Logistic Regression

```
> no.yes=c("No","Yes")
> smoking=gl(2,1,8,no.yes)
> obesity=gl(2,2,8,no.yes)
> snoring=gl(2,4,8,no.yes)
> n.tot=c(60,17,8,2,187,85,51,23)
> n.hyp=c(5,2,1,0,35,13,15,8)
> hypertension=data.frame(smoking,obesity,snoring,n.tot,n.hyp)
> hypertension
  smoking obesity snoring n.tot n.hyp
1      No      No      No    60     5
2     Yes      No      No    17     2
3      No     Yes      No     8     1
4     Yes     Yes      No     2     0
5      No      No     Yes   187    35
6     Yes      No     Yes    85    13
7      No     Yes     Yes    51    15
8     Yes     Yes     Yes    23     8
```

Logistic Regression

```
> hyp.table=cbind(hypertension$n.hyp,hypertension$n.tot-
+  hypertension$n.hyp)
> hyp.table
     [,1] [,2]
[1,]    5   55
[2,]    2   15
[3,]    1    7
[4,]    0    2
[5,]   35  152
[6,]   13   72
[7,]   15   36
[8,]    8   15
```

```
> M1=glm(hyp.table~smoking+obesity+snoring,family =
binomial("logit"))
> summary(M1)
```

Logistic Regression

```
Call:
glm(formula = hyp.table ~ smoking + obesity + snoring, family =
binomial("logit"))
Deviance Residuals:
       1         2         3         4         5         6
7         8
-0.04344   0.54145  -0.25476  -0.80051   0.19759  -0.46602
-0.21262   0.56231

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.37766    0.38018  -6.254    4e-10 ***
smokingYes  -0.06777    0.27812  -0.244   0.8075
obesityYes   0.69531    0.28509   2.439   0.0147 *
snoringYes   0.87194    0.39757   2.193   0.0283 *
---
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 14.1259  on 7  degrees of freedom
Residual deviance:  1.6184  on 4  degrees of freedom
AIC: 34.537
Number of Fisher Scoring iterations: 4
```

Logistic Regression

Alternatively, one can also provide the proportions instead of the counts:

```
> prop.hyp=n.hyp/n.tot
> M1_2=glm(prop.hyp~smoking+obesity+snoring,family=binomial,
weights=n.tot)
> summary(M1_2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.37766    0.38018  -6.254    4e-10 ***
smokingYes  -0.06777    0.27812  -0.244   0.8075
obesityYes   0.69531    0.28509   2.439   0.0147 *
snoringYes   0.87194    0.39757   2.193   0.0283 *
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 14.1259  on 7  degrees of freedom
Residual deviance:  1.6184  on 4  degrees of freedom
AIC: 34.537
Number of Fisher Scoring iterations: 4
```

Logistic Regression

- Null deviance corresponds to the deviance of a model that contains only the interval (and so, a fixed probability of success)

```
> M2=glm(hyp.table~obesity+snoring,family = binomial("logit"))
> summary(M2)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3921     0.3757  -6.366 1.94e-10 ***
obesityYes    0.6954     0.2851   2.440   0.0147 *
snoringYes    0.8655     0.3967   2.182   0.0291 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14.1259  on 7  degrees of freedom
Residual deviance:  1.6781  on 5  degrees of freedom
AIC: 32.597

Number of Fisher Scoring iterations: 4
```

```
> anova(M2)
Analysis of Deviance Table

Model: binomial, link: logit

Response: hyp.table

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev
NULL                        7    14.1259
obesity  1   6.8260         6     7.2999
snoring  1   5.6218         5     1.6781
```

Logistic Regression

```
> M2=glm(hyp.table~obesity+snoring,family = binomial("logit"))
> anova(M2,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: hyp.table

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                        7    14.1259
obesity  1   6.8260         6     7.2999 0.008984 **
snoring  1   5.6218         5     1.6781 0.017738 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Odds-estimates

