Arias, Heidi H.

Cancer Incidence

INTRODUCTION

Cancer, also known as *malignant tumor* or *malignant neoplasm*, is a group of almost a hundred diseases that are characterized by the uncontrolled growth of some abnormal cells in a body and the ability of these cells to spread from its original area to other parts such that if the spread is left uncontrolled, could result in death.

Cancer is one of the most common *dreaded diseases* in the world today. It often has a huge impact one on the person's lifestyle and longevity and is also expensive to treat. There are various types of cancers and also many known causes – environmental factors like tobacco usage, diet and obesity, infections, radiation, and inherited genetics - although it is said that the actual cause of a cancer in an individual is nearly impossible to pinpoint, since most of the cases have multiple possible causes.

This study aims to illustrate how some of the known causes of cancer are related to its incidence among several countries around the world.

SOURCES AND DISCLAIMER

1. The data, definitions and descriptions were obtained from the following sources and were used for the purpose of this study only.

- http://stats.oecd.org/
- https://en.wikipedia.org/wiki/Cancer
- http://medical-dictionary.thefreedictionary.com

2. This study has been done for the specific purpose of statistical data analysis student project and should not be taken as an actual medical study.

AN OVERVIEW OF THE CANCER INCIDENCE OF DIFFERENT COUNTRIES

In this study, we used the data as of 2012 we obtained from *Organisation for Economic Co-operation and Development (OECD)*. We will look into the relationship between cancer incidence and some of its causes – obesity, alcohol consumption and tobacco consumption - among several countries. To simplify our study, we trimmed it down to include 29 countries that have the updated data on the causes we will look into.

VARIABLES

We will use the variables as defined below:

N = number of countries in the study = 29

Response Variable
Y = Cancer, Incidence of Malignant Neoplasms, per 100,000 population

Quantitative Explanatory Variable
X_1 = Obese population, self-reported, % of total population
X_2 = Alcohol consumption, Liters per capita (age 15+)
X_3 = Tobacco consumption, % of population 15+ who are daily smokers

X_1, X_2 and X_3 are assumed to be independent from each other.

OUTCOME AND ANALYSIS

DATA SUMMARY

Table 1: Data Summary

	Country	Y - Cancer Incidence	Country	X1 - Obesity	Country	X2 - Alcohol Consumption	Country	X3 - Tobacco Consumption
Highest	Denmark	338.10	United States	28.70	Austria	12.20	Greece	38.90
Lowest	Germany	163.00	Italy	2.40	Switzerland	1.40	Spain	10.70
Mean		275.86		15.50		8.88		20.41
Standard Dev		42.42		4.79		2.52		5.89
Upper Bound		318.28		20.29		11.40		26.30
Lower Bound		233.43		10.71		6.36		14.53
No of Countries within Bound		21.00		22.00		21.00		22.00

From *Table 1* above, we see that those countries that have the highest and lowest in cancer incidence among the 29 countries, which are Denmark and Germany respectively, are not the highest and lowest in terms of the causes X_1, X_2 and X_3.

We computed for the mean and standard deviation of the causes and found that both countries are within bound, that is, within one standard deviation of the mean.

Country	Y - Cancer Incidence	X1 - Obesity	X2 - Alcohol Consumption	X3 - Tobacco Consumption	Within Bound Y	Within Bound X1	Within Bound X2	Within Bound X3
Denmark	338.10	14.20	9.50	17.00	no	yes	yes	yes
Germany	283.80	15.70	10.90	20.90	yes	yes	yes	yes

RESULTS AND ANALYSIS

Let us now check the relationship between our response variable Y – Cancer Incidence and our explanatory variables $X_1 - Obesity$, $X_2 - Alcohol\ Consumption$ and $X_3 - Tobacco\ Consumption$ by running several scenarios and see the resulting model and graphs for each scenario.

