MSDS 6372 - Project 1

Life Expectancy - Objective 1

Jenna Ford, Christian Nava, Shane Weinstock
October 05, 2019

I. Introduction

The World Health Organization's health-related statistics, housed in the Global Health Observatory, tracks health-related factors for many countries around the world. Using these data, along with data available from the World Bank, we seek to determine the factors that affect life expectancy. We will use linear regression selection techniques to obtain the most relevant predictors.

II. Data Description

We initiated the project with the Kaggle Life Expectancy (WHO) dataset[1]. During our preliminary exploratory data analysis, we realized there were data description and data quality issues (possibly transcription errors) with the dataset. We used the data sources listed in Kaggle to recreate the dataset.

The dataset we used was compiled using the World Health Organization's health-related statistics, housed in the Global Health Observatory, and the World Bank. The data source and description for each variable are provided in Figure 1.1 (see the appendix for Figures). Population, GDP, and HIV data come from the World Bank. Data for the categorical variable, Status, comes from the Kaggle Life Expectancy (WHO) dataset. All other data comes from the World Health Organization. We are considering only data from 2015, the most recent year for which all data are available, for the purpose of this assignment.

Several variables needed to be normalized as a percentage of the total population to enable comparison across countries. The variables we normalized are: adult_mortality, under_5_deaths, infant_deaths, neonatal_deaths, hepatitis, and tuberculosis variables.

[1] Rajarshi, Kumar. https://www.kaggle.com/kumarajarshi/life-expectancy-who.

III. Exploratory Data Analysis

Missing Values

We restricted our dataset to countries where the life expectancy, the response variable, value for 2015 was present. Missing values do occur in this dataset for explanatory variables and we did not remove these observations.

Transformations

We used several scatterplot matrices (Figures 1.2-1.4) to evaluate the need for data transformations and to evaluate multicollinearity between variables. SAS and readability limitations prevented us from putting all variables into one matrix. Variables that we expect could exhibit multicollinearity are grouped into the same matrix.

The scatterplot of GDP shows a curved relationship with Life_Expectancy, which suggests the need for a transformation. Figure 1.5 shows the scatterplots before and after the log-transformation on GDP. The log-transformation resulted in a linear correlation with Life_Expectancy and appears to be appropriate.

Multicollinearity

The scatterplot matrices reveal strong multicollinearity between Thinnes_5_9 and Thinness_5_19 (Figure 1.3), and between HIV_new and HIV_perc (Figure 1.4). We removed Thinness_5_9 and HIV_new from further steps because they had the lower correlation with Life_Expectancy between the pair.

Correlation

Per the correlation heatmap (Figure 1.6), we can see that several variables display moderate to strong correlations with life expectancy. These ranges, however, are arbitrary and we will use the classifications of Dancey & Reidy where values between 0.10 - 0.39 are considered weak correlations, values between 0.40 - 0.60 are considered moderate correlations, values between 0.70 - 0.90 are considered strong correlations, and a value of 1.00 is a perfect correlation[2].

IV. Linear Regression Model

Problem

We seek to determine the main factors that affect life expectancy, taking into consideration demographics, GDP, mortality rates, and immunization variables.

[2] Akoglu, Haldun. “User's guide to correlation coefficients.” Turkish journal of emergency medicine vol. 18,3 91-93. 7 Aug. 2018, doi:10.1016/j.tjem.2018.08.001

Solution Outline

In this project we will use linear regression selection techniques to select the variables that are able to best predict (or explain) life expectancy. We use cross validation with 5 folds for each selection technique to help prevent overfitting.

Several variable selection metrics will be used to help identify the best model: SBC, CV Press, AIC/AICC, R^2 , and an Adjusted R^2. Typically, it is advised to choose one selection metric as different metrics can sometimes select different models. Our models do in fact show that this is the case. No one selection metric is unilaterally better than another and one could spend an entire project discussing the different pros and cons of each one.

Model Selection

To determine the best selection method, we ran forward, backward, stepwise, LASSO, and Ridge regression selection techniques. The stepwise selection technique had the lowest SBC. The forward selection technique had the lowest CV Press and AIC/AICC. The backward selection technique had the highest R^2 and Adjusted R^2. Variable selection did differ slightly between forward, backward and stepwise (the 3 best models). Ultimately, we chose to use the model from the forward selection technique with the lowest CV Press and AIC/AICC. (See Table 1 below for selection metric values).

Check Assumptions

Linear regression is a method to determine whether one or more predictor variables explain the dependent variable. Assumptions of a linear relationship between the response and explanatory variables, normality in the distribution of residuals, and constant variance in the residuals must be met for a valid linear regression model.

Linear Relationship

Per the residual plot (Figure 1.8, top left), the residuals appear randomly distributed around 0, which suggests a linear relationship and satisfies the assumption of linearity.

Normality

Per the histogram of the residuals and the QQ plot (Figure 1.8, middle left and bottom left), there does not appear to be a significant deviation from normality. We also have the Central Limit Theorem and a large enough sample to assume normality.

Constant Variance/Multicollinearity

There should be little to no multicollinearity among the predictors. Per the variance inflation factor (VIF), where a value of VIF>10 would indicate the presence of multicollinearity, the assumption is satisfied. All VIF values are less than 10 (Figure 1.7).

Comparison of Competing Models

A comparison of competing models led us to choose a forward selection model as it performed the best in terms of CV Press and AIC/AICC, as shown in Table 1 below.

Table 1: Selection Model Comparative Metrics

Selection Model	SBC	CV Press	AIC/AICC	R^2	Adjusted R^2
Forward	165.03	430.72	252.51/254.03	0.9098	0.9042
Backward	177.97	463.74	254.88/258.31	0.9145	0.9053
Stepwise	163.87	442.43	254.01/255.17	0.9067	0.9019
Ridge	358.03	760.12	456.10/456.51	0.3099	0.2963
LASSO	358.03	760.02	456.10/456.51	0.3099	0.2963

Parameter Interpretation

Figure 1.7 shows the parameter estimates and the respective confidence intervals.

Interpretation of Parameters

β_0 : **Intercept** - The intercept in the model provides an estimate of the life expectancy (69.03 years) of a person given all other predictor values are zero. A 95% confidence interval for the mean life expectancy, when all other predictor values are zero, is (62.46, 75.60) years.