```
> exp(cbind(OR=coef(M2),confint(M2)))
Waiting for profiling to be done...
                    OR      2.5 %    97.5 %
(Intercept) 0.09143963 0.04035218 0.1794079
obesityYes  2.00454846 1.13362951 3.4791517
snoringYes  2.37609483 1.15143343 5.5609161
```

- Odds ratio per unit change in the covariate. For example, if we consider obesity, the odds ratio associated with obesity is approx 2.0. We refer to the “odds ratio” as the ratio of the odds of developing the disease given exposure and the odds of developing the disease given the non-exposure.
- An odds ratio of 1 indicates the condition is equally likely to occur in both groups.

Example: experiment on the toxicity to the tobacco budworm Heliothis virescens of doses of a pyrethroid (insecticide). Batches of 20 moths of each sex were exposed for 3 days and the number in each batch that were dead or knocked down was recorded.

	Dose					
Sex	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

Logistic Regression

```
>ldose=rep(0:5,2)
>numdead=c(1,4,9,13,18,20,0,2,6,10,12,16)
>sex=factor(rep(c("M","F"),c(6,6)))
>SF=cbind(numdead,numalive=20-numdead)
>M1=glm(SF~sex*ldose,family=binomial)
>summary(M1)

Call:
glm(formula = SF ~ sex * ldose, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.39849  -0.32094  -0.07592   0.38220   1.10375
```

Logistic Regression

increase in "intercept" for males

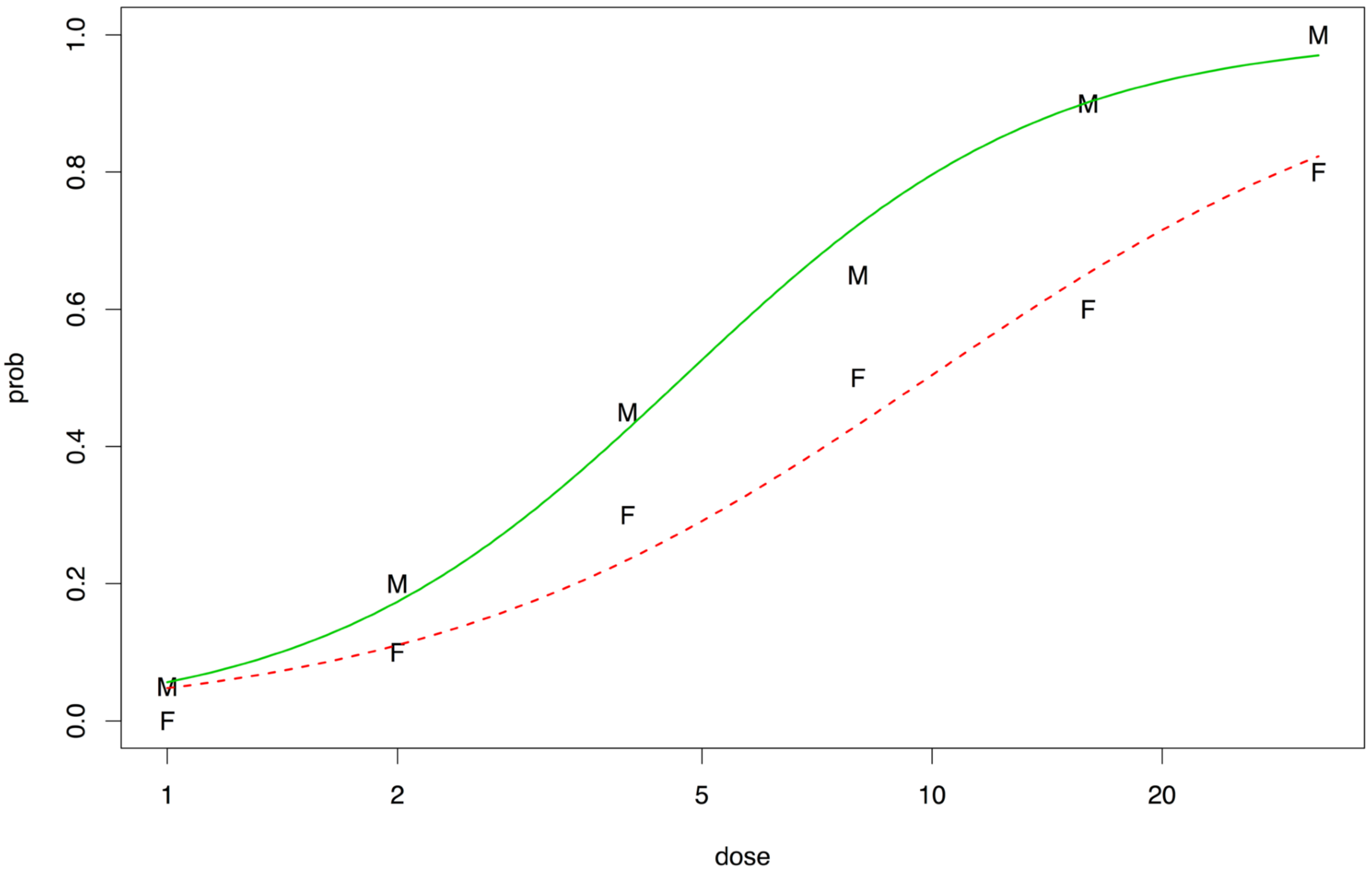
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9935     0.5527  -5.416 6.09e-08 ***
sexM          0.1750     0.7783   0.225    0.822
ldose         0.9060     0.1671   5.422 5.89e-08 ***
sexM:ldose    0.3529     0.2700   1.307    0.191
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   4.9937  on  8  degrees of freedom
AIC: 43.104

Number of Fisher Scoring iterations: 4
```

