STAT 560 Test 1

Yuchi Hu

October 6, 2018

1.

Table B.22 contains data from the Danish Energy Agency on Danish crude oil production.

1)

Plot the data and comment on any features that you observe from the graph.

Answer

```
# Load the B.22 data
B.22 <- read.csv('C:\\Users\\George\\Desktop\\TimeSeriesAnalysis\\B.22.csv')

# Crude oil production plot
nrb <- length(B.22$Month)
tt <- 1:nrb

plot(tt, B.22$Oil.Production, type='l', main='Denmark Crude Oil Production',
     ylab='Crude Oil Production (In Thousands of Tons)', xlab='', xaxt='n')
axis(1, seq(1, nrb, 6), labels=B.22$Month[seq(1, nrb, 6)], las=2)
points(tt, B.22$Oil.Production, pch=16)
```

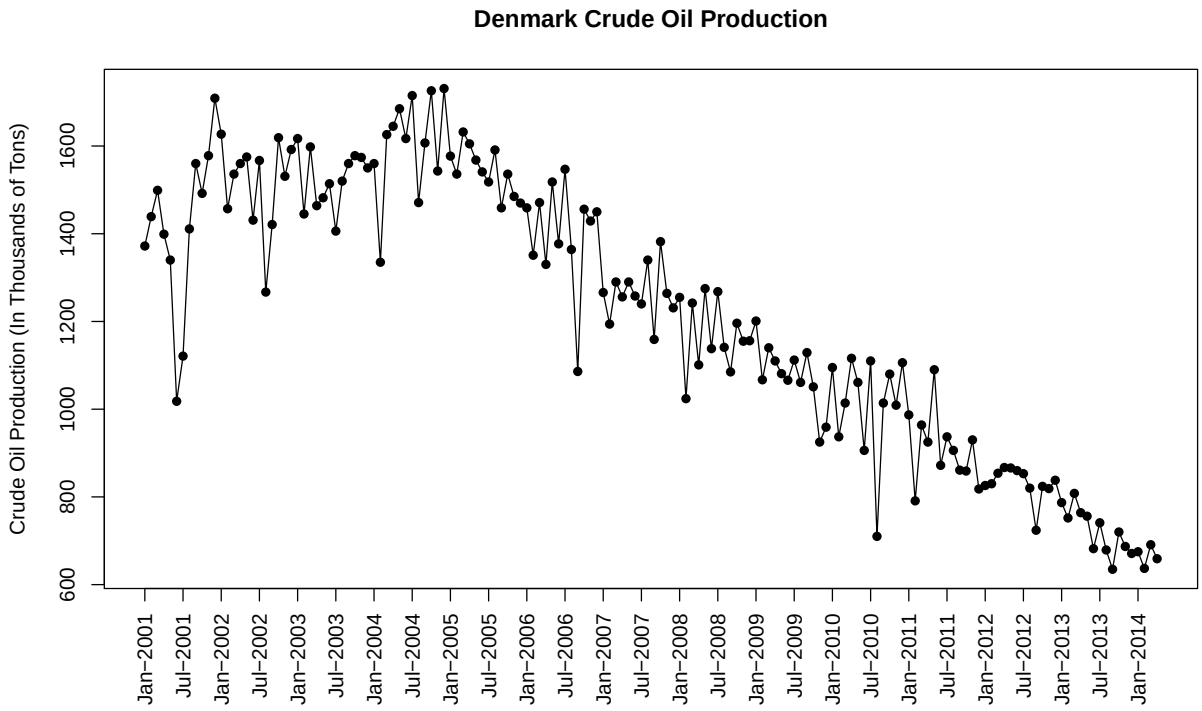

Figure 1: Plot of Denmark crude oil production.

Figure 1 shows the plot of Denmark crude oil production. We can see that crude oil production is more or less a constant process from 2001 to 2003. It briefly exhibits an increasing trend from 2003 to 2005, then exhibits a decreasing trend from 2005 to 2014.

2)

Calculate the sample Autocorrelation Function, variogram and Ljung-Box Statistics (Q_{LB}). Present your results in a table.

Answer

```
# Sample ACF
acf.oil <- acf(B.22$Oil.Production, lag.max=20, plot=F)

# Variogram
G <- NULL
d <- NULL
s.squared <- NULL
variogram.function <- function(y) {
  for (k in 1:20) {
    T <- length(y)
    t <- 1:(T-k)

    d[t] <- y[t+k] - y[t]
```

```
    d.bar <- (1/(T-k)) * sum(d[t])
    s.squared[k] <- sum((d[t]-d.bar)^2) / (T-k-1)
    G[k] <- s.squared[k] / s.squared[1]
  }
  return(G)
}
vario.oil <- variogram.function(B.22$Oil.Production)

# Ljung-Box statistic
QLB <- NULL
ljung.function <- function(y, acf) {
  T <- length(y)
  for (k in 1:20) {
    QLB[k] <- T * (T+2) * sum((1/(T-1:k)) * (acf$acf[2:(k+1)]^2))
  }
  return(QLB)
}
ljung.oil <- ljung.function(B.22$Oil.Production, acf.oil)

# Create a table
library(knitr)

dt <- cbind(Lag=1:20, 'Sample ACF'=acf.oil$acf[2:21], Variogram=vario.oil,
            'Ljung-Box Statistic'=ljung.oil)
kable(dt, align='c', caption='Sample ACF, Variogram, and Ljung-Box Statistic for
      Denmark Crude Oil Production (First 20 Lags)')
```

Table 1: Sample ACF, Variogram, and Ljung-Box Statistic for Denmark Crude Oil Production (First 20 Lags)

Lag	Sample ACF	Variogram	Ljung-Box Statistic
1	0.9088884	1.000000	134.6663
2	0.8949751	1.043710	266.0677
3	0.8691730	1.196885	390.7910
4	0.8459242	1.354737	509.6887
5	0.8485752	1.186795	630.1046
6	0.8388484	1.178588	748.5399
7	0.8307094	1.174340	865.4473
8	0.8104798	1.265067	977.4621
9	0.7907463	1.329850	1084.7949
10	0.7800515	1.321427	1189.9403
11	0.7536279	1.462668	1288.7416
12	0.7566345	1.185929	1389.0057
13	0.7262114	1.391980	1481.9973
14	0.7172971	1.382402	1573.3414
15	0.6974213	1.469012	1660.2890
16	0.6713591	1.645282	1741.4191
17	0.6606957	1.622257	1820.5420
18	0.6431989	1.731705	1896.0577
19	0.6317656	1.708782	1969.4293
20	0.6092918	1.871582	2038.1611

3)

Plot the ACF and variogram. Interpret these graphs.