β_1 : **under_5_deaths1** - Holding all other variables constant, it is estimated that for every 0.01% (of the total population) increase of under-5 deaths, life expectancy will decrease 35.82 years. A 95% confidence interval for the change in mean life expectancy is (-40.42, -31.22) years. Keep in mind this metric is evaluating under-5 deaths as a % of the population. To put this in perspective, the mean for under-5 deaths is 0.09% of the population.

β_2 : **health_expenditure_p** - Holding all other variables constant, it is estimated that for every 1% increase in health care expenditure (as a % of GDP), life expectancy will increase 0.33 years. A 95% confidence interval for the change in mean life expectancy is (0.17, 0.49) years.

β_3 : **alcohol** - Holding all other variables constant, it is estimated that for every 1 liter increase in alcohol consumption per person, life expectancy will decrease 0.15 years. A 95% confidence interval for the change in mean life expectancy is (-0.27, -0.04) years.

β_4 : **bmi** - Holding all other variables constant, it is estimated that for every 1 point increase in BMI, life expectancy will decrease 0.50 years. A 95% confidence interval for the change in mean life expectancy is (-0.72, -0.28) years.

β_5 : **log_gdp** - Holding all other variables constant, it is estimated that a doubling of GDP results in a 2.22*log(2) = .67 year increase in mean life expectancy. A 95% confidence interval for the change in mean life expectancy for each doubling of GDP is [1.80*log(2), 2.65*log(2)] = (0.54, 0.80) years.

β_6 : **tuberculosis1** - Holding all other variables constant, it is estimated that for every 0.01% (of the total population) increase in tuberculosis cases, life expectancy decreases 11.87 years. A 95% confidence interval for the change in mean life expectancy is (-14.59, -9.14) years. Keep in mind this metric is evaluating the number of incidents of tuberculosis as a % of the total population. To put this in perspective, the mean for tuberculosis cases is 0.12% of the population.

Conclusions

The forward selection technique identified the best model in our comparisons. We reviewed the assumptions for linear regression and our final model was appropriate. It was surprising that other infectious diseases did not play a larger role in explaining life expectancy, especially HIV. Several factors that we assumed should be in the model were in the final model: GDP, expenditure on healthcare, and population health markers such as BMI and alcohol consumption. Further research could take a repeated measures approach to see how (or if) countries have increasing life expectancy over time. Due to the interest of the group in time series analysis, we opted for a time series project for Objective 2 instead.

Sales Forecast - Objective 2

I. Introduction

Ten stores provided data for 50 different products and their sales for a 5-year period. How accurately can we forecast product sales for the next 30 days? We will forecast product sales using time series analysis.

II. Data Description

The dataset comes from Kaggle's Store Item Demand Forecasting Challenge[3]. The dataset includes daily data for 50 items at 10 different stores for the period 2013-2017. The dataset

[3] https://www.kaggle.com/c/demand-forecasting-kernels-only/overview

includes the date, store ID, item ID and number of sales. To simplify this in relation to our current classwork, we have chosen one store and one item to forecast in order to better understand our results and fit the project guidelines.

III. Exploratory Data Analysis

There is a slight increase of sales of the observed item as the years increase; as may be expected as economies and businesses generally scale upward, especially in our observed years. However, this trend is not so extensive to have considered this model additive. The shape of our time series model seems to have regular peaks and troughs in our data, and there is no irregular variation. There is a possible outlier in July of 2017; however, it does not seem too far out of line with previous trends.

A time series analysis is required in this instance because the assumption of independence required for linear regression is violated since the data was collected over time. The observations are correlated with one another and estimates and standard errors would be biased without correcting for this correlation. Before dealing with correlation, we need to determine whether the dataset is stationary.

Stationarity

Figure 2.1 shows the plot of the time series. There is clearly a seasonal pattern with higher sales in July than sales in January. There is also a day of the week pattern that is difficult to observe due to the amount of data. For a time series to be stationary it must have constant mean, constant variance, and constant autocorrelations.

ACF and PACF plots that do not show decreasing correlation over lags indicate nonstationarity. The middle, bottom plot in Figure 2.2 is an ACF plot that does not show correlation significantly decreasing over up to 25 lags. As such, we conclude this time series is nonstationary. For completeness, we will evaluate constant mean, constant variance and constant autocorrelation below.

Constant Mean

Figure 2.2 shows the residual diagnostic plots for fitting this time series, assuming stationarity. The plot in the upper left shows that the mean may not be constant. At the beginning of the series there are more residuals below 0 and at the end of the series there are more above 0.

Constant Variance

It is more difficult to assess constant variance from the upper left plot in Figure 2.2. The residuals appear to be smaller at the beginning of the series compared to the residuals at the end of the series, calling constant variance into question.

Constant Autocorrelation

There is no evidence from the residual diagnostic plots in Figure 2.2 to indicate this assumption is violated.

IV. Time Series Model

Time Series Analysis

To forecast this data in relation to what materials we have utilized in our course thus far, we evaluated an ARIMA model based on residual diagnostic plots.

Seasonality and Model Selection

After taking the first-order difference on sales, the ACF and PACF plots still indicated auto-regressive and moving average behavior (see Figure 2.3). The ACF plot shows lags in increments of 7 are still significant and the PACF plot shows lags up to 7 and at intervals of 7 are significant. The rule of thumb states that if both the ACF and PACF plots tail off gradually then an ARMA model is appropriate. In this case it would be ARIMA since we are already working with a first-order difference.

We ran the data series through SAS Forecast Server to help narrow down the auto-regressive and moving average values. The winning ARIMA model was ARIMA(7 14, 1, 1). The winning model from SAS Forecast Server was actually an Exponential Smoothing model, but that is outside the scope of this course. We took the ARIMA(7 14,1,1) model and were able to fine-tune it to produce a better result. Our final model was an ARIMA with p=(7 14 21 28 35), q=1 and d=1 with a 30-day holdout. We used the AIC selection criteria to choose the best model. Figure 2.4 shows the ACF and PACF plots for this model. The model was able to remove all significant lags from the ACF and PACF plots, indicating the model is appropriate. Figure 2.5 shows the parameter estimates for the moving average and auto-regressive components of the model.

