

Lab 5

The Sound of Gunfire, Off in the Distance

Alice Chang

Part 1-2: Estimate

```
#load data  
library(MASS)  
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
```

```
##
```

```
##      select
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
##      filter, lag
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##      intersect, setdiff, setequal, union
```

```
war <- read.csv("http://www.stat.cmu.edu/~cshalizi/uADA/15/hw/06/ch.csv", row.names = 1)
```

```
war_clean <- na.omit(war)
```

```
war_clean <- war_clean %>% mutate(exports2 = exports^2)
```

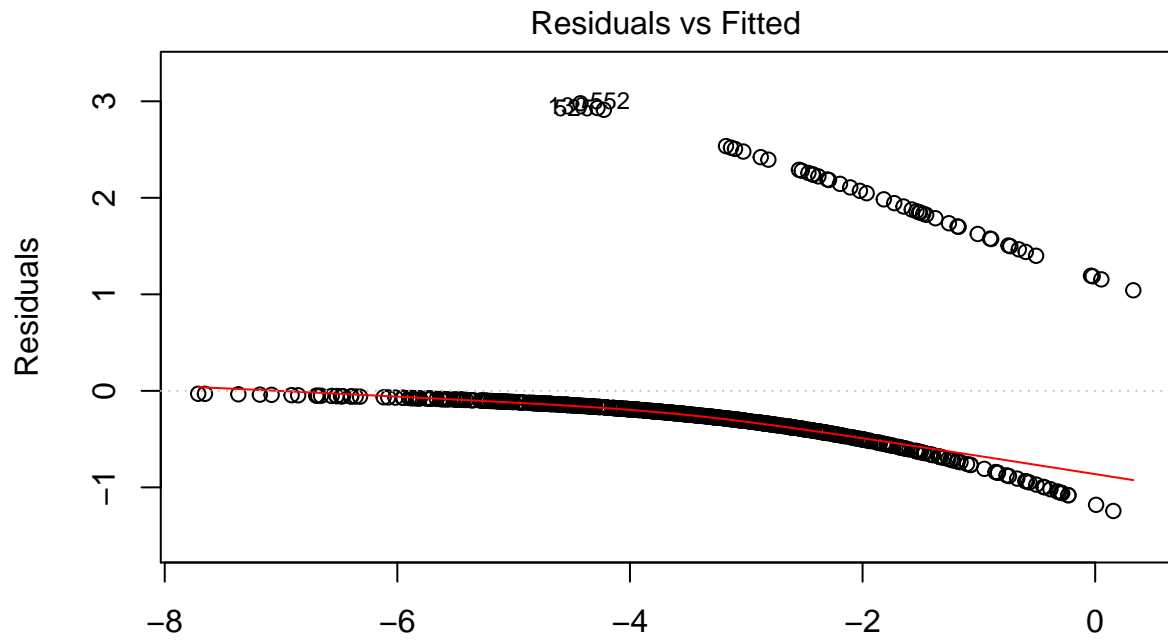
```
row.names(war_clean) <- (1:nrow(war_clean))
```

```
#include quadratic term
```

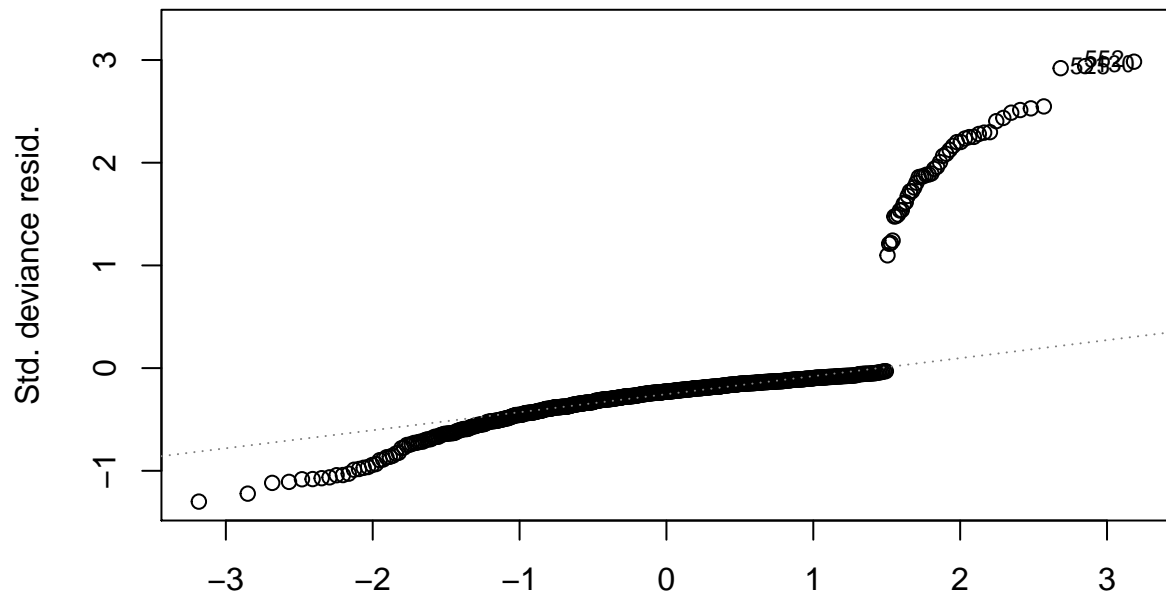
```
#fit logistic regression model
```

```
glm.fits <- glm(start ~ exports2 + schooling + growth + peace + concentration + lnpop + fractionalizati
```

