National University of Lesotho

B.A. Examinations – Supplementary

EC4401 – Data Sciences for Economists

August 2023 **100 Marks** **3 Hours**

INSTRUCTIONS:

1. Answer ALL questions
2. All questions have 25 marks each.
3. Refer to Appendix 1 for the description of the variables used.

Question One
Clearly compare and contrast the following

a) R and R studio
b) Data Science and Data management
c) R script and R Markdown
d) Text Mining and Web Scraping
e) R Pipes and Argument

[25 Marks]

Question Two

a) Given the data collected for 118 firms in Leribe, an OLS regression has been estimated and some OLS assumptions have been tested and presented in Appendix 2 below. Interpret the results in detail.

[20 Marks]

b) Why is there so much popularity and interest in data science now, when organisations and countries have been gathering data for centuries? **[5 Marks]**

Question Three
a) Appendix 3 below, contains a correlation matrix and graph, resulting from a pairwise estimation of average sales, age of business, total assets of firms and average annual profit of firms using R studio. Clearly interpret the correlation matrix and graph.

[15 Marks]

b) Clearly state the R codes the following:
i. Create a vector with (1,2,5,5,5,6,7,8,9) and name it with your surname
ii. Create another vector which includes the figures in your student ID and name it with your middle name
iii. Multiply middlename by the square of surname
iv. Divide middle name by half of surname
v. Calculate the summary statistics of middlename and surname

[10 Marks]

Question Four
Assume that you have been tasked with the responsibility to improve female ownership of firms and you have estimated a logit regression as presented in Appendix 5. Wherein the dependent variable is gender ownership (1 if female & 0 otherwise) and the independent variables are capital, age of business, total credit, and education. Interpret the results. **[17 Marks]**

b) Interpret the graphs in Appendix 4 **[8 Marks]**

Good Luck

APPENDICES

Appendix 1: Variable Descriptions

No	Variable name	Description
1	totalasset	Total assets of the firm
2	capital	Capital of the firms
3	agebus	Longevity or age of firm in years
4	employoperatives	Number of workers
5	education	Educational level where 2 = primary, 3 = secondary and 4 = tertiary
6	totalcredit	Total credit approved for firms
7	avsales	Average annual sales of firms
8	avprofit	Average annual profit of firms

