Segunda Lista de Exercícios

Mestrado Professional em Economia — Econometria Avançada
Professor: Hedibert Freitas Lopes — Monitor: Guilherme Piantino
Insper Instituto de Educação e Pesquisa — Outubro/2024

Regressão Linear Múltipla Gaussiana

Variáveis resposta: As variáveis `macro`, `micro`, `estat` e `mat` são as notas padronizadas (média zero e variância unitária) das provas de 2022 da ANPEC de macroeconomia, microeconomia, estatística e matemática, para $n = 779$ candidatos(as).

Variáveis explicativas/exógenas/regressors: As variáveis regressoras são quase todas binárias, menos a variável `idade`.

- `form20.21=1` se candidato(a) é formado(a) em 2020 ou 2021 (417 ou 53,5%)
- `SE=1` se candidato(a) reside na região sudeste do Brasil (392 ou 50,3%)
- `CO=1` se candidato(a) reside na região centro-oeste do Brasil (57 ou 7,3%)
- `naobranco=1` se candidato(a) é não-branco(a) (298 ou 38,3%)
- `femin=1` se candidata (235 ou 30,2%)
- `idade:` idade do candidato(a) em anos (centrada em 24 anos)

Notas:

a) São 66 pretos, 195 pardos, 12 amarelos, 1 indígena e 24 não declarados.

b) Os quartis de idade são 23, 26 e 30.

Dados e script do R: Os dados estão disponíveis no arquivo `anpec2022-data.txt`. O script do R abaixo mostra as 5 primeiras e as 5 últimas observações, além de estatística básicas para cada uma das variáveis.

```
data = read.table("anpec2022-data.txt",header=TRUE)
n = nrow(data)
data[c(1:5,(n-4):n),]
summary(data[,1:4])
summary(data[,5:10])
```

```
         macro       micro       estat         mat form20.21 SE CO naobranco idade femin
1   -1.2971250 -0.30069440 -0.77094800 -0.53402970         0  0  0         0     2     0
2   -0.4863862  0.07626237 -0.83511130 -0.61540860         1  1  0         1     1     1
3   -0.1326095  0.83017590 -0.06515169 -0.04575607         0  1  0         0     7     0
4   -1.0170510 -0.90382520 -0.64262140 -0.85954550         0  0  0         1     2     0
5   -0.8549037 -0.82843390 -1.09176400 -1.51057700         1  1  0         0     3     1
775 -1.0170510 -0.75304250 -1.02760100 -0.77816650         1  0  0         1    -1     0
776 -0.9875699 -0.97921660 -0.77094800 -0.53402970         0  0  0         1    23     0
```

```
777 -0.2800165 -0.14991170 -0.06515169  0.19838080         0  1  0         0     4     0
778  0.2064265  0.90556730  0.70480790  1.09354900         0  1  0         0    13     0
779  0.5159812 -0.30069440  0.83313450  1.01217000         1  0  0         0    -1     0
```

```
     macro              micro              estat              mat
 Min.   :-1.8868   Min.   :-1.9593   Min.   :-1.9259   Min.   :-2.2430
 1st Qu.:-0.7222   1st Qu.:-0.7530   1st Qu.:-0.7068   1st Qu.:-0.6154
 Median :-0.2800   Median :-0.2253   Median :-0.3218   Median :-0.2899
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.5160   3rd Qu.: 0.4532   3rd Qu.: 0.5765   3rd Qu.: 0.3611
 Max.   : 3.2135   Max.   : 3.6950   Max.   : 3.3355   Max.   : 4.3487

   form20.21            SE                CO             naobranco          idade             femin
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :-4.000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:-1.000   1st Qu.:0.0000
 Median :1.0000   Median :1.0000   Median :0.00000   Median :0.0000   Median : 2.000   Median :0.0000
 Mean   :0.5353   Mean   :0.5032   Mean   :0.07317   Mean   :0.3825   Mean   : 3.913   Mean   :0.3017
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.: 6.000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :38.000   Max.   :1.0000
```

Responda de forma detalhada aos seguintes itens:

I) Usando as notas de estatística (`estat`), encontre o subconjunto das variáveis explicativas que tem o menor *erro absoluto mediano (EAM)* na amostra teste. Faça a amostra treino ter tamanho 650 e a amostra teste ter tamanho 120, aleatoriamente alocadas. Repita o exercício 100 vezes.

