = \beta_0 + \beta_1 x + \epsilon$$

- ▶ x : 독립변수(independent variable), 예측변수(predictor)
- ▶ y : 종속변수(dependent variable), 반응변수(response variable)
- ▶ ϵ : 오차변수(error variable)
- ▶ 오차항은 보통 $E(\epsilon) = 0$ 혹은 $E(\epsilon|x) = 0$ 을 가정한다. 추가적으로 대칭인 분포가정을 하기도 한다.

선형회귀모형

- ▶ 변수관계를 나타내는 회귀식 $\beta_0 + \beta_1 x$ 은 사실 주어진 x 에서의 y 의 평균을 나타낸다.

$$E(y|x) = \beta_0 + \beta_1 x$$

- ▶ 주어진 x 에서의 y 의 변동성은 오차항의 분산으로 표현된다.

$$Var(y|x) = \sigma^2 = Var(\epsilon)$$

회귀모형

- ▶ 회귀모형은 변수들간의 정확한 관계를 나타낸다고보다는 실제 함수관계를 근사하기 위한 것으로 이해하는 것이 좋다.
- ▶ 회귀모형은 경험적 모형(empirical model)로 간주될 수 있다.
- ▶ 회귀모형은 변수들간의 원인-결과 혹은 원인-효과 관계를 증명해주지는 않음을 유의해야 한다.

데이터 수집

- ▶ 후향적연구(retrospective study) : 이미 알고 있는 결과에 대해서 과거를 추적하는 방식으로 수집
 - 데이터 수집이 상대적으로 용이하고 비용이 적게 든다
 - 분석에 한계가 있을 수 있으며 연구결과의 신뢰도가 상대적으로 떨어진다
- ▶ 관측연구(observational study) : 발생하는 일이나 과정을 관측하여 데이터를 수집
 - 대량의 데이터를 수집가능
 - 의미있는 정보의 추출을 위한 노력이 필요
- ▶ 실험계획(designed experiment) : 특정 목적을 위하여 잘 계획된 실험을 통하여 데이터를 수집
 - 목적 달성을 위해 가장 효율적인 방식
 - 특정 목적을 벗어나는 분석은 불가할 수 있으며, 실험계획에 많은 노력을 기울여야 함