Multiple Regression &Principle Component Analysis

胡懷源

目錄

1 Joint Hypotheses and the F-statistic
2 Testing Simultaneous Hypotheses
3 Prediction and Forecasting
4 Omitted Variable Bias
5 Irrelevant Variables
6 Model Selection Criteria
7 Collinearity
Computer Exercise

載入套件

```r
library(PoEdata)
library(knitr)
library(xtable)
library(printr)
library(effects)
library(car)
library(AER)
library(broom)
library(stats)
library(tidyverse)
```

01

Joint Hypotheses and the F-statistic

Unrestricted model

$$y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e$$

$$H_0: \beta_2 = 0_{and} \beta_3 = 0$$

$$H_A: \beta_2 \neq 0_{or}\ \beta_3 \neq 0$$

Restricted model

$$y = \beta_1 + \beta_4 X_4 + e$$

$$\mathrm{F} = \frac{(\mathrm{SSE_R} - \mathrm{SSE_U})/\mathrm{J}}{\mathrm{SSE_U}/(\mathrm{N} - \mathrm{K})} \sim \mathrm{F}_{(\mathrm{J}, N - \mathrm{K})}$$

02

Testing Simultaneous Hypotheses

$$Sales = \beta_1 + \beta_2 price + \beta_3\ advert + \beta_4\ advert^2 + e$$

$$H_0\colon \beta_3 = 0_{and} \beta_4 = 0$$

$$H_A\colon \beta_3 \neq 0_{or}\ \beta_4 \neq 0$$

```r
alpha <- 0.05
data("andy", package="PoEdata")
N <- NROW(andy) #Number of observations in dataset
K <- 4 #Four Betas in the unrestricted model
J <- 2 #Because Ho has two restrictions
fcr <- qf(1-alpha, J, N-K)
mod1 <- lm(sales~price+advert+I(advert^2), data=andy)
anov <- anova(mod1)
anov # prints 'anova' table for the unrestricted model
```

Anova table for the unrestricted model

```
Analysis of Variance Table

Response: sales
              Df  Sum Sq Mean Sq F value    Pr(>F)
price          1 1219.09 1219.09 56.4952 1.315e-10 ***
advert         1  177.45  177.45  8.2233  0.005441 **
I(advert^2)    1  186.86  186.86  8.6594  0.004393 **
Residuals     71 1532.08   21.58
```

```
SSEu <- anov[4, 2]
```

Values

alpha	0.05
fcr	3.12576423681303
J	2
K	4
N	75L
SSEu	1532.08445870452

Anova table for the restricted model

```
mod2 <- lm(sales~price, data=andy)
anov <- anova(mod2)
anov
```

```
Analysis of Variance Table

Response: sales
          Df Sum Sq Mean Sq F value    Pr(>F)
price      1 1219.1 1219.09  46.928 1.971e-09 ***
Residuals 73 1896.4   25.98
```

```
SSEr <- anov[2,2]
```

$$Sales = \beta_1 + \beta_2 price + \beta_3\ advert + \beta_4\ advert^2 + e$$

$$H_0: \beta_3 = 0_{and} \beta_4 = 0$$

$$H_A: \beta_3 \neq 0_{or}\ \beta_4 \neq 0$$

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} \sim F_{(J,N-K)}$$

$=8.441$

```
fval <- ((SSEr-SSEu)/J) / (SSEu/(N-K))

pval <- 1-pf(fval, J, N-K)
```

Values	
alpha	0.05
fcr	3.12576423681303
fval	8.44135997806643
J	2
K	4
N	75L
pval	0.000514159058423669
SSEr	1896.39083709117
SSEu	1532.08445870452

結果：F= 8.441，F_{cr} =3.1257

F> F_{cr}

落在拒絕域

Reject H_0

廣告支出對營業額有顯著影響

linearHypothesis function

```r
Hnull <- c("advert=0", "I(advert^2)=0")
linearHypothesis(mod1,Hnull)
```

```
Linear hypothesis test

Hypothesis:
advert = 0
I(advert^2) = 0

Model 1: restricted model
Model 2: sales ~ price + advert + I(advert^2)

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     73 1896.4
2     71 1532.1  2    364.31 8.4414 0.0005142 ***
```

產生相同結果

Exam content

F= 8.441 , F_{cr} =3.1257

F> F_{cr}

落在拒絕域

Reject H_0

廣告支出對銷售量有影響

$$Sales = \beta_1 + \beta_2 price + \beta_3\, advert + \beta_4\, advert^2 + e$$

```
Call:
lm(formula = sales ~ price + advert + I(advert^2), data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2553  -3.1430  -0.0117   2.8513  11.8050

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.7190     6.7990  16.137  < 2e-16 ***
price        -7.6400     1.0459  -7.304 3.24e-10 ***
advert       12.1512     3.5562   3.417  0.00105 **
I(advert^2)  -2.7680     0.9406  -2.943  0.00439 **
```

0

```
> fval
   value    numdf    dendf
24.45932  3.00000 71.00000
```

y 'summary(mod1)' output")

Table: Tidy 'summary(mod1)' output

term	estimate	std.error	statistic	p.value
(Intercept)	109.719036	6.799046	16.137418	0.0000000
price	-7.640000	1.045939	-7.304442	0.0000000
advert	12.151236	3.556164	3.416950	0.0010516
I(advert^2)	-2.767963	0.940624	-2.942688	0.0043927

Glance

```
library(broom)
kable(tidy(mod1), caption="'Tidy(mod1)' output")
glance(mod1)$statistic #Retrieves the F-statistic
```

```
> glance(mod1)$statistic
   value
24.45932
```

