Chicago Booth BUS 41100
Final **SAMPLE #1**

Instructor: Max H. Farrell

This exam is designed to be **50%** longer than your final will be.

Name: _____ **Section (circle):** 01 – Morning / 02 – Afternoon / 03 – Evening

I pledge my honor that I have not violated the Chicago Booth Honor Code during this exam:

Signed: _____

- You have 3 hours to complete the exam.
- This exam has 18 pages.
- Do not spend an inordinate amount of time on any one problem. Some questions are harder than others. Many questions on the exam are independent of each other.
- The exam is meant to be too long for everyone to finish. Don't worry.
- You may use a calculator and one 8.5×11 size (both sides) "cheat sheet" of your own notes, otherwise the exam is closed book, closed notes, etc.
- Throughout, when calculating probabilities or intervals, you can assume that:
  - 95% of observations will fall within 2 standard deviations of the mean.
  - 90% of observations will fall within 1.6 standard deviations of the mean.
- Present your answers in a clear and concise manner.
- Do **not** write your name on any page except this one.

Good Luck!!

1 Simple Linear Regression Mechanics

You run simple linear regression of Y on X. The **estimated** regression line is $\hat{Y} = 1 + 2X$. The t-statistic for testing the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$ is 2.1, and you reject the null at the 5% level. The sample mean of Y is 7.

(a) What is the standard error of b_1?

(i) 1.48

(ii) 0.95

(iii) 1.96

(iv) 0.72

(v) None of these

(b) Suppose you obtain one more data point, and re-estimate the regression using the original data together with the new data point. The **new** estimated regression line is $\hat{Y} = 1 + 2X$. Using the original data together with the new data point you re-construct the t-statistic for testing the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$. Call this new t-statistic t'. Which of the following is TRUE?

(i) $t' > 2.1$

(ii) $t' = 2.1$

(iii) $t' < 2.1$

(iv) Cannot determine the relationship between t' and 2.1 with the information given.

(c) Which of the following could potentially correspond the the new observation in part **(e)**?

(i) $X = 0, Y = 1$

(ii) $X = 2, Y = 5$

(iii) $X = -3, Y = -5$

(iv) $X = 1, Y = 3$

(v) All of the above

(vi) None of the above

2 Multiple Linear Regression

We want to assess the effect of participating in a job training program on future earnings. For 2,675 men, we observe their participation status (`job.training` = 1 if they received training, 0 if not) and `earnings.after` the program completed (in dollars, regardless of whether they received training). We also observe their `earnings.before` the program, `education` in years, `age`, and binary indicators for being `black`, `hispanic`, `married`, and a `high.school.grad`.

The goal of the study is to assess the association between `job.training` and `earnings.after`. The question is what else to control for, i.e. we must build a model.

(a) First consider the below output from a simple linear regression of `earnings.after` on `job.training`.

```
Call:
lm(formula = earnings.after ~ job.training)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      21553.9      303.6   70.98   <2e-16 ***
job.trainingYes -15204.8     1154.6  -13.17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15150 on 2673 degrees of freedom
Multiple R-squared:  0.06092,   Adjusted R-squared:  0.06057
F-statistic: 173.4 on 1 and 2673 DF,  p-value: < 2.2e-16
```

(i) Give a precise interpretation of the coefficient estimate for `job.training`. Based on this output, what do you conclude about the relationship between participation in the training program and earnings afterward? State a formal hypothesis test, including null and alternative hypotheses, significance level, test statistic value, and conclusion.

(ii) Give a reasonable explanation for why the coefficient is negative.

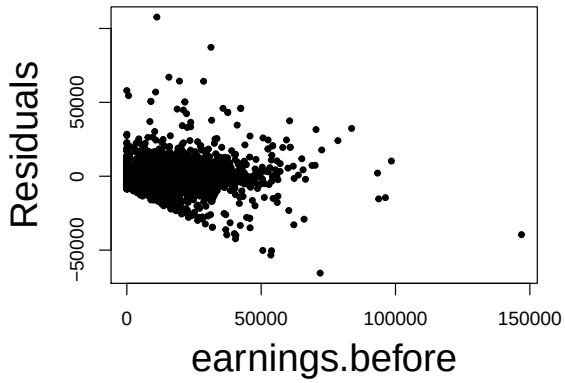
(b) Now consider the following multiple regression output, which controls for `earnings.before` the program and years of `educ`ation.

