3.2 Logistic Regression

1. General logistic regression

We now focus on models where the *outcome variables are measured on a binary scale*.

We define a **binary random variable**

$$Y = \begin{cases} 1 & \text{if the outcome is a “success”} \quad \pi \\ 0 & \text{if the outcome is a “failure”} \quad (1-\pi) \end{cases} \tag{3.2.1}$$

i.e. Y has a Bernoulli distribution.

Goal:

The goal of the analysis is to relate the probability of the Bernoulli distribution, π_i, to a set of explanatory variables $\mathbf{x}_i^\top$, i.e.

$$\Pr(\mathrm{Y}_i = 1 | \mathrm{X}_{i1}, \dots, \mathrm{X}_{ip}) = \pi_i \qquad \text{for } i = 1, \dots, N \tag{3.2.2}$$

The **joint likelihood function** is

$$\begin{aligned} f(\mathrm{Y}_1, \dots, \mathrm{Y}_N | \boldsymbol{\pi}) &= \prod_{i=1}^{N} \pi_i^{\mathrm{Y}_i} (1-\pi_i)^{1-\mathrm{Y}_i} \\ &= \exp\left[\sum_{i=1}^{N} \mathrm{Y}_i \log\left(\frac{\pi_i}{1-\pi_i} \right) + \sum_{i=1}^{N} \log(1-\pi_i) \right] \end{aligned} \tag{3.2.3}$$

We want to *describe the probability of success with respect to some predictors*:

$$g(\pi_i) = \mathbf{x}_i^\top \boldsymbol{\beta} \tag{3.2.4}$$

so to take into consideration that

- The response variable is **binary** and not continuous
- The response variable is **bounded** (in $[0,1]$)
- The *variance is not constant* $\mathbb{V}ar(Y_i) = \pi_i(1-\pi_i)$

Similar considerations apply to ordinal response variables.

Consider the `Default` dataset from the `ISLR` package in `R`. We want to estimate the probability of `default` as a function of `balance`.

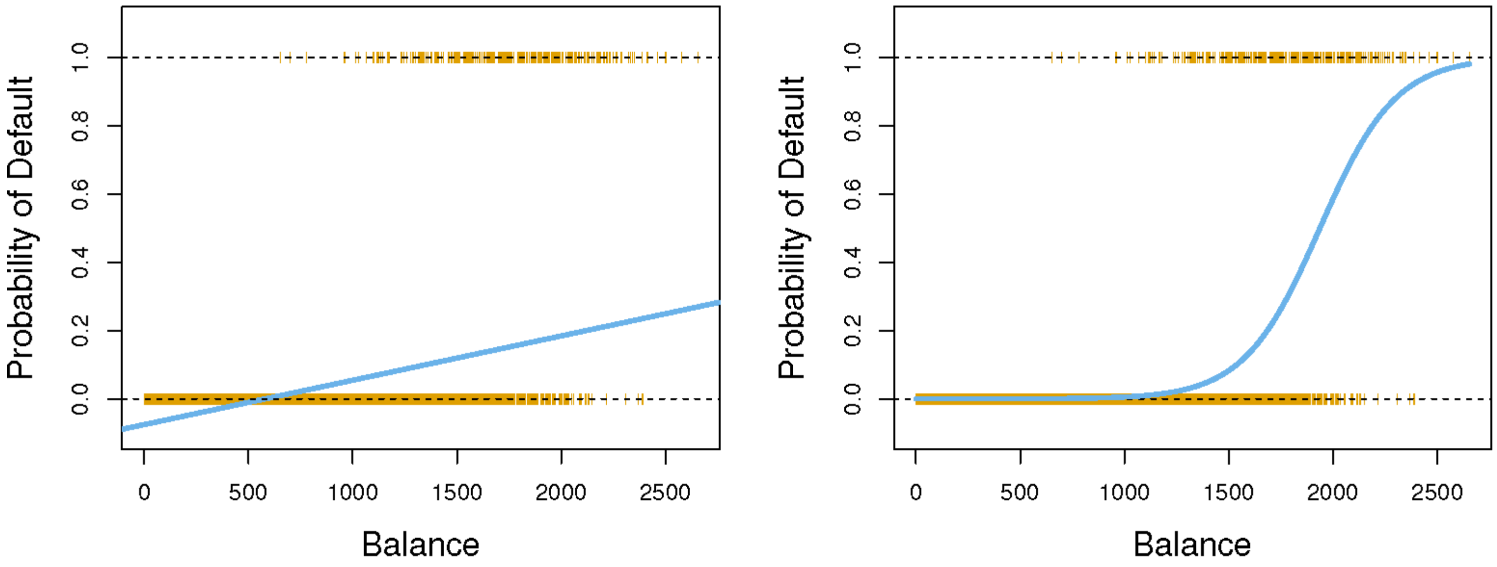


Figure 3.2.1: Estimated probability of `default` using linear (left) and logistic (right) regression.