Table: 'Tidy(mod1)' output

term	estimate	std.error	statistic	p.value
(Intercept)	109.719036	6.799046	16.137418	0.0000000
price	-7.640000	1.045939	-7.304442	0.0000000
advert	12.151236	3.556164	3.416950	0.0010516
I(advert^2)	-2.767963	0.940624	-2.942688	0.0043927

F= 24.45932 , $F_{(3,71)} = 2.734$, **Reject** H_0

至少***price***、***advertise***、$advertise^2$ 其中之一對營業額有影響

names

```r
names(glance(mod1)) #Shows what is available in 'glance'
kable(glance(mod1),
      caption="Function 'glance(mod1)' output", digits=2,
      col.names=(c("Rsq","AdjRsq","sig","F","pF","K",
                   "logL","AIC","BIC","dev","df.res“,”Nobs”)))
```

```
> names(glance(mod1)) #Shows what is available in 'glance
 [1] "r.squared"     "adj.r.squared" "sigma"         "statistic"     "p.value"       "df"
 [7] "logLik"        "AIC"           "BIC"           "deviance"      "df.residual"   "nobs"
```

```
Table: Function 'glance(mod1)' output


|  Rsq| AdjRsq|  sig|     F| pF|  K|    logL|    AIC|   BIC|     dev| df.res| Nobs|
|----:|------:|----:|-----:|--:|--:|-------:|------:|-----:|-------:|------:|----:|
| 0.51|   0.49| 4.65| 24.46|  0|  3| -219.55| 449.11| 460.7| 1532.08|     71|   75|
```

線性回歸的聯合檢定

$$Sales = \beta_1 + \beta_2 price + \beta_3\, advert + \beta_4\, advert^2 + e$$

利潤最大化條件：$\beta_3 + 2\beta_4 advert_0 = 1$

假設當$price$=6, $advert$=1.9,平均 $Sales$=80

$$H_0: \beta_3 + 2\beta_4 advert_0 = 1 \text{且} \beta_1 + 6\beta_2 + 1.9\beta_3 + 1.9^2\beta_4 = 80$$

Table: Joint hypotheses with the 'linearHypothesis' function

res.df	rss	df	sumsq	statistic	p.value
73	1779.860	NA	NA	NA	NA
71	1532.084	2	247.776	5.741229	0.0048847

03

Prediction and Forecasting

預測漢堡店的SALES

$$Sales = \beta_1 + \beta_2 price + \beta_3\, advert + \beta_4\, advert^2 + e$$

假設當$price$**=6,** $advert$**=1.9,**$advert^2 = 3.61$ $\widehat{SALES} = ?$

```r
predpoint <- data.frame(price=6, advert=1.9)
mod3 <- lm(sales~price+advert+I(advert^2), data=andy)
pre<-data.frame(predict(mod3, newdata=predpoint,interval="prediction"))
kable(pre, caption="Forecasting in the quadratic 'andy' model")
```

Table: Forecasting in the quadratic 'andy' model

fit	lwr	upr
76.97404	67.53258	86.41549

04

Omitted Variable Bias

假設 $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$

True model $y = \beta_1 + \beta_2 x_2 + u$

忽略變數偏誤

$$bias(b_2^*) = E(b_2) - \beta_2 = \beta_3 \frac{\widehat{cov(x_2, x_3)}}{var(x_2)}$$

$$\ln(FAMINC) = \beta_1 + \beta_2 HEDU + \beta_3 WEDU + e$$

Table: The incorrect model ('we' omitted)

term	estimate	std.error	statistic	p.value
(Intercept)	26191.270	8541.1084	3.066495	0.0023038
he	$\boxed{5155.483}$	658.4574	7.829639	0.0000000

05

Irrelevant Variables

加入 *xtra_x5* 和 *xtra_x6*

加入非相關變數

ln(FA

ln(FAMIN ... x6+e

Table: Correct 'faminc' model

term	estimate	std.error	statistic	p.value
(Intercept)	-7755.330	11162.9346	-0.6947393	0.4875992
he	3211.526	796.7026	4.0310216	0.0000658
we	4776.907	1061.1637	4.5015744	0.0000087
kl6	-14310.921	5003.9284	-2.8599372	0.0044466

Table: Incorrect 'faminc' with irrelevant variables

term	estimate	std.error	statistic	p.value
(Intercept)	-7558.6131	11195.411	-0.6751528	0.4999484
he	3339.7921	1250.039	2.6717496	0.0078378
we	5868.6772	2278.067	2.5761650	0.0103294
kl6	-14200.1839	5043.720	-2.8154190	0.0050996
xtra_x5	888.8426	2242.491	0.3963640	0.6920369
xtra_x6	-1067.1856	1981.685	-0.5385243	0.5904991

c)

variables"

