Integrating Statistical Excellence in Nursing Practice: Empowering Research with JASP

Norberto E. Milla, Jr.

Department of Statistics

2025-05-29

Outline of presentation

- Quick introduction to JASP
- Descriptive statistics
- Comparing two groups
- Comparing more than 2 groups
- Correlation and regression analyses

Quick intro to JASP

What is and why JASP?

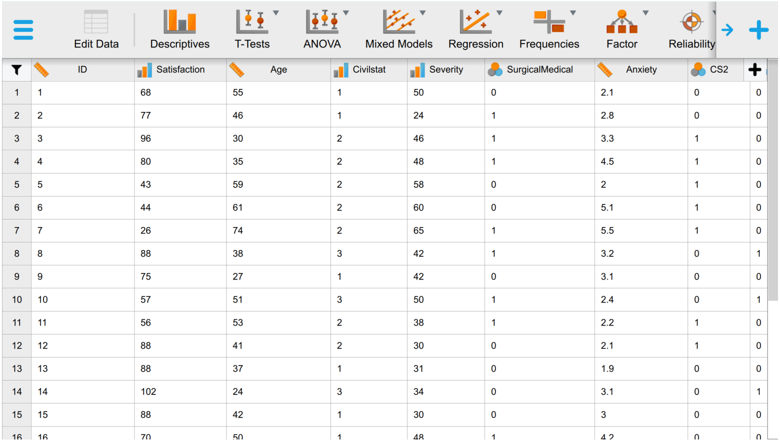
- **J**effrey's **A**mazing **S**tatistics **P**rogram
- **free**, multi-platform, and open-source statistics software
- **user-friendly**: simple drag and drop interface, easy access menus, intuitive analysis with real-time computation and display of all results
- Uses R and C++ in the background

How to get JASP?

https://jasp-stats.org/download

Quick intro to JASP

Figure 1: JASP GUI



	ID	Satisfaction	Age	Civilstat	Severity	SurgicalMedical	Anxiety	CS2	
1	1	68	55	1	50	0	2.1	0	0
2	2	77	46	1	24	1	2.8	0	0
3	3	96	30	2	46	1	3.3	1	0
4	4	80	35	2	48	1	4.5	1	0
5	5	43	59	2	58	0	2	1	0
6	6	44	61	2	60	0	5.1	1	0
7	7	26	74	2	65	1	5.5	1	0
8	8	88	38	3	42	1	3.2	0	1
9	9	75	27	1	42	0	3.1	0	0
10	10	57	51	3	50	1	2.4	0	1
11	11	56	53	2	38	1	2.2	1	0
12	12	88	41	2	30	0	2.1	1	0
13	13	88	37	1	31	0	1.9	0	0
14	14	102	24	3	34	0	3.1	0	1
15	15	88	42	1	30	0	3	0	0
16	16	70	50	1	48	1	4.2	0	0

Figure 2: Sample data set

Can easily read (import) data from other softwares

- Can read data from various sources:
  - text files (.txt)
  - SPSS data files (.sav)
  - SAS data files (.sas7bdat, .sas7bcat)
  - Stata data files (.dta)
  - MS Excel files (.csv)

- JASP data files have **.jasp** file extension

Quick intro to JASP

Provides easy mechanism for data management and manipulation

- Editing data
  - JASP already has a built-in data editor
  - JASP spreadsheet is *synced* with the source *CSV* file
  - Edit data in the csv file, save, and sync it in the JASP spreadsheet
- Assigning value labels
  - Sometimes data for categorical variables are number-coded during data entry (say, 1=Female, 2=Male)
- Recoding quantitative data into qualitative form
  - Age in years to age categories/groups
  - Reverse coding (1 to 5, 2 to 4, 4 to 2, 5 to 1)
- Creating new variables from existing ones
  - Computing the logarithm of income
- Filtering observations

Descriptive statistics using JASP

- Frequency distribution tables (frequency and percent) for qualitative data
- Data visualization (bar charts, histogram, correlation plots) both qualitative and quantitative data
- Summary statistics (mean, median, sd, std. error of the sample mean) for quantitative data

