Review on Linear Regression

AIE1901 - AI Exploration - LLM for Optimization

Jie Wang, 2025/11/18

Least Squares in Action with "Communities and Crime" Dataset

- Data with socio-economic data, law enforcement data and crime data
- Conducted by US census, LEMAS survey and FBI
- $n = 100$ features and $m = 1993$ cities

$$\min_{\beta \in \mathbb{R}^{n \times 1}} \quad \|y - X\beta\|_2^2$$

Solution to Linear Regression

$$\beta^* = \underset{\beta \in \mathbb{R}^{n \times 1}}{\arg\min} \ \|y - X\beta\|_2^2$$

$$\text{Fitting Error} = \|y - X\beta^*\|_2^2$$

Solution to Linear Regression (Credits to Minhe Cui)

每个特征的系数:
['0' 'state']: -0.000736
['1' 'population']: 0.170312
['2' 'householdsize']: 0.002164
['3' 'racepctblack']: 0.234512
['4' 'racePctWhite']: -0.010162
['5' 'racePctAsian']: 0.012891
['6' 'racePctHisp']: 0.094121
['7' 'agePct12t21']: 0.193153
['8' 'agePct12t29']: -0.109809
['9' 'agePct16t24']: -0.256304
['10' 'agePct65up']: 0.149932
['11' 'numbUrban']: -0.280876
['12' 'pctUrban']: 0.049135
['13' 'medIncome']: -0.167449
['14' 'pctWWage']: -0.114211
['15' 'pctWFarmSelf']: 0.046754
['16' 'pctWInvInc']: -0.112418
['17' 'pctWSocSec']: 0.124333
['18' 'pctWPubAsst']: 0.016643
['19' 'pctWRetire']: -0.068565
['20' 'medFamInc']: 0.257096
['21' 'perCapInc']: 0.087673
['22' 'whitePerCap']: -0.316445

['23' 'blackPerCap']: -0.025197
['24' 'indianPerCap']: -0.031712
['25' 'AsianPerCap']: 0.020247
['26' 'HispPerCap']: 0.048151
['27' 'NumUnderPov']: 0.088915
['28' 'PctPopUnderPov']: -0.128540
['29' 'PctLess9thGrade']: -0.109829
['30' 'PctNotHSGrad']: 0.073128
['31' 'PctBSorMore']: 0.067733
['32' 'PctUnemployed']: 0.015884
['33' 'PctEmploy']: 0.297863
['34' 'PctEmplManu']: -0.065194
['35' 'PctEmplProfServ']: -0.025940
['36' 'PctOccupManu']: 0.082755
['37' 'PctOccupMgmtProf']: 0.122962
['38' 'MalePctDivorce']: 0.499765
['39' 'MalePctNevMarr']: 0.241067
['40' 'FemalePctDiv']: 0.199440
['41' 'TotalPctDiv']: -0.562766
['42' 'PersPerFam']: -0.175245
['43' 'PctFam2Par']: -0.002757
['44' 'PctKids2Par']: -0.206555
['45' 'PctYoungKids2Par']: -0.018752
['46' 'PctTeen2Par']: 0.000549

['47' 'PctWorkMomYoungKids']: 0.054809
['48' 'PctWorkMom']: -0.190325
['49' 'NumIlleg']: -0.120749
['50' 'PctIlleg']: 0.142114
['51' 'NumImmig']: -0.142770
['52' 'PctImmigRecent']: 0.030405
['53' 'PctImmigRec5']: 0.013433
['54' 'PctImmigRec8']: -0.078797
['55' 'PctImmigRec10']: 0.051420
['56' 'PctRecentImmig']: -0.036438
['57' 'PctRecImmig5']: -0.202392
['58' 'PctRecImmig8']: 0.462944
['59' 'PctRecImmig10']: -0.244173
['60' 'PctSpeakEnglOnly']: 0.028557
['61' 'PctNotSpeakEnglWell']: -0.128188
['62' 'PctLargHouseFam']: 0.084508
['63' 'PctLargHouseOccup']: -0.272050
['64' 'PersPerOccupHous']: 0.587513
['65' 'PersPerOwnOccHous']: 0.030498
['66' 'PersPerRentOccHous']: -0.206516
['67' 'PctPersOwnOccup']: -0.713655
['68' 'PctPersDenseHous']: 0.198755
['69' 'PctHousLess3BR']: 0.141285
['70' 'MedNumBR']: 0.033666

