Logistic Regression

Aim

To describe the relationship between a binary response variable and one or more continuous or categorical explanatory variables.

Objectives

By the end of this session, you will be able to:

- Understand the basic ideas behind modeling categorical data using binary logistic regression.
- Understand how to fit the model and interpret the parameter estimates, especially in terms of odds and odd ratios.

Logistic Regression Models

Logistic regression models are part of a larger model family called *Generalized Linear Models (GLM).* GLMs are an extension of the traditional linear regression models where the response variable is discrete, and the error terms (residuals) do not follow a normal distribution.

Logistic regression can be considered analogous to linear regression with the difference that the outcome here is binary (i.e. yes/no, disease/no disease, alive/dead) whether or not the subject has a particular characteristic such as a disease. We want a regression equation that will predict the proportion of individuals who have the characteristic or, equivalently, estimate the probability that the individual will have the disease.

In a logistic regression model we model the probability that Y takes the value of 1, when a characteristic is present, as a function of the explanatory variables $X_1, X_2, \ldots, X_k$, where k is the number of explanatory variables. We denote as p the probability that $Y = 1$ given the explanatory variables $X_1, X_2, \ldots, X_k$.

Here, we cannot use the ordinary regression equation because this might result in predicted probabilities less than zero or greater than one.

To overcome this problem, we use the logit function. The logit function is the natural log of the odds that Y equals one of the two categories.

$$logit(p) = \log\left(\frac{p}{1-p}\right)$$

Odds can be defined as the ratio of favorable to unfavorable cases. If the probability that, say, $Y = 1$ equals one, then the odds are one-to-one. If the probability is 1/3 then the odds are one-to-two, etc.

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

$$logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

We, thus, assume that relationships are linear on the logistic scale. The estimation of the logistic regression equation coefficients is computer intensive, and the coefficients are given as log odds ratios (OR).

Assumptions for Logistic Regression

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.

Simple Logistic Regression

We use simple (or univariate) logistic regression when we have a binary outcome (the two groups are usually coded as 0 and 1) and one explanatory variable that we consider to be related to the outcome. The explanatory variable can be continuous or categorical, the latter nominal or ordinal. In logistic regression we usually take the group coded as 0 as the reference group.

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

Example - Risk Factors Associated With Low Infant Birth Weight

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

In order to find the appropriate model to fit our data, we can draw a scatterplot of the response variable against an explanatory variable. If this is a linear shape we can model the relationship with a linear regression model. If the outcome variable is dichotomous this graphical approach is not very helpful.

In this example the response variable is variable LOW which takes value 0 if a baby was not born with a low birthweight and the value 1 if a baby was born with a low birthweight.

We explain more in the following example.

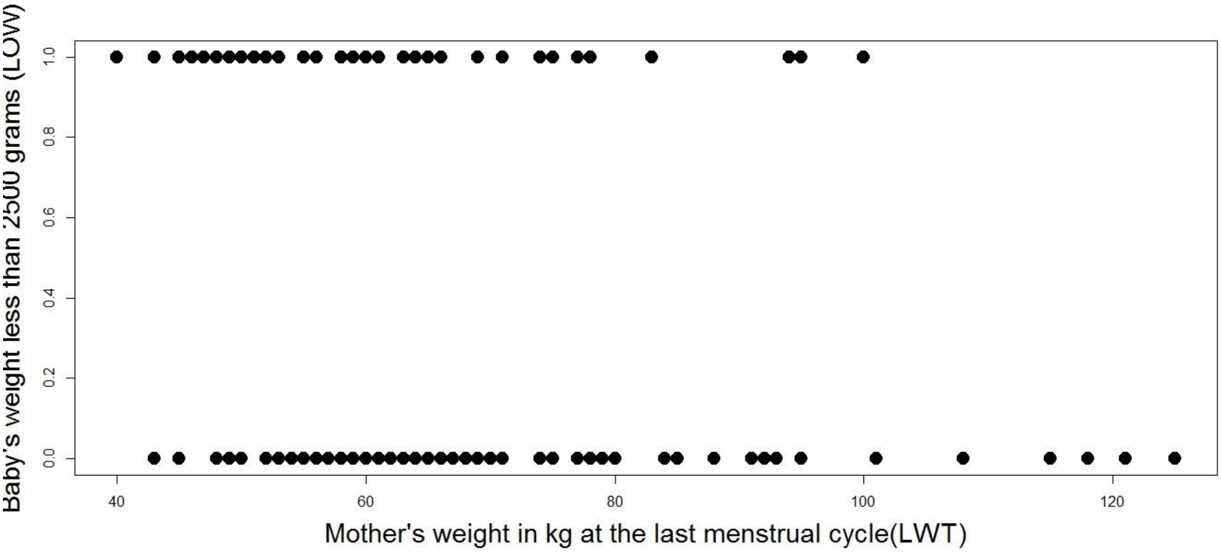
Example 1: Simple logistic regression with a continuous explanatory variable:

Baby’s low birthweight and mother’s weight in kg at the last menstrual period.

