Logistic Regression example with many independent variables:

Health study investigated epidemic outbreak of a disease spread by mosquitoes. 196 individuals from two sectors of a city were selected and the following variables were recorded:

Response variable Y=1 if disease was present, 0 otherwise
P = Pr(Y=1) probability of disease
Age
Socioeconomic status (upper, middle, lower)
D_{lower} =1 for status = lower *(dummy variable)*
D_{middle} =1 for status = middle *(dummy variable)*
Sector of the city (sector 1, sector 2)
$D_{sector2}$ = 1 for sector =2 *(dummy variable)*

Summary of analysis

```
> # logistic regression example
> # Dataset for programmers project success and months of experience
> myd= read.table("diseaseoutbreak_all.txt", header=T)
> myd[1,]
  case age status sector disease account
1    1  33      1      1       0       1
> #create dummy variables for sector and status;
> st.m=(myd$status==2)*1
> st.l=(myd$status==3)*1
> sec.2 = (myd$sector==2)*1
> myd=cbind(myd, st.m, st.l,sec.2)
> # FIT FULL LOGISTIC MODEL
```

The full logistic regression model to predict probability of disease p=Pr(disease=1) is as follows:

log(p/(1-p)= -2.293+0.027 age + 0.045 middle + 0.253 lower +1.243 sector2 +e
Note that the dummy variables for status are not significant.

```
> # logistic regression model fitted using glm() function with family=binomial
> full <- glm(disease~age +st.m+st.l+sec.2, data=myd, family=binomial())
> summary(full) # display results
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6576  -0.8295  -0.5652   1.0092   2.0842
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.293933   0.436769  -5.252  1.5e-07 ***
age          0.026991   0.008675   3.111 0.001862 **
st.m         0.044609   0.432490   0.103 0.917849
st.l         0.253433   0.405532   0.625 0.532011
sec.2        1.243630   0.352271   3.530 0.000415 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 236.33  on 195  degrees of freedom
Residual deviance: 211.22  on 191  degrees of freedom
AIC: 221.22
Number of Fisher Scoring iterations: 3

> # APPLY VARIABLE SELECTION PROCEDURES
> # Run backward selection procedure for variable selection
> # step() function works for both lm and glm objects
> step.mod=step(full, direction=c("backward"))
Start:  AIC=221.22
disease ~ age + st.m + st.l + sec.2
        Df Deviance    AIC
- st.m   1   211.23 219.23
- st.l   1   211.61 219.61
<none>       211.22 221.22
- age    1   221.26 229.26
- sec.2  1   224.22 232.22

Step:  AIC=219.23
disease ~ age + st.l + sec.2

        Df Deviance    AIC
- st.l   1   211.64 217.64
<none>       211.23 219.23
- age    1   221.30 227.30
- sec.2  1   224.22 230.22

Step:  AIC=217.64
disease ~ age + sec.2

        Df Deviance    AIC
<none>       211.64 217.64
- age    1   221.60 225.60
- sec.2  1   224.32 228.32

> #fit selected model
> fit <- glm(disease~age +sec.2, data=myd, family=binomial())
> summary(fit) # display results

Call:
glm(formula = disease ~ age + sec.2, family = binomial(), data = myd)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6839  -0.8199  -0.5607   1.0093   2.0275
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

Selected model

Variables that were removed

```
(Intercept) -2.15966    0.34388  -6.280 3.38e-10 ***
age          0.02681    0.00865   3.100 0.001936 **
sec.2        1.18169    0.33696   3.507 0.000453 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 236.33  on 195  degrees of freedom
Residual deviance: 211.64  on 193  degrees of freedom
AIC: 217.64

Number of Fisher Scoring iterations: 3
```

The selected logistic regression model M1 for p=Pr(disease=1) is
log(p/(1-p)= -2.16+0.027 age +1.181 sector2 +e
since age and sector 2 have a positive parameter, they both have a positive association with p.

```
> #compute goodness of fit test (likelihood ratio test)
> library(lmtest) #in lmtest package
> lrtest(fit)
Likelihood ratio test

Model 1: disease ~ age + sec.2
Model 2: disease ~ 1
  #Df  LogLik Df Chisq Pr(>Chisq)
1   3 -105.82
2   1 -118.17 -2 24.69  4.351e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.0
```

Goodness of fit test:
Ho: null model b1=b2=0
H1: Model M1 with b1 and b2 ≠ 0

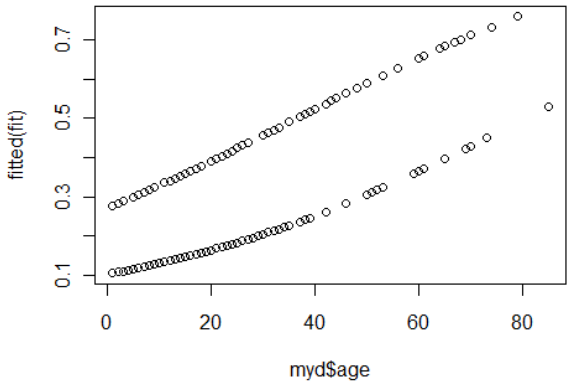
LR statistic = 24.69 with chi-square distribution with 2 DF.
P-value is almost zero (4.351e-06). So we can conclude that null hypothesis can be rejected and current model is better than the null model.

