Question 1: Multiple Choice (10 marks)

1. Consider a random variable Y. What is the difference between the sample average $\bar{Y}$ and the population mean?

   a. Both the population mean and the sample average $\bar{Y}$ are true measures of the central tendency of the distribution of Y.

   b. The sample average $\bar{Y}$ is a true measure of the central tendency of the distribution of Y whereas the population mean is an estimator of the sample average.

   c. Both the population mean and the sample average $\bar{Y}$ are estimators of the central tendency of the distribution of Y.

   d. The population mean is a true measure of the central tendency of the distribution of Y whereas the sample average $\bar{Y}$ is an estimator of the population mean.

2. In which circumstance are you necessarily in a dummy variable trap with a multiple linear regression model?

   a. You have a regressor that always equals 1

   b. You have two dummy variables that are highly but not perfectly correlated

   c. You have a collection of dummy variables whose sum always equals the value of another regressor

   d. When two or more dummy variables sum to one

3. For a single restriction ($q = 1$) in a regression, the F-statistic

   a. is the square of the t-statistic

   b. has a critical value of 3.84

   c. is the square root of the t-statistic

   d. is normally distributed as sample size n becomes large

4. You compute a sample mean of $\bar{X} = 14$ with a standard error of $SE(\bar{X}) = 4$. What is the 90% confidence interval for the sample mean?

   a. [6.16, 21.84]

   b. [3.68, 24.32]

   c. [7.40, 20.60]

   d. [4.68, 23.32]

5. Which of the following is correct about the value of $\bar{R}^2$ in a multiple linear regression model

   a. It is bounded between 0 and 1

   b. It can be larger than R^2

   c. It can be negative

   d. It strictly increases as sample size n grows

6. Consider the following multiple linear regression:

$$\ln(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Which of the following is the correct interpretation of β_2?

   a. It is the elasticity of Y with respect to X_2, holding X_1 fixed
   b. A 1-unit increase in X_2 yields a $100 \times \beta_2$ % increase in Y, holding X_1 fixed
   c. A 1% increase in X_2 yields a 1 unit increase in Y, holding X_1 fixed
   d. A β_2 unit increase in X_2 yields a 1% increase in Y holding X_2 fixed

7. Consider the estimates from the single linear regression of $Y = \beta_0 + \beta_1 X + u$:

$$\widehat{Y}_i = \underset{(8.99)}{35.15} + \underset{(2.41)}{32.12} X_{i1}, \bar{R}^2 = 0.32$$

What is the t-statistic for the test of the null that $\hat{\beta}_0 = 40$ versus the alternative that $\hat{\beta}_0 \neq 40$, and would you reject or fail to reject the null at the 1% level of significance?

   a. 13.328, reject the null
   b. -3.270, reject the null
   c. -0.539, fail to reject null
   d. 3.910, reject the null

8. Consider the following polynomial regression model of degree r:

$$Y_i = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \ldots + \beta_r X_i^r + u_i$$

Which of the following states the null hypothesis that the regression function is linear and the alternative that the regression function is nonlinear:

   a. $H_0 : \beta_2 = 0, \beta_3 = 0, \ldots, \beta_r = 0$ vs. H_1 : all of $\beta_j \neq 0$, $j = 2, \ldots, r$
   b. $H_0 : \beta_r = 0$ vs. $H_1 : \beta_r \neq 0$
   c. $H_0 : \beta_2 + \beta_3 + \ldots + \beta_r = 0$ vs. $H_1 : \beta_2 + \beta_3 + \ldots + \beta_r \neq 0$
   d. $H_0 : \beta_2 = 0, \beta_3 = 0, \ldots, \beta_r = 0$ vs. H_1 : at least one of $\beta_j \neq 0$ for $j = 2, \ldots, r$

9. The AIC statistic:

   a. is commonly used to test for heteroskedasticity in time series data
   b. is the 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 determinings the number of lags to include in a time series model

10. Consider the following multiple linear regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

where X_{1i} is a variable of interest and X_{2i} is the control variable. Conditional mean independence requires that:

   a. $E[u_i] = E[u_i|X_{2i}]$
   b. $E[u_i|X_{1i}, X_{2i}] = E[u_i|X_{2i}]$
   c. $E[u_i|X_{1i}, X_{2i}] = 0$
   d. $E[u_i|X_{1i}] = E[u_i|X_{2i}]$

