Logistic.Regression

By Saba Akram

3/1/2021

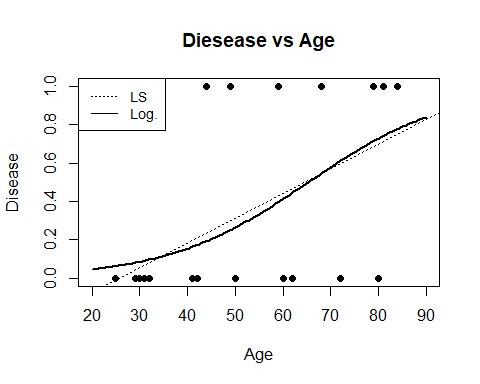
## logistic Regression  
 patients<-data.frame(age=c(25,29,30,31,32,41,41,42,44,49,50,59,60,62,68,72,79,80,81,84),disease=c(0,0,0,0,0,0,0,0,1,1,0,1,0,0,1,0,1,0,1,1))  
lm1<-lm(disease~age,data = patients)  
lm1

##   
## Call:  
## lm(formula = disease ~ age, data = patients)  
##   
## Coefficients:  
## (Intercept) age   
## -0.33357 0.01291

lr1<-glm(disease~age,data = patients,family = binomial)  
lr1

##   
## Call: glm(formula = disease ~ age, family = binomial, data = patients)  
##   
## Coefficients:  
## (Intercept) age   
## -4.37210 0.06696   
##   
## Degrees of Freedom: 19 Total (i.e. Null); 18 Residual  
## Null Deviance: 25.9   
## Residual Deviance: 20.2 AIC: 24.2

plot(patients$age,patients$disease,xlab = "Age",ylab = "Disease",main = "Diesease vs Age",xlim = c(20,90),pch=16)  
abline(lm1,lty=3)  
curve(predict(lr1,data.frame(age=x),type = "resp"),add = TRUE,lwd=2)  
legend("topleft",legend = c("LS","Log."),lty = c(3,1),cex = .9)



anova(lr1)

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: disease  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev  
## NULL 19 25.898  
## age 1 5.6964 18 20.201

with(lr1,null.deviance-deviance)

## [1] 5.696407

## Deviance is a measure of goodness of fit of a model. Higher numbers always indicates bad fit. The null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean) where as residual with inclusion of independent variables.  
with(lr1,df.null-df.residual)

## [1] 1

with(lr1,pchisq(null.deviance-deviance,df.null-df.residual,lower.tail = FALSE))

## [1] 0.01699968

with

## function (data, expr, ...)   
## UseMethod("with")  
## <bytecode: 0x0000000015256488>  
## <environment: namespace:base>

## Make Predictions. [g(x) and π(x)]  
newd<-with(patients,data.frame(age=c(50,72)))  
predict.glm(lr1,newdata = newd) ## Log odds

## 1 2   
## -1.0242945 0.4487388

## It is also called estimated logit g(x)=ln[π(x)/(1-π(x))]= β0 + β1x.  
  
predict(lr1,newdata = newd,type = "resp")

## 1 2   
## 0.2641917 0.6103393

## now by adding ome more parameter i.e. type="resp" gives us the probablity.  
## Hence we observe that the estimated probablity that a 50 year old patient has the disease is 26% and the estimated probablity that the 50 year old does not have any disease is 100-26=74%  
  
## Similarly the probability that a 72 year old has the disease is 61% and the 72 year old does not have any disease is 100-61=39%.

## Odds Ratios  
  
##urred compared to it is not occurring.  
  
