AIFFD Chapter 4 - Recruitment

Derek H. Ogle

Contents

This document contains R versions of the boxed examples from **Chapter 4** of the *Analysis and Interpretation of Freshwater Fisheries Data* book. Some sections build on descriptions from previous sections, so each section may not stand completely on its own. More thorough discussions of the following items are available in linked vignettes:

- the use of linear models in R in the linear models vignette,
- differences between and the use of type-I, II, and III sums-of-squares in the preliminaries vignette, and
- the use of "least-squares means" is found in the preliminaries vignette.

The following additional packages are required to complete all of the examples (with the required functions noted as a comment and also noted in the specific examples below).

```
library(FSA)          # fitPlot, Subset
library(NCStats)      # addSigLetters
library(car)          # Anova, durbinWatsonTest, vif
library(Hmisc)        # rcorr
library(Kendall)      # kendall
library(lsmeans)      # lsmeans
library(multcomp)     # glht, mcp, cld
library(nlstools)     # overview, nlsBoot
library(plotrix)      # thigmophobe.labels, rescale
```

In addition, external tab-delimited text files are used to hold the data required for each example. These data are loaded into R in each example with `read.table()`. Before using `read.table()` the working directory of R must be set to where these files are located on **your** computer. The working directory for all data files on **my** computer is set below.

```
setwd("c:/aaaWork/web/fishR/BookVignettes/AIFFD/")
```

In addition, I prefer to not show significance stars for hypothesis test output, reduce the margins on plots, alter the axis label positions, and reduce the axis tick length. In addition, contrasts are set in such a manner as to force R output to match SAS output for linear model summaries. All of these options are set below.

```
options(width=90,continue=" ",show.signif.stars=FALSE,contrasts=c("contr.sum","contr.poly"))
par(mar=c(3.5,3.5,1,1),mgp=c(2.1,0.4,0),tcl=-0.2)
```

0.1 Log-Linear Model to Test for Year-Class Abundance Differences

Below we conduct a test for year-class abundance differences among the 1990 to 1997 year-classes (`yearcl`) based on catch rates of age-0, age-1, and age-2 crappies (*Pomoxis* spp.) (`age` in years) from Weiss Lake (Table 4.1 in text). Trap-net catch rates (`catch`) were transformed to natural log values to homogenize variances as recommended by (Kimura 1988) for log-linear analysis. The data in Table 4.1 were rearranged to conduct the analysis. Year of collection (`yearcol`) was included in the data file, and the following R code was used to conduct the analysis.

0.1.1 Preparing Data

The `box4_1.txt` is read, the structure of the data frame is observed, and a new variable, `lcatch`, that is the natural log of the catch variable is created.

```
d1 <- read.table("data/box4_1.txt",header=TRUE)
str(d1)
```

```
## 'data.frame':    24 obs. of  4 variables:
##  $ yearcol: int  1990 1991 1992 1991 1992 1993 1992 1993 1994 1993 ...
##  $ yearcl : int  1990 1990 1990 1991 1991 1991 1992 1992 1992 1993 ...
##  $ age    : int  0 1 2 0 1 2 0 1 2 0 ...
##  $ catch  : num  8.03 5.32 2.43 0.47 0.39 0.39 0.61 0.97 0.61 1.38 ...
```

```
d1$lcatch <- log(d1$catch)
```

In addition, R must be told explicitly that `age` and `yearcl` are group factor rather than numeric variables.

```
d1$age <- factor(d1$age)
d1$yearcl <- factor(d1$yearcl)
```

0.1.2 Two-Way ANOVA Model

The authors of Box 4.1 fit a two-way ANOVA with**OUT** an interaction term. While a two-way ANOVA is generally fit with an interaction term, not using an interaction term is appropriate here because an interaction cannot be estimated as there are not multiple observations for each combination of the two factors. This fact is best illustrated with a two-way frequency table constructed from the two group factor variables with `table()`.

```
table(d1$age,d1$yearcl)
```

```
##
##     1990 1991 1992 1993 1994 1995 1996 1997
##   0    1    1    1    1    1    1    1    1
##   1    1    1    1    1    1    1    1    1
##   2    1    1    1    1    1    1    1    1
```