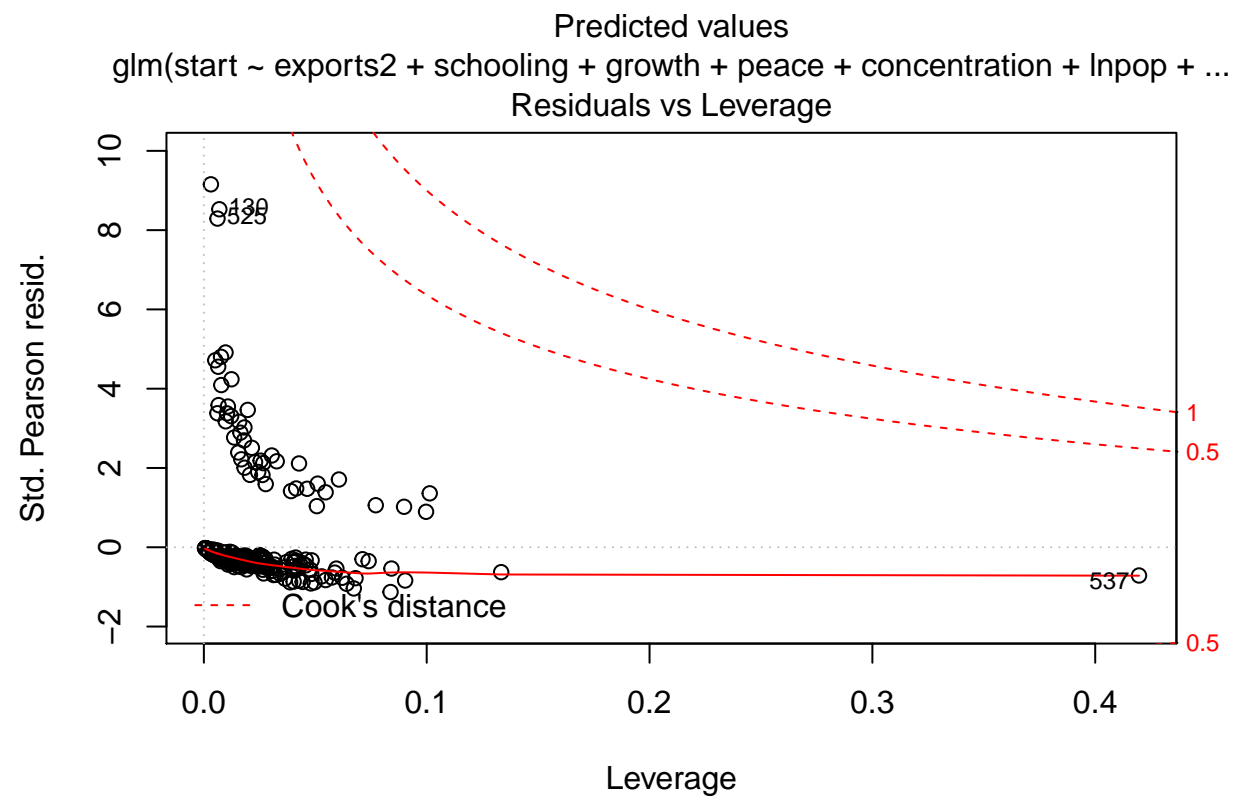
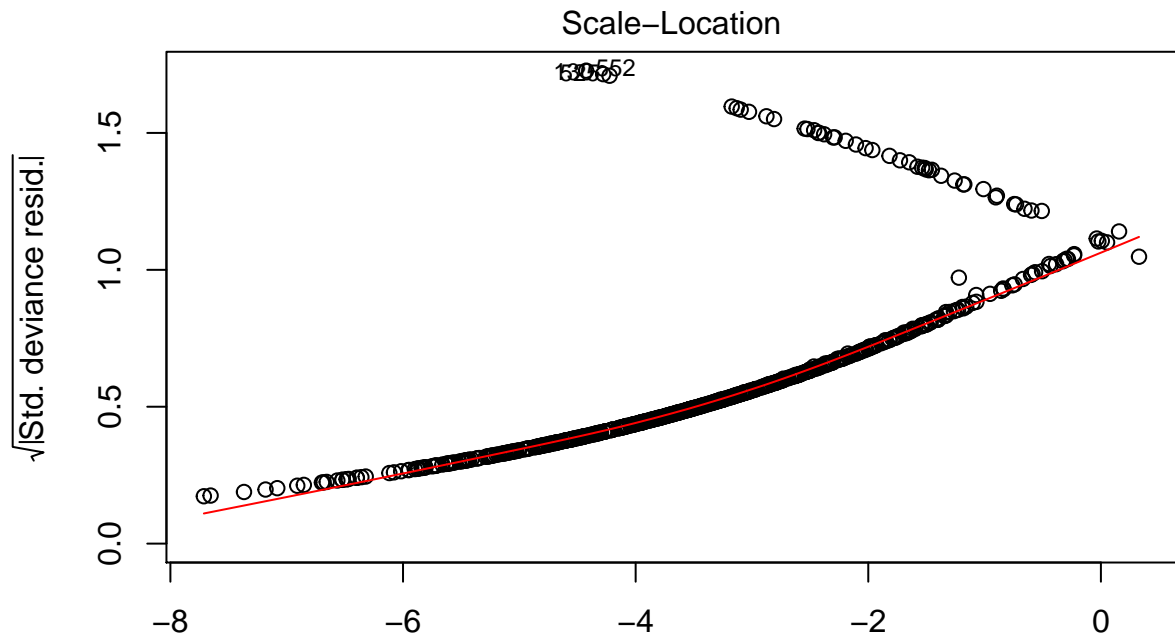
```
plot(glm.fits)
```



Predicted values
`glm(start ~ exports2 + schooling + growth + peace + concentration + lnpop + ...`
 Normal Q-Q



Theoretical Quantiles
`glm(start ~ exports2 + schooling + growth + peace + concentration + lnpop + ...`



glm(start ~ exports2 + schooling + growth + peace + concentration + lnpop + ...)

```
summary(glm.fits)$coef
```

	Estimate	Std. Error	z value	Pr(> z)
## (Intercept)	-7.469538e+00	2.046470e+00	-3.649963	0.0002622786
## exports2	2.799199e+00	1.870619e+00	1.496403	0.1345487717
## schooling	-2.388170e-02	8.684965e-03	-2.749776	0.0059636089

```
## growth          -1.364391e-01 4.335592e-02 -3.146953 0.0016498130
## peace           -4.088964e-03 1.096575e-03 -3.728849 0.0001923564
## concentration   -1.563426e+00 9.220441e-01 -1.695609 0.0899599576
## lnpop            4.776614e-01 1.249363e-01  3.823240 0.0001317096
## fractionalization -9.910336e-05 8.230627e-05 -1.204080 0.2285585239
## dominance        5.375078e-01 3.451495e-01  1.557319 0.1193948020
```

The effect of a country's dependency on commodity exports ("exports2") ($\beta_1 = 2.799199e+00$, SE = 1.870619e+00, P = 0.1345487717), geographic concentration of the population ("concentration") ($\beta_5 = -1.563426e+00$, SE = 9.220441e-01, P = 0.0899599576), social fractionalization ("fractionalization") ($\beta_7 = -9.910336e-05$, SE = 8.230627e-05, P = 0.2285585239), and extent of ethnic dominance ("dominance") ($\beta_8 = 5.375078e-01$, SE = 3.451495e-01, P = 0.1193948020) on the likelihood of a war starting are statistically insignificant.

