Salinity

Rashnil

21 November 2016

Load Salinity Data

sal <- read.table(file.choose(),header = FALSE)

```
sal <- read.table(file.choose(),header = FALSE);
names(sal) <-c("obs","sal","lagsal","fflow","period","year");
summary(sal)
```

```
##       obs             sal             lagsal           fflow      
##  Min.   : 1.00   Min.   : 4.300   Min.   : 4.300   Min.   :20.77  
##  1st Qu.: 7.75   1st Qu.: 8.075   1st Qu.: 8.075   1st Qu.:21.84  
##  Median :14.50   Median :11.150   Median :10.650   Median :22.97  
##  Mean   :14.50   Mean   :10.554   Mean   :10.332   Mean   :23.73  
##  3rd Qu.:21.25   3rd Qu.:13.025   3rd Qu.:13.025   3rd Qu.:24.86  
##  Max.   :28.00   Max.   :15.100   Max.   :15.000   Max.   :33.44  
##      period         year     
##  Min.   :0.0   Min.   :1972  
##  1st Qu.:1.0   1st Qu.:1973  
##  Median :2.5   Median :1974  
##  Mean   :2.5   Mean   :1975  
##  3rd Qu.:4.0   3rd Qu.:1976  
##  Max.   :5.0   Max.   :1977
```

```
n<-length(sal$obs)
```

Build the initial model with all the variables

```
model_full<-lm(sal~lagsal+fflow+period+year,data = sal);
summary(model_full);
```

```
##
## Call:
## lm(formula = sal ~ lagsal + fflow + period + year, data = sal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.75229 -0.51900  0.05684  0.51389  2.85992
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -831.88289  448.54450  -1.855   0.0765 .
## lagsal         0.61639    0.11860   5.197 2.86e-05 ***
## fflow         -0.26339    0.10299  -2.557   0.0176 *
## period         0.06002    0.15987   0.375   0.7108
## year           0.42648    0.22733   1.876   0.0734 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.266 on 23 degrees of freedom
## Multiple R-squared:  0.8494, Adjusted R-squared:  0.8232
## F-statistic: 32.44 on 4 and 23 DF,  p-value: 3.768e-09
```

Build full model with all the variables and the initial model with no variable

```
model_0 <- lm(sal~1,data = sal);
model_full<-lm(sal~lagsal+fflow+period+year,data = sal)
```

Run the stepwise forward selection process to identify the best model

```
sal.step<- step(model_0,scope = list(lower=model_0,upper=model_full),direction = "forward");
```

```
## Start:  AIC=62.69
## sal ~ 1
##
##          Df Sum of Sq     RSS    AIC
## + lagsal  1   185.815  58.835 24.791
## + year    1   136.824 107.825 41.753
## + fflow   1    55.687 188.962 57.462
## <none>                244.650 62.693
## + period  1     3.621 241.028 64.276
##
## Step:  AIC=24.79
## sal ~ lagsal
##
##          Df Sum of Sq    RSS    AIC
## + fflow   1   16.3161 42.518 17.697
## + year    1    6.0325 52.802 23.762
## <none>                58.835 24.791
## + period  1    2.8557 55.979 25.398
##
## Step:  AIC=17.7
## sal ~ lagsal + fflow
##
##          Df Sum of Sq    RSS    AIC
## + year    1    5.4556 37.063 15.851
## <none>                42.518 17.697
## + period  1    0.0444 42.474 19.667
##
## Step:  AIC=15.85
## sal ~ lagsal + fflow + year
##
##          Df Sum of Sq    RSS    AIC
## <none>                37.063 15.851
## + period  1   0.22573 36.837 17.680
```