06

Model Selection Criteria

R^2 、Adjusted R^2、AIC、SC

模型選取指標

$\bar{R}^2 = 1 - \frac{SSE\,/\,(N-K)}{SST\,/\,(N-1)}$

$AIC = ln\left(\frac{SSE}{N}\right) + \frac{2K}{N}$

$SC = ln\left(\frac{SSE}{N}\right) + \frac{K\,ln(N)}{N}$

```
mod1 <- lm(faminc~he, data=edu_inc)
mod2 <- lm(faminc~he+we, data=edu_inc)
mod3 <- lm(faminc~he+we+kl6, data=edu_inc)
mod4 <- lm(faminc~he+we+kl6+xtra_x5+xtra_x6, data=edu_inc)
r1 <- as.numeric(glance(mod1))
r2 <- as.numeric(glance(mod2))
r3 <- as.numeric(glance(mod3))
r4 <- as.numeric(glance(mod4))
tab <- data.frame(rbind(r1, r2, r3, r4))[,c(1,2,8,9)]
row.names(tab) <- c("he","he, we","he, we, kl6", "he, we, kl6, xtra_x5, xtra_x6")
kable(tab,caption="Model comparison, 'faminc' ", digits=4,
      col.names=c("Rsq","AdjRsq","AIC","BIC"))
```

tab

	X1	X2	X8	X9
r1	0.1258010	0.1237489	10316.65	10328.83
r2	0.1613004	0.1573536	10300.91	10317.15
r3	0.1771733	0.1713514	10294.73	10315.03
r4	0.1777965	0.1680547	10298.41	10326.82

各模型的選取指標

```
Table: Model comparison, 'faminc'

|                             |    Rsq|  AdjRsq|       AIC|       BIC|
|:----------------------------|------:|-------:|---------:|---------:|
|he                           | 0.1258|  0.1237|  10316.65|  10328.83|
|he, we                       | 0.1613|  0.1574|  10300.91|  10317.15|
|he, we, kl6                  | 0.1772|  0.1714|  10294.73|  10315.03|
|he, we, kl6, xtra_x5, xtra_x6 | 0.1778|  0.1681|  10298.41|  10326.82|
```

Mod1: FAMINC=β1 + β2*he

$\bar{R}^2 = 1 - \frac{SSE\,/\,(N-K)}{SST\,/\,(N-1)}$ $AIC = ln\left(\frac{SSE}{N}\right) + \frac{2K}{N}$ $SC = ln\left(\frac{SSE}{N}\right) + \frac{K\,ln(N)}{N}$

```
library(stats)
smod1 <- summary(mod1)
Rsq <- smod1$r.squared                   #R平方值
AdjRsq <- smod1$adj.r.squared            #調整R平方值
aic <- AIC(mod1)
bic <- BIC(mod1)
c(Rsq, AdjRsq, aic, bic)
```

```
> c(Rsq, AdjRsq, aic, bic)
[1] 1.258010e-01 1.237489e-01 1.031665e+04 1.032883e+04
```

Reset Test

```
mod3 <- lm(faminc~he+we+kl6, data=edu_inc)
resettest(mod3, power=2, type="fitted")
#Power 代表幾次函數應放入模型
resettest(mod3, power=2:3, type="fitted")
```

Quadratic

```
> resettest(mod3, power=2, type="fitted")

        RESET test

data:  mod3
RESET = 5.984, df1 = 1, df2 = 423, p-value = 0.01484
```

Quadratic cubic

```
> resettest(mod3, power=2:3, type="fitted")

        RESET test

data:  mod3
RESET = 3.1226, df1 = 2, df2 = 422, p-value = 0.04506
```

07

Collinearity

自變數X間線性重合

檢驗模型共線性

```
data("cars", package="PoEdata")
mod1 <- lm(mpg~cyl, data=cars)
kable(tidy(mod1), caption="A simple linear 'mpg' model")

mod2 <- lm(mpg~cyl+eng+wgt, data=cars)
kable(tidy(mod2), caption="Multivariate 'mpg' model")
tab <- tidy(vif(mod2)) #以VIF>10代表有共線性存在
kable(tab,caption="Variance inflation factors for the 'mpg' regression
      model",col.names=c("regressor","VIF"))
```

$$\text{VIF}=\frac{1}{1-R^2}$$

MPG= 44.37-0.26CYL-0.01ENG-0.05WGT

Table: Multivariate 'mpg' model

term	estimate	std.error	statistic	p.value
(Intercept)	44.3709616	1.4806851	29.9665086	0.0000000
cyl	-0.2677968	0.4130673	-0.6483126	0.5171663
eng	-0.0126740	0.0082501	-1.5362247	0.1252983
wgt	-0.0057079	0.0007139	-7.9951428	0.0000000

Table: Variance inflation factors for the 'mpg' regression model

regressor	VIF
cyl	10.515508
eng	15.786455
wgt	7.788716

VIF>10

Computer Exercise

Computer exercise

模型1：$PIZZA = \beta_1 + \beta_2 AGE + \beta_3 INCOME + \beta_4 (AGE * INCOME)$

(a) Test the hypothesis that age does not affect pizza expenditure—that is, test the joint hypothesis H0:b2 =0, b4 =0. What do you conclude?

R-code

```
library(PoEdata)
data(pizza4)
mod <- lm(pizza~age+income+age:income,data= pizza4)
summary(mod)
Hnull <- c("age=0", "age:income=0")
linearHypothesis(mod,Hnull)
```

(a)

```
Call:
lm(formula = pizza ~ age + income + age:income, data = pizza4)

Residuals:
    Min      1Q  Median      3Q     Max 
-200.86  -83.82   20.70   85.04  254.23 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) 161.46543  120.66341   1.338   0.1892  
age          -2.97742    3.35210  -0.888   0.3803  
income        6.97991    2.82277   2.473   0.0183 *
age:income   -0.12324    0.06672  -1.847   0.0730 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 127 on 36 degrees of freedom
Multiple R-squared:  0.3873,	Adjusted R-squared:  0.3363 
F-statistic: 7.586 on 3 and 36 DF,  p-value: 0.0004681
```

