



# LOGISTIC REGRESSION

Department of Epidemiology and Biostatistics

Hanna Melesse (MD, MPH)

Sep. 23, 2022

# OUT LINE

---

- Introduction to logistic regression
- Assumptions of logistic regression
- Steps in conducting logistic regression
- Advantages and Limitations of logistic regression
- Demonstration

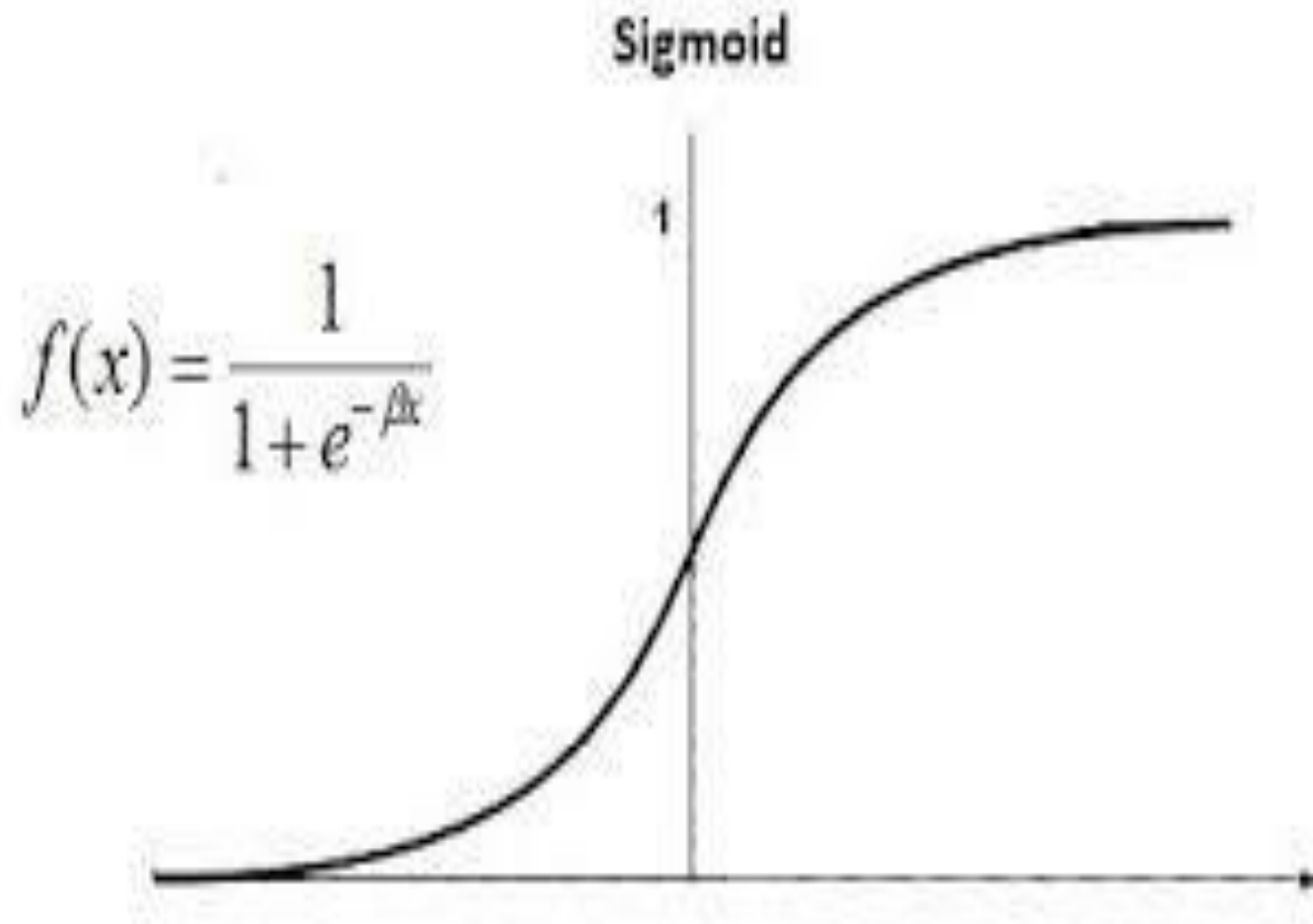


# Introduction: what is logistic regression

- A logistic regression model is a method used to model a data with dependent variable (out put data) that is categorical or binary.



