Logistic Regression
R-Studio

Eirini Pagkalidou
MSc, Phd
pagalidou@auth.gr

THESSALONIKI 2023

Logistic Regression

- Logistic regression is a GLM used to model a binary categorical variable using continuous and categorical explanatory variables.

- We only need to establish a *link* function that connects y to p.

$$logit(p) = \log\left(\frac{p}{1-p}\right), 0 \leq p \leq 1.$$

- The logistic regression model can be given by the following equation:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_n X_n$$

We assume that relationships are linear on the logistic scale

When to use simple logistic regression

- When we have a binary outcome Y (i.e. yes/no, treated/untreated)

- We have **one** independent variable X that we think it is related to the outcome Y.

The independent variable can be continuous, categorical or ordinal.

We will look at the interpretation of the simple logistic regression in three examples.

Example:
Risk Factors Associated With Low Infant Birth Weight

We want to examine whether several confounders have an effect on the birth of babies with low weight (<2500 grams). For this reason, the data of 189 women was collected, 59 of which had given birth to a baby with a low weight.

The confounders that were taken into account are:

- **Mother's age(AGE)**,
- **Mother's weight at the last menstrual period (LWT)**,
- **Mother's race (RACE, 1=White, 2=Black, 3=Other)**,
- **Smoking during pregnancy (SMOKE, 1= Yes, 0=No)**,
- History of premature births (PTL, 0=zero, 1=one etc),
- History of hypertension (HT, 1= Yes, 0=No),
- Uterus abnormalities (UI, 1= Yes, 0=No),
- Number of visits to the doctor the first trimester of pregnancy (FTV)

How do we use a logistic model in this example?

- Outcome: Baby's low birth weight (LOW)

$$LOW = \begin{cases} 0, & baby\ with\ a\ birth\ weight\ of\ 2500\ grams\ or\ more \\ 1, & baby\ with\ a\ birth\ weight\ less\ than\ 2500\ grams \end{cases}$$

- Explanatory variable: Mother's weight at the last menstrual cycle (LWT) (continuous variable)

- Model: We have the following logistic model equation:

$$logit(odds\ of\ LOW = 1) = \beta_0 + \beta_1 LWT$$

Results and Interpretation

```
Coefficients:
                  Estimate Std. Error z value  p-value
(Intercept)        1.02328    0.79043   1.295   0.1955
LWT               -0.02842    0.01239  -2.295   0.0218
```

β_1

- The intercept (β_0=1.023) is the estimated log odds of LOW for mothers whose weight is 0. (sometimes is not quite meaningful)
- The estimated coefficient (β_1= -0.028) of LWT is negative. β_1 is the estimated change in the log odds of LOW for one kg increase in LWT.
- To convert these values to odds (OR) we take the exponential value of log odds.
- So, the OR for β_1 is $e^{-0.02842} = 0.9719$.
- This means that the odds that baby is born with a low weight are reduced by about 2.8% as mother's weight increases by one kg ((0.9719-1)x100).
- p-value= 0.0218, 95%CI: (0.9471, 0.9944)
- To express the OR for every 10 kg increase in mother's weight raise the odds to the power of 10.
- $0.97198^{10} = 0.7526$
- The probability that a baby will be born with a low weight is reduced by about 25% for every 10 kg increase in mother's weight.

Example 2: Explanatory variable with two categories: Baby's low birth weight and mother's smoking status during pregnancy

Variables in the model:

- Outcome: Baby's low birth weight (LOW)

$$LOW = \begin{cases} 0, & baby\ with\ a\ birth\ weight\ of\ 2500\ grams\ or\ more \\ 1, & baby\ with\ a\ birth\ weight\ less\ than\ 2500\ grams \end{cases}$$

- Explanatory variable: Smoking status during pregnancy (SMOKE).

$$SMOKE = \begin{cases} 0, & no \\ 1, & yes \end{cases}$$

We consider the groups LOW=0 and SMOKE=0 as the **reference** groups.

- Model: We have the following logistic model equation:

$$logit(odds\ of\ LOW = 1) = \beta_0 + \beta_1 SMOKE$$

Results and Interpretation

```
Coefficients:
              Estimate Std. Error z value  p-value
(Intercept)   -1.0871   0.2147   -5.062   4.14e-07
SMOKE          0.7041   0.3196    2.203   0.0276
```

β_1

- β_1=0.704 is positive, so low birth weight is positively associated with smoking during pregnancy.
- OR=exp(β_1)= 2.021: the odds that a baby is born with low weight are almost two times higher for smokers than for non-smokers.
- p-value=0.027, 95%CI: (1.082, 3.800)

Chi-square test

- Chi-square test can be considered as a special case of logistic regression where both dependent and independent variables are binary.

