Models for binary response

Logistic regression

Dae-Jin Lee
dlee@bcamath.org

Basque Center for Applied Mathematics

http://idaejin.github.io/bcam-courses/

Outline

Models for binary response: Logistic regression

Interpretation of the parameters

Variable selection

Logistic regression model predictions

Model diagnostics

Interpretation of the results

Binary data

Introduction

- Suppose that for each individual, the response $\boldsymbol{y}$, can take only two possible values, 0 and 1 (also known as *dichotomous*)
- **Examples:** *in biomedicine where we might want to predict if a patient respond or not to a drug; in business management, a bank may want to predict whether or not an individual is likely to pay his credit card bills, etc.*
- We may write

$$\mathbb{P}\mathsf{r}(y_i = 1) = \pi_i \qquad \mathbb{P}\mathsf{r}(y_i = 0) = 1 - \pi_i$$

- Normally, we will have a set of covariates $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ associated with each individual, and our goal will be to investigate the relationship between the response probability $\pi = \pi(\boldsymbol{X})$ and the explanatory variables.

Binary data

Example: Health perception survey Comunidad de Madrid

- Let us consider the data `salud`
- **Aim:** describe how health perception is different depending on the variables `sexo`, `edad`, `bebedor`

Binary data

Example: Health perception survey Comunidad de Madrid

- Let us consider the data `salud`
- **Aim:** describe how health perception is different depending on the variables `sexo`, `edad`, `bebedor`
- We have that

$$y_i = \begin{cases} 1 & \text{Good health perception} \\ 0 & \text{Bad health perception} \end{cases}$$

Binary data

Example: Health perception survey Comunidad de Madrid

- Let us consider the data `salud`
- **Aim:** describe how health perception is different depending on the variables `sexo`, `edad`, `bebedor`
- We have that

$$y_i = \begin{cases} 1 & \text{Good health perception} \\ 0 & \text{Bad health perception} \end{cases}$$

- $y \sim \mathcal{B}ernoulli(p_i)$

$$\mathbb{P}\mathsf{r}[Y_i = y_i] = p_i^{y_i}(1-p_i)^{1-y_i} \quad \text{with } y_i = 0, 1.$$

Binary data

Example: Health perception survey Comunidad de Madrid

- Let us consider the data `salud`
- **Aim:** describe how health perception is different depending on the variables `sexo`, `edad`, `bebedor`
- We have that

$$y_i = \begin{cases} 1 & \text{Good health perception} \\ 0 & \text{Bad health perception} \end{cases}$$

- $y \sim \mathcal{B}ernoulli(p_i)$

$$\mathbb{P}\mathrm{r}[Y_i = y_i] = p_i^{y_i}(1-p_i)^{1-y_i} \quad \text{with } y_i = 0, 1.$$

$*$ Alternatively, we can clasify the individuals according to the variables of interest in k groups.

Binary data

Grouping by categorical predictors

- Let us consider the clasification the individuals in $k = 12$ groups according to `sexo` and `bebedor` variables.

Binary data

Grouping by categorical predictors

- Let us consider the clasification the individuals in $k = 12$ groups according to `sexo` and `bebedor` variables.

See `Logreg.R`

```
> ftable(list(g02,sexo,bebedor))
                x.3 poco/nada ocasional frecuente
x.1  x.2
good male              223       335        36
     female            516       281        15
bad  male              792      2090       190
     female           1446      1348        85
```

Logistic regression

Models for Binary data

- When binary data are grouped by covariate class, the responses have the form y_i/n_i, where $0 \leq y_i \leq n_i$ is the number of successes out of the n_i individuals in the i^{th} covariate class, and the total number of individuals is $n = \sum_i n_i$.
- If data are ungrouped, there are as many covariate classes as individuals and $n_i = 1$.
- In this context arises the **Binomial distribution**.

Logistic regression

Models for Binary data

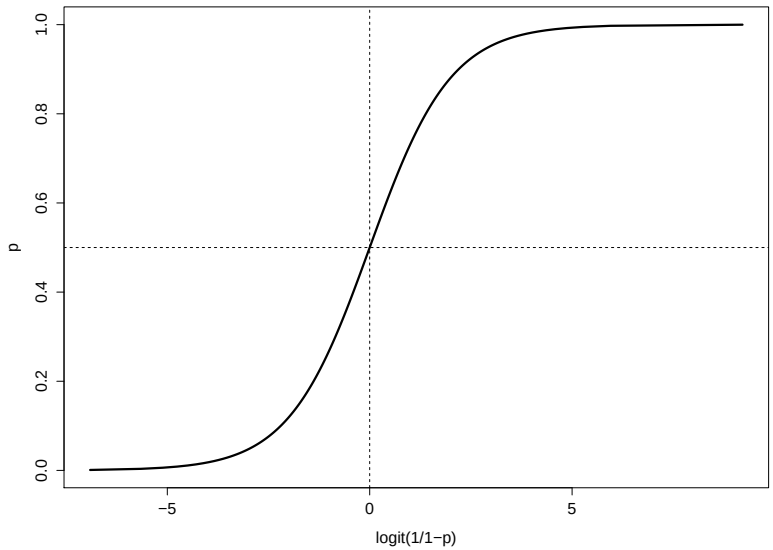
- When binary data are grouped by covariate class, the responses have the form y_i/n_i, where $0 \leq y_i \leq n_i$ is the number of successes out of the n_i individuals in the i^{th} covariate class, and the total number of individuals is $n = \sum_i n_i$.
- If data are ungrouped, there are as many covariate classes as individuals and $n_i = 1$.
- In this context arises the **Binomial distribution**.
- In the Health perception data example:
  - n_i is the number of observations in group i
  - y_i is the number of individuals that perceived themselves as Healthy people within group i

Logistic regression

Models for Binary data (cont.)

- With binary data, working with probabilities we need that $\pi_i \in (0, 1)$
- **Solution: Logit** or **log-odds**, i.e.:

$$\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} = \eta_i = \log\left(\frac{p_i}{1 - p_i}\right)$$



- **Odds*** is $\frac{p_i}{1-p_i}$
- **logit** transforms the probability $p \in (0, 1) \rightarrow \in (-\infty, \infty)$
- when $p = 1/2$, odds $= 1$ and logit $= 0$
- logit < 0 corresponds to $p < 1/2$ and viceversa.

Odds: *is the possibility that something will happen, the chance that one thing will happen instead of a different thing*

Logistic regression

Models for Binary data (cont.)

- The **logit** transformation is *one-to-one*, its inverse is known as the *antilogit* and allows the calculation of the probability from the logit, i.e.:

$$p_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

- The **logistic regression model** assumes that the logit of the probability can be modelled by a **linear model** of the form:

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

and then

$$p_i = \frac{e^{\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}}}$$

- The probability is a *non-linear function* of the predictors, and then it is difficult to interpret the effect in the probability given changes in the predictors. That is the main reason why it is better to work with the **logit**.

Logistic regression

as a Generalized Linear Model

- The estimates of the β's are obtained maximizing the Likelihood, i.e. in logistic regression:

$$\log \mathcal{L}(\beta) = \sum (y_i \log(p_i) + (n_i - p_i) \log(1 - p_i))$$

  as we saw previously the maximization problem cannot be solve analytically, and hence an iterative method based on Newton-Raphson must be used.
- Remind from previous session that for binary data, logistic regression is a particular case of a GLM
- A GLM has 3 components:

  1. **Probability distribution** of the response as a member of the Exponential family (Bernoulli or Binomial)
  2. **Systematic component:** the linear predictor $\boldsymbol{\eta} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$
  3. **Link function** to relate the mean with the linear predictor, i.e. $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ (**logit**)

GLM's for binary data

The function `glm()` in `R`

- The `R` procedure to fit GLM's is the function `glm()`
- Arguments of function `glm()`
  - `formula` similar to `lm()`
  - `family` from the Exponential Family
  - `link` link function
- Let us fit a logistic regression model to the `Health perception` data example
- Open the `R` script `Logreg.R`

GLM's for binary data

The function `glm()` in `R`

- The `R` procedure to fit GLM's is the function `glm()`
- Arguments of function `glm()`
  - `formula` similar to `lm()`
  - `family` from the Exponential Family
  - `link` link function
- Let us fit a logistic regression model to the `Health perception` data example
- Open the `R` script `Logreg.R`

```
> logistic1 <- glm(g02~sexo+bebedor,family=binomial(link=logit))
> summary(logistic1)
```

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.26858    0.06027  21.049  < 2e-16 ***
sexofemale       -0.23868    0.06230  -3.831 0.000128 ***
bebedorocasional  0.55151    0.06288   8.771  < 2e-16 ***
bebedorfrecuente  0.49377    0.15984   3.089 0.002007 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7177.9  on 7356  degrees of freedom
Residual deviance: 7059.2  on 7353  degrees of freedom
AIC: 7067.2

Number of Fisher Scoring iterations: 4
```

GLM's for binary data

The function `glm()` in `R`

```
> attributes(logistic1)

$names
 [1] "coefficients"      "residuals"         "fitted.values"
 [4] "effects"           "R"                 "rank"
 [7] "qr"                "family"            "linear.predictors"
[10] "deviance"          "aic"               "null.deviance"
[13] "iter"              "weights"           "prior.weights"
[16] "df.residual"       "df.null"           "y"
[19] "converged"         "boundary"          "model"
[22] "call"              "formula"           "terms"
[25] "data"              "offset"            "control"
[28] "method"            "contrasts"         "xlevels"

$class
[1] "glm" "lm"
```

GLM's for binary data

The function `glm()` in `R`

- The model is

$$\log\left(\frac{p}{1-p}\right) = 1,27 - 0,238 * \text{sexo}_{\text{female}} + 0,55 * \text{bebedor}_{\text{ocasional}} + 0,49 * \text{bebedor}_{\text{frec}}$$

- Confidence intervals are also obtained as:

```
> confint(logistic1)
                       2.5 %     97.5 %
(Intercept)        1.1512601  1.3875492
sexofemale        -0.3609674 -0.1167008
bebedorocasional   0.4283962  0.6749056
bebedorfrecuente   0.1891390  0.8169365
```

GLM's for binary data

The function `glm()` in `R` (cont.)

- In terms of **odds**

See `Logreg.R`

```
> exp(coefficients(logistic1))
     (Intercept)       sexofemale bebedorocasional bebedorfrecuente
        3.555815         0.787666         1.735871         1.638488
> exp(confint(logistic1))
                       2.5 %    97.5 %
(Intercept)        3.1621750 4.0050226
sexofemale         0.6970017 0.8898514
bebedorocasional   1.5347941 1.9638475
bebedorfrecuente   1.2082089 2.2635549
```

GLM's for binary data

Interpretation of the parameters

- The key to interpreting logistic regression coefficients is to think in terms of **odds**. Remember that we have defined the *odds* as $p_i/(1-p_i)$, the ratio of the probability to its complement. If we know the odds we can calculate the probability according to: $\text{odds}/(1+\text{odds})$.

GLM's for binary data

Interpretation of the parameters

- The key to interpreting logistic regression coefficients is to think in terms of **odds**. Remember that we have defined the *odds* as $p_i/(1-p_i)$, the ratio of the probability to its complement. If we know the odds we can calculate the probability according to: $\text{odds}/(1+\text{odds})$.
- In a logistic model with canonical link function we assume that

$$\text{logit}(p_i) = \ln(\text{odds}) = \ln\left(\frac{p}{1-p}\right) = \beta x_i.$$

  Hence, a unit increase in the explanatory variable will result in a β increase in the log-odds. It is difficult to think in terms of log-odds, so, if we take exponential in both sides:

$$\text{odds} = \frac{p_i}{1-p_i} = \exp(\beta x_i).$$

- Hence if x_i increases in one unit, the odds will be increased by $\exp(\beta)$.

Interpretation of the parameters

- When interpreting parameters in logistic regression we have to consider:
  1. functional relationship between the dependent variable and independent variable/s
  2. The unit of change of the independent variable/s
- The interpretation also will depend on the type of independent variables: *dichotomous, polytomous or continuous*.

Interpretation of the parameters

Dichotomous independent variable

- Let us first consider the case where X takes two possible values (coded as 0 or 1 or defined as a factor).
- The model is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1$$

the difference in the **logit** for an individual such that $x = 0$ and $x = 1$ is β_1.

- Hence,

$$p = \frac{e^{(\beta_0+\beta_1 x_1)}}{1+e^{(\beta_0+\beta_1 x_1)}} \quad \text{and} \quad 1-p = \frac{1}{1+e^{(\beta_0+\beta_1 x_1)}}$$

Interpretation of the parameters

Dichotomous independent variable

The different values for the probabilities and all possible combinations:

Y	$X = 1$	$X = 0$
$y = 1$	$p(y = 1\|x = 1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$p(y = 1\|x = 0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
$y = 0$	$p(y = 0\|x = 1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$p(y = 0\|x = 0) = \frac{1}{1+e^{\beta_0}}$
Total	1	1

Interpretation of the parameters

Dichotomous independent variable

The different values for the probabilities and all possible combinations:

Y	$X=1$	$X=0$
$y=1$	$p(y=1\|x=1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$p(y=1\|x=0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$
$y=0$	$p(y=0\|x=1)=\frac{1}{1+e^{\beta_0+\beta_1}}$	$p(y=0\|x=0)=\frac{1}{1+e^{\beta_0}}$
Total	1	1

The **odds** of the response variable when $X=1$ is:

$$\frac{p(y=1|x=1)}{p(y=0|x=1)}=\frac{p(y=1|x=1)}{1-p(y=1|x=1)}$$

and similarly when $X=0$, i.e.

$$p(y=1|x=0)/1-p(y=1|x=0)$$

Interpretation of the parameters

Dichotomous independent variable

- The **odds-ratio (OR)** is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group:

Interpretation of the parameters

Dichotomous independent variable

- The **odds-ratio (OR)** is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group:

$$\text{OR} = \frac{p(y=1|x=1)/1-p(y=1|x=1)}{p(y=1|x=0)/1-p(y=1|x=0)}$$

Interpretation of the parameters

Dichotomous independent variable

- The **odds-ratio (OR)** is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group:

$$
\begin{aligned}
\text{OR} &= \frac{p(y=1|x=1)/1-p(y=1|x=1)}{p(y=1|x=0)/1-p(y=1|x=0)} \\
\text{OR} &= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) \Big/ \left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) \Big/ \left(\frac{1}{1+e^{\beta_0}}\right)} \\
&= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \\
&= \boxed{e^{\beta_1}}
\end{aligned}
$$

Interpretation of the parameters

Dichotomous independent variable

- **OR** is one way to quantify how strongly the presence or absence of property A is associated with the presence or absence of property B in a given population.
- In medical research, **OR** is commonly used for case-control studies

Interpretation of the parameters

Dichotomous independent variable

- **OR** is one way to quantify how strongly the presence or absence of property A is associated with the presence or absence of property B in a given population.
- In medical research, **OR** is commonly used for case-control studies
- Sometimes it is common to interpret **OR** as the **Relative Risk (RR)**

$$\text{RR} = \frac{p(y=1|x=1)}{p(y=1|x=0)} = \frac{\text{Prob. of an event when exposed}}{\text{Prob. of an event when not-exposed}}$$

Interpretation of the parameters

Dichotomous independent variable

- **OR** is one way to quantify how strongly the presence or absence of property A is associated with the presence or absence of property B in a given population.
- In medical research, **OR** is commonly used for case-control studies
- Sometimes it is common to interpret **OR** as the **Relative Risk (RR)**

$$\text{RR} = \frac{p(y=1|x=1)}{p(y=1|x=0)} = \frac{\text{Prob. of an event when exposed}}{\text{Prob. of an event when not-exposed}}$$

- Odds ratios have often been confused with relative risk in medical literature
- **Relative risk (RR)** is the risk of an event (or of developing a disease) relative to exposure. Relative risk is a ratio of the probability of the event occurring in the exposed group versus a non-exposed group.
- To approximate OR to RR, $p(y=1|x=1)$ and $p(y=1|x=0)$ should be very small.

Interpretation of the parameters

Dichotomous independent variable

- **e.g.:** Let us consider the Health perception survey data

```
> table(g02,sexo)

      sexo
g02    male female
  good  594    812
  bad  3072   2879
```

$$p(y=1|x=1) = 2879/(2879+812) = 0{,}78 \qquad p(y=0|x=1) = 0{,}22$$
$$p(y=1|x=0) = 3072/(3072+594) = 0{,}838 \qquad p(y=0|x=0) = 0{,}162$$
$$\text{RR} = \frac{0{,}78}{0{,}838} = 0{,}93 \qquad \text{OR} = \frac{0{,}78/0{,}22}{0{,}838/0{,}162} = 0{,}68$$

- The event of healthy women is 0,68 times lower.

Interpretation of the parameters

Dichotomous independent variable

- **e.g.:** Confidence intervals for the estimated **odds-ratio** is computed with exponentials, i.e.

$$\exp\left(\hat{\beta}_1 \pm z_{\alpha/2}\widehat{s.e.(\hat{\beta}_1)}\right)$$

