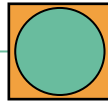


Video Tutorial: Logistic Regression

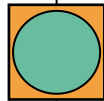
Data for Impact



D4I works to:

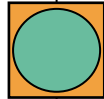
- Generate strong evidence for program and policy decision making
- Build individual and organizational capacity
- Enhance the use of data for global health programs and policies

Outline

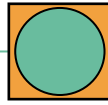


- Logistic Regression Overview
- STATA Demonstration

Resources

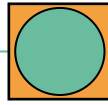


- Background information sheet
 - Overview of logistic regression
 - Useful resources
- PowerPoint presentation
- Sample dataset
- Sample do file



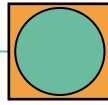
Video tutorial objective

- The team was tasked with creating video tutorials on how to use specific methods and commands to analyze routine data for evaluation and research.
 - This video is not intended for beginners but for those who have some experience with data analysis and STATA
- This tutorial focuses on logistic regression and is part of a part of a series that outlines the use of two statistical methods in the evaluation of routine data
- For information on different methods used to analyze routine data, as well as the materials used in this presentation see <https://www.data4impactproject.org/>



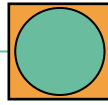
Logistic Regression

- Dependent variable is a nonmetric, dichotomous/binary variable
 - Analytical approach used to model dichotomous outcome variables
- Extensive use in health care and social sciences
- The log odds of the outcome is modeled as a linear combination of the predictor variables.



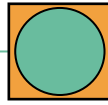
Benefits

- Gained widespread application in situations involving a binary outcome
- Easy to use and interpretation
- Robust estimation properties
- No assumptions on distribution of classes
- It can be extended to multi-class classification (multinomial logistic regression)



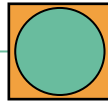
Limitations

- Assumption of linearity between exploratory variables and log odds
- Can lead to an overfit model
- Requires no/moderate collinearity between covariates
- Cannot be used for continuous outcomes



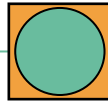
Assumptions

1. Dependent variable should consist of two categorical, independent groups
 - Dichotomous variable
 - Categories are mutually exclusive and exhaustive
2. Two or more independent variables
 - Measured at the continuous or nominal level
3. Independence of observations
 - No relationship between the observations



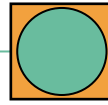
Assumptions

4. Data must not show multicollinearity
 - High correlation between two or more independent variables
5. Linear relationship between any continuous independent variables and the logit transformation of the dependent variable
6. No significant outliers, high leverage points or highly influential points



Sample Size

- Peduzzi Concato, Kemper, Holford, & Feinstein (1996)
 - $n = 10 (k / p)$
 - k = number of covariates/predictors
 - p = proportion of successes
 - Minimum $n = 100$
- Long (1997)
 - $n < 100$ cases \rightarrow risky
 - $n \geq 500$ cases \rightarrow adequate
 - at least 10 cases per predictor



Dependent Variable/Outcome

0 1

Failure Success

Regression Model

Odds/logit

$$\ln \left(\frac{p}{(1-p)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

where

- p = probability when outcome occurs ($y=1$)
- X = independent variables
- β = regression coefficients

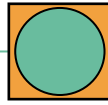
$\beta \rightarrow$ change in the expected log odds relative to a one unit change in X_1 (a given predictor) holding all other predictors constant



Amount of change

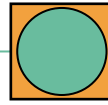


Direction of change



Odds

- **Odds > 0** → positive association
 - Log odds of being a success/falling into target group **increases** with increasing values in the covariate variable
- **Odds < 0** → negative association
 - Log odds of being a success/falling into target group **decreases** with increasing scores in the covariate variable
- **Odds = 0** → lack of association
 - **No systematic increase or decrease** in the log odds of being a success/falling into target group with increasing values in the covariate variable



Regression Model

Dependent Variable/Outcome

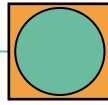
0	1
Failure	Success

Odds Ratio $\left(\frac{p}{(1-p)} \right) = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}$

where

- p = probability when outcome occurs ($y=1$)
- X = independent variables
- β = regression coefficients

$\beta \rightarrow$ change in the odds of success relative to a one unit change in X_1 (a given predictor) holding all other predictors constant



