lab 5

Zeren Li

10/3/2019

Roadmap

- Matrix
- Multivariate regression
- Omitted variable bias
- Multicollinearity

Matrix

Vector

```r
# unit Vector
matrix(1,3,1)
```

```
##      [,1]
## [1,]    1
## [2,]    1
## [3,]    1
```

```r
#zero vector
matrix(0,3,1)
```

```
##      [,1]
## [1,]    0
## [2,]    0
## [3,]    0
```

```r
# construct a 3*2 matrix
A =  matrix(c(2,3,-2,1,2,2),3,2)
A
```

```
##      [,1] [,2]
## [1,]    2    1
## [2,]    3    2
## [3,]   -2    2
```

```r
# Is something a matrix
is.matrix(A)
```

```
## [1] TRUE
```

Operation

```r
# multiplication by a scalar
c <- 3
c*A
```

```
##      [,1] [,2]
## [1,]    6    3
## [2,]    9    6
## [3,]   -6    6
```

```
# matrix addition & subtraction
B <- matrix(c(4,-2,1,1,2,1),3,2)
A + B
```

```
##      [,1] [,2]
## [1,]    6    2
## [2,]    1    4
## [3,]   -1    3
```

```
A - B
```

```
##      [,1] [,2]
## [1,]   -2    0
## [2,]    5    0
## [3,]   -3    1
```

```
# matrix multiplication
E <- matrix(c(2,-2,1,2,3,1),2,3) # 2*3 matrix

E %*% A # 2*3 matrix * 3*2 matrix
```

```
##      [,1] [,2]
## [1,]    1   10
## [2,]    0    4
```

```
A %*% E # 3*2 matrix * 2*3 matrix
```

```
##      [,1] [,2] [,3]
## [1,]    2    4    7
## [2,]    2    7   11
## [3,]   -8    2   -4
```

Transpose

```
# recall A
A
```

```
##      [,1] [,2]
## [1,]    2    1
## [2,]    3    2
## [3,]   -2    2
```

```
# T(A)
t(A)
```

```
##      [,1] [,2] [,3]
## [1,]    2    3   -2
## [2,]    1    2    2
```

```
# T(T(A)) = A
t(t(A))
```

```
##      [,1] [,2]
## [1,]    2    1
```

```
## [2,]    3    2
## [3,]   -2    2
```

Common Matrices

```r
# unit matrix
matrix(1,3,2)
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    1
## [3,]    1    1
```

```r
#zero matrix
matrix(0,3,2)
```

```
##      [,1] [,2]
## [1,]    0    0
## [2,]    0    0
## [3,]    0    0
```

```r
# diagonal Matrix
S <- matrix(c(2,3,-2,1,2,2,4,2,3),3,3)
S
```

```
##      [,1] [,2] [,3]
## [1,]    2    1    4
## [2,]    3    2    2
## [3,]   -2    2    3
```

```r
diag(S)
```

```
## [1] 2 2 3
```

```r
diag(diag(S))
```

```
##      [,1] [,2] [,3]
## [1,]    2    0    0
## [2,]    0    2    0
## [3,]    0    0    3
```

```r
# identity matrix
I = diag(c(1,1,1))
I
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

```r
# Symmetric Matrix
C = matrix(c(2,1,5,1,3,4,5,4,2),3,3)
C
```

```
##      [,1] [,2] [,3]
## [1,]    2    1    5
## [2,]    1    3    4
## [3,]    5    4    2
```

```r
CT <- t(C)

# inverse of a matrix
A <- matrix(c(4,4,-2,2,6,2,2,8,4),3,3)

AI <- solve(A)
AI
```

```
##      [,1] [,2] [,3]
## [1,]  1.0 -0.5  0.5
## [2,] -4.0  2.5 -3.0
## [3,]  2.5 -1.5  2.0
```

