ECOM20001 PRACTICE FINAL EXAM SOLUTIONS

Question 1: Multiple Choice (10 marks)

1. A discrete random variable X takes on one of ten discrete values $X = 0, 1, 2, \ldots, 10$ has the following density function: $P(X = 1) = 0.1$, $P(X = 3) = 0.2$, $P(X = 5) = 0.5$, $P(X = 10) = 0.2$. All other discrete values of X occur with probability 0. What is the value of the cumulative density function for X at $X = 4$?

   b. 0.3

2. You estimate a regression model with $n = 323$ observations and obtain the following estimation results:

$$\hat{Y}_i = \underset{(2.5)}{10.22} + \underset{(0.71)}{31.59}X_{1i} - \underset{(1.44)}{13.01}X_{2i} + \underset{(22.84)}{94.18}X_{3i}$$

   where the regression standard errors are in brackets. Which of the following hypothesized values for the regression coefficient on X_{3i}, β_3, do not belong in its 90% confidence interval?

   d. 45.10

3. Why is accounting for heteroskedasticity important?

   d. You can obtain incorrect standard errors if it is not accounted for

4. Suppose you estimate an ARDL(3,6) model with $T = 100$ observations, where all variables in the model are in terms of first differences. How many observations are used to estimate the model?

   c. 93

5. Suppose you estimated the following regression model using a cross-section of $n = 428$ observations:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_2 + \sum_{j=1}^{4} \gamma_j Z_{ji} + u_i$$

   Using your estimates, suppose you run the following hypothesis test:

$$H_0 : \gamma_1 = \gamma_2 \text{ and } \gamma_3 = \gamma_4 \text{ vs } H_0 : \gamma_1 \neq \gamma_2 \text{ or } \gamma_3 \neq \gamma_4$$

   What would be the distribution for corresponding F-statistic for this test?

   c. $F_{2,421}$

6. In which regression model does β_1 represent the expected change in Y for a 1-unit change in X?

   c. $Y = \beta_0 + \beta_1 X + u$

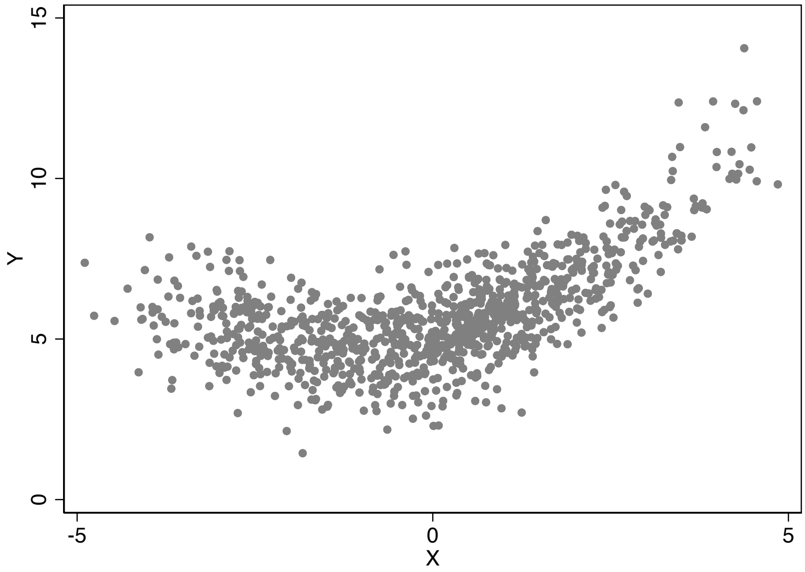
7. Suppose that the first difference of Y_t, ΔY_t, follows an AR(1) model: $\Delta Y_t = \beta_0 + \beta_1 \Delta Y_{t-1} + u_t$. The model for Y_t can alternatively be written as:

   d. $Y_t = \beta_0 + (1 + \beta_1)Y_{t-1} - \beta_1 Y_{t-2} + u_t$

8. When testing a joint hypothesis with a multiple linear regression model, you should:

   b. use the F-statistics and reject at least one of the hypotheses if the statistic exceeds the critical value.

9. Consider the following scatter plot:



   Which regression model would most likely yield the best trade-off for model fit and precision?

   c. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$

10. Which, if any, of the following models cannot be estimated by multiple linear regression

   d. None of these models can be estimated using multiple regression

Question 2: Short Answer Questions (10 Marks)

a. Consider the following joint probability table that describes the distribution of students' tastes for econometrics and microeconomics:

	Likes Econometrics	Does Not Like Econometrics	Total
Likes Microeconomics	0.21	0.12	0.33
Does Not Like Microeconomics	0.07	0.60	0.67
Total	0.28	0.72	1.00

Carefully explain whethere students' tastes for econometrics and microeconomics independently distributed. (2 points)

If two variables X and Y are independent, then $P(X, Y) = P(X) \times P(Y)$. Consider then, for example, $P(Likes\ Micro, Likes\ Metrics) = 0.21$. $P(Likes\ Micro) = 0.21 + 0.12 = 0.33$ and $P(Likes\ Metrics) = 0.21 + 0.07 = 0.28$. Therefore, $P(Likes\ Micro) \times P(Likes\ Metrics) = 0.33 \times 0.28 = 0.0924$. Since this does not equal the joint probability $P(Likes\ Micro, Likes\ Metrics) = 0.21$, we have an example that implies the students' tastes for econometrics and microeconomics are not independently distributed.

b. Carefully explain the trade-off inherent to using the AIC and BIC in selecting a time series regression model. Which of these information criterion is more likely to suggest an econometric model with more regression parameters? (3 points)

The formulas (from the formula sheet provided) are:

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

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

In both formulas, the first part with SSR included falls as a regression model has more regressors and parameters because the SSR can only fall with larger models. Bigger models have better fit.

On the other hand, the second part of the formula (with $\ln(T)$ for example), rises with the number of regression parameters K. That is, it rises with model size/complexity.

Since we use the AIC/BIC to select time series models by minimising them, the trade-off inherent to them is model fit versus model size/compexity. The AIC will tend to suggest larger models because the second part has 2 and not $\ln(T)$ in the numerator, which means the second part of the formula tends to rise slower for the AIC as K rises as 2 is less than $\ln(T)$ for $T > \exp(2) = 7.389$

c. Consider the following regression model:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

Suppose you were interested in testing the following null hypothesis:

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

Carefully describe two separate ways you could test this hypothesis using a F-statistic and t-statistic. Where necessary, state the degrees of freedom either statistic (or both). (5 points)

One option is to use an F-test and directly test the hypothesis. You would compute the F-statistic that corresponds to the test, F^{act}. This statistic would have a F distribution with df1=1 and df2=n-3-1 degrees of freedom where n is sample size. You would reject the null assuming significance level α if $F^{act} > F^{\alpha}$, where F^{α} is the critical value for the test defined where $\alpha = 1 - G(F^{\alpha})$, where $G()$ is the cumulative density of the $F_{1,n-3-1}$ distribution.

The second option is to transform the regression and use a t-statistic. Notice that we can transform the regression as follows:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u + \beta_2 X_1 - \beta_2 X_1 + \beta_3 X_1 - \beta_3 X_1 \\ &= \beta_0 + (\beta_1 + \beta_2 + \beta_3) X_1 + \beta_2 (X_2 - X_1) + \beta_3 (X_3 - X_1) + u \\ &= \beta_0 + \gamma X_1 + \beta_2 W_1 + \beta_3 W_2 + u \end{aligned}$$

where $\gamma = \beta_1 + \beta_2 + \beta_3$, and where $W_1 = X_2 - X_1$ and $W_2 = X_3 - X_1$ are new regressors made from the original regressors. By running the latter regression, we can test the null by computing the t-statistic corresponding to the equivalent test that $H_0 : \gamma = 0$ vs $H_1 : \gamma \neq= 0$, and we reject the null if t-statistic we obtain, t^{act}, is bigger in absolute value than the critical value, t^{α}, where the critical value is defined where $\alpha = 2\Phi(-|t^{\alpha}|)$, where $\Phi()$ is the cumulative density function of the normal distribution.