The following code can be used to replicate the previous plots.

```
library(ISLR)
data("Default")
attach(Default)

default.bin <- rep(0, length(default)) # initialise binary vector
default.bin[default=="Yes"] <- 1

par(mfrow=c(1,2)) # To put the plots next to each other

# Linear model
lm <- lm(default.bin ~ balance)
plot(balance, default.bin, pch=3, col="orange",
     xlab="Balance", ylab="Probability of Default")
abline(h=c(0,1), lty=2)
abline(a=lm$coefficients[1], b=lm$coefficients[2], col="blue", lwd=3)
# Logistic model
log <- glm(default.bin ~ balance, family="binomial")
o <- order(balance)

plot(balance, default.bin, pch=3, col="orange", xlab="Balance",
     ylab="Probability of Default")
abline(h=c(0,1), lty=2)
lines(balance[o], log$fitted.values[o], col="blue", lwd=3)
```

1.1 Example

We want to predict the medical condition of a patient in the emergency room on the basis of the symptoms. Let's suppose we have three possible diagnoses:

$$Y = \begin{cases} 1 & \text{stroke} \\ 2 & \text{drug overdose} \\ 3 & \text{epileptic seizure} \end{cases}$$

Using a linear regression would assume

- The ordering is meaningful: numbers 1, 2 and 3 are just labels!
- The difference between "stroke" and "drug overdose" has the same meaning than that between "drug overdose" and "epileptic seizure"

The general logistic regression model is

$$\operatorname{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^\top \boldsymbol{\beta} \tag{3.2.5}$$

where $\mathbf{x}_i$ is a vector of either continuous measurements or categorical variables and $\boldsymbol{\beta}$ is a parameter vector. Recall that $\frac{\pi_1}{1-\pi_i}$ is an **odds** taking value between 0 and ∞, *indicating very low and very high probabilities of default*. This means that

$$\begin{aligned} \frac{\pi_i}{1-\pi_i} &= \exp[\mathbf{x}_i^\top \boldsymbol{\beta}] \\ \pi_i &= \exp[\mathbf{x}_i^\top \boldsymbol{\beta}] - \pi_i \exp[\mathbf{x}_i^\top \boldsymbol{\beta}] \\ (1+\exp[\mathbf{x}_i^\top \boldsymbol{\beta}])\pi_i &= \exp[\mathbf{x}_i^\top \boldsymbol{\beta}] \\ \pi_i &= \frac{\exp[\mathbf{x}_i^\top \boldsymbol{\beta}]}{1+\exp[\mathbf{x}_i^\top \boldsymbol{\beta}]} \end{aligned}$$

And the **log-likelihood** can be rewritten with respect to $\boldsymbol{\beta}$

$$\ell(\boldsymbol{\beta}; \mathbf{y}, \mathbf{x}) = \sum_{i=1}^{N} \left[y_i \log\left(\frac{\exp[\mathbf{x}_i^\top \boldsymbol{\beta}]}{1+\exp[\mathbf{x}_i^\top \boldsymbol{\beta}]}\right) + (1-y_i)\log\left(\frac{1}{1+\exp[\mathbf{x}_i^\top \boldsymbol{\beta}]}\right)\right] \tag{3.2.6}$$

The estimation process is the same if Y_i is binomially distributed instead of Bernoulli distributed, with the corresponding modification to consider the number of trials.

If the goal is **prediction**, one might predict

$$Y_{N+1} = 1 \qquad \text{if } \pi_{N+1} | \mathbf{x}_{N+1}^\top > 0.5.$$

However, other thresholds could be used, e.g. if we want to be particularly conservative, we can set the threshold to 0.1.

1.2 Example

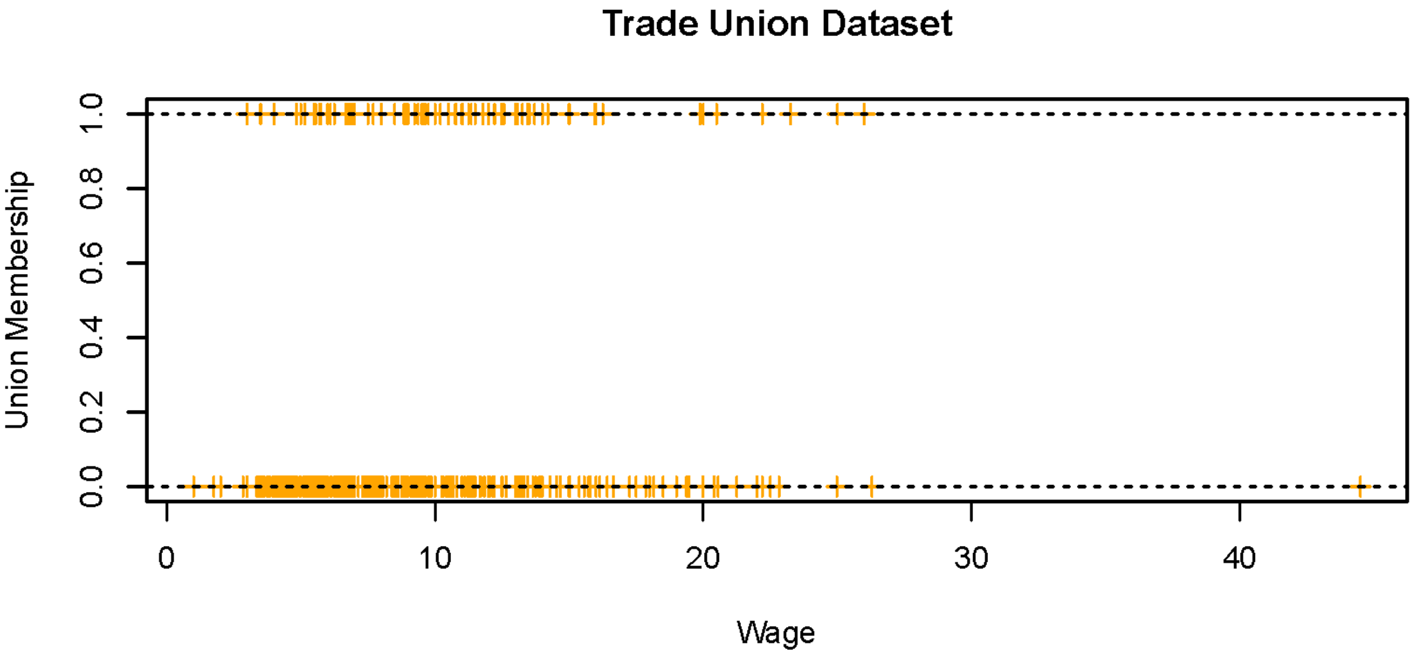
The trade union data are collected from 1985 and available in the **R** package **SemiPar**. The variable **union.member** is binary while variables **age** and **wages** are continuous. We illustrate the model comparison with this data set.

