Non-Parametric Tests, Correlation, and Regression

Data Analysis Group A

Contents

Data Set Description

1

Dataset

- 14 variables: Age, Sex, Chest pain type, BP, Cholesterol, FBS over 120, EKG results, Max HR, Exercise angina, ST depression, Slope of ST, Number of vessels fluro, Thallium, heart disease presence
- 270 rows of data
- The variables consist of a mix of numerical and categorical variables

Non-Parametric Tests

2

Kruskal-Wallis Test

- Research Objective: Explore potential relationships
  - Variables:
    - Chest Pain Type
    - Maximum Heart Rate (Max HR)
    - EKG Results
    - Blood Pressure (BP)
  - Grouped Dependent Variable:
    - Heart Disease Presence/Absence
- Hypothesis:
  - Null: Chest pain, Max HR, EKG, and BP have no effect on heart disease
  - Alternative: At least one factor affects heart disease

Kruskal-Wallis Test

- Descriptive statistics:
  - Chest pain type
  - Max HR
  - EKG results
  - BP (Blood pressure)
  - Heart disease presence

- Ranks
  - Mean ranks and their correlation to each other.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Chest pain type	270	3.17	.950	1	4
Max HR	259	149.32	22.995	71	202
EKG results	270	1.02	.998	0	2
BP	255	131.09	18.116	94	200
HeartDisease	270	.44	.498	0	1

Ranks

	HeartDisease	N	Mean Rank
Chest pain type	0	150	105.08
	1	120	173.53
	Total	270	
Max HR	0	144	157.98
	1	115	94.97
	Total	259	
EKG results	0	150	124.47
	1	120	149.29
	Total	270	
BP	0	145	121.47
	1	110	136.60
	Total	255	

Hypothesis

H0: There is no difference between groups.

H1: There is a difference between groups

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Chest pain type is the same across categories of Heart Disease.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
2	The distribution of EKG results is the same across categories of Heart Disease.	Independent-Samples Kruskal-Wallis Test	.003	Reject the null hypothesis.
3	The distribution of Max HR is the same across categories of Heart Disease.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
4	The distribution of BP is the same across categories of Heart Disease.	Independent-Samples Kruskal-Wallis Test	.104	Retain the null hypothesis.

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

Wilcoxon Test

- The Wilcoxon test is a non-parametric statistical test used to compare two paired groups when the data does not meet the assumptions of the paired t-test. It is specifically designed for situations where each observation in one group is paired with a corresponding observation in the other group.

- Key points:
  - Non-parametric: It doesn't assume a specific distribution of the data.
  - Paired Observations: It requires each data point in the two groups to be matched or paired.
  - Hypothesis Testing: It assesses whether the medians of the two groups are significantly different.
  - Assumptions: Unlike the paired t-test, it doesn't require the data to be normally distributed or the variances to be equal.

- Similarly to the paired t-test done in our previous presentation, this test cannot be done

Mann-Whitney U Test BP

- sex 0 and sex 1.
- Difference between the BP/Max HR/Cholesterol/Heart disease (changed to numeric and ordinal, 1 and 2) of the two sexes (changed to nominal).
- H0: The sum of the two rankings does not differ in the population.
- H1: The sum of the two rankings differs in the population.

Ranks

	Sex	N	Mean Rank	Sum of Ranks
SMEAN(BP)	0	86	141,19	12142,50
	1	183	132,09	24172,50
	Total	269		

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of SMEAN(BP) is the same across categories of Sex.	Independent-Samples Mann-Whitney U Test	,370	Retain the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

Test Statistics[a]

	SMEAN(BP)
Mann-Whitney U	7336,500
Wilcoxon W	24172,500
Z	-,897
Asymp. Sig. (2-tailed)	,370

a. Grouping Variable: Sex

Mann-Whitney U Test Max HR

- H0: The sum of the two rankings does not differ in the population.
- H1: The sum of the two rankings differs in the population.

H1 Rejected.

Ranks

	Sex	N	Mean Rank	Sum of Ranks
SMEAN(MaxHR)	0	86	144,98	12468,00
	1	183	130,31	23847,00
	Total	269		

Test Statistics[a]

	SMEAN (MaxHR)
Mann-Whitney U	7011,000
Wilcoxon W	23847,000
Z	-1,442
Asymp. Sig. (2-tailed)	,149

a. Grouping Variable: Sex

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of SMEAN (MaxHR) is the same across categories of Sex.	Independent-Samples Mann-Whitney U Test	,149	Retain the null hypothesis.