Question 3: Estimating Cereal Demand at Amazon (10 Marks)

In June 2017, Amazon purchased a supermarket chain in the U.S. called Whole Foods as it further enhanced its presence in the supermarket industry. Suppose that after Amazon made this purchase, that it started using randomized control trials to estimate demand for products. One of the first experiments it ran was to randomize prices for two types of cereals across its supermarkets: Corn Flakes and Coco Pops.

Using these data from $i = 1, \ldots, 2459$ of its supermarkets in a dataset called dat_demand.csv, Amazon attempts to estimate the following demand equation:

$$\ln(q_i^{CF}) = \beta_0 + \beta_1 \ln(p_i^{CF}) + \beta_2 \ln(p_i^{CP}) + \beta_3 Income_i + \beta_4 Age_i + u_i$$

where

q_i^{CF}: quantity of Corn Flakes sold in store i (in 1000s)

p_i^{CF}: price of Corn Flakes in store i

p_i^{CP}: price of Coco Pops in store i

$Income_i$: average income of shoppers at store i (in \$10000s)

Age_i: average age of shoppers at store 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 the question, only conduct hypothesis tests based on regressions with heteroskedasticity-robust standard errors. Please answer the following questions using information from the regression output:

a. What is the 99% confidence interval for β_3? (1 mark)

99% CI is: [-0.0742551-2.58*0.0043838,-0.0742551+2.58*0.0043838]=[-0.0855,-0.0629]

b. Interpret the coefficient estimates on p_i^{CF} and p_i^{CP} and comment on whether they are statistically significantly different from 0 using the 5% level. (2 marks)

The coefficient on p_i^{CF} implies the elasticity of q^{CF} with respect to p^{CF} is -3.08. Its p-value for the test of the null that it equals 0 is less than 0.05, so it is statistically significant at the 5% level.

The coefficient on p_i^{CP} implies the elasticity of q^{CF} with respect to p^{CP} is -2.38. Its p-value for the test of the null that it equals 0 is less than 0.05, so it is statistically significant at the 5% level.

Now suppose Amazon estimates a richer demand model:

$$\ln(q_i^{CF}) = \beta_0 + \beta_1 \ln(p_i^{CF}) + \beta_2 \ln(p_i^{CP}) + \beta_3 \left(\ln(p_i^{CF}) \times Age_i\right) + \beta_4 \left(\ln(p_i^{CP}) \times Age_i\right) \\ + \beta_3 Income_i + \beta_4 Age_i + u_i$$

c. The estimation results are reported in Table 3. Interpret the coefficients estimates $\hat{\beta}_3$ and $\hat{\beta}_4$ and comment on whether each is statistically significantly different from 0 using a 5% level of significance. (2 marks)

The coefficient on $\left(\ln(p_i^{CF}) \times Age_i\right)$ implies that the elasticity of q^{CF} with respect to p^{CF} declines by 0.0304 with each additional year of average age at store i. That is, demand becomes more elastic with respect to p^{CF} with age. The p-value for the test that the coefficient equals 0 is 0.138, meaning we fail to reject the null that the coefficient equals 0 at the 5% level.

The coefficient on $\left(\ln(p_i^{CP}) \times Age_i\right)$ implies that the elasticity of q^{CF} with respect to p^{CP} declines by 0.058 with each additional year of average age at store i. The p-value for the test that the coefficient equals 0 is 0.037, meaning we reject the null that the coefficient equals 0 at the 5% level.

d. Using only the raw data provided, provide the **pseudo-code**[1] for estimating the elasticity of q_i^{CF} with respect to p_i^{CF} and its standard error based on the regression model in part b. when $p_i^{CP} = 6$, $Age_i = 50$, and $Income_i = 40$

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 5 variables in the original dataset provided in the question with 5 variables: $q_i^{CF}, p_i^{CF}, p_i^{CP}, Income_i, Age_i$. 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)

1. Compute the natural logs of $q_i^{CF}, p_i^{CF}, p_i^{CP}$ and construct the variables $\log(q_i^{CF}), \log(p_i^{CF}), \log(p_i^{CP})$
2. Compute the interactions of $\left(\ln(p_i^{CF}) \times Age_i\right)$ and $\left(\ln(p_i^{CP}) \times Age_i\right)$
3. Run the regression described in part b. using the lm() and coeftest() commands (the latter provides heteroskedastic standard errors)
4. Compute the elasticity at $Age_i = 50$ using the estimated regression model:

$$\Delta \log(q^{CF}) = \hat{\beta}_1 + \hat{\beta}_3 \times 50$$

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

$$H_0 : \hat{\beta}_1 + \hat{\beta}_3 \times 50 = 0 \text{ vs } H_1 : \hat{\beta}_1 + \hat{\beta}_3 \times 50 \neq 0$$

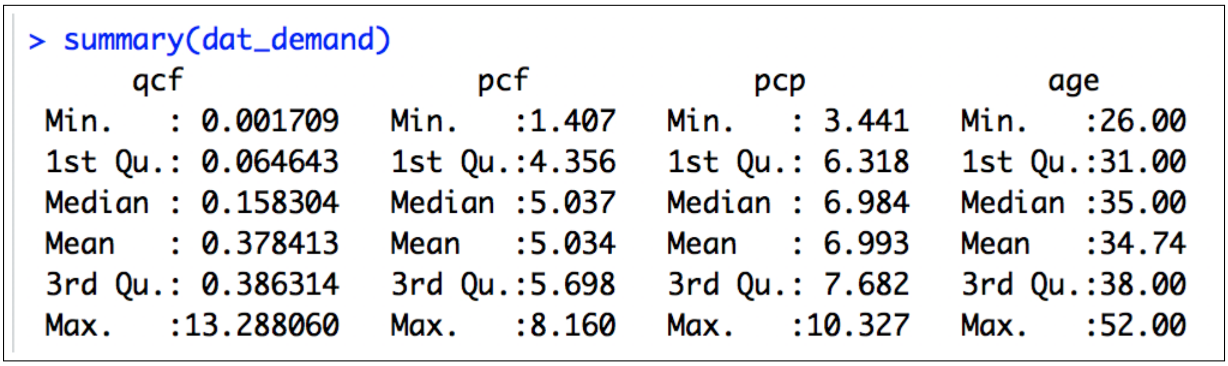
   Call this F-statistic F^{act}

6. Compute the standard error for the elasticity as:

$$SE(\Delta \log(q^{CF})) = \frac{|\hat{\beta}_1 + \hat{\beta}_3 \times 50|}{\sqrt{F^{act}}}$$