Odds Ratio

- **Odds Ratio (OR)**

- Values are between 0 and $+\infty$

- **OR > 1.0** → positive association

≈ Odds > 0

- **OR < 1.0** → negative association

≈ Odds < 0

- **OR = 1.0** → lack of association

≈ Odds = 0

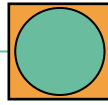
- **95% Confidence Interval**

- Range includes 1

- association is not statistically significant ($p \geq 0.05$)

- Range does not include 1

- statistically significant association ($p < 0.05$)



STATA Syntax

Dependent Variable/Outcome

0

Failure

1

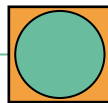
Success

Odds

- `Logit depvar indepvar1 indepvar2 indepvar3`

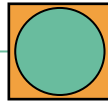
Odds Ratio

- `Logit depvar indepvar1 indepvar2 indepvar3, or`
- `Logitistic depvar indepvar1 indepvar2 indepvar3`



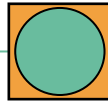
STATA Syntax

- Continuous covariates
 - Entered as is or add prefix "c." before the variable name;
 - e.g.; age or c.age
- Categorical covariates
 - Add prefix "i." before the variable name
 - e.g.; i.education
 - 1. No education [Reference]
 - 2. Primary education
 - 3. Secondary education
 - 4. Higher education
 - Change the reference group → b[level].varname
 - e.g. b2.education --> reference is now primary education



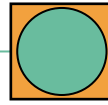
Inclusion of Covariates

- A priori
- Selection based on
 - Theory
 - Previous research



What to report

- I. Brief summary of the analysis conducted
- II. Information about your sample, including any missing values
- III. Examination of all the assumptions of the logistic regression, including any remedies that were taken for violations of any of these assumptions.
- IV. Assessment of how well the model fits the data e.g. use of measures, such as the Hosmer-Lemeshow test
- V. Regression coefficients and/or odds ratios and 95% confidence intervals.



Examples of studies using logistic regression

PLOS ONE

RESEARCH ARTICLE

Recent HIV infection among pregnant women in the 2017 antenatal sentinel cross-sectional survey, South Africa: Assay-based incidence measurement

Selamawit Woldesenbet^{1,2*}, Tendes Samuel Manda^{4,5}, Mireille Cheyip⁶, Ka Karidia Diallo⁸, Bharat Parekh⁹, Adriar

OPEN ACCESS Freely available online



Patient Enrolment into HIV Care and Treatment within 90 Days of HIV Diagnosis in Eight Rwandan Health Facilities: A Review of Facility-Based Registers

Fikse², Veronicah Mugisha³, Anita Asimwe⁴,

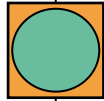


¹, Amsterdam, The Netherlands, ³ICAP, Mailman School of Public Health, Columbia University, ⁴Centers for Disease Prevention and Control, Ministry of Health, Kigali, Rwanda, ⁵Amsterdam Institute of Public Health, ⁶Department of Infectious Disease Epidemiology, Centre for Infection and Immunity Amsterdam (CINIMA), AMC, Amsterdam, The Netherlands,

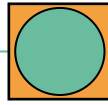
RESEARCH ARTICLE

Socio-demographic determinants of low birth weight: Evidence from the Kassena-Nankana districts of the Upper East Region of Ghana

Isaiah Awintuen Agorinya^{1,2,3*}, Edmund Wedam Kanmiki⁴, Engelbert Adamwaba Nonterah^{1,5}, Fabrizio Tediosi³, James Akazili^{1,2}, Paul Welaga¹, Daniel Azongo¹, Abraham Rexford Oduro¹

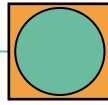


Tutorial Example



Tutorial Example

- Research question:
 - What factors are associated with women's satisfaction with maternity care services?
- Data Source
 - Health facility registries (routine data)
 - Exit interviews



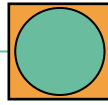
Tutorial Example: Data Source

Health facility's antenatal care and delivery records

- Age
- Gravida
- Parity
- Type of delivery
- Gestational age of pregnancy
- Syphilis test
- HIV test

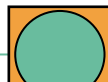
Exit interviews conducted before discharge

- Socio-demographic characteristics
 - Occupational status, marital status, educational level
- Rating of satisfaction
 - Staff friendliness
 - Staff helpfulness
 - Speed of service
 - Recommendation of service to a friend/family member