```r
A %*% AI
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

```r
AI %*% A
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

Regression in Matrix Notation

Simple Linear Regression:

$$Y_i = \beta_0 + x_i\beta_1 + c + u_i \text{ for } i = 1, \ldots, n$$

Rewrite in vectors:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \beta_1 + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Ordinary Least Squares

- OLS estimates of parameters β_0 and β minimize sum of squared errors

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - (\beta_0 + X_i\beta_1))^2$$

$$L(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

- OLS estimate of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Summarizing Model Fit

- Fitted values

$$\hat{Y}_i = x_i\hat{\beta}$$

- Residuals (estimates of errors)

$$u_i = Y_i - \hat{Y}_i = \hat{u}_i$$

$$\begin{aligned} ESS &= \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 \\ TSS &= \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \\ SSR &= \sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2 \\ R^2 &= \frac{ESS}{TSS} \end{aligned}$$

- $MSE = SSE/(n-p)$ is an estimate of σ^2
- degrees of freedom $n-p$ where p is the number of parameters in the mean function

Example: Discrimination Analysis

`discrim` is a zip code-level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of blacks.

```
data("discrim", package="wooldridge")  #from wooldridge package

dim(discrim)
```

```
## [1] 410  37
```

```
names(discrim)
```

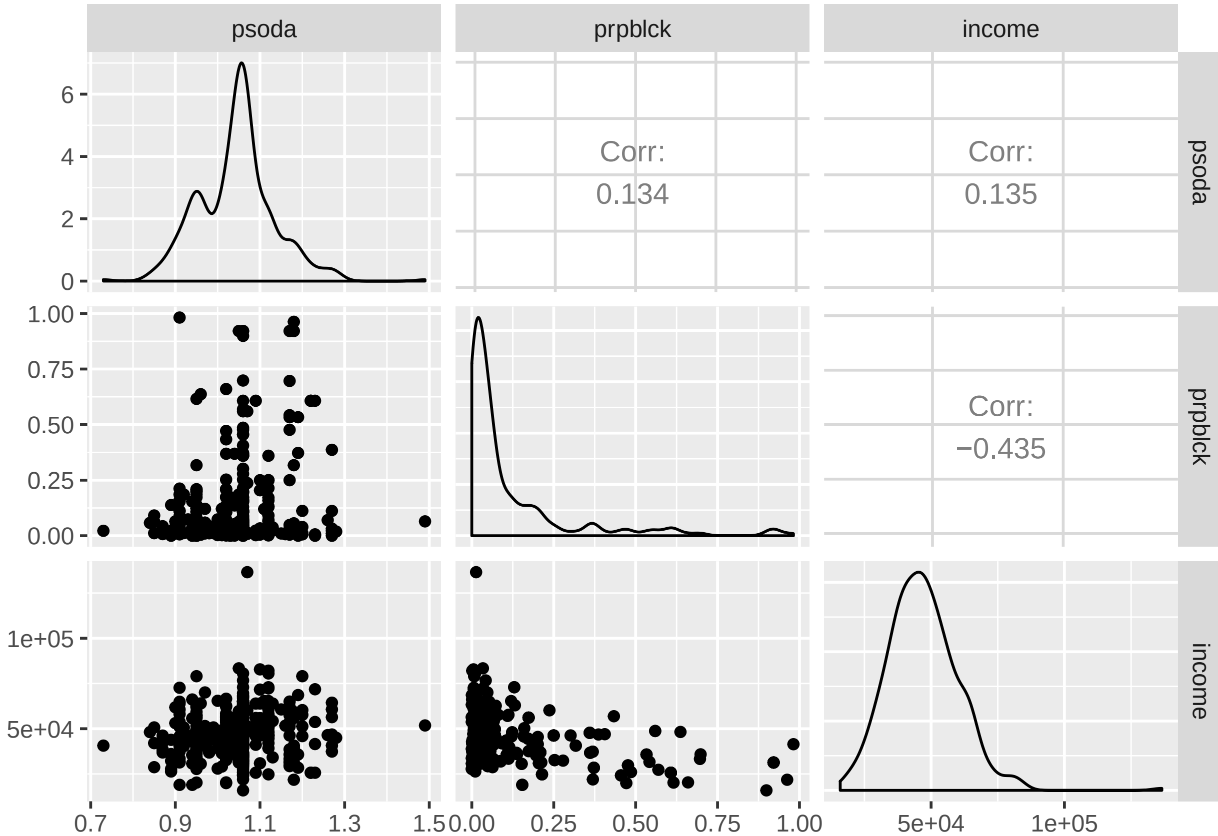
```
##  [1] "psoda"    "pfries"   "pentree"  "wagest"   "nmgrs"    "nregs"
##  [7] "hrsopen"  "emp"      "psoda2"   "pfries2"  "pentree2" "wagest2"
## [13] "nmgrs2"   "nregs2"   "hrsopen2" "emp2"     "compown"  "chain"
## [19] "density"  "crmrte"   "state"    "prpblck"  "prppov"   "prpncar"
## [25] "hseval"   "nstores"  "income"   "county"   "lpsoda"   "lpfries"
## [31] "lhseval"  "lincome"  "ldensity" "NJ"       "BK"       "KFC"
## [37] "RR"
```

$$psoda = \beta_0 + \beta_1 prpblck + income + u.$$

Pairs Plots