One-way and two-way (crosstab) frequency tables

Frequencies for Sex

Sex	Frequency	Percent	Valid Percent	Cumulative Percent
F	159	50.64	50.64	50.64
M	155	49.36	49.36	100.00
Missing	0	0.00		
Total	314	100.00		

Contingency Tables

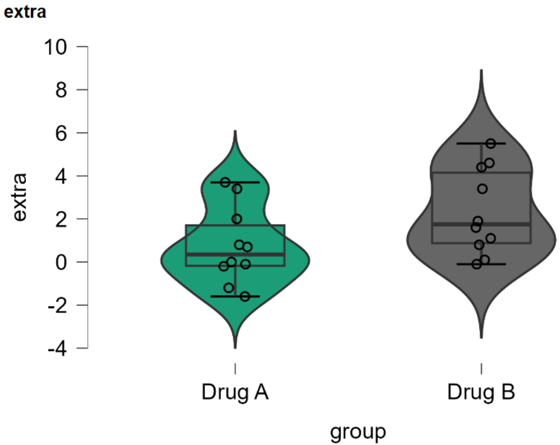
		Overall Satisfaction					
Sex		1	2	3	4	5	Total
F	Count	8.00	22.00	36.00	54.00	39.00	159.00
	% within row	5.03 %	13.84 %	22.64 %	33.96 %	24.53 %	100.00 %
M	Count	19.00	23.00	37.00	54.00	22.00	155.00
	% within row	12.26 %	14.84 %	23.87 %	34.84 %	14.19 %	100.00 %
Total	Count	27.00	45.00	73.00	108.00	61.00	314.00
	% within row	8.60 %	14.33 %	23.25 %	34.39 %	19.43 %	100.00 %

Summary statistics for numerical data

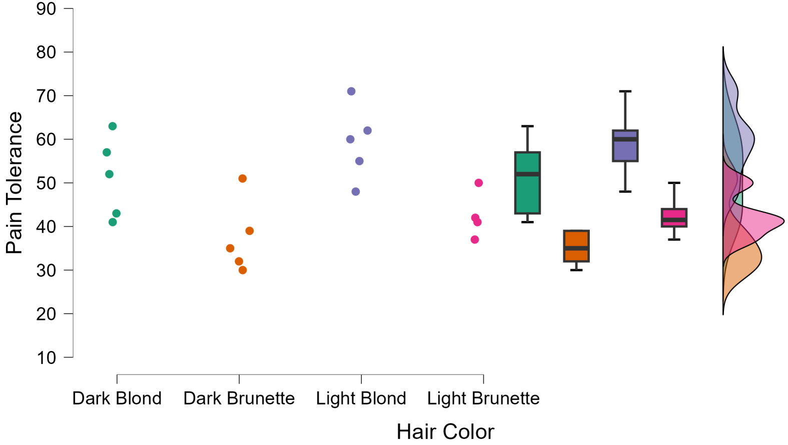
Descriptive Statistics ▼

	extra	
	Drug A	Drug B
Valid	10	10
Missing	0	0
Median	0.350	1.750
Mean	0.750	2.330
Std. Deviation	1.789	2.002
IQR	1.875	3.275
Minimum	−1.600	−0.100
Maximum	3.700	5.500

Graphical display



Graphical display



Reporting results of descriptive analysis

Demographic variables	No. of Students	Percent
Grade Level		
7	77	24.52
8	88	28.03
9	79	25.16
10	70	22.29
Sex		
Male	155	49.36
Female	159	50.64
Age *(mean=14.4, sd=1.6)*		
12 - 13	106	33.76
14 - 15	124	39.49
16 - 17	70	22.29
18 - 19	14	4.46
General Point Average *(mean=81.9, sd=4.9)*		
Outstanding (90-100)	34	10.83
Very Satisfactory (85-89)	48	15.29
Satisfactory (80-84)	131	41.72
Fairly Satisfactory (75-79)	99	31.53
Did Not Meet Expectations (Below 75)	2	0.64

Objective: Compare means (or medians) of two *independent* or *dependent* groups

- **Research question 1**: Is there a significant difference in the number of cigarettes smoked per day between males and females?
- **Research question 2**: Is there a difference in the pain tolerance of blonde-haired and brunette-haired women?
- **Research question 3**: Is there a significant reduction in the average blood sugar after 6 months of intermittent fasting?