Scenario 1

In this scenario we will analyze the relationship between the cancer incidence rates and all the independent variables $X_1, X_2\ and\ X_3$. The results are as follows:

Table 2

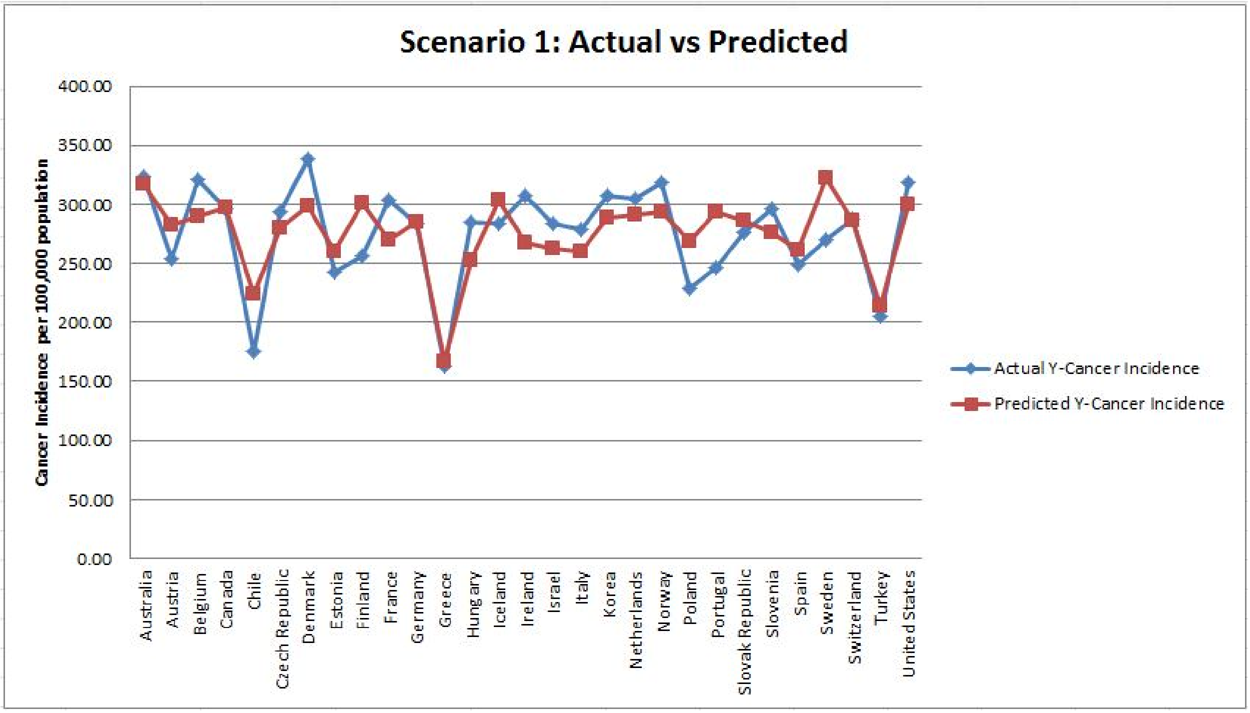
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.753141401							
R Square	0.56722197							
Adjusted R Square	0.515288606							
Standard Error	29.53618446							
Observations	29							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	28584.89691	9528.298969	10.92211116	9.02822E-05			
Residual	25	21809.65482	872.3861926					
Total	28	50394.55172						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	346.0331543	31.76052312	10.895071	5.53025E-11	280.6211325	411.4451761	280.6211325	411.4451761
X1 - Obesity	-0.83259235	1.167449406	-0.713172105	0.48234393	-3.23699941	1.571814709	-3.23699941	1.571814709
X2 - Alcohol Consumption	5.707071217	2.252107806	2.534102143	0.017913585	1.068768365	10.34537407	1.068768365	10.34537407
X3 - Tobacco Consumption	-5.28879983	0.964086959	-5.485812023	1.06623E-05	-7.27437409	-3.303225571	-7.27437409	-3.303225571

The Scenario 1 will then be modeled as:

$$\boldsymbol{Y = 346.03 - 0.83259X_1 + 5.70707X_2 - 5.2888X_3}$$

Using the equation above, we computed for the actual vs. predicted cancer incidence rates of the countries. Then, using *Graph 1* below to see it better, we see that the predicted is quite far from the actual incidence rates, except on certain points.

