SESSION-12: LINEAR REGRESSION ANALYSIS

Course detail: http://julhas.com/jsedutech/stata-level-one.html
Mentor: Julhas Sujan

Recap-Session 11

- Statistical tests – One sample proportion, Chi-squared, t-Test, ANNOVA
- Hypothesis testing

Session-12

Agenda:

- What is Statistical model?
- What is regression? Use of regression.
- Linear regression
- Multiple linear regression

Before starting:
Watch the video first: https://www.youtube.com/watch?v=HafqFSB9x70

What is scatterplot and correlation? https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/introduction-to-scatterplots/a/scatterplots-and-correlation-review

Statistical model: A statistical model is usually specified as a mathematical relationship between one or more random variables and other non-random variables.

Regression: Regression analysis is a statistical method that helps us to analyse and understand the relationship between two or more variables of interest. The process that is adapted to perform

regression analysis helps to understand which factors are important, which factors can be ignored and how they are influencing each other.

For the regression analysis is a successful method, we understand the following terms:

- **Dependent Variable:** This is the variable that we are trying to understand or forecast.
- **Independent Variable:** These are factors that influence the analysis or target variable and provide us with information regarding the relationship of the variables with the target variable.

Use of regression:

- Financial Industry- Understand the trend in the stock prices, forecast the prices, evaluate risks in the insurance domain
- Marketing- Understand the effectiveness of market campaigns, forecast pricing and sales of the product.
- Manufacturing- Evaluate the relationship of variables that determine to define a better engine to provide better performance
- Medicine- Forecast the different combination of medicines to prepare generic medicines for diseases.

Types of regression:

- Linear regression
- Logistics regression

What is Linear Regression?

Linear Regression is a predictive model used for finding the ***linear*** relationship between a dependent variable and one or more independent variables.

$$Y = a + bx$$

Dependent Variable is continuous

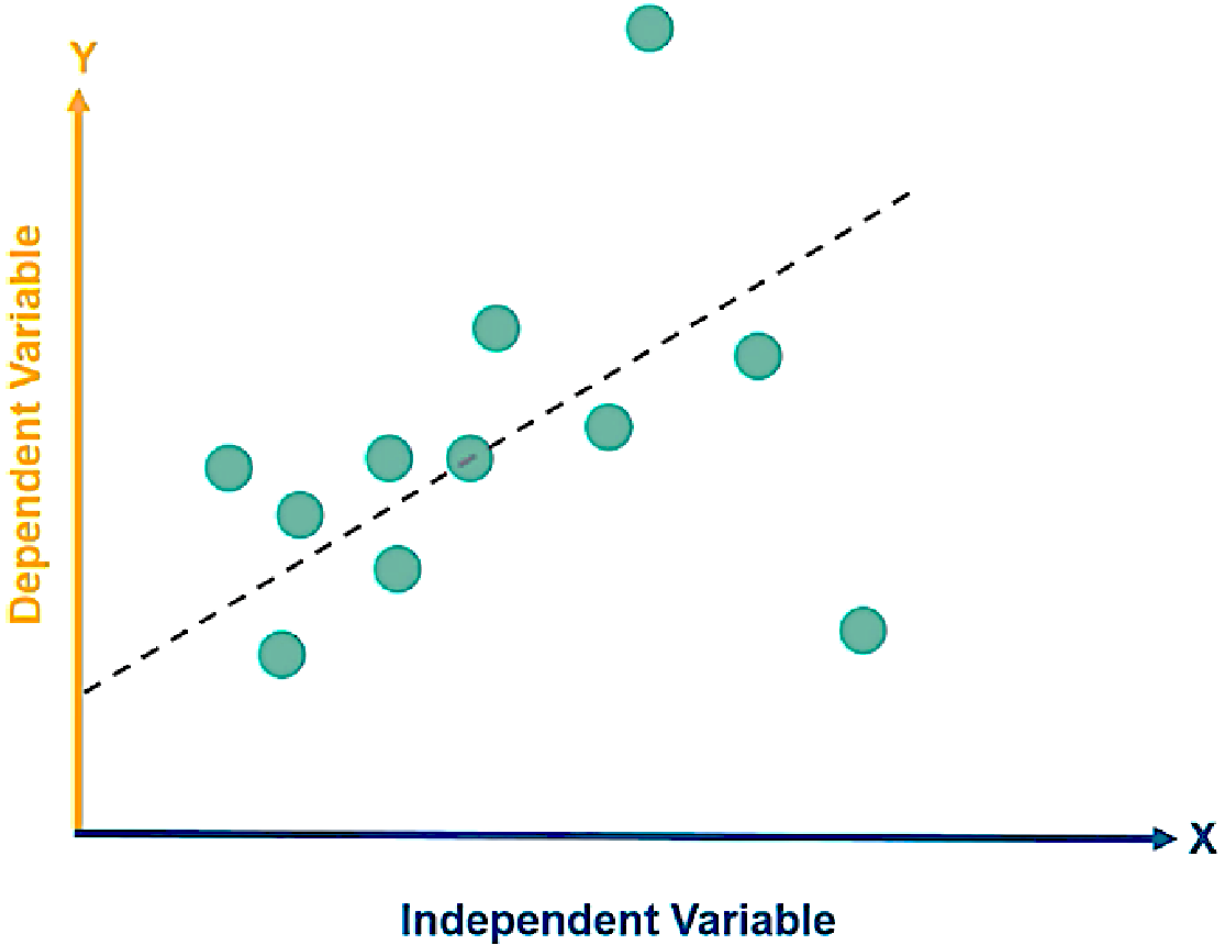
Here, 'Y' is our dependent variable, which is a continuous numerical and we are trying to understand how does 'Y' change with 'X'.