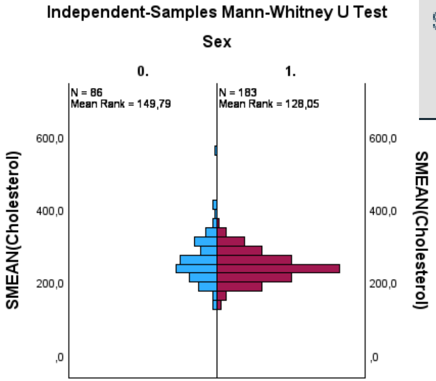
a. The significance level is ,050.

b. Asymptotic significance is displayed.

- H0: The sum of the two rankings does not differ in the population.
- H1: The sum of the two rankings differs in the population.

H1 Accepted.

Mann-Whitney U Test Cholesterol



Ranks

	Sex	N	Mean Rank	Sum of Ranks
SMEAN(Cholesterol)	0	86	149,79	12882,00
	1	183	128,05	23433,00
	Total	269		

Test Statistics[a]

	SMEAN (Cholesterol)
Mann-Whitney U	6597,000
Wilcoxon W	23433,000
Z	-2,138
Asymp. Sig. (2-tailed)	,033

a. Grouping Variable: Sex

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of SMEAN (Cholesterol) is the same across categories of Sex.	Independent-Samples Mann-Whitney U Test	,033	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

- H0: The sum of the two rankings does not differ in the population.
- H1: The sum of the two rankings differs in the population.
- H1 Accepted.

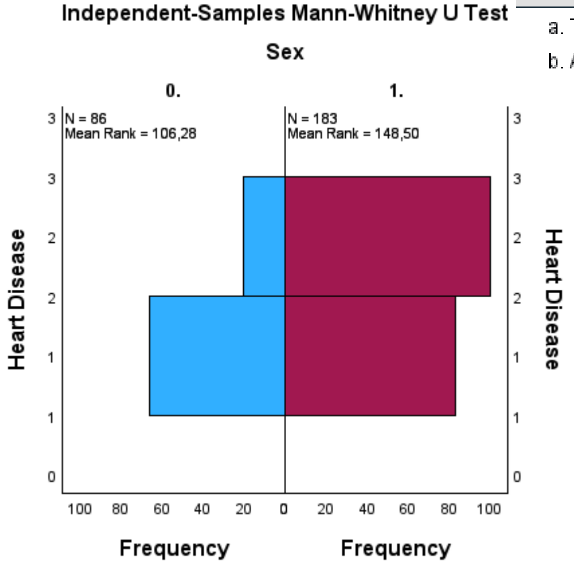
Mann-Whitney U Test Presence of Heart Disease

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Heart Disease is the same across categories of Sex.	Independent-Samples Mann-Whitney U Test	<,001	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.



Ranks

	Sex	N	Mean Rank	Sum of Ranks
Heart Disease	0	86	106,28	9140,00
	1	183	148,50	27175,00
	Total	269		

Test Statistics[a]

	Heart Disease
Mann-Whitney U	5399,000
Wilcoxon W	9140,000
Z	-4,821
Asymp. Sig. (2-tailed)	<,001

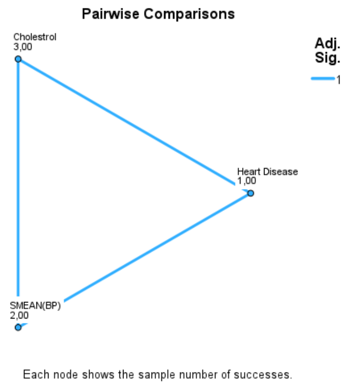
a. Grouping Variable: Sex

Friedman Test

- Measuring the effect of the patient’s BP, Cholesterol and presence of heart disease. Were **tests** for BP and cholesterol be effective for determining presence of heart disease?
- H0: there is no significant difference between the ranks of the dependent groups.
- H1: there is a significant difference between the ranks of the dependent groups.
- Null hypothesis is rejected.
- H1: is supported, there is a significant difference between the rank sums of BP, cholesterol readings and heart disease determination.
- Therefore, indeed the tests applied for BP and cholesterol are effective in giving us information useful to the diagnosis of heart disease.
- Comparatively, the Cholesterol proved with highest mean-ranking a more effective test from which we draw out sources of information on diagnosing heart disease.

