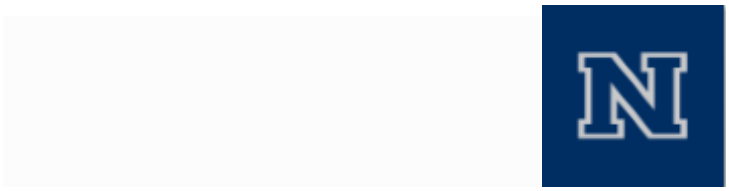
Assignment 9

Biplav Timalsina

April 23, 2018



STAT 757 Applied Regression Analysis

Instructions [20 points]

Modify this file to provide responses to the Ch.9 Exercises in @sheather2009. You can find some helpful code here: http://www.stat. tamu.edu/~sheather/book/docs/rcode/Chapter9.R. Also address the project milestones indicated below. Please email **both** your .Rmd (or roxygen .R) and one of the following either .HTML, .PDF, or .DOCX using the format SURNAME-FIRSTNAME-Assignment9.Rmd and SURNAME-FIRSTNAME-Assignment9.pdf.

Exercise 9.4.1 [60 points]

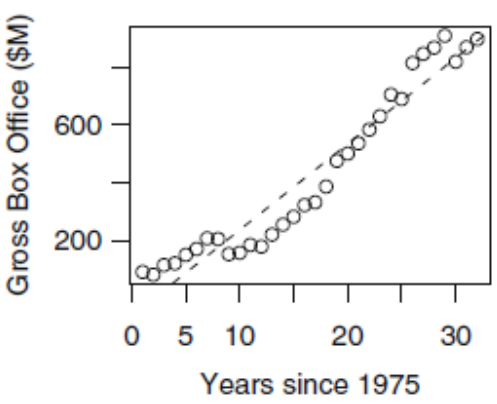
1. Senior management at the Australian Film Commission (AFC) has sought your help with the task of developing a model to predict yearly gross box office receipts from movies screened in Australia. Such data are publicly available for the period from 1976 to 2007 from the AFC's web site (www.afc.gov.au). The data are given in Table 9.2 and they can be found on the book web site in the file boxoffice.txt. Interest centers on predicting gross box office results for 1 year beyond the latest observation, that is, predicting the 2008 result. In addition, there is interest in estimating the extent of any trend and autocorrelation in the data. A preliminary analysis of the data has been undertaken by a staffer at the AFC and these results appear below. In this analysis the variable Year was replaced by the number of years since 1975, which we shall denote as YearsS1975 (i.e., YearsS1975 = Year – 1975).

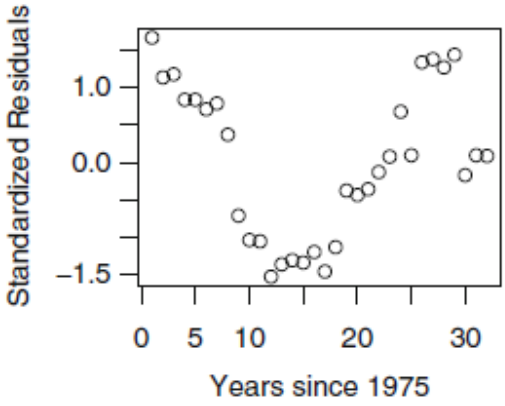
   The first model fit to the data by the staffer was

$$\text{GrossBoxOffice} = \beta_0 + \beta_1 \text{YearsS1975} + e \tag{9.8}$$

Table 9.2 Australian gross box office results

Year	Gross box office ($M)	Year	Gross box office ($M)
1976	95.3	1992	334.3
1977	86.4	1993	388.7
1978	119.4	1994	476.4
1979	124.4	1995	501.4
1980	154.2	1996	536.8
1981	174.3	1997	583.9
1982	210.0	1998	629.3
1983	208.0	1999	704.1
1984	156.0	2000	689.5
1985	160.6	2001	812.4
1986	188.6	2002	844.8
1987	182.1	2003	865.8
1988	223.8	2004	907.2
1989	257.6	2005	817.5
1990	284.6	2006	866.6
1991	325.0	2007	895.4





