Statistics MS Proficiency Exam
31 May, 2023

This exam consists of eight problems. You must complete five of them satisfactorily in order to pass the exam. You are free to attempt as many as you have time to complete. You have three hours to complete the exam.

1. Consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where the ε_i are independent and identically distributed normal random variables with mean 0 and variance σ^2.

   (a) Derive the least squares estimators (LSE) of β_0 and β_1. Argue that they minimize the error sum of squares.

   (b) Show that the LSEs are unbiased for their respective parameters and find their variances.

   (c) suppose the sample standard error of our slope is 1.3. We observe $\hat{\beta}_1 = 1.63$. Construct a 95% confidence interval for β_1 based on a sample of size 18.

   (d) Carry out a hypothesis test for the significance of the regression at the 5% level without relying on part c.

2. Consider the regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where the ε_i are independent and identically distributed normal random variables with mean 0 and variance σ^2. Here, y_i is the percentage grade earned on the MathStat 215 final exam, and x_i is the indicator for whether student i was in class at least 80% of the time. We have six students in our analysis. See Table 1 for the data. Construct the design matrix and use it to find estimates of β_0 and β_1.

i	y_i	x_i
1	91	1
2	72	0
3	64	0
4	87	1
5	39	0
6	98	1

Table 1: Data for Problem 2

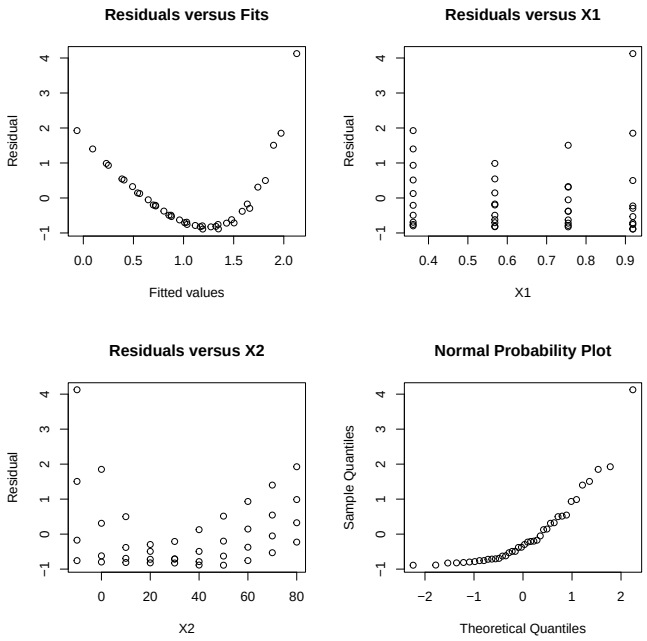

Figure 1: Residual Plots for Problem 3

3. A particular data set yielded the following estimated regression equation:

$$y_i = 0.6794 + 1.4073x_{1i} - 0.0156x_{2i}.$$

A partial ANOVA table is given in Table 2. This table is based on a sample size of 40 observations.

Source of Variation	df	SS	MS	F
Regression		11.4955		
Error				
Total		13.9840		

Table 2: ANOVA Table for Problem 3

(a) Complete the ANOVA table and test for the significance of the regression using the ANOVA table.

(b) Using the plots in Figure 1, carry out a full residual analysis. Suggest, but do not implement, possible data transformations to fix any violations of model assumptions.

4. Smith et al. (1992) discuss a study of the ozone layer over the Antarctic. These scientists developed a measure of the degree to which oceanic phytoplankton production is inhibited by exposure to ultraviolet radiation (UVB). The response is INHIBIT. The regressors are UVB and SURFACE, which is depth below the oceans surface from which the sample was taken. The data consist of 17 observations. SURFACE is an indicator for whether the depth is deep (1) or surface-level (0).

   (a) Write out a statistical model to describe how we would relate the two predictors to the response.

   (b) Using the plots in Figure 2, carry out a full residual analysis for this problem.

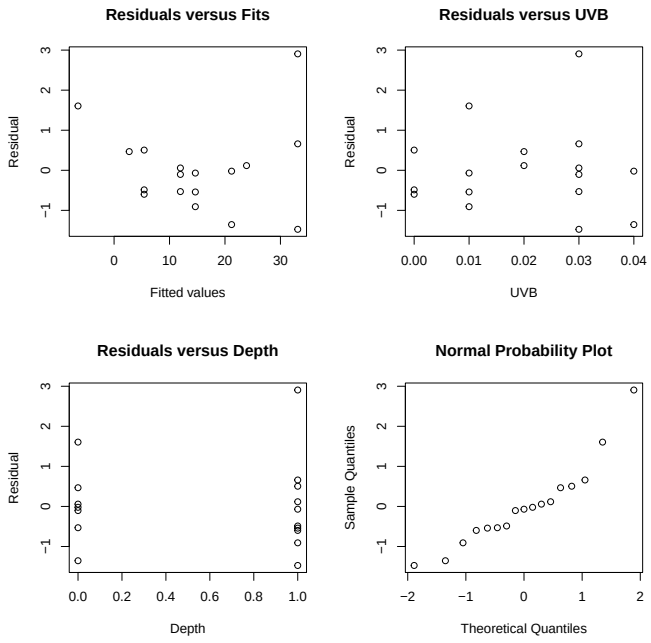

Figure 2: Residual Plots for Problem 4

   (c) If we wanted to add an intermediate depth to the model, how would your model change? Write down the new model. Discuss how regression and error degrees of freedom change.

5. Let $Y_t = 0.7Y_{t-1} + \varepsilon_t$, where the ε_t represent a normally distributed white noise process with mean 0 and variance σ^2. Find the mean function, autocovariance function, and autocorrelation function for Y_t and use it to determine whether the series is stationary. Assume that $\mathbb{E}[Y_0] = 0$.

