REPORT & ASSIGNMENTS

Statistical Data Analysis with SPSS

(Agere-008 SPSS KL20)

Spring 2020

Diana Crowe

Student nr.: 012056152

Introduction

SPSS is a statistical analysis program used in social science, market research, health research, surveys, education research, etc. It is very user friendly in that it does not require programming skills. The base software handles statistical data analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary is stored in the datafile).

Statistics included in the base software [1]:

- Descriptive statistics: Cross tabulation, Frequencies, Descriptives, Explore, Descriptive Ratio Statistics
- Bivariate statistics: Means, t-test, ANOVA, Correlation (bivariate, partial, distances), Nonparametric tests, Bayesian
- Prediction for numerical outcomes: Linear regression
- Prediction for identifying groups: Factor analysis, cluster analysis (two-step, K-means, hierarchical), Discriminant
- Geo spatial analysis, simulation
- R extension (GUI), Python

In this report I will showcase some of the statistical abilities of SPSS (version: IBM SPSS Statistics 25) by going through the paces for seven data sets which were provided as assignments.

Data Sets and analyses performed:

1. Smoking data: cross tabulation - Descriptive statistics
2. Heart data: correlations and graphical presentation - Bivariate statistics
3. Table 7.2: t-test - Bivariate statistics
4. ttestpr data: t-test - Bivariate statistics
5. Table 8.1: one-way parametric ANOVA - Bivariate statistics
6. Bodyfat data: linear regression analysis - Prediction for numerical outcomes
7. Disease data: logistic regression analysis - Prediction for numerical outcomes

[1] Wikipedia: https://en.wikipedia.org/wiki/SPSS

Analyses of Data Sets

I have divided the treatment of the problem(s) into the following:

- File name and (short) description
- Problem(s) to be solved in the assignment
- Conclusions – aka solution(s) to the problem(s)
- Analysis – with screenshots of data (charts, tables…) and comments

Assignment 1:

File name and description:
Smoking.sav - Relationship between smoking and lung cancer

Problem:
Smoking data: Examine with cross table analysis the association between variables smoking and cancer.

Conclusions:
In our crosstable (fig 1.1) we see that the counts of cancer for people who don't smoke (170a) and people who smoke (330b) have a different subscript letter. A different subscript letter indicates that the corresponding percentages are statistically dissimilar.

The amount of non-smokers who got lung cancer (30.4%) is a lot lower than the number of smokers who got lung cancer (75.0%).

Analysis:

lungcancer * smoking Crosstabulation

			Smoking		
			nosmoking	Smoking	Total
lungcancer	nocancer	Count	390$_a$	110$_b$	500
		Expected Count	280.0	220.0	500.0
		% within smoking	69.6%	25.0%	50.0%
	cancer	Count	170$_a$	330$_b$	500
		Expected Count	280.0	220.0	500.0
		% within smoking	30.4%	75.0%	50.0%
Total		Count	560	440	1000
		Expected Count	560.0	440.0	1000.0
		% within smoking	100.0%	100.0%	100.0%

Each subscript letter denotes a subset of smoking categories whose column proportions do not differ significantly from each other at the .05 level.

***Figure 1.1**: crosstable comparing smokers, non-smokers, and cancer counts*

Assignment 2:

File name and description:
Heart.sav – heart data

Problem:
Heart data: calculate Pearson correlations for variables age, both blood pressures, height and weight. Make also graphical presentation with scatter plots in one picture (matrix plot).

Conclusions:
- The values for the different Pearson correlations can be seen in the table below (fig. 2.1)
- A matrix plot with all the scatter plots in one picture can be found after the table (fig 2.2)

There is not a high correlation between most of the chosen variables. We do however see a positive correlation between weight and height (the taller you are, the heavier you are) and between weight and each of the blood pressures (heavier people have higher blood pressures). There is also a positive correlation between the two blood pressures.

Analysis:

Correlations

		age	cholesterole	systolic bloodpr	diastolic bloodpr	height	weight
age	Pearson Correlation	1	.019	.182	.071	-.182	-.142
	Sig. (2-tailed)		.856	.073	.486	.075	.164
	N	98	97	98	98	97	97
cholesterole	Pearson Correlation	.019	1	.170	.139	-.126	.087
	Sig. (2-tailed)	.856		.095	.176	.220	.400
	N	97	97	97	97	96	96
systolic bloodpr	Pearson Correlation	.182	.170	1	.551**	.071	.589**
	Sig. (2-tailed)	.073	.095		.000	.491	.000
	N	98	97	98	98	97	97
diastolic bloodpr	Pearson Correlation	.071	.139	.551**	1	.052	.439**
	Sig. (2-tailed)	.486	.176	.000		.613	.000
	N	98	97	98	98	97	97
height	Pearson Correlation	-.182	-.126	.071	.052	1	.460**
	Sig. (2-tailed)	.075	.220	.491	.613		.000
	N	97	96	97	97	97	97
weight	Pearson Correlation	-.142	.087	.589**	.439**	.460**	1
	Sig. (2-tailed)	.164	.400	.000	.000	.000	
	N	97	96	97	97	97	97

**. Correlation is significant at the 0.01 level (2-tailed).

***Figure 2.1**: correlations table for variables age, cholesterol, systolic blood pressure, diastolic blood pressure, height, and weight*