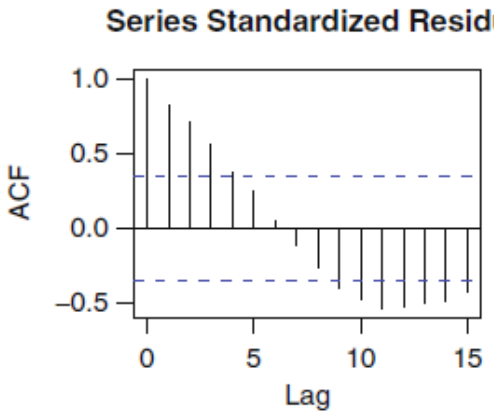


Figure 9.16 Plots associated with the LS fit of model (9.8)

Figure 9.16 shows plots associated with the least squares fit of model (9.8) that were produced by the staffer. The staffer noted that a number of statistically significant autocorrelations in the standardized residuals as well as the existence of an obvious roller coaster pattern in the plot of standardized residuals against

YearsS1975. As such, the staffer decided to fit model (9.8) assuming that the errors are AR(1). Given below is the output from R.

Output from R

```
Generalized least squares fit by maximum likelihood
Model: GrossBoxOffice ~ YearsS1975
Data: boxoffice
     AIC        BIC       logLik
330.3893   336.2522    -161.1947

Correlation Structure: AR(1)
Formula: ~YearsS1975
Parameter estimate(s):
      Phi
0.8782065

Coefficients:
                 Value   Std. Error     t-value     p-value
(Intercept)    4.514082     72.74393    0.062054      0.9509
YearsS1975    27.075395      3.44766    7.853259      0.0000

Correlation:
              (Intr)
YearsS1975    -0.782

Residual standard error: 76.16492
Degrees of freedom: 32 total; 30 residual
```

Given below is the output from R associated with fitting model (9.8) assuming that the errors are AR(1) using least squares based on the transformed versions of the response and predictor variables in (9.6). The staffer was delighted that the results match those in the previous R output.

Output from R

```
Call:
lm(formula = ystar ~ xstar - 1)

Coefficients:
                   Estimate  Std. Error  t value  Pr(>|t|)
xstar(Intercept)      4.514      72.744    0.062      0.95
xstarYearS1975       27.075       3.448    7.853  9.17e-09  ***
```

Figure 9.17 shows diagnostic plots based on the least squares residuals from (9.6). The staffer is relieved that none of the autocorrelations in the right-hand plot are statistically significant indicating that an AR(1) process provides a valid model for the errors in model (9.8). However, the staffer is concerned about the distinct nonrandom pattern in the left-hand plot of Figure 9.17. The dashed line is from a cubic LS fit which is statistically significant (p-value = 0.027). At this stage, the staffer is confused about what to do next and has sought your assistance.