Figure 2.6 shows predictions compared to actuals for the last 30 days of data (holdout sample).

Conclusions

Our analysis determined the product sales time series has observations that are correlated with one another. The time series was also nonstationary. There were cyclic and seasonal patterns present in the data. We used an ARIMA model to fit the data, account for nonstationarity with a first-order difference and include the observed day of week patterns. While not included in the dataset provided on Kaggle, further research could include other model types (such as Exponential Smoothing) and other explanatory variables. Another possible avenue to research to increase predictions would be to include ‘events’ such as holidays or sales promotions. We might also have to transform the explanatory variables to make them stationary as nonstationary data can produce spurious relationships and give us inaccurate predictions.

Appendix

Life Expectancy - Objective 1

Figure 1.1 - Variable Descriptions

Variable	Source	Description
country	World Health Organization	
year	World Health Organization	
life_expectancy	World Health Organization, Life Expectancy	Life expectancy at birth (years)
status	https://www.kaggle.com/augustus0498/life-expectancy-who	Developed or Developing Country
adult_mortality	World Health Organization, Adult Mortality	Adult mortality rate (probability of dying between 15 and 60 years per 1000 population)
under_5_deaths	World Health Organization, Child Mortality	Number of under-five deaths (thousands)
infant_deaths	World Health Organization, Child Mortality	Number of infant deaths (thousands)
neonatal_deaths	World Health Organization, Child Mortality	Number of neonatal deaths (thousands)
health_expenditure_perc	World Heath Organization, Health Financing	Current health expenditure (CHE) as percentage of gross domestic product (GDP) (%)
tuberculosis	World Health Organization, Tuberculosis	Number of incident tuberculosis cases
bmi	World Health Organization, Body Mass Index (BMI)	Mean BMI (kg/m^2) (age-standardized estimate)
alcohol	World Health Organization, Global Information System on Alcohol and Health	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
diphtheria	World Health Organization, Immunization	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
hepatitis	World Health Organization, Immunization	Hepatitis B (HepB3) immunization coverage among 1-year-olds (%)
measles1	World Health Organization, Immunization	Measles-containing-vaccine first-dose (MCV1) immunization coverage among 1-year-olds (%)

measles2	World Health Organization, Immunization	Measles-containing-vaccine second-dose (MCV2) immunization coverage by the nationally recommended age (%)
polio	World Health Organization, Immunization	Polio (Pol3) immunization coverage among 1-year-olds (%)
obesity	World Health Organization, Body Mass Index (BMI)	Prevalence of obesity among adults, BMI >- 30 (age-standardized estimate)(%)
population	World Bank, SP.POP.TOTL	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.
thinness_18_plus	World Health Organization, Body Mass Index (BMI)	Prevalence of underweight among adults, BMI < 18 (age-standardized)(%)
thinness_5_9	World Health Organization, Body Mass Index (BMI)	Prevalence of thinness among children and adolescents ages 5-19, BMI <- 2 standard deviations below the median (crude estimate)(%)
thinness_5_19	World Health Organization, Body Mass Index (BMI)	Prevalence of thinness among children and adolescents ages 5-9, BMI <- 2 standard deviations below the median (crude estimate)(%)
gdp	World Bank, NY.GDP.PCAP.CD	GDP per capita is gross domestic product divided by midyear population. Data are in current U.S. dollars.
hiv_new	World Bank, SH.HIV.INCD.ZS	Incidence of HIV: # of new HIV infections among uninfected populations ages 15-49 expressed per 1,000 uninfected population in the year before the period.
hiv_perc	World Bank, SH.DYN.AIDS.ZS	Prevalence of HIV: % of population ages 15-49 who are infected with HIV.