Independent samples: units in one sample are selected independent of the units in the other sample

- For example, random sample of frontline medical personnel in Samar and another random sample of frontline medical personnel in Leyte
- A random sample of brunette-haired women and a separate random sample of blonde-haired women

Dependent (or related) samples: a unit in one sample is *paired* or *matched* with a unit in the other sample

- Use of pairs of twins: a member of a pair is given the experimental drug and the other a placebo
- A group of diabetic men participated in an experiment on the effect of intermittent fasting on blood sugar level: pre-experiment blood glucose vs post-experiment blood glucose
- One group of participants are given a pre-training evaluation and after the training they are given the (similar/parallel) post-training evaluation

T tests: compare the means of two groups/samples (IV); continuous DV

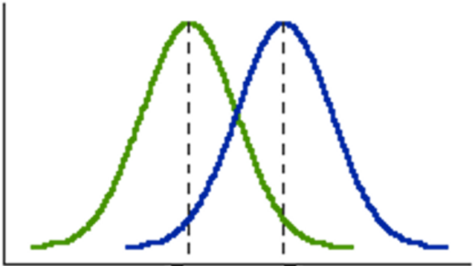
- Independent samples *Student's t* test: normal distribution and equal variance
- Independent samples *Welch's t* test: normal distribution, unequal variance
- *Paired/dependent samples t* test: normally distributed pairwise differences

Nonparametric alternatives to t tests

- **Wilcoxon rank sum/Mann Whitney U** test: alternative to Student's t test or Welch t test if normality assumption is not meet
- **Wilcoxon signed rank** test: alternative to paired samples t test if normality assumption about the pairwise difference is not meet

Checking assumptions of t tests

- **Normality assumption**: *histogram*, *box plot*, *Shapiro-Wilk test*
- **Equal variance assumption**: *Box plot*, *Levene's test*



Comparing two group means

RQ: What is the effect of smoking on infant's birth weight?

Test of Normality (Shapiro-Wilk) ▼

		W	p
BWT	Nonsmoker	0.987	0.344
	Smoker	0.983	0.410

Note. Significant results suggest a deviation from normality.

- $p > 0.05$ indicates that the distributions of birth weights for smokers and nonsmokers are NORMAL

Test of Equality of Variances (Levene's)

	F	df_1	df_2	p
BWT	1.508	1	187	0.221

- $p > 0.05$ indicates that the birth weight distributions for smokers and nonsmokers have EQUAL variances

Independent samples t tests

Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
BWT	Nonsmoker	115	3054.957	752.409	70.163	0.246
	Smoker	74	2773.243	660.075	76.732	0.238

Independent Samples T-Test

	t	df	p	Cohen's d	SE Cohen's d
BWT	2.634	187	0.005	0.392	0.151

Note. For all tests, the alternative hypothesis specifies that group *Nonsmoker* is greater than group *Smoker* .

Note. Student's t-test.

- **Interpretation**: A Student's t test showed that mothers who smoked during pregnancy gave birth to infants who have significantly lower birth weight (M=2773g) than mothers who did not smoke during pregnancy (M=3055g), t(187)=2.63, p=0.005<0.05, Cohen's d=0.39.

Paired-samples t tests

RQ: Is there a significant reduction in body mass after a 4-week diet?