```
> logistica0 <- glm(g02~sexo,family=binomial(logit),data=salud)
> exp(coefficients(logistica0))

(Intercept)       sexo2
  5.1717172   0.6855685

> exp(confint(logistica0))

                2.5 %   97.5 %
(Intercept) 4.7408181 5.651618
sexo2       0.6094711 0.770800
```

Interpretation of the parameters

Polytomous independent variable

- Now we consider the case with a **polytomous independent variable**.
- We grouped the variable edad into 3 categories, i.e.

  We create groups of ages $18 - 29$, $30 - 44$ and $45 - 64$ in edad2

```
> # Create a new variable edad2
> salud$edad2 <- salud$edad
> salud$edad2[salud$edad >= 18 & salud$edad <= 29] <- 1
> salud$edad2[salud$edad >= 30 & salud$edad <= 44] <- 2
> salud$edad2[salud$edad >= 45 & salud$edad <= 64] <- 3
> # make it factor
> salud$edad2<-factor(salud$edad2)
```

Interpretation of the parameters

Polytomous independent variable

- Now we consider the case with a **polytomous independent variable**.
- We grouped the variable edad into 3 categories, i.e.

  We create groups of ages $18-29$, $30-44$ and $45-64$ in edad2

```
> # Create a new variable edad2
> salud$edad2 <- salud$edad
> salud$edad2[salud$edad >= 18 & salud$edad <= 29] <- 1
> salud$edad2[salud$edad >= 30 & salud$edad <= 44] <- 2
> salud$edad2[salud$edad >= 45 & salud$edad <= 64] <- 3
> # make it factor
> salud$edad2<-factor(salud$edad2)

> logistic2=glm(g02~edad2,family=binomial(link=logit),data=salud)
> exp(coefficients(logistic2))

(Intercept)       edad22       edad23
  9.0186914    0.6407686    0.2403349
```

- **How do we interpret the odds ratios?**

Interpretation of the parameters

Continuous independent variable

- When the predictor is continuous, the interpretation of the parameters depends on how the predictor is included in the model and the units of measure.
- β_1 represents the change in the log-odds when the variable X changes a unit, or equivalently, e^{β_1} is the change in the odds.

Interpretation of the parameters

Continuous independent variable

- When the predictor is continuous, the interpretation of the parameters depends on how the predictor is included in the model and the units of measure.
- β_1 represents the change in the log-odds when the variable X changes a unit, or equivalently, e^{β_1} is the change in the odds.
- If the change is of c units, then the change is given by $e^{c\beta_1}$.

Interpretation of the parameters

Continuous independent variable

- When the predictor is continuous, the interpretation of the parameters depends on how the predictor is included in the model and the units of measure.
- β_1 represents the change in the log-odds when the variable X changes a unit, or equivalently, e^{β_1} is the change in the odds.
- If the change is of c units, then the change is given by $e^{c\beta_1}$.
- Let us study the relationship between **Health perception** (g02) and **Body Mass Index** (`imc`)

See `Logreg.R`

```
> logistic3 <- glm(g02~imc,family=binomial(link=logit))
```

```
> summary(logistic3)


Call:
glm(formula = g02 ~ imc, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2860   0.4988   0.5881   0.6689   1.9949

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.120383   0.195110   21.12   <2e-16 ***
imc         -0.107795   0.007637  -14.12   <2e-16 ***
---
Signif. codes:  0
```

Interpretation of the parameters

Continuous independent variable

odds

```
> exp(coefficients(logistic3))


(Intercept)         imc
 61.5828259   0.8978114
```

Interpretation of the parameters

Continuous independent variable

odds

```
> exp(coefficients(logistic3))


(Intercept)         imc
 61.5828259   0.8978114
```

How do we interpret the coefficient for `imc`?

- The coefficient and intercept estimates give us the following equation:

$$\log(p/(1-p)) = logit(p) = 4{,}1203830 - 0{,}1077952 * \text{imc}$$

Interpretation of the parameters

Continuous independent variable

- Let's fix `imc` at some value. We will use 54. Then the conditional logit of being healthy when the `imc` is 54 and 55 is

```
> logit54 <- predict(logistic3,data.frame(imc=54))
> logit55 <- predict(logistic3,data.frame(imc=55))
> # Take the difference
> logit55-logit54
         1
-0.1077952
```

Interpretation of the parameters

Continuous independent variable

- Let's fix `imc` at some value. We will use 54. Then the conditional logit of being healthy when the `imc` is 54 and 55 is

```
> logit54 <- predict(logistic3,data.frame(imc=54))
> logit55 <- predict(logistic3,data.frame(imc=55))
> # Take the difference
> logit55-logit54
         1
-0.1077952
```

- We can say now that the coefficient for `imc` is the difference in the log odds. In other words, for a one-unit increase in the `imc`, the expected change in log odds is $-0{,}1077952$.

Interpretation of the parameters

Continuous independent variable

- **Can we translate this change in log odds to the change in odds?**

Interpretation of the parameters

Continuous independent variable

- **Can we translate this change in log odds to the change in odds?**
- Indeed, we can.

```
> exp(logit55-logit54)
        1
0.8978114
```

Interpretation of the parameters

Continuous independent variable

- **Can we translate this change in log odds to the change in odds?**
- Indeed, we can.

```
> exp(logit55-logit54)
        1
0.8978114
```

- So we can say for a one-unit increase in `imc`, we expect to see about $0{,}89$ times less perception in health.

Interpretation of the parameters

Continuous independent variable

- **Can we translate this change in log odds to the change in odds?**
- Indeed, we can.

```
> exp(logit55-logit54)
        1
0.8978114
```

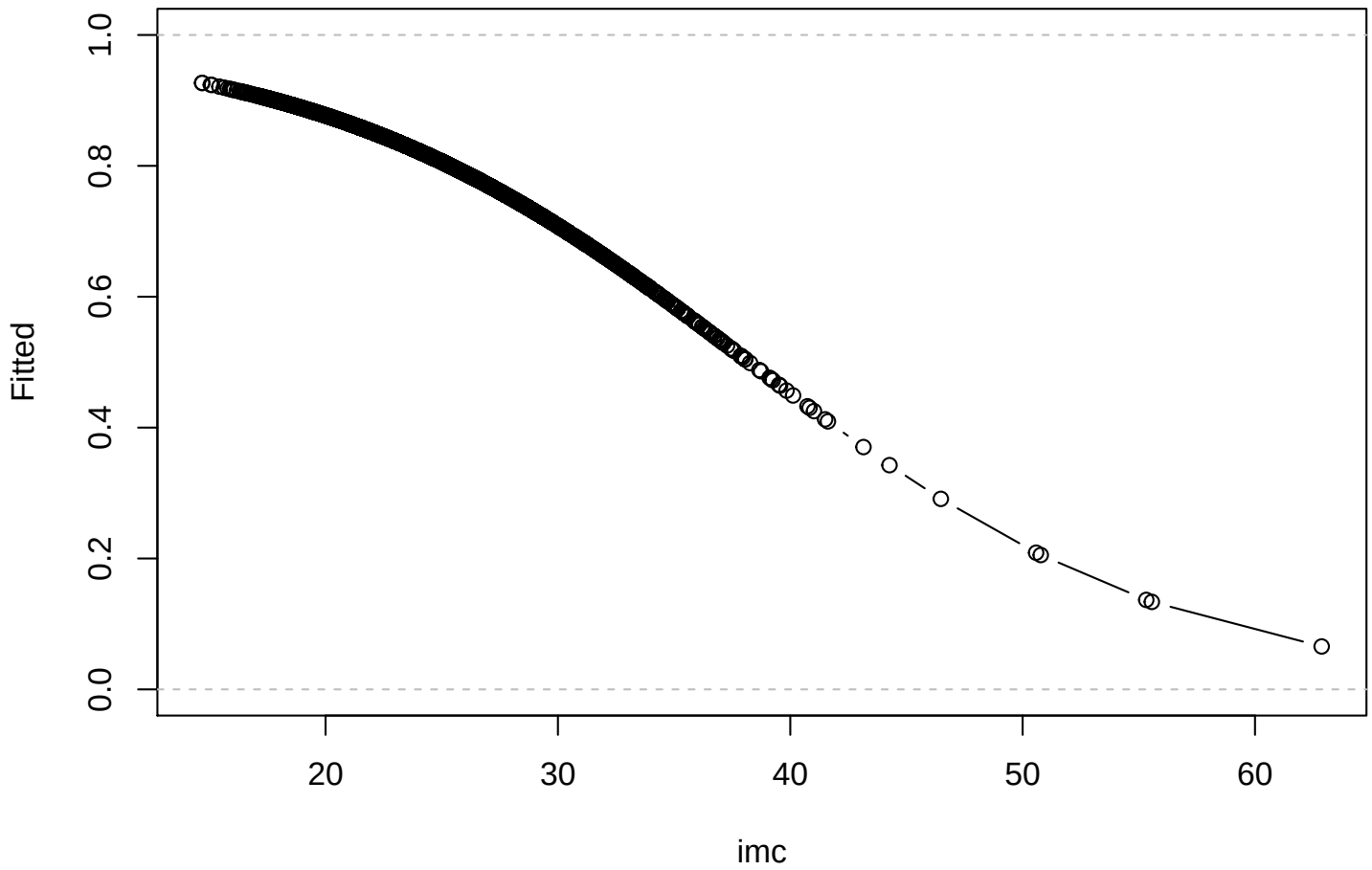
- So we can say for a one-unit increase in `imc`, we expect to see about $0{,}89$ times less perception in health.
- If we consider an increase of c units in `imc`, then:

```
> c = 10
> exp(logit55-logit54)^c # is equivalent to exp(c*logit55-c*logit54)
        1
0.3402917
```

Interpretation of the parameters

Continuous independent variable

- Let us plot the fitted values (See `Logreg.R`)



Interpretation of the parameters

Categorical and Continuous independent variables

- Generally, we will have more than one predictor in logistic regression models.
- In order to interpret a logistic regression model with **several predictors**, we need to understand how to interpret the model fit according to the rest of variables in the model.
- For simplicity, let us consider a **LM** case, where we have 2 predictors: one dichotomous and the other continuous. But we are interested in studying y by factor (the **dichotomous** vble)

Interpretation of the parameters

Categorical and Continuous independent variables

- Generally, we will have more than one predictor in logistic regression models.
- In order to interpret a logistic regression model with **several predictors**, we need to understand how to interpret the model fit according to the rest of variables in the model.
- For simplicity, let us consider a **LM** case, where we have 2 predictors: one dichotomous and the other continuous. But we are interested in studying y by factor (the **dichotomous** vble)

Interpretation of the parameters

Categorical and Continuous independent variables

- **Example:** consider the lm `imc ~ peso` where `imc` (y) of `males` and `females`.

Interpretation of the parameters

Categorical and Continuous independent variables

- **Example:** consider the lm `imc ~ peso` where `imc` (y) of `males` and `females`.
  - If the weight (`peso`) distribution is the same for both groups of individuals, we can directly compare the average BMI of both groups, and this estimation would be an estimate of the BMI difference between males and females.
  - However, if a group has lower average weight than the other, then it does not make sense to do this direct comparison.
  - Because a proportion of the observed `BMI` difference is due to the difference in the `weight` of the groups.
  - To estimate the effect of the variable `sexo` in the `BMI`, we need first to remove the effect due to the weight difference between the groups.

Interpretation of the parameters

Categorical and Continuous independent variables

- **Example:** consider the lm `imc ~ peso` where `imc` (y) of `males` and `females`.
  - If the weight (`peso`) distribution is the same for both groups of individuals, we can directly compare the average BMI of both groups, and this estimation would be an estimate of the BMI difference between males and females.
  - However, if a group has lower average weight than the other, then it does not make sense to do this direct comparison.
  - Because a proportion of the observed `BMI` difference is due to the difference in the `weight` of the groups.
  - To estimate the effect of the variable `sexo` in the `BMI`, we need first to remove the effect due to the weight difference between the groups.
  - **Let us illustrate the idea graphically!!!**

Interpretation of the parameters

Categorical and Continuous independent variables

- For simplicity, let us consider a model with **no-interaction**:

$$y = \beta_0 + \beta_1 \text{sexo2} + \beta_2 \text{peso} + \epsilon,$$

where **sexo2** is a dichotomous variable (‘`males`’, or ‘`female`’) and continuous variable **peso** (for weight in kgr).

Interpretation of the parameters

Categorical and Continuous independent variables

- For simplicity, let us consider a model with **no-interaction**:

$$y = \beta_0 + \beta_1 \text{sexo2} + \beta_2 \text{peso} + \epsilon,$$

  where **sexo2** is a dichotomous variable (‘`males`’, or ‘`female`’) and continuous variable **peso** (for weight in kgr).
- Recall that **sexo2** is a **factor** and `lm` codifies x as a dummy variable ($\text{sexo2} = 1$ or 0)
  - β_1 represents the difference in BMI between the two groups (males/females)

Interpretation of the parameters

Categorical and Continuous independent variables

- For simplicity, let us consider a model with **no-interaction**:

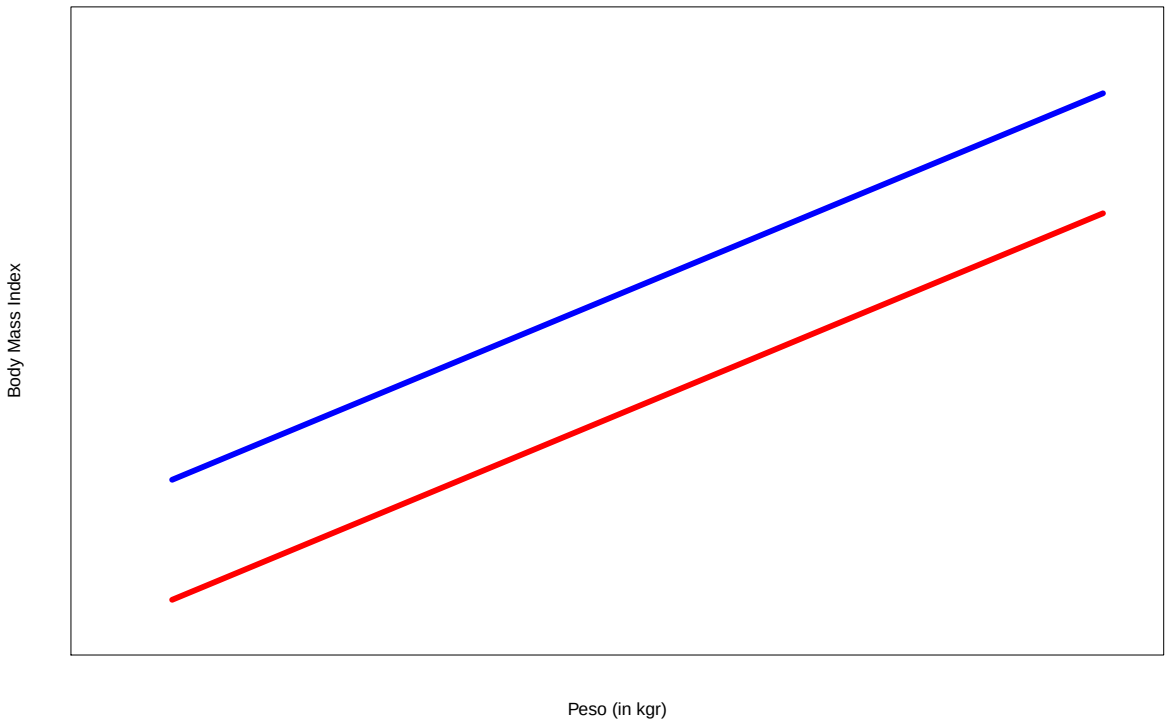
$$y = \beta_0 + \beta_1 \text{sexo2} + \beta_2 \text{peso} + \epsilon,$$

  where **sexo2** is a dichotomous variable (‘`males`’, or ‘`female`’) and continuous variable **peso** (for weight in kgr).
- Recall that **sexo2** is a **factor** and `lm` codifies x as a dummy variable ($\text{sexo2} = 1$ or 0)
  - β_1 represents the difference in BMI between the two groups (males/females)
  - β_2 is the rate of change in BMI (y) per kgr of `peso`.

Interpretation of the parameters

Categorical and Continuous independent variables

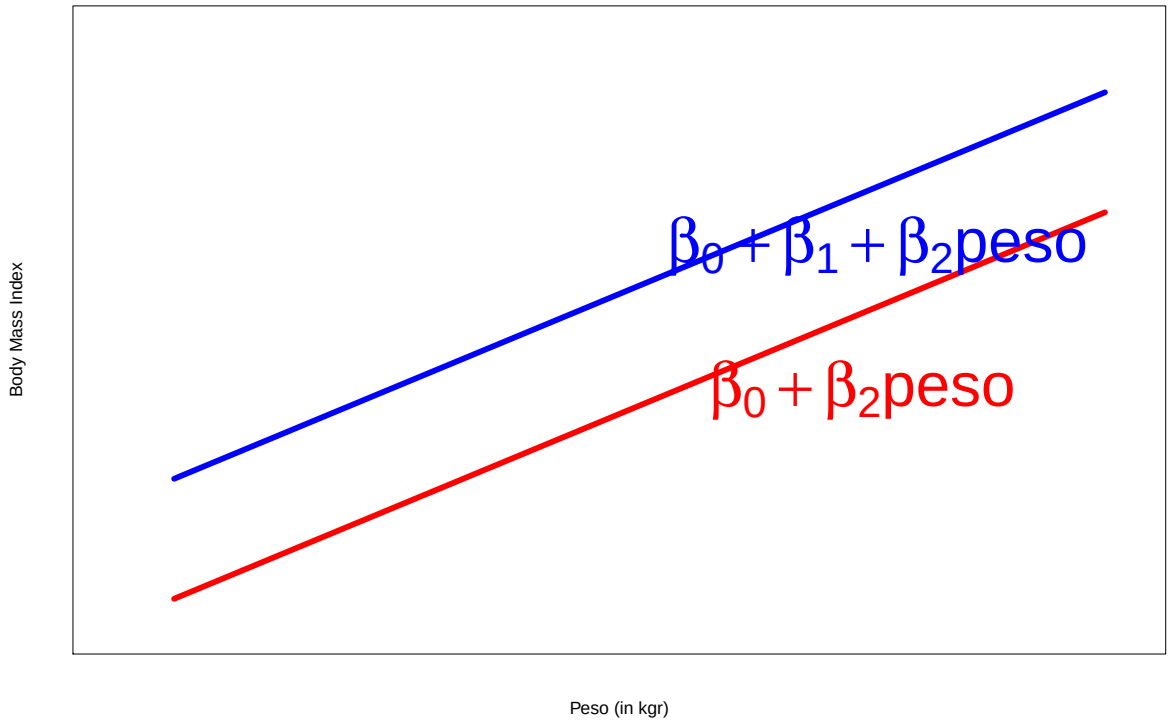
- Let us see the fitted lines for each group
- $\overline{peso}_1$ and $\overline{peso}_2$ are respectively the mean peso of each group (males/females).
- $\overline{peso}_1$ is the overall mean of peso



Interpretation of the parameters

Categorical and Continuous independent variables

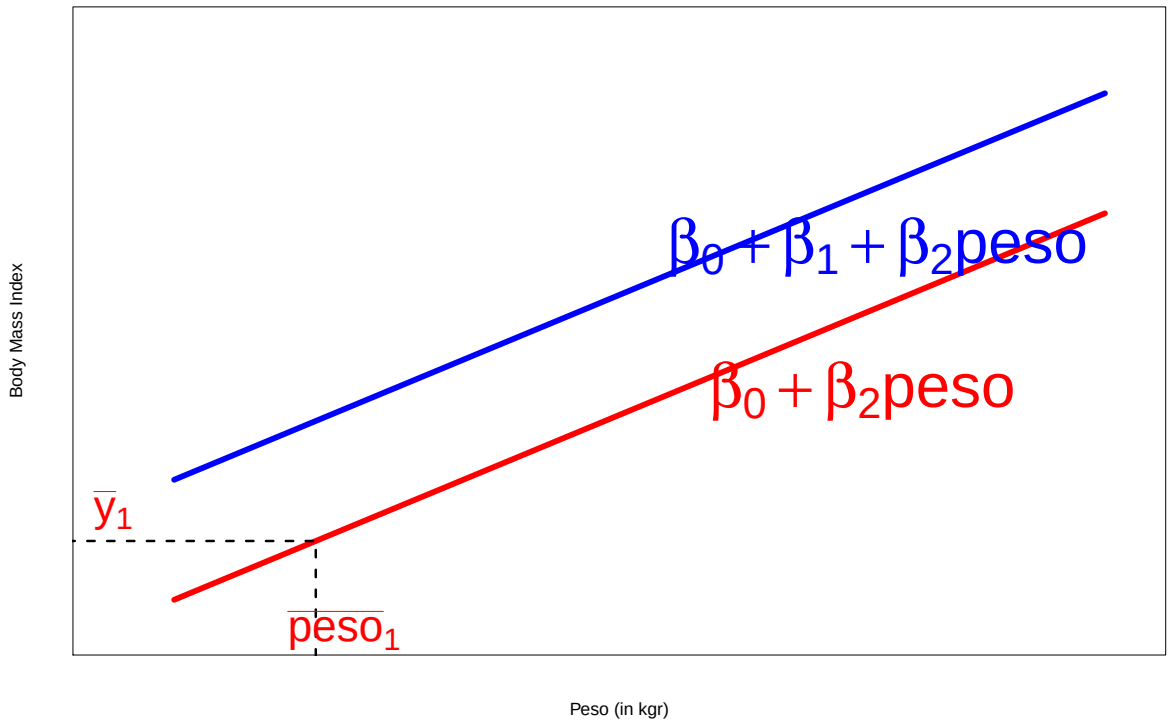
- Let us see the fitted lines for each group
- $\overline{peso}_1$ and $\overline{peso}_2$ are respectively the mean peso of each group (males/females).
- $\overline{peso}_1$ is the overall mean of peso



Interpretation of the parameters

Categorical and Continuous independent variables

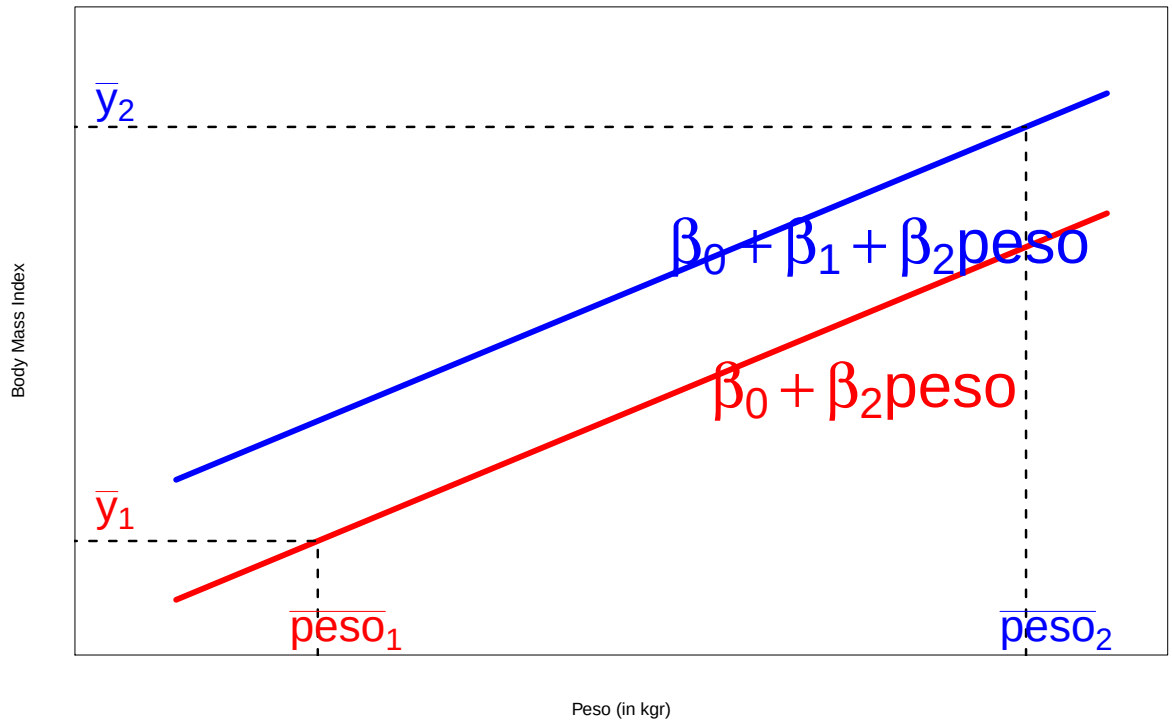
- Let us see the fitted lines for each group
- $\overline{peso}_1$ and $\overline{peso}_2$ are respectively the mean peso of each group (males/females).
- $\overline{peso}_1$ is the overall mean of peso



Interpretation of the parameters

Categorical and Continuous independent variables

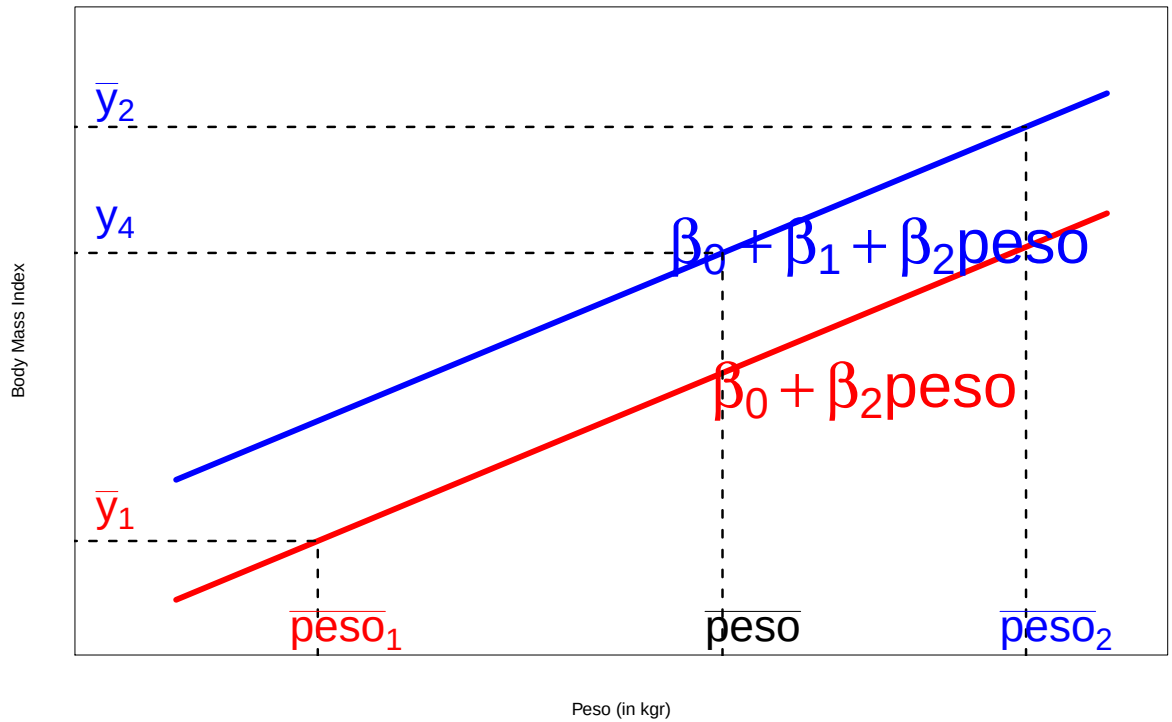
- Let us see the fitted lines for each group
- $\overline{peso}_1$ and $\overline{peso}_2$ are respectively the mean peso of each group (males/females).
- $\overline{peso}_1$ is the overall mean of peso



Interpretation of the parameters

Categorical and Continuous independent variables

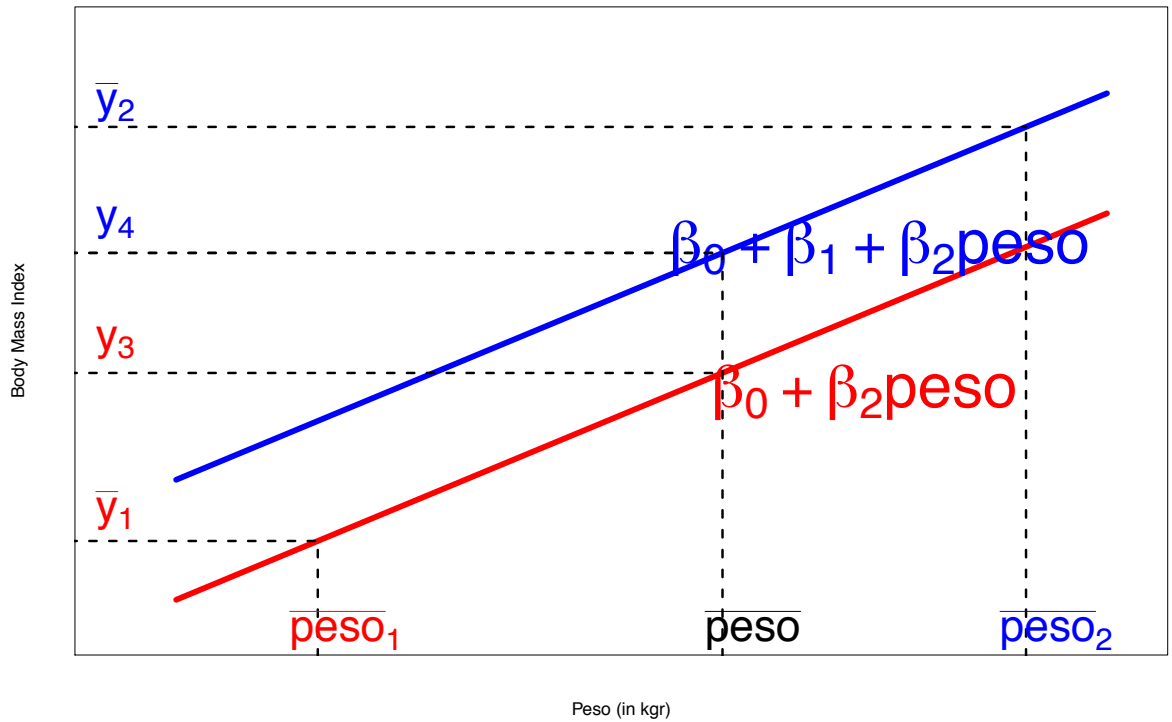
- Let us see the fitted lines for each group
- $\overline{peso}_1$ and $\overline{peso}_2$ are respectively the mean peso of each group (males/females).
- $\overline{peso}_1$ is the overall mean of peso



Interpretation of the parameters

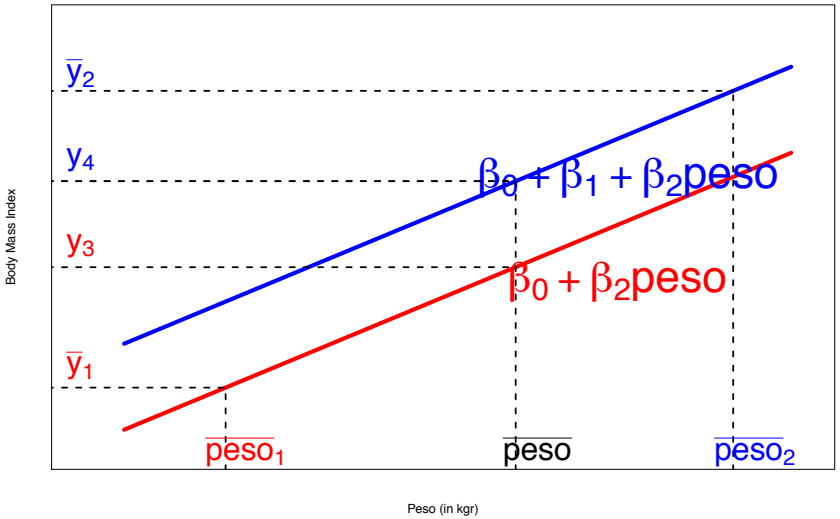
Categorical and Continuous independent variables

- Let us see the fitted lines for each group
- $\overline{peso}_1$ and $\overline{peso}_2$ are respectively the mean peso of each group (males/females).
- $\overline{peso}_1$ is the overall mean of peso



Interpretation of the parameters

Categorical and Continuous independent variables

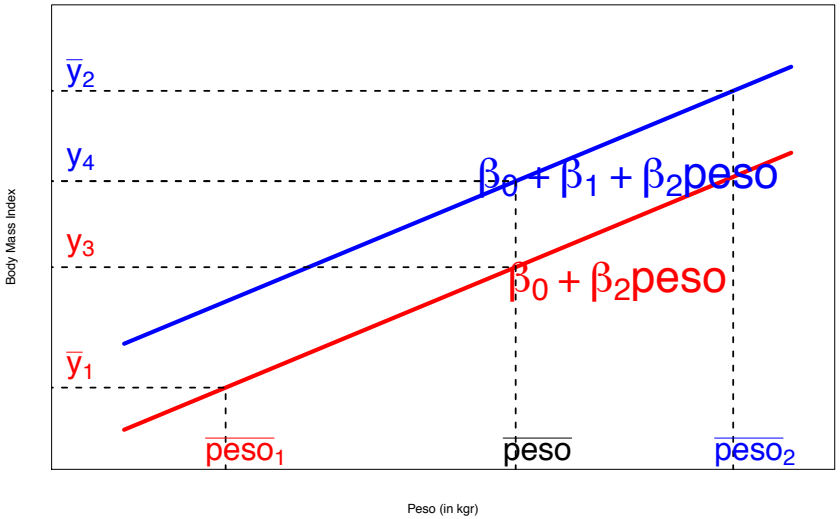


- Each line passes though the points $(\overline{peso}_1, y_1)$ and $(\overline{peso}_2, y_2)$.
- $(\bar{y}_1, \bar{y}_2)$ represent the average BMI of each group.
- Then

$$(\bar{y}_2 - \bar{y}_1) = \beta_1 + \beta_2(\overline{peso}_2 - \overline{peso}_1)$$

Interpretation of the parameters

Categorical and Continuous independent variables



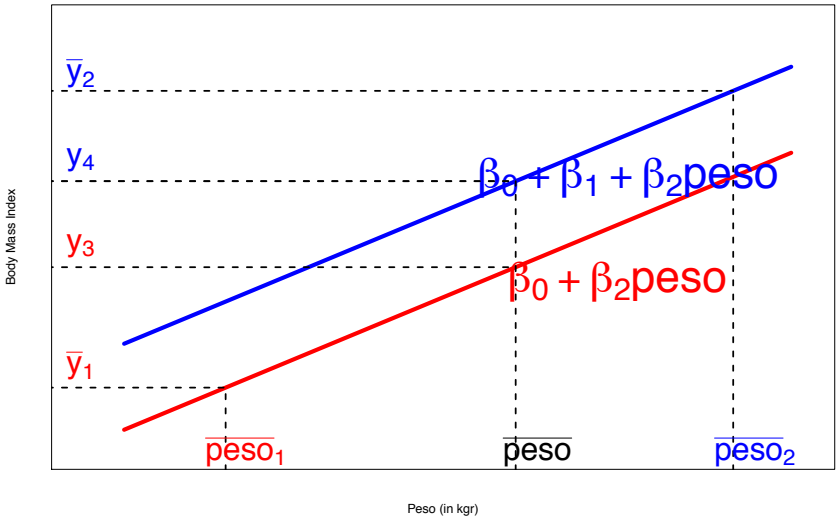
- Each line passes though the points $(\overline{peso}_1, y_1)$ and $(\overline{peso}_2, y_2)$.
- $(\bar{y}_1, \bar{y}_2)$ represent the average BMI of each group.
- Then

$$(\bar{y}_2 - \bar{y}_1) = \beta_1 + \beta_2(\overline{peso}_2 - \overline{peso}_1)$$

- This comparison includes:
  - Not only the difference between the groups β_1, and also
  - $\beta_2(\overline{peso}_2 - \overline{peso}_1)$ that represents the difference between group weights.

Interpretation of the parameters

Categorical and Continuous independent variables



- Each line passes though the points $(\overline{peso}_1, y_1)$ and $(\overline{peso}_2, y_2)$.
- $(\bar{y}_1, \bar{y}_2)$ represent the average BMI of each group.
- Then

$$(\bar{y}_2 - \bar{y}_1) = \beta_1 + \beta_2(\overline{peso}_2 - \overline{peso}_1)$$

- This comparison includes:
  - Not only the difference between the groups β_1, and also
  - $\beta_2(\overline{peso}_2 - \overline{peso}_1)$ that represents the difference between group weights.
- Modelling `imc` by `sexo` and `peso` allows for comparison for the same value of `peso`, and the value used is the mean `peso`, $\overline{peso}$, and then is equivalent to compare

$$(y_4 - y_3) = \beta_1 + \beta_2(\overline{peso}_2 - \overline{peso}_1) = \beta_1$$

Interpretation of the parameters

Categorical and Continuous independent variables

- Let us consider the health perception data and the logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * \text{sexo} + \beta_2 * \text{edad}$$

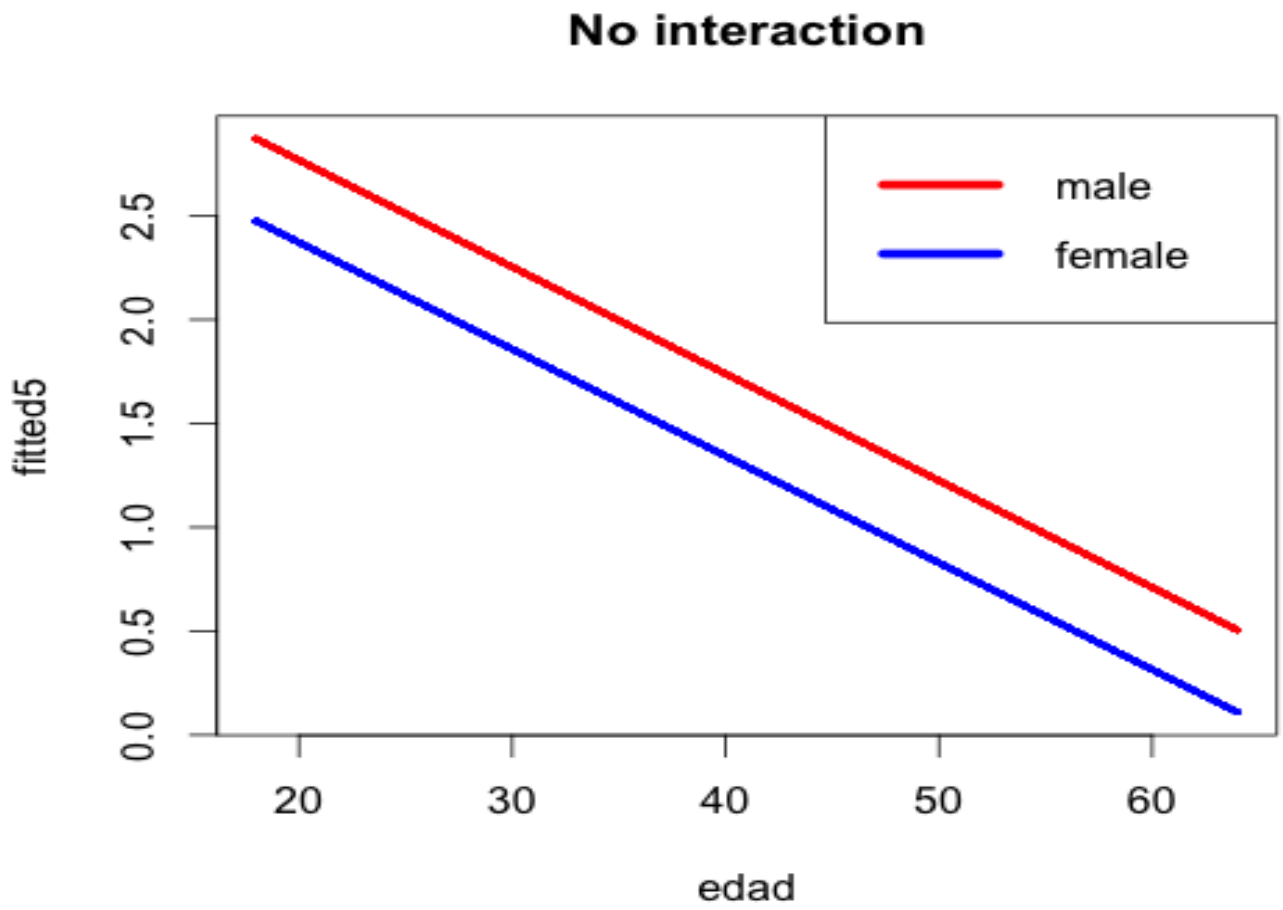
```
> logistic4<-glm(g02~sexo,family=binomial(link=logit))
> logistic5<-glm(g02~sexo+edad,family=binomial(link=logit))
> coefficients(logistic4)