Answer

```
par(mfrow=c(1,2))
# ACF plot
acf(B.22$Oil.Production, lag.max=20, main='Sample ACF of Denmark Crude Oil Production')

# Variogram plot
plot(vario.oil, pch=16, main='Variogram of Denmark Crude Oil Production',
     ylab='Variogram', xlab='Lag')
```

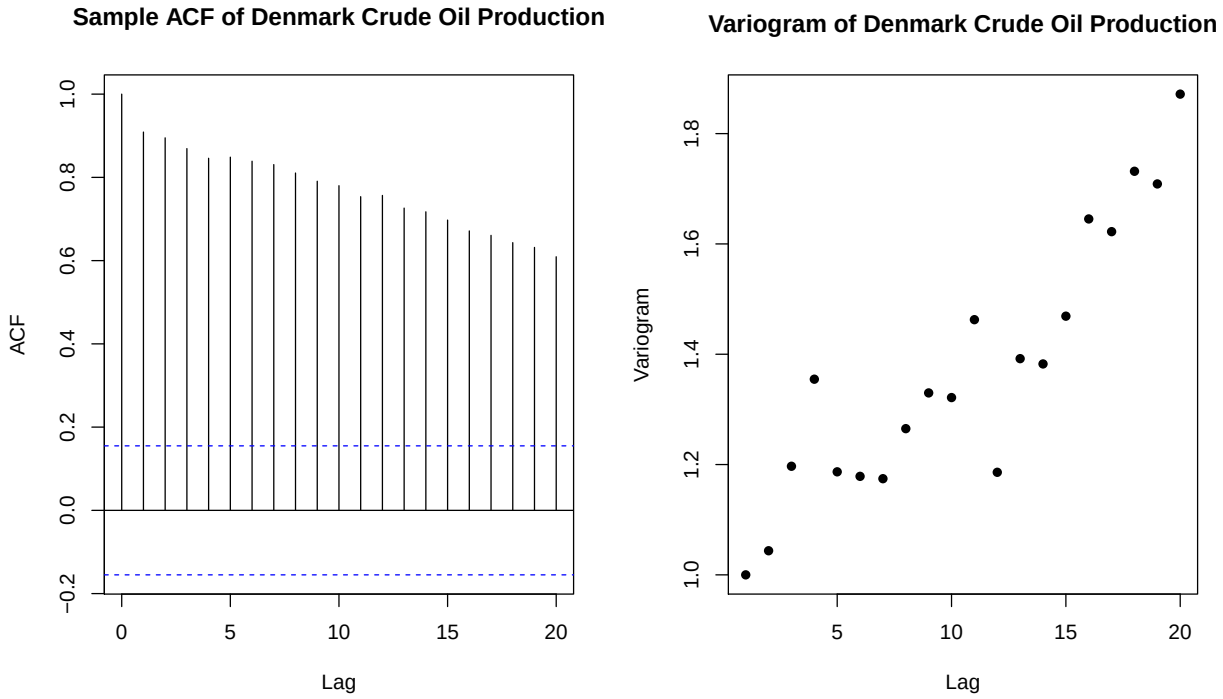

Figure 2: Sample ACF and variogram plots of Denmark crude oil production.

Figure 2 shows the sample ACF and variogram plots of Denmark crude oil production. We can see that the sample ACF decays very slowly, so the data appears to be nonstationary. The variogram does not converge to a generally stable level, so this is another indication that the data is nonstationary.

4)

Plot the first difference of the data and comment on any features that you observe from the graph.

Answer

```
# First difference of the data
dB.22 <- B.22
dB.22$Oil.Production <- c(NA, diff(dB.22$Oil.Production,1))

# Crude oil production plot (first difference)
plot(tt, dB.22$Oil.Production, type='l', main='Denmark Crude Oil Production, d = 1',
     ylab='Crude Oil Production (In Thousands of Tons), d = 1', xlab='', xaxt='n')
axis(1, seq(1, nrb, 6), labels=dB.22$Month[seq(1, nrb, 6)], las=2)
points(tt, dB.22$Oil.Production, pch=16)
```

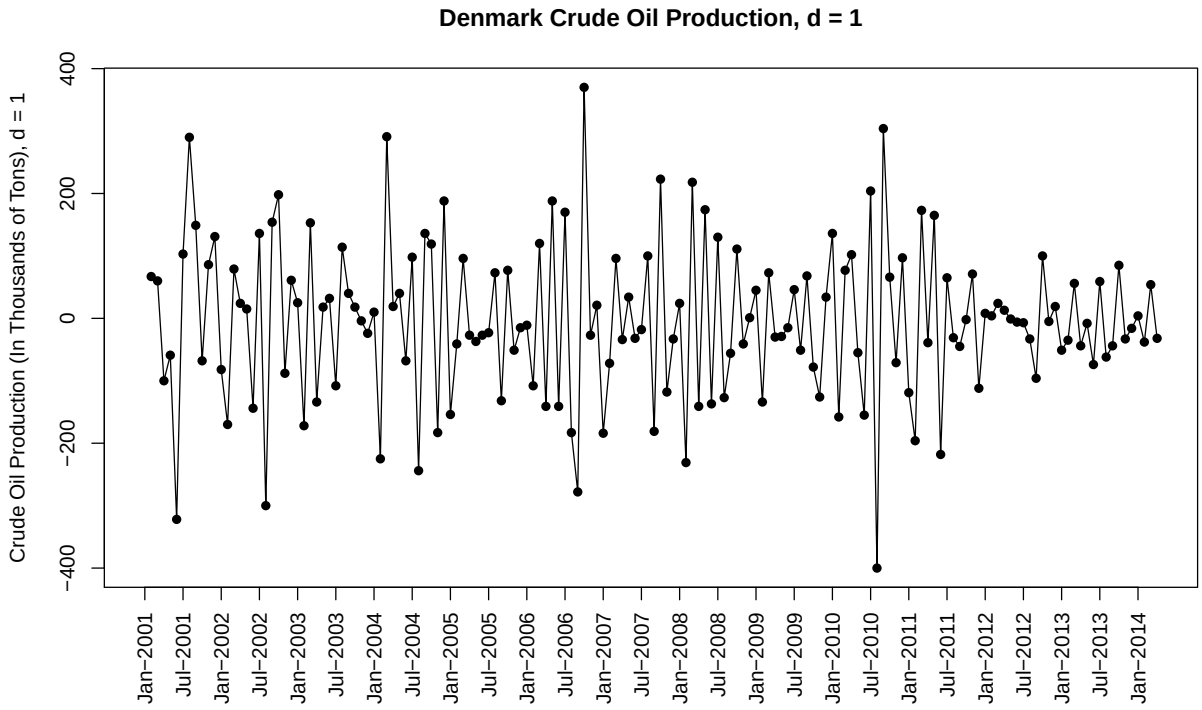

Figure 3: Plot of Denmark crude oil production, d = 1.

Figure 3 shows the plot of the first difference of the data. We can see that the data now fluctuates around a constant level: that is, it appears to be stationary.

5)

Calculate the sample Autocorrelation Function, variogram and Ljung-Box Statistics (Q_{LB}) for the differenced data. Present your results in a table.

