SMA 2018, Ex #9 Solution

Guy Ashiri-Prossner

Q1

For logistic regression we get the following results:

$$g(u) = \frac{e^u}{1+e^u} \rightarrow g'(u) = \frac{e^u}{(1+e^u)^2}$$

$$H(u) = \frac{g'(u)}{g(u)} = \frac{\frac{e^u}{(1+e^u)^2}}{\frac{e^u}{1+e^u}} = \frac{1}{1+e^u} \rightarrow H'(u) = -\frac{e^u}{(1+e^u)^2}$$

$$Q(u) = \frac{g'(u)}{1-g(u)} = \frac{\frac{e^u}{(1+e^u)^2}}{\frac{1}{1+e^u}} = \frac{e^u}{1+e^u} = g(u) \rightarrow Q'(u) = g'(u) = \frac{e^u}{(1+e^u)^2}$$

$$\Omega(u) = H(u) + Q(u) = \frac{1}{1+e^u} + \frac{e^u}{1+e^u} = 1 \rightarrow \Omega'(u) = 0$$

$$R_i(\beta) = \Omega(\beta^T X_i)(y_i - g(\beta^T X_i)) = y_i - \frac{e^{\beta^T X_i}}{1+e^{\beta^T X_i}}$$

$$W_i(\beta) = -\Omega'(\beta^T X_i)(y_i - g(\beta^T X_i)) + \Omega(\beta^T X_i) g'(\beta^T X_i) = -0 \cdot (y_i - g(\beta^T X_i)) + 1 \cdot \ g'(\beta^T X_i) = \frac{e^{\beta^T X_i}}{(1+e^{\beta^T X_i})^2}$$

a

```
Q <- function(u){
  return((exp(u))/(1 + exp(u)))
}

Q.tag <- function(u){
  return((exp(u)) / ((1 + exp(u))^2))
}

R <- function(X, y, beta){
  return((y - Q(X %*% beta)))
}

W <- function(X, beta){
  return(diag(as.numeric(Q.tag(X %*% beta))))
}

Z <- function(X, y, beta){
  return(X %*% beta + solve(W(X, beta)) %*% R(X, y, beta))
}

NR <- function(X, y, beta, threshold){
  e <- 2*threshold
  while(abs(sqrt(e)) > threshold){
    beta.old <- beta
    beta <- (solve(t(X) %*% W(X, beta) %*% X) %*% t(X) %*% W(X, beta) %*% Z(X, y, beta))
    e <- t(beta - beta.old) %*% (beta - beta.old)
  }
  return(beta)
}
```

b

```
D <- read.csv('mrfit.csv')
X <- as.matrix(cbind(1, D[,-6]))
y <- D$died10
threshold <- 1e-10
beta <- rep(0, ncol(X))
my.beta <- NR(X, y, beta, threshold)
print(as.numeric(my.beta))
```

```
## [1] -5.82077088  0.05463912  0.65391230  0.00130552  0.01218442  0.40978817
```

c

```
glm_object <- glm(data = D, formula = died10 ~ ., family = binomial(link = "logit"))
print(glm_object$coefficients)
```

```
## (Intercept)         age        diab        chol         map       smoke
## -5.82077088  0.05463912  0.65391230  0.00130552  0.01218442  0.40978817
```

```
all.equal(as.numeric(my.beta), as.numeric(glm_object$coefficients))
```

```
## [1] TRUE
```

Q2

a

In logistic regression, the odds ratio for a change of size Δ in the k^{th} component is given by $e^{\beta_k \Delta}$ so

$$\hat{\psi} = \hat{\beta}_k \Delta = 0.0013 \cdot 30 = 0.0392 \rightarrow e^{\hat{\psi}} = 1.0399$$

The variances matrix is given by $(X^T W(\beta) X)^{-1}$, we'll use the relevant element for *chol*: $(X^T W(\beta) X)^{-1}_{chol,chol} = 7.4073 \times 10^{-7}$:

$$\sqrt{\Delta^T \, V \Delta} = \sqrt{\Delta \, V \Delta} = \sqrt{\Delta^2 \, V} = \sqrt{30^2 \cdot 7.4073334 \times 10^{-7}} = 30 \cdot \sqrt{7.4073334 \times 10^{-7}} = 0.0258$$

$$\alpha = 0.05 \rightarrow z_{1-\frac{\alpha}{2}} = 1.96 \rightarrow z_{1-\frac{\alpha}{2}} \sqrt{\Delta^T V \Delta} = 0.0506$$

$$e^{\beta_k \Delta} \in [e^{\hat{\psi} - z_{1-\frac{\alpha}{2}} \sqrt{\Delta^T V \Delta}}, e^{\hat{\psi} + z_{1-\frac{\alpha}{2}} \sqrt{\Delta^T V \Delta}}] \rightarrow e^{\beta_k \Delta} \in [e^{-0.0114}, e^{0.0898}] \rightarrow e^{\beta_k \Delta} \in [0.9886, 1.0939]$$

b

For $x = [54, 0, 214, 85, 1]$:

$$\hat{\theta} = x^T \hat{\beta} = -1.1454 \rightarrow \hat{p}(x) = g(\hat{\theta}) = 0.2413, x^T \hat{V} x = 0.0056, \alpha = 0.05 \rightarrow z_{1-\frac{\alpha}{2}} = 1.96$$

$$z_{1-\frac{\alpha}{2}} \sqrt{x^T \hat{V} x} = 0.1472 \rightarrow p(x) \in [g(\hat{\theta} - z_{1-\frac{\alpha}{2}} \sqrt{x^T \hat{V} x}), g(\hat{\theta} + z_{1-\frac{\alpha}{2}} \sqrt{x^T \hat{V} x})]$$

$$p(x) \in [g(-1.2926), g(-0.9982)] \rightarrow p(x) \in [0.2154, 0.2693]$$

Q3

a

```
D <- read.csv("spam.train.csv")
l <- glm(data = D, spam ~ ., family = binomial(link = "logit"))
s <- summary(l)
cols <- c("Intercept",colnames(D))
alpha <- 0.001
signif <- which(s$coefficients[,4] < alpha)
```