Descriptives ▼

	N	Mean	SD	SE	Coefficient of variation
Pre diet body mass	78	72.53	8.723	0.988	0.120
Post 4 weeks diet	78	68.74	9.009	1.020	0.131

Test of Normality (Shapiro-Wilk) ▼

			W	p
Pre diet body mass	-	Post 4 weeks diet	0.975	0.124

Note. Significant results suggest a deviation from normality.

- $p = 0.124 > 0.05$ means that the pairwise differences have normal distribution

Paired-samples t tests

Paired Samples T-Test

Measure 1		Measure 2	t	df	p	Mean Difference	SE Difference	Cohen's d	SE Cohen's d
Pre diet body mass	-	Post 4 weeks diet	13.04	77	< .001	3.782	0.290	1.476	0.047

Note. Student's t-test.

- **Interpretation**: On average, participants lost 3.78kg (SE=0.29kg) body mass following a 4-week diet plan. A paired samples t-test showed that this decrease is significant ($t(77) = 13.04$, $p<.001$, Cohen's $d = 1.48$).

Comparing three (or more) group means

- **Objective**: Compare means (or medians) of three or more independent or dependent populations
- **Research question**: Is there a significant difference in the average waiting times (in hours) in the emergency room of three hospitals?
- **Statistical method**: Analysis of Variance (ANOVA)
  - method of partitioning the total variance in the DV into different components which can be attributed to different sources: effect of manipulated factors (**Systematic variation**), and experimental error (**Random variation**)
  - **Assumptions**: Normal distribution, Equal variance, Independence
  - **Kruskal-Wallis** test: if assumption of normality is not meet, or if DV is in ordinal scale

One-way ANOVA

Descriptive Statistics ▼

	Waiting time		
	A	B	C
Valid	20	20	20
Missing	0	0	0
Mean	1.150	1.925	1.620
Std. Deviation	0.495	0.599	0.542
Shapiro-Wilk	0.952	0.920	0.952
P-value of Shapiro-Wilk	0.405	0.098	0.394
Minimum	0.400	0.700	0.300
Maximum	2.200	2.700	2.400

Test for Equality of Variances (Levene's)

F	df1	df2	p
0.532	2.000	57.000	0.590

One-way ANOVA

ANOVA - Waiting time

Cases	Sum of Squares	df	Mean Square	F	p	η^2
Hospital	6.097	2	3.049	10.198	< .001	0.264
Residuals	17.040	57	0.299			

Note. Type III Sum of Squares

- $p<0.001$ indicates significant differences in the mean waiting time for the emergency room of three hospitals
- If the ANOVA reports no significant difference you can go no further in the analysis
- If the ANOVA is significant, post hoc testing can now be carried out

Post hoc test

- Researchers might want to answer two additional questions:
  - Which means are different?
  - What is the magnitude of the difference between means?
- ANOVA by itself answers neither of these questions
- There are two approaches to figuring out which means are different and by how much: *planned* and *unplanned* comparisons of means
- These procedures are generally referred to as *post-hoc* tests

Post hoc test

Post Hoc Comparisons - Hospital

		Mean Difference	SE	t	p_{tukey}
A	B	−0.775	0.173	−4.482	< .001
	C	−0.470	0.173	−2.718	0.023
B	C	0.305	0.173	1.764	0.191

Note. P-value adjusted for comparing a family of 3

Comparing three (or more) group means

- **Interpretation**: One-way analysis of variance showed that there is a significant difference in the mean waiting times among the three hospitals, F(2,57) = 10.2, p < 0.001, $\eta^2 = 0.26$. Post hoc analysis using Tukey's method indicated that the average waiting time of Hospital A (M = 1.15, SD = 0.49) is significantly lower than Hospital B (M = 1.93, SD = 0.60) and Hospital C (M = 1.62, SD = 0.54). Further, there is no significant difference in the mean waiting time between Hospital B and Hospital C.
- **Interpretation of** η^2: 0.01 (Small effect size), 0.06 (Medium effect size), 0.14 or higher (Large effect size)