# Introduction: logistic regression



$$\log \left[ \frac{P(x)}{1 - P(x)} \right] = \beta_0 + \beta_1 x.$$

# Introduction: logistic regression

---

## Logistic regression

### Binary logistic regression

- Number of categories are 2

**Eg.** Pass or fail

### Ordinal logistic regression

- Number of categories are 3 or more
- Characteristics are at natural ordering of the level

**Eg** strongly agree, agree, neutral, disagree, strongly disagree

### Nominal logistic regression

- Number of categories are 3 or more
- Characteristics are not as per at natural ordering of the level

**Eg.** Disease a, disease b, disease c



# Assumptions in logistic regression

---

- The outcome variable should be categorical (binary, ordinal, nominal)
- The predictor or the independent variable can be continuous or categorical
- Assumes linearity of independent variables and log odds.
- Independent observations: the observations should not come from repeated measurements or matched data.



# Assumptions in logistic regression

---

- ❑ logistic regression typically requires a large sample size.
- ❑ The model should provide a good fit to the data
- ❑ The correlation among the predictors(multi-collinearity) should not be severe.



# Steps for conducting the binary logistic regression in SPSS

---

- Step 1 : Check the assumptions
- Step 2: conduct the univariate analysis
- Step 3: conduct the multivariate analysis
- Step 4: interpret the result





# Step 1 : Check the assumptions

## Model fit test

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	33.732	2	.000
	Block	33.732	2	.000
	Model	33.732	2	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	212.083 <sup>a</sup>	.099	.186

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.205	1	.651

## Omnibus test of model coefficients

- Statistical significance indicate good model fit

**Model Summary:** Nagelkerke R Square, which tells us the percentage of the variation in the response variable that can be explained by the predictor variables.

- In this case, wealth, marital status and age are able to explain 18.6% of the variability in anxiety disorder.

## Hosmer and Lemeshow test

- Statistical significance indicate the model is not good fit for the data

# Step 1 : Check the assumptions

## □ How to check multi collinearity

The screenshot displays the IBM SPSS Statistics Data Editor interface. The 'Analyze' menu is open, and the 'Regression' option is selected, leading to the 'Linear...' dialog box. The 'Dependent' variable is 'anxiety\_ [anxiety\_]'. The 'Independent(s)' variables are 'Marital Status [MaritalStatus]' and 'Wealth Final [Wealth\_Final]'. The 'Method' is set to 'Enter'. The 'Linear Regression: Statistics' sub-dialog box is also open, showing the 'Collinearity diagnostics' checkbox checked under the 'Regression Coefficients' section. The 'Residuals' section has 'Outliers outside: 3 standard deviations' selected.

Label	Values	Missing	Column
row old ...	None	None	12
ave you...	{0, NO}...	None	12
/that is t...	None	None	14
/that is y...	{1, Married ...	None	12
What is ...	{1, House ...	None	12
ther Sp...	None	None	18

# Step 1 : Check the assumptions

- How to check **multi collinearity**

Coefficients<sup>a</sup>

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics		
	B	Std. Error	Beta			Tolerance	VIF	
1	(Constant)	.312	.081		3.855	.000		
	Marital Status	.396	.074	.286	5.373	.000	.949	1.053
	Wealth Final	.120	.036	.178	3.341	.001	.949	1.053

a. Dependent Variable: anxiety\_

# How to perform logistic regression in SPSS

## ■ **Step 2: conduct the univariate analysis**

### **Variables inclusion and selection**

- limited sample size in relation to the number of candidate variables needs pre-selection
- test all variables previously, using models univariate models and include in the multivariate model all variables that have shown a relaxed P-value (for instance,  $P \leq 0.25$ ).

