

Regression Analysis

Part II

Prepared by Dr. Hesham Mousli

MHA 613
Statistics in Healthcare

Lecture 10



Intended Objectives

1. Understand the concepts of simple and multiple linear regression in healthcare settings.
2. Learn the steps involved in performing regression analysis, including preparing data and interpreting results.
3. Explore the applications of regression models, including assessing significant associations and making predictions.
4. Analyze case scenarios to apply regression models to real-world healthcare problems.



Chapter

10

1. Simple Linear Regression

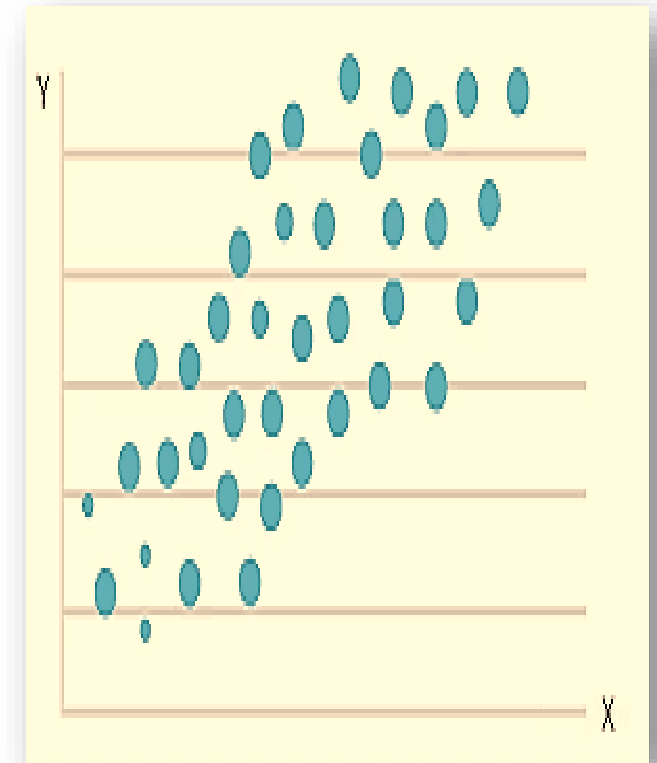
**Understanding Concepts and Interpreting
Results**

Simple Linear Regression

- Simple linear regression examines the relationship between two numerical variables.
- **Key assumptions:**
 - Data points should roughly align along a straight line.
 - Random scatter around the line with no extreme outliers.

Steps in Performing Regression Analysis

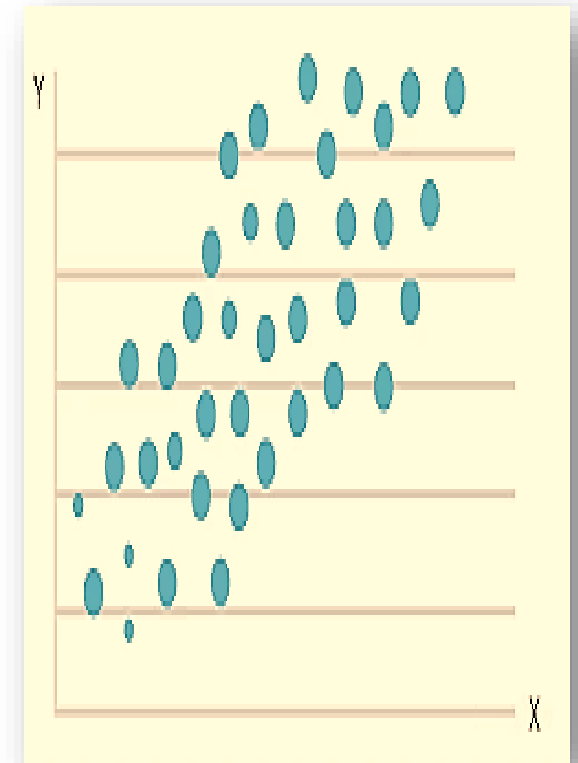
- 1. Prepare data:** Two columns for X (independent) and Y (dependent).
- 2. Set variables:** Assign roles for X and Y.



Steps in Performing Regression Analysis

3. Use software to:

- Generate graphs and residual plots.
- Provide regression table and goodness-of-fit measures.



