Workforce Management Projects

Parikshit Patil

2024-08-19

Impact of Workplace Smoking Rules

a)

Likelihood:

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

$$= \frac{e^{-\lambda}\lambda^{3}}{3!} * \frac{e^{-\lambda}\lambda^{0}}{0!} * \frac{e^{-\lambda}\lambda^{0}}{0!} * \frac{e^{-\lambda}\lambda^{1}}{1!} * \frac{e^{-\lambda}\lambda^{2}}{2!} * \frac{e^{-\lambda}\lambda^{1}}{1!} = \frac{e^{-6\lambda}\lambda^{7}}{12}$$

Log Likelihood:

$$\log L = \sum_{i=1}^{n} -\lambda + y_i \log(\lambda) - log(y_i!)$$

$$\log L = [-\lambda + 3\log(\lambda) - log(3!)] + [-\lambda + 0\log(\lambda) - log(0!)] + [-\lambda + 0\log(\lambda) - log(0!)] +$$

$$[-\lambda + 1\log(\lambda) - log(1!)] + [-\lambda + 2\log(\lambda) - log(2!)] + [-\lambda + 1\log(\lambda) - log(1!)]$$

$$= -6\lambda + 7\log\lambda - \log(12)$$

b)

Since there is no difference between home and work, the sample mean of y is a reasonable estimate for lambda.

$$\lambda = \frac{(3+0+0+1+2+1)}{6} = 1.166$$

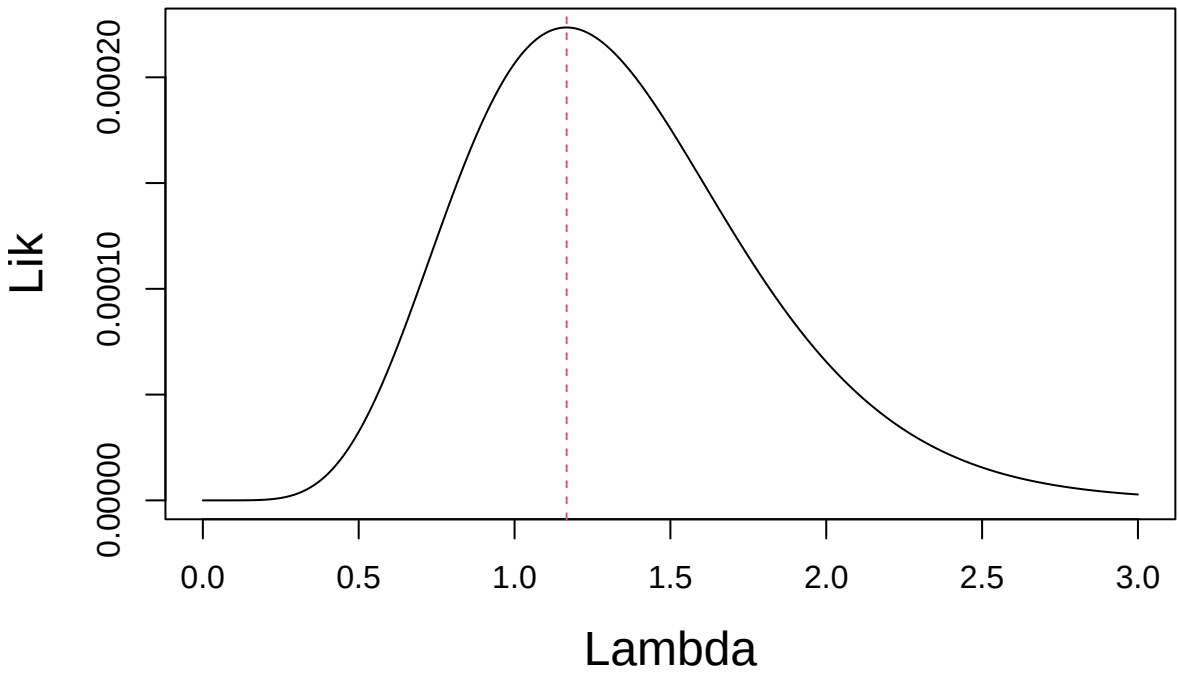
c)

```r
#Calculate and plot the likelihood function

Lik = function(lamb){
Lik = (exp(-6*lamb)*lamb^7)/12
return(Lik)
}

Lambda=seq(0,3,0.001)
df = Lik(Lambda)

plot(Lambda,df,type="l",cex.lab=1.5,ylab="Lik")
abline(v=Lambda[which(df==max(df))],lty=2,col=2)
```




```r
#Optimize
optimize(Lik, interval=c(0,3), maximum=TRUE)
```

```
## $maximum
## [1] 1.166678
##
## $objective
## [1] 0.0002235553
```

