ABSTRACT

CHEN, ZHENGYUAN. Analysis of Risk Behavior and Economic Factors for State Mortality Rates in the U.S. (Under the direction of Mehmet Caner.)

To determine the Behavioral Risk and Regional economic Factors for state Mortality Rate (MR) in the U.S., age-adjusted Mortality Rate are obtained from the U.S Centers for Disease Control and Prevention of the year 2001 to 2006. Multivariate Linear Regression (MLR) and Time Series Cross Section Regression are used to investigate factors which included Personal Income, Unemployment rate, Medical Insurance Coverage, Smocking and Heavy Drinking Behaviors. Panel data were first transformed by the Box-Cox transformation. MLR model (R^2 =0.72, $p < .0001$), One-way Fixed Effect Model (R^2 =0.977, $p < .0001$), Two-way Fixed Effect Model (R^2 =0.987, $p < .0001$), One-way Random Effect Model (R^2 =0.785), Two-way Random Effect Model (R^2 =0.126) were tested. The results of Hausman test were to reject Fixed Effects Model. I preferred One-way Random Effect Model. Smoking and Heavy Drinking significantly contributed to MR. However, Excises have negative impact on MR. States with higher personal income tend to have low Mortality Rates. Medical Insurance Coverage and unemployment rate are not significant in the model.

Analysis of Risk Behaviors and Economic Factors for State Mortality Rates in the U.S.

by
Zhengyuan Chen

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Master of Science

Economics

Raleigh, North Carolina

2011

APPROVED BY:

Mehmet Caner
Committee Chair

Lexin Li

Yichao Wu

DEDICATION

To my parents who have provided support throughout my college career.

BIOGRAPHY

Zhengyuan Chen, raised in Kaili, Guizhou, China, graduated from Zhejiang University with a Bachelors degree in Bioinformatics. She spent a memorable and enjoyable time in Hangzhou during her college life.

She entered North Carolina State University in 2008 and completed her Masters degree in Statistics and Economics in 2011.

ACKNOWLEDGMENTS

I would like to thank Dr. Mehmet Caner, Dr. Yichao Wu and Dr. Lexin Li for all your guidance and help on my thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Cause of death is classified according to International Classification of Disease (Table A.1). In this study, Age-adjusted Death Rates(ADR) were used as the dependent variable, which are weighted averages of the age-specific death rates, where the weights represent a fixed population by age (Centers for Disease Control and Prevention, 2010).

Mortality Rates in the U.S. have big differences among states. For example, Mississippi has two times higher mortality rate than Minnesota. The state-specific factors such as average income, Health Care Coverage, Unemployment, and other Behavioral Risk Factors contributed a lot to different Mortality Rates. It is really valuable to understand factors which influenced to a high or low mortality rate among states. Smoking, Heavy Drinking, Excises, Unemployment, Health Care Coverage and Income are used to predict state-base mortality rate.

The purpose of this study is to determine the factors that contributed to mortality rate from 2001 to 2006. The important aspect of my study applied panel data analysis. Not only multivariate linear method but also Box-Cox method is utilized as well. Panel data is very flexible in modeling the individual differences.

2. Literature Review

Most of the state level mortality research is only focus on the specific disease mortality rates, like oral cancer or infant mortality rates. In oral cancer mortality study, the author chose the behavior risk factors like smoking and tobacco use, which was found to relate with oral cancer rate (Anthony).

In previous study, oral cancer incidence and mortality rates exhibit geographic specificity within the U.S. (Anthony and Kingsley). Some of the studies have demonstrated the tobacco and alcohol were highly correlated to oral cancer (LA).

One study about state infant mortality rates for 2001 and 2002 indicated that infant mortality rates are correlated with Percentage of non-Hispanic, Smocking rate/pregnancy, Teen birth rate, Percentage Hispanic ethnicity, Body Percentage of women with normal Body Mass Index (BMI) (David).

