Name: Raj Shah

PSID: 1499521

**Assignment 5**

Prob 1)

I have looked at Kidney dataset, containing infection to patient(delta) as a binary response variable, alongside with their time and type of catheter place which act as the predictors. Type 1 Surgically, Type 2 Percutaneously

Code:

*library(readr)*

*dataset<-read.csv("C:/Users/Raj Shah/Desktop/Studies UH/MATH 6359 Stastical Computing/Homework/Answers/HW5/dataset.csv")*

*attach(dataset)*

*dataset$sex<-as.factor(dataset$sex) #Converting Sex as Factor,1=Male 2=Female*

*dataset$type<-as.factor(dataset$type) #Converting Type of Catheter Placement as factor*

*k2.glm<-glm(delta~type+time+sex,family = binomial,data=dataset)*

*summary(k2.glm)*

Output:

Call:

glm(formula = delta ~ type + time + sex, family = binomial, data = dataset)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5423 -0.6518 -0.4774 0.8010 2.4288

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.13331 0.67851 -0.196 0.84424

type2 -1.24897 0.58233 -2.145 0.03197 \*

time -0.09764 0.03771 -2.589 0.00962 \*\*

sex2 1.30135 0.53889 2.415 0.01574 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.51 on 118 degrees of freedom

Residual deviance: 110.17 on 115 degrees of freedom

AIC: 118.17

Number of Fisher Scoring iterations: 5

Analysis:

1. Response: Let yi denote whether the ith patient is infected (yi = 1) or not infected (yi = 0).

2. Predictors: Let (w1,i ; x1,i ; z1,i) denote the observed values of dummy variables w1,i Type of catheter placement, age x1,i and sex z1,i for the ith patient.

3. Probability: Let Pi = P(yi = 1) - probability of infection for a patient with predictor values (w1,i ; x1,i ; z1,i).

4. Logistic regression formula:

logit(Pi) = B0 + B1w1;i + B2x1;i + B3z1;i ; i = 1,….., n

5. Logistic regression revealed all the predictors – Type of Catheter placement, time and sex to play a role in patients infection. Time was more significant as compared to type and gender

6. The smallest p-values correspond to B2(time) followed by B3(gender) and B1 (Type2 of Catheter Placement).

B1= -1.245 🡺 log-odds of infection are 1.245 less for Type 2 as compared to reference group Type 1 of catheter placement

B3= 1.30 🡺 log-odds of infection are more by 1.30 for females compared to reference group males

Odds Ratio

Code:

> exp(coef(k2.glm))

(Intercept) type2 time sex2

0.8751977 0.2867997 0.9069750 3.6742460

Odds

> k2.glm<-glm(delta~type+time+sex-1,family = binomial,data=dataset)

> exp(cbind(Odds=coef(k2.glm),confint(k2.glm)))

Odds 2.5 % 97.5 %

type1 0.8751977 0.2304551 3.3688053

type2 0.2510064 0.1053194 0.5469849

time 0.9069750 0.8358631 0.9710693

sex2 3.6742460 1.2912218 10.8887692

> k2.glm<-glm(delta~sex+time+type-1,family = binomial,data=dataset)

> exp(cbind(Odds=coef(k2.glm),confint(k2.glm)))

