**A Statistical Model to Predict**

**Cigarette Consumption**

**Caifeng Ye (Celine)**

**X03386581**

**Prof. Manuel G. Russon**

**Predicted Analytics & Business Forecast**

**DS633**

**I. Introduction**

The use of cigars is associated with many economic factors. Data on cigar sales are examined to predict cigar consumption for a specific state based on cigarette price, state population, consumer price index and income. The conclusions of the research may draw attention to policy makers and cigarette companies.

**II. Previous Research**

There is no previous research has been done at present.

**III. Methodology**

The research is based on a time-series dataset Cigar. It contains information about yearly per capita cigarette sales for 46 geographic regions in the United States over the period 1963–1992. The variables *price, pop, pop16, cpi, ndi, pimin, sales* denote price per pack of cigarettes, population, population above the age of 16, consumer price index with (1983=100), per capita disposable income, minimum price in adjoining states per pack of cigarettes and per capita cigarette sales. In the research, graphical techniques (histograms, time-series plots and scatter-plots) and descriptive statistics, correlation and multiple log-linear regression were applied. All the analyses are conducted on R statistical software.

The functional specification is given below:

- + + + +

**Eqn. 1** Functional Specificationsales = f (price, pop16, cpi, ndi, year)

**Eqn. 2** Population Regressionsales = α + β1\*price + β2\*pop16 + β3\*cpi + β4\*ndi+ β5\*year + Ɛ

**Eqn. 3** Sample Regressionsales = a + b1\*price + b2\*pop16 + b3\*pi + b4\*ndi + b5\*year + e

The research assumes that cpi, per capita disposable income, population above the age of 16 and year have positive impact on sales while cigar price have negative impact.

**IV. Results**

**a. Histogram & Time-Series Plot**

The following are the histograms of all variables(Fig.1 to Fig. 9). The graphs show that most variables are skewed to the right except sales.

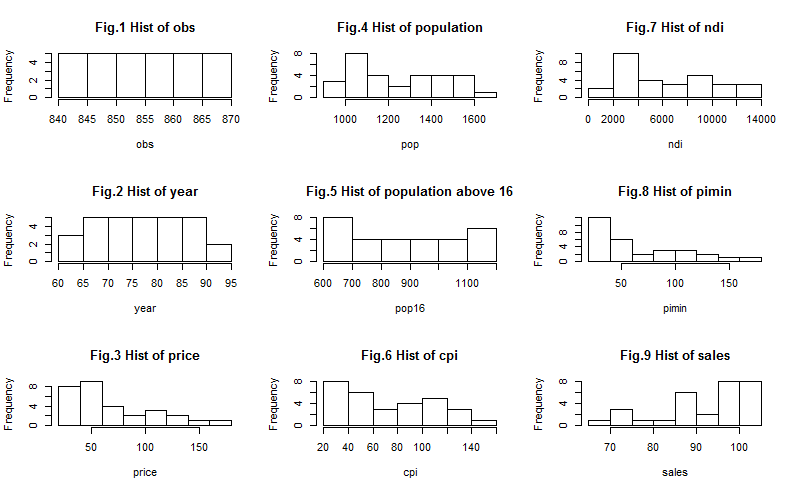
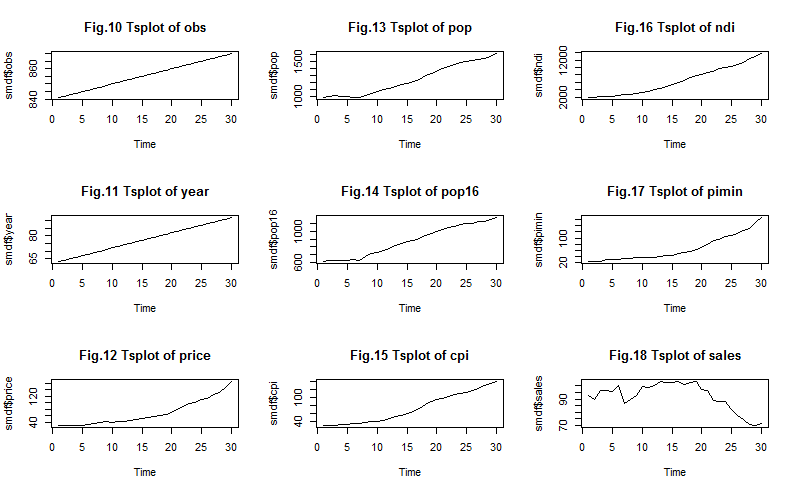
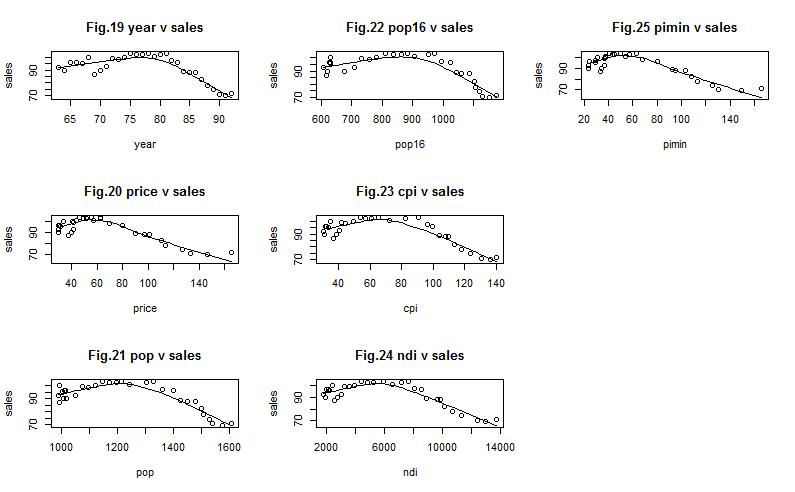


