Dragos Ailoae

Applied Statistics and Econometrics I

2021-11-11

Homework 3 - Suggested Solutions

Problem 1 (3 points)

A multiple regression of y on a constant, x_1, and x_2 produces the following results:

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 5 & 15 & 25 \\ 15 & 55 & 81 \\ 25 & 81 & 129 \end{bmatrix} \quad , \quad \boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} 20 \\ 76 \\ 109 \end{bmatrix}$$

$$RSS = 1.5, ESS = 26.5, n = 5, R^2 = 0.95$$

a. Calculate (i) the adjoint and determinant of $\boldsymbol{X}'\boldsymbol{X}$

In [1]:
```
options(warn=-1) # Suppress warnings
qui <- suppressPackageStartupMessages # quiet! - suppress library load messages

# install (if not already installed) and load package
qui(if(!require(RConics)){install.packages('RConics')})  # for adjoint()
XpX <- matrix(c(  5, 15,  25,
                 15, 55,  81,
                 25, 81, 129),
                 byrow=T,nrow = 3)

print("Determinant (XpX)")
print(det(XpX))
print("Adjoint (XpX)")
adjoint(XpX)
```

```
[1] "Determinant (XpX)"
[1] 20
[1] "Adjoint (XpX)"
```

534	90	-160
90	20	-30
-160	-30	50

a. Calculate (ii) regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$.

In [2]:
```
# method 1
XpX.inverse <- 1 / det(XpX) * adjoint(XpX)
print("Inverse (XpX) method 1 (adjoint)")
XpX.inverse

# method 2
XpX.inverse <- solve(XpX)
print("Inverse (XpX) method 2 (direct)")
XpX.inverse
XpY <- matrix(c( 20,
                 76,
                109),
                 byrow=T,nrow = 3)

# calculate betas
beta.hat <- XpX.inverse %*% XpY
print("Estimated coefficients vector:")
beta.hat
```

```
[1] "Inverse (XpX) method 1 (adjoint)"
```

26.7	4.5	-8.0
4.5	1.0	-1.5
-8.0	-1.5	2.5

```
[1] "Inverse (XpX) method 2 (direct)"
 26.7   4.5  -8.0
  4.5   1.0  -1.5
 -8.0  -1.5   2.5
[1] "Estimated coefficients vector:"
 4.0
 2.5
-1.5
```

b. Compute a 95% confidence interval for β_1.

The sampling variance of the coefficients is $\text{Est. Var}\left[\hat{\beta} \mid \mathbf{X}\right] = s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ where $s^2 = \frac{e'e}{n-K}$ is an unbiased estimator for σ_ϵ^2.

In [3]:
```
n <- 5
K <- length(beta.hat)
RSS <- 1.5
paste0("K= ", K)
s.squared <- RSS / (n-K)
var.beta <- s.squared * XpX.inverse
paste0("Estimated coefficients VCM:")
var.beta
```

'K= 3'

'Estimated coefficients VCM:'

```
20.025   3.375  -6.000
 3.375   0.750  -1.125
-6.000  -1.125   1.875
```

Therefore the square root of the k th diagonal element of this matrix, $SE(\hat{\beta}_k) = \left\{\left[s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]_{kk}\right\}^{1/2}$, is the standard error of the estimator $\hat{\beta}_k$.

In [4]:
```
beta.std.errors <- diag(var.beta)^.5
print("Coefficients standard errors are:")
print(beta.std.errors)
```

```
[1] "Coefficients standard errors are:"
[1] 4.4749302 0.8660254 1.3693064
```

With α = 5%, the 95% interval can be calculated as:

$$\text{Prob}\left[\hat{\beta}_k - t_{(1-\alpha/2),[n-K]}SE(\hat{\beta}_k) \leq \beta_k \leq \hat{\beta}_k + t_{(1-\alpha/2),[n-K]}SE(\hat{\beta}_k)\right] = 1-\alpha$$

In [5]:
```
alpha <- 0.05
CV <- qt(c(alpha/2, 1 - alpha/2) , n-K)

paste0("The 95% critical values for a t-distribution with ", n-K," degrees of freedom")
round(CV,4)
```

'The 95% critical values for a t-distribution with 2 degrees of freedom'

1. -4.3027
2. 4.3027

The 95% interval for β_1 is therefore:

In [6]:
```
beta.1.min <- beta.hat[2] + CV[1]*beta.std.errors[2]
beta.1.max <- beta.hat[2] + CV[2]*beta.std.errors[2]
paste0('[',round(beta.1.min,3),',',round(beta.1.max,3),']')
```