Basic ideas

- Correlation analysis is concerned with the analysis of linear relationship between two or more variables
- It is used to determine the strength and direction, as well as statistical significance, of the correlation between variables
- The correlation between two variables could be positive or negative
- Positive correlation: $X \uparrow$ and $Y \uparrow$ or $X \downarrow$ and $Y \downarrow$
- Negative correlation: $X \uparrow$ and $Y \downarrow$ or $X \downarrow$ and $Y \uparrow$

Scatter plot

- It is a chart of the x-values (X-axis) and y-values (Y-axis)
- It is a visual representation of the relationship of X and Y
- Also known as *scatter diagram*

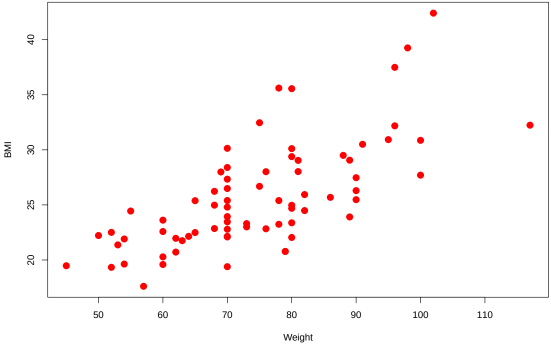
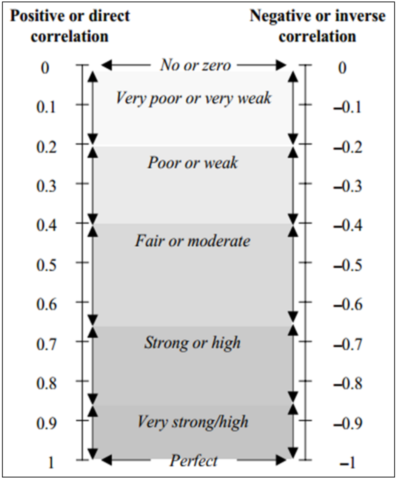


Figure 3: Scatter Plot of Weight and BMI

Correlation Coefficient

- The strength or magnitude of the correlation between variables is measured by a *correlation coefficient*
  - *Pearson r*: both variables are measured in at least interval scale; bivariate normal distribution
  - *Spearman rho*: both variables are measured in at ordinal scale
  - *Point-biserial*: one variable is binary, the other is interval or ratio
  - *Rank-biserial*: one variable is binary and the other is ordinal
- The value of a correlation coefficient ranges from -1 to +1
- A zero correlation coefficient indicates that the variables are NOT LINEARLY independent

Guide in interpreting correlation coefficients



Test of significance of a correlation coefficient

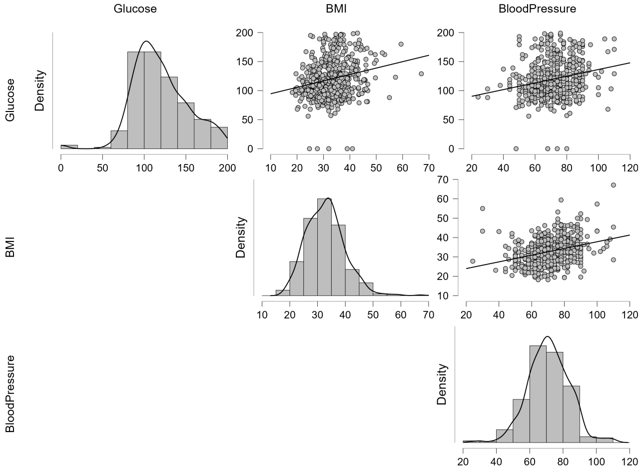
- H_0: Correlation coefficient is equal to zero. (There is no linear relationship between the variables.) $\implies \rho = 0$
- H_1: Correlation coefficient is not equal to zero. (There is linear relationship between the variables.) $\implies \rho \neq 0$
- Test statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- Reject H_0 if p-value associated with t is less than the significance level (α)

Correlation analysis

Research question: Is there a significant relationship between blood glucose, blood pressure, and BMI?



