SS3859A Assignment 4

Bradley Assaly-Nesrallah

11/11/2020

Question 1

##a. We obtain the probability of Y = 1 at x1=1 and x2=0.5 as follows: ## P(Y-1|x1=1,x2=0.5)=1/(1+e^-(-2.7399+3.0287*1-1.2081*0.5))=0.42183379341

##b. We test H0: B2=0 vs H1: B2=/=0 at alpha=0.05 as follows: ## We do z test z=*B2-0/SE(B2)=-1.208/0.4620=-2.61471 ## p-val = 2*P(z>|z|) = 2P(z>2.61)=2*0.005=0.01<0.05=alpha ## thus reject H0 at the given confidence interval so B2=/=0 so done

##c. We test H0: B1=B2=0: vs H1:H0 false at alpha=0.05 as follows: ## We use test stat D=Dr-Df=110.216-56.436=53.78 ## p-val = P(chisqr>D)<0.05 so reject H0: one of the predictors is statistically sig so done

Question 2

##a.We obtain Y_hat values at cutoff 0.5 which are 1,0,1,0,0,1,0,1,1,0: ## Now we make the confustion matrix as follows ## Y ## 0 1 ## Y_hat 0 #TN=3 #FN=2 ## 1 #FP=3 #TP=2 ## accuracy=(TP+TN)/all=5/10=0.5 ## sensitivity=TP/(TP+FN)=2/4=0.5 ## precision=TP/(TP+FP)=2/5=0.4

##b.We obtain Y_hat values at cutoff 0.8 which are 0,0,1,0,0,0,0,1,0,0: ## Now we make the confustion matrix as follows ## Y ## 0 1 ## Y_hat 0 #TN=5 #FN=3 ## 1 #FP=1 #TP=1 ## accuracy=(TP+TN)/all=6/10=0.6 ## sensitivity=TP/(TP+FN)=1/4=0.25 ## precision=TP/(TP+FP)=1/2=0.5

##c.If we want to increase the sensitivit of prediction you want decrease ##the cutoff value. This will increase the TP values and decrease the false negative ##value as well so sensitivity=TP/(TP+FN) will increase

Question 3

```
##a. We obtain y_hat values and make confusion matrix and report acc,sens,spec,prec:
library(bestglm)
```

```
## Loading required package: leaps
```

```
fit_full=glm(chd~.,data=SAheart,family=binomial)
fit_full
```

```
##
## Call:  glm(formula = chd ~ ., family = binomial, data = SAheart)
```

```
##
## Coefficients:
##    (Intercept)             sbp         tobacco             ldl       adiposity
##     -6.1507209       0.0065040       0.0793764       0.1739239       0.0185866
## famhistPresent           typea         obesity         alcohol             age
##      0.9253704       0.0395950      -0.0629099       0.0001217       0.0452253
##
## Degrees of Freedom: 461 Total (i.e. Null);  452 Residual
## Null Deviance:       596.1
## Residual Deviance: 472.1     AIC: 492.1
```

```r
n = nrow(SAheart)
cutoff = 0.5
y_hat = rep(0,n)
idx = which(fitted(fit_full)>cutoff)
y_hat[idx] = 1
conf_mat = table(predicted = y_hat, actual = SAheart$chd)
conf_mat
```

```
##          actual
## predicted   0   1
##         0 256  77
##         1  46  83
```

```r
mean(y_hat == SAheart$chd) # Accuracy
```

```
## [1] 0.7337662
```

```r
conf_mat[2, 2] / sum(conf_mat[, 2]) # Sensitivity
```

```
## [1] 0.51875
```

```r
conf_mat[1, 1] / sum(conf_mat[, 1]) # Specificity
```

```
## [1] 0.8476821
```

```r
conf_mat[2, 2] / sum(conf_mat[2, ]) # Precision
```

```
## [1] 0.6434109
```

```r
##b. We use backward selection with BIC to find best subset of predictors chd
fit_back_bic = step(fit_full, direction = "backward", k=log(n),trace=0)
fit_back_bic
```

```
##
## Call:  glm(formula = chd ~ tobacco + ldl + famhist + typea + age, family = binomial,
##     data = SAheart)
##
## Coefficients:
```

```
##    (Intercept)         tobacco            ldl  famhistPresent           typea
##       -6.44644         0.08038        0.16199         0.90818         0.03712
##            age
##        0.05046
##
## Degrees of Freedom: 461 Total (i.e. Null);  456 Residual
## Null Deviance:       596.1
## Residual Deviance: 475.7     AIC: 487.7
```

```r
## thus the final model is chd ~ tobacco + ldl + famhist + typea + age

##c. We want to see if predictors the predictors in b are significant with the lr test
## Full model= chd~ sbp + tobacco + ldl + adiposity+ famhist + typea +obesity+ alcohol + age
## Reduced model = chd ~ tobacco + ldl + famhist + typea + age
## we test H0: Btobacco=Bldl=Bfamhist=Btypea=Bage=0 vs H1: H0 is false as follows:

##d.we obtain the test staitstic and make a conclusion at alpha=0.05 as follows
D_stat = deviance(fit_back_bic) - deviance(fit_full)
D_stat
```

```
## [1] 3.545546
```

