Logistic Regression

Example

$$x = \begin{cases} 1 & \text{if expoesed to factor} \\ 0 & \text{if not} \end{cases}$$

$$y = \begin{cases} 1 & \text{if develops disease} \\ 0 & \text{if not} \end{cases}$$

Let $\pi_i(x_i)$ be the probability of disease for a given value of x_i

Odds

odds of disease among unexposed= $\pi_i(0)/(1-\pi_i(0))$

odds of disease among exposed= $\pi_i(1)/(1-\pi_i(1))$

Odds Ratio

$\text{OR} = \frac{\text{odds of disease among exposed}}{\text{odds of disease among unexposed}} = \frac{\pi_i(1)/(1-\pi_i(1))}{\pi_i(0)/(1-\pi_i(0))}$

$\text{OR} > 1$ indicates a risk factor.

$\text{OR} < 1$ indicates a predictive factor.

Logistic regression

In a logistic regression we model the logarithm of the odds $\log(\pi_i(x_i)/(1-\pi_i(x_i)))$ as a linear regression on covariates. Specifically, let Y_i be a 0/1 Bernoulli random variable and x_i a vector of covariates for the ith individual, then we model $\text{logit}(\pi_i(x_i)) = \log(\pi_i(x_i)/(1-\pi_i(x_i))) = x_i'\beta$, where here β is a vector of regression coefficients. Solving for the success probability, this yields $\pi_i(x_i) = \exp(x_i'\beta)/(1+\exp(x_i'\beta))$.

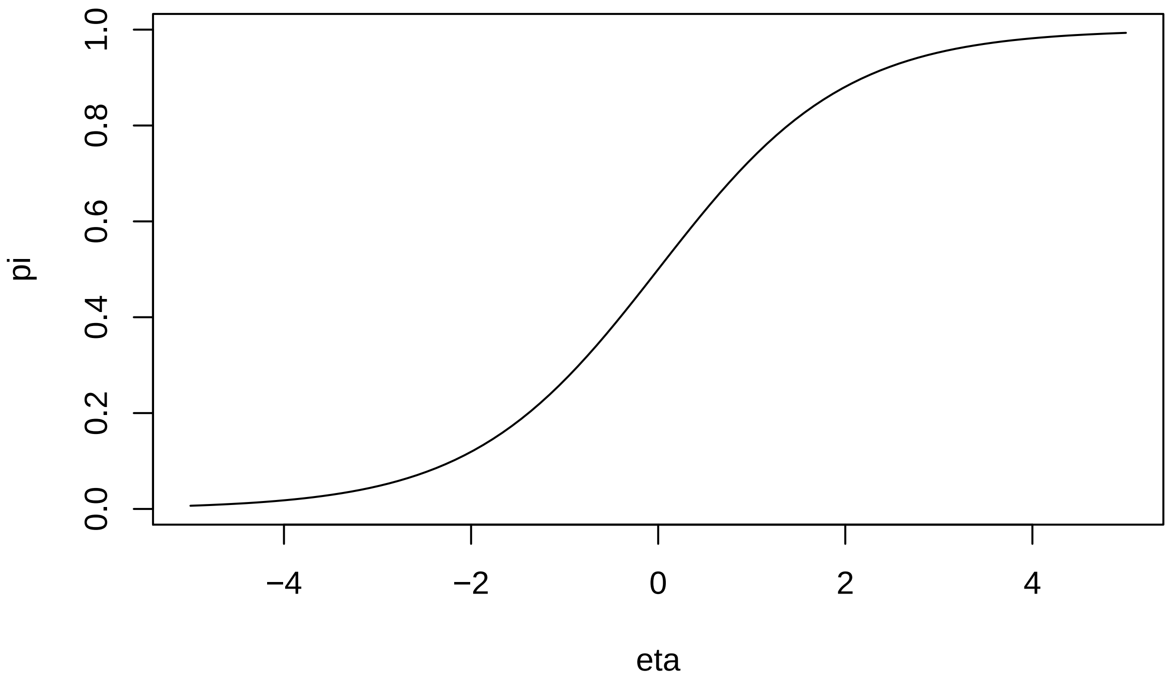
Equivalently,

$$\text{logit}(\pi_i(x_i)) = \log\left(\frac{\pi_i(x_i)}{1-\pi_i(x_i)}\right) = x_i'\beta$$

NOTE: We chose logit function such that its range is the entire real line. I.e., $\text{logit}(\pi_i(x_i)) = \log\left(\frac{\pi_i(x_i)}{1-\pi_i(x_i)}\right)$ takes on values in $(-\infty,\infty)$. The benefit of such a choice is that the estimated linear predictor, $x_i'\beta$, can take on any value in $(-\infty,\infty)$.

Plot of π vs. $\eta = x_i'\beta$

```
eta=seq(-5,5,,100)
pi=exp(eta)/(1+exp(eta))
plot(eta,pi,type="l")
```



Example: Attendance data

Assume that there are 30 students. There are 20 males and 10 females. We wish to study whether the gender (binary) affects the probability of attendance.