H0: AGE = 0
INCOME*AGE = 0

```
> linearHypothesis(mod,Hnull)
Linear hypothesis test

Hypothesis:
age = 0
age:income = 0

Model 1: restricted model
Model 2: pizza ~ age + income + age:income

  Res.Df    RSS Df Sum of Sq      F   Pr(>F)   
1     38 819286                                
2     36 580609  2    238677 7.3995 0.002033 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(b) Construct point estimates and 95% interval estimates of the marginal propensity to spend on pizza for individuals of ages 20, 30, 40, 50, and 55. Comment on these estimates.

R-code

$$MPS = {}^{dpizza}/_{dIncome} = \beta_3 + \beta_4 AGE$$

```
age<-c(20,30,40,50,55)
MPS<- mod$coefficients[3]+mod$coefficients[4]*age
df <- df.residual(mod)
alpha<-0.05
tcr1 <- qt(1-(alpha/2),df)
seMPS <-sqrt(vcov(mod)[3,3]+(age^2)*vcov(mod)[4,4]+2*age*vcov(mod)[4,3])
```

(b)

```
upper<- MPS+tcr1*seMPS
lower<- MPS-tcr1*seMPS
Interval<-cbind(MPS,seMPS,lower,upper)

rownames(Interval)<-
c("AGE=20","AGE=30","AGE=40“,
   "AGE=50","AGE=55")

colnames(Interval)<-c("Point
estimates","standard errors","Lower","Upper")
kable(Interval)
```

95%CI of

$$\beta_3 + \beta_4 AGE$$

	Point estimates	standard errors	Lower	Upper
AGE=20	4.5151180	1.5203944	1.4316153	7.598621
AGE=30	3.2827245	0.9048794	1.4475441	5.117905
AGE=40	2.0503310	0.4650721	1.1071211	2.993541
AGE=50	0.8179375	0.7099684	-0.6219452	2.257820
AGE=55	0.2017408	0.9908536	-1.8078035	2.211285

(c) Modify the equation to permit a ''life-cycle'' effect in which the marginal effect of income on pizza expenditure increases with age, up to a point, and then falls. Do so by adding the term (AGE^2 × INC) to the model. What sign do you anticipate on this term? Estimate the model and test the significance of the coefficient for this variable. Did the estimate have the expected sign?

R-code

$$\text{模型2:}PIZZA = \beta_1 + \beta_2 AGE + \beta_3 INCOME + \beta_4(AGE * INCOME) + \beta_5(AGE^\wedge 2 * INCOME) + e$$

```
mod2<-lm(pizza~age+income+age:income+income:I(age^2),data= pizza4)
summary(mod2)
Hnull2 <- c("income:I(age^2)=0")
linearHypothesis(mod2,Hnull2)
```

(c)

```
Call:
lm(formula = pizza ~ age + income + age:income + income:I(age^2),
    data = pizza4)

Residuals:
     Min       1Q   Median       3Q      Max
-212.080  -79.979    7.395   81.429  260.074

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     109.720767 135.572473   0.809    0.424
age              -2.038273   3.541904  -0.575    0.569
income           14.096163   8.839862   1.595    0.120
age:income       -0.470371   0.413908  -1.136    0.264
income:I(age^2)   0.004205   0.004948   0.850    0.401

Residual standard error: 127.5 on 35 degrees of freedom
Multiple R-squared:  0.3997,    Adjusted R-squared:  0.3311
F-statistic: 5.826 on 4 and 35 DF,  p-value: 0.001057
```

H0:INCOME*AGE^2=0

```
> linearHypothesis(mod2,Hnull2)
Linear hypothesis test

Hypothesis:
income:I(age^2) = 0

Model 1: restricted model
Model 2: pizza ~ age + income + age:income + income:I(age^2)

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     36 580609
2     35 568869  1     11740 0.7223 0.4012
```

(d) Using the model in (c), construct point estimates and 95% interval estimates of the marginal propensity to spend on pizza for individuals of ages 20, 30, 40, 50 and 55. Comment on these estimates. In light of these values, and of the range of age in the sample data, what can you say about the quadratic function of age that describes the marginal propensity to spend on pizza?

$$\text{MPS} = {dpizza}/_{dIncome}$$

$$= \beta_3 + \beta_4 AGE + \beta_5 AGE\text{^}2$$

R-code

```
mps2<-
mod2$coefficients[3]+mod2$coefficients[4]*age+mod2$coefficients[5]*age^2
df2 <- df.residual(mod2)
tcr2 <- qt(1-(alpha/2),df2)
seMPS2<-
   sqrt(vcov(mod2)[3,3]+(age^2)*vcov(mod2)[4,4]+(age^4)*vcov(mod2)[5,5]+
         2*age*vcov(mod2)[3,4]+2*age^2*vcov(mod2)[3,5]+
         2*age^3*vcov(mod2)[4,5])
```

(d)

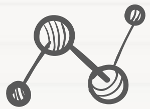
R-code