Correlation analysis

Research question: Is there a significant relationship between blood glucose, blood pressure, and BMI?

Shapiro-Wilk Test for Bivariate Normality ▼

			Shapiro-Wilk	p
Glucose	-	BMI	0.976	< .001
Glucose	-	BloodPressure	0.976	< .001
BMI	-	BloodPressure	0.981	< .001

Correlation Table

			Pearson		Spearman	
			r	p	rho	p
Glucose	-	BMI	0.230	< .001	0.234	< .001
Glucose	-	BloodPressure	0.215	< .001	0.237	< .001
BMI	-	BloodPressure	0.310	< .001	0.317	< .001

- Glucose, blood pressure, and BMI have weak but significant correlations.

Regression analysis

- *Regression analysis* is a technique of studying the dependence of one variable (called dependent variable), on one or more independent variables (called explanatory variables)
- Regression analysis is useful for:
  - Estimating the relationship between the dependent variable and the explanatory variable(s)
  - Determining the effect of each of the explanatory variables on the dependent variable, controlling the effects of all other explanatory variables
  - Predicting the value of the dependent variable for a given value of the explanatory variable

Linear regression model

- In regression analysis we wish to express the relationship between the dependent variable and the explanatory variable in a functional form

$$Y = f(X) + \epsilon$$

- Suppose we observe pairs (X,Y) and the scatterplot shows a fairly linear pattern, then we can let

$$f(X) = \beta_0 + \beta_1 X$$

- Thus, the (simple) linear regression model is given by

$$Y = \beta_0 + \beta_1 X + \epsilon$$

  where:
  - Y is the response variable
  - X is the regressor (predictor)
  - β_0 - y-intercept
  - β_1 - slope of the regression line; regression coefficient
  - ϵ is the random error term

Multiple linear regression model

- Oftentimes, a regression model with a single predictor variable is not enough to provide an adequate description of the response variable
- This occurs when several key variables affect the response variable in important and predictive ways
- The multiple linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

where:

- $\beta_0, \beta_1, \cdots, \beta_p$ are the regression coefficients
- ϵ is the random error, $\epsilon \sim N(0, \sigma^2)$

Assumptions

- Standard assumptions:
  - Y is a continuous random variable
  - $\epsilon_i \sim N(0, \sigma^2), \forall\ i \Longrightarrow$ *normality* and *homogenous variance* assumptions
  - For two different trials, i and j, the error terms ϵ_i and ϵ_j are independent $\Longrightarrow$ independence assumption
- Assumptions on the predictor variables:
  - The predictor variables are assumed fixed or selected in advance.
  - The values of the predictors are assumed to be measured without error.
  - The predictor variables are assumed to be linearly independent of each other (*No multicollinearity*) $\Longrightarrow$ multicollinearity occurs if at least 2 of the regressors are strongly and significantly correlated
- Assumption about the observations:
  - Absence of outliers and influential observations

Interpretation of regression parameters

- β_0 is the regression constant or intercept and is the value of Y when all X's are equal to zero; must be interpreted with caution
- β_1 is the estimated mean response for every 1 unit change in X_1, holding all other X's constant
- β_2 is the estimated mean response for every 1 unit change in X_2, holding all other X's constant

$\vdots$

and so on

Estimation of regression parameters

- The regression parameters β_0 and β_1 are unknown quantities that must be estimated from the data
- The most common and popular method of estimating regression parameters is the *Method of Least Squares*
- The resulting estimators are called Ordinary Least Squares (OLS) estimators
- The OLS estimators of β_0, β_1, β_2, etc. are $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, etc.