However, the secondary school enrollment rate for males ("schooling") ($\beta_2 = -2.388170e-02$, SE = 8.684965e-03, P = 0.00596), annual GDP growth rate ("growth") ($\beta_3 = -1.364391e-01$, SE = 4.335592e-02, P = 0.00165), number of months since the country's last war ("peace") ($\beta_4 = -4.088964e-03$, SE = 1.096575e-03, P = 0.00019), and natural logarithm of the country's population ("lnpop") ($\beta_6 = 4.776614e-01$, SE = 1.249363e-01, P = 0.00013) are all statistically significant at the 5% level.

The negative coefficient estimates for "schooling," "growth," and "peace," suggest that as a country's school enrollment rate for males, annual GDP growth rate, and number of months since the last war increases, while holding each other and all other variables fixed, the likelihood of a war starting decreases. The positive coefficient estimate of "lnpop" suggests that, holding all other variables constant, the larger a country's population, the greater the likelihood of a war starting.

Part 2: Interpretation

```
#India, 1975
```

```
which((war_clean$country == "India") & (war_clean$year == 1975))
```

```
## [1] 272
```

```
war_clean[272,]
```

```
##      country year start exports schooling growth peace concentration
## 272   India 1975      0   0.026         36  0.322  112          0.537
##      lnpop fractionalization dominance exports2
## 272 20.23462          2937          0 0.000676
```

```
#Predict for India in 1975
```

```
probs_a1 <- predict(glm.fits, newdata = data.frame(exports2 = 0.026^2, schooling = 36, growth = 0.322, lnpop = 20.23462, fractionalization = 2937, dominance = 0, exports2 = 0.000676))
probs_a1
```

```
##      1
## -0.2946148
```

```
#Predict for country like India with higher schooling
```

```
probs_b1 <- predict(glm.fits, newdata = data.frame(exports2 = 0.026^2, schooling = 66, growth = 0.322, lnpop = 20.23462, fractionalization = 2937, dominance = 0, exports2 = 0.000676))
probs_b1
```

```
##      1
## -1.011066
```

```
#Predict for country like India with higher export to GDP ratio
```

```
probs_c1 <- predict(glm.fits, newdata = data.frame(exports2 = (0.026 + 0.1)^2, schooling = 36, growth = 0.322, lnpop = 20.23462, fractionalization = 2937, dominance = 0, exports2 = 0.000676))
probs_c1
```

```
##      1
```

```
## -0.252067
#Nigeria, 1965
which((war_clean$country == "Nigeria") & (war_clean$year == 1965))

## [1] 464
war_clean[464,]

##      country year start exports schooling growth peace concentration
## 464 Nigeria 1965      1   0.123          7  1.916   232          0.539
##      lnpop fractionalization dominance exports2
## 464 17.65479          6090          0 0.015129
#Predict for Nigeria in 1965
probs_a2 <- predict(glm.fits, newdata = data.frame(exports2 = 0.123 ^2, schooling = 7, growth = 1.916
probs_a2

##          1
## -1.817633
#Predict for country like Nigeria, with higher schooling
probs_b2 <- predict(glm.fits, newdata = data.frame(exports2 = 0.123 ^2, schooling = 37, growth = 1.916
probs_b2

##          1
## -2.534084
#Predict for country like Nigeria with higher export to GDP ratio
probs_c2 <- predict(glm.fits, newdata = data.frame(exports2 = (0.123 + 0.1)^2, schooling = 7, growth = 
probs_c2

##          1
## -1.720781
```

1. The model returns a very small (“negative”) probability estimate for civil war in India in 1975 of -0.2946148, indicating that it predicts the probability that a civil war will not begin.

For a country just like India in 1975, but with a male secondary school enrollment rate 30 points higher, the model returns an extremely small probability estimate for civil war of -1.011066, indicating that it predicts the probability that a civil war will not begin.

For a country just like India in 1975, but with a ratio of commodity exports to GDP that is 0.1 higher, the model also returns a small probability estimate for civil war of -0.252067, indicating that it predicts the probability that a civil war will not begin.

2. The model returns a very small (“negative”) probability estimate for civil war in Nigeria in 1965 of -1.817633, indicating that it predicts the probability that a civil war will not begin under these conditions.

For a country just like Nigeria in 1965, but with a male secondary school enrollment rate 30 points higher, the model returns an extremely small probability estimate for civil war of -2.534084, indicating that it predicts the probability that a civil war will not begin under these conditions.

For a country just like Nigeria in 1965, but with a ratio of commodity exports to GDP that is 0.1 higher, the model also returns a small probability estimate for civil war of -1.720781, indicating that it still predicts the probability that a civil war will not begin under these conditions.

3. The changes in predicted probabilities of war were not equal between India and Nigeria after increasing the two predictor variables because male school enrollment rate and the ratio of community exports to GDP vary in the extent of their impact in the two countries. While increasing the rate of male school enrollment rate in India had similar effects on increasing the predicted probability of war, the impact

was slightly higher in India due to the higher male school enrollment rate in the country. Similarly, the impact of raising the ratio of community exports to GDP in Nigeria on the predicted probability of war was slightly higher than in India due to the originally higher extent of the country's dependency on its exports.

Part 3: Confusion

```
#fit logistic regression model
glm.probs = predict(glm.fits, type = "response")
#Convert to war/no war
nrow(war_clean)
```

```
## [1] 688
```

```
glm.pred = rep("No war", 688)
glm.pred[glm.probs>0.5] = "War"
#Confusion matrix
conf_log <- table(glm.pred,war_clean$start)
conf_log
```

```
##
## glm.pred    0    1
##   No war 640  44
##   War    2    2
```

```
#Calculate misclassification rate
log_mis = (2 + 44)/688
log_mis
```

```
## [1] 0.06686047
```

```
#If always predict no war
642/688
```

```
## [1] 0.9331395
```

```
#Fraction correct where glm_fits also makes prediction
#glm makes only 644 no war and four war
# pundit says 688 no war
644/688
```

```
## [1] 0.9360465
```

The misclassification rate of the logistic regression model is 0.06686047.

Considering a pundit that always predicts no war, their predictions will be correct 0.9331395 of the time.