Answer

```
diff.oil <- diff(B.22$Oil.Production)

# Sample ACF
acf.diff.oil <- acf(diff.oil, lag.max=20, plot=F)

# Variogram
vario.diff.oil <- variogram.function(diff.oil)

# Ljung-Box statistic
ljung.diff.oil <- ljung.function(diff.oil, acf.diff.oil)

# Create a table
dt <- cbind(Lag=1:20, 'Sample ACF'=acf.diff.oil$acf[2:21], Variogram=vario.diff.oil,
            'Ljung-Box Statistic'=ljung.diff.oil)
kable(dt, align='c', caption='Sample ACF, Variogram, and Ljung-Box Statistic for
      Denmark Crude Oil Production (First 20 Lags, d = 1)')
```

Table 2: Sample ACF, Variogram, and Ljung-Box Statistic for Denmark Crude Oil Production (First 20 Lags, d = 1)

Lag	Sample ACF	Variogram	Ljung-Box Statistic
1	-0.4802516	1.0000000	37.36832
2	0.0565774	0.6400771	37.89024
3	-0.0023284	0.6831280	37.89113
4	-0.1600871	0.7958788	42.12370
5	0.1006687	0.6058270	43.80828
6	-0.0097453	0.6852283	43.82417
7	0.0363386	0.6438573	44.04656
8	-0.0042469	0.6732135	44.04962
9	-0.0394991	0.7017839	44.31588
10	0.0776380	0.6207176	45.35146
11	-0.1770721	0.8053262	50.77474
12	0.2140139	0.5256698	58.75082
13	-0.1147956	0.7657684	61.06139
14	0.0491749	0.6486096	61.48831
15	0.0434937	0.6570635	61.82460
16	-0.0869885	0.7587938	63.17920
17	0.0462264	0.6609258	63.56442
18	-0.0468449	0.7330806	63.96283
19	0.0641321	0.6384838	64.71488
20	-0.0220612	0.7041217	64.80451

6)

Plot the sample ACF and variogram for the differenced data. Interpret these graphs.

Answer

```
par(mfrow=c(1,2))
# ACF plot
acf(diff.oil, lag.max=20, main='Sample ACF of Denmark Crude Oil Production \nd = 1')

# Variogram plot
plot(vario.diff.oil, pch=16, main='Variogram of Denmark Crude Oil Production \nd = 1',
     ylab='Variogram', xlab='Lag')
```

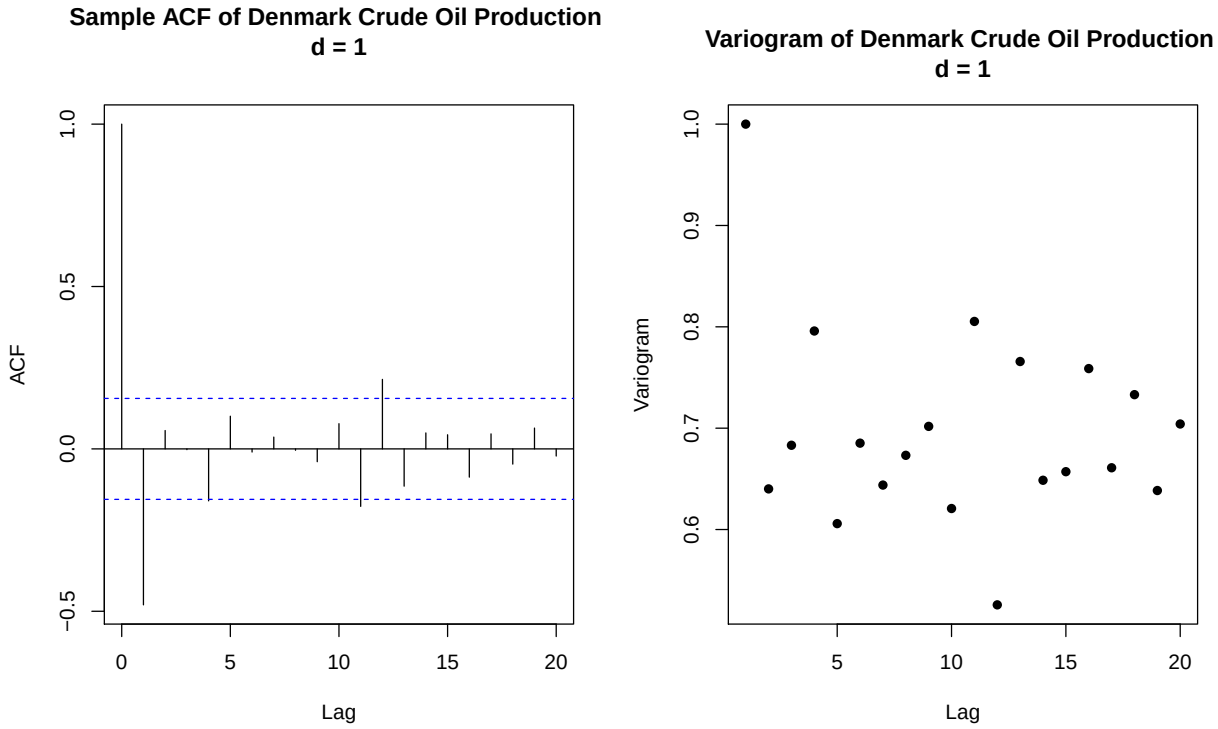

Figure 4: Sample ACF and variogram plots of Denmark crude oil production, d = 1.

Figure 4 shows the sample ACF and variogram plots of the first difference of the data. In the ACF plot, there are a couple of autocorrelations that cross the blue dashed lines, which means these autocorrelations are significantly different from 0. But other than that, the ACF fluctuates about 0, and the variogram converges to a generally stable level, indicating a stationary time series.

7)

What impact did differencing have?

Answer

Differencing removed the trend in the time series and made it stationary.

2.

Table B.25 contains data from the National Highway Traffic Safety Administration on motor vehicle fatalities from 1966 to 2012, along with several other variables. These data are used by a variety of governmental and industry groups, as well as research organizations.

1)

Plot the fatalities data. Comment on the graph.

Answer

```
# Load the B.25 data
B.25 <- read.csv('C:\\Users\\George\\Desktop\\TimeSeriesAnalysis\\B.25.csv')

# Fatalities plot
nrb <- length(B.25$Year)
tt <- 1:nrb

plot(tt, B.25$Fatalities, type='l', main='Motor Vehicle Fatalities',
     ylab='Fatalities', xlab='Year', xaxt='n')
axis(1, seq(1, nrb, 2), labels=B.25$Year[seq(1, nrb, 2)], las=2)
points(tt, B.25$Fatalities, pch=16)
```