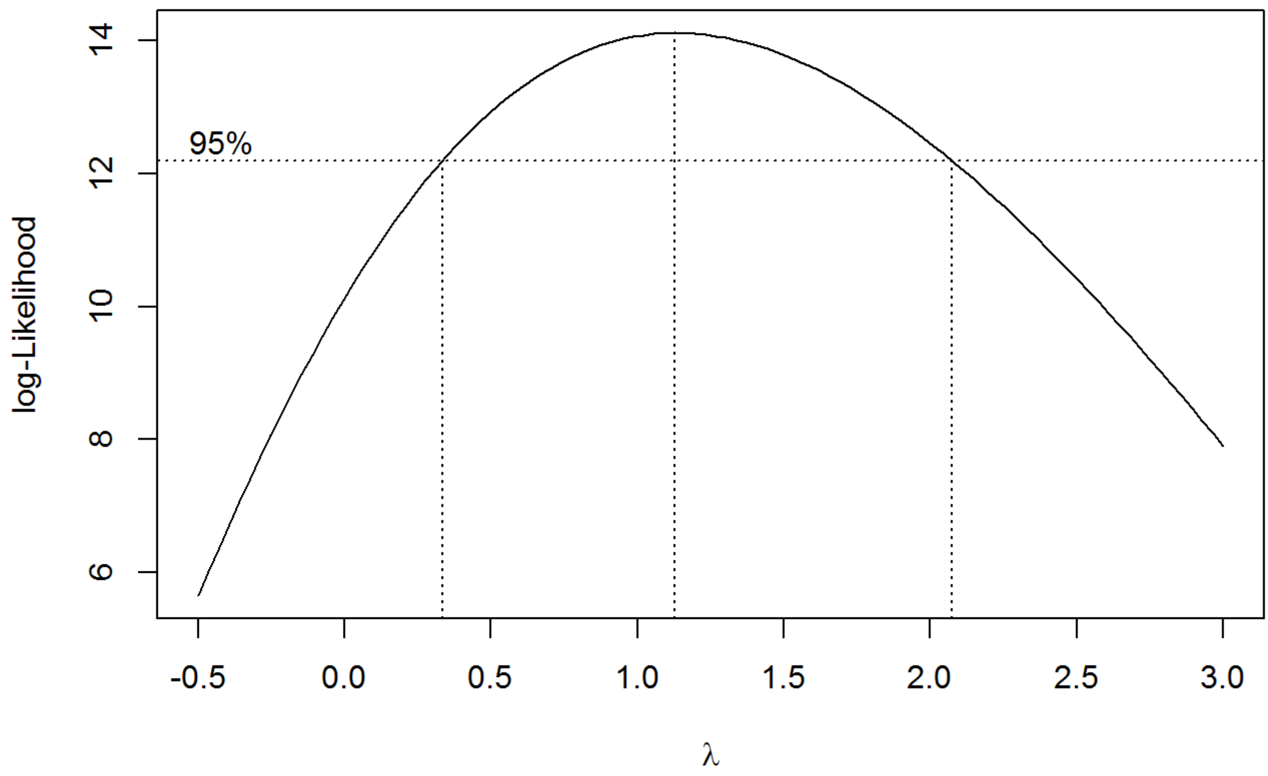
```
summary(sal.step)
```

```
##
## Call:
## lm(formula = sal ~ lagsal + fflow + year, data = sal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.75514 -0.50421  0.04945  0.57970  2.66638
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -783.28581  421.70351  -1.857  0.07556 .
## lagsal         0.62204    0.11552   5.385 1.57e-05 ***
## fflow         -0.28218    0.08839  -3.192  0.00391 **
## year           0.40214    0.21395   1.880  0.07236 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.243 on 24 degrees of freedom
## Multiple R-squared:  0.8485, Adjusted R-squared:  0.8296
## F-statistic: 44.81 on 3 and 24 DF,  p-value: 5.461e-10
```

Above model selection leads to model with legsal,fflow,year

Check whether Transformation is required using BOX COX

```
library("MASS")
boxcox(sal.step,lambda = seq(-.5,3,0.1),interp = TRUE)
bc <- boxcox(sal.step,lambda = seq(-.5,3,0.1),interp = TRUE)
```




```
bc$x[which.max(bc$y)]
```

```
## [1] 1.126263
```

The possible transformation candidates from this can be 0.5(Square root) and 1(No transformation)

Testing for square root and no transformation

Square Root

```
mean<-1;
for(i in 1:n){mean <- mean*sal$sal[i]};
lambda <- 0.5;p<-3;
h_sal <- (mean^(1-lambda))*((sal$sal^lambda-1)/lambda);
Sq_rt_sse <- (summary(lm(h_sal ~ sal$lagsal+sal$fflow+sal$year))$sigma)^2/(n-p);
AIC_sqrt<-n*log(Sq_rt_sse/n) +2*p
BIC_sqrt<-n*log(Sq_rt_sse/n) +p*log(n)
print(AIC_sqrt)
```

```
## [1] 1582.231
```

```
print(AIC_sqrt)
```

```
## [1] 1582.231
```

No Transformation lambda =1

```
mean<-1;
for(i in 1:n){mean <- mean*sal$sal[i]};
lambda <- 1;p<-3;
no_trans_sse <- (summary(lm(sal$sal ~ sal$lagsal+sal$fflow+sal$year))$sigma)^2/(n-p);
AIC_no_trans<-n*log(no_trans_sse/n) +2*p
BIC_no_trans<-n*log(no_trans_sse/n) +p*log(n)
print(AIC_no_trans)
```

```
## [1] -165.2626
```

```
print(BIC_no_trans)
```

```
## [1] -161.266
```

This gives lowet AIC and BIC with No Transformation. Hence, we can go ahead with our model of no transformation.