3. DATA

State-based Mortality Data is obtained from the Centers for Disease Control and Prevention (CDC). Smoke, Drink, Excises behavior data were also downloaded from the Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS). State Annual Personal Income is obtained from Bureau of Economic Analysis

(BEA) and Unemployment rate are downloaded from Bureau of Labor Statistics (BLS).

Table A2 contains all the variables name and definitions.

As shown in figure 1, the states with lowest mortality rates are Minnesota, California, North Dakota, Connecticut, Utah; the states with highest mortality rates are Mississippi, Louisiana, Alabama, West Virginia, Kentucky

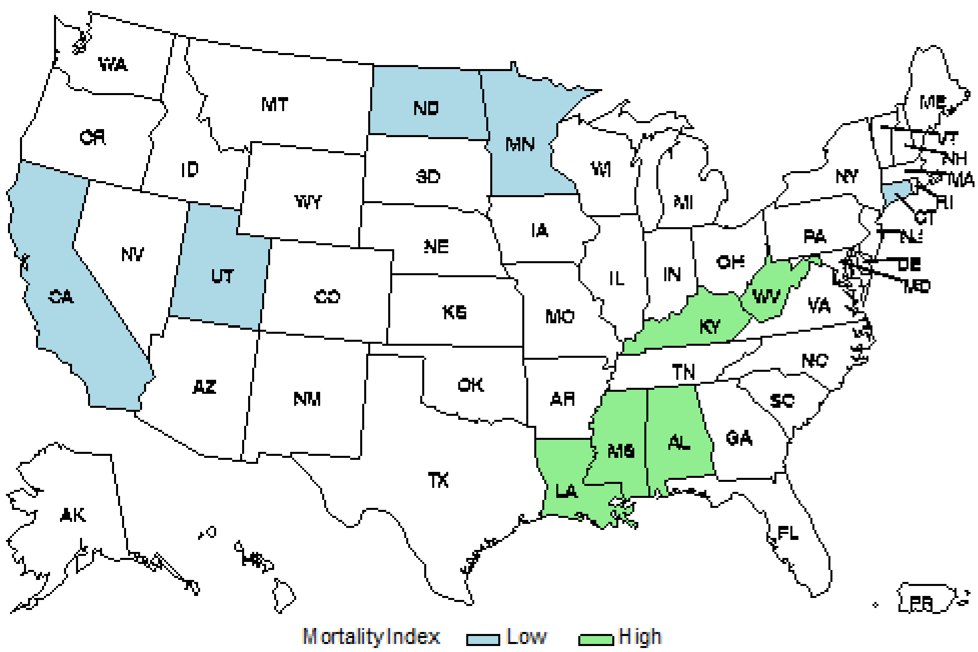

Figure 3.1 Maps of Average Mortality Rates for Top and Bottom 5 States, 2001-2008

1. Centers for Disease Control and Prevention. Available at: http://wonder.cdc.gov/cmf-icd10.html. Mortality for 1999 - 2006 with ICD 10 codes.

2. Behavioral Risk Factor Surveillance System**.** Available at: http://apps.nccd.cdc.gov/brfss/page.asp?yr=2009&state=All&cat=TU#TU.

3. State Annual Personal Income. Available at: http://www.bea.gov/regional/spi/default.cfm?selTable=SA04&selSeries=ancillary. Revised Sep.20, 2001

4. Unemployment Rates for States. Available at: http://www.bls.gov/lau/.

4. Methodology

4.1Model details

The state-based Mortality data combines time series and cross sections, which can be viewed as Panel Data in Economics. The basic panel data model is a regression model of the form (Greene, 2002)

$$Y_{it} = \sum_{k=1}^{K} X_{itk}\beta_k + u_{it} \qquad i = 1,\ldots,N; \qquad t = 1,\ldots,T$$

Where K is the number of predict variables, N is the number of *i*th unit cross sections and T is the length of time series within the *i*th unit cross sections.