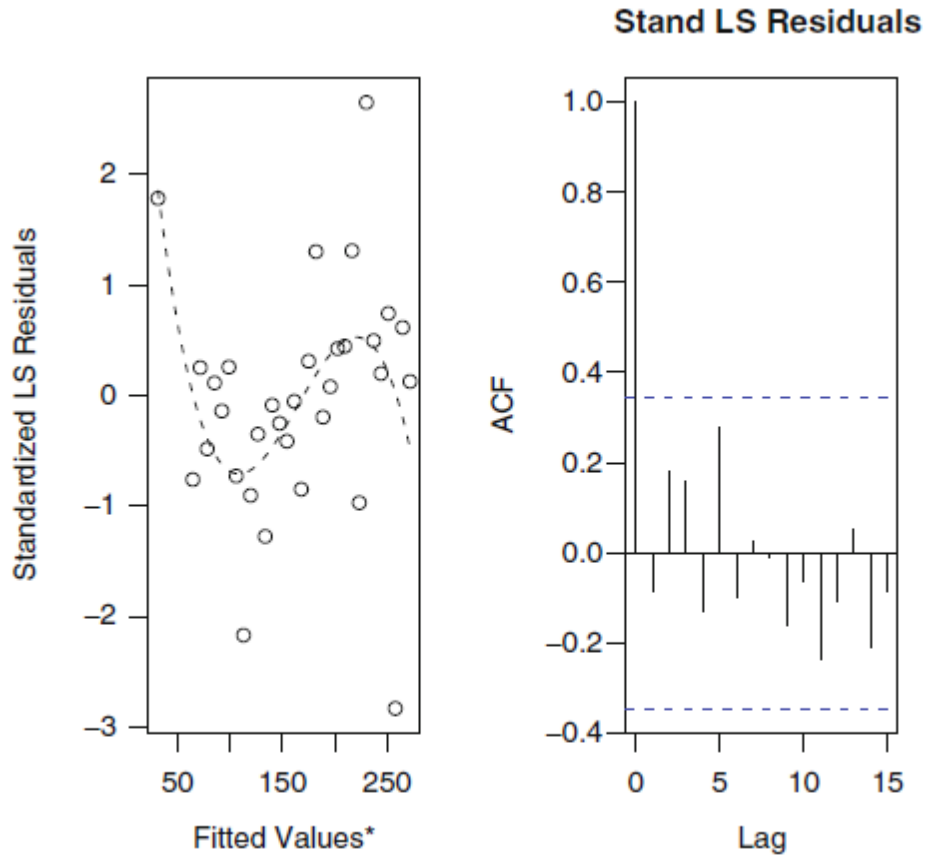
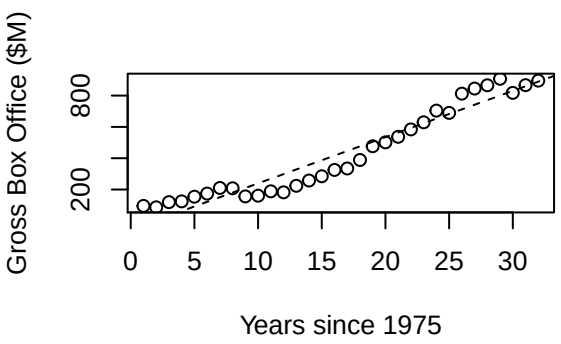


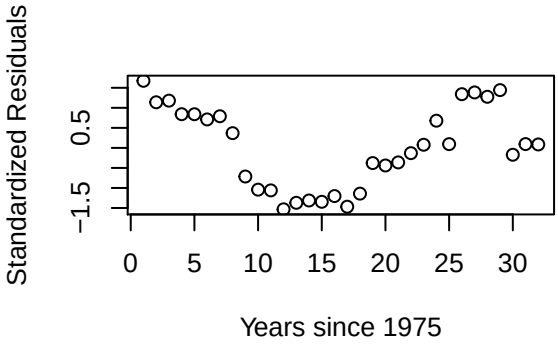
Figure 9.17 Plots of standardized LS residuals from model (9.6)

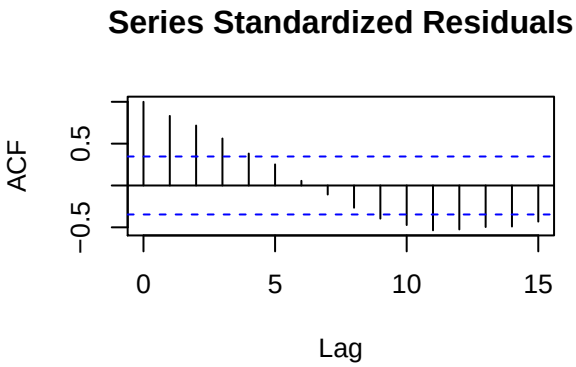
(a) Comment on the analysis performed by the staffer.
(b) Obtain a final model for predicting GrossBoxOffice from YearsS1975. Ensure that you produce diagnostic plots to justify your choice of model. Describe any weaknesses in your model.
(c) Use your model from (b) to predict GrossBoxOffice in 2008.
(d) Use your model from (b) to identify any outliers. In particular, decide whether the year 2000 is an outlier. There is some controversy about the year 2000. In one camp are those that say that fewer people went to the movies in Australia in 2000 due to the Olympics being held in Sydney. In the other camp are those that point to the fact that a 10% Goods and Services Tax (GST) was introduced in July 2000 thus producing an increase in box office receipts.