Examples of Independent & Dependent Variables:

• x is Rainfall and y is Crop Yield

• x is Advertising Expense and y is Sales

• x is sales of goods and y is GDP



If the relationship with the dependent variable is in the form of single variables, then it is known as Simple Linear Regression

Types of Linear regression:

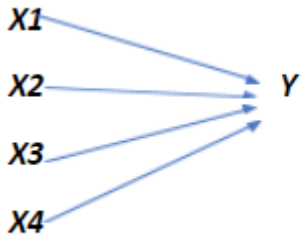
1. Simple Linear Regression
2. Multiple Linear Regression
3. Polynomial Linear Regression

Simple Linear Regression

X* ——> *Y

If the relationship between Independent and dependent variables are multiple in number, then it is called Multiple Linear Regression

Multiple Linear Regression



Simple Linear Regression Model

As the model is used to predict the dependent variable, the relationship between the variables can be written in the below format.

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Where,

Y_i – Dependent variable

β_0 — Intercept

β_1 – Slope Coefficient

X_i – Independent Variable

ε_i – Random Error Term

The main factor that is considered as part of Regression analysis is understanding the variance between the variables. For understanding the variance, we need to understand the measures of variation.

Practical exercise:

Research Project Title: Effect of body mass index (BMI) on blood pressure and hypertension among adult women in Nepal

Main Exposure: BMI - Calculated by weight in kilogram divided by Height in meter squared
Outcome variable: Blood pressure (Systolic and diastolic blood pressure)-average of three reading and Hypertension
Covariates: Residence type, Wealth index, Age, Marital Status, Smoking status, Education level, Medicine BP
Confounders: Age, Marital Status, Smoking status, Education level, Type of residence

Research question I: Is BMI associated with blood pressure outcome in adult women?
Hypothesis: H0 = There is no association with BMI and blood pressure outcome
H1 = There is an association with BMI and blood pressure outcome

Research question II: Is BMI associated with hypertension in adult women?

Hypothesis: H0 = There is no association with BMI and hypertension
H1 = There is an association with BMI and hypertension.

Statistical analysis method:

This study presents the association between the body mass index (BMI) and blood pressure in adult women. At first we identified our exposure, outcome, confounders and covariates. We prepared our dataset by dropping missing observations from exposure, outcome, confounders and covariates. We changed the variables to a meaningful name, categorized age as age groups, generated new variables for the average of systolic and diastolic. The average systolic and diastolic blood pressure are indicating the measurement of the hypertension status. By summarizing these we measured hypertension. We recoded BMI as underweight, ideal, overweight and obese.

While the outcome variable is continuous we did bivariate linear regression among the exposure BMI and blood pressure. We considered age, education, residence type, marital and smoking status as confounders. These variables are associated with both the exposure and outcome. We adjusted the confounders by using the multivariate linear regression. We did linear regressions between average systolic blood pressure and bmi, age, residence type, educational level, marital and smoking status. We also did a logistic regression to see the association between the BMI and hypertension. We adjusted the confounders by using the multivariate logistic regression.

Results:

Table-1 shows that the socio-demographic information of 8,645 Nepali adult women. The total number of type of residence is higher in urban areas than the rural areas. Most of the participants they didn't complete preschool (47.47%, n=4,104) and it was almost half of the participants. That means they are uneducated or illiterate. Only 13.23% (n=1,144) participants were completed primary education and 28.17% (n=2,435) were completed secondary level. In marital status, 74.12% (n=6,408) participants were married. A greater portion of adult women were participated with the age range 15-39 years (61.53%, n=5,319). As we saw a majority of the participants didn't smoke (93.43%, n=8,345). The BMI results indicating that 61.40% (n=5,308) of the participants were normal weight or in the ideal stage, 18.76% (n=1,622) underweight, 15.74% (n=1,361) overweight and 4.09% (n=354) obese. Finally, we found the hypertension status as 64.19% (n=5,549) participants were normotensive and 35.81% (n=3,096) having hypertension.