```r
discrim1 <- discrim %>%
        select(psoda,prpblck, income) %>%
        na.omit()

ggpairs(discrim1)
```



Summary Statistics

```r
stargazer(discrim1, header = FALSE)
```

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
psoda	401	1.045	0.089	0.730	0.980	1.090	1.490
prpblck	401	0.115	0.184	0.000	0.012	0.121	0.982
income	401	46,999.400	13,215.330	15,919	37,883	54,981	136,529

```r
# indepedent variable
X_ncons <- as.matrix(discrim1[2:3])

# constant
cons <-  as.matrix( rep(1,length(discrim1$psoda)))
```

```r
X <- cbind(cons,X_ncons)
y <- as.matrix(discrim1[1])

b<- solve(t(X) %*% X) %*% t(X)%*%y
b
```

```
##                 psoda
##          9.563196e-01
## prpblck 1.149882e-01
## income  1.602674e-06
```

```r
# y hat
y_hat_byhand <- X %*% b

colnames(y_hat_byhand) <- "y_hat_byhand"

# u hat
u_hat = y-y_hat_byhand

m1 <- lm(psoda ~ prpblck + income, discrim1)
y_hat <- predict(m1)

data.frame(y, y_hat_byhand,  y_hat ) %>% head()
```

```
##   psoda y_hat_byhand    y_hat
## 1  1.12     1.047374 1.047374
## 2  1.06     1.047374 1.047374
## 3  1.06     1.027738 1.027738
## 4  1.12     1.043116 1.043116
## 5  1.12     1.076137 1.076137
## 6  1.06     1.034462 1.034462
```

Exercise: perform OLS regression using matrix operation

```r
X1 <- matrix(c(1,8,3.8, 1,4.5,2.7,3,4,1),3,3)
y1 <- matrix(c(1,0,1),3,1)

cons <-  as.matrix( rep(1,3))
X2 <- cbind(cons, X1 )