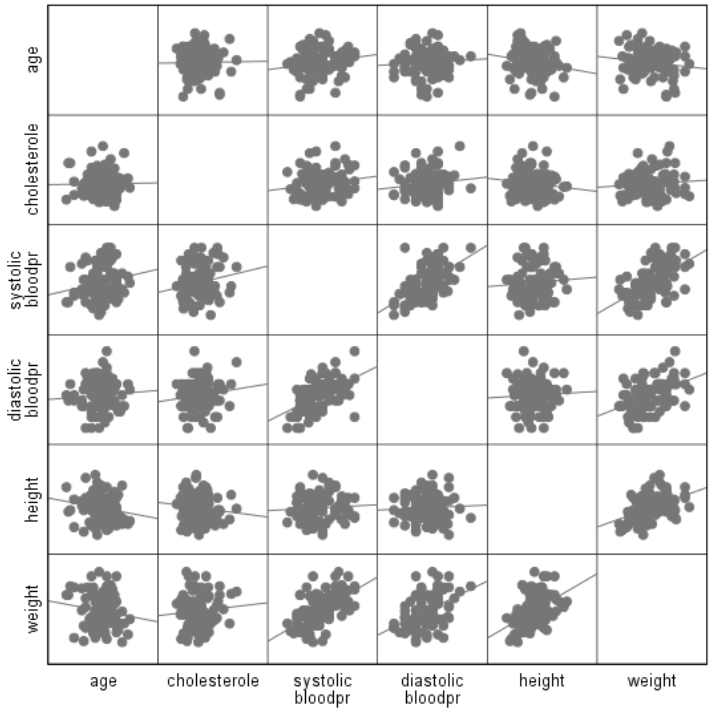


Figure 2.2*: Matrix plot displaying the scatter plots of the different variable combinations. The central empty spaces would have contained graphs of the variable plotted against itself, which would yield no information as it would be a simple oblique line.*

Assignment 3:

File name and description:
Table 7.2.sav - body weight of ewes in two groups: Treatment and control group

Problem:
Table 7.2: 2 independent samples, body weights of ewes, flushed and non-flushed: Examine with t-test whether the body weights of ewes are different in flushed and non-flushed groups.

Conclusions:
The treatment group (flushed ewes) has a higher mean weight (67.36 Kg) than the control group (65.77 Kg).

Analysis:

Since there is a small amount of data (fewer than 30 samples), we should check the normality first in order to make sure that we can use a t-test:

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Treatment	Statistic	df	Sig.	Statistic	df	Sig.
ewe body weight	Flushed	.108	24	.200*	.985	24	.965
	Control	.065	30	.200*	.986	30	.952

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

***Figure 3.1**: checking the normality of the distribution of ewe body weights*

From the Sig. values in the Shapiro-Wilks test we see that both groups are normally distributed so we can use the t-test. The normality is also evident qualitatively when having a look at the corresponding histograms:

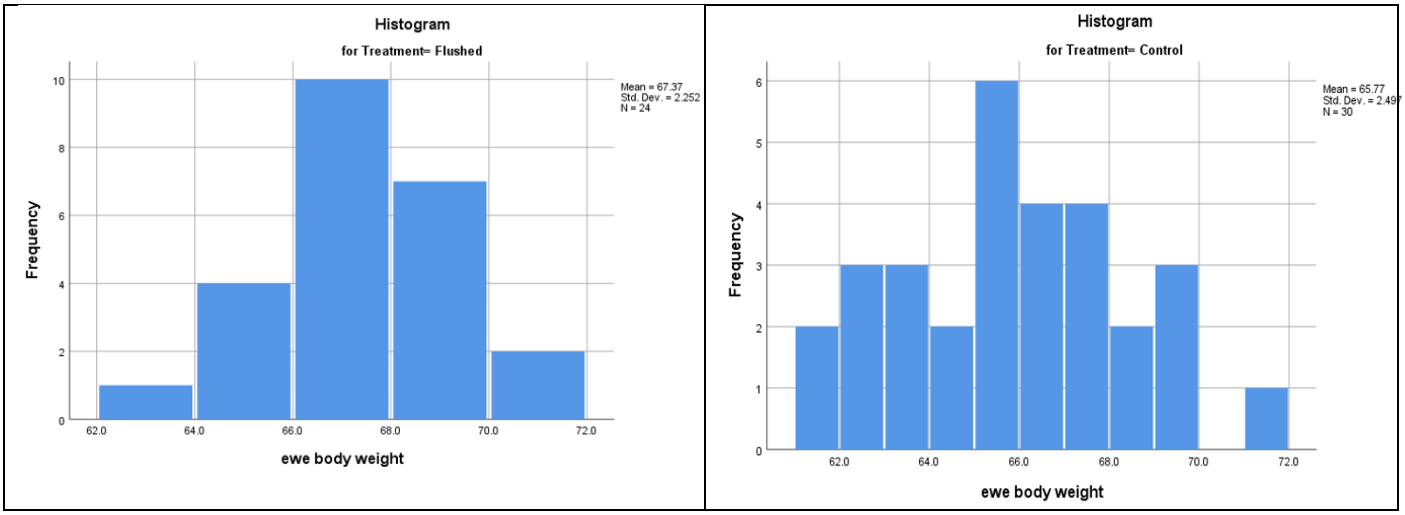


***Figure 3.2**: histograms of the body weights of Flushed ewes and non-Flushed (Control) ewes.*

Now we can get some basic statistics and do the t-test:

Group Statistics

	Treatment	N	Mean	Std. Deviation	Std. Error Mean
ewe body weight	Flushed	24	67.367	2.2525	.4598
	Control	30	65.773	2.4972	.4559

***Figure 3.3**: Group statistics for two groups of ewes (treatment and control)*

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
ewe body weight	Equal variances assumed	.253	.617	2.432	52	.018	1.5933	.6551	.2788	2.9079
	Equal variances not assumed			2.461	51.204	.017	1.5933	.6475	.2935	2.8931

***Figure 3.4**: t-test for independent samples*

Levene's test for equality of variables indicates that the variances are similar.

Since we can assume similar variances, we use the first row of the Independent Samples Test table. We have a t = 2.432 with 52 degrees of freedom and the p-value of 0.018 (=Sig. 2-tailed). This means that we must reject the Null Hypothesis and the two groups are dissimilar – that is, the difference in weights cannot be wholly attributed to random factors.

We therefore must conclude that the treatment group (flushed ewes) has a higher mean weight (67.36 Kg) than the control group (65.77 Kg).

Assignment 4:

File name and description:

ttestpr.sav – study of blood pressure before and after treatment

Problem:

ttestpr data: Examine with t-test whether the blood pressures are different before and after of using the blood pressure drug as treatment.