The figure below shows logistic regression fitted to the two variables separately with union membership as the response.

```
library(SemiPar)

data("trade.union")
attach(trade.union)

plot(wage, union.member, main="Trade Union Dataset", pch=3,
     ylab="Union Membership", xlab="Wage", col="orange")
abline(h=c(0,1), lty=2)
```



The code below fits the logistic regression model displays the fitted line and prints the output.

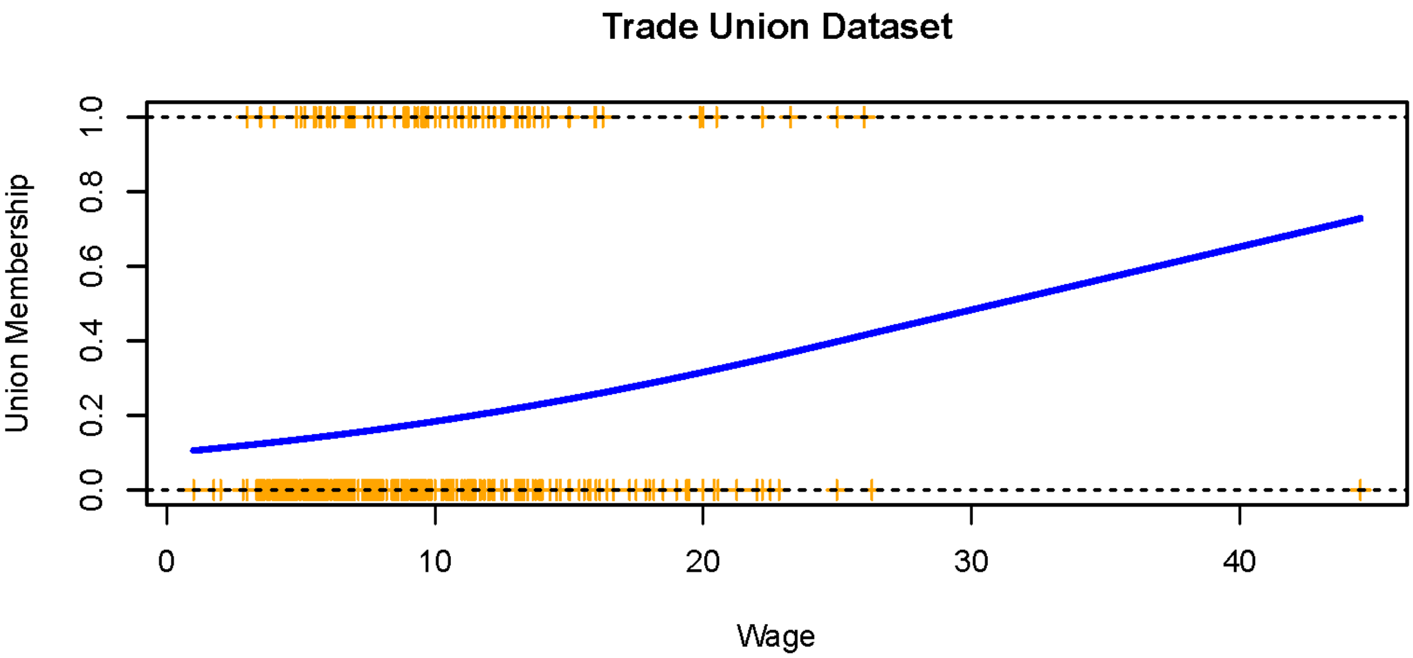
```
library(SemiPar)

data("trade.union")
attach(trade.union)

union.wage.glm <- glm(union.member ~ wage, family="binomial")
o <- order(wage)
```

```
plot(wage, union.member, main="Trade Union Dataset", pch=3,
     ylab="Union Membership", xlab="Wage", col="orange")
abline(h=c(0,1), lty=2)
lines(wage[o], union.wage.glm$fitted.values[o], col="blue", lwd=3)

summary(union.wage.glm)
```