The significant predictors for $\alpha = 0.001$ are:

Intercept, word_freq_our, word_freq_remove, word_freq_internet, word_freq_free, word_freq_business, word_freq_your, word_freq_000, word_freq_hp, word_freq_george, word_freq_re, word_freq_edu, char_freq_..3, char_freq_..4, capital_run_length_longest

b

```
test_set <- read.csv("spam.test.csv")
theta.hat <- predict.glm(object = l, newdata = test_set)
pi.hat <- Q(theta.hat)
y.hat <- 0 + (pi.hat > 0.5)
```

c

```
pc <- mean((0 + (y.hat == test_set$spam)))
```

The model has 91.8478% accuracy rate.

d

```
sig_cols <- signif[-1]-1
D2 <- cbind(D[,sig_cols], spam = D$spam)
reduced <- glm(data = D2, spam ~ ., family = binomial(link = "logit"))
theta.hat2 <- predict.glm(object = reduced, newdata = test_set)
pi.hat2 <- Q(theta.hat2)
y.hat2 <- 0 + (pi.hat2 > 0.5)
```

e

```
pc2 <- mean((0 + (y.hat2 == test_set$spam)))
```

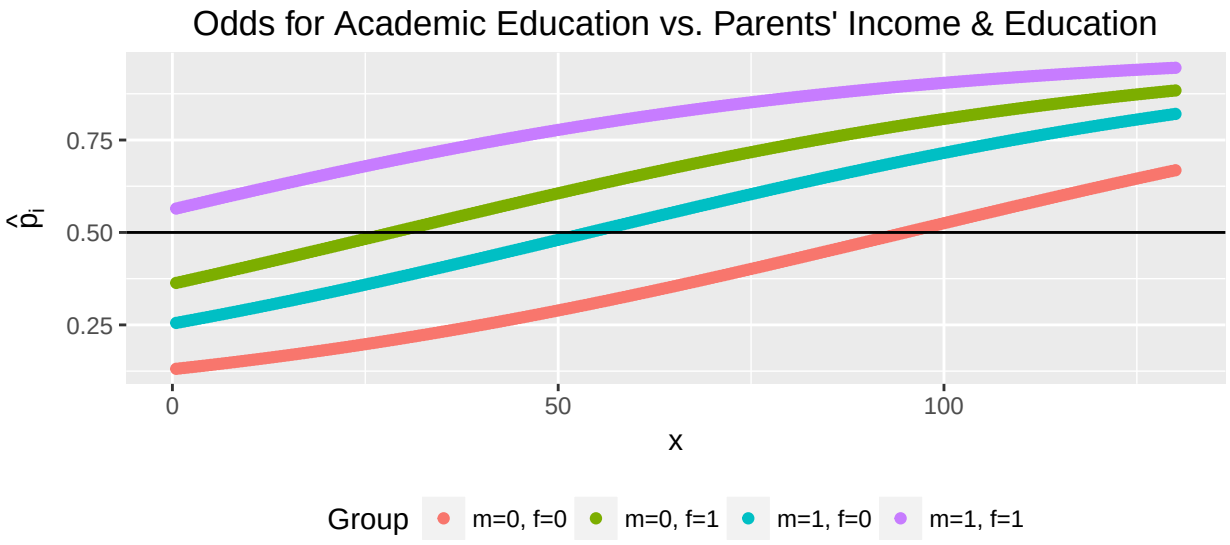
The reduced model has 91.9565% accuracy rate.

Q4

a

```
x <- seq(0.5, 130, by = 0.5)
grp00 <- data.frame(x = x, m = 0, f = 0, Group = "m=0, f=0")
grp01 <- data.frame(x = x, m = 0, f = 1, Group = "m=0, f=1")
grp10 <- data.frame(x = x, m = 1, f = 0, Group = "m=1, f=0")
grp11 <- data.frame(x = x, m = 1, f = 1, Group = "m=1, f=1")
df <- rbind(grp00, grp01, grp10, grp11)
df$Group <- factor(df$Group)
df$lp <- -1.9 + 0.02 * df$x + 0.82 * df$m + 1.33 * df$f
df$p.hat <- Q(df$lp)

ggplot(data = df, aes(x = x, y = p.hat, color = Group)) + geom_point() +
  geom_hline(yintercept = 0.5) + labs(ylab(expression(hat(p)[i]))) +
  ggtitle("Odds for Academic Education vs. Parents' Income & Education") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "bottom")
```



It can be seen that the paernts' education does have a positive effect, with the father's education having stronger effect than the mother's.

b

```
LD50_idx <- aggregate(df$p.hat, by = list(df$Group), FUN = function(x){min(which(x >= 0.5))})$x
LD50 <- x[LD50_idx]
```

m=0, f=0: $P(y = 1) = P(y = 0)$ for x=95.

m=0, f=1: $P(y = 1) = P(y = 0)$ for x=28.5.

m=1, f=0: $P(y = 1) = P(y = 0)$ for x=54.

m=1, f=1: $P(y = 1) = P(y = 0)$ for x=0.5.