Odds 2.5 % 97.5 %

sex1 0.8751977 0.23045508 3.3688053

sex2 3.2156916 1.17458314 9.7531223

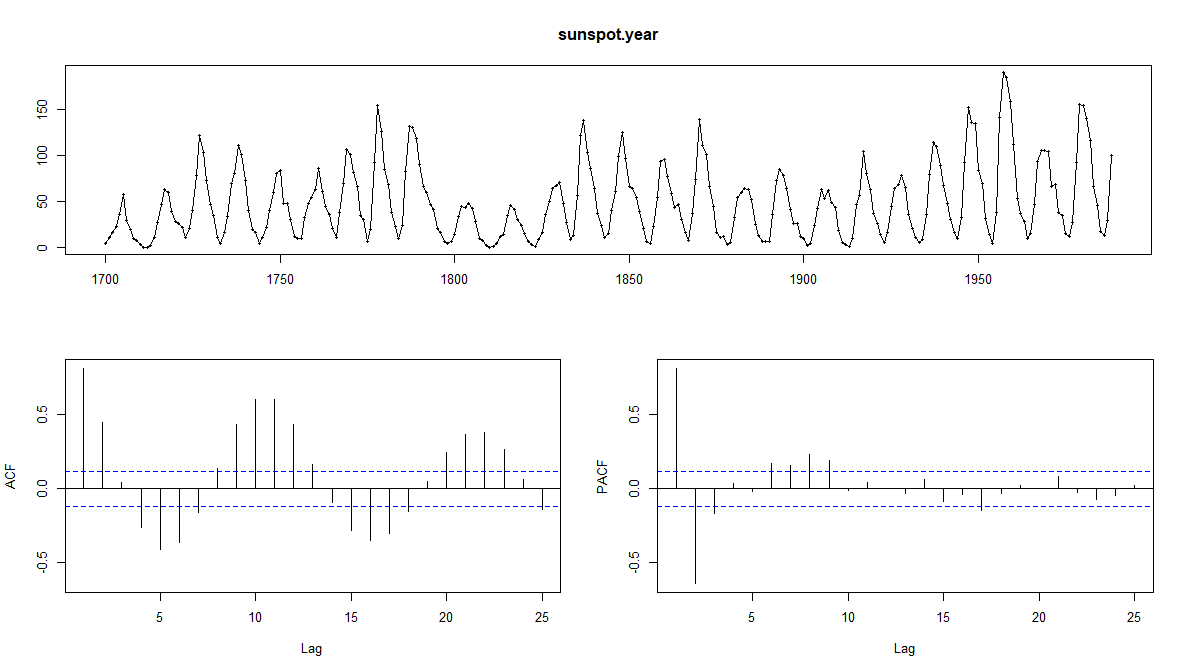
time 0.9069750 0.83586309 0.9710693

type2 0.2867997 0.08870974 0.8859110

An **odds ratio** is a relative measure of effect, which allows the comparison of the intervention group of a study relative to the comparison group. So here we see the odds ratio for Type i.e. odds for type 2/odds for type 1 (0.2510064/0.8751977) is 0.2868 which means the type of infection is not significantly different but odds ratio for sex the odds ratio i.e. odds of Female/Odds of Male (3.2156916/0.8751977) is 3.6742 which implies the infection for Females is more as compared to males.

Prob 2)

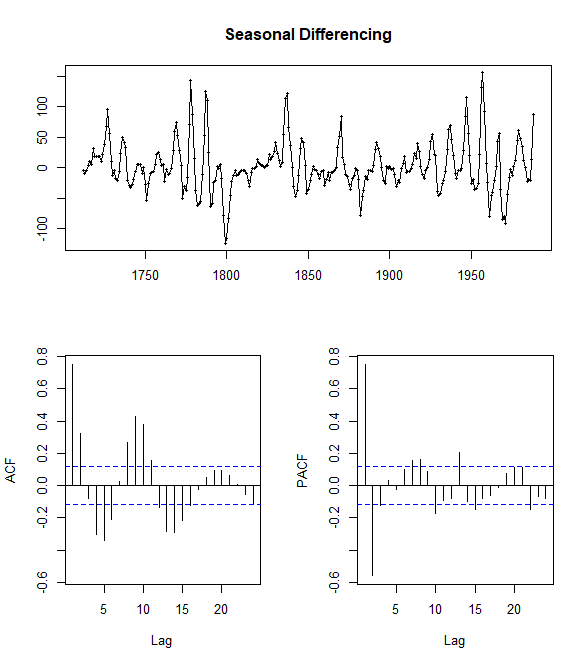
tsdisplay(sunspot.year)