We would like to check whether the mother’s weight at the last menstrual cycle affects whether the baby will be born weighing less than 2500 grams.

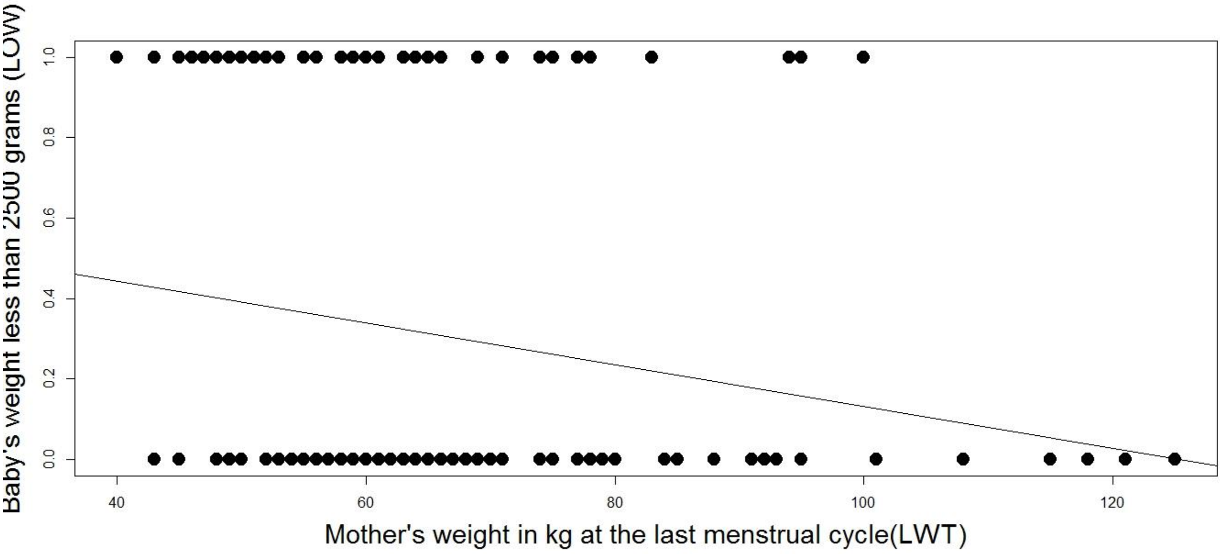
The following figure (Figure 1) shows the sample of 189 babies with a low birthweight (variable LOW) as a function of their mother’s weight at their last menstrual period (LWT). Babies are coded as 1 or 0 depending on whether their birthweight was less than 2500 grams or not, respectively.

Figure 1



Note that fitting a linear regression model in this case does not make sense since the response variable LOW can only take the values 0 and 1. Thus, if, for example, we extrapolate the regression line in the plot below it seems that that the variable LOW can also take negative values which is not possible under this context (Figure 2).

Figure 2



Logistic regression will allow for the estimation of an equation that fits a curve that is the probability of a baby with a low weight at birth as a relationship to mother's weight at the last menstrual cycle (Figure 3). This is done using of the logit function mentioned earlier.

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

For this example, X_1 denotes the mother's weight at the last menstrual cycle (LWT). If we solve for p (the probability a baby is born with a low birth weight) we get the following function