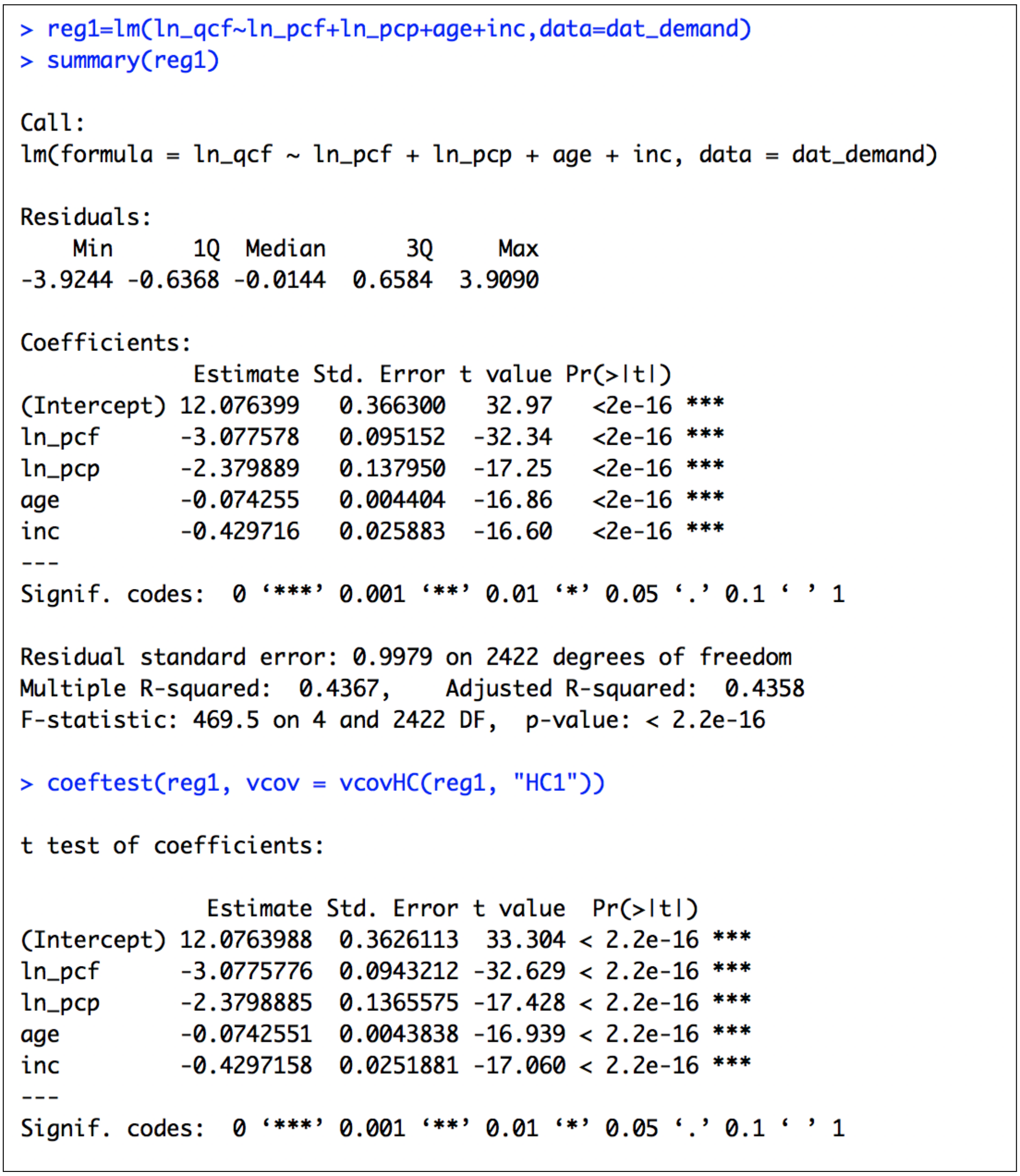
[1] A pseudo-code consists of all the steps you would take in an R program for conducting a particular analysis or calculation. It is primarily written in words and not R commands or syntax.

Figure 1: Cereal Demand Data Summary Statistics



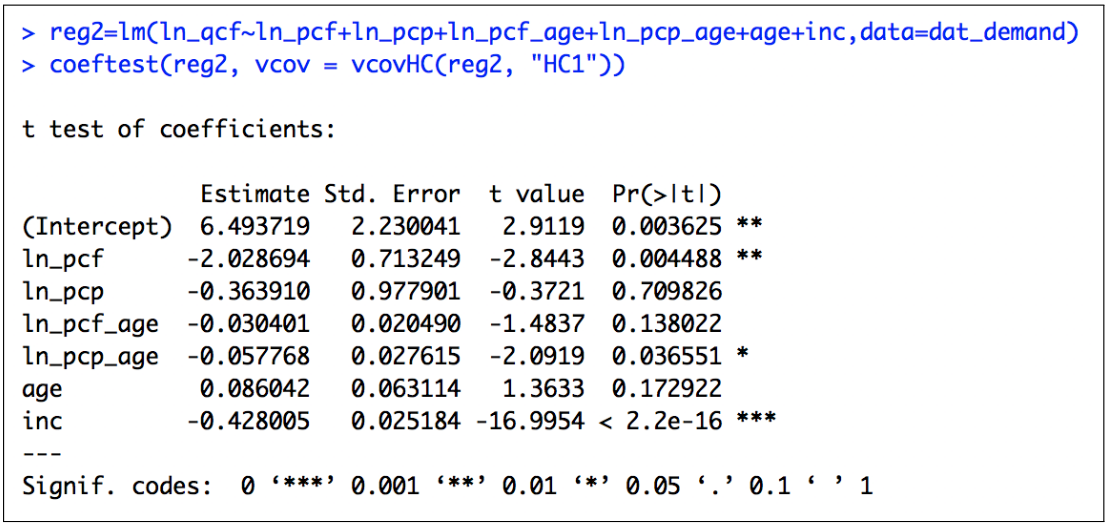
```
> summary(dat_demand)
      qcf                  pcf              pcp               age       
 Min.    : 0.001709   Min.   :1.407   Min.   : 3.441   Min.   :26.00  
 1st Qu.: 0.064643   1st Qu.:4.356   1st Qu.: 6.318   1st Qu.:31.00  
 Median : 0.158304   Median :5.037   Median : 6.984   Median :35.00  
 Mean   : 0.378413   Mean   :5.034   Mean   : 6.993   Mean   :34.74  
 3rd Qu.: 0.386314   3rd Qu.:5.698   3rd Qu.: 7.682   3rd Qu.:38.00  
 Max.   :13.288060   Max.   :8.160   Max.   :10.327   Max.   :52.00  
```

Figure 2: Cereal Demand Regression Output 1



```
> reg1=lm(ln_qcf~ln_pcf+ln_pcp+age+inc,data=dat_demand)
> summary(reg1)

Call:
lm(formula = ln_qcf ~ ln_pcf + ln_pcp + age + inc, data = dat_demand)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9244 -0.6368 -0.0144  0.6584  3.9090 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.076399   0.366300   32.97   <2e-16 ***
ln_pcf      -3.077578   0.095152  -32.34   <2e-16 ***
ln_pcp      -2.379889   0.137950  -17.25   <2e-16 ***
age         -0.074255   0.004404  -16.86   <2e-16 ***
inc         -0.429716   0.025883  -16.60   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9979 on 2422 degrees of freedom
Multiple R-squared:  0.4367,	Adjusted R-squared:  0.4358 
F-statistic: 469.5 on 4 and 2422 DF,  p-value: < 2.2e-16

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

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 12.0763988  0.3626113  33.304 < 2.2e-16 ***
ln_pcf      -3.0775776  0.0943212 -32.629 < 2.2e-16 ***
ln_pcp      -2.3798885  0.1365575 -17.428 < 2.2e-16 ***
age         -0.0742551  0.0043838 -16.939 < 2.2e-16 ***
inc         -0.4297158  0.0251881 -17.060 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Cereal Demand Regression Output 2



```
> reg2=lm(ln_qcf~ln_pcf+ln_pcp+ln_pcf_age+ln_pcp_age+age+inc,data=dat_demand)
> coeftest(reg2, vcov = vcovHC(reg2, "HC1"))

t test of coefficients:

             Estimate Std. Error  t value  Pr(>|t|)    
(Intercept)  6.493719   2.230041   2.9119  0.003625 ** 
ln_pcf      -2.028694   0.713249  -2.8443  0.004488 ** 
ln_pcp      -0.363910   0.977901  -0.3721  0.709826    
ln_pcf_age  -0.030401   0.020490  -1.4837  0.138022    
ln_pcp_age  -0.057768   0.027615  -2.0919  0.036551 *  
age          0.086042   0.063114   1.3633  0.172922    
inc         -0.428005   0.025184 -16.9954 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Question 4: Speeding and Speed Enforcement (10 Marks)

The Commonwealth Government commissioned an inquiry into policies aimed at reducing traffic speed. For this, the government randomly sampled traffic speed from $n = 750$ 1-kilometer road segments across Australia and constructed the following dataset

$speed_i$: average speed of a given car on road segment i

$limit_i$: speed limit on road segment i

$camera_i$: dummy equals 1 if there is a road camera on road segment i, 0 otherwise

$police_i$: dummy equals 1 if there is a sign stating police monitor highways in road segment i, 0 otherwise

$state_i$: state in which road segment i is in, 0 otherwise

For all parts of the question, only conduct hypothesis tests based on regressions with heteroskedasticity-robust standard errors.

a. Using these data, you first run the following single linear regression:

$$speed_i = \beta_1 limit_i + u_i$$

Suppose the regression coefficient for β_1 equalled 1 and you computed the average of the residuals. How would you interpret this average in simple, non-econometric terms? (1 mark)

It is important to notice that the regression does not have an intercept. Given this, the residual would correspond to whether average speed on segment i is above (positive residual) or below (negative residual) the speed limit. Therefore, the average of the residuals is the average number of km/hr a road segment is above or below the speed limit.

b. Now suppose you ran the following regression:

$$speed_i = \beta_1 limit_i + \beta_2 qld_i + \beta_3 nsw_i + \beta_4 vic_i + \beta_5 tas_i + \beta_6 sa_i + \beta_7 nt_i + \beta_8 wa_i + u_i$$

where $qld_i = 1$ is road segment i is in Queensland and 0 otherwise, $nsw_i = 1$ is road segment i is in New South Wales and 0 otherwise, and similarly for the other state dummy variables vic_i, tas_i, sa_i, nt_i. The regression results are reported in Figure 4. Notice that the regression coefficient on $limit_i$is almost equal to 1, and is not statistically significantly different from 1 in a two-tailed test at the 5% level.