a) The preliminary data analysis that has been done by the staffer at AFC is good. He has recognized the presence of pattern in the plot of Standardized Residuals and years since 1975. Also the pLot of ACF and lag in Fig 9.16 indicates the presence of autocorelation of different levels (1, 2,3,4 and 9,10,11,12,13). This is very interesting pattern. Thence, he fit a model with errors AR(1), which is also reasonable. Again he fit a model assuming the erros are AR(1) using least squares based on the transformed versions of the response and the predictor variables in 9.6. The results were also consistent, which is natural. Finally, he observed that there is some pattern in Standardized LS Residuals and fittes values in Fig 9.17, but the figure of lags is correct in the same. And the pattern has raised some concern from him, which is natural. I think because he has not considered the Lag(2), lag(3) and other lags seen in the model, the pattern has been generated. We might need to consider the other models to address that.

b) I think I will go with the model with transformed values of the response and the predictor variables. So the final model is GrossBoxOffice=4.514 + 27.075 * YearsS1975.





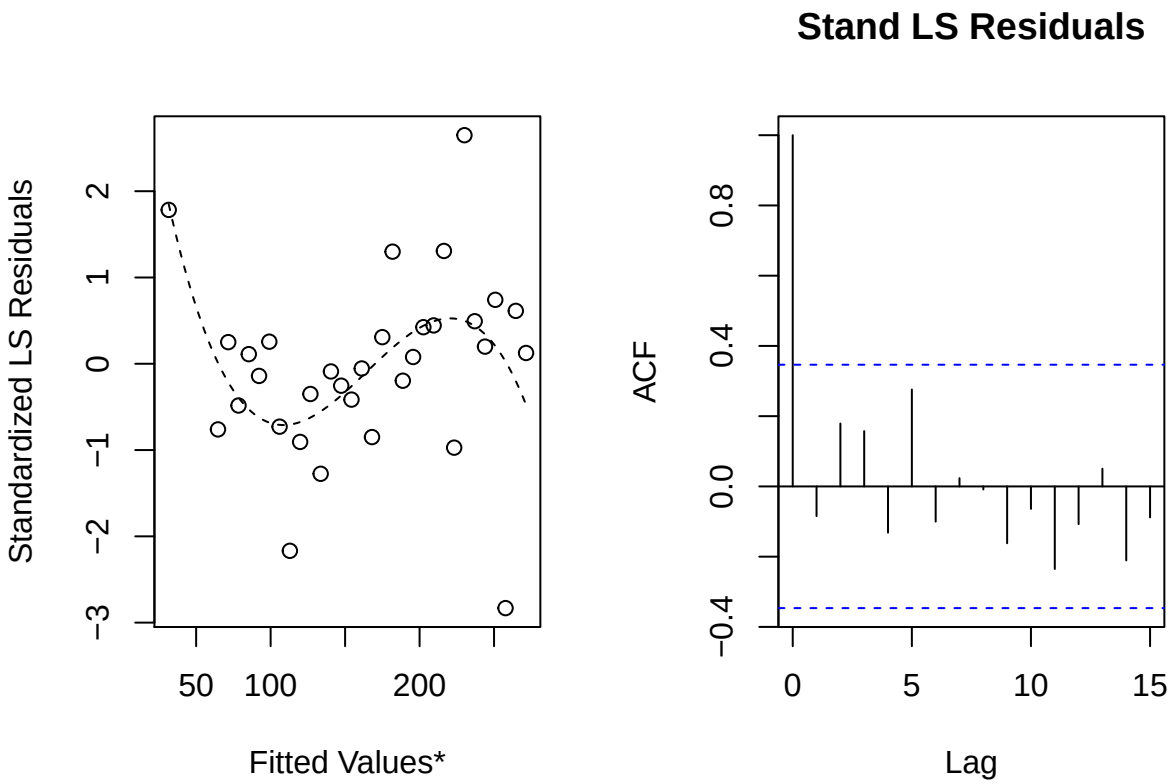



```
#R output on page 327
g <- lm(GrossBoxOffice~YearsS1975,data=boxoffice)
rho <- 0.8782065
x <- model.matrix(g)
Sigma <- diag(length(YearsS1975))
Sigma <- rho^abs(row(Sigma)-col(Sigma))
sm <- chol(Sigma)
smi <- solve(t(sm))
xstar <- smi %*% x
ystar <- smi %*% GrossBoxOffice
m1tls <- lm(ystar ~ xstar-1)
summary(m1tls)
```