['71' 'HousVacant']: 0.169255
['72' 'PctHousOccup']: -0.047132
['73' 'PctHousOwnOcc']: 0.605640
['74' 'PctVacantBoarded']: 0.050369
['75' 'PctVacMore6Mos']: -0.060656
['76' 'MedYrHousBuilt']: -0.011060
['77' 'PctHousNoPhone']: 0.023730
['78' 'PctWOFullPlumb']: -0.018002
['79' 'OwnOccLowQuart']: -0.427449
['80' 'OwnOccMedVal']: 0.291401
['81' 'OwnOccHiQuart']: 0.018587
['82' 'RentLowQ']: -0.248969
['83' 'RentMedian']: 0.010708
['84' 'RentHighQ']: -0.079179
['85' 'MedRent']: 0.367479
['86' 'MedRentPctHousInc']: 0.051856
['87' 'MedOwnCostPctInc']: -0.036408
['88' 'MedOwnCostPctIncNoMtg']: -0.075512
['89' 'NumInShelters']: 0.146848
['90' 'NumStreet']: 0.169658
['91' 'PctForeignBorn']: 0.154510
['92' 'PctBornSameState']: 0.033369
['93' 'PctSameHouse85']: 0.002175
['94' 'PctSameCity85']: 0.020577

['94' 'PctSameCity85']: 0.020577
['95' 'PctSameState85']: 0.006768
['96' 'LandArea']: 0.012123
['97' 'PopDens']: 0.000476
['98' 'PctUsePubTrans']: -0.037276
['99' 'LemasPctOfficDrugUn']: 0.025286

Fitting Error ≈ 33

Sparse Linear Regression

$$\min_{\beta \in \mathbb{R}^{n \times 1}} \quad \|y - X\beta\|_2^2$$

$$\text{s.t.} \quad \|\beta\|_0 \leq k$$

Sparse Linear Regression

$$\min_{\beta \in \mathbb{R}^{n\times 1}, q \in \{0,1\}^n} \quad \|y - X\beta\|_2^2$$

$$\text{s.t.} \quad \sum_{i=1}^{n} q_i \leq k$$

$$\text{s.t.} \quad -M \cdot q_i \leq \beta_i \leq M \cdot q_i$$

Sparse Linear Regression (with Quadratic Regularization)

$$\min_{\beta \in \mathbb{R}^{n \times 1}, q \in \{0,1\}^n} \quad \|y - X\beta\|_2^2 + \lambda \cdot \|\beta\|_2^2$$

$$\text{s.t.} \quad \sum_{i=1}^{n} q_i \leq k$$

$$\text{s.t.} \quad -M \cdot q_i \leq \beta_i \leq M \cdot q_i$$

Sparse Linear Regression (with Quadratic Regularization)

$$\min_{\beta \in \mathbb{R}^{n \times 1}, q \in \{0,1\}^n} \quad \|y - X\beta\|_2^2 + \lambda \cdot \sum_{i=1}^{n} \beta_i^2$$

$$\text{s.t.} \quad \sum_{i=1}^{n} q_i \leq k$$

$$\text{s.t.} \quad -M \cdot q_i \leq \beta_i \leq M \cdot q_i$$

Sparse Linear Regression (with Quadratic Regularization)

$$\min_{\beta \in \mathbb{R}^{n\times 1}, q \in \{0,1\}^n} \quad \|y - X\beta\|_2^2 + \lambda \cdot \sum_{i=1}^{n} t_i$$

$$\text{s.t.} \quad \beta_i^2 \leq t_i$$

$$\sum_{i=1}^{n} q_i \leq k$$

$$\text{s.t.} \quad -M \cdot q_i \leq \beta_i \leq M \cdot q_i$$

Solution to Sparse Linear Regression

- $\lambda = 5, M = 0.4, k = 4$
- Selected Variables:
  - “racepctblack”
  - “MalePctDivorce”
  - “PctIlleg”
  - “PctPersDenseHous”