Ranks

	Mean Rank
Heart Disease	1,00
SMEAN(BP)	2,00
SMEAN(Cholesterol)	3,00



Pairwise Comparisons

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
Heart Disease-SMEAN(BP)	-1,004	,086	-11,641	<,001	,000
Heart Disease-SMEAN (Cholesterol)	-,993	,086	-11,511	<,001	,000
SMEAN(BP)-SMEAN (Cholesterol)	-1,996	,086	-23,152	<,001	,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 ,050.

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

Related-Samples Friedman's Two-Way Analysis of Variance by Ranks Summary

Total N	269
Test Statistic	536,007
Degree Of Freedom	2
Asymptotic Sig.(2-sided test)	<,001

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distributions of Heart Disease, SMEAN(BP) and SMEAN (Cholesterol) are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	<,001	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

3 Correlation

Correlation

- In general, people might assume that variables such as age and blood pressure correlate to cholesterol level
- This might be useful since age and blood pressure are easier to measure than cholesterol
- In this analysis, we want to see if our data set agrees with these assumptions
- We will perform a correlation analysis for Cholesterol vs. Age and Cholesterol vs. BP
- In this case, cholesterol will be the dependant variable

Correlation

- **Null Hypothesis (H0):** there is no correlation between the variables under consideration
- **Alternate Hypothesis (H1):** there is a correlation between the variables under consideration

Correlation

- Correlation coefficients:
  - -1 (strong negative correlation) to 1 (strong positive correlation)
  - As these values get closer to zero, the strength of the correlation decreases, with 0 being no correlation
- P-values:
  - $P < 0.05 \rightarrow$ the correlation coefficient is statistically significant
  - $P > 0.05 \rightarrow$ the correlation coefficient is not statistically significant

Pearson Correlation

- In a previous presentation, we determined that the variables Cholesterol and Age are approximately normally distributed so we will be using Pearson correlation analysis for Cholesterol vs. Age

	Histograms	Box Plots	Q-Q Plots	P-P Plots	Skewness	Kurtosis	Shapio-Wilk	Kolmogorov-Smirnov	Anderson-Darling
Age	✓	✓	✓	✓	✓	✓	✗	✗	✗
BP	✓	✓	✓	✓	✗	✗	✗	✗	✗
Cholesterol	✓	✓	✓	✓	✓	✓	✗	✓	✗
Max HR	✗	✗	✓	✓	✗	✗	✗	✗	✗
ST Depression	✗	✗	✗	✗	✗	✗	✗	✗	✗

