MAASAI MARA UNIVERSITY

UNIVERSITY EXAMINATIONS 2022/2023

FOURTH YEAR SECOND SEMESTER EXAMINATIONS FOR THE DEGREE OF BACHELOR OF SCIENCE IN ECONOMICS AND STATISTICS

ECO 4220: LINEAR MODELLING

DATE: APRIL 2023 **TIME: 2 HOURS**

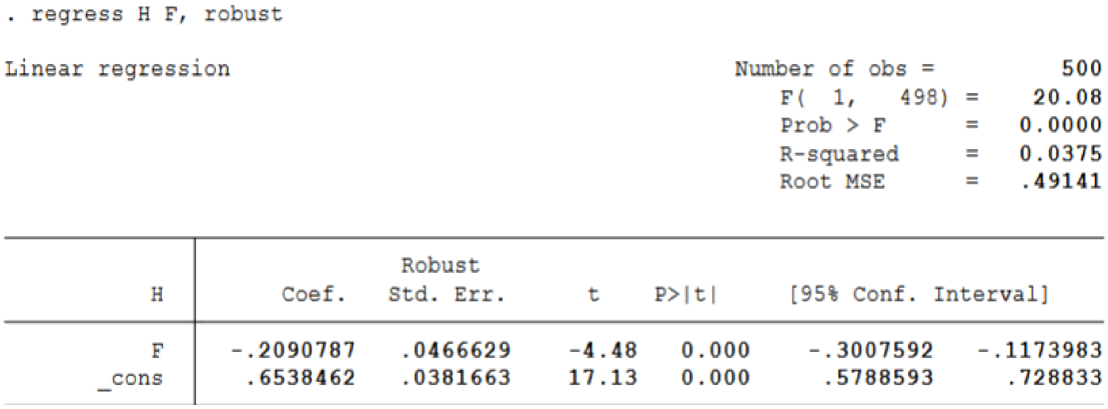
INSTRUCTIONS TO CANDIDATES:

1. *Answer questions ONE (section A) and any THREE questions in section B*

SECTION A

QUESTION ONE {25 MARKS}

a) The government of a developing country wants to implement a program where poor families receive food stamps that can be used to purchase prepackaged foods with high nutritional value. The government decides to set up an experiment where 500 families (each with 1 child) are randomly assigned to a treatment group (eligible for food stamps, $T_i = 1$) and to a control group (ineligible for food stamps, $T_i = 0$). The government has hired a researcher to investigate the effect of food stamps on the probability that a child has poor health. After the experiment the researcher performs a regression of H_i (a binary variable that equals 1 if a child has poor health) on F_i (a binary variable that equals 1 if a family received food stamps). She obtains the following OLS estimation results.

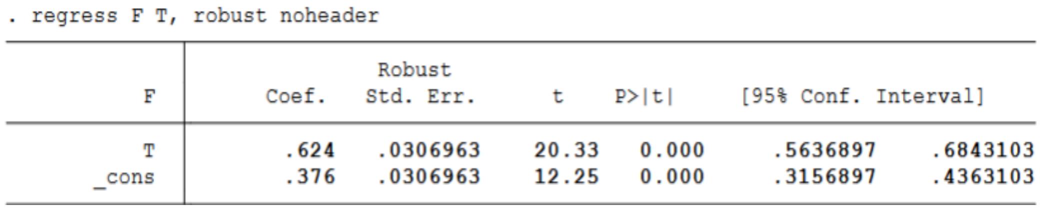


```
. regress H F, robust

Linear regression                               Number of obs =        500
                                                   F(  1,   498) =    20.08
                                                   Prob > F      =   0.0000
                                                   R-squared     =   0.0375
                                                   Root MSE      =   .49141

------------------------------------------------------------------------------
             |               Robust
           H |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           F |  -.2090787   .0466629    -4.48   0.000    -.3007592   -.1173983
       _cons |   .6538462   .0381663    17.13   0.000     .5788593     .728833
------------------------------------------------------------------------------
```

i. Interpret the two estimated coefficients. (2 marks)

ii. The researcher finds out that some of the families in the control group received food stamps. Explain whether we can interpret the estimated OLS coefficient on F as the causal effect of food stamps on child health? (3 marks)