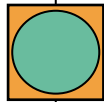
Outcome



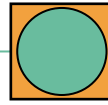


age	work	mstatus	gestage	gravida	parity	numANC	ethnic~y	delivery	satisf~i
28	1	1	37	1	0	3	5	2	0
44	1	0	39	3	2	3	5	1	0
43	1	1	36	4	1	1	3	1	0
32	1	0	39	3	0	3	5	1	0
49	1	0	37	2	2	3	4	2	0
31	0	1	40	1	1	2	4	2	0
34	1	0	36	5	3	4	1	2	0
19	1	0	37	5	3	4	5	2	0
17	1	0	36	1	1	3	2	1	0
16	1	0	36	7	3	1	2	1	0
47	1	1	39	2	2	3	5	2	0
49	0	1	40	8	2	3	5	1	0
21	0	0	40	6	0	1	3	1	0
17	1	0	38	6	1	2	4	1	0
27	1	1	39	1	1	1	5	1	0
36	1	0	37	4	4	2	5	1	1
43	0	1	37	1	1	3	3	1	1
43	1	1	37	1	1	6	2	2	1
29	1	1	36	3	3	4	1	2	1
19	0	1	38	6	1	5	5	2	1
38	1	1	38	1	1	1	2	1	1
39	1	0	40	6	1	3	2	1	1
21	0	1	40	5	1	3	3	2	1
24	1	0	38	5	2	2	3	2	1
32	1	1	40	1	1	2	4	1	1
16	0	1	41	4	4	5	4	1	1
30	0	1	36	3	1	4	5	1	1

Sample
Dataset



STATA Demonstration



Univariate Analysis

- Categorical variables
 - Percentages
- Continuous variables
 - Normally distributed
 - Means
 - Standard deviations
 - Not normally distributed
 - Median
 - Interquartile ranges

```
. tab satisfied_bi
```

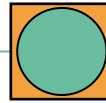
Level of satisfaction (binary)	Freq.	Percent	Cum.
Low satisfaction	235	44.76	44.76
High satisfaction	290	55.24	100.00
Total	525	100.00	

```
-> tabulation of delivery
```

Type of delivery	Freq.	Percent	Cum.
Vaginal	267	50.86	50.86
C-section	258	49.14	100.00
Total	525	100.00	

```
. sum age gestage gravida parity numANC
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	525	30.93524	9.598207	15	49
gestage	525	38.32952	1.534584	36	41
gravida	525	3.693333	1.927912	1	8
parity	525	2.062857	1.506301	0	7
numANC	525	2.952381	1.496695	1	6



Bivariate Analysis

STATA Code:

- `tab work satisfied_bi, col chi`

Working status	Level of satisfaction (binary)		Total
	Low satis	High sati	
Not employed	112 47.66	141 48.62	253 48.19
Working	123 52.34	149 51.38	272 51.81
Total	235 100.00	290 100.00	525 100.00

Pearson $\chi^2(1) = 0.0480$

Pr = 0.827

Covariates

Outcome

Categorical

Chi-square test or Fischer's exact test

Continuous

One-way ANOVA or T-test if assumption of normality holds

STATA Code:

- `oneway age satisfied_bi, tabulate`

Level of satisfaction (binary)	Summary of Age			Freq.
	Mean	Std. Dev.		
Low satis	32.153191	10.064001		235
High sati	29.948276	9.1019265		290
Total	30.935238	9.5982065		525

Source	Analysis of Variance				
	SS	df	MS	F	Prob > F
Between groups	631.088851	1	631.088851	6.93	0.0087
Within groups	47642.7092	523	91.0950464		
Total	48273.7981	524	92.1255689		

Bartlett's test for equal variances: $\chi^2(1) = 2.6199$ Prob> $\chi^2 = 0.106$

Statistical difference

```
. logit satisfied_bi gestage gravida parity numANC i.hiv i.syph delivery age i.education mstatus work i.ethnic
> ity
```

