**Q1**

**PROC** **IMPORT** OUT= WORK.Npda

DATAFILE= "C:\Users\ayoyi\OneDrive - De Montfort University\

Analytics programming\week 9\psa09-npda.xlsx"

DBMS=EXCEL REPLACE;

RANGE="npda$";

GETNAMES=YES;

MIXED=NO;

SCANTEXT=YES;

USEDATE=YES;

SCANTIME=YES;

**RUN**;

ods exclude enginehost;

**proc** **contents**

data = npda;

**run**;

ods select all;

/\*transpose the data from wide to long\*/

**data** NPDA\_;

set NPDA;

year = **1** ;

patients = patients1;

target = target1;

output;

year = **2** ;

patients = patients2;

target = target2;

output;

year = **3** ;

patients = patients3;

target = target3;

output;

year = **4** ;

patients = patients4;

target = target4;

output;

year = **5** ;

patients = patients5;

target = target5;

output;

year = **6**;

patients = patients6;

target = target6;

output;

year = **7**;

patients = patients7;

target = target7;

output;

keep

unit

region

year

patients

target

;

**run**;

/\*analyse the independent variables\*/

**proc** **freq**

data= NPDA\_;

tables region unit;

**run**;

/\*produce descriptive statistics\*/

**proc** **means**

data = NPDA\_;

class region;

var year

patients

target

;

**run**;

Q2

/\*requirements:

create an informat

create a format

create new variable\*/

**proc** **format**;

value ns

**1** = 'north'

**2** = 'middle'

**3** = 'south'

;

**run**;

**data** NPDA\_NEW;

set NPDA\_;

if region = 'London and South East' then ns = **1**;

else if region = 'South Central' then ns = **1**;

else if region = 'South West' then ns = **1**;

else if region = 'East Midlands' then ns = **2**;

else if region = 'East of England' then ns = **2**;

else if region = 'Wales' then ns = **2**;

else if region = 'West Midlands' then ns = **2**;

else if region = 'Yorkshire and Humber' then ns = **3**;

else if region = 'North East' then ns = **3**;

else if region = 'North West' then ns = **3**;

keep unit

region

year

patients

target

ns

;

format

ns ns.

;

label

ns = 'North/South';

**run**;

**proc** **tabulate**

data = NPDA\_NEW;

class year ns;

var target;

table year\*ns\*(target), N NMISS MIN P25 MEDIAN\*f=**5.2** P75 MAX MEAN STD;

**run**;

**proc** **freq**

data = NPDA\_NEW;

table unit;

**run**;

There are 182 number of PDUs in the dataset.

There are slight differences in the percentiles of the regions, but the largest difference is in the north region for year 2012/13 with maximum value of proportion of patients below the target of 58 mmol/mol being 0.73 compared to 0.35 and 0.29 for middle and south regions respectively﻿﻿﻿. ﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿

The north region also has the highest number of proportion of patients below the target of 58 mmol/mol across all the years.

The middle region has the highest number of patients with no proportion in year 2011/12 at 16.

Q3

**proc** **sql**;

create table NPDA1 as

select \*

from NPDA\_NEW

where ns in (**1**, **3**) /\*selects year 2011/12 and 2013/14\*/

order by year; /\*sorts by year\*/

**quit**;

ODS HTML;

ods graphics on;

/\*performs T test\*/

title "Test of the difference between North and South Region as determined by year";

**proc** **ttest**

data = NPDA1;

class ns;

var target;

by year;

**run**;

title;

ods noproctitle;

ods graphics off;

ODS HTML CLOSE;

/\*test for normality\*/

title "Plot to test normality of North and South Region";

**proc** **univariate**

data = NPDA1 normal ;

qqplot /Normal(mu=est sigma=est);

class ns;

var target;

by year;

**run**;

title;

ods noproctitle;