Conclusions:

The t-test analysis shows us that there are indeed differences in the two groups. When we look at the two group means we see that people have lower blood pressures after the treatment.

Doing additionally a qualitative study and a non-parametric test confirms the conclusion.

Analysis:

This is a case of "related" samples because it's the same people/cases in all groups. It is a small data set, so we would be justified in using non-parametric tests (Wilcoxon), but we shall study the normality first to see whether we can use a parametric t-test.

We want to study the normality of the differences, so we need to first create a new variable to calculate the row-wise differences ("diff").

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
diff	.159	10	.200*	.955	10	.730

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Figure 4.1*: testing the normality of the variable "diff" containing the row-wise difference of our before/after data*

With a sig. of 0.73 we can accept the Null Hypothesis: the difference between the rows is normally distributed so a t-test is reliable. We should still however use both the t-test and non-parametric tests so that we can compare the two results since it is such a small amount of data.

Parametric t-test: if the mean of the differences is around zero, then the two groups are similar.

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Before treatment	142.50	10	17.044	5.390
	After treatment	116.40	10	13.615	4.306

***Figure 4.2**: basic statistics of our treatment data. We are using it here to see the group Means.*

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Before treatment & After treatment	10	.199	.582

***Figure 4.3**: Paired Samples Correlations - the correlation value of 0.199 is effectively zero, which means no correlation between the two groups.*

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Before treatment - After treatment	26.100	19.587	6.194	12.088	40.112	4.214	9	.002

***Figure 4.4**: Paired samples test for the paired differences*

Conclusion: The sig. (2-tailed) of 0.002 is tiny (Fig 4.4) so we need to reject the Null Hypothesis. There are indeed differences in the two groups. When we look at the two group means (Fig 4.2) we see that **people have lower blood pressures after the treatment**.

Qualitative Study:

If we want to do a qualitative study of the differences, we can do a nice line chart. Visually we can see that the two curves are not really parallel, so we do not have a high correlation.

***Figure 4.5**: simple line chart of the blood pressures before and after treatment.*

Doing the additional **non-parametric analysis**:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Before treatment and After treatment equals 0.	Related-Samples Wilcoxon Signed Rank Test	.007	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

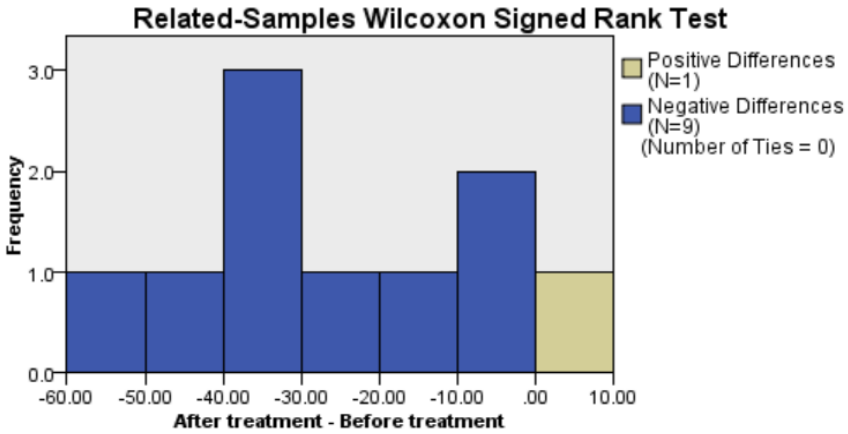


***Figure 4.6**: Table and histogram for the non-parametric test (Related-Samples Wilcoxon Signed Rank Test)*

Conclusion: we reject the Null Hypothesis – which means that there is a difference between the blood pressure value before and after treatment (same conclusion as in the care of the parametric t-test). In the Related-Samples Wilcoxon Signed Rank Test we see that most cases have negative differences and only one case has a positive difference – which tells us that, in all cases except one, the medication lowered the blood pressure.

The parametric test is more accurate since we lose information in the non-parametric analysis due to grouping into ranks, but when having a small amount of data points it is useful to do both tests.

Assignment 5:

File name and description:
table 8.1.sav – teeth calculus in 3 groups of dogs (1 control group and 2 treatment groups)

Problem:
Table 8.1: one-way ANOVA, 3 diets of dogs for affecting on the build-up of calculus of teeth data: Examine with parametric ANOVA whether there are differences in calculus among the diets. Perform also post hoc tests if necessary.

Conclusions:
The ANOVA test tells us that there are differences in the means of at least two groups. The Post-Hoc tests tell us that those differences are between the HMP group and the control group. Looking at the means, the control group has the most amount of calculus and the HMP group has the least amount of calculus.
A qualitative visual analysis of histograms with fitted normal curves illustrates these findings and running non-parametric tests (Kruskal-Wallis) confirms them.

Analysis:

As it is a small data set, we need to first check whether the data is normally distributed so that we can use a parametric test:

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Group	Statistic	df	Sig.	Statistic	df	Sig.
Calculus	Control	.176	9	.200*	.941	9	.592
	P207	.122	9	.200*	.978	9	.955
	HMP	.131	8	.200*	.974	8	.929

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Figure 5.1*: two different tests to check the normality of our data (Kolmogorov-Smirnov and Shapiro-Wilks)*

The normality tests show us that we can accept a normal distribution for all 3 groups of dogs.

We next perform the one-way ANOVA parametric test. Due to the small number of data points, we will also want to do the non-parametric test and then compare the two.

Oneway

Descriptives

Calculus

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Control	9	1.0889	.42254	.14085	.7641	1.4137	.49	1.66
P207	9	.7467	.36953	.12318	.4626	1.0307	.22	1.38
HMP	8	.4375	.29065	.10276	.1945	.6805	.05	.95
Total	26	.7700	.44347	.08697	.5909	.9491	.05	1.66

***Figure 5.2**: general statistics for our data – we can (among other things) see the Means for our three groups of dogs*

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Calculus	Based on Mean	.855	2	23	.438
	Based on Median	.744	2	23	.486
	Based on Median and with adjusted df	.744	2	22.232	.487
	Based on trimmed mean	.861	2	23	.436

***Figure 5.3**: test of homogeneity of Variances*

In fig 5.3, we look at the row where we have the "Calculus based on Mean". With this value of Sig. we will want to accept the Null Hypothesis – that is, the variances are similar.

