

1.) For an AR(1) process:

$$e_t = \rho e_{t-1} + v_t \quad \text{for } t = 2, 3, \dots, n \text{ and where } v_t \stackrel{iid}{\sim} N(0, \sigma_v^2)$$

Show that  $\text{corr}(e_t, e_{t-2}) = \rho^2$

$$\boxed{\text{NOTE: } \text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}; \text{cov}(x, y) = E[(x - E[x])(y - E[y])] = E[xy] - E[x]E[y]}$$

• For our example:  $\text{WTF } \text{corr}(e_t, e_{t-2}) = \frac{\text{cov}(e_t, e_{t-2})}{\sigma_{e_t} \sigma_{e_{t-2}}}$

$$\begin{aligned} \text{var}(e_t) &= \text{var}(\rho e_{t-1} + v_t) = \rho^2 \text{var}(e_{t-1}) + \text{var}(v_t) + 2\text{cov}(e_{t-1}, v_t) \\ &= \rho^2 \sigma_e^2 + \sigma_v^2 = \sigma_e^2 \Leftrightarrow \sigma_e^2 = \frac{\sigma_v^2}{1 - \rho^2} \end{aligned}$$

• Alternatively, we can do the following:

$$\begin{aligned} \text{var}(e_t) &= E[e_t^2] - E[e_t]^2 = E[e_t^2] \\ &= E[(\rho e_{t-1} + v_t)^2] = E[\rho^2 e_{t-1}^2 + 2\rho e_{t-1} v_t + v_t^2] \\ &= \rho^2 E[e_{t-1}^2] + 2\rho E[e_{t-1} v_t] + E[v_t^2] \end{aligned}$$

$$\boxed{\text{NOTE: } e_t \text{ \& } v_t \text{ are independent} \Rightarrow E[e_{t-1} v_t] = E[e_{t-1}] E[v_t]} \\ = \rho^2 \sigma_e^2 + \sigma_v^2 = \sigma_e^2 \Leftrightarrow \sigma_e^2 = \sigma_v^2 / (1 - \rho^2)$$

$$\begin{aligned} \text{cov}(e_t, e_{t-2}) &= E[e_t e_{t-2}] - E[e_t] E[e_{t-2}] \\ &= E[(\rho e_{t-1} + v_t) e_{t-2}] = E[\rho e_{t-1} e_{t-2} + v_t e_{t-2}] \\ &= \rho E[e_{t-1} e_{t-2}] + E[v_t e_{t-2}] \\ &= \rho E[(\rho e_{t-2} + v_{t-1}) e_{t-2}] + E[v_t] E[e_{t-2}] \quad (\text{since } v_t \text{ \& } e_{t-2} \text{ are ind.}) \\ &= \rho E[\rho e_{t-2}^2 + v_{t-1} e_{t-2}] \\ &= \rho (\rho E[e_{t-2}^2] + E[v_{t-1}] E[e_{t-2}]) \\ &= \rho^2 \sigma_e^2 \end{aligned}$$

$$\boxed{\text{corr}(e_t, e_{t-2}) = \frac{\text{cov}(e_t, e_{t-2})}{\sigma_e \cdot \sigma_e} = \frac{\rho^2 \sigma_e^2}{\sigma_e^2} = \rho^2}$$

- 2.) Consider the regression model:  $y_t = \beta_0 + \beta_1 x_t + e_t$  where the  $e_t$  follow an AR(1) process. Conduct a simulation to examine the coverage probabilities of nominal 95% CI for the mean response when  $\rho = 0.1, 0.2, \dots, 0.9$  and when the usual LS is used to fit the model (i.e. assuming no serial correlation). Provide either a table or plot showing the actual coverage probabilities as a function of  $\rho$ . In your simulation, use  $x_1, x_2, \dots, x_n \stackrel{iid}{\sim} U(0,1)$ ;  $\beta_0 = 0.5$ ,  $\beta_1 = 1.5$ ,  $n = 50$  and  $\sigma_e^2 = 0.5$ . In each simulation, compute a CI for the mean response when  $x = 0.5$ . Use  $M = 1000$  simulations for each value of  $\rho$ .

