ECOM20001 2018 SEMESTER 1 FINAL EXAM SOLUTIONS

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?

   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?

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

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

   a. is the square of the t-statistic

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?

   c. [7.40, 20.60]

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

   c. It can be negative

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

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

   Which of the following is the correct interpretation of β_2?

   b. A 1-unit increase in X_2 yields a $100 \times \beta_2$ % increase in Y, holding X_1 fixed

7. Consider the following multiple linear regression:

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

   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?

   c. -0.539, fail to reject 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:

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:

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:

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

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)

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 $\hat{\beta}$ as n gets big. In other words, the $\hat{\beta}$, tends to recover an estimate that is close true value of β for sufficiently large samples. It gets the "right" answer.

Central Limit Theorem (CLT): As n gets large, the marginal distribution of $\hat{\beta}$ is approximately $N(\hat{\beta}, \sigma^2_{\beta})$ The CLT is important because we can exploit the fact that $\hat{\beta}$ is approximately normally distributed around the true value β as n grows to test hypotheses about the true value β and to construct confidence intervals for the true value β.

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)

We would expect that $Earnings_i$ and $Crime_i$ to have a negative relationship (poorer postcodes have higher crime rates), and $Crime_i$ and $Educ_i$ to have a negative relationship (less educated postcodes have higher crime rates).

The regression equation for equation (1) can be written as:

$$Earnings_i = \beta_0 + \beta_1 Educ_i + u$$

Therefore, the omitted variable $Crime_i$ in regression from equation (1) in the question would enter the error term u with a negative relationship with the dependent variable, and it would have a negative relationship with the regressor $Educ_i$. Together, this implies a positive relationship between $Educ_i$ and u in the regression model: $\rho_{Xu} > 0$, yielding an upward bias in the regression coefficient on $Educ_i$ in equation (1) (11.83) as compared its coefficient in equation (2) (8.49).

Alternatively, an answer can more simply state that when $Crime_i$ increases, both $Earnings_i$ and $Educ_i$ fall, which the implies β_1 in the first regression will be driven by positive correlation between $Earnings_i$ and $Educ_i$ driven by variation in the ommitted variable $Crime_i$, which creates upward bias in the regression coefficient in equation (1) due to $Crime_i$ being an omitted variable in the error term u.

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)

Deriving the partial effect $\Delta \ln(Y)$ from a $\Delta \ln(X_1)$ in $\ln(X_1)$:

$$E[\ln(Y)|X] = \ln(Y) = \beta_0 + \beta_1 \ln(X)$$
$$E[\ln(Y)|X + \Delta X] = \ln(Y + \Delta Y) = \beta_0 + \beta_1 \ln(X + \Delta X)$$

which implies:

$$\begin{aligned}\Delta \ln(Y) &= E[\ln(Y)|X + \Delta X] - E[\ln(Y)|X] \\ &= \ln(Y + \Delta Y) - \ln(Y) \\ &= \beta_1 (\ln(X + \Delta X) - \ln(X))\end{aligned}$$

If ΔX is small, then $\ln(X + \Delta X) - \ln(X) \approx \frac{\Delta X}{X}$ and $\ln(Y + \Delta Y) - \ln(Y) \approx \frac{\Delta Y}{Y}$ and the above equation is approximated by:

$$\frac{\Delta Y}{Y} \approx \beta_1 \frac{\Delta X}{X}$$

or

$$\beta_1 = \frac{\Delta Y / Y}{\Delta X / X} = \frac{100 \times \Delta Y / Y}{100 \times \Delta X / X} = \frac{\text{percent change in } Y}{\text{percent change in } X}$$

which is the definition of the elasticity of Y with respect to X

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

   Having one additional plant in a town is associated with a 2.42 percentage reduction in air quality, holding other regressors fixed. It is statistically significantly different from 0 at the 5% level as the p-value is less than 0.05.

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

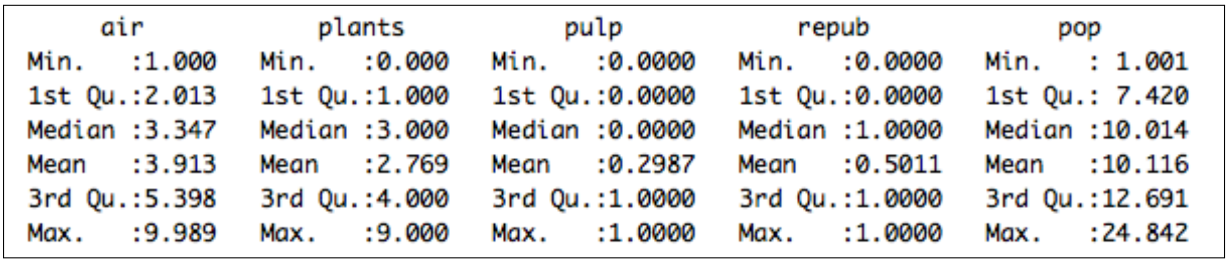
   Having a pulp and paper plant in a market is associated with a 7.12 percentage reduction in air quality, holding other regressors fixed. It is statistically significantly different from 0 at the 5% level as the p-value is less than 0.05.