ANOVA

Calculus

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.805	2	.902	6.668	.005
Within Groups	3.112	23	.135		
Total	4.917	25			

***Figure 5.4**: ANOVA ("**An**alysis **of** **Va**riance") table.*

In the ANOVA table:

- the Sum of Squares Between Groups gives us the total variance
- F indicates Fisher's F statistics
- Sig. is very low (0.005) so we will have to reject the Null Hypothesis: there are indeed some differences between the group means (or at least two of the groups are different)

SPSS only calculates partial eta square. We can calculate eta-square (which ranges from zero to one) directly from our ANOVA table:

Eta-square = sum of squares (between groups) / total sum of squares = 1.805/ 3.112 = 0.37

This result of 0.37 means that the grouping variable explains about 37% of the test variable.

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Calculus

	(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Bonferroni	Control	P207	.34222	.17340	.182	-.1055	.7899
		HMP	.65139*	.17874	.004	.1899	1.1129
	P207	Control	-.34222	.17340	.182	-.7899	.1055
		HMP	.30917	.17874	.291	-.1523	.7707
	HMP	Control	-.65139*	.17874	.004	-1.1129	-.1899
		P207	-.30917	.17874	.291	-.7707	.1523
Dunnett T3	Control	P207	.34222	.18711	.229	-.1546	.8390
		HMP	.65139*	.17435	.006	.1830	1.1198
	P207	Control	-.34222	.18711	.229	-.8390	.1546
		HMP	.30917	.16041	.196	-.1196	.7379
	HMP	Control	-.65139*	.17435	.006	-1.1198	-.1830
		P207	-.30917	.16041	.196	-.7379	.1196

*. The mean difference is significant at the 0.05 level.

***Figure 5.5**: Post-Hoc tests (analysis of the results of our experimental data)*

In fig 5.5 (above) we have a table with multiple comparisons using two tests: Bonferroni and Dunnett. The Bonferroni test is good if there are not a lot of groups (eg., 3 or 4 groups). Dunnett's test is good for control group comparing.

In our table, if there is a difference between the Means, this is represented by the presence of an asterisk. This means that there is a difference between the Means for the Control group and the HMP group.

We can also visualize our data for a qualitative approach:

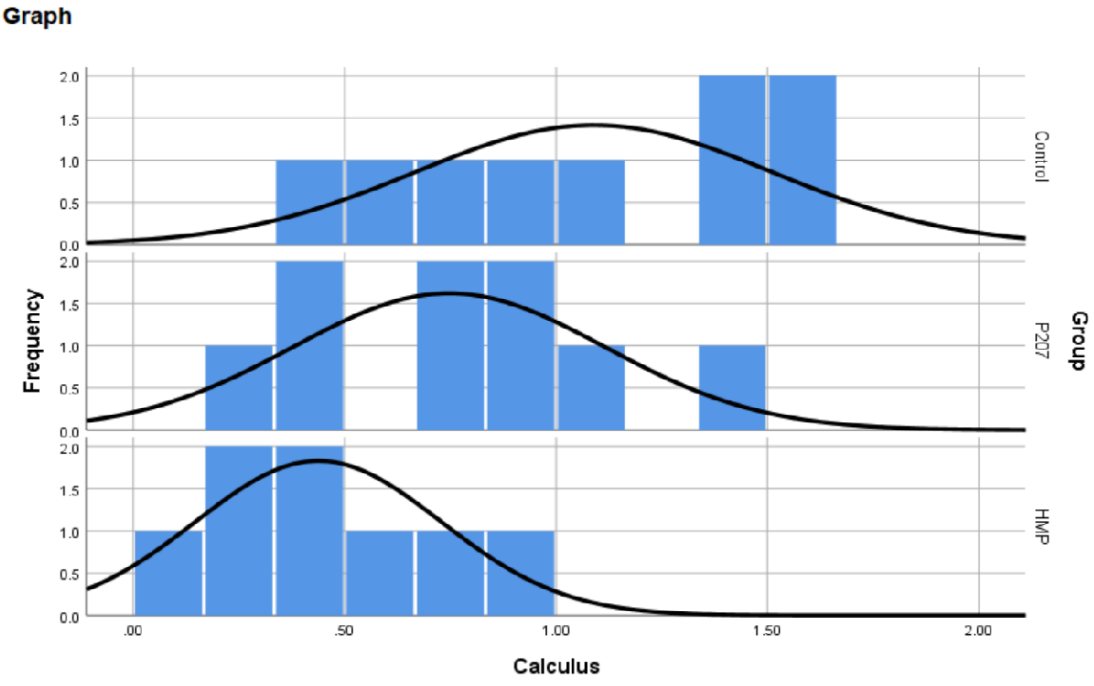


***Figure 5.6**: histograms of the 3 data groups with corresponding normal curves. HMP has clearly less calculus than the control group*

Conclusion: the control group has the most amount of calculus and the HMP group has the least amount of calculus.

We should now do the non-parametric analysis for comparison:

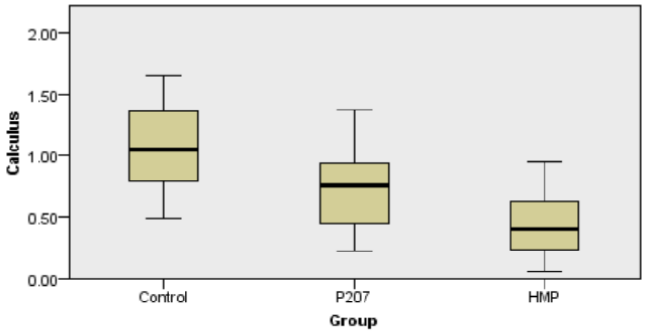
Nonparametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Calculus is the same across categories of Group.	Independent-Samples Kruskal-Wallis Test	.010	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Independent-Samples Kruskal-Wallis Test



Total N	26
Test Statistic	9.198
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	.010

1. The test statistic is adjusted for ties.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
HMP-P207	5.708	3.715	1.536	.124	.373
HMP-Control	11.264	3.715	3.032	.002	.007
P207-Control	5.556	3.604	1.541	.123	.370

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.
Significance values have been adjusted by the Bonferroni correction for multiple tests.

