



ARISTOTLE UNIVERSITY
OF THESSALONIKI

FACULTY OF HEALTH SCIENCES - SCHOOL OF MEDICINE
MSc Health Statistics and Data Analytics

Logistic Regression

Eleni Verykoui, PhD

Biostatistician

**Research Associate of the Laboratory of Hygiene,
Social-Preventive Medicine and Medical Statistics, AUTH
everykoui@auth.gr**



THESSALONIKI 2021-22

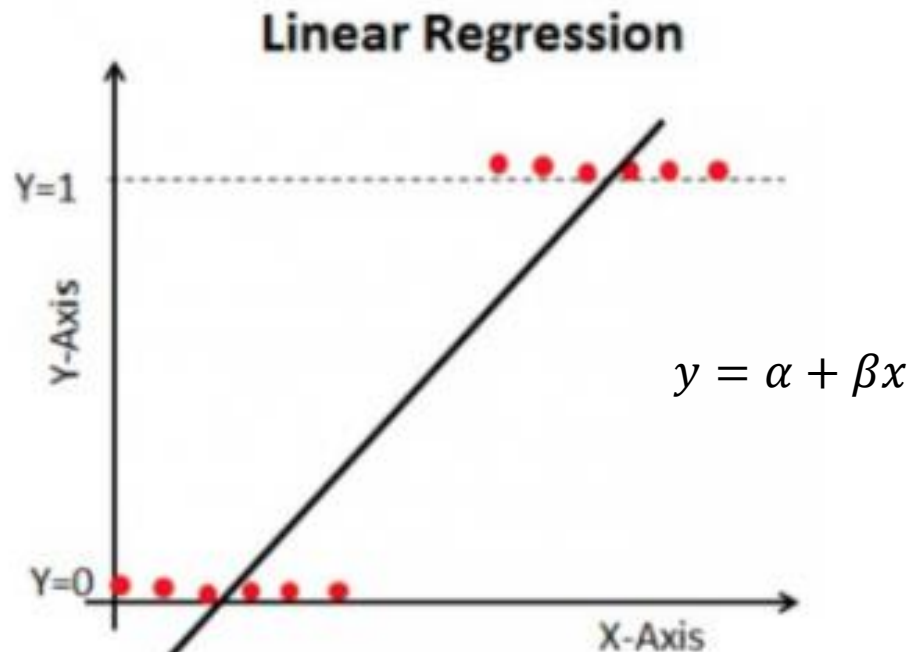


Intro

Suppose we have a binary outcome variable y (i.e., patients with/without disease) and a continuous explanatory variable x .

What model to use to describe their relationship?

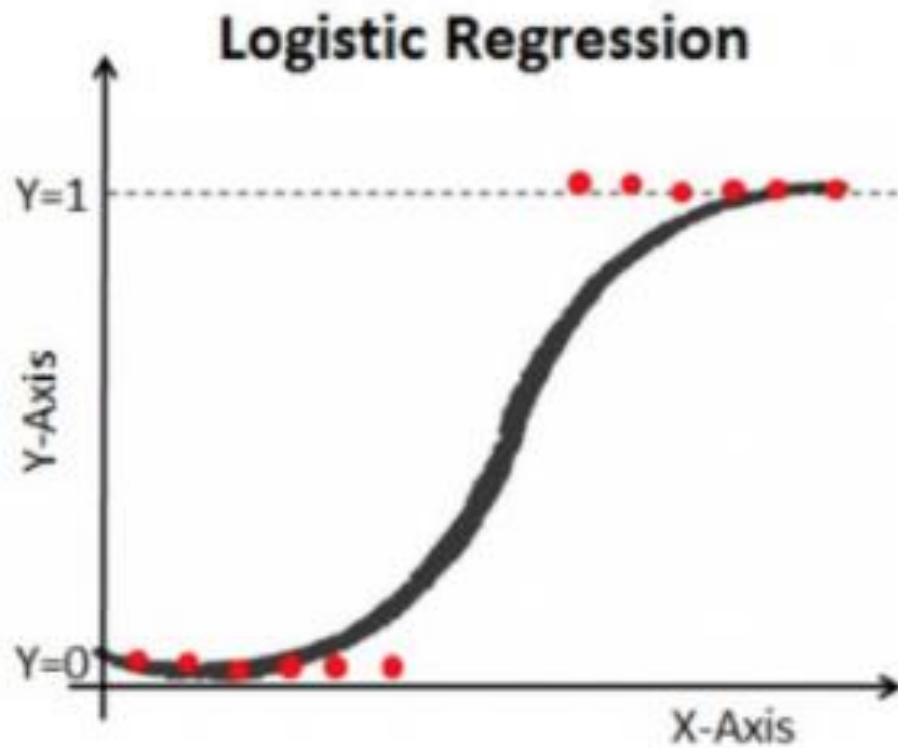
Attempt 1: Use Linear regression



<https://odsc.medium.com>

Linear regression predicts probabilities less than zero or greater than one.

Attempt 2: Use Logistic Regression



When to use Logistic Regression

Logistic regression belongs to a larger model family called *Generalized Linear Models (GLM)*.