Given this, interpret the magnitude of the coefficient on β_2, and comment on whether it is statistically significantly different from 0 at the 5% level of significance. Provide a simple, non-econometric interpretation of the coefficient, similar to the interpretation that you provided in part a. (1 mark)

The estimated regression coefficient on $\hat{\beta}_2$ of 2.01 is statistically significantly different from 0 at the 5% level as it has a p-value for this test of less than 0.05. Given the regression does not have an intercept, and that speed limit is a regressor, the coefficient implies that Queensland on average has road segments with average speed that is 2 km/hr above the speed limit.

c. What test is being performed in Figure 5 on the next page? Carefully describe the outcome of the using the 5% significance level, noting the relevant test statistic and degrees of freedom (if necessary). (2 marks)

The test is testing whether all of the coefficients on the state dummies jointly equal 0, that is, none of the states exhibit speed above or below the local speed limit across their road segments. The alternative is that at least one of the coefficients is not equal to 0. The corresponding F-statistic for the test is 72.668, which has an F-distribution with df1=7 and df2=742 degrees of freedom. The p-value is less than 0.05 implying the null is rejected at the 5% level. At least one of the coefficients on the state dummy variables is different from 0, implying systematic differences between local average traffic speed and the local speed limit.

d. Building further on your regression model, you now estimate a third regression model:

$$\begin{aligned} speed_i =& \beta_0 + \beta_1 limit_i + \beta_2 camera_i + \beta_3 police_i + \beta_4 camera_i \times police_i \\ &+ \beta_5 qld_i + \beta_6 nsw_i + \beta_7 vic_i + \beta_8 tas_i + \beta_9 sa_i + \beta_{10} nt_i + u_i \end{aligned}$$

The regression results are reported in Figure 6. What is the base category in this regression specification? (1 mark)

The base category is that category that corresponds to the situation when all the dummies equal 0. In the regression, the base category is a road segment in Western Australia without a speed camera and without a police sign.

e. Interpret the magnitude of the regression coefficient estimates on β_2, β_3. Also comment on whether either estimate is statistically significantly different from 0 at the 5% level. (2 marks)

$\hat{\beta}_2$ is statistically significantly different from 0 as the p-value for this test is less than 0.05. Interpreting the coefficient, holding other regressors fixed, having a speed camera reduces the average speed by 3.13 km/hr.

$\hat{\beta}_3$ is statistically significantly different from 0 as the p-value for this test is less than 0.05. Interpreting the coefficient, holding other regressors fixed, having a police sign reduces the average speed by 0.62 km/hr.

f. Compare the regression coefficient estimates on vic_i in Figure 6 to the sum of the intercept and the coefficient on vic_i in Figure 4. Is there omitted variable bias with the regression intercept for vic_i (Victoria) in Figure 4? If so, carefully explain a potential source of the bias. (2 marks)

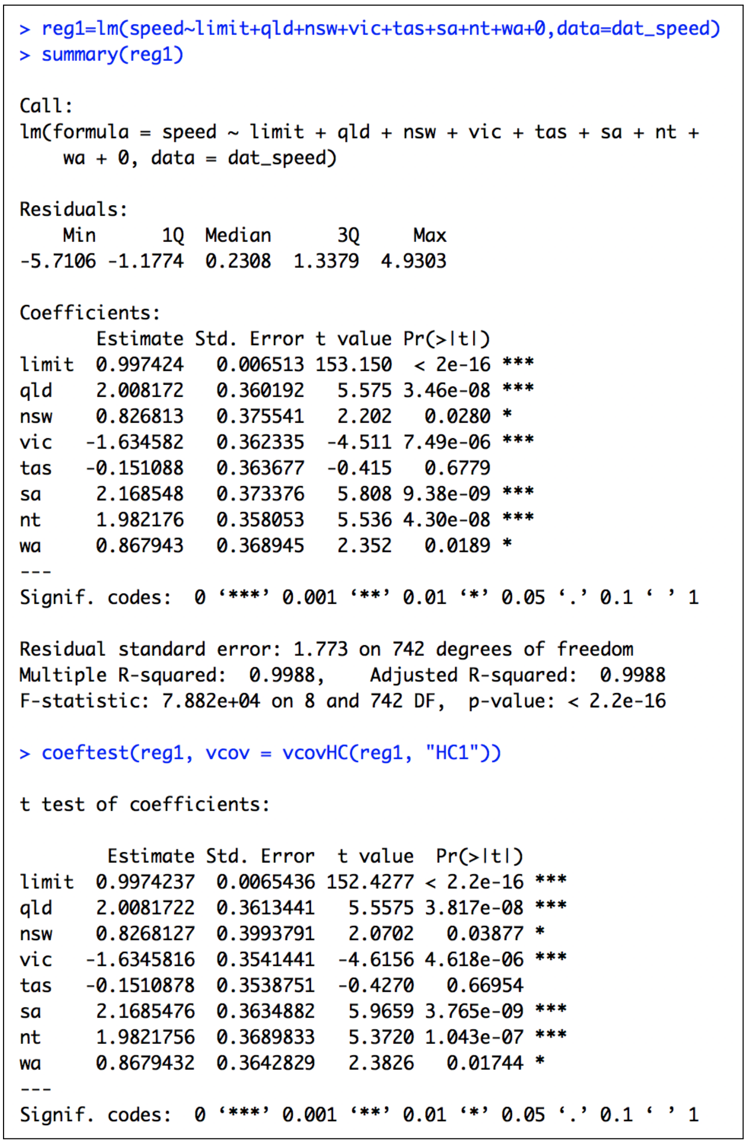
The estimated intercept for Victoria in Figure 4 is -1.63 km/hr, while the estimated intercept for Victoria (holding other regressors fixed) in Figure 5 is 1.98-0.77=1.21. This is a large, 1.21-(-1.63)=2.84 change in the intercept for Victoria, suggesting omitted variable bias.

One source of the bias could be that Victoria has more speeding cameras (or police signs) relative to other states. Given speed cameras and police signs have a negative impact on speed, if there were a positive correlation between the Victoria dummy and the camera or police signs dummies, then this would yield a negative omitted variable bias in Figure 4. Simply put, in this earlier regression, part of the negative Victoria effect on speed could be driven by Victoria having more speed cameras and/or police signs.

g. What is the partial effect on $speed_i$ from having a police sign on a road segment where the speed limit is 40 km/hr, there is a speeding camera, and where the road segment is in Tasmania. (1 mark)

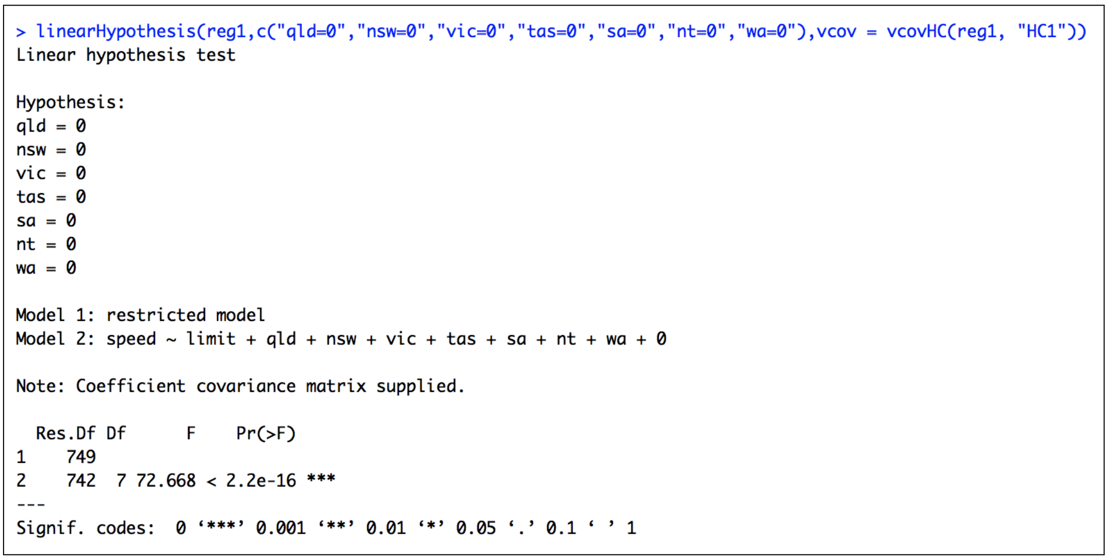
The partial effect is -0.6210010-0.3821607=-1.00 km/hr.

