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 $X = 4$?

   a. 0.7

   b. 0.3

   c. 0.5

   d. 0.4

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?

   a. 101.05

   b. 131.09

   c. 99.33

   d. 45.10

3. Why is accounting for heteroskedasticity important?

   a. Ignoring it can lead to biased estimation results

   b. Model fit is improved if heteroskedasticity is also modeled

   c. It creates inconsistent estimation results if not accounted for

   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?

   a. 96

   b. 94

   c. 93

   d. 91

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?

a. $F_{6,428}$

b. $F_{4,424}$

c. $F_{2,421}$

d. $F_{8,420}$

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

a. $Y = \beta_0 + \ln(X) + u$

b. $\ln(Y) = \beta_0 + \ln(X) + u$

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

d. $\ln(Y) = \beta_0 + \ln(X) + u$

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:

a. $Y_t = \beta_0 + (1 + \beta_1)Y_{t-1} + \beta_2 Y_{t-2} + u_t$

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

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

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:

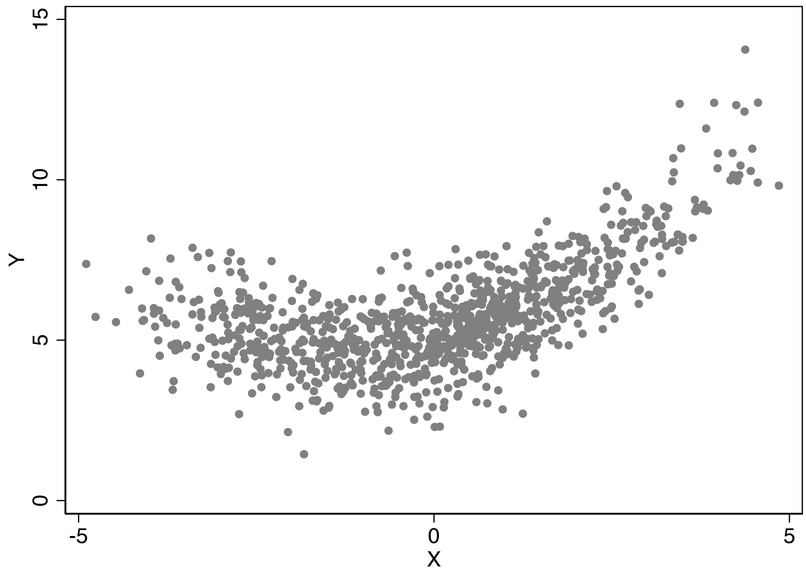
a. use t-statistics for each hypothesis and reject the null hypothesis if all of the restrictions fail.

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

c. use t-statistics for each hypothesis and reject the null hypothesis once the statistic exceeds the critical value for a single hypothesis.

d. use the F-statistic and reject all 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?

a. $Y = \beta_0 + \beta_1 \ln(X) + u$

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

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

d. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + u$

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

a. $Y = \beta_0 X^{\beta_1} + u$

b. $Y = \beta_0 exp(\sqrt{\beta_1} u)$

c. $Y = exp(1/\beta_0 + \beta_1 X + u)$

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)

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)

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)

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:

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

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)

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)

Now suppose Amazon estimates a richer demand model:

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

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)

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] 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)

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.

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)

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)

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)

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)

f. Compare the regression coefficient estimates on vic_i in Figure 5 to the sum of the intercept and the coefficient on vic_i in Figure 6. Is there omitted variable bias with the regression intercept for vic_i (Victoria) in Figure 5? If so, carefully explain a potential source of the bias. (2 marks)

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)

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)

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)

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.

