Final Exam for Data-based statistical decision model

From rags to riches

Sungkyu Jung

8/30/2018

Final exam

마지막 시험 답안은 R markdown을 이용해서 작성하세요. 2018년 9 월 9 일 자정 까지 김보영 조교 (by0705@gmail.com) 에게 .Rmd 와 .html (또는 .pdf, .docx) 을 보내세요.

Late submission policy

I will deduct 5% of points for each hour. So if you submit at 3:30am on Monday, then you will get only (1-0.05)^3 = 0.857375 of your earned points.

The data

이 시험에서는 현대 미국에서 여러 세대에 걸친 경제적 이동성(저소득층 출신이 고소득층이 되는 경향)을 조사합니다. 이 데이터는 세금 기록을 바탕으로 한 대규모 연구에서 나온 것으로, 연구원은 수십 년 전에 성인 소득을 부모 소득에 연계 할 수있었습니다. 사생활 보호를 위해 개인 수준의 데이터는 없지만 대부분의 미국 인구를 포함하는 수백 개의 커뮤니티의 경제적 이동성에 대한 통계자료를 수집했습니다. 우리는 지역 사회의 특성으로부터 경제적 이동성을 예측하는 데 관심이 있습니다.

This take-home exam will look at economic mobility across generations in the contemporary USA. The data come from a large study[1], based on tax records, which allowed researchers to link the income of adults to the income of their parents several decades previously. For privacy reasons, we don't have that individual-level data, but we do have aggregate statistics about economic mobility for several hundred communities, containing most of the American population, and covariate information about those communities. We are interested in predicting economic mobility from the characteristics of communities.

데이터는 다음의 R 코드를 사용하여 읽을 수 있습니다. 741 개의 커뮤니티 (관찰값들)과 43 개의 변수가 있습니다.

[1] Chetty, Raj, Nathaniel Hendren, Patrick Kline and Emmanuel Saez (2014). "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." Quarterly Journal of Economics, 129: 1553-1623. Finding and reading this paper does not actually help you

Data can be read using the following R code. There are 741 communities (observations) and 43 variables.

```
dat <- read.csv("mobility.csv")
```

우리가 예측하고자하는 변수는 경제적 이동성입니다 (economic mobility). 나머지 변수는 예측 변수입니다.

The variable we want to predict is economic mobility; the rest are predictor variables or covariates.

1. Mobility: The probability that a child born in 1980 – 1982 into the lowest quintile (20%) of household income will be in the top quintile at age 30. Individuals are assigned to the community they grew up in, not the one they were in as adults. (가계 소득의 최저 5 분위수 (20 %)에 속해 있는 1980-1982 년 출생한 아이가 30세에 되었을 때 상위 1 분위에 속할 확률)
2. Population in 2000. (2000년 기준 인구)
3. Is the community primarily urban or rural? (커뮤니티가 도시인가 시골인가?)
4. Black: percentage of individuals who marked black (and nothing else) on census forms. (흑인의 비율)
5. Racial segregation: a measure of residential segregation by race. (인종별 주거지 분리의 정도)
6. Income segregation: Similarly but for income. (소득별 주거지 분리의 정도)
7. Segregation of poverty: Specifically a measure of residential segregation for those in the bottom quarter of the national income distribution. (저소득층과 중상류층의 주거지 분리의 정도)
8. Segregation of affluence: Residential segregation for those in the top qarter. (상류층과 중하층의 주거지 분리의 정도)
9. Commute: Fraction of workers with a commute of less than 15 minutes. (15 분 미만 통근하는 주민의 비율)
10. Mean income: Average income per capita in 2000. (평균 소득)
11. Gini: A measure of income inequality, which would be 0 if all incomes were perfectly equal, and tends towards 100 as all the income is concentrated among the richest individuals. (지니 계수)
12. Share 1%: Share of the total income of a community going to its richest 1%. (상위 1% 가 차지하는 수입의 비율)
13. Gini bottom 99%: Gini coefficient among the lower 99% of that community. (상위 1 %를 제외한 나머지의 지니 계수)
14. Fraction middle class: Fraction of parents whose income is between the national 25th and 75th percentiles. (중산층 비율)
15. Local tax rate: Fraction of all income going to local taxes. (지방세율)
16. Local government spending: per capita. (1 인당 지방정부 지출)