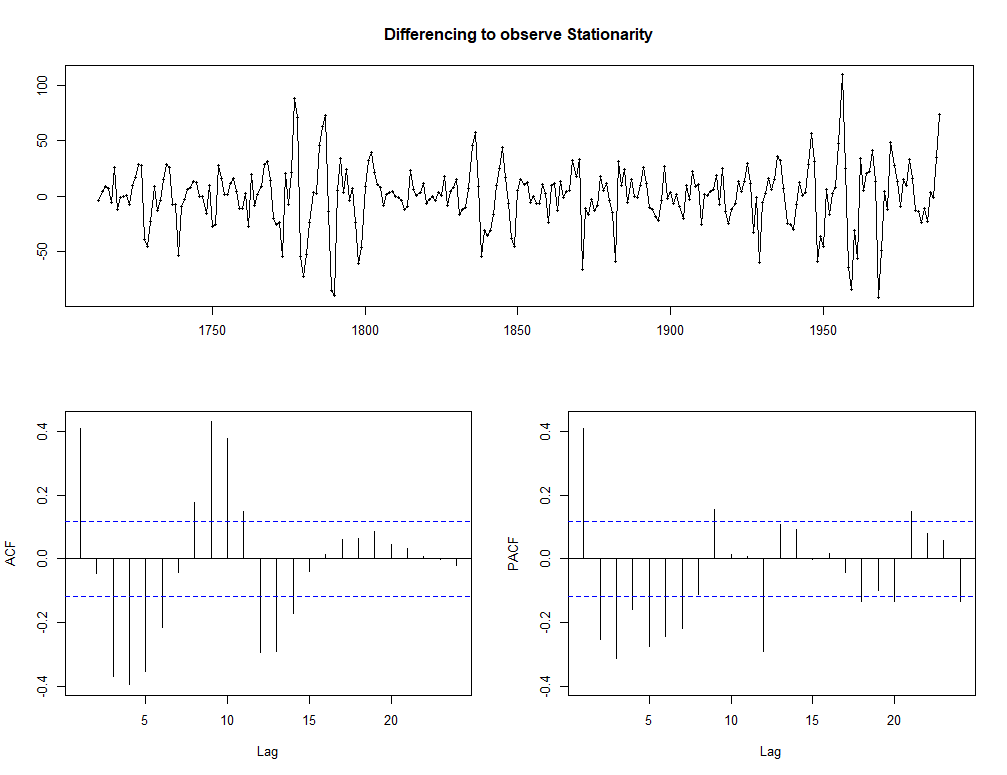
We can observe seasonality from the first graph which can be verified from the ACF plot, where we see sinusoidal plots and peak at monthly lags m=12,24….

Applying Seasonal Differencing

tsdisplay(diff(sunspot.year,12),main="Seasonal Differencing")



tsdisplay(diff(tsadjs),main="Differencing to observe Stationarity")



arima.param <- list(c(3,1,0),c(4,1,0),c(7,1,0),c(6,1,0),c(5,1,0))

arima.aic <- sapply(arima.param,function(x) Arima(diff(tsadjs),order=x)$aic)

print(arima.param[[which.min(arima.aic)]])

[1] 7 1 0

On observing sinusoidal behavior of ACF plot we can conclude it’s of the form (p,d,0)

my.arima.fit <-Arima(diff(tsadjs),c(7,1,0))

summary(my.arima.fit)

Series: diff(tsadjs)

ARIMA(7,1,0)

Coefficients:

ar1 ar2 ar3 ar4 ar5 ar6 ar7

-0.3762 -0.3965 -0.5360 -0.4123 -0.3840 -0.3384 -0.3025

s.e. 0.0579 0.0590 0.0591 0.0628 0.0589 0.0586 0.0577

sigma^2 estimated as 650.2: log likelihood=-1278.17

AIC=2572.34 AICc=2572.88 BIC=2601.28

Training set error measures:

ME RMSE MAE MPE MAPE MASE ACF1

Training set 0.2251112 25.12763 18.35868 NaN Inf 0.8592694 -0.128098

auto.arima(diff(tsadjs),seasonal=F)

Series: diff(tsadjs)