Building model with No Transformation

```
final_model <-lm(sal~lagsal+fflow+year, data = sal);
summary(final_model)
```

```
##
## Call:
## lm(formula = sal ~ lagsal + fflow + year, data = sal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.75514 -0.50421  0.04945  0.57970  2.66638
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -783.28581  421.70351  -1.857  0.07556 .
## lagsal         0.62204    0.11552   5.385 1.57e-05 ***
## fflow         -0.28218    0.08839  -3.192  0.00391 **
## year           0.40214    0.21395   1.880  0.07236 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.243 on 24 degrees of freedom
## Multiple R-squared:  0.8485, Adjusted R-squared:  0.8296
## F-statistic: 44.81 on 3 and 24 DF,  p-value: 5.461e-10
```

Diagnostics for influential values and outliers

```
sigma<-summary(final_model)$sigma
inf<- lm.influence(final_model)
shapiro.test(final_model$residual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  final_model$residual
## W = 0.93874, p-value = 0.1027
```

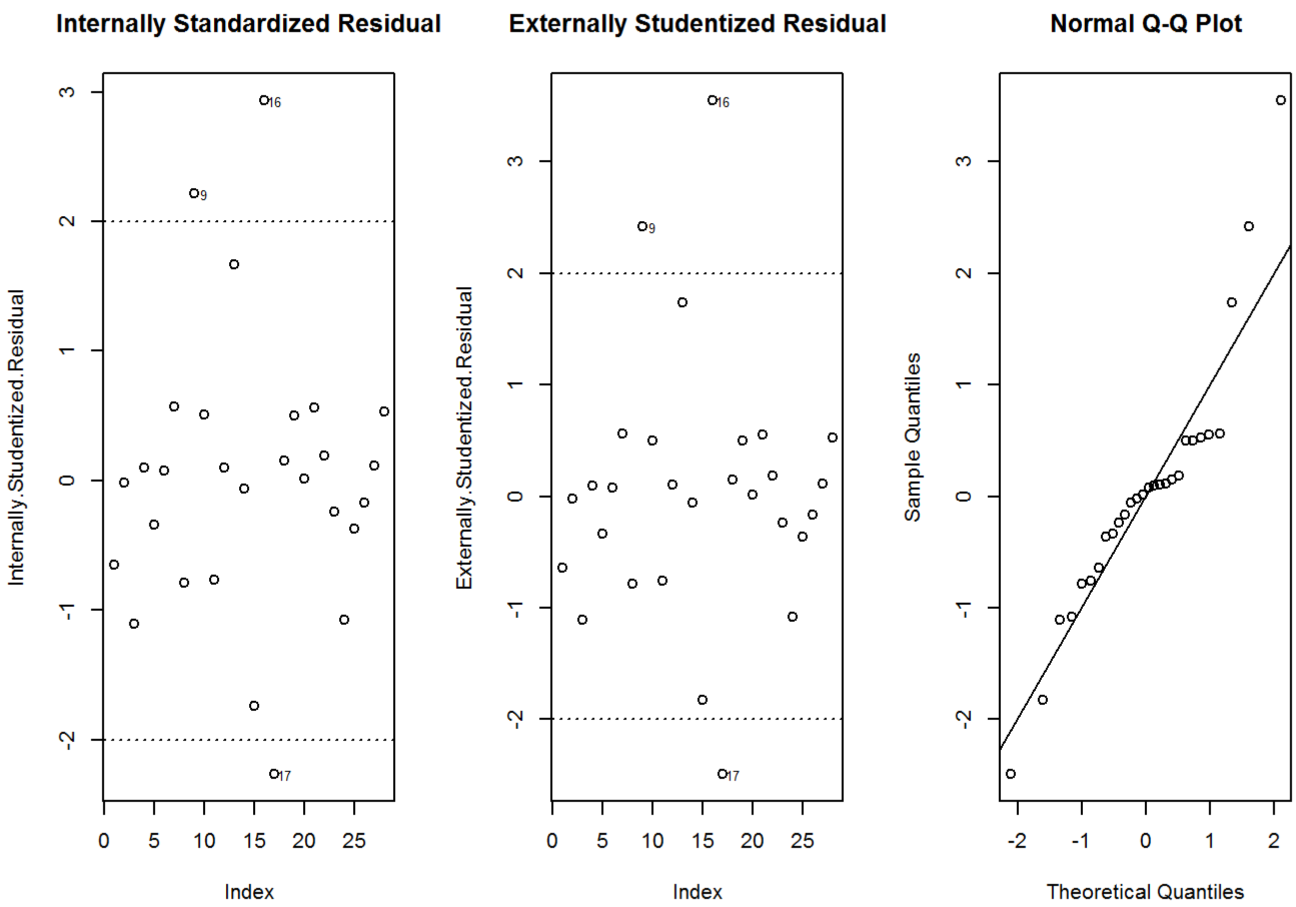
Shapiro Wilk test gives values on the higher side. Therefore, we need to test further with plots to get the influential observations

```
par(mfrow=c(1,3))

#Internal Studentized Residuals
Internally.Studentized.Residual<-final_model$residuals/
(sigma*sqrt(1-inf$hat))
plot(Internally.Studentized.Residual);title("Internally Standardized Residual")
abline(h=c(-2,2),lty="dotted")
for(i in 1:n){
if(abs(Internally.Studentized.Residual[i])> 2)text(i+1,Internally.Studentized.Residual[i],i,c
ex=0.6)}

#External Studentized Residuals
Externally.Studentized.Residual<-final_model$residuals*
sqrt((n-p-1)/((1-inf$hat)*(n-p)*sigma^2
- (final_model$residuals)^2))
plot(Externally.Studentized.Residual);title("Externally Studentized Residual")
abline(h=c(-2,2),lty="dotted")
for(i in 1:n){
if(abs(Externally.Studentized.Residual[i])> 2)text(i+1,Externally.Studentized.Residual[i],i,c
ex=0.6)}

#qq plot
qqnorm(Externally.Studentized.Residual);abline(c(0,0),c(1,1))
```