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)

   The overall regression F-statistic is 18.98. It has an F-distribution with df1=4 and df2=7347 degrees of freedom. It has a p-value of less than 0.05, which implies that we reject the null at the 5% level that all the regression coefficients jointly equal 0. In other words, we reject the null that the regression model is statistically useless for explaining air quality.

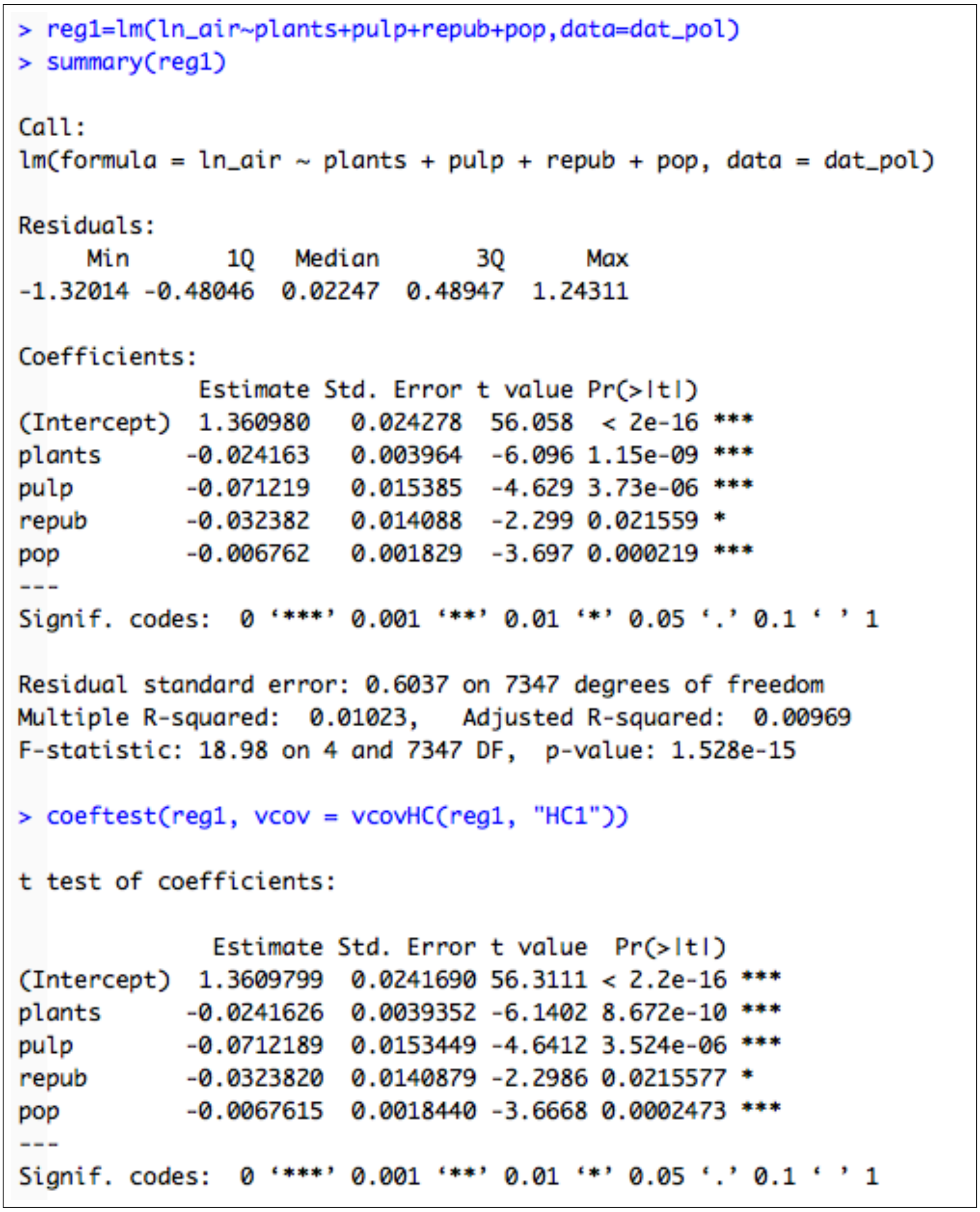
[1] The markets consist of towns and cities across the United States.

