# SESSION-13 (FINAL): LOGISTICS REGRESSION ANALYSIS

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

## Review the full course

- Session-1 # Stata introduction, download and installation
- Session-2 # Stata Window and Command's
- Session-3 # Data Management - Dataset preparation and import to Stata
- Session-4 # Data Management (Import to excel, Variables, Operators)
- Session-5 # Data Management (Log and DO file)
- Session-6 # Data analysis and visualization (Append, Merge, Variables grouping)
- Session-7 # Writing style and data visualization
- Session-8 # Descriptive Statistics - Cross Tabulation and Frequency table
- Session-9 # Descriptive Statistics - One way table
- Session-10 # Descriptive Statistics - Mean, Median, Mode, Percentile, Quartile, Range, Variance, SD
- Session-11 # Statistical Test - One sample proportion, t-Test, ANNOVA, Chi-Squared, Correlation, Hypothesis Testing
- Session-12 # Statistical Test - Linear Regression

## Session-13

**Agenda:**

- Logistics regression

**Before starting:**

Watch the video first: https://www.youtube.com/watch?v=rSU1L3-xRk0

**What is logistic regression?**

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables.

So: Logistic regression is the correct type of analysis to use when you're working with binary data. You know you're dealing with binary data when the output or dependent variable is dichotomous or categorical in nature; in other words, if it fits into one of two categories (such as "yes" or "no", "pass" or "fail", and so on).

**What are log odds?**

In very simplistic terms, log odds are an alternate way of expressing probabilities. In order to understand log odds, it's important to understand a key difference between odds and probabilities: odds are the ratio of something happening to something not happening, while probability is the ratio of something happening to everything that could possibly happen.

**What is logistic regression used for?**

Now we know, in theory, what logistic regression is—but what kinds of real-world scenarios can it be applied to? Why is it useful?

Logistic regression is used to calculate the probability of a binary event occurring, and to deal with issues of classification. For example, predicting if an incoming email is spam or not spam, or predicting if a credit card transaction is fraudulent or not fraudulent.

**What are the different types of logistic regression?**

In this post, we've focused on just one type of logistic regression—the type where there are only two possible outcomes or categories (otherwise known as binary regression). In fact, there are three different types of logistic regression, including the one we're now familiar with.

The three types of logistic regression are:

1. **Binary logistic regression** is the statistical technique used to predict the relationship between the dependent variable (Y) and the independent variable (X), where the dependent variable is binary in nature. For example, the output can be Success/Failure, 0/1 , True/False, or Yes/No. This is the type of logistic regression that we've been focusing on in this post.
2. **Multinomial logistic regression** is used when you have one categorical dependent variable with two or more unordered levels (i.e two or more discrete outcomes). It is very

similar to logistic regression except that here you can have more than two possible outcomes. For example, let's imagine that you want to predict what will be the most-used transportation type in the year 2030. The transport type will be the dependent variable, with possible outputs of train, bus, tram, and bike (for example).

3. **Ordinal logistic regression** is used when the dependent variable (Y) is ordered (i.e., ordinal). The dependent variable has a meaningful order and more than two categories or levels. Examples of such variables might be t-shirt size (XS/S/M/L/XL), answers on an opinion poll (Agree/Disagree/Neutral), or scores on a test (Poor/Average/Good).

## Advantages of logistic regression

- **Logistic regression is much easier to implement than other methods, especially in the context of machine learning**
- **Logistic regression works well for cases where the dataset is linearly separable**
- **Logistic regression provides useful insights**

## Disadvantages of logistic regression

- **Logistic regression fails to predict a continuous outcome**
- **Logistic regression assumes linearity between the predicted (dependent) variable and the predictor (independent) variables**
- **Logistic regression may not be accurate if the sample size is too small**

## Final thoughts

So there you have it: A complete introduction to logistic regression. Here are a few takeaways to summarize what we've covered:

- Logistic regression is used for classification problems when the output or dependent variable is dichotomous or categorical.
- There are some key assumptions which should be kept in mind while implementing logistic regressions (see section three).
- There are different types of regression analysis, and different types of logistic regression. It is important to choose the right model of regression based on the dependent and independent variables of your data.

Source: https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/

## Final Project:

# 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**<br>**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.

Table-4: Association of BMI and hypertension by using multivariate logistics regression

| **Variables** | **Unadjusted OR (95% confidence interval)** | **Adjusted OR (95% CI)** |
|---|---|---|
| **BMI** | | |
| Ideal | Reference | |
| Underweight | 0.69 *** (.61, .78) | 0.52 *** (.45, .59) |
| Overweight | 2.54 *** (2.25, 2.87) | 2.53 *** (2.22, 2.88) |
| Obese | 4.37 *** (3.47, 5.49) | 4.33 *** (3.40, 5.52) |
| **Age** | | |
| 15-39 Years | Reference | |
| 40-59 Years | 3.03 *** (2.73, 3.36) | |
| 60-79 Years | 4.02 *** (3.48, 4.63) | |
| >= 80 Years | 6.83 *** (4.66, 10.01) | |
| **Type of residence** | | |
| Urban | Reference | |
| Rural | .97 ( .88, 1.06) | |
| **Educational level** | | |
| No education | Reference | |
| Primary | .65 *** (.57, .75) | |
| Secondary | .45 *** (.40, .50) | |
| Higher | .38 *** (.32, .45) | |
| **Marital status** | | |

| | |
|---|---|
| Never married | Reference |
| Married | 2.33 *** (2.02, 2.69) |
| Widowed | 5.43 *** (4.47, 6.59) |
| Divorced | 3.31 *** (2.12, 5.15) |
| **Smoking status** | |
| Yes | 2.23 *** (1.76, 2.81) |
| No | Reference |

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