Fig.10 to Fig.18 show the time-series plots of all variables.



**b. Scatter-plots**

Fig.19 to Fig.26 are the scatter-plots of the dependent variable sales and each independent variables. All the graphs show curvature shapes which indicate that the relationships between sales and each independent variables are not linear.



**c. Descriptive Statistics**

**Table 1 Descriptive Statistics**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **mean** | **sd** | **Median** | **trimmed** | **mad** | **min** | **max** | **range** | **skew** | **kurtosis** | **se** | **n** |
| **price** | 68.45 | 38.95 | 53.80 | 63.21 | 31.43 | 30.10 | 165.4 | 135.30 | 0.93 | -0.37 | 7.11 | 30 |
| **pop** | 1244.37 | 212.16 | 1205.00 | 1234.79 | 289.11 | 990.00 | 1606.2 | 616.2 | 0.24 | -1.54 | 38.74 | 30 |
| **pop16** | 874.70 | 195.62 | 872.55 | 872.39 | 285.85 | 609.30 | 1174.8 | 565.50 | 0.01 | -1.57 | 35.72 | 30 |
| **cpi** | 73.60 | 37.14 | 62.90 | 71.15 | 44.48 | 30.60 | 140.3 | 109.70 | 0.34 | -1.46 | 6.78 | 30 |
| **ndi** | 6330.48 | 3753.10 | 5593.61 | 6043.71 | 4588.64 | 1834.69 | 13709.0 | 11874.31 | 0.40 | -1.25 | 685.22 | 30 |
| **pimin** | 65.42 | 40.89 | 47.10 | 60.23 | 26.24 | 23.90 | 165.7 | 141.80 | 0.9 | -0.47 | 7.46 | 30 |
| **sales** | 92.06 | 10.43 | 95.90 | 93.35 | 9.64 | 69.90 | 103.1 | 32.20 | -0.84 | -0.52 | 1.90 | 30 |

As shown in Table 1, most variables can be considered as normal distribution except price, pimin and sales. Variable price and pimin are skewed to the right while variable sales is skewed to the left. Besides, the value of kurtosis show that all the data distributed a little bit heavily in the mid-region.

**d. Correlation Matrix**

Table 2 presents the correlation between dependent variable and all independent variables where it shows that sales has negative correlation with cpi and ndi The correlation is not in line with the original hypotheses that variables cpi and ndi have positive affection on sales. The values also suggest that there are major correlation between independent variables. The impact of multicollinearity on parameter should be considered. Therefore, variables pop and pimin were dropped during the regression analysis.

