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

1. Which of the following is an example of time series data?
A) Data on the unemployment rates in different parts of a country during a year.
B) Data on the consumption of wheat by 200 households during a year.
C) Data on the gross domestic product of a country over a period of 10 years.
D) Data on the number of vacancies in various departments of an organisation in a particular month.
E) None of the above.
Answer: C

2. The probability of an outcome
A) is the number of times that the outcome occurs in the long run.
B) equals $M \times N$, where M is the number of occurrences and N is the population size.
C) is the proportion of times that the outcome occurs in the long run.
D) equals the sample mean divided by the sample standard deviation.
E) none of the above.
Answer: C

3. An estimator $\hat{\mu}_Y$ of the population value μ_Y is unbiased if
A) $\hat{\mu}_Y = \mu_Y$.
B) $\overline{Y}$ has the smallest variance of all estimators.
C) $\overline{Y} \xrightarrow{p} \mu_Y$.
D) $E(\hat{\mu}_Y) = \mu_Y$.
E) none of the above.
Answer: D

4. What does <u>not</u> change if you multiply the dependent variable by 100 and the explanatory variable by 100,000 in a single linear regression?
A) The OLS estimate of the slope.
B) The OLS estimate of the intercept.
C) The regression R^2.
D) The variance of the OLS estimator.
E) None of the above.
Answer: C

5. The t-statistic is calculated by dividing
A) the OLS estimator by its variance.
B) the slope by the standard deviation of the explanatory variable.
C) the estimator minus its hypothesized value by the standard error of the estimator.
D) the slope by 1.96.
E) None of the above.
Answer: C

6. When there are omitted variables in the regression, which are also determinants of the dependent variable, then
A) you cannot measure the effect of the omitted variable, but the estimator of your included variable(s) is (are) unaffected.
B) this has no effect on the estimator of your included variable because the other variable is not included.
C) this will always bias the OLS estimator of the included variable(s).
D) the OLS estimator(s) of the included variable(s) is (are) biased if the omitted variable is correlated with the included variable(s).
E) None of the above.
Answer: D

7. The overall regression *F*-statistic tests the null hypothesis that
A) all slope coefficients are zero.
B) all slope coefficients and the intercept are zero.
C) the intercept in the regression and at least one, but not all, of the slope coefficients is zero.
D) the slope coefficient of the variable of interest is zero, but that the other slope coefficients are not.
E) None of the above.
Answer: A

8. The interpretation of the slope coefficient in the model $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$ is as follows:
A) a 1% change in X is associated with a β_1 % change in Y.
B) a change in X by one unit is associated with a β_1 change in Y.
C) a change in X by one unit is associated with a 100 β_1 % change in Y.
D) a 1% change in X is associated with a change in Y of 0.01 β_1.
E) None of the above.
Answer: A

9. In the equation $\widehat{TestScore} = 607.3 + 3.85\ Income - 0.0423 Income^2$, which of the following income levels approximately results in the maximum test score?
A) 607.3.
B) 91.02.
C) 45.5.
D) cannot be determined without a plot of the data.
E) None of the above.
Answer: C

10. The AIC statistic:
A) is commonly used to test for heteroskedasticity in time series data
B) is the correct measure for testing goodness of fit in time series models
C) is an alternative to the BIC when sample size is small (typically T < 50)
D) helps in determining the number of lags to include in a time series model
E) None of the above.
Answer: D

Question 2: Short Answer Questions (20 marks)

1. What are the Law of Large Numbers and the Central Limit Theorem and why are they important for regression-based empirical analysis? (6 marks)

   Law of Large Numbers (LLN): as sample size n gets large, the sample OLS estimate $\hat{\beta}$ will be very close to the population value β with very high probability. LLN is important because it implies OLS is a consistent estimator of $\hat{\beta}$. In other words, the $\hat{\beta}$ tends to recover an estimate that is close to the true value of β for sufficiently large samples.
   Central Limit Theorem (CLT): As the sample size n gets large, the marginal distribution of $\hat{\beta}$ is approximately $N(\beta, \sigma_{\hat{\beta}}^2)$. The CLT is important because through it we can exploit the fact that $\hat{\beta}$ is approximately normally distributed around the true value β for large n to test hypotheses about the population value β, and to construct confidence intervals for the population value β.