Interpreting the Results

A. Scatter plot: Visual trend between X and Y.

- **p-value for intercept & slope:**

- < 0.05 : Significant association between X and Y.
- ≥ 0.05 : No conclusive evidence of association.

B. Goodness-of-fit: Evaluates how well the model fits the data.

Predictive Model

- **Regression equation:** $Y = a + bX$.
- **Use to predict Y based on X values.**
- **Residual standard error:** Measures prediction accuracy.

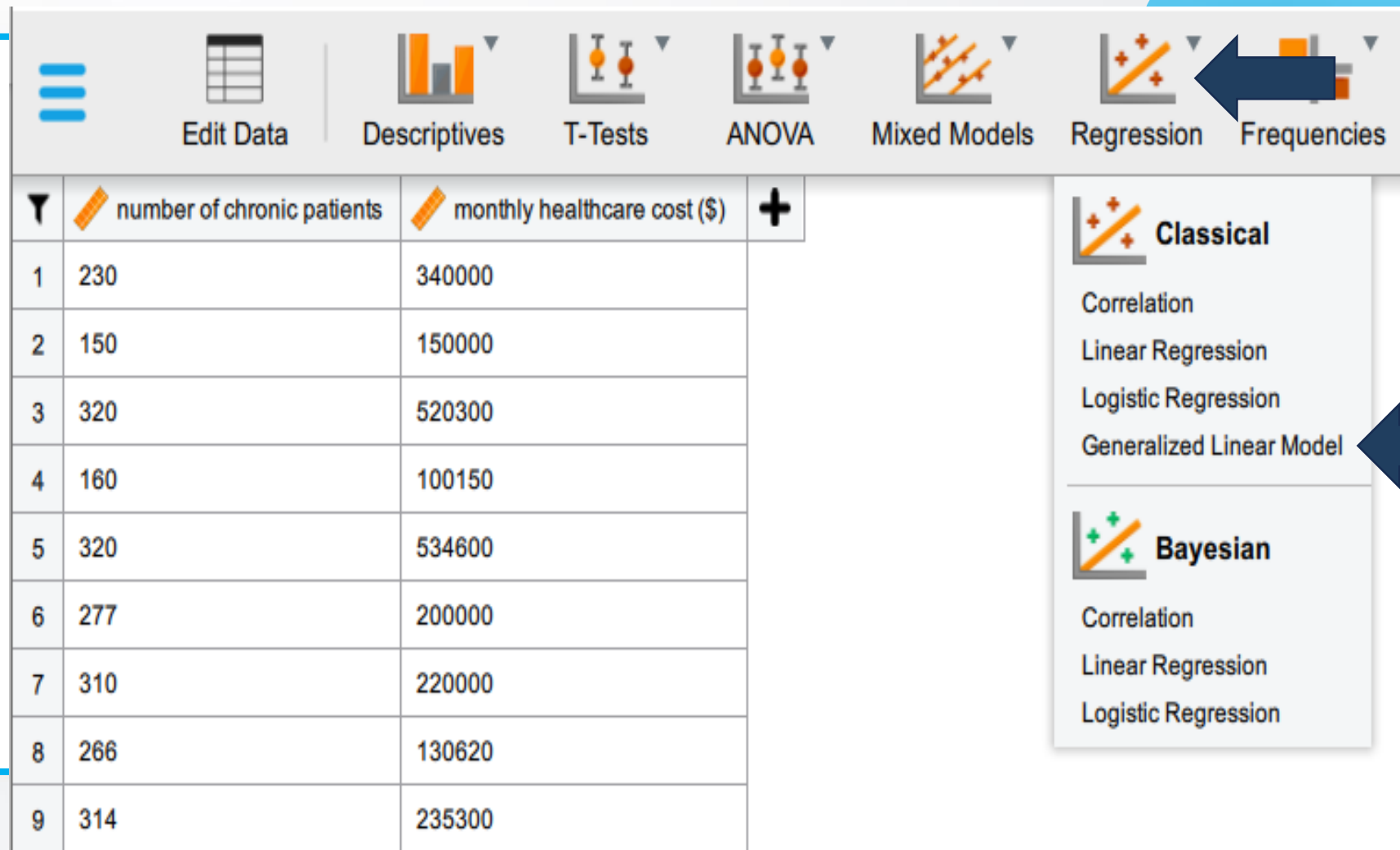
Case Example: Weight and Blood Pressure

- **Hypothesis:** Higher weight is associated with higher BP.
- **Observations:**
 - Low weight → Low BP (lower-left of scatter plot).
 - High weight → High BP (upper-right of scatter plot).
- **Example regression equation:** $BP = 76.9 + 0.487 \times \text{Weight}$.