Let

$$Y_i = \begin{cases} 1, & \text{a student } i \text{ is present} \\ 0, & \text{otherwise} \end{cases}$$

$$x_i = \begin{cases} 1, & \text{a student } i \text{ is a male} \\ 0, & \text{if not} \end{cases}$$

Intepretion of β

Let $x_i'\beta = \alpha + \beta_1 x_i$. Let $\pi_i(x_i)$ be the probability that the i^{th} student with the variable x_i attends class.

$$x_i = 0 \Rightarrow \text{logit}(\pi_i(0)) = \alpha \tag{1}$$

$$x_i = 1 \Rightarrow \text{logit}(\pi_i(1)) = \alpha + \beta_1 \tag{2}$$

$$(2)-(1) \Rightarrow \beta_1 = \text{logit}(\pi_i(1)) - \text{logit}(\pi_i(0)) = \log\left(\frac{\pi_i(1)}{1-\pi_i(1)}\right) - \log\left(\frac{\pi_i(0)}{1-\pi_i(0)}\right) = \log\left(\frac{\pi_i(1)/(1-\pi_i(1))}{\pi_i(0)/(1-\pi_i(0))}\right) = \log(\text{OR})$$

$\Rightarrow \exp(\beta_1) = \text{OR}$

x_i	Linear predictor	Prob attend
Female	α	$\frac{\exp(\alpha)}{1+\exp(\alpha)}$
Male	$\alpha+\beta_1$	$\frac{\exp(\alpha+\beta_1)}{1+\exp(\alpha+\beta_1)}$

Likelihood

Let π_i be the probability of attending class for subject i.

We first need to choose a distribution for Y_i. Because Y_i is binary, we have only one choice: the Bernoulli distribution. This distribution can be written as

$$\begin{aligned} \mathrm{P}[Y=y] &= \pi^y(1-\pi)^{1-y} \\ &= \exp\{y\log\pi+(1-y)\log(1-\pi)\} \\ &= \exp\{y\log\left(\frac{\pi}{1-\pi}\right)+\log(1-\pi)\}. \end{aligned}$$

Now, the likelihood is

$$\begin{aligned} \mathcal{L}(\beta;y) &= \prod_{i=1}^n \exp\left\{y_i\log\left(\frac{\pi_i}{1-\pi_i}\right)+\log(1-\pi_i)\right\} \\ \log\mathcal{L}(\beta;y) &= \sum_{i=1}^n y_i\log\left(\frac{\pi_i}{1-\pi_i}\right)+\sum_{i=1}^n\log(1-\pi_i) \\ &= \sum_{i=1}^n y_i\sum_{j=1}^p x_{ij}\beta_j-\sum_{i=1}^n\log\left(\exp\left\{\sum_{j=1}^p x_{ij}\beta_j\right\}+1\right). \end{aligned}$$

The MLE of β does not have a closed form, i.e. a numerical method such as Newton-Raphson is required.

Suggested excercise.

Develop an R-function to evaluate the log-likelihood of a logistic regression. As a template for the function you can use the following

```
negLogLik=function(y,X,b){
  eta=X%*%b
  pi=exp(eta)/(1+exp(eta))
  logLik=sum(ifelse(y==1,log(pi),log(1-pi)))
      return(-logLik)
}
```

Generate small test data set

```
set.seed(195021)
n=1000
X=cbind(1,runif(n))
b=c(.2,.25)
eta=X%*%b
p=exp(eta)/(1+exp(eta))
y=rbinom(n=n,size=1,prob=p)
```

Estimation using `glm` function

The `glm()` function can be used to fit logistic regression models via maximum likelihood.

```
fit=glm(y~X-1,family=binomial(link=logit))
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ X - 1, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4483  -1.3415   0.9528   1.0111   1.0744
##
## Coefficients:
##    Estimate Std. Error z value Pr(>|z|)
## X1   0.2460     0.1317   1.869   0.0617 .
## X2   0.3723     0.2327   1.600   0.1096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1386.3  on 1000  degrees of freedom
## Residual deviance: 1338.4  on  998  degrees of freedom
## AIC: 1342.4
##
## Number of Fisher Scoring iterations: 4
```

```
confint(fit) #95% CI for the coefficients
```

```
## Waiting for profiling to be done...
```

```
##          2.5 %    97.5 %
## X1 -0.01130526 0.5051387
## X2 -0.08304343 0.8297905
```

```
exp(coef(fit)) #exponentiated coefficients
```

```
##       X1       X2
## 1.278936 1.451117
```

```
exp(confint(fit)) #95% CI for exponentiated coefficients
```

```
## Waiting for profiling to be done...
```

```
##        2.5 %   97.5 %
## X1 0.9887584 1.657215
## X2 0.9203112 2.292838
```

```
#predict (fit, type="response") #predicted values
```