```
Iteration 0: log likelihood = -361.01602
Iteration 1: log likelihood = -315.51075
Iteration 2: log likelihood = -315.33041
Iteration 3: log likelihood = -315.33033
Iteration 4: log likelihood = -315.33033
```

Logistic regression

Log likelihood = **-315.33033**

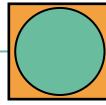
Number of obs	=	525
LR chi2(17)	=	91.37
Prob > chi2	=	0.0000
Pseudo R2	=	0.1265

Model chi-sq tests whether fit is an improvement relative to the null model

McFadden's Pseudo R2 -% or proportion of improvement over the null model

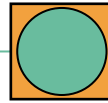
satisfied_bi	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
gestage	.1655198	.0644309	2.57	0.010	.0392376	.291802
gravida	-.1382945	.0602229	-2.30	0.022	-.2563292	-.0202598
parity	-.0795355	.0770031	-1.03	0.302	-.2304587	.0713877
numANC	.3150327	.0687648	4.58	0.000	.1802562	.4498093
hiv						
Tested but no results	.1968685	.2391047	0.82	0.410	-.2717681	.6655051
Tested and received results	-.0872409	.2351991	-0.37	0.711	-.5482226	.3737409
syph						
Tested but no results	-.183586	.2417375	-0.76	0.448	-.6573827	.2902107
Tested and received results	-.4147604	.2391187	-1.73	0.083	-.8834245	.0539036
delivery	-.1717185	.1950158	-0.88	0.379	-.5539425	.2105055
age	-.0126107	.0103747	-1.22	0.224	-.0329446	.0077233
education						
Secondary/higher education	.978413	.1964442	4.98	0.000	.5933896	1.363437
mstatus	.0861951	.1943371	0.44	0.657	-.2946986	.4670888
work	-.0934586	.1940561	-0.48	0.630	-.4738016	.2868844
ethnicity						
Storm	-.618606	.3252055	-1.90	0.057	-1.255997	.0187851
Strong	-.293877	.3263863	-0.90	0.368	-.9335824	.3458284
Twilight	-.3679474	.3156449	-1.17	0.244	-.9866	.2507051
Twinkle	-.3275213	.3083722	-1.06	0.288	-.9319197	.2768771
_cons	-5.732876	2.525066	-2.27	0.023	-10.68191	-.7838375

Logistic Regression: Odds



Odds: Interpretation

Covariate	Coeff	p value	Interpretation
Gestational age	0.1655	0.010	positive and significant predictor of the log odds of having high satisfaction
Gravida	-0.1383	0.022	negative and significant predictor of the log odds of having high satisfaction
# of ANC visits	0.3150	0.000	positive and significant predictor of the log odds of having high satisfaction
Secondary/higher education	0.9784	0.000	positive and significant predictor of the log odds of having high satisfaction