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Year** | **Variance** | **P-value** | **Mean** | **Mean Diff** | **95% CI** |
| 1 | p-value (shown under "Pr>F") > 0.05 then variance is equal | = 0.2311 (Using pooled method)  H0 is accepted | North = 0.1696  South = 0.1869 | 0.0173 | (-0.0457+ 0.0112) |
| 2 | p-value (shown under "Pr>F") < 0.05 then variance is unequal | = 0.7145 (Using **S**atterthwaite method)  H0 is accepted | North = 0.1654  South = 0.1601 | -0.00527 | (-0.0232 + 0.0337) |
| 3 | p-value (shown under "Pr>F") > 0.05 then variance is equal | = 0.1532 (Using pooled method)  H0 is accepted | North = 0.1893  South = 0.1710 | -0.0183 | (-0.00690 + 0.0435) |
| 4 | p-value (shown under "Pr>F") > 0.05 then variance is equal | = 0.5734 (Using pooled method)  H0 is accepted | North = 0.2272  South = 0.2188 | -0.00836 | (-0.0210 + 0.0377) |
| 5 | p-value (shown under "Pr>F") > 0.05 then variance is equal | = 0.2130 (Using pooled method)  H0 is accepted | North = 0.2662  South = 0.2474 | - 0.0188 | (-0.0109 + 0.0485) |
| 6 | p-value (shown under "Pr>F") > 0.05 then variance is equal | = 0.0346 (Using pooled method)  H0 is rejected | North = 0.2935  South = 0.2622 | -0.0313 | (-0.00231 + 0.0604) |
| 7 | p-value (shown under "Pr>F") > 0.05 then variance is equal | = 0.3946 (Using pooled method)  H0 is accepted | North = 0.2793  South = 0.2650 | 0.0144 | (-0.0189 + 0.0476) |

From the summary above, there is difference between north and south in year 2012/13. Null hypothesis is accepted in all the years except 2016/17 where the p-value is less than 0.05.

Two sample t-test assumes that

* There is one continuous dependent variable and one categorical independent variable (with 2 levels).
* The two samples are independent.
* The two samples follow normal distributions and can be done with Normality check.

The first 2 assumptions are met. The normality check will be conducted using univariate procedure.

|  |  |  |
| --- | --- | --- |
| **Year** | **North P- value** | **South P- value** |
| 1 | 0.6212 | 0.0018 |
| 2 | <0.0001 | 0.0803 |
| 3 | 0.0058 | 0.1135 |
| 4 | 0.0375 | 0.1066 |
| 5 | 0.6513 | 0.9631 |
| 6 | 0.7682 | 0.2314 |
| 7 | 0.8185 | 0.0423 |

According to SAS manual, If the sample size is less than 2000, the Shapiro-Wilk test is better. The 3rd assumption is that if p-value < than 0.05, null hypothesis that there is no significant departure from normality is rejected. From the table above, the two samples (north and south) does not have p-value > 0.05 in all the years except years 2015/16 and 2016/17 which shows that the null hypothesis assumption that there is no significant departure from normality is rejected. This has nullified the previous inferences; therefore, other possible methods should be used.

Q4

**proc** **sql**;

create table NPDA11 as

select \*

from NPDA\_NEW

where year in (**6**); /\*selects year 2016/17\*/

**quit**;

/\*performs one-way ANOVA test\*/

title "Test to examine the difference between the North, Middle and South regions";

**proc** **glm**

data = NPDA11;

class ns;

model target = ns;

means ns/ lsd /\* pairwise comparisons \*/

hovtest=bartlett; /\* homogeneity of variance\*/

output out=npdap p=predicted r=residual;

**quit**;

title;

ods noproctitle;

/\*tests for normality\*/

title "Residual plot";

**proc** **univariate**

data=npdap

plot normal;

var residual;

qqplot /

normal(mu=est sigma=est); /\* normally distributed? \*/

**run**;

The Corrected Total degrees of freedom is 171 (172-1).

F-test is significant (F= 5.15, p-value = 0.0399) and so the null hypothesis is rejected. This shows that the model provides a better fit than the intercept-only model. The middle and south, north and south have comparisons at significance level of 5%.

Type I and Type III are the same which shows the data is balanced.

R-squared provides an estimate of the strength of the relationship between the model and the response variable. R-square of 0.037406 indicates that the model accounts only for 3.7% of the variation in proportion of patients with HbA1c less than 58 mmol/mol for the year 2016. A good model will have R-square of 100%.

Assumptions of One-Way ANOVA

* Normality – Samples should be approximately normally distributed.
* Independence – Samples must be independent from each other.
* Homogeneity of variance – Variances of the samples should be equal.

Plotting the residual, the samples are normally distributed. The data meets the assumptions and there is no alteration in the previous interpretation.