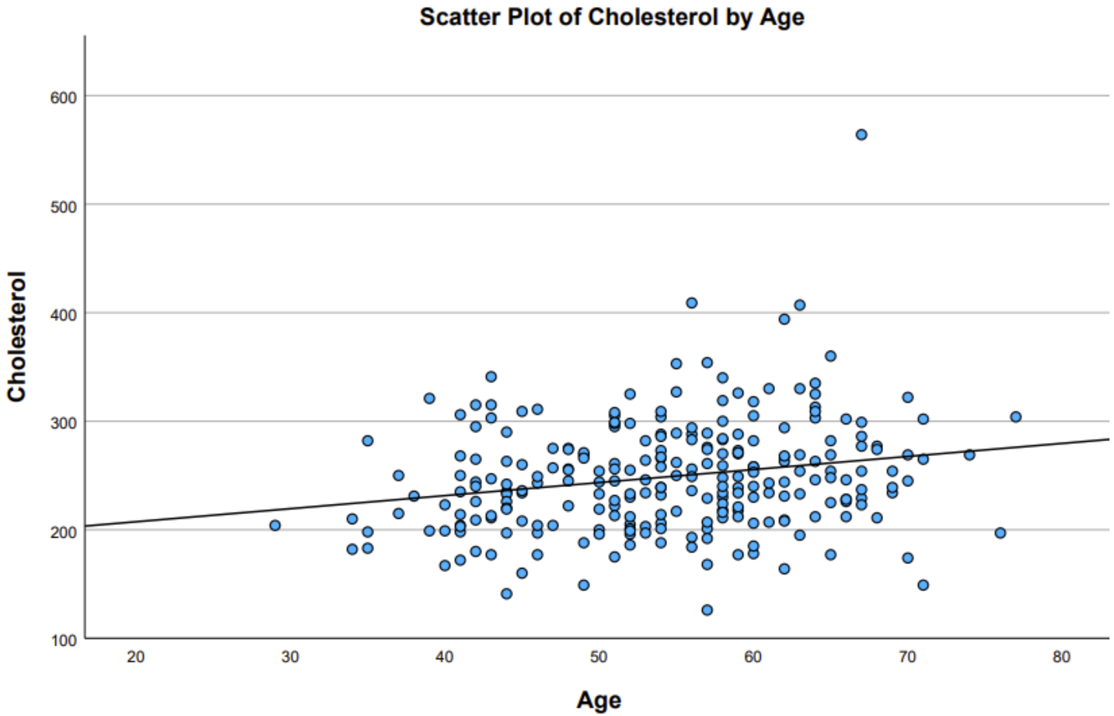
Pearson Correlation

Correlations

		Age	Cholesterol
Age	Pearson Correlation	1	.215**
	Sig. (2-tailed)		<.001
	N	270	253
Cholesterol	Pearson Correlation	.215**	1
	Sig. (2-tailed)	<.001	
	N	253	253

- Correlation coefficient = 0.215 → weak positive correlation
- P-value < 0.05 → correlation coefficient is statistically significant
- Accept alternate hypothesis: there is a correlation between Cholesterol and Age

Pearson Correlation



Spearman Correlation

- In a previous presentation, we determined that the variable BP is not approximately normally distributed so we will be using Spearman correlation analysis for Cholesterol vs. BP

	Histograms	Box Plots	Q-Q Plots	P-P Plots	Skewness	Kurtosis	Shapio-Wilk	Kolmogorov-Smirnov	Anderson-Darling
Age	✓	✓	✓	✓	✓	✓	✗	✗	✗
BP	✓	✓	✓	✓	✗	✗	✗	✗	✗
Cholesterol	✓	✓	✓	✓	✓	✓	✗	✓	✗
Max HR	✗	✗	✓	✓	✗	✗	✗	✗	✗
ST Depression	✗	✗	✗	✗	✗	✗	✗	✗	✗

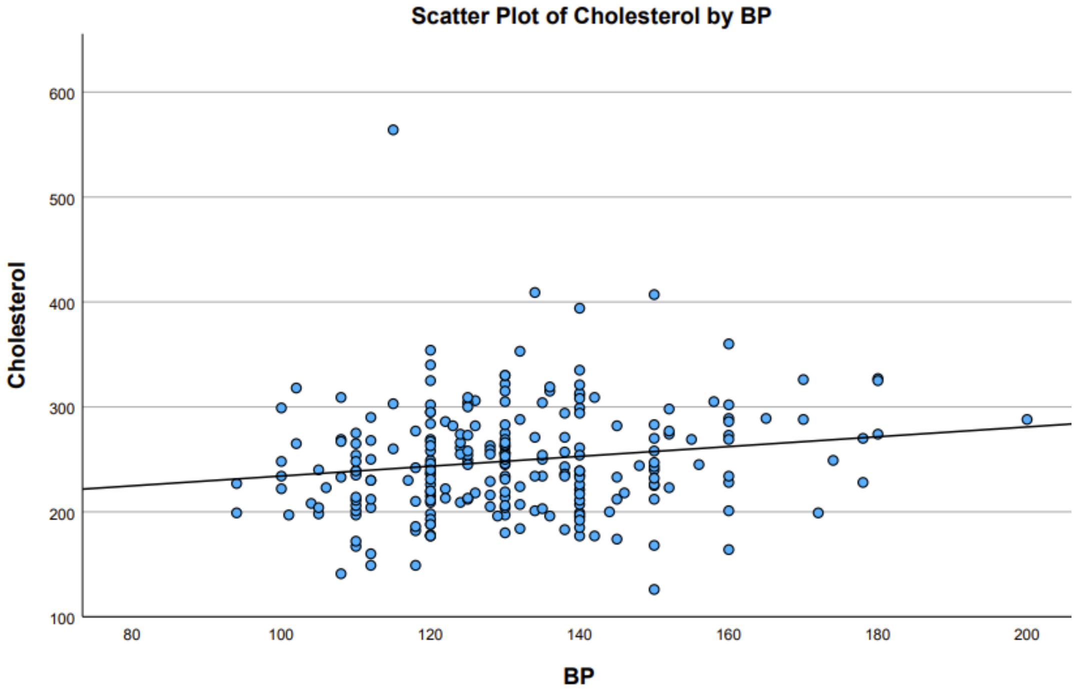
Spearman Correlation

Correlations

			Cholesterol	BP
Spearman's rho	Cholesterol	Correlation Coefficient	1.000	.174**
		Sig. (2-tailed)	.	.007
		N	253	240
	BP	Correlation Coefficient	.174**	1.000
		Sig. (2-tailed)	.007	.
		N	240	255

