National College of Ireland

STATISTICS FOR DATA ANALYTICS

Continuous Assessment 2

Prashanth A R
X16137591

First Analysis: Linear Regression

Overview of Regression

According to Lind, Marchal and Mathen (2011), The technique of developing an equation to estimate the value of dependent variable based on the selected value of independent variable is called regression analysis.

The equation which represents the linear relationship between the dependent and independent variable is called *regression equation*.

Least squares principle is a mathematical method which determines the regression line by minimizing the sum of squares between observed and predicted values of the dependent variable.

General form of linear regression equation is $Y' = a + bX$ where:

Y' is the predicted value for a given value of the independent variable (X).

a is Y the intercept, i.e. the predicted value of Y' when X=0.

b is the slope of the line which signifies the change in Y' for every unit change of value X.

X is any arbitrary value of the independent variable.

Objective: This analysis is conducted to predict the electricity bill amount using kilowatt-hour power consumption as independent variable for a city.

Null Hypothesis: The slope of the predictor variable is zero.

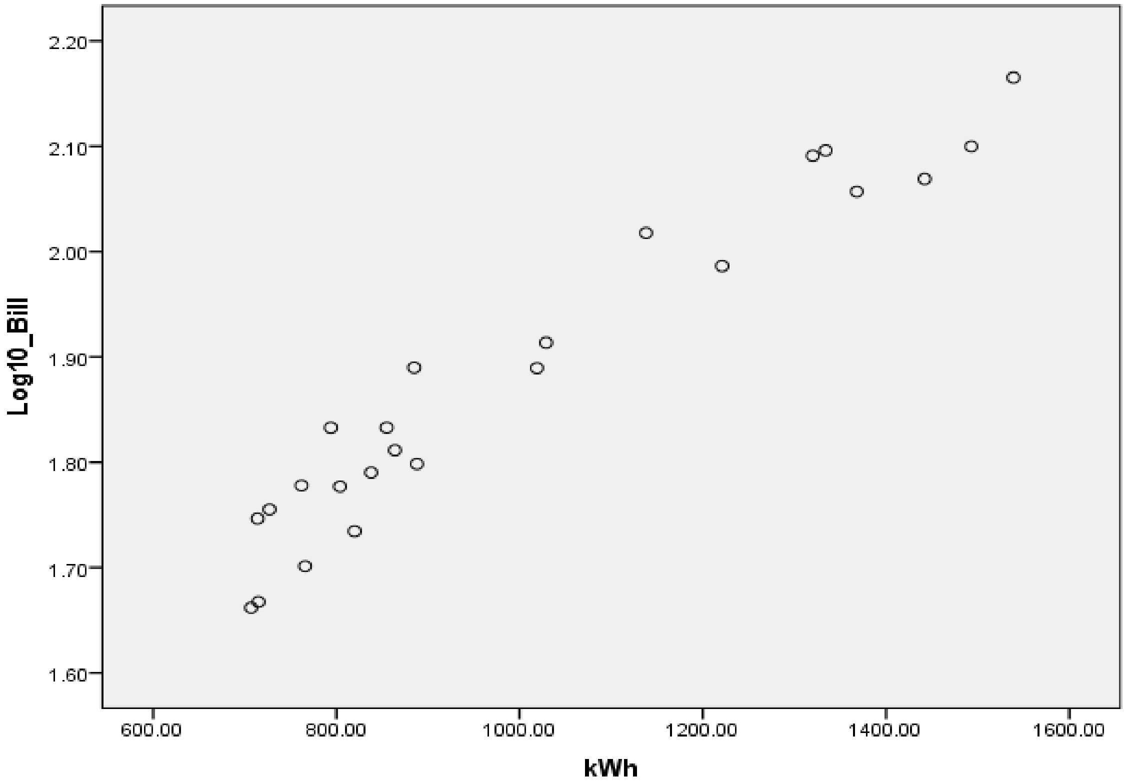
Alternate Hypothesis: The slope of the predictor variable is non-zero.

Levels of measurement: There is one dependent and independent variable with each of them being continuous.