2. Females, it is said, make 70% of the male wage. To investigate this phenomenon, you collect data on weekly earnings from 1,744 individuals, 850 females and 894 males. Next, you calculate their average weekly earnings and find that the females in your sample earned $346.98, while the males made $517.70.
   (a) How would you test whether this gender-wage gap is statistically significant? Give two approaches. (2 marks)

   We could test the difference in means using a t-statistic for comparison of two means. The alternative is to run a regression of earnings on a constant and a binary variable, which takes on the value of one for females and 0 otherwise. Using a t-test on the slope of the binary variable amounts to the same test as the difference in means.

   (b) A peer suggests that this is consistent with the idea that there is discrimination against females in the labor market. What is your response? (2 marks)

   Gender differences in attributes of the individuals, such as education, ability, and tenure with an employer, have not been taken into account. These attributes are potential determinants of wages too, and can create omitted variable bias if they also differ between men and women. Hence, in itself, this is weak evidence, at best, for discrimination.

   (c) Assuming, heroically, that education is constant across the 1,744 individuals, you consider regressing earnings on age and a binary variable for gender. You estimate two specifications initially:

$$\widehat{Earn} = 323.70 + 5.15 \times Age - 169.78 \times Female,\ R^2=0.13$$
$$\quad (21.18)\quad (0.55)\quad\quad (13.06)$$
$$\widehat{Ln(Earn)} = 5.44 + 0.015 \times Age - 0.421 \times Female,\ R^2=0.17$$
$$\quad (0.08)\quad (0.002)\quad\quad (0.036)$$

where *Earn* are weekly earnings in dollars, *Age* is measured in years, and *Female* is a binary variable, which takes on the value of one if the individual is a female and is zero otherwise. Interpret the linear regression carefully. (2 marks)

The linear specification suggests that for every additional year individuals receive $5.15 of additional weekly earnings on average. Females make $169.78 less than males at a given age. The intercept should not be interpreted, though it describes the predicted wage of a man at age zero. The regression explains 13 percent of the variation in earnings.

(d) Interpret the non-linear regression carefully. Based on the above regression results, for a given age, how much less do females earn on average? Should you choose the second specification on grounds of the higher regression R^2? (4 marks)

The log-linear specification says that earnings increase by 1.5 percent for every additional year of age. Females earn 42.1 percent less than males at a given age. Again, the intercept should not be interpreted, but it represents the predicted log wage of a man at age zero. The regression explains 17 percent of the variation in the log of earnings.
We should not choose the log specification over the linear one based on the higher regression R^2 since these cannot be compared because the dependent variable is different in both models (i.e., earnings vs log-earnings).

(e) Your peer points out to you that age-earning profiles typically take on an inverted U-shape. To test this idea, you add the square of age to your log-linear regression.

$$\widehat{Ln(Earn)} = \underset{}{3.04} + \underset{(0.18)}{0.147} \times Age - \underset{(0.009)}{0.421} \times Female - \underset{(0.033)}{0.0016}\ Age^2,\ \underset{(0.0001)}{R^2 = 0.28}$$

Are there strong reasons to assume that this specification is superior to the previous one? (2 marks)

The coefficient on Age^2 is statistically significant and including Age^2 substantially increases the regression R^2. These are good reasons to prefer this specification.
- 1 mark for preferring the model which includes Age^2
- 1 mark for explaining that this choice is based on the statistical significance of the coefficient of Age^2 and on the higher R^2.

(f) What other factors may play a role in earnings determination? (2 marks)

Students' answers will differ, but education, ability, regional differences, race, and professional choice can be mentioned.

