Question 1: Multiple Choice (20 marks, 2 marks per question)

Answer Question 1 using the multiple choice form provided.

1. Which of the following statements about intercept in the single linear regression model is correct?

   c. The intercept corresponds to the average value of outcome Y for the case in which the regressor X equals zero

2. In a regression model with two correlated regressors ($\rho_{X_1X_2} \neq 0$), if you exclude one of the regressors then...

   d. Statements b. and c. are both correct

3. You estimate a single linear regression:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

   and obtain a 90% confidence interval for β_1 of [-0.50,-0.10]. Which of the following statements is necessarily true?

   a. The value of R^2 for the model is greater than zero

4. The standard deviation of $\hat{\beta}_1$ increases...

   c. With greater dispersion of the error term u

5. What constitutes the dummy variable trap?

   c. A linear combination of one or more dummies in our regression is equal to the constant

6. Consider the following regression model:

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

   Assuming that the zero conditional mean assumption holds, which is the correct interpretation of β_0?

   c. It is the population mean for group 1 ($D_i = 0$)

7. Which of these statements is a direct consequence of heteroskedasticity?

   b. The OLS estimators are not BLUE

8. What problem does the classical measurement error in a regressor create for regression analysis?

   a. It causes regression coefficients to be smaller in magnitude than the population parameters

9. Consider the estimation results from the following regression (standard errors are in brackets):

$$Y_i = \underset{(12.702)}{420.21} + \underset{(80.503)}{129.31} X_{1i} - \underset{(7.929)}{30.34} X_{2i} + \underset{(0.002)}{0.01} X_{3i}; \qquad \bar{R}^2 = 0.53, SER = 10.11$$

What is the conclusion that can be drawn from these regression results?

e. None of the above

10. In order to evaluate whether Australia is becoming subject to more extreme heat-waves, you were given daily temperature data for years 1999 and 2019. Which descriptive statistic should you pay special attention to in your analysis of the yearly temperature distribution?

c. Skewness

Question 2: Short Answer Questions (20 marks)

Answer Questions 2-5 using exam booklets. You do not have to answer the questions in the order in which they are asked.

a. List the Least Squares assumptions corresponding to the multiple linear regression model. (3 marks)

Zero Conditional Mean

$(X_1, ...X_k, Y)$ are IID

Large outliers are unlikely

No perfect multicollinearity

b. Explain the difference between the probability distribution and the cumulative probability distribution of a discrete random variable (3 marks)

Probability distribution lists the probabilities corresponding to each of the possible outcomes for a given random variable.

Cumulative probability distribution quantifies the probability that a random variable is less than or equal to some value.

c. Consider the estimation results from the following regression (standard errors are in brackets):

$$Y_i = \underset{(0.50)}{5.59} + \underset{(0.33)}{4.53} X_{1i} - \underset{(0.35)}{3.34} X_{2i}$$

and the following statistics:

$$n = 100, s_Y = 3.28, SER = 1.73$$

Using this information, compute R^2, $\bar{R}^2$, and the overall homoskedasticity-only F-statistic corresponding to this regression. (5 marks)

$$SSR = (n-k-1) \cdot SER^2 = 97 \cdot 1.73 \cdot 1.73 = 290.31$$
$$TSS = (n-1) \cdot s_Y^2 = 99 \cdot 3.28 \cdot 3.28 = 1065.08$$
$$R^2 = 1 - \frac{SSR}{TSS} = 0.73$$
$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right) \frac{SSR}{TSS} = 0.72$$
$$F^{act} = \frac{\left(R^2_{unr} - R^2_{res}\right)/q}{\left(1-R^2_{unr}\right)/(n-k-1)} = \frac{(0.73-0)/2}{(1-0.73)/(97)} = 131.13$$

d. What is the consequence of independence of two random variables X and Y for the relationship between their joint and marginal probability distributions? (3 marks)

Their joint probability distribution is the product of their marginal probability distributions.

e. What is the interpretation of the standard error of the regression? (3 marks)

It is the typical size of the residual (regression error).

Alternatively: It is the standard error (sample standard deviation) of the residuals.

f. Explain the issue of simultaneous causality (3 marks)

It is a type of omitted variable bias which arises when causality "also runs backwards", i.e., the regressor X has a causal effect on the outcome variable Y and the outcome variable Y has a causal effect on the regressor X.

Question 3: Malawi Trial (20 marks)

You are a researcher evaluating the results of a randomized control trial in Malawi which aims at identifying the effects of insecticide-treated bed nets on the incidence of malaria among Malawian families. The experiment involved 2000 families who were followed over the period of one year. The experimenters randomly distributed bed nets to half of the families and tracked the outbreaks of malaria in all families over the period of observation. The resulting dataset contains 2000 observations of two variables: $BN_i = 1$ if the family received a bed net and 0 otherwise; and $M_i = 1$ if the family members contracted Malaria within the period of observation and 0 otherwise.
Below is a table of descriptive statistics corresponding to the two groups and their outcomes