'[-1.226,6.226]'

For each hypothesis test in c,d, and e below make sure you show: (i) the value of your test statistic, (ii) the relevant critical value(s), (iii) comparison of the test statistic vs. critical value(s), (iv) your conclusion. Unless specified, assume 95% significance levels i.e. α=0.05.

c. Test $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$.

$$t_k = \frac{\hat{\beta}_k - \beta_k}{SE(\hat{\beta}_k)}$$

(i) Test-statistic:

In [7]:
```
test.value <- 0
test.statistic <- (beta.hat[3] - test.value) / beta.std.errors[3]

paste0('Beta hat: ', beta.hat[3])
paste0('Test value: beta = ', test.value)
paste0('Std error: ', round(beta.std.errors[3],3))

paste0('Test statistic: t=', round(test.statistic,3))
```

'Beta hat: -1.5'

'Test value: beta = 0'

'Std error: 1.369'

'Test statistic: t=-1.095'

(ii) Critical values:

Since H_a is "not equal" that means we perform a two tailed test:

In [8]:
```
alpha <- 0.05
n <- 5
K <- length(beta.hat)

CV <- qt(c(alpha/2, 1 - alpha/2) , n-K)

paste0("The two tail 95% critical values for a t-distribution with ", n-K," degrees of freedom are")
round(CV,3)
```

'The two tail 95% critical values for a t-distribution with 2 degrees of freedom are'

1. -4.303
2. 4.303

(iii) comparison of test statistic vs. critical value

In [9]:
```
test.statistic < CV[1]
test.statistic > CV[2]
```

FALSE

FALSE

The test statistic does not exceed the critical values at either tail i.e. the test value is well within the "acceptance" region.

(iv) conclusion

Since the test statistic is within the 95% acceptance region **we cannot reject the null hypothesis**. The data appear to be consistent with the null hypothesis meaning that the variable is not relevant to the regression. Another way of phrasing it is that the variable is statistically insignificant at the 95% level.

d. Test $H_0 : \beta_1 + \beta_2 = 0$ versus $H_a : \beta_1 + \beta_2 \neq 0$.

(i) Test-statistic:

$$t = \frac{\left(\hat{\beta}_1 + \hat{\beta}_2\right) - (\beta_1 + \beta_2)}{\text{se}\left(\hat{\beta}_1 + \hat{\beta}_2\right)}$$

Per Greene Appendix the variance of the sum of two random variables is:

$$\begin{aligned}\text{Var}[ax + by + c] &= a^2\,\text{Var}[x] + b^2\text{Var}[y] + 2ab\,\text{Cov}[x, y]\\ &= \text{Var}[ax + by],\end{aligned} \qquad \textbf{(B-55)}$$

In [10]:
```
var.sum <- var.beta[2,2] + var.beta[3,3] + 2*var.beta[2,3]
print(var.sum)
```

```
[1] 0.375
```

Therefore the standard error is:

In [11]:
```
se.sum <- sqrt(var.sum)
print(se.sum)
```

```
[1] 0.6123724
```

In [12]:
```
test.value <- 0
test.statistic <- (beta.hat[2] + beta.hat[3] - test.value) / se.sum

paste0('Sum of betas: ', beta.hat[2]+beta.hat[3])
paste0('Test value: beta = ', test.value)
paste0('Std error: ', round(se.sum,3))

paste0('Test statistic: t=', round(test.statistic,3))
```

'Sum of betas: 1'

'Test value: beta = 0'

'Std error: 0.612'

'Test statistic: t=1.633'

(ii) Critical values:

Since H_a is "not equal" that means we perform a two tailed test:

In [13]:
```
alpha <- 0.05
n <- 5
K <- length(beta.hat)

CV <- qt(c(alpha/2, 1 - alpha/2) , n-K)

paste0("The two tail 95% critical values for a t-distribution with ", n-K," degrees of freedom are")
round(CV,3)
```

'The two tail 95% critical values for a t-distribution with 2 degrees of freedom are'

1. -4.303
2. 4.303

(iii) comparison of test statistic vs. critical value

In [14]:
```
test.statistic < CV[1]
test.statistic > CV[2]
```

FALSE

FALSE

The test statistic does not exceed the critical values at either tail i.e. the test value is well within the "acceptance" region.