**Table 2 Correlation Matrix**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **price** | **pop** | **pop16** | **cpi** | **ndi** | **pimin** | **sales** |
| **price** | 1.0000 | 0.9514 | 0.9212 | 0.9627 | 09726 | 0.9985 | -0.8066 |
| **pop** | 0.9514 | 1.0000 | 0.9929 | 0.9959 | 0.9932 | 0.9536 | -0.6317 |
| **pop16** | 0.9212 | 0.9929 | 1.0000 | 0.9841 | 0.9820 | 0.9221 | -0.5537 |
| **cpi** | 0.9627 | 0.9959 | 0.9841 | 1.0000 | 0.9973 | 0.9656 | -0.6679 |
| **ndi** | 0.9726 | 0.9932 | 0.9820 | 0.9973 | 1.0000 | 0.9732 | -0.6766 |
| **pimin** | 0.9985 | 0.9536 | 0.9221 | 0.9656 | 0.9732 | 1.0000 | -0.8039 |
| **sales** | -0.8066 | -0.6317 | -0.5537 | -0.6679 | -0.6766 | -0.8039 | 1.0000 |

**e. Log-linear Regression Results**

In order to avoid heteroscedasticity and improve the overall model, the dependent variable was transformed to logarithms. Table 3 shows the regression analysis results for predicting sales based on price, pop16, cpi, per capita disposable income and year.

**Table 3 Regression Analysis**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Eqn.4** | log(sales) = -6.827e-03\*price+9.571e-04\*pop16--8.712e-03\*cpi  +9.122e-05\*ndi-3.310e-03\*year+ 4.466 | | | | | |
| t-stat | -6.051 | 2.640 | -3.119 | 2.373 | -0.527 | 14.354 |
| p-value | 3.01e-06 | 0.01433 | 0.00467 | 0.02600 | 0.60280 | 2.82e-13 |
| r(corr) | -0.8066 | -0.5537 | -0.6679 | -0.6766 | **--** | **--** |

n= 30, r-sq. = 0.9351 F= 69.13 SE(RMSE) = 0.03376 DW=1.459

The coefficient of determination of .9351 indicates that 93.51% of total variation in the dependent variable sales can be explained by, or attributed to, variation in the price per pack of cigarettes, population above the age of 16, consumer price index with (1983=100), per capita disposable income and year.

F-statistics is 69.13 for the null hypothesis Ho (β1= β2= β3= β4= β5=0) which implies that the linear model is significant at the 1% level of significant. Therefore, The null hypothesis was rejected and the alternative hypothesis Ha(at least 1 independent variable has significant linear relationship with the dependent variable) was taken by the research. The Durbin-Watson test reveals there is evidence of positive serial correlation remaining in the residuals.

For the t-statistic, the hypothesis and alternative hypothesis are set out below.

for price Ho : β1=0

Ha : β1<0

for pop16 Ho : β2=0

Ha : β2>0

for cpi (1983=100) Ho : β3=0

Ha : β3>0

for ndi Ho : β4=0

Ha : β4>0

for year Ho : β5=0

Ha : β5>0

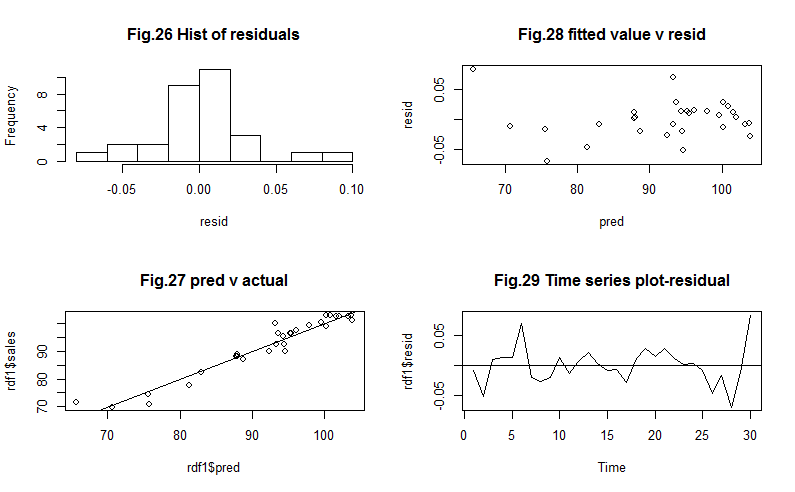
From the t-test, the t values for variables price, pop16, cpi and ndi are greater than 2.33 which indicate that the significant is at the 1% level of significance. So the null hypothesis β1=0, β2=0, β3=0 and β4=0 were rejected and the alternative hypothesis was taken. However, the t value of year is less than 1.28 ( significant at the 10% level of significance). Thus, the null hypothesis was accepted and it shows no evidence of a linear relationship between sales and year.