# compute beta
```

Measure of Fit

```r
# tss
 tss <- sum( (y - mean(y))^2)
 tss
```

```
## [1] 3.154017
```

```r
# ESS
ess <- sum( (y_hat - mean(y))^2)
ess
```

```
## [1] 0.2025522
```

```r
# SSR
ssr <- sum( (y- y_hat)^2 )
ssr
```

```
## [1] 2.951465
```

```r
# R^2
r_2 = ess/tss
r_2
```

```
## [1] 0.06422039
```

double-check with the result from `lm()`

```r
summary(m1)
```

```
##
## Call:
## lm(formula = psoda ~ prpblck + income, data = discrim1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29401 -0.05242  0.00333  0.04231  0.44322
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.563e-01  1.899e-02  50.354  < 2e-16 ***
## prpblck     1.150e-01  2.600e-02   4.423 1.26e-05 ***
## income      1.603e-06  3.618e-07   4.430 1.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08611 on 398 degrees of freedom
## Multiple R-squared:  0.06422,    Adjusted R-squared:  0.05952
## F-statistic: 13.66 on 2 and 398 DF,  p-value: 1.835e-06
```

Omitted variable bias

OVB is the bias in the OLS estimator that arises when the regressor, X, is *correlated* with an omitted variable. For omitted variable bias to occur, two conditions must be fulfilled: - X is correlated with the omitted variable. - Omitted variable is a determinant of the Y.

Together, result in a violation of the first OLS assumption $E(u_i|X_i) = 0$. Formally, the resulting bias can be expressed as

Direction of Bias

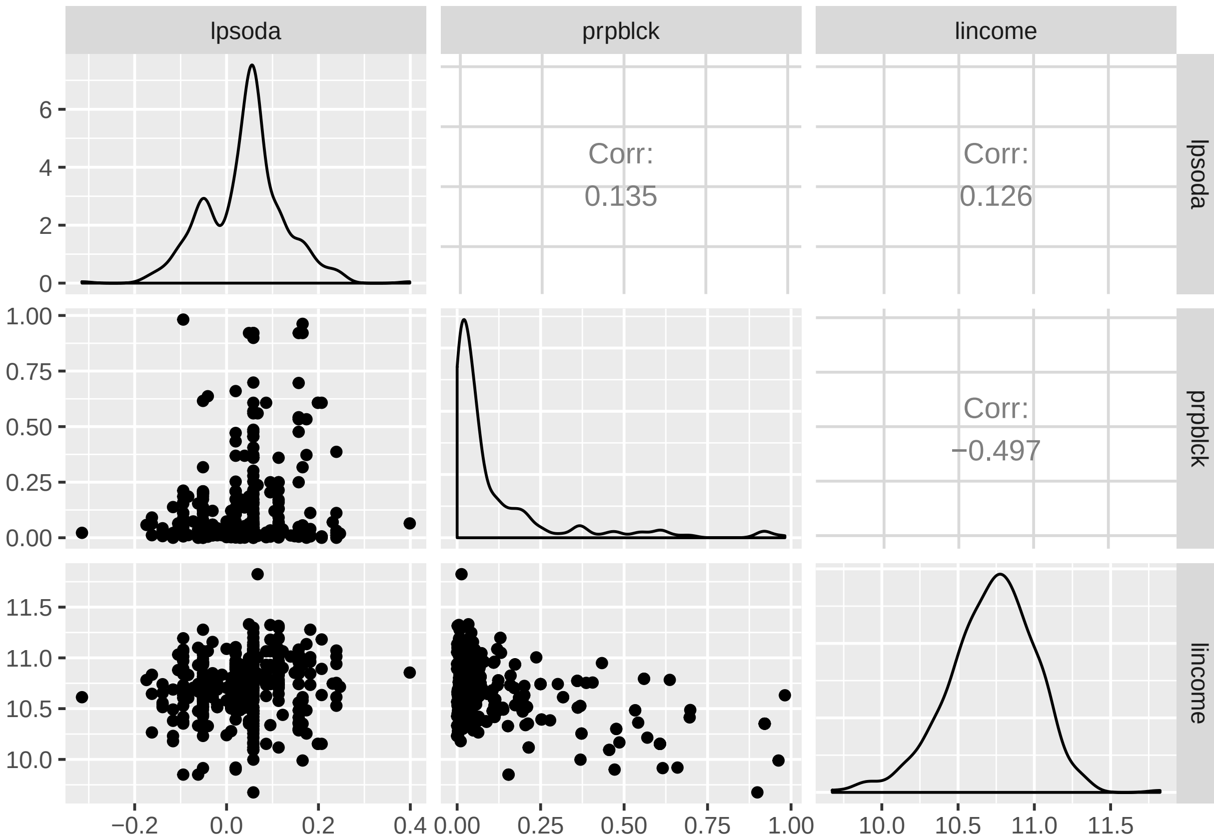
See details in the chapter 3 of Wooldridge's book.

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive Bias	Negative Bias
$\beta_2 < 0$	Negative Bias	Positive Bias

Example: discrimination analysis

$$log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 log(income) + u.$$

```
discrim %>%
      select(lpsoda, prpblck, lincome)   %>%
      ggpairs()
```




```
m1 <- lm(lpsoda ~ prpblck , discrim)
m2 <- lm(lpsoda ~ lincome , discrim)
m3 <- lm(lpsoda ~ prpblck + lincome, discrim)

stargazer(m1, m2, m3, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Dec 29, 2019 - 01:27:37

Yes, one unit shift in prpblck equals a 12% increase in the price of soda, but what does one unit change in proportion black really mean.