Graph 1



Based on the resulting regression, the adjusted R^2 value is at 51.53% which shows that the model has high variation between the causes of cancer incidence.

Let us now look at each of the causes separately in the next 3 scenarios.

Scenario 2

This scenario shows the relationship between Y and X_1.

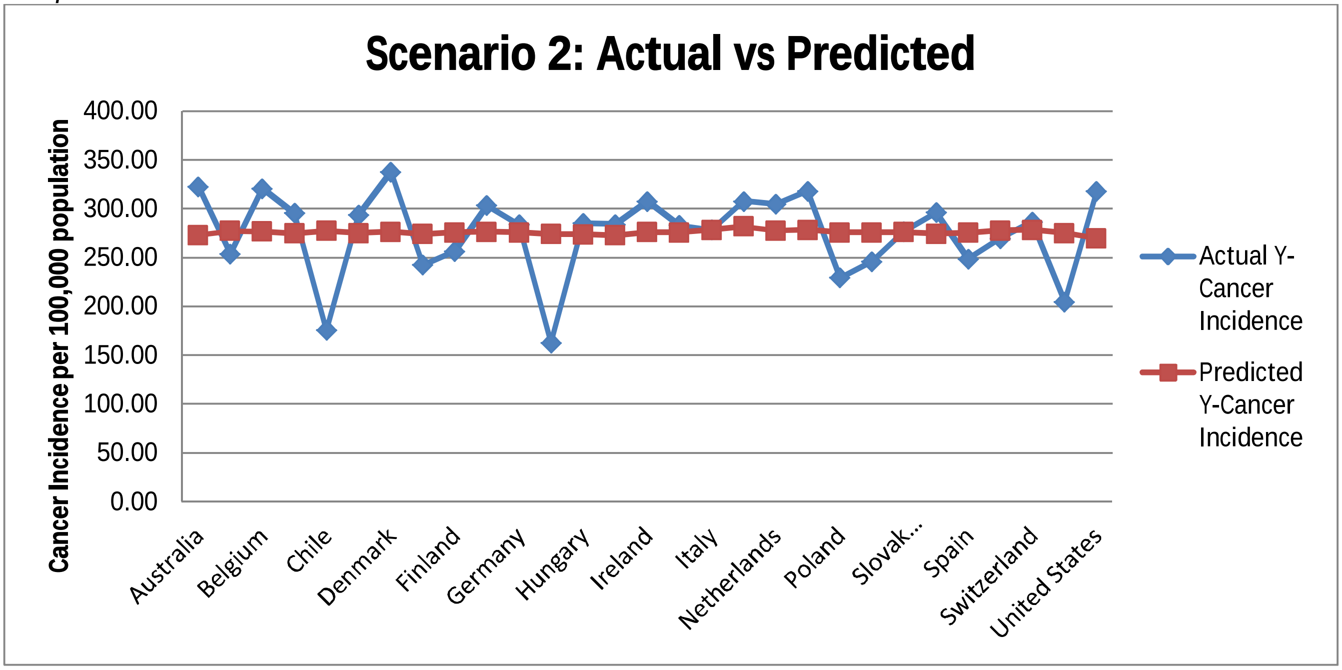
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.051569392							
R Square	0.002659402							
Adjusted R Square	-0.034279138							
Standard Error	43.14511789							
Observations	29							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	134.0193838	134.0193838	0.071995325	0.790492887			
Residual	27	50260.53234	1861.501198					
Total	28	50394.55172						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	282.93647	27.58061148	10.25852781	8.18826E-11	226.3457297	339.5272103	226.3457297	339.5272103
X1 - Obesity	-0.456756293	1.702285464	-0.268319446	0.790492887	-3.949557556	3.036044971	-3.949557556	3.036044971

The equation for Scenario 2 model will then be given by:

$$Y = 282.936 - 0.45676X_1$$

Again, the graph using the above formula is shown below.

Graph 2



This model shows that the actual cancer incidence rates is very far from the predicted rates. R square is also very low (less than 1%), so we can conclude that this is not a very good model. We can conclude from this model that the variable X_1 by itself will not cause a high change in cancer incidence rates.