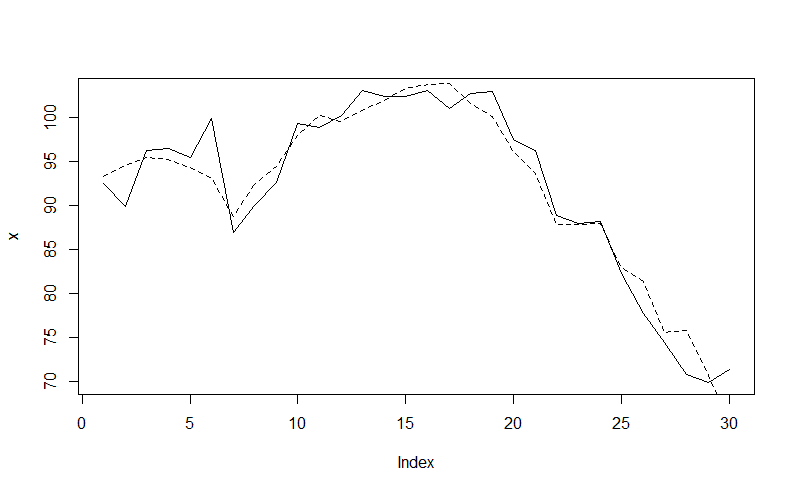
From Table 3, b1 equals to -0.006827 indicates that sales will decrease by 0.006827 unit by increasing one dollar in price on average (everything is equal) . Also, holding the other predictors constantly, increasing 1 unit of cpi that corresponds to decrease sales by 0.008712 unit.

**f. Residual Analysis**

The results of residuals in Fig. 26 shows that it is normally distributed. Fig. 27 shows the strong positive linear relationship between predicted and actual sales value. Fig.28 shows a plot of the residuals against the fitted values for the cigarette sales model. The plot shows no obvious systematic patterns and the variation in the residuals does not seem to change with the size of the fitted value. Fig.29 shows a time plot of the residuals. There is an outlier in the residuals (obs:30). It would be worth investigating that outlier to see if there were any unusual circumstances or events that may have increased cigarette sales for that year.

Fig.30 shows a time-series plot of both fitted value and actual value. The predicted value was close to the actual value.



****

**Fig.30 Time-series plot of pred & actual value**

**V. Conclusion**

The research was successful, with good coefficient of determination equals to .9351. The independent variables price, population above 16 years old, cpi and ndi have strong effect on sales while the year has no effect on cigarette sales. The research results can be improved by adding more observations.