(Intercept)  sexofemale
  1.6432048  -0.3775068

> coefficients(logistic5)

(Intercept)  sexofemale        edad
 3.79723317 -0.39588681 -0.05144576
```

- Both β_1 coefficients are similar because mean values of edad for male and females are similar ($39{,}16$ and $39,32$)

Interpretation of the parameters

Categorical and Continuous independent variables (cont.)

- Let us plot the fitted values (See `Logreg.R` script)



Interpretation of the parameters

Interaction and confusion

- **Confounding** term is used to describe that a covariate is related at the same time to the response variable and to the factor of interest.
- When these two relationships are present both the response and the factor are **confounded**.
- **NOTE:** an indicative of confounding is when the estimated parameters of the factor changes significantly when the covariate is included in the model. In our example, there is no significant change (age is not confounding).

Interpretation of the parameters

Interaction and confusion

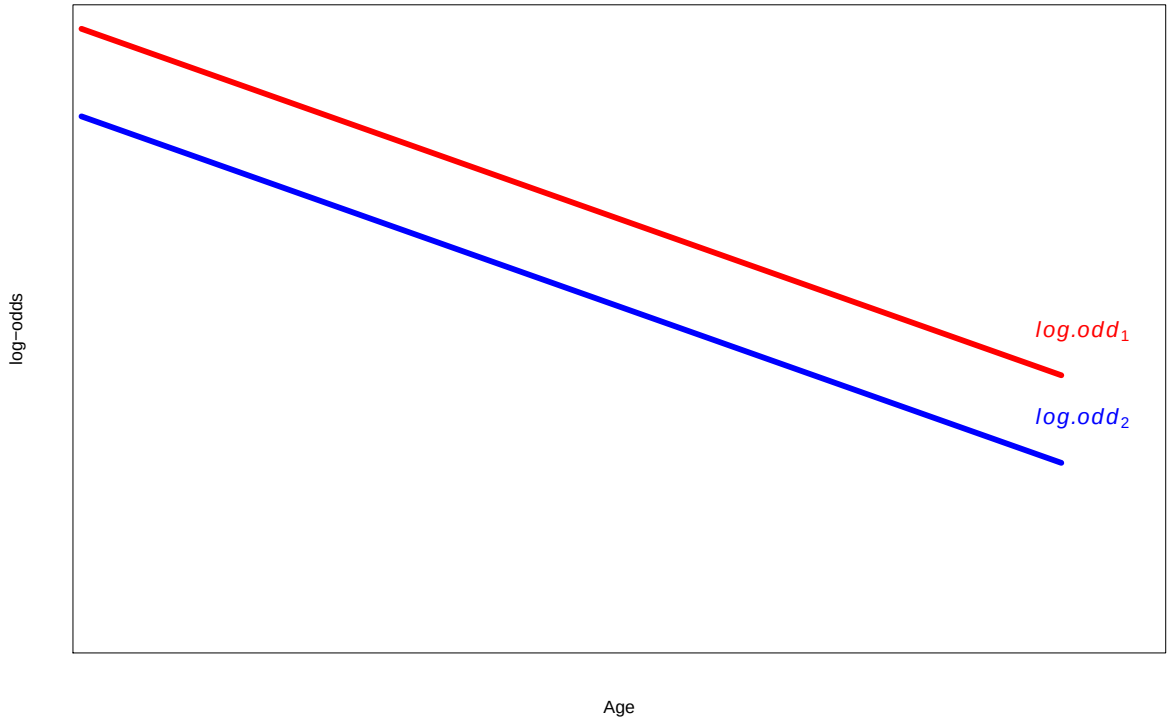
- **Confounding** term is used to describe that a covariate is related at the same time to the response variable and to the factor of interest.
- When these two relationships are present both the response and the factor are **confounded**.
- **NOTE:** an indicative of confounding is when the estimated parameters of the factor changes significantly when the covariate is included in the model. In our example, there is no significant change (age is not confounding).

```
> coefficients(logistic5)[2]

sexofemale
-0.3958868

> coefficients(logistic4)[2]

sexofemale
-0.3775068
```

Interpretation of the parameters

Categorical and Continuous independent variables

- Let us consider the factor `sexo` and the covariate `edad`.

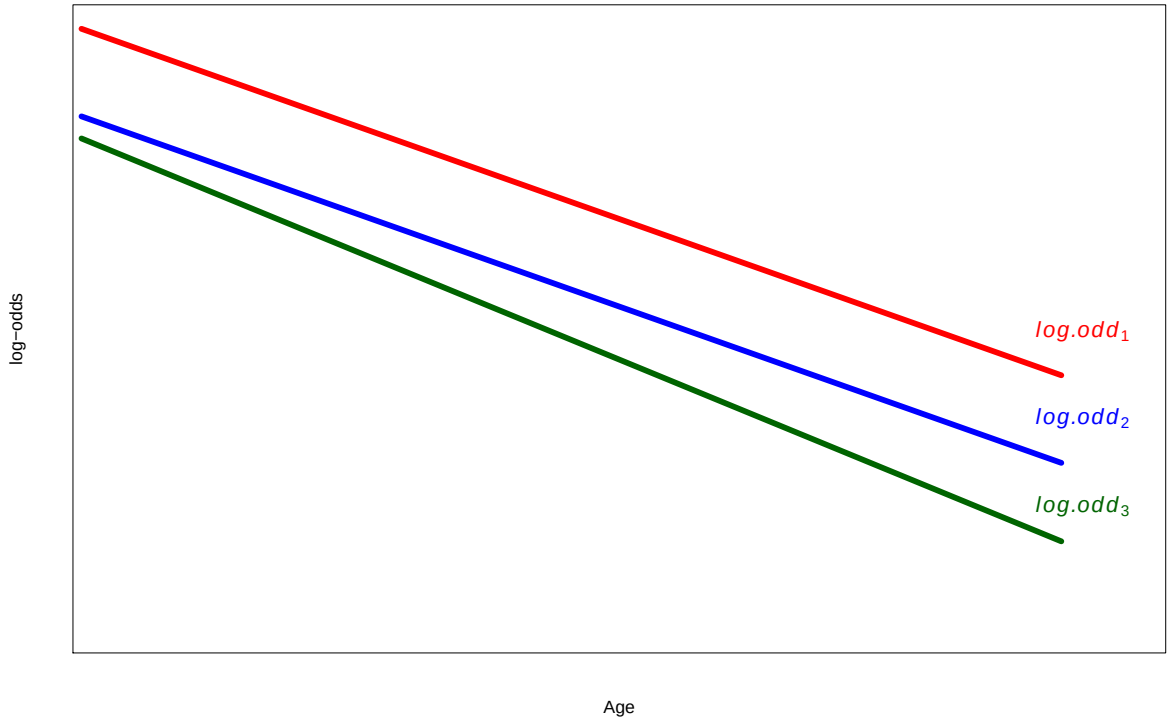


- two parallel lines (for each factor level), indicate **no-interaction**

Interpretation of the parameters

Categorical and Continuous independent variables

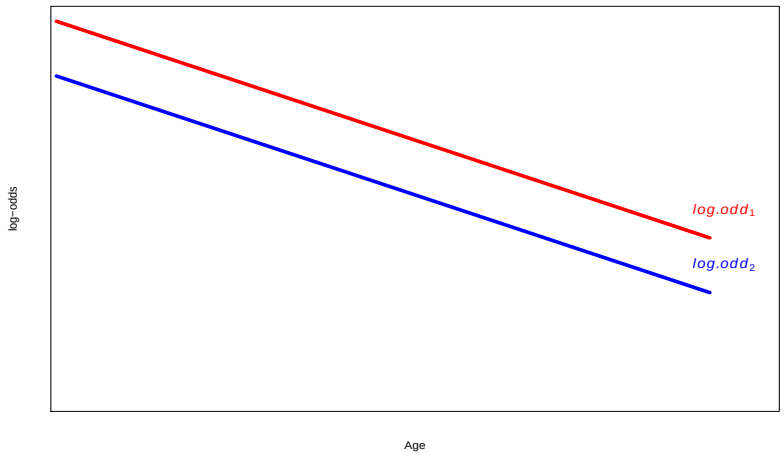
- Let us consider the factor `sexo` and the covariate `edad`.



- two parallel lines (for each factor level), indicate **no-interaction**
- When there is **interaction**, the association between the factor `sexo` and the response depends on the level of the covariate (`edad`), i.e. the effect of the covariate `edad` modifies the effect of the factor `sexo`.

Interpretation of the parameters

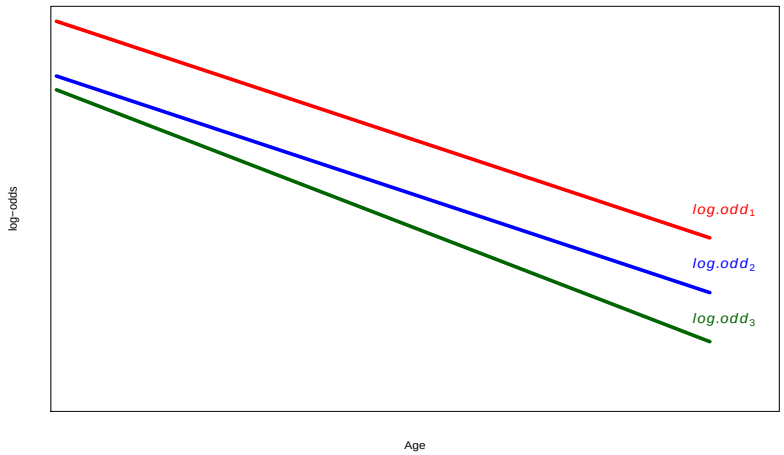
Categorical and Continuous independent variables



- $log.odds_1$ is the logit for females given age, $log.odds_2$ the equivalent for males. They are parallel, i.e. the relationship between age and health perception is the same for both sexes. The log-odds ratio by age is the difference $log.odds_1 - log.odds_2$

Interpretation of the parameters

Categorical and Continuous independent variables



- $log.odds_1$ is the logit for females given age, $log.odds_2$ the equivalent for males. They are parallel, i.e. the relationship between age and health perception is the same for both sexes. The log-odds ratio by age is the difference $log.odds_1 - log.odds_2$
- Suppose $log.odds_3$ is the logit for females, this means that the relationship between the health perception is different for males and females (**interaction**). The log-odds ratio by age depends on the value of age (effect modification of age).

Interpretation of the parameters

Interaction and confusion

- Let us fit a model for health perception **with interaction**

Interpretation of the parameters

Interaction and confusion

- Let us fit a model for health perception **with interaction**