Let us look at the effect of the next variable.

Scenario 3

This scenario shows the relationship between Y and X_2.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.206846824
R Square	0.042785609
Adjusted R Squa	0.007333224
Standard Error	42.26827466
Observations	29

ANOVA

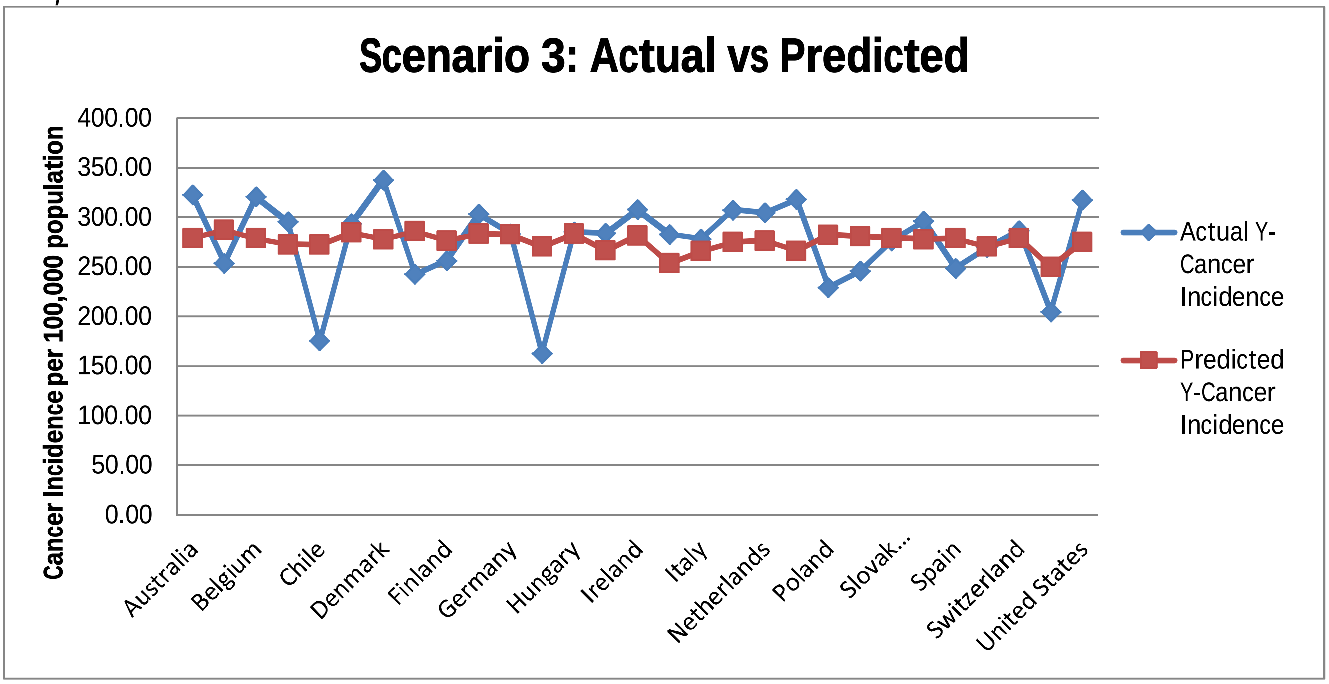
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2156.161562	2156.161562	1.206847119	0.281657512
Residual	27	48238.39016	1786.607043		
Total	28	50394.55172			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	244.9222503	29.23105534	8.378837078	5.45664E-09	184.9450789	304.8994217	184.9450789	304.8994217
X2 - Alcohol Cor	3.48235536	3.16991019	1.098565937	0.281657512	-3.021763103	9.986473823	-3.021763103	9.986473823

Equation for Scenario 3:

$$Y = 244.922 + 3.482355X_2$$

Graph 3



The adjusted R^2 value is still significantly lower than Scenario 1, but a little higher than Scenario 2. We can observe this in the way the graph of the Actual vs Predicted Incidence rates are moving a little more similarly with each other than in Scenario 1. With this, we can conclude that the by itself, alcohol consumption does not directly or significantly affect cancer incidence rates.