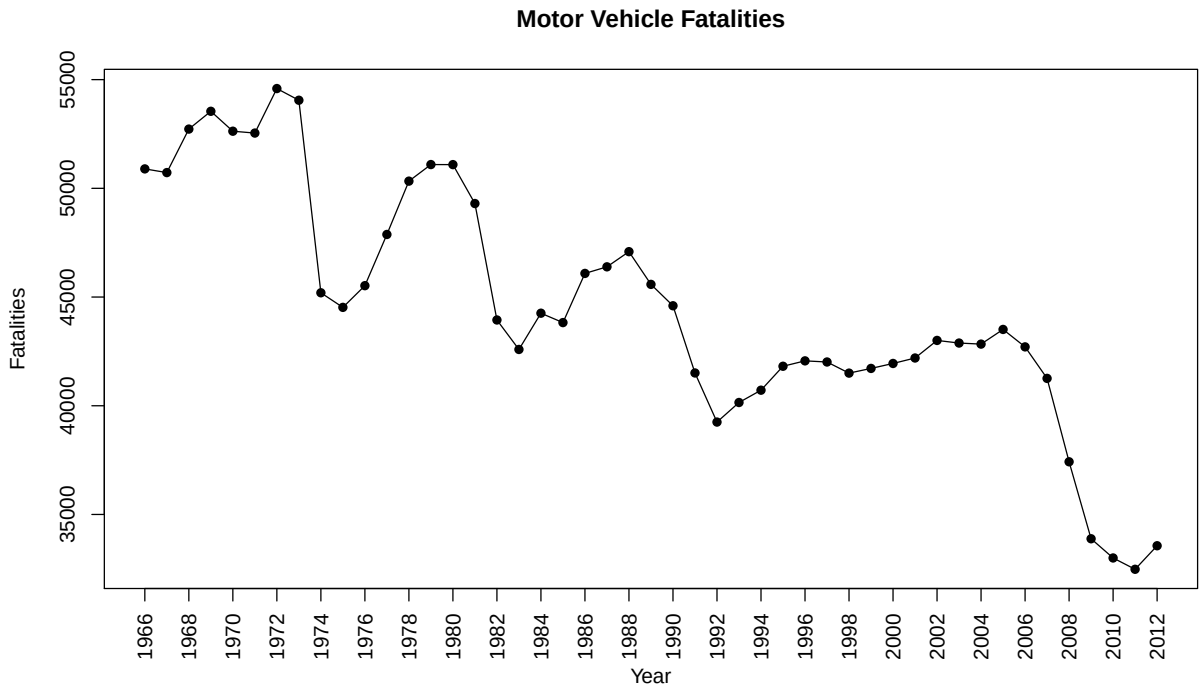

Figure 5: Plot of motor vehicle fatalities.

Figure 5 shows the plot of motor vehicle fatalities. We can see that the number of fatalities fluctuates but exhibits an overall decreasing trend.

2)

Construct a scatter plot of fatalities versus number of licensed drivers. Comment on the apparent relationship between these two factors.

Answer

```
# Scatter plot of fatalities vs. number of licensed drivers
plot(B.25$Drivers, B.25$Fatalities, main='Motor Vehicle Fatalities vs. Number of Licensed Drivers',
     ylab='Fatalities', xlab='Licensed Drivers (In Thousands)', pch=16)
```

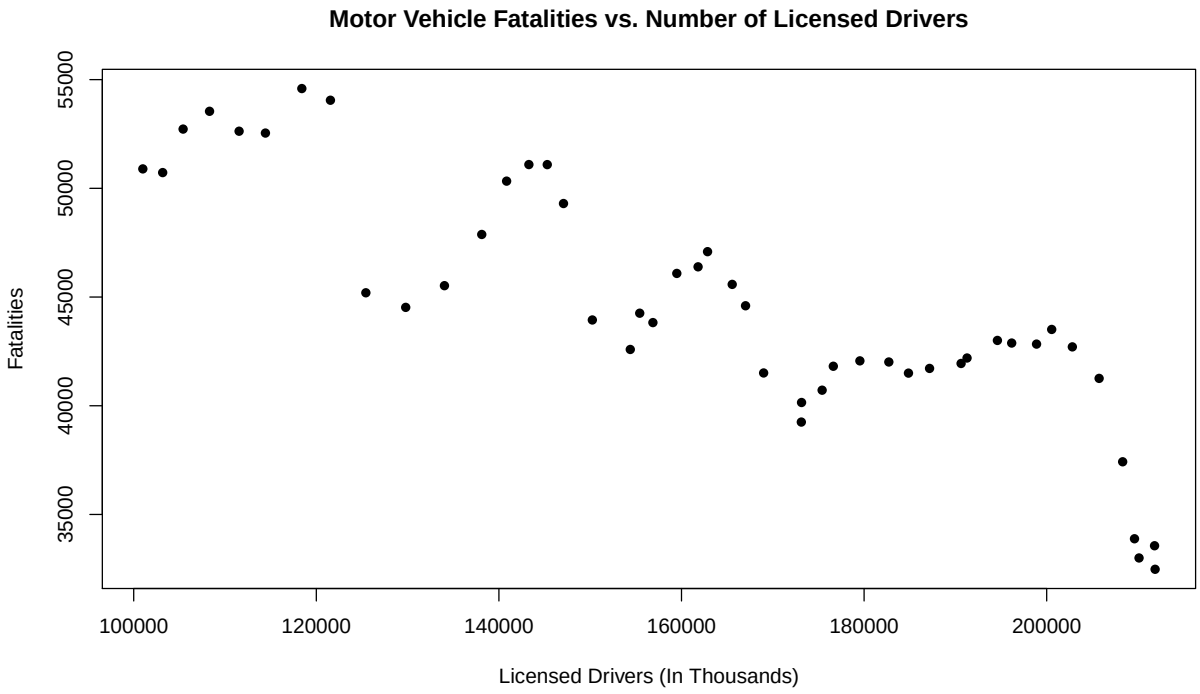

Figure 6: Scatter plot of motor vehicle fatalities vs. number of licensed drivers.

Figure 6 shows the scatter plot of motor vehicle fatalities vs. number of licensed drivers. It appears that the two factors have a negative relationship.

3)

Fit a simple linear regression model to the fatalities data, using the number of licensed drivers as the predictor variable. Discuss the summary statistics from this model.

Answer

We fit a simple linear regression model with fatalities as the response and the number of licensed drivers as the predictor. The model summary is below:

```
# Fitted simple linear regression model
model <- lm(Fatalities ~ Drivers, data=B.25)
# Model summary
summary(model)
```

```
##
## Call:
## lm(formula = Fatalities ~ Drivers, data = B.25)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5000.9 -2382.8   342.6  2537.3  4402.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.792e+04  2.081e+03   32.63  < 2e-16 ***
## Drivers     -1.437e-01  1.252e-02  -11.47 5.91e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2843 on 45 degrees of freedom
## Multiple R-squared:  0.7452, Adjusted R-squared:  0.7396
## F-statistic: 131.6 on 1 and 45 DF,  p-value: 5.909e-15
```

```
# Model coefficients
model$coefficients
```

```
##   (Intercept)       Drivers
## 67921.7776701    -0.1436785
```

```
# Model AIC
AIC(model)
```

```
## [1] 884.8693
```

Using the estimated coefficients, the fitted simple linear regression model is

$$\hat{y} = 67921.778 - 0.144x$$

where $\hat{y}$ = predicted value of fatalities and x = number of licensed drivers.

We can see that the coefficient for drivers is significant (p-value = 5.91e-15) at an alpha level of 0.05. The AIC of the model is 884.8693. The multiple and adjusted R^2 are 0.7452 and 0.7396 respectively, so the model explains about 74% of the variability in fatalities.