The base Mortality Rate model:

$$DR_{it} = \beta_1 + \beta_2 EXS_{it} + \beta_3 INS_{it} + \beta_4 SMK_{it} + \beta_5 ALC_{it} + \beta_6 UNEMP_{it} + \beta_7 LOG\_INC_{it}$$

Where DR denotes Mortality Rate in state i and time t; EXS_{it} is Excises which is the percentage of participating in physical activities; INS_{it} is percent of people covered by health care insurance; SMK_{it} is the percent of smoker adults, ALC_{it} is the percent of heavy drinkers (adult men having more than two drinks per day and adult women having more than one drink per day), $UNEMP_{it}$ is unemployment rate, LOG_INC_{it} is log transform of Personal income.

4.1.1 Fixed Effects Model

A major motivation for using panel data is that it allows us to control for unobserved heterogeneity. We hope that our independent variables have explained much of what is different about a company, or a year, but there is probably some un-modeled heterogeneity. Since we haven't modeled it, it goes into u_{it} .

Fixed-effects explore the relationship between predictor and outcome variables within an individual. The fixed model controls for all time-invariant differences between the individuals, which are designed to study the causes of changes within an individual. Ordinal least squares (OLS) estimation is best linear unbiased.

Fixed model is to fit this Lease squares dummy variable (LSDV) model.

$$y_i = X_i \beta + i\alpha_i + \varepsilon_i$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \beta + \begin{bmatrix} i & 0 & \dots & 0 \\ 0 & i & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & i \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

a. One-way fixed effect model

$$u_{it} = v_i + \varepsilon_{it}$$

Where the $v_i s$ are fixed.

b. Two-way fixed effect model

$$u_{it} = v_i + e_t + \varepsilon_{it}$$

Where the $v_i s$ and $e_t s$ are fixed effects.

4.1.2 Random effects model

The variation across entities is assumed to be random and uncorrelated with the independent variables included in the model. There is an individual effect, but the effect is random. This individual effect may reflect omitted variables which are not fixed.

Random effects allows to generalize the inferences beyond the sample used in the model

Random effects assume that the individual's error term is not correlated with the predictors which allows for time-invariant variables to play a role as explanatory variables.

The estimation method is a GLS procedure.

a. One-way random effect model

$$u_{it} = v_i + \varepsilon_{it}$$

$E(v_i) = 0, E(v_i^2) = \sigma_v^2$ and $E(v_i v_t) = 0 \, for i \neq j$, and v_i is uncorrelated with ε_{it} for all i and t.

b. Two-way random effect model

$$u_{it} = v_i + e_t + \varepsilon_{it}$$

In addition to all of the preceding, $E(e_i) = 0, E(e_t^2) = \sigma_t^2$

The model is a variance components model, with the variance components σ_v^2 and σ_t^2, as well as σ_ε^2, to be estimated.

4.2 Hausman test

To decide between fixed or random effects we can run a Hausman test where the null hypothesis is that the preferred model is random effects.

It basically tests whether the unique errors are correlated with the independent variables, the null hypothesis is they are not. Under H0: OLS in the LSDV model and GLS are consistent, but OLS is inefficient. Under alternative: OLS is consistent, but GLS is not.

H0: both of b and $\widehat{\beta}$ are consistent, but $\widehat{\beta}$ is more efficient than b (b is estimators from OLS, and $\widehat{\beta}$ is estimators from GLS).

$$
\begin{aligned}
&\operatorname{var}(b-\widehat{\beta}) = \operatorname{var}(b) + \operatorname{var}(\widehat{\beta}) - 2\operatorname{cov}(b,\widehat{\beta}) \\
&\operatorname{cov}((b-\widehat{\beta}),\widehat{\beta}) = \operatorname{cov}(b,\widehat{\beta}) - \operatorname{var}(\widehat{\beta}) = 0 \\
&\operatorname{cov}(b,\widehat{\beta}) = \operatorname{var}(\widehat{\beta}) \\
&\operatorname{var}(b-\widehat{\beta}) = \operatorname{var}(b) - \operatorname{var}(\widehat{\beta}) = \psi \\
&W = \chi^2(k-1) = [b-\widehat{\beta}]'\widehat{\psi}^{-1}[b-\widehat{\beta}]
\end{aligned}
$$

4.3 Box-Cox Transform

The usual Box-Cox method is to find the maximum likelihood power transformations of the dependent variable in a regression model.