Scenario 4

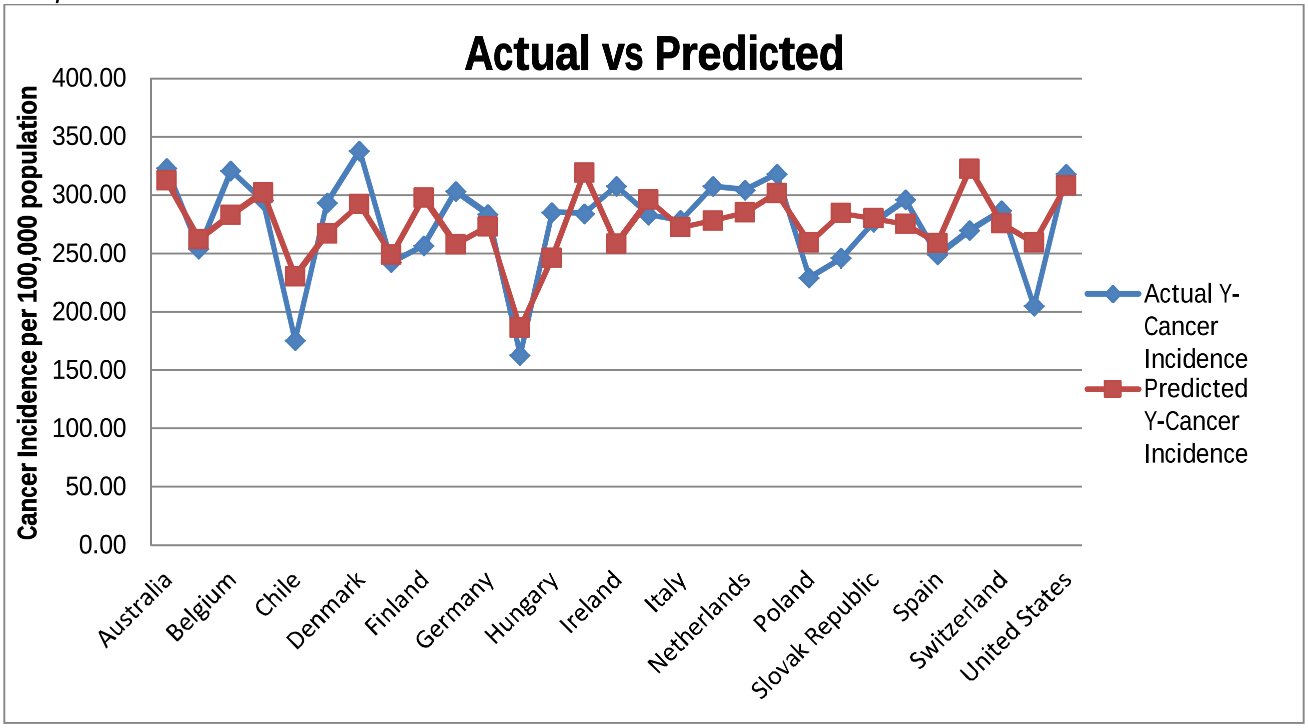
This scenario shows the relationship between Y and X_3.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.670630777							
R Square	0.449745639							
Adjusted R Squa	0.429365848							
Standard Error	32.04731563							
Observations	29							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	22664.72986	22664.72986	22.06821628	6.86376E-05			
Residual	27	27729.82186	1027.030439					
Total	28	50394.55172						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	374.4962259	21.8248262	17.15918479	4.75917E-16	329.7153815	419.2770703	329.7153815	419.2770703
X3 - Tobacco Coı	-4.832078633	1.028609136	-4.697682011	6.86376E-05	-6.942610246	-2.721547019	-6.942610246	-2.721547019

Equation for Scenario 4:

$$Y = 374.496 - 4.832X_3$$

Graph 4



The adjusted R^2 is now much higher 42.94% which leads us to conclude that Scenario 4 is a better model for this study, but not as good as Scenario 1, which includes all 3 variables. This leads us to conclude that the variable X_3 has higher significance to the movement of cancer incidence rates than the other two variables.