Case Scenario (1)

- An insurance company wants to estimate a patient's monthly healthcare costs based on the number of chronic conditions they have. Using patient data, they build a regression model where **monthly healthcare cost (in dollars)** is the dependent variable and the **number of chronic patients** is the independent variable.

T	number of chronic patients	monthly healthcare cost (\$)
1	230	340000
2	150	150000
3	320	520300
4	160	100150
5	320	534600
6	277	200000
7	310	220000
8	266	130620
9	314	235300



The screenshot displays a statistical software interface. At the top, there is a menu bar with icons for various statistical tests: Edit Data, Descriptives, T-Tests, ANOVA, Mixed Models, Regression, and Frequencies. A blue arrow points to the 'Regression' icon. Below the menu bar is a data table with two columns: 'number of chronic patients' and 'monthly healthcare cost (\$)'. The table contains 9 rows of data. To the right of the table, a dropdown menu is open, showing options for 'Classical' and 'Bayesian' regression models. A blue arrow points to the 'Generalized Linear Model' option under the 'Classical' section.

	number of chronic patients	monthly healthcare cost (\$)
1	230	340000
2	150	150000
3	320	520300
4	160	100150
5	320	534600
6	277	200000
7	310	220000
8	266	130620
9	314	235300

Classical

- Correlation
- Linear Regression
- Logistic Regression
- Generalized Linear Model

Bayesian

- Correlation
- Linear Regression
- Logistic Regression

▼ Generalized Linear Model



Dependent variable

 monthly healthcare cost (\$)



Covariates

 number of chronic patients



Generalized Linear Model

Model Summary - monthly healthcare cost (\$)

Model	Deviance	AIC	BIC	df	X ²	p
H ₀	2.088×10 ⁺¹¹	244.350	244.744	8		
H ₁	1.319×10 ⁺¹¹	242.215	242.806	7	7.693×10 ⁺¹⁰	0.083

Coefficients



	Estimate	Standard Error	t	p
(Intercept)	-111442.388	194310.040	-0.574	0.584
number of chronic patients	1463.124	724.161	2.020	0.083



Case Scenario (1)

➤ **The regression model's p-values indicate the following:**

- **Model ($p = 0.83$):** No significant overall relationship between the number of chronic patients and monthly healthcare costs.
- **Intercept ($p = 0.584$):** Baseline costs when chronic conditions are zero are not statistically reliable.

Case Scenario (1)

- **Number of Chronic Conditions ($p = 0.083$):** It does not provide strong evidence of a meaningful relationship with monthly healthcare costs.
- **Implication:**
 - **The model is not reliable for estimating monthly healthcare costs based solely on the number of chronic conditions.**



Chapter

10

2. Multiple Linear Regression

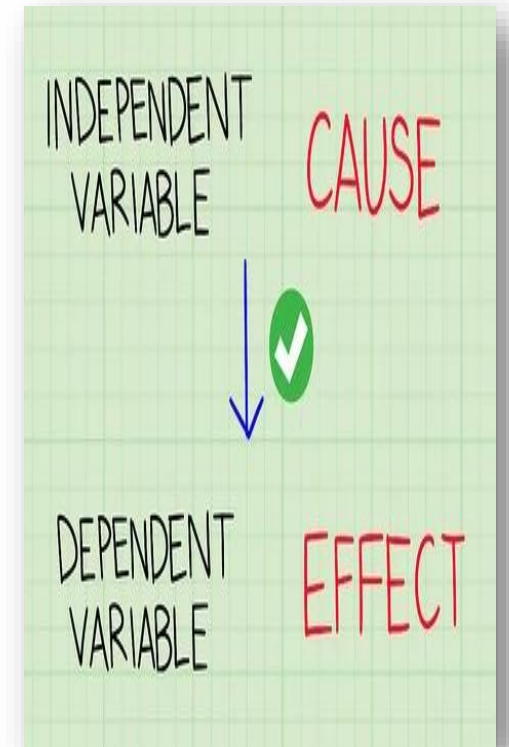
**Understanding Concepts and Interpreting
Results**

What is Multiple Linear Regression?