Appendix 2: OLS Regression model

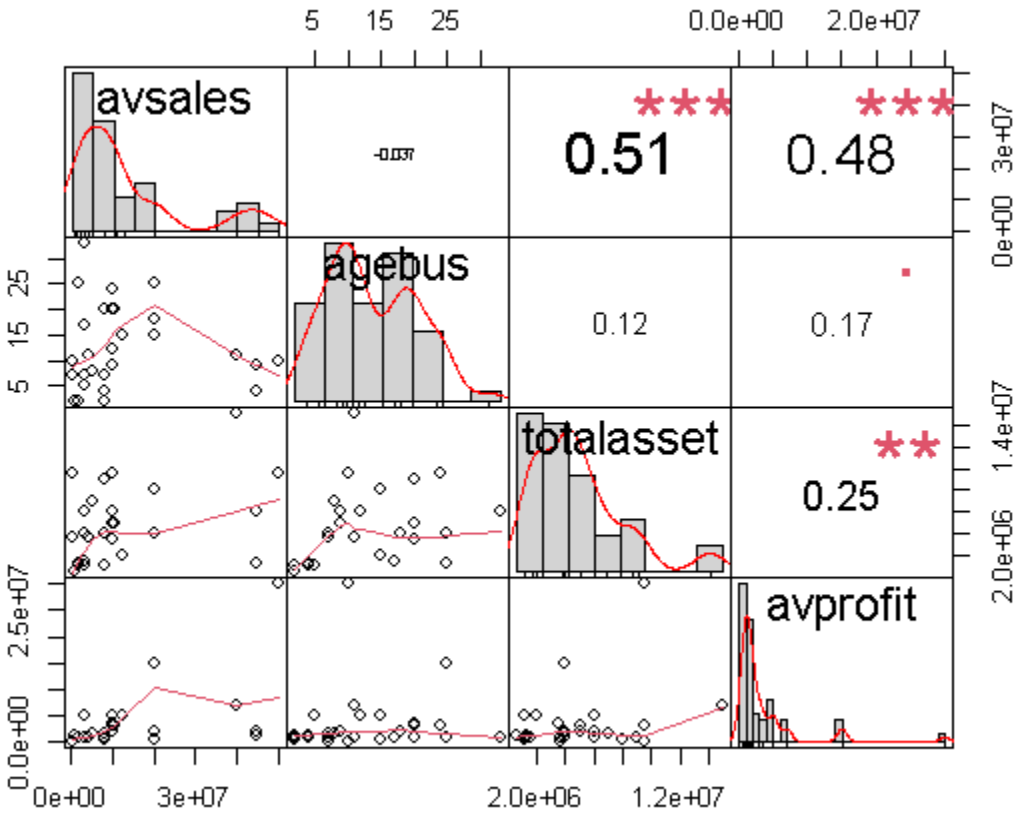
```
MODEL INFO:
Observations: 118
Dependent Variable: totalasset
Type: OLS linear regression

MODEL FIT:
F(5,112) = 18.709, p = 0.000
R² = 0.455
Adj. R² = 0.431

Standard errors: OLS
------------------------------------------------------------
                             Est.    S.E.   t val.       p
------------------------- -------- ------- -------- -------
(Intercept)                 0.174   0.268    0.649   0.518
capital                     0.309   0.076    4.098   0.000
agebus                      0.174   0.074    2.361   0.020
employoperatives            0.280   0.075    3.730   0.000
factor(education)`3        -0.475   0.283   -1.681   0.096
factor(education)`4         0.247   0.297    0.832   0.407
------------------------------------------------------------

Continuous variables are mean-centered and scaled by 1 s.d.

> export_summs(reg_exam)
```

	Model 1
(Intercept)	2506147.92 *
	(1061539.03)
capital	0.13 ***
	(0.03)
agebus	79929.57 *
	(33860.61)
employoperatives	191810.62 ***
	(51425.39)
factor(education)3	-1640706.67
	(976277.78)
factor(education)4	852469.70
	(1024793.69)
N	118
R2	0.46

```
                    *** p < 0.001; ** p < 0.01; * p < 0.05.

> # Outlier Test
> outlierTest(reg_exam)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferroni p
42 -2.363171           0.019859           NA
>

> # Normality Test
> shapiro.test(reg_exam$residuals)

        Shapiro-Wilk normality test

data:  reg_exam$residuals
W = 0.95751, p-value = 0.0009034



> # Heteroscdasticity
> ols_test_breusch_pagan(reg_exam)

 Breusch Pagan Test for Heteroskedasticity
 -----------------------------------------
 Ho: the variance is constant
 Ha: the variance is not constant

                 Data
 -------------------------------------
 Response : totalasset
 Variables: fitted values of totalasset

         Test Summary
 -----------------------------
 DF            =    1
 Chi2          =    24.03363
 Prob > Chi2   =    9.466748e-07
>

> # Multicolinearity
> vif(reg_exam)
                     GVIF Df GVIF^(1/(2*Df))
capital            1.172407  1        1.082778
agebus             1.117553  1        1.057144
employ operatives  1.154445  1        1.074451
factor(education)  1.220212  2        1.051015


> # Linearity
> raintest(reg_exam)

        Rainbow test

data:  reg_exam
Rain = 0.89831, df1 = 59, df2 = 53, p-value = 0.6569
```