### Relaxed P-value criterion

- ✓ reduce the initial number of variables in the model
- ✓ reduce the risk of missing important variables



# How to perform logistic regression in SPSS

- Step 2: conduct the univariate analysis
- Click the **Analyze** tab, then **Regression**, then **Binary Logistic Regression**:

The screenshot shows the IBM SPSS Statistics Data Editor interface. The 'Analyze' menu is open, and the 'Regression' option is selected. The 'Binary Logistic...' option is highlighted in the submenu. The background shows a data grid with columns for 'Name' and 'Type'.

Name	Type
Chron_Dis	Numeric
anxiety_	Numeric
depression_	Numeric
age_	Numeric
Education_3	Numeric
Education_...	Numeric
Residence	Numeric
Wealth_Final	Numeric
Alcohol_useF	Numeric

# How to perform logistic regression in SPSS

## Step 2: conduct the univariate analysis

The image shows two overlapping dialog boxes from the SPSS software. The background dialog box is the 'Logistic Regression: Options' dialog, and the foreground dialog box is the main 'Logistic Regression' dialog.

**Logistic Regression: Options Dialog (Background):**

- Statistics and Plots:**
  - Classification plots
  - Hosmer-Lemeshow goodness-of-fit
  - Casewise listing of residuals
  - Outliers outside 2 std. dev.
  - All cases
  - Correlations of estimates
  - Iteration history
  - CI for exp(B): 95 %
- Display:**
  - At each step
  - At last step
- Probability for Stepwise:**
  - Entry: 0.05
  - Removal: 0.10
  - Classification cutoff: 0.5
  - Maximum iterations: 20
- Include constant in model

Buttons: Continue, Cancel, Help

**Main Logistic Regression Dialog (Foreground):**

- Dependent:** anxiety\_ [anxiety\_]
- Block 1 of 1:** Previous, Next
- Covariates:** Wealth\_Final(Cat)
- Method:** Enter
- Selection Variable:** Rule...

Buttons: Categorical..., Save..., Options..., Bootstrap..., >a\*b>, OK, Paste, Reset, Cancel, Help

q121...	Numeric	12	0	Q12
Repr...	String	12	0	gp_c

# How to perform logistic regression in SPSS

## ■ Step 2: conduct the univariate analysis

a. The cut value is .500

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 <sup>a</sup> Wealth_Final(1)	1.525	.373	16.692	1	.000	4.597	2.212	9.557
Constant	1.253	.207	36.620	1	.000	3.500		

a. Variable(s) entered on step 1: Wealth\_Final.



# How to perform logistic regression in SPSS

## Step 3: conduct the multivariate analysis

The screenshot shows the IBM SPSS Statistics Data Editor interface. The 'Analyze' menu is open, and the 'Binary Logistic...' option is selected. The background shows a data table with columns for Name and Type, and rows for variables like Chron\_Dis, anxiety\_, depression\_, age\_, Education\_3, Education\_..., Residence, Wealth\_Final, and Alcohol\_useF.