Question 3: Lead mortality (40 marks)

Lead is toxic, particularly for young children, and for this reason government regulations severely restrict the amount of lead in our environment. But this was not always the case. In the early part of the 20th century, the underground water pipes in many U.S. cities contained lead, and lead from these pipes leaked into drinking water.

You will investigate the effect of these lead water pipes on infant mortality. You are provided with a dataset called Lead_Mortality.csv on 172 U.S. cities in the year 1900 that includes the following variables:

- *infrate*: Infant mortality rate (deaths per 1,000 infants in population)
- *lead*: Indicator = 1 if city had lead pipes.
- *pH*: Water pH (a measure of water acidity)
- *pop*: City population (in 100s)
- *age*: Average age of city population

You learn that the amount of lead leaked from lead pipes depends on the chemistry of the water running through the pipes. The more acidic the water (that is, the lower its pH), the more lead is leaked. You create a new variable (*lead_pH*) which is the interaction term *lead* x *pH* and you run the following regression:

$$infrate_i = \beta_0 + \beta_1 lead_i + \beta_2 pH_i + \beta_3 lead\_pH_i + u_i$$

Figures 1 and 2 present some summary statistics for variable *infrate* and the regression results from R-Studio. For all parts of question 3, only conduct hypothesis tests based on regressions with heteroskedasticity-robust standard errors. Also, round to three decimal points for all calculations.

a. Using information from Figure 1, compute the 95% confidence intervals of the average infant mortality rate for cities with lead pipes and for cities with non-lead pipes. Note that there are 55 cities with lead pipes (lead=1) and 172 - 55 = 117 cities with non-lead pipes in the sample. (4 marks)

The average mortality rate is 3.812 (SD 1.478) in cities with non-lead pipes and 4.033 (SD 1.531) in cities with lead pipes.
The respective standard errors of the sample means are:
SE (lead=0) = SD(lead = 0) / $\sqrt{\#\text{ cities with non} - \text{lead pipes}}$ = 1.478 / $\sqrt{117}$ = 0.137
SE (lead=1) = SD(lead = 1) / $\sqrt{\#\text{ cities with lead pipes}}$ = 1.531 / $\sqrt{55}$ = 0.206
The 95% confidence intervals are:
non-lead pipe cities: [3.812 – 1.96*0.137, 3.812 + 1.96*0.137] = [3.543, 4.081]
lead pipe cities: [4.033 – 1.96*0.206, 4.033 + 1.96*0.206] = [3.629, 4.437]

b. Turning to Figure 2, what is the 95% confidence interval for β_3? (2 marks)

The 95% CI is: [-0.569-1.96*0.281, -0.569+1.96*0.281] = [-1.120, -0.018]

c. What is the overall regression *F*-statistic for the regression in Figure 2, and what are its corresponding degrees of freedom. Interpret the statistical significance of this test at the 1% level, and its implication for the model. (2 marks)

The overall regression *F*-statistic is 20.974. It has an F-distribution with df1=3 and df2=168. It has a p-value of less than 0.01, which implies that we reject the null at the 1% level that all the regression coefficients jointly equal zero. This implies that we reject the null that the regression model is statistically useless for explaining infant mortality.

d. Interpret the signs of the coefficient estimates on *lead*, *pH*, and *lead_pH* in Figure 2. Does the effect of *lead* on infant mortality depend on *pH*? Discuss statistical significance at the 5% level. (4 marks)

Cities with lead pipes have higher mortality rates than cities with non-lead pipes. Cities with higher water pH have lower mortality rates. The effect of *lead* on infant mortality depends on *pH* (or, equivalently, the effect of *pH* depends on *lead)*. The positive effect of having lead pipes on mortality rates decreases with the water pH (or, equivalently, the negative effect higher water pH on mortality rates increases with lead pipes). All three coefficients are statistically significant at the 5% level (or, more precisely, the p-values of the tests of the null hypotheses that each coefficient is equal 0 are all less than 0.05).

e. The average value of *pH* in the sample is 7.323. At this pH level, what is the estimated effect of *lead* on infant mortality implied by Figure 2? Interpret this result. (2 marks)