Figure 1.2 - Scatterplot Matrix 1

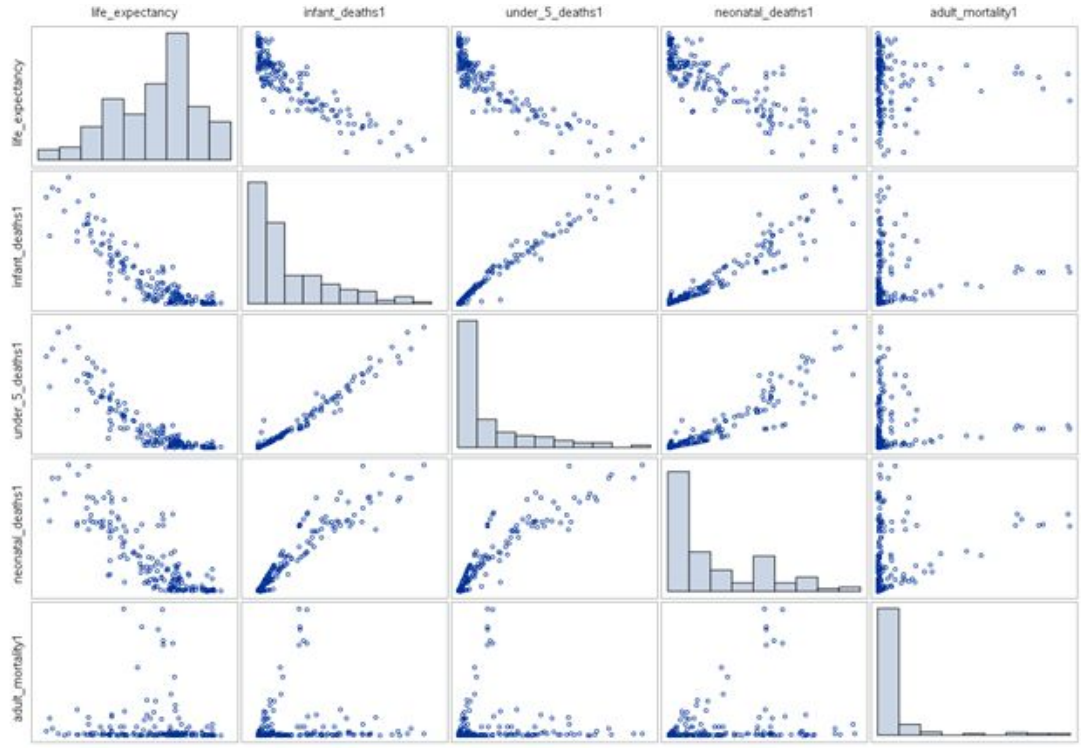


Figure 1.3 - Scatterplot Matrix 2

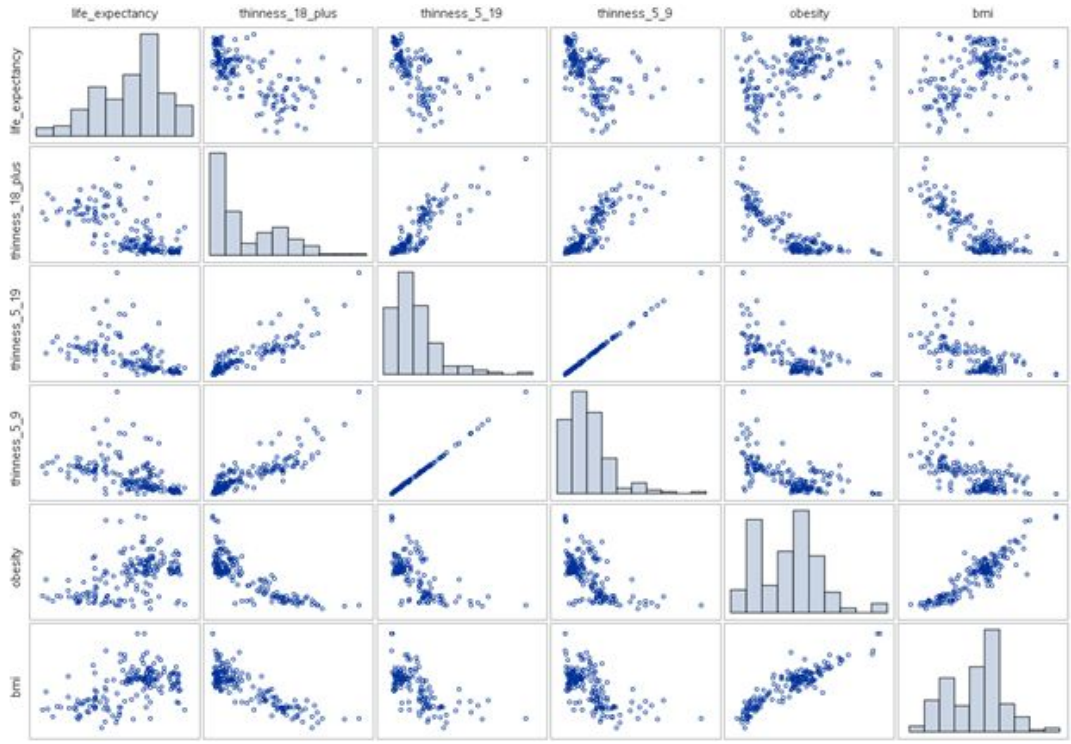