All generalized linear models have the following three characteristics:

- A probability distribution describing the outcome variable
- A linear model $y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$
- A link function that relates the linear model to the parameter of the outcome distribution

$$g(p) = y \text{ or } p = g^{-1}(y)$$

Logistic Regression

- Logistic regression is a GLM used to model a binary categorical variable using continuous and categorical explanatory variables.
- We assume that the outcome variable is produced by a binomial distribution and we want to model p the probability of success for a given set of explanatory variables.

We only need to establish a *link* function that connects y to p .
The most commonly used is the logit function.

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

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

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

- We assume that relationships are linear on the logistic scale.

Linear vs. Logistic Regression

Linear Regression	Logistic Regression
Response is a continuous variable	Response is a binary variable
Requires a linear relationship among dependent and independent variable	Does not require a linear relationship between dependent and independent variable
Used to solve regression problems	Used to solve classification problems
Estimates the dependent variable when there is a change in the independent variable	Calculates the probability of an event occurring

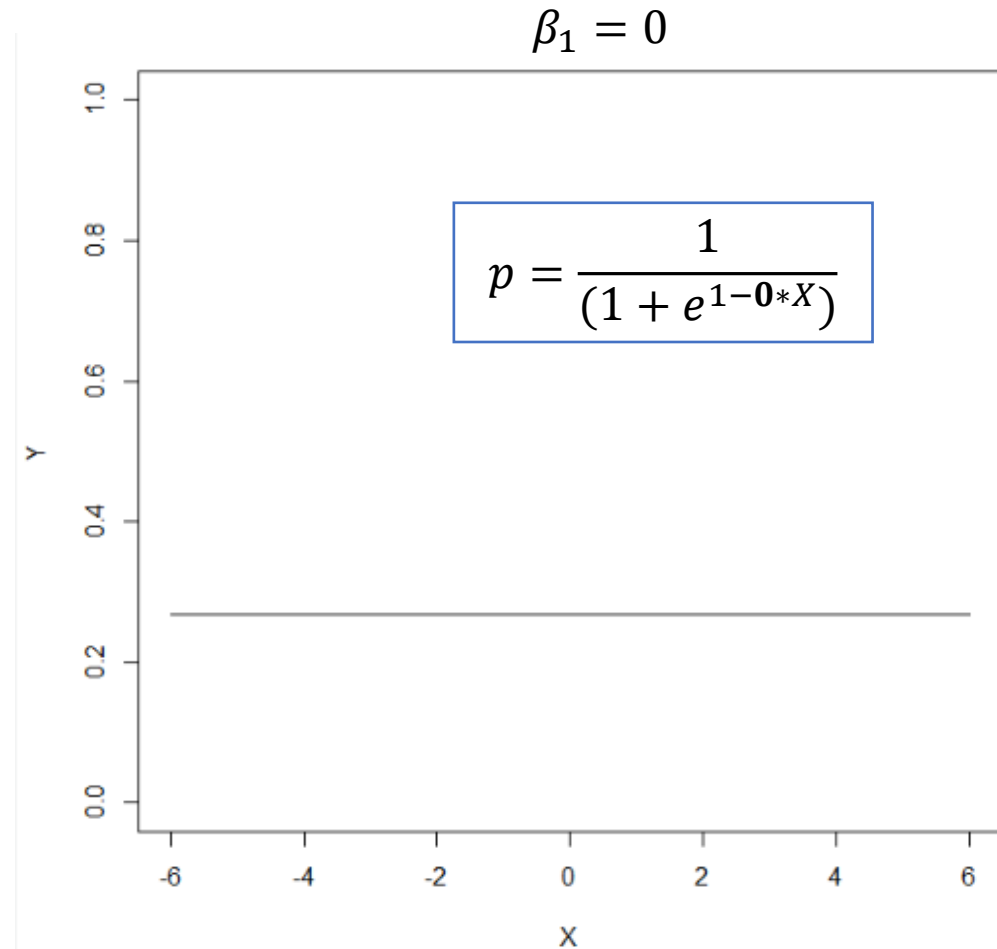
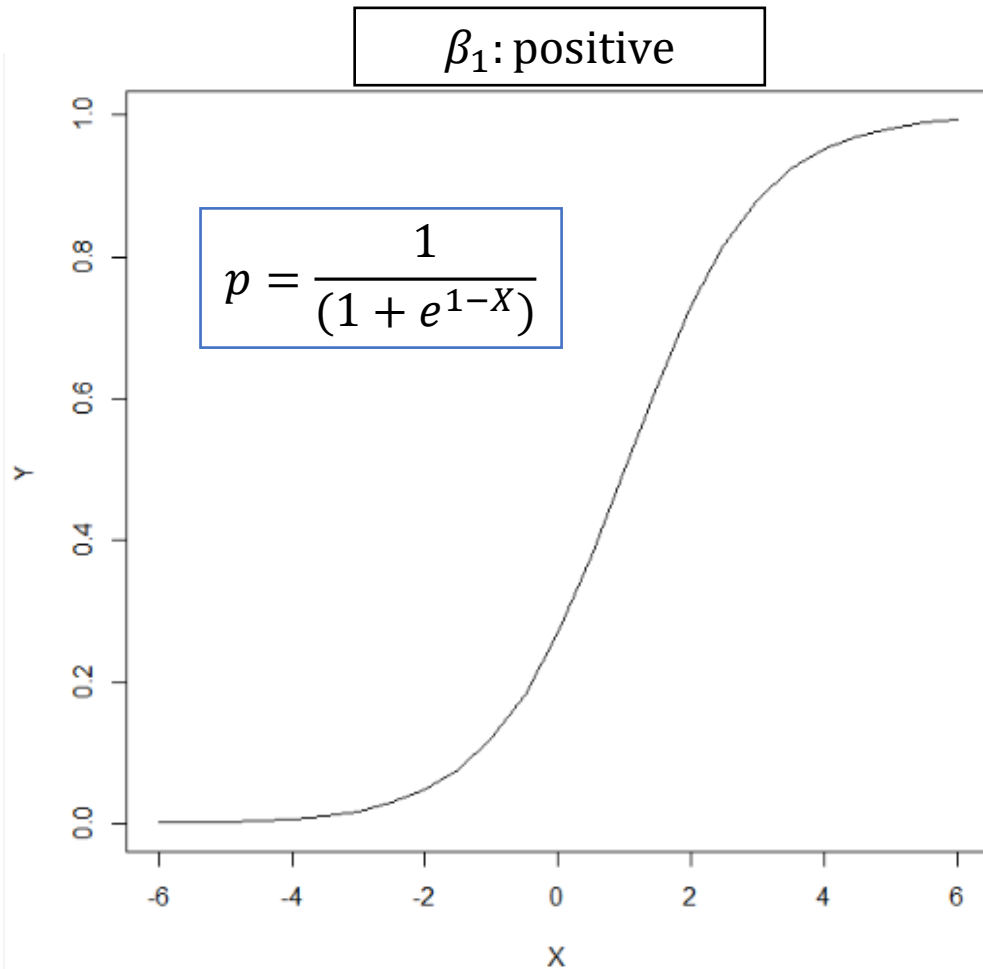
Logistic regression assumptions