Appendix: Variable Information

Variable Name	Description
state	US state (by number) - not counted as predictive above, but if considered, should be considered nominal
population	Population for community (numeric - decimal)
householdsize	Mean people per household (numeric - decimal)
racepctblack	Percentage of population that is African American (numeric - decimal)
racePctWhite	Percentage of population that is Caucasian (numeric - decimal)
racePctAsian	Percentage of population that is of Asian heritage (numeric - decimal)
racePctHisp	Percentage of population that is of Hispanic heritage (numeric - decimal)
agePct12t21	Percentage of population that is 12-21 in age (numeric - decimal)
agePct12t29	Percentage of population that is 12-29 in age (numeric - decimal)
agePct16t24	Percentage of population that is 16-24 in age (numeric - decimal)
agePct65up	Percentage of population that is 65 and over in age (numeric - decimal)
numbUrban	Number of people living in areas classified as urban (numeric - decimal)
pctUrban	Percentage of people living in areas classified as urban (numeric - decimal)
medIncome	Median household income (numeric - decimal)
pctWWage	Percentage of households with wage or salary income in 1989 (numeric - decimal)
pctWFarmSelf	Percentage of households with farm or self employment income in 1989 (numeric - decimal)
pctWInvInc	Percentage of households with investment / rent income in 1989 (numeric - decimal)
pctWSocSec	Percentage of households with social security income in 1989 (numeric - decimal)
pctWPubAsst	Percentage of households with public assistance income in 1989 (numeric - decimal)
pctWRetire	Percentage of households with retirement income in 1989 (numeric - decimal)
medFamInc	Median family income (differs from household income for non-family households) (numeric - decimal)
perCapInc	Per capita income (numeric - decimal)
whitePerCap	Per capita income for Caucasians (numeric - decimal)
blackPerCap	Per capita income for African Americans (numeric - decimal)
indianPerCap	Per capita income for Native Americans (numeric - decimal)

Appendix: Variable Information

Variable Name	Description
AsianPerCap	Per capita income for people with Asian heritage (numeric - decimal)
HispPerCap	Per capita income for people with Hispanic heritage (numeric - decimal)
NumUnderPov	Number of people under the poverty level (numeric - decimal)
PctPopUnderPov	Percentage of people under the poverty level (numeric - decimal)
PctLess9thGrade	Percentage of people 25 and over with less than a 9th grade education (numeric - decimal)
PctNotHSGrad	Percentage of people 25 and over that are not high school graduates (numeric - decimal)
PctBSorMore	Percentage of people 25 and over with a bachelor's degree or higher education (numeric - decimal)
PctUnemployed	Percentage of people 16 and over, in the labor force, and unemployed (numeric - decimal)
PctEmploy	Percentage of people 16 and over who are employed (numeric - decimal)
PctEmplManu	Percentage of people 16 and over who are employed in manufacturing (numeric - decimal)
PctEmplProfServ	Percentage of people 16 and over who are employed in professional services (numeric - decimal)
PctOccupManu	Percentage of people 16 and over who are employed in manufacturing (numeric - decimal)
PctOccupMgmtProf	Percentage of people 16 and over who are employed in management or professional occupations (numeric - decimal)
MalePctDivorce	Percentage of males who are divorced (numeric - decimal)
MalePctNevMarr	Percentage of males who have never married (numeric - decimal)
FemalePctDiv	Percentage of females who are divorced (numeric - decimal)
TotalPctDiv	Percentage of population who are divorced (numeric - decimal)
PersPerFam	Mean number of people per family (numeric - decimal)
PctFam2Par	Percentage of families (with kids) that are headed by two parents (numeric - decimal)
PctKids2Par	Percentage of kids in family housing with two parents (numeric - decimal)
PctYoungKids2Par	Percent of kids 4 and under in two parent households (numeric - decimal)
PctTeen2Par	Percent of kids age 12-17 in two parent households (numeric - decimal)
PctWorkMomYoungKids	Percentage of moms of kids 6 and under in labor force (numeric - decimal)
PctWorkMom	Percentage of moms of kids under 18 in labor force (numeric - decimal)
NumIlleg	Number of kids born to never married (numeric - decimal)