***Figure 5.7**: results from running non-parametric tests. We used the Kruskal-Wallis test because it is very good for studying 3 groups.*

Conclusion: In our non-parametric analysis we reach the same conclusion as in the parametric analysis: the Control group has the most amount of calculus and the HMP group has the least amount of calculus.

When looking at the table of pairwise comparisons, we again get the conclusion that the only difference is between HMP and Control (adjusted sig. = 0.007).

Assignment 6:

File name and description:
Bodyfat.sav – body fat and several different body measurements (midarm, thigh, triceps)

Problem:
Bodyfat data: Perform linear regression analysis in which bodyfat is the dependent, and the other variables are independent variables. a) What is the value of Rsquare? b) Is there multicollinearity? c) What is the regression equation now?

Conclusions:

a) R square = 0.741
b) There is no multicollinearity
c) Regression equation: Y = 105.744 + 0.852.(midarm) - 0.097.(thigh)

Analysis:

The dependent variable in the file Bodyfat.sav is the bodyfat. The independent variables are triceps skinfold thickness, thigh circumference and midarm circumference.

We want to find independent variables with sufficiently high linear correlation with the dependent variable, and with no inter correlations (multi collinearity) between independent variables. We want also to use the smallest amount of independent variables possible to get a suitable but simple model.

We check the correlations from the following correlations matrix:

Correlations

		triceps_skinfold_thickness	thigh_circumference	midarm_circumference	bodyfat
triceps_skinfold_thickness	Pearson Correlation	1	-.101	.366**	.417**
	Sig. (2-tailed)		.371	.001	.000
	N	80	80	80	80
thigh_circumference	Pearson Correlation	-.101	1	.146	-.051
	Sig. (2-tailed)	.371		.195	.655
	N	80	80	80	80
midarm_circumference	Pearson Correlation	.366**	.146	1	.843**
	Sig. (2-tailed)	.001	.195		.000
	N	80	80	80	80
bodyfat	Pearson Correlation	.417**	-.051	.843**	1
	Sig. (2-tailed)	.000	.655	.000	
	N	80	80	80	80

**. Correlation is significant at the 0.01 level (2-tailed).

From the Correlations matrix we can see that:

- The midarm has a high positive correlation (0.843 **)
- The thigh (in practise) has correlation zero (-0.051)
- The triceps only have a bit of correlation (0.417 **)

We could guess from this that a final model for bodyfat will for sure contain midarm, might contain triceps and will probably not contain thigh measurements.

We can best decide which variables to use by using using automatic methods of obtaining a model. We should try to use at least 2 models to compare.

- Forward – adds one variable at a time to the model
- Backward – starts with all the variables and then removes them one at a time
- Stepwise – mixture of the previous two methods

We will start with the **"Stepwise" Method**:

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.843[a]	.710	.706	26.95233
2	.861[b]	.741	.734	25.63731

a. Predictors: (Constant), midarm_circumference

b. Predictors: (Constant), midarm_circumference, thigh_circumference

c. Dependent Variable: bodyfat

It shows us only midarm and thigh as important variables. Model Summary shows what the model would look like with 1 variable or with 2 variables. We choose which is the most worth-while (2 variables is more accurate, but is also a more complicated model).

The Model Summary gives us the goodness criteria (R square and adjusted R square).

R square gives us multiple correlation (if many variables), or the correlation (if single). It varies from 0 to 1. Closer to 1 is better but sometimes even 0.3 is acceptable for a model.

In this case R square = 0.71, which is pretty good. It means that the variable (midarm) explains 71% of variance of the body fat. R increased in the 2 variables model, so more of the total variance is explained (74.1%).

Adjusted R square is usually more reliable. The Adjusted R square decreases if irrelevant variables are added. Here it increased, so the extra variable is relevant. **We want to accept the 2-variable model**.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	87.580	12.705		6.893	.000		
	midarm_circumference	.826	.060	.843	13.821	.000	1.000	1.000
2	(Constant)	105.744	13.487		7.841	.000		
	midarm_circumference	.852	.057	.869	14.817	.000	.979	1.022
	thigh_circumference	-.097	.032	-.178	-3.034	.003	.979	1.022

a. Dependent Variable: bodyfat

In the coefficients table we see that the Collinearity Tolerance and Statistics VIF values are very close to one, so **there is no multicollinearity** (multicollinearity would be a value close to zero).

The coefficients table is the most interesting one. In the "Unstandardized B" column we get constant = 87.58 (this is the y intercept of the linear regression equation) and midarm = 0.826 (this is the slope of the linear regression equation).

The regression equation is now (2-variable model): **Y = 105.744 + 0.852.(midarm) - 0.097.(thigh)**

This means that, if the bodyfat increases by 1 unit, the midarm radius increases by 0.852 units.

The method used to obtain the regression equation was the least squares method, where the sum of the squares of the residuals should be minimized.

The standardized coefficients Beta ranges from 0 to 1 and, the higher the value, the more important the variable is for the model. The most important one here is obviously midarm.