(iv) conclusion

Since the test statistic is within the 95% acceptance region **we cannot reject the null hypothesis**. The data appear to be consistent with the null hypothesis meaning that the variables are joint irrelevant to the regression. Another way of phrasing it is that the variables are joint insignificant at the 95% level.

e. Test $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$. Note that most software packages include the test in the default regression output as the *F Statistic*.

(i) Test-statistic:

$$F = \frac{\left(\mathrm{RSS_R} - \mathrm{RSS_{UR}}\right)/J}{\mathrm{RSS_{UR}}/(n-K)}$$

In this case, RSS_R is the RSS if all coefficients except for the intercept were set to 0 i.e. $\sum(Y-\bar{Y})^2$ i.e. the TSS_{UR}. The number of restrictions, J, is just the number of regression coefficients excluding the constant, $K-1$

$$F = \frac{\left(\mathrm{TSS_{UR}} - \mathrm{RSS_{UR}}\right)/J}{\mathrm{RSS_{UR}}/(n-K)} = \frac{\mathrm{ESS}/(K-1)}{\mathrm{RSS}/(n-K)} = \frac{26.5/2}{1.5/2}$$

In [15]:
```
n <- 5
K <- length(beta.hat)
RSS <- 1.5
ESS <- 26.5

test.statistic <- (ESS / (K-1) /(RSS / (n - K)))
paste0('Test statistic: F=', round(test.statistic,3))
```

'Test statistic: F=17.667'

Note that solutions based on the provided R-squared are inaccurate since the figure of 0.95 is rounded. You could easily compute R-squared with more significant digits from the provided data as:

In [16]:
```
myR.squared <- ESS / (ESS + RSS)
myR.squared
```

0.946428571428571

Then, using

$$F = \frac{R^2/(K-1)}{\left(1-R^2\right)/(n-K)}$$

In [17]:
```
test.statistic.alternate <- (myR.squared / (K-1)) / ((1 - myR.squared) / (n-K))
paste0('Test statistic alternate: F=', round(test.statistic.alternate,3))
```

'Test statistic alternate: F=17.667'

(ii) Critical value(s):

Note that this is a **one tailed test**: we are checking whether the variance of the restricted model is less than the variance of the unrestricted model. In other words we're checking for the following conditions: (a) "less than" vs (b) "not less than". Compare that to a two tail test like above where we check (a) "equal" vs. (b) "not equal" i.e. "not equal" implies checking both conditions of "less than" or "greater than."

Since we're performing a one tailed test we need the critical value for an F-distribution with 2 degrees of freedom in the numerator and 2 degrees of freedom in the denominator and α = 0.05.

In [18]:
```
alpha <- 0.05
n <- 5
K <- length(beta.hat)

CV <- qf(1-alpha, K-1, n-K)

paste0("The 95% critical value for a F-distribution with [",K-1,",", n-K,"] degrees of freedom is")
round(CV,3)
```

'The 95% critical value for a F-distribution with [2,2] degrees of freedom is'
19

(iii) comparison of test statistic vs. critical value

In [19]:
```
test.statistic > CV
```

FALSE

The test statistic does not exceed the critical value i.e. test statistic falls in the "acceptance" region.

(iv) conclusion

Since the test statistic is within the 95% acceptance region, **we cannot reject the null hypothesis**. In other words, the data appear to be consistent with the null hypothesis i.e. restricted model has less variance than the unrestricted model i.e. the independent variables in the regression are statistically irrelevant in explaining variation in the dependent variable.

f. What inferences do you make about this regression based on the R-squared value and how do these compare with inferences made based on the F-statistic in (e) above?

Despite the high R-squared we cannot reject the null hypothesis that all independent variables are irrelevant in explaining the variation in y. This is a consequence of the small sample size (n=5).

Problem 2 (4 points)

In the fourth quarter of 1966, UK's Labor government liberalized the National Insurance Act by replacing the flat-rate system of short-term unemployment benefits by a mixed system of flat-rate and (previous) earnings-related benefits, which increased the level of unemployment benefits. Dataset https://www747.github.io/NYU/HW3/problem_2_data.csv includes the UK unemployment rate U, vacancy rate V, and a dummy variable D = 0 before 1966Q4 and 1 after.

a. Import the dataset and estimate the model $U = \beta_0 + \beta_1 V + \beta_2 D + \beta_3 D \cdot V + \epsilon$