Figure 4: Speed Regression Output 1



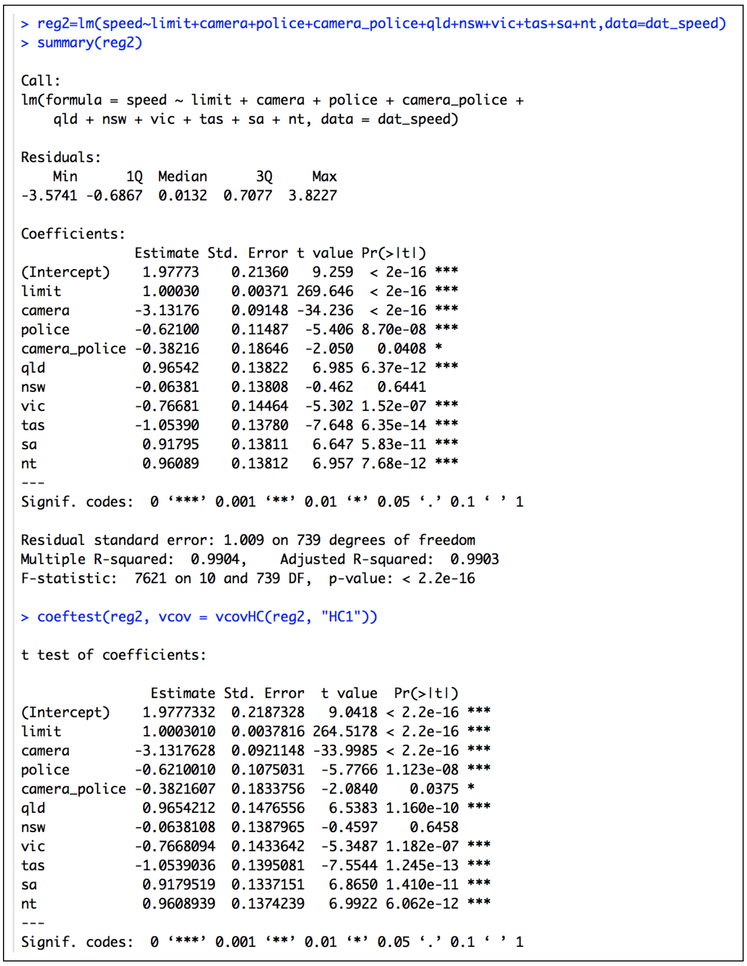
```
> reg1=lm(speed~limit+qld+nsw+vic+tas+sa+nt+wa+0,data=dat_speed)
> summary(reg1)

Call:
lm(formula = speed ~ limit + qld + nsw + vic + tas + sa + nt +
    wa + 0, data = dat_speed)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7106 -1.1774  0.2308  1.3379  4.9303

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
limit  0.997424   0.006513 153.150  < 2e-16 ***
qld    2.008172   0.360192   5.575 3.46e-08 ***
nsw    0.826813   0.375541   2.202   0.0280 *
vic   -1.634582   0.362335  -4.511 7.49e-06 ***
tas   -0.151088   0.363677  -0.415   0.6779
sa     2.168548   0.373376   5.808 9.38e-09 ***
nt     1.982176   0.358053   5.536 4.30e-08 ***
wa     0.867943   0.368945   2.352   0.0189 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.773 on 742 degrees of freedom
Multiple R-squared:  0.9988,	Adjusted R-squared:  0.9988
F-statistic: 7.882e+04 on 8 and 742 DF,  p-value: < 2.2e-16

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

t test of coefficients:

        Estimate Std. Error  t value  Pr(>|t|)
limit  0.9974237  0.0065436 152.4277 < 2.2e-16 ***
qld    2.0081722  0.3613441   5.5575 3.817e-08 ***
nsw    0.8268127  0.3993791   2.0702   0.03877 *
vic   -1.6345816  0.3541441  -4.6156 4.618e-06 ***
tas   -0.1510878  0.3538751  -0.4270   0.66954
sa     2.1685476  0.3634882   5.9659 3.765e-09 ***
nt     1.9821756  0.3689833   5.3720 1.043e-07 ***
wa     0.8679432  0.3642829   2.3826   0.01744 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: Speed Test 1



```
> linearHypothesis(reg1,c("qld=0","nsw=0","vic=0","tas=0","sa=0","nt=0","wa=0"),vcov = vcovHC(reg1, "HC1"))
Linear hypothesis test

Hypothesis:
qld = 0
nsw = 0
vic = 0
tas = 0
sa = 0
nt = 0
wa = 0

Model 1: restricted model
Model 2: speed ~ limit + qld + nsw + vic + tas + sa + nt + wa + 0