Table-1: participants socio-demographic characteristics

Variables	**Data frequency N= 8645**	**Percentage (%)**
Type of residence		
Urban	5,460	63.16
Rural	3,185	36.84
Highest educational level		
Preschool	4,104	47.47
Primary	1,144	13.23
Secondary	2,435	28.17
Higher	958	11.08
Marital status		
Never married	1,358	15.71
Married	6,408	74.12
Widowed	792	9.16
Divorced	87	1.01

Age in years		
15-39 Years	5,319	61.53
40-59 Years	2,233	25.83
60-79 Years	965	11.16
80 Years and Above	128	1.48
Smoking status		
Yes	300	3.47
No	8,345	96.53
Taking medicine to lower bp		
Yes	309	3.57
No	8,336	96.43
BMI		
Underweight	1,622	18.76
Ideal	5,308	61.40
Overweight	1,361	15.74
Obese	354	4.09
hypertension status		
Normotensive	5,549	64.19
Hypertensive	3,096	35.81

Table-2 shows that the result of the association among the body mass index (BMI) and the systolic blood pressure (SBP). To find the association between the BMI and SBP, we did bivariate linear analysis and to adjust confounders, we performed multivariate linear regression. The result shows that the BMI has significant association with systolic blood pressure that means if BMI increases then systolic blood pressure will also increase. For each unit of increasing BMI, the systolic blood pressure of the participants increased by 5.03 mmHg (Coefficient: 5.03, 95% CI: 4.52, 5.53). The participants age, marital and smoking status are positive association with systolic blood pressure as opposed to the education level and type of residence are negative association.

Table-2: Association of BMI and systolic blood pressure

Variables	**Unadjusted systolic coefficient (95% confidence interval)**	**Adjusted coefficient (95% CI)**
BMI	4.68 *** (4.13, 5.24)	5.03 *** (4.52, 5.53)
Age	0.57 *** (.556, .597)	
Type of residence		
Urban	Reference	
Rural	0.661 (-.17, 1.49)	
Educational level		

No education	Reference
Primary	-7.81 *** (-9.01, -6.63)
Secondary	-11.82 *** (-12.72, -10.91)
Higher	-13.59 *** (-14.87, -12.32)
Marital status	
Never married	Reference
Married	8.31 *** (7.24, 9.36)
Widowed	24.92 *** (23.34, 26.51)
Divorced	11.66 *** (7.74, 15.57)
Smoking status	
Yes	12.37 *** (10.19, 14.54)
No	Reference

P-value: *** < 0.001, ** < 0.01, * < 0.05

Table-3 shows that the result of the association among the body mass index (BMI) and the diastolic blood pressure (DBP). To find the association between the BMI and DBP, we did bivariate linear analysis and to adjust confounders, we performed multivariate linear regression. The result shows that the BMI has significant association with diastolic blood pressure that means if BMI increases then diastolic blood pressure will also increase. For each unit of increasing BMI, the diastolic blood pressure of the participants increased by 4.63 mmHg (Coefficient: 4.63, 95% CI: 4.31, 4.95). The participants age, marital and smoking status are positive association with diastolic blood pressure as opposed to the education level and type of residence are negative association.

Table-3: Association of BMI and diastolic blood pressure

Variables	**Unadjusted diastolic coefficient (95% confidence interval)**	**Adjusted coefficient (95% CI)**
BMI	4.48 *** (4.17, 4.81)	4.63*** (4.31, 4.95)
Age	0.19 *** (.18, .21)	
Type of residence		
Urban	Reference	
Rural	-0.15 (-.63, .33)	
educational level		
No education	Reference	
Primary	-1.89 *** (-2.61, -1.17)	
Secondary	-4.03*** (-4.57, -3.48)	
Higher	-4.39*** (-5.16, -3.63)	
Marital status		

	Never married	Reference
	Married	4.3*** (3.67, 4.95)
	Widowed	7.666857*** (6.71, 8.62)
	Divorced	7.17*** (4.82, 9.53)
Smoking status		
	Yes	4.14 *** (2.87, 5.41)
	No	Reference

P-value: *** < 0.001, ** < 0.01, * < 0.05

Table-4 represents the association between the body mass index (BMI) and the hypertension among the adult women. To find the association, we did bivariate logistic regression and to adjust the confounders we performed multivariate logistic regression. We found the significant positive association between the BMI and Hypertension. Evidence shows that overweight women 2.53 times more likely (Coefficient: 3.53, 95% CI: 2.22, 2.88) to have hypertension compare to ideal measurement. We also see the obese women 4.33 times more likely (Coefficient: 4.33, 95% CI: 3.40, 5.52) to have hypertension compare to normal hypertension. We see that age, marital and smoking status are also positive association.

Reference:

1. https://www.mygreatlearning.com/blog/what-is-regression/