When we employ logistic regression, we assume that:

- The outcome is a binary/dichotomous variable.
- The observations are independent.
- There is a linear relationship between the logit of the outcome and each predictor variables.
- There are no extreme values or influential values in the continuous predictors.
- There is no multicollinearity among the predictors.

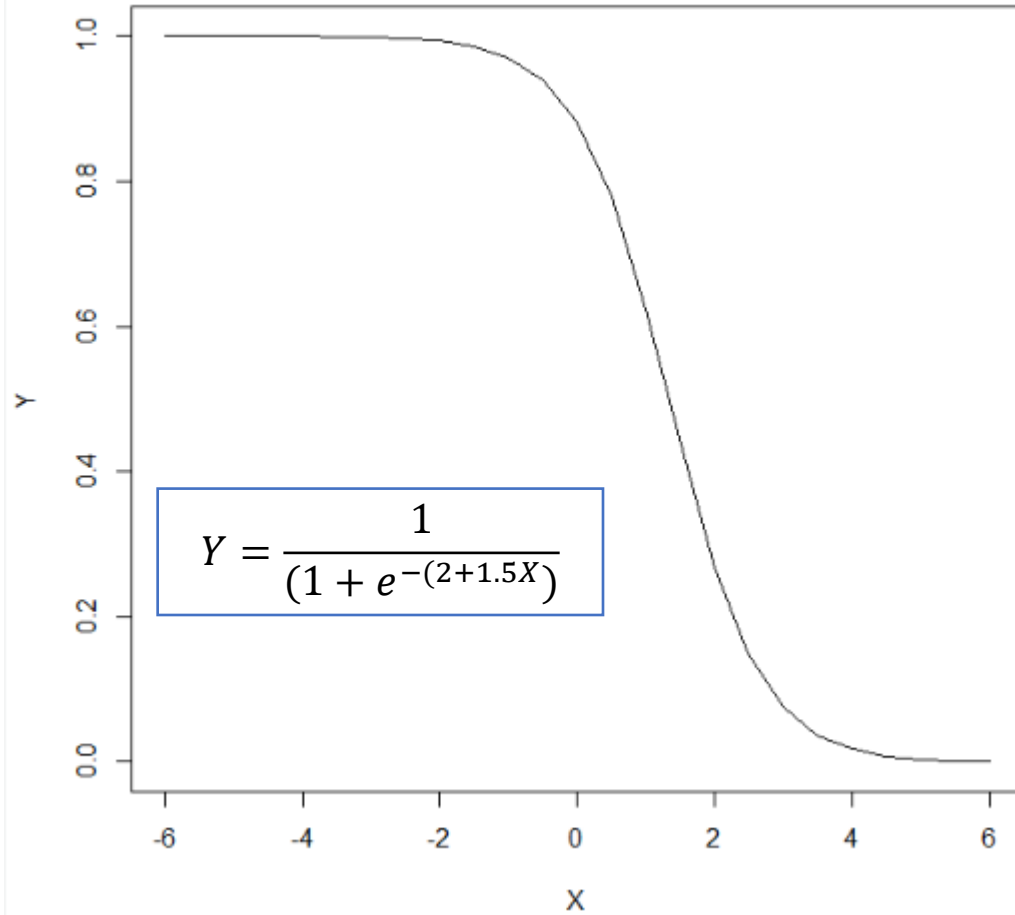
Graphical Representation of Logistic Regression (1)