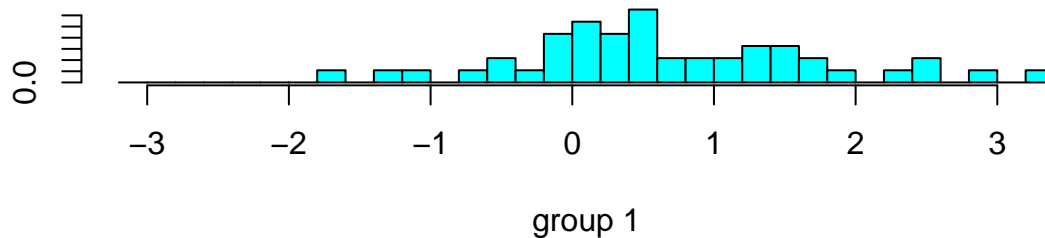
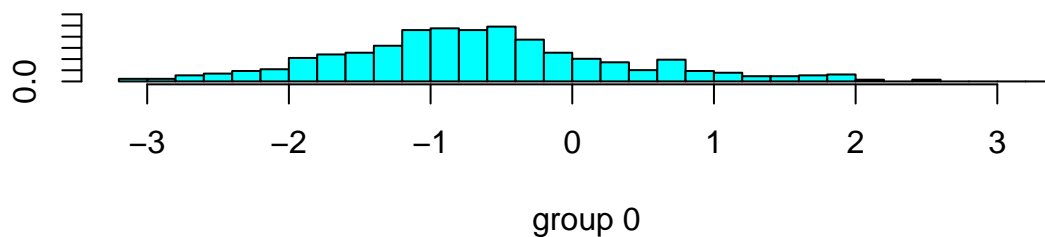
Considering the pundit that always predicts no war, their predictions will be correct on data points where the logistic regression model also makes a prediction 0.9360465 of the time. ##### Discriminant Analysis

```
library(ggplot2)
#Fit LDA
lda.fit <- lda(start ~ exports2 + schooling + growth + peace + concentration + lnpop + fractionalization +
lda.fit
```

```
## Call:
## lda(start ~ exports2 + schooling + growth + peace + concentration +
##      lnpop + fractionalization + dominance, data = war_clean)
##
## Prior probabilities of groups:
```

```
##           0           1
## 0.93313953 0.06686047
##
## Group means:
##   exports2 schooling    growth    peace concentration    lnpop
## 0 0.04505594 45.64548 1.73095794 357.7850      0.6038349 15.68224
## 1 0.04127454 28.34783 0.04384783 204.2826      0.5762391 16.58465
## fractionalization dominance
## 0      1764.882 0.4376947
## 1      2146.696 0.4565217
##
## Coefficients of linear discriminants:
##                               LD1
## exports2      1.858921e+00
## schooling     -6.409013e-03
## growth        -1.415680e-01
## peace         -4.496371e-03
## concentration -1.098846e+00
## lnpop         3.053406e-01
## fractionalization -6.507113e-05
## dominance     3.137136e-01
```

```
plot(lda.fit)
```



```
lda.pred = predict(lda.fit, war_clean)
conf_lda <- table(lda.pred$class, war_clean$start)
conf_lda
```

```
##
##      0    1
## 0 638  41
## 1   4   5
```

```
#Find LDA misclassification rate
```

```
lda_mis = (4 + 41)/688
```

```
lda_mis
```

```
## [1] 0.06540698
```

```
#Fit QDA
```

```
qda.fit <- qda(start ~ exports2 + schooling + growth + peace + concentration + lnpop + fractionalization + dominance, data = war_clean)
```

```
## Call:
```

```
## qda(start ~ exports2 + schooling + growth + peace + concentration +  
## lnpop + fractionalization + dominance, data = war_clean)
```

```
##
```

```
## Prior probabilities of groups:
```

```
## 0 1
```

```
## 0.93313953 0.06686047
```

```
##
```

```
## Group means:
```

```
## exports2 schooling growth peace concentration lnpop
```

```
## 0 0.04505594 45.64548 1.73095794 357.7850 0.6038349 15.68224
```

```
## 1 0.04127454 28.34783 0.04384783 204.2826 0.5762391 16.58465
```

```
## fractionalization dominance
```

```
## 0 1764.882 0.4376947
```

```
## 1 2146.696 0.4565217
```

```
qda.pred = predict(qda.fit, war_clean)
```

```
conf_qda <- table(qda.pred$class, war_clean$start)
```

```
conf_qda
```

```
##
```

```
## 0 1
```

```
## 0 623 31
```

```
## 1 19 15
```

```
#Find QDA misclassification rate
```

```
qda_mis = (19 + 31)/688
```

```
qda_mis
```

```
## [1] 0.07267442
```

```
#Why of different rates
```

```
x <- data.frame(war_clean$exports2, war_clean$schooling, war_clean$growth, war_clean$peace, war_clean$lnpop, war_clean$fractionalization, war_clean$dominance)
```

```
cor_matrix <- cor(x)
```

```
cor_matrix
```

```
## war_clean.exports2 war_clean.schooling
```

```
## war_clean.exports2 1.00000000 -0.08906098
```

```
## war_clean.schooling -0.08906098 1.00000000
```

```
## war_clean.growth 0.016181048 0.13602810
```

```
## war_clean.peace 0.062880876 0.39800518
```

```
## war_clean.concentration 0.001536338 0.07052828
```

```
## war_clean.lnpop -0.303299110 0.12845771
```

```
## war_clean.fractionalization 0.181536821 -0.36567075
```

```
## war_clean.dominance 0.093155843 0.04546414
```

```
## war_clean.growth war_clean.peace
```

```
## war_clean.exports2 0.016181048 0.06288088
```



```

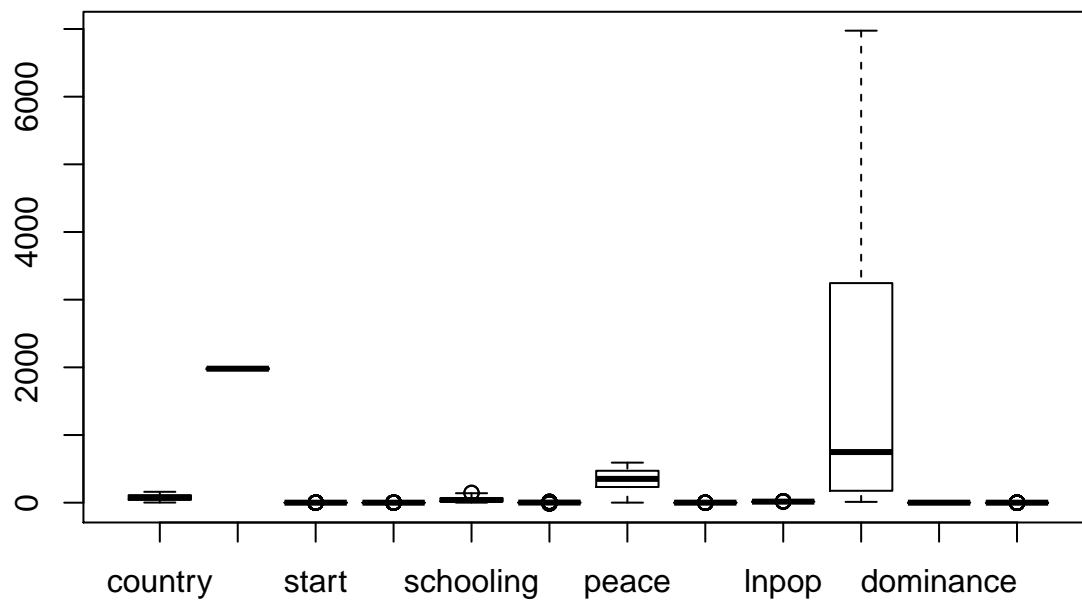
## war_clean.schooling      0.136028097      0.39800518
## war_clean.growth        1.000000000      -0.10362462
## war_clean.peace         -0.103624621      1.00000000
## war_clean.concentration  0.014257650      -0.04075917
## war_clean.lnpop         0.049958724      -0.18824162
## war_clean.fractionalization -0.181648026      -0.14505851
## war_clean.dominance     -0.000916039      0.09226036
##
## war_clean.concentration war_clean.lnpop
## war_clean.exports2      0.001536338      -0.30329911
## war_clean.schooling     0.070528281      0.12845771
## war_clean.growth        0.014257650      0.04995872
## war_clean.peace         -0.040759167      -0.18824162
## war_clean.concentration  1.000000000      0.19446482
## war_clean.lnpop         0.194464821      1.00000000
## war_clean.fractionalization -0.136267701      -0.07710752
## war_clean.dominance     -0.026922925      -0.14090610
##
## war_clean.fractionalization
## war_clean.exports2      1.815368e-01
## war_clean.schooling     -3.656707e-01
## war_clean.growth        -1.816480e-01
## war_clean.peace         -1.450585e-01
## war_clean.concentration  -1.362677e-01
## war_clean.lnpop         -7.710752e-02
## war_clean.fractionalization 1.000000e+00
## war_clean.dominance     -9.528842e-05
##
## war_clean.dominance
## war_clean.exports2      9.315584e-02
## war_clean.schooling     4.546414e-02
## war_clean.growth        -9.160390e-04
## war_clean.peace         9.226036e-02
## war_clean.concentration  -2.692293e-02
## war_clean.lnpop         -1.409061e-01
## war_clean.fractionalization -9.528842e-05
## war_clean.dominance     1.000000e+00