```
> logistic7<-glm(g02~sexo+edad+sexo:edad,family=binomial(link=logit))
> summary(logistic7)

Call:
glm(formula = g02 ~ sexo + edad + sexo:edad, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3857   0.3889   0.5161   0.6780   1.1544

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.653489   0.170571  21.419   <2e-16 ***
sexofemale     -0.135450   0.230287  -0.588    0.556
edad           -0.048200   0.003721 -12.952   <2e-16 ***
sexofemale:edad -0.005918   0.005046  -1.173    0.241
---
Signif. codes:  0
```

Interpretation of the parameters

Interaction and confusion

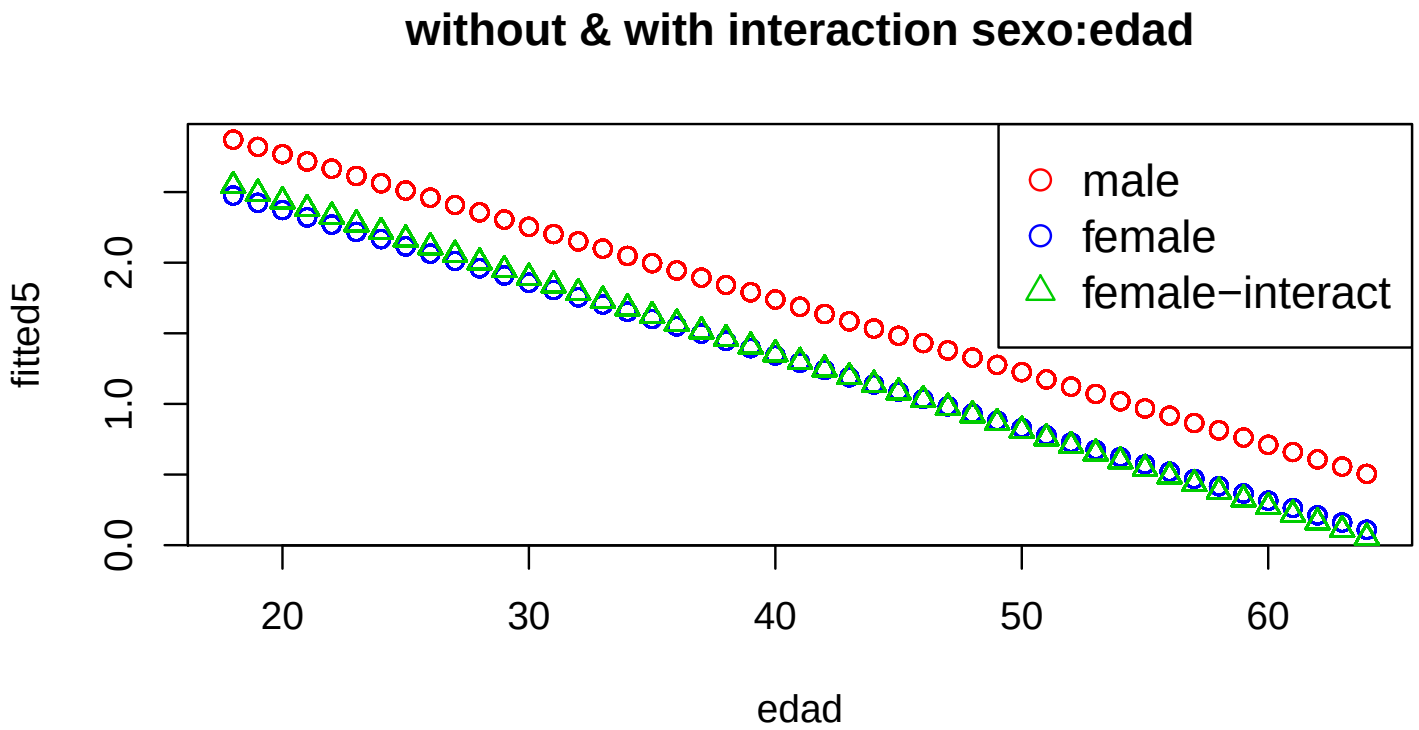
- The `sexo:edad` coefficient **is not significant**
- The coefficient for `sexofemale` changed and became not significant.
- **NOTE:** In this case, we cannot use the criteria and say that `sexofemale` is confounding, because including the interaction term (specially when the covariate is continuous) usually changes the estimation of the parameters, even if the interaction is not significative.

```
> coefficients(logistic7)

    (Intercept)      sexofemale            edad sexofemale:edad
    3.653488923    -0.135449547    -0.048200283    -0.005918381
```

Interpretation of the parameters

Interaction and confusion

- Graphically, the inclusion of an interaction term does not change the slope.



* Hence, the covariate edad is not counfounding or modication effect.

Interpretation of the parameters

Interaction and confusion

- Let us include now the variable peso instead of edad:

without interaction

```
> logistic8<-glm(g02~sexo+peso,family=binomial(link=logit))
> summary(logistic8)

Call:
glm(formula = g02 ~ sexo + peso, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2849   0.5139   0.6055   0.6773   1.7336

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.712536   0.218700  16.975   <2e-16 ***
sexofemale  -0.825937   0.076422 -10.808   <2e-16 ***
peso        -0.026188   0.002671  -9.803   <2e-16 ***
---
Signif. codes:  0
```

Interpretation of the parameters

Interaction and confusion

- Compare with `logistic4`:

```
> coefficients(logistic4)

(Intercept)  sexofemale
  1.6432048  -0.3775068

> coefficients(logistic8)

(Intercept)  sexofemale        peso
 3.71253575 -0.82593673 -0.02618773
```

- The inclusion of the covariate peso changes the estimation of the coefficient of sexo from $-0{,}3775$ to $-0{,}8259$ (a decrease of $> 50\,\%$). This is an indicative that peso can be confounding.

Interpretation of the parameters

Interaction and confusion

- Compare with `logistic4`:

```
> coefficients(logistic4)

(Intercept)  sexofemale
  1.6432048  -0.3775068

> coefficients(logistic8)

(Intercept)  sexofemale        peso
 3.71253575 -0.82593673 -0.02618773
```

- The inclusion of the covariate peso changes the estimation of the coefficient of sexo from $-0{,}3775$ to $-0{,}8259$ (a decrease of $> 50\,\%$). This is an indicative that peso can be confounding.

with interaction

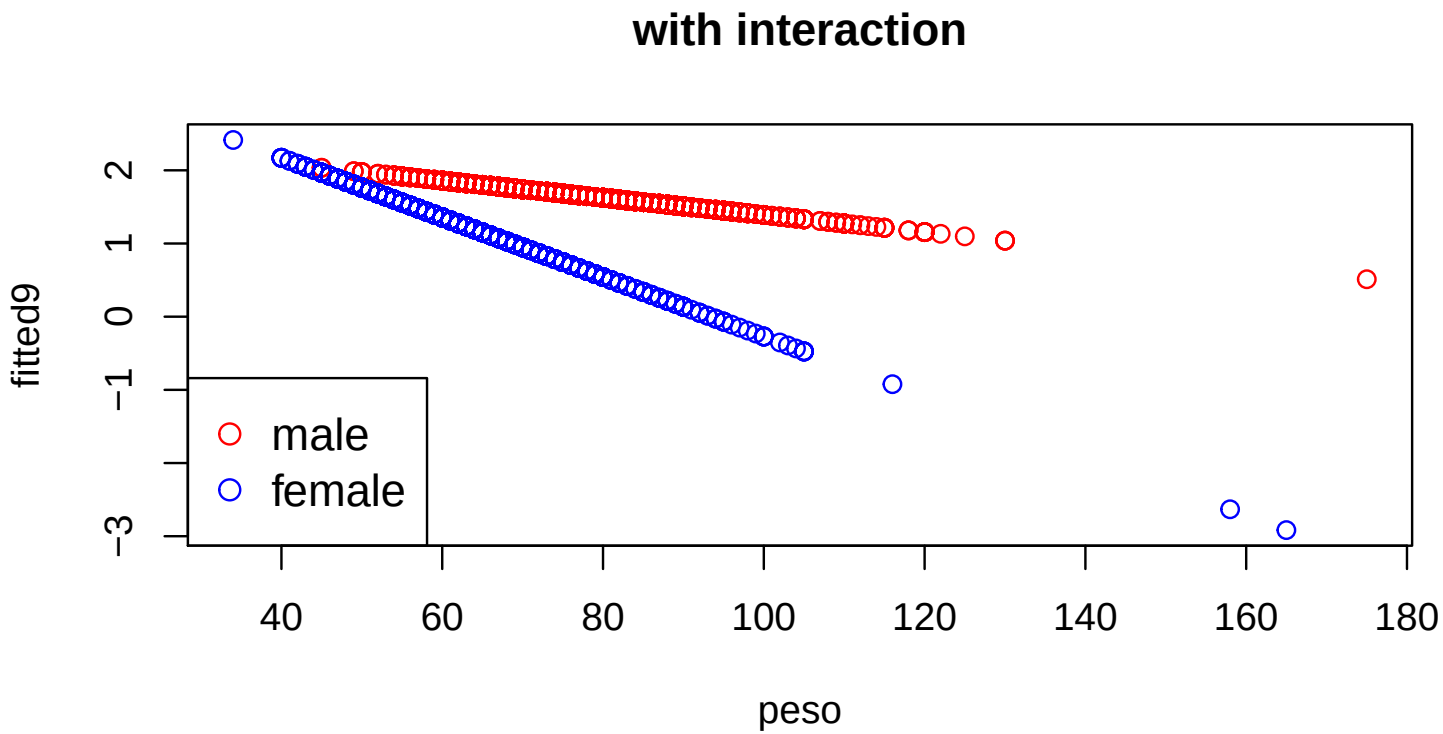
```
> logistic9<-glm(g02~sexo+peso+sexo:peso,family=binomial(link=logit))
```

basque center for applied mathematics

Interpretation of the parameters

Interaction and confusion

- Let us plot the fitted values



- The health perception by `peso` is different for `males` and `females`. Females are more affected by covariate `peso`.

Interpretation of the parameters

Interaction and confusion

- In summary, to determine if a variable is a confounder and/or an effect modifier depends on various aspects.
  - A **confounder** variable must verify 2 conditions:
    1. The covariate has to be associated to the response, i.e. the estimated coefficient must be significantly $\neq 0$.
    2. The covariate has to be associated to the risk factor.
  - To determine an **effect modifier**, we must look at the parametric structure of the logit.

Interpretation of the parameters

Interaction and confusion

- In summary, to determine if a variable is a confounder and/or an effect modifier depends on various aspects.
  - A **confounder** variable must verify 2 conditions:
    1. The covariate has to be associated to the response, i.e. the estimated coefficient must be significantly $\neq 0$.
    2. The covariate has to be associated to the risk factor.
  - To determine an **effect modifier**, we must look at the parametric structure of the logit.
- **In practice**
  - To evaluate a confounder: compare the estimated coefficient for the risk factor for the models with and without the covariate. Any significant change in the coefficient suggests that the covariate is a confounder. If this happens and the interaction is not statistically significant, do not include the variable in your model.
  - A variable is an effect modifier, only when the interaction term is statistically significant.

Interpretation of the parameters

Odds-ratio in the presence of an interaction

- When there exists an interaction between a risk factor and another variable $\Rightarrow$ the estimated parameter for the risk factor depends on the variable that interacts with it.

Interpretation of the parameters

Odds-ratio in the presence of an interaction

- When there exists an interaction between a risk factor and another variable $\Rightarrow$ the estimated parameter for the risk factor depends on the variable that interacts with it.
- We **cannot** obtain the odds-ratio (OR) by taking exponentials

Solution:

- Write the equation of the logit for both levels of the risk factor
- Compute the difference between the logits
- Take the exponential of the obtain difference

Interpretation of the parameters

Odds-ratio in the presence of an interaction

- Consider a model with 2 variables and their interaction.
- Let us say, a factor `F` and a covariate `X` and `F` $\times$ `X`
- The logit for `F` $= f$ and `X` $= x$ is

$$\log\left(\frac{p(f,x)}{1-p(f,x)}\right) = \beta_0 + \beta_1 f + \beta_2 x,$$

- The **OR** for the two levels of the factor `F`, i.e. f_1 versus f_2 is

Interpretation of the parameters

Odds-ratio in the presence of an interaction

- Consider a model with 2 variables and their interaction.
- Let us say, a factor `F` and a covariate `X` and `F` $\times$ `X`
- The logit for `F` $= f$ and `X` $= x$ is

$$\log\left(\frac{p(f,x)}{1-p(f,x)}\right) = \beta_0 + \beta_1 f + \beta_2 x,$$

- The **OR** for the two levels of the factor `F`, i.e. f_1 versus f_2 is

$$\log\left(\frac{p(f_1,x)}{1-p(f_1,x)}\right) = \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 x, \quad (1)$$

Interpretation of the parameters

Odds-ratio in the presence of an interaction

- Consider a model with 2 variables and their interaction.
- Let us say, a factor `F` and a covariate `X` and `F` $\times$ `X`
- The logit for `F` $= f$ and `X` $= x$ is

$$\log\left(\frac{p(f,x)}{1-p(f,x)}\right) = \beta_0 + \beta_1 f + \beta_2 x,$$

- The **OR** for the two levels of the factor `F`, i.e. f_1 versus f_2 is

$$\log\left(\frac{p(f_1,x)}{1-p(f_1,x)}\right) = \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 x, \quad (1)$$

$$\log\left(\frac{p(f_0,x)}{1-p(f_0,x)}\right) = \beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 x, \quad (2)$$

Interpretation of the parameters

Odds-ratio in the presence of an interaction

- Consider a model with 2 variables and their interaction.
- Let us say, a factor `F` and a covariate `X` and `F` $\times$ `X`
- The logit for `F` $= f$ and `X` $= x$ is

$$\log\left(\frac{p(f,x)}{1-p(f,x)}\right) = \beta_0 + \beta_1 f + \beta_2 x,$$

- The **OR** for the two levels of the factor `F`, i.e. f_1 versus f_2 is

$$\log\left(\frac{p(f_1,x)}{1-p(f_1,x)}\right) = \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 x, \quad (1)$$

$$\log\left(\frac{p(f_0,x)}{1-p(f_0,x)}\right) = \beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 x, \quad (2)$$

$$\log\left(\frac{p(f_1,x)}{1-p(f_1,x)}\right) - \log\left(\frac{p(f_0,x)}{1-p(f_0,x)}\right) = \beta_1(f_1 - f_0) + \beta_3 x(f_1 - f_0) \quad (3)$$

Interpretation of the parameters

Odds-ratio in the presence of an interaction

- Consider a model with 2 variables and their interaction.
- Let us say, a factor `F` and a covariate `X` and `F` $\times$ `X`
- The logit for `F` $= f$ and `X` $= x$ is

$$\log\left(\frac{p(f,x)}{1-p(f,x)}\right) = \beta_0 + \beta_1 f + \beta_2 x,$$

- The **OR** for the two levels of the factor `F`, i.e. f_1 versus f_2 is

$$\log\left(\frac{p(f_1,x)}{1-p(f_1,x)}\right) = \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 x, \quad (1)$$

$$\log\left(\frac{p(f_0,x)}{1-p(f_0,x)}\right) = \beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 x, \quad (2)$$

$$\log\left(\frac{p(f_1,x)}{1-p(f_1,x)}\right) - \log\left(\frac{p(f_0,x)}{1-p(f_0,x)}\right) = \beta_1(f_1-f_0) + \beta_3 x(f_1-f_0) \quad (3)$$

$$\text{OR} = \exp\left(\beta_1(f_1-f_0) + \beta_3 x(f_1-f_0)\right) \quad (4)$$

Interpretation of the parameters

Odds-ratio in the presence of an interaction

- In our health perception example, we have:

```
> coefficients(logistic9)
    (Intercept)       sexofemale             peso sexofemale:peso
     2.56152098       1.23683081      -0.01171294      -0.02898403
```

- The comparison between `females` and `males` is $f_1 - f_0$, hence

$$\text{OR} = \exp\left(1{,}237 - 0{,}029x\right)$$

Interpretation of the fitted values

- In general, in logistic regression we are interested in the coefficients and the Odd-Ratios.
- Sometimes, fitted values are also of interest.

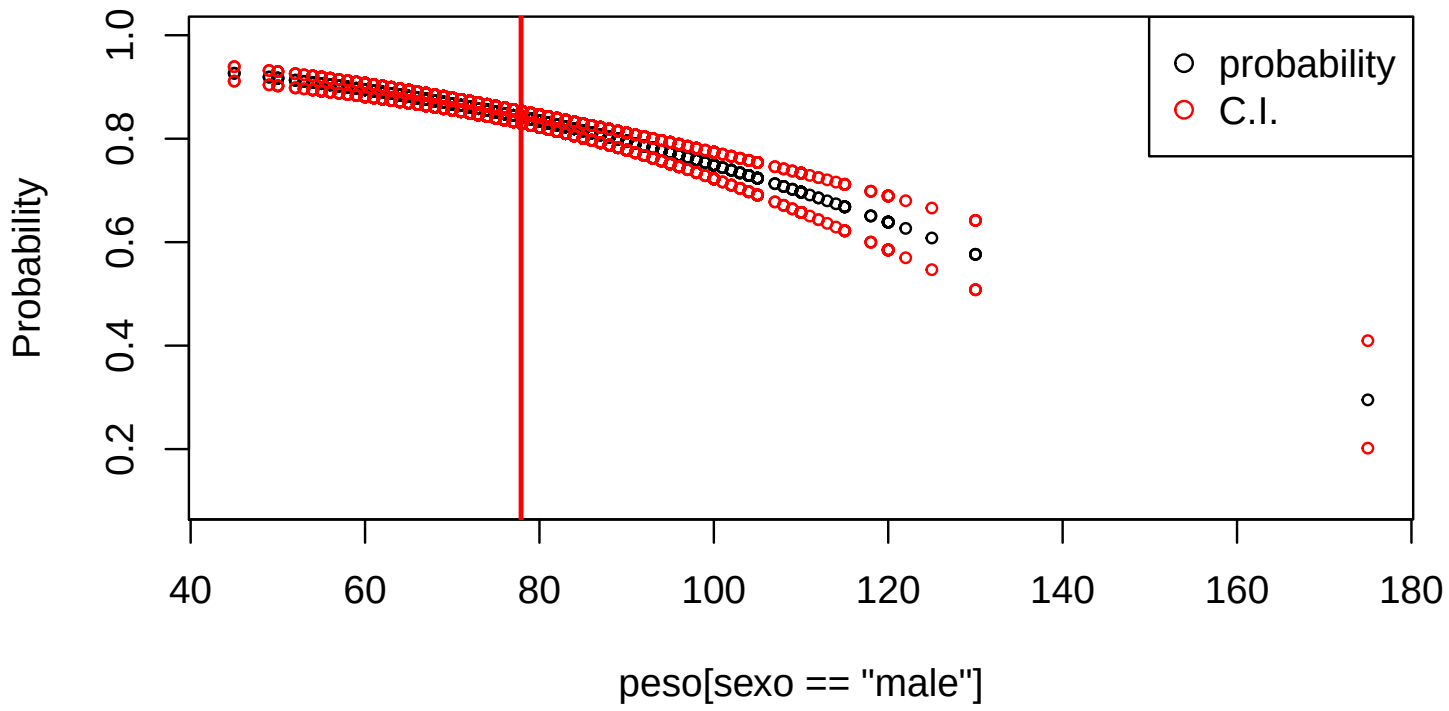
Interpretation of the fitted values

- In general, in logistic regression we are interested in the coefficients and the Odd-Ratios.
- Sometimes, fitted values are also of interest.
- For instance, the computation of the **confidence intervals (C.I.'s)** of the logit is easy. We can obtain the standard errors from the fitted model.
- To compute the **C.I.**'s of the probabilities, we only need to use the relationship between the logit and the probability, i.e.:

$$\frac{e^{\text{C.I. logit}}}{1 + e^{\text{C.I. logit}}}$$

Interpretation of the parameters

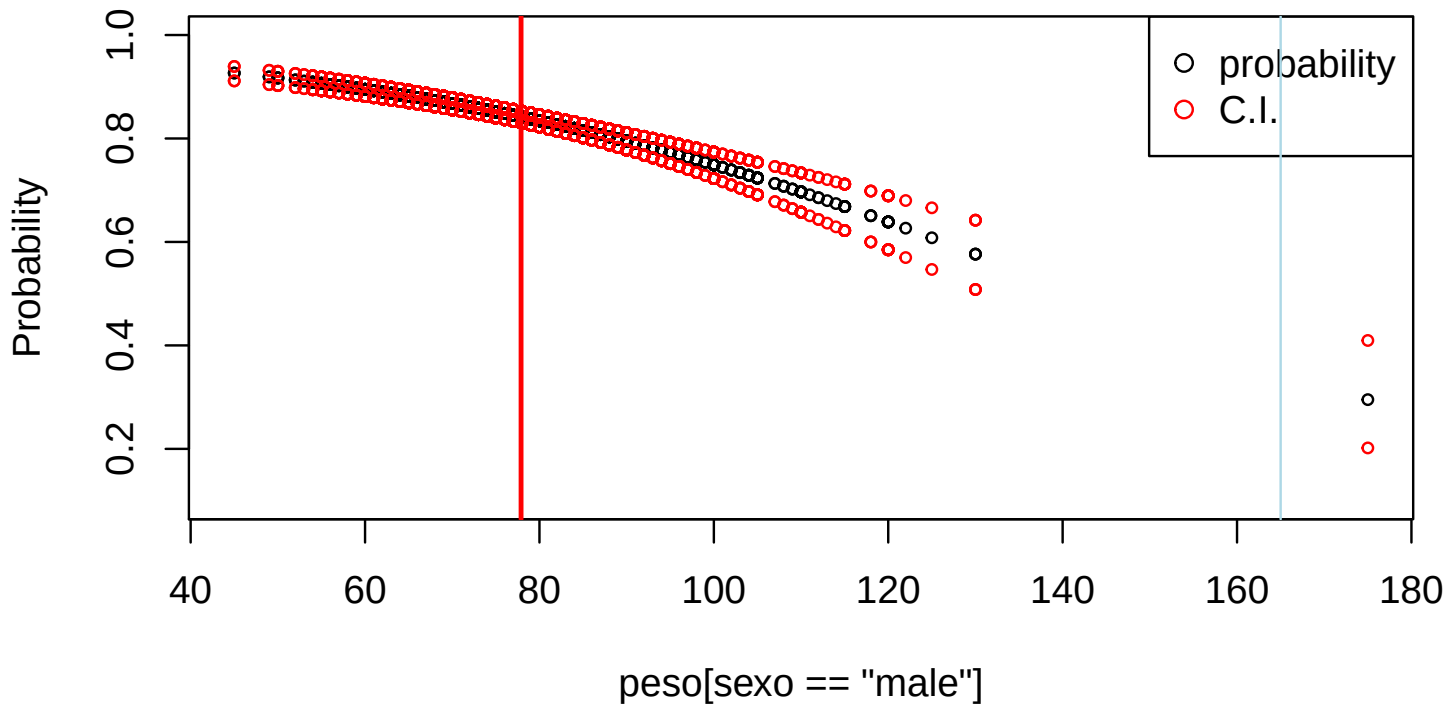
Interaction and confusion



- The vertical line is the average weight for males.
- Each black point represents the average of the response (health perception) for males given a value of weight.
- C.I.'s are wider at the end of the range because there are less observations.

Interpretation of the parameters

Interaction and confusion



- The estimated proportion of males with good health perception and 80 kgr of weight is $0{,}834$ within a $95\,\%$ confidence interval $\in (0{,}822, 0{,}847)$. **Q: And for males with weight of 165 kgr.?**
- A common error is to assume that these estimates of the probability corresponds to individual subjects, however, they represent a $\%$ of subjects (from an *unknown* population).

Variable selection

- Once we have estimated the parameters, we need to determine which variables of the model are significant or not.

Variable selection

- Once we have estimated the parameters, we need to determine which variables of the model are significant or not.
- This means the formulation of hypothesis tests to see if the $X's$ are significantly related to the response y

Variable selection

- Once we have estimated the parameters, we need to determine which variables of the model are significant or not.
- This means the formulation of hypothesis tests to see if the $X's$ are significantly related to the response y
- An important question to solve is: **Is the model that includes the variable giving more information than the one that does not include the variable?**

Variable selection

- Once we have estimated the parameters, we need to determine which variables of the model are significant or not.
- This means the formulation of hypothesis tests to see if the $X's$ are significantly related to the response y
- An important question to solve is: **Is the model that includes the variable giving more information than the one that does not include the variable?**
- **A:** The answer is based in a criteria that compares y_i and $\hat{y}_i$
- **A:** If the fitted values $\hat{y}_i$ obtained from a model that includes X are *'better'* than the ones obtain excluding the X predictor, we say that the variable X is statistically *significant*. But what is 'better'?

* **NOTE:** Do not confuse significance of a predictor variable with goodness-of-fit of the model (in the latter we look at the adecuacy of the fitted values in a absolute sense).

Variable selection

in a GLM: logistic regression

- In a GLM the comparison of the models is based on the log-likelihood,

$$\log \mathcal{L}(\beta) = \sum (y_i \log(p_i) + (n_i - p_i) \log(1 - p_i))$$

- The **saturated model** is the model that has as many parameters as data values, i.e. $\hat{y}_i = y_i$. Then the comparison between the observed values and the fitted values is done with the **Deviance** ($\mathcal{D}$):

$$\mathcal{D} = -2 \log \underbrace{\left[\frac{\text{Likelihood of the fitted model}}{\text{Likelihood of the saturated model}} \right]}_{(*)}$$

the value $(*)$ is called **Likelihood ratio**. We take logarithms and multiply by -2 to obtain that $\mathcal{D} \sim \chi^2$ (Chi-square distribution)

- Then, we can use the χ^2 for Hypothesis Testing: **Likelihood Ratio Test (LRT)**

Variable selection

LRT: logistic regression

- In logistic regression, the Deviance ($\mathcal{D}$) is:

$$-2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{p}(x_i)}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{p}(x_i)}{1-y_i}\right)\right]$$

- When the response variable is dichotomous, the likelihood of the saturated model is 1, and hence in logistic regression the Deviance is

$$\mathcal{D} = -2\log\left[\text{Likelihood of the fitted model}\right]$$

Variable selection

LRT: logistic regression

- In logistic regression, the Deviance ($\mathcal{D}$) is:

$$-2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{p}(x_i)}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{p}(x_i)}{1-y_i}\right)\right]$$

- When the response variable is dichotomous, the likelihood of the saturated model is 1, and hence in logistic regression the Deviance is

$$\mathcal{D} = -2\log\left[\text{Likelihood of the fitted model}\right]$$

- In order to check if a variable is significant, we compare the value of the Deviance $\mathcal{D}$ with and without the predictor:

$$\begin{aligned} G &= -2\log\left[\frac{\text{Likelihood of the model without the predictor}}{\text{Likelihood of the model with the predictor}}\right] \\ &= \mathcal{D}(\text{model without the predictor}) - \mathcal{D}(\text{model with the predictor}) \end{aligned}$$

Variable selection

Example: variable bebedor

```
> anova(logistic4,logistic4a,test="Chisq")


Analysis of Deviance Table

Model 1: g02 ~ sexo
Model 2: g02 ~ sexo + bebedor
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      7355     7137.8
2      7353     7059.2  2   78.579 < 2.2e-16 ***
---
Signif. codes:  0
```

Another alternative test is the **Wald test**

Variable selection

Wald test

- The Wald test can be used to test the true value of the parameter based on the sample estimate. (Theoretical results skipped)

`wald.test`

```
> library(aod)
> wald.test(coef(logistic4a), Sigma=vcov(logistic4a),Terms=3:4)

Wald test:
----------


Chi-squared test:
X2 = 78.5, df = 2, P(> X2) = 0.0
```

- The order in which the coefficients are given in the table of coefficients is the same as the order of the terms in the model. `Terms=3:4` indicates the coefficients for the variable `bebedor` levels.
- The χ^2 statistic of $78{,}5$ with 2 degrees of freedom is associated to a p-value of 0, indicating that the overall effect of `bebedor` is statistically significant.

Variable selection

Wald test

- We can also test additional hypotheses about the differences in the coefficients for the different levels of `bebedor`.

`wald.test`

```
> l <- cbind(0, 0, 1, -1)
> wald.test(coef(logistic4a), Sigma=vcov(logistic4a),L=l)

Wald test:
----------


Chi-squared test:
X2 = 0.13, df = 1, P(> X2) = 0.72
```

- `l` is a vector that defines the test we want to perform. We want to test the difference of the terms `bebedorocasional` and `bebedorfrecuente`
- The χ^2 statistic of $0{,}13$ with 1 degree of freedom is associated to a p-value of $0{,}72$, indicating that the difference between the coefficient for `bebedorocasional` and `bebedorfrecuente` is statistically not significant.

Variable selection

Interactions

- Recall that **interaction** between two variables is present when the effect of one of the two variables is not constant.
- Before including all possible interactions think of its scientific meaning.
- Include interaction one by one, and use LRT to test its statistical significance.
- An not-significant interaction does not change too much the estimated parameters but increases the standard error.

Variable selection

Interactions

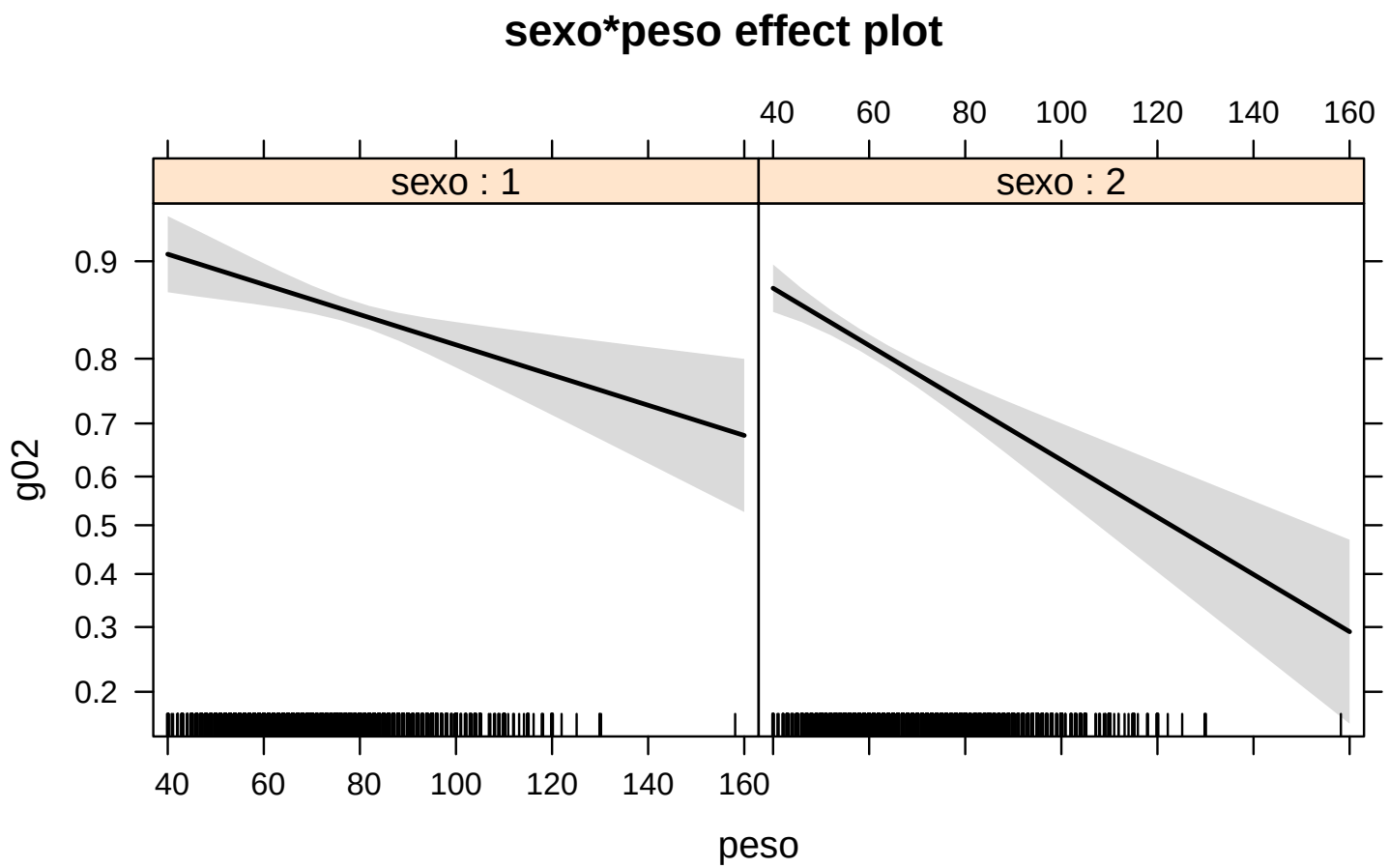
- Recall that **interaction** between two variables is present when the effect of one of the two variables is not constant.
- Before including all possible interactions think of its scientific meaning.
- Include interaction one by one, and use LRT to test its statistical significance.
- An not-significant interaction does not change too much the estimated parameters but increases the standard error.