d)

$$\log L = \sum_{i=1}^{n} -\lambda + y_i \log(\lambda) - log(y_i!)$$

$$\log L = [-\lambda_H + 3\log(\lambda_H) - log(3!)] + [-\lambda_W + 0\log(\lambda_W) - log(0!)] +$$
$$[-\lambda_W + 0\log(\lambda_W) - log(0!)] + [-\lambda_W + 1\log(\lambda_W) - log(1!)] +$$
$$[-\lambda_H + 2\log(\lambda_H) - log(2!)] + [-\lambda_H + 1\log(\lambda_H) - log(1!)]$$
$$= -3\lambda_H - 3\lambda_W + 6\log(\lambda_H) + 1\log(\lambda_W) - \log(12)$$

e)

For lambda_H, a reasonable estimate would be the sample mean of y which only corresponds to working from home (i.e. mean of 3,2,1).

$$\lambda_H = \frac{(3+2+1)}{3} = 2$$

For lambda_W, a reasonable estimate would be the sample mean of y which only correspond to working from the office (i.e. mean of 0,0,1).

$$\lambda_W = \frac{(0+0+1)}{3} = 0.333$$

f)

```
#Calculate and optimize the likelihoodSS

Lik_H = function(lamb_H){
Lik_H = -3*lamb_H + 6*log(lamb_H) - log(12)
return(Lik_H)
}

Lik_W = function(lamb_W){
Lik_W = -3*lamb_W + log(lamb_W) - log(12)
return(Lik_W)
}

optimize(Lik_H, interval=c(0,3), maximum=TRUE)
```

```
## $maximum
## [1] 2.000008
##
## $objective
## [1] -4.326024
```

```
optimize(Lik_W, interval=c(0,3), maximum=TRUE)
```

```
## $maximum
## [1] 0.3333148
##
## $objective
## [1] -4.583519
```

g)

$$Given : \log(\lambda) = \beta_0 + \beta_1 x => \lambda = e^{\beta_0 + \beta_1 x}$$

$$\log L = \sum_{i=1}^{n} -\lambda_i + y_i \log(\lambda_i) - log(y_i!) = \sum_{i=1}^{n} -e^{\beta_0+\beta_1 x_i} + y_i(\beta_0 + \beta_1 x_i) - log(y_i!)$$

$$= [-e^{\beta_0+\beta_1 0}+3(\beta_0+\beta_1 0)-log(3!)]+[-e^{\beta_0+\beta_1 1}+0(\beta_0+\beta_1 1)-log(0!)]+[-e^{\beta_0+\beta_1 1}+0(\beta_0+\beta_1 1)-log(0!)]+[-e^{\beta_0+\beta_1 1}+1(\beta_0+$$