Question 2: Short Answer Questions (10 Marks)

a. What are the Law of Large Numbers and the Central Limit Theorem and why are they important for regression-based empirical analysis? (3 marks)

b. Suppose you had a dataset consisting of postcode-level data on average household earnings ($Earnings_i$), share of households with bachelor's degrees ($Educ_i$), and local crime rates ($Crime_i$). You run a single linear regression of $Earnings_i$ on $Educ_i$, and a separate multiple linear regression of $Earnings_i$ on $Educ_i$ and $Crime_i$ and obtain the following results:

$$\widehat{Earnings}_i = \underset{(12.72)}{25.90} + \underset{(2.78)}{11.83} Educ_i$$

$$\widehat{Earnings}_i = \underset{(6.55)}{14.11} + \underset{(3.55)}{8.49} Educ_i - \underset{(0.324)}{0.961} Crime_i$$

Comparing these regression results, carefully describe the sign of the omitted variable bias in the first regression, and explain how this bias could arise. (3 marks)

c. Consider the following single linear regression model:

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$$

Prove that β_1 is the elasticity of Y with respect to X. (4 marks)

Question 3: Pollution and Carbon Taxes (10 Marks)

The United States government has approached you to evaluate the impact of carbon taxes on market-level pollution. You are provided a dataset called dat_pol.csv that includes the following variables from $n = 7352$ markets[1] across the United states:

air_i: continuous air quality measure in city i based on the amount of sulphur dioxide in the air ranging between 1 and 10 (10=very little pollution, 1=extreme pollution)

$plants_i$: number of manufacturing plants in city i

$pulp_i$: dummy variable equalling 1 if market i has at least one pulp and paper mill, and equals 0 otherwise

$repub_i$: dummy variable equalling 1 if market i is in a state where the Republican party is currently in power, and equals 0 otherwise

pop_i: population of market i (in terms of 1000's of people)

The following regression is estimated to investigate the determinants of air pollution across markets:

$$\ln(air_i) = \beta_0 + \beta_1 plants_i + \beta_2 pulp_i + \beta_3 repub_i + \beta_4 pop_i + u_i$$

Figures 1 and 2 on the next page respectively present summary statistics for the dataset 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. Based on the output in Figures 1 and 2 on the next page, please answer the following questions:

a. Interpet the coefficient estimate on $plants_i$ in Figure 2, and comment on whether it is statistically significantly different from 0 at the 5% level. (1 mark)

b. Interpet the coefficient estimate on $pulp_i$ in Figure 2, and comment on whether it is statistically significantly different from 0 at the 5% level. (1 mark)

c. From Figure 2, what is the overall regression F-statistic for this regression, and what is its corresponding degrees of freedom. Interpret the statistical significance of this test at the 5% level and its implication for the model. (2 marks)

[1] The markets consist of towns and cities across the United States.

Figure 1: Pollution Data Summary Statistics

```
      air             plants            pulp              repub              pop        
 Min.   :1.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   : 1.001  
 1st Qu.:2.013   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 7.420  
 Median :3.347   Median :3.000   Median :0.0000   Median :1.0000   Median :10.014  
 Mean   :3.913   Mean   :2.769   Mean   :0.2987   Mean   :0.5011   Mean   :10.116  
 3rd Qu.:5.398   3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:12.691  
 Max.   :9.989   Max.   :9.000   Max.   :1.0000   Max.   :1.0000   Max.   :24.842  
```