When the dependent variable Y is known to be positive, the following transformation can be used:

$$
y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda}}{\lambda} & \text{when } \lambda \neq 0 \\ \log(y_i) & \text{when } \lambda = 0 \end{cases}
$$

5. Results and Discussion

1. Box-Cox Transform

1.1 Check Normality

Based on the following basic model, I checked the residuals as shown in the plot (Figure 5.1).

The residuals do not have any patterns. The scatter plot shows the nature relationship about the response variable death rate against several independent variables.

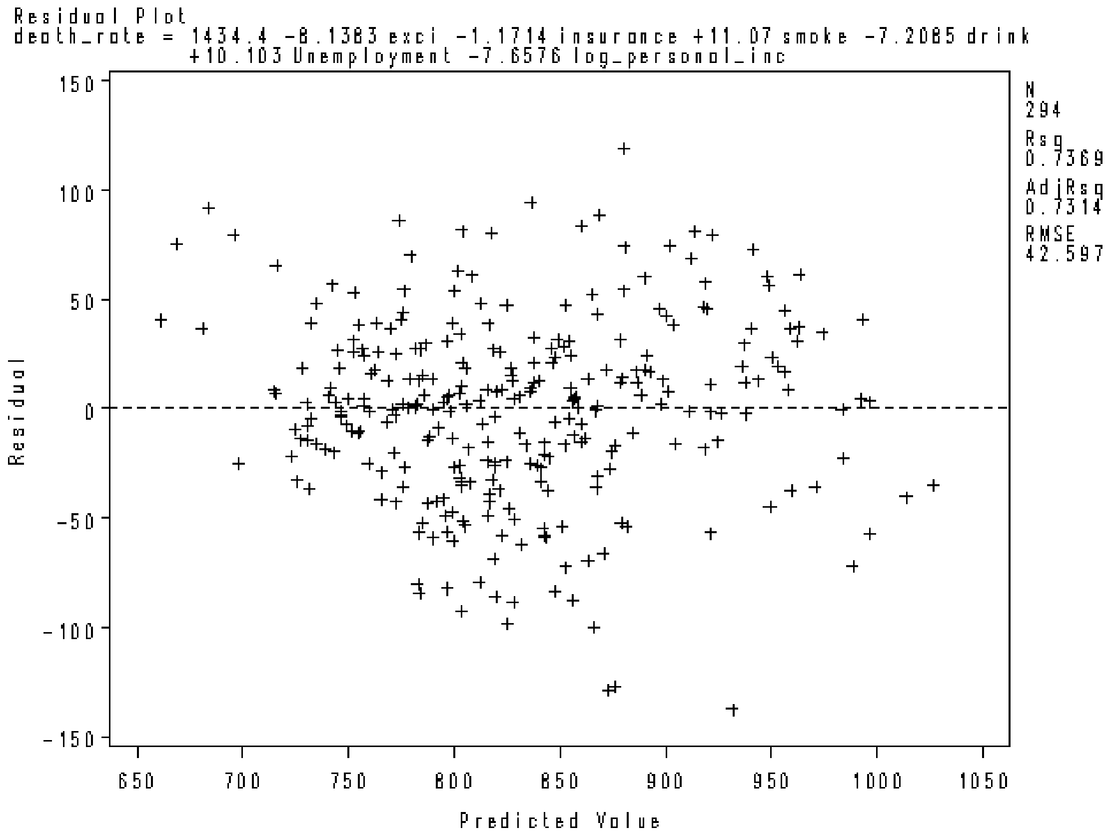

Figure 5.1.Residual Plot of Box-Cox transformed model

1.2 Quantile-Quantile plot

The line Q-Q plot (Figure 5.2) is almost straight, except few points in the left tail.