From these plots observation 16 and 9 seems outliers. Testing Leverage, Cook's D, DFFITS

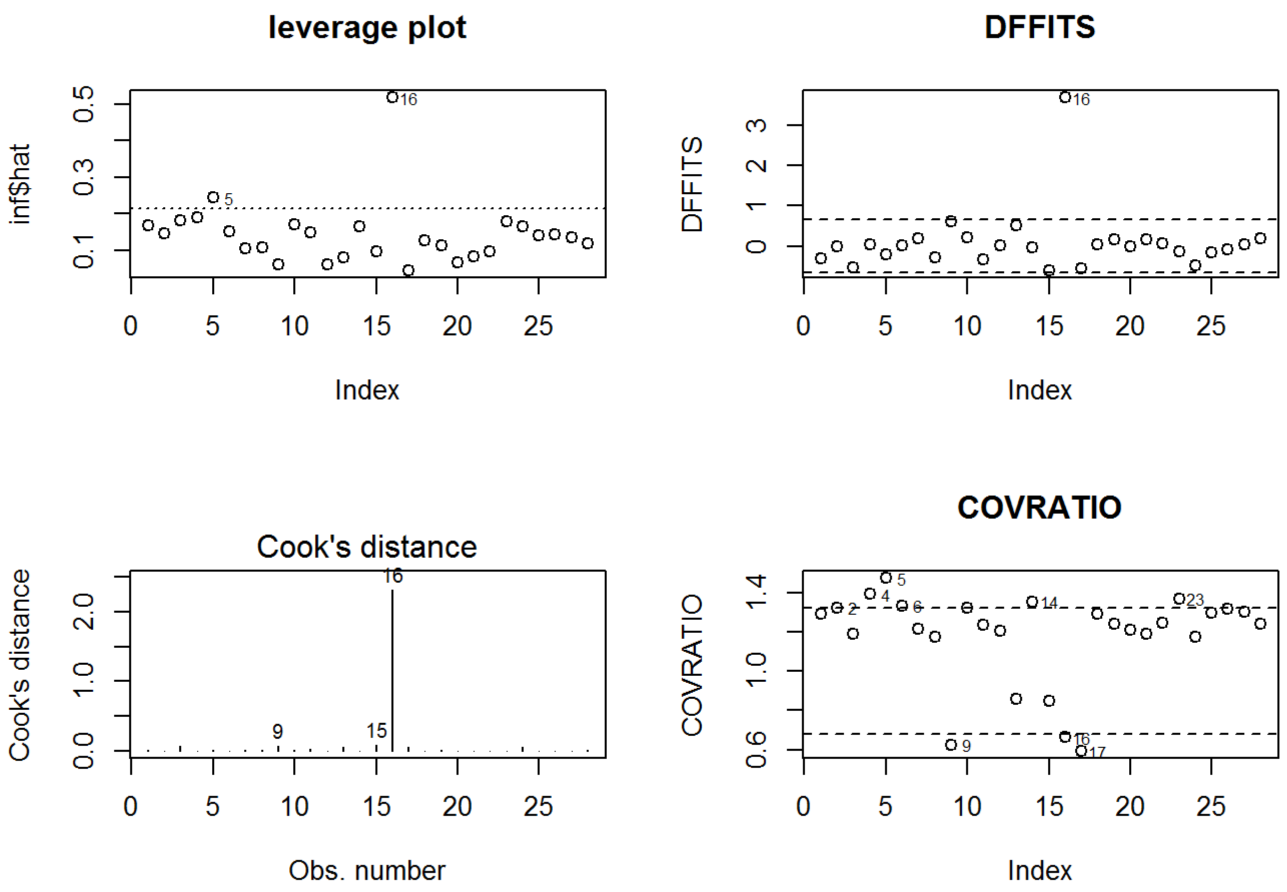
```
par(mfrow=c(2,2))

plot(inf$hat);
title("leverage plot")
abline(h=2*p/n,lty=3) ### high leverage points
leverage<-c(inf$hat>2*p/n)
for(i in 1:n){ if(leverage[i]==T)text(i+1,inf$hat[i],i,cex=0.6)}

DFFITS <- Externally.Studentized.Residual*sqrt(inf$hat/(1-inf$hat))
plot(DFFITS);abline(h=2*sqrt(p/n),lty=2);abline(h=-2*sqrt(p/n),lty=2);title("DFFITS")
DF.detected <- c(abs(DFFITS)> 2*sqrt(p/n))
for(i in 1:n){ if(DF.detected[i]==T)text(i+1,DFFITS[i],i,cex=0.6)}

#Cooks D
plot(final_model, which=4)

first <- ((n-p-1) / (n-p))+ Externally.Studentized.Residual^2/(n-p)
COVRATIO <- first^(-p) /(1-inf$hat)
plot(COVRATIO);abline(h=3*p/n+1,lty=2);
abline(h=-3*p/n+1,lty=2);title("COVRATIO")
CR.detected <- c(abs(COVRATIO-1)>3*p/n)
for(i in 1:n){ if(CR.detected[i]==T)text(i+1,COVRATIO[i],i,cex=0.6)}
```



From all these plots, clearly observation 16 seems to be an influential observation. Therefore, we can remove obs 16 from data and check normality

Removing Obervation 16 and test

```
sal_new<-subset(sal,obs !=16)
new_model <-lm(sal~lagsal+fflow+year, data = sal_new);
summary(new_model)
```

```
##
## Call:
## lm(formula = sal ~ lagsal + fflow + year, data = sal_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25275 -0.27150 -0.06741  0.40462  2.42491
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -660.19890  346.80024  -1.904   0.0695 .
