Assignment 2

Jesse Smart

31 August 2019

Contents

a) Mean and Variance expressions of the Bernoulli distribution

$$
\begin{aligned}
f(y,\pi) &= \pi^y(1-y)^{1-y} \\
&= \exp(ylog(\frac{\pi}{1-\pi}) + log(1-\pi)) \\
a(y) &= y \\
b(\pi) &= log(\frac{\pi}{1-\pi}) \\
c(\pi) &= log(1-\pi)
\end{aligned}
$$

$$
\begin{aligned}
\text{mean}(a(y)) &= \frac{-c\prime(\pi)}{b\prime(\pi)} \\
&= \pi
\end{aligned}
$$

$$
\begin{aligned}
\text{var}(a(y)) &= \frac{b\prime\prime(\pi)c\prime(\pi) - c\prime\prime(\pi)b\prime(\pi)}{b\prime(\pi)^3} \\
&= \pi(1-\pi)
\end{aligned}
$$

b) Weights expression

$$
\begin{aligned}
g(\mu_i) = \eta_i &= logit(\mu_i) = log(\frac{\mu_i}{1-\mu_i}) \\
\mu_i &= \frac{e^{\eta_i}}{1+e^{\eta_i}} \\
\frac{\partial \mu_i}{\partial \eta_i} &= \frac{e^{\eta_i}}{(1+e^{\eta_i})^2} \\
\text{var}(Y_i) &= \mu_i(1-\mu_i) \\
w_{ii} &= \frac{1}{\text{var}(Y_i)}(\frac{\partial \eta_i}{\partial \mu_i})^2 \\
&= \frac{1}{\mu_i(1-\mu_i)}(\frac{e^{\eta_i}}{(1+e^{\eta_i})^2})^2 \\
&= \frac{1}{\mu_i(1-\mu_i)}\mu_i^2(1-\mu_i)^2 \\
&= \mu_i(1-\mu_i)
\end{aligned}
$$

$$
\begin{aligned}
U_j &= \sum_{i=1}^{N} \frac{(yi-\mu_i)}{\text{var}(Yi)} x_{ij}(\frac{\partial \mu_i}{\partial \eta_i}) \\
&= \sum_{i=1}^{N} \frac{(yi-\mu_i)}{\mu_i(1-\mu_i)} x_{ij}(\mu_i(1-\mu_i)) \\
&= \sum_{i=1}^{N} x_{ij}(y_i - \mu_i)
\end{aligned}
$$

$$
\therefore U^{(m-1)} = X^T(\mathrm{Y} - \mu)
$$

c) glm_fit function

```
#glm_fit
glm_fit = function(y, x, beta_start, k = 1){
  w = matrix(0, nrow = length(y), ncol = length(y))
  beta = beta_start

  for( i in 2:k ){
    eta = x %*% beta
    mu = exp(eta)/(1 + exp(eta))
    diag(w) = mu * (1- mu)

    info = t(x) %*% w %*% x
    inverseJ  = solve(info)
    U = t(x) %*% (y-mu)

    beta = beta + inverseJ %*% U
    stder = diag(inverseJ)^0.5

  }
  return(list(beta, stder))
}
```

d) glm_fit output

The following table displays the output Betas and Standard errors from the glm_fit function written in c)

	rows	1	2
columns	0	Estimates	Stnd.err.
X	(intercept)	7.54692218343798	1.37366524017017
X.1	Sexmale	-6.45065649269823	1.1812424691531
X.2	Age	-0.0390785768403112	0.0119939113151842
X.3	Pclass	-2.20955369237628	0.442874175878738
X.4	Fare	-0.00478445263476093	0.00393700788797861
X.5	Sexmale:Pclass	1.60045179226125	0.445606381257609

e) Odds ratio of survival

The odds ratio of a 1st class vs 3rd class, male, 25 year old is 2.471459. The 95% confidence interval for this odds ratio is (0.9777183, 6.2473095). Further, The odds ratio of a 2nd class vs 3rd class, male, 25 year old is 1.572087. The 95% confidence interval for this odds ratio is (0.9887964, 2.4994618). We can conclude that 1st class has about 2.5 times more chance of survival than 3rd class. Further, 2nd class has about 1.6 times more chance of survival than 3rd class. The confidence intervals both have a lower bound close to 1, this means that we can be 95% confidnent that 3rd class passengers will not have agreater chance of survival than 1st or 2nd.

