A comparison of Alzheimer's across five states

by

Travis Keep, Danielle Novinski, Valli Meenakshi Sundaram, and Jamielee Jimenez

Introduction

In this report, we seek to compare the effects of various factors on Alzheimer's death rate across five states. The factors we consider include smoking, inactivity, obesity, age greater than 65, heart disease, diabetes and cancer. The five states we consider are California, Texas, Florida, New Jersey, and Washington. We chose the states in such a way that there is both geographic and population diversity. We apply various statistics methods and modelling to the dataset. The rest of the report discusses these findings.

Analysis of Variance

Statistical methods such as the Anova F-test assume that the data being analyzed are normally distributed and that the data across the five states have equal variances. Therefore we test each factor for normality, and we check whether or not each factor has equal variance across the five states.

We tested for normality of Alzheimer's death rate as well as the other factors across the five states using the Shapiro-Wilk test at alpha = 0.05. The null hypothesis is that the data are normally distributed; the alternative hypothesis is that the data is not normally distributed. If the p-value from Shapiro-Wilk is less than 0.05, we can reject the null hypothesis and conclude that the data is not normally distributed; if the p-value from Shapiro-Wilk is greater than 0.05, we cannot reject the null hypothesis, and we must conclude that the data are normally distributed. We conclude that the smoking variable is normally distributed across the five states, but we conclude that all other variables are not normally distributed across the five states.

We used Hartley's test to see if smoking has equal variances across the five states. Hartley assumes that the data are normally distributed and that the sample sizes are equal. We used the Brown-Forsythe test for the other variables since they are not normally distributed. We carried out both tests at a significance level of 0.05. The null hypothesis is the variances are equal; the alternate hypothesis is that at least two variances differ. If the P-value from the test is less than 0.05. We can reject the null hypothesis that all variances are equal. We conclude that none of the variables have equal variances across the five states.

We applied the Anova F-test to the smoking variable since it is normally distributed. We applied the Anova F-test with normal quantiles as well as the Kruskal-Wallis Test to the rest of the variables since they are not normally distributed. We used significance level 0.05 for our tests. The null hypothesis is that the variable selected has equal means across all five states; the alternate hypothesis is that at least two states have different means for the variable selected. For all variables we rejected the null hypothesis. That is, for each variable the means differ in at least two states.

Based on the results of the Anova F-test, we apply post-hoc analysis to see where the differences are. We use Tukey's test, Duncan's test, SNK test, and Fisher's LSD test.

From Tukey's test, we conclude that the Alzheimer's death rate was significantly different between Florida and Texas, Florida and Washington, and Florida and New Jersey. The cancer, smoking and diabetes rate in Florida was significantly different from all the other 4 states. The inactivity rate in California is significantly different from all the four states.
According to Bonferroni's test, the Alzheimer's death rate is not significantly different between Washington, Texas, and California; Texas, California, and New Jersey; and California, New Jersey, and Florida. Cancer rates in Texas, Washington, New Jersey, and California are not significantly different from Florida. For the heart variable, Texas is significantly different from California and Washington, and Florida and Washington are also significantly different from each other. For diabetes, all states are significantly different from each other, with the exception of Washington and New Jersey. Obesity in California and New Jersey is significantly lower than Washington, Texas, and Florida. For age greater than 65, New Jersey, California, Texas, and Washington are not significantly different from each other, and Texas, Washington and Florida are not significantly different from each other but are higher than the other states. Inactivity in Texas and Florida is significantly different from California, Washington and New Jersey. The smoking variable is significantly different in Florida from all other groups.