# In binary logistic regression with a dichotomous predictor, the odds that the response variable occurred (y=1) for records with x=1 can be denoted as:  
  
# 𝜋(1)/[1-𝜋(1)] = [e^(𝛽0+𝛽1)/[1+e^(𝛽0+𝛽1)]]/[1/(1+e^(𝛽0+𝛽1))] = e^(𝛽0+𝛽1)  
  
# Correspondingly, the odds that the response variable occurred for records with x=0 can be denoted as: e^(𝛽0)  
  
# Hence Odds Ratio = e^(𝛽0+𝛽1)/e^(𝛽0) = e^(𝛽1)  
  
round(exp(coef(lr1)),3)

## (Intercept) age   
## 0.013 1.069

# Logistics Regression using Dichotomous Example  
  
churn<-read.csv("D:/M.Sc in Banking and Financial Analytics/Sem 3/Data Analytic and Machine learning/Data Mining and Predictive Analysis/Data sets/churn.txt", stringsAsFactors=TRUE)  
table(churn$Churn.,churn$VMail.Plan)

##   
## no yes  
## False. 2008 842  
## True. 403 80

churn$VMail.Plan.ind<-ifelse(churn$VMail.Plan=="yes",1,0)  
lr2<-glm(Churn.~VMail.Plan.ind,data = churn,family = "binomial")  
summary(lr2)

##   
## Call:  
## glm(formula = Churn. ~ VMail.Plan.ind, family = "binomial", data = churn)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.6048 -0.6048 -0.6048 -0.4261 2.2111   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.60596 0.05458 -29.422 < 2e-16 \*\*\*  
## VMail.Plan.ind -0.74780 0.12910 -5.792 6.94e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2758.3 on 3332 degrees of freedom  
## Residual deviance: 2720.3 on 3331 degrees of freedom  
## AIC: 2724.3  
##   
## Number of Fisher Scoring iterations: 5

# Dichotomous Example : Odds Ratio and Prediction.  
  
# Odds Ratio  
round(exp(coef(lr2)),3)

## (Intercept) VMail.Plan.ind   
## 0.201 0.473

# Making Predictions  
newp<-with(churn,data.frame(VMail.Plan.ind=c(0,1)))  
predict.glm(lr2,newdata = newp) #g(x)

## 1 2   
## -1.605958 -2.353753

predict.glm(lr2,newdata = newp,type = "resp") #π(x)

## 1 2   
## 0.1671506 0.0867679

# Polychotomous Example  
  
# Redefining customer service calls  
  
churn$CSC<-factor(churn$CustServ.Calls)  
levels(churn$CSC)

## [1] "0" "1" "2" "3" "4" "5" "6" "7" "8" "9"

levels(churn$CSC)[1:2]<-"Low"  
levels(churn$CSC)[2:3]<-"Medium"  
levels(churn$CSC)[3:8]<-"High"  
churn$CSC\_Medium<-ifelse(churn$CSC=="Medium",1,0)  
churn$CSC\_High<-ifelse(churn$CSC=="High",1,0)  
table(churn$Churn.,churn$CSC)

##   
## Low Medium High  
## False. 1664 1057 129  
## True. 214 131 138

lr3<-glm(Churn.~CSC\_Medium+CSC\_High,data = churn,family = "binomial")  
summary(lr3)

##   
## Call:  
## glm(formula = Churn. ~ CSC\_Medium + CSC\_High, family = "binomial",   
## data = churn)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.2062 -0.4919 -0.4919 -0.4834 2.0999   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.05100 0.07262 -28.243 <2e-16 \*\*\*  
## CSC\_Medium -0.03699 0.11770 -0.314 0.753   
## CSC\_High 2.11844 0.14238 14.879 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2758.3 on 3332 degrees of freedom  
## Residual deviance: 2526.7 on 3330 degrees of freedom  
## AIC: 2532.7  
##   
## Number of Fisher Scoring iterations: 4

lr4<-glm(Churn.~Day.Mins,data = churn,family = "binomial")  
summary(lr4)

##   
## Call:  
## glm(formula = Churn. ~ Day.Mins, family = "binomial", data = churn)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0241 -0.6001 -0.4902 -0.3738 2.8102   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.929289 0.202823 -19.37 <2e-16 \*\*\*  
## Day.Mins 0.011272 0.000975 11.56 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2758.3 on 3332 degrees of freedom  
## Residual deviance: 2614.3 on 3331 degrees of freedom  
## AIC: 2618.3  
##   
## Number of Fisher Scoring iterations: 5