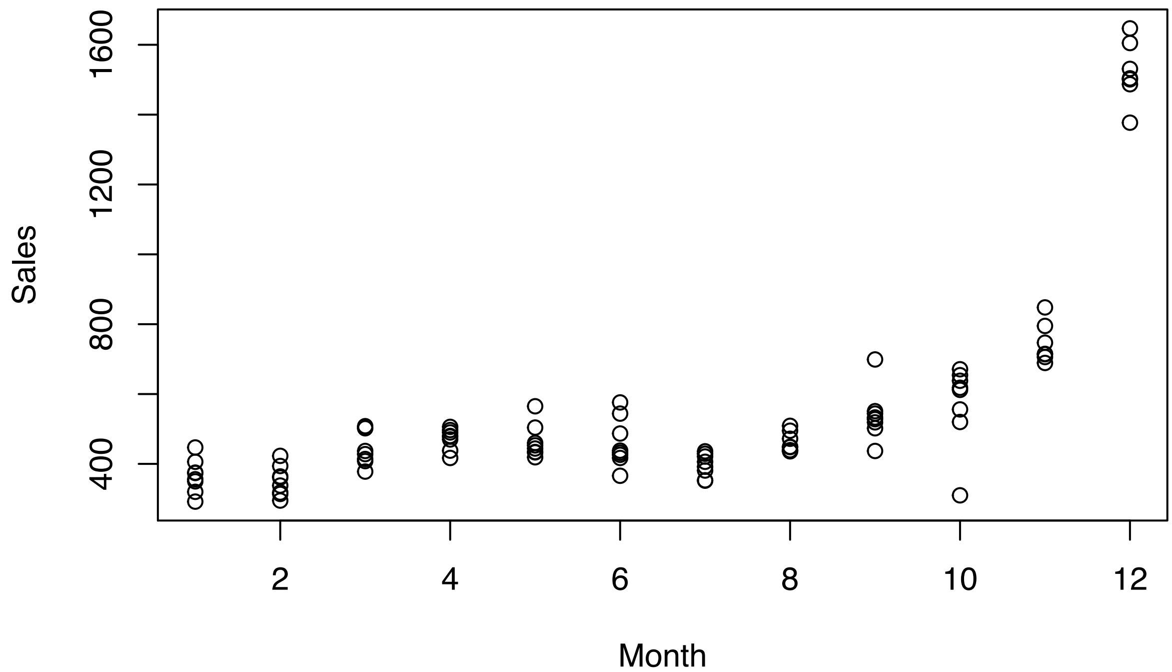
```r
k=9-5
pval=1-pchisq(D_stat,k)
pval
```

```
## [1] 0.4709869
```

```r
## p val is greater that 0.05 so we do not reject H0 so we are done
## thus the reduced model is not stastically different from the full
```