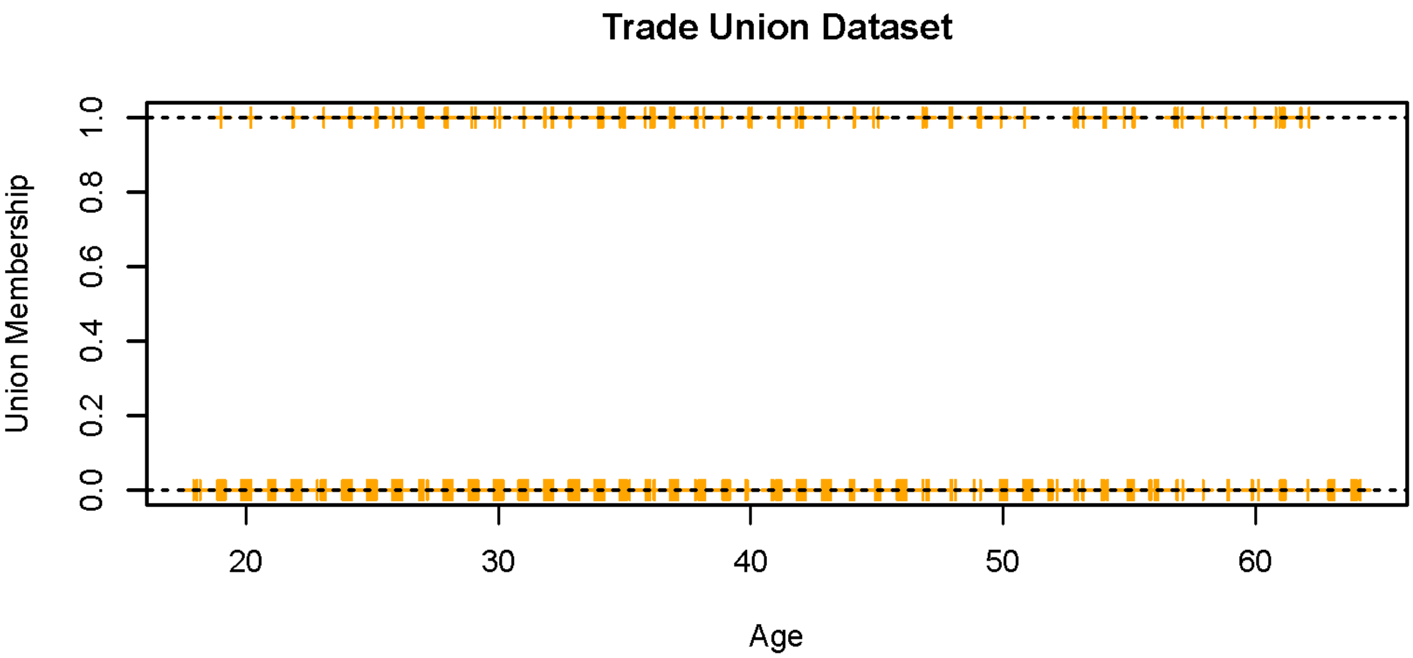


Let's now model **`union.member`** as function of **`age`**

```
library(SemiPar)

data("trade.union")
attach(trade.union)

plot(jitter(age), union.member, main="Trade Union Dataset",
     col="orange", pch=3, ylab="Union Membership", xlab="Age")
abline(h=c(0,1), lty=2)
```



The code below fits the logistic regression model displays the fitted line and prints the output.

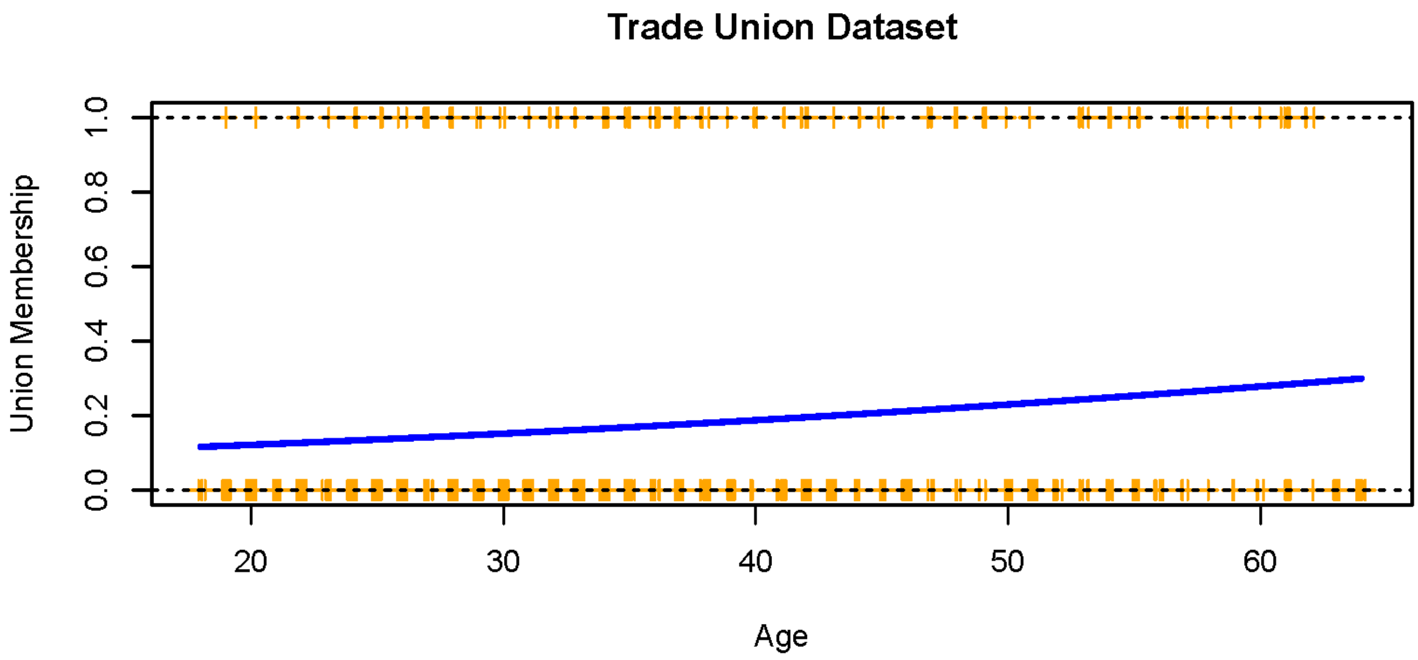
```
library(SemiPar)

data("trade.union")
attach(trade.union)

union.age.glm <- glm(union.member ~ age, family="binomial")
o <- order(age)

plot(jitter(age), union.member, main="Trade Union Dataset", pch=3,
     ylab="Union Membership", xlab="Wage", col="orange")
abline(h=c(0,1), lty=2)
lines(age[o], union.age.glm$fitted.values[o], col="blue", lwd=3)

summary(union.age.glm)
```