17. Progressivity: Measure of how much state income tax rates increase with income. (세금 가중의 정도)

18. EITC: Measure of how much the state contributed to the Earned Income Tax Credit (a sort of negative income tax for very low-paid wage earners). (저소득층을 위한 세금 공제의 정도)

19. School expenditures: Average spending per pupil in public schools. (공립학교의 학생 1 인당 평균 지출.)

20. Student/teacher ratio: Number of students in public schools divided by number of teachers.(학생 / 교사 비율)

21. Test scores: Residuals from a linear regression of mean math and English test scores on household income per capita. (시험 점수: 언어+수학 점수를 평균 가정 소득에 회귀한 잔차)

22. High school dropout rate: Also, residuals from a linear regression of the dropout rate on per-capita income. (고등학교 중퇴율 : 실제 중퇴율를 평균 가정 소득에 회귀한 잔차)

23. Colleges per capita (1 인당 대학의 갯수)

24. College tuition: in-state, for full-time students (대학 등록금)

25. College graduation rate: Again, residuals from a linear regression of the actual graduation rate on household income per capita. (대학 졸업율: 실제 졸업율를 평균 가정 소득에 회귀한 잔차)

26. Labor force participation: Fraction of adults in the workforce. (노동인구 중 성인의 비율)

27. Manufacturing: Fraction of workers in manufacturing. (제조업 근로자의 비율)

28. Chinese imports: Growth rate in imports from China per worker between 1990 and 2000. (중국산 수입 증가율)

29. Teenage labor: fraction of those age 14 – 16 who were in the labor force. (노동인구 중 10 대의 비율)

30. Migration in: Migration into the community from elsewhere, as a fraction of 2000 population. (이사오는 비율)

31. Migration out: Ditto for migration into other communities. (이사 나가는 비율)

32. Foreign: fraction of residents born outside the US. (외국 태생 인구 비율)

33. Social capital: Index combining voter turnout, participation in the census, and participation in community organizations. (사회 참여의 정도)

34. Religious: Share of the population claiming to belong to an organized religious body. (종교 생활 참여의 정도)

35. Violent crime: Arrests per person per year for violent crimes. (폭력 범죄율)

36. Single motherhood: Number of single female households with children divided by the total number of households with children. (전체 아이가 있는 가정 중 엄마 혼자 아이 키우는 집의 비율)
37. Divorced: Fraction of adults who are divorced. (이혼한 비율)
38. Married: Ditto. (결혼한 비율)
39. Longitude: Geographic coordinate for the center of the community (경도: 동서)
40. Latitude: Ditto (위도: 남북)
41. ID: A numerical code, identifying the community. (커뮤니티 식별 코드)
42. Name: the name of principal city or town. (동네 이름)
43. State: the state of the principal city or town of the community. (동네가 속한 미국의 주)

The questions

1. A map of mobility

a. Make a plot where the x and y coordinates are longitude and latitude, and mobility is indicated by color (possibly grey scale), by a third coordinate, or some other suitable device. Make sure your map is legible. Describe the geographic pattern in words. [logitude, latitude, mobility 변수의 값을 나타내는 그래프를 그리세요. X 축과 Y 축이 longitude 와 latitude 로 된 그래프를 만들어 Mobility 변수의 값을 색을 이용하거나, 차원을 늘리거나 또는 다른 방법을 이용해서 표시하세요. 그래프는 읽기 편하도록 만들고, 그래프를 통해서 발견된 mobility의 지리적 경향을 말로 설명하세요.]

b. Discretizing the `Mobility` values may enhance visualizing. Create a new variable, called `MobilityCat` with values `high` if `Mobility > 0.1`, and `low` otherwise. Make a plot where the x and y coordinates are longitude and latitude, and the categorized mobility (i.e. `MobilityCat`) is indicated by color. This time, filter your observations so that only the continental part of USA is visible (that is, remove data corresponding to Alaska and Hawaii). Has the geographic pattern become clearer? [`Mobility` 변수의 값을 몇 개의 카테고리로 나누면 그래프의 가독성을 높일 수가 있습니다. `MobilityCat`이라는 이름의 새 변수를 만드세요. 새 변수의 값은 `mobility > 0.1` 이면 `high`, 그렇지 않으면 `low` 로 주면 됩니다. 새로운 변수의 값이 색으로 표현된 그래프를 새로 만드세요. 이 그래프는 위 문제와 같이 x, y축이 위도와 경도이 되도록 그리면 됩니다. 이번에는 알래스카와 하와이를 제외한 미 대륙 본토만을 표시합니다. Mobility의 지리적 경향이 더 확실하게 보이나요?]