Duncan's test for the Alzheimer's death rate shows Washington and Texas are not significantly different, Texas and California are not significantly different, California and New Jersey are not significantly different, and New Jersey and Florida are not significantly different. For cancer, Florida, Texas, Washington, and New Jersey are not significantly different from one another, and California has the lowest rate. For the heart variable, Texas, Florida and New Jersey are significantly different from California and Washington. For diabetes, Florida, Texas and California are significantly different from each other, but not Washington and New Jersey. For the obesity variable, California and New Jersey are significantly lower than Washington, Texas, and Florida. For age greater than 65, New Jersey, California, and Texas are not significantly different from each other, Texas and Washington are not significantly different from each other but have higher rates than New Jersey and California. Washington and Florida are not significantly different from each other but have higher rates than the rest of the states. For smoking, Texas, Florida, and Washington are significantly different from California and New Jersey. For inactivity, Texas and Florida are significantly different from California, Washington and New Jersey.

The SNK test for Alzheimer's death rate indicates Washington, Texas, and California are not significantly different, Texas, California, and New Jersey are not significantly different, California, New Jersey, and Florida are not significantly different. For the cancer variable, Florida, Texas, Washington, and New Jersey are not significantly different from one another. For heart, Texas, Florida and New Jersey are not significantly different from each other, Florida, New Jersey and California are not significantly different from each other, California and Washington are not significantly different from each other. For the diabetes variable, Florida, Texas and California are significantly different from each other, but not Washington and New Jersey. For obesity,

California and NJ are significantly lower than Washington, Texas, and Florida. For age greater than 65, New Jersey, California, Texas, and Washington are not significantly different from each other. Texas, Washington and Florida are not significantly different from each other but are higher. For inactivity, Texas and Florida are significantly different from California, Washington and New Jersey. For smoking, Texas, Florida, Washington are significantly different from California and New Jersey.

Analysis of Contingency Tables

We use contingency tables to see whether or not variables are independent of each other. To determine whether or not two variables are independent, we split each variable into quartiles. Our contingency table has 16 cells: four rows for the quantiles of the first variable, and four columns for the quantiles of the second variable. We then perform a chi squared test with 9 degrees of freedom (4-1)(4-1) and a significance level of 0.05. High chi square values with p-value < 0.05 mean that the two variables are not independent whereas low chi square values with p-value > 0.05 mean that the two variables are independent. We conclude that smoking vs. Alzheimer's death rate and diabetes vs. Alzheimer's death rate are independent while the rest of the variables vs. Alzheimer's death rate are dependent.

Variable	Chi-Square Value	P value	Comments
Smoking vs. Alzheimer's death rate	8.6898	0.4664	The p value is greater than 0.05, so smoking and Alzheimer's death rate are independent
Inactivity vs. Alzheimer's death rate	24.2864	0.0039	The p value is less than than 0.05, so inactivity and Alzheimer's death rate are dependent
Smoking vs. inactivity	129.5495	<.0001	The p value is less than than 0.05, so inactivity and smoking are dependent
Age65 vs. Alzheimer's death rate	25.1202	0.0028	The p value is less than 0.05 so these statistics are not independent.
Obesity vs. Alzheimer's death rate	21.8327	0.0094	The p value is less than 0.05 so these statistics are not independent.
Diabetes vs. Heart	96.7915	<.0001	The p value is less than 0.05 so these statistics are not independent.
Diabetes vs. Alzheimer's Death Rate	6.8298	0.0775	The p value is higher than 0.05 so these statistics are independent.
Heart vs. Alzheimer's Death Rate	13.0972	0.0044	The p value is less than 0.05 so these statistics are not independent.
Cancer vs. Alzheimer's Death Rate	31.6465	0.0002	The p value is less than 0.05 so these statistics are not independent.

Regression Analysis and Correlation Analysis

We performed simple linear regression making Alzheimer's death rate the response variable and making each of the other variables in turn be the regressor. We test the null hypothesis that slope equals zero at significance level of 0.05. We also perform the lack of fit test where the null hypothesis is that there is no lack of fit. We found that when the regressor variable is obesity, heart, or cancer that the slope P-value is less than 0.05 which means we can reject the null hypothesis that slope is 0. For these same regressors, the lack of fit p-value was always above 0.9 indicating no lack of fit in the linear model. For the regressor age65, the slope P-value was 0.0559 which is borderline, and the lack of fit P-value was 0.2341 so we have to assume no lack of fit for age65 as well. For smoking, inactivity, and diabetes both the slope P-value and the lack of fit P-value were well above 0.05 indicating no relationship between Alzheimer's death rate and these regressors.