Let's see what we would get with another method:

"Backward" Method:

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.865[a]	.749	.739	25.41999
2	.861[b]	.741	.734	25.63731

a. Predictors: (Constant), triceps_skinfold_thickness, thigh_circumference, midarm_circumference

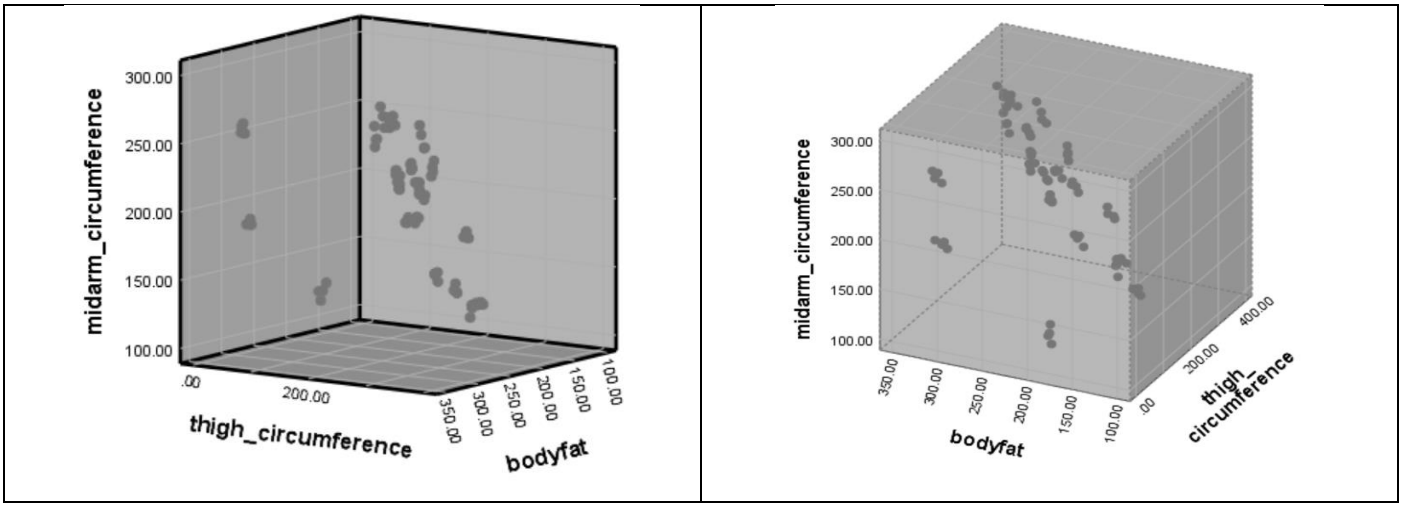
b. Predictors: (Constant), thigh_circumference, midarm_circumference

c. Dependent Variable: bodyfat

In this "Backward" Method we again obtain thigh and midarm as the most important variables. The difference between the R Square in the 1st model (with all 3 variables) and the second model (with only midarm and thigh) is insignificant, whereas the addition of the variable triceps would make the model a lot more complicated and doesn't explain much more of the total variance.

This Method agrees that the best option is a model with only two variables (midarm and thigh). (We already found the equation and checked multicollinearity when examining with the "stepwise" method.)

Using only two independent variables, our regression equation gives a 3D model for studying bodyfat.



As a final step, we should study the residuals. The mean of the residuals should be zero and the residuals should be normally distributed around zero.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	178.7312	325.7867	258.1620	42.81327	80
Std. Predicted Value	-1.855	1.580	.000	1.000	80
Standard Error of Predicted Value	2.909	8.541	4.702	1.604	80
Adjusted Predicted Value	177.7183	326.8779	258.0905	42.76599	80
Residual	-50.49762	61.24629	.00000	25.31071	80
Std. Residual	-1.970	2.389	.000	.987	80
Stud. Residual	-2.008	2.476	.001	1.007	80
Deleted Residual	-52.49277	65.81510	.07151	26.33238	80
Stud. Deleted Residual	-2.050	2.565	.003	1.021	80
Mahal. Distance	.029	7.780	1.975	2.116	80
Cook's Distance	.000	.152	.014	.027	80
Centered Leverage Value	.000	.098	.025	.027	80

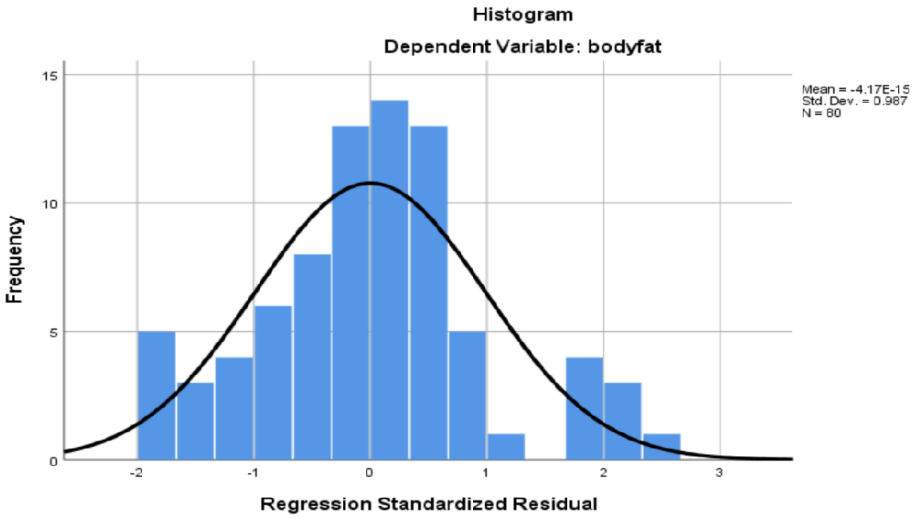
a. Dependent Variable: bodyfat

The distance between the points and the fitted curve is the “residuals”. When a residual is too far from the curve we call it an “outlier” (which is usually removed and studied separately).