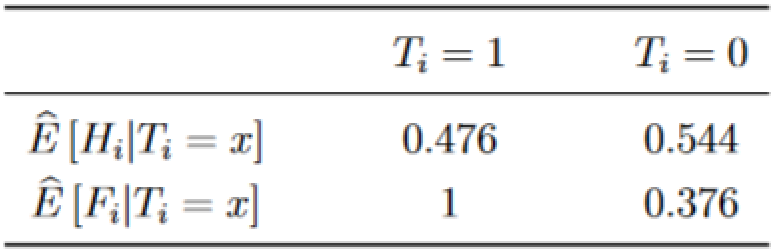
iii. The researcher decides to estimate the effect of food stamps using an instrumental variable approach. She uses assignment to the treatment group as instrument for the actual receipt of food stamps. She obtains the following first stage estimation results. Do you think that the instrument relevance condition holds? Is T a weak instrument? (3 marks)



```
. regress F T, robust noheader

------------------------------------------------------------------------------
             |               Robust
           F |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |       .624   .0306963    20.33   0.000     .5636897    .6843103
       _cons |       .376   .0306963    12.25   0.000     .3156897    .4363103
------------------------------------------------------------------------------
```

iv. Do you think that the instrument exogeneity condition holds? (2 marks)

v. The following table shows the averages of Hi and Fi for those assigned to treatment group (T_i = 1) and for those assigned to the control group (T_i = 1). Use the results in the table below to obtain the instrumental variable estimate of the effect of food stamps on the probability that a child has poor health. (3 marks)



	$T_i = 1$	$T_i = 0$
$\widehat{E}\left[H_i \vert T_i = x\right]$	0.476	0.544
$\widehat{E}\left[F_i \vert T_i = x\right]$	1	0.376

b) With standard notation, the multiple linear regression model is given by $E(Y) = X\beta$ and $Var(Y) = \sigma^2 I$ where X is a non-stochastic $n \times k$ design matrix.

i) Derive the least squares estimator $\hat{\beta}$ of β and show that it is given by $\hat{\beta} = (X^T X)^{-1} X^T Y$ (4 marks)

ii) Show that $\hat{\beta}$ is an unbiased estimator of β and obtain its variance-covariance matrix. (3 marks)

iii) Show that estimator of σ^2 is given by $\hat{\sigma}^2 = \frac{1}{n-k}(Y - X\hat{\beta})^T (Y - X\hat{\beta})$ is unbiased. (3 marks)

c) A simple linear regression model is given by the following equation

$$y_i = \beta x_i + e_i \qquad i = 1, \dots, n$$

Such that $e_i \sim iid\ N(0, \sigma^2)$.

i) Compute the least squares estimate of β (2 marks)

ii) Suppose that $\tilde{\beta} = \frac{\bar{Y}}{\bar{X}}$ is another estimator of β. Show that $\tilde{\beta}$ is an unbiased estimator of β. (3 marks)