2. A bunch of simple regression models

Make scatter plots of mobility against each of the following variables. Include on each plot a line for the simple or univariate regression, and give a table of the regression coefficients. Carefully explain the interpretation of each coefficient. Do any of the results seem odd? [Y축을 Mobility로 하는 하면서 다음의 리스트에 있는 변수를 X축의 값으로 하는 그래프를 각각 그리세요. 적합된 회귀선을 각각의 그래프에 추가하며, 적합된 coefficient 의 값을 테이블에 정리하고, 각각의 coefficient를 해석하세요. 적합 결과에 대해 간단히 논하세요.]

a. Population

b. Mean household income per capita

c. Racial segregation

d. Income share of the top 1%

e. Mean school expenditures per pupil

f. Violent crime rate

g. Fraction of workers with short commutes.

3. All things considered

Run a linear regression of mobility against all appropriate covariates. [적절한 설명변수들을 이용하여 다중회귀분석을 식시합니다.]

a. Report all regression coefficients and their standard errors; you may use either a table or a figure as you prefer. [적합된 계수의 값과 그 standard error를 보여주세요.]

b. Explain why the ID variable must be excluded. [Variable `ID`는 설명변수로 쓰이지 않습니다. 이유를 설명하세요.]

c. Explain which other variables, if any, you excluded from the regression, and why. (If you think they can all be used, explain why.) [For this question, do not use any automated variable selection, and try to keep as many variables as possible.] [다른 설명변수 중에 모형 적합에 쓰지 않은 변수가 있다면, 왜 제외했는지 설명하세요. 만약 모두 쓰였다면, 그렇게 할 수 있는 이유를 설명하세요.]

d. Compare the coefficients you found in problem 2 to the coefficients for the same variables in this regression. Are they significantly different? Have any changed sign? [2 번 문제의 답과 다중회귀 분석에서 적합한 coefficient 들을 비교하세요. 다른 값들이 나왔나요? 계수의 부호가 바뀌었나요?]

e. Take a look at the variation inflation factor for each variable. Report those variables with VIF greater than 10. Do you suspect a (nearly) multicollinearity? If so, give a reason for, and suggest a way to avoid it. [Variation inflation factor (VIF) 를 보고, VIF 가 10 보다 큰 변수들을 찾으세요. Multicollinearity가 있다고 의심이 되나요? 그렇다면, 이유를 대고, multicollinearity 를 피할 수 있는 방법을 제시하세요.]

4. Please in my front yard

a. Inspect the missingness pattern in variables `Colleges`, `Tuition` and `Graduation`. [Note: `NA` is a missing value.] How many observations have no measurements for these variables? [변수 `Colleges`, `Tuition` and `Graduation`에는 기록되지 않은 값(missing value)가 많이 있습니다. 얼마나 많은 관측값에서 이 변수들의 값이 기록되지 않았나요?]

b. Did the missing values happen at random? To answer this, plot a scatter of `Mobility` and `Population` (choose a suitable scale for `Population`), and inspect which data points have missing values in (all of, or some of) variables `Colleges`, `Tuition` and `Graduation`. [누락된 값이 임의로 형성되었나요? 대답을 하기 위해서, `Mobility`와 `Population`의 scatter plot을 그리고, 어떤 관측값에서 전술한 세 변수의 값이 누락되었는지 조사하세요.]

c. Create a new variable, called `HE`, whose value is `TRUE` if there is a higher education institution in the community, is `FALSE` if not. Replace all `NA` values in variables `Colleges`, `Tuition` and `Graduation` with `0`. [`HE` 라는 이름의 새 변수를 만드세요. 각 community (관측값) 에 고등 교육기관이 있다면 `HE` 의 값은 `TRUE`, 그렇지 않다면 `FALSE`로 줍니다. 전술한 세 변수의 `NA` 값은 모두 0 으로 바꿉니다.]