```
upper<- mps2+tcr2*seMPS2
lower<- mps2-tcr2*seMPS2
Interval2<-cbind(mps2,seMPS2,lower,upper)
rownames(Interval2)<-
  c("AGE=20","AGE=30","AGE=40",
    "AGE=50","AGE=55")
colnames(Interval2)<-
  c("Point estimates","standard
    errors","Lower","Upper")
kable(Interval2)
```

$$MPS = {}^{dpizza}/_{dIncome}$$

$$= \beta_3 + \beta_4 AGE + \beta_5 AGE^2$$

MPS in (d)



	Point estimates	standard errors	Lower	Upper
AGE=20	6.3706583	2.6639225	0.962608	11.778709
AGE=30	3.7693368	1.0737844	1.589438	5.949235
AGE=40	2.0089691	0.4694063	1.056024	2.961914
AGE=50	1.0895552	0.7811004	-0.496163	2.675273
AGE=55	0.9452059	1.3246506	-1.743978	3.634390

MPS in (b)

	Point estimates	standard errors	Lower	Upper
AGE=20	4.5151180	1.5203944	1.4316153	7.598621
AGE=30	3.2827245	0.9048794	1.4475441	5.117905
AGE=40	2.0503310	0.4650721	1.1071211	2.993541
AGE=50	0.8179375	0.7099684	-0.6219452	2.257820
AGE=55	0.2017408	0.9908536	-1.8078035	2.211285

(e) For the model in part(c),are each of the coefficient estimates for AGE, (AGE × INC)and(AGE^2 × INC) significantly different from zero at a 5% significance level ? Carry out a joint test for the significance of these variables. Comment on your results.

R-code

```
Hnull3 <- c("age=0", "age:income=0","income:I(age^2)=0")
linearHypothesis(mod2,Hnull3)
```

HO:

- AGE=0
- INCOME*AGE=0
- INCOME*AGE^2=0

```
> linearHypothesis(mod2,Hnull3)
Linear hypothesis test

Hypothesis:
age = 0
age:income = 0
income:I(age^2) = 0

Model 1: restricted model
Model 2: pizza ~ age + income + age:income + income:I(age^2)

  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     38 819286
2     35 568869  3    250417 5.1357 0.004763 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(f) Check the model used in part (c) for collinearity. Add the term (AGE^3×INC) to the model in (c) and check the resulting model for collinearity.

$$\text{模型3:PIZZA} = \beta_1 + \beta_2 AGE + \beta_3 INCOME + \beta_4(AGE * INCOME) + \beta_5(AGE\hat{}2 * INCOME) + \beta_6(AGE\hat{}3 * INCOME) + e$$

R-code

```
mod3<-lm(pizza~age+income+age:income+income:I(age^2)+I(age^3):income
           ,data= pizza4)
library(Hmisc)
mydata<- data.frame(pizza4$age,pizza4$income,(pizza4$age*pizza4$income)
           ,(pizza4$age^2*pizza4$income),(pizza4$age^3*pizza4$income))
colnames(mydata)<-c("AGE","INC","AGE*INC","AGE2*INC","AGE3*INC")
res2 <- rcorr(as.matrix(mydata))
tab <- tidy(vif(mod3))
kable(tab,caption="Variance inflation factors for the 'mpg' regression model",
      col.names=c("regressor","VIF"))
```

(f)

檢驗共線性

1. 相關係數法

```
> res2
          AGE  INC AGE*INC AGE2*INC AGE3*INC
AGE      1.00 0.47    0.59     0.65     0.69
INC      0.47 1.00    0.98     0.94     0.90
AGE*INC  0.59 0.98    1.00     0.99     0.96
AGE2*INC 0.65 0.94    0.99     1.00     0.99
AGE3*INC 0.69 0.90    0.96     0.99     1.00

n= 40


P
         AGE    INC    AGE*INC AGE2*INC AGE3*INC
AGE             0.0023 0.0000  0.0000   0.0000
INC      0.0023        0.0000  0.0000   0.0000
AGE*INC  0.0000 0.0000         0.0000   0.0000
AGE2*INC 0.0000 0.0000 0.0000           0.0000
AGE3*INC 0.0000 0.0000 0.0000  0.0000
```

2. 變異數膨脹因子法

```
Table: Variance inflation factors for the 'mpg' regression model

|regressor          |          VIF|
|:------------------|------------:|
|age                | 5.746482e+00|
|income             | 5.854281e+03|
|age:income         | 7.290751e+04|
|income:I(age^2)    | 1.020606e+05|
|income:I(age^3)    | 1.603378e+04|
```

VIF>10

補充

主成分分析方法

Principal **C**omponent **A**nalysis

解決自變數間共線性的問題

PCA的目的

- 維度縮減：新變數為原變數的線性組合
- 獨立性：主成分之間彼此不相關
- 代表性：主成分能解釋原變數的最大變異程度

原始變數
X1, X2, X3...XN

線性降維

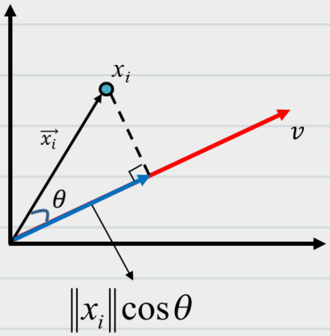
新變數
X1*=f(x1, x2..xn)