Let us create 3 more scenarios to see this further. The following scenarios show 2 explanatory variables X_i are modelled with our response variable Y.

Scenario 5

This scenario shows the relationship between Y, X_1 and X_2.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.215078536
R Square	0.046258777
Adjusted R Square	-0.027105933
Standard Error	42.99524187
Observations	29

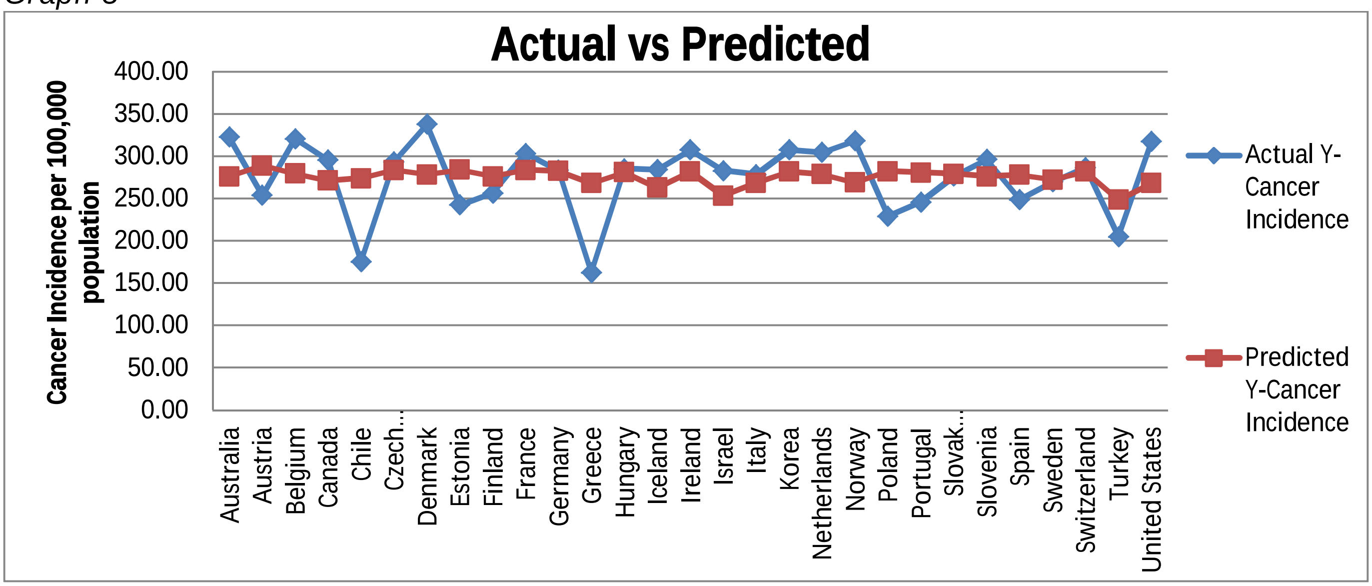
ANOVA

	df	SS	MS	F	Significance F
Regression	2	2331.190314	1165.595157	0.630531723	0.540253026
Residual	26	48063.36141	1848.590823		
Total	28	50394.55172			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	252.7074612	39.04144653	6.472799645	7.35396E-07	172.4566186	332.9583039	172.4566186	332.9583039
X1 - Obesity	-0.522309685	1.697437443	-0.307704821	0.760758408	-4.011442319	2.966822949	-4.011442319	2.966822949
X2 - Alcohol Consu	3.51752352	3.226454049	1.090213425	0.285620435	-3.11454776	10.1495948	-3.11454776	10.1495948

As we have observed from Scenarios 2 and 3, we can see that the adjusted R square is very low and that the values will behave more similarly to Scenario 3. Hence, this is not a very good model for our study. We can see this more clearly through the graph below.

Graph 5



Scenario 6

This scenario shows the relationship between Y, X_2 and X_3.