increase in "slope" for males

Logistic Regression

```
>plot(c(1,32),c(0,1),type="n",xlab="dose",ylab="prob",log="x")
> text(2^ldose,numdead/20,labels=as.character(sex))
> ld=seq(0,5,0.1)
> lines(2^ld,predict(M1,
+   data.frame(ldose=ld,sex=factor(rep("M",length(ld)),
+   levels=levels(sex))),type="response"),col=3,lwd=2)
> lines(2^ld,predict(M1,
+   data.frame(ldose=ld,sex=factor(rep("F",length(ld)),
+   levels=levels(sex))),type="response"),lty=2,col=2,lwd=2)
```

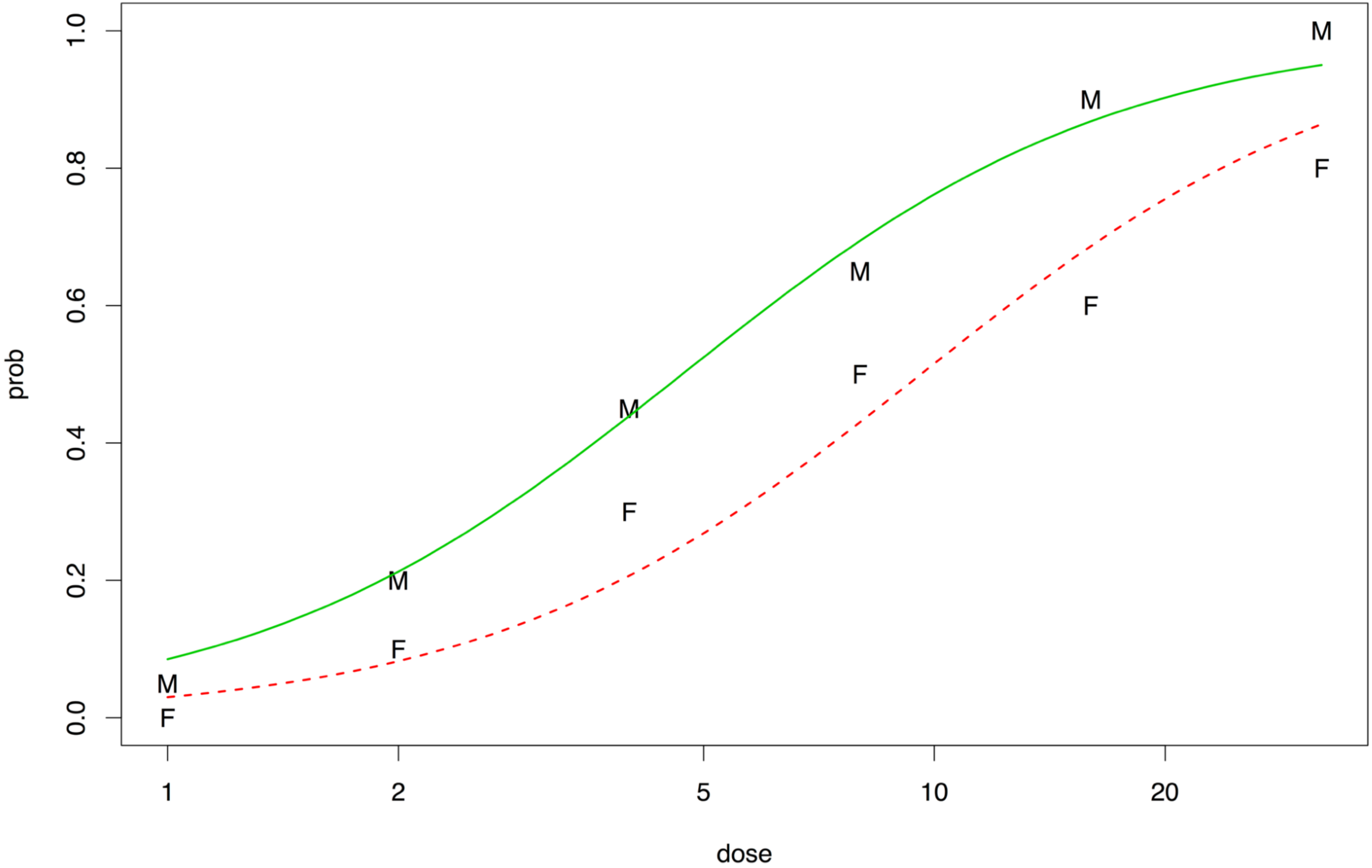
Logistic Regression



Logistic Regression

```
> M2=glm(SF~sex+ldose,family=binomial)
> summary(M2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.10540  -0.65343  -0.02225   0.48471   1.42944

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.4732     0.4685  -7.413 1.23e-13 ***
sexM          1.1007     0.3558   3.093  0.00198 **
ldose         1.0642     0.1311   8.119 4.70e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   6.7571  on  9  degrees of freedom
AIC: 42.867

Number of Fisher Scoring iterations: 4
```

Logistic Regression



Titanic Data. The Titanic sank on April 15, 1912, killing 1502 out of 2224 passengers and crews. There were not enough lifeboats. Some groups of people were more likely to survive than others, such as women, children, and the upper class. The data is from Kaggle, but available in **R**.

```
> library(titanic)
> str(titanic_train)
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs.
John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina"
"Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
 $ Sex        : chr  "male" "female" "female" "female" ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282"
"113803" ...
```

Variables:

- Survival: 1 (Yes) and 0 (No)
- Pclass: 1st, 2nd and 3rd
- Sex: male, female
- Age
- Passenger ID
- Name
- SibSp: #siblings/spouses
- Parch: #of parents/children aboard
- Ticket: Ticket number
- Embarked: Port of embarkation (C=Cherbourg, Q=Queenstown, S=Southampton)
- Fare
- Cabin: cabin number