Figure 1.4 - Scatterplot Matrix 3

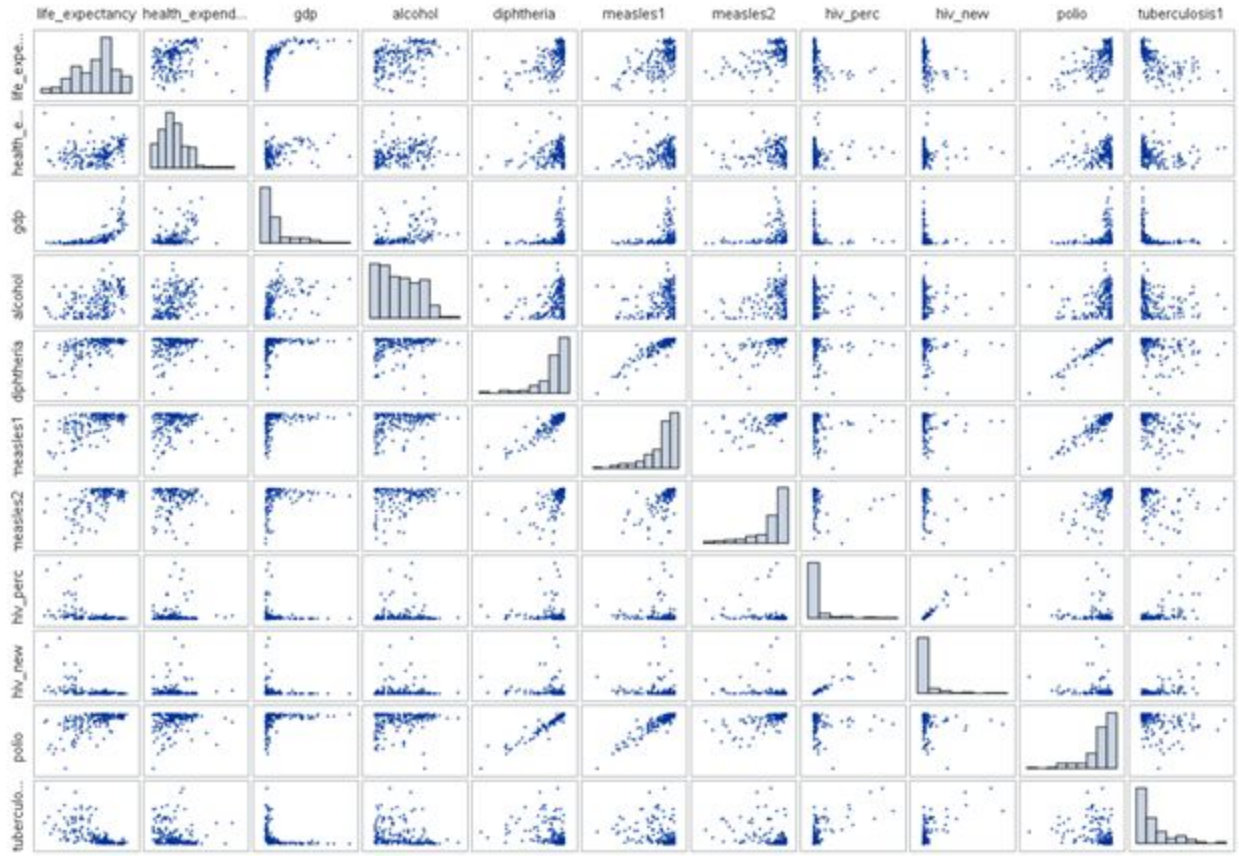


Figure 1.5 - GDP Before and After LOG Transformation