	Average contraction rate of malaria in the sample, $\bar{M}$	Sample standard deviation of malaria contractions, s_M
Families without bed nets	0.37	0.48
Families with bed nets	0.32	0.47

a. Compute the 99% Confidence Interval for the population mean of malaria contraction rate among families with and without bed nets. (4 marks)

$$
\begin{aligned}
&Let\ Y = M,\\
&SE(\bar{Y}) = s_Y/\sqrt{n} = 0.48/\sqrt{1000} = 0.0152\\
&CI = [\bar{Y} - t_{0.01} \cdot SE(\bar{Y}), \bar{Y} + t_{0.01} \cdot SE(\bar{Y})]\\
&\quad = [0.37 - 2.576 \cdot 0.0152, 0.37 + 2.576 \cdot 0.0152]\\
&\quad = [0.3309, 0.4092]
\end{aligned}
$$

$$
\begin{aligned}
&SE(\bar{Y}) = s_Y/\sqrt{n} = 0.47/\sqrt{1000} = 0.0149\\
&CI = [\bar{Y} - t_{0.01} \cdot SE(\bar{Y}), \bar{Y} + t_{0.01} \cdot SE(\bar{Y})]\\
&\quad = [0.32 - 2.576 \cdot 0.0149, 0.32 + 2.576 \cdot 0.0149]\\
&\quad = [0.2816, 0.3584]
\end{aligned}
$$

b. To evaluate whether the intervention influenced the incidence of malaria outbreaks, conduct a formal three-step hypothesis test using an appropriate null hypothesis and a two-sided alternative hypothesis. Formulate the null and the alternative hypotheses, list each step of the test, and decide what is the outcome of this test (present the formulas and calculations of the relevant statistics, use 1% significance level). (6 marks)

$$
\begin{aligned}
&H_0 : \mu_1 - \mu_2 = 0\\
&H_1 : \mu_1 - \mu_2 \neq 0\\
&t_{act} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{SE_1^2 + SE_2^2}} = \frac{0.05}{\sqrt{0.0005}} = 2.3491\\
&t_{act} < t_{0.01} = 2.576
\end{aligned}
$$

Hence we fail to reject the null hypothesis of no influence at 1% significance level.

c. The same evaluation can be also done using a single linear regression model. Write down the specification of this regression model and outline this alternative testing procedure. (4 marks)

$$M_i = \beta_0 + \beta_1 BN_i + u_i$$

We can evaluate the claim by testing the statistical significance of the OLS parameter $\hat{\beta}_1$ corresponding to the regression model defined above. If the parameter is significant at the selected confidence level, we can conclude that the intervention had an effect on the incidence of malaria outbreaks.

d. Do you expect the error terms corresponding to this particular regression model to be homoskedastic, or heteroskedastic? Define the concept of heteroskedasticity (in words or using a formula), answer the question, and motivate your answer. (6 marks)

Heteroskedasticity arises when the variance of the error term u_i is dependent on the value of at least one of the regressors $X_1, ..., X_k$:

$$var(u|X) \neq var(u)$$

Since the bed nets were randomly distributed among the surveyed families, the dispersion of other factors influencing the malaria risks should not be related to the receipt of bed nets. The variance of u conditional on BN should therefore be the same as the unconditional variance of u, and we can expect the error terms to be homoskedastic.

Question 4: Smoking in Australia (20 Marks)

Australian Institute for Health and Welfare (AIHW) has approached you to study the determinants of smoking in Australia. You have been given a dataset which contains responses to a telephone survey which AIHW has been collecting over the last five years. Each survey respondent has been interviewed once, and the dataset contains the following information for a total sample of $n = 2421$ respondents:

$cigs_i$: number of cigarettes smoked by individual i per week

$educ_i$: number of years of education of individual i

$cigpric_i$: local price of cigarettes per pack (at the time of the interview)

$lcigpric_i$: natural logarithm of price of cigarettes per pack.

$cauca_i$: dummy variable equaling 1 if individual i is Caucasian

age_i: age of individual i in years

$agesq_i$: age of individual i in years, squared

$income_i$: annual income of individual i in AUD$

$lincome_i$: natural logarithm of $income_i$

The AIHW's leadership wants to know two things - first, they want to understand which personal attributes are associated with higher incidence of smoking. Second, they want to know whether people smoke less if they are facing higher prices of cigarettes. Figure 1 on the next page produces summary statistics for these data, and regression output from R for three different regressions that explore the two associations (the dependent variable in these regressions is $cigs_i$). Based on this output, answer the following questions. Throughout assume a 5% level of confidence for assessing statistical significance.

a. Suppose you ranked all people in the dataset by their income. What is the nominal value of income for the person who is in the exact middle of this ranking? (2 marks)

It is the median income value, 80,000.

b. Provide an interpretation of the regression coefficient estimate on $lincome_i$ in Regression 1. Comment on its sign, magnitude, and statistical significance (3 marks)

The coefficient estimate on $lincome_i$ is 1.239, it is positive and highly statistically significant (significant at the 5% confidence level). The interpretation of the coefficient is that a 1% change of annual income is associated with an increase of cigarette consumption by**0.01239**cigarettes per week, *holding other regressors fixed.*

c. The regression coefficient on age_i changes substantially between Regressions 1 and 2. Carefully explain what drives this large change. (4 marks)

The difference between the two regressions is that in Regression 2, we add the quadratic term *agesq* to the set of regressors. *age* and *agesq* are collinear, and this collinearity is affecting the magnitude of the coefficient on *age* in Regression 2.