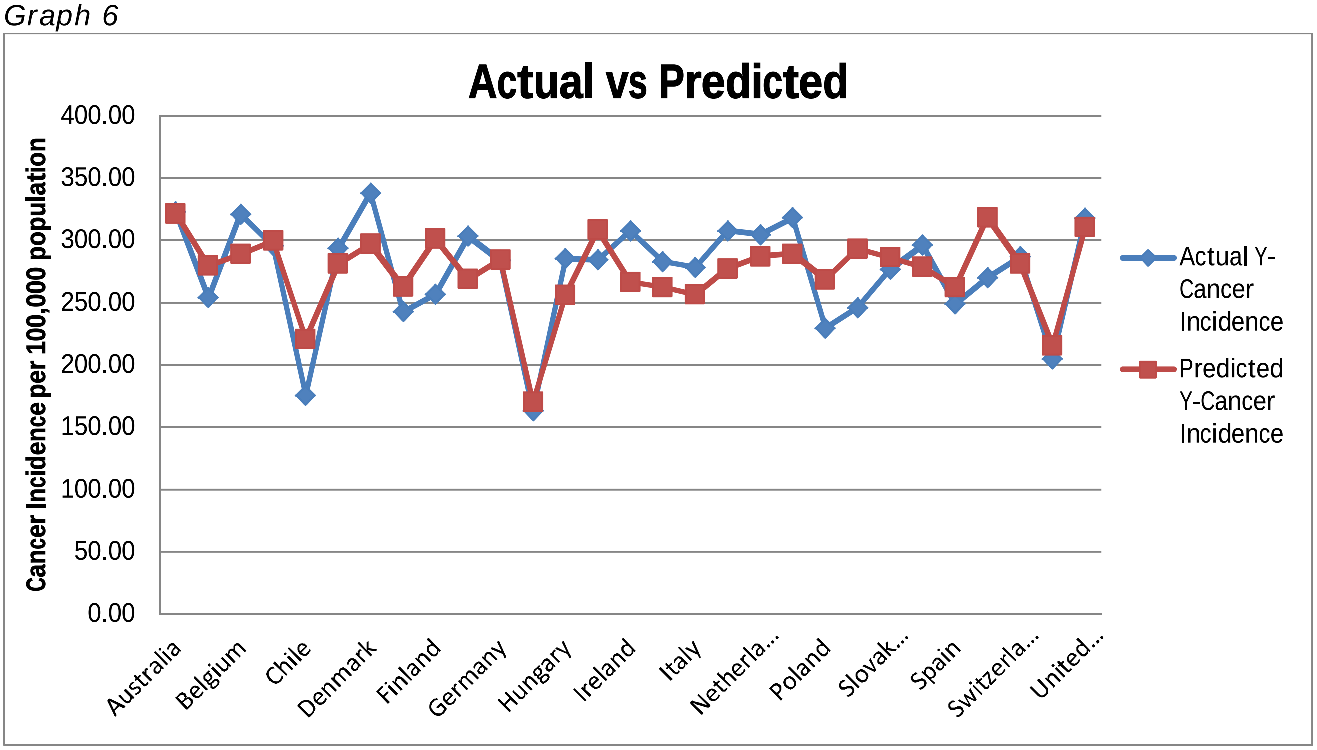
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.747273232							
R Square	0.558417284							
Adjusted R Squa	0.524449382							
Standard Error	29.25574437							
Observations	29							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	28141.18868	14070.59434	16.43955802	2.42749E-05			
Residual	26	22253.36304	855.8985785					
Total	28	50394.55172						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	333.0643949	25.79216672	12.91339337	8.10797E-13	280.0478369	386.0809529	280.0478369	386.0809529
X2 - Alcohol Cor	5.637351986	2.228621888	2.529523746	0.017822821	1.056354088	10.21834988	1.056354088	10.21834988
X3 - Tobacco Coi	-5.255488722	0.953811762	-5.50998523	8.79423E-06	-7.216076879	-3.294900566	-7.216076879	-3.294900566

The high adjusted R square 52.44% is an indicator that this is a good model for this study. We note that this is even higher than Scenario 1, which was our best model so far.