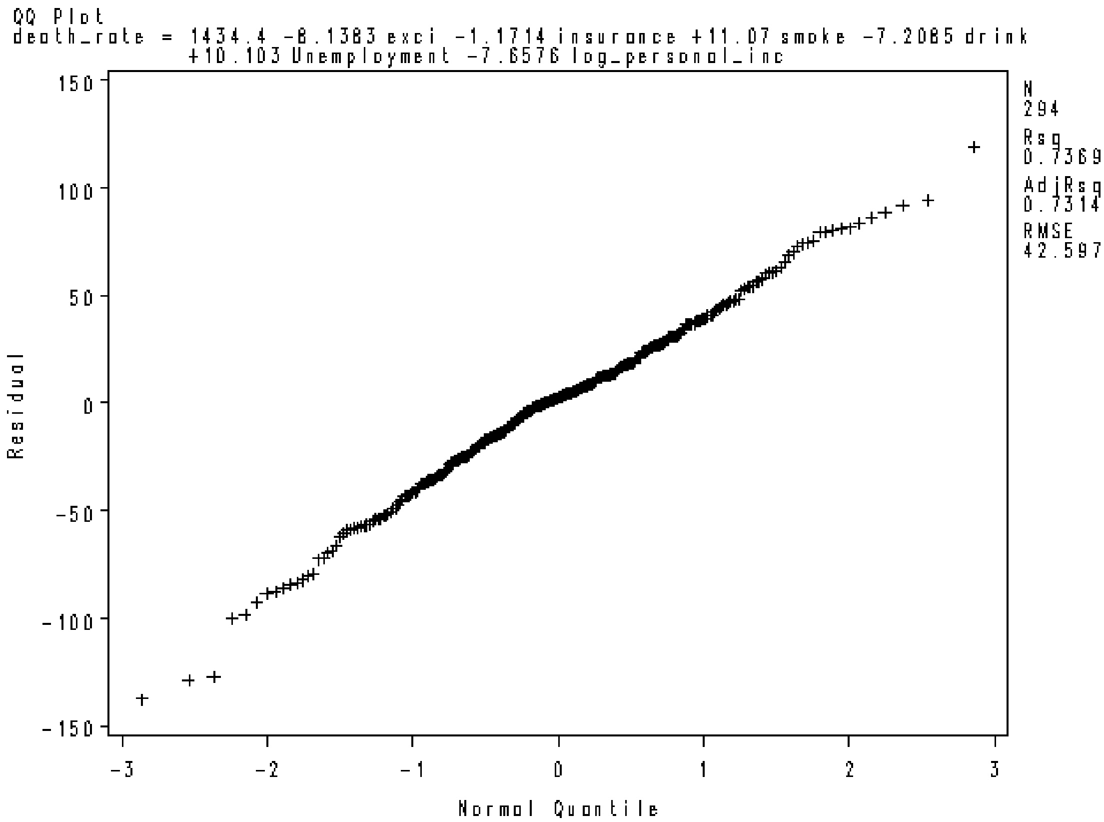

Figure 5.2.Quantile-Quantile plot for OLS model

1.3 Box-cox Transformation

I transformed the data using Box-Cox transformation to get a better linear relationship (Table 1).

Choosing the Lambda = -0.5 and then transformed Death Rate from to transformed death rate

(TDEATH_RATE).

Table 1. Box-Cox Transformation Output for Death Rate

Box-Cox Transformation Information for Death Rate			
Lambda	R-Square	Log Like	
-1.75	0.72	-1102.54	
-1.5	0.72	-1101.23	
-1.25	0.72	-1100.17	*
-1	0.72	-1099.38	*
-0.75	0.73	-1098.85	*
-0.5	0.73	-1098.6	<
-0.25	0.73	-1098.63	*
0	0.73	-1098.93	*
0.25	0.73	-1099.52	*
0.5	0.73	-1100.4	*
0.75	0.74	-1101.57	
< - Best Lambda			
* - 95% Confidence Interval			
+ - Convenient Lambda			

2. Simple OLS model Results

I first provided Ordinary Least Squares (OLS) regression results with the U.S. state mortality data from 2001 to 2006, a total of 294 records. These results are shown in the below table 2.