# Adult data example  
  
adult<-read.csv("D:/M.Sc in Banking and Financial Analytics/Sem 3/Data Analytic and Machine learning/Data Mining and Predictive Analysis/Data sets/adult/Clem3Training", stringsAsFactors=TRUE)  
  
adult$over50K<-ifelse(adult$income==">50K.",1,0)  
adult$capnet<-adult$capital.gain-adult$capital.loss  
lr5<-glm(over50K~capnet,data = adult,family = "binomial")  
summary(lr5)

##   
## Call:  
## glm(formula = over50K ~ capnet, family = "binomial", data = adult)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.3015 -0.6853 -0.6853 -0.5596 2.1775   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.329e+00 1.599e-02 -83.13 <2e-16 \*\*\*  
## capnet 2.561e-04 7.871e-06 32.54 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 27517 on 24999 degrees of freedom  
## Residual deviance: 25455 on 24998 degrees of freedom  
## AIC: 25459  
##   
## Number of Fisher Scoring iterations: 6

# Adult Data Example: Catgorizing variables  
  
adult$cap\_level<-factor(adult$capnet)  
levels(adult$cap\_level)

## [1] "-4356" "-3900" "-3770" "-3683" "-3004" "-2824" "-2603" "-2559" "-2547"  
## [10] "-2489" "-2467" "-2457" "-2444" "-2415" "-2392" "-2377" "-2352" "-2339"  
## [19] "-2282" "-2267" "-2258" "-2246" "-2238" "-2231" "-2206" "-2205" "-2201"  
## [28] "-2179" "-2174" "-2163" "-2149" "-2129" "-2080" "-2057" "-2051" "-2042"  
## [37] "-2002" "-2001" "-1980" "-1977" "-1974" "-1944" "-1902" "-1887" "-1876"  
## [46] "-1848" "-1844" "-1825" "-1816" "-1762" "-1755" "-1741" "-1740" "-1735"  
## [55] "-1726" "-1721" "-1719" "-1672" "-1669" "-1668" "-1651" "-1648" "-1628"  
## [64] "-1617" "-1602" "-1594" "-1590" "-1579" "-1573" "-1564" "-1539" "-1504"  
## [73] "-1485" "-1408" "-1380" "-1340" "-1258" "-1138" "-1092" "-974" "-880"   
## [82] "-810" "-653" "-625" "-419" "-323" "-213" "-155" "0" "114"   
## [91] "401" "594" "914" "991" "1055" "1086" "1111" "1151" "1173"   
## [100] "1409" "1424" "1455" "1471" "1506" "1639" "1797" "1831" "1848"   
## [109] "2009" "2036" "2050" "2062" "2105" "2174" "2176" "2202" "2228"   
## [118] "2290" "2329" "2346" "2354" "2407" "2414" "2463" "2538" "2580"   
## [127] "2597" "2635" "2653" "2829" "2885" "2907" "2936" "2961" "2964"   
## [136] "2977" "2993" "3103" "3137" "3273" "3325" "3411" "3418" "3432"   
## [145] "3456" "3464" "3471" "3674" "3781" "3818" "3887" "3908" "3942"   
## [154] "4064" "4101" "4386" "4416" "4508" "4650" "4687" "4787" "4865"   
## [163] "4931" "4934" "5013" "5178" "5455" "5556" "5721" "6097" "6360"   
## [172] "6418" "6497" "6514" "6723" "6767" "6849" "7298" "7430" "7443"   
## [181] "7688" "7896" "7978" "8614" "9386" "9562" "10520" "10566" "10605"  
## [190] "11678" "13550" "14084" "14344" "15020" "15024" "15831" "18481" "20051"  
## [199] "22040" "25124" "25236" "27828" "34095" "41310" "99999"

levels(adult$cap\_level)[1:88]<-"Loss"  
levels(adult$cap\_level)[2]<-"None"  
levels(adult$cap\_level)[3:77]<-"Gain<$5K."  
levels(adult$cap\_level)[4:44]<-"Gain>=$5K."  
adult$cap\_loss<-ifelse(adult$cap\_level=="Loss",1,0)  
adult$cap\_5K<-ifelse(adult$cap\_level=="Gain<$5K.",1,0)  
adult$cap\_ge5k<-ifelse(adult$cap\_level=="Gain>=$5K.",1,0)