Logistic Regression

We consider a subset of 800 passengers and some variables:

```
> sapply(titanic_train,function(x) sum(is.na(x)))
PassengerId    Survived      Pclass        Name         Sex         Age
          0           0           0           0           0         177
      SibSp       Parch      Ticket        Fare       Cabin    Embarked
          0           0           0           0           0           0
```

```
>my_titanic=data.frame(Survived=Survived[1:800],
Pclass=Pclass[1:800],Age=Age[1:800],SibSp=SibSp[1:800],
Sex=Sex[1:800],Parch=Parch[1:800],Fare=Fare[1:800])

>M1=glm(Survived~.,family=binomial,data=my_titanic)
>summary(M1)
```

We will also consider some test data:

```
> my_titanic_test=data.frame(Survived=Survived[801:891],
+Pclass=as.factor(Pclass[801:891]),Age=Age[801:891],SibSp=SibSp[801:891],
+Sex=Sex[801:891],Parch=Parch[801:891],Fare=Fare[801:891])

> missing_index=(1:91)[apply(apply((my_titanic_test),2,is.na),1,sum)==1]

> my_titanic_test=my_titanic_test[-missing_index,]

> length(my_titanic_test$Survived)
> 7
```

Logistic Regression

```
> summary(M1)
Call:
glm(formula = Survived ~ ., family = binomial, data = my_titanic)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7382  -0.6477  -0.3944   0.6375   2.4249
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.119120   0.528061   7.800 6.17e-15 ***
Pclass2     -1.233680   0.338641  -3.643 0.000269 ***
Pclass3     -2.482686   0.355952  -6.975 3.06e-12 ***
Age         -0.041966   0.008630  -4.863 1.16e-06 ***
SibSp       -0.338228   0.133493  -2.534 0.011287 *
Sexmale     -2.645685   0.233589 -11.326  < 2e-16 ***
Parch       -0.097362   0.132937  -0.732 0.463927
Fare         0.001470   0.002518   0.584 0.559173

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 860.71  on 636  degrees of freedom
Residual deviance: 569.54  on 629  degrees of freedom
  (163 observations deleted due to missingness)
AIC: 585.54
Number of Fisher Scoring iterations: 5
```

Logistic Regression

```
> M2=glm(Survived~Pclass+Age+SibSp+Sex,family=binomial,data=my_titanic)
> summary(M2)
Deviance Residuals:
    Min        1Q    Median        3Q       Max
-2.7543  -0.6436  -0.3904   0.6268   2.4376

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.213846   0.469989   8.966  < 2e-16 ***
Pclass2     -1.326588   0.298966  -4.437 9.11e-06 ***
Pclass3     -2.601180   0.299387  -8.688  < 2e-16 ***
Age         -0.042482   0.008573  -4.955 7.23e-07 ***
SibSp       -0.358688   0.127083  -2.822  0.00477 **
Sexmale     -2.619322   0.227509 -11.513  < 2e-16 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 860.71  on 636  degrees of freedom
Residual deviance: 570.28  on 631  degrees of freedom
  (163 observations deleted due to missingness)
AIC: 582.28

Number of Fisher Scoring iterations: 5
```

Logistic Regression

```
> exp(cbind(OR=coef(M2),confint(M2)))
Waiting for profiling to be done...
                     OR       2.5 %      97.5 %
(Intercept) 67.61612046 27.76619768 175.7418514
Pclass2      0.26538108  0.14614501   0.4726936
Pclass3      0.07418596  0.04057979   0.1314865
Age          0.95840741  0.94204916   0.9743052
SibSp        0.69859232  0.53995155   0.8898612
Sexmale      0.07285225  0.04608315   0.1125910
```

```
> fitted_results=predict(M2,my_titanic_test,type='response')
> fitted_results<-ifelse(fitted_results>0.5,1,0)
> mean(fitted_results==my_titanic_test$Survived)
[1] 0.8181818
```