Table 2. OLS Estimates with U.S. State Mortality Rate data from 2001 to 2006

Model	R-Square	Parameters	Estimate	Standard Error	P-value
OLS	0.7281	Intercept	1.95329	0.0043	<0.0001
		EXS	-0.00032	0.00003678	<0.0001
		INS	-0.00004734	0.00002859	0.0988
		SMK	0.00046509	0.00004179	<0.0001
		ALC	-0.000254	0.00008768	0.0041
		UNEMP	0.00041606	0.00011539	0.0004
		INC	-0.00029862	0.00011391	0.02338

From the results, the INS is not significant at 5% level for the p-value 0.09, which means the insurance coverage do not have any significant effect on the mortality rate. For the reasons that, most employers have health insurance provided by the company, also, low-income and older citizens have Medicaid and Medicare Insurance; That’s why insurance coverage do not have a necessary correlation with mortality rates. From the results, smoking harms your body with the significant positive coefficients. Also, heavy drinking does the same thing (adult men having more than two drinks per day and adult women having more than one drink per day). The more heaving drinks, the higher mortality. And states with the higher personal income tend to have lower mortality rates. The more excises, the healthier they are. An interesting result is unemployment rate; the result showed that mortality rates among states have a statistical significant positive correlation with unemployment rates.

3. One-way fixed Model Results

In the one-way fixed model, states are considered as fixed effects. Compared with the simple OLS model (1), the Degree Of Freedom (DF) of error decreased from 284 to 238 since we treated 49 states as fixed and then we lost 48 DFs. At the same time, R-Squares increased from 0.7281 to 0.9771.

The major different in this model is that when controlled the state effects, Percent of Healthy Care Coverage, Unemployment rates were no longer statistically significant, for which p-value were 0.4997, 0.2862.

For the regional effects, coefficient of North Dakota is -77.9359 and California is 801.96. Personal income still has high negative correlation with mortality rates. Controlled for state effects, behavior factors such as Excises, Smoking, and Heavy Drinking determined the mortality rates. Excises are good for health and smoking and heavy drinking are not, which is the same results as simple OLS model.

Onew-way fixed model results showed below in Table 3.

Table 3. One-way fixed model results with U.S. state-based mortality data

Intercept	EXS	INS	SMK	ALC	UNEMP	INC	R-square
4216.33 (317.6)	-1.8821 (0.598)	0.5260 (0.7780)	4.508 (0.9427)	3.67333 (1.5184)	-1.7277 (1.6164)	-203.10 (16.89)	0.9771
CS1- CS7	524.28 (37.50)	-10.06 (12.22)	387.098 (38.053)	355.214 (28.4)	801.964 (70.6117)	400.292 (37.669)	370.013 (38.233)
CS8- CS14	84.38 (15.57)	616.346 (59.487)	627.27 (47.89)	121.32 (13.45)	640.95 (57.88)	496.80 (45.15)	246.88 (30.66)
CS15- CS21	311.93 (29.34)	458.82 (40.33)	533.38 (36.95)	145.92 (17.71)	523.64 (43.81)	491.69 (47.71)	587.8 (53.787)
CS22- CS28	364.39 (41.78)	426.96 (30.09)	494.08 (44.11)	64.63 (9.23)	175.65 (21.39)	337.48 (28.12)	127.03 (19.60)
CS29- CS35	572.93 (52.11)	185.27 (18.91)	673.39 (64.99)	584.69 (48.70)	-77.94 (9.34)	634.41 (56.23)	90.15 (17.80)
CS36- CS42	359.80 (32.14)	634.49 (57.66)	90.15 (17.80)	445.24 (36.13)	-12.3 (10.63)	556.92 (45.82)	744.61 (61.48)
CS43- CS48	243.60 (18.76)	-44.74 (9.94)	546.86 (47.55)	460.52 (43.10)	310.48 (23.06)	402.15 (42.24)	

4. Two-way fixed effects

As I fixed the state and time effects, I lose 5 more error DFs compared with One-way fixed effects model. However, all the risk behavior and economic factors became none statistically significant. Instead, 21 states effects and 4 time effects are significant (Table 4).