- A statistical method to **model** the **relationship between** a **dependent variable** and **multiple predictors**.
- Used to **predict an outcome** variable based on **multiple input variables**.
- Extends simple regression by **incorporating multiple independent variables**.

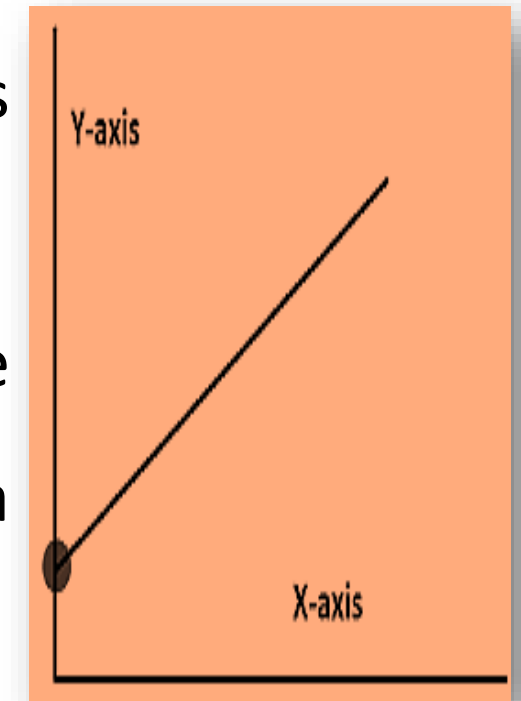
Key Terms

- **Dependent Variable:** The outcome variable being predicted.
- **Independent Variables:** Predictors that influence the dependent variable.



Key Terms

- **Intercept:** The predicted value when all predictors are zero.
- **Coefficients:** Represent the change in the dependent variable for a one-unit change in a predictor, holding others constant.