Alternatively 1: The coefficient on *agesq* in Regression 2 is highly statistically significant (significant at the 5% confidence level), which suggests that the relationship

between smoking and age is quadratic. The linear *age* coefficient in Regression 1 is unable to capture this relationship, which reflects in the substantial change of the magnitude of the *age* coefficient in Regression 2.

Alternatively 2: Regression 1 models a linear relationship captured by the coefficient on *age*, whereas Regression 2 models a quadratic relationship captured by the coefficients on *age* and *agesq*. Thus, the age coefficients in Regressions 1 and 2 are not directly comparable.

d. Interpret the sign, magnitude and statistical significance of the regression coefficient estimate on $cauca_i$ in Regression 3. (3 marks)

The coefficient estimate on $cauca_i$ is -0.2049. The coefficient is negative, and the interpretation is that Caucasians respondents smoke on average -0.2049 less cigarettes per week than the respondents with different racial background, holding other regressors constant. The coefficient does not attain statistical significance at the conventional levels of confidence (not significant at the 5% confidence level).

e. Based on Regression 3, quantify the expected cigarette consumption among 17-year-olds with median values of each of income, race, and years of education who are subject to the median cigarette price. Work with the OLS coefficients rounded to the third decimal. At which age will the adult respondents with the same characteristics reach the same expected cigarette consumption as the 17-year-olds? (5 marks)

The consumption is computed by plugging the median values into the OLS equation:

$$2.712+0.782\cdot 17-0.009\cdot 17^2+0.754\cdot 11.290-0.205-0.514\cdot 12-2.901\cdot 3.419 = 5.626$$

This calculation yields the result that a 17-year-old with median characteristics smokes on average 5.626 cigarettes per week

The signs of the *age* and *agesq* coefficients tell us that the cigarette consumption is a concave function of respondents' age (it is first increasing in age, then it peaks and starts declining afterward). To find the age at which the average cigarette consumption equals the average consumption of 17-year-olds, we can proceed in two ways:

Either, we can use our OLS estimates to isolate the contribution of *age* and *agesq* regressors to the cigarette consumption at the age of 17, use this value as a constant in an age polynomial equation where the linear and quadratic coefficients are the OLS estimates of *age* and *agesq* coefficients, and find the second root of this polynomial (with the first one being 17).

$$\begin{aligned} &0.782 \cdot 17 - 0.009 \cdot 17^2 = 10.693 \\ &solve: \ 0.009 \cdot age^2 - 0.782 \cdot age + 10.693 = 0 \\ &x = \frac{-b+\sqrt{b^2-4ac}}{2a} = \frac{0.782+\sqrt{0.782^2-4\cdot 0.009\cdot 10.693}}{2\cdot 0.009} = 69.89 \end{aligned}$$

Alternatively, since the quadratic function is symmetric around its peak, we can compute the age when the cigarette consumption peaks, quantify the difference between the peak age and the age of 17, and add this difference to the peak age.

$$\beta_{age} + 2\beta_{agesq} \cdot age_{peak} = 0$$
$$age_{peak} = -\frac{\beta_{age}}{2\beta_{agesq}} = \frac{0.782}{2 \cdot 0.009} = 43.44$$
$$age_{samecon} = age_{peak} + (age_{peak} - 17) = 69.89$$

Accordingly, people reach the same cigarette consumption as 17-year-olds shortly prior to their 70^{th} birthdays

f. What is your response to the AIHW's second question? Is there an association between cigarette prices and demand for cigarettes? Would you say that this regression captures the causal effect of cigarette prices on cigarette demand? Motivate your answer! (3 marks)

There is no significant association between cigarette prices and demand for cigarettes (P-value corresponding to the coefficient *lcigpric* is 0.40>0.05). The regression is unlikely to capture the causal effect of cigarette prices on cigarette demand, because the demand affects the prices as well. This means that we are facing the problem of simultaneous causality, and our OLS estimates should not be interpreted causally.