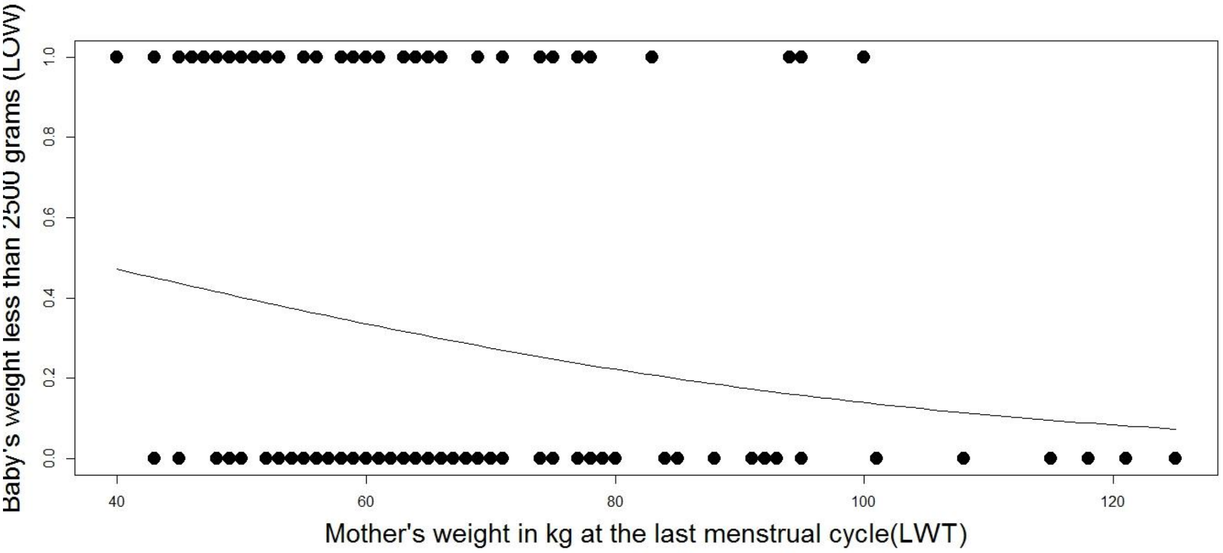
$$p = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}}$$

which is the formula used to derive the curve in Figure 3.

Note that all logistic regression equations have an S -shape when represented graphically.

Now, looking at Figure 3 it seems clear that the less the mother weighs at the last menstrual cycle, the more possible is that a baby will be born with a low weight.

Figure 3



How do we use a logistic model that will let us explore this relationship in more detail?

We have the following variables:

The outcome: Baby's low birthweight (LOW)

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

The explanatory variable: Mothers weight at the last menstrual cycle (LWT).

In logistic regression the OR of a continuous explanatory variable is the factor of the odds that outcome=1 when the explanatory variable is increased by one unit.

Model

We have the following logistic model equation:

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

The results when fitting a logistic regression to the data are the following

```
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
```

Interpretation

The intercept ($\hat{\beta}_0 = 1.0232$) is the estimated log odds of LOW for individuals whose weight is 0. The intercept is mathematically necessary to specify the entire equation and use the entire equation to estimate the log odds of the outcome for any group given LWT. In many cases the intercept might not have a meaningful practical explanation.

We notice that the estimated coefficient ($\hat{\beta}_1 = -0.02842$) of LWT is negative. $\hat{\beta}_1$ is the estimated change in the log odds of LOW for one kg increase in LWT. Here, we notice a negative association between LWT and log odds of LOW.

To convert these values to odds ratio (OR) we just need to take the exponential value of log odds. So, the OR for $\hat{\beta}_1$ is $\mathrm{e}^{-0.02842} = 0.97198$. This means that the odds of a baby to be born with a low weight are reduced by about 2.8% (0.972-1=-0.028) as the mother’s weight increases by one kg (derived as 1-0.97198 times 100%). The related p-value is 0.0218 which means that this is a statistically significant result at a 5% level. The 95% confidence interval for the odds of LWT is (0.9471, 0.9944) does not contain 1, which also indicates that the result is significant.

If we subtract 1 from the odds ratio and multiply by 100 (that is, (odds ratio-1) x 100), we obtain the percentage change in the odds for a 1-unit change in the explanatory variable. Thus, the odds ratios allow us to see what the effect of the explanatory variables is on the odds. It is usually best thought of in terms of the percentage change in the odds for a one-unit change in the explanatory variable.

In case we would like to express the OR for every 10 kg increase in mother's weight, we just need to raise the odds on the power of 10. So, $0.97198^{10} = 0.7526$, which means that the probability that a baby will be born with a low weight is reduced by about 25% (0.75-1=-0.25) for every 10 kg increase in the mother's weight.

Example 2: Simple logistic regression with categorical explanatory variable with two categories: Baby's low birthweight and whether the mother was smoking during pregnancy.

Using the same sample of 189 babies with the low birthweight (LOW) is affected by the mother's smoking status (SMOKE). Babies are coded as 1 or 0 depending on whether their birthweight was less than 2500 grams or not respectively.

In logistic regression the OR of a binary explanatory variable is the factor of the odds that outcome=1 within one category, compared to the odds that outcome=1 within the other category which is used as a reference group.

