STAT 757 Assignment 9

DUE 4/29/2018 11:59PM

AG Schissler

2/14/2018

Instructions [20 points]

Modify this file to provide responses to the Ch.9 Exercises in Sheather (2009). 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.

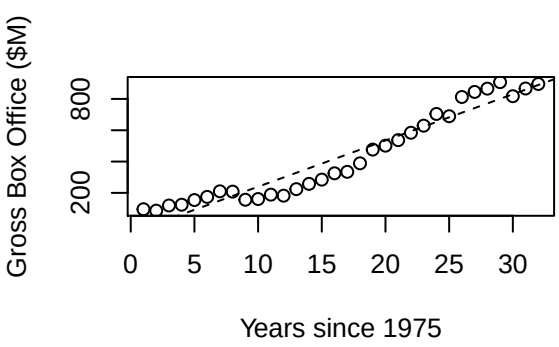
Exercise 9.4.1 [60 points]

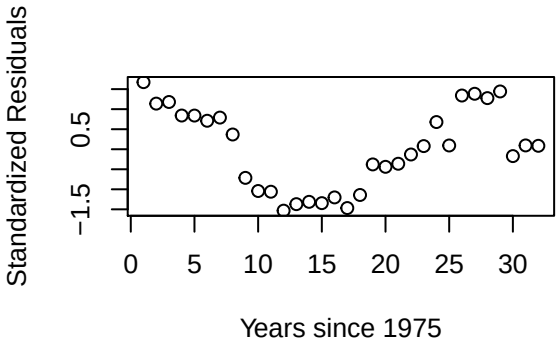
I will reproduce the plots and model summaries here and comment on each in turn.

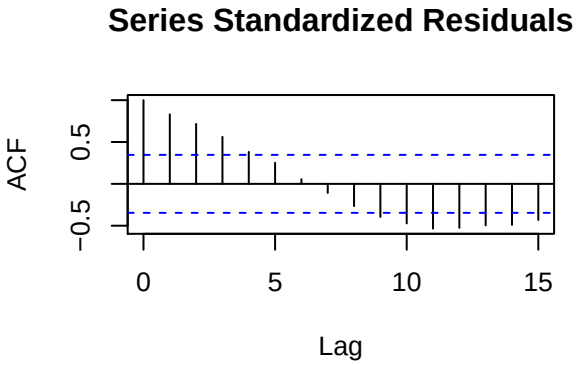
Part A

```
myDir <- "~/OneDrive - University of Nevada, Reno/Teaching/STAT_757/Sheather_data/Data/"
boxoffice <- read.table(file.path(myDir,"boxoffice.txt"), header = T)
str(boxoffice)
attach(boxoffice)
library(nlme)
```

```
YearsS1975 <- year - 1975
lsm1 <- lm(GrossBoxOffice~YearsS1975,data=boxoffice)
StanRes1 <- rstandard(lsm1)
par(mfrow=c(2,2))
plot(YearsS1975,GrossBoxOffice,ylab="Gross Box Office ($M)",xlab="Years since 1975")
abline(lsm1,lty=2)
plot(YearsS1975,StanRes1,ylab="Standardized Residuals",xlab="Years since 1975")
acf(StanRes1,main="Series Standardized Residuals")
```







The staffer corrected identified autocorrelation in the data, resulting in a poor fit for the ordinary least squares model.

```
m1 <- gls(GrossBoxOffice~YearsS1975,correlation=corAR1(form=~YearsS1975),data=boxoffice,method="ML")
summary(m1)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: GrossBoxOffice ~ YearsS1975
##   Data: boxoffice
##       AIC    BIC  logLik
##   330.39 336.25 -161.19
##
## Correlation Structure: AR(1)
##  Formula: ~YearsS1975
##  Parameter estimate(s):
##     Phi
## 0.87821
##
## Coefficients:
##                Value Std.Error t-value p-value
## (Intercept)  4.5141    72.744  0.0621  0.9509
## YearsS1975  27.0754     3.448  7.8533  0.0000
##
##  Correlation:
##             (Intr)
## YearsS1975 -0.782
##
## Standardized residuals:
##       Min        Q1       Med        Q3       Max
## -1.934208 -1.385920  0.018223  0.332026  1.542698
```

```
##
## Residual standard error: 76.165
## Degrees of freedom: 32 total; 30 residual
```