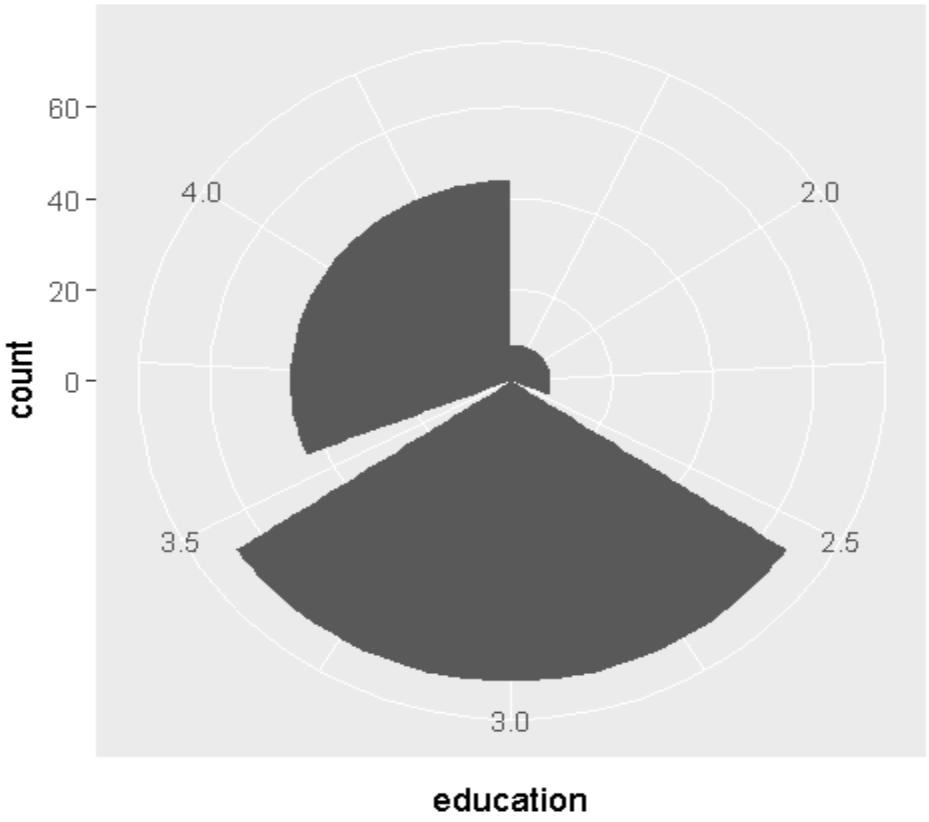
Appendix 3

```
Correlation matrix
            avsales agebus totalasset avprofit
avsales        1.00  -0.04        0.51     0.48
agebus        -0.04   1.00        0.12     0.17
totalasset     0.51   0.12        1.00     0.25
avprofit       0.48   0.17        0.25     1.00
Sample Size
[1] 118
Probability values (Entries above the diagonal are adju
sted for multiple tests.)
            avsales agebus totalasset avprofit
avsales        0.00   0.69        0.00     0.00
agebus         0.69   0.00        0.42     0.20
totalasset     0.00   0.21        0.00     0.03
avprofit       0.00   0.07        0.01     0.00
```