Figure 1: Summary Statistics and Estimation Results for the Cigarette Regressions

```
                                q4_cig_output.txt
      educ          cigpric          cauca             age            income
 Min.   : 6.00   Min.   :22.00   Min.   :0.0000   Min.   :17.00   Min.   :  2000
 1st Qu.:10.00   1st Qu.:29.07   1st Qu.:1.0000   1st Qu.:28.00   1st Qu.: 50000
 Median :12.00   Median :30.53   Median :1.0000   Median :38.00   Median : 80000
 Mean   :12.47   Mean   :30.15   Mean   :0.8786   Mean   :41.24   Mean   : 77219
 3rd Qu.:13.50   3rd Qu.:31.59   3rd Qu.:1.0000   3rd Qu.:54.00   3rd Qu.:120000
 Max.   :18.00   Max.   :35.06   Max.   :1.0000   Max.   :88.00   Max.   :120000
      cigs           lincome           agesq         lcigpric
 Min.   : 0.000   Min.   : 7.601   Min.   : 289   Min.   :3.091
 1st Qu.: 0.000   1st Qu.:10.820   1st Qu.: 784   1st Qu.:3.370
 Median : 0.000   Median :11.290   Median :1444   Median :3.419
 Mean   : 8.686   Mean   :11.074   Mean   :1990   Mean   :3.403
 3rd Qu.:20.000   3rd Qu.:11.695   3rd Qu.:2916   3rd Qu.:3.453
 Max.   :80.000   Max.   :11.695   Max.   :7744   Max.   :3.557

Regression 1
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -3.746343   3.605091 -1.0392 0.2988244
age         -0.031315   0.013334 -2.3485 0.0189313 *
lincome      1.239360   0.324421  3.8202 0.0001367 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Regression 2
t test of coefficients:

               Estimate  Std. Error t value Pr(>|t|)
(Intercept) -6.38273256  3.61262217 -1.7668  0.07739 .
age          0.72575406  0.08092398  8.9683  < 2e-16 ***
agesq       -0.00833493  0.00085303 -9.7710  < 2e-16 ***
lincome      0.15607113  0.33058031  0.4721  0.63689
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Regression 3
t test of coefficients:

               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  2.71157331 12.59356936   0.2153   0.82954
age          0.78208284  0.08058579   9.7050 < 2.2e-16 ***
agesq       -0.00912285  0.00085147 -10.7143 < 2.2e-16 ***
lincome      0.75352761  0.34595379   2.1781   0.02949 *
cauca       -0.20488617  0.79582724  -0.2575   0.79685
educ        -0.51430203  0.09373105  -5.4870 4.515e-08 ***
lcigpric    -2.90086752  3.46614254  -0.8369   0.40272
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


                                    Page 1
```

Question 5: Forecasting crypto currency prices (20 Marks)

Westpac bank has hired you to develop time series models for crypto currency prices. They provide you with a dataset that has the daily prices in dollars for two crypto currencies – Bitcoin $bitcoin_t$ and Ethereum eth_t – over the last year and thus a total of $T = 365$ observations. For all parts of this question, only conduct hypothesis tests based on regressions with heteroskedasticity-robust standard errors.

a Westpac asks you to compute the elasticity of Ethereum with respect to Bitcoin because they want to know by how much the Ethereum price changes in percentage terms on a given day for a 1% change in the Bitcoin price that day. A Westpac analyst speculates that the elasticity is 1 and Westpac wants you to test this hypothesis.

Starting from the raw data in crypto_daily.csv, write down the pseudo-code in R you would develop to estimate this elasticity and test this hypothesis assuming a 5% level of significance. List the steps you would include in your code, and if it helps in describing your answer, you may state explicit R code though this is not necessary for obtaining full marks. Be precise and explicitly describe any variable transformation you would perform, regressions run in your code, hypothesis tests required, test statistics used, or any other calculations necessary. (4 marks)

Steps in the pseudo-code are as follows:

- Load the AER() package, load the data
- Compute 2 new logarithmic variables

$$log\_bitcoin_t = \ln(bitcoin_t)$$
$$log\_eth_t = \ln(eth_t)$$

- NB: an equivalently valid approach would be to use the daily percentage changes in these prices instead of the logarithms.
- Run the following regression:

$$log\_eth_t = \beta_0 + \beta_1 log\_bitcoin_t + u_t$$

- From the estimated log-log regression, β_1 is the elasticity of Ethereum prices with respect to Bitcoin prices.
- We then test, assuming heteroskedasticity, the linear hypothesis:

$$H_0 : \beta_1 = 1 \quad H_1 : \beta_1 \neq 1$$

- A p-value under 0.05 means we would reject H_0, while a p-value larger than 0.05 means we would fail to reject it.

b Figure 2 presents time series plots of the Ethereum prices eth_t and their first difference $\Delta eth_t = eth_t - eth_{t-1}$. What does it tell you about the stationarity of Ethereum prices and their daily changes? (4 marks)

Stationarity refers to when future values of a time series look like past value of a time series; that is, when a time series appears to be mean-reverting. The daily

Ethereum price change plot does not exhibit a persistent trend and suggests stationary. By contrast, the Ethereum price plot shows an initial upward trend, followed by a downward trend and another upward trend, suggesting the time series is not stationary.

c To estimate the relationship between Ethereum and Bitcoin prices you estimate the following ADL(1,2) model:

$$eth_t = \beta_0 + \beta_1 eth_{t-1} + \beta_2 bitcoin_{t-1} + \beta_3 bitcoin_{t-2} + u_t$$

The regression results are reported in Figure 3 below. How many observations are used in estimating this model? Briefly explain why this is the number of observations used in estimation. (2 marks)

The time series has T = 365 observations, and the maximum number of lags used in the ADL model is 2. This means that the first 2 observations are not used in estimation because the first two observations have lagged values that start before the first period t = 1 in the sample. Therefore, 363 observations are used in estimating the ADL model.

d Interpret the regression coefficients on 1st and 2nd lags of the Bitcoin price. Comment on their statistical significance at the 5% level. (4 marks)

A 1 dollar increase in lagged Bitcoin price leads to a 0.2-cent decrease in the Ethereum price, holding constant the 2-day-lagged Bitcoin and the 1-day lagged Ethereum prices. This effect is not statistically different from zero at the 5% level as the p-value is 0.11.