The two-way ANOVA model without the interaction term is fit with `lm()`. The ANOVA table using type-III SS is then extracted from the `lm()` object with `Anova()` from the `car` package. From this, there is evidence for a significant difference in means among ages and among year-classes.

```
lm1 <- lm(lcatch~age+yearcl,data=d1)
Anova(lm1,type="III")
```

```
##               Sum Sq     Df F value    Pr(>F)
## (Intercept)  2.7457  1.000 12.4839  0.003309
## age          4.0952  2.000  9.3097  0.002683
## yearcl      19.0150  7.000 12.3508 4.994e-05
## Residuals    3.0792 14.000
## Total       24.0000 28.935
```

0.1.3 Least-Squares Means for Year-Class

Least-squares means are computed with `lsmeans()` from the `lsmeans` package using a right-hand-sided formula as the second argument to isolate the least-square means for each factor variable. From this, it appears that CPE generally decreases (not surprisingly) with age and that 1990 and 1996 are relatively strong year-classes and 1991 is a relatively poor year-class.

```
lsmeans(lm1,~age)
```

```
##  age     lsmean        SE df   lower.CL  upper.CL
##  0    0.6973961 0.1658088 14  0.3417717 1.0530205
##  1    0.5576609 0.1658088 14  0.2020365 0.9132853
##  2   -0.2403432 0.1658088 14 -0.5959676 0.1152812
##
## Results are averaged over the levels of: yearcl
## Confidence level used: 0.95
```

```
lsmeans(lm1,~yearcl)
```

```
##  yearcl     lsmean        SE df    lower.CL   upper.CL
##  1990    1.5475164 0.2707646 14  0.96678413  2.1282486
##  1991   -0.8794132 0.2707646 14 -1.46014545 -0.2986810
##  1992   -0.3396840 0.2707646 14 -0.92041618  0.2410483
##  1993    0.6259558 0.2707646 14  0.04522358  1.2066880
##  1994    0.6280314 0.2707646 14  0.04729921  1.2087637
##  1995   -0.2701080 0.2707646 14 -0.85084020  0.3106243
##  1996    1.7311522 0.2707646 14  1.15041997  2.3118844
##  1997   -0.3375473 0.2707646 14 -0.91827955  0.2431849
##
## Results are averaged over the levels of: age
## Confidence level used: 0.95
```

0.1.4 Multiple Comparisons

The authors of Box 4.1 use Fisher's LSD multiple comparison procedure. In general, this procedure does not guard well against an inflated experimentwise error rate. In general, Tukey's HSD procedure performs better in this regard and will be illustrated below.