$$= -3e^{\beta_0} + -3e^{\beta_0+\beta_1} + 7\beta_0 + \beta_1 - \log(12)$$

h)

```
#Lik_b0 = function(b0){
#Lik_b0 = -3*exp(b0) + -3*exp(b0) + 7*b0 - log(12)
#return(Lik_b0)
#}

#Lik_b1 = function(b1){
#Lik_b1 = -3*exp(b1) + b1 - log(12)
#return(Lik_b1)
#}

#optimize(Lik_b0, interval=c(0,100), maximum=TRUE)
#optimize(Lik_b1, interval=c(0,100), maximum=TRUE)
```

```
# Model 3 Likelihood
lik_b0_b1 = function(params) {
  b_0 = params[1]
  b_1 = params[2]

  Lik <- -3 * exp(b_0) - 3 * exp(b_0 + b_1) + 7 * b_0 + b_1 - log(12)

  return(-Lik)
}

# Set initial values for optimization
initial_values = c(0, 0)

# Optimize
result = optim(par = initial_values, fn = lik_b0_b1, method = "L-BFGS-B")

# MLE estimates b_0 b_1
b_0 = result$par[1]
b_1 = result$par[2]

cat("Maximum Likelihood Estimators:\n")
```

```
## Maximum Likelihood Estimators:
```

```
cat("b_0 =", b_0, "\n")
```

```
## b_0 = 0.693147
```

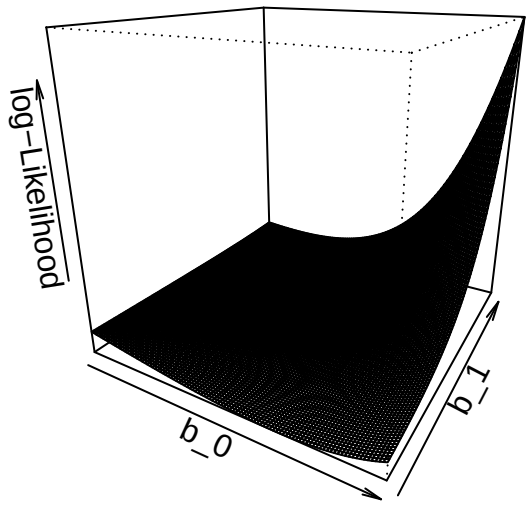
```r
cat("b_1 =", b_1, "\n")
```

```
## b_1 = -1.79176
```

```r
# 3D plot
b_0_val = seq(-2, 2, length.out = 100)
b_1_val = seq(-2, 2, length.out = 100)
logLik_values = matrix(0, nrow = length(b_0_val), ncol = length(b_1_val))

#log-likelihood values
for (i in 1:length(b_0_val)) {
  for (j in 1:length(b_1_val)) {
    logLik_values[i, j] = lik_b0_b1(c(b_0_val[i], b_1_val[j]))
  }
}

# 3D Plot
persp(x = b_0_val, y = b_1_val, z = logLik_values, main = "Log-Likelihood Function (Model 3)",
      xlab = "b_0", ylab = "b_1", zlab = "log-Likelihood", theta = 30, phi = 20)
```



i)

```r
#create dataset
x = c(0, 1, 1, 1, 0, 0)
y = c(3, 0, 0, 1, 2, 1)

SmokeData = data.frame(x = x, y = y)

#Fit model 1
mod1_poisson <- glm(SmokeData$y ~ 1, family=poisson, SmokeData)

#Fit model 3
mod3_poisson <- glm(SmokeData$y ~ ., family=poisson, SmokeData)

summary(mod1_poisson)
```

```
##
## Call:
## glm(formula = SmokeData$y ~ 1, family = poisson, data = SmokeData)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1542     0.3780   0.408    0.683
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 7.2062  on 5  degrees of freedom
## Residual deviance: 7.2062  on 5  degrees of freedom
## AIC: 18.812
##
## Number of Fisher Scoring iterations: 5
```

```r
summary(mod3_poisson)
```

```
##
## Call:
## glm(formula = SmokeData$y ~ ., family = poisson, data = SmokeData)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.6931     0.4082   1.698   0.0895 .
## x            -1.7918     1.0801  -1.659   0.0971 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 7.2062  on 5  degrees of freedom
## Residual deviance: 3.2437  on 4  degrees of freedom
## AIC: 16.849
##
## Number of Fisher Scoring iterations: 5
```