Appendix: Variable Information

Variable Name	Description
PctIlleg	Percentage of kids born to never married (numeric - decimal)
NumImmig	Total number of people known to be foreign born (numeric - decimal)
PctImmigRecent	Percentage of immigrants who immigrated within last 3 years (numeric - decimal)
PctImmigRec5	Percentage of immigrants who immigrated within last 5 years (numeric - decimal)
PctImmigRec8	Percentage of immigrants who immigrated within last 8 years (numeric - decimal)
PctImmigRec10	Percentage of immigrants who immigrated within last 10 years (numeric - decimal)
PctRecentImmig	Percent of population who have immigrated within the last 3 years (numeric - decimal)
PctRecImmig5	Percent of population who have immigrated within the last 5 years (numeric - decimal)
PctRecImmig8	Percent of population who have immigrated within the last 8 years (numeric - decimal)
PctRecImmig10	Percent of population who have immigrated within the last 10 years (numeric - decimal)
PctSpeakEnglOnly	Percent of people who speak only English (numeric - decimal)
PctNotSpeakEnglWell	Percent of people who do not speak English well (numeric - decimal)
PctLargHouseFam	Percent of family households that are large (6 or more) (numeric - decimal)
PctLargHouseOccup	Percent of all occupied households that are large (6 or more people) (numeric - decimal)
PersPerOccupHous	Mean persons per household (numeric - decimal)
PersPerOwnOccHous	Mean persons per owner occupied household (numeric - decimal)
PersPerRentOccHous	Mean persons per rental household (numeric - decimal)
PctPersOwnOccup	Percent of people in owner occupied households (numeric - decimal)
PctPersDenseHous	Percent of persons in dense housing (more than 1 person per room) (numeric - decimal)
PctHousLess3BR	Percent of housing units with less than 3 bedrooms (numeric - decimal)
MedNumBR	Median number of bedrooms (numeric - decimal)
HousVacant	Number of vacant households (numeric - decimal)
PctHousOccup	Percent of housing occupied (numeric - decimal)
PctHousOwnOcc	Percent of households owner occupied (numeric - decimal)

Appendix: Variable Information

Variable Name	Description
PctVacantBoarded	Percent of vacant housing that is boarded up (numeric - decimal)
PctVacMore6Mos	Percent of vacant housing that has been vacant more than 6 months (numeric - decimal)
MedYrHousBuilt	Median year housing units built (numeric - decimal)
PctHousNoPhone	Percent of occupied housing units without phone (in 1990, this was rare!) (numeric - decimal)
PctWOFullPlumb	Percent of housing without complete plumbing facilities (numeric - decimal)
OwnOccLowQuart	Owner occupied housing - lower quartile value (numeric - decimal)
OwnOccMedVal	Owner occupied housing - median value (numeric - decimal)
OwnOccHiQuart	Owner occupied housing - upper quartile value (numeric - decimal)
RentLowQ	Rental housing - lower quartile rent (numeric - decimal)
RentMedian	Rental housing - median rent (Census variable H32B from file STF1A) (numeric - decimal)
RentHighQ	Rental housing - upper quartile rent (numeric - decimal)
MedRent	Median gross rent (Census variable H43A from file STF3A - includes utilities) (numeric - decimal)
MedRentPctHousInc	Median gross rent as a percentage of household income (numeric - decimal)
MedOwnCostPctInc	Median owners cost as a percentage of household income - for owners with a mortgage (numeric - decimal)
MedOwnCostPctIncNoMtg	Median owners cost as a percentage of household income - for owners without a mortgage (numeric - decimal)
NumInShelters	Number of people in homeless shelters (numeric - decimal)
NumStreet	Number of homeless people counted in the street (numeric - decimal)
PctForeignBorn	Percent of people foreign born (numeric - decimal)
PctBornSameState	Percent of people born in the same state as currently living (numeric - decimal)
PctSameHouse85	Percent of people living in the same house as in 1985 (5 years before) (numeric - decimal)
PctSameCity85	Percent of people living in the same city as in 1985 (5 years before) (numeric - decimal)
PctSameState85	Percent of people living in the same state as in 1985 (5 years before) (numeric - decimal)
LandArea	Land area in square miles (numeric - decimal)
PopDens	Population density in persons per square mile (numeric - decimal)
PctUsePubTrans	Percent of people using public transit for commuting (numeric - decimal)

Appendix: Variable Information

Variable Name	Description
LemasPctOfficDrugUn	Percent of officers assigned to drug units (numeric - decimal)

The variable "ViolentCrimesPerPop" is our target variable, which represents *Total number of violent crimes per 100K population (numeric - decimal).*

Sparse Linear Regression in Reality

In a MRI system, we have measurement in the form of

$$b = Ax + \varepsilon$$

- $x \in \mathbb{R}^{n^2}$: image of interest
- $\varepsilon \in \mathbb{R}^{m}$: measurement noise
- $A \in \mathbb{R}^{m \times n^2}$: measurement matrix