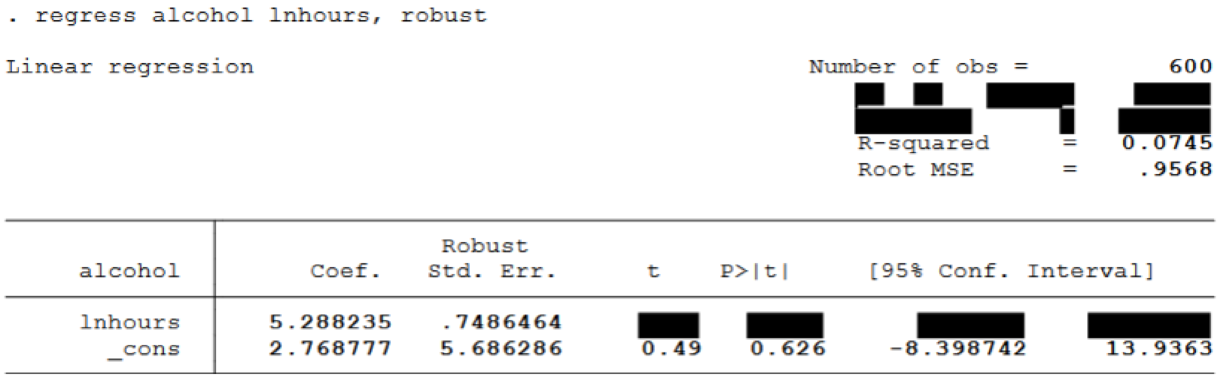
SECTION B

QUESTION TWO { 15 MARKS}

The Norwegian government wants to know whether restricting the opening hours of liquor stores reduces alcohol consumption. Holger, an employee of Statistics Norway, is asked to investigate this research question. He uses panel data for n = 60 municipalities observed in T = 10 time periods. The data set contains information on per capita alcohol consumption (in liters per year) in municipality i in year t (alcohol$_{it}$) and on the number of hours that liquor stores were open during year t in municipality i (hours$_{it}$). Holger estimates

$$alcohol_{it} = \beta_0 + \beta_1 ln(hours_{it}) + u_{it}$$

by OLS and obtains the following estimation results.



```
. regress alcohol lnhours, robust

Linear regression                               Number of obs =        600
                                                [redacted]
                                                R-squared     =     0.0745
                                                Root MSE      =      .9568

------------------------------------------------------------------------------
             |               Robust
     alcohol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lnhours |   5.288235   .7486464  [redacted] [redacted] [redacted] [redacted]
       _cons |   2.768777   5.686286    0.49   0.626    -8.398742     13.9363
------------------------------------------------------------------------------
```

a) Test the null hypothesis that β_1= 0 at a 1% significance level. (3 marks)

b) Use the above estimation results to predict the change in alcohol consumption if the opening hours of liquor stores are reduced by 20 percent. (2 marks)

c) Marit, Holger's colleague, suggests to augment the model with municipality fixed effects α_i

$$alcohol_{it} = \beta_0 + \beta_1 ln(hours_{it}) + \alpha_i + u_{it} \qquad (1)$$

Explain how you could estimate model (1) (5 marks)

d) Holger decides to estimate a model that includes both municipality and year fixed effects. Both Holger and Marit are confident that by including municipality and year fixed effects the estimated coefficient on $ln(\text{hours}_{it})$ cannot suffer from omitted variable bias problems. Do you agree with Holger and Marit? (5 marks)

QUESTION THREE {15 MARKS}

The data given in Appendix R.1 represents the impact of three advertising media- YouTube, Facebook and Newspaper, on sales. The values for all the four columns are given in millions of Kenyan shillings (KSHs). The advertising experiment was repeated 200 times, hence there are 200 data values for each column but Appendix 1 shows the first 6 observations.

a) Based on the scatter plots in Appendix R.2, comment on the relationship between (both strength and direction) each of the three media channels with sales. (3 marks)

b) Using the R Output in Appendix R.3 answer the following questions

(i) Write the equation for the fitted model (1 mark)

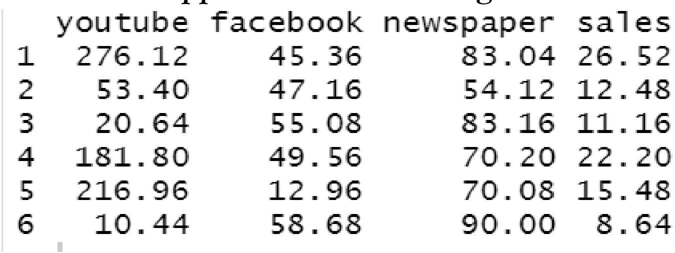
(ii) Interpret the p-value corresponding to the F- statistic (1 mark)