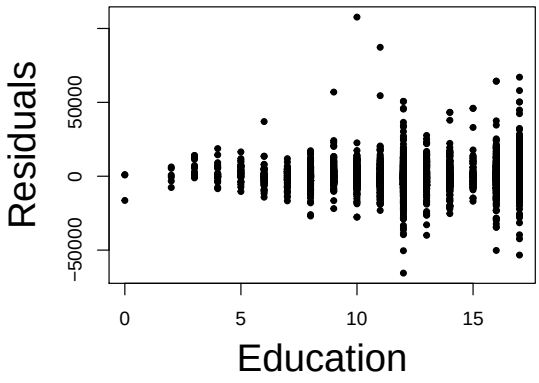
```
Call:
lm(formula = earnings.after ~ job.training + earnings.before +
    educ)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.138e+03  8.160e+02  -2.620  0.00883 **
job.trainingYes  4.792e+02  8.219e+02   0.583  0.55992
earnings.before  8.360e-01  1.657e-02  50.436  < 2e-16 ***
educ             6.275e+02  6.879e+01   9.121  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10160 on 2671 degrees of freedom
Multiple R-squared:  0.5777,    Adjusted R-squared:  0.5772
F-statistic:  1218 on 3 and 2671 DF,  p-value: < 2.2e-16
```

(i) Give a precise interpretation of the coefficient estimate for `job.training`. Based on this output, what do you conclude about the relationship between participation in the training program and earnings afterward? State a formal hypothesis test, including null and alternative hypotheses, significance level, test statistic value, and conclusion.

(ii) What does this conclusion say about the training program?

(iii) Consider the two plots below. What do you conclude about this regression? What problems do you see, and what fixes do you propose for these problems?





(c) Now consider expanding the regression to include all the observed variables described above. We have the following regression output.

```
Call:
lm(formula = earnings.after ~ job.training + earnings.before +
    educ + age + black + hispanic + married + high.school.grad)

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -43.50987 1697.68742  -0.026   0.9796
job.trainingYes     714.92999  920.24279   0.777   0.4373
earnings.before       0.84832    0.01743  48.658  < 2e-16 ***
educ                597.11088  103.86503   5.749  1.0e-08 ***
age                 -88.46638   20.90856  -4.231  2.4e-05 ***
blackYes           -621.57753  497.80243  -1.249   0.2119
hispanicYes        1904.11397 1097.24813   1.735   0.0828 .
marriedYes         1184.57915  589.37459   2.010   0.0445 *
high.school.gradNo  610.11290  650.31497   0.938   0.3482
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10130 on 2666 degrees of freedom
Multiple R-squared:  0.5817,    Adjusted R-squared:  0.5804
F-statistic: 463.4 on 8 and 2666 DF,  p-value: < 2.2e-16
```

We also have the Analysis of Variance Table below, where "Model 2" is the current regression and "Model 1" is the regression in part **(b)**.

```
Model 1: earnings.after ~ job.training + earnings.before + educ
Model 2: earnings.after ~ job.training + earnings.before + educ + age +
    black + hispanic + married + high.school.grad
  Res.Df        RSS Df Sum of Sq      F    Pr(>F)
1   2671 2.7597e+11
2   2666 2.7335e+11  5 2.622e+09 5.1144 0.0001133 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(i) Give a precise description of the p-value in the Analysis of Variance Table above. Include the null and alternative hypotheses of the relevant test, what these mean conceptually for comparing "Model 1" and "Model 2", and what you conclude based on this output.

(ii) "Model 1" from part **(b)** has a BIC value of 49390.32, whereas "Model 2" has a BIC value of 49404.24. Based on this information **and** the Analysis of Variance Table, which model do you prefer? Why?

3 Understanding regression output

We have 70 observations of flat panel TV `price` collected from an online retailer, and we also know the `size` in inches of the diagonal length of the viewing area, the `brand` (indicating LG, Panasonic, or Samsung), and the `type` (indicating an LED or plasma). Our goal is to build a model to predict `price` using the other three variables.

Consider the following regression output:

```
Call:
lm(formula = log(price) ~ size + type + brand)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.975399   0.190335  26.140  < 2e-16 ***
size            0.045329   0.003923  11.554  < 2e-16 ***
typeplasma     -0.266354   0.070914  -3.756 0.000371 ***
brandPanasonic -0.017702   0.085992  -0.206 0.837546
brandSamsung    0.174208   0.064959   2.682 0.009272 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.231 on 65 degrees of freedom
Multiple R-squared:  0.7309,    Adjusted R-squared:  0.7143
F-statistic: 44.13 on 4 and 65 DF,  p-value: < 2.2e-16
```

(a) Conceptually, what do the coefficient estimates for `typeplasma`, `brandPanasonic`, and `brandSamsung` add to our understanding of the relationship between `price` and `size`? That is, visually, what do these variables represent if the relationship were plotted?

(b) Numerically, give an interpretation of the results for `brandPanasonic`. Interpret both the coefficient estimate itself and the associated significance testing.

(c) What would be the p-value for the partial F test of whether `type` is worthwhile to add to the model beyond `size` and `brand`? What do you conclude based on this p-value?

For parts **(d) - (f)** below, consider the following expanded regression.

```
Call:
lm(formula = log(price) ~ size * type + brand)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.992840   0.223091  22.380  < 2e-16 ***
size             0.044945   0.004680   9.603 5.07e-14 ***
typeplasma      -0.331737   0.433194  -0.766  0.44661
brandPanasonic  -0.013657   0.090589  -0.151  0.88064
brandSamsung     0.174483   0.065477   2.665  0.00974 **
size:typeplasma  0.001255   0.008200   0.153  0.87886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2327 on 64 degrees of freedom
Multiple R-squared:  0.731,     Adjusted R-squared:   0.71
F-statistic: 34.78 on 5 and 64 DF,  p-value: < 2.2e-16
```

(d) Conceptually, what does the coefficient estimate for `size:typeplasma` add to our understanding of the relationship between `price` and `size`? That is, visually, what does this variable represent if the relationship were plotted? How is this different from part **(a)**?

(e) What would be the p-value for the partial F test of whether this expanded model is worthwhile beyond the original regression? What do you conclude based on this p-value?

(f) Comparing this regression output to the original, why does your conclusion in part **(e)** make sense? Cite specific values from both outputs.

4 Classification & Model Building

Our goal for this question is to build an email spam filter: based on observed characteristics of an email message we want to build a classification rule for assigning the message either as spam (marked with a "1") or not spam ("0"). To build our filter we have a data of 4601 emails, and for each message we have a human-assigned label `spam` (1 for spam, 0 for not spam) and the following characteristics:

- `caps_avg` = the average of the lengths of strings of capital letters used in the email (e.g. "The" = 1, "HELLO" = 5)
- `c_paren`, `c_exclaim`, `c_dollar` = the percentage of characters in the message which are parentheses ("(", "]", ")", "]"), exclamation point ("!"), and dollar sign ("$") respectively. (Percentages are between 0 and 100.)

(a) In our current context, what are the two types of errors that a classifier can make?

(i)

(ii)

In the present context, is one type of mistake "worse" than the other? Explain your reasoning.

Use the following output to answer parts **(b)** – **(f)**.

```
Call:
glm(formula = spam ~ caps_avg + c_paren + c_exclaim + c_dollar,
    family = "binomial", data = spam)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.75       0.07     -25   <2e-16 ***
caps_avg        0.21       0.02      12   <2e-16 ***
c_paren        -1.66       0.23      -7    2e-13 ***
c_exclaim       1.38       0.11      12   <2e-16 ***
c_dollar       11.86       0.62      19   <2e-16 ***

    Null deviance: 6170.2  on 4600  degrees of freedom
Residual deviance: 4160.7  on 4596  degrees of freedom
AIC: 4171