The estimated effect of *lead* on infant mortality at the average *pH* level is: $4.618 - 0.569 \times 7.323 = 0.451$. At the average water pH level of 7.323, cities with lead pipes have infant mortality rates 0.451 infants per 1,000 infants higher than cities with non-lead pipes.

f. Carefully explain the steps required to compute the standard error of the estimated effect of *lead* on infant mortality at the average *pH* level that you derived in question e. above. How would you then derive the 90% confidence interval? (4 marks)

We can compute the standard error for the effect in two steps:
Step 1: Compute the *F*-statistic corresponding to the following test:

$$H_0: \hat{\beta}_1 + \hat{\beta}_3 \times 7.323 = 0 \text{ vs } H_1: \hat{\beta}_1 + \hat{\beta}_3 \times 7.323 \neq 0$$

Call this *F*-statistic F^{act}.

Step 2: Then compute the standard error as follows:

$$SE(\Delta \widehat{infrate}) = \frac{|\Delta \widehat{infrate}|}{\sqrt{F^{act}}} = \frac{0.451}{\sqrt{F^{act}}}$$

Step 3: The 90% confidence interval is:

$$\left[0.451 - 1.64 \times \frac{0.451}{\sqrt{F^{act}}}\,, 0.451 + 1.64 \times \frac{0.451}{\sqrt{F^{act}}}\right]$$

Note: Since this is a single restriction test, one could replace F^{act} by $(t^{act})^2$ in Steps 1 through 3 above without changing the results, where t^{act} is the actual t-statistic from the single restriction t-test in Step 1.

g. The standard deviation of *pH* is 0.691. Suppose that we could decrease the pH level of water to one standard deviation lower than the average level of *pH* in the sample; what would the estimated effect of *lead* on infant mortality be? What if the *pH* level were one standard deviation higher than the average value? Interpret these results.
(4 marks)

At one standard deviation lower than the average, $pH = 7.323 - 0.691 = 6.632$. At this level, the estimated effect of *lead* on infant mortality is: $4.618 - 0.569 \times 6.632 = 0.844$. Similarly, at one standard deviation higher, the estimated effect of *lead* on infant mortality is: $4.618 - 0.569 \times (7.323 + 0.691) = 0.058$.
At a pH level one standard deviation below the average, lead pipes increase mortality by 0.844 infants per 1,000 infants. At a pH level one standard deviation above the average the increase is much smaller, at only 0.058 per 1,000 infants.

h. You are concerned about omitted variable bias. Building further on your regression model, you now estimate a second regression model:

$$infrate_i = \beta_0 + \beta_1 lead_i + \beta_2 pH_i + \beta_3 lead\_pH_i + \beta_4 age_i + \beta_5 age_i^2 + \beta_6 \log(pop) + u_i$$

The regression results are reported in Figure 3. Comparing regression results in Figures 2 and 3, should you be concerned about omitted variable bias? (2 marks)

The coefficient estimates on *lead*, *pH*, and *lead_pH* do not change almost at all in size or statistical significance between Figures 2 and 3. Hence, we should not be concerned about omitted variable bias due to the omission of *age*, *age*^{2} and *log(pop)* in Figure 2.

i. Interpret the coefficient on log(pop) in Figure 3 and comment on whether it is statistically significant at the 10% level. (2 marks)

An 1% increase in the population size leads to an increase in infant mortality rate by 0.01*0.088 = 0.00088 infants per 1,000 infants (>0.001 is also fine here). This (tiny) effect is not statistically different from 0 at the 10% level as the p-value is larger than 0.1.

j. Looking at the coefficient estimates on *age* and *age*^{2} in Figure 3, is the model depicting a U-shape or an inverted U-shape relationship between *infrate* and *age*? At what value

of *age* is infant mortality predicted to be the lowest? What does it say about the sign of the effect of *age* on infant mortality in your sample? (4 marks)