6. Suppose that for the process in problem 5, we estimate that $Y_t = 0.5893Y_{t-1} + \varepsilon_t$. Suppose $Y_0 = 2$. Find the first three ψ-weights of the process and use them to construct 95% forecast intervals for lead times $l = 1, 2, 3$.

7. Using the characteristic equation, verify that the process $Y_t = 0.9Y_{t-1} - 0.2Y_{t-2} + \varepsilon_t$, where ε_t is normally distributed white noise with mean 0 and variance σ^2, is stationary.

8. Using the plots in Figure 3, suggest a model to fit to the data at hand. Next, consider the output in Table 3 that shows the results of fitting an AR(1) model and that of Table 4, where an AR(4) model is fit. Is the AR(4) an overfit? Based on your results from the tables, what type of model would you suggest? Based on the plots in Figures 4 and 5, respectively, carry out a full residual analysis of each model.

Table 3: AR(1) Table for Problem 8

AR(1)	SE	AIC
0.2364	0.0660	-515.11

Table 4: AR(4) Table for Problem 8

	AR(1)	AR(2)	AR(3)	AR(4)
Coefficient	0.2673	-0.1550	0.0238	-0.0970
SE	0.0669	0.0691	0.0691	0.0681
AIC				-515.64

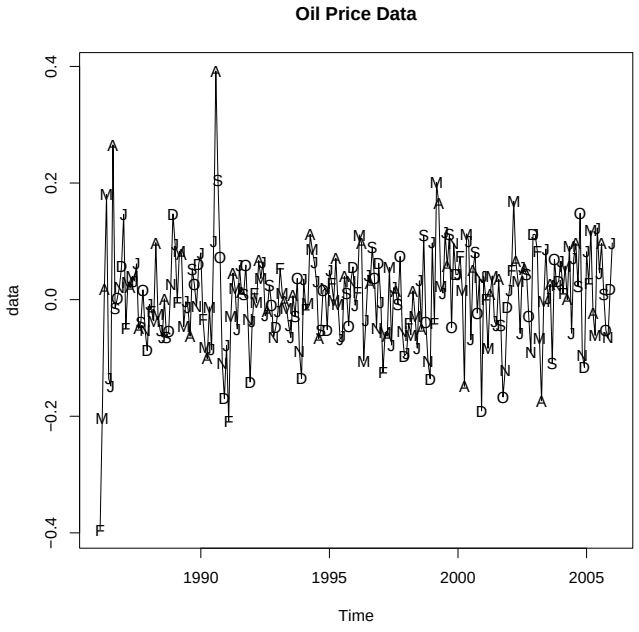

Figure 3: ACF and PACF for Problem 8

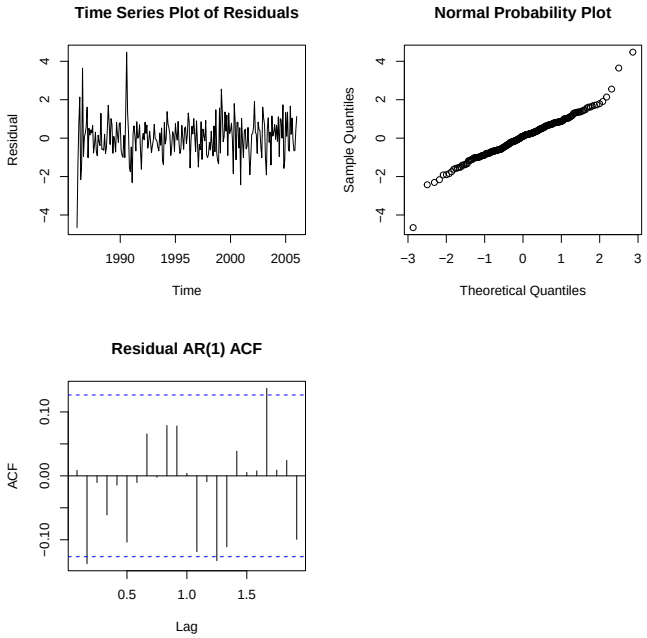

Figure 4: Residual Plots for AR(1) Problem 8

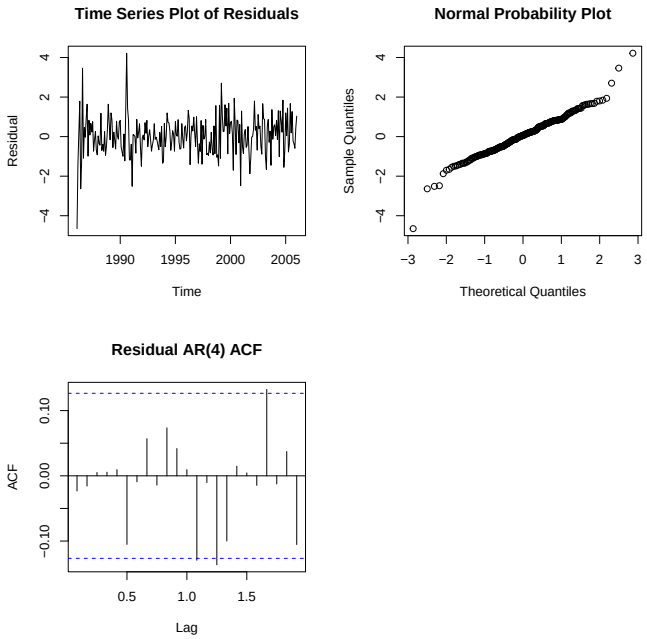

Figure 5: Residual Plots for AR(4) Problem 8