```
> anova(logistic12,logistic13,test="Chisq")

Analysis of Deviance Table

Model 1: g02 ~ educa + edad2 + sexo + peso
Model 2: g02 ~ educa + edad2 + sexo * peso
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      7349       6464
2      7348       6460  1   4.0151  0.04509 *
---
Signif. codes:  0
```

Variable selection

Interactions

- We can plot the interaction effects of the variables using the function `effect` in `library(effects)`



Variable selection

stepwise procedures

- We can choose a model by AIC in a Stepwise Algorithm
- See `?stepAIC` in `library(MASS)`

let us include the variable `anio`

```
> mod <- glm(formula=g02~educa+edad2+bebedor+sexo*peso+anio, family = binomial(link =
> library(MASS)
> stepAIC(mod,trace=FALSE)$anova
```

Goodness-of-fit

- The **goodness of fit** of a statistical model describes how well it fits a set of observations.
- Measures of goodness of fit typically summarize the discrepancy between observed values (y_i) and the values expected under the model in question ($\hat{y}_i$).

Goodness-of-fit

- The **goodness of fit** of a statistical model describes how well it fits a set of observations.
- Measures of goodness of fit typically summarize the discrepancy between observed values (y_i) and the values expected under the model in question ($\hat{y}_i$).
- **Pearson-χ^2 and Deviance:**
  - Both are measures of the difference between y_i and $\hat{y}_i$ based on the model residuals. The **Pearson residuals** are defined as

$$r(y_j, \hat{p}_j) = \frac{y_j - m_j \hat{p}_j}{\sqrt{m_j \hat{p}_j (1 - \hat{p}_j)}}$$

    where m_j es the number of individuals that share the same covariates pattern, i.e. the same combinations of predictors.
  - The **Pearson-χ^2** is

$$\chi^2 = \sum_{j=1}^{J} r(y_j, \hat{p}_j)^2,$$

    where J es the number of different patterns.
    The **standardized Pearson residuals** are distributed as a $\mathcal{N}(0, 1)$, and hence should be within the interval $(-3, 3)$.

Goodness-of-fit

GoF

- **Deviance residuals:**
  - are based on the definition of Deviance ($\mathcal{D}$)

- In both cases, in general a value larger of 4 of the statistic indicates a bad fitting of the model.
- The use of these tests is recommended when the number of observations n is much larger than the number of patterns.

Goodness-of-fit

GoF

- **Deviance residuals:**
  - are based on the definition of Deviance ($\mathcal{D}$)

- In both cases, in general a value larger of 4 of the statistic indicates a bad fitting of the model.
- The use of these tests is recommended when the number of observations n is much larger than the number of patterns.

Consider `logistic1`

```
> sum(residuals(logistic1, type = "pearson")^2)
> sum(residuals(logistic1, type = "deviance")^2)
```

Goodness-of-fit (cont.)

- If any of the covariates is continuous, we cannot use these test, and instead we use the so-called **Hosmer-Lemeshow** test for logistic regression models.

Consider `logistic13`

```
> library(ResourceSelection)
> g02num <- as.numeric(g02)-1 # transform to numeric 0/1
> hoslem.test(g02num,fitted(logistic13))