Preliminary Tests:

1. **Linear relationship**:

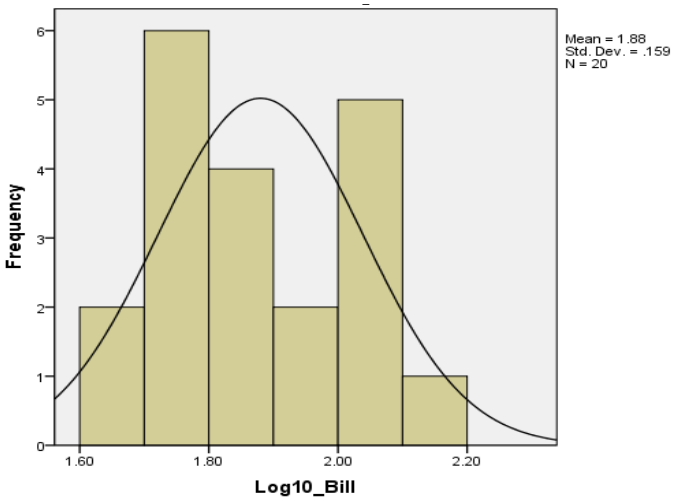
   Linear regression assumes linearity between the dependent and independent variable. In this case, a scatter plot is drawn with kWh consumption as X-axis and corresponding electricity bill as Y-axis. The following scatter plot is obtained:



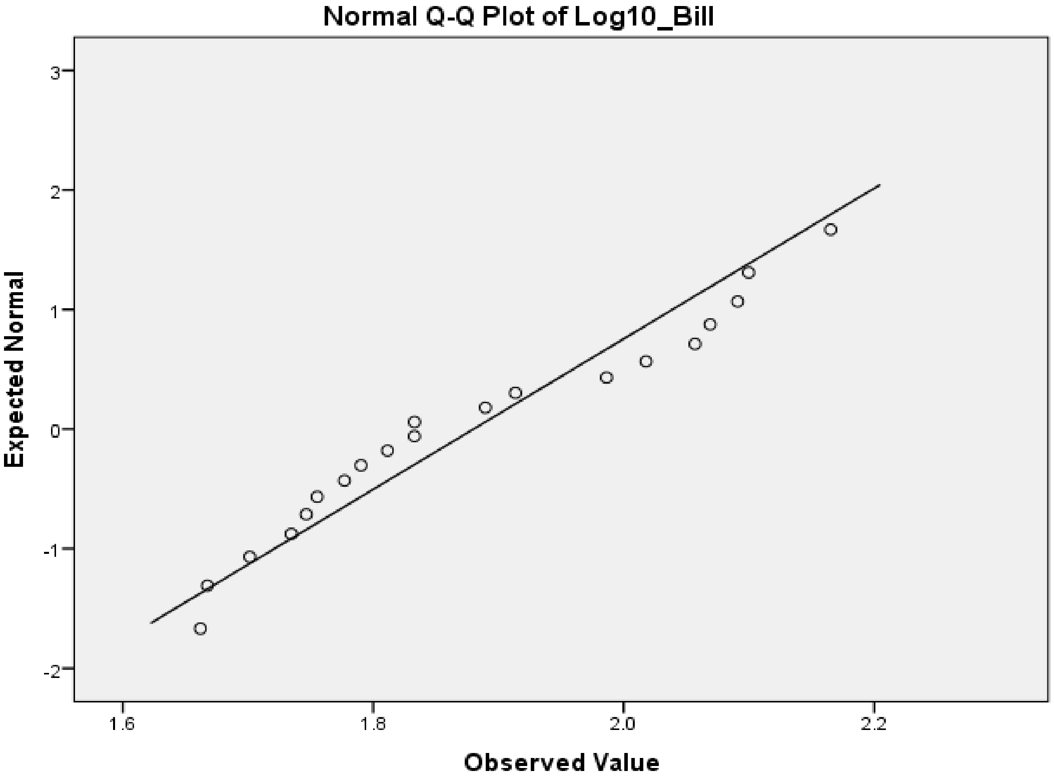
By observing the plot, we can conclude there exists a positive linear relationship between the dependent and independent variable.

2. Normality:

Linear regression assumes the normal distribution of dependent variables. In this case, a Histogram and Q-Q Plot is drawn for the kWh consumption which is the dependent variable. Due to the skewness of the graph to towards the left side, a log10 transformation was applied to the kWh variable and plots were drawn.



From the histogram, we can assume that dependent variable is normally distributed with reasonably less outliers.