Number of Fisher Scoring iterations: 15
```

(b) Provide a precise, numerical interpretation of the coefficient estimate for `c_dollar`. Do you find this result credible? Why or why not?

(c) Provide a precise, numerical interpretation of the coefficient estimate for `caps_avg`. Do you find this result credible? Why or why not?

(d) For a message that is 2% parentheses, 2% exclamation points, has zero dollar signs, and never strings together more than one capital letter, what is the estimated probability that this message is spam?

(e) We will use the regression above to build a classification rule based on the predicted probabilities. For some number K, we will flag a message as spam if the estimated $\mathbb{P}[\texttt{spam} = 1|X] > K$. Referring to your answer in part **(a)**, would you prefer to choose $K = 1/4$, $K = 1/2$, or $K = 3/4$? Why?

(f) For $K = 0.5$, the table below compares the classification results to the human-assigned labeling. Use the table to compute the *rates* of the two types of errors in part **(a)**.

		Classifier ≤ 0.5	Classifier > 0.5
Human-assigned:	`spam`==0	2658	130
	`spam`==1	628	1185

List error rate results:

(i)

(ii)

In addition to the variables defined above, we also have word counts for 43 different words, which are stored in variables like `w_<word>`. For example `w_credit` is the number of times the word "credit" appears in the message. We want to build a classifier that includes these word counts only if that word is relevant for predicting `spam`. We will take the train/test approach, holding out 1000 messages for testing. We will choose variables using (i) forward stepwise selection based on AIC, (ii) on BIC, and (iii) selection using the LASSO.

(g) Briefly explain in words the three approaches to model building, and how they differ from each other and what advantages and disadvantages they each have.

(i) Stepwise AIC:

(ii) Stepwise BIC:

(iii) LASSO:

On the 1000 held-out messages, we obtain predictions from each selected model and flag a message as spam if the estimated $\mathbb{P}[\texttt{spam} = 1|X] > 0.5$ (just like part **(f)**). For each of the three models, we then computed the following for each of the 1000 held-out messages:

$$\text{error} = \texttt{spam} - \mathbb{1}\left\{\mathbb{P}[\texttt{spam} = 1|X] > 0.5\right\}$$

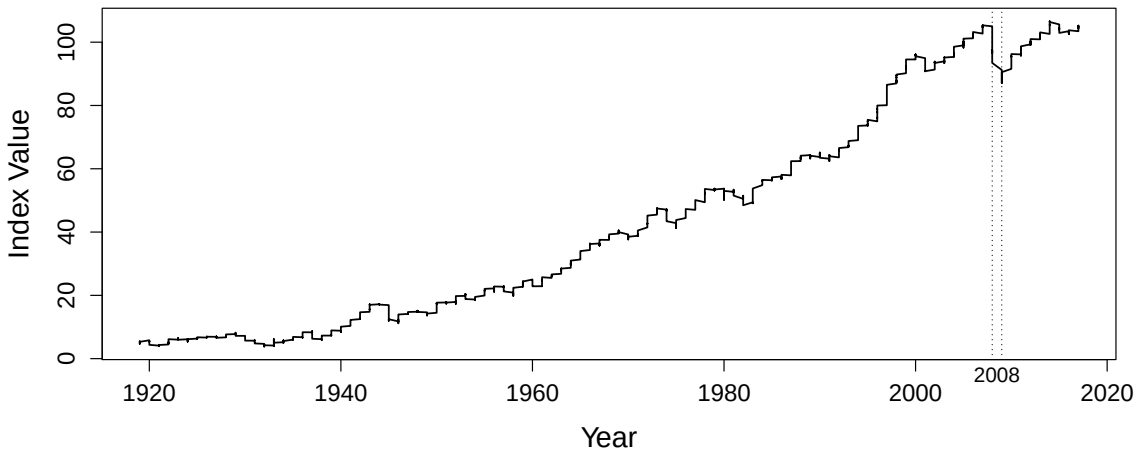
where $\mathbb{1}\left\{\mathbb{P}[\texttt{spam} = 1|X] > 0.5\right\}$ indicates if the estimated $\mathbb{P}[\texttt{spam} = 1|X] > 0.5$. We obtained the following results.

	error		
	-1	0	1
AIC	40	940	20
BIC	50	900	50
LASSO	20	920	60

(h) Based on these results, which method of model selection do you prefer and why? Refer to your answer in part **(a)**.

5 Time Series #1

The Industrial Production Index (`index.prod`) is an economic indicator that measures real output for all facilities located in the United States, including manufacturing, mining, and electric, and gas utilities. It measures movements in production output and highlights structural developments in the economy. The index is compiled on a monthly basis to bring attention to short-term changes in industrial production. We have the index value from January 1919 to September 2017. The series is plotted below, with 2008 between the dotted lines.

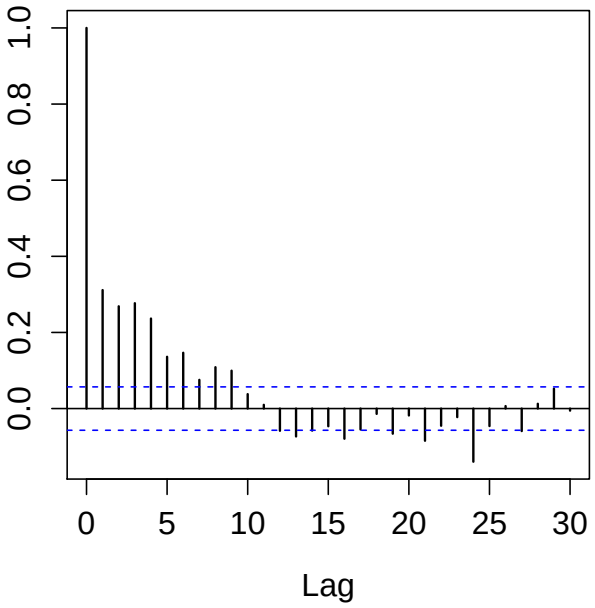


Consider the autoregressive model of order 1, AR(1): $\texttt{index.prod}_t = \beta_0 + \beta_1 \texttt{index.prod}_{t-1} + \varepsilon_t$. The (abbreviated) `summary` output from this regression is below.