As the coefficient on age^2 is positive the model describes a U-shape relationship between *infrate* and *age*. By taking the derivative of the estimated regression function with respect to *age*, setting the derivative equal to 0, and then solving for *age* we obtain the value of *age* at which the marginal effect of age on *infrate* is smallest:

$$-1.002 + 0.013 \times 2 \times age = 0$$

$$age = \frac{-1.002}{-0.026} = 38.538$$

And, since we've established that there is a U-shape relationship between *infrate* and *age* with a minimum at 38.538, the effect of *age* on *infrate* must be negative for countries with average age below 38.538 and positive for countries with average age above 38.538.

k. What test is being conducted on Figure 4? Describe the outcome of the test using a 5% significance level, noting the relevant test statistic and degrees of freedom (if necessary). Interpret the result of this test. Based on this test alone, would you prefer the first model (Figure 2) or the second model (Figure 3)? (4 marks)

The test is $H_0: (\beta_4 = 0, \beta_5 = 0 \; and \; \beta_6 = 0)$ vs $H_1: (\beta_4, \beta_5 \; or \; \beta_6 \neq 0)$ or, in words, whether age, age squared, and the log of the population are statistically significant predictors of the infant mortality rate over and above lead pipes, water PH, and their interaction.
The corresponding test statistic for this joint test is $F = 15.415$ which has an F-distribution with df1=3 and df2=165 degrees of freedom. The test has a p-value smaller than 0.05, implying we reject the null hypothesis at the 5% level. In words, we reject the hypothesis that there is no relationship between *infrate* and *age*, age^2 and *log(pop)*, after controlling for *lead* and *pH* and their interaction. Based on this test, we would prefer the second model.

l. Would the adjusted R^2 presented in Figures 2 and 3 make you change your mind? (2 marks)

The adjusted R^2 increases from 0.259 in Figure 2 to 0.399 in Figure 3. This is a substantial increase that leads us to prefer the second model, a result in line with the results of the F-test from part *k*.

m. Using only the raw data provided, provide the **pseudo-code**[1] for an R program (e.g., the .R code) that you would write in R-Studio for estimating the elasticity of *infrate* with respect to *pop*, and its standard error, based on the model in Figure 3 (from part h) at the median of average age (27.59).
Your pseudo-code can be written in a series of bullet points. It should explicitly state all steps required in R-script to generate this test. You do not need to cite explicit R

[1] A pseudo-code consists of all the steps you would take in an R program for conducting an analysis or calculation. It is primarily written in words and not R commands or syntax.

commands, syntax, or equations, but you may do so if it helps clarify what each part of your pseudo-code does. (4 marks)

Line 1. Compute the natural log of *infrate, log_infrate = log(inf).*

Line 2. Compute the interaction *lnpop_age = log(pop) x age*

Line 3. Run the regression

$$\log\_infrate_i = \beta_0 + \beta_1 lead_i + \beta_2 pH_i + \beta_3 lead\_pH_i + \beta_4 age_i + \beta_5 age_i^2 + \beta_6 \log(pop) + \beta_7 lnpop\_age + u_i$$

using the *lm()* command

Line 4. Compute the elasticity at age = 27.59 using the estimated regression model:

$$\Delta\log\_infrate = \hat{\beta}_6 + \hat{\beta}_7 \times 27.59$$

Line 5. Compute the F-statistic corresponding to the following test:

$$H_0: \hat{\beta}_6 + \hat{\beta}_7 \times 27.59 = 0 \ \ vs\ H_1: \hat{\beta}_6 + \hat{\beta}_7 \times 27.59 \neq 0$$

Call this F-statistic F^{act}.