Model 1 Estimates: Beta0 = 0.1542

Model 3 Estimates: Beta0 = 0.6931, Beta1 = -1.7918

Proof for MLE of Model 3 agrees with MLE of Model 2:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i$$

Consider Subject 3 from the data, where x = 1, lambda_i is from model 2, and betas are from model 3

$$\log(\lambda_W) = \beta_0 + \beta_1 x_3$$

$$\log(0.3333148) = 0.6931 + (-1.7918)1$$

$$-1.098 = -1.098$$

$$LHS = RHS Hence\ Proved$$

Impact of Gender on Admissions

```
suppressMessages(library(faraway))
str(UCBAdmissions)
```

```
##  'table' num [1:2, 1:2, 1:6] 512 313 89 19 353 207 17 8 120 205 ...
##  - attr(*, "dimnames")=List of 3
##   ..$ Admit : chr [1:2] "Admitted" "Rejected"
##   ..$ Gender: chr [1:2] "Male" "Female"
##   ..$ Dept  : chr [1:6] "A" "B" "C" "D" ...
```

a)

```
library(broom)
#ct = xtabs( ~ Gender + Admit, data = UCBAdmissions)
ucb_tidy <- tidy(UCBAdmissions)
```

```
## Warning: 'tidy.table' is deprecated.
## Use 'tibble::as_tibble()' instead.
## See help("Deprecated")
```

```
# Contingency Table
ct <- xtabs(n~Admit+Gender, ucb_tidy)

data = UCBAdmissions
#get overall percentages
prop.table(apply(data, c(1, 2), sum),1)
```

```
##           Gender
## Admit           Male    Female
##   Admitted 0.6826211 0.3173789
##   Rejected 0.5387947 0.4612053
```

```
#show each department
data
```

```
## , , Dept = A
##
##           Gender
## Admit      Male Female
##   Admitted  512     89
##   Rejected  313     19
##
## , , Dept = B
##
##           Gender
## Admit      Male Female
##   Admitted  353     17
##   Rejected  207      8
##
## , , Dept = C
##
##           Gender
## Admit      Male Female
##   Admitted  120    202
##   Rejected  205    391
##
## , , Dept = D
##
##           Gender
## Admit      Male Female
##   Admitted  138    131
##   Rejected  279    244
##
## , , Dept = E
##
##           Gender
## Admit      Male Female
##   Admitted   53     94
##   Rejected  138    299
##
## , , Dept = F
##
##           Gender
## Admit      Male Female
##   Admitted   22     24
##   Rejected  351    317
```

We can see that the proportion of male applicants who were admitted (68%) and rejected (53%) are both larger than female applicants. This cannot be possible, since only the proportion of being either admitted or rejected can be larger than the other gender, not both. Additionally, we can see that departments C and E have more admissions for females than males. Therefore, this is an example of Simpson's paradox.

b)

```r
# Convert the UCBAdmissions dataset to a data frame
UCB_df <- as.data.frame(UCBAdmissions)

# Add a column for department
UCB_df$Dept <- rep(LETTERS[1:6], each = 4)

# Reshape the data to long format
UCB_reformatted <- reshape2::melt(UCB_df, id.vars = c("Admit", "Gender", "Dept"))

# Convert Admit to a binary variable (0 for Rejected, 1 for Admitted)
#UCB_reformatted$Admit <- ifelse(UCB_reformatted$Admit == "Admitted", 1, 0)

#UCB_reformatted

#Mantel Haenszel Test
mantelhaen.test(data,exact=TRUE)
```

```
##
##  Exact conditional test of independence in 2 x 2 x k tables
##
## data:  data
## S = 1198, p-value = 0.2278
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.769714 1.063417
## sample estimates:
## common odds ratio
##         0.9050762
```