Tukey's multiple comparison procedure, implemented through `glht()` from the `multcomp` package, can be used to identify where the differences in means occur. The `glht()` function requires the `lm()` object as the first argument. The second argument is also required and uses `mcp()` to declare a "multiple comparison procedure." In this instance the argument to `mcp()` is the factor variable in the `lm()` object for which you are testing for differences set equal to the `"Tukey"` string. The result from `glht()` is saved to an object that can be submitted to `summary()` to extract p-values for each difference in pairs of means, to `confint()` to extract confidence intervals for each difference in pairs of means, and to `cld()` to identify significance letters that depict significant differences among means.

```
mc1a <- glht(lm1,mcp(age="Tukey"))
summary(mc1a)
```

```
##            Estimate Std. Error t value Pr(>|t|)
## 1 - 0 == 0  -0.1397     0.2345  -0.596  0.82454
## 2 - 0 == 0  -0.9377     0.2345  -3.999  0.00356
## 2 - 1 == 0  -0.7980     0.2345  -3.403  0.01121
```

```
mc1yc <- glht(lm1,mcp(yearcl="Tukey"))
summary(mc1yc)
```

```
## Warning in RET$pfunction("adjusted", ...): Completion with error > abseps

## Warning in RET$pfunction("adjusted", ...): Completion with error > abseps

##                  Estimate Std. Error t value Pr(>|t|)
## 1991 - 1990 == 0 -2.426930   0.382919  -6.338  < 0.001
## 1992 - 1990 == 0 -1.887200   0.382919  -4.928  0.00405
## 1993 - 1990 == 0 -0.921561   0.382919  -2.407  0.30867
## 1994 - 1990 == 0 -0.919485   0.382919  -2.401  0.31060
## 1995 - 1990 == 0 -1.817624   0.382919  -4.747  0.00566
## 1996 - 1990 == 0  0.183636   0.382919   0.480  0.99959
## 1997 - 1990 == 0 -1.885064   0.382919  -4.923  0.00413
## 1992 - 1991 == 0  0.539729   0.382919   1.410  0.83899
## 1993 - 1991 == 0  1.505369   0.382919   3.931  0.02426
## 1994 - 1991 == 0  1.507445   0.382919   3.937  0.02414
## 1995 - 1991 == 0  0.609305   0.382919   1.591  0.74818
## 1996 - 1991 == 0  2.610565   0.382919   6.818  < 0.001
## 1997 - 1991 == 0  0.541866   0.382919   1.415  0.83635
## 1993 - 1992 == 0  0.965640   0.382919   2.522  0.26178
## 1994 - 1992 == 0  0.967715   0.382919   2.527  0.25951
## 1995 - 1992 == 0  0.069576   0.382919   0.182  1.00000
## 1996 - 1992 == 0  2.070836   0.382919   5.408  0.00181
## 1997 - 1992 == 0  0.002137   0.382919   0.006  1.00000
## 1994 - 1993 == 0  0.002076   0.382919   0.005  1.00000
## 1995 - 1993 == 0 -0.896064   0.382919  -2.340  0.33832
## 1996 - 1993 == 0  1.105196   0.382919   2.886  0.14949
## 1997 - 1993 == 0 -0.963503   0.382919  -2.516  0.26389
## 1995 - 1994 == 0 -0.898139   0.382919  -2.346  0.33523
## 1996 - 1994 == 0  1.103121   0.382919   2.881  0.15076
## 1997 - 1994 == 0 -0.965579   0.382919  -2.522  0.26220
## 1996 - 1995 == 0  2.001260   0.382919   5.226  0.00233
## 1997 - 1995 == 0 -0.067439   0.382919  -0.176  1.00000
## 1997 - 1996 == 0 -2.068700   0.382919  -5.402  0.00168
```

```
cld(mc1yc)
```

```
## Warning in RET$pfunction("adjusted", ...): Completion with error > abseps
```

```
## Warning in RET$pfunction("adjusted", ...): Completion with error > abseps
```

```
## 1990 1991 1992 1993 1994 1995 1996 1997
##  "c"  "a" "ab" "bc" "bc" "ab"  "c" "ab"
```

An plot of the group means, with appropriate confidence intervals is constructed with `fitPlot()` from the `FSA` package. The significance letters can be added to the means plot with `addSigLetters()` from the `NCStats` package. See `?addSigLetters` for a description of the arguments.

```
fitPlot(lm1,which="yearcl",xlab="Year Class",ylab="Loge(Catch)",main="")
addSigLetters(lm1,which="yearcl",lets=c("c","a","ab","bc","bc","ab","c","ab"),
              pos=c(4,2,2,2,4,2,2,2),cex=0.75)
```