```

```

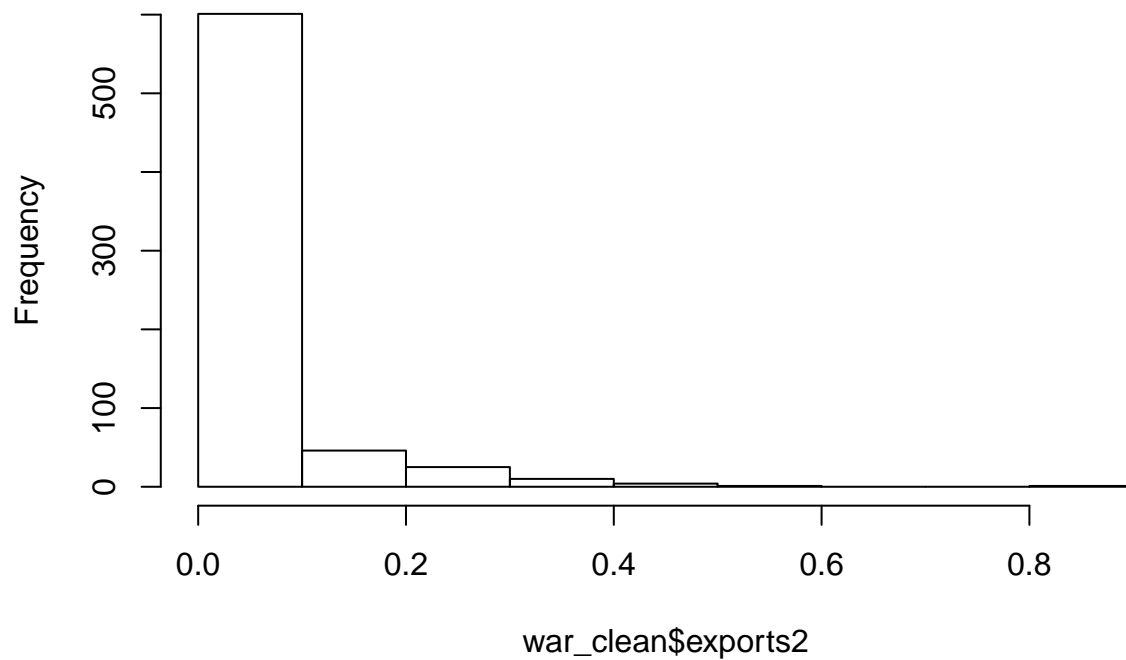
#Explore data, normality etc.
boxplot(war_clean)