##Question 4

```r
data_1<-read.csv("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw4-data1.csv")
##a. We want to check scatterplot sales~month and compare two models A-year and month both numerical,
## B year num moth cat and fit and compare in terms of adjusted R^2
plot(Sales~Month,data=data_1)
```




```r
##it appears that sales tend to increase as the month increases
##make model A
lm_a=lm(Sales~Month+Year,data=data_1)
##make model B
data_1$Month <-as.factor(data_1$Month)
lm_b=lm(Sales~Month+Year,data=data_1)
summary(lm_a)$adj.r.squared
```

```
## [1] 0.4321569
```

```r
summary(lm_b)$adj.r.squared
```

```
## [1] 0.9581081
```

```r
## the r^2 of model B is significantly higher than model A so done

##b.Using model B we describe the yearly trend and seasonal patern, we predict sales
lm_b
```

```
##
## Call:
## lm(formula = Sales ~ Month + Year, data = data_1)
##
## Coefficients:
```

```
## (Intercept)       Month2       Month3       Month4       Month5       Month6
##  -10368.909      -14.125       82.250      107.000       99.000       95.750
##      Month7       Month8       Month9      Month10      Month11      Month12
##      31.250       95.875      174.125      207.375      382.549     1159.407
##        Year
##       5.384
```

```
##we can see from lm_b that sales increase per year by the beta for year
##we can see sales are low in jan-july, so winter to end of summer,
##aug-dec increasing, so higher fall-start of winter
m1=predict(lm_b,newdata=data.frame(Month='1',Year=1998))
m2=predict(lm_b,newdata=data.frame(Month='2',Year=1998))
m3=predict(lm_b,newdata=data.frame(Month='3',Year=1998))
m4=predict(lm_b,newdata=data.frame(Month='4',Year=1998))
m5=predict(lm_b,newdata=data.frame(Month='5',Year=1998))
m6=predict(lm_b,newdata=data.frame(Month='6',Year=1998))
m7=predict(lm_b,newdata=data.frame(Month='7',Year=1998))
m8=predict(lm_b,newdata=data.frame(Month='8',Year=1998))
m9=predict(lm_b,newdata=data.frame(Month='9',Year=1998))
m10=predict(lm_b,newdata=data.frame(Month='10',Year=1998))
m11=predict(lm_b,newdata=data.frame(Month='11',Year=1998))
m12=predict(lm_b,newdata=data.frame(Month='12',Year=1998))
##predictions for sales next 12 months
m1
```

```
##      1
## 389.23
```

```
m2
```

```
##       1
## 375.105
```

```
m3
```

```
##      1
## 471.48
```

```
m4
```

```
##      1
## 496.23
```

```
m5
```

```
##      1
## 488.23
```

```
m6
```

```
##      1
## 484.98
```

```r
m7
```

```
##      1
## 420.48
```

```r
m8
```

```
##       1
## 485.105
```

```r
m9
```

```
##       1
## 563.355
```

```r
m10
```

```
##       1
## 596.605
```

```r
m11
```

```
##        1
## 771.7794
```

```r
m12
```

```
##        1
## 1548.637
```

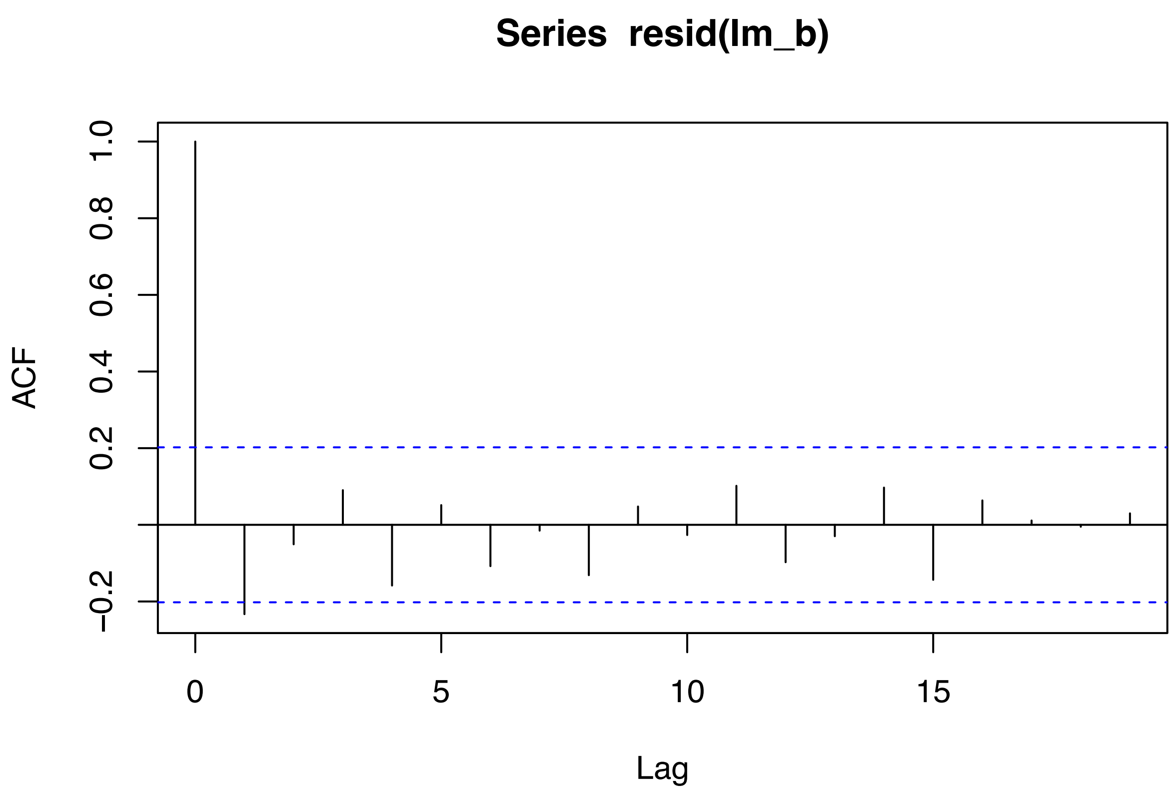
```r
## for the next 12 months and discuss the model assumptions
## model assumptions are adjacent residuals are correlated
## and error independence assumption

##c. We check the model assuptions, and check if adjacent residuals are correlated
## Durbin-Watson test:
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
acf(resid(lm_b))
```