We see in the Residuals Statistics table that:

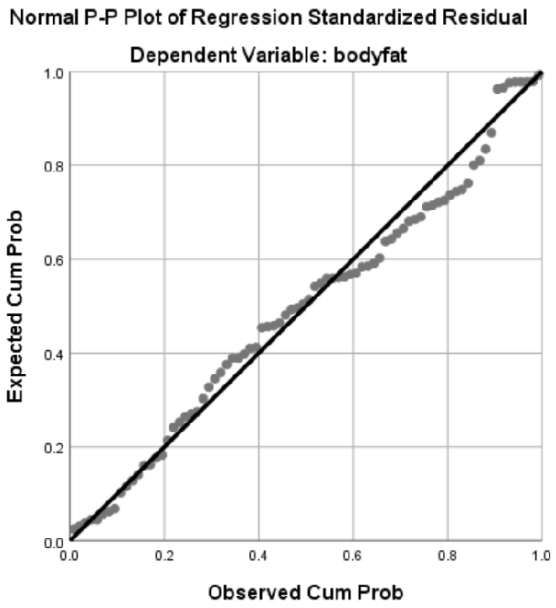
- the Mean is 0.000 (which is ideal)
- Stud. Residual is between -2 and +2, so there are no outliers

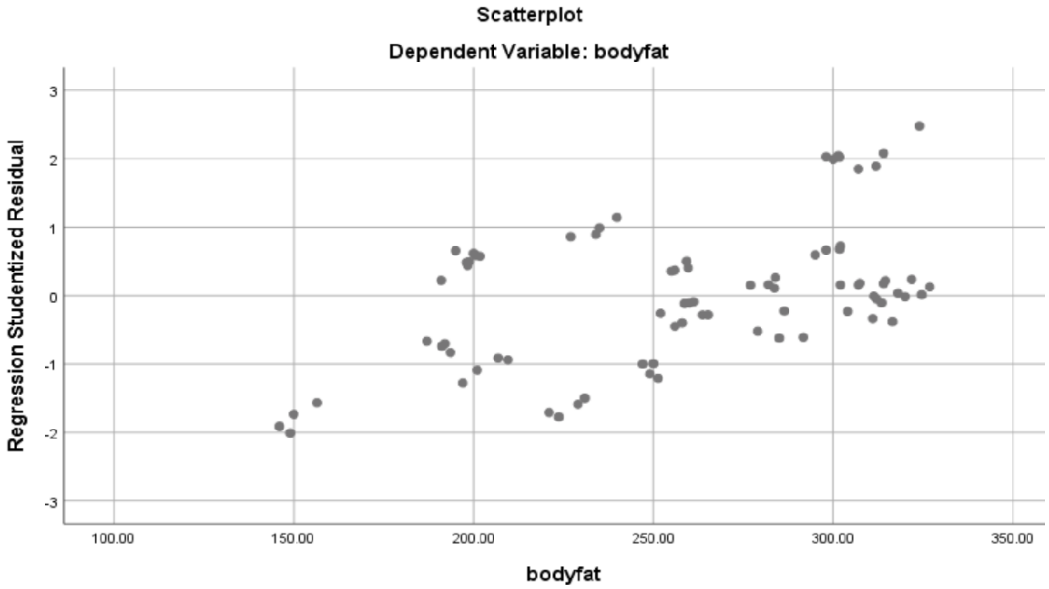
In the histogram we see the Mean is zero and the residuals are more or less normally distributed around zero (as they should).



The normal p-p plot shows our data very close to the line.

We could also have used Shapiro-Wilks to study the normality of the residuals, but these graphics are sufficient in this instance.





The scatterplot is important for see the location around the horizontal zero line. If random location of residuals (acceptable to a point), then the residuals are due to random errors. If we could see all points above or below the zero horizontal line, then we would be dealing with a systematic error. The ideal distribution of residuals would be a box-like figure (the narrower the better).

Assignment 7:

File name and description:

disease.sav – disease (0 = healthy, 1 = sick), age, social class (1 to 3, where 1 is the lowest and 3 is the highest) and sector (where the person lives: there are 3 city sectors)

Problem:

Disease data: Perform logistic regression analysis (classification cutoff = 0.5) in which age and sector are independent variables, and disease is dependent variable. a) How much is pseudo Rsquare? b) Odds ratio value in the case of sector? c) Does the model contain insignificant independent variables according to Wald's test?

Conclusions:

a) The Nagelkerke R square (or "pseudo R square) is 0.169
b) Exp(B) = 3.26 is the odds ratio in the case of city sector
c) Yes. According to Wald's test, Socioeconomical Status is an insignificant independent variable

Also:

- Sector 2 has a higher probability of having the disease.
- The older the people, the higher the probability that they will have the disease (in either sector)

Analysis:

In Logistic Regression Analysis the dependent variable is always dichotomous (it has two values: 0 or 1). If the two values are different, SPSS will transform them into 0 ("control") and 1 ("treatment"). We can use both categorical (discrete) and continuous variables in this kind of analysis.

In the file disease.sav our depend variable is "disease" (dichotomous; 0 = healthy and 1 = sick) and the predictors (independent variables) are discreet (**age**, social **class** and **sector** of residence).

We use at least two automatic methods to help us choose the significant predictors for a good model.

Starting with **Method: Forward Stepwise (Conditional)**

Block 1: Method = Forward Stepwise (Conditional)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 2	Step	9.957	1	.002
	Block	24.690	2	.000
	Model	24.690	2	.000

"Step" compares our last model to Block 0 which would be the regression model without any of the independent variables included (contains only the constant). With a Sig. = 0.002, the Null Hypothesis is rejected so our model is better than the zero model (Block 0).

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
2	211.639[a]	.118	.169

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Model summary gives us the model goodness criteria:

- -2 Log likelihood - must be minimized (as small as possible)
- Nagelkerke R square – equivalent to R square in the linear case. Ranges from 0 to 1 and should be as high as possible.

Classification Table[a]

Observed			Predicted: Disease status: no	Predicted: Disease status: yes	Percentage Correct
Step 2	Disease status	no	130	9	93.5
		yes	40	17	29.8
	Overall Percentage				75.0

a. The cut value is .500

The Classification Table compares the predicted values with the original disease values with a default cut value of 0.5

We see that this model is good at predicting healthy people (93.5% correct) but bad at predicting sick people (only 29.8% correct), which is not good.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 2[a]	City sector(1)	1.182	.337	12.298	1	.000	3.260	1.684	6.310
	Age	.027	.009	9.608	1	.002	1.027	1.010	1.045
	Constant	-2.160	.344	39.436	1	.000	.115		

a. Variable(s) entered on step 2: Age.