4)

Analyze the residuals from the model in part 3). Discuss the adequacy of the fitted model. (Limit the pages to 4)

Answer

Residual Plots and Normality Test

To analyze the residuals and check the adequacy of the model, we can look at the residual plots of the model.

```
par(mfrow=c(2,2), oma=c(0,0,0,0))
# Residual plots
qqnorm(model$res, datax=TRUE, pch=16,
       xlab='Residual', main='Normal Q-Q Plot of the Residuals')
qqline(model$res, datax=TRUE)
plot(model$fit, model$res, pch=16, xlab='Fitted Value', ylab='Residual',
     main='Residuals vs. Fitted Values') ; abline(h=0)
hist(model$res, col='gray', xlab='Residual', main='Histogram of the Residuals')
plot(model$res, type='l', xlab='Observation Order', ylab='Residual',
     main='Residuals vs. Order of the Data')
points(model$res, pch=16, cex=0.8) ; abline(h=0)
```

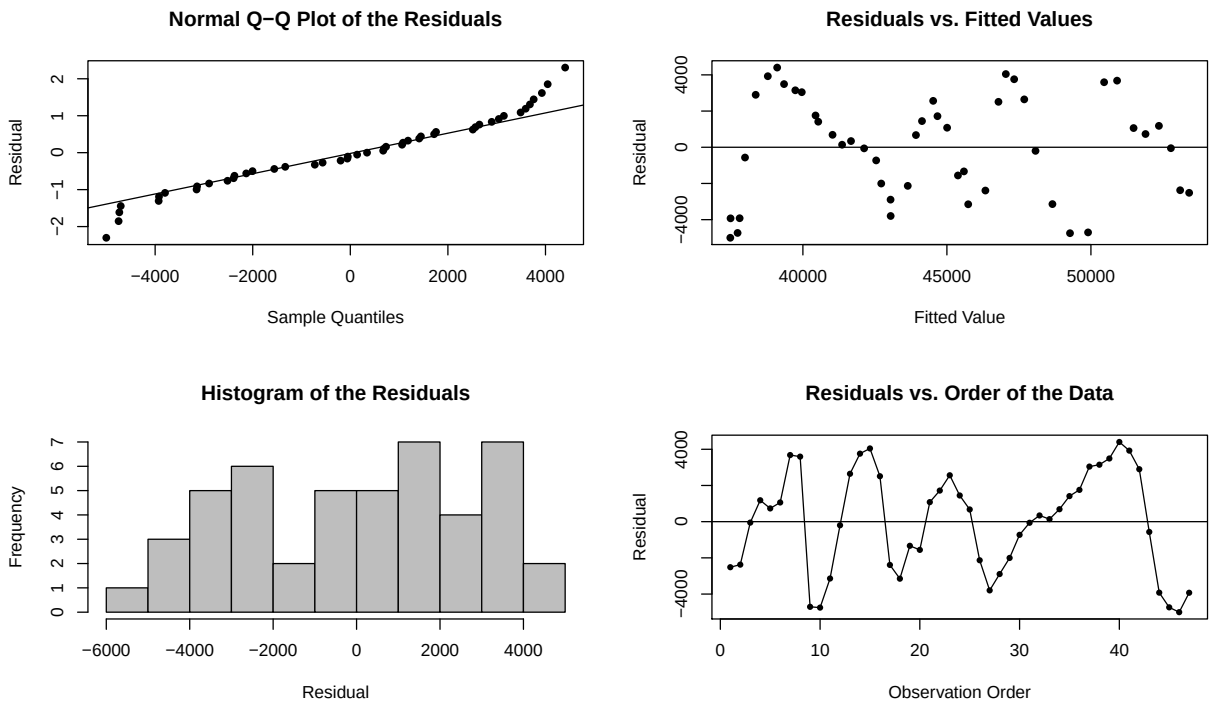

Figure 7: Residual plots for the fatalities model.

Figure 7 contains the residual plots for the fatalities model. The normal q-q plot of the residuals shows that most of the residuals follow a normal distribution except at the tails. The histogram of the residuals does not look normal. The residuals vs. fitted values plot indicates nonconstant variance. The residuals vs. order of the data plot indicates that the residuals are positively autocorrelated.

We can also use the Shapiro-Wilk test to test the normality of the residuals.

```
# Shapiro-Wilk test
shapiro.test(model$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$res
## W = 0.94735, p-value = 0.03413
```

The null hypothesis is that the residuals are normally distributed. Since the p-value of 0.03413 is less than an alpha level of 0.05, we reject the null hypothesis and conclude that the residuals are not normally distributed.

Prediction R^2

To get an idea for how well the model will predict new data, we can calculate the prediction R^2

$$R^2_{Prediction} = 1 - \frac{PRESS}{SS_T}$$

where PRESS = prediction error sum of squares and SS_T = total sum of squares.

The prediction R^2 for the fatalities model is:

```
# Prediction R-squared
anova <- anova(model)
SST <- sum(anova$`Sum Sq`)
PRESS <- sum(rstandard(model, type='pred')^2)
1 - PRESS/SST
```

```
## [1] 0.7194071
```

Thus, this model accounts for about 72% of the variability in new data, which is pretty good.

Studentized Residuals, Hat Diagonals, and Cook's Distance

In addition to residual plots, we can examine the studentized residuals, hat diagonals, and Cook's distance.

```
# Vectors of residuals and other diagnostics
a <- round(residuals(model), 5)
b <- round(rstandard(model), 5)
c <- round(rstudent(model), 5)
d <- round(hatvalues(model), 5)
e <- round(cooks.distance(model), 5)

# Combine above vectors into a table
table <- cbind(Year=1966:2012, Residuals=a, 'Studentized Residuals'=b,
```

```
              'R-Student'=c, 'h[i,i]'=d, "Cook's Distance"=e)
kable(table, align='c',
      caption='Residuals and Other Diagnostics for the Fatalities Model')
```

Table 3: Residuals and Other Diagnostics for the Fatalities Model

Year	Residuals	Studentized Residuals	R-Student	h[i,i]	Cook's Distance
1966	-2516.53530	-0.93090	-0.92949	0.09559	0.04580
1967	-2374.17822	-0.87576	-0.87345	0.09046	0.03814
1968	-51.62571	-0.01899	-0.01878	0.08537	0.00002
1969	1182.46726	0.43347	0.42953	0.07908	0.00807
1970	731.55461	0.26721	0.26444	0.07242	0.00279
1971	1060.77976	0.38630	0.38262	0.06684	0.00534
1972	3680.76966	1.33529	1.34733	0.05965	0.05655
1973	3593.77076	1.30013	1.31045	0.05443	0.04865
1974	-4704.61293	-1.69669	-1.73411	0.04850	0.07337
1975	-4748.59991	-1.70720	-1.74560	0.04252	0.06472
1976	-3140.68462	-1.12613	-1.12958	0.03742	0.02465
1977	-198.75790	-0.07111	-0.07032	0.03317	0.00009
1978	2645.47869	0.94527	0.94413	0.03070	0.01415
1979	3758.05426	1.34145	1.35381	0.02873	0.02661
1980	4044.99175	1.44280	1.46087	0.02728	0.02919
1981	2510.73950	0.89502	0.89300	0.02612	0.01074
1982	-2391.38008	-0.85171	-0.84906	0.02438	0.00906
1983	-3150.39586	-1.12106	-1.12434	0.02268	0.01458
1984	-1333.68860	-0.47451	-0.47039	0.02236	0.00257
1985	-1558.21683	-0.55429	-0.54998	0.02198	0.00345
1986	1079.93351	0.38406	0.38039	0.02150	0.00162
1987	1717.70444	0.61081	0.60650	0.02130	0.00406
1988	2563.84274	0.91168	0.90994	0.02128	0.00903
1989	1446.77472	0.51450	0.51025	0.02142	0.00290
1990	673.68903	0.23960	0.23707	0.02161	0.00063
1991	-2132.82752	-0.75870	-0.75507	0.02200	0.00648
1992	-3797.43526	-1.35175	-1.36464	0.02331	0.02181
1993	-2893.98698	-1.03016	-1.03088	0.02332	0.01267
1994	-2004.13561	-0.71377	-0.70982	0.02432	0.00635
1995	-727.12943	-0.25905	-0.25634	0.02495	0.00086
1996	-60.88129	-0.02171	-0.02147	0.02666	0.00001
1997	342.57960	0.12230	0.12095	0.02891	0.00022
1998	139.77576	0.04994	0.04939	0.03065	0.00004
1999	687.52944	0.24592	0.24334	0.03273	0.00102
2000	1411.93870	0.50595	0.50173	0.03622	0.00481
2001	1756.47341	0.62964	0.62537	0.03693	0.00760
2002	3043.34814	1.09315	1.09558	0.04081	0.02542
2003	3147.06134	1.13157	1.13520	0.04278	0.02861
2004	3490.29792	1.25740	1.26578	0.04644	0.03850
2005	4402.80425	1.58811	1.61631	0.04882	0.06472
2006	3925.66137	1.41854	1.43515	0.05222	0.05544
2007	2897.92677	1.04978	1.05100	0.05693	0.03326
2008	-567.52635	-0.20607	-0.20386	0.06135	0.00139
2009	-3921.17532	-1.42556	-1.44258	0.06367	0.06910
2010	-4733.76710	-1.72181	-1.76159	0.06458	0.10234
2011	-5000.89292	-1.82218	-1.87221	0.06787	0.12087