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 \Rightarrow p = \frac{e^{\beta_0 + \beta_1 X_1}}{(1 + e^{\beta_0 + \beta_1 X_1})} \text{ or } p = \frac{1}{(1 + e^{\beta_0 + \beta_1 X_1})}$$

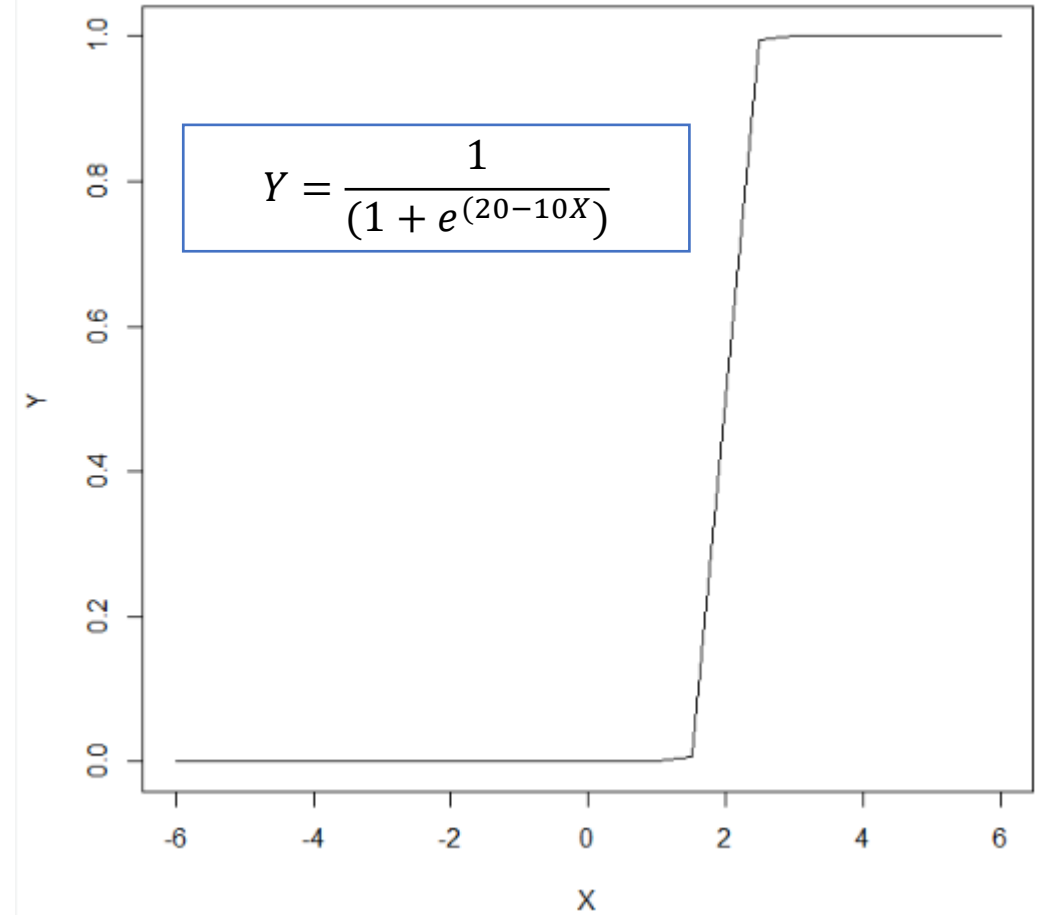


Graphical Representation of Logistic Regression (2)

β_1 : negative



β_1 : very large



Simple Logistic Regression

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 wish 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 considered 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)

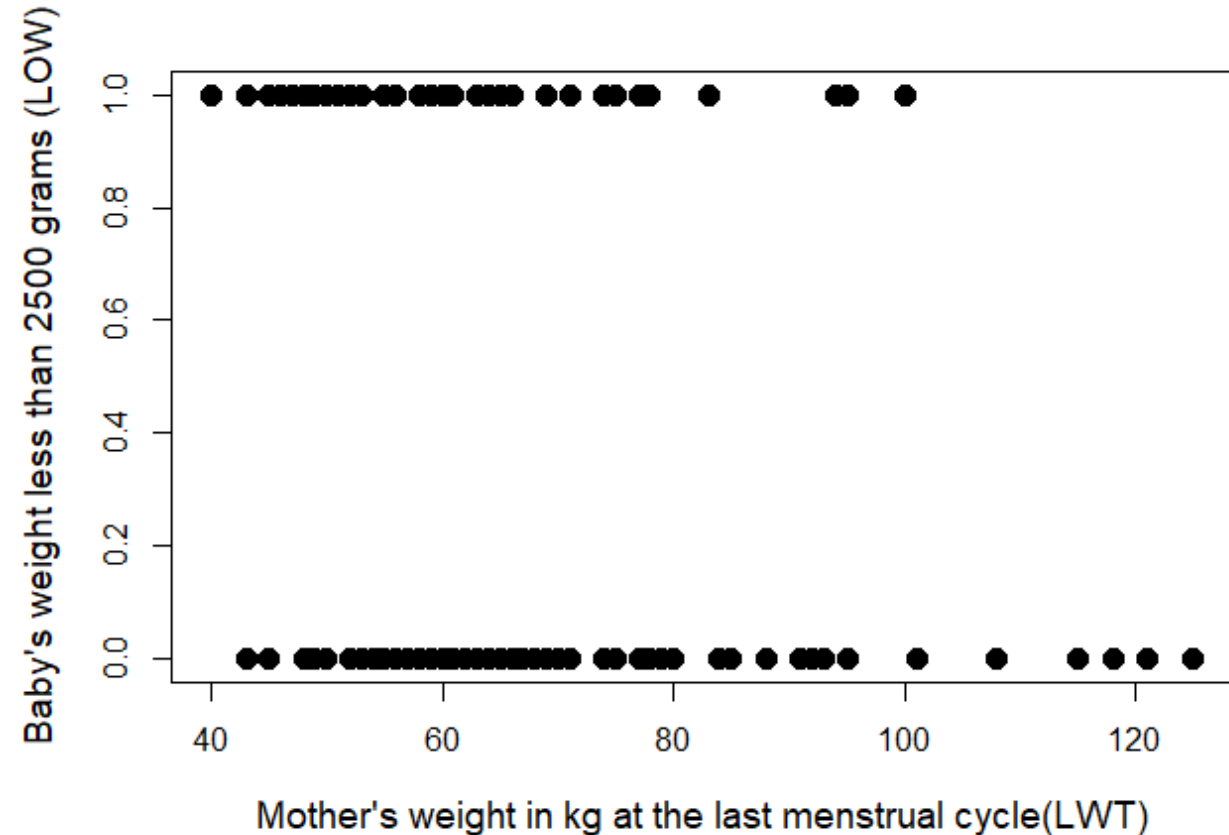
Example 1:

Simple logistic regression with a continuous explanatory variable

Baby's low birth weight and mother's weight in kg at the last menstrual period.

This figure shows low birth weight (variable LOW) (y axis), as a function of their mother's weight at her last menstrual period (LWT) (x axis).

Babies are coded as 1 or 0 depending on whether their birth weight was less than 2500 grams or not respectively.



How do we use a logistic model in this example?

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

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

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

Model

We have the following logistic model equation:

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

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



We can notice that:

- The intercept ($\beta_0=1.023$) is the estimated log odds of LOW for mothers whose weight is 0. (sometimes is not quite meaningful)
- The estimated coefficient ($\beta_1= -0.028$) 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 $((0.9719-1) \times 100)$.
- 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 is 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)

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

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

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

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

Model

We have the following logistic model equation:


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

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



- $\beta_1=0.704$ is positive, so low birth weight is positively associated with smoking during pregnancy.
- OR=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	
		No	Yes
SMOKE	No	86	29
	Yes	44	30

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

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)

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

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

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

Model

We have the following logistic model equation:

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

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

Ordinal explanatory variables can be treated either as continuous or as categorical unordered categories. In the former case we can make assumptions about the differences between the scale items and in the latter we just throw the information about the ordering.

Multiple Logistic Regression

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.

- For example, in a sepsis mortality study, assume that 30 patients died and 50 patients lived. The logistic regression model could reasonably accommodate, at most, three independent variables (since 30 are the fewest event in the outcome).

Example:

Risk Factors Associated With Low Infant Birth Weight

- We would like to see if any of the variables (AGE, LWT, RACE, SMOKE, PTL, HT, UI, FTV) 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
LWT	0.971(0.947, 0.994)	0.021
RACE – Black	2.327 (0.925, 5.774)	0.068
RACE - Other	1.889 (0.9565, 3.7578)	0.067
SMOKE	2.021 (1.082, 3.800)	0.027
AGE	0.950 (0.891, 1.009)	0.105
HT	3.365 (1.028, 11.829)	0.046
UI	2.577 (1.133, 5.881)	0.023
FTV	0.873 (0.632, 1.174)	0.388

Results when some of the variables are included in the model

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	OR	95%CI
(Intercept)	0.38306	0.92870	0.412	0.68000	1.4667	(0.2511, 9.7353)
RACEblack	1.29140	0.52202	2.474	0.01337	3.6378	(1.3115, 10.3313)
RACEother	0.94437	0.42348	2.230	0.02574	2.5712	(1.1365, 6.0364)
SMOKE	1.07118	0.38774	2.763	0.00573	2.9188	(1.3858, 6.3945)
HT	1.75649	0.69180	2.539	0.01112	5.7920	(1.5488, 24.6577)
LWT	-0.03621	0.01364	-2.654	0.00794	0.9644	(0.9373, 0.9890)

Final Model

Coefficients:

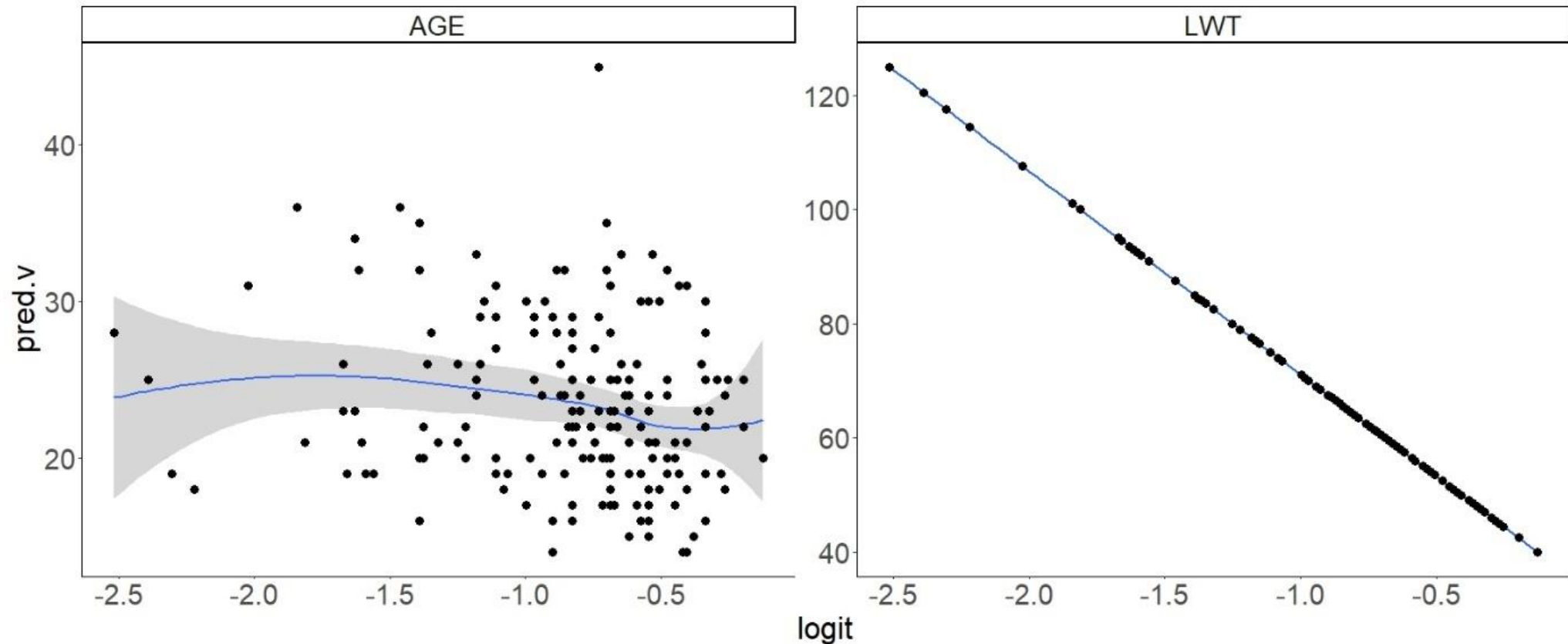
	Estimate	Std. Error	z value	Pr(> z)	OR	95%CI
(Intercept)	0.38306	0.92870	0.412	0.68000	1.4667	(0.2511, 9.7353)
RACEblack	1.29140	0.52202	2.474	0.01337	3.6378	(1.3115, 10.3313)
RACEother	0.94437	0.42348	2.230	0.02574	2.5712	(1.1365, 6.0364)
SMOKE	1.07118	0.38774	2.763	0.00573	2.9188	(1.3858, 6.3945)
HT	1.75649	0.69180	2.539	0.01112	5.7920	(1.5488, 24.6577)
LWT	-0.03621	0.01364	-2.654	0.00794	0.9644	(0.9373, 0.9890)

Interpretation

- The interpretation of the variables is similar to simple logistic regression
- For example,
 - “Black” mothers are 3.6 ($p=0.013$) 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 2.5 ($p=0.025$) times more likely to have a baby with a low weight than white mothers adjusted for all the other variables in the model.