```
. logistic satisfied_bi gestage gravida parity numANC i.hiv i.syph delivery age education mstatus work i.ethnicity  
> city
```

Logistic regression

Log likelihood = **-315.33033**

Number of obs = **525**

LR chi2(17) = **91.37**

Prob > chi2 = **0.0000**

Pseudo R2 = **0.1265**

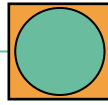
Model chi-sq tests whether fit is an improvement relative to the null model

McFadden's Pseudo R2 -% or proportion of improvement over the null model

satisfied_bi	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
gestage	1.180006	.0760288	2.57	0.010	1.040018	1.338838
gravida	.8708422	.0524446	-2.30	0.022	.7738872	.9799441
parity	.9235452	.0711158	-1.03	0.302	.7941692	1.073998
numANC	1.370304	.0942287	4.58	0.000	1.197524	1.568013
hiv						
Tested but no results	1.217584	.2911301	0.82	0.410	.7620309	1.945473
Tested and received results	.9164563	.2155497	-0.37	0.711	.5779762	1.453161
syph						
Tested but no results	.8322803	.2011933	-0.76	0.448	.5182059	1.336709
Tested and received results	.6604985	.1579375	-1.73	0.083	.4133649	1.055383
delivery	.8422162	.1642455	-0.88	0.379	.5746797	1.234302
age	.9874685	.0102446	-1.22	0.224	.9675922	1.007753
education	2.660231	.5225869	4.98	0.000	1.810114	3.909606
mstatus	1.090019	.2118311	0.44	0.657	.744756	1.595343
work	.9107757	.1767416	-0.48	0.630	.6226308	1.33227
ethnicity						
Storm	.5386949	.1751865	-1.90	0.057	.2847917	1.018963
Strong	.7453681	.243278	-0.90	0.368	.3931428	1.41316
Twilight	.6921536	.2184747	-1.17	0.244	.3728422	1.284931
Twinkle	.7207079	.2222463	-1.06	0.288	.393797	1.319004
_cons	.0032378	.0081755	-2.27	0.023	.000023	.4566503

Note: **_cons** estimates baseline odds.

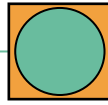
Logistic Regression: Odds Ratio



Odds Ratio: Interpretation

Gestational age

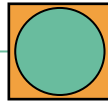
- OR = 1.18
 - Positive association, $OR > 1$
 - Odds of a woman being highly satisfied increased by a factor of 1.18 with every unit increase in the gestational age of the pregnancy.
 - Odds is increasing with each increase in the predictor, gestational age.
 - For every increase in gestational age, the predicted odds increases by 18% $[(1.18-1)*100\% = 18\%]$.
 - The predicted odds are increasing by 18% per unit increase on gestational age.



Odds Ratio: Interpretation

Gravida

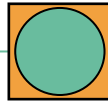
- $OR = 0.87$
 - Negative association, $OR < 1$
 - Odds of a woman being highly satisfied changes by a factor of 0.87 with every unit increase in the gravida of the woman.
 - Odds is decreasing with each increase in the predictor, gravida.
 - For every increase in gravida, the predicted odds of high satisfaction decreases by 13% $[(0.87-1)*100\%=-13\%]$.
 - The predicted odds are decreasing by 13% per unit increase on gravida.



Odds Ratio: Interpretation

Number of ANC visits

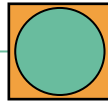
- $OR = 1.37$
 - Positive association, $OR > 1$
 - Odds of a woman being highly satisfied increased by a factor of 1.37 with every unit increase in the number of ANC visits.
 - Odds is increasing with each increase in our predictor, number of ANC visits.
 - For every increase in the number of ANC visits, the predicted odds increases by 37% $[(1.37-1)*100\% = 37\%]$.
 - The predicted odds are increasing by 37% per unit increase on number of ANC visits.



Odds Ratio: Interpretation

Educational level

- $OR = 2.66$
 - Positive association, $OR > 1$
 - Odds of a more educated woman (secondary/higher education) being highly satisfied are 2.7 times the odds of a woman who is less educated.
 - Odds increases with more education compared to those with less education.
 - Women who have a secondary or higher education are more likely to be highly satisfied compared to the women with lower education (primary or no education).
 - For more educated women compared to the less educated, the predicted odds increases by 166% $[(2.66-1)*100\%=166\%]$.
 - The predicted odds are increasing by 166% for more educated compared to the less educated women.



Hosmer-Lemeshow goodness of fit test

- Compares the estimated model to the saturated model (model that has a perfect fit)
 - **Test not significant → model is adequately specified/fit**
 - Test significant → model is misspecified

```
. estat gof, group(10)
```

Logistic model for satisfied_bi, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

number of observations =	525
number of groups =	10
Hosmer-Lemeshow chi2(8) =	5.97
Prob > chi2 =	0.6509

```
. collin gestage gravida parity numANC hiv syph delivery age education mstatus work ethnicity
(obs=525)
```

Collinearity Diagnostics

Variable	VIF	SQRT VIF	Tolerance	R- Squared
gestage	1.02	1.01	0.9786	0.0214
gravida	1.46	1.21	0.6849	0.3151
parity	1.46	1.21	0.6831	0.3169
numANC	1.03	1.02	0.9695	0.0305
hiv	1.02	1.01	0.9839	0.0161
syph	1.05	1.03	0.9502	0.0498
delivery	1.02	1.01	0.9802	0.0198
age	1.06	1.03	0.9393	0.0607
education	1.04	1.02	0.9628	0.0372
mstatus	1.02	1.01	0.9842	0.0158
work	1.01	1.01	0.9893	0.0107
ethnicity	1.01	1.00	0.9918	0.0082
Mean VIF	1.10			

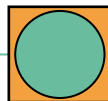
	Eigenval	Cond Index
1	10.3266	1.0000
2	0.5293	4.4169
3	0.4896	4.5926
4	0.4624	4.7257
5	0.3303	5.5917
6	0.1908	7.3564
7	0.1703	7.7866
8	0.1529	8.2192
9	0.1305	8.8953
10	0.0998	10.1731
11	0.0861	10.9519
12	0.0307	18.3455
13	0.0008	115.3394

Condition Number **115.3394**

Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)

Det(correlation matrix) **0.5925**

Collinearity Diagnostics



Overall effect of factor variables