A 1 dollar increase in the 2-day-lagged Bitcoin price leads to a 0.3-cent increase in the Ethereum price, holding constant the 1-day-lagged Bitcoin and the 1-day-lagged Ethereum prices. This effect is statistically significant at the 5% level as the p-value is less than 0.05.

e The Ethereum price today is $220.59, the Bitcoin price today is $6,517.18, and was $6,281.2 yesterday and $6,371.3 the day before yesterday. Based on the ADL(1,2) model from Figure 3, what is your forecast for the Ethereum price tomorrow? Compute the 95% forecast interval assuming IID normal errors in the regression equation. Round to 3 digits after the decimal in conducting your calculations. (4 marks)

Using the coefficient estimates of the regression, the out-of-sample forecast is:

$$\widehat{eth}_{366} = 2.702 + 0.966 \cdot 220.59 - 0.003 \cdot 6517.18 + 0.003 \cdot 6281.2 = 215.084$$

The forecast Ethereum price for tomorrow is $215.084.

The SER in the regression is 8.929, which implies a 95% forecast interval of:

$$[215.084 - 1.96 \cdot 8.929, 215.084 + 1.96 \cdot 8.929] = [197.583, 232.585]$$

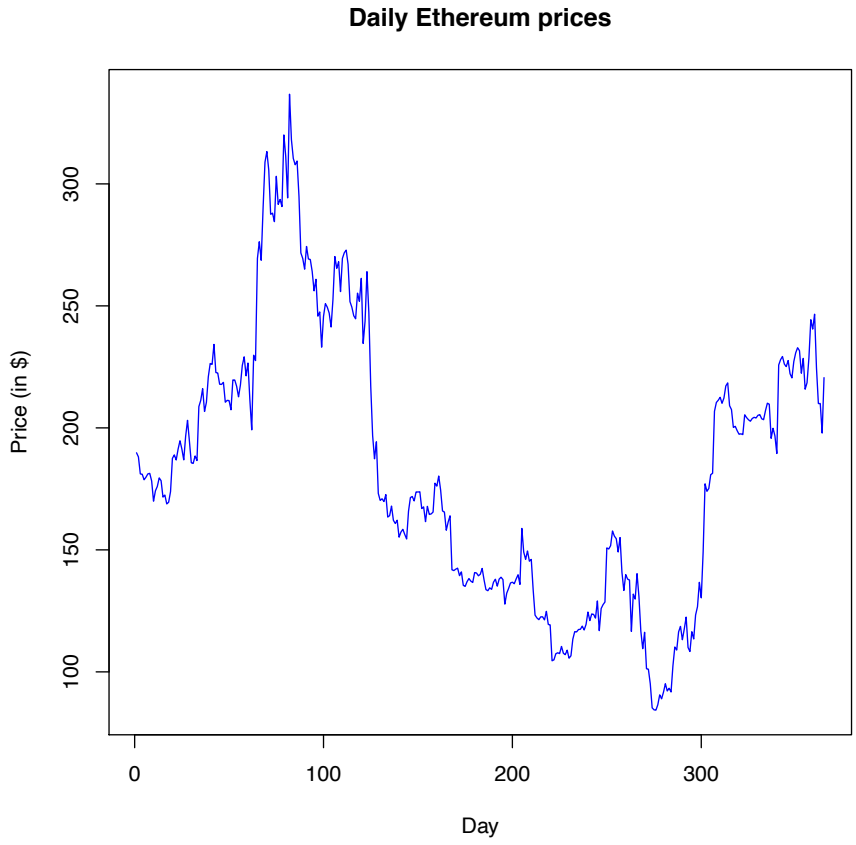
f The results of 3 tests are reported in Figure 4 based on the ADL(1,2) model from question c. Choose the appropriate set of results and describe the results of the Granger Causality test determining whether $bitcoin_t$ Granger Causes eth_t. Assume a 5% level of significance. (2 marks)

The Granger Causality test to determine whether $bitcoin_t$ Granger Causes eth_t has the following null and alternative hypotheses:

$H_0 : \beta_2 = 0 \ \& \ \beta_3 = 0 \quad H_1 : \beta_2 \neq 0 \ or \ \beta_3 \neq 0$

It is thus the last test in the figure. The corresponding test statistic for this joint test is F = 2.46 which has an F-distribution with df1=2 and df2=359 degrees of freedom. The test has a p-value larger than 0.05, implying we fail to reject the null hypothesis at the 5% level and thus cannot conclude that the Bitcoin price "Granger causes" the Ethereum price

Figure 2: Time Series Plots of Ethereum Prices and Changes of Ethereum Prices by Day