Model Diagnostics – checking assumptions

Linearity Assumption

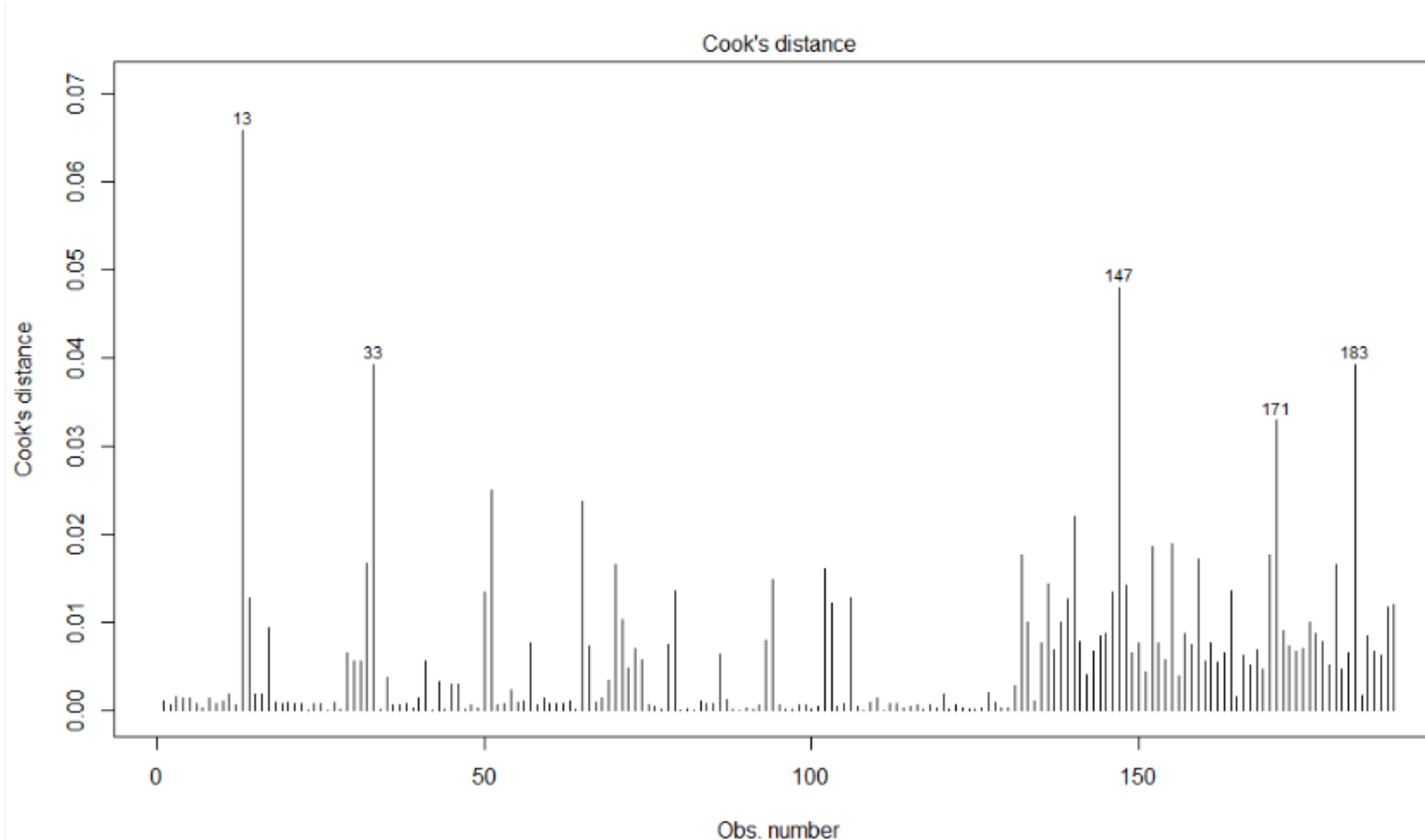


AGE and LWT are all quite linearly associated with the outcome LOW in logit scale.