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    749
2    742  7 72.668 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: Speed Regression Output 2



```
> reg2=lm(speed~limit+camera+police+camera_police+qld+nsw+vic+tas+sa+nt,data=dat_speed)
> summary(reg2)

Call:
lm(formula = speed ~ limit + camera + police + camera_police +
    qld + nsw + vic + tas + sa + nt, data = dat_speed)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5741 -0.6867  0.0132  0.7077  3.8227

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.97773    0.21360   9.259  < 2e-16 ***
limit          1.00030    0.00371 269.646  < 2e-16 ***
camera        -3.13176    0.09148 -34.236  < 2e-16 ***
police        -0.62100    0.11487  -5.406 8.70e-08 ***
camera_police -0.38216    0.18646  -2.050   0.0408 *
qld            0.96542    0.13822   6.985 6.37e-12 ***
nsw           -0.06381    0.13808  -0.462   0.6441
vic           -0.76681    0.14464  -5.302 1.52e-07 ***
tas           -1.05390    0.13780  -7.648 6.35e-14 ***
sa             0.91795    0.13811   6.647 5.83e-11 ***
nt             0.96089    0.13812   6.957 7.68e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.009 on 739 degrees of freedom
Multiple R-squared:  0.9904,    Adjusted R-squared:  0.9903
F-statistic:  7621 on 10 and 739 DF,  p-value: < 2.2e-16

> coeftest(reg2, vcov = vcovHC(reg2, "HC1"))

t test of coefficients:

                Estimate Std. Error  t value  Pr(>|t|)
(Intercept)    1.9777332  0.2187328   9.0418 < 2.2e-16 ***
limit          1.0003010  0.0037816 264.5178 < 2.2e-16 ***
camera        -3.1317628  0.0921148 -33.9985 < 2.2e-16 ***
police        -0.6210010  0.1075031  -5.7766 1.123e-08 ***
camera_police -0.3821607  0.1833756  -2.0840    0.0375 *
qld            0.9654212  0.1476556   6.5383 1.160e-10 ***
nsw           -0.0638108  0.1387965  -0.4597    0.6458
vic           -0.7668094  0.1433642  -5.3487 1.182e-07 ***
tas           -1.0539036  0.1395081  -7.5544 1.245e-13 ***
sa             0.9179519  0.1337151   6.8650 1.410e-11 ***
nt             0.9608939  0.1374239   6.9922 6.062e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Question 5: Modeling Unemployment Time Series (10 Marks)

The Reserve Bank of Australia has hired you to develop time series models for the national unemployment rate. They provide you with a time series for just one variable, $unemp_t$ which is the Australian unemployment rate in month t. These data are provided from January 2001 to April 2018 for a total of $T = 208$ observations.

a. The time series is plotted in Figure 7 on the next page. Does the time series exhibit seasonality? Briefly explain why or why not. (1 mark)

The regularly spaced spikes in the unemployment rate does indeed suggest the presence of seasonality in the data.

b. Figure 8 contains R-Studio output for three different time series models, reg1, reg2, and reg3. The SSR for each regression is also reported after the coefficient estimates. What types of time series models are each of these? (1 mark)

These are AR(1), AR(2), and AR(3) models, respectively.

c. Interpret the magnitude of the regression coefficient in the first regression model, labeled reg1, in Figure 8. (1 mark). Also comment on whether it is statistically significantly different from 0 at the 5% level.

A 1 percentage point increase in lagged unemployment rate has a corresponding 0.90 percentage point increase in current unemployment rate. The p-value for the test of null that this coefficient equals 0 versus the null that it is not equal 0 is less than 0.05, implying a statistically significant coefficient.

d. Using an information criterion, select the "best" time series model for $unemp_t$ from Figure 8. (1 mark)

You could use either the BIC or AIC to choose the model.
The BIC's for the 3 AR models are: -2.287, -2.345, and -2.458.
The AIC's for the 3 AR models are: -2.319, -2.393, and -2.522.
Thus, the AR(3) model is what minimizes both the BIC and AIC, implying both information criterion woudl select this model.

e. Now consider a richer time series model in Figure 9. This model also includes month of year dummy variables, $jan = 1$ if t is January and 0 otherwise, $feb = 1$ if t is February and 0 otherwise, and so on for all months of the year. What is wrong with the R code as inputted into the lm() command, and what does R do to fix the problem? (1 mark)

The code includes all 12 month dummies in the regression as well as an intercept, which would lead to a dummy variable trap. From the regression output, R drops the December dummy automatically to produce results.

f. Compare the regression coefficient estimates on the lagged regressors in the reg3 model from Figure 9 and the reg4 model in Figure 10. Is there omitted variable bias from not including month-of-the-year dummies in the time series model? Provide intuition for the potential source of the bias. (2 marks)

Unemployment's first lag is much smaller in magnitude in reg3 in Figure 8 compared to that in reg4 in Figure 9. Seasonal persistence in unemployment, which was not controlled for in the Figure 9 results, for instance like that implied by the March through July dummies in Figure 9, could have caused such upward bias in the coefficient on the first lag of unemployment in reg3. This is because, for example, these consecutive similar-impact monthly dummies would be in the error term in reg3, causing upward bias in the one-period lag of unemployment in reg3.

An alternative bias one could discuss is the negative coefficient on lagged twice unemployment in reg3 versus its positive and near 0 coefficient in reg 4. In reg 3, this could be driven by the large change in the month dummy coefficient estimates from January to March, and from November to January. Both of these swings in seasonal unemployment in the error term would have driven the negative coefficient on unemployment lagged twice in reg3.

g. Which months respectively tend to exhibit the highest and lowest levels of unemployment? Interpret the coefficients estimates on the dummy variables for these months and comment on whether they are statistically different from 0 at the 5% level. (1 mark)

January has the highest unemployment rate, as implied by its coefficient 0.539. Holding other regressors fixed, unemployment is 0.539 percentage points higher in January. The p-value for the test of the null that this coefficient is equal to 0 against the alternative that it is not is less than 0.05, implying it is statistically significant.

April has the lowest unemployment rate, as implied by its coefficient -0.402. Holding other regressors fixed, unemployment is 0.402 percentage points lower in April. The p-value for the test of the null that this coefficient is equal to 0 against the alternative that it is not is less than 0.05, implying it is statistically significant.

h. What series of tests are being conducted in Figure 10 on the next page? Carefully describe the outcome of each test at the 5% significance level, noting the relevant test statistic and degrees of freedom (if necessary). (1 mark)

The first test tests whether the coefficients on the dummies for January, February, and March are all equal versus the alternative that at least two are not .The F-stat is 46.977, with an F-distributon with df1=2 and df=208-14-1=193 degrees of freedom, and a p-value less than 0.05 meaning we reject the null.

The second test tests whether the coefficients on the dummies for April, May, and June are all equal versus the alternative that at least two are not .The F-stat is 2.6235, with an F-distributon with df1=2 and df=208-14-1=193 degrees of freedom, and a p-value of 0.07517 meaning we fail reject the null.

The third test tests whether the coefficients on the dummies for July, August, and September are all equal versus the alternative that at least two are not .The F-stat is 10.896, with an F-distributon with df1=2 and df=208-14-1=193 degrees of freedom, and a p-value less than 0.05 meaning we reject the null.

The fourth test tests whether the coefficients on the dummies for Oct and Nov are jointly equal to 0 versus the alternative that at least two are not .The F-stat is 16.346, with an F-distributon with df1=2 and df=208-14-1=193 degrees of freedom, and a p-value less than 0.05 meaning we reject the null.

i. Based on the test results from question g., would it be problematic to use quarter-of-the-year dummies (e.g., for summer, fall, winter, spring) as opposed to month-of-the-year dummies to control for seasonality? Explain. (1 mark)

The test results suggest that it would be problematic to use quarter-of-the-year dummies because these would ignore important differences in monthly seasonal effects within quarters. The first test in particular shows that between January, February, and March there are large differences in seasonality month-to-month, which is clear, for example, when compare the coefficient estimate on January and March in Figure 10.

Figure 7: Unemployment Rate: Jan 2001 - Apr 2018

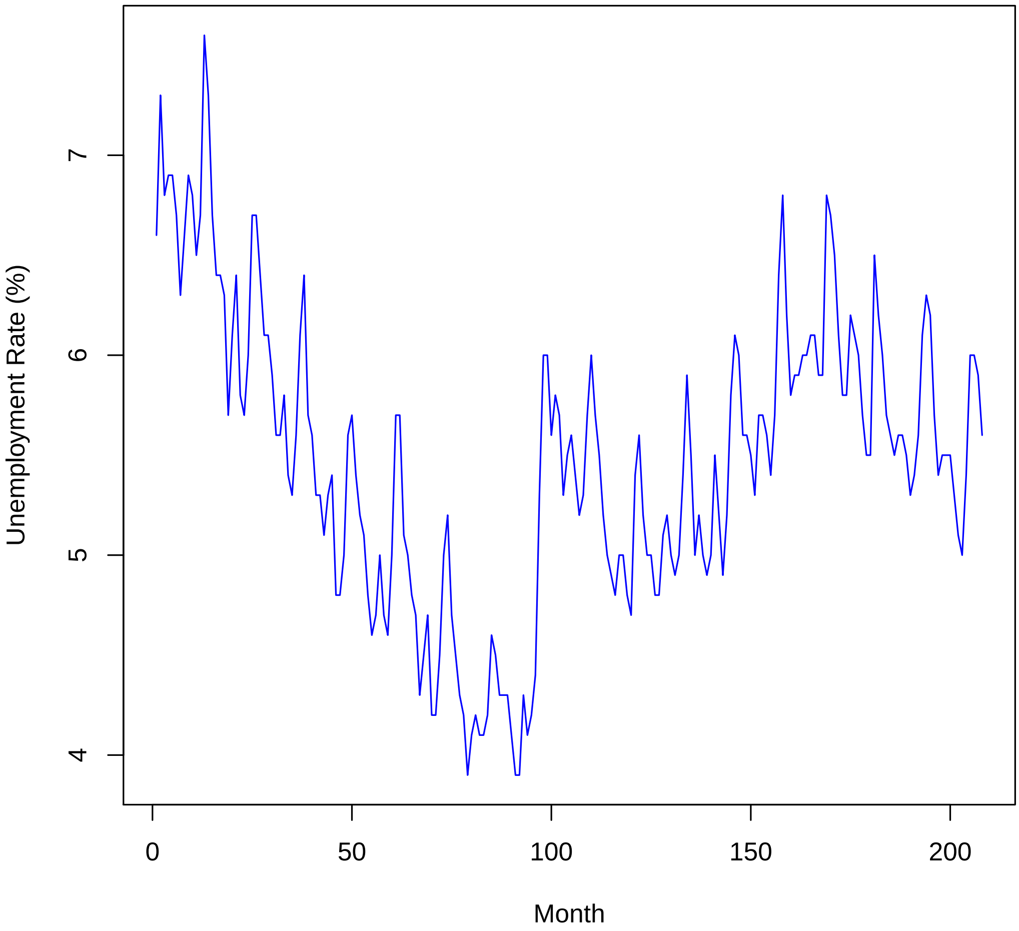
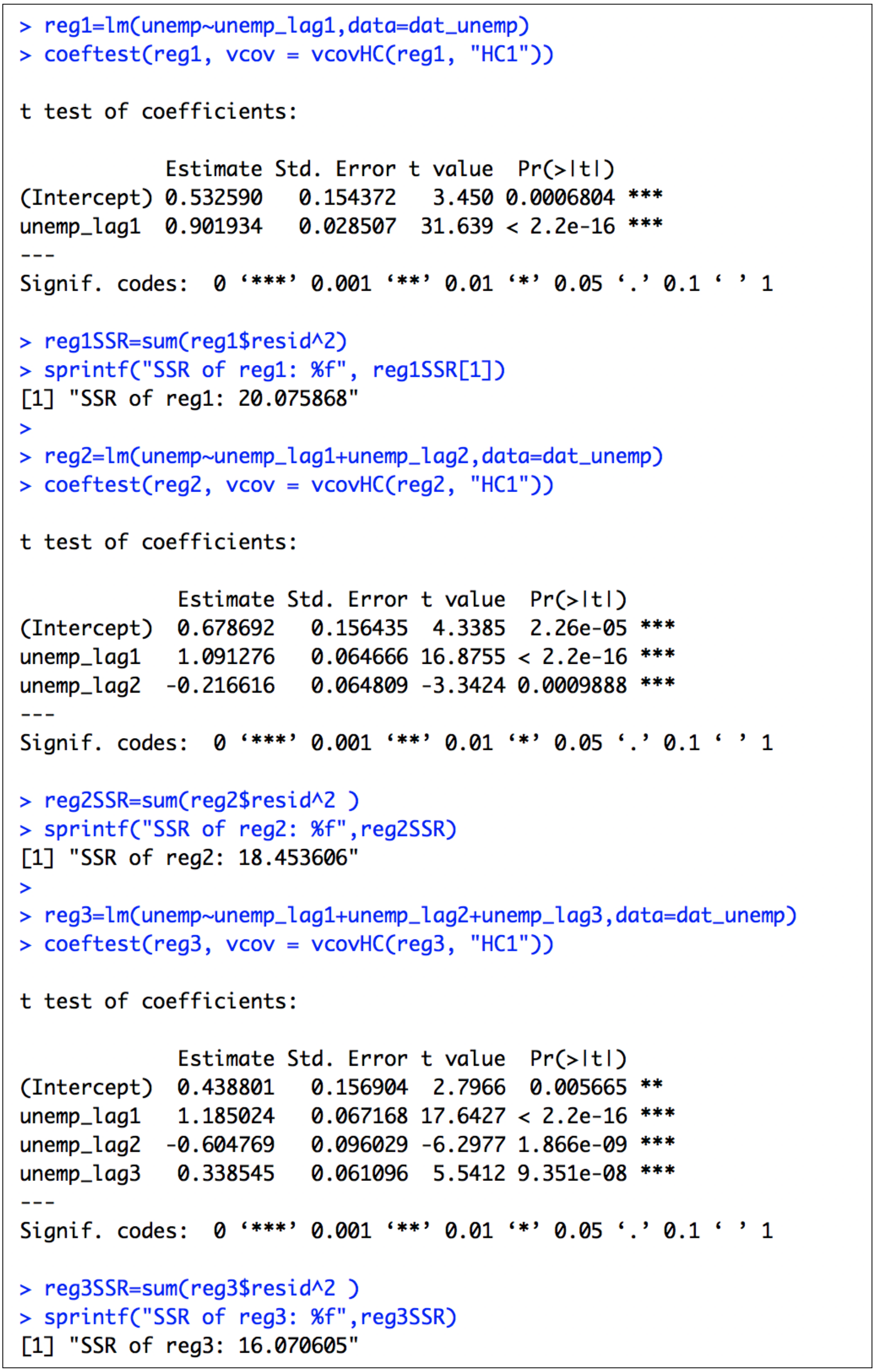
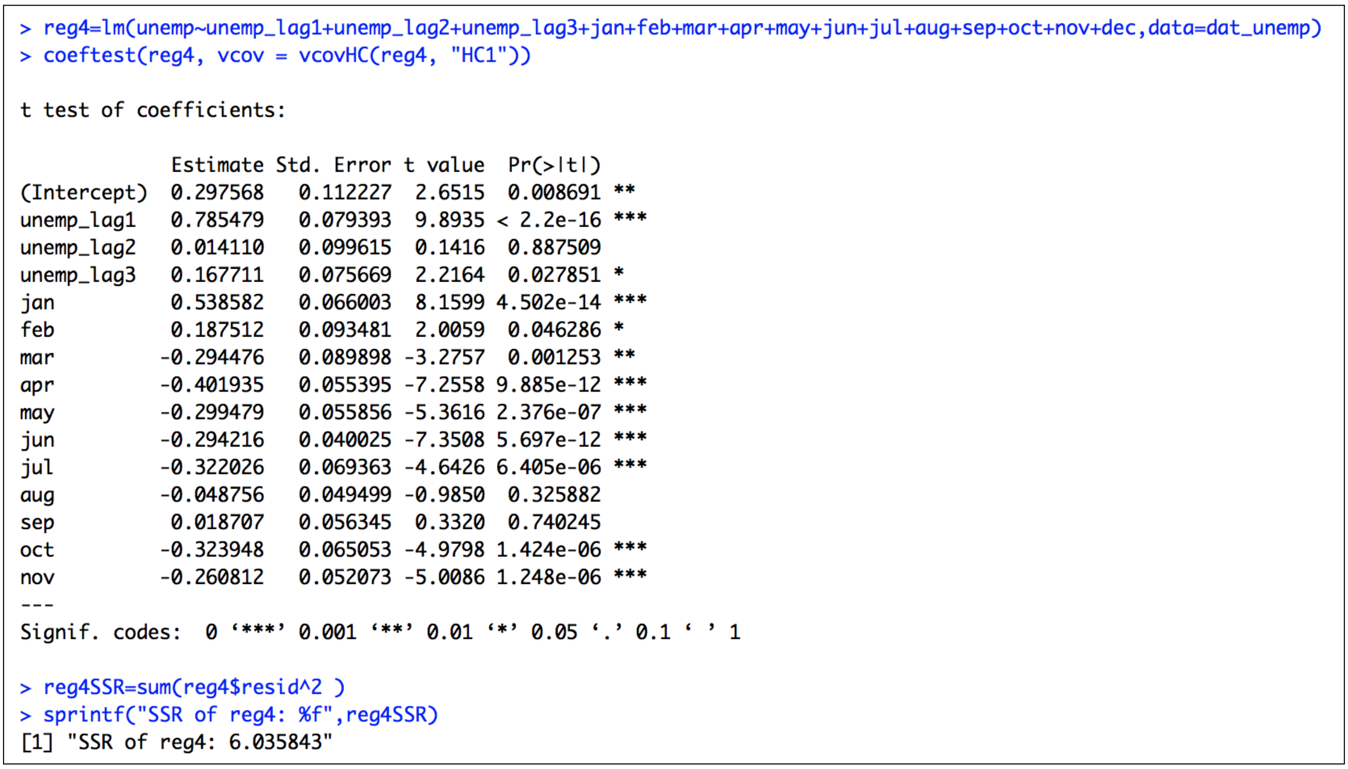


Figure 8: Unemployment Regression Output 1



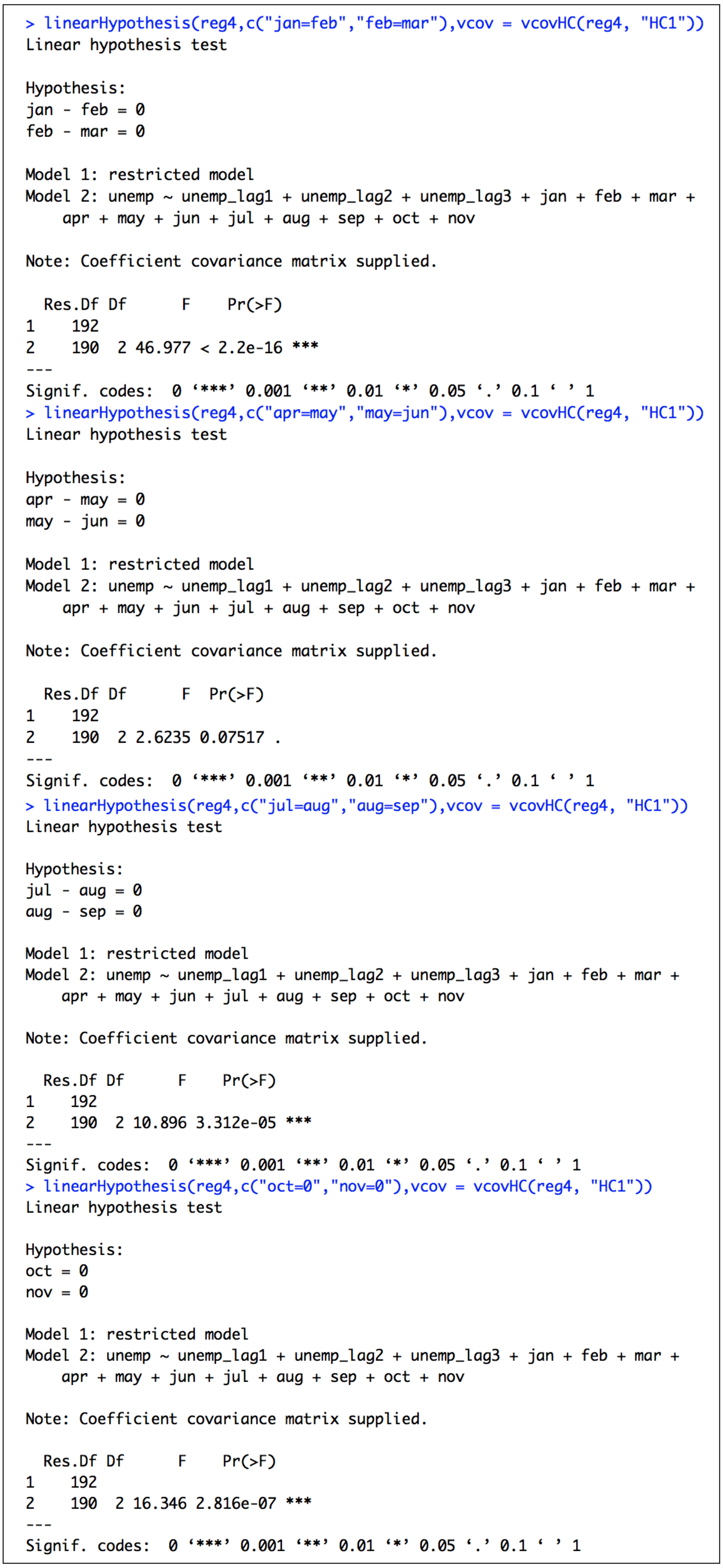
```
> reg1=lm(unemp~unemp_lag1,data=dat_unemp)
> coeftest(reg1, vcov = vcovHC(reg1, "HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 0.532590   0.154372   3.450 0.0006804 ***
unemp_lag1  0.901934   0.028507  31.639 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> reg1SSR=sum(reg1$resid^2)
> sprintf("SSR of reg1: %f", reg1SSR[1])
[1] "SSR of reg1: 20.075868"
>
> reg2=lm(unemp~unemp_lag1+unemp_lag2,data=dat_unemp)
> coeftest(reg2, vcov = vcovHC(reg2, "HC1"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.678692   0.156435  4.3385  2.26e-05 ***
unemp_lag1   1.091276   0.064666 16.8755 < 2.2e-16 ***
unemp_lag2  -0.216616   0.064809 -3.3424 0.0009888 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> reg2SSR=sum(reg2$resid^2 )
> sprintf("SSR of reg2: %f",reg2SSR)
[1] "SSR of reg2: 18.453606"
>
> reg3=lm(unemp~unemp_lag1+unemp_lag2+unemp_lag3,data=dat_unemp)
> coeftest(reg3, vcov = vcovHC(reg3, "HC1"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.438801   0.156904  2.7966  0.005665 **
unemp_lag1   1.185024   0.067168 17.6427 < 2.2e-16 ***
unemp_lag2  -0.604769   0.096029 -6.2977 1.866e-09 ***
unemp_lag3   0.338545   0.061096  5.5412 9.351e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> reg3SSR=sum(reg3$resid^2 )
> sprintf("SSR of reg3: %f",reg3SSR)
[1] "SSR of reg3: 16.070605"
```

Figure 9: Unemployment Regression Output 2