Year	Residuals	Studentized Residuals	R-Student	h[i,i]	Cook's Distance
2012	-3927.51363	-1.43098	-1.44833	0.06775	0.07441

Table 3 contains the studentized residuals, hat diagonals, and Cook's distance for the fatalities model.

Absolute values of the studentized residuals that are greater than three or four indicate potential unusual values or outliers. The largest studentized residual is -1.82218 (observation from 2011); similarly, the largest R-student is -1.87221 (observation from 2011). So there does not appear to be any unusual values or outliers.

Hat diagonals that are greater than twice their average value $2p/n$ (p = number of parameters and n = number of observations) indicate high-leverage observations. There are $p = 2$ parameters in the fatalities model and $n = 47$ observations, so $2p/n = 2(2)/47 = 0.0851$. Hat diagonals that exceed 0.0851 correspond to the first three observations in the data. However, hat diagonals will identify points that are potentially influential due to their location in the predictor variable space (observations with extreme predictor values are high leverage). We want to consider both the location of the point and the response variable in measuring influence. This can be accomplished by examining Cook's distance.

Observations with Cook's distance greater than one are considered influential. The largest Cook's distance is 0.1209 (observation from 2011), so there does not appear to be any influential observations.

Conclusion

Since we found that the residuals are not normally distributed, have nonconstant variance, and are positively autocorrelated, the fitted model is not adequate.

5)

Calculate the Durbin-Watson test statistic for the model in part 3). Is there evidence of autocorrelation in the residuals? Is a time series regression model more appropriate than an OLS model for these data?

Answer

We will use the "dwt" function from the **car** package to calculate the Durbin-Watson test statistic for the model. We will test for positive autocorrelation with a one-sided test.

```
# Durbin-Watson test
library(car)
dwt(model, alternative='positive')
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.7110256     0.5181106       0
##  Alternative hypothesis: rho > 0
```

The Durbin-Watson test statistic is 0.518. The null hypothesis is that the errors are uncorrelated. Since the p-value is extremely small, we reject the null hypothesis and conclude that the errors are positively autocorrelated. Thus, a time series regression model is more appropriate than an OLS model for these data.

6)

There are several candidate predictors that could be added to the model. Use stepwise regression to find an appropriate model.

Answer

We perform stepwise "mixed" selection with fatalities as the response and resident population, number of licensed drivers, number of registered motor vehicles, vehicle miles traveled, and annual unemployment rate as the predictors.

```
# Fitted full model
model.full <- lm(Fatalities ~ Population + Drivers + Vehicles + Miles.Traveled +
                   Unemployment.Rate, data=B.25)
# Stepwise selection
step.both <- step(model.full, direction='both')
```

```
## Start:  AIC=735.94
## Fatalities ~ Population + Drivers + Vehicles + Miles.Traveled +
##     Unemployment.Rate
##
##                     Df Sum of Sq       RSS    AIC
## - Miles.Traveled     1    153305 230048768 733.97
## - Drivers            1    423545 230319008 734.03
## - Vehicles           1   2775566 232671028 734.50
## - Population         1   8572377 238467839 735.66
## <none>                           229895463 735.94
## - Unemployment.Rate  1  38421336 268316799 741.20
##
## Step:  AIC=733.97
## Fatalities ~ Population + Drivers + Vehicles + Unemployment.Rate
##
##                     Df Sum of Sq       RSS    AIC
## - Drivers            1   2860632 232909399 732.55
## - Vehicles           1   3478693 233527461 732.68
## <none>                           230048768 733.97
## - Population         1  13688139 243736907 734.69
## + Miles.Traveled     1    153305 229895463 735.94
## - Unemployment.Rate  1 114634503 344683271 750.98
##
## Step:  AIC=732.55
## Fatalities ~ Population + Vehicles + Unemployment.Rate
##
##                     Df Sum of Sq       RSS    AIC
## - Vehicles           1   1339976 234249376 730.82
## <none>                           232909399 732.55
## - Population         1  13124754 246034154 733.13
## + Drivers            1   2860632 230048768 733.97
## + Miles.Traveled     1   2590392 230319008 734.03
## - Unemployment.Rate  1 148250130 381159530 753.70
##
## Step:  AIC=730.82
## Fatalities ~ Population + Unemployment.Rate
```

```
##
##                     Df Sum of Sq        RSS    AIC
## <none>                            234249376 730.82
## + Miles.Traveled     1   1508526  232740850 732.52
## + Vehicles           1   1339976  232909399 732.55
## + Drivers            1    721914  233527461 732.68
## - Unemployment.Rate  1 148967669  383217045 751.96
## - Population         1 878560434 1112809809 802.06
```

The stepwise "mixed" selection is a combination of forward selection and backward elimination. It removes or adds the predictor that decreases AIC the most until no predictors remain that can be removed or added to decrease AIC. As we can see, stepwise selection sequentially removed vehicle miles traveled (*Miles.Traveled*), number of licensed drivers (*Drivers*), and number of registered motor vehicles (*Vehicles*). Thus, the best model contains resident population (*Population*) and annual unemployment rate (*Unemployment.Rate*).

The summary of this new model is below:

```
# Fitted new model
model2 <- lm(Fatalities ~ Population + Unemployment.Rate, data=B.25)
# Model summary
summary(model2)
```

```
##
## Call:
## lm(formula = Fatalities ~ Population + Unemployment.Rate, data = B.25)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4593.6 -1676.7  -424.8  1724.8  4612.0
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.259e+04  2.566e+03   32.18  < 2e-16 ***
## Population        -1.247e-01  9.709e-03  -12.85  < 2e-16 ***
## Unemployment.Rate -1.095e+03  2.071e+02   -5.29 3.68e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2307 on 44 degrees of freedom
## Multiple R-squared:  0.8359, Adjusted R-squared:  0.8284
## F-statistic:   112 on 2 and 44 DF,  p-value: < 2.2e-16
```

```
# Model coefficients
model2$coefficients
```

```
##       (Intercept)        Population Unemployment.Rate
##     82594.6101191        -0.1247182     -1095.2794261
```

```
# Model AIC
AIC(model2)
```

```
## [1] 866.2024
```

Using the estimated coefficients, the new fitted multiple linear regression model is

$$\hat{y} = 82594.610 - 0.125x_1 - 1095.279x_2$$

where $\hat{y}$ = predicted value of fatalities, x_1 = resident population, and x_2 = annual unemployment rate.