Model Diagnostics – checking assumptions

Influential Values

Cook's Distance

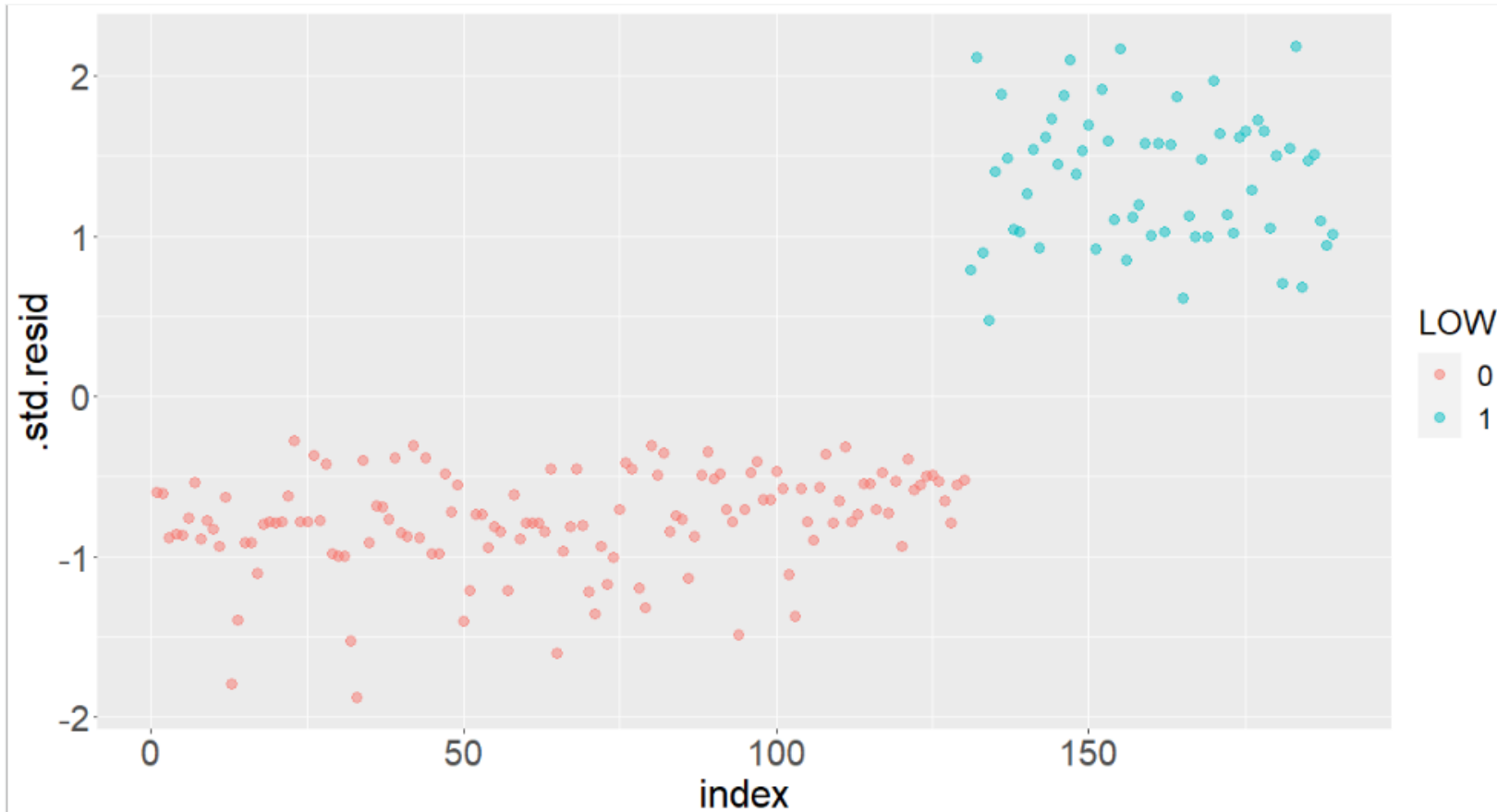


There are 5 observations that might influence the model.

Model Diagnostics – checking assumptions

Influential Values

Check Standardized Residuals



Standardized residuals above 3 represent possible outliers and may deserve closer attention. There are no standardized residuals > 3 in our case.

Multicollinearity Diagnostics

Same as in linear regression:

- We have,

	GVIF	Df		$GVIF^{(1 / (2 * Df))}$
RACE	1.429498	2		1.093442
SMOKE	1.309420	1		1.144299
HT	1.134127	1		1.064954
LWT	1.233284	1		1.110533

All variables have a quite low IVF

Model Fit

Likelihood Ratio Test and ANOVA test

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

Final model: LOW~RACE+SMOKE+HT+LWT

Reduced model: LOW~RACE+SMOKE+LWT

Likelihood Ratio test

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	5	-107.45			
2	6	-104.05	1	6.8117	0.009056 **

ANOVA TEST

Model 1: LOW ~ RACE + SMOKE + LWT

Model 2: LOW ~ RACE + SMOKE + HT + LWT

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	184	214.91			
2	183	208.10	1	6.8117	0.009056 ***

anova(Final model, Reduced model, test="Chisq")

Reduced model does not fit better than Final model

Hosmer and Lemeshow test can be also found in the literature as a goodness of fit test for logistic regression but is not recommended anymore.

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

Model Fit

Nagelkerke's pseudo R-squared

- R-squared cannot be defined for logistic regression but there are pseudo R-squared measures instead.
- A pseudo R-squared cannot be interpreted in the same way as R-squared.
- Pseudo R-squared measures are relative measures among similar models indicating how well the model explains the data.

Stepwise Logistic Regression

Backward Selection

We can also use automatic procedures like *Forward Selection* or *Backward Selection*, which will carry out the choice of the predictive variables.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	OR	95%CI
(Intercept)	0.38306	0.92870	0.412	0.68000	1.4667	(0.2511, 9.7353)
RACEBlack	1.29140	0.52202	2.474	0.01337	3.6378	(1.3115, 10.3313)
RACEOther	0.94437	0.42348	2.230	0.02574	2.5712	(1.1365, 6.0364)
SMOKE	1.07118	0.38774	2.763	0.00573	2.9188	(1.3858, 6.3945)
HT	1.75649	0.69180	2.539	0.01112	5.7920	(1.5488, 24.6577)
LWT	-0.03621	0.01364	-2.654	0.00794	0.9644	(0.9373, 0.9890)

Final Results

Variables	Univariate Analysis			Multivariable Analysis ^a		
	OR	95%CI	p-value	OR	95% CI	p-value
Age	0.95	0.89, 1.01	0.105			
Weight (for every 10 kg increase)	0.75	0.59 , 0.96	0.021	0.66	0.52 , 0.95	0.007
Race						
Black/White	2.33	0.94 , 5.77	0.068	3.63	1.31 , 10.33	0.013
Other/White	1.89	0.96 , 3.74	0.067	2.57	1.13 , 6.03	0.026
Smoking (Yes / No)	2.02	1.08 , 3.78	0.027	2.91	1.38 , 6.39	0.005
History of premature births	4.32	1.92 , 9.73	0.011			
Hypertension (Yes / No)	3.37	1.02 , 11.09	0.046	5.79	1.54 , 24.65	0.011
Uterus abnormalities (Yes / No)	2.58	1.14 , 5.84	0.023			
Number of visits to the doctor the first trimester of pregnancy	0.87	0.64 , 1.19	0.388			

OR:OddsRatio, CI: Confidence Interval

^aUsing the Backward Elimination

Special Case

Complete/Quasi Complete Separation

- Complete separation or quasi-complete separation occurs when a linear combination of the predictors yields a perfect prediction of the response variable for all or the most values of the predictors.
- When we have an unexpectedly large OR with an infinite 95%CI we suspect that there might be a complete or quasi-complete separation in the data.
- The simplest way to check is by using a two-way table between the dependent and the independent variable.

Note: We don't report OR and 95%CI in such cases!

Special Case

Complete/Quasi Complete Separation

What to do?

We usually cannot do much to fix the problem. Some remedies (depending on the independent variable) are:

- Increasing the number of observations.
- Consider what the separation means. Complete separation and quasi-complete separation can indicate important relationships.
- Consider an alternative model. – this applies when a complete separation or a quasi-complete separation is detected in a multivariate model.
- Check to see whether you can combine categories in problematic variables.

Thank you!