The results from this two-way fixed effects model are really interesting. The states in the mid-west are more likely to have lower mortality rates than other places. However, Arkansas, Maine, Montana, South Dakota and West Virginia tend to have higher mortality rates. Also, fixed the time effects, in the year 2001, 2002, 2003 and 2005 have higher death rates than other years.

Table 4. Two-way fixed model with state-based Mortality Data

Intercept	EXS	INS	SMK	ALC	UNEMP	INC	R-square
-208.82 (545.5)	1.24 (0.5571)	-0.96 (0.6648)	1.0062 (0.80)	0.3365 (0.29)	2.1936 (1.78)	31.84 (31.64)	0.9865
CS1- CS7	98.74 (61.53)	-64.58 (12.79)	-141.52 (71.58)	53.07 (44.64)	-221.72 (137.0)	-145.914 (72.39)	-147.48 (70.27)
CS8- CS14	4.69 (17.25)	-181.73 (109.4)	-0.65 (86.12)	-77.41 (25.66)	-103.95 (103.1)	-24.88 (74.56)	-116.46 (51.45)
CS15- CS21	-54.97 (50.79)	68.6 (59.16)	102.162 (62.09)	-30.95 (26.46)	-60.41 (79.54)	-142.79 (86.99)	-75.62 (92.72)
CS22- CS28	190.118 (74.80)	127.84 (45.59)	-17.97 (73.21)	-30.83 (13.42)	-86.33 (36.41)	7.57 (47.57)	-96.10 (31.76)
CS29- CS35	-127.64 (95.67)	-68.65 (34.33)	-188.71 (118.3)	-32.79 (85.16)	-82.12 (7.69)	-42.88 (95.65)	67.51 (54.39)
CS36- CS42	-94.4317 (60.816)	-80.87 (99.89)	-58.06 (24.46)	-18.54 (60.69)	-69.24 (11.16)	59.47 (72.66)	-116.86 (116.6)
CS43- CS48	-114.94 (43.49)	-67.34 (8.40)	-79.50 (85.60)	-151.04 (81.40)	113.82 (-127.0)	-126.96 (73.91)	
TS1-	77.41	67.55	53.65	17.82	23.78		

TS5	(8.77)	(7.39)	(6.01)	(4.22)	(3.18)

5. The One-way Random Effects Model

In this part, the cross sections-state effects are considered as random. So we need to estimate the variance components. The variance component estimates are shown below (Table 5.),

Table 5. Variance Component Estimate of One-way Random Model

Variance Component Estimates	
Variance Component for Cross Sections	0.002437
Variance Component for Error	3.717E-7

Compared to the fixed effects model, R-squares decreased a lot. And random the state effects, personal Income became none statistical significant.

The null hypothesis for Hausman Test(Table 6.) is that random model is preferred. Based on the table, we reject the null hypothesis at 5% level and believe that random model is better.

Table 7 showed One-way random effects model results.

Table 6. One-way Random Hausman Test

Hausman Test for Random Effects		
DF	m Value	Pr > m
6	0.12	1.0000

Table 7. One-way Random Effects Results

R-Square	Parameters	Estimate	Standard Error	P-value
0.7847	Intercept	2.09	0.0155	<0.0001
	EXS	-0.00008	0.000024	0.0013
	INS	0.000029	0.000032	0.3550
	SMK	0.000197	0.000038	<0.0001
	ALC	0.000159	0.000062	0.0101
	UNEMP	-0.00004	0.000066	0.5654
	INC	-0.00852	0.000681	<0.0001

6. Two-way Random Effects Model

The time series-year effects and cross sections-state effects are considered as random effects.