# Adult Data Example : Regression Model  
  
lr6<-glm(over50K~cap\_loss+cap\_5K+cap\_ge5k,data = adult,family = "binomial")  
summary(lr6)

##   
## Call:  
## glm(formula = over50K ~ cap\_loss + cap\_5K + cap\_ge5k, family = "binomial",   
## data = adult)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1576 -0.6489 -0.6489 -0.6099 1.8835   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.45088 0.01728 -83.954 <2e-16 \*\*\*  
## cap\_loss 1.46472 0.06131 23.890 <2e-16 \*\*\*  
## cap\_5K -0.13689 0.09435 -1.451 0.147   
## cap\_ge5k 3.67595 0.09685 37.957 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 27517 on 24999 degrees of freedom  
## Residual deviance: 24313 on 24996 degrees of freedom  
## AIC: 24321  
##   
## Number of Fisher Scoring iterations: 4

# Multiple logistic regression  
  
churn$Int.l.Plan.ind<-ifelse(churn$Int.l.Plan=="yes",1,0)  
churn$VMail.Plan.ind<-ifelse(churn$VMail.Plan=="yes",1,0)  
lr7<-glm(Churn.~Int.l.Plan.ind+VMail.Plan.ind+Day.Mins+Eve.Mins+Night.Mins+Intl.Mins,data = churn,family = "binomial")  
summary(lr7)

##   
## Call:  
## glm(formula = Churn. ~ Int.l.Plan.ind + VMail.Plan.ind + Day.Mins +   
## Eve.Mins + Night.Mins + Intl.Mins, family = "binomial", data = churn)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5557 -0.5512 -0.4173 -0.2726 3.0784   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.718625 0.453342 -14.820 < 2e-16 \*\*\*  
## Int.l.Plan.ind 1.800469 0.136405 13.199 < 2e-16 \*\*\*  
## VMail.Plan.ind -0.892620 0.138569 -6.442 1.18e-10 \*\*\*  
## Day.Mins 0.011608 0.001024 11.331 < 2e-16 \*\*\*  
## Eve.Mins 0.006069 0.001070 5.671 1.42e-08 \*\*\*  
## Night.Mins 0.003087 0.001052 2.934 0.003350 \*\*   
## Intl.Mins 0.074214 0.019412 3.823 0.000132 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2758.3 on 3332 degrees of freedom  
## Residual deviance: 2356.3 on 3326 degrees of freedom  
## AIC: 2370.3  
##   
## Number of Fisher Scoring iterations: 5

# Higher Order Terms.  
adult$age\_sq<-adult$age^2  
lr8<-glm(over50K~age+age\_sq,data = adult,family = "binomial")  
summary(lr8)

##   
## Call:  
## glm(formula = over50K ~ age + age\_sq, family = "binomial", data = adult)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0466 -0.8746 -0.4929 -0.1756 3.3850   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.080e+00 1.945e-01 -46.68 <2e-16 \*\*\*  
## age 3.478e-01 8.946e-03 38.88 <2e-16 \*\*\*  
## age\_sq -3.450e-03 9.922e-05 -34.77 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 27517 on 24999 degrees of freedom  
## Residual deviance: 24442 on 24997 degrees of freedom  
## AIC: 24448  
##   
## Number of Fisher Scoring iterations: 5

# Validating the model  
# Prepare the data  
levels(adult$marital.status)

## [1] "Divorced" "Married-AF-spouse" "Married-civ-spouse"   
## [4] "Married-spouse-absent" "Never-married" "Separated"   
## [7] "Widowed"

levels(adult$marital.status)[2:4]<-"Married"  
levels(adult$marital.status)