Both **age** and **wage** seem to be significant, so let's fit a model with **3** parameters.

```
library(SemiPar)

data("trade.union")
attach(trade.union)

union.wage.age.glm <- glm(union.member ~ wage + age, family="binomial")
summary(union.wage.age.glm)
```

2. Prediction

Once the coefficients have been estimated, **predictions** are obtained by using those estimates with the desired level of predictors.

Example: Analysis of trade union dataset

If we want to predict the probability of **union membership** for someone who is **56 years old** and has a **wage of 6.5**, we compute

$$\pi_{new} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 6.5 + \hat{\beta}_2 56)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 6.5 + \hat{\beta}_2 56)}$$

```
library(SemiPar)

data("trade.union")
attach(trade.union)

union.wage.age.glm <- glm(union.member ~ wage + age, family="binomial")

p_pred <- exp( sum(union.wage.age.glm$coefficients * c(1,6.5,56))) /
          (1+exp( sum(union.wage.age.glm$coefficients * c(1,6.5,56))))
p_pred
```

The $\pi_{new} < 0.5$, therefore we can classify the new individual as not a union member.

3. Goodness of fit

In an assessment of goodness-of-fit for a **linear model**, **residual plots** are useful in exhibiting violations of model assumptions (e.g. independence, homoscedasticity).

In a GLM, we would like to *assign a residual e_i to each observation which measures the discrepancy between Y_i and the value predicted by the fitted model*.

There are two main difficulties associated with generalised linear models:

- The model variances depend on the expectations;
- It is not obvious that data and fitted values should be compared on the original scale of the responses.

3.1 Pearson chi-squared statistic

For calculation of Pearson residuals, we take the *difference between observed and fitted values* and *divide by an estimate of the standard deviation of the observed values*.

Residuals for the Binomial model: For $Y_i \sim \mathrm{Bin}(n_i, \pi_i)$, the **Pearson Residuals** are

$$P_i = \frac{(y_i - n_i\hat{\pi}_i)}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}, \quad i = 1, \ldots, N$$

Instead of maximising the likelihood, we can estimate the parameters by *minimising the weighted sum of squares*

$$S_w = \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}$$

where $\mathbb{E}(\mathrm{Y}_i) = n_i\pi_i$ and $\mathbb{V}\mathrm{ar}(\mathrm{Y}_i) = n_i\pi_i(1-\pi_i)$, which is also called **Pearson chi-squared statistic**

$$P^2 = \sum_{i=1}^{N} \frac{(o_i - e_i)^2}{e_i}$$

where o_i represents the observed frequencies and e_i represents the expected frequency.

The reason for the equivalence is

$$\begin{aligned} P^2 &= \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i} + \sum_{i=1}^{N} \frac{[(n_i - y_i) - n_i(1-\pi_i)]^2}{n_i(1-\pi_i)} \\ &= \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}(1 - \pi_i + \pi_i) = S_w \end{aligned}$$

When the **Pearson chi-squared statistic** is evaluated at the estimated expected frequencies

$$P^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

3.2 Deviance

The **deviance** for the logistic model is

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right] \tag{3.2.7}$$

ℹ Check this assertion.

It is possible to prove that the deviance is asymptotically equivalent to the *Pearson chi-squared statistic* evaluated at the estimated expected frequencies.

ℹ Check this assertion.

Under the null hypothesis (H_0), the asymptotic distribution of D is

$$D \sim \chi^2(N - p) \tag{3.2.8}$$

therefore $P^2 \sim \chi^2(N - p)$.

The adequacy of the approximation depends on how well D or P^2 are χ^2-distributed. There is some evidence that P^2 is better than D, however both of them are influenced by small frequencies. This is typical of continuous covariates.

Example: Analysis of trade union dataset

We fitted the logistic model (M_1)

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1\texttt{wage} + \beta_2\texttt{age},$$

where π_i is the probability of union membership (3 params). The observed deviance is $d_1 = 485.5239$.