Line 6. Compute the standard error for the elasticity as:

$$SE(\Delta loginfrate) = \frac{|\hat{\beta}_6 + \hat{\beta}_7 \times 27.59|}{\sqrt{F^{act}}}$$

Figure 1. Mean and S.D. of *infrate* by type of water pipe

```
> mean(infrate[lead==0])
[1] 3.811679
> sd(infrate[lead==0])
[1] 1.477588
>
> mean(infrate[lead==1])
[1] 4.032576
> sd(infrate[lead==1])
[1] 1.530873
```

Figure 2. Infant mortality regression output 1

```
> lead_pH=lead*pH
> reg1=lm(infrate~lead+pH+lead_pH,data=mydata)
> coeftest(reg1, vcov = vcovHC(reg1, "HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  9.18904    1.50494  6.1059 6.866e-09 ***
lead         4.61798    2.07614  2.2243 0.0274596 *
pH          -0.75179    0.20953 -3.5879 0.0004369 ***
lead_pH     -0.56862    0.28084 -2.0247 0.0444778 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(reg1)$adj.r.squared
[1] 0.2588579
> waldtest(reg1, vcov = vcovHC(reg1, "HC1"))
Wald test

Model 1: infrate ~ lead + pH + lead_pH
Model 2: infrate ~ 1
  Res.Df Df      F    Pr(>F)
1    168
2    171 -3 20.974 1.366e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3. Infant mortality regression output 2

```
> age2=age*age
> reg2=lm(infrate~lead+pH+lead_pH+age+age2+log(pop),data=mydata)
> coeftest(reg2, vcov = vcovHC(reg2, "HC1"))

t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.212831  12.437970  2.1075  0.03659 *
lead         4.362605   1.884854  2.3146  0.02187 *
pH          -0.738779   0.182627 -4.0453 8.01e-05 ***
lead_pH     -0.570530   0.253902 -2.2470  0.02596 *
age         -1.002287   0.878252 -1.1412  0.25543
age2         0.013322   0.015423  0.8638  0.38897
log(pop)     0.088242   0.085663  1.0301  0.30447
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(reg2)$adj.r.squared
[1] 0.3994807
> waldtest(reg2, vcov = vcovHC(reg2, "HC1"))
Wald test