From the Mantel Haenszel Test since the p-value is large, we reject the null hypothesis of independence between Gender and Admits.

goodness of fit using deviance to check mutual independence and joint independence:

```r
#Mutual Independance
poisson_model_admi = glm(n ~ Admit + Gender + Dept, ucb_tidy, family=poisson)
c(deviance(poisson_model_admi),df.residual(poisson_model_admi))
```

```
## [1] 2097.671   16.000
```

```r
pchisq(poisson_model_admi$deviance,df.residual(poisson_model_admi),lower=FALSE)
```

```
## [1] 0
```

```r
#Joint Independence
poisson_model_admi_2 = glm(n ~ Gender*Admit + Dept, ucb_tidy, family=poisson)
c(deviance(poisson_model_admi_2),df.residual(poisson_model_admi_2))
```

```
## [1] 2004.222   15.000
```

```r
pchisq(deviance(poisson_model_admi_2),df.residual(poisson_model_admi_2),lower=FALSE)
```

```
## [1] 0
```

The model with main effects-only (under independence), and the model with Dept, Admit, Gender, and the interaction do not fit the data well.

Feature Selection Process:

```r
# Backward Selection process
model1 <- glm(n ~ Gender*Dept*Admit, ucb_tidy, family=poisson)
coefficients(model1)
```

```
##                      (Intercept)                       GenderMale
##                       4.48863637                       1.74968826
##                            DeptB                            DeptC
##                      -1.65542303                       0.81963133
##                            DeptD                            DeptE
##                       0.38656095                       0.05465841
##                            DeptF                   AdmitRejected
##                      -1.31058254                      -1.54419739
##                GenderMale:DeptB                GenderMale:DeptC
##                       1.28356646                      -2.27046421
##                GenderMale:DeptD                GenderMale:DeptE
##                      -1.69763189                      -2.32269112
##                GenderMale:DeptF       GenderMale:AdmitRejected
##                      -1.83669963                       1.05207596
##             DeptB:AdmitRejected             DeptC:AdmitRejected
##                       0.79042559                       2.20463725
##             DeptD:AdmitRejected             DeptE:AdmitRejected
##                       2.16616829                       2.70134618
##             DeptF:AdmitRejected GenderMale:DeptB:AdmitRejected
##                       4.12504533                      -0.83205342
## GenderMale:DeptC:AdmitRejected GenderMale:DeptD:AdmitRejected
##                      -1.17699758                      -0.97008876
## GenderMale:DeptE:AdmitRejected GenderMale:DeptF:AdmitRejected
##                      -1.25226298                      -0.86318013
```

```r
# Drop three-way interaction
drop1(model1,test="Chi")
```

```
## Single term deletions
##
## Model:
## n ~ Gender * Dept * Admit
##                    Df Deviance    AIC    LRT Pr(>Chi)
## <none>                  0.000 207.06
## Gender:Dept:Admit  5   20.204 217.26 20.204 0.001144 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Drop two-way interaction terms
model2 <- glm(n ~ (Gender+Dept+Admit)^2, ucb_tidy, family=poisson)
drop1(model2,test="Chi")
```

```
## Single term deletions
##
## Model:
## n ~ (Gender + Dept + Admit)^2
##              Df Deviance     AIC     LRT Pr(>Chi)
## <none>             20.20  217.26
## Gender:Dept   5  1148.90 1335.96 1128.70   <2e-16 ***
## Gender:Admit  1    21.74  216.80    1.53   0.2159
## Dept:Admit    5   783.61  970.67  763.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value for the interaction term Gender:Admit reflects the hypothesis of independence of the Mantel-Haenszel test (similar p-values) As a result, we reject the hypothesis of independence between Gender and Admissions.

c)

For some 3-way tables, we may regard one variable as response the other two as predictors For the UCBAdmissions data set, we could model Admit (response) as a Binomial GLM:

```r
# Fit the logistic regression model
logistic_model = glm(Admit ~ Gender + Dept, data = UCB_reformatted, family = binomial)

summary(logistic_model)
```

```
##
## Call:
## glm(formula = Admit ~ Gender + Dept, family = binomial, data = UCB_reformatted)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.537e-15  1.080e+00       0        1
## GenderFemale -5.439e-16  8.165e-01       0        1
## DeptB        -7.238e-16  1.414e+00       0        1
## DeptC        -1.806e-15  1.414e+00       0        1
## DeptD        -1.446e-15  1.414e+00       0        1
## DeptE        -1.386e-15  1.414e+00       0        1
## DeptF        -1.413e-15  1.414e+00       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33.271  on 23  degrees of freedom
## Residual deviance: 33.271  on 17  degrees of freedom
## AIC: 47.271
##
## Number of Fisher Scoring iterations: 2
```

This is a model containing only the main effects, and is the same as the Poisson model of uniform association seen previously.

High School Program Selection

a)

```
library(faraway)
library(nnet)
data("hsb")
hsb <- hsb[,-1] ## removing first column corresponding to student ID
str(hsb)
```

```
## 'data.frame':    200 obs. of  10 variables:
##  $ gender : Factor w/ 2 levels "female","male": 2 1 2 2 2 2 2 2 2 2 ...
##  $ race   : Factor w/ 4 levels "african-amer",..: 4 4 4 4 4 4 1 3 4 1 ...
##  $ ses    : Factor w/ 3 levels "high","low","middle": 2 3 1 1 3 3 3 3 3 3 ...
##  $ schtyp : Factor w/ 2 levels "private","public": 2 2 2 2 2 2 2 2 2 2 ...
##  $ prog   : Factor w/ 3 levels "academic","general",..: 2 3 2 3 1 1 2 1 2 1 ...
##  $ read   : int  57 68 44 63 47 44 50 34 63 57 ...
##  $ write  : int  52 59 33 44 52 52 59 46 57 55 ...
##  $ math   : int  41 53 54 47 57 51 42 45 54 52 ...
##  $ science: int  47 63 58 53 53 63 53 39 58 50 ...
##  $ socst  : int  57 61 31 56 61 61 61 36 51 51 ...
```

```
hsb_mnom = multinom(prog ~ ., data = hsb)
```

```
## # weights:  42 (26 variable)
## initial  value 219.722458
## iter  10 value 171.814970
## iter  20 value 153.793692
## iter  30 value 152.935260
## final  value 152.935256
## converged
```

```
summary(hsb_mnom)
```

```
## Call:
## multinom(formula = prog ~ ., data = hsb)
##
## Coefficients:
##          (Intercept)  gendermale raceasian racehispanic racewhite     seslow
## general     3.631901 -0.09264717  1.352739   -0.6322019 0.2965156 1.09864111
## vocation    7.481381 -0.32104341 -0.700070   -0.1993556 0.3358881 0.04747323
##          sesmiddle schtyppublic        read       write       math    science
## general  0.7029621    0.5845405 -0.04418353 -0.03627381 -0.1092888 0.10193746
## vocation 1.1815808    2.0553336 -0.03481202 -0.03166001 -0.1139877 0.05229938
##                socst
## general  -0.01976995
## vocation -0.08040129
##
## Std. Errors:
##          (Intercept) gendermale raceasian racehispanic racewhite    seslow
## general     1.823452  0.4548778  1.058754    0.8935504 0.7354829 0.6066763
## vocation    2.104698  0.5021132  1.470176    0.8393676 0.7480573 0.7045772
```

```
##          sesmiddle schtyppublic       read      write       math    science
## general  0.5045938    0.5642925 0.03103707 0.03381324 0.03522441 0.03274038
## vocation 0.5700833    0.8348229 0.03422409 0.03585729 0.03885131 0.03424763
##               socst
## general  0.02712589
## vocation 0.02938212
##
## Residual Deviance: 305.8705
## AIC: 357.8705
```

b)