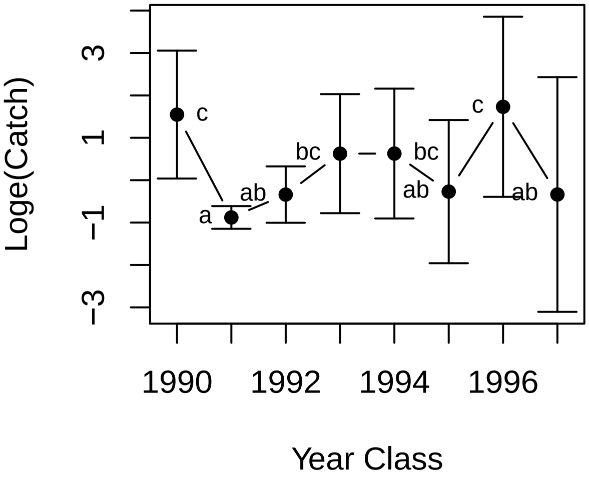

Figure 1:

0.2 Evaluation of Time Series Trends in Recruit Abundance

The following code presents a plot and computes the Pearson correlation coefficient between age-0 C/f (`age0cpe`) of white bass (*Morone chrysops*) and year and the Kendall tau-b non-parametric correlation coefficient for ranks between these two variables (data published in (Madenjian et al. 2000)). In addition, the simple linear regression between C/f and year was computed along with the Durbin-Watson statistic to determine temporal autocorrelation. Finally, the residuals from the regression were plotted against year.

0.2.1 Preparing Data

The box4_2.txt data file is read and the structure of the data frame is observed below. It appears that the authors of Box 4.2 only used years from 1972-1997 in Box 4.2 even though the data provided on the CD with the AIFFD book is for 1969-1997. Thus, a new data frame, called d1, restricted to years after 1971 is created with Subset() from the FSA package, with the original data frame as the first argument and a conditional statement from which to create the subset as the second argument. Note that Subset() is very similar to subset() from base R with the exception that Subset() will remove unused levels from a factor variable after the subsetting. This feature is useful in many situations but is irrelevant in this situation as the subsetting is conducted on a non-factor variable.

```
d2 <- read.table("data/box4_2.txt",header=TRUE)
str(d2)
```

```
## 'data.frame':    29 obs. of  2 variables:
##  $ year   : int  1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 ...
##  $ age0cpe: num  206.08 202.51 75.17 24.38 4.29 ...
```

```
d2a <- Subset(d2,year>=1972)
str(d2a)
```

```
## 'data.frame':    26 obs. of  2 variables:
##  $ year   : int  1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 ...
##  $ age0cpe: num  24.38 4.29 10.06 18.16 23.44 ...
```

0.2.2 Summary Plot and Statistics

The authors of Box 4.2 initially plot age-0 CPE against year. This plot is constructed in R with the first use of plot() below. I prefer to connect the points to more easily see the year-to-year pattern. This modification is accomplished with the second use of plot() below (note the use of type="b" where "b" is for “both”" points and lines.)

```
plot(age0cpe~year,data=d2a,ylab="Age-0 CPE",xlab="Year",main="",pch=19)
```

```
plot(age0cpe~year,data=d2a,type="b",ylab="Age-0 CPE",xlab="Year",main="",pch=19)
```

The summary statistics presented in Box 4.2 are efficiently computed with Summarize() from the FSA package.

```
Summarize(d2a$age0cpe,digits=4)
```

```
##        n     mean       sd      min       Q1   median       Q3      max percZero
##  26.0000   8.5535   8.0782   0.5700   2.9150   4.6850  11.0400  25.2400   0.0000
```

```
Summarize(d2a$year,digits=4)
```

```
##         n      mean        sd       min        Q1    median        Q3       max  percZero
##   26.0000 1984.5000    7.6485 1972.0000 1978.0000 1984.0000 1991.0000 1997.0000    0.0000
```

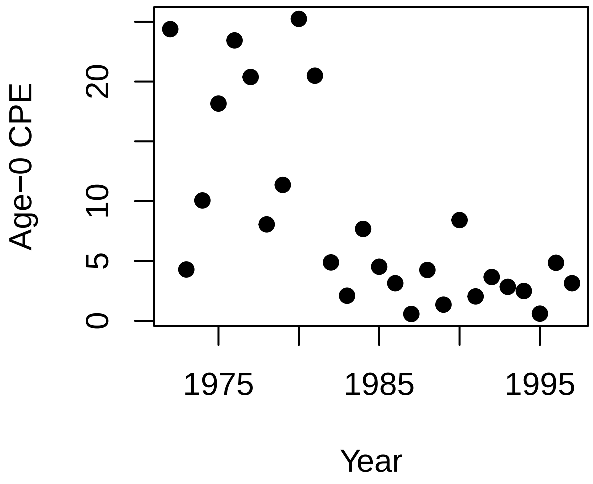

Figure 2:

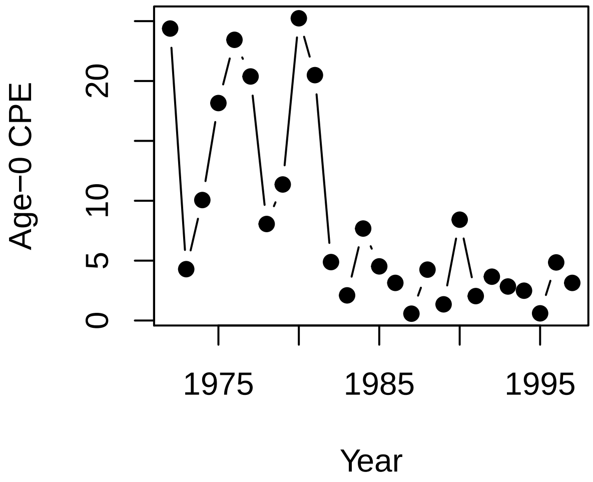

Figure 3:

0.2.3 Pearson Correlation Analysis

The authors of Box 4.2 examined the correlation coefficient between `age0cpe` and `year`, with p-values corresponding to a test of whether the correlation is equal to zero or not. The `rcorr()` function from the `Hmisc` package is needed to compute both the correlation coefficients AND the corresponding p-values. [*NOTE: The correlation coefficients alone are computed with* `cor` *in base R.*] The `rcorr()` function requires a **matrix** as the first argument so the two variables in the data frame must first be isolated and then coerced to a matrix with `as.matrix()`.

```
rcorr(as.matrix(d2a[,c("age0cpe","year")]))
```

```
##          age0cpe  year
## age0cpe     1.00 -0.67
## year       -0.67  1.00
##
## n= 26
##
##
## P
##          age0cpe year
## age0cpe          2e-04
## year     2e-04
```

It should be noted that the relationship between `age0cpe` and `year` is clearly non-linear (as observed from the previous plot) indicating that the value of the correlation coefficient is not strictly interpretable (i.e., correlation coefficients assume a linear relationship).

0.2.4 Kendall's Tau Correlation Analysis}

Kendall's tau-b correlation coefficient is computed by submitting the two variables as the first two arguments to `Kendall()` from the `Kendall` package.

```
Kendall(d2a$age0cpe,d2a$year)
```

```
## tau = -0.487, 2-sided pvalue =0.00053745
```

0.2.5 Analysis of Residuals from Regression}

The simple linear regression described in Box 4.2 is fit with `lm()`. The ANOVA table is extracted from the `lm()` object with `anova()` and the coefficients or estimated parameters, among other summary information, is extracted with `summary()`. Because of the evident non-linearity, this result is of dubious value. However, if it were to be interpreted, it shows a significant negative relationship between age-0 CPE and year.

```
lm1 <- lm(age0cpe~year,data=d2a)
anova(lm1)
```

```
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## year         1  736.99  736.99  19.775 0.0001695
## Residuals 24  894.44   37.27
## Total     25 1631.43
```

```
summary(lm1)
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1417.3042   316.7934   4.474 0.000158
## year          -0.7099     0.1596  -4.447 0.000170
##
## Residual standard error: 6.105 on 24 degrees of freedom
## Multiple R-squared: 0.4517,  Adjusted R-squared: 0.4289
## F-statistic: 19.78 on 1 and 24 DF,  p-value: 0.0001695
```