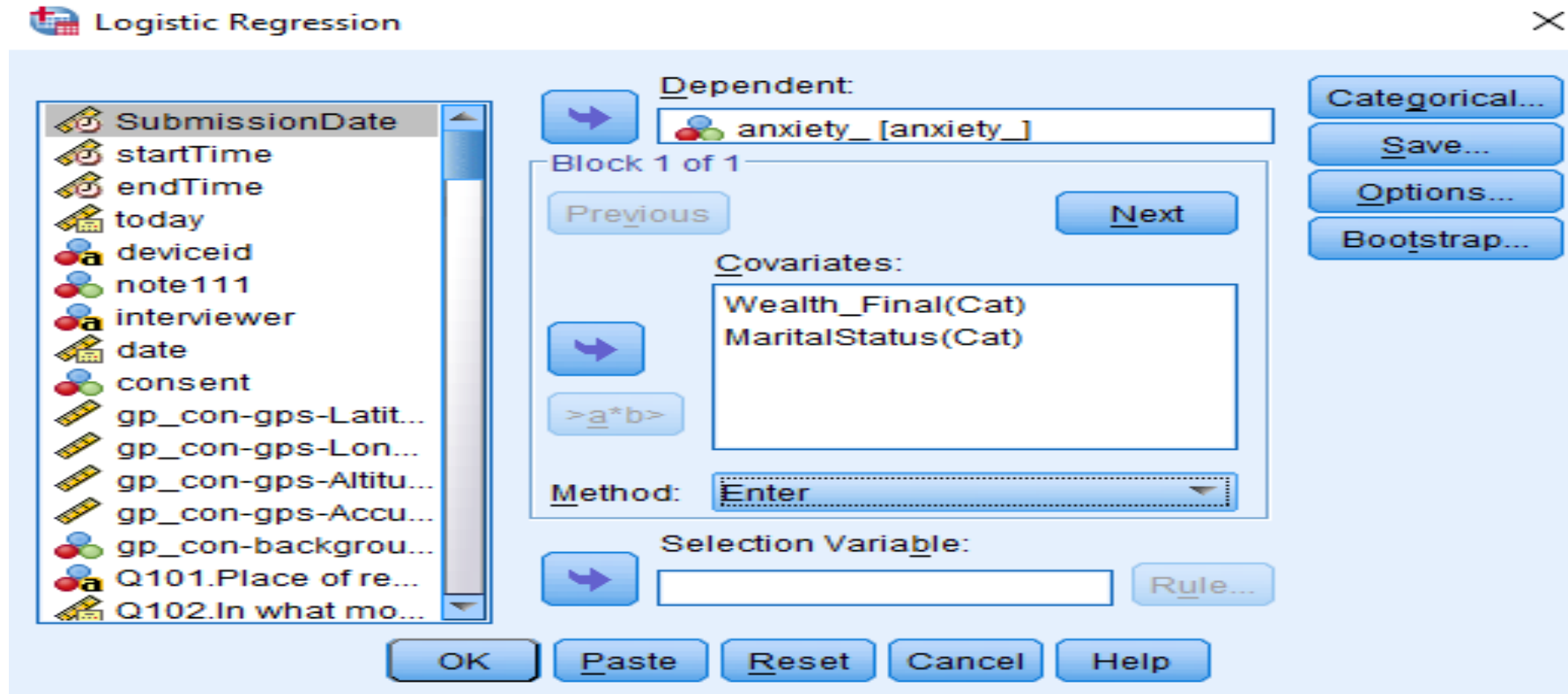
Name	Type
134 Chron_Dis	Numeric
135 anxiety_	Numeric
136 depression_	Numeric
137 age_	Numeric
138 Education_3	Numeric
139 Education_...	Numeric
140 Residence	Numeric
141 Wealth_Final	Numeric
142 Alcohol_useF	Numeric
143	
144	
145	
146	
147	
148	
149	
150	
151	
152	
153	
154	
155	
156	
157	
158	

The 'Analyze' menu path is: Analyze > Regression > Binary Logistic... The 'Binary Logistic...' option is highlighted in yellow.



# How to perform logistic regression in SPSS

- **step 3: conduct the multivariate analysis**
- In the new window that pops up, drag the binary response variable **ANXIETY** into the box labelled Dependent. Then drag the two predictor variables **wealth status** and **marital status** into the box labelled Block 1 of 1. Leave the **Method** set to Enter. Then click **OK**.