```



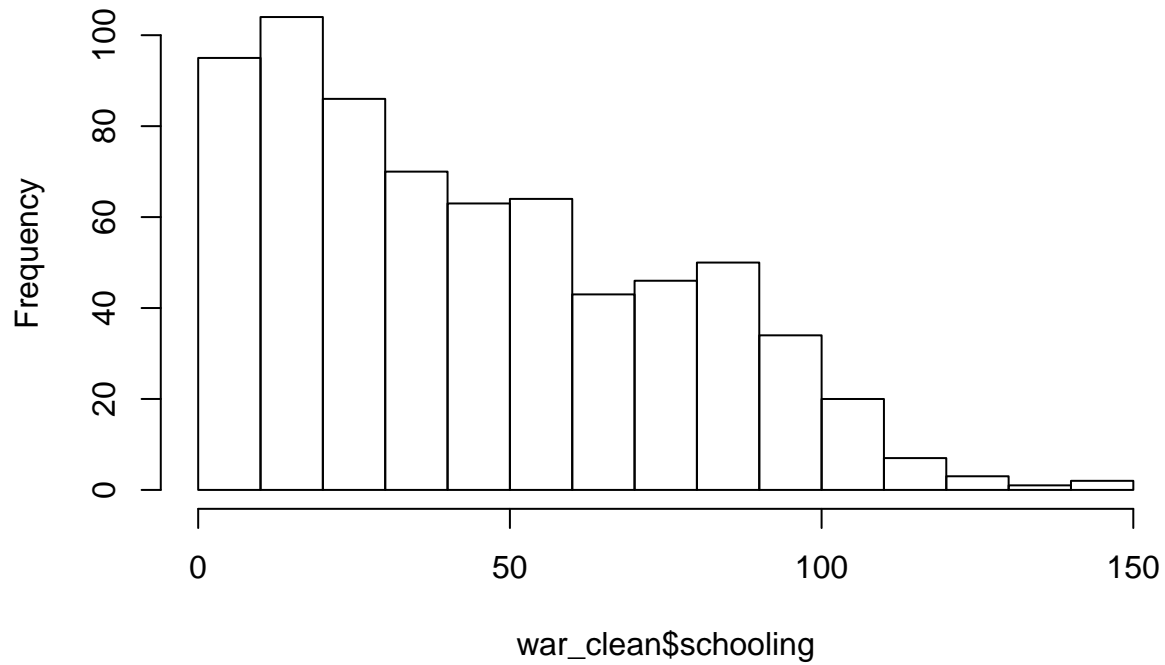
```
hist(war_clean$exports2)
```

Histogram of war_clean\$exports2



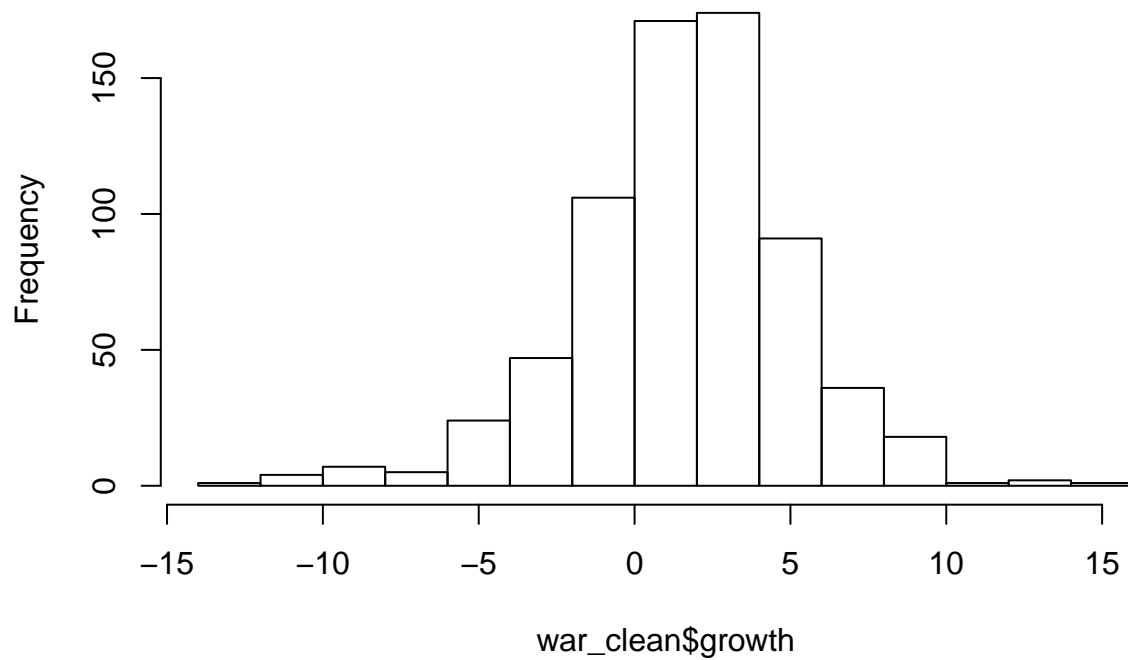
```
hist(war_clean$schooling)
```

Histogram of war_clean\$schooling



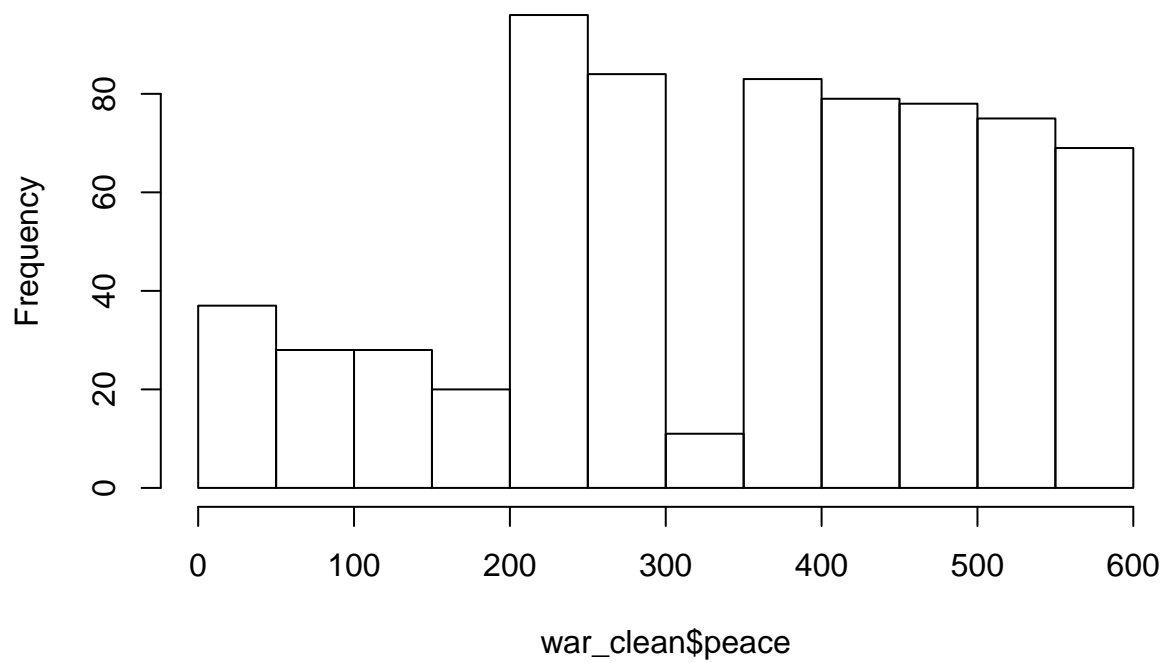
```
hist(war_clean$growth)
```

Histogram of war_clean\$growth



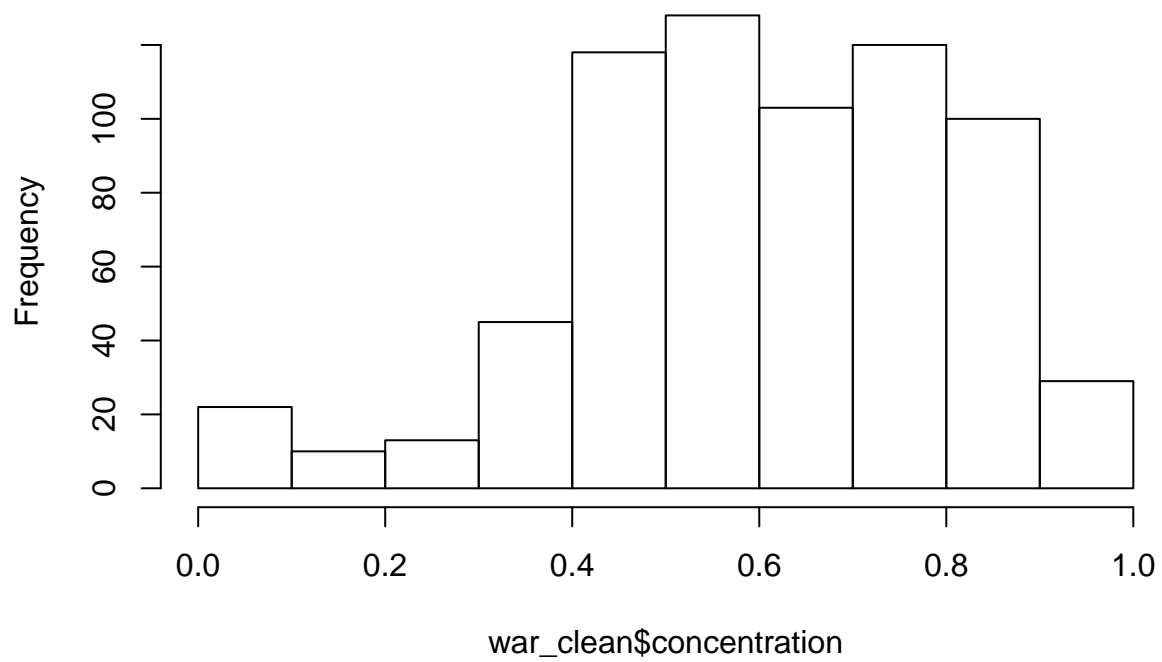
```
hist(war_clean$peace)
```

Histogram of war_clean\$peace



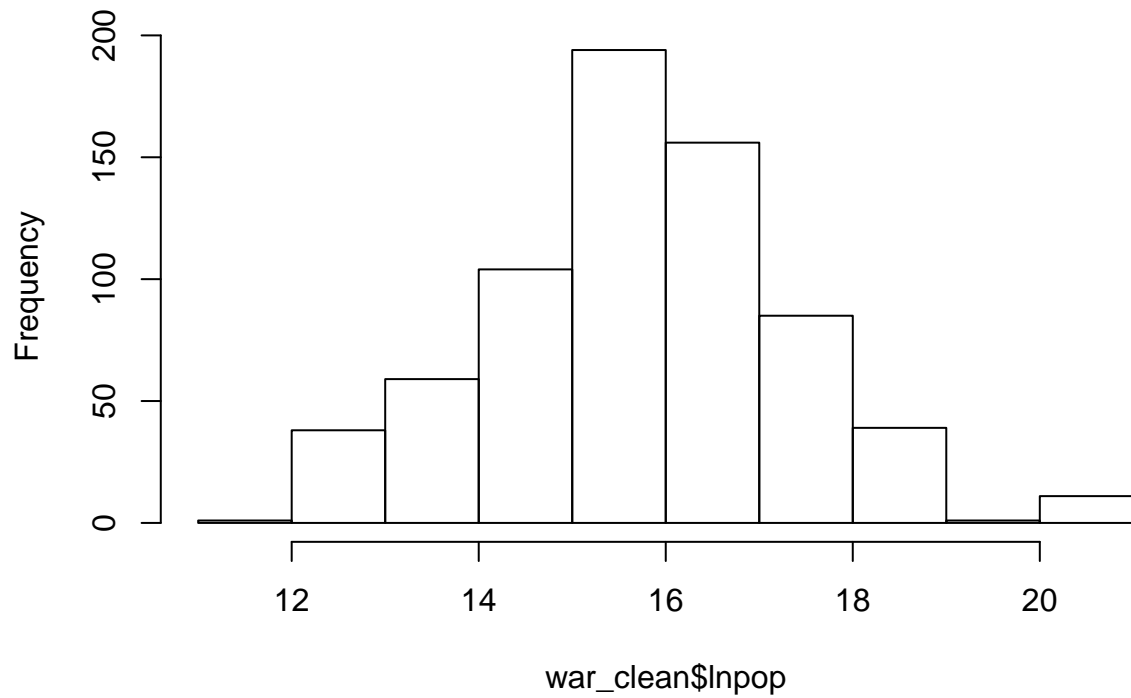
```
hist(war_clean$concentration)
```

Histogram of war_clean\$concentration



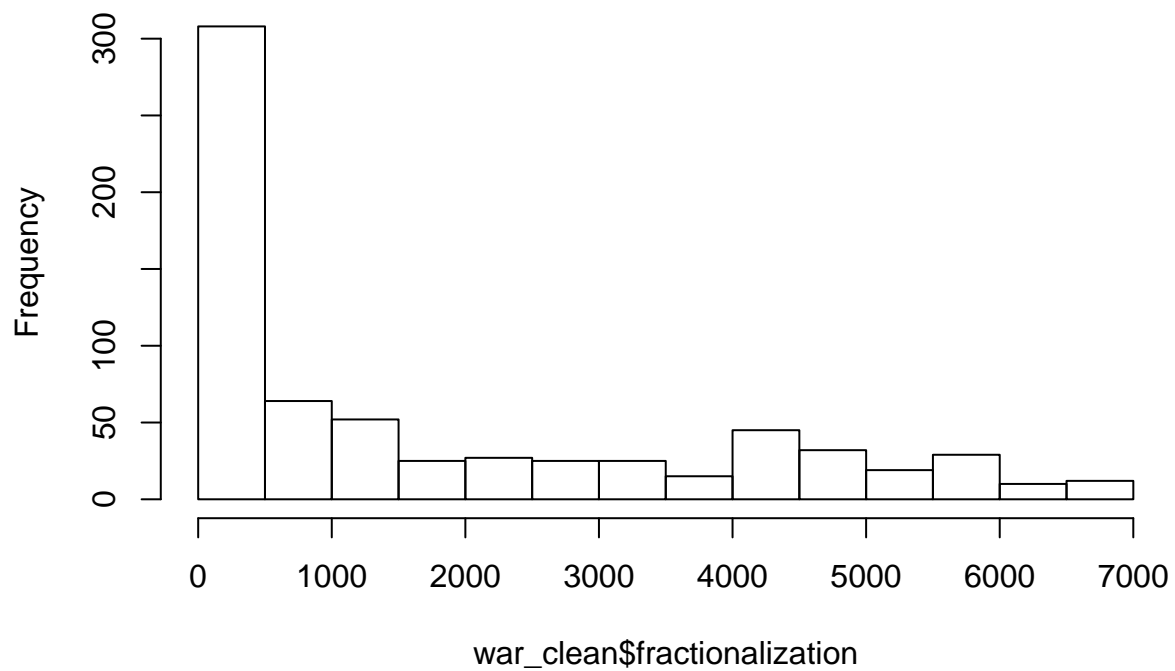
```
hist(war_clean$lnpop)
```

Histogram of war_clean\$lnpop



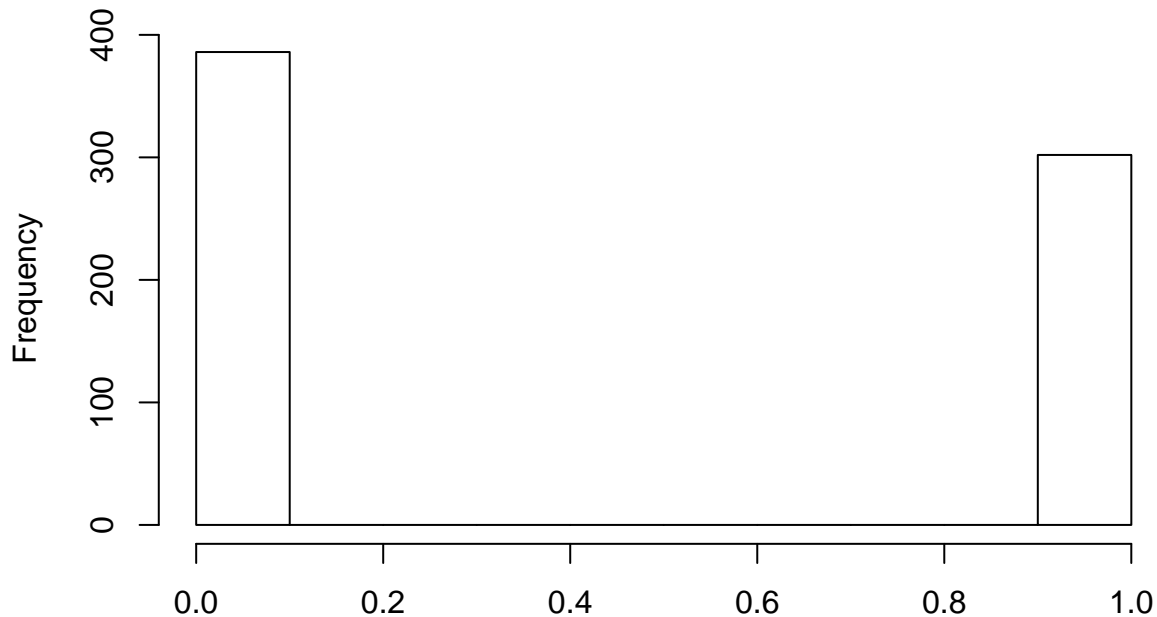
```
hist(war_clean$fractionalization)
```

Histogram of war_clean\$fractionalization



```
hist(war_clean$dominance)
```

Histogram of war_clean\$dominance



war_clean\$dominance

