Started on	Tuesday, 8 November 2022, 10:44 PM
State	Finished
Completed on	Monday, 14 November 2022, 8:55 PM
Time taken	5 days 22 hours
Grade	**30.50** out of 33.00 (**92.43**%)

Information

Declaration

The University regards plagiarism as a form of academic misconduct, and has very strict rules regarding plagiarism. See UNSW policies, penalties, and information to help you avoid plagiarism, as well as the guidelines in the online ELISE tutorials for all new UNSW students.

By submitting this assessment item, you declare that this assessment item is your own work, except where acknowledged, and acknowledge that the assessor of this item may, for the purpose of assessing this item:

1. Reproduce this assessment item and provide a copy to another member of the University; and/or
2. Communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking)

You certify that you have read and understood the University Rules with respect to Student Conduct.

Information

Software

You may use any software of your choice to help you in this assessment. Use of R and computer algebra systems is permitted and encouraged. You are not required to derive exact answers where numerical answers are requested.

However, any code you use to get your results must be submitted as a part of your working.

Information

Data on $n = 20$ consecutive years have been collected reflecting annual average prices of beef steers X_1 and of hogs X_2 in dollars per pound and the annual per capita consumption of beef X_3 and of pork X_4 in pounds. We are interested in the relationship of livestock prices to meat production.

The file `price-cons.csv` contains the variables Y (year index) and X_1, X_2, X_3, X_4. We could proceed by calculating $U = (X_1 + X_2)/2, V = X_3 + X_4$ and then regressing U on V.

A perhaps better procedure would be to construct a (weighted) price index $U = a_1X_1 + a_2X_2$ and consumption index $V = b_3X_3 + b_4X_4$ and to look at the maximal correlation between U and V. This is the canonical correlation analysis approach.

(a)i Recall that when performing canonical correlation, we may use raw coefficients or standardised coefficients. We obtain the latter by first standardising (dividing by its standard deviation) each variable. In this case, the prices are in dollar units but the consumption is in pounds. Does it make sense to standardise here?

Select one:

- Yes
- No ✔

Your answer is correct.

Although the four variables are not all on the same scale, within each pair of variables (consumption and prince) the scale is the same.

The correct answer is: No

Question 2

Complete

Mark 0.00 out of 1.00

(a)ii Why, or why not? (Give a 1-2 sentence explanation.)

Standardisation only makes sense when we're comparing variables within the same group. In this case, we're not - we're comparing a group of prices to a group of masses (in pounds).

Although the two sets of variables are on different scales, each set of variables is on the same scale, so there is no need to standardise.

Comment:

(b)i Perform a canonical correlation analysis of the type you have selected above. Report the following:

- The first (0.8466626 ✔) and the second (0.7291162 ✔) canonical correlations.
- The estimated weighted price index $U =$ 0.4848242 ✔ X_1+ 0.4095221 ✔ X_2+ 0 ✔ X_3+ 0 ✔ X_4 and consumption index $V =$ 0 ✔ X_1+ 0 ✔ X_2+ -0.05040897 ✔ X_3+ -0.14820099 ✔ X_4.
- The p-value for the permutation test of the null hypothesis that price of meat is not linearly associated with its consumption: 0 ✔ .

The code I used for this analysis is straightforward:

```
library(CCA)
meat <- read.csv("price-cons.csv")
(meat.cc <- with(meat, cc(cbind(steer_price, hog_price),
                          cbind(beef_consumption, pork_consumption))))
```

```
# or
```

```
(meat.cc <- with(meat, cancor(cbind(steer_price, hog_price),
                              cbind(beef_consumption, pork_consumption))))
```

```
library(CCP)
p.perm(meat[,2:3], meat[,4:5]) # Other tests also accepted.
```

- The canonical correlations are 0.847 and 0.729, regardless of scaling.
- U is the price index, so X_3 and X_4 are not a part of it, and vice versa for V. Their coefficients are therefore set to zero. The correct answers are, therefore,
  - $U = -0.111X_1 + -0.094X_2 + 0X_3 + 0X_4$
    or
    $U = 0.485X_1 + 0.410X_2 + 0X_3 + 0X_4$.
  - $V = 0X_1 + 0X_2 + 0.015X_3 + 0.034X_4$
    or
    $V = 0X_1 + 0X_2 + -0.050X_3 + -0.148X_4$.

▲ Back to Top

Question 4

Correct

Mark 1.00 out of 1.00

(b)ii Based on the appropriate hypothesis test, is only one canonical variable pair enough to explain the correlations in the data?

Select one:

- Yes
- No ✔

Your answer is correct.

Based on the second ("2-2") hypothesis p-value, the second canonical correlation is highly significant at any reasonable α, so one correlation is not enough.

The correct answer is: No

(b)ii (continued) What hypothesis did you test to answer the above question? (Write down the H_0 and H_1, and paste in the relevant (and *only* the relevant) R output.)