## lagsal         0.58264    0.09517   6.122 3.03e-06 ***
## fflow         -0.54098    0.10222  -5.292 2.27e-05 ***
## year           0.34303    0.17586   1.951   0.0634 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.017 on 23 degrees of freedom
## Multiple R-squared:  0.9028, Adjusted R-squared:  0.8901
## F-statistic: 71.19 on 3 and 23 DF,  p-value: 8.623e-12
```

It can be seen from this model that p-value of year does not seems significant. Therefore, we can remove year and for a new model with lagsal and fflow

```
new_model_final <-lm(sal~lagsal+fflow, data = sal_new);
summary(new_model_final)
```

```
##
## Call:
## lm(formula = sal ~ lagsal + fflow, data = sal_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1474 -0.5500  0.2067  0.4770  2.1943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.24142    2.87288   5.653 8.04e-06 ***
## lagsal       0.71311    0.07155   9.967 5.24e-10 ***
## fflow       -0.55841    0.10761  -5.189 2.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 24 degrees of freedom
## Multiple R-squared:  0.8867, Adjusted R-squared:  0.8772
## F-statistic: 93.91 on 2 and 24 DF,  p-value: 4.479e-12
```

This looks to be the optimal model we can get. The R square value is able to explain the maximum variability in the data of around 87%