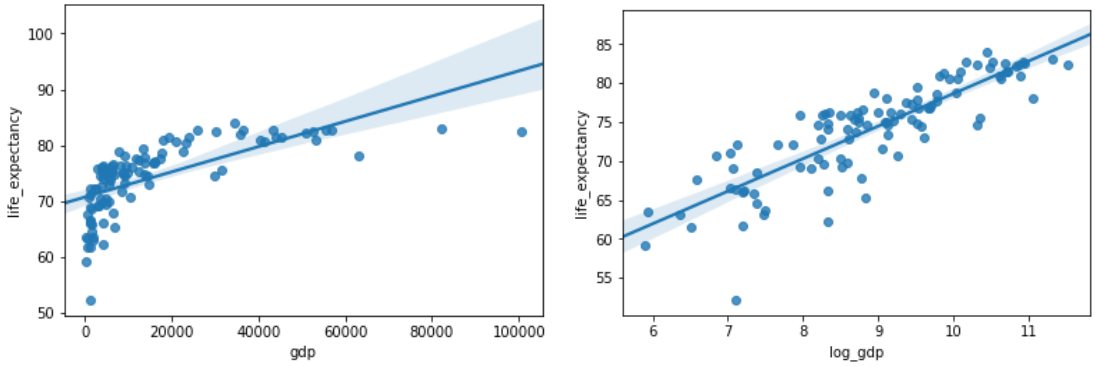


Figure 1.6 - Correlation Heatmap

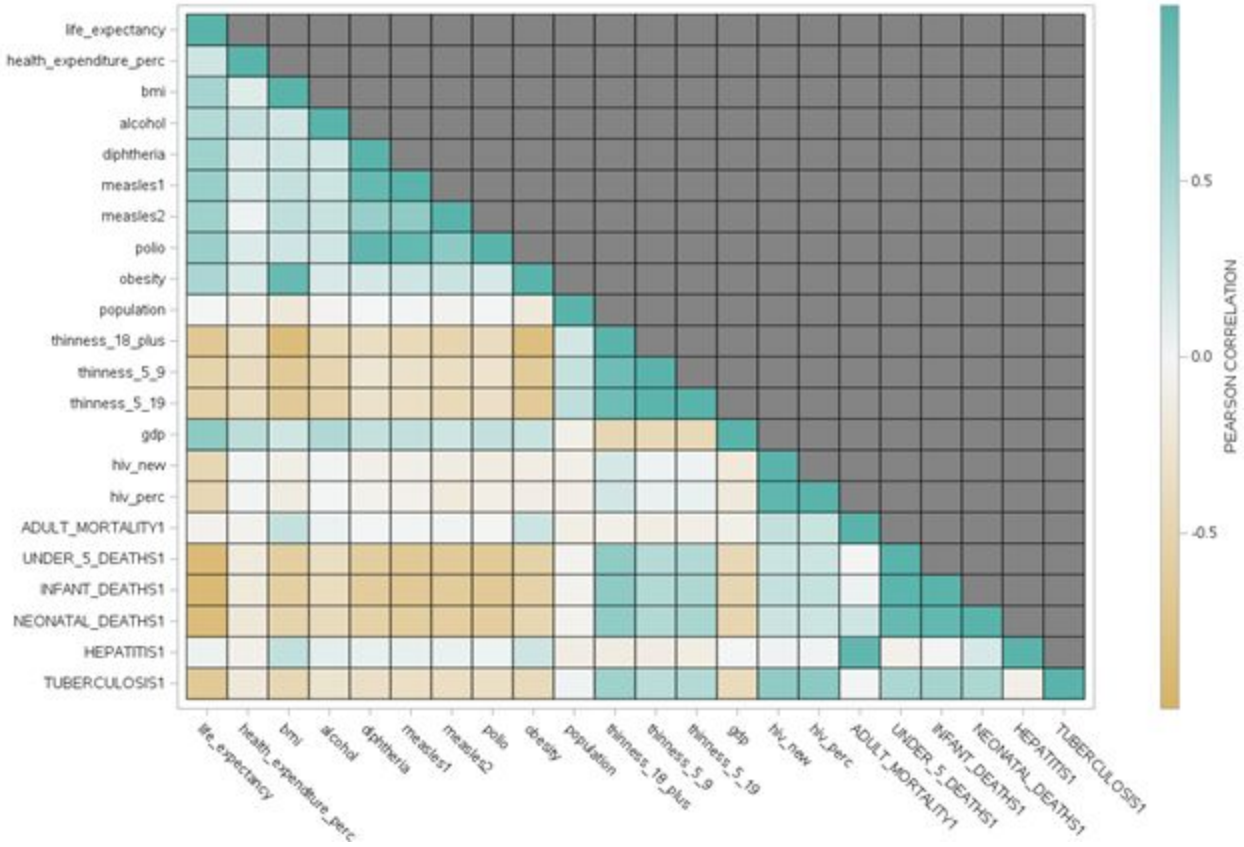
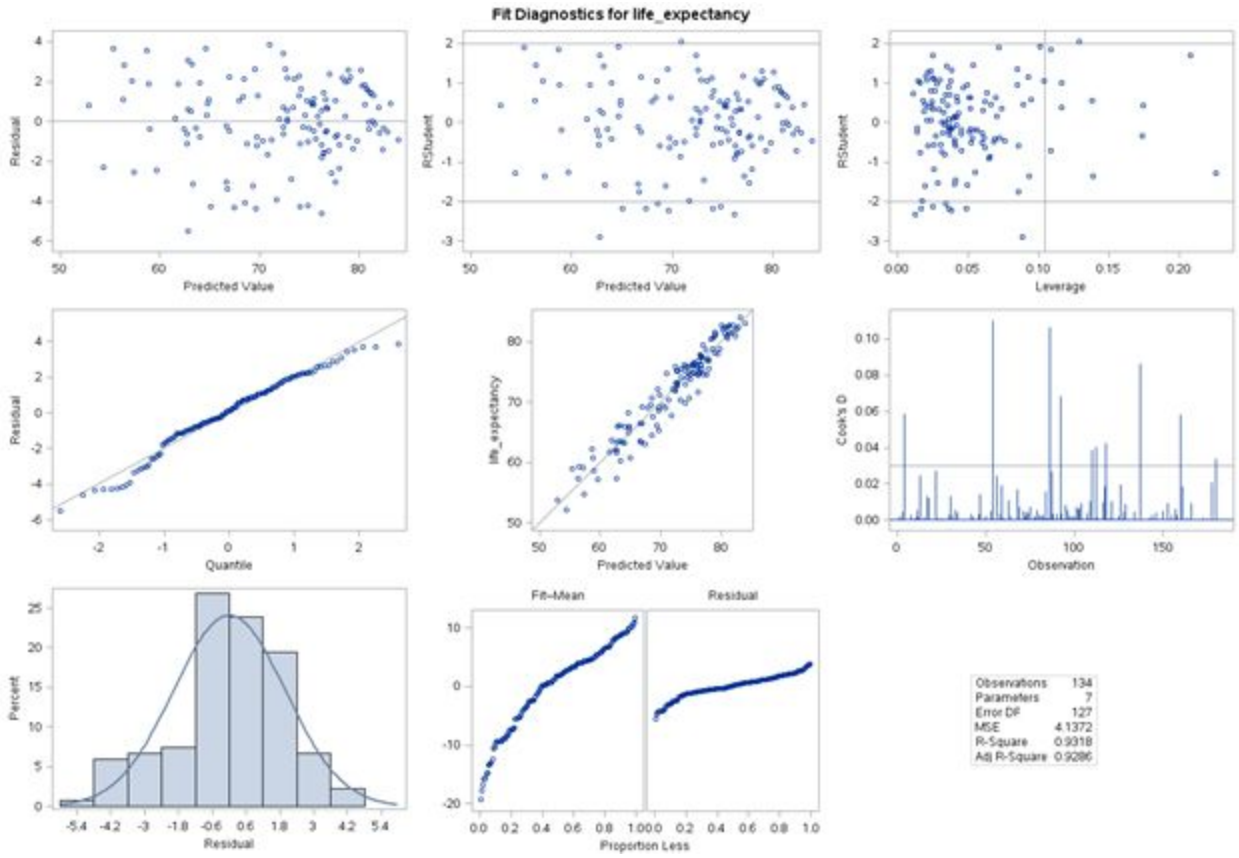