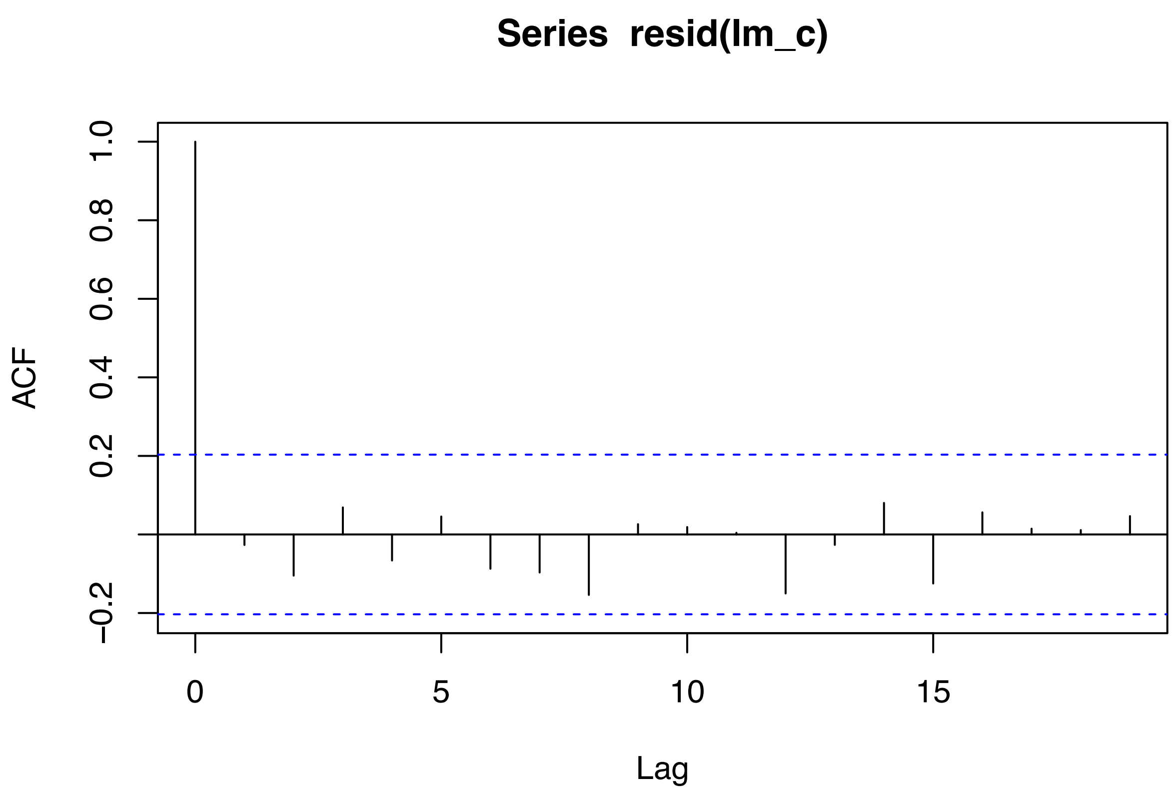



```r
## error independence assumption
dwtest(lm_b,alternative="two.sided")
```

```
##
##  Durbin-Watson test
##
## data:  lm_b
## DW = 2.4509, p-value = 0.03902
## alternative hypothesis: true autocorrelation is not 0
```

```r
## p-value is 0.03<0.05 thus reject H0, so autocorrelation is nonzero so done

##d. We estimate lag 1 autocorrelation rho, and fit another model
##we then use acf to check error independence assumption is met and compare model b,c AIC as follows:
num_obs=nrow(data_1)
rho_hat_dw = (1-dwtest(lm_b)$statistic/2)
sales_t = data_1$Sales[-1]
sales_t_1 = data_1$Sales[-num_obs]
sales_new = sales_t - rho_hat_dw*sales_t_1 # transformed sales

year_t = data_1$Year[-1]
year_t_1 = data_1$Year[-num_obs]
```

```r
year_new = year_t - rho_hat_dw*year_t_1 #transformed year
data_1<-data_1[-c(93),]
data_1$Sales<-sales_new
data_1$Year<-year_new
##transformed data
##create lm c with transformed data
lm_c= lm(Sales~Year+Month, data=data_1)
acf(resid(lm_c))
```




```r
##thus error independence assumption appears to hold in model c
AIC(lm_b)
```

```
## [1] 1054.002
```

```r
AIC(lm_c)
```

```
## [1] 1040.732
```

```r
##thus model c performs better than model b in terms of AIC so we are done
```