5. All things considered, again.

Fit a linear regression model, incorporating your findings in problems 3 and 4. If you have removed, created, or modified variables, explain. Report all regression coefficients and their standard errors. Use this model for all problems below. [3번과 4번 문제의 답을 포함하여 새로운 다중 회귀 분석 모형을 세우고, 적합합니다. 새로운 변수를 만들거나, 어떤 변수를 지웠다면, 설명하세요. 모든 coefficient의 값과 그 standard error를 보여주세요. 아래의 문제에는 이 모형을 이용합니다.]

6. Make a map of predicted mobility.

Make a map of the model's predicted mobility. How does it compare, qualitatively, to the map of actual mobility? [5번의 모형을 이용하여 예측한 mobility를 (1번과 같이) 지도 위에 표시합니다. 실제 관측값의 지도와 비교하세요.]

7. Just because I was there

Find `Pittsburgh` in the data set. For this question, assume that the model (you have fitted just before) is well-fitted.

a. What its actual mobility? What is its predicted mobility, according to the model?

b. Holding all else fixed, what is the predicted mobility if the violent crime rate is doubled? If it is halved?

c. Provide a 95% confidence interval for the expected mobility at Pittsburgh.

d. Provide a 95% prediction interval for the expected mobility at Pittsburgh. Explain the difference.

8. After making proper allowances

a. Make a map of the model's residuals. [5번의 모형에서 나온 잔차를 지도 위에 표시하세요.]

b. What are the five communities with the largest positive residuals? The five with the most negative residuals? Provide the names of the communies. (Can you mark these on the map?) [잔차가 (양으로, 그리고 음으로) 가장 큰 다섯 개의 관측값의 이름을 찾으세요. (지도 위에 표시할 수 있나요?)]

9. Expectations and reality

a. Make a scatterplot of actual mobility against predicted mobility. Is the relationship linear? Should it be, is the model right? Is the relationship flat? Should it be, is the model right? [실제 mobility의 값과 예측된 mobility의 값을 이용하여 scatter plot을 그리세요. 관계가 선형인가요? 그렇다면, 좋은 모형인가요? 서로 관계가 없나요? 그렇다면, 좋은 모형인가요?]

b. Make a scatterplot of the model's residuals against predicted mobility. Is the relationship linear? Should it be, is the model right? Is the relationship flat? Should it be, is the model right? [모형의 잔차와 예측된 mobility의 값을 이용하여 scatter plot을 그리세요. 관계가 선형인가요? 그렇다면, 좋은 모형인가요? 서로 관계가 없나요? 그렇다면, 좋은 모형인가요?]

10. Cross-validation, bootstrap and smoothing

For this question, focus on predicting mobility by the fraction of middle class in the community.[이 문제에서는 각 동네의 중산층의 비율을 이용하여 mobility를 예측합니다.]

a. Fit a simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$ to predict `Mobility` by `Middle_class`. Create a plot showing the data points, the fitted regression line, and 95% confidence and prediction intervals. For the intervals, assume that the error ϵ is i.i.d. $N(0, \sigma^2)$. (Is the assumption right?) [간단한 회귀분석 모형을 적합하고, 결과를 그림으로 그리세요. 그래프에는 자료의 값, 적합된 회귀분석 모형, 95% confidence and prediction intervals 을 보여줍니다. Interval을 계산하기 위하여, noise의 분포가 i.i.d. $N(0, \sigma^2)$ 이라고 가정합니다.(이 가정이 이 데이타에 잘 맞나요?)]

b. Use a resampling method to obtain 95% confidence interval. Can you build a 95% prediction interval? (resampling 방법을 이용하여 95% confidence interval을 계산합니다. 95% prediction interval 도 resampling method 를 이용하여 계산할 수 있나요?)

c. Use a smoothing spline to do a nonparametric regression of `Mobility` on `Middle_class`. Use cross-validation to choose the degree of flexibility. Then use a resampling method to obtain 95% confidence interval. Create a plot showing the data points, the spline and the confidence interval. [Smoothing spline을 이용하여 비모수적 회귀분석을 합니다. crossvalidation을 이용하여 smoothing 의 정도를 정하고, resampling 방법을 이용하여 95% confidence interval을 구합니다. 문제 a 와 마찬가지로 그래프를 이용하여 자료, smoothing spline와 그 confidence interval을 모두 표시합니다.]

d. Test whether smoothing is needed here. [Smoothing이 필요한지 가설 검정을 하세요.]