H0: Σ12 = 0, H1: Σ12 /= 0 (1 and 2 denote the variable sets)

Wilks' Lambda, using F-approximation (Rao's F):

stat approx df1 df2 p.value

1 to 2: 0.1326303 13.96689 4 32 1.069829e-06

2 to 2: 0.4683896 19.29457 1 17 3.974817e-04

If ρ_1 and ρ_2 are the first and second canonical correlations, then it's $H_0 : \rho_1 = \text{anything}, \rho_2 = 0$ vs $H_1 : \rho_1 = \text{anything}, \rho_2 \neq 0$, and we look at the second row of the asymptotic tests. E.g.,

```
> p.asym(meat.cc$cor, nrow(meat), 2,2)
Wilks' Lambda, using F-approximation (Rao's F):
              stat    approx df1 df2      p.value
1 to 2:  0.1326303 13.96689   4  32 1.069829e-06
2 to 2:  0.4683896 19.29457   1  17 3.974817e-04
```

Since the p-value is very small, we see that there is evidence that $\rho_2 \neq 0$ and one correlation is not enough. Alternatively, other test statistics could be used:

```
> p.asym(meat.cc$cor, nrow(meat), 2,2, tstat="Hotelling")
 Hotelling-Lawley Trace, using F-approximation:
             stat    approx df1 df2      p.value
1 to 2:  3.666517 13.74944   4  30 1.759234e-06
2 to 2:  1.134975 19.29457   1  34 1.037369e-04
> p.asym(meat.cc$cor, nrow(meat), 2,2, tstat="Pillai")
 Pillai-Bartlett Trace, using F-approximation:
              stat    approx df1 df2      p.value
1 to 2:  1.2484480 14.11986   4  34 6.897527e-07
2 to 2:  0.5316104 13.75738   1  38 6.624381e-04
```

Comment:

(c) In microeconomics, two goods (such as butter and jam) are called *complementary* if they are often consumed together, or if consumption of one improves the value of consuming the other. Conversely, two goods (such as tea and coffee) are called *substitutes* if they can (as the name suggests) take each other's place.

A quick way to check empirically if two goods are complementary, substitutes, or neither is to look at the correlation in their consumption. However, such a correlation may be spurious, because, for example, the goods have common production inputs and so their prices fluctuate together, creating a positive correlation in consumption. We would thus want to control for their price.

i What is the correlation of beef and pork consumption, after controlling for their prices?

-0.1881614

✔

ii Is there sufficient evidence to believe that it is different from 0? Report the p-value for the appropriate test:

0.4546362

✔ .

iii Based on the appropriate linear regression, and using $\alpha = 0.05$, an increase in the price of a steer will [decrease] ✔ the consumption of beef and will [decrease] ✔ the consumption of pork, whereas an increase in the price of a hog will [increase] ✔ the consumption of beef and will [decrease] ✔ the consumption of pork.

i

I used:

```
library(ggm)
pcor(c(3,4,1,2), cov(meat[2:5]))
```

and got -0.1881614. Notice that consumption variables are the second two (i.e., 4 and 5), so we need to list them as the first two and then the conditioned-on price variables. (Partial marks were given for switching them.)

ii

Either the *t*-test or the test based on the Fisher's *Z* transformation is acceptable. We are conditioning on two variables, so:

```
pcor.test(pcor(c(3,4,1,2), cov(meat[2:5])), 2, nrow(meat))
```

gives the p-value 0.455, or get the z score (standardised) and compute the p-value as follows:

```
r <- pcor(c(3,4,1,2), cov(meat[2:5]))
z <- (log((1+r)/(1-r))-log((1+0)/(1-0)))/2*sqrt(nrow(meat)-3-2)
pnorm(-abs(z))*2
```

for a p-value of 0.461. (**Note:** An earlier version has a typo in the code that resulted in the wrong p-value. Assessments have been regraded.)