## [1] "Divorced" "Married" "Never-married" "Separated"   
## [5] "Widowed"

adult$msmarried<-ifelse(adult$marital.status=="Married",1,0)  
adult$msnevermarried<-ifelse(adult$marital.status=="Never-married",1,0)  
adult$msseparated<-ifelse(adult$marital.status=="Separated",1,0)  
adult$mswidowed<-ifelse(adult$marital.status=="Widowed",1,0)  
adult$capnet<-adult$capital.gain-adult$capital.loss  
levels(adult$sex)

## [1] "Female" "Male"

adult$male<-ifelse(adult$sex=="Male",1,0)  
  
## Creating a holdout sample  
  
hold<-runif(dim(adult)[1],0,1)  
trainA<-adult[which(hold<.5),]  
trainB<-adult[which(hold>=.5),]  
dim(trainA)

## [1] 12555 27

dim(trainB)

## [1] 12445 27

# Validating the models : Running the models  
lr11A<-glm(over50K~age+education.num+msmarried+msnevermarried+msseparated+mswidowed+male+hours.per.week+capnet,data = trainA,family = "binomial")  
lr11B<-glm(over50K~age+education.num+msmarried+msnevermarried+msseparated+mswidowed+male+hours.per.week+capnet,data = trainB,family = "binomial")  
summary(lr11A)

##   
## Call:  
## glm(formula = over50K ~ age + education.num + msmarried + msnevermarried +   
## msseparated + mswidowed + male + hours.per.week + capnet,   
## family = "binomial", data = trainA)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9167 -0.5804 -0.2434 -0.0684 3.2360   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -8.964e+00 2.302e-01 -38.946 < 2e-16 \*\*\*  
## age 2.410e-02 2.329e-03 10.347 < 2e-16 \*\*\*  
## education.num 3.728e-01 1.212e-02 30.771 < 2e-16 \*\*\*  
## msmarried 2.016e+00 1.013e-01 19.893 < 2e-16 \*\*\*  
## msnevermarried -6.319e-01 1.281e-01 -4.933 8.09e-07 \*\*\*  
## msseparated 8.603e-02 2.352e-01 0.366 0.71450   
## mswidowed 1.666e-01 2.289e-01 0.728 0.46666   
## male 2.117e-01 7.493e-02 2.826 0.00472 \*\*   
## hours.per.week 3.229e-02 2.327e-03 13.880 < 2e-16 \*\*\*  
## capnet 2.300e-04 1.279e-05 17.988 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 13781.6 on 12554 degrees of freedom  
## Residual deviance: 8792.1 on 12545 degrees of freedom  
## AIC: 8812.1  
##   
## Number of Fisher Scoring iterations: 6

summary(lr11B)

##   
## Call:  
## glm(formula = over50K ~ age + education.num + msmarried + msnevermarried +   
## msseparated + mswidowed + male + hours.per.week + capnet,   
## family = "binomial", data = trainB)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.8598 -0.5554 -0.2423 -0.0706 3.2701   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -8.948e+00 2.325e-01 -38.492 < 2e-16 \*\*\*  
## age 2.624e-02 2.351e-03 11.159 < 2e-16 \*\*\*  
## education.num 3.704e-01 1.215e-02 30.482 < 2e-16 \*\*\*  
## msmarried 2.030e+00 1.026e-01 19.791 < 2e-16 \*\*\*  
## msnevermarried -4.555e-01 1.253e-01 -3.634 0.000279 \*\*\*  
## msseparated -2.896e-01 2.553e-01 -1.135 0.256548   
## mswidowed -4.324e-01 2.526e-01 -1.711 0.087004 .   
## male 2.692e-01 7.536e-02 3.572 0.000354 \*\*\*  
## hours.per.week 3.020e-02 2.393e-03 12.621 < 2e-16 \*\*\*  
## capnet 2.447e-04 1.342e-05 18.237 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 13735.3 on 12444 degrees of freedom  
## Residual deviance: 8728.2 on 12435 degrees of freedom  
## AIC: 8748.2  
##   
## Number of Fisher Scoring iterations: 6