Compare with the nested model (M_0)

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0,$$

where the probability of trade union membership is **constant** (1 param). The observed deviance is

$d_0 = 503.0841.$

Example: Analysis of trade union dataset

We wish to test

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = 0 \\ H_1 &: \beta_1, \beta_2 \text{ not both zero} \end{aligned}$$

If H_0 was true, then both models describe the data well. We would have $D_0 \sim \chi^2(N-1)$ and $D_1 \sim \chi^2(N-3)$, so that $D_0 - D_1 \sim \chi^2(2)$. However, we observe

$$d_0 - d_1 = 503.0841378 - 485.5239 = 17.56029,$$

which is **larger** than the 95%-quantile $\chi^2_{2,0.95} = 5.9914645$. Hence we reject H_0 at the 5% significance level.

This code reproduce the calculations of the previous example

```
library(SemiPar)

data("trade.union")
attach(trade.union)

union.wage.age.glm <- glm(union.member ~ wage + age, family="binomial")

d0 <- union.wage.age.glm$null.deviance
d1 <- union.wage.age.glm$deviance

alpha <- 0.05

crit.val <- qchisq(1-alpha, df=2)
if(d0-d1 > crit.val){
  cat("We reject H0 at alpha=", alpha, "significance level.")
}else{
  cat("We cannot reject H0 at alpha=", alpha, "significance level.")
}
```

3.3 Hosmer-Lemeshow Statistic

A possible solution is to **group observations**, with approximately equal numbers of observations in each group. Then the *Pearson chi-squared statistic* is computed on the contingency table obtained by grouping observations. This statistic is called **Hosmer-Lemeshow statistic**.

```
library(SemiPar)
library(doBy)

data("trade.union")
attach(trade.union)

# Run the functions sum and length on the value of "union.member" for each group, broken down by wa
uniongrp <- summaryBy(union.member ~ wage, data=trade.union ,
                      FUN=c(sum, length))
names(uniongrp) = c("x","y","n")
head(uniongrp)

union.grp.glm <- glm(cbind(uniongrp$y, uniongrp$n-uniongrp$y) ~ uniongrp$x,
                     family=binomial)
summary(union.grp.glm)
```

The estimates and standard errors are the same but the goodness of fit differ.

3.4 Likelihood ratio, Pseudo R^2, AIC and BIC

Likelihood ratio χ^2 statistic

Sometimes the log-likelihood of the fitted model is compared with the log-likelihood of the **minimal model**, the model for which all π_i are equal; therefore the estimate is $\tilde{\pi} = \sum_{i=1}^N y_i / \sum_{i=1}^N n_i$.

The statistic is defined as

$$\begin{aligned} C &= 2[\ell(\hat{\boldsymbol{\pi}};\mathbf{y}) - \ell(\tilde{\boldsymbol{\pi}};\mathbf{y})] \\ &= 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{\hat{y}_i}{n_i\tilde{\pi}}\right) + (n_i - y_i)\log\left(\frac{n_i - \hat{y}_i}{n_i - n_i\tilde{\pi}}\right)\right] \end{aligned}$$

Therefore $C \sim \chi^2(p-1)$.

```
library(SemiPar)

data("trade.union")
attach(trade.union)

union.wage.age.glm <- glm(union.member ~ wage + age, family="binomial")

# Minimal model
union0.glm <- glm(union.member ~ 1, family=binomial)

Cstat <- 2*(logLik(union.wage.age.glm) - logLik(union0.glm))
alpha <- 0.05
p <- length(union.wage.age.glm$coefficients)

if(Cstat[1] > qchisq(1-alpha,p-1)){
  cat("We reject H0 at alpha=", alpha, "significance level.")
}else{
  cat("We cannot reject H0 at alpha=", alpha, "significance level.")
}
```

In this example the fitted model is preferred compared to the minimal model.

Pseudo-R^2

Analogously to the multiple linear regression, the likelihood ratio statistic can be **normalised**

$$\text{pseudo-}R^2 = \frac{\ell(\tilde{\boldsymbol{\pi}};\mathbf{y}) - \ell(\hat{\boldsymbol{\pi}};\mathbf{y})}{\ell(\tilde{\boldsymbol{\pi}};\mathbf{y})} \qquad (3.2.9)$$

representing the proportional improvement in the log-likelihood function due to the terms in the model of interest, compared with the minimal model.

As for the R^2, the distribution of the pseudo-R^2 cannot be determined, and it increases as the number of predictors increases. Therefore, several adjustments have been proposed.

```
library(SemiPar)

data("trade.union")
attach(trade.union)

union.wage.age.glm <- glm(union.member ~ wage + age, family="binomial")

# Minimal model
union0.glm <- glm(union.member ~ 1, family=binomial)

R2 <- (logLik(union0.glm) - logLik(union.wage.age.glm)) / logLik(union0.glm)
R2
```

AIC and BIC

The **Akaike information criterion** (AIC) and the **Bayesian information criterion** (BIC) are very popular goodness of fit statistics based on the log-likelihood, with an adjustment for the number of parameters and the sample size.

$$\mathrm{AIC} = -2\ell(\hat{\boldsymbol{\pi}}; \mathbf{y}) + 2p \tag{3.2.10}$$

where p is the number of parameters.

$$\mathrm{BIC} = -2\ell(\hat{\boldsymbol{\pi}}; \mathbf{y}) + p \times \log N \tag{3.2.11}$$

where N is the sample size.