```
. testparm i.syph
```

```
( 1) [satisfied_bi]2.syph = 0  
( 2) [satisfied_bi]3.syph = 0
```

```
chi2( 2) = 3.01  
Prob > chi2 = 0.2217
```

```
. test 2.syph 3.syph
```

```
( 1) [satisfied_bi]2.syph = 0  
( 2) [satisfied_bi]3.syph = 0
```

```
chi2( 2) = 3.01  
Prob > chi2 = 0.2217
```

```
. testparm i.hiv
```

```
( 1) [satisfied_bi]2.hiv = 0  
( 2) [satisfied_bi]3.hiv = 0
```

```
chi2( 2) = 1.45  
Prob > chi2 = 0.4849
```

```
. test 2.hiv 3.hiv
```

```
( 1) [satisfied_bi]2.hiv = 0  
( 2) [satisfied_bi]3.hiv = 0
```

```
chi2( 2) = 1.45  
Prob > chi2 = 0.4849
```

```
. testparm i.ethnicity
```

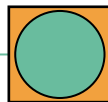
```
( 1) [satisfied_bi]2.ethnicity = 0  
( 2) [satisfied_bi]3.ethnicity = 0  
( 3) [satisfied_bi]4.ethnicity = 0  
( 4) [satisfied_bi]5.ethnicity = 0
```

```
chi2( 4) = 3.68  
Prob > chi2 = 0.4512
```

```
. test 2.ethnicity 3.ethnicity 4.ethnicity 5.ethnicity
```

```
( 1) [satisfied_bi]2.ethnicity = 0  
( 2) [satisfied_bi]3.ethnicity = 0  
( 3) [satisfied_bi]4.ethnicity = 0  
( 4) [satisfied_bi]5.ethnicity = 0
```

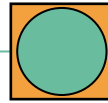
```
chi2( 4) = 3.68  
Prob > chi2 = 0.4512
```



Publication of quality tables in STATA