- Correlation coefficient = 0.174 → weak positive correlation
- P-value < 0.05 → correlation coefficient is statistically significant
- Accept alternate hypothesis: there is a correlation between Cholesterol and BP

Spearman Correlation



Regression 4

Linear Regression

- Is used to **predict** the value of a variable based on the value of another variable.
- The variable you want to predict is called the **dependent** variable.
- The variable you are using to predict the other variable's value is called the **independent** variable.
- Estimates the **coefficients** of the **linear equation**, involving one or more independent variables that best predict the value of the dependent variable.

Simple Linear Regression

- With the Anova table we can check the effectiveness of variables
- As p-value is less than 5% we can say that ST depression is effective on MaxHR.

- R shows the correlation between the predictor variable, x, and the response variable.
- R Square shows the percentage of the model variance that can be predict by the independent variable

Descriptive Statistics

	Mean	Std. Deviation	N
SMEAN(MaxHR)	149,346	22,9499	255
SMEAN(STdepression)	,9713	1,01550	255

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16839,863	1	16839,863	36,433	<,001[b]
	Residual	116941,767	253	462,220		
	Total	133781,630	254			

a. Dependent Variable: SMEAN(MaxHR)

b. Predictors: (Constant), SMEAN(STdepression)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,355[a]	,126	,122	21,4993	,126	36,433	1	253	<,001

a. Predictors: (Constant), SMEAN(STdepression)

b. Dependent Variable: SMEAN(MaxHR)

Simple Linear Regression

- In the correlation table we can see the correlations of all the variables used in the linear regression.
- ST depression is negatively correlated with maximum heart rate by 35%.

- Our linear model is written like this:

  ŷ= -8,018 x1 + 157,134

- Standardized beta show the power of the independent variable to explain the dependent variable

Correlations

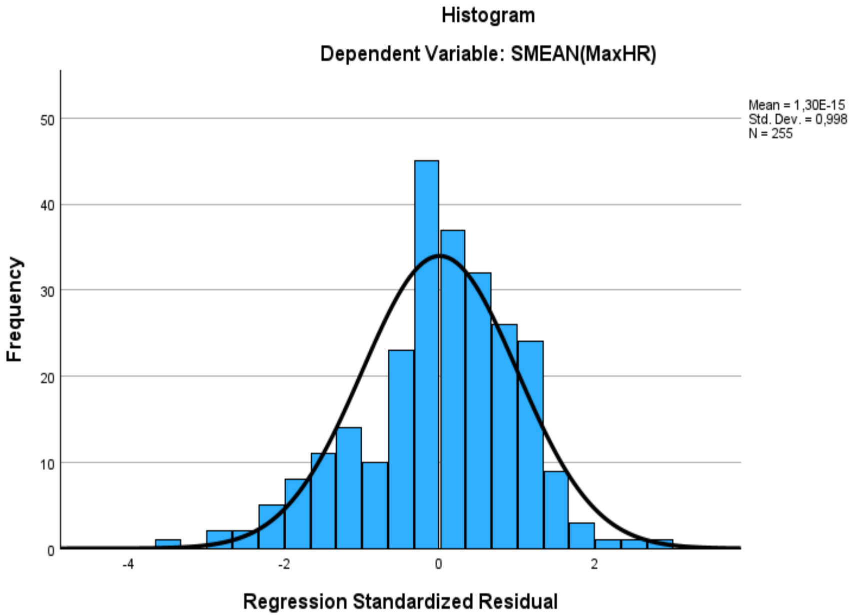
		SMEAN (MaxHR)	SMEAN (STdepression)
Pearson Correlation	SMEAN(MaxHR)	1,000	-,355
	SMEAN(STdepression)	-,355	1,000
Sig. (1-tailed)	SMEAN(MaxHR)	.	<,001
	SMEAN(STdepression)	,000	.
N	SMEAN(MaxHR)	255	255
	SMEAN(STdepression)	255	255