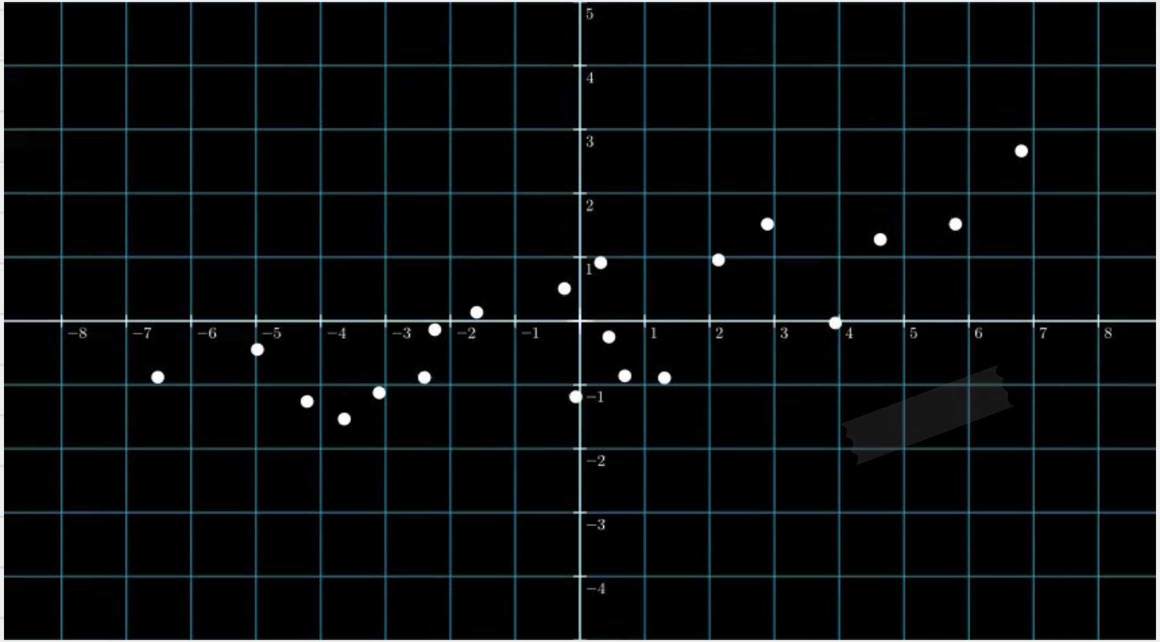
數學的本質不是將簡單的事情變複雜，而是將複雜的事物簡化。
—Stan Gudder

什麼是線性降維(一)？

- 將原始數據拆解成更具代表性的主成分，並以其作為新的基準由此獲得更能描述數據本質的新成分。

投影

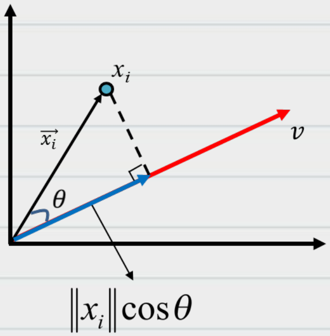


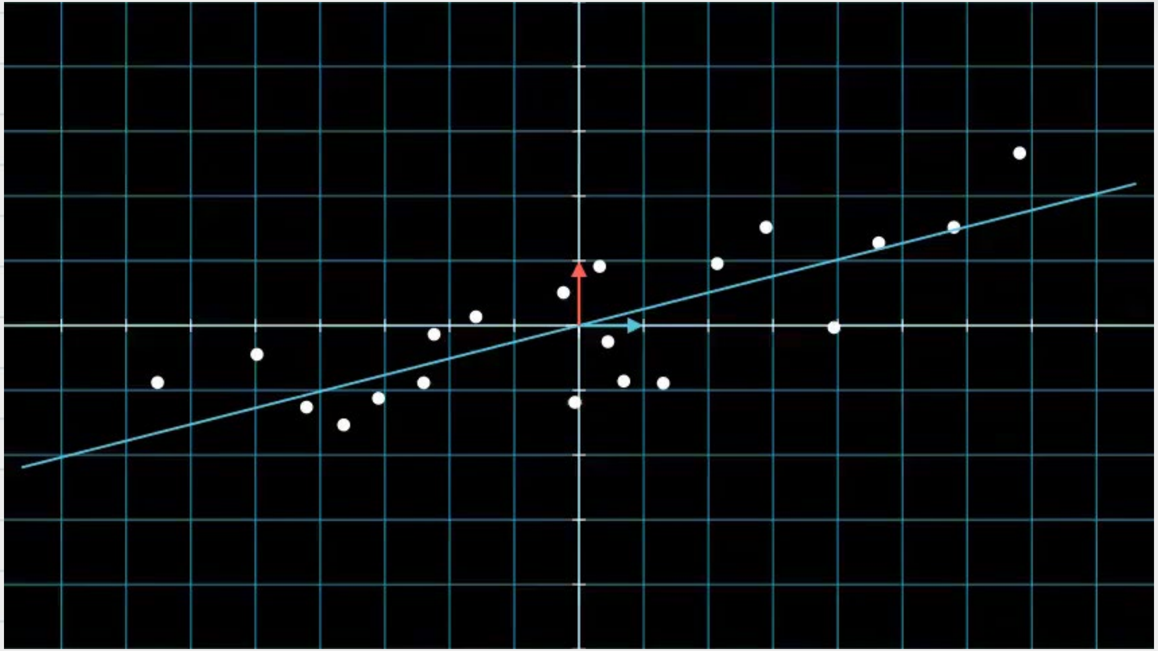


什麼是線性降維(二)?