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 2 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. Interpret the coefficient estimate on $(plants_i \times pulp_i)$ in Figure 3, and comment on whether it is statistically significant at the 5% level. (1 mark)

Relative to markets without a pulp and paper plant, markets with a pulp and paper plant realize an additional 1.33% reduction in air quality from having one additional plant, holding other regressors fixed. With a p-value of 0.127, this coefficient is not statistically significant at the 5% level.

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

Relative to markets without a pulp and paper plant and without being a state with a Republican party, markets with a pulp and paper plant and in a state with a republican party realize an additional 3.24% reduction in their air quality, holding other regressors fixed. With a p-value of 0.290, this coefficient is not statistically significant at the 5% level.

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

The test is: $H_0 : \beta_5 = \beta_6$ versus $H_0 : \beta_5 \neq \beta_6$. The corresponding test statistic for this joint test is $F = 1.220$ which has an F-distribution with df1=1 and df2=7344 degrees of freedom. The test has a p-value of 0.269, implying we do not reject the null at the 5% level.

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)

Calculating the partial effect in steps:
- Values for the other covariates are: $pulp = 1$, $plants = 3$ and $pop = 10.014$.
- Predicted value from the regression if $repub = 0$:

$$\begin{aligned}E[\ln(air_i)|repub = 0] =& \beta_0 + \beta_1 3 + \beta_2 1 + \beta_3 0 + \beta_4 10.014 + \beta_5 (3 \times 1) + \beta_6 (3 \times 0) + \beta_7 (1 \times 0) \\ =& \beta_0 + 3\beta_1 + \beta_2 + 10.014\beta_4 + 3\beta_5\end{aligned}$$

- Predicted value from the regression if $repub = 1$:

$$\begin{aligned}E[\ln(air_i)|repub = 1] =& \beta_0 + \beta_1 3 + \beta_2 1 + \beta_3 1 + \beta_4 10.014 + \beta_5 (3 \times 1) + \beta_6 (3 \times 1) + \beta_7 (1 \times 1) \\ =& \beta_0 + 3\beta_1 + \beta_2 + \beta_3 + 10.014\beta_4 + 3\beta_5 + 3\beta_6 + \beta_7\end{aligned}$$

- Partial effect is therefore:

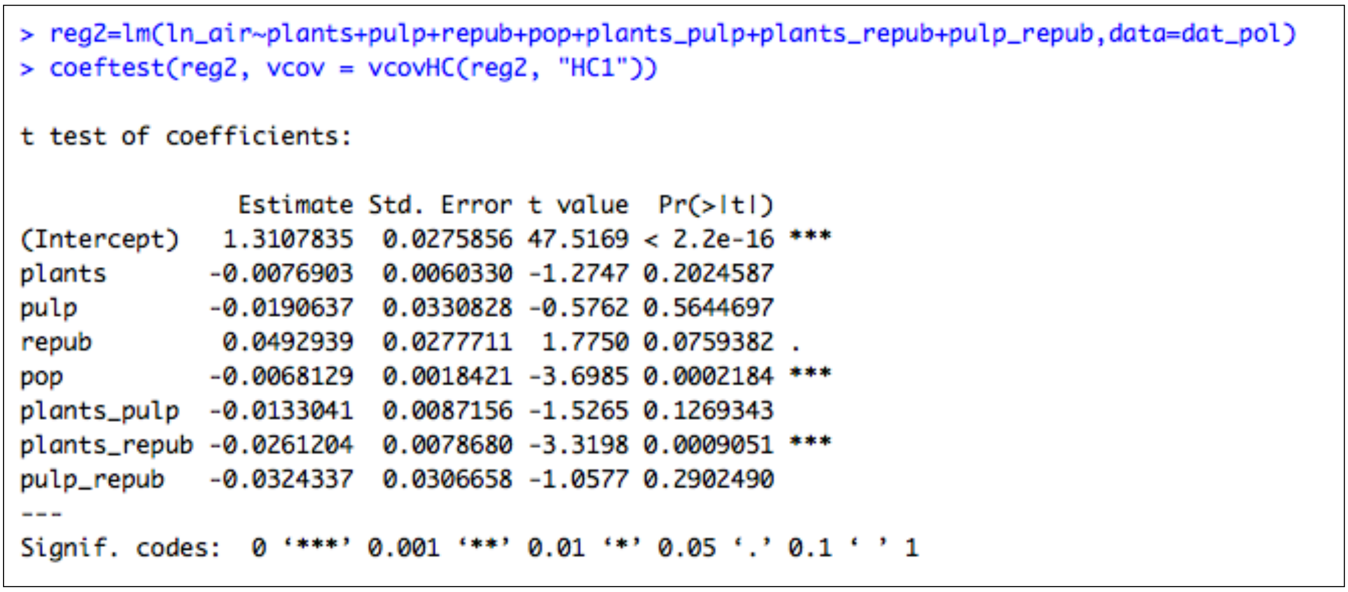
$$E[\ln(air_i)|repub = 1] - E[\ln(air_i)|repub = 0] = \beta_3 + 3\beta_6 + \beta_7$$

which implies an estimated partial effect of

$$\hat{\beta}_3 + 3 \times \hat{\beta}_6 + \hat{\beta}_7 = 0.0493 - 3 \times 0.0261 - 0.0324 = -0.0614$$