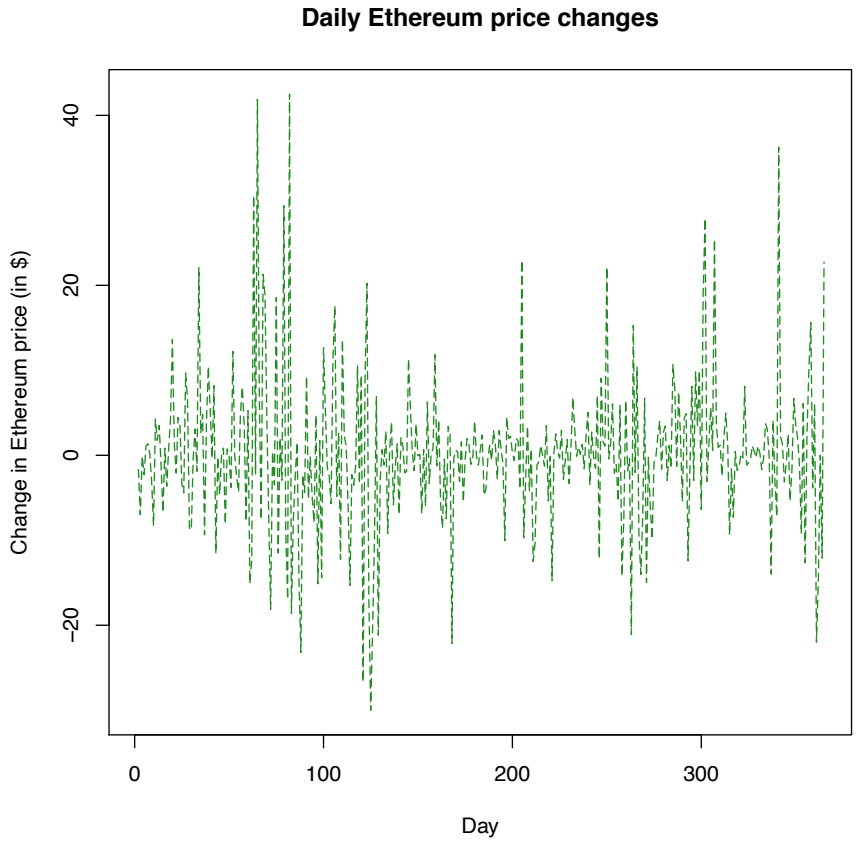


Figure 3: Ethereum price regression results

```
> reg1=lm(eth~eth_lag1+bitcoin_lag1+bitcoin_lag2,data=mydata)
> coeftest(reg1, vcov = vcovHC(reg1, "HC1"))

t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.701750   1.619441   1.668   0.0961 .
eth_lag1      0.965649   0.014392  67.095   <2e-16 ***
bitcoin_lag1 -0.002552   0.001571  -1.625   0.1051
bitcoin_lag2  0.003105   0.001556   1.995   0.0468 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(reg1)
# (note: regression output for reg1 with homoscedastic standard
# errors are omitted)

Residual standard error: 8.929 on 359 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.9747,   Adjusted R-squared:  0.9745
F-statistic:  4608 on 3 and 359 DF,  p-value: < 2.2e-16
```