```
tabout [ varlist ] [ if exp ] [ in range ] [ weight = exp ] using  
filename [ , options ]
```

- Program that produces publication-quality tables of descriptive statistics.
- Used for univariate and bivariate analysis
 - Tabulations of counts and percentages
 - Tabulations of summary tables
 - means, medians, standard errors, etc.
- Free user-written program
 - Version 2 → `ssc install tabout`
 - Version 3 beta → `tabout` home page



Tabulation of Univariate Analysis

Table 1.2. Univariate analysis of independent variables of interest

	Percentage	# observed
Educational level		
No/primary education	45.1429	237
Secondary/higher education	54.8571	288
Marital status		
Married	47.0476	247
Unmarried	52.9524	278
Working status		
Not employed	48.1905	253
Working	51.8095	272
Ethnicity		
Clay	17.7143	93
Storm	18.8571	99
Strong	18.6667	98
Twilight	20.9524	110
Twinkle	23.8095	125
HIV testing and results		
Not tested	34.8571	183
Tested but no results	32.0000	168
Tested and received results	33.1429	174
Syphilis testing and results		
Not tested	34.4762	181
Tested but no results	31.2381	164
Tested and received results	34.2857	180
type of delivery		
Vaginal	50.8571	267
C-section	49.1429	258
IPTOTAL!		
Total	100.0000	525

Source: D4I_Logistic_Tutorial.dta

Table 1.3. Univariate analysis of independent variables of interest (Mean)

	age	gestage	gravida	ndeliver	numANC	Obs
tot						
1	30.9352	38.3295	3.6933	2.0629	2.9524	525
IPTOTAL!						
Total	30.9352	38.3295	3.6933	2.0629	2.9524	525

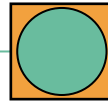
Source: D4I_Logistic_Tutorial.dta

Univariate analysis of independent variables of interest

	Percentage	N
Educational level		
No/primary education	45.1	237
Secondary/higher education	54.9	288
Marital status		
Married	47.0	247
Unmarried	53.0	278
Working status		
Not employed	48.2	253
Working	51.8	272
Ethnicity		
Clay	17.7	93
Storm	18.9	99
Strong	18.7	98
Twilight	21.0	110
Twinkle	23.8	125
HIV testing and results		
Not tested	34.9	183
Tested but no results	32.0	168
Tested and received results	33.1	174
Syphilis testing and results		
Not tested	34.5	181
Tested but no results	31.2	164
Tested and received results	34.3	180
Type of delivery		
Vaginal	50.9	267
C-section	49.1	258
Mean Age (SD)	30.94 (9.60)	525
Mean gestational age (SD)	38.33 (1.53)	525
Mean gravida (SD)	3.69 (1.93)	525
Mean parity (SD)	2.06 (1.51)	525
Mean number of ANC visits (SD)	2.95 (1.50)	525
Total		
	100.0	525

Source: D4I_Logistic_Tutorial.dta

ANC = antenatal care; SD = standard deviation



Tabulation of Bivariate Analysis

Table 1.1. Bivariate analysis

	Level of satisfaction (binary)			# observed
	Low satisfaction-high satisfaction		Total	
	Percentage	Percentage	Percentage	
Educational level				
No/primary education	59.5745	33.4483	45.1429	237
Secondary/higher education	40.4255	66.5517	54.8571	288
Marital status				
Married	46.3830	47.5862	47.0476	247
Unmarried	53.6170	52.4138	52.9524	278
Working status				
Not employed	47.6596	48.6207	48.1905	253
Working	52.3404	51.3793	51.8095	272
Ethnicity				
Clay	14.8936	20.0000	17.7143	93
Storm	22.1277	16.2069	18.8571	99
Strong	18.7234	18.6207	18.6667	98
Twilight	22.1277	20.0000	20.9524	110
Twinkle	22.1277	25.1724	23.8095	125
HIV testing and results				
Not tested	37.0213	33.1034	34.8571	183
Tested but no results	28.5106	34.8276	32.0000	168
Tested and received results	34.4681	32.0690	33.1429	174
Syphilis testing and results				
Not tested	32.7660	35.8621	34.4762	181
Tested but no results	29.7872	32.4138	31.2381	164
Tested and received results	37.4468	31.7241	34.2857	180
type of delivery				
Vaginal	49.7872	51.7241	50.8571	267
C-section	50.2128	48.2759	49.1429	258
IPTOTAL!				
Total	100.0000	100.0000	100.0000	525

Source: D4I_Logistic_Tutorial.dta



Table 1.2. Bivariate analysis (Mean)

	age	gestage	gravida	ndeliver	numANC	Obs
Level of satisfaction (binary)						
Low satisfaction	32.1532	38.1277	4.0681	2.3149	2.5787	235
High satisfaction	29.9483	38.4931	3.3897	1.8586	3.2552	290
IPTOTAL!						
Total	30.9352	38.3295	3.6933	2.0629	2.9524	525

Source: D4I_Logistic_Tutorial.dta

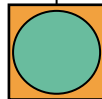
Bivariate Analysis

	Level of satisfaction		
	Low	High	Total
Educational level			
No/primary education	59.6	33.4	45.1
Secondary/higher education	40.4	66.6	54.9
Marital status			
Married	46.4	47.6	47.0
Unmarried	53.6	52.4	53.0
Working status			
Not employed	47.7	48.6	48.2
Working	52.3	51.4	51.8
Ethnicity			
Clay	14.9	20.0	17.7
Storm	22.1	16.2	18.9
Strong	18.7	18.6	18.7
Twilight	22.1	20.0	21.0
Twinkle	22.1	25.2	23.8
HIV testing and results			
Not tested	37.0	33.1	34.9