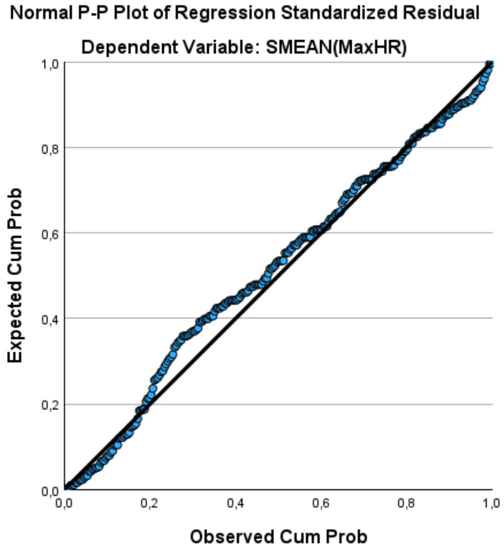
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	157,134	1,865		84,263	<,001
	SMEAN(STdepression)	-8,018	1,328	-,355	-6,036	<,001

a. Dependent Variable: SMEAN(MaxHR)

Simple Linear Regression





Multiple Linear Regression

- In our multiple linear regression we decide to predict the value of maximum heart rate with the variables ST depression, age, blood pressure and exercise angina.

- Significant level in Anova table is less than 5%.The model is effective on the dependent variable.

Descriptive Statistics

	Mean	Std. Deviation	N
SMEAN(MaxHR)	149,346	22,9499	255
SMEAN(STdepression)	,9713	1,01550	255
Age	54,20	9,226	255
SMEAN(BP)	129,896	16,4039	255
Exercise angina	,33	,469	255

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	41406,870	4	10351,718	28,016	<,001[b]
	Residual	92374,760	250	369,499		
	Total	133781,630	254			

a. Dependent Variable: SMEAN(MaxHR)

b. Predictors: (Constant), Exercise angina, SMEAN(BP), Age, SMEAN(STdepression)

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
1	,556[a]	,310	,298	19,2224	,310	28,016	4	250	<,001

a. Predictors: (Constant), Exercise angina, SMEAN(BP), Age, SMEAN(STdepression)

b. Dependent Variable: SMEAN(MaxHR)

Multiple Linear Regression

- We can see the correlations of the independent variables with the dependent variable. All of the variables are negatively correlated.
- Our multiple linear regression is written like this:

  ŷ= -4,801 x1- 0,861 x2 +0,147 x3 -13,265x4 + 185,955
- Significant levels less than 5% indicate that the coefficient is meaningful for the prediction.

Correlations

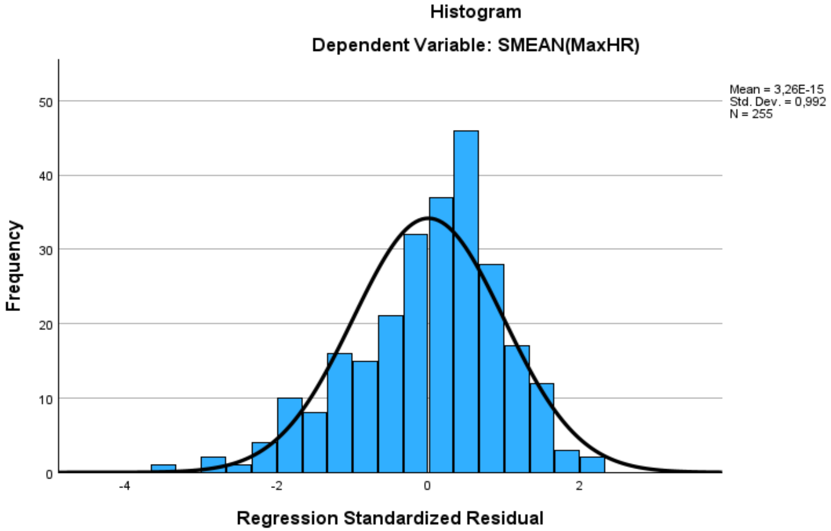
		SMEAN (MaxHR)	SMEAN (STdepression)	Age	SMEAN(BP)	Exercise angina
Pearson Correlation	SMEAN(MaxHR)	1,000	-,355	-,388	-,017	-,374
	SMEAN(STdepression)	-,355	1,000	,193	,127	,327
	Age	-,388	,193	1,000	,256	,103
	SMEAN(BP)	-,017	,127	,256	1,000	,024
	Exercise angina	-,374	,327	,103	,024	1,000
Sig. (1-tailed)	SMEAN(MaxHR)	.	<,001	<,001	,393	<,001
	SMEAN(STdepression)	,000	.	,001	,022	,000
	Age	,000	,001	.	,000	,050
	SMEAN(BP)	,393	,022	,000	.	,352
	Exercise angina	,000	,000	,050	,352	.
N	SMEAN(MaxHR)	255	255	255	255	255
	SMEAN(STdepression)	255	255	255	255	255
	Age	255	255	255	255	255
	SMEAN(BP)	255	255	255	255	255
	Exercise angina	255	255	255	255	255