**VI. Appendix I Dataset Used**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| obs | state | year | price | pop | pop16 | cpi | ndi | sales | pimin |
| 841 | 32 | 63 | 30.1 | 990 | 609.3 | 30.6 | 1834.692 | 92.5 | 23.9 |
| 842 | 32 | 64 | 30.2 | 1008 | 621.9 | 31 | 1931.034 | 89.9 | 24 |
| 843 | 32 | 65 | 30.2 | 1014 | 629.6 | 31.5 | 2031.565 | 96.3 | 24.2 |
| 844 | 32 | 66 | 30.4 | 1009 | 628.6 | 32.4 | 2156.182 | 96.5 | 29.6 |
| 845 | 32 | 67 | 31.5 | 1002 | 629.6 | 33.4 | 2254.618 | 95.5 | 29.2 |
| 846 | 32 | 68 | 33.7 | 994 | 632.5 | 34.8 | 2433.689 | 99.9 | 31.2 |
| 847 | 32 | 69 | 37.9 | 994 | 619 | 36.7 | 2563.542 | 87 | 33.3 |
| 848 | 32 | 70 | 39.7 | 1016 | 675.9 | 38.8 | 2785.548 | 90 | 34.6 |
| 849 | 32 | 71 | 41.7 | 1050 | 708.7 | 40.5 | 3026.404 | 92.6 | 36.6 |
| 850 | 32 | 72 | 41.1 | 1073 | 732.8 | 41.8 | 3286.109 | 99.3 | 37.2 |
| 851 | 32 | 73 | 41.8 | 1097 | 758.8 | 44.4 | 3605.504 | 98.9 | 36.5 |
| 852 | 32 | 74 | 43.7 | 1120 | 782.9 | 49.3 | 3928.042 | 100.3 | 37.8 |
| 853 | 32 | 75 | 46.3 | 1146 | 810.9 | 53.8 | 4432.791 | 103.1 | 40.5 |
| 854 | 32 | 76 | 49.5 | 1172 | 837.9 | 56.9 | 4847.482 | 102.4 | 43.4 |
| 855 | 32 | 77 | 51.6 | 1195 | 862.9 | 60.6 | 5266.361 | 102.4 | 44.7 |
| 856 | 32 | 78 | 56 | 1215 | 882.2 | 65.2 | 5920.86 | 103.1 | 49.5 |
| 857 | 32 | 79 | 57.6 | 1241 | 905.3 | 72.6 | 6555.463 | 101 | 53.7 |
| 858 | 32 | 80 | 62.6 | 1303 | 951.6 | 82.4 | 7168.074 | 102.7 | 57.2 |
| 859 | 32 | 81 | 63 | 1328 | 971.9 | 90.9 | 7683.295 | 103 | 62.7 |
| 860 | 32 | 82 | 69.4 | 1359 | 994 | 96.5 | 8046.673 | 97.5 | 68.1 |
| 861 | 32 | 83 | 79.6 | 1399 | 1023.9 | 99.6 | 8546.187 | 96.3 | 80.5 |
| 862 | 32 | 84 | 90.2 | 1424 | 1042.3 | 103.9 | 8896.999 | 88.9 | 92.8 |
| 863 | 32 | 85 | 97.5 | 1450 | 1061.5 | 107.6 | 9619.565 | 88 | 95.1 |
| 864 | 32 | 86 | 101.2 | 1479 | 1085.6 | 109.6 | 9889.743 | 88.2 | 103.5 |
| 865 | 32 | 87 | 110.2 | 1500 | 1103 | 113.6 | 10162.01 | 82.3 | 108.6 |
| 866 | 32 | 88 | 113.7 | 1507 | 1105 | 118.3 | 10672 | 77.7 | 113.5 |
| 867 | 32 | 89 | 127.2 | 1528 | 1120 | 124 | 11302 | 74.4 | 125.6 |
| 868 | 32 | 90 | 133.6 | 1539.2 | 1128.2 | 130.7 | 12398 | 70.8 | 130.2 |
| 869 | 32 | 91 | 146.9 | 1572.7 | 1150.5 | 136.2 | 12961 | 69.9 | 149.1 |
| 870 | 32 | 92 | 165.4 | 1606.2 | 1174.8 | 140.3 | 13709 | 71.4 | 165.7 |

**VII. Appendix II R Script**

# install packages and load libraries

install.packages("psych")

install.packages("nlme")

library(psych)

library(mgcv)

install.packages("zoo")

library(lmtest)

# this function is courtesy of Peter Yin

ftsplot2 <- function(x,y){plot(x,type="l",lty=1); lines(y,lty=2)}

# import data set

cigar <- read.csv("C:/Users/aden/Dropbox/cf/DS633/project/Cigar.csv")

dim(cigar)

cgdf<-cigar

names(cgdf)

names(cgdf)[1]<-paste("obs")

# small data set(time series)

smdf<-cgdf[cgdf$state==32,]

dim(smdf)

names(smdf)

# Histgram of all variables

par(mfcol=c(3,3))

hist(smdf$obs,main="Fig.1 Hist of obs",xlab="obs")

hist(smdf$year,main="Fig.2 Hist of year",xlab="year")

hist(smdf$price,main="Fig.3 Hist of price",xlab="price")

hist(smdf$pop,main="Fig.4 Hist of population",xlab="pop")

hist(smdf$pop16,main="Fig.5 Hist of population above 16",xlab="pop16")

hist(smdf$cpi,main="Fig.6 Hist of cpi",xlab="cpi")

hist(smdf$ndi,main="Fig.7 Hist of ndi",xlab="ndi")