iii

Here's the regression and output I get (with extraneous output trimmed):

```
> summary(lm(cbind(beef_consumption, pork_consumption)~cbind(steer_price, hog_price),data=meat))
Response beef_consumption :
[ ... ]
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                66.3336     6.1292  10.822  4.8e-09 ***
cbind(steer_price, hog_price)steer_price  -2.2535     0.5839  -3.859  0.00126 **
cbind(steer_price, hog_price)hog_price     1.2252     0.4363   2.808  0.01209 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[ ... ]
Response pork_consumption :
[ ... ]
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               107.6163     7.4913  14.365 6.14e-11 ***
cbind(steer_price, hog_price)steer_price  -2.0033     0.7136  -2.807   0.0121 *
cbind(steer_price, hog_price)hog_price    -2.7563     0.5332  -5.169 7.71e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

▲ Back to Top

- hog_price $\rightarrow$ beef_ consuption : postive
- hog_price $\rightarrow$ pork_ consuption : negative

Information

Unless otherwise stated, multiple choice and True/False questions in the following section have a *guessing penalty*. This means that a correct answer gains you 1 mark, but an incorrect answer *loses* you around $1/(\# \text{ choices} - 1)$ marks. This means that if guessing randomly, your *expected* mark will be 0.

Do **not** let it prevent you from answering questions, because unless you are completely randomly guessing, you are far more likely to gain marks by answering than to lose them.

Question 7

Correct

Mark 2.00 out of 2.00

Standardising the variables before performing canonical correlation analysis affects which of the following?

A. the coefficients to construct the variates (i.e., the $\boldsymbol{a}$ in $Z_1 = \boldsymbol{a}^\top \boldsymbol{X}_{(1)}$ and similarly for $\boldsymbol{X}_{(2)}$)

B. the canonical correlations

C. the conclusions of the tests for the significance of the canonical correlations

More marks will be given the closer you are to the correct answer.

Select one:

- None
- A only ✔
- B only
- C only
- A and B only
- A and C only
- B and C only
- A, B, and C

Your answer is correct.

The correct answer is: A only

Back to Top

True or False? Unlike ordinary correlation, canonical correlation can detect and represent nonlinear relationships between variables.

Select one:

- True
- False ✔

Your answer is correct.

The correct answer is: False

Information

Instructions: Please read carefully

The following instructions apply to all CodeRunner questions.

1. Into the space provided, write R code that (ultimately) assigns to each of these variables the answer to the question.
   - You may either implement your answer in the space provided, or you may work elsewhere, and only provide your final answers as constants.
   - Note that not all R packages may be available in the testing environment.
2. Click "Precheck" to check that your answers are formatted correctly, and address any issues.
   - Passing tests get marked with ✔, failing tests with ✘.
   - **IMPORTANT:** If your code fails the first pre-check (i.e., code produces errors when executing), your assessment **cannot** be marked, and you may receive 0 **for the whole question**.
3. The code will be submitted when you submit the quiz.
4. When you submit the quiz, you **may** get a weird error message along the lines of "Expected 0 test results, got 1." Please disregard it.

Suppose that random vector $\boldsymbol{X} \in \mathbb{R}^4$ has mean $\boldsymbol{\mu}$ and variance-covariance matrix

$$\Sigma = \begin{bmatrix} 177 & 40 & -14 & 4 \\ 40 & 98 & -37 & -4 \\ -14 & -37 & 314 & 5 \\ 4 & -4 & 5 & 1 \end{bmatrix}.$$

For your convenience, you can construct it in R by running the following code:

```
Sigma <- matrix(c(177, 40, -14, 4, 40, 98, -37, -4, -14, -37, 314, 5, 4, -4, 5, 1), 4, 4)
```

(c)

(i)

Calculate the two canonical correlations between (X_1, X_2) and (X_3, X_4). Store their values (in order: first, then second) in a vector `ans_c_i`.

(ii)

Suppose that the Σ above is actually $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$, for a random sample of size 30. Assuming multivariate normality, obtain the Wilks's Lambda test statistic and the p-value for the null hypothesis that sets of variables (X_1, X_2) and (X_3, X_4) have no linear association between them. Store the result in variables `ans_c_ii_lambda` and `ans_c_ii_pval`, respectively.

Answer: (penalty regime: 100 %)

Reset answer

```
Sigma <- matrix(c(177, 40, -14, 4, 40, 98, -37, -4, -14, -37, 314,

ans_c_i <- c(0.6011624, 0.1590759)

ans_c_ii_lambda <- 3.477574
ans_c_ii_pval <- 0.01365727

```

	Criterion	Result	Mark	
✔	Quiz results available?	Yes	0.00/0.00	✔
✔	(c)i is at least partially correct	Yes	1.00/1.00	✔
✔	(c)i is completely correct	Yes	1.00/1.00	✔
✘	(c)ii lambda test statistic is correct (or consistent with (c)i)	Correct answer: 0.622443795249587 See Sec. 7.2 for motivation and 8.3.1 for the formulas.	0.00/0.50	✘
✔	(c)ii p-value is correct (or consistent with (c)i)	Yes	0.50/0.50	✔

I will go over the details in lecture, but here is an outline of how to approach the questions. Your specific checklist will also provide some directions.

(c)

The expressions are given in the lecture. Essentially, you needed to compute $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$ (or vice versa with partitions 1 and 2, or the version without square roots), take its eigenvalues, and take their square roots (common mistake) to get the correlations.

▲ Back to Top

```
TE: These are for validation purposes only. These are *not* the cor

a <- matrix(c(177, 40, -14, 4, 40, 98, -37, -4, -14, -37, 314, 5, 4

c_i <- c(0.5500807, 0.2616556)

c_ii_lambda <- 0.6496639
c_ii_pval <- 0.0383037

```

Partially correct

Marks for this submission: 5.00/6.00.

Back to Week

Previous Activity

Next Activity

Back to Top