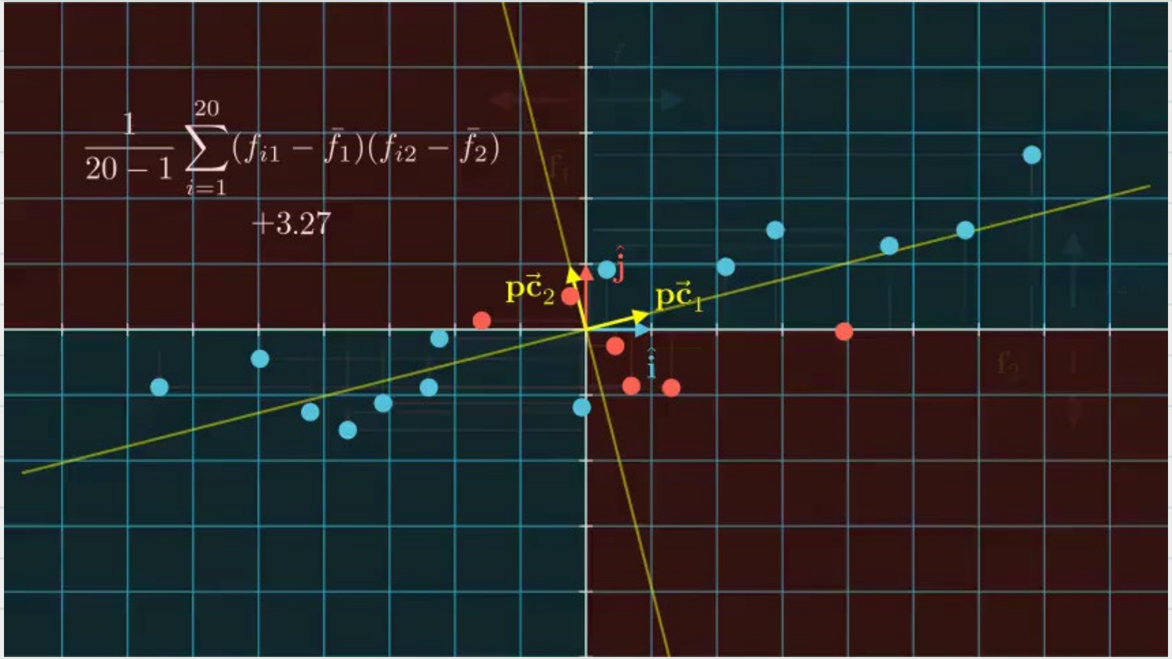
- 將原始數據拆解成更具代表性的主成分，並以其作為新的基準
  由此獲得更能描述數據本質的新成分。

投影





透過轉軸找到變異程度最大的comp



橫看成嶺側成峰 遠近高低各不同

新變數與原變數的線性組合

第一個主成分(PC1)到第n個主成分(PCn)可透過以下公式表示：

$$PC_1 = \Phi_{11}X_1 + \Phi_{12}X_2 + \Phi_{13}X_3 \ldots\ldots \Phi_{1n}X_n$$

$$PC_2 = \Phi_{21}X_1 + \Phi_{22}X_2 + \Phi_{23}X_3 \ldots\ldots \Phi_{2n}X_n$$

$$\vdots$$

$$PC_n = \Phi_{n1}X_1 + \Phi_{n2}X_2 + \Phi_{n3}X_3 \ldots\ldots \Phi_{nn}X_n$$

1. Φ為每一個主成分的特徵向量
2. X為原變數的數值
3. PC為新變數數值(主成份計分)

主成分分析步驟

Step 1
資料維度標準化

Step 2
計算共變數矩陣

Step 3
計算矩陣的特徵值
與特徵向量

Step 4
萃取主成分

如何決定主成份個數?

累積解釋變異>70%

陡坡圖(Scree plot)
找開始平坦的點

特徵值(eigenvalue)>=1

決定新變數個數

Rcode

```
data(iris) #print data
iris$Species<-as.numeric(iris$Species)
dat<- scale(iris) #standardized data
pca<- princomp(dat, cor=F)
pca #eigenvalue
summary(pca) #summary eigenvalue
pca$loadings #eigenvector
plot(pca,type="line")
print(-1*pca$loadings, digits=8, cutoff=0)
-1*pca$scores #principal components scores
cor(cbind(-1*pca$scores,dat), method='pearson') #loading
biplot(pca)
```

Sepal length: 花萼長度(cm)
Sepal width: 花萼寬度(cm)
Petal length: 花瓣長度(cm)
Petal width: 花瓣寬度(cm)
Species:花種

Eigenvalue

```
> pca #eigenvalue
Call:
princomp(x = dat, cor = F)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
1.9522904 0.9529124 0.4300984 0.2044639 0.1428167
```

累積變異

```
> summary(pca) #summary eigenvalue
Importance of components:
                          Comp.1    Comp.2     Comp.3      Comp.4
Standard deviation     1.9522904 0.9529124 0.43009840 0.204463940
Proportion of Variance 0.7674036 0.1828273 0.03724523 0.008417215
Cumulative Proportion  0.7674036 0.9502309 0.98747608 0.995893297
                            Comp.5
Standard deviation     0.142816749
Proportion of Variance 0.004106703
Cumulative Proportion  1.000000000
```