Simple
Linear
Regression

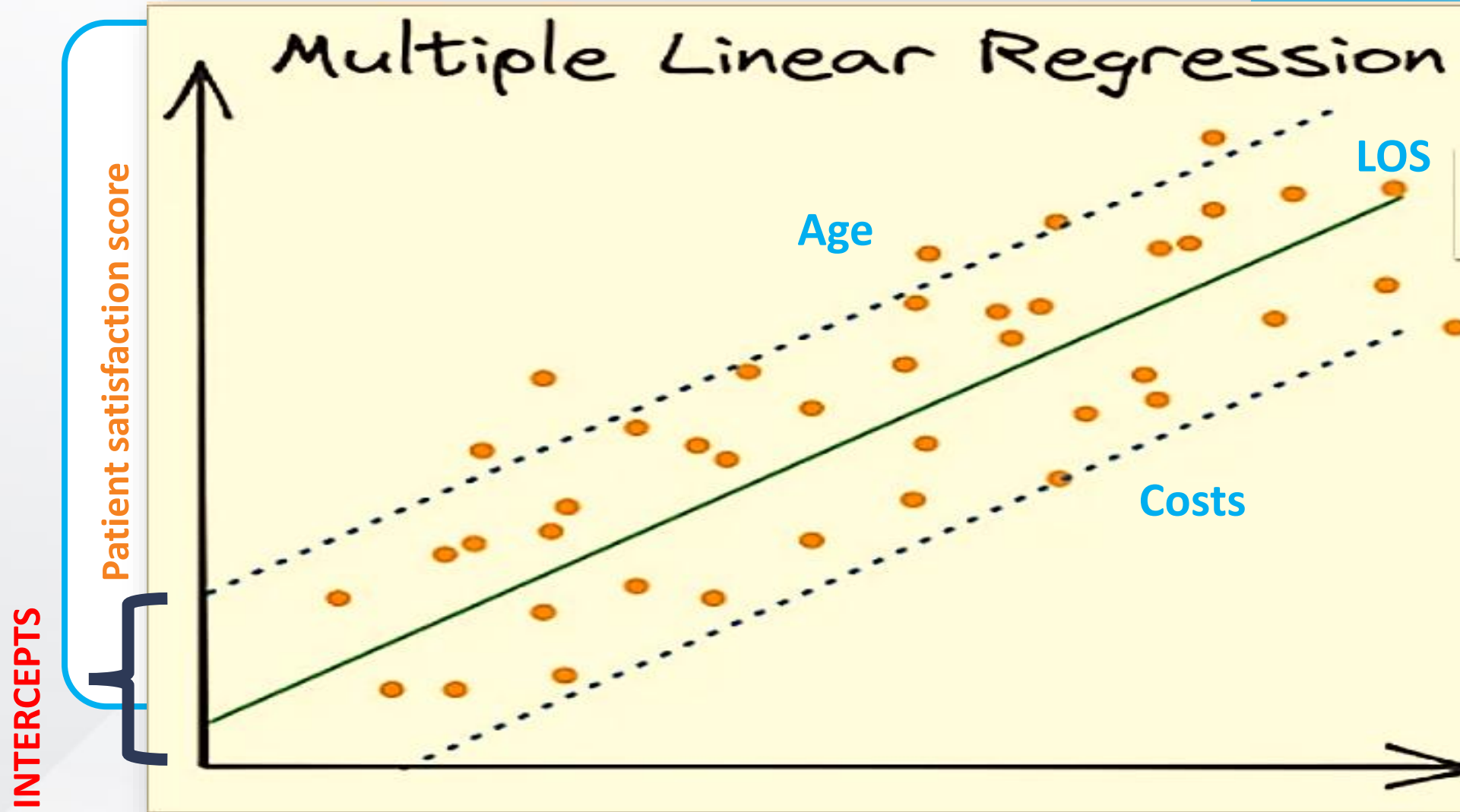
$$y = b_0 + b_1x_1$$

Multiple
Linear
Regression

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

Formulas and Interpretation

- **General Model:** $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_nX_n$
- **Coefficients (β):** Reflect the impact of each predictor.
- **Constant Term (β_0):** Baseline prediction (Intercept).
- **Example:** $Y = a + bX_1 + cX_2$ predicts Y using X1 and X2.



Applications of Multiple Linear Regression

- Assessing significant associations.
- Making predictions and forecasting.
- Deriving values of parameters with practical meanings.

Interpreting the Results

- The **P-value** in the context of multiple linear regression results evaluates the statistical significance of the coefficients and the overall model.



Interpreting the Results

I. P-values

Chapter

10

1. R-squared:

- **Role:** The p-value is indirectly related to R^2 through the F-test, which tests the overall model fit.
- **Interpretation:** A small p-value for the F-test indicates that the model as a whole is significant, meaning at least one predictor explains a significant portion of the variance in the dependent variable.

Interpreting the Results

Chapter

10

I. P-values

2. Intercept:

- **Role:** Tests whether the intercept is significantly different from zero.
- **Interpretation:** A small p-value (<0.05) suggests that the intercept has a meaningful contribution, though its practical importance depends on the context of the predictors.

Interpreting the Results

II. R^2 & Coefficients

1. R-squared (Coefficient of Determination):

- **Definition:** Represents the proportion of variance in the dependent variable explained by the predictors.
- **Interpretation:** Higher values (closer to 1) indicate a better fit of the model to the data.

Interpreting the Results

II. R^2 & Coefficients

2. Intercept (Constant Term):

- **Definition:** The predicted value of the dependent variable when all predictors are zero.
- **Interpretation:** Serves as the baseline prediction; its relevance depends on whether a zero value for predictors is meaningful.