(iii)Give an interpretation of the estimates and p-values of the intercept, and the coefficients of the three advertising media (4 marks)

(iv) Based on the output, do you think one or more of the independent variables can be removed? If yes, which one and why? If No, why? (2marks)

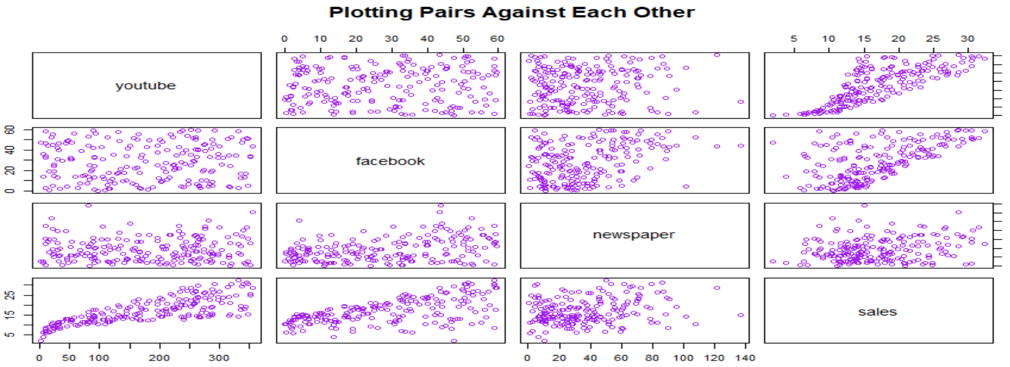
d) What patterns or problems do you see in the diagnostic plots in Appendix R.4. Is the multiple linear regression a good fit for the marketing dataset? (4 marks)

Appendix R.1 Marketing Data



```
  youtube facebook newspaper sales
1  276.12    45.36     83.04 26.52
2   53.40    47.16     54.12 12.48
3   20.64    55.08     83.16 11.16
4  181.80    49.56     70.20 22.20
5  216.96    12.96     70.08 15.48
6   10.44    58.68     90.00  8.64
```