```
> reg4=lm(unemp~unemp_lag1+unemp_lag2+unemp_lag3+jan+feb+mar+apr+may+jun+jul+aug+sep+oct+nov+dec,data=dat_unemp)
> coeftest(reg4, vcov = vcovHC(reg4, "HC1"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  0.297568   0.112227  2.6515  0.008691 ** 
unemp_lag1   0.785479   0.079393  9.8935 < 2.2e-16 ***
unemp_lag2   0.014110   0.099615  0.1416  0.887509    
unemp_lag3   0.167711   0.075669  2.2164  0.027851 *  
jan          0.538582   0.066003  8.1599 4.502e-14 ***
feb          0.187512   0.093481  2.0059  0.046286 *  
mar         -0.294476   0.089898 -3.2757  0.001253 ** 
apr         -0.401935   0.055395 -7.2558 9.885e-12 ***
may         -0.299479   0.055856 -5.3616 2.376e-07 ***
jun         -0.294216   0.040025 -7.3508 5.697e-12 ***
jul         -0.322026   0.069363 -4.6426 6.405e-06 ***
aug         -0.048756   0.049499 -0.9850  0.325882    
sep          0.018707   0.056345  0.3320  0.740245    
oct         -0.323948   0.065053 -4.9798 1.424e-06 ***
nov         -0.260812   0.052073 -5.0086 1.248e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> reg4SSR=sum(reg4$resid^2 )
> sprintf("SSR of reg4: %f",reg4SSR)
[1] "SSR of reg4: 6.035843"
```

Figure 10: Unemployment Regression Testing