```
exp(coef(hsb_mnom))
```

```
##          (Intercept) gendermale raceasian racehispanic racewhite   seslow
## general     37.78456  0.9115151 3.8680061    0.5314204  1.345164 3.000086
## vocation  1774.69006  0.7253918 0.4965505    0.8192585  1.399182 1.048618
##          sesmiddle schtyppublic      read     write      math  science
## general   2.019726     1.794166 0.9567783 0.9643762 0.8964715 1.107314
## vocation  3.259523     7.809443 0.9657869 0.9688359 0.8922690 1.053691
##              socst
## general  0.9804242
## vocation 0.9227460
```

NOTE - All the below interpretations are relative to the Academic program category.

Read Score: For every one-unit increase in read score, the odds of being in the general category decrease by approximately 4.4%. For every one-unit increase in read score, the odds of being in the vocation category decrease by approximately 3.5%.

Write Score: For every one-unit increase in write score, the odds of being in the general category decrease by approximately 3.6%. For every one-unit increase in write score, the odds of being in the vocation category decrease by approximately 3.2%.

Math Score: For every one-unit increase in math score, the odds of being in the general category decrease by approximately 10.4%. For every one-unit increase in math score, the odds of being in the vocation category decrease by approximately 10.8%.

Science Score: For every one-unit increase in science score, the odds of being in the general category increase by approximately 10.7%. For every one-unit increase in science score, the odds of being in the vocation category **increase** by approximately 5.3%.

Socst Score: For every one-unit increase in social studies score, the odds of being in the general category decrease by approximately 2%. For every one-unit increase in social studies score, the odds of being in the vocation category decrease by approximately 7.8%.

c)

Out of the 5 subjects, the results for science is very unexpected. The reason for this might be that students who are better in science might prefer a lower workload in order to focus on other tasks such as sports or projects.

S&P 500 Market Direction

a)

The logistic model is fitted to lag 2, volume and interaction of lag2 and volume. There are 4 coefficients, including the coefficients.

p = probability of predicted Direction.

$$\eta = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{Lag2} + \beta_2 x_{Volume} + \beta_3 x_{Lag2} x_{Volume}$$

glm(direction~Lag2*Volume, family = binomial, data = Weekly)

It can be seen that the Direction increases as the Lag2 increases. We can also see that both the categories of Volume have different slopes since they cross, indicating that there is an interaction term between Lag2 and Volume.Therefore there are 4 predictors - Lag2+Volume+Lag2:Volume.

b)

```
(1-exp(0.065))*100
```

```
## [1] -6.715902
```

```
(1-exp(0.05))*100
```

```
## [1] -5.12711
```

For the black line (high volume), every unit increase in Lag2 results in a 6.7% increase in the odds of direction going up.

For the red line (low volume), every unit increase in Lag2 results in a 5.1% increase in the odds of direction going up.

C)

```
(1-exp(0.138))*100
```

```
## [1] -14.79756
```

```
(1-exp(0.296))*100
```

```
## [1] -34.44702
```

There is a practical interpretation for the intercepts, since Lag2 can be 0 practically.

The intercept of the black line represents an odds of 14.79% of the direction going up when Lag2 is 0 and the volume is high.

The intercept of the red line represents an odds of 34.44% of the direction going up when Lag2 is 0 and the volume is low.

d) When Volume = 0 (low- red line):

$$\eta = \beta_0 + \beta_1 x_{lag2}$$

When Volume = 1 (high- black line):

$$\eta = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{lag2}$$

$$\therefore \beta_0 = 0.296$$

$$\beta_1 = 0.05$$

$$\beta_2 = 0.138 - 0.296 = -0.158$$

$$\beta_3 = 0.065 - 0.05 = 0.015$$