Figure 1.7 - Final Model Parameter Estimates

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	69.02940	3.32004	20.79	<.0001	0	62.45964	75.59917
UNDER_5_DEATHS1	1	-3581.67445	232.44932	-15.41	<.0001	2.34300	-4041.64970	-3121.69920
health_expenditure_perc	1	0.33143	0.08073	4.11	<.0001	1.20797	0.17169	0.49118
alcohol	1	-0.15110	0.05783	-2.61	0.0101	1.67878	-0.26554	-0.03666
bmi	1	-0.50344	0.11120	-4.53	<.0001	1.68828	-0.72348	-0.28339
LOG_GDP	1	2.22226	0.21471	10.35	<.0001	2.97297	1.79739	2.64714
TUBERCULOSIS1	1	-1186.80302	137.78705	-8.61	<.0001	1.54600	-1459.45873	-914.14731

Figure 1.8 - Final Model Residual Diagnostics



Product Sales - Objective 2

Figure 2.1 - Daily Product Sales 2013-2017

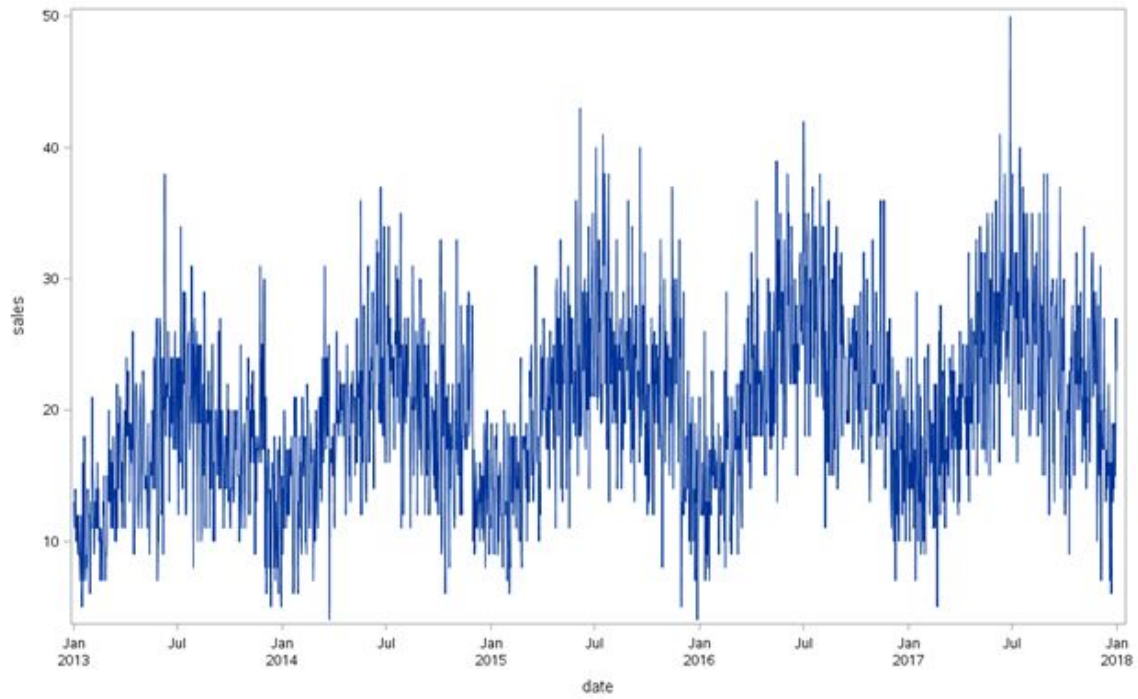


Figure 2.2 - Residual Diagnostics to Determine Stationarity