```
Call:
lm(formula = index.prod[2:T] ~ index.prod[1:(T - 1)])

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.051139   0.019251    2.66    0.008
index.prod[1:(T - 1)] 1.000758   0.000349 2864.51   <2e-16
---

Residual standard error: 0.406 on 1182 degrees of freedom
Multiple R-squared:     1,      Adjusted R-squared:     1
F-statistic: 8.21e+06 on 1 and 1182 DF,  p-value: <2e-16
```

(a) Interpret the coefficient, standard error, and p-value corresponding to `index.prod[1:(T - 1)]`, in the context of this specific data set.

(b) Below is a plot based on the residuals from the AR(1). Use the space below, describe what is being shown, how this plot is useful as a diagnostic tool, and what you conclude in this example.

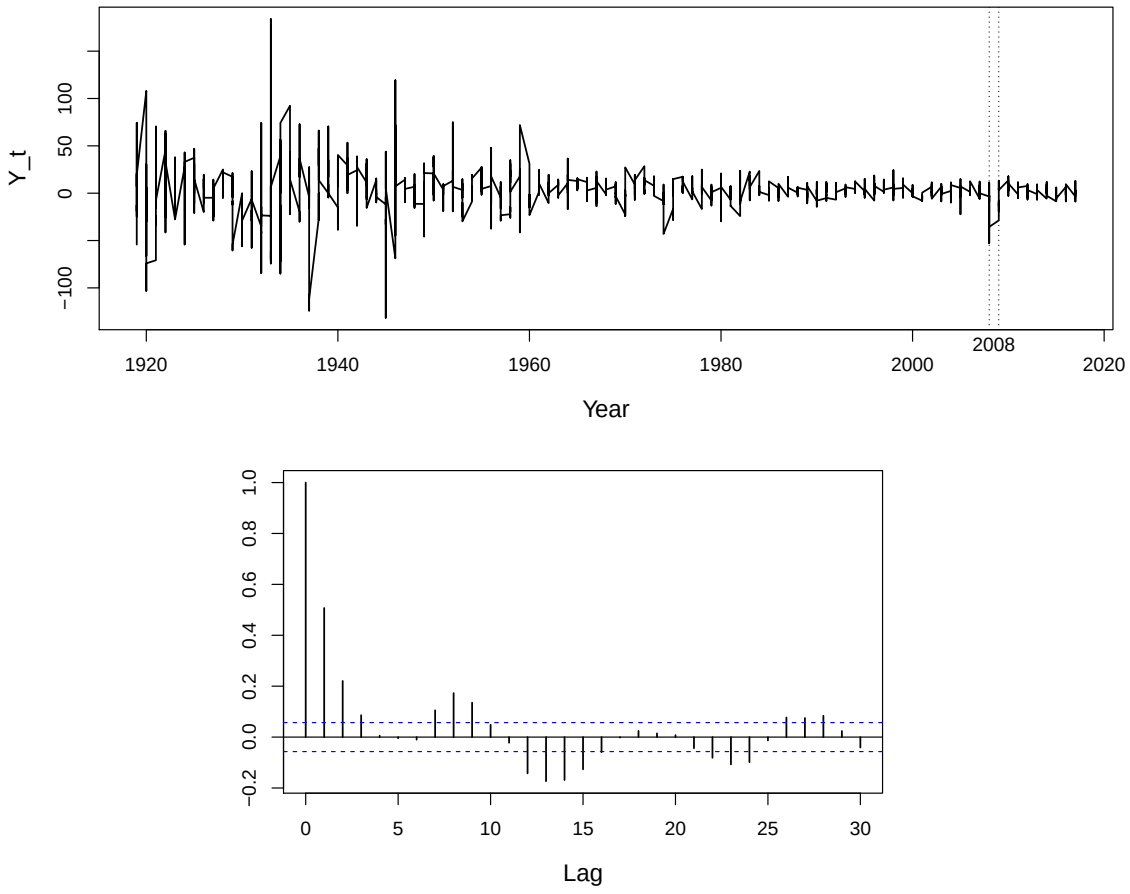


6 Time Series #2

Continuing our study of the Industrial Production Index, we will now consider a transformed series, defined by

$$Y_t = 1200 \times \log\left(\frac{\texttt{index.prod}_t}{\texttt{index.prod}_{t-1}}\right).$$

This new series is plotted below, with 2008 between the dotted lines. Below that is the autocorrelation function (ACF) for Y_t.



(a) A forecaster states that Y_t measures the monthly percentage change in industrial production, measured in percentage points per annum. Is this correct? Explain.

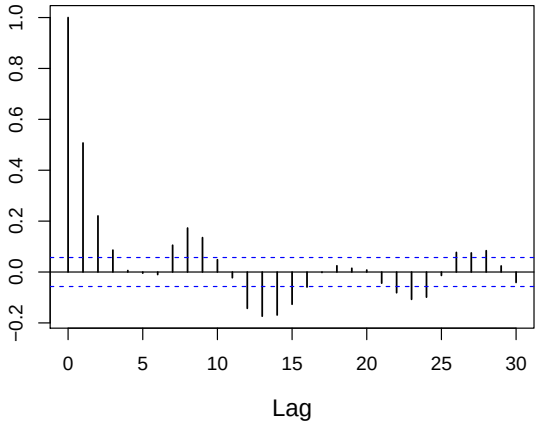
(b) Motivated by the ACF plot, we run the regression model given by

$$Y_t = \beta_0 + \beta_1 \sin(2\pi t/12) + \beta_2 \cos(2\pi t/12) + \varepsilon_t.$$

(i) Give a precise explanation of how this regression models the time series behavior of Y_t.

Abbreviated `summary` output and the ACF plot of the residuals from this model are below.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.07       0.67     4.6    4e-06
sin12           0.30       0.94     0.3      0.8
cos12          -0.54       0.94    -0.6      0.6
---

Residual standard error: 23 on 1181 degrees of freedom
Multiple R-squared: 0.00036,  Adjusted R-squared: -0.0013
F-statistic: 0.22 on 2 and 1181 DF,  p-value: 0.81
```