```
> confint(fit) # 95% CI for the coefficients
Waiting for profiling to be done...
                  2.5 %      97.5 %
(Intercept) -2.86990940 -1.51605906
age          0.01010532  0.04421365
sec.2        0.52854584  1.85407936
> exp(coef(fit)) # compute exp(coefficients) to analyze change in odds for
changes in X
(Intercept)         age       sec.2
  0.1153644   1.0271756   3.2598900
> exp(confint(fit)) # 95% CI for exp(coefficients), that is change in odds
Waiting for profiling to be done...
                 2.5 %    97.5 %
(Intercept) 0.05670406 0.2195755
age         1.01015655 1.0452056
sec.2       1.69646359 6.3858165
```

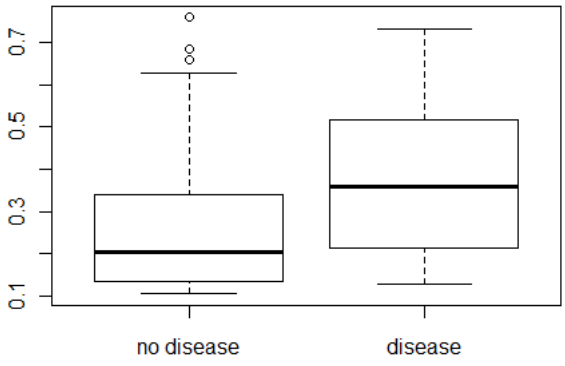
The selected logistic regression model M1 for p=Pr(disease=1) is
log(p/(1-p)= -2.16+0.027 age +1.181 sector2 +e
since age and sector 2 have a positive parameter, they both have a positive association with p.

The coefficient beta for a certain variable X represents the change in log(odds) for any 1-unit increase in X. Exp(beta) is the estimated odds ratio for a unit increase of X, and is equal to the odds at X+1 divided by the odds at X: odds_ratio=odds(x+1)/odds(x).
Thus, exp(beta) represents the change (increase or decrease) in odds for any 1-unit increase in X. If exp(beta)>1, the odds increase, if exp(beta)<1, the odds decrease (this is explained in week 7 slides).
INTERPRETATION OF MODEL 1 PARAMETERS:
AGE: for people living in the same sector, any additional year of age increases the average odds of disease by 2.7%, and we are 95% confidence that the average increase is between 1% and 4.5%
SECTOR2: for people with the same age, if they live in sector 2 the odds of disease are 326% the odds of disease in sector 1, or in other words, living in sector 2 increases the odds of disease by 226%. Also, we are 95% confidence that the average increase is between 69% and 538%.

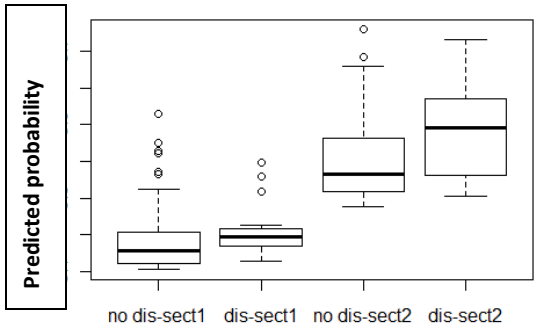
```
> predict(fit, type="response") # predicted values
> residuals(fit, type="deviance") # residuals
> #plot of predicted probabilities vs age
> plot(myd$age, fitted(fit))
```




```
> #boxplot of predicted probabilities by disease
> # useful visualization for classification purposes
> boxplot(fitted(fit)~myd$disease, names=c("no disease", "disease"))
```



```
> #boxplot of predicted probabilities by task success and sector
> # useful visualization for classification purposes
> boxplot(fitted(fit)~myd$disease*sec.2, names=c("no dis-sect1", "dis-
sect1", "no dis-sect2", "dis-sect2"))
```



The boxplot shows the distribution of predicted probabilities for each observed patient by disease (whether sick or not) and sector (1 or 2). The predicted probabilities for people in sector 1 are all below 0.3 (indicating low chance of disease in sector 1) and the predicted probabilities for sector 2 are all above 0.3 (indicating higher chance of disease in sector 2).

```
> #compute predicted probability of contracting disease for a 20 year old
> # individual living in the second sector of the city
> newd = data.frame(age=c(20), sec.2=c(1))
> predict(fit,newdata=newd,type="response", se.fit=T) #type="response" for
probabilities
$fit
        1
0.3913341

$se.fit
         1
0.05817779

$residual.scale
[1] 1
```

The predicted probability for a 20 year old individual living in sector 2 of the city is computed as $\hat{p} = 0.39$ with a standard error of 0.058. So we can say that the 95% prediction interval is

0.39+/- 1.96*0.058 or (0.27, 0.50) (using the asymptotic normality of predictions)

Note that the predicted value of p can be computed manually from the model:

$\log(\frac{\hat{p}}{1-\hat{p}})$ = *-2.16+0.027*20 +1.181 *1 = -0.439 and therefore*

$\hat{p} = \exp(-0.439)/(1+\exp(-0.439) = 0.39$