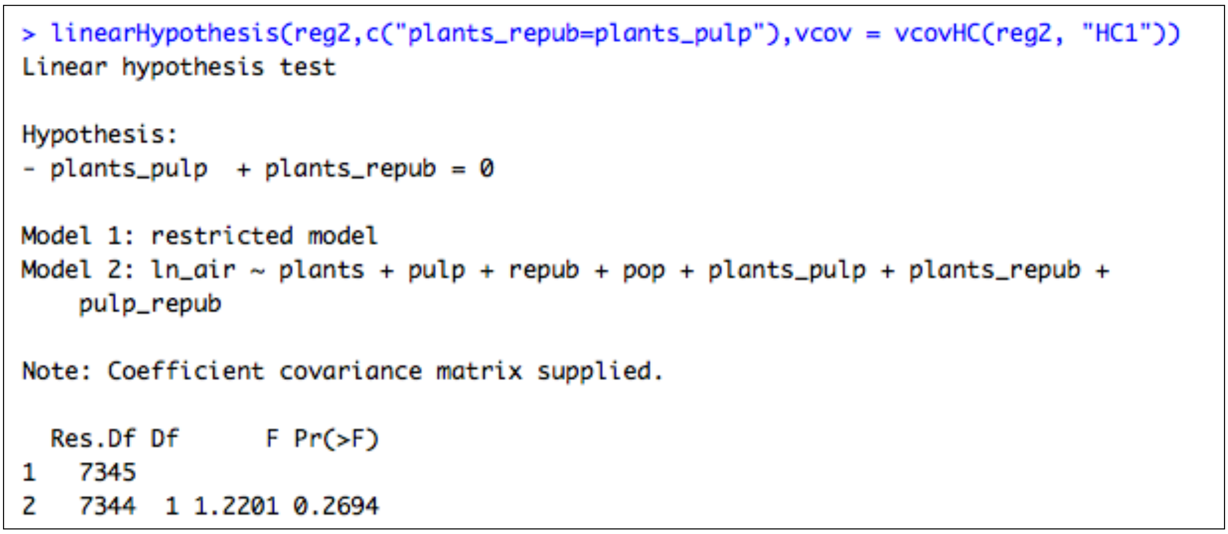
or a 6.14% reduction in air quality from changing $repub = 0$ to $repub = 1$ at median plants and population in a market with a pulp and paper plant, holding other factors fixed.

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:

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

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)

Holding other regressors fixed, having a store open one additional hour is associated with a $10000 \times 0.123801 = \$1,238$ increase in daily profits. The p-value for the test is 0.1867, which implies we fail to reject the null that the coefficient is different from 0 at the 5% level.

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)

The coefficient on $price_i^2$ can be used to determine if there is such a nonlinear relationship. Formally stating the test, $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$ yields a t-statistic of -4.7179 and p-value less than 0.05. This implies that we reject the null of a linear relationship between $profit_i$ and $price_i$ at the 5% level.

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)

Holding other factors fixed, evaluating the partial effect on $profit_i$ at mean $income_i$ of 39.88 from changing $price_i$ from 8 to 12 we obtain:

$$\begin{aligned}\Delta profit &= \left(\beta_1 12 + \beta_2 12^2 + \beta_3 12 \times 39.88\right) - \left(\beta_1 8 + \beta_2 8^2 + \beta_3 8 \times 39.88\right)\\ &= 4\beta_1 + 80\beta_2 + 39.88 \times 4 \times \beta_3\end{aligned}$$

Inserting our regression coefficient estimates and computing the partial effect we obtain:

$$\begin{aligned}\Delta\widehat{profit} &= 4\hat{\beta}_1 + 80\hat{\beta}_2 + 39.88 \times 4 \times \hat{\beta}_3 \\ &= 4 \times 0.782854 - 80 \times 0.070015 + 39.88 \times 4 \times 0.080440 \\ &= 10.36\end{aligned}$$

In words, increasing prices from 8 to 12 in a market with mean income, holding other factors fixed, increases daily profits by $10.36 \times \$10,000$=\$103,600 per day.