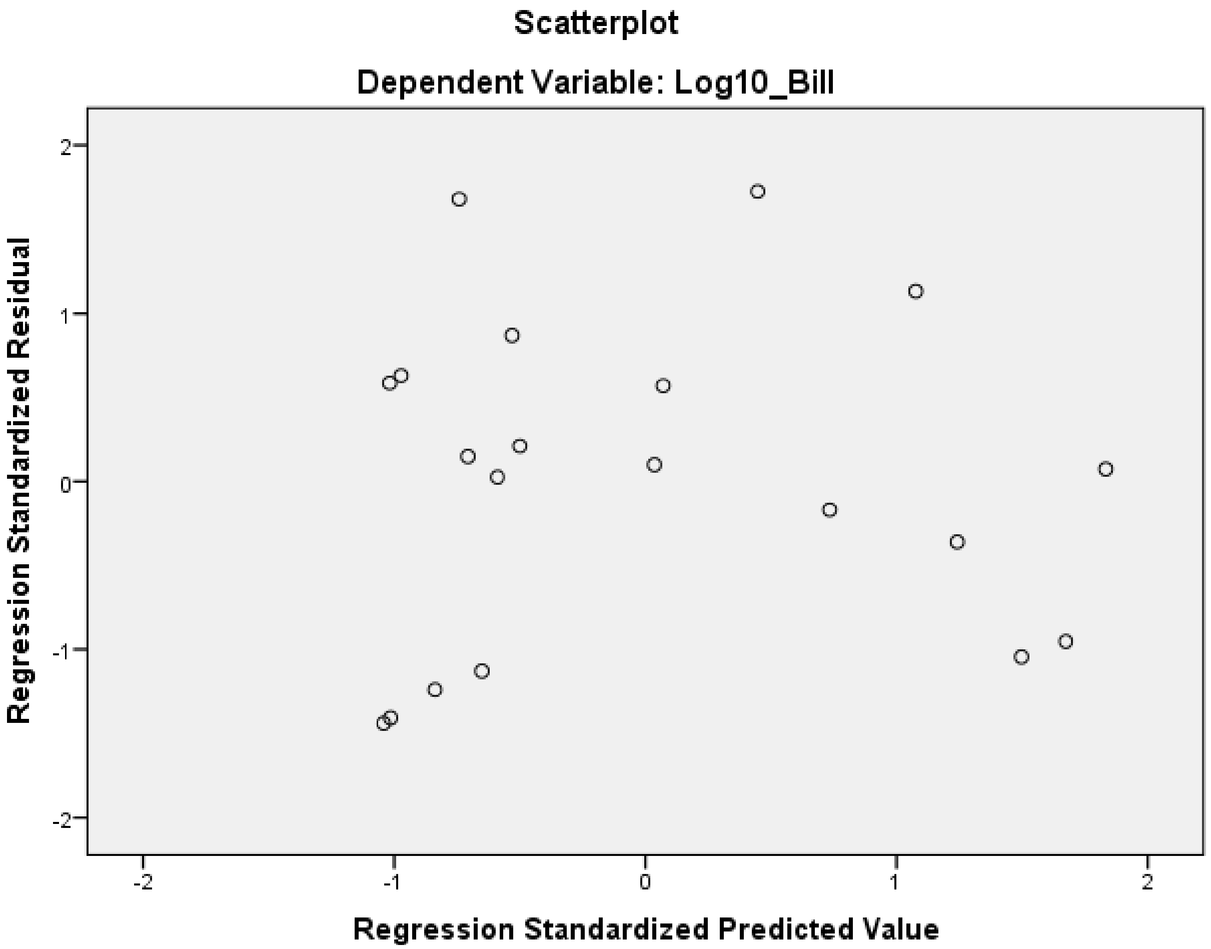
As we can see from the above Q-Q plot, the points are close to the line indicating that there are no major deviations from the normality.

3. Multicollinearity:

Multicollinearity exists only if there are more than one predictor variables. However, in this case there is only one predictor variable and hence there is no need to test this factor.