Overall measures of fit

- Coefficient of determination (R^2):
  - the percentage of variation in Y that can be explained or attributed to its linear relation with X
  - the closer to 1 or 100% the better the fit
  - Use Adjusted R^2 in multiple linear regresssion
- F test of overall effect of all X's: $F = \frac{MSR}{MSE}$
  - A significant F value indicates that Y is significantly with all X's
- Root Mean Square Error (RMSE):
  - standard deviation of the residuals (=difference between the observed Y and the predicted Y based on the equation)
  - the closer to zero the better the fit

Test of the overall significance of the regression

- $H_0 : \beta_i = 0, \forall i$
- $H_1 : \beta_i \neq 0,$ for at least one i
- Test statistic: $F = \frac{MSR}{MSE}$

Test of the significance of each β_i

- $H_0 : \beta_i = 0$
- $H_1 : \beta_i \neq 0$
- Test statistic: $t = \frac{\beta_i}{se(\beta_i)}$

Linear regression analysis

Research question: What are the factors that affect blood glucose level in women?

Model Summary - Glucose

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	32.975
H_1	0.546	0.298	0.291	27.760

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	173634.150	5	34726.830	45.064	< .001
	Residual	409190.814	531	770.604		
	Total	582824.965	536			

Note. The intercept model is omitted, as no meaningful information can be shown.

- 29.1% of the variance in blood glucose level can be explained by BMI, blood pressure, age, insulin, and diabetespedigreefunction
  - Residual (unexplained) variance = 100 -29.1 = 70.9% ⇒ Very High!
- RMSE = 27.76 is quite acceptable but not ideal

Linear regression analysis

Research question: What are the factors that affect blood glucose level in women?

Coefficients ▼

Model		Unstandardized	Standard Error	Standardized	t	p	Collinearity Statistics	
							Tolerance	VIF
H_0	(Intercept)	119.903	1.423		84.262	< .001		
H_1	(Intercept)	49.141	8.172		6.014	< .001		
	BloodPressure	0.321	0.110	0.120	2.925	0.004	0.791	1.264
	BMI	0.426	0.189	0.089	2.256	0.025	0.851	1.175
	DiabetesPedigreeFunction	7.168	3.572	0.075	2.007	0.045	0.954	1.048
	Insulin	0.111	0.010	0.415	11.055	< .001	0.940	1.064
	Age	0.556	0.120	0.181	4.635	< .001	0.866	1.155

- Age and diabetespidegreefunction have the largest effects on blood glucose level
  - Blood glucose level increases by 7.168 points for every point change in diabetespedigreefunction, assuming all else are equal
  - Blood glucose level increases by 0.556 point for every year change in age, assuming all else are equal

Other important activities

- Addition of categorical explanatory variables in the model, say sex and marital status
- Residual analysis: Checking if all assumptions are met
- Checking for outliers and influential points
- Remediation,if issues arise

Logistic regression model (Logit model)

- A form of regression model where the dependent variable is binary, such whether one is diabetic or not, whether an infant has low birth weight or not, whether a lump is cancerous or not
- **Research question**: What are the risk factors of heart attack?

Model Summary - Heart Attack

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox & Snell R^2
H_0	55.452	57.452	59.141	39						
H_1	34.195	40.195	45.261	37	21.257	< .001	0.383	0.550	0.446	0.412

- This result suggests a significant relationship ($\chi^2(37) = 21.257, p < .001$) between the outcome (heart attack) and the predictor variables (exercise prescription and stress levels)
- McFadden's $R^2 = 0.383$. It is suggested that a range from 0.2 to 0.4 indicates a good model fit

Logistic regression analysis

Coefficients ▼

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	−4.368	2.550	0.013	−1.713	2.933	1	0.087
Stress level	0.089	0.041	1.093	2.159	4.662	1	0.031
Exercise (1)	−2.043	0.890	0.130	−2.295	5.268	1	0.022

Note. Heart Attack level '1' coded as class 1.

- Both stress level and exercise prescription are significant predictor variables ($p = 0.031$ and $p = 0.022$, respectively)
- For stress level $OR = 1.093, p < 0.05$ suggests that high-stress levels are significantly related to an increased probability of having a heart attack
- Having an exercise intervention is related to a significantly reduced probability of a heart attack.
  - $OR = 0.130$ can be interpreted as only having a 13% probability of heart attack if undergoing an exercise intervention