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

Using the calculations from part c., we can compute the standard error for the partial effect in two steps: - Compute the F-statistic corresponding to the following test:

$$H_0 : 4\beta_1 + 80\beta_2 + 39.88 \times 4 \times \beta_3 = 0 \ \text{ vs } \ H_1 : 4\beta_1 + 80\beta_2 + 39.88 \times 4 \times \beta_3 \neq 0$$

Call this F-statistic F^{act}
- Then compute the standard error as follows:

$$SE(\Delta\widehat{profit}) = \frac{|\Delta\widehat{profit}|}{F^{act}} = \frac{10.36}{F^{act}}$$

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)

- The estimated regression function implies an quadratic profit function in profits that increases at a decreasing rate, meaning we can use calculus to maximize it
- Taking the derivate of the estimated profit function (or the predicted value) with respect $price_i$ at $income_i = 35$, setting the derivative equal to 0, and solving we obtain:

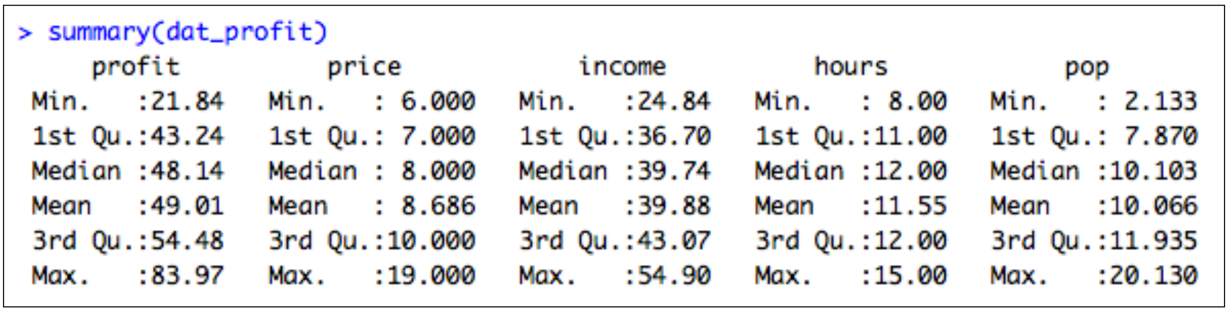
$$\begin{aligned}&0.782854 - 2 \times 0.070015 \times price + 35 \times 0.080440 = 0 \\ &\Rightarrow price = \frac{0.782854 + 35 \times 0.080440}{2 \times 0.070015} = \frac{4.914149}{0.14003} = 25.70\end{aligned}$$

- Performing a similar calculation for maximising profits with respect to price at $income_i = 45$ we obtain:

$$\begin{aligned}&0.782854 - 2 \times 0.070015 \times price + 45 \times 0.080440 = 0 \\ &\Rightarrow price = \frac{0.782854 + 45 \times 0.080440}{2 \times 0.070015} = \frac{6.094519}{0.14003} = 31.44\end{aligned}$$

