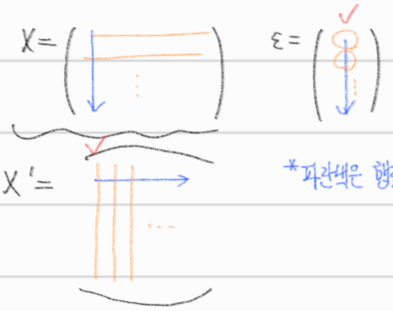
QnA.

1. A3Rsru vs. A3Rmi

이 위치에는 항상 x의 첫번째 행과 ε의 첫번째 원소가 곱해진다.

A3Rsru : $E(X'\varepsilon)=0 \rightarrow X=$



$\varepsilon=$ → $X'\varepsilon = \begin{pmatrix} \vdots \\ x_{1i}\varepsilon_1 + x_{2i}\varepsilon_2 + \cdots + x_{ni}\varepsilon_n \\ \\ \end{pmatrix}$

*파란색은 행렬곱 방향을 의미.

$$= \begin{pmatrix} \sum_j^n x_{j1}\varepsilon_j \\ \vdots \\ \sum_j^n x_{ji}\varepsilon_j \end{pmatrix}$$: Data 상에서 같은 행의 곱의 합

$$= \varepsilon_1 x_{\cdot 1} + \varepsilon_2 x_{\cdot 2} + \cdots + \varepsilon_n \underbrace{x_{\cdot n}}_{\text{열 벡터.}}$$

(ε_i들이 가중치 역할)

- $E(X'\varepsilon)=0$은 Data 상에서 같은 행의 x와 ε의 곱의 합이 0이 나와야 하며 이는 x와 ε간에 covariance 혹은 correlation이 없음을 의미하기도 함. 대신에 Data 수가 커지고 $\frac{1}{n}$로 나눠줘야 함.
- A3Rmi는 어떤 Data(X)가 와도 ε은 평균적으로 $E(\varepsilon|X)=0$이 되는 것을 의미하기에 소표본에서도 x와 ε간 관계가 없음을 생각하면 된다.
  A3Rsru는 대표본에서 상관이 없다고 받아들일 수 있다.

2. Non indep 부분. → P.11 $Var(\bar{x})=\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

*초기하 분포를 따르는 변수의 분산 $V(X)=\underbrace{n\cdot\frac{k}{M}\cdot\frac{M-k}{M}}_{n\cdot p\cdot(1-p)}\cdot\frac{M-n}{M-1}$

: 이항 분포의 분산 또한 indep이 깨졌을 때 수정된다.

- 변수들의 sample간 iid가 깨져서 $Var(X_i)+Var(X_j)$ 뿐만 아니라 $2Cov(X_i,X_j)$가 생기고 Variance가 커진다. (참고로 Cov는 음수가 나올 수 있으나 이전 샘플의 성격을 거의 유지할 것이므로 양수로 볼 수 있음)
- $n \ll N$ 일 때 Correction factor가 1에 가까워진다는 것은 수식에서도 알 수 있지만 (정확히는 $n\rightarrow 1$)
  위 현상을 바탕으로 해석할 때는 Cov가 생길 표본 수가 줄어들어 분산의 크기가 많이 증가하지 않기 때문으로 생각할 수 있다.
- Fancier CLT? $x_i \sim iid$ If $x_i \sim$ not iid, Not fancier.

→ Lindberg-Levy's CLT

Not fancier CLT : L-F CLT

Question 6.

문제의 의미: Regression의 해석 순서

→ Regression 혹은 Statistics의 결점. ① Predictive ② Analytic

선호. Why? 예측은 하면할수록 Data의 특성과 불확실성이 커지기 때문.
관리는 마케팅 측면에서 특정 고객층이나 요인에 대한 상관성
혹은 인과성을 고려할 수 있음.

· 복습: 1강의 R^2 부분. / 제 1종 오류와 2종 오류 ($\alpha \uparrow$ $(1-\beta) \downarrow$ 무슨 의미?)

Multicollinearity : $\dfrac{1}{1-corr(X_i, X_j)}$ 은 두 수의 관계밖에 확인 못함. Data 자체의 Multicollinearity를 확인하고 싶으면 $Det(X)$를 확인하자.