Remark: all these statistics (except the pseudo-R^2) summarise how well a particular model fits the data: *a small value (or a large p-value) indicates that the model fits well*.

```
library(SemiPar)

data("trade.union")
attach(trade.union)

union.wage.glm <- glm(union.member ~ wage, family="binomial")
union.wage.age.glm <- glm(union.member ~ wage + age, family="binomial")

# Minimal model
union0.glm <- glm(union.member ~ 1, family=binomial)

BIC(union0.glm)
AIC(union0.glm)
```

```
BIC(union.wage.glm)
AIC(union.wage.glm)

BIC(union.wage.age.glm)
AIC(union.wage.age.glm)
```

4. Residuals

The residuals correspond to some of the statistics we have already analysed.

For $\mathrm{Y}_i \sim \mathrm{Bin}(n_i, \pi_i)$, the **Pearson Residuals** are

$$P_i = \frac{(\mathrm{Y}_i - n_i\hat{\pi}_i)}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}, \quad i = 1, \ldots, N$$

which can be **standardised** by dividing by the leverage h_{ii}

$$e_{iP} = \frac{P_i}{\sqrt{1-h_{ii}}}$$

Notice that $\sum_{i=1}^{N} P_i^2 = P^2$.

```
library(SemiPar)

data("trade.union")
attach(trade.union)

union.wage.age.glm <- glm(union.member ~ wage + age, family="binomial")

pr <- residuals(union.wage.age.glm,type="pearson")
prss <- sum(pr^2)
prss
```

The **Deviance Residuals** are defined as

$$d_i = \mathrm{sign}(\mathrm{Y}_i - n_i\hat{\pi}_i)\left\{2\left[\mathrm{Y}_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - \mathrm{Y}_i}{n_i - n_i\hat{\pi}_i}\right)\right]\right\}^{1/2}$$

(the sign term makes sure that the signs of d_i and P_i match).

Note that $\sum_{i=1}^{N} d_i^2 = D$, the deviance.

```
library(SemiPar)

data("trade.union")
attach(trade.union)

union.wage.age.glm <- glm(union.member ~ wage + age, family="binomial")
union.wage.age.glm$deviance
```

The residuals can be used in the usual way: they should be plotted

- Plotted *against each continuous explanatory variable* to check if the assumption of linearity is appropriate
- Plotted *against other possible explanatory variables* not included in the model
- Plotted *in the order of the measurements* to check for correlation
- Through *normality plots*

In general, for GLM residual plots are less informative than for multiple linear regression, therefore it is important to check all the other goodness-of-fit statistics.

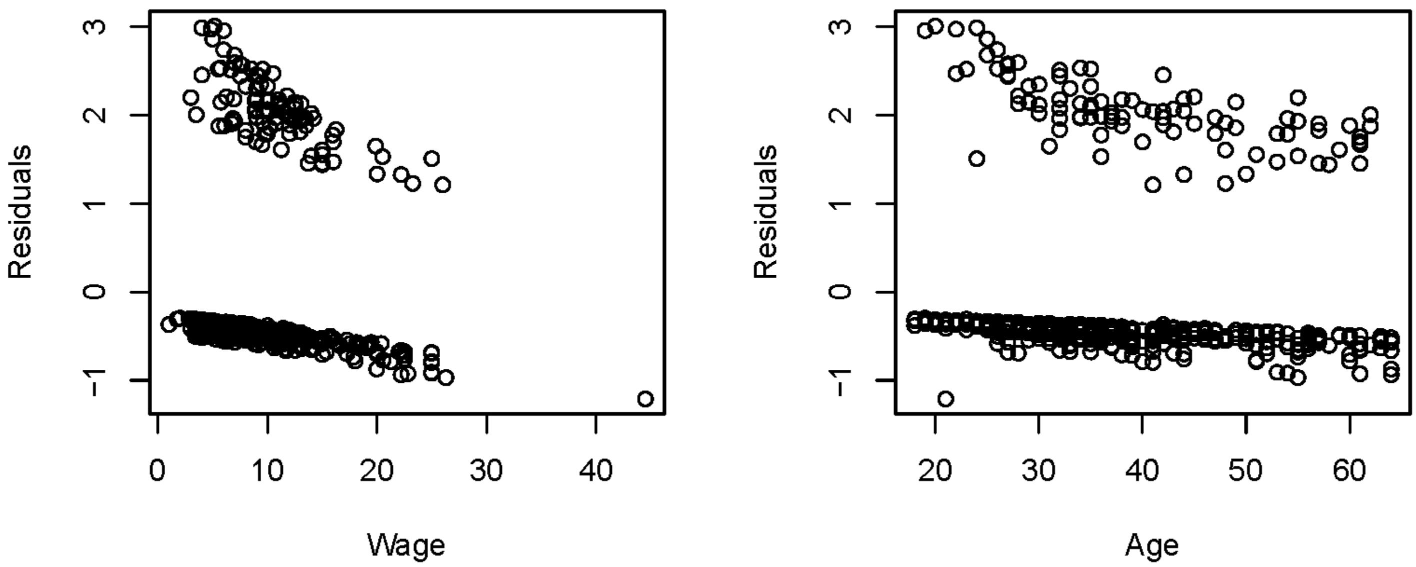


Figure 3.2.2: Plot of the residuals against each continuous explanatory variable

Usually, we plot residuals against the values of the linear predictors in a generalised linear model to look for patterns in the residuals that may be related to the mean.

But sometimes it can be hard to see patterns in residual plots for generalised linear models. For instance, *in logistic regression where the responses are binary, the residuals can only take on two possible values* (depending on whether the response is zero or one) and when residuals are plotted against linear predictor values, all points lie on one of the two smooth curves.