Variance Component Estimates are shown in Table 8.:

Table 8. Variance Component Estimates of Two-way Random Effects Model

Variance Component Estimates	
Variance Component for Cross Sections	2.956E-6
Variance Component for Time Series	2.112E-7
Variance Component for Error	2.382E-7

The Hausman test (Table 9.) indicates that two-way random effects model is not preferred.

Table 9. Two-way Random Hausman Test

Hausman Test for Random Effects		
DF	m Value	Pr > m
3	200.73	<0.0001

In the two-way random effects model (Table 10.), the only significant predict variable is Tobacco use. That's why this model have very small p-value.

Table 10. Two-way Random Effects Results

R-Square	Parameters	Estimate	Standard Error	P-value
0.1264	Intercept	1.93	0.00607	<0.0001
	EXS	-0.00003	0.000027	0.2393
	INS	-0.00005	0.000032	0.1327
	SMK	0.000202	0.000037	<0.0001
	ALC	-0.00004	0.000065	0.5358
	UNEMP	0.000198	0.000080	0.0143
	INC	-0.0001	0.000269	0.7060

6. Conclusion

After discussing and comparing four different models, One Way Random Model is the best model to fit the data which random state effects. The results from Hausman Test showed that states should be considered as a random factor. Incorporated income variable in the regression model has already taken state effects.

Personal income plays an important role in this model; higher income states tend to have lower death rates;

Higher percent of heaving drinkers and smokers will lead to higher death rates;

Unemployment and percentage of health care coverage do not have significant effects on the death rates;

It is interesting that the more people participated in the physical activity, the fewer death rates in that state.

References

Kingsley K, O'Malley S, Chino M (2008): Analysis of oral cancer epidemiology in the US reveals state-specific trends: implications for oral cancer prevention. BMC Public Health, 8(1)**:**87

La Vecchia C, Franceschi S, Levi F (2003): Epidemiological research on cancer with a focus on Europe. Eur J Cancer Prev , 12:5-14

Anthony Bunnell (2009): Analysis of primary risk factors for oral cancer from select US states with increasing rates. Tobacco Induced Diseases, 8:5

David A. Paul (2009), State Infant Mortality: An Ecologic Study to Determine Modifiable Risks and Adjusted Infant Mortality Rates. Matern Child Health J 13:343–348

The TSCSREG Procedure in SAS (1999), SAS Institute Inc.

Greene, W. H. (2003): Econometric Analysis. 5th edition.

Appendix

Table A1 The International Classification of Disease (ICD) 10th revision (Centers for Disease Control and Prevention)

ICD-10 Code	Title
B33.4	Hantavirus (cardio)-pulmonary syndrome [HPS][HCPS]
G90.4	Autonomic dysreflexia
I15.0	Renovascular hypertension
I15.9	Secondary hypertension, unspecified
K22.7	Barrett’s esophagus
K85.0	Idiopathic acute pancreatitis
K85.1	Biliary acute pancreatitis
K85.2	Alcohol-induced acute pancreatitis
K85.3	Drug-induced acute pancreatitis
K85.8	Other acute pancreatitis
K85.9	Acute pancreatitis, unspecified
M31.7	Microscopic polyangiitis
M79.7	Fibromyalgia
R29.6	Tendency to fall, not elsewhere classified
R50.2	Drug-induced fever
R50.8	Other specified fever
W46	Contact with hypodermic needle

Table A2 Variable Name and Definition

Variable Name	Variable Description
DEATH_RATE	Age-adjusted Death Rates
TDEATH_RATE	The Box-Cox transform of Age-adjusted Death Rates
EXCI	Percentage of participating in physical activities
INS	Percent of people covered by health care insurance
SMK	Percentage of smoker adults
ALC	Percentage of heavy drinkers
UNEMP	Unemployment rate
INC	Personal income
LOG_INC	Log Transform of Personal Income