One point improvement is actually a 100% improvement, so a 20% improvement is actually a .20 point improvement. Therefore, a 2.4% increase in the price of soda.

Table 2:

	Dependent variable:		
	lpsoda		
	(1)	(2)	(3)
prpblck	0.062***		0.122***
	(0.023)		(0.026)
lincome		0.037**	0.077***
		(0.015)	(0.017)
Constant	0.033***	−0.361**	−0.794***
	(0.005)	(0.158)	(0.179)
Observations	401	401	401
R^2	0.018	0.016	0.068
Adjusted R^2	0.016	0.013	0.063
Residual Std. Error	0.084 (df = 399)	0.084 (df = 399)	0.082 (df = 398)
F Statistic	7.451*** (df = 1; 399)	6.437** (df = 1; 399)	14.540*** (df = 2; 398)
Note:			*p<0.1; **p<0.05; ***p<0.01

Multicollinearity

Consider the following model

$$log(psoda) = \beta_0 + \beta_1 prpblck + \beta 2 log(income) + beta_3 prppov + u.$$

```r
m4 <- lm(lpsoda ~ prpblck + lincome + prppov, discrim)
stargazer(m1,m2,m3,m4, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Dec 29, 2019 - 01:27:37

Variance inflation factor (VIF)

$$VIF_j = 1/(1 - R_j^2)$$

- we would like VIF_j to be smaller
- Rule of thumb: If VIF_j is above 10 (equivalently, R_j^2 is above .9), then we conclude that multicollinearity is a "problem" for estimating bj. But a VIF_j above 10 does not mean that the standard deviation of b^j is too large to b

```r
# fit the model
mv1 <- lm( prpblck ~ prppov + lincome, discrim)

# auxilliary R-squared
a_r_2 <- summary(mv1)$r.squared

# compute VIF
1/(1-a_r_2)
```

Table 3:

	Dependent variable:			
	lpsoda			
	(1)	(2)	(3)	(4)
prpblck	0.062***		0.122***	0.073**
	(0.023)		(0.026)	(0.031)
lincome		0.037**	0.077***	0.137***
		(0.015)	(0.017)	(0.027)
prppov				0.380***
				(0.133)
Constant	0.033***	−0.361**	−0.794***	−1.463***
	(0.005)	(0.158)	(0.179)	(0.294)
Observations	401	401	401	401
R^2	0.018	0.016	0.068	0.087
Adjusted R^2	0.016	0.013	0.063	0.080
Residual Std. Error	0.084 (df = 399)	0.084 (df = 399)	0.082 (df = 398)	0.081 (df = 397)
F Statistic	7.451*** (df = 1; 399)	6.437** (df = 1; 399)	14.540*** (df = 2; 398)	12.604*** (df = 3; 397)
Note:				*p<0.1; **p<0.05; ***p<0.01

```
## [1] 1.927172
```

Exercise: calculate the VIF of *lincome*

```
VIF()
```

```r
vif(mv1)
```

```
##   prppov  lincome
## 3.367308 3.367308
```

What if adding more control variables?

Okay, what happened. The sign on poverty has flipped. Now, the poorer you are, the less you pay for soda, after controlling for the impact of cars.

```r
mv2 <- lm( prpblck ~ prppov + lincome + prpncar, discrim)
summary(mv2)
```

```
##
## Call:
## lm(formula = prpblck ~ prppov + lincome + prpncar, data = discrim)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.46562 -0.04715 -0.02166  0.02273  0.82361
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.74272    0.46079  -3.782 0.000179 ***
## prppov       1.97748    0.30777   6.425  3.7e-10 ***
## lincome      0.15724    0.04199   3.745 0.000207 ***
## prpncar      0.25817    0.15259   1.692 0.091434 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1314 on 405 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4847, Adjusted R-squared:  0.4809
## F-statistic:   127 on 3 and 405 DF,  p-value: < 2.2e-16
```

```r
vif(mv2)
```

```
##    prppov   lincome   prpncar
## 10.176091  3.370050  7.582398
```

- It is easy to misuse such statistics because we cannot specify how much correlation among explanatory variables is "too much."
- Some multicollinearity "diagnostics" are omnibus statistics in the sense that they detect a strong linear relationship among any subset of explanatory variables.