Predictor Variable	Test for Slope P-value	Lack of Fit P-value	Regression Line	R^2	PRESS	Comments
age65	0.0559	0.2341	y=1.25231-0.00989X	.0083	137.18498	Borderline correlation between age65 and Alzheimer's death rate.
obesity	0.0312	0.9491	Y=0.21985+0.9491X	0.0106	136.84974	There is a linear model possible between obesity and Alzheimer's death.
heart	0.0077	0.9974	Y=0.81411+0.0077X	0.0162	136.26493	There appears to be a linear model possible between heart and Alzheimer's death. 16% of variance in the death rate is explained by the variance in heart disease rate.
cancer	0.0039	0.9235	Y=0.56750+0.00305X	0.0189	136.48536	There does appear to be a linear model between cancer and Alzheimer's death.18% of variance in the death rate is explained by the variance in cancer rate.

Smoking, inactivity, diabetes	There seems to be a linear model NOT possible between smoking and Alzheimer's death, Inactivity and Alzheimer's death, diabetes and Alzheimer's death.

We performed multiple regression to see what variables predict Alzheimer's death rate. We picked three candidate models as well as the full model by running proc rsquare in SAS. We picked the models based on high adjusted R square, low mallow Cp, low MSE, and low PRESS value. Below are the models we chose:

Model	Adj R square	MSE	Mallow Cp	PRESS	R^2 pred
age65, obesity, diabetes, heart, cancer	0.0330	0.30288	4.96	135.40123	0.013045898561429
age65, heart, cancer	0.0289	0.30418	4.8130	135.57439	0.01178
age65, diabetes, obesity,cancer	0.0299	0.30386	5.3627	135.35974	0.0133483
age65, obesity, smoking, inactivity, diabetes, heart, cancer	0.0307	0.30361	8.0000	136.60091	0.0043013

Of these models, we chose age65 heart and cancer as the best model because it has only three variables; it has the lowest Cp value of the candidates, from the contingency tables; we concluded that Alzheimer's death rate and variables Age65, heart, and cancer are dependent on each other.

We also ran proc corr on Alzheimer's death rate against the other variables and found only weak correlations. Based on the Pearson correlation coefficient, age65, smoking, inactivity, diabetes, heart, and cancer were weakly correlated with Alzheimer's death rate.
However, of these results, heart and cancer had the highest correlation coefficients with Alzheimer's death rate, potentially showing a relationship.

Variable	Correlation Coefficient	Comments
cancer	0.13739	Cancer and Alzheimer's death rate are weakly correlated.
age65	-0.09133	Age65 and Alzheimer's death rate are weakly correlated.
obesity	0.10284	Obesity and Alzheimer's death rate are weakly correlated.
diabetes	0.05144	Diabetes and Alzheimer's death rate are weakly correlated.
heart	0.12710	Heart Disease and Alzheimer's death are rate weakly correlated.
smoking	0.04036	Smoking and Alzheimer's death rate are weakly correlated.
obesity	0.10284	Obesity and Alzheimer's death rate are weakly correlated.

We found other observations not directly related to Alzheimer's death rate. For instance, we found that there is a strong linear relationship between obesity and inactivity. The R square value was about 0.60 with no evidence of lack of fit. Obesity and diabetes also showed a strong linear relationship with R squared value of about 0.68 and no evidence of lack of fit.

Conclusion

In conclusion, we found Alzheimer's death rate very difficult to predict with the variables we had. From contingency tables and linear regression, we found indicators of possible relationships between Alzheimer's death rate and age greater than 65, heart disease, and cancer. However, the model has a low adjusted R-squared value and is not very strong for future prediction. Yet, between obesity and inactivity as well as between obesity and diabetes, we found rather strong linear relationships. Further research could expand on the links between obesity, inactivity, and diabetes.