# How to perform logistic regression in SPSS

## Step 3: Interpret the out put

Classification Table<sup>a</sup>

Observed		Predicted		Percentage Correct
		anxiety_		
		anxiety	no anxiety	
Step 1	anxiety_ anxiety	9	32	22.0
	no anxiety	8	274	97.2
Overall Percentage				87.6

a. The cutvalue is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1 <sup>a</sup>	gp_conq103age	-.010	.039	.072	1	.788	.990	.917	1.068
	Wealth_Final(1)	1.246	.390	10.197	1	.001	3.478	1.618	7.474
	MaritalStatus(1)	1.908	.513	15.184	1	.000	7.371	2.690	20.134
	Constant	-.114	1.109	.011	1	.918	.892		

a. Variable(s) entered on step 1: gp\_conq103age, Wealth\_Final, MaritalStatus.

## What does the boxes indicate

**Classification Table:** Overall Percentage, which tells us the percentage of observations that the model was able to classify correctly.

## Variables in the Equation:

- **Wald:** is used to determine whether or not each predictor variable is statistically significant.
- **Sig:** The p-value that corresponds to the Wald test statistic for each predictor variable.
  - p-value for **wealth** is .001 and the p-value for **marital status** is .0001.
- **Exp(B):** The odds ratio for each predictor variable. This tells us the change in the odds of a pregnant mother getting anxiety disorder associated with an increase in a given predictor variable.

# Step 4: interpret the result

---

## Example : report

- Logistic regression was performed to determine how wealth status, marital status and age affect a pregnant mothers probability of getting anxiety disorder. A total of 323 pregnant mothers were used in the analysis.
- The model explained 18.6% of the variation in anxiety disorder due to this predictors and correctly classified **87.6%** of cases.
- We can use the coefficients (the values in the column labeled B) or the odds ratio(the values in the column labeled Exp(B)) to predict the probability that a pregnant women will get anxiety disorder.



# Step 4: interpret the result

- ✓ The odds of anxiety disorder in a pregnant women who are single is 7.25 times higher compared to the married pregnant women keeping other variables constant.
- ✓ The odds of anxiety disorder in a pregnant women who are in low wealth status is 3.5 times higher compared to the high wealth status pregnant women keeping other variables constant.
- ✓ For each increment of age there is decrement of anxiety disorder by 0.005times or 0.01% keeping others constant (but the  $p > 0.05$  so no need to interpret this variable)

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 <sup>a</sup>								
gp_conq103age	-.010	.039	.072	1	.788	.990	.917	1.068
Wealth_Final(1)	1.246	.390	10.197	1	.001	3.478	1.618	7.474
MaritalStatus(1)	1.998	.513	15.184	1	.000	7.371	2.699	20.134
Constant	-.114	1.109	.011	1	.918	.892		

a. Variable(s) entered on step 1: gp\_conq103age, Wealth\_Final, MaritalStatus.

# Advantages of logistic regression

---

- It does not only provides a measure of how appropriate a predictor (coefficient size) is, but also its direction of association (positive or negative).
- Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.
- It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.
- It makes no assumptions about distributions of classes.
- Logistic regression is easier to implement, interpret, and efficient to train

# Limitations of logistic regression

---

- If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.
- It can only be used to predict discrete functions.
- Non-linear problems can't be solved with logistic regression because it has a linear decision surface.
- Logistic Regression requires average or no multi collinearity between independent variables

---

- Demonstration with SPSS

# Summary

---

- logistic regression model is a method used to model a data with dependent variable that is categorical or binary.
- Check the assumptions of logistic regression.
- Conduct the bivariate and multivariate analysis and interpret.



# Reference

---

- Principle of Biostatistics 2<sup>nd</sup> Edition (Marcello Pagano)
- Bujang MA, Sa'at N, Sidik TMITAB, Joo LC. Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data. Malays J Med Sci. 2018 Jul;25(4):122–30.
- Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014 Feb 15;24(1):12–8.
- Schober P, Vetter TR. Logistic Regression in Medical Research. Anesth Analg. 2021 Feb;132(2):365–6.



---

■ Thank you