1. The misclassification rate for the LDA model is 0.06540698. 2. The misclassification rate for the QDA model is 0.07267442 3. While the discrepancies are not substantial, the misclassification rate for the QDA model is the highest and the misclassification rate for the LDA model is the lowest. The misclassification rate for the logistic regression model is slightly lower than the rate of the QDA model and higher than the LDA model (0.06686047).

The greater prediction accuracy of the LDA and logistic regression model suggest that the true decision boundaries (or set of points where both response classes do just as well) are more linear since the QDA model provides a non-linear quadratic decision boundary.

Problem Set: Chapter 4

4)

- 10% on average ($0.10^1 * 100$)
- 1% on average ($0.10^2 * 100$) ($0.10^{100} * 100$) d, As p increases, nearby observations decreases exponentially
- p=1, side = 0.1 p=2, side = $\sqrt{0.1} = 0.316$ p=100, side = $0.1^{1/100} = 0.977$ As p increases, it becomes necessary to use the entire range of of each p to include 10% of the training set.

6)

a. Logistic Regression: logistic regression, $p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}$ Plug in: $p(X) = \frac{e^{-6 + 0.05 \times 40 + 1 \times 3.5}}{1 + e^{-6 + 0.05 \times 40 + 1 \times 3.5}} = \exp(-6 + 0.05 * 40 + 1 * 3.5) / (1 + \exp(-6 + 0.05 * 40 + 1 * 3.5))$

```
## [1] 0.3775407
```

37.75%

b.

Solve: $0.5 = \frac{e^{-6 + 0.05 X_1 + 1 \times 3.5}}{1 + e^{-6 + 0.05 X_1 + 1 \times 3.5}} = \log\left(\frac{0.5}{1 - 0.5}\right) = -6 + 0.05 X_1 + 1 \times 3.5$

Which equates to solving the logit equation $\log(\frac{0.5}{1-0.5}) = -6 + 0.05X_1 + 1 \times 3.5$

```
(log(0.5/(1-0.5)) + 6 - 3.5*1)/0.05
```

```
## [1] 50
```

The student needs to study for 50 hours

7. Constant variance: $p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu_k)^2)}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu_l)^2)} = p_{yes}(4) = \frac{0.8 \exp(-\frac{1}{2 \times 36}(4-10)^2)}{0.8 \exp(-\frac{1}{2 \times 36}(4-10)^2) + (1-0.8) \exp(-\frac{1}{2 \times 36}(4-0)^2)}$

```
(0.8*exp(-1/(2*36)*(4-10)^2))/(0.8*exp(-1/(2*36)*(4-10)^2)+(1-0.8)*exp(-1/(2*36)*(4-0)^2))
```

```
## [1] 0.7518525
```

The probability is 75.2%