We can see that the coefficients for both predictors are significant at an alpha level of 0.05. The AIC of the model is 866.2024, which is smaller than that of the original model (884.8693). The adjusted R^2 is 0.8284, which is greater than that of the original model (0.7396). Thus, the new model is a better fit than the original model.

7)

Calculate the Durbin-Watson test statistic for the model in part 6). Is there evidence of autocorrelation in the residuals? Is a time series regression model more appropriate than an OLS model for these data?

Answer

We will use the "dwt" function from the **car** package to calculate the Durbin-Watson test statistic for the model. We will test for positive autocorrelation with a one-sided test.

```
# Durbin-Watson test
dwt(model2, alternative='positive')
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.6936623     0.5692754       0
##  Alternative hypothesis: rho > 0
```

The Durbin-Watson test statistic is 0.569. The null hypothesis is that the errors are uncorrelated. Since the p-value is extremely small, we reject the null hypothesis and conclude that the errors are positively autocorrelated. Thus, a time series regression model is more appropriate than an OLS model for these data (Cochrane-Orcutt method is applied in part 8).

8)

Analyze the residuals from the model that you obtained in part 6). Discuss the adequacy of the fitted model. (Limit the pages to 4)

Answer

Residual Plots and Normality Test

To analyze the residuals and check the adequacy of the new model, we can look at the residual plots of the model.

```
par(mfrow=c(2,2), oma=c(0,0,0,0))
# Residual plots
qqnorm(model2$res, datax=TRUE, pch=16,
       xlab='Residual', main='Normal Q-Q Plot of the Residuals')
qqline(model2$res, datax=TRUE)
```

```r
plot(model2$fit, model2$res, pch=16, xlab='Fitted Value', ylab='Residual',
     main='Residuals vs. Fitted Values') ; abline(h=0)
hist(model2$res, col='gray', xlab='Residual', main='Histogram of the Residuals')
plot(model2$res, type='l', xlab='Observation Order', ylab='Residual',
     main='Residuals vs. Order of the Data')
points(model2$res, pch=16, cex=0.8) ; abline(h=0)
```

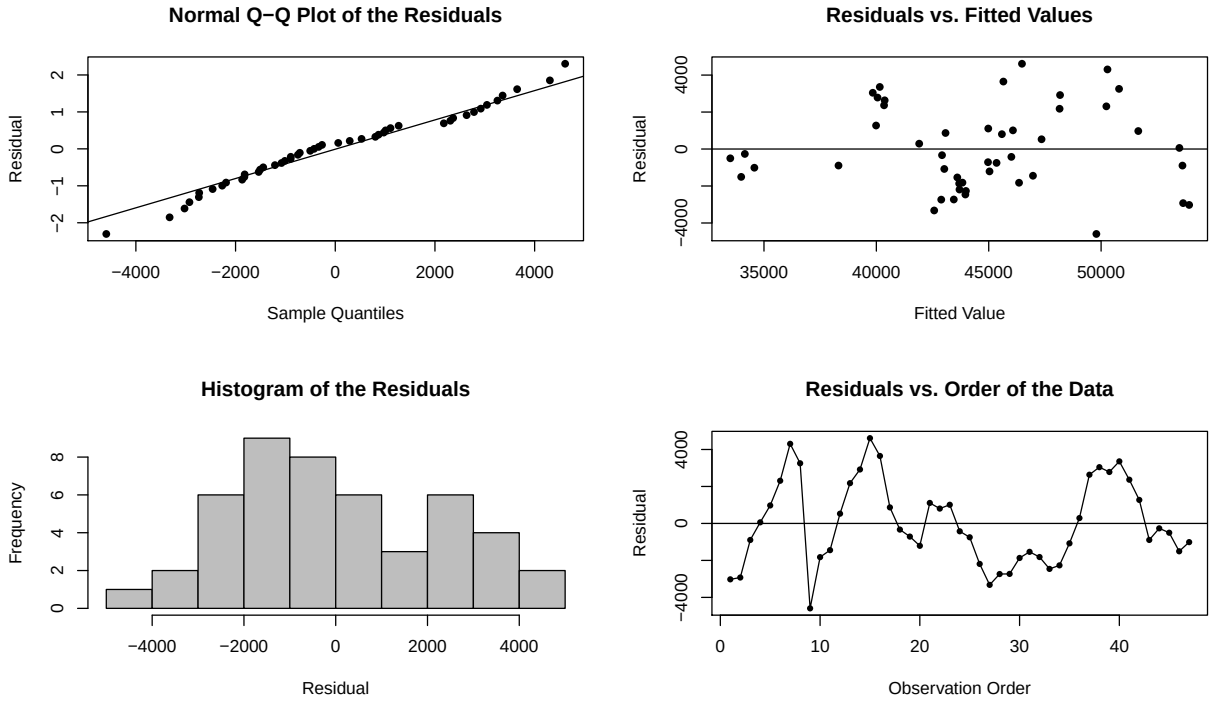

Figure 8: Residual plots for the new fatalities model.

Figure 8 contains the residual plots for the new fatalities model. The normal q-q plot of the residuals shows that most of the residuals follow a normal distribution except at the tails. The histogram of the residuals looks mostly normal as well. The residuals vs. fitted values plot does not show any obvious patterns in the residuals, so the equal variance assumption does not appear to be violated. The residuals vs. order of the data plot indicates that the residuals are positively autocorrelated.

We can also use the Shapiro-Wilk test to test the normality of the residuals.

```r
# Shapiro-Wilk test
shapiro.test(model2$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$res
## W = 0.97022, p-value = 0.27
```

The null hypothesis is that the residuals are normally distributed. Since the p-value of 0.27 is greater than an alpha level of 0.05, we fail to reject the null hypothesis and conclude that there is insufficient evidence that the residuals are not normally distributed.

Prediction R^2

To get an idea for how well the model will predict new data, we can calculate the prediction R^2

$$R^2_{Prediction} = 1 - \frac{PRESS}{SS_T}$$

where PRESS = prediction error sum of squares and SS_T = total sum of squares.

The prediction R^2 for the new fatalities model is:

```
# Prediction R-squared
anova <- anova(model2)
SST <- sum(anova$`Sum Sq`)
PRESS <- sum(rstandard(model2, type='pred')^2)
1 - PRESS/SST
```

```
## [1] 0.8163319
```

Thus, this model accounts for about 81.6% of the variability in new data, which is pretty good.