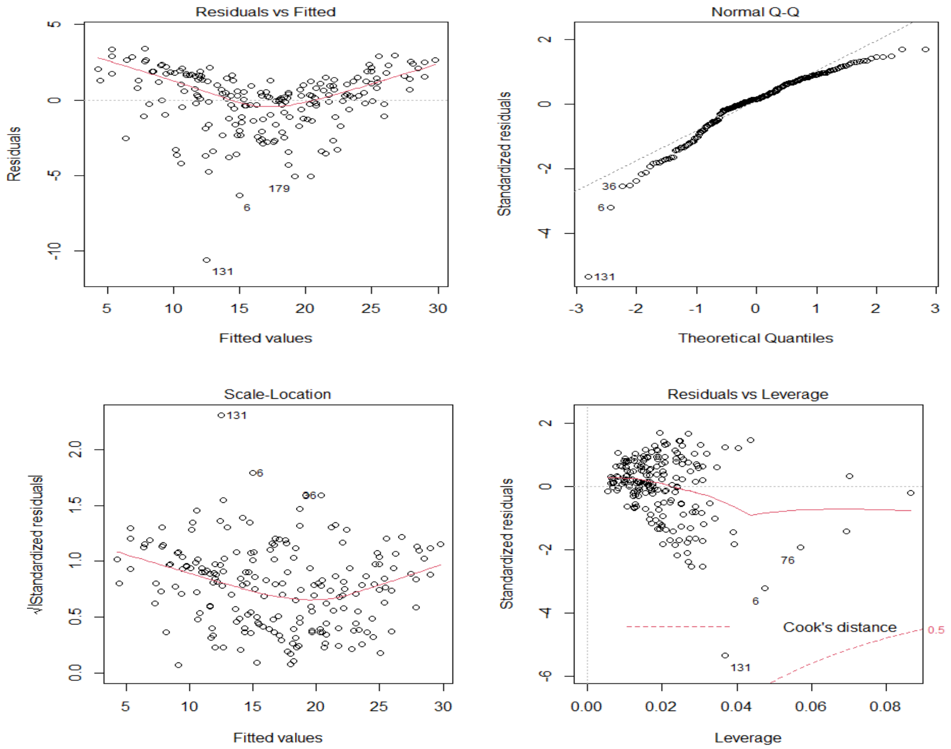
Appendix R.2 Scatter Plots



Appendix R.3 Model Summary

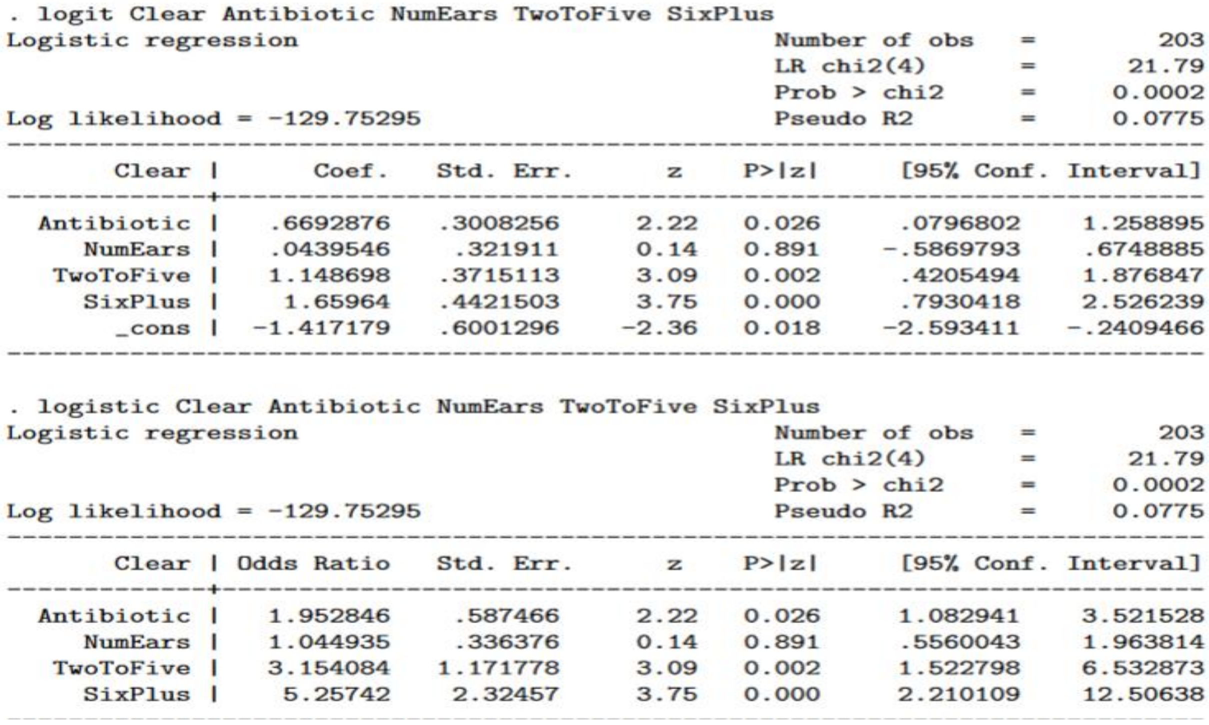
```
Call:
lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.5932  -1.0690   0.2902   1.4272   3.3951 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.526667   0.374290   9.422   <2e-16 ***
youtube      0.045765   0.001395  32.809   <2e-16 ***
facebook     0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Appendix R.4 Diagnostic plots



QUESTION FOUR {15 MARKS}

An experiment was designed to assess the impact of two different antibiotics on the chances a child will be cured of an ear infection after adjusting for agd and whether one or both ears were infected. The variables are “Clear”–whether the infection has been cleared from both ears after 14 days treatment, “Antibiotic”–the treatment type (1 = Ceftriaxone, 0 = Amoxicillin), Age (categories under two years old, 2-5 years old and 6 year or older), and “NumEars”–the number of ears infected (either1 or 2). STATA outputs for the pertinent logistic regression model are below. There are two versions, logit which gives the raw coefficients and their standard errors and logistic which gives the odds ratios and their standard errors.