We have the following variables:

The outcome: Baby's low birthweight (LOW)

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

The 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:

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

The results when fitting a logistic regression to the data are the following

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

Interpretation

We notice that coefficient $\hat{\beta}_1 = 0.704$ is positive, so low birthweight is positively associated with smoking during pregnancy.

The OR of coefficient $\hat{\beta}_1$ is $\mathrm{e}^{0.7041} = 2.021$ meaning that the odds for the birth of a baby with low weight are almost two times higher for smokers than for non-smokers. The related p-value is 0.027 which means that this is a statistically significant result. The 95% confidence interval for the odds of SMOKE is $(1.082, 3.800)$ does not contain 1, which also indicates that the result is significant.

Relation to the Chi-square test

Chi-square test can be considered as a special case of logistic regression where both dependent and independent variables are binary. The same way as t-test and one-way ANOVA are special cases of simple linear regression.

In the above example, let Table 1 below be the $2x2$ contingency table of the variables LOW and SMOKE.

Table 1: Contingency Table of variables LOW and SMOKE.

		LOW	
		No	Yes
SMOKE	No	86	29
	Yes	44	30

The Chi-square test of independence gives $\chi^2 = 4.923,\ df = 1,\ p - value = 0.0264$. If we calculate the OR we have

$$OR = \frac{30/44}{29/86} = 2.021$$

which is the same as the OR found from the simple logistic regression.

Example 3: Simple logistic regression with a categorical explanatory variable with more than two categories: Baby's low birthweight and mother's race.

Using the same sample of 189 babies with the low birthweight (less than 2500 grams) (LOW) is affected by the mother's race. Babies are coded as 1 or 0 depending on whether their birthweight was less than 2500 grams or not respectively.

The explanatory variable here has three categories so we need to create dummy variables for each of these categories. With dummy coding we recode the original categorical variable into a set of binary variables that have values of one or zero meaning whether or not the original variable has that particular category value respectively.

We are including all the categories to the logistic regression model except one which is going to be used as the reference group.

In logistic regression, the ORs of a dummy variable is the factor of the odds that outcome=1 within that category of the explanatory variable compared to the odds that outcome=1 within the reference category.

So, we have the following variables:

The outcome: Baby’s low birthweight (LOW)

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

The explanatory variable: Mother’s race (RACE).

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

We consider the groups LOW=0 and RACE=1 as the reference groups.

Model

We have the following logistic model equation:

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

The results when fitting a logistic regression to the data are the following

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(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
```

Interpretation

We notice that both coefficients for the race categories (0.8448 for “black” and 0.6362 for “other”) are positive, so low birthweight is positively associated with each of these two categories.

The OR of the coefficient for black mothers is $\mathrm{e}^{0.8448} = 2.3275$ meaning that the odds for the birth of a baby with low weight are 2.3 times higher for black mothers compared to white mothers. Similarly, the OR of the coefficient for mothers with other race is $\mathrm{e}^{0.6362} = 1.8892$ meaning that the odds for the birth of a baby with low weight are 1.88 times higher for mothers of ‘other’ race compared to white mothers. The related p-values are 0.068 and 0.067 for black mothers and mothers with other race respectively, meaning that the results are non-significant. The 95% confidence intervals for black mothers and mothers with other race are (0.9255, 5.7746) and (0.9565, 3.7578) respectively. Both confidence intervals contain 1 which also implies the non-significance of the results.

NOTE: 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. In any case, we need to follow the most meaningful approach to our problem.

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 and/or categorical, nominal or ordinal for the latter.

Sample size for the Multiple Logistic Regression

The rule of thumb is that for every independent variable, there should be no fewer than 10 outcome events for each binary category (e.g., alive ⁄ deceased), with the least common outcome determining the maximum number of independent variables. 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 events in the outcome).

Example for Multiple Logistic Regression

We will explain multiple logistic regression using an example. Consider the same example described in the Simple Logistic Regression section. Now we would like to see if any of the variables (AGE, LWT, RACE, SMOKE, PTL, HT, UI and FTV) have an effect on low birthweight (LOW). We firstly need to perform a separate univariate logistic regression for each of the explanatory variables. This begins to investigate confounding, as well as providing an initial “unadjusted” view of the importance of each variable, by itself. Similarly to the linear regression, the variables that have a $p \leq 0.2$ in the univariate analysis will be included in the multivariable model.

The following table gives us the results of the univariate logistic regression for each variable.

Table 2: Univariate analysis results

Variable Name	OR (95% CI)	p
LWT	0.971 (0.947, 0.994)	0.021
RACE – Black RACE - Other	2.327 (0.925, 5.774) 1.889 (0.9565, 3.7578)	0.068 0.067
SMOKE	2.021 (1.082, 3.800)	0.027
AGE	0.950 (0.891, 1.009)	0.105
PTL	2.229 (1.218, 4.283)	0.011
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

We can see that all the variables have a p-value less that 0.2 except from FTV. Thus FTV will not be included in the model.

Moreover, since we have 59 outcome events we can include at most 6 independent variables in the model. Note that if a categorical variable has more than two categories, then these categories will count as separate independent variables.

So, in this example, we wish to include variable RACE in the model. RACE has three categories, “White”, “Black”, “Other”. We will use “White” as the reference group, so the rest two categories “Black” and “Other” will account for two independent variables in the model. This means than we can now include at most four more independent variables in the model. We run the multivariable model including the most important variables (derived from our clinical knowledge or background literature) and the results are shown below:

```
Coefficients:
          Estimate Std.Error z value  p-value   OR       (95%CI)
(Intercept) 0.15108 0.94821   0.159    0.8734   1.163  (0.191,  7.976)
RACE Black  1.29413 0.52277   2.476    0.0133   3.648  (1.312, 10.363)
RACE Other  0.91036 0.42834   2.125    0.0336   2.485  (1.087,  5.889)
SMOKE       0.94527 0.39519   2.392    0.0168   2.573  (1.201,  5.707)
PTL         0.60296 0.33524   1.799    0.0721   1.827  (0.959,  3.619)
HT          1.75304 0.69598   2.519    0.0118   5.772  (1.530, 24.753)
LWT        -0.03359 0.01375  -2.443    0.0146   0.967  (0.939,  0.992)
```

We notice that PTL is also not significant. If we calculate the variable frequency table we get:

Table 3: Frequency Table for variable PTL

PTL	
Values	**Frequency (%)**
0	159 (84.1)
1	24 (12.7)
2	5 (2.6)
3	1 (0.5)

We notice at Table 3 that only very few women had more than one premature births so it is better if we transform this variable to binary with values:

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

Running the model again with the binary PTL we get,

```
Coefficients:
              Estimate Std.Error z value p-value  OR      (95%CI)
(Intercept)     0.12701  0.96113  0.132  0.8949  1.135  (0.182,  7.993)
RACE.cat Black  1.26725  0.52971  2.392  0.0167  3.551  (1.259, 10.216)
RACE.cat Other  0.86434  0.43513  1.986  0.0470  2.373  (1.024,  5.698)
SMOKE           0.87527  0.40093  2.183  0.0290  2.399  (1.105,  5.372)
PTL_bin         1.23172  0.44648  2.759  0.0058  3.427  (1.441,  8.394)
HT              1.77516  0.70949  2.502  0.0123  5.901  (1.523, 25.988)
LWT            -0.03387  0.01394 -2.430  0.0151  0.966  (0.939,  0.992)
```

Now we notice that there is a significant association between premature birth as binary (PTL_bin) and low birthweight (LOW) (p=0.006).

This yields the final model whose results are shown below:

```
Coefficients:
              Estimate Std.Error z value p-value  OR      (95%CI)
(Intercept)     0.12701  0.96113  0.132  0.8949  1.135  (0.182,  7.993)
RACE.cat Black  1.26725  0.52971  2.392  0.0167  3.551  (1.259, 10.216)
RACE.cat Other  0.86434  0.43513  1.986  0.0470  2.373  (1.024,  5.698)
SMOKE           0.87527  0.40093  2.183  0.0290  2.399  (1.105,  5.372)
PTL_bin         1.23172  0.44648  2.759  0.0058  3.427  (1.441,  8.394)
HT              1.77516  0.70949  2.502  0.0123  5.901  (1.523, 25.988)
LWT            -0.03387  0.01394 -2.430  0.0151  0.966  (0.939,  0.992)
```

Interpretation

Interpretation of the variables' effects is similar to simple logistic regression. For variable RACE we can say that Black mothers are 3.5 (p=0.017) times more likely to have a baby with a low weight than white mothers (white mothers is the reference group here), adjusted for all the other variables in the model. Similarly, mothers of other race are 2.37 times (p=0.047) more likely to have a baby with a low weight than white mothers, adjusted for all the other variables in the model. All the other variables can be explained in a similar way.

Model Diagnostics – checking assumptions

Linearity assumption

The linear relationship between continuous predictor variables and the logit of the outcome can be checked by visually inspecting the scatter plot between each predictor and the logit values.

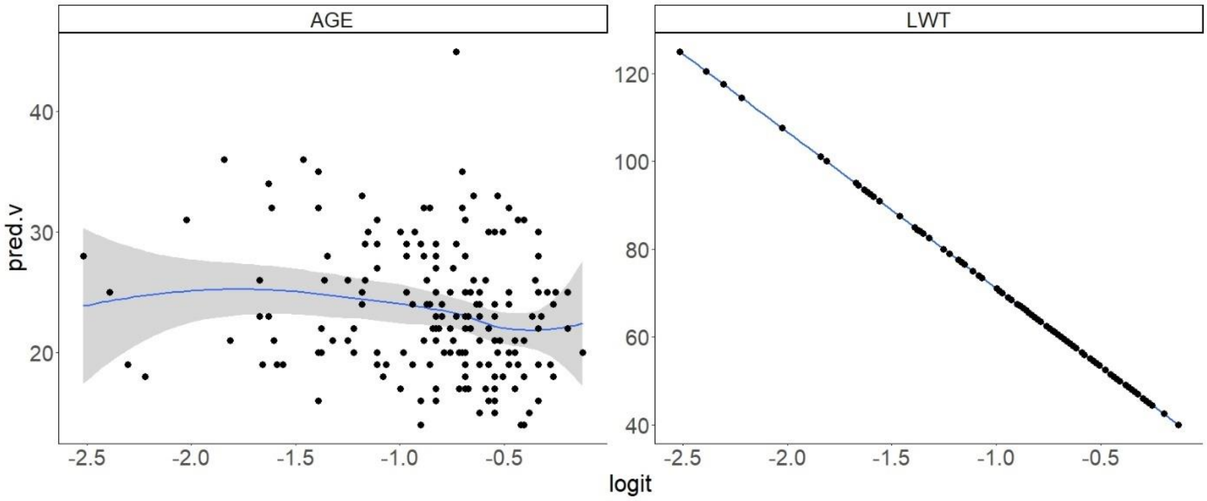


Figure 4 Linearity of continuous explanatory variables AGE and LWT to the outcome LOW in logit scale.

The smoothed scatter plots in Figure 4 show that variables AGE and LWT are all quite linearly associated with the outcome LOW in logit scale.