Equation for Scenario 6:

$$Y = 333.064 + 5.63735X_2 - 5.25549X_3$$

Graph 6



Scenario 7

The last scenario shows the relationship between Y, X_1 and X_3.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.675318887
R Square	0.456055599
Adjusted R Square	0.414213722
Standard Error	32.47000643
Observations	29

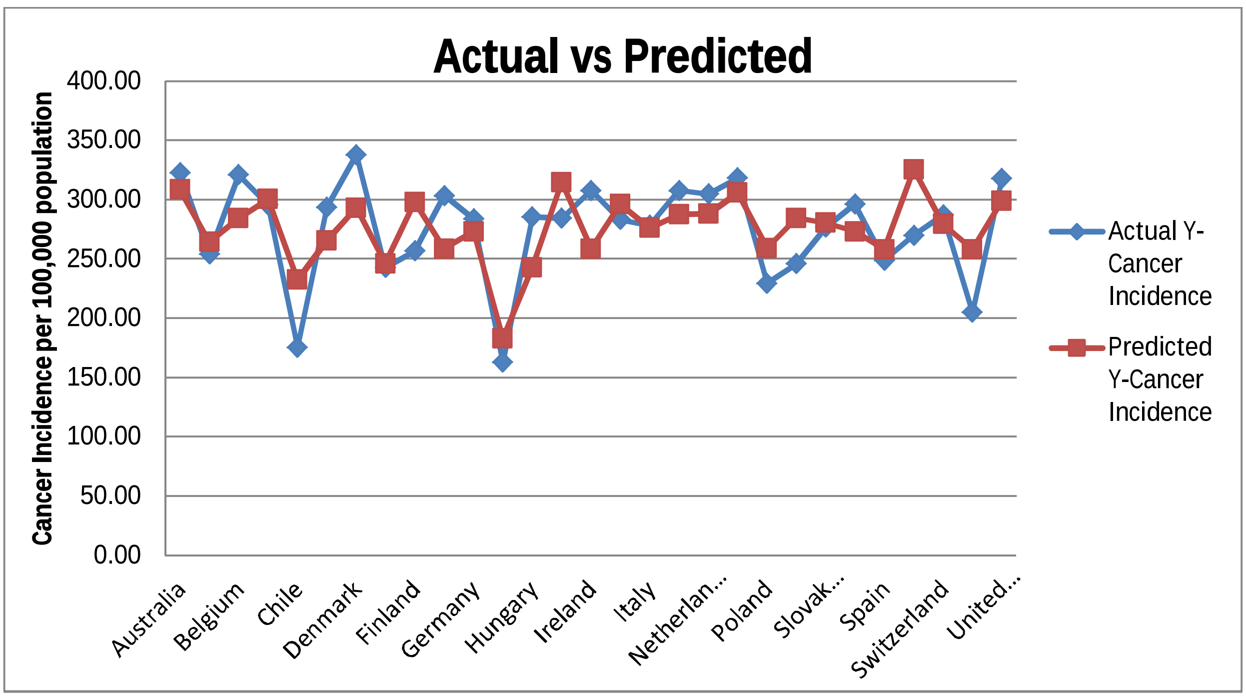
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	22982.71747	11491.35873	10.89950144	0.000364929
Residual	26	27411.83426	1054.301318		
Total	28	50394.55172			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	385.898045	30.33144076	12.72270737	1.13437E-12	323.5508756	448.2452144	323.5508756	448.2452144
X1 - Obesity	-0.704172769	1.282202166	-0.549190126	0.587561217	-3.339777067	1.931431529	-3.339777067	1.931431529
X3 - Tobacco Cons	-4.85582302	1.043072498	-4.655307306	8.3495E-05	-6.999889246	-2.711756794	-6.999889246	-2.711756794

Although the adjusted R square is significantly high, it is not as high as Scenario 1 and Scenario 6.

Graph 7



As expected, the values do not fit as well as in Scenario 6.

CONCLUSION

Therefore, based on the resulting scenarios, we may take out the variable X_1 - Obesity from the model since this variable does not significantly affect the cancer incidence rates.

It is conservative to conclude that statistically, the best predictive model is from Scenario 6, which is given by:

$$\boldsymbol{Y = 333.064 + 5.63735X_2 - 5.25549X_3}$$

This model has the highest adjusted R^2 value and has independent variables with P values close to zero.