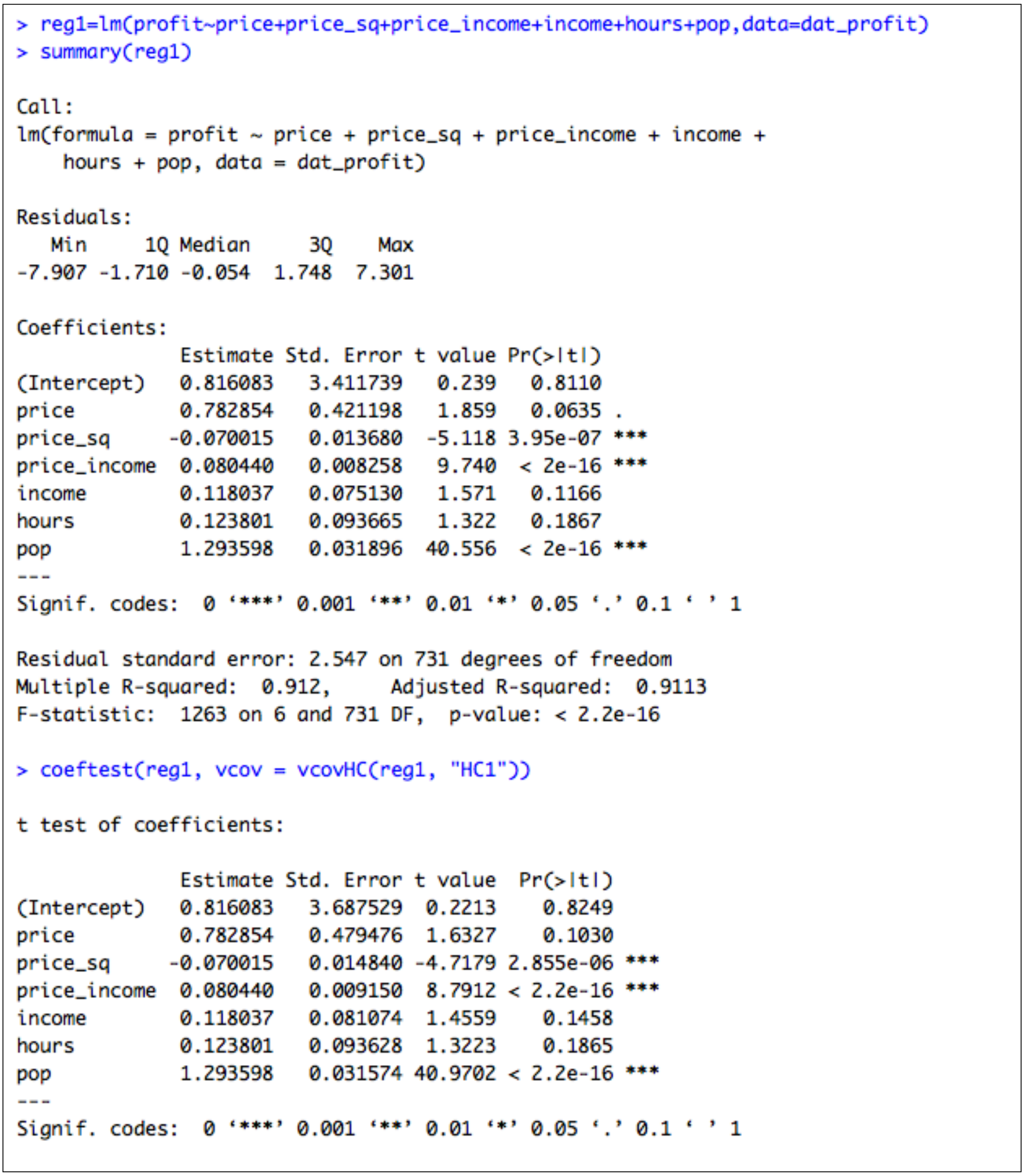
- In words, we maximize profits by setting $price = 25.70$ in markets with $income_i = 35$, and by setting $price = 31.44$ in markets with $income_i = 45$. Intuitively, the $price_i \times income_i$ income interaction in the regression implies a more positive growth rate in profits with higher income levels, implying markets with higher incomes should be charged higher prices if a store profit maximizes. Profits do not diminish as fast at lower price levels in markets with higher incomes.

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 the following 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)

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. Visually, $rate_t$ appears to be stationary, while $unemp_t$ is clearly trending downward suggesting it is non-stationarity.

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 mark)

The time series has $T = 137$ 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, 135 observations are used in estimating the ADL model.

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)

- Using the estimation results and the last 2 observations in the sample from the latter figure, we obtain an out-of-sample forecast for period $t = 138$ of:

$$\widehat{unemp}_t = 0.094801 + 0.991467 \times 5.3 + 0.030862 \times 4.75 - 0.043939 \times 4.61 = 5.29$$

The SER from the regression output is 0.07455. Therefore, the 95% CI for the out-of-sample forecast is:

$$[5.29 - 1.96 \times 0.07455, 5.29 + 1.96 \times 0.07455] = [5.14, 5.44]$$

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)

Step 1. Comptue the first three lags of $unemp_t$ $(unemp_{t-1}, unemp_{t-2}, unemp_{t-3})$ and the first three lags of $rate_t$ $(rate_{t-1}, rate_{t-2}, rate_{t-3})$

Step 2. Compute the natural logs of $unemp_t$ and $rate_t$. Call these ln_unemp_t and ln_rate_t. Similarly construct the natural logs of the three lagged values of $unemp_t$ and $rate_t$

Step 3. Compute the growth rate in $unemp_t$ and $rate_t$ as the first difference in their natural logs.
- Growth in $unemp_t$, the dependent variable, is computed as

$$g\_unemp_t = ln\_unemp_t - ln\_unemp_{t-1}$$

- Growth in $unemp_t$ lagged one and two periods are computed as:

$$g\_unemp_{t-1} = ln\_unemp_{t-1} - ln\_unemp_{t-2}$$

$$g\_unemp_{t-2} = ln\_unemp_{t-2} - ln\_unemp_{t-3}$$

- Similarly, growth in $rate_t$ lagged one and two periods are computed as

$$g\_rate_{t-1} = ln\_rate_{t-1} - ln\_rate_{t-2}$$

$$g\_rate_{t-2} = ln\_rate_{t-2} - ln\_rate_{t-3}$$

Step 4. Estimate the ADL(2,2) model using the lm() command by regressing g_unemp_t on $g\_unemp_{t-1}, g\_unemp_{t-2}, g\_rate_{t-1}, g\_rate_{t-2}$