Tested but no results	28.5	34.8	32.0
Tested and received results	34.5	32.1	33.1
Syphilis testing and results			
Not tested	32.8	35.9	34.5
Tested but no results	29.8	32.4	31.2
Tested and received results	37.4	31.7	34.3
type of delivery			
Vaginal	49.8	51.7	50.9
C-section	50.2	48.3	49.1
Total	100.0	100.0	100.0
Mean (SD)			
Age	32.15 (10.06)	29.95 (9.10)	30.94 (9.60)
Gestational age	38.13 (1.38)	38.49 (1.63)	38.33 (1.53)
Gravida	4.07 (2.10)	3.39 (1.72)	3.69 (1.93)
Parity	2.31 (1.64)	1.86 (1.36)	2.06 (1.51)
Number of ANC visits	2.58 (1.19)	3.26 (1.64)	2.95 (1.50)
N	235	290	525

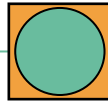
Source: D4I_Logistic_Tutorial.dta

ANC = antenatal care; SD = standard deviation

Logistic Regression Tabulation



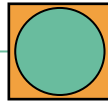
	OR	p value	95% CI
Educational level			
No/primary education	[Ref]		
Secondary/higher education	2.660	0.000	[1.810, 3.910]
Marital status			
Married	[Ref]		
Unmarried	1.090	0.657	[0.745, 1.595]
Working status			
Not employed	[Ref]		
Working	0.911	0.630	[0.623, 1.332]
Ethnicity			
Clay	[Ref]		
Storm	0.539	0.057	[0.285, 1.019]
Strong	0.745	0.368	[0.393, 1.413]
Twilight	0.692	0.224	[0.373, 1.285]
Twinkle	0.721	0.288	[0.394, 1.319]
HIV testing and results			
Not tested	[Ref]		
Tested but no results	1.218	0.410	[0.762, 1.945]
Tested and received results	0.916	0.711	[0.578, 1.453]
Syphilis testing and results			
Not tested	[Ref]		
Tested but no results	0.832	0.448	[0.518, 1.337]
Tested and received results	0.660	0.083	[0.413, 1.055]
Type of delivery			
Vaginal	[Ref]		
C-section	0.842	0.379	[0.575, 1.234]
Age	0.987	0.224	[0.968, 1.008]
Gestational age	1.180	0.010	[1.040, 1.339]
Gravida	0.871	0.022	[0.774, 0.980]
Parity	0.924	0.302	[0.794, 1.074]
Number of ANC visits	1.370	0.000	[1.198, 1.568]
Constant	0.003	0.023	[0.000, 0.457]
Source: D4I_Logistic_Tutorial.dta			
ANC - antenatal care ; SD - standard deviation			



Conclusion

What factors are associated with women's satisfaction with maternity care services?

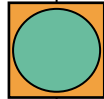
- Findings
 - Gestational age ↑
 - Gravida ↓
 - Number of ANC visits ↑
 - Educational level ↑



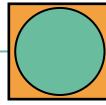
Summary

- Analytical approach used to model dichotomous outcome variables which has extensive use in health care and social sciences
- It predicts the probability of the success of an event or falling into a target group
- It's easy to use and produces robust estimates
- Produces odds
 - Change in expected log odds relative to the unit change in the predictor, holding all other predictors constant
 - Interpret odds for direction of change, and not the amount of change
- Produces odds ratios
 - Change in the odds of success relative to the unit change in the predictor, holding all other predictors constant
 - Values are between 0 and $+\infty$
 - Present the OR and 95% Confidence interval

Resources

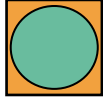


- Background information sheet
 - Overview of logistic regression
 - Useful resources
- PowerPoint presentation
- Sample dataset
- Sample do file



References

- Hair, J., Black, W., Babin, B., & Anderson, R. (2019). Multivariate Data Analysis (8th ed.). Cengage Learning EMEA.
- Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). Applied Logistic Regression (3rd ed.). John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118548387>
- Logistic Regression. UCLA: Statistical Consulting Group. <https://stats.idre.ucla.edu/stata/dae/logistic-regression/>
- Long, J., & Freese, J. (2014). Regression Models for Categorical Dependent Variables using Stata (3rd ed.). Stata Press
- Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables. SAGE.
- Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, 49(12), 1373–1379. [https://doi.org/10.1016/S0895-4356\(96\)00236-3](https://doi.org/10.1016/S0895-4356(96)00236-3)
- Pituch, K., & Stevens, J. (2016). Applied Multivariate Statistics for the Social Sciences (6th ed.). Routledge Taylor and Francis Group. <https://doi.org/10.4324/9781315814919>



This presentation was produced with the support of the United States Agency for International Development (USAID) under the terms of the Data for Impact (D4I) associate award 7200AA18LA00008, which is implemented by the Carolina Population Center at the University of North Carolina at Chapel Hill, in partnership with Palladium International, LLC; ICF Macro, Inc.; John Snow, Inc.; and Tulane University. The views expressed in this publication do not necessarily reflect the views of USAID or the United States government.

www.data4impactproject.org