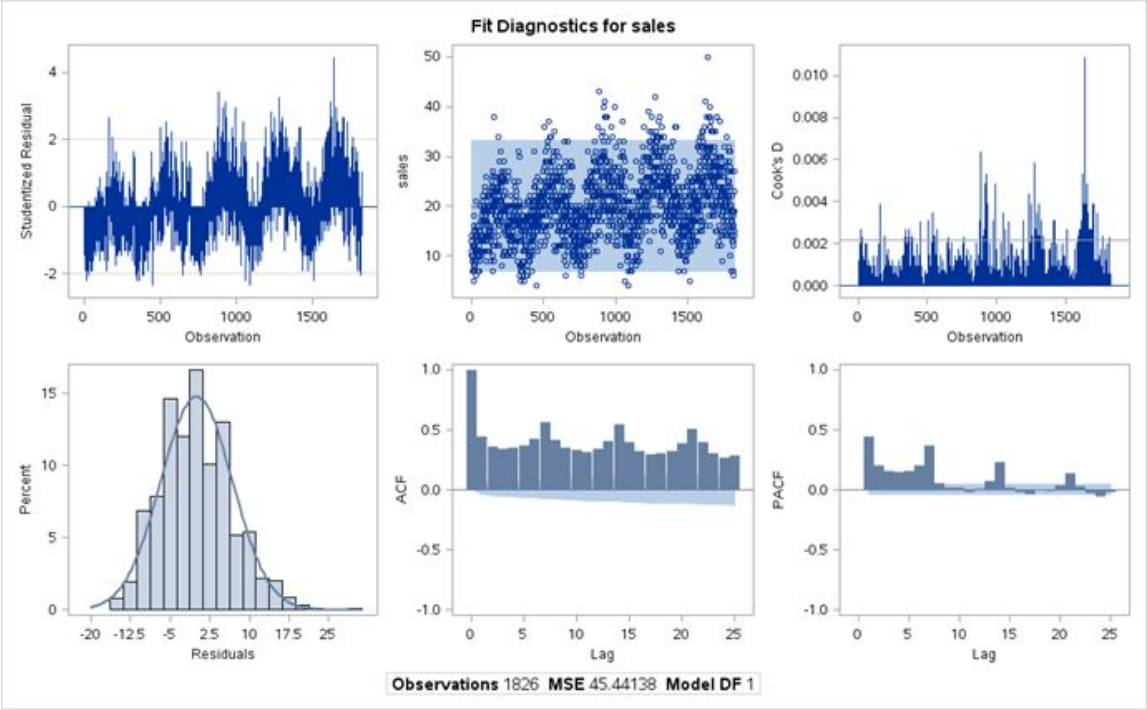


Figure 2.3 - Residual Diagnostics after First-Order Differencing

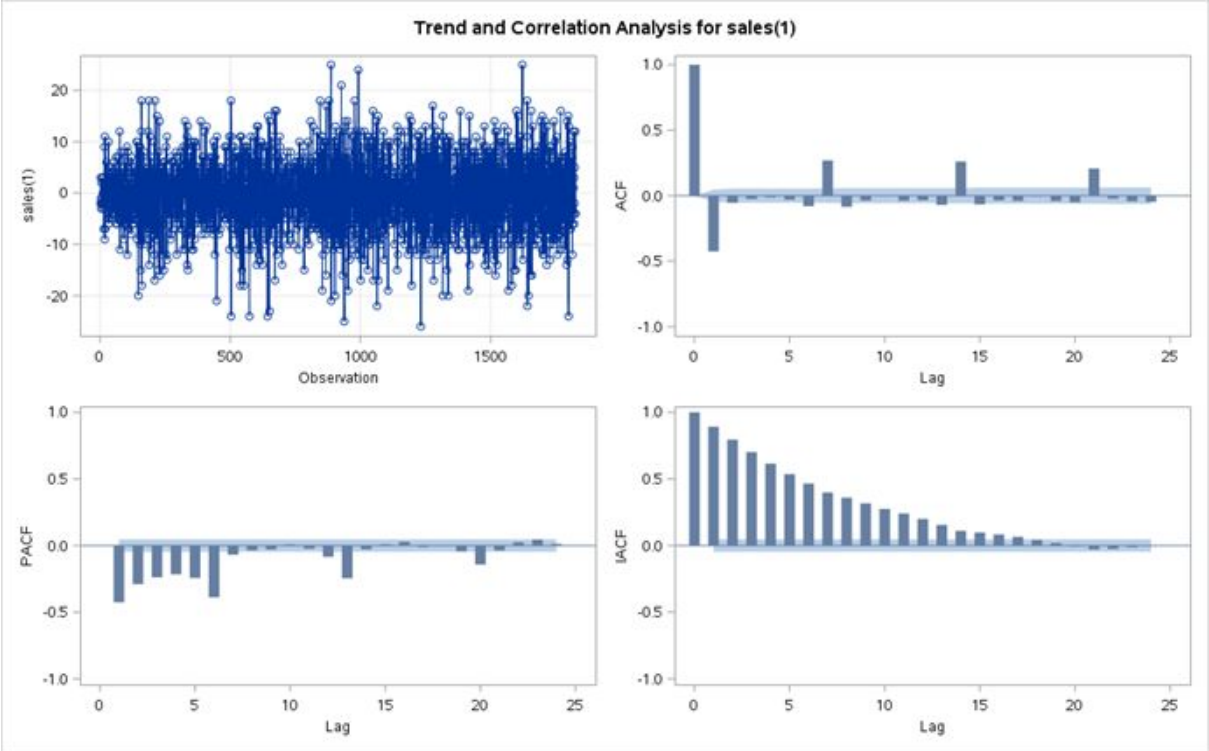


Figure 2.4 - Residual Diagnostics 1 for ARIMA Model

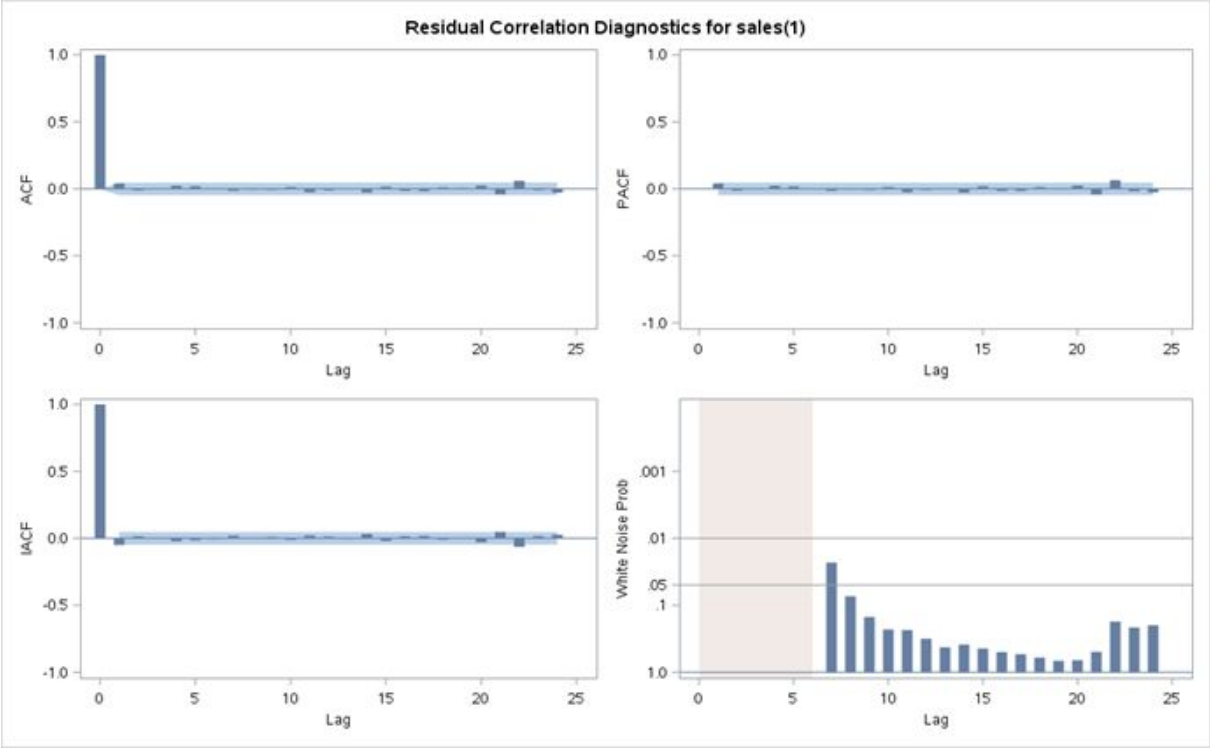
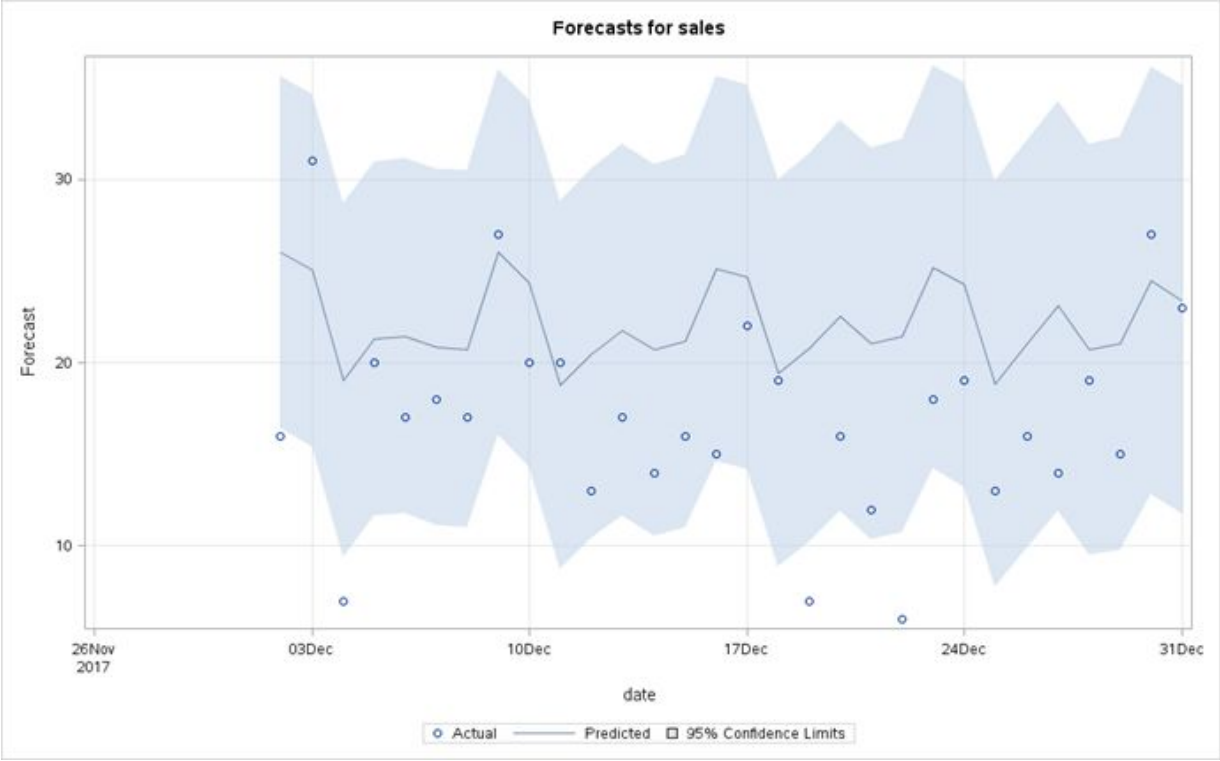


Figure 2.5 - ARIMA Model Parameter Estimates

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.0008206	0.02377	0.03	0.9725	0
MA1,1	0.92807	0.0090944	102.05	<.0001	1
AR1,1	0.15954	0.02414	6.61	<.0001	7
AR1,2	0.15216	0.02359	6.45	<.0001	14
AR1,3	0.10286	0.02370	4.34	<.0001	21
AR1,4	0.13077	0.02364	5.53	<.0001	28
AR1,5	0.11401	0.02382	4.79	<.0001	35

Figure 2.6 - ARIMA Model 30-Day Forecast


Forecasts for sales
Forecast
30
20
10
26Nov
2017
03Dec
10Dec
17Dec
24Dec
31Dec
date
Actual
Predicted
95% Confidence Limits