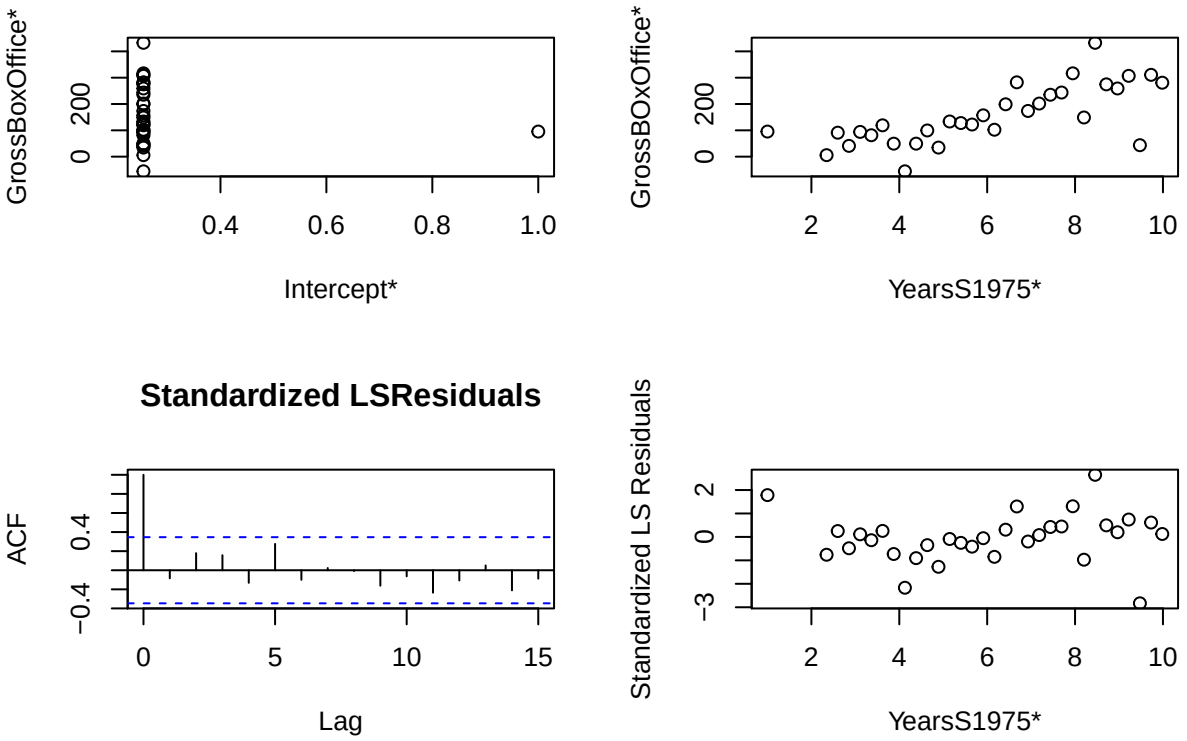
```
##
## Call:
## lm(formula = ystar ~ xstar - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -214.235  -42.370    0.902   33.011  202.415
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## xstar(Intercept)     4.514     72.744   0.062    0.951
## xstarYearsS1975     27.075      3.448   7.853 9.17e-09 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.66 on 30 degrees of freedom
## Multiple R-squared:  0.8427, Adjusted R-squared:  0.8322
## F-statistic: 80.37 on 2 and 30 DF,  p-value: 8.919e-13
```

```
#Figure 9.17 on page 328
StanRes1 <- rstandard(m1tls)
mres2 <- lm(StanRes1~m1tls$fitted.values+I(m1tls$fitted.values^2)+I(m1tls$fitted.values^3))
b1 <- mres2$coeff[1]
b2 <- mres2$coeff[2]
b3 <- mres2$coeff[3]
b4 <- mres2$coeff[4]
mres3 <- lm(StanRes1~m1tls$fitted.values+I(m1tls$fitted.values^2)+I(m1tls$fitted.values^3)+I(m1tls$fitte
par(mfrow=c(1,2))
plot(m1tls$fitted.values,StanRes1,ylab="Standardized LS Residuals",xlab="Fitted Values*")
curve(b1 + b2*x + b3*x^2 + + b4*x^3, add = TRUE,lty=2)
acf(StanRes1,main="Stand LS Residuals")
```




```
detach(boxoffice)
```