(\* Done in R; see Figure 1. \*)

### 3.) Exercise 9.2 in Textbook:

(a) Model:  $y_i = \beta_0 + \beta_1 \text{time} + \beta_2 \text{Month 2} + \dots + \beta_{12} \text{Month 12} + e_i$

- Looking at Figure 2, it is hard to tell if there is a pattern in the time series plot of the <sup>standardized</sup> residuals. To investigate further, we can make an acf plot of the standardized residuals (Figure 3). Looking at the acf plot we can see that our residuals seem to have an AR(1) structure w/  $p = -0.2231149 < -2/\sqrt{35} = -0.2073903$ . Thus, the assumption that

the residuals are independent seems to be violated.

it is interesting to note that

- However, if we conduct a 95% CI for the Spearman-Rank correlation

coefficient (use Spearman rank b/c residuals aren't normally distributed)

we get  $(-0.31900551, 0.8476474)$ . Thus, we don't have evidence at the  $\alpha=0.05$

level to conclude that  $\rho \neq 0$ . However, b/c tests of this sort were not discussed

in lecture, we will continue as though we do have significant evidence that the lag 1 auto correlation is significant.

Ask about note below:

- Is it common to conduct a hypothesis on the lag(1) auto correlation to test if it is significantly different from 0?
- Is one of the assumptions for the acf test that our residuals are  $i.i.d. N(0, \sigma^2)$ ? And if they're not should we be using something like Spearman-Rank cor. as our estimates of auto correlation at lag(1)?

- The assumption of constant variance doesn't seem to be violated, however (as seen in Figure 4) the assumption that the errors are normally distributed does seem to be violated.
- Additionally, there are a few influential points (obs 11, 32, 33, 39, 89) all have Cook's D values  $\geq \frac{4}{33-12-1} = 0.05$ .

(b) Model:  $y_i = \beta_0 + \beta_1 \text{Advert} + \beta_2 \text{log(Advert)} + \beta_3 \text{time} + \beta_4 \text{Month 2} + \dots + \beta_{12} \text{Month 12} + e_i$

- Looking at figures 5, 6, 7 we can see that the model in (b) has the same violations of assumptions as the model in (a). Additionally, performing a partial F-test, we see that the variables added to the model in (b) do not significantly reduce RSS. Thus, I would use model (a).

for the correlation between standardized residuals and the lag 1 version of the standardized residuals



Figure 1: (A very nice figure 1)