hist(smdf$pimin,main="Fig.8 Hist of pimin",xlab="pimin")

hist(smdf$sales,main="Fig.9 Hist of sales",xlab="sales")

par(mfcol=c(3,3))

plot.ts(smdf$obs); title("Fig.10 Tsplot of obs")

plot.ts(smdf$year); title("Fig.11 Tsplot of year")

ts.plot(smdf$price); title("Fig.12 Tsplot of price")

ts.plot(smdf$pop); title("Fig.13 Tsplot of pop")

ts.plot(smdf$pop16); title("Fig.14 Tsplot of pop16")

ts.plot(smdf$cpi); title("Fig.15 Tsplot of cpi")

ts.plot(smdf$ndi); title("Fig.16 Tsplot of ndi")

ts.plot(smdf$pimin); title("Fig.17 Tsplot of pimin")

ts.plot(smdf$sales); title("Fig.18 Tsplot of sales")

# scatter plot

par(mfcol=c(3,3))

plot(smdf$year,smdf$sales,main="Fig.19 year v sales",xlab="year",ylab="sales")

plot(smdf$price,smdf$sales,main="Fig.20 price v sales",xlab="price",ylab="sales")

plot(smdf$pop,smdf$sales,main="Fig.21 pop v sales",xlab="pop",ylab="sales")

plot(smdf$pop16,smdf$sales,main="Fig.22 pop16 v sales",xlab="pop16",ylab="sales")

plot(smdf$cpi,smdf$sales,main="Fig.23 cpi v sales",xlab="cpi",ylab="sales")

plot(smdf$ndi,smdf$sales,main="Fig.24 ndi v sales",xlab="ndi",ylab="sales")

plot(smdf$pimin,smdf$sales,main="Fig.25 pimin v sales",xlab="pimin",ylab="sales")

par(mfcol=c(3,3))

scatter.smooth(smdf$year,smdf$sales, main="Fig.19 year v sales",xlab="year",ylab="sales")

scatter.smooth(smdf$price,smdf$sales, main="Fig.20 price v sales",xlab="price",ylab="sales")

scatter.smooth(smdf$pop,smdf$sales, main="Fig.21 pop v sales",xlab="pop",ylab="sales")

scatter.smooth(smdf$pop16,smdf$sales, main="Fig.22 pop16 v sales",xlab="pop16",ylab="sales")

scatter.smooth(smdf$cpi,smdf$sales, main="Fig.23 cpi v sales",xlab="cpi",ylab="sales")

scatter.smooth(smdf$ndi,smdf$sales, main="Fig.24 ndi v sales",xlab="ndi",ylab="sales")

scatter.smooth(smdf$pimin,smdf$sales, main="Fig.25 pimin v sales",xlab="pimin",ylab="sales")

# Table 1 Descriptive Statistics

describe(smdf[,c("price","pop","pop16","cpi","ndi","pimin","sales")])

#Table 2 Correlation matrix

round(cor(smdf[,c("price","pop","pop16","cpi","ndi","pimin","sales")]),4)

# Log-linear model

fit1<-lm(log(sales)~price+cpi+ndi+pop16+year,data=smdf)

summary(fit1)

dwtest(fit1)

rdf1<-data.frame(smdf,pred=exp(fit1$fitted.values),resid=fit1$residuals)

par(mfcol=c(2,2))

hist(rdf1$resid,main="Fig.26 Hist of residuals",xlab="resid")

plot(rdf1$pred,rdf1$sales,main="Fig.27 pred v actual"); abline(lm(sales~pred,data=rdf1))

plot(rdf1$pred,rdf1$resid,main="Fig.28 pred v resid",xlab="pred",ylab="resid")

ts.plot(rdf1$resid,main="Fig.28 Time series plot-residual"); abline(h=0)

par(mfcol=c(1,1));

ftsplot2(rdf1$sales,rdf1$pred)

**VIII. Bibliography**

Baltagi, B.H. and D. Levin (1992) “Cigarette taxation: raising revenues and reducing consumption”, Structural Changes and Economic Dynamics, 3, 321–335.

Baltagi, B.H., J.M. Griffin and W. Xiong (2000) “To pool or not to pool: homogeneous versus heterogeneous estimators applied to cigarette demand”, Review of Economics and Statistics, 82, 117–126.