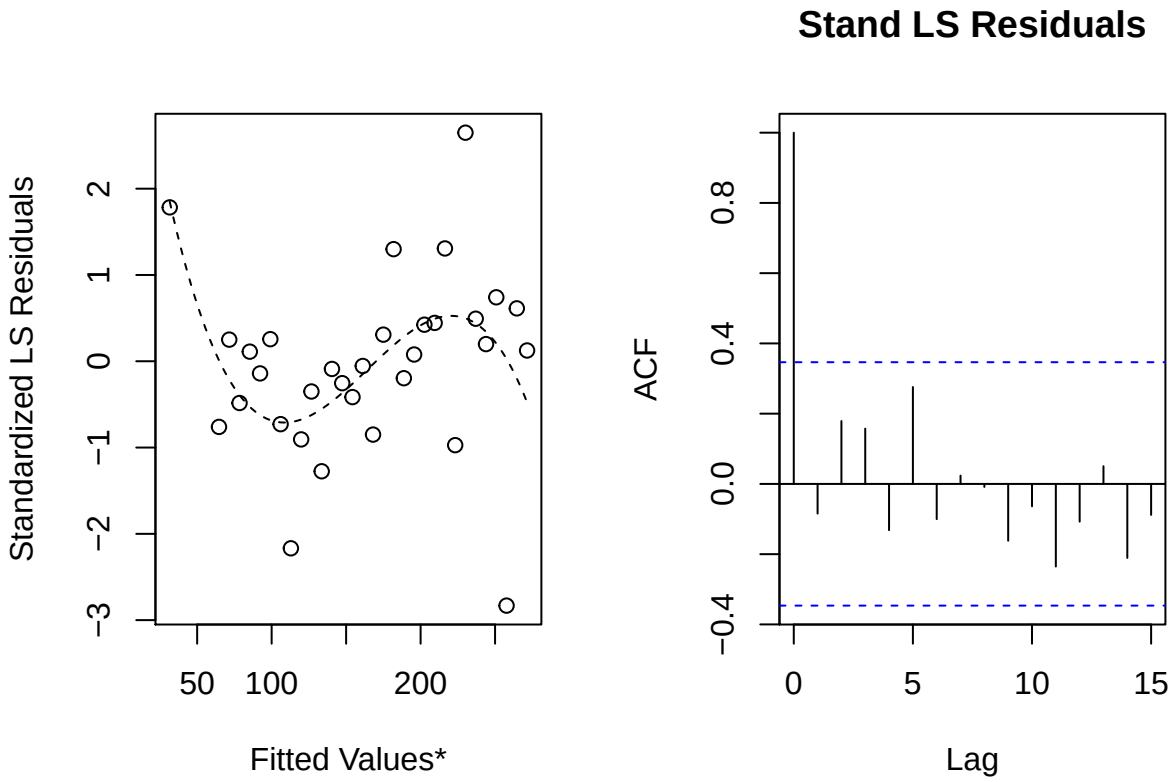
Nice job fitting an autocorrelation model with a lag 1 structure.

```
#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
## -214.2  -42.4    0.9   33.0  202.4
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## xstar(Intercept)     4.51      72.74    0.06     0.95
## xstarYearsS1975     27.08       3.45    7.85  9.2e-09
##
## Residual standard error: 78.7 on 30 degrees of freedom
## Multiple R-squared:  0.843,  Adjusted R-squared:  0.832
## F-statistic: 80.4 on 2 and 30 DF,  p-value: 8.92e-13
```

The transformation is correctly executed and the verification strategy is appropriate.

```
#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")
```



The analyst was doing a fine job up to this point. However, as pointed out on page 318 in Sheather (2009), the first point in an AR(1) model is usually a high leverage point. The polynomial model was overfit to the first point, incorrectly showing a siginficant model. By removing the first point (only for demonstration/diagnostic purposes; i.e., the model includes the first point),

```
mres4 <- lm(StanRes1[-1] ~ m1tls$fitted.values[-1]+
               I(m1tls$fitted.values[-1]^2)+
               I(m1tls$fitted.values[-1]^3)+
               I(m1tls$fitted.values[-1]^4)+
               I(m1tls$fitted.values[-1]^5))
summary(mres4)
```

```
##
## Call:
## lm(formula = StanRes1[-1] ~ m1tls$fitted.values[-1] + I(m1tls$fitted.values[-1]^2) +
##     I(m1tls$fitted.values[-1]^3) + I(m1tls$fitted.values[-1]^4) +
##     I(m1tls$fitted.values[-1]^5))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.615 -0.345  0.112  0.480  2.167
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -3.37e+01   2.81e+01   -1.20     0.24
## m1tls$fitted.values[-1]        1.30e+00   1.02e+00    1.28     0.21
## I(m1tls$fitted.values[-1]^2) -1.88e-02   1.39e-02   -1.35     0.19
## I(m1tls$fitted.values[-1]^3)  1.26e-04   8.99e-05    1.40     0.17
## I(m1tls$fitted.values[-1]^4) -3.92e-07   2.78e-07   -1.41     0.17
## I(m1tls$fitted.values[-1]^5)  4.61e-10   3.30e-10    1.39     0.18
```

```
##
## Residual standard error: 0.954 on 25 degrees of freedom
## Multiple R-squared:  0.256,  Adjusted R-squared:  0.107
## F-statistic: 1.72 on 5 and 25 DF,  p-value: 0.168
```

We see that the "distinct pattern" is no longer distinct even this highly responsive model is not significant by the ANOVA results.