Linear Regression with L1-Norm Regularization

- x is typically sparse $\Longrightarrow$ finding the sparsest solution
- L_0-norm minimization

$$\min_{x \in \mathbb{R}^n} \ \lambda \cdot \|x\|_0 + \|Ax - b\|_2^2$$

- Nonconvex, difficult to solve
- Compressive sensing: L_1-norm minimization

$$\min_{x \in \mathbb{R}^n} \ \lambda \cdot \|x\|_1 + \|Ax - b\|_2^2$$

MRI and Compressive Sensing Results

Use 1/4 Fourier coefficients

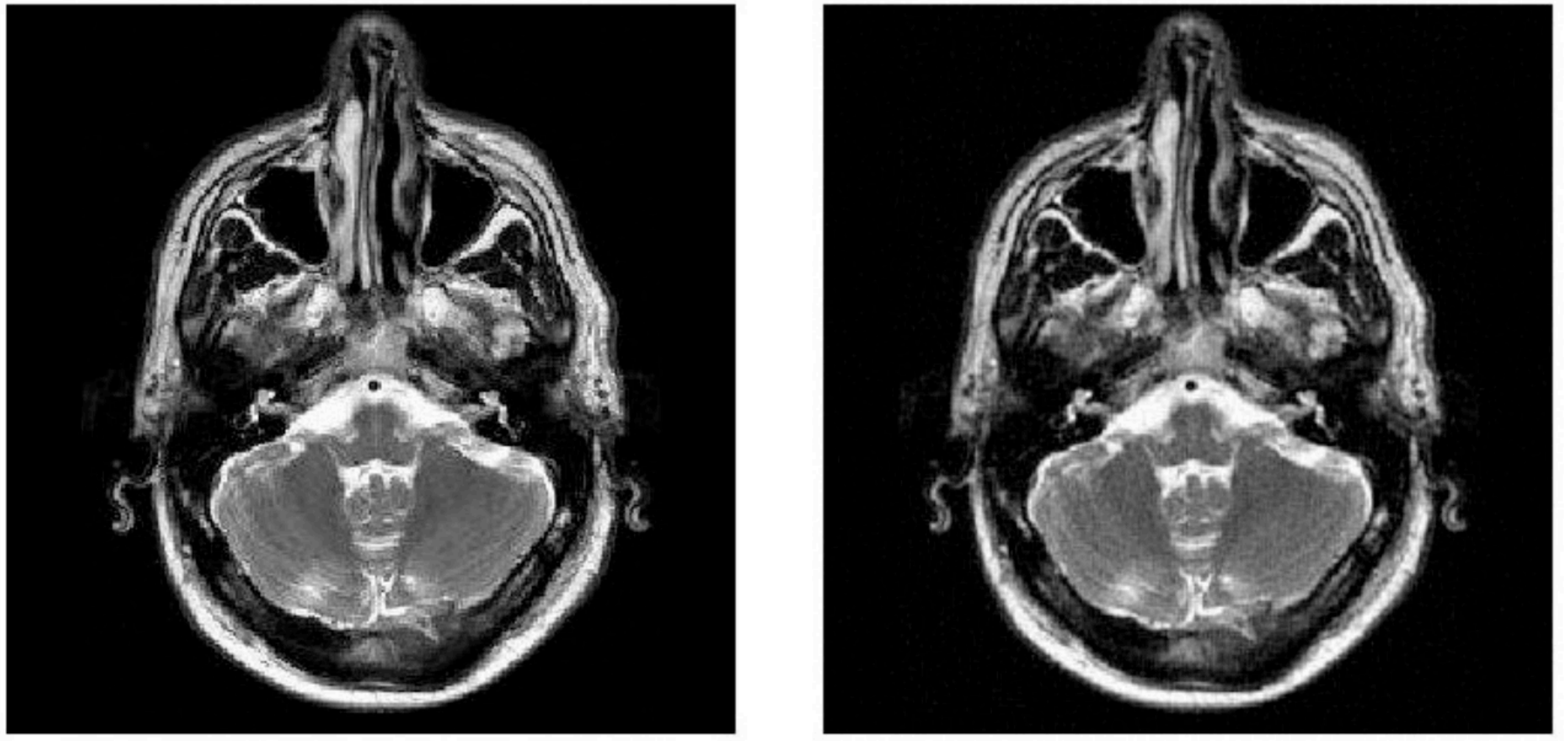

Figure 2: Original vs. Reconstructed Image (courtesy of Y. Zhang, Rice University)

Code Demo (Run test2d_lasso.m)

$n^2 = 128^2 = 16384, m = 12288, k = 8040$