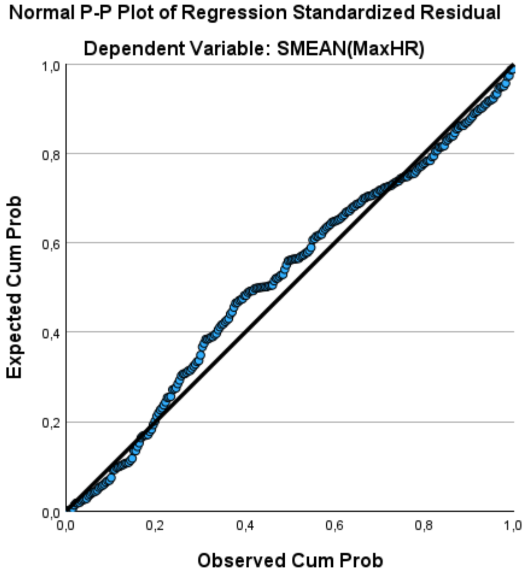
Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	185,955	10,802		17,215	<,001
	SMEAN(STdepression)	-4,801	1,280	-,212	-3,750	<,001
	Age	-,861	,137	-,346	-6,270	<,001
	SMEAN(BP)	,147	,076	,105	1,920	,056
	Exercise angina	-13,265	2,722	-,271	-4,872	<,001

a. Dependent Variable: SMEAN(MaxHR)

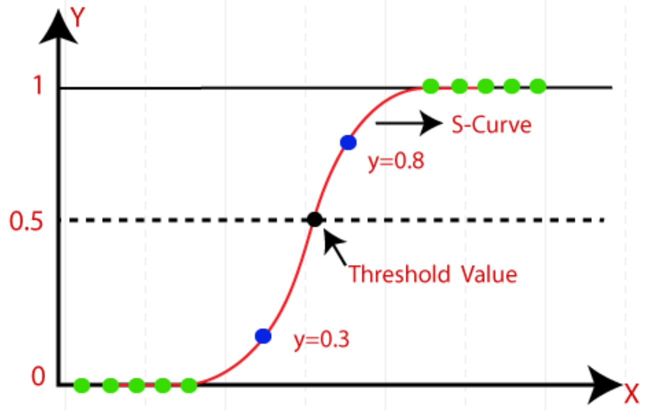
Multiple Linear Regression





Logistic Regression

- A type of statistical model that is often used for classification and predictive analytics.
- Logistic regression maps the predicted values to probabilities of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables.
- Since the outcome is a probability, the dependent variable is bounded between 0 and 1.



Binary Logistic Regression

Assumptions:

1. The dependent variable should be measured on a **dichotomous scale**. (e.g: heart disease or no heart disease)
2. One or more independent variables, which can be either **continuous** (i.e., an interval or ratio variable) or **categorical** (i.e., an ordinal or nominal variable).
3. Independence of observations; and the dependent variable should have **mutually exclusive** and **exhaustive categories**.
4. There needs to be a linear relationship between any continuous independent variables and the logit transformation of the dependent variable. ->> Box Tidwel Test

Example:

We want to test whether we can predict the presence of heart disease using the independent variables:

- Sex (Binary)
- Age (Continuous)
- Cholesterol (Continuous) (From previous research)

Baseline Model

Classification Table[a,b]

Observed			Predicted: Heart Disease: Absence	Predicted: Heart Disease: Presence	Predicted: Percentage Correct
Step 0	Heart Disease	Absence	144	0	100.0
		Presence	109	0	.0
	Overall Percentage				56.9

a. Constant is included in the model.

b. The cut value is .500

Model after including predictors

Classification Table[a]

Observed			Predicted: Heart Disease: Absence	Predicted: Heart Disease: Presence	Predicted: Percentage Correct
Step 1	Heart Disease	Absence	110	$\boxed{34}$	76.4
		Presence	$\boxed{44}$	65	59.6
	Overall Percentage				69.2

a. The cut value is .500

False negatives

False positives

What if we include more independent variables?

(BP, Chest pain type, FBS over 120, Max HR, Exercise Angina… etc.)