Model 1: infrate ~ lead + pH + lead_pH + age + age2 + log(pop)
Model 2: infrate ~ 1
  Res.Df Df      F    Pr(>F)
1    165
2    171 -6 20.343 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4. Infant mortality regression test

```
> linearHypothesis(reg2,c("age=0","age2=0","log(pop)=0"),vcov = vcovHC(reg2, "HC1"))
Linear hypothesis test

Hypothesis:
age = 0
age2 = 0
log(pop) = 0

Model 1: restricted model
Model 2: infrate ~ lead + pH + lead_pH + age + age2 + log(pop)

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    168
2    165  3 15.415 6.907e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 4: Forecasting inflation (20 marks)

The Reserve Bank of Australia has hired you to develop time series models for the inflation rate. They provide you with a time series for just one variable, $PCEC_t$, which is the price index in quarter t. These data are provided from 1963:Q1 to 2013:Q4 for a total of T = 204 observations.
For all parts of question 4, only conduct hypothesis tests based on regressions with heteroskedasticity-robust standard errors. Also, round to three decimal points for all calculations.

a. You compute the inflation rate, $Infl_t = 100 \times [\ln(PCEC_t) - \ln(PCEC_{t-1})]$. What are the units of measurement of $Infl_t$? Explain. (2 marks)

We have: $\ln(PCEC_t) - \ln(PCEC_{t-1}) \approx \frac{\Delta PCEC_t}{PCEC_{t-1}}$
100 times this ratio thus corresponds to the percentage change from the previous quarter, or the quarterly inflation rate measured in percentage points.

b. Figure 5 plots the autocorrelations of the inflation rate and of the first difference in inflation rate ($\Delta Infl_t = Infl_t - Infl_{t-1}$). What does this tell you about the persistence of inflation? (4 marks)

The inflation rate has a very persistent time series, with a strong positive autocorrelation decreasing only slowly as the lag increases. The autocorrelation function (ACF) for the first differences shows that first differencing essentially removes the persistence in the time series. In other words, the inflation rate has a very persistent time series but quarter-on-quarter changes in the inflation rate (i.e., inflation growth) show very little persistence.

c. You now estimate autoregressive models of order 1 and 2 (AR(1) and AR(2) models) for the inflation rate. Regression results are reported in Figure 6. Interpret the regression coefficients on 1st and 2nd lags in the AR(2) model of Figure 6. Comment on their statistical significance at the 1% level. (4 marks)

A 1 percentage point increase in lagged inflation rates lead to a 0.672 percentage point increase in the current inflation rate, holding constant twice-lagged inflation (i.e. inflation 2 quarters earlier). A 1 percentage point increase in the twice-lagged inflation rate leads to a 0.191 percentage point increase in the current inflation rate, holding constant lagged inflation. Both coefficients are statistically different from 0 at the 1% level as both p-values are less than 0.01.

d. Figure 6 also reports the BIC and AIC for both regressions. Based on these information criteria, which model should you prefer? (2 marks)

The AR(2) model minimises both the BIC and AIC and is thus the preferred model.

e. Using information from Figure 6, what is your best guess of the Residual Mean Square Forecast Error (RMFSE) for the AR(2) model? Explain the required condition for your guess to be valid. (4 marks)

The SER can be a good approximation of the RMFSE. If this is the case, our best approximation of the RMSFE is 0.351 percentage points. This approximation will be good if most of the forecast error comes from future error, and little of it comes from the uncertainty from estimating the unknown coefficients of the AR(2) model.

f. You now consider the potential effect of seasonality and include quarterly dummies in your AR(2) model (*q2* = 1 if *Quarter* = 2 and 0 otherwise, *q3* = 1 if *Quarter* = 3 and 0 otherwise, *q4* = 1 if *Quarter* = 4 and 0 otherwise). Regression results are reported in Figure 7. Comment on the statistical significance of the individual quarterly dummies at the 10% level. Interpret only the regression coefficients that are statistically different from 0. (2 marks)

Quarterly dummies *q2* and *q3* have p-values above 0.1 and are thus not statistically significant at the 10% level. The dummy *q4* is statistically different from 0 at the 10% level, with a p-value of 0.097, indicating that inflation is lower by 0.128 percentage point in the 4th quarter of the year compared to the first quarter of the year, holding inflation rate in the two previous quarters fixed.

g. The inflation rate was 0.173% in 2013:Q4, 0.475% in 2013:Q3, -0.029% in 2013:Q2 and 0.269% in 2013:Q1. Based on the AR(2) model from Figure 7, what is your forecast for the inflation rate in 2014:Q1? (2 marks)

Using the coefficient estimates of the regression from Figure 7:

$$Infl_{2014:Q1} = 0.165 + 0.690 \times\ 0.173 + 0.174 \times 0.475 = 0.367$$

The forecast inflation rate for 2014:Q1 is 0.367%.

Figure 5. Autocorrelations in inflation rates

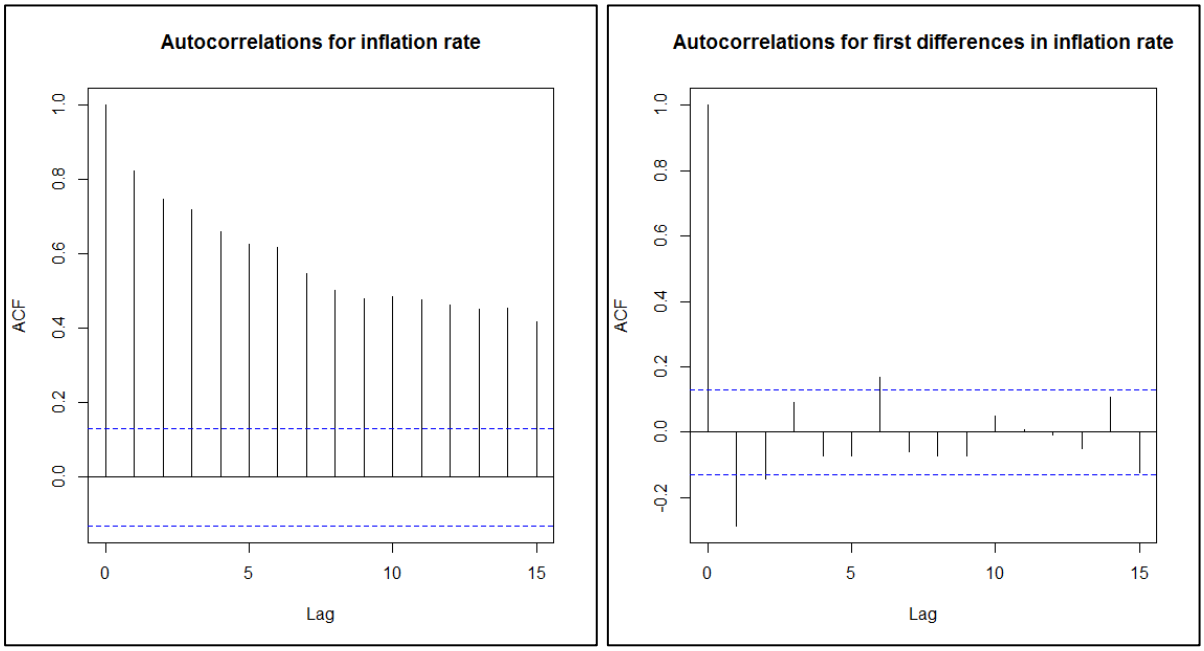

Figure 6. Inflation regression output 1