II) Repita I), mas utilizando a idéia de‘ ‘leave-one-out cross validation (LOOCV)”.

III) Repita I), mas agora utilize a idéia de "10-fold cross validation". Ou seja, as amostras de teste terão 78 observações, menos a última que terá 77, de forma que $9 \times 78 + 77 = 779 = n$.

Fold	Observações na amostra de teste	Folds na amostra treino
1	$i = 1, \ldots, 78$	2–10
2	$i = 79, \ldots, 156$	1,3–10
3	$i = 157, \ldots, 234$	1–2,4–10
4	$i = 235 \ldots, 312$	1–3,5–10
5	$i = 313, \ldots, 390$	1–4,6–10
6	$i = 391, \ldots, 468$	1–5,7–10
7	$i = 469, \ldots, 546$	1–6,8–10
8	$i = 547, \ldots, 624$	1–7,9–10
9	$i = 625, \ldots, 702$	1–8,10
10	$i = 703, \ldots, n$	1–9

IV) Repita I), II) e III), mas assumindo que a variável resposta a média das notas das 4 matérias, macroeconomia, microeconomia, estatística e matemática. Não se esqueca de padronizar a média final. Isso é feito no script abaixo.

```
y = scale(apply(data[,1:4],1,mean))
```

Trabalho junto ao monitor

Guilherme usará as variáveis que aparecem nos dados usados no seguinte exemplo para fazer sua monitoria.

`https://hedibert.org/wp-content/uploads/2024/01/regressaomultipla-return-to-education.html`

The data is a 1976 Panel Study of Income Dynamics, based on data for the previous year, 1975. Of the 753 observations, the first 428 are for women with positive hours worked in 1975, while the remaining 325 observations are for women who did not work for pay in 1975. A more complete discussion of the data is found in Mroz [1987], Appendix 1. Thomas A. Mroz (1987) The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions. Econometrica, Vol. 55, No. 4 (July 1987), pp. 765-799. Stable URL: `http://www.jstor.org/stable/1911029`.

```
data = read.table("http://hedibert.org/wp-content/uploads/2020/01/mroz-data.txt",header=TRUE)
attach(data)
y = scale(log(FAMINC))
X = scale(cbind(LFP,WHRS,KL6,K618,WA,WE,WW,RPWG,HHRS,HA,HE,HW,MTR,WMED,WFED,UN,CIT,AX))
```

The variables in the dataset are as follows:

LFP "A dummy variable = 1 if woman worked in 1975, else 0";
WHRS "Wife's hours of work in 1975";
KL6 "Number of children less than 6 years old in household";
K618 "Number of children between ages 6 and 18 in household";
WA "Wife's age";
WE "Wife's educational attainment, in years";
WW "Wife's average hourly earnings, in 1975 dollars";
RPWG "Wife's wage reported at the time of the 1976 interview (not the same as the 1975 estimated wage). To use the subsample with this wage, one needs to select 1975 workers with LFP=1, then select only those women with non-zero RPWG. Only 325 women work in 1975 and have a non-zero RPWG in 1976.";
HHRS "Husband's hours worked in 1975";
HA "Husband's age";
HE "Husband's educational attainment, in years";
HW "Husband's wage, in 1975 dollars";
FAMINC "Family income, in 1975 dollars. This variable is used to construct the property income variable.";
MTR "This is the marginal tax rate facing the wife, and is taken from published federal tax tables (state and local income taxes are excluded). The taxable income on which this tax rate is calculated includes Social Security, if applicable to wife.";
WMED "Wife's mother's educational attainment, in years";
WFED "Wife's father's educational attainment, in years";
UN "Unemployment rate in county of residence, in percentage points. This taken from bracketed ranges.";
CIT "Dummy variable = 1 if live in large city (SMSA), else 0";
AX "Actual years of wife's previous labor market experience";