It is nearly always helpful to superimpose a scatterplot smoother on the residual plot to help identify any trends. In the figure, the black line is a smoothing spline (more on this later) suggests that there may be a trend in the residuals, namely the mean is underestimated in the middle, and overestimated on the two extremes.

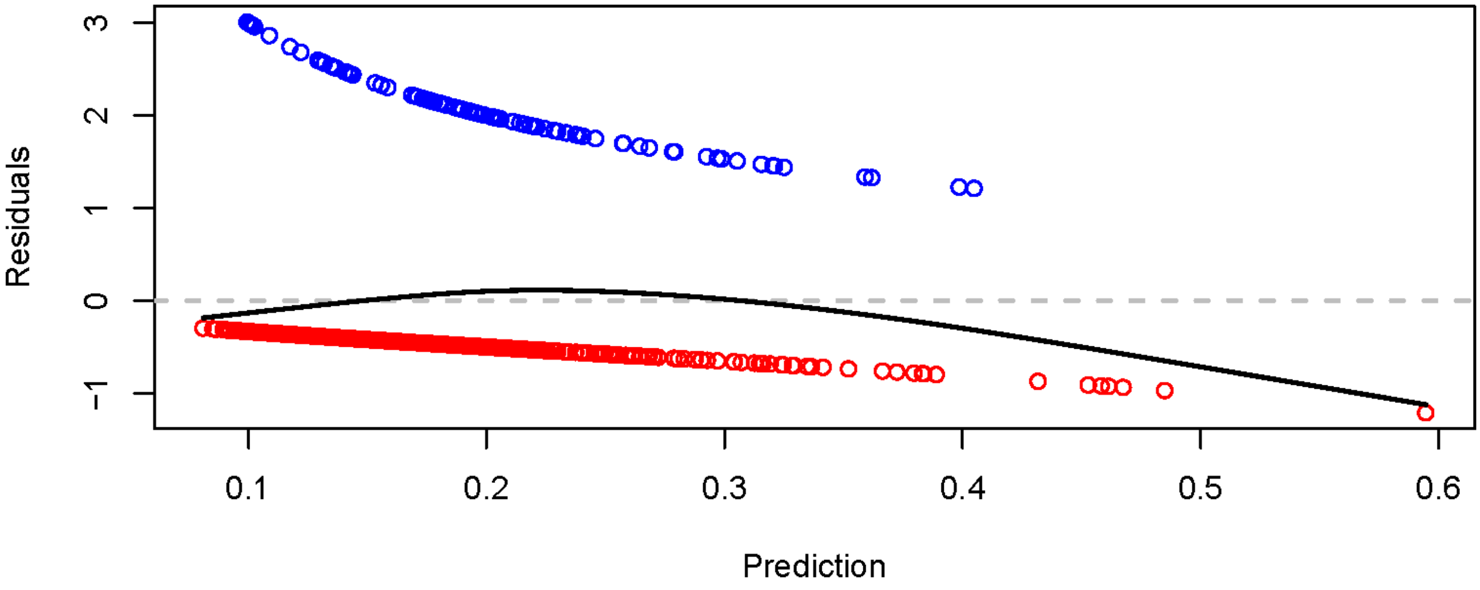


Figure 3.2.3: Pearson residuals against the predicted probability of being a member. **Blue** illustrates positive residuals (the individual is a member) and **red** negative residuals. The black line represents the smoothing spline.

```
library(SemiPar)

data("trade.union")
attach(trade.union)

union.wage.age.glm <- glm(union.member ~ wage + age, family="binomial")

pr <- residuals(union.wage.age.glm,type="pearson")

# Prediciton
pred.memb <- predict(union.wage.age.glm)
pred.memb <- exp(pred.memb ) / (1+exp(pred.memb )) # original scale

plot(pred.memb, pr, col=c("red","blue")[1+union.member],
     xlab="Prediction", ylab="Residuals")
abline(h=0,lty=2,col="grey", lwd=2)
ss <- smooth.spline(pred.memb, pr)
lines(ss, lwd=2)
```

5. Activity

Senility and WAIS (Dobson & Barnett, Section 7.8)

A sample of $N = 54$ elderly people was given a psychiatric examination to determine whether symptoms of senility were present. Other measurements taken at the same time included the score on a subset of the Wechsler Adult Intelligence Scale (WAIS).The data are binary although some people have the same WAIS scores so there are $m = 17$ different covariate patterns. Let Y_i denote the number of people with symptoms among n_i people with the i-th covariate pattern. The dataset is available in the `dobson` package in `R` under `senility`.

1. Group the observations per covariate patterns ($m = 17$) and fit the regression model :

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i,$$

   where $Y_i \sim \text{Bin}(n_i, \pi_i), i = 1, \ldots, m$. Write some `R` code to compute the estimated regression coefficients, their standard errors.
2. Compute the Pearson χ^2 statistic and the deviance and test if the model fits well according to those statistics
3. Reproduce Table 7.9 from Dobson & Barnett (2018, p.169)
4. Display the standardised residuals (Pearson and Deviance) as function of the WAIS score. What conclusion can you make?
5. Conduct some research to learn how to use the `R` package `ResourceSelection` in order to compute the Hosmer-Lemeshow statistic for a $g \times 2$ contingency table, with $g = 3$. Display the observed and expected frequencies to match Table 7.10 (Dobson & Barnett (2018, p.170). What is the value of the statistic and its sampling distribution. What conclusions can be drawn?