Interpreting the Results

II. R^2 & Coefficients

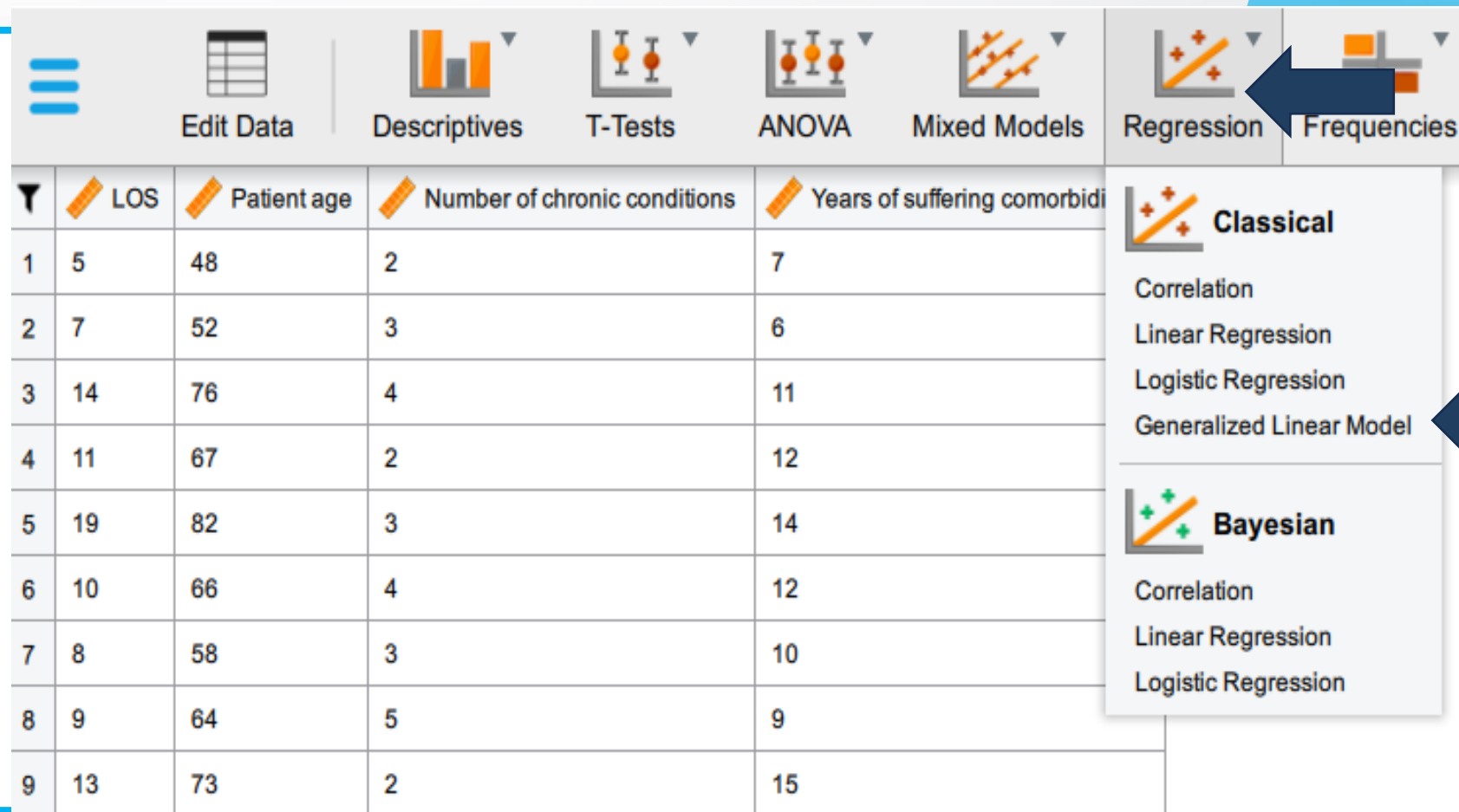
3. Coefficients of Multiple Predictors:

- **Definition:** Measure the change in the dependent variable for a one-unit change in the predictor, holding other predictors constant.
- **Interpretation:** Positive coefficients indicate an increase, and negative coefficients indicate a decrease in the outcome variable with an increase in the predictor.

Case Scenario (2)

- A hospital administrator wants to predict the length of stay (LOS) of patients. Using historical data, the regression model includes the following predictors:
- **Dependent Variable (Y):** Patient's length of stay (LOS)
 - **Independent Variables (X):** Patient age, number of chronic conditions, and years of suffering comorbidities.
 - **Question:** How can the results guide the number of assigned beds for each department?