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

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)

f. Compare the regression coefficient estimates on the lagged regressors in the reg3 model from Figure 9 and the reg4 model in Figure 10. Focusing on just one of the regressors from the reg3 model, 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)

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)

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)

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)

Figure 7: Unemployment Rate: Jan 2001 - Apr 2018

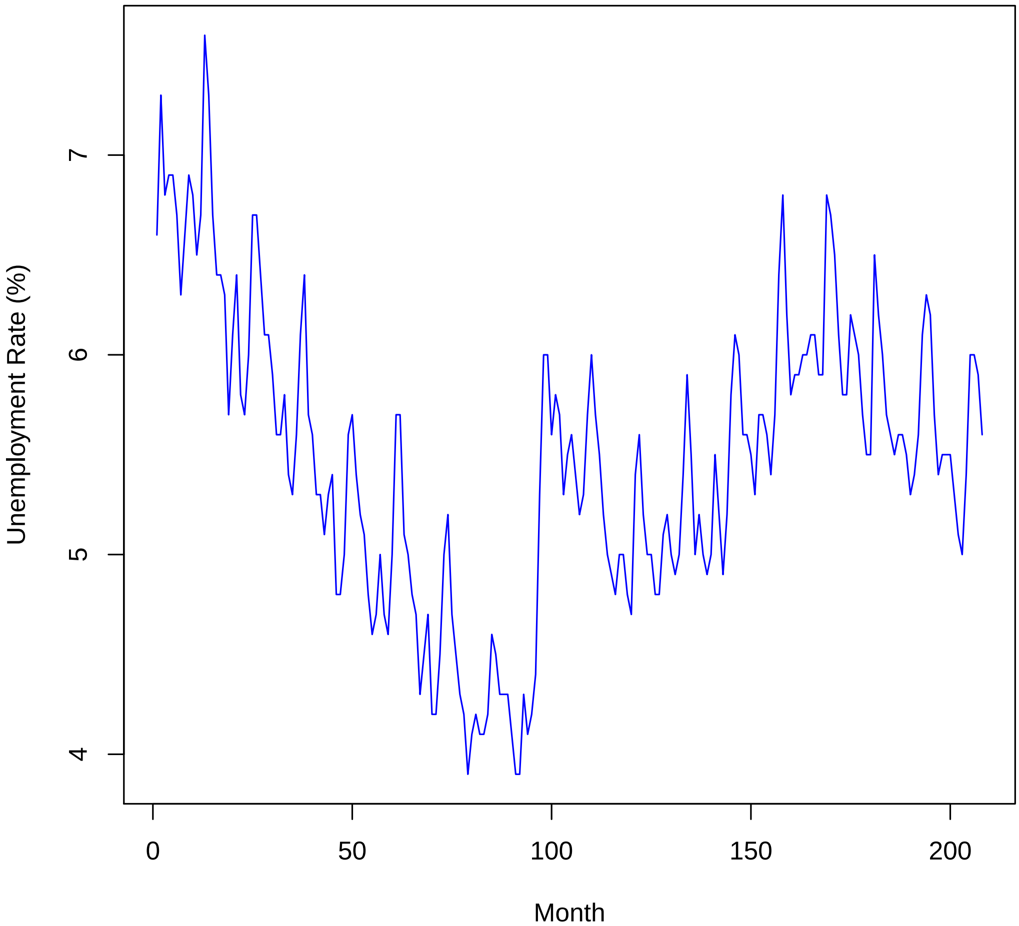


Figure 8: Unemployment Regression Output 1

```
> 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
```

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

$E(c) = c$

$E(cx) = cE(x)$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$F = \frac{1}{2}\left(\frac{(t_1^{act})^2+(t_2^{act})^2-2\hat{\rho}_{t_1^{act},t_2^{act}}t_1^{act}t_2^{act}}{1-\hat{\rho}_{t_1^{act},t_2^{act}}}\right)$ 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: $[\hat{Y}_{T+1|T} - t^\alpha \times SE(Y_{T+1} - \hat{Y}_{T+1|T}), \hat{Y}_{T+1|T} + t^\alpha \times SE(Y_{T+1} - \hat{Y}_{T+1|T})]$

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

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