* Multicollinearity가 아닐 때 X의 변동이 낮으면 ($X \rightarrow$ Constant : 변동이 없음) $|X'X| \downarrow$ $(X'X)^{-1} \uparrow$ $V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \uparrow$ $t\text{-value} = \dfrac{\hat{\beta} - \beta_0}{s.e(\hat{\beta})} \downarrow$

X의 2nd moment를 고려해보자.

단측 검정을 하면.

* 단측 검정이 양측 검정보다 좋은 이유? 부호를 알고 있을 때, $H_0 : \beta = 0$ 이라는 기각할 확률이 높아지니까.

양측 검정을 한다면 이는 사실상 유의 확률을 $\frac{\alpha}{2}$로 두는 것과 동일하다.

Question 7.

문제의 의미: Dummy variable의 추가와 Interaction term의 의미.

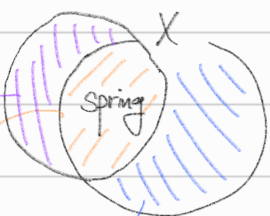
/ test나 C.I를 구하기 위해서는 분산(2nd moment)이 필요하다. → 통계학은 다 평균과 분산 얘기다! (대부분 정규 분포에 기반한다는 얘기)

* Dummy variable interaction.

ex) X : 고객 방문량. sp : 봄 dummy 변수

y : 옷 판매량 (dependent)

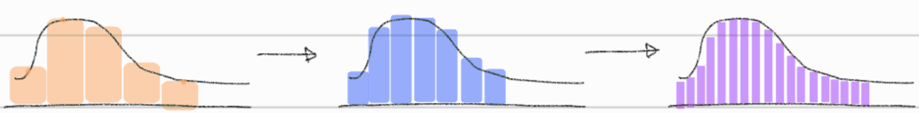
$y = \beta_0 + \beta_1 X + \beta_2 sp + \beta_3 X \cdot sp$



: 변수 X에서 sp가 같이 설명하는 비중을 따로 떼어놓은 것이라고 생각하자.

* 복습 : 분산에 $\frac{1}{n}$이 추가되는 이유는? 오차와 문제 다시 보기 혹은 오차와 강의 자료.

* 이산형이 Data 수가 커질수록 연속형으로 수렴?



수업 내용.

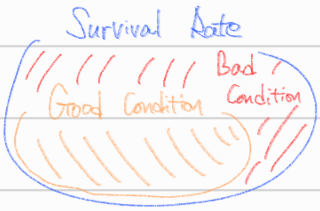
6. Categorical Data.

6.1 Categorical input with categorical input.

- odds ratio : odds 비교. odd? 전체 분의 비율이 아닌(Risk) 반대 경우의 수에 대한 비율 언제 쓰나? 콘텐트 영역에 대해 알아보기.

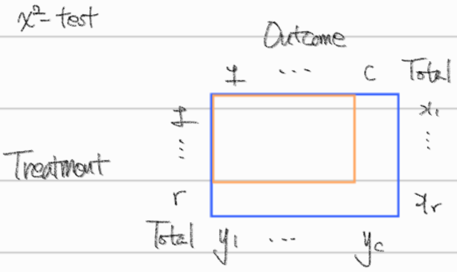
6.1.1 Simpson's paradox. (Question 7)

☆ Categorical Variable 내에 고려할 변수가 남아있다는 것 → Multivariate Regression.



6.1.2 χ^2-test

- 자유도 :



$\chi^2((r-1)(c-1))$ 인 이유는

각 행과 열의 합이 정해져 있기 때문에

경우의 수가 1씩 줄어들기 때문이다.

6.2 ANOVA

6.2.3 Interpretations and assumptions.

- 모든 Data의 2nd moment인 분산이 동일하다는 전제 하에 (Homogeneity) 2nd moment가 얼마나 다른지 보겠다는 것.

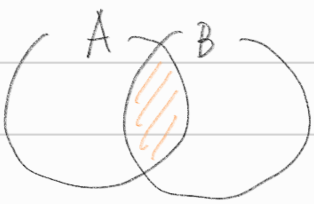
→ 분산을 보겠다는 것.

교재 읽어보기.

6.2.4 Some extensions of ANOVA (Question 7번)

- $SS_A + SS_B + SS_{AB}$

↳ Interaction term이 차지 혹은 설명하는 2nd moment를 의미.



6.2.6 Kruskal-Wallis : the non-parametric version.

- Wilcoxon signed rank test / Mann-Whitney U test : comparing medians

  ↓ ANOVA.

  Kruskal-Wallis