ARIMA(0,0,2) with zero mean

Coefficients:

ma1 ma2

0.5472 0.2491

s.e. 0.0610 0.0625

sigma^2 estimated as 613.4: log likelihood=-1276.61

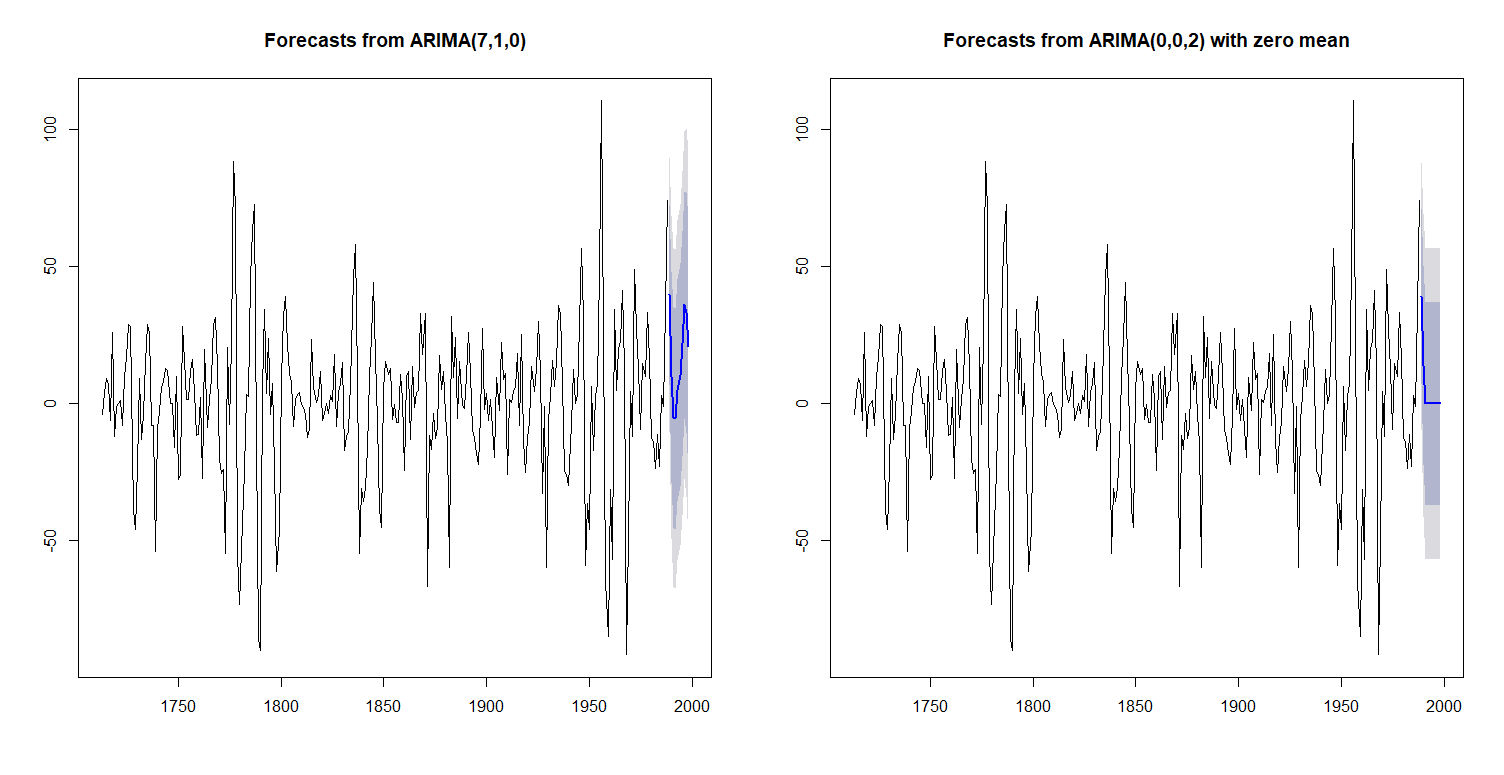
AIC=2559.22 AICc=2559.31 BIC=2570.08

par(mfrow=c(1,2))

plot(forecast(my.arima.fit))

plot(forecast(auto.arima(diff(tsadjs),seasonal=F)))

par(mfrow=c(1,1))



Thus we observe the differences in the forecast from the auto.arima model and user model. Both the models decrease initially and then increase.