4. Homoscedasticity:

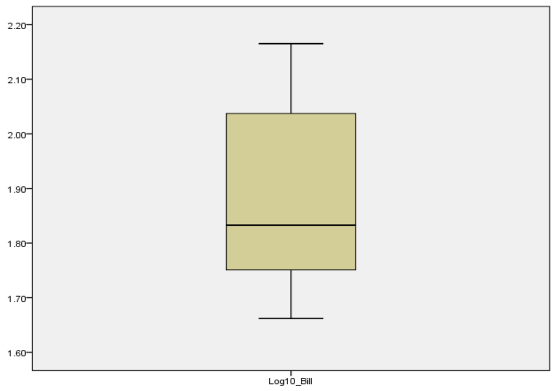
The scatter plot obtained from the test can be visualized to determine the homoscedasticity.



As the points in the distribution do not cross the range (-2,2) and not form any specific pattern the homoscedasticity factor is satisfied.

5.Outliers:

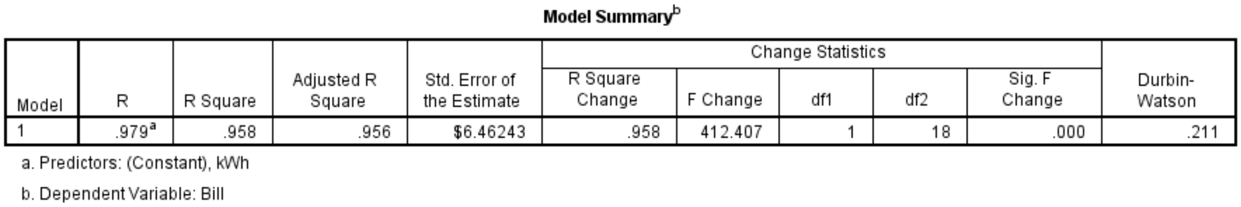
The presence of outliers can be determined by examining the boxplot, in this case a boxplot is drawn for the kWh consumption which is the dependent variable.



Observing the boxplot, we can conclude that there are no outliers in the dependent variable that is kWh.

6. Auto-correlation:

The Durbin-Watson test determines the correlation between the residuals.



Model Summary[b]

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.979[a]	.958	.956	$6.46243	.958	412.407	1	18	.000	.211

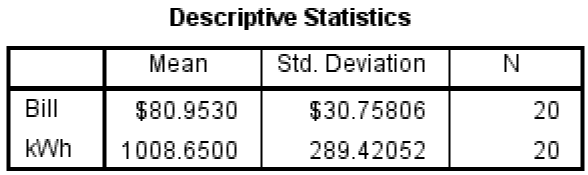
a. Predictors: (Constant), kWh

b. Dependent Variable: Bill

From the table above we can observe that the Durbin-Watson value is 0.211 which somewhat close to 0 indicating that there is a positive correlation. This is obvious as increase in kWh consumption leads to increase in the bill amount.

Test Results

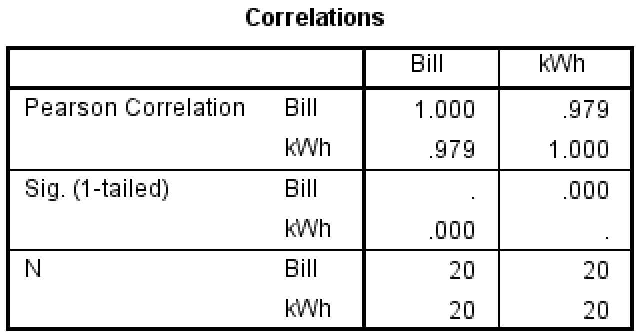
A Linear regression was performed with kWh as independent variable and bill as the dependent variable with 0.05 significance level. The following results were obtained. A random sample of size 20 was selected for the analysis.



Descriptive Statistics

	Mean	Std. Deviation	N
Bill	$80.9530	$30.75806	20
kWh	1008.6500	289.42052	20

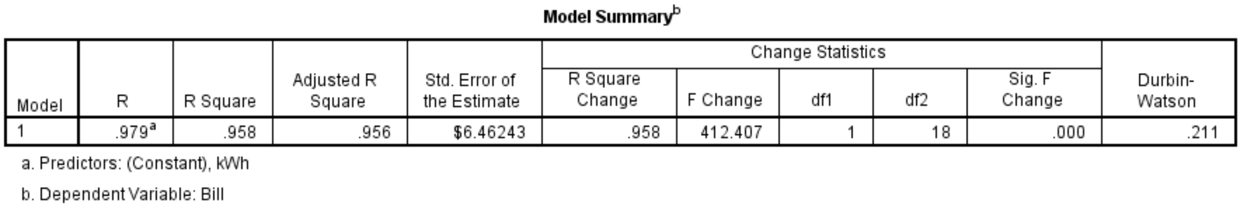
Above table represents the mean and standard deviation of bill and kWh consumption with mean bill amount of $80.95 and kWh consumption of 1008.65.