		LOW	
		Yes	No
SMOKE	Yes	30	44
	No	29	86

- $\chi^2=4.923$, df=1, p-value=0.0264
- $OR=(30/44)\big/(29/86)=2.021$

Example 3: Categorical explanatory variable with more than two categories: Baby's low birth weight and mother's race

Variables in the model:

- Outcome: Baby's low birth weight (LOW)

$$LOW = \begin{cases} 0, & baby\ with\ a\ birth\ weight\ of\ 2500\ grams\ or\ more \\ 1, & baby\ with\ a\ birth\ weight\ less\ than\ 2500\ grams \end{cases}$$

- Explanatory variable: Mother's race (RACE).

$$RACE = \begin{cases} 1, & white \\ 2, & black \\ 3, & other \end{cases}$$

- Model: We have the following logistic model equation:

$$logit(odds\ of\ LOW = 1) = \beta_0 + \beta_1 RACE$$

Results and Interpretation

```
Coefficients:
               Estimate Std. Error  z value  p-value
(Intercept)    -1.1550    0.2391   -4.830   1.36e-06
RACE.Black      0.8448    0.4634    1.823   0.0683
RACE.Other      0.6362    0.3478    1.829   0.0674
```

- Black mothers:
  - OR=exp(0.8448)=2.3257, p-value=0.068, 95%CI: (0.9255, 5.7746)
- Mothers with other race:
  - OR=exp(0.6362)= 1.8892, p-value=0.0674, 95%CI: (0.9565, 3.7578)

Multiple Logistic Regression

- We use multiple logistic regression when we have a binary outcome and two or more explanatory variables.

- We want to investigate how the explanatory variables affect the binary outcome.

  - Explanatory variables can be continuous, categorical or ordinal.

How many explanatory variables can we include in the model?

A minimum of 10 **events** per explanatory variable; where **event** denotes the cases belonging to the less frequent category in the dependent variable.

In our example, the data of 189 women were collected, 59 of which had given birth to a baby with a low weight. The logistic regression model could reasonably accommodate, at most, six (59/10) independent variables (since 59 are the fewest event in the outcome).

```
> table(lowbwt$LOW)

 No Yes
130  59
```

Example: Risk Factors Associated With Low Infant Birth Weight

- We would like to see if any of the variables (AGE, LWT, RACE, SMOKE) have an effect on low birth weight (LOW).

- Firstly, we perform a separate univariate logistic regression for each of the explanatory variables.
  - variables that have a $p<0.2$ in the univariate analysis will be included in the multivariable model.

Univariate analysis results

Variable Name	OR (95% CI)	P-value
LWT	0.97 (0.95,0.99)	0.021
RACE – (Black/White) RACE – (Other/White)	2.33 (0.93,5.77) 1.89 (0.96,3.76)	0.068 0.067
SMOKE (yes/no)	2.02 (1.08,3.80)	0.027
AGE	0.95 (0.89,1.01)	0.105

Multicollinearity Diagnostics

Same as in linear regression:

- We have,

	GVIF	Df	GVIF^(1/(2*Df))
LWT	1.128124	1	1.062132
AGE	1.051659	1	1.025504
SMOKE	1.302165	1	1.141125
RACE	1.461758	2	1.099560

All variables have a quite low VIF

Model Fit
Likelihood Ratio Test and ANOVA test

- Both tests are equivalent.
- This test asks whether the model with predictors fits significantly better than a model with fewer predictors (**only makes sense for nested models**).

Full model: LOW~RACE+SMOKE+AGE+LWT
Reduced model: LOW~RACE+SMOKE+LWT

```
Likelihood ratio test

Model 1: LOW ~ LWT + SMOKE + RACE
Model 2: LOW ~ LWT + AGE + SMOKE + RACE
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   5 -107.45
2   6 -107.24  1 0.4279      0.513
```

```
Analysis of Deviance Table

Model 1: LOW ~ LWT + SMOKE + RACE
Model 2: LOW ~ LWT + AGE + SMOKE + RACE
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       184     214.91
2       183     214.48  1  0.42788    0.513
```

This means that adding parameter AGE to the model did not lead to a significantly improved fit over the model 1.

