Jack Rodoni
STAT 608
Assignment #7
4/29/22

1.) For an AR(1) process:

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

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

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

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

• $var(e_t) = var(\rho e_{t-1} + v_t) = \rho^2 var(e_{t-1}) + var(v_t) + 2cov(e_{t-1}, v_t)$

$= \rho^2 \sigma_e^2 + \sigma_v^2 = \sigma_e^2 \iff \sigma_e^2 = \frac{\sigma_v^2}{1-\rho^2}$

• Alternatively, we can do the following:

$var(e_t) = E[e_t^2] - E[e_t]^2 \overset{(E[e_t]=0)}{=} E[e_t^2]$

$= E[(\rho e_{t-1} + v_t)^2] = E[(\rho e_{t-1})^2 + 2\rho e_{t-1} v_t + v_t^2]$

$= \rho^2 E[e_{t-1}^2] + 2\rho E[e_{t-1} v_t] + E[v_t^2]$

[NOTE: e_t & v_t are independent $\Rightarrow E[e_{t-1} \cdot v_t] = E[e_{t-1}]E[v_t]$]

$= \rho^2 \sigma_e^2 + \sigma_v^2 = \sigma_e^2 \iff \sigma_e^2 = \sigma_v^2 / 1-\rho^2$

• $cov(e_t, e_{t-2}) = E[e_t \cdot e_{t-2}] - \overset{0}{\cancel{E[e_t]}}\overset{0}{\cancel{E[e_{t-2}]}}$

$= E[(\rho e_{t-1} + v_t) e_{t-2}] = E[\rho e_{t-1} e_{t-2} + v_t e_{t-2}]$

$= \rho E[e_{t-1} e_{t-2}] + E[v_t e_{t-2}]$

$= \rho E[(\rho e_{t-2} + v_t) e_{t-2}] + \overset{0}{\cancel{E[v_t]}}\overset{0}{\cancel{E[e_{t-2}]}}$ (since v_t & e_{t-2} are ind)

$= \rho E[\rho e_{t-2}^2 + v_t e_{t-2}]$

$= \rho(\rho E[e_{t-2}^2] + E[v_t]E[e_{t-2}])$

$= \rho^2 \sigma_e^2$

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

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% CIs for the mean response when $\rho = 0.1, 0.2, \ldots, 0.9$ 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 ρ. In your simulation, use $x_1, x_2, \ldots x_n \overset{iid}{\sim} U(0,1)$; $\beta_0 = 0.5$, $\beta_1 = 1.5$, $n = 50$ and $\sigma_v^2 = 0.5$. In each simulation, compute a CI for the mean response when $x = 0.5$. Use $M = 1000$ simulations for each value of ρ.

(* Done in R; see figure 1. *)

Ask about note below:

① Is it common to conduct a hypo test on the lag(1) auto correlation to test if it is significantly different from 0?

② Is one of the assumptions for the acf stuff that our residuals are $\sim N(0, \sigma_v^2)$? And if they're not should we be using something like spearman-rank cor. as our estimates of auto correlation at lag(1)

3.) Exercise 9.2 in Textbook:

(a) Model: $y_i = \beta_0 + \beta_1 \text{time} + \beta_2 \text{Month 2} + \cdots + \beta_{13} \text{Month 12} + e_i$

- Looking at Figure 2, it is hard to tell if there is a pattern in the time series plot of the standardized 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/ $\rho = -0.2231149 < -2/\sqrt{93} = -0.2073903$. Thus, the assumption that the residuals are independent seems to be violated.

- However, it is interesting to note that if we conduct a 95% CI for the spearman-Rank correlation coefficient (use spearman rank b/c residuals aren't normally distributed) for the correlation btwn standardized residuals and the lag1 version of the standardized residuals we get $(-0.31900851, 0.8476474)$. Thus, we dont 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 lag1 auto correlation is significant.

- 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, 59, 89) all have cooks D values $\geq \frac{4}{93-12-1} = 0.05$.

(b) Model: $y_i = \beta_0 + \beta_1 \text{Advert} + \beta_2 \text{lag1 Advert} + \beta_3 \text{time} + \beta_4 \text{Month 2} + \cdots + \beta_{15} \text{Month 12} + e_i$

- 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) to not significantly reduce RSS. Thus, I would use model 1.