In [20]:
```
mydata <- read.csv("https://www747.github.io/NYU/HW3/problem_2_data.csv")
#mydata <- read.csv("problem_2_data.csv")
head(mydata)
tail(mydata)
```

	t	U	V	D	DV
	12/31/1958	1.915	0.510	0	0
	3/31/1959	1.876	0.541	0	0
	6/30/1959	1.842	0.541	0	0
	9/30/1959	1.750	0.690	0	0
	12/31/1959	1.648	0.771	0	0
	3/31/1960	1.450	0.836	0	0

	t	U	V	D	DV
46	3/31/1970	2.225	0.757	1	0.757
47	6/30/1970	2.241	0.746	1	0.746
48	9/30/1970	2.366	0.739	1	0.739
49	12/31/1970	2.324	0.707	1	0.707
50	3/31/1971	2.516	0.583	1	0.583
51	6/30/1971	2.909	0.524	1	0.524

In [21]:
```
my.model <- lm(U ~ V + D + DV, mydata)
my.model
```

```
Call:
lm(formula = U ~ V + D + DV, data = mydata)

Coefficients:
(Intercept)            V            D           DV
     2.7331      -1.5126       1.1667      -0.8679
```

b. Holding the job vacancy rate constant, what is the average unemployment rate in the period beginning in the fourth quarter of 1966?

Note that dummy variables have to be interpreted with respect to the regression intercept (bias) which is itself a sort of dummy variable.

In [22]:
```
beta.hat <- unname(my.model$coeff)  # remove names from coefficient vector
avg.ur <- beta.hat[1] + beta.hat[3] # compute average UR post 1966Q4 by summing the intercept and D coefficient
print(avg.ur)
```

```
[1] 3.89982
```

b. Is it statistically different from the period before 1966 fourth quarter? How do you know?

In [23]:
```
summary(my.model)
```

```
Call:
lm(formula = U ~ V + D + DV, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33629 -0.07242  0.01457  0.07464  0.25658

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.7331     0.1014  26.949  < 2e-16 ***
V            -1.5126     0.1210 -12.499  < 2e-16 ***
D             1.1667     0.3178   3.671 0.000616 ***
DV           -0.8679     0.4306  -2.016 0.049556 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1344 on 47 degrees of freedom
Multiple R-squared:  0.9122,    Adjusted R-squared:  0.9066
F-statistic: 162.8 on 3 and 47 DF,  p-value: < 2.2e-16
```

Looking at the standard errors of the regression coefficients we note that the dummy coefficient (D) is highly significant therefore we conclude that the unemployment rate post 1966Q4 is higher.

c. Are the slopes in the pre- and post-1966 fourth quarter statistically different? How do you know?

Since the differential dummy coefficient is significant at the 5% level (t = -2.016, p = 0.0496), we conclude that the slopes in the two periods are different.

d. Is it safe to conclude from this study that generous unemployment benefits lead to higher unemployment rates? Does this make economic sense?

Yes and agrees with the prevaling economic theory ie. more generous unemployment benefits reduces the opportunity cost of remaining unemployed.

Problem 3 (3 points)

Dataset https://www747.github.io/NYU/HW3/problem_3_data.csv includes quarterly data on the following variables:

$Y =$ quantity of donuts sold, dozens

$X_1 =$ the trend variable

$X_2 =$ average price of donuts, $\$/$ dozen

$X_3 =$ average price of cupcakes, $\$/$ dozen

$X_4 =$ average weekly family disposable income, $\$/$ week

a. Estimate the demand function $Y_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{4t} + \epsilon_t$ and interpret the slope coefficients of this linear model.

In [24]:
```
mydata <- read.csv("https://www747.github.io/NYU/HW3/problem_3_data.csv")
# mydata <- read.csv("problem_3_data.csv")
head(mydata)
tail(mydata)
```

	t	Y	X1	X2	X3	X4
	3/31/1971	11484	1	2.26	3.49	158.11
	6/30/1971	9348	2	2.54	2.85	173.36
	9/30/1971	8429	3	3.07	4.06	165.26
	12/31/1971	10079	4	2.91	3.64	172.92
	3/31/1972	9240	5	2.73	3.21	178.46
	6/30/1972	8862	6	2.77	3.66	198.62

	t	Y	X1	X2	X3	X4
11	9/30/1973	5911	11	3.77	3.65	181.87

	t	Y	X1	X2	X3	X4
12	12/31/1973	7950	12	3.64	3.60	185.00
13	3/31/1974	6134	13	2.82	2.94	184.00
14	6/30/1974	5868	14	2.96	3.12	188.20
15	9/30/1974	3160	15	4.24	3.58	175.67
16	12/31/1974	5872	16	3.69	3.53	188.00