```
> linearHypothesis(reg4,c("jan=feb","feb=mar"),vcov = vcovHC(reg4, "HC1"))
Linear hypothesis test

Hypothesis:
jan - feb = 0
feb - mar = 0

Model 1: restricted model
Model 2: unemp ~ unemp_lag1 + unemp_lag2 + unemp_lag3 + jan + feb + mar +
    apr + may + jun + jul + aug + sep + oct + nov

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    192
2    190  2 46.977 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> linearHypothesis(reg4,c("apr=may","may=jun"),vcov = vcovHC(reg4, "HC1"))
Linear hypothesis test

Hypothesis:
apr - may = 0
may - jun = 0

Model 1: restricted model
Model 2: unemp ~ unemp_lag1 + unemp_lag2 + unemp_lag3 + jan + feb + mar +
    apr + may + jun + jul + aug + sep + oct + nov

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F  Pr(>F)
1    192
2    190  2 2.6235 0.07517 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> linearHypothesis(reg4,c("jul=aug","aug=sep"),vcov = vcovHC(reg4, "HC1"))
Linear hypothesis test

Hypothesis:
jul - aug = 0
aug - sep = 0

Model 1: restricted model
Model 2: unemp ~ unemp_lag1 + unemp_lag2 + unemp_lag3 + jan + feb + mar +
    apr + may + jun + jul + aug + sep + oct + nov

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    192
2    190  2 10.896 3.312e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> linearHypothesis(reg4,c("oct=0","nov=0"),vcov = vcovHC(reg4, "HC1"))
Linear hypothesis test

Hypothesis:
oct = 0
nov = 0

Model 1: restricted model
Model 2: unemp ~ unemp_lag1 + unemp_lag2 + unemp_lag3 + jan + feb + mar +
    apr + may + jun + jul + aug + sep + oct + nov

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    192
2    190  2 16.346 2.816e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```