Figure 2: Pollution Regression 1 Output

```
> reg1=lm(ln_air~plants+pulp+repub+pop,data=dat_pol)
> summary(reg1)

Call:
lm(formula = ln_air ~ plants + pulp + repub + pop, data = dat_pol)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.32014 -0.48046  0.02247  0.48947  1.24311 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.360980   0.024278  56.058  < 2e-16 ***
plants      -0.024163   0.003964  -6.096 1.15e-09 ***
pulp        -0.071219   0.015385  -4.629 3.73e-06 ***
repub       -0.032382   0.014088  -2.299 0.021559 *  
pop         -0.006762   0.001829  -3.697 0.000219 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6037 on 7347 degrees of freedom
Multiple R-squared:  0.01023,	Adjusted R-squared:  0.00969 
F-statistic: 18.98 on 4 and 7347 DF,  p-value: 1.528e-15

> coeftest(reg1, vcov = vcovHC(reg1, "HC1"))

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  1.3609799  0.0241690 56.3111 < 2.2e-16 ***
plants      -0.0241626  0.0039352 -6.1402 8.672e-10 ***
pulp        -0.0712189  0.0153449 -4.6412 3.524e-06 ***
repub       -0.0323820  0.0140879 -2.2986 0.0215577 *  
pop         -0.0067615  0.0018440 -3.6668 0.0002473 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Question 3 continued)

Building on the first regression, the following modified regression is estimated:

$$\begin{aligned}\ln(air_i) =& \beta_0 + \beta_1 plants_i + \beta_2 pulp_i + \beta_3 repub_i + \beta_4 pop_i + \\ & \beta_5 (plants_i \times pulp_i) + \beta_6 (plants_i \times repub_i) + \beta_7 (pulp_i \times repub_i) + u_i\end{aligned}$$

Figure 3 on the next page contains the regression output from R-Studio for this regression. In the regression output, plants_pulp is $(plants_i \times pulp_i)$, plants_repub is $(plants_i \times repub_i)$, and pulp_repub is $(pulp_i \times repub_i)$.

d. Interpet the coefficient estimate on $(plants_i \times pulp_i)$ in Figure 3, and comment on whether it is statistically significantly different from 0 at the 5% level. (1 mark)

e. Interpet the coefficient estimate on $(pulp_i \times repub_i)$ in Figure 3, and comment on whether it is statistically significantly different from 0 at the 5% level. (1 mark)

f. What test is being conducted on Figure 4 on the next page? Carefully describe the outcome of the test using a 5% significance level, noting the relevant test statistic and degrees of freedom (if necessary). (1 mark)

g. What is the partial effect on air_i from $repub_i$ changing from 0 to 1 for a market with the median number of plants and population, and that has a pulp and paper mill? (3 marks)

Figure 3: Pollution Regression 2 Output

```
> reg2=lm(ln_air~plants+pulp+repub+pop+plants_pulp+plants_repub+pulp_repub,data=dat_pol)
> coeftest(reg2, vcov = vcovHC(reg2, "HC1"))

t test of coefficients:

               Estimate Std. Error t value  Pr(>|t|)
(Intercept)   1.3107835  0.0275856 47.5169 < 2.2e-16 ***
plants       -0.0076903  0.0060330 -1.2747 0.2024587
pulp         -0.0190637  0.0330828 -0.5762 0.5644697
repub         0.0492939  0.0277711  1.7750 0.0759382 .
pop          -0.0068129  0.0018421 -3.6985 0.0002184 ***
plants_pulp  -0.0133041  0.0087156 -1.5265 0.1269343
plants_repub -0.0261204  0.0078680 -3.3198 0.0009051 ***
pulp_repub   -0.0324337  0.0306658 -1.0577 0.2902490
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: Pollution Regression Test

```
> linearHypothesis(reg2,c("plants_repub=plants_pulp"),vcov = vcovHC(reg2, "HC1"))
Linear hypothesis test

Hypothesis:
- plants_pulp  + plants_repub = 0

Model 1: restricted model
Model 2: ln_air ~ plants + pulp + repub + pop + plants_pulp + plants_repub +
    pulp_repub

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
1   7345
2   7344  1 1.2201 0.2694
```

Question 4: Profiting from Hamburgers (10 Marks)

A restaurant chain that only sells one type of hamburger approaches you with a dataset called dat_profit.csv that consists of the following store-level variables from a cross-section of $n = 738$ stores:

$profit_i$: daily total profit earned in store i (in \$10,000's of dollars)

$price_i$: price charged in store i in dollars

$hours_i$: number of hours store i is open each day

$income_i$: average household income around store i (in terms of \$1000's of dollars)

pop_i: total potential customers served by store i (in terms of 1000's of people)