The Durbin-Watson statistic is computed by submitting the `lm()` object as the first argument to `durbinWatsonTest()` from the `car` package.}. By default `durbinWatsonTest()` computes the statistics only for times lags of one unit. The `max.lag=` argument is used to compute the Durbin-Watson statistic for other time lags (illustrated below for a time lag of five years). The autocorrelation value and test statistic for a time-lag of 1 are the same as shown in Box 4.2; however, the interpretation from the p-value shown here for a time-lag of 1 and what was described in Box 4.2 are different. These results, if a 5% significance level is used, suggest that there is no significant autocorrelation up to fifth order time lags.

```
durbinWatsonTest(lm1,max.lag=5)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2162394      1.500086   0.112
##    2      -0.3226973      2.383062   0.256
##    3      -0.1726707      2.043186   0.716
##    4       0.2916256      1.104914   0.044
##    5       0.2046703      1.191092   0.132
##  Alternative hypothesis: rho[lag] != 0
```

The residuals from the model fit are stored in the `$residuals` object of the `lm()` object. These residuals can be accessed to construct a plot of residuals versus year.

```
plot(lm1$residuals~d2a$year,ylab="Residuals",xlab="Year",main="",pch=19)
abline(h=0,lty=2)                                # adds horizontal reference line at 0
```

```
plot(lm1$residuals~d2a$year,type="b",ylab="Residuals",xlab="Year",main="",pch=19)
abline(h=0,lty=2)
```

0.2.6 An Alternative Residual Analysis}

The results from above indicate that a strong overall trend in the CPE data by year – i.e., high and variable CPE in the early years followed by a much lower and less variable CPE in the later years. The simple linear regression does not represent this trend very well as exhibited by the residual plot (above) and the following fitted-line plot constructed with `fitPlot()` from the `FSA` package.}.

```
fitPlot(lm1,ylab="Age-0 CPE",xlab="Year",main="")
```

If the age-0 CPE is transformed to the log scale (new variable called `logage0cpe`) and a new linear model is fit then trends in the residuals with the overall trend removed more appropriately can be examined.

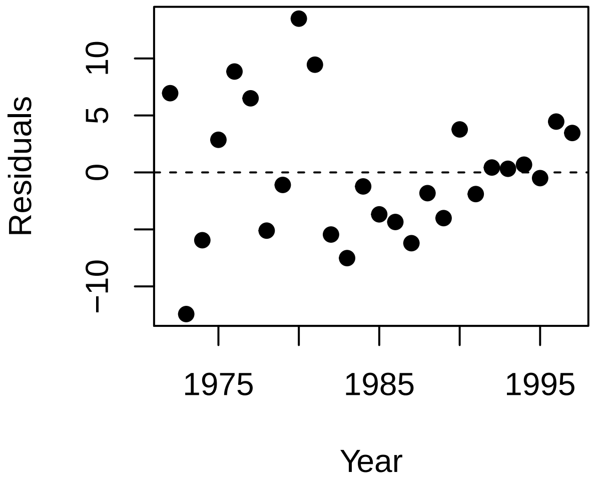

Figure 4:

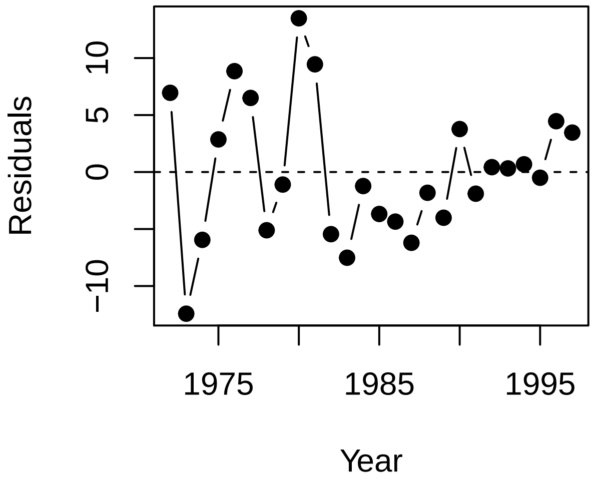

Figure 5:

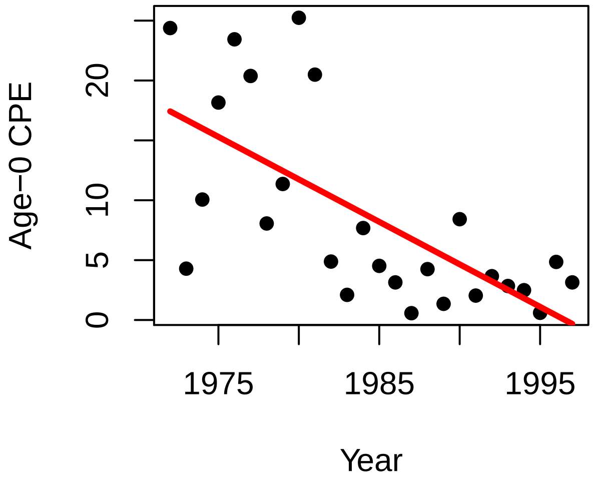

Figure 6:

```
d2a$logage0cpe <- log(d2a$age0cpe)
lm2 <- lm(logage0cpe~year,data=d2a)
plot(lm2$residuals~d2a$year,ylab="Residuals",xlab="Year",main="",pch=19)
abline(h=0,lty=2)                           # adds horizontal reference line at 0
```

```
plot(lm2$residuals~d2a$year,type="b",ylab="Residuals",xlab="Year",main="",pch=19)
abline(h=0,lty=2)
```

With these changes, there is a significant negative trend in the relationship between log CPE of age-0 fish by year and neither the Durbin-Watson statistic (there is a weak suggestion for a time-lag of four years) or the auto-correlation function test (implemented with `ccf()`) found a significant auto-correlation in the data.

```
anova(lm2)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## year       1 12.983 12.9833  19.714 0.0001725
## Residuals 24 15.806  0.6586
## Total     25 28.789
```

```
summary(lm2)
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 188.64581   42.11253    4.48 0.000156
## year         -0.09422    0.02122   -4.44 0.000173
##
```

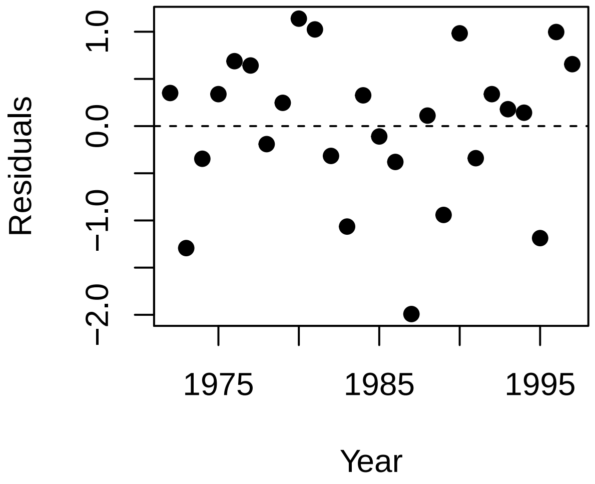

Figure 7:

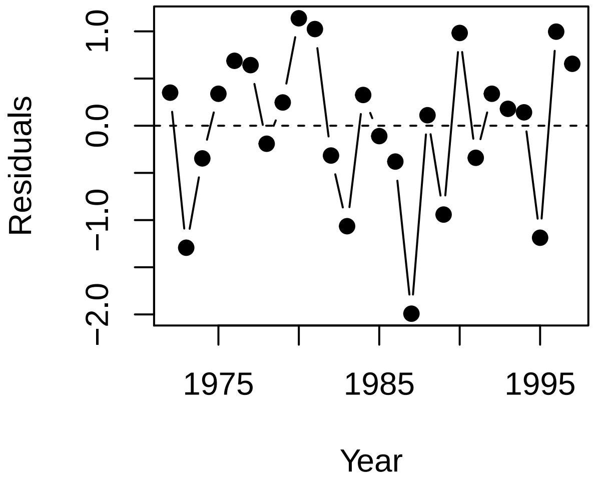

Figure 8:

```
## Residual standard error: 0.8115 on 24 degrees of freedom
## Multiple R-squared: 0.451,   Adjusted R-squared: 0.4281
## F-statistic: 19.71 on 1 and 24 DF,  p-value: 0.0001725
```

```
durbinWatsonTest(lm2,max.lag=5)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1    -0.003428693      1.971829   0.794
##    2    -0.008081435      1.812510   0.644
##    3    -0.220896027      2.141396   0.538
##    4     0.291742801      1.107594   0.034
##    5    -0.050812342      1.760729   0.922
##  Alternative hypothesis: rho[lag] != 0
```

```
ccf(lm2$residuals,d2a$year)
```

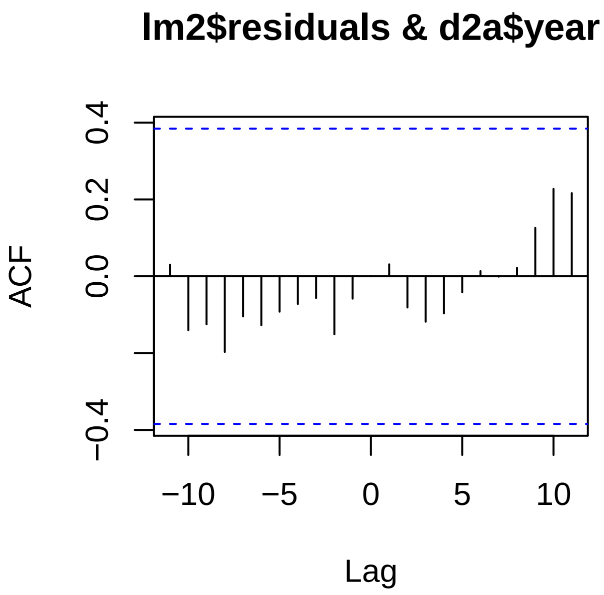

Figure 9:

For completeness, a fitted-“line” plot of the this alternative model to the original data is constructed by predicting log age-0 CPE for each year, back-transforming these values (i.e., using as the power of e), and the plotting the back-transformed values against year on top of a plot that already has the original values plotted against year.

```
plot(age0cpe~year,data=d2a,ylab="Age-0 CPE",xlab="Year",pch=19)
pCPE <- exp(predict(lm2))
lines(pCPE~d2a$year,lwd=2,col="red")
```

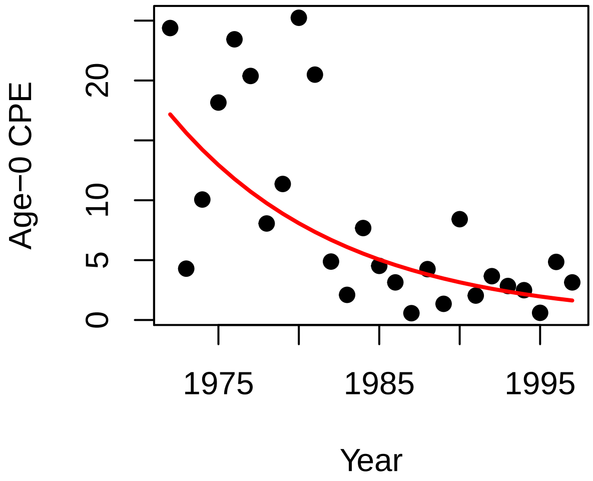

Figure 10:

Kimura, D. K. 1988. Analyzing relative abundance indices with log-linear models. Transactions of the American Fisheries Society 8:175–180.

Madenjian, D. P., R. L. Knight, M. T. Bur, and J. L. Forney. 2000. Reduction in recruitment of white bass in Lake Erie after invasion of white perch. Transactions of the American Fisheries 129:1340–1353.