Correlations

		Bill	kWh
Pearson Correlation	Bill	1.000	.979
	kWh	.979	1.000
Sig. (1-tailed)	Bill	.	.000
	kWh	.000	.
N	Bill	20	20
	kWh	20	20

Pearson Correlation for this analysis has a value of 0.979 which signifies a strong positive correlation between the kWh consumption and the bill amount.



Model Summary[b]

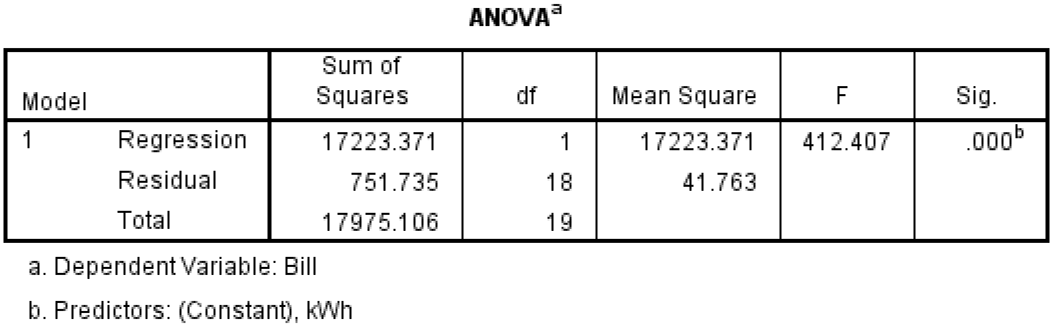
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.979[a]	.958	.956	$6.46243	.958	412.407	1	18	.000	.211

a. Predictors: (Constant), kWh

b. Dependent Variable: Bill

R Square signifies the ability to predict the variance of dependent variable with respect to dependent variable for a model. In this case, we must look at Adjusted R square value which is 0.956 which means that 95.6% of variations in bill amount can be explained by the kWh consumption.

Also, the standard error estimate of bill amount is obtained as $6.46 which is the average distance which the observed value falls from the regression line.



ANOVA[a]

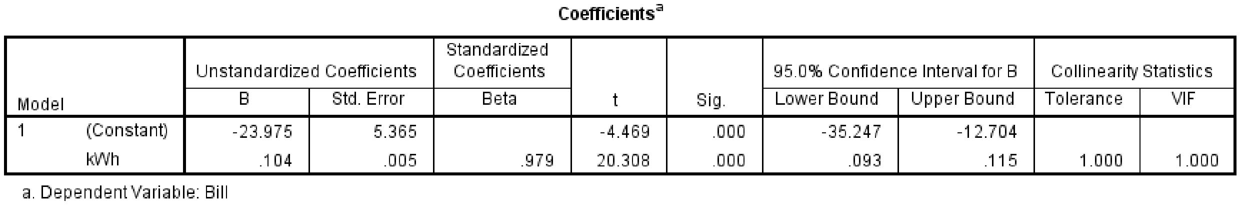
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	17223.371	1	17223.371	412.407	.000[b]
	Residual	751.735	18	41.763		
	Total	17975.106	19			

a. Dependent Variable: Bill

b. Predictors: (Constant), kWh

In the ANOVA table, we can see that the p value is less than 0.05 indicating that there is a very strong evidence to reject the null hypotheses that coefficient on the independent variable is 0, which is not in this case.

The VIF and Tolerance value is checked to determine the multicollinearity between the independent variables. As we are using only one predictor variable in our analysis, these fields do not matter.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-23.975	5.365		-4.469	.000	-35.247	-12.704		
	kWh	.104	.005	.979	20.308	.000	.093	.115	1.000	1.000

a. Dependent Variable: Bill

A regression equation can be constructed from the above coefficients table by using unstandardized coefficients. It roughly estimates the value of dependent variable for a given value of predictor variable. Column B gives the intercept and the corresponding coefficient of independent variable. The regression equation can be represented as:

Bill amount = 0.104(kWh) - 23.975. Suppose the kWh consumption was 900 then the bill amount would roughly be 0.104(900) - 23.975 = $69.625.