T	LOS	Patient age	Number of chronic conditions	Years of suffering comorbidities
1	5	48	2	7
2	7	52	3	6
3	14	76	4	11
4	11	67	2	12
5	19	82	3	14
6	10	66	4	12
7	8	58	3	10
8	9	64	5	9
9	13	73	2	15



The screenshot displays a statistical software interface. At the top, there is a menu bar with icons for various statistical functions: Edit Data, Descriptives, T-Tests, ANOVA, Mixed Models, Regression, and Frequencies. The 'Regression' menu is currently open, showing two main categories: 'Classical' and 'Bayesian'. The 'Classical' category includes Correlation, Linear Regression, Logistic Regression, and Generalized Linear Model. The 'Bayesian' category includes Correlation, Linear Regression, and Logistic Regression. A blue arrow points to the 'Regression' menu, and another blue arrow points to the 'Generalized Linear Model' option.

	LOS	Patient age	Number of chronic conditions	Years of suffering comorbidities
1	5	48	2	7
2	7	52	3	6
3	14	76	4	11
4	11	67	2	12
5	19	82	3	14
6	10	66	4	12
7	8	58	3	10
8	9	64	5	9
9	13	73	2	15

Generalized Linear Model



Dependent variable

LOS

Covariates

Patient age
Number of chronic conditions
Years of suffering comorbidities

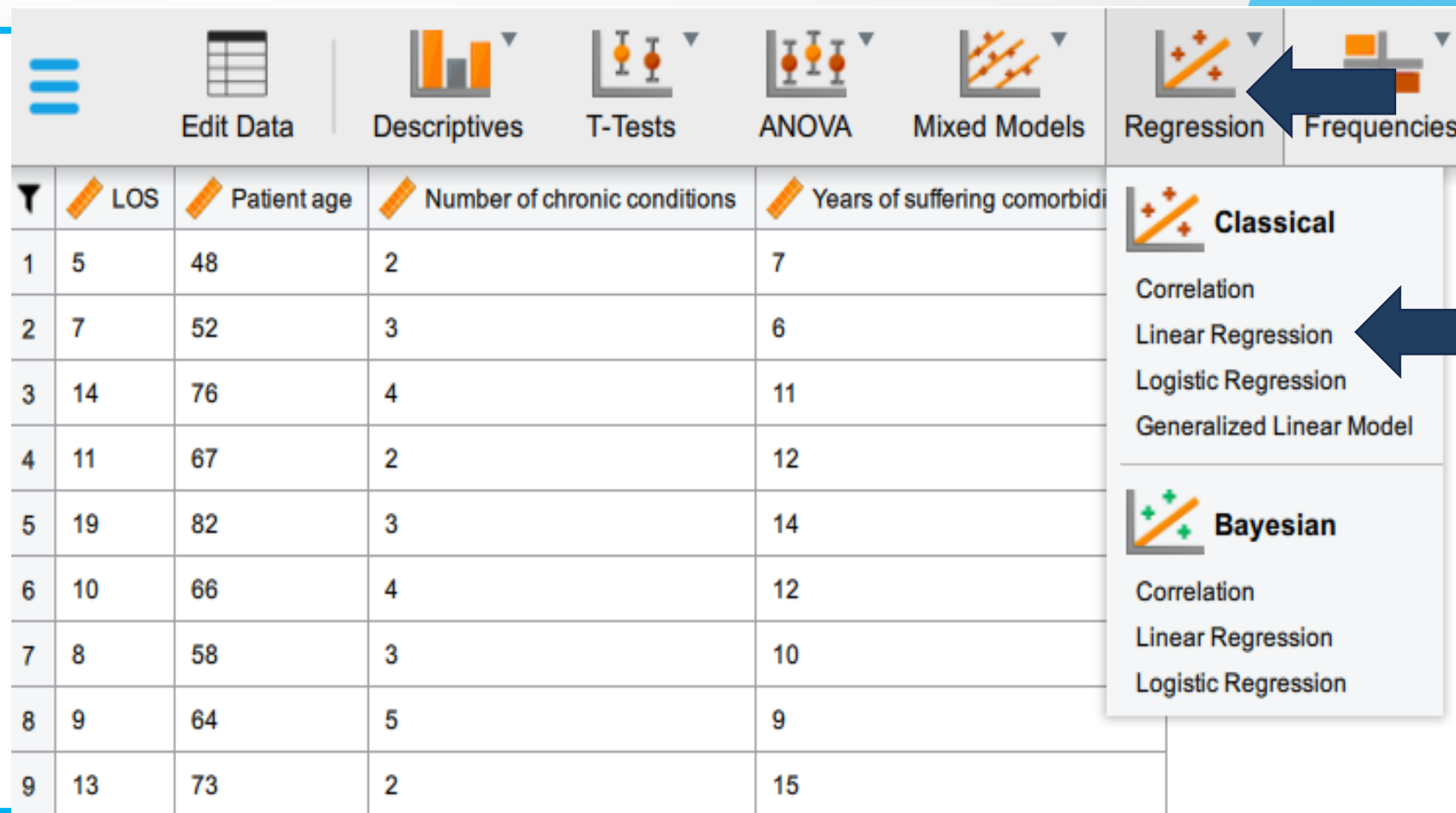
Generalized Linear Model

Model Summary - LOS

Model	Deviance	AIC	BIC	df	X ²	p
H ₀	142.000	54.368	54.763	8		
H ₁	5.108	30.443	31.429	5	136.892	< .001

Coefficients

	Estimate	Standard Error	t	p
(Intercept)	-13.544	2.355	-5.752	0.002
Patient age	0.492		6.333	0.001
Number of chronic conditions	-0.870	0.427	-2.038	0.097
Years of suffering comorbidities	-0.479	0.286	-1.677	0.154



The screenshot displays a statistical software interface. At the top, there is a menu bar with icons for various statistical functions: Edit Data, Descriptives, T-Tests, ANOVA, Mixed Models, Regression, and Frequencies. A blue arrow points to the 'Regression' menu. Below the menu bar is a data table with 5 columns and 9 rows. The columns are labeled 'LOS', 'Patient age', 'Number of chronic conditions', and 'Years of suffering comorbidities'. The first column contains row numbers 1 through 9. The 'Regression' menu is open, showing two sections: 'Classical' and 'Bayesian'. The 'Classical' section includes 'Correlation', 'Linear Regression', 'Logistic Regression', and 'Generalized Linear Model'. A blue arrow points to 'Linear Regression'. The 'Bayesian' section includes 'Correlation', 'Linear Regression', and 'Logistic Regression'.