Model Fit AIC

- It's useful for comparing models
- Can be used for comparing non-nested models
- We select the model that has the smallest AIC

- Full model AIC=226.48
- **Reduced model AIC=224.9**

Final Results

Variables	Univariate Analysis			Multivariable Analysis		
	OR	95%CI	p-value	OR	95% CI	p-value
Age in years	0.95	0.89, 1.01	0.105	0.98	0.91, 1.04	0.515
Weight in Kg	0.97	0.95, 0.99	0.022	0.97	0.95, 0.99	0.047
Race						
Black/White	2.33	0.94, 5.77	0.068	3.44	1.25, 9.67	0.017
Other/White	1.89	0.96, 3.74	0.067	2.57	1.15, 5.94	0.023
Smoking (Yes/No)	2.02	1.08, 3.78	0.027	2.87	1.38, 6.18	0.006

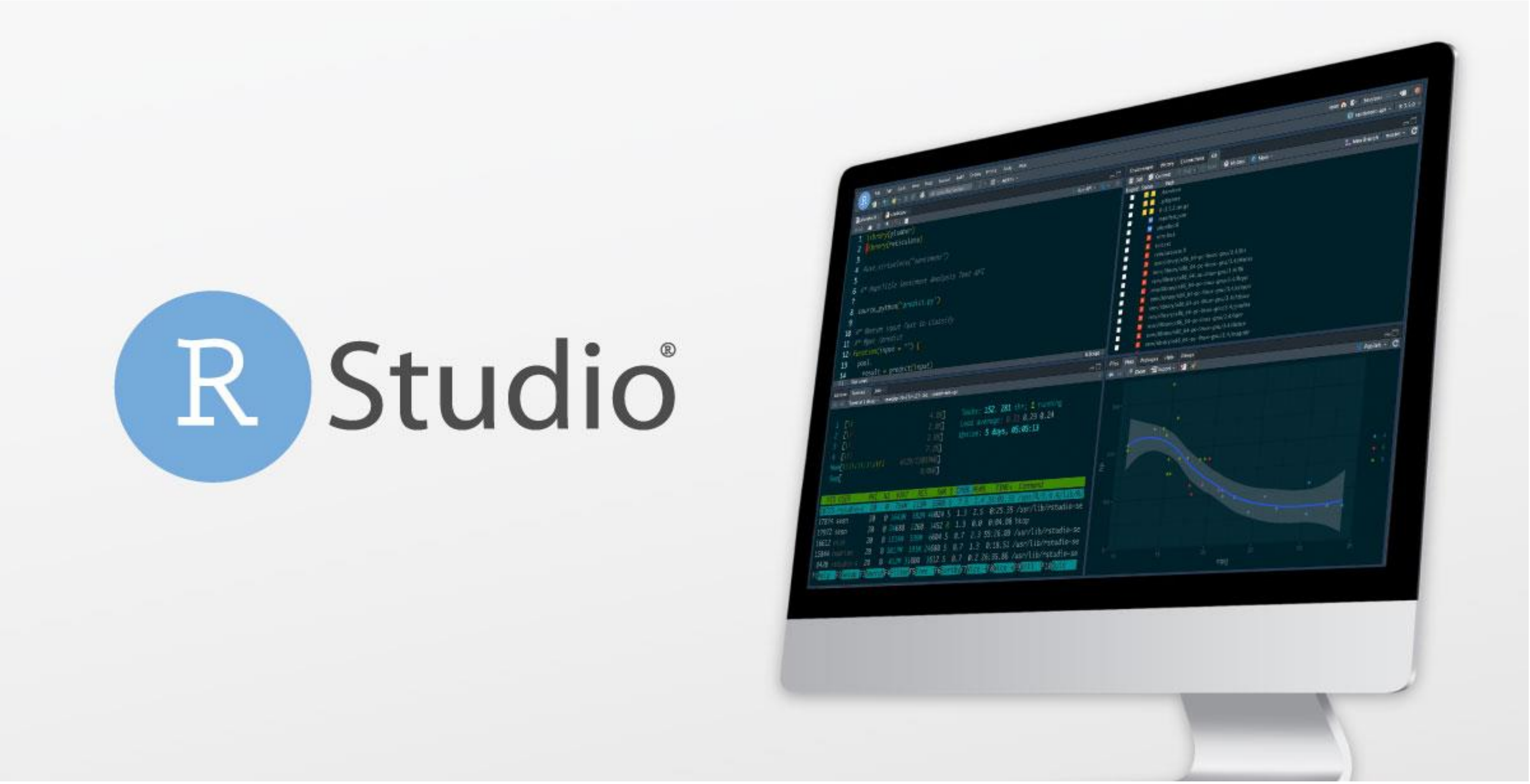
OR:Odds Ratio, CI: Confidence Interval

The interpretation of the variables is similar to simple logistic regression

For example,

“Black” mothers are 3.4 (p=0.017) times more likely to have a baby with a low weight than white mothers adjusted for all the other variables in the model.

Mothers of “other” race are 2.6 (p=0.023) times more likely to have a baby with a low weight than white mothers adjusted for all the other variables in the model.


RStudio