Hosmer and Lemeshow goodness of fit (GOF) test

data:  g02num, fitted(logistic13)
X-squared = 6.1108, df = 8, p-value = 0.6348
```

- The p-value of indicates that the GoF is ok (a large p-value indicates no evidence of poor fit)

Logistic regression model predictions

Classification or Contingency tables

Given that we estimate probabilities. how can we translate this into a predicted outcome?

Logistic regression model predictions

Classification or Contingency tables

Given that we estimate probabilities. how can we translate this into a predicted outcome?
Two possibilities for prediction rules are:

1. Use $0{,}5$ as a **cutoff**. That is if the predicted probability is greater than $0{,}5$, its predicted outcome is 1, otherwise is 0. This approach is reasonable when:
   a) it is equally likely in the population of interest that an outcome 0 or 1 occur.
   b) The cost of incorrectly 0 and 1 are approximately the same.

Logistic regression model predictions

Classification or Contingency tables

Given that we estimate probabilities. how can we translate this into a predicted outcome?
Two possibilities for prediction rules are:

1. Use $0{,}5$ as a **cutoff**. That is if the predicted probability is greater than $0{,}5$, its predicted outcome is 1, otherwise is 0. This approach is reasonable when:
   a) it is equally likely in the population of interest that an outcome 0 or 1 occur.
   b) The cost of incorrectly 0 and 1 are approximately the same.
2. Find the **best cutoff** for the data set. Using this approach, we evaluate different cutoff values and calculate for each of them the proportion of observations incorrectly classified. Then we select the cutoff that minimized the proportion of misclassified observations. This approach is reasonable when:
   a) The data set is a random sample from the population of interest.
   b) The cost of incorrectly 0 and 1 are approximately the same.

Logistic regression model predictions

Classification or Contingency tables (cont.)

- Contingency tables must not be used to compare models, because they depend on the probability distribution of the sample probabilities they are based.
- The same model evaluated in different samples may lead to different classifications.

Logistic regression model predictions

Classification or Contingency tables (cont.)

- Contingency tables must not be used to compare models, because they depend on the probability distribution of the sample probabilities they are based.
- The same model evaluated in different samples may lead to different classifications.

- We create the following table, where $\hat{y}_i = 1$ if $\hat{\pi}_i > s$, where s is the cutoff point.

Classification or Contigency Table:

Logistic regression model predictions

Classification or Contingency tables (cont.)

- Contingency tables must not be used to compare models, because they depend on the probability distribution of the sample probabilities they are based.
- The same model evaluated in different samples may lead to different classifications.

- We create the following table, where $\hat{y}_i = 1$ if $\hat{\pi}_i > s$, where s is the cutoff point.

Classification or Contigency Table:

	$y = 1$	$y = 0$	**Total**
$\hat{y}_i = 1$	a	b	$a + b$
$\hat{y}_i = 0$	c	d	$c + d$
	$a + c$	$b + d$	n

Logistic regression model predictions

Area under the ROC curve

	$y = 1$	$y = 0$	**Total**
$\hat{y}_i = 1$	a	b	$a + b$
$\hat{y}_i = 0$	c	d	$c + d$
	$a + c$	$b + d$	n

We define

- The **sensitivity** is the proportion of true 1's estimated as 1's: $Ss = a/(a + c)$.

Logistic regression model predictions

Area under the ROC curve

	$y = 1$	$y = 0$	**Total**
$\hat{y}_i = 1$	a	b	$a + b$
$\hat{y}_i = 0$	c	d	$c + d$
	$a + c$	$b + d$	n

We define

- The **sensitivity** is the proportion of true 1's estimated as 1's: $Ss = a/(a + c)$.
  (probability of predicting a 1 correctly)

Logistic regression model predictions

Area under the ROC curve

	$y = 1$	$y = 0$	**Total**
$\hat{y}_i = 1$	a	b	$a + b$
$\hat{y}_i = 0$	c	d	$c + d$
	$a + c$	$b + d$	n

We define

- The **sensitivity** is the proportion of true 1's estimated as 1's: $Ss = a/(a + c)$.
  (probability of predicting a 1 correctly)
- The **specificity** is the proportion of true 0's estimated as 0's: $Sp = d/(b + d)$.

Logistic regression model predictions

Area under the ROC curve

	$y = 1$	$y = 0$	**Total**
$\hat{y}_i = 1$	a	b	$a + b$
$\hat{y}_i = 0$	c	d	$c + d$
	$a + c$	$b + d$	n

We define

- The **sensitivity** is the proportion of true 1's estimated as 1's: $Ss = a/(a + c)$.
  (probability of predicting a 1 correctly)
- The **specificity** is the proportion of true 0's estimated as 0's: $Sp = d/(b + d)$.
  (probability of predicting a 0 correctly)

Logistic regression model predictions

Area under the ROC curve

	$y = 1$	$y = 0$	**Total**
$\hat{y}_i = 1$	a	b	$a + b$
$\hat{y}_i = 0$	c	d	$c + d$
	$a + c$	$b + d$	n

We define

- The **sensitivity** is the proportion of true 1's estimated as 1's: $Ss = a/(a + c)$.
  (probability of predicting a 1 correctly)
- The **specificity** is the proportion of true 0's estimated as 0's: $Sp = d/(b + d)$.
  (probability of predicting a 0 correctly)
- The **false positive rate** is the proportion of true 0's estimated as 1's: $F_+ = b/(b + d)$.

Logistic regression model predictions

Area under the ROC curve

	$y = 1$	$y = 0$	**Total**
$\hat{y}_i = 1$	a	b	$a + b$
$\hat{y}_i = 0$	c	d	$c + d$
	$a + c$	$b + d$	n

We define

- The **sensitivity** is the proportion of true 1's estimated as 1's: $Ss = a/(a + c)$.
  (probability of predicting a 1 correctly)
- The **specificity** is the proportion of true 0's estimated as 0's: $Sp = d/(b + d)$.
  (probability of predicting a 0 correctly)
- The **false positive rate** is the proportion of true 0's estimated as 1's: $F_+ = b/(b + d)$.
  (probability of predicting a 0 incorrectly)

Logistic regression model predictions

Area under the ROC curve

	$y = 1$	$y = 0$	**Total**
$\hat{y}_i = 1$	a	b	$a + b$
$\hat{y}_i = 0$	c	d	$c + d$
	$a + c$	$b + d$	n

We define

- The **sensitivity** is the proportion of true 1's estimated as 1's: $Ss = a/(a + c)$.
  (probability of predicting a 1 correctly)
- The **specificity** is the proportion of true 0's estimated as 0's: $Sp = d/(b + d)$.
  (probability of predicting a 0 correctly)
- The **false positive rate** is the proportion of true 0's estimated as 1's: $F_+ = b/(b + d)$.
  (probability of predicting a 0 incorrectly)
- The **false negative rate** is the proportion of true 1'ss estimated as 0's: $F_- = c/(a + c)$.

Logistic regression model predictions

Area under the ROC curve

	$y = 1$	$y = 0$	**Total**
$\hat{y}_i = 1$	a	b	$a + b$
$\hat{y}_i = 0$	c	d	$c + d$
	$a + c$	$b + d$	n

We define

- The **sensitivity** is the proportion of true 1's estimated as 1's: $Ss = a/(a + c)$.
  (probability of predicting a 1 correctly)
- The **specificity** is the proportion of true 0's estimated as 0's: $Sp = d/(b + d)$.
  (probability of predicting a 0 correctly)
- The **false positive rate** is the proportion of true 0's estimated as 1's: $F_+ = b/(b + d)$.
  (probability of predicting a 0 incorrectly)
- The **false negative rate** is the proportion of true 1'ss estimated as 0's: $F_- = c/(a + c)$.
  (probability of predicting a 1 incorrectly)

Logistic regression model predictions

Contingency Table

		DISEASE	
		+	-
TEST	+	True + (a)	False + (b)
	-	False - (c)	True - (d)
		Sensitivity = a / a + c = TP / TP + FN	Specificity = d / b + d = TN / FP + TN

ROC curve

- The **ROC (Receiver Operating Characteristic) curve** is a graphic display that gives a measure of the predictive accuracy of the model.
- It allows to represent the impact of the cutoff point on the sensitivity and specificity (you have to keep in mind that as you increase the cutoff point from 0 to 1, the sensitivity decreases and the specificity increases).
- Ideally, we would like to have high values for both sensitivity and specificity.
- The ROC curve is a plot of **Sensitivity** against $1 -$ **Specificity**, i.e., $1 - F_+$.
- Then compute the **area under the curve** (**AUC**).
- For a model with high predictive accuracy, the ROC curve rises quickly.

ROC curve

AUC

- Measuring the **area under the ROC curve**, we can obtain the accuracy of the classfication test.
- The larger area, the better the diagnostic test is.
- If the **area is 1.0**, we have an ideal test because test achieves $100\,\%$ sensitivity and $100\,\%$ specificity.
- If the **area is 0.5**, we have a test which has effectively $50\,\%$ sensitivity and $50\,\%$ specificity.
- In a few words, the area measures the ability of the test to correctly classify those with and without the disease.

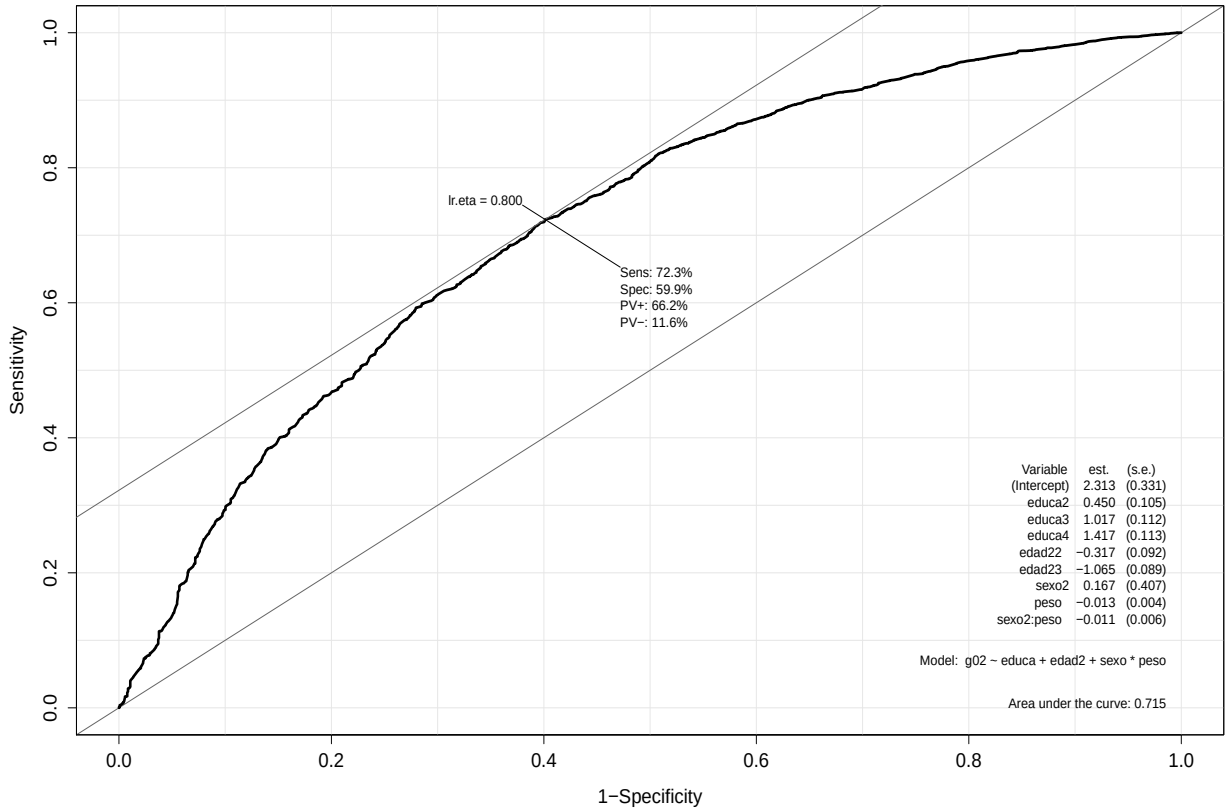
$$\text{AUC} = \int_0^1 \text{ROC}(t)dt$$

where $t = 1 - \text{specificity}$ (false positive rate) and $\text{ROC}(t)$ is **sensitivity (true positive rate)**.

ROC curve

- Let us plot the ROC curve

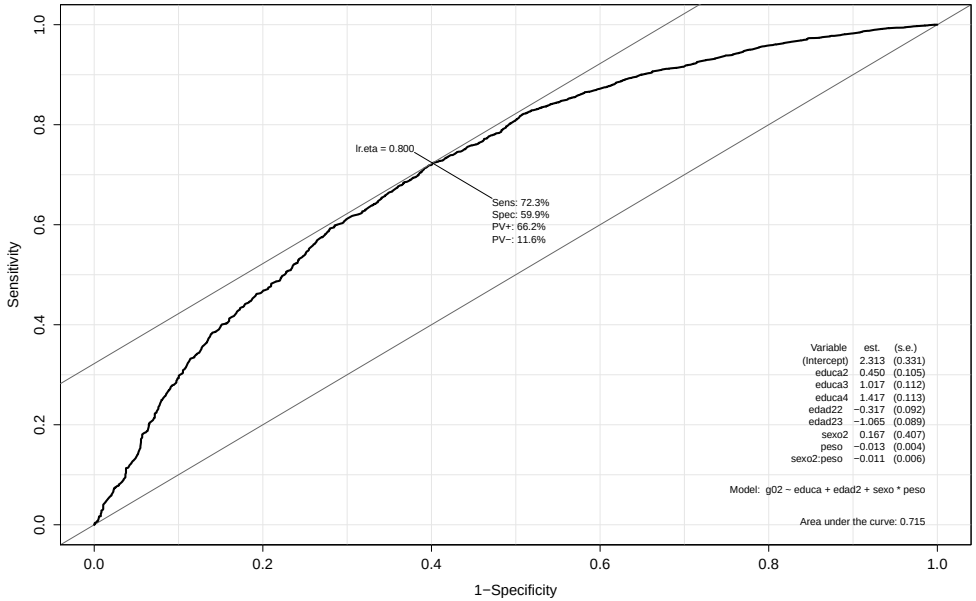
```
> library(Epi)
> ROC(form=g02~educa+edad2+sexo*peso, data=salud,plot="ROC",lwd=3,cex=1.5)
```



ROC curve

In general:

- AUC $\leq 0{,}5$: the model is worthless, it does not help us to discriminate
- $0{,}6$AUC $< 0{,}8$: acceptable (fair)
- $0{,}8$AUC $< 0{,}9$: excellent (good)
- AUC $> 0{,}9$: outstanding



Model diagnostics in logistic regression

When the hypothesis of the logistic regression model are violated, the fitted model can give the next kind of errors:

Model diagnostics in logistic regression

When the hypothesis of the logistic regression model are violated, the fitted model can give the next kind of errors:

- Biased coefficientes
- Unefficient estimators
- Statistical Inference on the parameters is not valid

Model diagnostics in logistic regression

When the hypothesis of the logistic regression model are violated, the fitted model can give the next kind of errors:

- Biased coefficientes
- Unefficient estimators
- Statistical Inference on the parameters is not valid

  ✓ **Bias**: systematic tendency of over/under estimate the coefficients of the model.

Model diagnostics in logistic regression

When the hypothesis of the logistic regression model are violated, the fitted model can give the next kind of errors:

- Biased coefficientes
- Unefficient estimators
- Statistical Inference on the parameters is not valid

  ✓ **Bias**: systematic tendency of over/under estimate the coefficients of the model.
  ✓ **Unefficiency**: large standard errors for the coefficient's estimates.

Model diagnostics in logistic regression

When the hypothesis of the logistic regression model are violated, the fitted model can give the next kind of errors:

- Biased coefficientes
- Unefficient estimators
- Statistical Inference on the parameters is not valid

  ✓ **Bias**: systematic tendency of over/under estimate the coefficients of the model.
  ✓ **Unefficiency**: large standard errors for the coefficient's estimates.

- **Other issues:** *high-leverage points* of the predictors or *outliers* in the response, that may affect the estimation of the parameters.

bcam
basque center for applied mathematics

Model diagnostics in logistic regression

When the hypothesis of the logistic regression model are violated, the fitted model can give the next kind of errors:

- Biased coefficientes
- Unefficient estimators
- Statistical Inference on the parameters is not valid

  ✓ **Bias**: systematic tendency of over/under estimate the coefficients of the model.
  ✓ **Unefficiency**: large standard errors for the coefficient's estimates.

- **Other issues:** *high-leverage points* of the predictors or *outliers* in the response, that may affect the estimation of the parameters.

We will focus on the hypotheses of the model and how to detect when they are violated

Model diagnostics in logistic regression

Specification error

- To check if the model is well specified. We need to study:
  - The functional relationship between the predictors and the response.
  - The presence of irrelevant predictors and the absence of important variables.

Model diagnostics in logistic regression

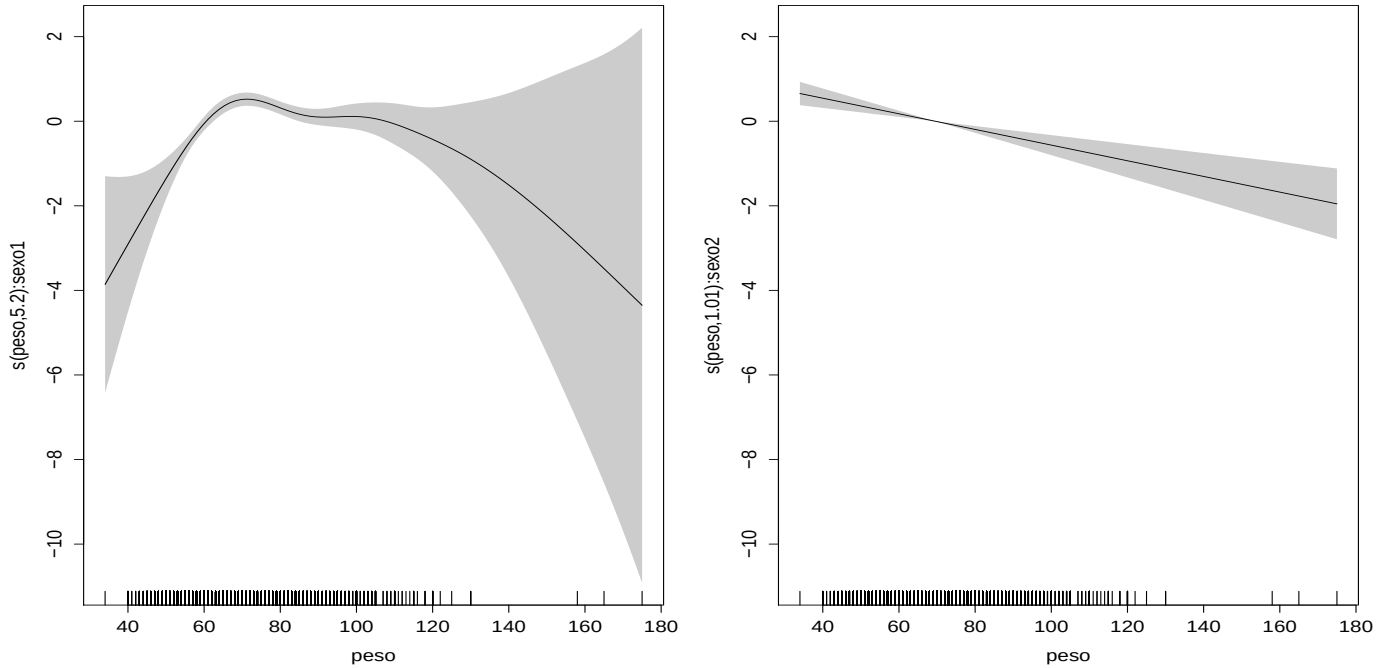
Specification error

- To check if the model is well specified. We need to study:
  - The functional relationship between the predictors and the response.
  - The presence of irrelevant predictors and the absence of important variables.
- **If the model is not correctly specified:**
  - Coefficient's estimates are biased.
  - The relation of the logit and the predictors is not linear.
  - The relation between the variables is multiplicative, not additive (there is interaction, contrast statistical significance).
  - If we include more variables than needed, the standard errors of the estimated coefficients increases, the efficiency of the estimadors are reduced, although there is no bias.
  - Presence of multicolinearity, i.e. correlated predictors. (`cor(model.matrix(logistic13)[,-1])`)

Model diagnostics in logistic regression

Specification error

- If the relationship between the logit and the predictors is **not linear** a unit increment of X is not constant, and do depends on the value of X.
- We can detect the non-linearity using a smoothing technique and the function `gam()` in `library(mgcv)`



Model diagnostics in logistic regression

Residual analysis

- We had defined different type of residuals in the case of a glm:

1. **Response residuals:** $y_i - \hat{\mu}_i$, they are not appropriate since $Var(y_i)$ is not constant.
2. **Pearson residuals:**

$$r_{i,P} = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{Var(y_i)}}}$$

They have constant variance and mean zero. Mostly useful for detecting variance misspecification.

3. **Deviance residuals:**

$$sign(y_i - \hat{\mu}_i)\sqrt{d_i^2}$$

where d_i is the contribution to the model deviance of the i^th observation. For many models the deviance residuals are closer to a Normal distribution than the Pearson residuals, and so they are more appropriate for constructing diagnostic plots.

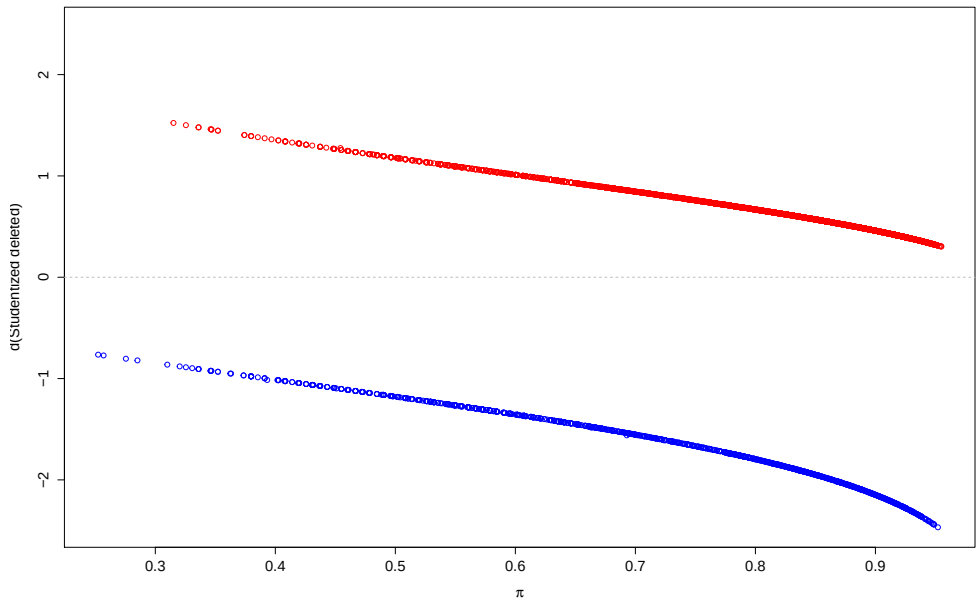
Model diagnostics in logistic regression

Residual analysis (cont.)

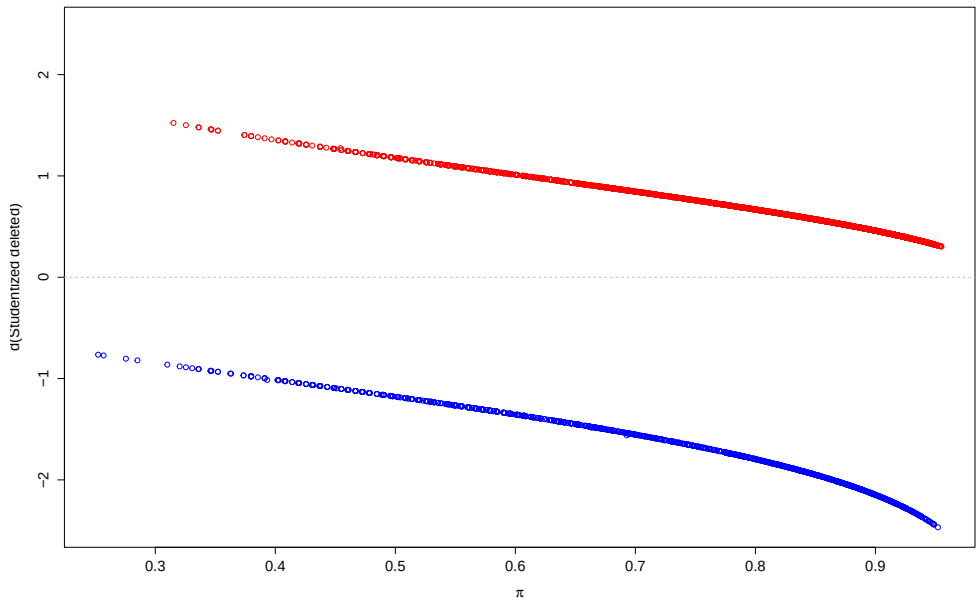
4. **Standardized residuals:** Both the Pearson and the deviance residuals can be variance-standardized and corrected for the effects of leverage by dividing them by $\sqrt{\phi(1-h_{ii})}$, in most cases $\phi = 1$, and when it is not, it is replaced by an estimate. These residuals should be approximately $N(0,1)$ for Poisson and binomial models with large counts and should lie within -2 and $+2$.

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(y_i)}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

See `Logreg.R`



- In logistic regression, the data are discrete and so are the residuals.
- **Why do we have those two lines of points?**



- In logistic regression, the data are discrete and so are the residuals.
- **Why do we have those two lines of points?**
- Because we predict a probability for a variable taking values 0 or 1. If the true value is 0, then we always predict more, and residuals have to be negative (the blue points) and if the true value is 1, then we underestimate, and residuals have to be positive (the red points). Points are exactly on a smooth curve, as a function of the predicted value,

Logistic regression

Interpretation of the model

- In logistic regression, the data are discrete and so are the residuals.
- **Why do we have those two lines of points?**

Logistic regression

Interpretation of the model

- In logistic regression, the data are discrete and so are the residuals.
- **Why do we have those two lines of points?**
- Because we predict a probability for a variable taking values 0 or 1. If the true value is 0, then we always predict more, and residuals have to be negative (the blue points) and if the true value is 1, then we underestimate, and residuals have to be positive (the red points). Points are exactly on a smooth curve, as a function of the predicted value,

Interpretation of the results

Once we check that the model is ok, we can extract some conclusions

Interpretation of the results

Once we check that the model is ok, we can extract some conclusions

- The main effect is the continuous variable `peso`

Interpretation of the results

Once we check that the model is ok, we can extract some conclusions

- The main effect is the continuous variable `peso`
- There is a nominal dichotomous variable `sexo`

Interpretation of the results

Once we check that the model is ok, we can extract some conclusions

- The main effect is the continuous variable `peso`
- There is a nominal dichotomous variable `sexo`
- And two polytomous `edad2` and education level (`educa`), we also include the `sexo:peso` interaction.

Interpretation of the results

```
> summary(logistic13)


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.313414   0.330517   6.999 2.57e-12 ***
educa2       0.449794   0.104774   4.293 1.76e-05 ***
educa3       1.017172   0.111848   9.094  < 2e-16 ***
educa4       1.417463   0.113447  12.495  < 2e-16 ***
edad22      -0.316745   0.092114  -3.439 0.000585 ***
edad23      -1.064558   0.089488 -11.896  < 2e-16 ***
sexo2        0.166893   0.406530   0.411 0.681417
peso        -0.012573   0.003967  -3.170 0.001527 **
sexo2:peso  -0.011252   0.005626  -2.000 0.045488 *
```

Interpretation of the results

- We saw that the estimated Odds-ratios are obtained exponentiating the coefficients and the C.I.'s
- Recall that when the variables are dichotomous or polytomous, the lower level is taken as the baseline.

Interpretation of the results

- We saw that the estimated Odds-ratios are obtained exponentiating the coefficients and the C.I.'s
- Recall that when the variables are dichotomous or polytomous, the lower level is taken as the baseline.

```
> exp(logistic13$coeff)

(Intercept)      educa2      educa3      educa4      edad22      edad23
 10.1088729   1.5679890   2.7653644   4.1266397   0.7285166   0.3448804
      sexo2        peso  sexo2:peso
  1.1816273   0.9875053   0.9888107

> exp(confint(logistic13))
```

Interpretation of the results

Results for variables `edad` and `educa`:

Variable values	Odds Ratio	C.I. 95 %
Age		
< 29	1.00	
$30 - 44$	0.73	0.61 0.87
$45 - 64$	0.34	0.29 0.41
Education Level		
Low	1.00	
Low-Mid	1.56	1.27 1.92
Mid-High	2.76	2.22 3.44
High	4.12	3.30 5.15

- The estimated OR for individuals of mid-age is $0{,}73$.
- The odd of having a good health perception for individuals between 30-44 years is $0{,}73$ lower than for the individual (same sex, education and weight) but younger.
- The odds for indivuals aged $45 - 64$ is much lower.
- For the variable Education levels, all the odds values are greater than 1, that means that a higher Education level is an advantage in good health perception

References

D. W. Hosmer, S. Lemeshow (2005). *Applied Logistic Regression, Second Edition*. John Wiley Sons, Inc.