f) GLM model creation using deviance

A general linear model is built with a logit link, it attempts to predict survival using the variables "Pclass" , "Sex", "Age", "Parch", "Fare" and "Embarked". The first variable added to the model is "Sex" which is seen to be significant to our model. Variables will be added to this initial model in no specific order, the significance of the variable will be tested using Deviance statistics which are tested to be significant under the Chi squared distribution using the anova() function. "Pclass" is the next variable added, we see a decrease of 19.432 in deviance which is significant and the variable is included. "Age" is then added and the deviance decreases by 9.7864 which is significant and the variable is included. The next three variables added seperately to the model were all found to have insignificant changes in the deviance statistic, thus "Fare", "Embarked" and "parch" were not included in the model.

Mathematical Description:

The underlying distribution is binomial:

$$Y_i \sim Binomial(\pi_i), \pi \in [0, 1],$$

with a 'logit' link

$$log(\frac{\pi_i}{1-\pi_i}) = X_i^T \beta = \eta_i$$

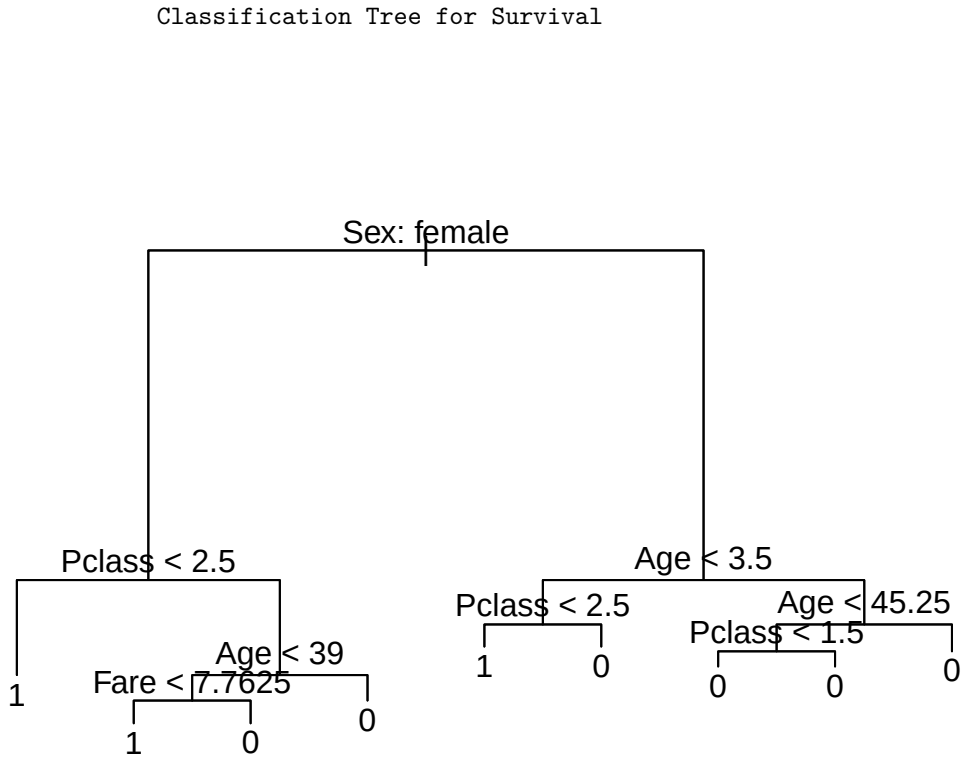
where

$$X_i^T \beta = \eta_i = \beta_1 + \beta_2 x_i^{Sex} + \beta_3 x_i^{Pclass} + \beta_4 x_i^{Age}$$

X2 is 1 if male and 0 if female. X3 is 1 if first class, 2 if second class and 3 if third class. X4 is a covariate for age.