Studentized Residuals, Hat Diagonals, and Cook's Distance

In addition to residual plots, we can examine the studentized residuals, hat diagonals, and Cook's distance.

```
# Vectors of residuals and other diagnostics
a <- round(residuals(model2), 5)
b <- round(rstandard(model2), 5)
c <- round(rstudent(model2), 5)
d <- round(hatvalues(model2), 5)
e <- round(cooks.distance(model2), 5)

# Combine above vectors into a table
table <- cbind(Year=1966:2012, Residuals=a, 'Studentized Residuals'=b,
               'R-Student'=c, 'h[i,i]'=d, "Cook's Distance"=e)
kable(table, align='c',
      caption='Residuals and Other Diagnostics for the New Fatalities Model')
```

Table 4: Residuals and Other Diagnostics for the New Fatalities Model

Year	Residuals	Studentized Residuals	R-Student	h[i,i]	Cook's Distance
1966	-3023.93874	-1.38265	-1.39755	0.10155	0.07202
1967	-2925.54517	-1.33513	-1.34745	0.09813	0.06465
1968	-894.91296	-0.40921	-0.40531	0.10165	0.00632
1969	59.37867	0.02716	0.02685	0.10240	0.00003
1970	972.97559	0.43586	0.43181	0.06396	0.00433
1971	2308.64481	1.02874	1.02943	0.05402	0.02015
1972	4305.80616	1.91561	1.97800	0.05100	0.06573
1973	3253.16830	1.45015	1.46911	0.05472	0.04057
1974	-4593.55920	-2.03802	-2.11713	0.04577	0.06640
1975	-1823.97100	-0.83279	-0.82984	0.09897	0.02539
1976	-1445.02561	-0.64841	-0.64409	0.06713	0.01008

Year	Residuals	Studentized Residuals	R-Student	h[i,i]	Cook's Distance
1977	527.68565	0.23459	0.23205	0.04958	0.00096
1978	2177.99512	0.96125	0.96040	0.03569	0.01140
1979	2919.46525	1.28684	1.29677	0.03322	0.01897
1980	4611.96700	2.04203	2.12173	0.04187	0.06075
1981	3649.10020	1.62337	1.65514	0.05090	0.04711
1982	867.31760	0.40509	0.40121	0.13896	0.00883
1983	-332.81001	-0.15470	-0.15297	0.13068	0.00120
1984	-711.34470	-0.31515	-0.31190	0.04300	0.00149
1985	-1210.14503	-0.53396	-0.52958	0.03521	0.00347
1986	1108.30159	0.48785	0.48359	0.03058	0.00250
1987	803.97049	0.35248	0.34894	0.02279	0.00097
1988	1009.90212	0.44324	0.43916	0.02489	0.00167
1989	-424.80754	-0.18664	-0.18458	0.02693	0.00032
1990	-749.34407	-0.32867	-0.32531	0.02363	0.00087
1991	-2190.64152	-0.96133	-0.96049	0.02463	0.00778
1992	-3323.13166	-1.46654	-1.48657	0.03555	0.02642
1993	-2736.95010	-1.20186	-1.20812	0.02591	0.01281
1994	-2729.89054	-1.19685	-1.20292	0.02280	0.01114
1995	-1867.72799	-0.82058	-0.81748	0.02690	0.00620
1996	-1536.21752	-0.67623	-0.67200	0.03063	0.00482
1997	-1817.20223	-0.80440	-0.80112	0.04141	0.00932
1998	-2460.00835	-1.09601	-1.09858	0.05372	0.02273
1999	-2267.90562	-1.01680	-1.01720	0.06555	0.02418
2000	-1077.75542	-0.48810	-0.48383	0.08421	0.00730
2001	290.02417	0.12998	0.12852	0.06479	0.00039
2002	2635.08308	1.17006	1.17512	0.04733	0.02267
2003	3042.81426	1.35223	1.36545	0.04891	0.03134
2004	2783.53953	1.24349	1.25146	0.05879	0.03219
2005	3357.66352	1.50977	1.53274	0.07097	0.05805
2006	2365.09202	1.07429	1.07622	0.08961	0.03787
2007	1271.66361	0.57937	0.57495	0.09510	0.01176
2008	-892.93287	-0.40218	-0.39832	0.07409	0.00431
2009	-265.45954	-0.12351	-0.12212	0.13236	0.00078
2010	-502.34543	-0.23615	-0.23360	0.15004	0.00328
2011	-1506.92845	-0.69796	-0.69383	0.12442	0.02307
2012	-1011.05746	-0.46319	-0.45902	0.10503	0.00839

Table 4 contains the studentized residuals, hat diagonals, and Cook's distance for the new fatalities model.

Absolute values of the studentized residuals that are greater than three or four indicate potential unusual values or outliers. The largest studentized residual is 2.04203 (observation from 1980); similarly, the largest R-student is 2.12173 (observation from 1980). So there does not appear to be any unusual values or outliers.

Hat diagonals that are greater than twice their average value $2p/n$ (p = number of parameters and n = number of observations) indicate high-leverage observations. There are $p = 3$ parameters in the new fatalities model and $n = 47$ observations, so $2p/n = 2(3)/47 = 0.1277$. There are four observations with hat diagonals that exceed 0.1277. However, hat diagonals will identify points that are potentially influential due to their location in the predictor variable space (observations with extreme predictor values are high leverage). We want to consider both the location of the point and the response variable in measuring influence. This can be accomplished by examining Cook's distance.

Observations with Cook's distance greater than one are considered influential. The largest Cook's distance is 0.072 (observation from 1966), so there does not appear to be any influential observations.

Conclusion

Since we found that the residuals are positively autocorrelated, the fitted model is not adequate.

Cochrane-Orcutt Method

The autocorrelation problem may be fixed by applying the Cochrane-Orcutt method, which will adjust (transform) the parameter estimates and their standard errors. The model summary is below:

```
# Cochrane-Orcutt method
library(orcutt)
model2.cochrane <- cochrane.orcutt(model2)
summary(model2.cochrane)
```

```
## Call:
## lm(formula = Fatalities ~ Population + Unemployment.Rate, data = B.25)
##
##                      Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)        8.5962e+04  5.8195e+03  14.771 < 2.2e-16 ***
## Population        -1.3526e-01  2.1901e-02  -6.176 2.029e-07 ***
## Unemployment.Rate -1.1702e+03  2.4262e+02  -4.823 1.803e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1605.843 on 43 degrees of freedom
## Multiple R-squared:  0.604 ,  Adjusted R-squared:  0.5855
## F-statistic: 32.8 on 2 and 43 DF,  p-value: < 2.246e-09
##
## Durbin-Watson statistic
## (original):    0.56928 , p-value: 3.475e-10
## (transformed): 1.67701 , p-value: 9.142e-02
```

```
# Model coefficients
model2.cochrane$coefficients
```

```
##       (Intercept)        Population Unemployment.Rate
##     85961.9155747        -0.1352576     -1170.1552535
```

We can see that the Durbin Watson statistic for the transformed model is 1.67701. Since the p-value is $0.09142 > 0.05$, we fail to reject the null hypothesis and conclude that there is insufficient evidence of autocorrelation in the errors.

Using the estimated coefficients from the model summary, the fitted time series regression model is

$$\hat{y_t}' = 85961.916 - 0.135x'_{1,t} - 1170.155x'_{2,t}$$

where $\hat{y_t}'$ = predicted value of fatalities (transformed) at time t, $x'_{1,t}$ = resident population (transformed) at time t, and $x'_{2,t}$ = annual unemployment rate (transformed) at time t.