Q5

**proc** **sort**

data= NPDA\_NEW;

by year; /\* sorts year in ascending order\*/

**run**;

title "Relationship plot between number of patients and proportion with HbA1c less than 58 mmol/mol";

**proc** **sgscatter**

data= NPDA\_NEW;

plot target\*patients / reg= (cli clm);

by year

;

**run**;

title;

ods noproctitle;

The plot for year 2013/14 and 2017/18 ﻿﻿﻿﻿﻿﻿﻿shows that there is linear relationship between﻿﻿﻿﻿﻿﻿ number of patients and proportion with HbA1c less than 58 mmol/mol. There is a very slight relationship between both variables in year 2011/12 and 2014/15.

Q6

**proc** **sql**;

create table NPDA111 as

select \*

from NPDA\_NEW

where year in (**1**,**3**,**4**,**7**); /\*selects years 2011/12, 2013/14, 2014/15 and 2017/18\*/

**quit**;

title "Effect of the number of patients on the proportion of patients reaching the target of HbA1c less than 58 mmol/mol";

**proc** **reg**

data= NPDA111;

model target = patients/

p r /\* prediction residual\*/

cli /\* prediction interval for new observation\*/

clm /\* confidence interval for mean response\*/

;

by year;

output out=npda\_pred p=yhat r=resid;

**quit**;

title;

ods noproctitle;

/\*test for normality\*/

title "Residual plot";

**proc** **univariate**

data = npda\_pred

plot normal;

var resid;

qqplot /normal(mu=est sigma=est); \* residual vs normal QQ plot;

by year;

**run**;

title;

ods noproctitle;

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **R-Square** | **P-value** | **Regression Equation** | **Fit Plot** |
| 1 | 0.0321 shows that the independent variable (patient) is only able to explain 3.2% of the variation of the dependent variable (target). | 0.0248 indicates that the H0 should be rejected which means that the slope is not zero. | Target = 0.14223 + (patient \* 0.00019650). With one increase in number of patients, the proportion of patients with HbA1c less than 58 mmol/mol will increase by 0.00019650. | The fit plot further confirms inaccurate prediction of the model. Though the confident band is tight around the fit line, the prediction band is very wide. |
| 3 | 0.0322 Shows that the independent variable (patient) is only able to explain 3.2% of the variation of the dependent variable (target). | 0.0168 indicates that the H0 should be rejected which means that the slope is not zero. | Target = 0.15649 + (patient \* 0.00015489)  With one increase in number of patients, the proportion of patients with HbA1c less than 58 mmol/mol will increase by 0.00015489. | The fit plot further confirms inaccurate prediction of the model. Though the confident band is tight around the fit line, the prediction band is very wide. |
| 4 | 0.0141 shows that the independent variable (patient) is only able to explain 1.4% of the variation of the dependent variable (target). | 0.1177 indicates that the H0 should be accepted. Slope is zero. | Target = 0.19751 + (patient \* 0.00011924) | The fit plot further confirms inaccurate prediction of the model. Though the confident band is tight around the fit line, the prediction band is very wide. |
| 7 | 0.0099 shows that the independent variable (patient) is only able to explain 1% of the variation of the dependent variable (target). | 0.1946 indicates that the H0 should be accepted. Slope is zero. | Target = 0.26442+ (patient \* 0.00011173) | The fit plot further confirms inaccurate prediction of the model. Though the confident band is tight around the fit line, the prediction band is very wide. |

Assumptions

1. The relationship between the IVs and the DV is linear- Scatterplots show that this assumption had been met for all the years.
2. There is no multicollinearity in the data- This does not apply as we have only 2 variables
3. The values of the residuals are independent- the points in the **residual** plot are randomly dispersed around the horizontal axis which shows this assumption is met.
4. The variance of the residuals is constant (homoscedasticity)- is the assumption that the variation in the residuals (or amount of error in the model) is similar at each point of the model. This assumption is met except in year 2013/14 where plot of residuals vs predicted values showed obvious signs of funnelling.
5. The values of the residuals are normally distributed- The Q-Q plot for the model suggested that the assumption of normality of the residuals are met for only year 2014/14 and 2017/18 as there are no deviations from normality.
6. There are no influential cases biasing the model- Cook’s Distance values were under 1 for all the years, suggesting no influence on the model.

The assumptions do not alter the interpretation of the model.

**proc** **ttest**

data = NPDA1;

class ns;

var target;

by year;

**run**;

**proc** **univariate**

data = NPDA1 normal ;

qqplot /Normal(mu=est sigma=est);

class ns;

var target;

by year;

**run**;