Second Analysis: Kruskal-Wallis Test

Overview of non-parametric tests

According to Pallant (2013) and Douglas et al., (2011) parametric tests such as t-tests, analysis of variance etc., requires the population to follow normal probability distribution. However, non-parametric tests do not assume the normality of the distribution.

Kruskal-Wallis test is a non-parametric test which is used to determine if there is a statistically significant difference between two or more groups on an independent variable on a continuous or ordinal scale. It is an alternative to one-way between groups ANOVA. (Kruskal-Wallis H Test in SPSS Statistics, 2017)

Objective: The objective is to compare the medians of divorced, widowed and separated males for different areas and check if is there any statistically significant difference between them**.**

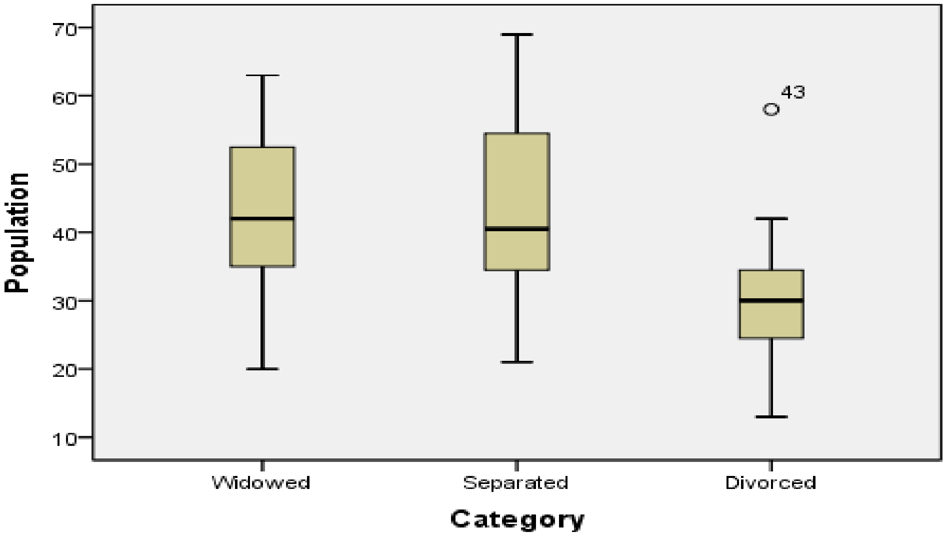
Null Hypothesis: Population medians are equal.

Alternate Hypothesis: Population medians are not equal.

Levels of Measurement: In this test, a continuous dependent variable and a categorical independent variable is chosen with 1,2,3 representing widowed males, separated males and divorced males respectively.

Preliminary Tests: Non-parametric tests do not assume the normality and homogeneity of the population. However, there are few assumptions to be made which are as follows:

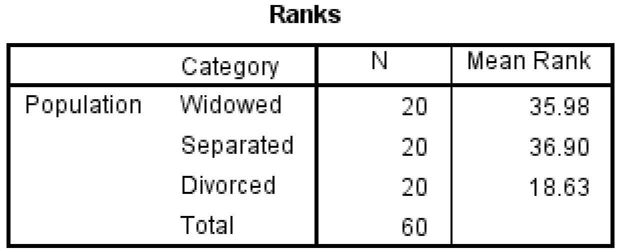
1. The samples collected are independent which satisfies assumption of independence.
2. The samples are randomly selected.
3. The below boxplot determines the similar shapes of distributions:



Results

The sample size was chosen as 20 for each categorical variable which means there were 60 records.

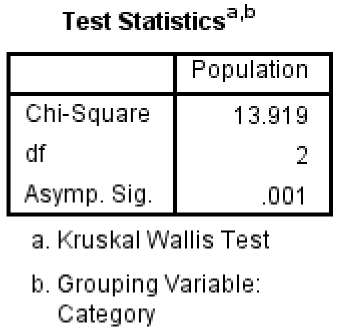
Analysis was done using Kruskal-Wallis test with 3 independent samples with 0.05 significance level. The results obtained are as follows:



Ranks

	Category	N	Mean Rank
Population	Widowed	20	35.98
	Separated	20	36.90
	Divorced	20	18.63
	Total	60	

The above table provide the mean ranks of the 3 categories, with widowed and separated means being almost identical.