Using the dataset, you estimate the following regression model for the firm's profits:

$$profit_i = \beta_0 + \beta_1 price_i + \beta_2 price_i^2 + \beta_3 price_i \times income_i + \beta_4 income_i + \beta_5 hours_i + \beta_6 pop_i + u_i$$

Summary statistics for the dataset, and regression results are presented in Figures 5 and 6 on the next page, respectively. In the regression output, price_sq is $price_i^2$ and price_income is $price_i \times income_i$. For all parts of question 4, only conduct hypothesis tests based on regressions with heteroskedasticity-robust standard errors.

a. Interpret the partial effect of $hours_i$ on $profit_i$ and comment on whether it is statistically significantly different from 0 at the 5% level. (1 mark)

b. Using the regression results, test using the 5% level whether there is a nonlinear relationship between $profit_i$ and $price_i$. Carefully state the test and describe its outcome, noting the relevant test statistic and degrees of freedom (if necessary) (1 mark)

c. What is the partial effect from changing $price_i$ from 8 to 12 in a market with mean $income_i$, mean $hours_i$, and mean pop_i. (2 marks)

d. Carefully explain the steps required to compute the standard error for the partial effect you computed in part c. (3 marks)

e. The company is looking to maximise profits at each of its stores. Based on the estimation results,

   - What would you recommend $price_i$ should be for a store in a market with $income_i = 35$?
   - What would you recommend $price_i$ should be for a store in a market with $income_i = 45$?

   Briefly provide intuition based on the estimated regression model for why your recommended prices differ in the two markets. (3 marks)

Figure 5: Profits Data Summary Statistics

```
> summary(dat_profit)
     profit          price            income          hours             pop        
 Min.   :21.84   Min.   : 6.000   Min.   :24.84   Min.   : 8.00   Min.   : 2.133  
 1st Qu.:43.24   1st Qu.: 7.000   1st Qu.:36.70   1st Qu.:11.00   1st Qu.: 7.870  
 Median :48.14   Median : 8.000   Median :39.74   Median :12.00   Median :10.103  
 Mean   :49.01   Mean   : 8.686   Mean   :39.88   Mean   :11.55   Mean   :10.066  
 3rd Qu.:54.48   3rd Qu.:10.000   3rd Qu.:43.07   3rd Qu.:12.00   3rd Qu.:11.935  
 Max.   :83.97   Max.   :19.000   Max.   :54.90   Max.   :15.00   Max.   :20.130  
```