- Between 56%-75% of the variance in the dependent variable is explained by these predictor variables.
- The Hosmer-Lemeshow tests the null hypothesis that predictions made by the model fit perfectly with observed group memberships.
- The model correctly predicts the presence/absence of heart disease 89.9% of the time.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	122.494[a]	.558	.752

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12.516	8	.130

Classification Table[a]

Observed			Predicted: Heart Disease Absence	Predicted: Heart Disease Presence	Percentage Correct
Step 1	Heart Disease	Absence	125	8	94.0
		Presence	15	79	84.0
	Overall Percentage				89.9

a. The cut value is .500

- The Sig column shows the significance of the variable in the model.
- We can use the information in this table to predict the probability of an event occurring based on a one-unit change in an independent variable.
- For example, the table shows that the odds of having heart disease ("Present" category) is 0.966 greater as the age increases by one unit.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sex(1)	-1.877	.730	6.619	1	.010	.153
	Age	-.034	.031	1.241	1	.265	.966
	Cholesterol	.007	.005	2.418	1	.120	1.008
	BP	.043	.016	7.754	1	.005	1.044
	Chest pain type			16.374	3	.001	
	Chest pain type(1)	-3.294	.994	10.986	1	.001	.037
	Chest pain type(2)	-1.497	.799	3.512	1	.061	.224
	Chest pain type(3)	-2.127	.649	10.724	1	.001	.119
	FBS over 120(1)	.668	.745	.805	1	.370	1.951
	EKG results			5.858	2	.053	
	EKG results(1)	-1.262	.521	5.858	1	.016	.283
	EKG results(2)	-.769	4.528	.029	1	.865	.464
	Max HR	-.041	.015	7.077	1	.008	.960
	Exercise angina(1)	-.545	.527	1.068	1	.301	.580
	ST depression	.512	.309	2.746	1	.098	1.668
	Slope of ST			4.949	2	.084	
	Slope of ST(1)	1.874	1.543	1.476	1	.224	6.516
	Slope of ST(2)	2.626	1.434	3.354	1	.067	13.819

- **Note:** How well the independent variables perform in the model can be checked by testing their correlation with the dependent (predicted) variable.

Example:

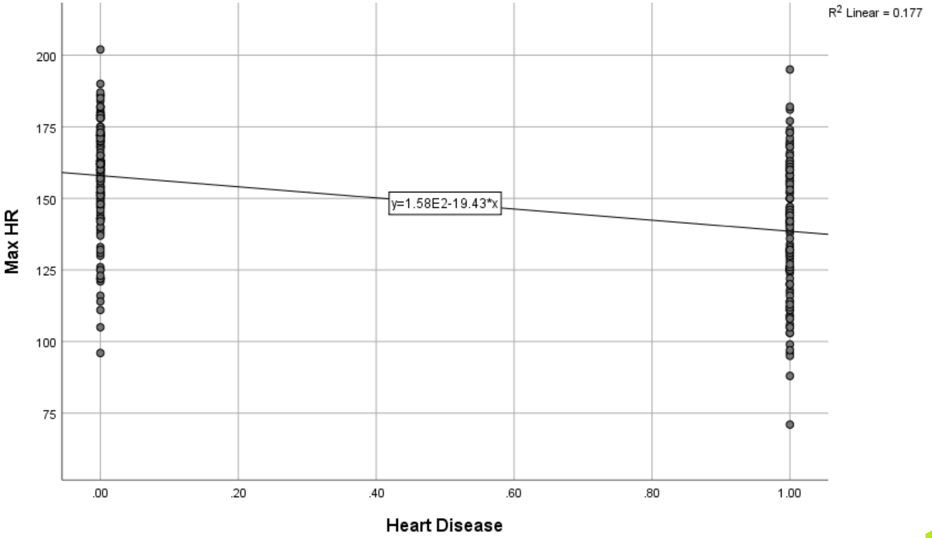
- Chi-square test shows that there is a correlation between 'Sex' and 'Heart Disease';
- Point-biserial correlation shows that there is a negative correlation between 'Max HR' and 'Heart Disease'.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	22.946[a]	1	.000		
Continuity Correction[b]	21.704	1	.000		
Likelihood Ratio	23.990	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	270				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 38.22.

b. Computed only for a 2x2 table



Correlations

		Heart Disease	Max HR
Heart Disease	Pearson Correlation	1	-.421**
	Sig. (2-tailed)		.000
	N	270	259
Max HR	Pearson Correlation	-.421**	1
	Sig. (2-tailed)	.000	
	N	259	259

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

Thank you