Test Statistics[a,b]

	Population
Chi-Square	13.919
df	2
Asymp. Sig.	.001

a. Kruskal Wallis Test

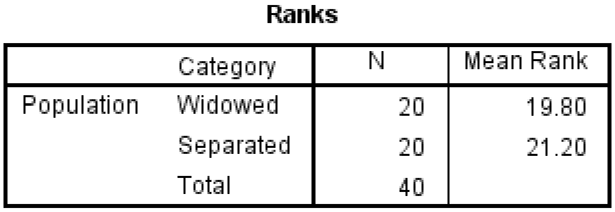
b. Grouping Variable: Category

The p value from the statistics table is less than 0.05 which signifies that there is a statistically significant difference between means of 3 categories. Hence the null hypothesis is rejected.

However, it is still not clear as to between which categories there is statistically significant difference as Kruskal-Wallis is an omnibus test. A workaround for this is changing the grouping variable combinations using select cases and performing Kruskal-Wallis test for each combination.

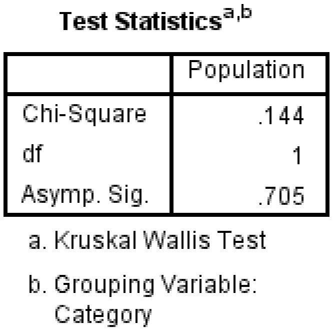
Notice the Chi-Square value in the test statistics table, this value can be used to calculate the effect size estimate by diving its value from the sample size minus one, i.e., 13.919 / (59-1) = 0.2359 which means 23.59% of the variability in mean ranks are accounted by the type of male population (male, widowed and divorced).

Case 1: Widowed and Separated



Ranks

	Category	N	Mean Rank
Population	Widowed	20	19.80
	Separated	20	21.20
	Total	40	



Test Statistics[a,b]

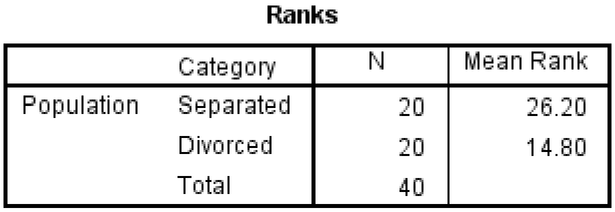
	Population
Chi-Square	.144
df	1
Asymp. Sig.	.705

a. Kruskal Wallis Test

b. Grouping Variable: Category

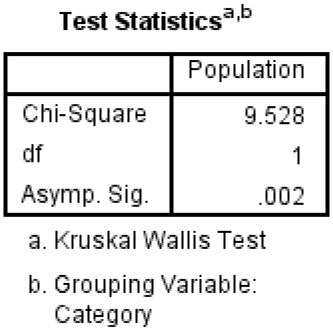
For this case, the p value is 0.705 which signifies there is no statistically significant difference between means of widowed and separated male population. Size effect estimate for this case is 0.144 / (40-1) = 0.00369 or 0.369% which is not statistically significant.

Case 2: Separated and Divorced



Ranks

	Category	N	Mean Rank
Population	Separated	20	26.20
	Divorced	20	14.80
	Total	40	



Test Statistics[a,b]

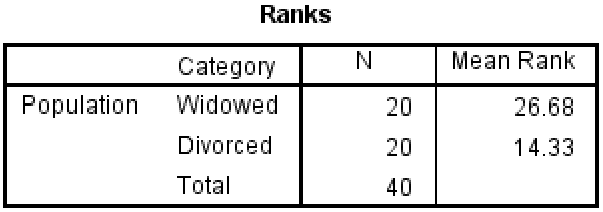
	Population
Chi-Square	9.528
df	1
Asymp. Sig.	.002

a. Kruskal Wallis Test

b. Grouping Variable: Category

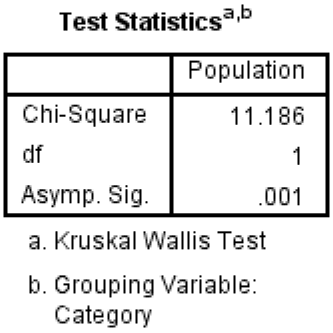
In this case, the p value is less than 0.05 indicating that there is statistically significant difference between separated and divorced male population. Size effect estimate for this case is 9.528 / (40-1) = 0.24430 or 24.43% which is statistically significant. It means 24.43% of variability in mean ranks is accounted by the widowed and divorced male population.