	LOS	Patient age	Number of chronic conditions	Years of suffering comorbidities
1	5	48	2	7
2	7	52	3	6
3	14	76	4	11
4	11	67	2	12
5	19	82	3	14
6	10	66	4	12
7	8	58	3	10
8	9	64	5	9
9	13	73	2	15

Regression

Classical

- Correlation
- Linear Regression
- Logistic Regression
- Generalized Linear Model

Bayesian

- Correlation
- Linear Regression
- Logistic Regression

▶ Linear Regression

▼ Linear Regression

Dependent Variable

LOS

Method

Enter

Covariates

- Patient age
- Number of chronic conditions
- Years of suffering comorbidities



The image shows a software interface for a Linear Regression analysis. It features a sidebar on the left with a 'Linear Regression' section. The main panel contains three fields: 'Dependent Variable' (set to 'LOS'), 'Method' (set to 'Enter'), and 'Covariates' (a list containing 'Patient age', 'Number of chronic conditions', and 'Years of suffering comorbidities'). A blue arrow points to the 'Dependent Variable' field, and an orange bracket with a blue arrow points to the 'Covariates' list. The interface includes standard window controls (minimize, maximize, close) and a list of icons (R, edit, copy, info, delete) at the top right of the main panel.

Linear Regression

Model Summary - LOS

Model	R	R ²	Adjusted R ²	RMSE
M ₀	0.000	0.000	0.000	4.213
M ₁	0.982	0.964	0.942	1.011

Note. M₁ includes Patient age, Number of chronic conditions, Years of suffering comorbidities

Case Scenario (2)

- The regression model is statistically significant ($p < 0.001$) and explains a very high proportion of the variability in length of stay (LOS), ($R^2 = 0.964$).
- **Implication:**
 - Allocating resources efficiently, focusing on patients predicted to have the highest length of stay.

Wrap up

1. **Simple Linear Regression:** Models the relationship between two numerical variables.
2. **Multiple Linear Regression:** Extends simple regression by including multiple predictors.
3. **Interpreting Multiple Linear Regression Results:**
 - P-values indicate the significance of predictors and the model as a whole.
 - R^2 reflects the model's goodness of fit.
 - Coefficients show the effect of each predictor on the outcome.



Thank You