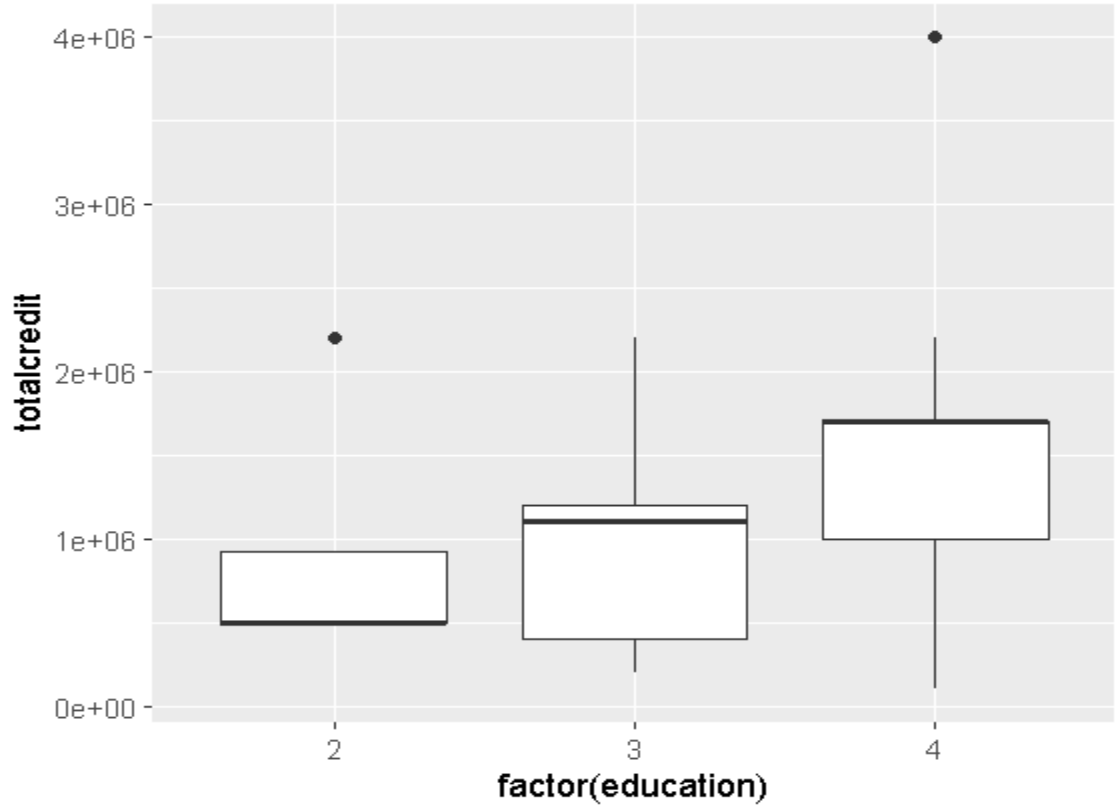


Appendix 4

a.



b.



Appendix 5: Logit Regression

```
glm(formula = genderfh ~ capital + agebus + totalcredit + factor(education),
    family = binomial(link = "logit"), data = Firm_Surv2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9094  -1.3156   0.8139   0.8968   1.4112

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.305e+00  8.844e-01   1.476    0.140
capital            -3.900e-08  3.101e-08  -1.258    0.208
agebus             -1.192e-02  2.861e-02  -0.417    0.677
totalcredit        -4.635e-07  2.988e-07  -1.551    0.121
factor(education)3  6.333e-02  7.918e-01   0.080    0.936
factor(education)4  5.051e-01  8.563e-01   0.590    0.555

(Dispersion parameter for binomial family taken to be 1)

Number of Fisher Scoring iterations: 4
```

```
> export_summs(model7)
```

	Model 1
(Intercept)	1.31
	(0.88)
capital	-0.00
	(0.00)
agebus	-0.01
	(0.03)
totalcredit	-0.00
	(0.00)
factor(education)3	0.06
	(0.79)
factor(education)4	0.51
	(0.86)
N	118
AIC	158.18
BIC	174.80
Pseudo R2	0.07

*** p < 0.001; ** p < 0.01; * p < 0.05.

Odds Ratio Estimates

```
> exp(model7$coefficients)
       (Intercept)             capital              agebus         totalcredit
         3.6878330           1.0000000           0.9881497           0.9999995
factor(education)3 factor(education)4
         1.0653825           1.6571209
```

Marginal Effect Estimates

```
> model8 <- mean(dlogis(predict(model7, type = "link")))
> model8* coef(model7)
       (Intercept)             capital              agebus         totalcredit
      2.803380e-01      -8.378708e-09      -2.560788e-03      -9.956127e-08
factor(education)3 factor(education)4
      1.360488e-02        1.084976e-01
```

Model Fitness

```
> chis=model7$null.deviance-model7$deviance
> dfdiff=model7$df.null-model7$df.residual
> #For pvalue of chi square pchisq(chis,dfdiff,lower.tail=F)
[1] 0.283242
```