Step 5. Using the estimated ADL(2,2) model, compute the predicted values for g_unemp_t for each value of $t = 1, \ldots, 137$, which yields the within-sample predictions for g_unemp_t, $g\_\widehat{unemp}_t$

Step 6. Compute the forecast error for each data point as $error_t = g\_unemp_t - g\_\widehat{unemp}_t$ and use the summary() commany in R to report the summary statistics

[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.

for $error_t$

Step 7. Use the plot() command to create a time series plots of realized values of g_unemp_t and their within-sample forecasts, $g\_\widehat{unemp}_t$

Figure 7: Macro Data Time Series Plot

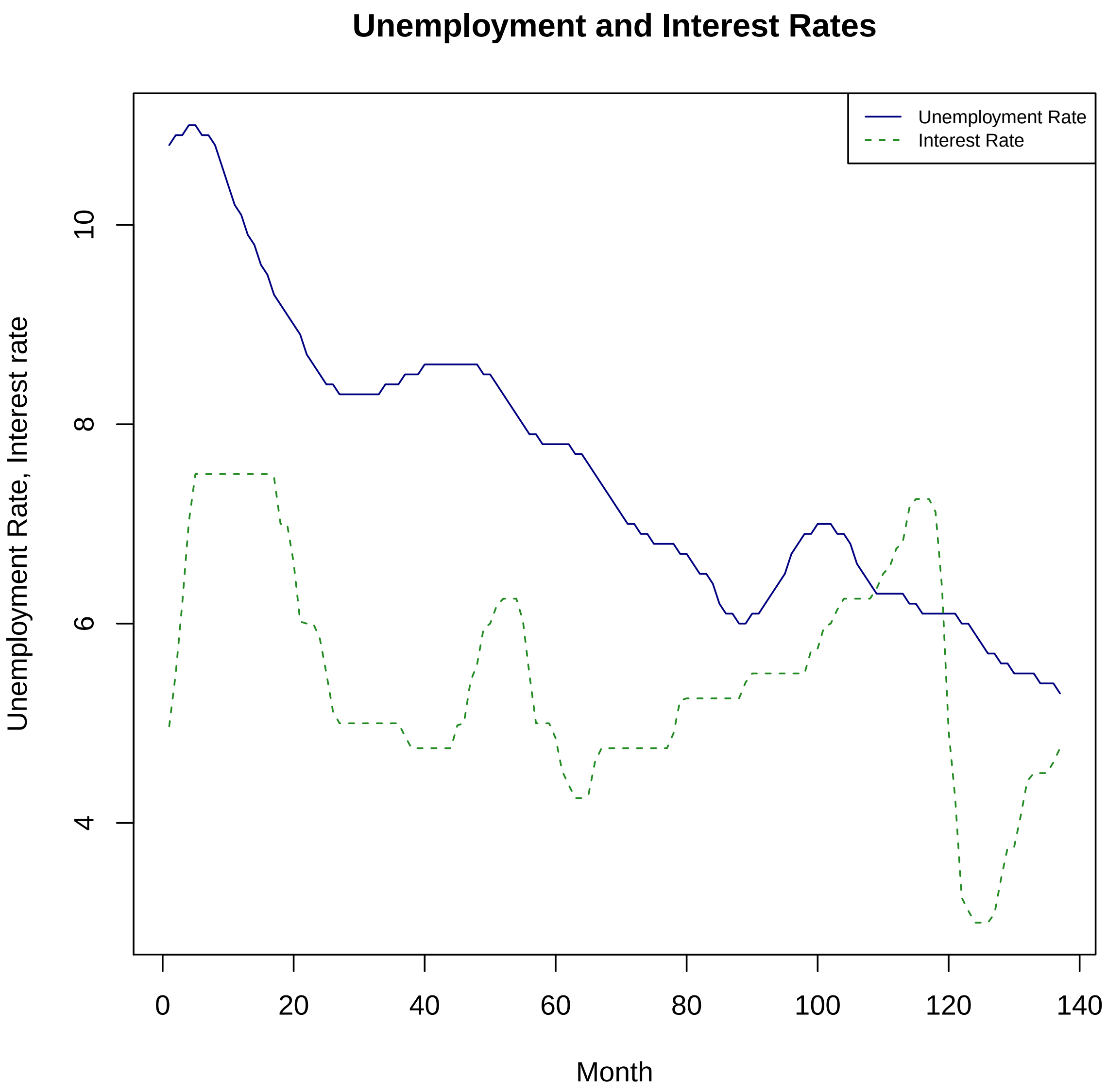
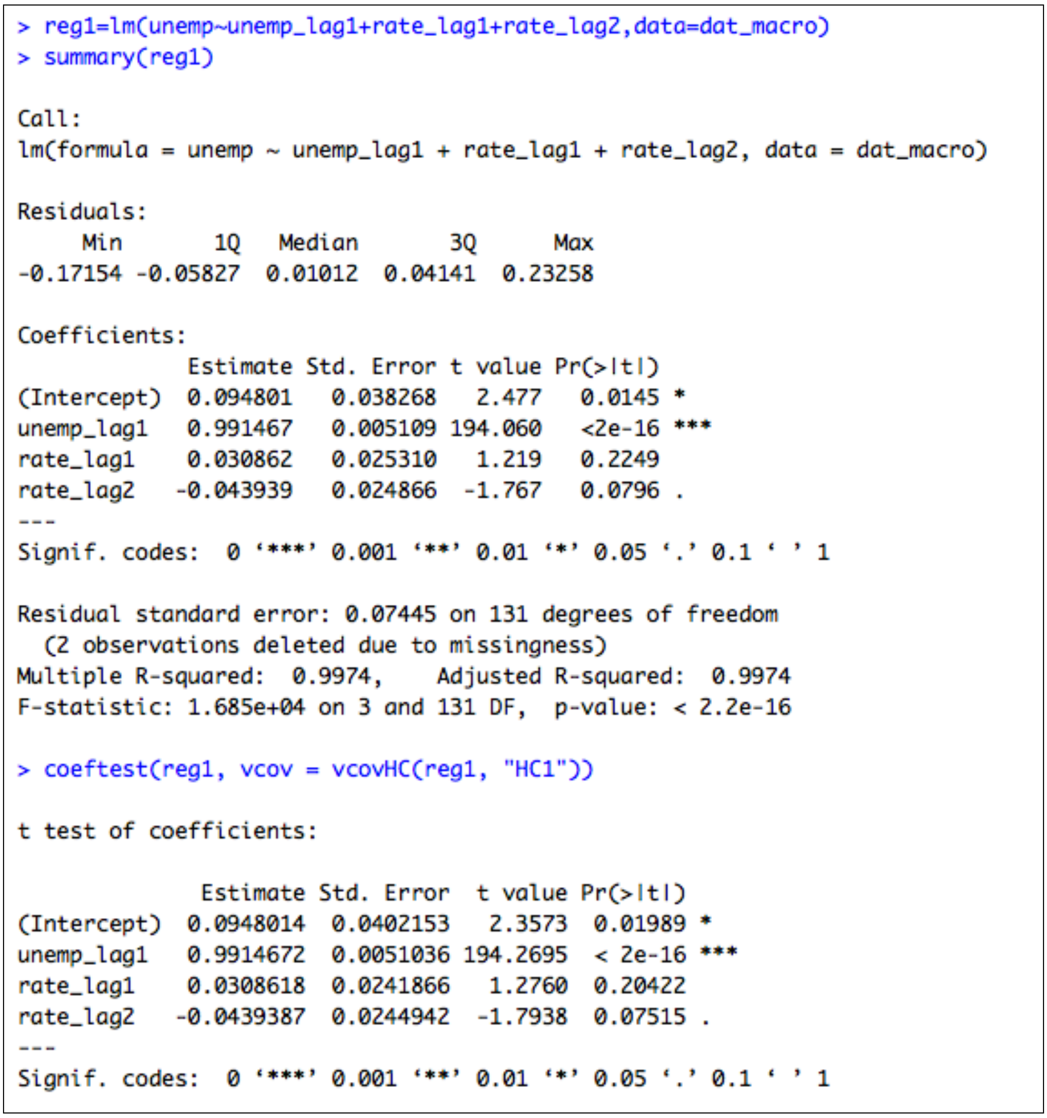
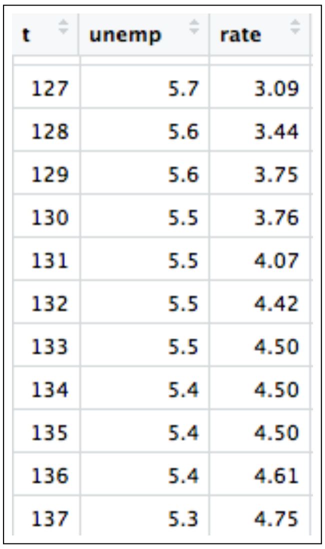


Figure 8: Macro Regression Output



```
> 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