```
> reg1=lm(infl~infl_lag1,data=mydata)
> coeftest(reg1, vcov = vcovHC(reg1, "HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 0.147134   0.041396  3.5543 0.0004718 ***
infl_lag1   0.831424   0.043236 19.2300 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>
>
> reg2=lm(infl~infl_lag1+infl_lag2,data=mydata)
> coeftest(reg2, vcov = vcovHC(reg2, "HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 0.119215   0.043494  2.7410  0.006678 **
infl_lag1   0.672156   0.058938 11.4045 < 2.2e-16 ***
infl_lag2   0.191355   0.068591  2.7898  0.005782 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>
> K1=2
> SER1=sd(residuals(reg1))
> ssr1=sum(reg1$resid^2)
> BIC1=log(ssr1/T)+K1*(log(T)/T)
> AIC1=log(ssr1/T)+K1*(log(2)/T)
>
> K2=3
> SER2=sd(residuals(reg2))
> ssr2=sum(reg2$resid^2)
> BIC2=log(ssr2/T)+K2*(log(T)/T)
> AIC2=log(ssr2/T)+K2*(log(2)/T)
>
> sprintf("BIC of reg1: %f", BIC1[1])
[1] "BIC of reg1: -2.005760"
> sprintf("BIC of reg2: %f", BIC2[1])
[1] "BIC of reg2: -2.016902"
> sprintf("AIC of reg1: %f", AIC1[1])
[1] "AIC of reg1: -2.051103"
> sprintf("AIC of reg2: %f", AIC2[1])
[1] "AIC of reg2: -2.084916"
> sprintf("SER of reg1: %f", SER1[1])
[1] "SER of reg1: 0.358261"
> sprintf("SER of reg2: %f", SER2[1])
[1] "SER of reg2: 0.351657"
```

Figure 7. Inflation regression output 2

```
> reg3=lm(infl~infl_lag1+infl_lag2+q2+q3+q4,data=mydata)
> coeftest(reg3, vcov = vcovHC(reg3, "HC1"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.165130   0.057055  2.8942  0.004227 **
infl_lag1    0.690027   0.057933 11.9108 < 2.2e-16 ***
infl_lag2    0.174432   0.067606  2.5801  0.010601 *
q2          -0.055402   0.071621 -0.7736  0.440118
q3          -0.003446   0.062268 -0.0553  0.955923
q4          -0.128206   0.076897 -1.6672  0.097048 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

END OF EXAMINATION