Figure 4: Ethereum price regression tests

```
> linearHypothesis(reg1,c("eth_lag1=0"),vcov = vcovHC(reg1, "HC1"))
Linear hypothesis test

Hypothesis:
eth_lag1 = 0

Model 1: restricted model
Model 2: eth ~ eth_lag1 + bitcoin_lag1 + bitcoin_lag2

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    360
2    359  1 4530.4 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> linearHypothesis(reg1,c("eth_lag1=0","bitcoin_lag1=0"),vcov = vcovHC(reg
1, "HC1"))
Linear hypothesis test

Hypothesis:
eth_lag1 = 0
bitcoin_lag1 = 0

Model 1: restricted model
Model 2: eth ~ eth_lag1 + bitcoin_lag1 + bitcoin_lag2

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    361
2    359  2 2270.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> linearHypothesis(reg1,c("bitcoin_lag1=0","bitcoin_lag2=0"),vcov = vcovHC
(reg1, "HC1"))
Linear hypothesis test

Hypothesis:
bitcoin_lag1 = 0
bitcoin_lag2 = 0

Model 1: restricted model
Model 2: eth ~ eth_lag1 + bitcoin_lag1 + bitcoin_lag2

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
1    361
2    359  2 2.4573 0.0871 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

END OF EXAMINATION

Formula Sheet

Expected Values, Variances, Correlation

$E(c) = c$

$E(cx) = cE(x)$

$E(a + cx) = a + cE(x)$

$E(x + y) = E(x) + E(y)$

$E(c_1x + c_2y) = c_1E(x) + c_2E(y)$

$var(x) = \sigma^2 = E(x - E(x))^2$

$std(x) = \sigma = \sqrt{E(x - E(x))^2}$

$var(a + cx) = c^2var(x)$

$cov(x, y) = E[(x - E(x))(y - E(y))]$

$corr(x, y) = \rho = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$

$P(y = y_1|x = x_1) = \frac{P(x=x_1,y=y_1)}{p(X=x_1)}$

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

$var(\bar{Y}) = \frac{\sigma_Y^2}{n}$

$std(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}$

$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$

$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$

$SE(\bar{y}) = \frac{s_y}{\sqrt{n}}$

$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

$r_{xy} = \frac{s_{xy}}{s_x s_y}$

Logarithms

$x = \ln(e^x)$

$\frac{d\ln(x)}{dx} = \frac{1}{x}$

$\ln(1/x) = -\ln(x)$

$\ln(ax) = \ln(a) + \ln(x)$

$\ln(x/a) = \ln(x) - \ln(a)$

$\ln(x^a) = a\ln(x)$

$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}$ (approximately equal for small Δx)

Quadratic Formula

The solution to the quadratic equation:

$$ax + bx^2 + c = 0$$

where a, b, and c are constants can be computed by the quadratic formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Calculus

x^* that maximizes (minimizes) a strictly concave (convex) function, $f(x)$, solves $\frac{df(x)}{dx} = 0$

OLS Estimator for Single Linear Regression

$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

$\sigma^2_{\hat{\beta}_1} = \frac{1}{n} \frac{var((X_i - \mu_X)u_i))}{(var(X_i))^2}$

$\sigma^2_{\hat{\beta}_0} = \frac{1}{n} \frac{var(H_i u_i)}{(E(H_i^2))^2}$; where $H_i = 1 - (\frac{\mu_X}{E(X_i^2)})X_i$

$\hat{\beta}_1 \to \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}$

Testing Differences in Means

$H_0 : \mu_w - \mu_m = d_0; \quad vs. \quad H_1 : \mu_w - \mu_m \neq d_0$

$SE(\bar{Y}_w - \bar{Y}_m) = \sqrt{s_w^2/n_w + s_m^2/n_m}$

$t^{act} = \frac{(\bar{Y}_w - \bar{Y}_m) - d_0}{SE(\bar{Y}_w - \bar{Y}_m)}$

Single Hypothesis Testing in Regression Models

$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$

$H_0 : \beta_1 = \beta_{1,0}$ vs. $H_1 : \beta_1 \neq \beta_{1,0}$, p-value $= 2\Phi(-|t^{act}|)$

$H_0 : \beta_1 = \beta_{1,0}$ vs. $H_1 : \beta_1 < \beta_{1,0}$, p-value $= \Phi(t^{act})$

$H_0 : \beta_1 = \beta_{1,0}$ vs. $H_1 : \beta_1 > \beta_{1,0}$, p-value $= 1 - \Phi(t^{act})$

t^α is the critical value for a two-sided test with α significance level

$\alpha = 2\Phi(-|t^\alpha|)$

$(1 - \alpha)$ CI: $[\hat{\beta}_1 - t^\alpha SE(\hat{\beta}_1), \hat{\beta}_1 + t^\alpha SE(\hat{\beta}_1)]$

For testing means, replace β with μ_X and $\hat{\beta}$ with $\bar{X}$

Joint Hypothesis Testing in Regression Models

$H_0 : \beta_j = \beta_{j,0},\ \beta_m = \beta_{m,0}, \ldots$ for a total of q restrictions

H_1 : one or more of the q restrictions under H_0 does not hold

F-statistic for the test is distributed $F_{q,n-k-1}$

p-value $= \Pr[F_{q,n-k-1} > F^{act}] = 1 - G(F^{act}; q, n-k-1)$

For $q = 2$ restrictions, relationship between F-statistic and individual t-statistics for testing coefficients jointly equal 0:

$$F = \frac{1}{2}\left(\frac{(t_1^{act})^2 + (t_2^{act})^2 - 2\hat{\rho}_{t_1^{act},t_2^{act}} t_1^{act} t_2^{act}}{1 - \hat{\rho}_{t_1^{act},t_2^{act}}}\right)$$

Homoskedasticity–Only F-statistic with q restrictions

$$F^{act} = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n-k-1)} = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n-k-1)}$$

$F^{act}_{1,n-k-1} = (t^{act})^2$ if $q = 1$ restriction