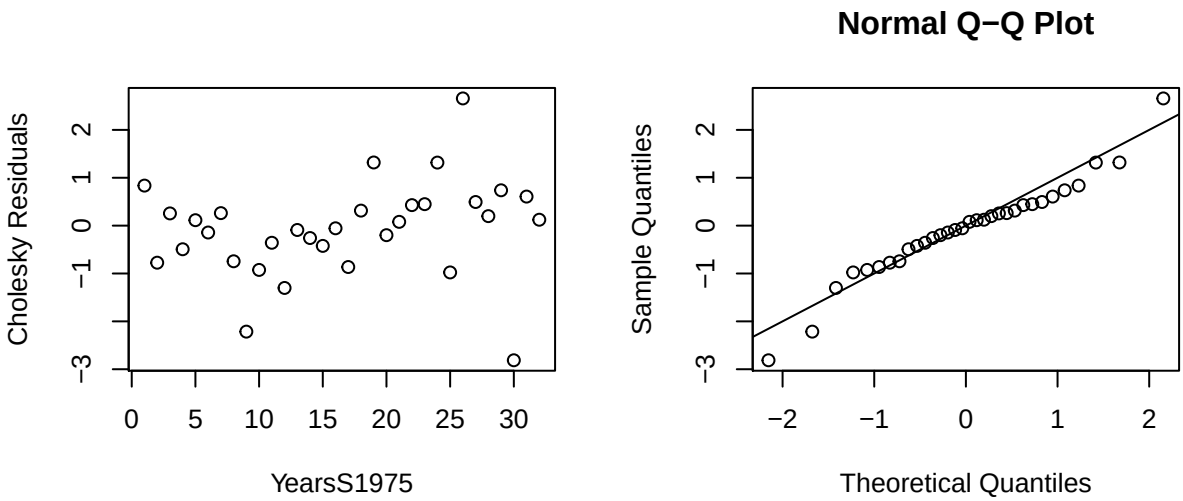
Part B

Note that the autocorrelation was removed by transformation above. Below are the Cholesky (normalized) residuals to help determine if the model is appropriately fit:

```
(chol_resid <- residuals(m1, type = "normalized"))
```

```
##         1         2         3         4         5         6         7
##  0.836481 -0.774559  0.255366 -0.493459  0.113524 -0.143643  0.261274
##         8         9        10        11        12        13        14
## -0.744800 -2.214570 -0.925211 -0.357997 -1.301972 -0.091080 -0.259049
##        15        16        17        18        19        20        21
## -0.423234 -0.055640 -0.864823  0.313796  1.319277 -0.199221  0.079332
##        22        23        24        25        26        27        28
##  0.428343  0.448626  1.316957 -0.977618  2.657584  0.493629  0.198488
##        29        30        31        32
##  0.738180 -2.812779  0.607022  0.123400
## attr(,"label")
## [1] "Normalized residuals"
```

```
par(mfrow=c(2,2))
plot(chol_resid, xlab="YearsS1975", ylab="Cholesky Residuals")
qqnorm(chol_resid);abline(0,1)
```



I retain the first model (m1) as my final model as the residuals appear adequate except for some outliers. Indeed, years 9, 26, and 30 (which correspond to 1985, 2001, 2005) are outlying based on large standardized residuals. Perhaps with some transformation on the sales variable, the fit could be improved a little, but I would be satisified with m1.

Part C

It is easier to interpret the prediction when using the non-transformed generalized least squares AR(1) model (m1).

```
as.numeric(predict(m1, newdata = data.frame( YearsS1975 = 33)))
```

```
## [1] 898
```

The predicted box office sales in 2008 is $898 million.

Part D

No, as discussed above, the year 2000 only has a standardized residual approximately equal to -1. I suppose the effects of the olympics and tax law "cancelled each other out".

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?

References

Sheather, Simon. 2009. *A Modern Approach to Regression with R*. Springer Science & Business Media.