In [25]:
```
lin.model <- lm(Y ~ X1 + X2 + X3 + X4, mydata)
lin.model
```

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = mydata)

Coefficients:
(Intercept)           X1           X2           X3           X4
  10816.043     -197.400    -2227.704     1251.141        6.283
```

In this model the slope coefficients measure the rate of change of Y with respect to the relevant variable.

b. Estimate the semi-log demand function $lnY_t = \gamma_0 + \gamma_1 X_{1t} + \gamma_2 X_{2t} + \gamma_3 X_{3t} + \gamma_4 X_{4t} + \epsilon_t$.

In [26]:
```
semi.log.model <- lm(log(Y)~ X1 + X2 + X3 + X4, mydata)
semi.log.model
```

```
Call:
lm(formula = log(Y) ~ X1 + X2 + X3 + X4, data = mydata)

Coefficients:
(Intercept)           X1           X2           X3           X4
   8.660426    -0.027866    -0.376531     0.213270     0.005077
```

c. Estimate the demand function $\ln Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + \beta_4 \ln X_{4t} + \epsilon_t$ and interpret the the slope coefficients of this log-linear model.

In [27]:
```
log.log.model <- lm(log(Y) ~ log(X1) + log(X2) + log(X3) + log(X4), mydata)
log.log.model
```

```
Call:
lm(formula = log(Y) ~ log(X1) + log(X2) + log(X3) + log(X4),
    data = mydata)

Coefficients:
(Intercept)      log(X1)      log(X2)      log(X3)      log(X4)
     0.6268      -0.1816      -1.2736       0.9373       1.7130
```

In this model all the slope coefficients are partial elasticities of Y with respect to the relevant variable.

d. β_2, β_3, and β_4 give, respectively, the own-price, cross-price, and income elasticities of demand. What are their expected signs? (hint: donuts are a normal good) Do the results concur with the a priori expectations?

The own-price elasticity is expected to be negative, the cross price elasticity is expected to be positive for substitute goods and negative for complimentary goods, and the income elasticity is expected to be positive, since roses are a normal good.

e. How would you compute the own-price, cross-price, and income elasticities for the linear model?

The general formula for elasticity for linear equation is:

$$\text{Elasticity} = \frac{\partial Y}{\partial X_i}\frac{\bar{X}_i}{\bar{Y}},$$

where X_i is the relevant regressor. So for a linear model, the elasticity can be computed at the mean values.

f. On the basis of your analysis, which model, if either, would you choose and why?

In [28]:
```
options(warn=-1) # Suppress warnings
qui <- suppressPackageStartupMessages # quiet! - suppress library load messages
# install (if not already installed) and load package
qui(if(!require(stargazer)){install.packages('stargazer')})

stargazer(lin.model,semi.log.model, log.log.model, column.labels=c("Linear","Semi-log","Log-linear"), type="text")
```

```
================================================================
                                Dependent variable:
                            --------------------------------
                                  Y             log(Y)
                               Linear    Semi-log  Log-linear
                                (1)        (2)       (3)
----------------------------------------------------------------
X1                           -197.400*    -0.028
                             (101.561)    (0.016)

X2                          -2,227.704** -0.377**
                             (920.466)    (0.149)

X3                           1,251.141     0.213
                            (1,157.021)   (0.187)

X4                             6.283       0.005
                              (30.622)    (0.005)

log(X1)                                                -0.182
                                                      (0.128)

log(X2)                                               -1.274**
                                                      (0.527)

log(X3)                                                0.937
                                                      (0.659)

log(X4)                                                1.713
                                                      (1.201)

Constant                    10,816.040*  8.660***    0.627
                            (5,988.348)   (0.967)   (6.148)

----------------------------------------------------------------
Observations                    16         16         16
R2                             0.835      0.809      0.778
Adjusted R2                    0.775      0.740      0.697
Residual Std. Error (df = 11)  969.874    0.157      0.169
F Statistic (df = 4; 11)      13.886***  11.650***  9.635***
================================================================
Note:                               *p<0.1; **p<0.05; ***p<0.01
```

All three models give similar results but the log-linear model gives direct estimates of the (constant) elasticity of the relevant variable with respect to the regressor under consideration. Note that the R-squareds of the three models are not directly comparable so you cannot chose based on that figure.