>70%

Eigenvector

```
> pca$loadings #eigenvector

Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
Sepal.Length  0.445  0.382  0.751  0.141  0.270
Sepal.Width  -0.233  0.921 -0.287        -0.122
Petal.Length  0.506               -0.243 -0.827
Petal.Width   0.497        -0.385 -0.613  0.474
Species       0.495        -0.452  0.739

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
SS loadings       1.0    1.0    1.0    1.0    1.0
Proportion Var    0.2    0.2    0.2    0.2    0.2
Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

代回

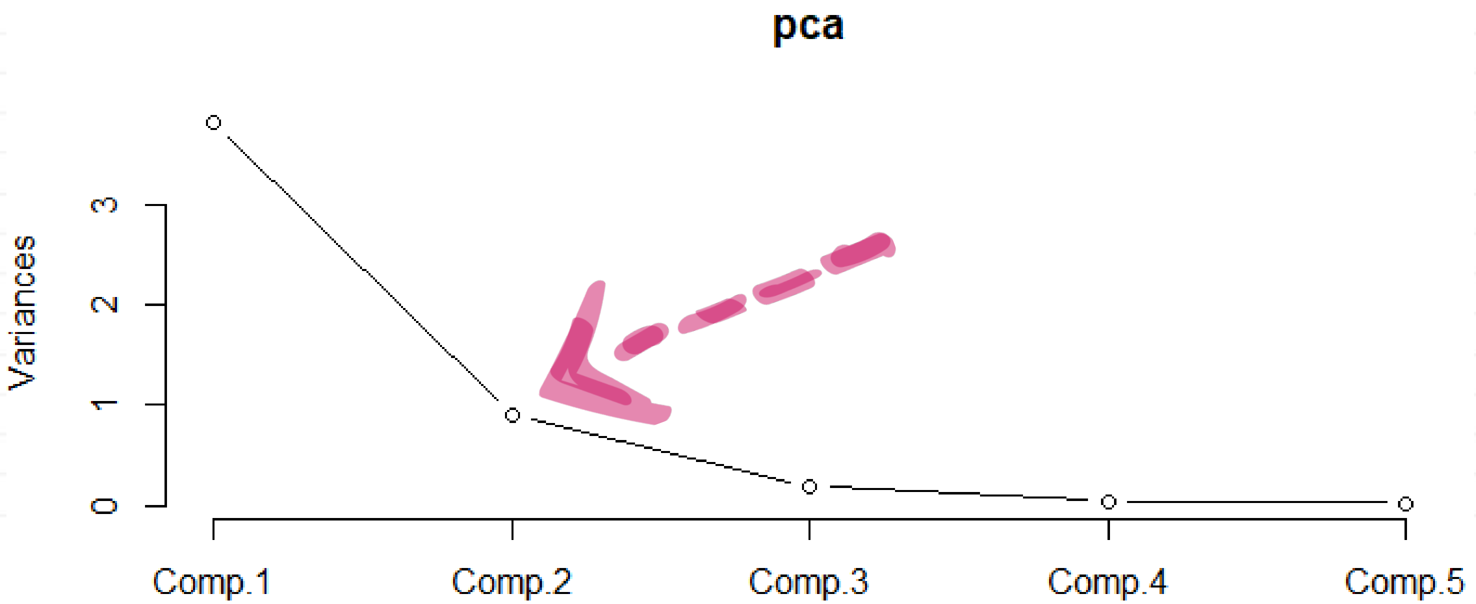
$$PC_1 = \Phi_{11}X_1 + \Phi_{12}X_2 + \Phi_{13}X_3 \ldots\ldots \Phi_{1n}X_n$$

$$PC_2 = \Phi_{21}X_1 + \Phi_{22}X_2 + \Phi_{23}X_3 \ldots\ldots \Phi_{2n}X_n$$

$$\vdots$$

$$PC_n = \Phi_{n1}X_1 + \Phi_{n2}X_2 + \Phi_{n3}X_3 \ldots\ldots \Phi_{nn}X_n$$

運用陡坡圖找軸點



Comp1

$$=0.445*X1+0.233*X2+0.506*X3+0.497*X4+0.495*X5$$

Comp2

$$=0.382*X1+0.921*X2$$

將原變數的資料
代回萃取的
Component

Sepal length: 花萼長度(cm)
Sepal width: 花萼寬度(cm)
Petal length: 花瓣長度(cm)
Petal width: 花瓣寬度(cm)
Species:花種

最終算出主成分計分(新變數)

```
> -1*pca$scores #principal components scores
```

	Comp.1	Comp.2	Comp.3	Comp.4
[1,]	2.56751880	-0.47291496	-0.0541826346	-0.102879414
[2,]	2.40725795	0.67582788	-0.2024022408	-0.064745216
[3,]	2.65045330	0.34711905	0.1123050460	-0.046022170
[4,]	2.59330274	0.60129156	0.1338393753	-0.000626848
[5,]	2.67478340	-0.63808276	0.1023683826	-0.086633559
[6,]	2.40413896	-1.47990450	0.0338055866	0.044779262
[7,]	2.71740435	-0.04023953	0.3837215969	0.063552647
[8,]	2.53903439	-0.21711569	-0.0310430528	-0.071252756
[9,]	2.62233233	1.11777041	0.1849342960	0.021320477
[10,]	2.49732208	0.47217129	-0.1886720964	-0.132148990
[11,]	2.48476604	-1.03552293	-0.1959254810	-0.141878663
[12,]	2.61781457	-0.12648421	0.1486475461	-0.023380243
[13,]	2.52616995	0.73122410	-0.1623220098	-0.128059660
[14,]	2.88078420	0.96675001	0.2957333758	-0.084106030
[15,]	2.51649613	-1.84904988	-0.3559921518	-0.253810563
[16,]	2.56800894	-2.67152292	0.0945877720	-0.037944189
[17,]	2.51880354	-1.47339742	0.0402265962	-0.010295394
[18,]	2.50232997	-0.48218205	-0.0036149399	-0.022513045
[19,]	2.25461687	-1.39772785	-0.3546089766	-0.085936810
[20,]	2.63428814	-1.11769958	0.1924890822	-0.011162591
[21,]	2.26680945	-0.40488571	-0.3968452603	-0.111923125
[22,]	2.51555787	-0.91566976	0.1771536854	0.070009847
[23,]	3.00434063	-0.44705920	0.4713810949	-0.073500518
[24,]	2.17887101	-0.08300271	-0.0391014909	0.181137824

變數縮減

5>>2

參考資料

世上最生動的PCA:
直觀理解並應用主成分分析

R筆記-
(7)主成份分析(2012美國職棒MLB)

Principal Components Analysis (PCA)
| 主成份分析 | R 統計

多變量分析
Applied Multivariate Techniques
by Subhash Sharma

Thanks for listening