Variables not in the Equation

			Score	df	Sig.
Step 2	Variables	Socioecon status	.420	2	.810
		Socioecon status(1)	.027	1	.869
		Socioecon status(2)	.410	1	.522
	Overall Statistics		.420	2	.810

In the “Variables in the Equation” we see that age and city sector have small p values (Sig. =0.02 and =0.000 respectively) which means that they are significant variables in the model, whereas social status had high p value and is therefore insignificant for the model.

Checking what another Method will suggest as to significant variables:

Method: Backward Stepwise (Conditional)

Block 1: Method = Backward Stepwise (Conditional)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	25.109	4	.000
	Block	25.109	4	.000
	Model	25.109	4	.000
Step 2[a]	Step	-.419	2	.811
	Block	24.690	2	.000
	Model	24.690	2	.000

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Socioecon status			.420	2	.811			
	Socioecon status(1)	.045	.433	.011	1	.918	1.046	.448	2.441
	Socioecon status(2)	.253	.406	.391	1	.532	1.288	.582	2.853
	City sector(1)	1.244	.352	12.462	1	.000	3.468	1.739	6.918
	Age	.027	.009	9.680	1	.002	1.027	1.010	1.045
	Constant	-2.294	.437	27.579	1	.000	.101		
Step 2[a]	City sector(1)	1.182	.337	12.298	1	.000	3.260	1.684	6.310
	Age	.027	.009	9.608	1	.002	1.027	1.010	1.045
	Constant	-2.160	.344	39.436	1	.000	.115		

a. Variable(s) entered on step 1: Socioecon status, City sector, Age.

Variables not in the Equation

			Score	df	Sig.
Step 2[a]	Variables	Socioecon status	.420	2	.810
		Socioecon status(1)	.027	1	.869
		Socioecon status(2)	.410	1	.522
	Overall Statistics		.420	2	.810

a. Variable(s) removed on step 2: Socioecon status.

This Method agrees with the findings of the previous method: Age and City Sector are significant variables in the model, whereas socioeconomical status is not.

We further see in the "Variables in the Equation" table above that the **Wald's test** score for the socioeconomical status is very low (which is not the case for age and city sector), hence **socioeconomical status is an insignificant variable** and can be excluded from the model.

We will now focus on just the two significant variables and use Method "Enter":

Method: Enter

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	211.639[a]	.118	.169

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

The Nagelkerke R square (or "pseudo R square") is 0.169

Variables in the Equation

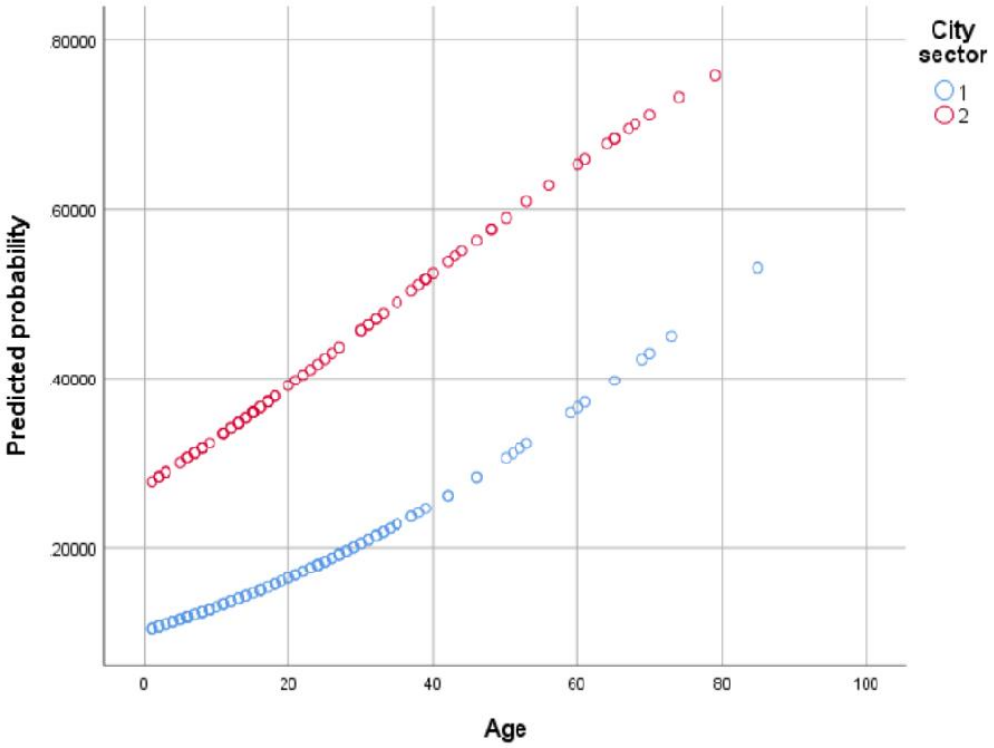
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	City sector(1)	1.182	.337	12.298	1	.000	3.260	1.684	6.310
	Age	.027	.009	9.608	1	.002	1.027	1.010	1.045
	Constant	-2.160	.344	39.436	1	.000	.115		

a. Variable(s) entered on step 1: City sector, Age.

Looking at the Variables in the Equation table:

- **Age**, which is a continuous variable: we look at the B values. If the B value is significant (limit point is zero) and greater than zero, it means that **the older the person, the higher the risk of having the disease** – this is what we have here. (Conversely, if negative B, the older the person the lower the risk of having the disease.)
- **City Sector**: Exp(B) = 3.26 is the odds ratio. Limit point is 1. If significant and above 1, the higher the sector, the higher the risk of disease. In sector 2, the odds for the disease is 3.26 higher than in sector 1.
- **95% confidence interval**: if the interval is not very large, it is reliable. The biggest range here is for sector (at least 1.7 to about 6.3) which is a small enough range.

Looking at a scatter plot of our predicted probability of getting sick, we see that:



- Sector 2 has a higher probability of having the disease.
- The older the people, the higher the probability that they will have the disease (in either sector)