Case 3: Widowed and Divorced



Ranks

	Category	N	Mean Rank
Population	Widowed	20	26.68
	Divorced	20	14.33
	Total	40	



Test Statistics[a,b]

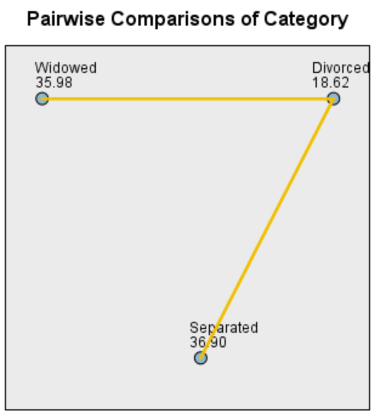
	Population
Chi-Square	11.186
df	1
Asymp. Sig.	.001

a. Kruskal Wallis Test

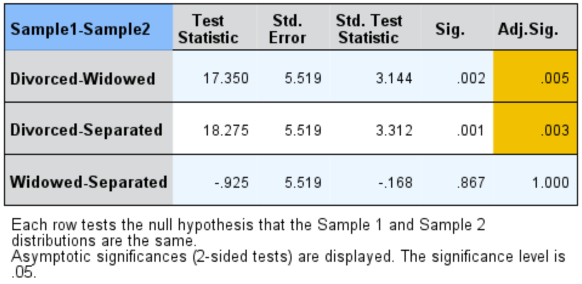
b. Grouping Variable: Category

In this case, the p value is less than 0.05 thereby indicating that there is a statistically significant difference in means of widowed and divorced male population. Size effect estimate for this case is 11.186 / (40-1) = 0.2868 or 28.68% which is statistically significant. It means 28.68% of variability in mean ranks is accounted by the widowed and divorced male population.

The above 3 cases can be represented by Pairwise Comparisons by Category as follows:



Each node shows the sample average rank of Category.



Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Divorced-Widowed	17.350	5.519	3.144	.002	.005
Divorced-Separated	18.275	5.519	3.312	.001	.003
Widowed-Separated	-.925	5.519	-.168	.867	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

The diagram shows the mean ranks of 3 categories. The table uses Bonferroni correction for significance level.

References

[1] Kruskal-Wallis H Test in SPSS Statistics |*Procedure, output and interpretation of the output using a relevant example*.. [ONLINE] Available at: https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php. [Accessed 11 April 2017].

[2] Douglas Lind, 2011. *Statistical Techniques in Business and Economics (Mcgraw-Hill/Irwin Series Operations and Decision Sciences)*. 15 Edition. McGraw-Hill/Irwin.

[3] Julie Pallant, 2010. *SPSS Survival Manual: A step by step guide to data analysis using SPSS, 4th Edition*. 4 Edition. Open University Press.

[4] YouTube. 2017. *When and How To Run Kruskal Wallis in SPSS - YouTube*. [ONLINE] Available at: https://www.youtube.com/watch?v=Ip_ZUChger8&t=584s. [Accessed 12 April 2017].

[5] YouTube. 2017. *Linear Regression in SPSS - YouTube*. [ONLINE] Available at: https://www.youtube.com/watch?v=U2p16pCHW3c&t=298s. [Accessed 11 April 2017].

[6] YouTube. 2017. *how2stats - YouTube*. [ONLINE] Available at: https://www.youtube.com/channel/UCr3OHuCSrwAO2KYP2CJB6zg. [Accessed 10 April 2017].

Data Sources:

Source 1: *Residential Average Monthly kWh and Bills* - Data.gov. 2017. [ONLINE] Available at: https://catalog.data.gov/dataset/residential-average-monthly-kwh-and-bills. [Accessed 13 April 2017].

Source 2: *Dublin Demographic Profile – Datasets* – Data.gov.ie. 2017 [ONLINE] Available at: https://data.gov.ie/dataset/dublin-demographic-profile. [Accessed 15 April 2017].