Influential Values

Influential values are extreme individual data points that can alter the quality of the logistic regression model. Influential values have a disproportionate effect on the model and can produce misleading results. The most extreme values in the data can be examined by visualizing the Cook's distance values. Not all outliers are influential observations. The standardized residual error can be employed to check whether the data contains potential influential observations. Data points with an absolute standardized residuals above 3 represent possible outliers and may deserve closer attention. Cook's distance plot for our model can be seen in Figure 5 where the top 5 largest values are labeled.

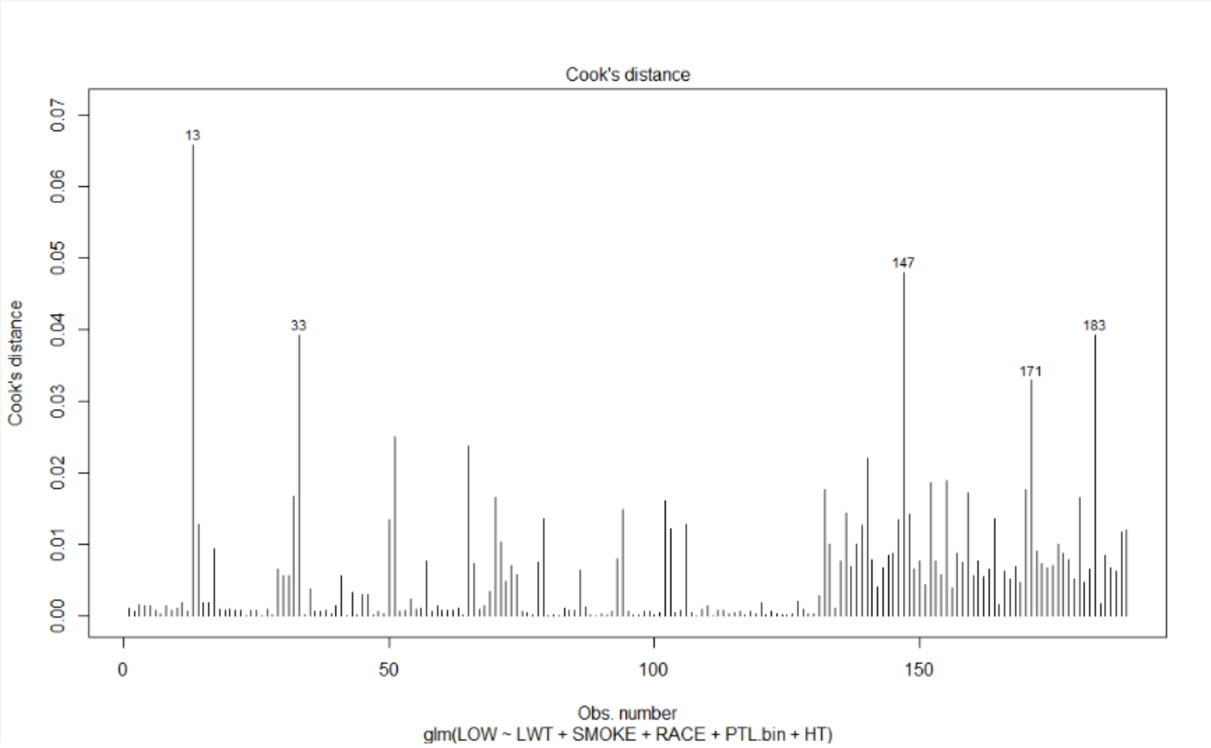


Figure 5 Cook's distance for the final model.

To check whether the data contains potential influential observations, the standardized residual error can be inspected. Data points with an absolute standardized residuals above 3 represent possible outliers and may deserve closer attention. Figure 6 displays such a plot where it can be seen that there are no standardized residuals > 3 in our case.

However, when there are outliers in a continuous predictor, we can either remove these observations, transform the data in the log scale or try using other methods for modelling such as parametric.

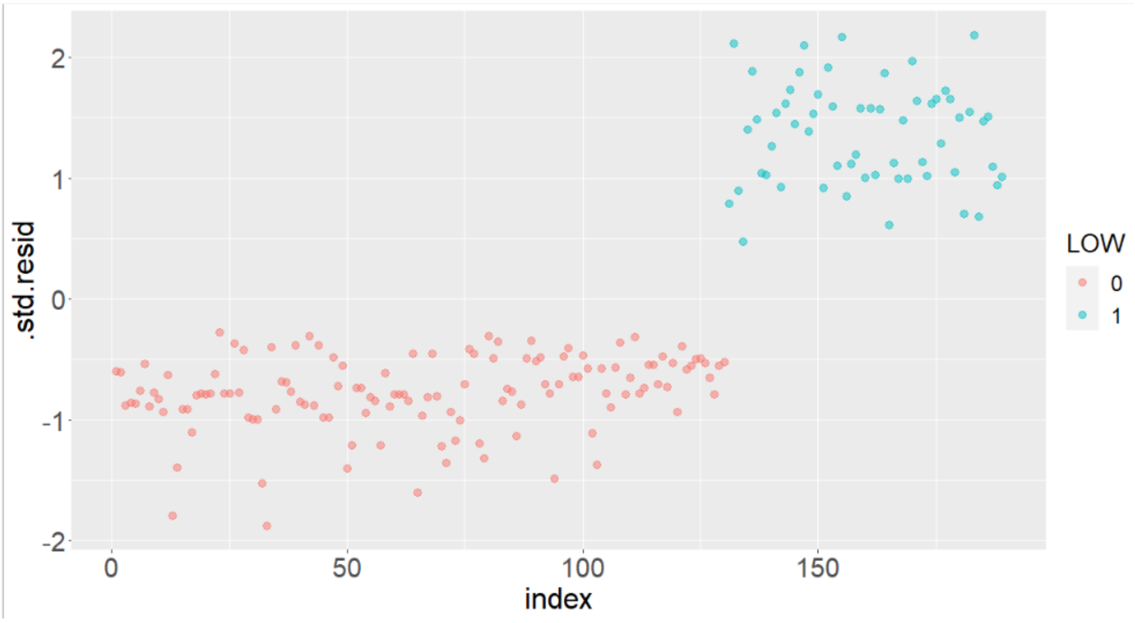


Figure 6 Standardised Residual plot to check for influential points.

Multicollinearity Diagnostics

To check for multicollinearity between the explanatory variables in our model we follow the same method as in the linear regression.

Here we have the multicollinearity diagnostics taken from R:

```
             GVIF      Df   GVIF^(1/(2*Df))
RACE      1.431055   2          1.093740
SMOKE     1.329747   1          1.153147
PTL_bin   1.032742   1          1.016239
HT        1.139632   1          1.067535
LWT       1.245690   1          1.116105
```

We notice that all the variables have a quite low IVF, so there are no any multicollinearity issues in our model.

Model Validation

There are many tests to evaluate the model fit but the most commonly used are the Likelihood Ratio test and the ANOVA test. These tests are equivalent and test whether the model with predictors fits significantly better than a model with fewer predictors (Note that it only makes sense for nested models).

In the bibliography you could also find the use of the Hosmer–Lemeshow goodness of fit test for assessing the model's fit in logistic regression. However, this test is not recommended anymore due to its low power.

Applying the Likelihood Ratio test to our model we have:

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

```
Resid. Df  Resid. Dev  Df  Deviance Pr(>Chi)
    182      200.34
    183      208.10    -1   -7.7619 0.005336 **
```

We notice that the p-value is smaller than the critical value 0.05 so the *Final* model fits better than the *Reduced* model.

Applying the ANOVA test to our model we have:

```
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 ***
```

We notice that the p-value is smaller than the critical value 0.05 so the *Final* model fits better than the *Reduced* model.

Akaike Information Criterion (AIC)

In case we are interested to compare different models we use the *Akaike Information Criterion (AIC)* as in linear regression.

The smallest the AIC the better the model.

Nagelkerke’s pseudo R-squared

For many types of models, like logistic regression, R-squared is not defined. For these models, there are some pseudo R-squared measures that can be calculated. A pseudo R-squared is not directly comparable to the R-squared for OLS models. It cannot be interpreted as the proportion of the variability in the dependent variable that is explained by model. Pseudo R-squared are relative measures among similar models and indicate how well the model explains the data.

Stepwise Logistic Regression

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

The following results are derived using a *Forward Selection*:

```
Coefficients:
            Estimate Std. Error p-value      OR      (95%CI)
(Intercept) 0.95076    1.20509  0.43014   2.587  (0.254, 29.299)
AGE        -0.04254    0.03759  0.25768   0.958  (0.888,  1.030)
RACE Black  1.17250    0.53301  0.02782   3.230  (1.138,  9.359)
RACE Other  0.81494    0.44274  0.06567   2.259  (0.959,  5.500)
SMOKE       0.85760    0.40499  0.03421   2.357  (1.077,  5.322)
PTL_bin     1.33375    0.45777  0.00357   3.795  (1.564,  9.532)
HT          1.74827    0.70421  0.01304   5.744  (1.498, 25.028)
LWT        -0.03128    0.01412  0.02676   0.969  (0.941,  0.995)
```

It can be seen that this model does not differ to the one we chose before without performing stepwise logistic regression.

Using the *Backward Selection* we get the following results:

```
Coefficients:

           Estimate Std. Error p-value  OR       (95%CI)
(Intercept) 0.12701   0.96113  0.8949  1.135  (0.181,  7.993)
RACE Black  1.26725   0.52971  0.0167  3.551  (1.258, 10.215)
RACE Other  0.86434   0.43513  0.0470  2.373  (1.024,  5.698)
SMOKE       0.87527   0.40093  0.0290  2.399  (1.104,  5.372)
PTL_bin     1.23172   0.44648  0.0058  3.427  (1.441,  8.394)
HT          1.77516   0.70949  0.0123  5.901  (1.523, 25.988)
LWT        -0.03387   0.01394  0.0151  0.966  (0.938,  0.991)
```

We notice now that using the backward elimination we ended up in a different model. In such cases we usually present the results with one stepwise method, but the backward elimination method is usually preferred. However, we need to mention whether other methods were used and their results.

Summary of Results

After performing a Logistic Regression we can summarise the results from the univariate and multivariate analysis in a table. Table 4 shows the summary of results for our example.

Table 4: Logistic models for the low birthweight prediction.

	Univariate Analysis			**Multivariable Analysis**[a]		
Variables	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	0.71	0.54 , 0.94	0.015
Race (Reference = White)						
Black/White	2.33	0.94 , 5.77	0.068	3.55	1.25 , 10.21	0.017
Other/White	1.89	0.96 , 3.74	0.067	2.37	1.02 , 5.69	0.047
Smoking (Yes / No)	2.02	1.08 , 3.78	0.027	2.40	1.10 , 5.37	0.029
History of premature births (Yes / No)	4.32	1.92 , 9.73	0.011	3.43	1.44 , 8.36	0.006

Hypertension (Yes / No)	3.37	1.02 , 11.09	0.046	5.90	1.52 , 25.99	0.012
Uterus abnormalities (Yes / No)	2.58	1.14 , 5.84	0.023			
Number of visit to the doctor the first trimester of pregnancy	0.87	0.64 , 1.19	0.388			

OR:Odds Ratio, CI: Confidence Interval

[α]Using the Backward Selection method