We can say that socio - economic status does play a large role is survival since Pclass is included in our model and has a beta value of approximately -1. This means that the lower class you are in the less chance of survival you have.

g) Tree model of Titanic Data



This tree makes sense because it shows similar results to our glm model fit. The variabel Sex with the initial split means that the distribution for survival of male and female could be different. With regards to the lower levels of the tree, it is fit to match our data set and this could be seen as the phenomenon "over fitting". The variables playing the largest roles in the lower cuts of the trees are Age, Pclass and Fare. All variables but Fare were included in our glm model.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
#glm_fit
glm_fit = function(y, x, beta_start, k = 1){
  w = matrix(0, nrow = length(y), ncol = length(y))
  beta = beta_start

  for( i in 2:k ){
    eta = x %*% beta
    mu = exp(eta)/(1 + exp(eta))
    diag(w) = mu * (1- mu)

    info = t(x) %*% w %*% x
    inverseJ  = solve(info)
    U = t(x) %*% (y-mu)
```

```
    beta = beta + inverseJ %*% U
    stder = diag(inverseJ)^0.5

  }
  return(list(beta, stder))
}

#tabulating function reu
dat = read.table('Assignment_2_Titanic_SetA.txt', h = TRUE)
attach(dat)
fit1 = glm( Survived ~ Sex + Age + Pclass + Fare + Sex*Pclass, family = binomial(link = 'logit'), data =
x = model.matrix(fit1)
y = matrix(Survived, length(Survived), 1)
ownfit = glm_fit(y, x, c(0,0,0,0,0,0), 35 )
rows = c("0","(intercept)","Sexmale","Age","Pclass", "Fare", "Sexmale:Pclass")
row2col=data.frame(rows)
estimates = matrix(unlist(ownfit[1]))
standard_errors = matrix(unlist(ownfit[2]))
columns = c("Estimates", "Stnd.err.")
final = cbind(estimates, standard_errors)
final2 = rbind(columns,final)
final3 = cbind(row2col, final2)
knitr::kable(final3)
#odds ratios
fit2 = glm(Survived ~ Sex + Age + Pclass + Sex*Pclass, family = binomial(link = 'logit'), data = dat)
summary(fit2)

cos = coef(fit2)
vars = vcov(fit2)
norm95 = 1.96

odds_first = exp(cos[4] + cos[5])
odds_second = exp(2*cos[4] + 2*cos[5])
odds_third =  exp(3*cos[4] + 3*cos[5])

odds_ratio1_3 = odds_first/odds_third
odds_ratio1_3
odds_ratio2_3 = odds_second/odds_third
odds_ratio2_3

var1 = 4*vars[4,4] + 4*vars[5,5] + 8*vars[4,5]
var2 = vars[4,4] + vars[5,5] + 2*vars[4,5]

CI1_3_upper = odds_ratio1_3 * exp(norm95*sqrt(var1))
CI1_3_lower = odds_ratio1_3 * exp(-norm95*sqrt(var1))
CIone = c(CI1_3_lower, CI1_3_upper)
CIone

CI2_3_upper = odds_ratio2_3 * exp(norm95*sqrt(var2))
CI2_3_lower = odds_ratio2_3 * exp(-norm95*sqrt(var2))
CItwo = c(CI2_3_lower, CI2_3_upper)
CItwo
```

```r
# model fitting using deviance
names(dat)

fitall = glm(Survived ~ ., family = binomial(link = 'logit'), data = dat)

fit3 = glm(Survived ~ Sex, family = binomial(link = 'logit'), data = dat)

fit3_1 = glm(Survived ~ Sex + Pclass , family = binomial(link = 'logit'), data = dat)
anova(fit3,fit3_1, test = "Chisq")

fit3_2 = glm(Survived ~ Sex + Pclass + Age, family = binomial(link = 'logit'), data = dat)
anova(fit3_1, fit3_2, test = "Chisq")

fit3_3 = glm(Survived ~ Sex + Pclass + Age + Fare, family = binomial(link = 'logit'), data = dat)
anova(fit3_2, fit3_3, test = "Chisq")

fit3_4 = glm(Survived ~ Sex + Pclass + Age + Embarked, family = binomial(link = 'logit'), data = dat)
anova(fit3_2, fit3_4, test = "Chisq")

fit3_5 = glm(Survived ~ Sex + Pclass + Age + Parch, family = binomial(link = 'logit'), data = dat)
anova(fit3_2, fit3_5, test = "Chisq")


#classification tree
library(tree)
tree1 <- tree(factor(Survived)~., data=dat)

plot(tree1, main = "hi")
text(tree1, pretty = 0)
```