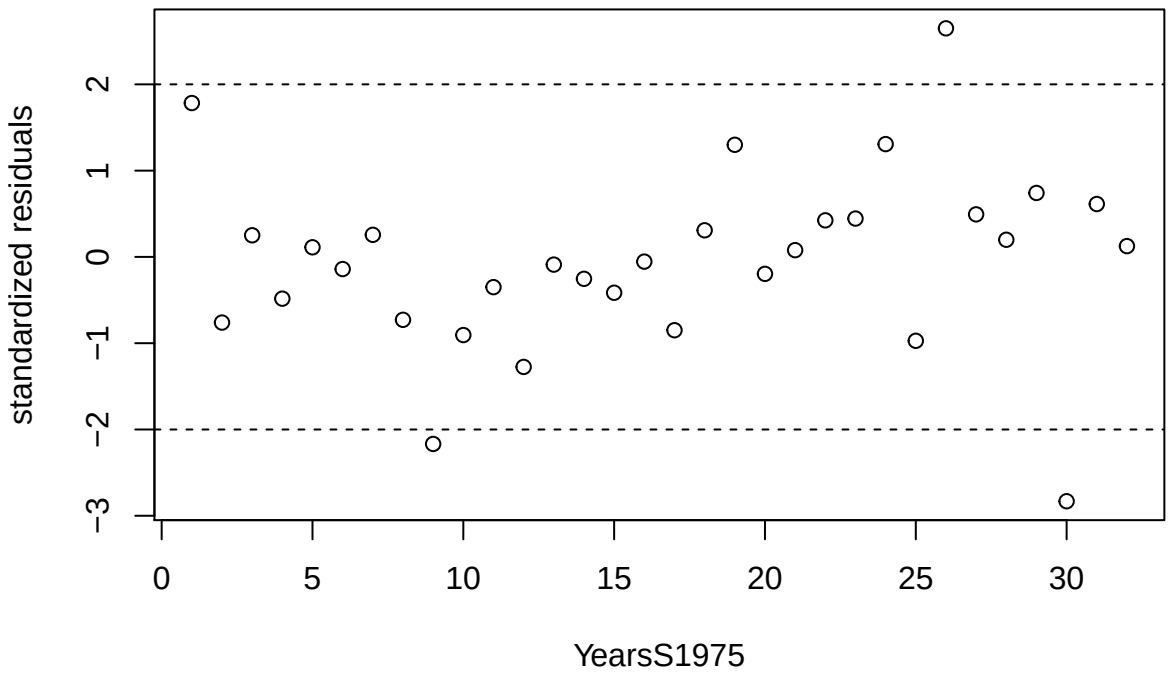
We can see the above plots for the model selected. THe standardized residuals don't have any pattern in the plot. But, the model also has pattern in the plot of fitted Values* and standardized residuals. This definitely is a deficit in our model.

c) The predicted values for year 2008 is 898.0021 millions of dollars.

```
predict(m1,data.frame(YearsS1975=33), interval="prediction")
```

```
## [1] 898.0021
## attr(,"label")
## [1] "Predicted values"
```

d) We can observe the following outliers from the following analysis.



The years 1984,2001 and 2005 are the outliers because their standardized residuals falls outside the range of +2 and -2. Under observation, we can see that year 2001 is an outlier but 2000 is not an outlier. Probably the service tax increment introduced in July raised the ticket sale after its introduction and it lasted until 2001 and slowly became normal. The year 2000 might have had less sale in earlier months but had more sale after July accordingly.

Project milestones [20 points]

1. Interpret the results of your model in a draft results sections with preliminary figures.

- What interesting patterns do you observe? Anything surprising?
- How do the model results relate to your research question and hypothesis?

Done with Dhurba Neupane

References