Figure 6: Profits Regression Output

```
> reg1=lm(profit~price+price_sq+price_income+income+hours+pop,data=dat_profit)
> summary(reg1)

Call:
lm(formula = profit ~ price + price_sq + price_income + income + 
    hours + pop, data = dat_profit)

Residuals:
   Min     1Q Median     3Q    Max 
-7.907 -1.710 -0.054  1.748  7.301 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.816083   3.411739   0.239   0.8110    
price         0.782854   0.421198   1.859   0.0635 .  
price_sq     -0.070015   0.013680  -5.118 3.95e-07 ***
price_income  0.080440   0.008258   9.740  < 2e-16 ***
income        0.118037   0.075130   1.571   0.1166    
hours         0.123801   0.093665   1.322   0.1867    
pop           1.293598   0.031896  40.556  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.547 on 731 degrees of freedom
Multiple R-squared:  0.912,	Adjusted R-squared:  0.9113 
F-statistic:  1263 on 6 and 731 DF,  p-value: < 2.2e-16

> coeftest(reg1, vcov = vcovHC(reg1, "HC1"))

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept)   0.816083   3.687529  0.2213    0.8249    
price         0.782854   0.479476  1.6327    0.1030    
price_sq     -0.070015   0.014840 -4.7179 2.855e-06 ***
price_income  0.080440   0.009150  8.7912 < 2.2e-16 ***
income        0.118037   0.081074  1.4559    0.1458    
hours         0.123801   0.093628  1.3223    0.1865    
pop           1.293598   0.031574 40.9702 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Question 5: Unemployment and Interest Rates (10 Marks)

Suppose as analyst at the Reserve Bank of Australia you developed a time series dataset called dat_macro.csv with the following variables:

$unemp_t$: unemployment rate in Australia in monthly t

$rate_t$: interest rate in Australia in monthly t

Your data span months $t = 1, 2, \ldots, 137$.

a. You begin your analysis with a plot of $unemp_t$ and $rate_t$ over time, and obtain the graph from R-Studio presented in Figure 7 on the next page. Explain whether or not the two time series appear to be stationary. (1 mark)

b. Suppose you estimate the following ADL(1,2) model:

$$unemp_t = \beta_0 + \beta_1 unemp_{t-1} + \beta_2 rate_{t-1} + \beta_3 rate_{t-2} + u_t$$

and obtain the estimation results presented Figure 8 on page 12 below. How many observations are used in estimating this model? Briefly explain why this is the number of observations used in estimation. (2 marks)

c. What is the out-of-sample forecast and 95% confidence interval for $unemp_t$ in month $t = 138$? In answering this question, you might need to use some of the last 10 observations in the sample, which are presented in Figure 9 on page 12 below. Assume u_t is i.i.d with a N(0,1) distribution in constructing the interval. (2 marks)

d. Suppose you wanted to estimate a different ADL(2,2) model where the growth rate in $unemp_t$ is the dependent variable, and where the regressors consist of lagged growth rates in $unemp_t$ and lagged growth rates in $rate_t$.

Provide the **pseudo-code**[2] for an R program (e.g., the .R code) that you would write in R-Studio for: (1) estimating this ADL(2,2) model; (2) computing within-sample forecast errors for the growth rate in $unemp_t$; (3) reporting summary statistics for these within-sample forecast errors, and (4) constructing a time series plot of within-sample forecasts for $unemp_t$ and the realised values of $unemp_t$.

Your pseudo-code can be written in a series of bullet points. It should explicitly state all steps required in R-script to generate these results given the 2 variables in the dataset dat_macro.csv listed above, $unemp_t$ and $rate_t$. You do not need to cite explicit R commands, syntax, or equations, but you may do so if it helps clarify what each part of your pseudo-code does. (5 marks)

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

Figure 7: Macro Data Time Series Plot

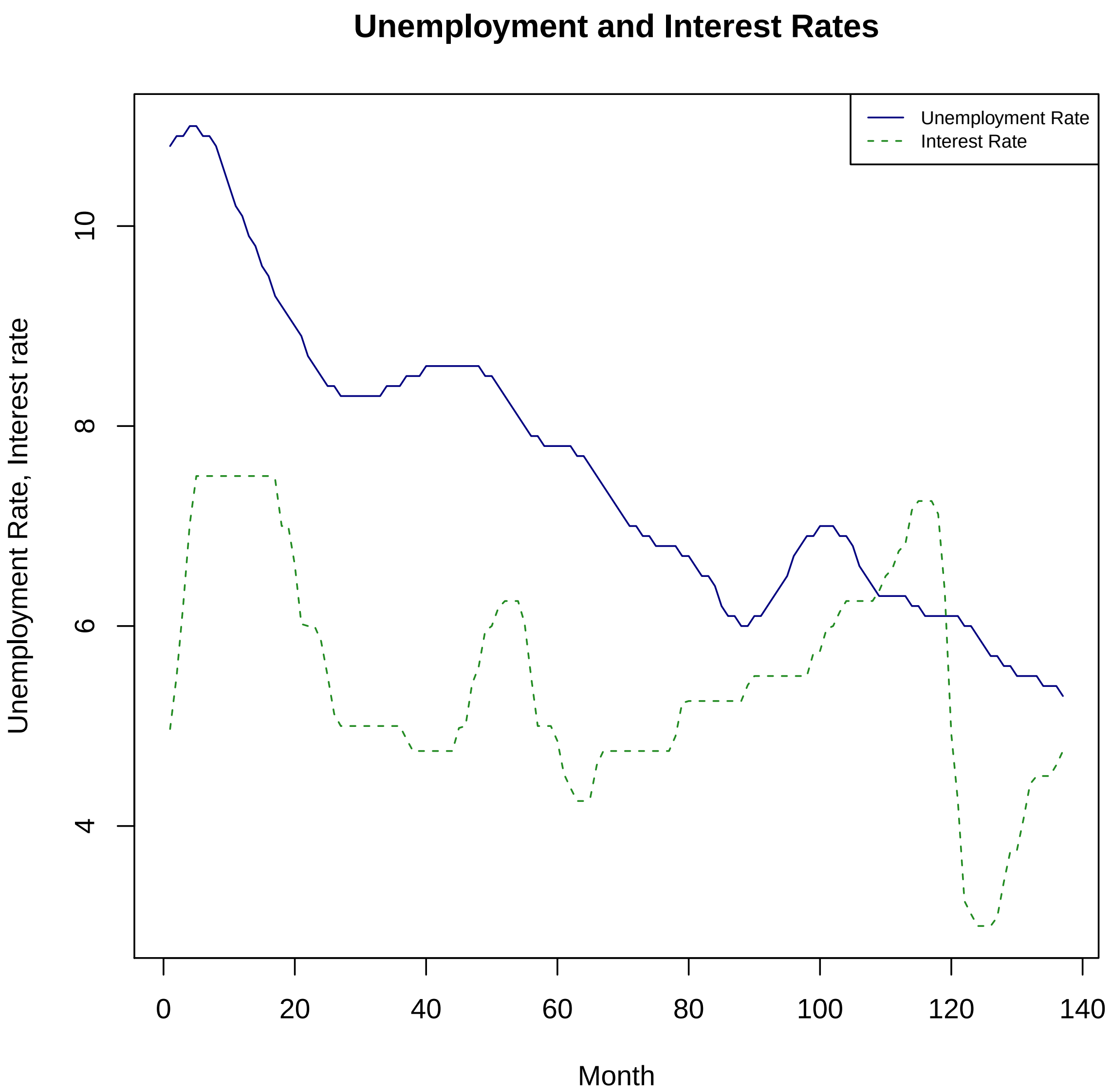

Figure 8: Macro Regression Output