```
. logit Clear Antibiotic NumEars TwoToFive SixPlus
Logistic regression                               Number of obs   =        203
                                                  LR chi2(4)      =      21.79
                                                  Prob > chi2     =     0.0002
Log likelihood = -129.75295                       Pseudo R2       =     0.0775
------------------------------------------------------------------------------
       Clear |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Antibiotic |   .6692876   .3008256     2.22   0.026     .0796802    1.258895
     NumEars |   .0439546    .321911     0.14   0.891    -.5869793    .6748885
   TwoToFive |   1.148698   .3715113     3.09   0.002     .4205494    1.876847
     SixPlus |    1.65964   .4421503     3.75   0.000     .7930418    2.526239
       _cons |  -1.417179   .6001296    -2.36   0.018    -2.593411   -.2409466
------------------------------------------------------------------------------

. logistic Clear Antibiotic NumEars TwoToFive SixPlus
Logistic regression                               Number of obs   =        203
                                                  LR chi2(4)      =      21.79
                                                  Prob > chi2     =     0.0002
Log likelihood = -129.75295                       Pseudo R2       =     0.0775
------------------------------------------------------------------------------
       Clear | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Antibiotic |   1.952846    .587466     2.22   0.026     1.082941    3.521528
     NumEars |   1.044935    .336376     0.14   0.891     .5560043    1.963814
   TwoToFive |   3.154084   1.171778     3.09   0.002     1.522798    6.532873
     SixPlus |    5.25742    2.32457     3.75   0.000     2.210109    12.50638
------------------------------------------------------------------------------
```

a) Overall do these variables help explain how likely a child is to have their ear infections cleared in14 days? Briefly justify your answer. (2 marks)

b) Do these variables explain a lot the “variability” in how likely an ear infection is to clear? Explain briefly. What are the practical implications of this statement for treating ear infections in smallc hildren with antibiotics? (3 marks)

c) Describe what you think would happen if you used backwards stepwise selection to find the best model for predicting whether a child’s ear-infection would clear. That is, say what variables would be included in the intial model, what would happen at each step, and what you think the final model would be, and what you would have to do to verify your answer.

(4 marks)

d) Explain briefly how you could figure out what variable to add first in a forwards stepwise model selection procedure for this data. (2 marks)

e) Which of the age categories have I used as the reference in this model? (2 marks)

f) Give brief interpretations of the odds ratios for the “Antibiotic” and “TwoToFive” Variables and show how you would compute them from the information given in the first (logit) printout. (2 marks)

QUESTION FIVE {15 MARKS}

a) Suppose the relationship between two variables X_j and Y_j can be given by the model

$Y_j = \alpha + \beta X_j + e_j$, $j = 1, 2, \ldots n$, where $(Y_j, X_j) \in \Re x \Re$, Y_j is a dependant variable, X_j is random and independent of Y_j , and e_j's are random error variables. The researcher uses the model and least squares method for estimation. Suppose the he observes X_{n+1} so that the forecast of Y_{n+1} is $\hat{Y}_{n+1} = \hat{\alpha} + \hat{\beta} X_{n+1}$, where the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are calculated on the basis of the observations for $j = 1, 2, \ldots, n$. Find the Variance of the forecast error (7 Marks)

b) Consider a general non-linear model $Y_j = \mu(x_j, \theta) + e_j$, $j = 1, 2, \ldots n$ and

$(x_j, y_j) \in \Re^{k+1} x \Re$, $E|Y| < \infty$, $\theta \in \Re^{k+1}$ are unknown parameters and $\mu(x_j, \theta)$ is the conditional mean at X_j, e_j is the error term. State and proof the necessary assumptions for the regression error e_j that enables estimation of $\mu(x_j, \theta)$ (8 Marks)