Goodness of Fit in Regression Models

$SSR = \sum_{i=1}^n \hat{u}_i^2; \qquad ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$

$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2 = (n-1)s_Y^2$

$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$

$SER = s_{\hat{u}} = \sqrt{s^2_{\hat{u}}},\ s^2_{\hat{u}} = \frac{SSR}{n-k-1}$

$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s^2_{\hat{u}}}{s^2_Y}$

Nonlinear Regression Partial Effects, Standard Errors, and CIs

$E[Y|X_1, X_2, \ldots, X_k] = f(X_1, X_2, \ldots, X_k)$

$\Delta\hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k); \qquad SE(\Delta\hat{Y}) = \frac{|\Delta\hat{Y}|}{\sqrt{F}}$

$(1-\alpha)$ CI: $[\Delta\hat{Y} - t^\alpha SE(\Delta\hat{Y}), \Delta\hat{Y} + t^\alpha SE(\Delta\hat{Y})]$

Time Series Regression

$\text{RMSFE} = \sqrt{E[(Y_{T+1} - \hat{Y}_{T+1|T})^2]}$

$SE(Y_{T+1} - \hat{Y}_{T+1|T}) = \widehat{RMSFE} = \sqrt{var(\hat{u}_t)} = SER$

$(1-\alpha)$ CI: $[\hat{Y}_{T+1|T} - t^\alpha \times SE(Y_{T+1} - \hat{Y}_{T+1|T}), \hat{Y}_{T+1|T} + t^\alpha \times SE(Y_{T+1} - \hat{Y}_{T+1|T})]$

$\text{BIC}(K) = \ln\left[\frac{SSR(K)}{T}\right] + K\frac{\ln(T)}{T}$

$$\text{AIC}(K) = \ln\left[\frac{SSR(K)}{T}\right] + K\frac{2}{T}$$

F-statistic for the Granger Causality test has q and $T-\ell-p-1$ degrees of freedom, where q is the number of restrictions imposed under the null, T is sample size, ℓ is the maximum lag length included in the time series model, p is the number of parameters in the time series model excluding the constant.

Statistical Distribution Tables

Critical Values of the t Distribution

	Significance Level					
	1- Tailed:	*.10*	*.05*	*.025*	*.01*	*.005*
	2- Tailed:	*.20*	*.10*	*.05*	*.02*	*.01*
	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
	9	1.383	1.833	2.262	2.821	3.250
	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
	12	1.356	1.782	2.179	2.681	3.055
	13	1.350	1.771	2.160	2.650	3.012
	14	1.345	1.761	2.145	2.624	2.977
	15	1.341	1.753	2.131	2.602	2.947
	16	1.337	1.746	2.120	2.583	2.921
	17	1.333	1.740	2.110	2.567	2.898
	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
Degrees of Freedom	**21**	1.323	1.721	2.080	2.518	2.831
	22	1.321	1.717	2.074	2.508	2.819
	23	1.319	1.714	2.069	2.500	2.807
	24	1.318	1.711	2.064	2.492	2.797
	25	1.316	1.708	2.060	2.485	2.787
	26	1.315	1.706	2.056	2.479	2.779
	27	1.314	1.703	2.052	2.473	2.771
	28	1.313	1.701	2.048	2.467	2.763
	29	1.311	1.699	2.045	2.462	2.756
	30	1.310	1.697	2.042	2.457	2.750
	35	1.306	1.690	2.030	2.438	2.724
	36	1.306	1.688	2.028	2.434	2.719
	37	1.305	1.687	2.026	2.431	2.715
	38	1.304	1.686	2.024	2.429	2.712
	39	1.304	1.685	2.023	2.426	2.708
	40	1.303	1.684	2.021	2.423	2.704
	60	1.296	1.671	2.000	2.390	2.660
	90	1.291	1.662	1.987	2.368	2.632
	120	1.289	1.658	1.980	2.358	2.617
	∞	1.282	1.645	1.960	2.326	2.576

95th Percentile for the F-distribution F_{ν_1, ν_2}

		Numerator ν_1											
	ν_2/ν_1	**1**	**2**	**3**	**4**	**5**	**7**	**9**	**10**	**15**	**20**	**60**	∞
	1	161.45	199.50	215.71	224.58	230.16	236.77	240.54	241.88	245.95	248.01	252.2	254.31
	2	18.51	19.00	19.16	19.25	19.30	19.35	19.41	19.40	19.43	19.45	19.48	19.50
	3	10.13	9.55	9.28	9.12	9.01	8.89	8.81	8.79	8.70	8.66	8.57	8.53
D	**4**	7.71	6.94	6.59	6.39	6.26	6.09	6.00	5.96	5.86	5.80	5.69	5.63
e	**5**	6.61	5.79	5.41	5.19	5.05	4.88	4.77	4.74	4.62	4.56	4.43	4.37
n	**6**	5.99	5.14	4.76	4.53	4.39	4.21	4.10	4.06	3.94	3.87	3.74	3.67
o	**7**	5.59	4.74	4.35	4.12	3.97	3.79	3.68	3.64	3.51	3.44	3.30	3.23
m	**8**	5.32	4.46	4.07	3.84	3.69	3.50	3.39	3.35	3.22	3.15	3.01	2.93
i	**9**	5.12	4.26	3.86	3.63	3.48	3.29	3.18	3.14	3.01	2.94	2.79	2.71
n	**10**	4.96	4.10	3.71	3.48	3.33	3.14	3.02	2.98	2.85	2.77	2.62	2.54
a	**15**	4.54	3.68	3.29	3.06	2.90	2.71	2.59	2.54	2.40	2.33	2.16	2.07
t	**20**	4.35	3.49	3.10	2.87	2.71	2.51	2.39	2.35	2.20	2.12	1.92	1.84
o	**30**	4.17	3.32	2.92	2.69	2.53	2.33	2.21	2.16	2.01	1.93	1.74	1.62
r	**40**	4.08	3.23	2.84	2.61	2.45	2.25	2.12	2.08	1.92	1.84	1.64	1.51
	50	4.03	3.18	2.79	2.56	2.40	2.20	2.07	2.03	1.87	1.78	1.58	1.44
ν_2	**60**	4.00	3.15	2.76	2.53	2.37	2.17	2.04	1.99	1.84	1.75	1.53	1.39
	120	3.92	3.07	2.68	2.45	2.29	2.09	1.95	1.91	1.75	1.66	1.43	1.25
	∞	3.84	3.00	2.60	2.37	2.21	2.01	1.88	1.83	1.67	1.57	1.32	1.00

Critical Values for the Chi-Squared Distribution

Degrees of Freedom	Critical Values		
	1%	5%	10%
1	6.64	3.84	2.71
2	9.21	5.99	4.61
3	11.35	7.81	6.25
4	13.28	9.49	7.78
5	15.09	11.07	9.24
6	16.81	12.59	10.65
7	18.48	14.07	12.02
8	20.09	15.51	13.36
9	21.67	16.92	14.68
10	23.21	18.31	15.99
11	24.73	19.68	17.28
12	26.22	21.0	18.55
13	27.69	22.4	19.81
14	29.14	23.7	21.06
15	30.58	25.0	22.31
16	32.00	26.3	23.54
17	33.41	27.6	24.77
18	34.81	28.9	25.99
19	36.19	30.1	27.20
20	37.57	31.4	28.41