```
> reg1=lm(unemp~unemp_lag1+rate_lag1+rate_lag2,data=dat_macro)
> summary(reg1)

Call:
lm(formula = unemp ~ unemp_lag1 + rate_lag1 + rate_lag2, data = dat_macro)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.17154 -0.05827  0.01012  0.04141  0.23258 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.094801   0.038268   2.477   0.0145 *  
unemp_lag1   0.991467   0.005109 194.060   <2e-16 ***
rate_lag1    0.030862   0.025310   1.219   0.2249    
rate_lag2   -0.043939   0.024866  -1.767   0.0796 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07445 on 131 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.9974,	Adjusted R-squared:  0.9974 
F-statistic: 1.685e+04 on 3 and 131 DF,  p-value: < 2.2e-16

> coeftest(reg1, vcov = vcovHC(reg1, "HC1"))

t test of coefficients:

              Estimate Std. Error  t value Pr(>|t|)    
(Intercept)  0.0948014  0.0402153   2.3573  0.01989 *  
unemp_lag1   0.9914672  0.0051036 194.2695  < 2e-16 ***
rate_lag1    0.0308618  0.0241866   1.2760  0.20422    
rate_lag2   -0.0439387  0.0244942  -1.7938  0.07515 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9: Macro Data Last 10 Observations

t	unemp	rate
127	5.7	3.09
128	5.6	3.44
129	5.6	3.75
130	5.5	3.76
131	5.5	4.07
132	5.5	4.42
133	5.5	4.50
134	5.4	4.50
135	5.4	4.50
136	5.4	4.61
137	5.3	4.75

END OF EXAMINATION

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	∞
Denominator ν_2	**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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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

Formula Sheet

Expected Values, Variances, Correlation

$E(c) = c$

$E(cx) = cE(x)$

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

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

$E(c_1 x + c_2 y) = c_1 E(x) + c_2 E(y)$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Logarithms

$x = \ln(e^x)$

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

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

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

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

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

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

Calculus

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

OLS Estimator

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

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

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

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

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

Hypothesis Testing

Different populations

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

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

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

Linear Regression

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

$H_0: \beta_1 = \beta_{1,0}$ 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-testing

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

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

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

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

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

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

Goodness of Fit

$$SSR = \sum_{i=1}^{n} u_i^2$$

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

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

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

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

$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2}$

Nonlinear and Time Series Regression

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

$\Delta\hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k)$

$SE(\Delta\hat{Y}) = \frac{|\Delta\hat{Y}|}{\sqrt{F}}$

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

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

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

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

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

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