(ii) Interpret the output and ACF plot. What is the pattern in the plot and why is it there?

(c) We now consider an AR(1) model of Y_t, given by $Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$. Motivated by the idea that the Great Recession in 2008 may have affected the time series, we also try the model $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 \mathbb{1}\{\texttt{year} = 2008\} + \varepsilon_t$. Below is abbreviated `summary` output from each.

AR(1) Model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.564      0.578    2.71   0.0069
Y[1:(TY - 1)]   0.507      0.025   20.29   <2e-16
---

Residual standard error: 19.7 on 1181 degrees of freedom
Multiple R-squared: 0.259, Adjusted R-squared: 0.258
F-statistic:  412 on 1 and 1181 DF,  p-value: <2e-16
```

AR(1) + 2008

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.663      0.581    2.86   0.0043
Y[1:(TY - 1)]        0.505      0.025   20.19   <2e-16
year2008[3:T]TRUE   -9.081      5.724   -1.59   0.1129
---

Residual standard error: 19.7 on 1180 degrees of freedom
Multiple R-squared: 0.26,  Adjusted R-squared: 0.259
F-statistic:  207 on 2 and 1180 DF,  p-value: <2e-16
```

(i) Interpret the p-value for `year2008[3:T]TRUE`. What does it mean for these two models?

(ii) Consider the following R commands and output.

```
> extractAIC(AR1.model)
[1]    2.000 7055.461
> extractAIC(AR1.plus.2008.model)
[1]    3.00 7054.94
```

Explain what these commands do, i.e. what information is being computed and why is it useful. Then, what does this specific output mean for these two models? Compare your answer to part **(i)** above.

(d) Concerned that the AR(1) model may not be adequate, we consider AR(p) models for lag orders p=1, 2, 3, and 4.

(i) Write down the AR(4) model, complete with assumptions.

(ii) The s^2 from each model is given in the table below. Use these numbers and AIC to estimate the number of lags that should be in the model. Circle your choice in the table. Show your work.

AR order:	1	2	3	4
$\mathbf{s^2}$	388.5	388.2	387.7	387.8