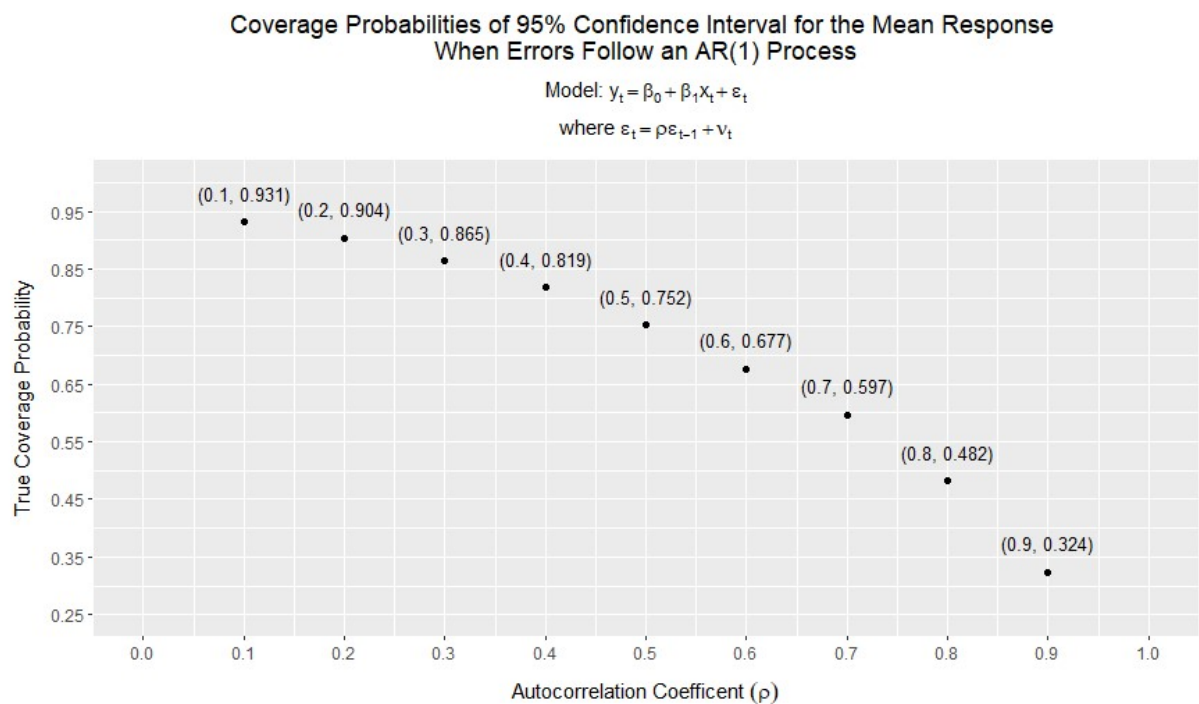


Figure 2:

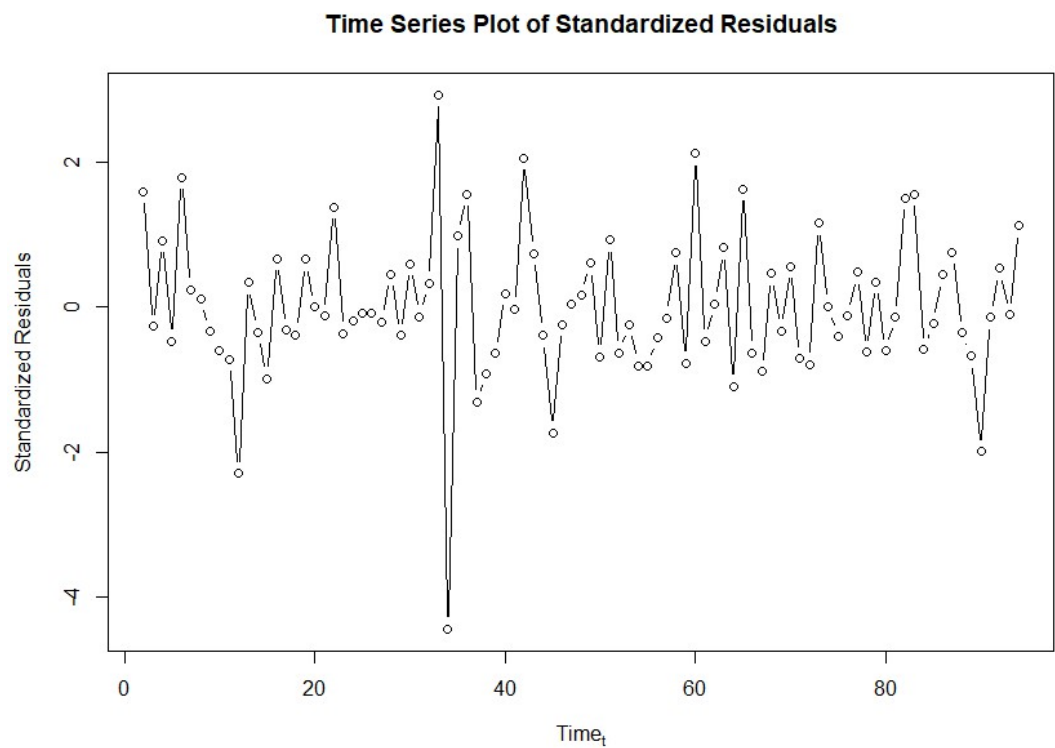


Figure 3:

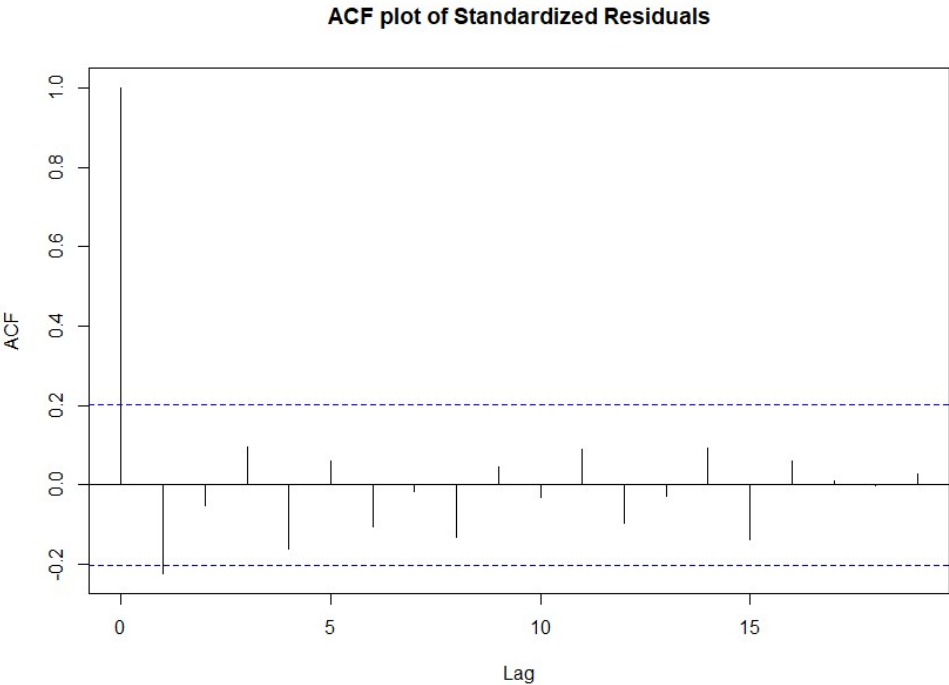


Figure 4:

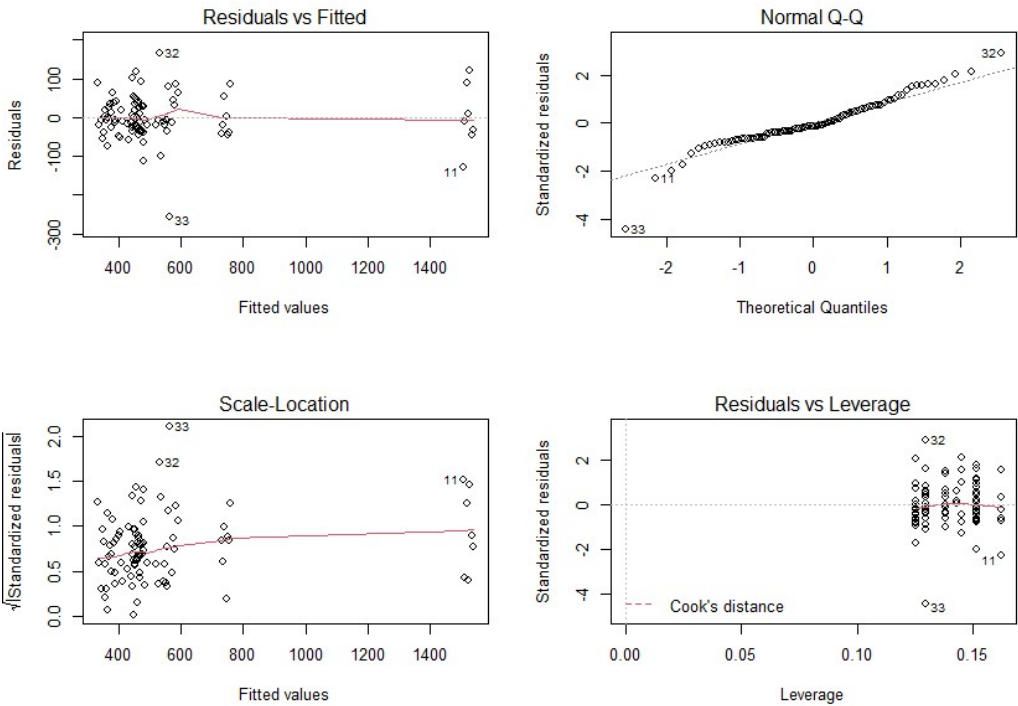


Figure 5:

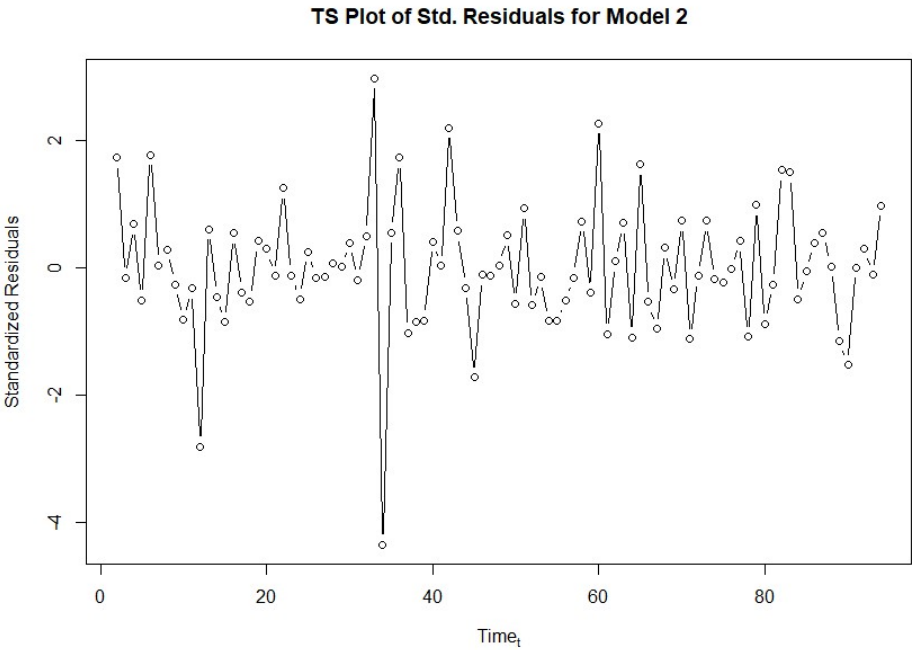


Figure 6:

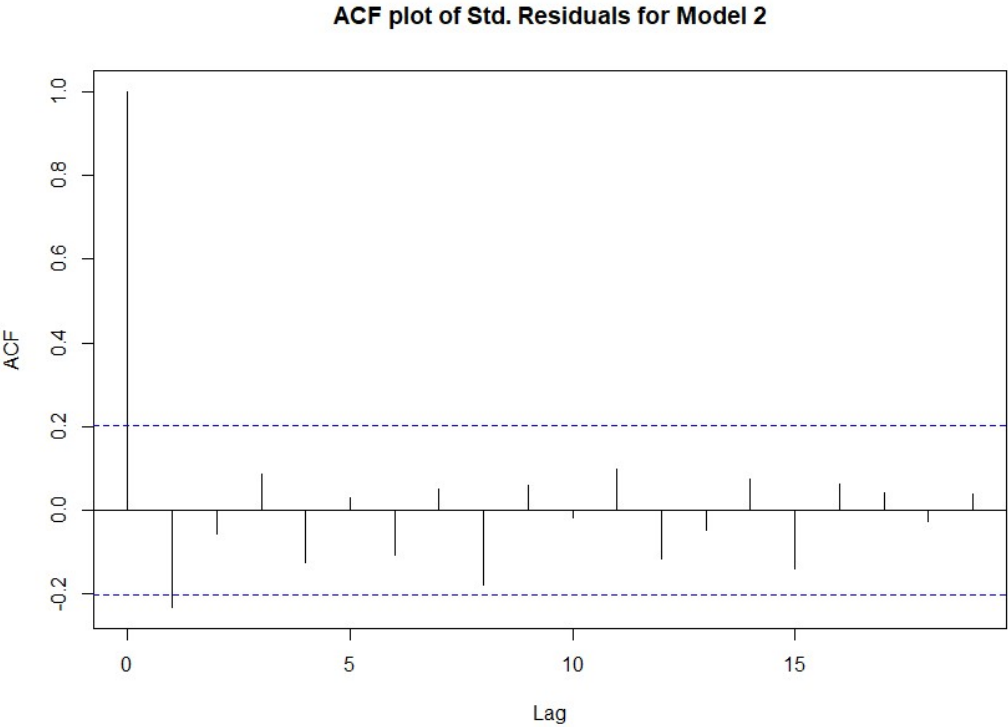


Figure 7:

