Lab 5: The Sound of Gunfire, Off in the Distance

Claire Jellison

10/12/2019

```
library(ggplot2)
library(dplyr)
```

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

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

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ISLR)
library(caret)
```

```
## Loading required package: lattice
```

```
library(lda)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

Chapter 4 Exercises

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

1)Estimate: Fit a logistic regression model for the start of civil war on all other variables except country and year (yes, this makes some questionable assumptions about independent observations); include a quadratic term for exports. Report the coefficients and their standard errors, together with R's p-values. Which ones are found to be significant at the 5% level?

```
head(war)
```

```
##       country year start exports schooling growth peace concentration
## 1 Afghanistan 1960     0   0.074         2     NA   172         0.492
## 2 Afghanistan 1965     0   0.074         4     NA   232         0.492
## 3 Afghanistan 1970     0   0.043        13     NA   292         0.492
## 4 Afghanistan 1975     1   0.085        13     NA   352         0.492
## 5 Afghanistan 1980    NA   0.240        16     NA    NA         0.492
## 6 Afghanistan 1985    NA      NA        11     NA    NA         0.492
##      lnpop fractionalization dominance
## 1 16.11969               132         1
## 2 16.22381               132         1
```

```
## 3 16.33779               132         1
## 4 16.45728               132         1
## 5 16.58497               132         1
## 6 16.71070               132         1
```

```
exports2 <- war$exports^2
logmodel <- glm(start ~ exports +
                  exports2 +
                  schooling +
                  growth +
                  peace +
                  concentration +
                  lnpop +
                  fractionalization +
                  dominance, data = war, family = binomial)
summary(logmodel)
```

```
##
## Call:
## glm(formula = start ~ exports + exports2 + schooling + growth +
##     peace + concentration + lnpop + fractionalization + dominance,
##     family = binomial, data = war)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3655  -0.3627  -0.1893  -0.0932   3.3636
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.307e+01  2.795e+00  -4.677 2.91e-06 ***
## exports            1.894e+01  5.865e+00   3.229 0.001243 **
## exports2          -2.944e+01  1.178e+01  -2.499 0.012449 *
## schooling         -3.156e-02  9.784e-03  -3.225 0.001259 **
## growth            -1.152e-01  4.307e-02  -2.675 0.007466 **
## peace             -3.713e-03  1.093e-03  -3.397 0.000681 ***
## concentration     -2.487e+00  1.005e+00  -2.474 0.013357 *
## lnpop              7.677e-01  1.658e-01   4.632 3.63e-06 ***
## fractionalization -2.135e-04  9.102e-05  -2.345 0.019020 *
## dominance          6.704e-01  3.535e-01   1.896 0.057920 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 337.73  on 687  degrees of freedom
## Residual deviance: 256.42  on 678  degrees of freedom
##   (600 observations deleted due to missingness)
## AIC: 276.42
##
## Number of Fisher Scoring iterations: 7
```

Exports, exports^2, schooling, growth, peace, concentration, lnpop and fractionalization are all significant at the 95% confidence level.

2)Interpretation: All parts of this question refer to the logistic regression model you just fit.

a)What is the model's predicted probability for a civil war in India in the period beginning 1975? What probability would it predict for a country just like India in 1975, except that its male secondary school enrollment rate was 30 points higher? What probability would it predict for a country just like India in 1975, except that the ratio of commodity exports to GDP was 0.1 higher?

```
logmodel.probs = predict(logmodel, type = "response")

india1975 <- filter(war, country == "India", year =="1975")
india1975$exports2 = india1975$exports^2

indiapred <- predict(logmodel, india1975, type = "response")
indiapred
```

```
##         1
## 0.3504199
```

```
likeindia <- india1975 %>%
  mutate(schooling = schooling + 30)

likeindiapred <- predict(logmodel, likeindia , type = "response")
likeindiapred
```

```
##       1
## 0.17309
```

```
likeindia2 <- india1975 %>%
  mutate(exports = exports + .1)
likeindia2$exports2 = likeindia2$exports^2

likeindiapred2 <- predict(logmodel, likeindia2 , type = "response")
likeindiapred2
```

```
##         1
## 0.6961378
```

The predicted probability for a civil war in India in the period begining in 1975 was 35%. If the only difference was a school enrollmentrat that wa 30 points higher then the predicted probability of civil war would go down to 17%. With the higher ratio of commodity exports the predicted probability went up to around 70%.

b)What is the model's predicted probability for a civil war in Nigeria in the period beginning 1965? What probability would it predict for a country just like Nigeria in 1965, except that its male secondary school enrollment rate was 30 points higher? What probability would it predict for a country just like Nigeria in 1965, except that the ratio of commodity exports to GDP was 0.1 higher?

```
nigeria1965 <- filter(war, country == "Nigeria", year =="1965")
nigeria1965$exports2 = nigeria1965$exports^2
nigeria1965
```

```
##   country year start exports schooling growth peace concentration    lnpop
## 1 Nigeria 1965     1   0.123         7  1.916   232         0.539 17.65479
##   fractionalization dominance exports2
## 1              6090         0 0.015129
```

```
nigeriapred <- predict(logmodel, nigeria1965, type = "response")
nigeriapred
```

```
##         1
## 0.1709917
```

```
likenigeria <- nigeria1965 %>%
  mutate(schooling = schooling + 30)

likenigeriapred <- predict(logmodel, likenigeria , type = "response")
likenigeriapred
```

```
##          1
## 0.07410315
```

```
likenigeria2 <- nigeria1965 %>%
  mutate(exports = exports + .1)
likenigeria2$exports2 = likenigeria2$exports^2

likenigeriapred2 <- predict(logmodel, likenigeria2 , type = "response")
likenigeriapred2
```

```
##         1
## 0.3310044
```

The predicted probability for a civil war in Nigeria in the period begining in 1965 was 17%. With the higher schooling rate the probability drops to around 7.4%. Finally with the greater exports to GDP ration the predicted probability goes to around 33%.

c)In the parts above, you changed the same predictor variables by the same amounts. If you did your calculations properly, the changes in predicted probabilities are not equal. Explain why not. (The reasons may or may not be the same for the two variables.)

Unlike a linear model, the slope is not constant on a logistic model for any given value of x, so they would not likely increment up the same amount.

3)Confusion: Logistic regression predicts a probability of civil war for each country and period. Suppose we want to make a definite prediction of civil war or not, that is, to classify each data point. The probability of misclassification is minimized by predicting war if the probability is ≥ 0.5, and peace otherwise.

a)Build a 2 by 2 confusion matrix (a.k.a. "classification table" or "contigency table") which counts: the number of outbreaks of civil war correctly predicted by the logistic regression; the number of civil wars not predicted by the model; the number of false predictions of civil wars; and the number of correctly predicted absences of civil wars. (Note that some entries in the table may be zero.)

Below is the code for splitting the data, but I think that it ended up being unecessary.

```
war_full <- na.omit(war) #includes only complete observations
war_full$exports2 = war_full$exports^2
smp_size = floor(0.5*nrow(war_full))
set.seed(11)
split= sample(seq_len(nrow(war_full)),size = smp_size)
war_test <- war_full[split, ]
war_train <- war_full[-split, ]
head(war_test)
```

```
##        country year start exports schooling growth peace concentration
## 360    Finland 1995     0   0.051       108 -1.897   592         0.787
## 10     Algeria 1965     0   0.190        10 -1.682    24         0.916
## 606     Kuwait 1985     0   0.386        95 -8.459   472         0.006
## 27   Argentina 1970     0   0.050        42  2.326   172         0.858
## 104     Belize 1995     0   0.213        47  4.165   592         0.333
## 1183    Uganda 1990     0   0.070        17  0.512    20         0.508
##         lnpop fractionalization dominance exports2
```

```
## 360  15.44632               224         0 0.002601
## 10   16.29398               132         1 0.036100
## 606  14.35317               180         1 0.148996
## 27   16.99198               434         1 0.002500
## 104  12.28627              2695         1 0.045369
## 1183 16.60851              5940         0 0.004900
```

Making a model on the dataset with complete observations,

```
logmodelfull <- glm(start ~ exports + exports2 + schooling + growth + peace + concentration + lnpop + fi
summary(logmodelfull)
```

```
##
## Call:
## glm(formula = start ~ exports + exports2 + schooling + growth +
##     peace + concentration + lnpop + fractionalization + dominance,
##     family = binomial, data = war_full)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3655  -0.3627  -0.1893  -0.0932   3.3636
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.307e+01  2.795e+00  -4.677 2.91e-06 ***
## exports            1.894e+01  5.865e+00   3.229 0.001243 **
## exports2          -2.944e+01  1.178e+01  -2.499 0.012449 *
## schooling         -3.156e-02  9.784e-03  -3.225 0.001259 **
## growth            -1.152e-01  4.307e-02  -2.675 0.007466 **
## peace             -3.713e-03  1.093e-03  -3.397 0.000681 ***
## concentration     -2.487e+00  1.005e+00  -2.474 0.013357 *
## lnpop              7.677e-01  1.658e-01   4.632 3.63e-06 ***
## fractionalization -2.135e-04  9.102e-05  -2.345 0.019020 *
## dominance          6.704e-01  3.535e-01   1.896 0.057920 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 337.73  on 687  degrees of freedom
## Residual deviance: 256.42  on 678  degrees of freedom
## AIC: 276.42
##
## Number of Fisher Scoring iterations: 7
```

Using the model to make predictions on the test data, we build the following confusion matrix,

```
my_log_probs = predict(logmodelfull, war_full, type = "response")
my_log_pred <- ifelse(my_log_probs < 0.5, "No", "Yes")
table(my_log_pred, war_full$start)
```

```
##
## my_log_pred   0   1
##         No  637  43
##         Yes   5   3
```

```
missclassificationlog = (5+43)/(637 + 43 + 5 + 3)
missclassificationlog
```

```
## [1] 0.06976744
```

b)What fraction of the logistic regression's predictions are incorrect, i.e. what is the misclassification rate? (Note that this is if anything too kind to the model, since it's looking at predictions to the same training data set).

We see that it guessed 48/688 of the observations incorrectly.

c)Consider a foolish (?) pundit who always predicts "no war". What fraction of the pundit's predictions are correct on the whole data set? What fraction are correct on data points where the logistic regression model also makes a prediction?

Based on the whole dataset, the pundit would correctly predict 642/688 of the observations (all actual instances of war not ocurring). Since the logistic regression model is making a prediction on every point the pundit would guess the same fraction correct for the second part of the question.

4)Comparison: Since this is a classification problem with only two classes, we can compare Logistic Regression right along side Discriminant Analysis.

a)Fit an LDA model using the same predictors that you used for your logistic regression model. What is the training misclassification rate?

```
#exports2 = war_train_full$exports^2
lda.fit= lda(start ~ exports  +
               exports2 +
               schooling +
               growth +
               peace +
               concentration +
               lnpop +
               fractionalization +
               dominance, data= war_full)
lda.fit
```

```
## Call:
## lda(start ~ exports + exports2 + schooling + growth + peace +
##     concentration + lnpop + fractionalization + dominance, data = war_full)
##
## Prior probabilities of groups:
##          0          1
## 0.93313953 0.06686047
##
## Group means:
##     exports   exports2 schooling     growth    peace concentration
## 0 0.1574330 0.04505594  45.64548 1.73095794 357.7850     0.6038349
## 1 0.1668478 0.04127454  28.34783 0.04384783 204.2826     0.5762391
##      lnpop fractionalization dominance
## 0 15.68224          1764.882 0.4376947
## 1 16.58465          2146.696 0.4565217
##
## Coefficients of linear discriminants:
##                              LD1
## exports             7.5279499420
## exports2           -9.3781631631
## schooling          -0.0063973381
```

```
## growth            -0.1242737735
## peace             -0.0041224852
## concentration     -1.1570459065
## lnpop              0.3813814561
## fractionalization -0.0001052021
## dominance          0.3644566472
```

```
lda.class=predict(lda.fit,war_full)$class
table(lda.class, war_full$start)
```

```
##
## lda.class   0   1
##         0 636  40
##         1   6   6
```

```
missclassificationlda = (6+40)/(636 + 40 + 6 + 6)
missclassificationlda
```

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

The missclassificaiton rate is around .067.

b)Fit a QDA model using the very same predictors. What is the training misclassification rate? How does the prediction accuracy of the three models compare? Why do you think this is?

```
qda.fit= qda(start ~ exports + exports2 + schooling + growth + peace + concentration + lnpop + fractiona
qda.fit
```

```
## Call:
## qda(start ~ exports + exports2 + schooling + growth + peace +
##     concentration + lnpop + fractionalization + dominance, data = war_full)
##
## Prior probabilities of groups:
##          0          1
## 0.93313953 0.06686047
##
## Group means:
##     exports   exports2 schooling     growth    peace concentration
## 0 0.1574330 0.04505594  45.64548 1.73095794 357.7850     0.6038349
## 1 0.1668478 0.04127454  28.34783 0.04384783 204.2826     0.5762391
##      lnpop fractionalization dominance
## 0 15.68224          1764.882 0.4376947
## 1 16.58465          2146.696 0.4565217
```

```
qda.class=predict(qda.fit,war_full)$class
table(qda.class, war_full$start)
```

```
##
## qda.class   0   1
##         0 618  26
##         1  24  20
```

```
missclassificationrateqda = (26+24)/(618 + 26 + 24 + 20)
missclassificationrateqda
```

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

The missclassification rate for the QDA is around .073.

c)How does the prediction accuracy of the three models compare? Why do you think this is?

```
missclassificationlog
```

```
## [1] 0.06976744
```

```
missclassificationlda
```

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

```
missclassificationrateqda
```

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

The LDA had the best prediction accuracy followed by the log and the QDA had the worst prediction rate. I think that the less flexible models probably preformed better because instances of civil war breaking out are so relatively rare in the dataset.

Challenge problem: Using the code available from class slides, construct an ROC curve for your logistic regression model. For an extra challenge, plot the ROC curves of all three models on the same plot.

Exercise 4

When the number of features p is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the curse of dimensionality, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigate this curse.

a) If we are only useing the 10% of the range of X closest to that test observation, then we will only be using on average 1/10 of the available observations.

b) Now we are only using 1/100 of the available observations because the observations must be in the ranges for both of the identified features. We can imagine all the observations as being scattered in a box [0,1] X [0,1] with area = 1. Then for the specified ranges the area inside the box of observations we can use is $.1 * .1 = .01$

c)With p = 100, we now can only use $.1^{100}$ of our available observations.

d) From the above, we see that KNN may not work well when the number of observations is small compared to the number of identified features. For any given test observation, we may end up with only a few training observations nearby to predict it's value. This will result in a high amount of variability and a scenario wherein outliers can have a very large influence due to the scarcity of observations. In order to have sufficient observations one may have to extend the range they use to predict however this could result in more bias as the range increases.

e) For p = 1 the hypercube is a line segement with length 0.1, for p = 2 the hypercube is a square with length $\sqrt{.1}$, for p = 100 the hypercube has side length $\sqrt[100]{.1}$. This is because $\sqrt{.1}^2 = .1$ meaning that it includes on average 10% of the training observations given that they lie in [0,1] along both axes. The same logic applies to p = 100. We see that the side length of the cube can decrease with added p, so long as the volume (or p dimensional equivalent) has 1/10 of the total observations on average.

Exercise 6

Suppose we collect data for a group of students in a statistics class with variables X1 = hours studied, X2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\beta^0 = -6, \beta^1 = 0.05, \beta^2 = 1$

(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

Our equation is $P(Y=1) = \frac{e^{\beta^0+\beta^1(X_1)+\beta^2(X_2)}}{1+e^{\beta^0+\beta^1(X_1)+\beta^2(X_2)}}$. Plugging in our values, we get

```
p = exp(-6 + .05*40 + 3.5) / (1 + exp(-6 + .05*40 + 3.5))
p
```

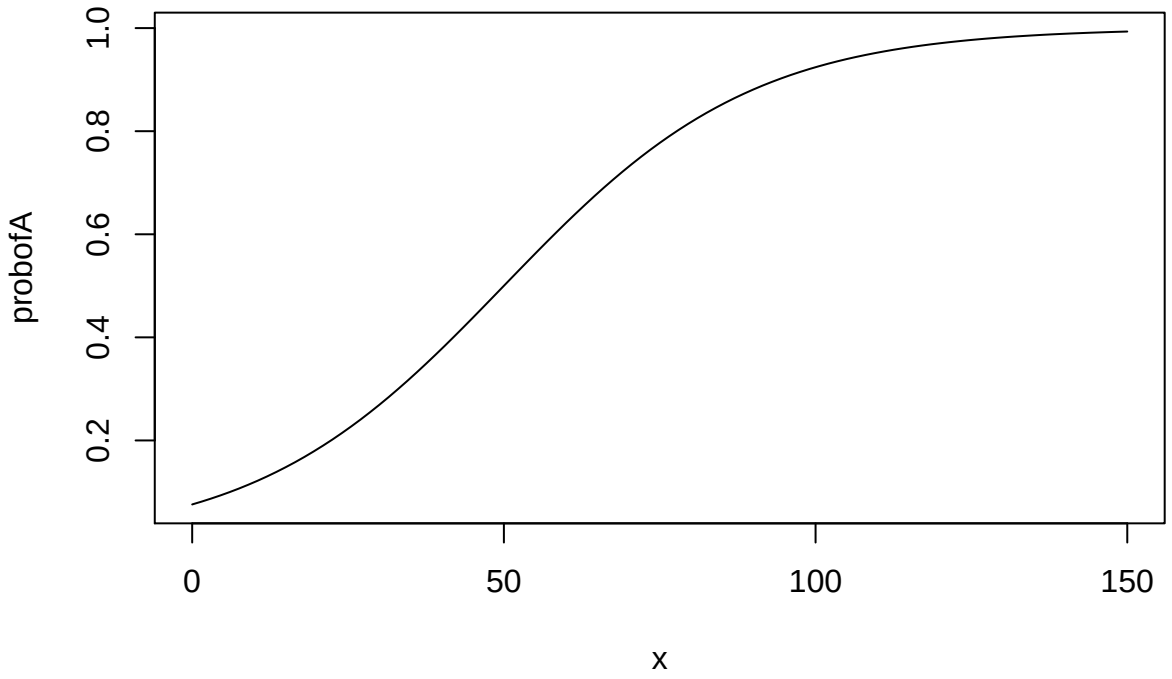
```
## [1] 0.3775407
```

This indicates that there is around a 38% chance that the student will recieve an A according to our model.

(b) How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?

Rearraging the formula above we can solve for X_2. Below, is the plot

```
probofA <- function(x) {exp(-6 + .05*x + 3.5) / (1 + exp(-6 + .05*x+ 3.5))}
plot(probofA, 0, 150)
```




```
x_vector <- seq(0,100, by = 1)
y_vector = probofA(x_vector)
mat <- cbind(x_vector,y_vector)
data <- data.frame("X" = x_vector, "Y" = y_vector)
data$Y
```

```
##   [1] 0.07585818 0.07943855 0.08317270 0.08706577 0.09112296 0.09534946
##   [7] 0.09975049 0.10433122 0.10909682 0.11405238 0.11920292 0.12455336
##  [13] 0.13010847 0.13587290 0.14185106 0.14804720 0.15446527 0.16110895
##  [19] 0.16798161 0.17508627 0.18242552 0.19000157 0.19781611 0.20587037
##  [25] 0.21416502 0.22270014 0.23147522 0.24048908 0.24973989 0.25922510
##  [31] 0.26894142 0.27888482 0.28905050 0.29943286 0.31002552 0.32082130
##  [37] 0.33181223 0.34298954 0.35434369 0.36586441 0.37754067 0.38936077
##  [43] 0.40131234 0.41338242 0.42555748 0.43782350 0.45016600 0.46257015
##  [49] 0.47502081 0.48750260 0.50000000 0.51249740 0.52497919 0.53742985
##  [55] 0.54983400 0.56217650 0.57444252 0.58661758 0.59868766 0.61063923
##  [61] 0.62245933 0.63413559 0.64565631 0.65701046 0.66818777 0.67917870
##  [67] 0.68997448 0.70056714 0.71094950 0.72111518 0.73105858 0.74077490
##  [73] 0.75026011 0.75951092 0.76852478 0.77729986 0.78583498 0.79412963
```

```
##  [79] 0.80218389 0.80999843 0.81757448 0.82491373 0.83201839 0.83889105
##  [85] 0.84553473 0.85195280 0.85814894 0.86412710 0.86989153 0.87544664
##  [91] 0.88079708 0.88594762 0.89090318 0.89566878 0.90024951 0.90465054
##  [97] 0.90887704 0.91293423 0.91682730 0.92056145 0.92414182
```

The student with the 3.5 gpa would need to study 50 hrs according to the model to a have a 50% chance of getting an A in the class.

Exercise 7

Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was mean X = 10, while the mean for those that didn't was 0. In addition, the variance of X for these two sets of companies was $\sigma^2 = 36$. Finally, 80 % of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was X = 4 last year.

look at pg 139

Using Bayes' theorem we can write that,

$$Pr(Y = Yes \mid X = 4) = \frac{\pi_y f_y(x)}{\pi_y f_k(y) + \pi_n f_k(n)}$$

We know that $f_(x) = \frac{e^{-(x-u)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$.

```
normaldist <- function(x, u, o){ (exp(-(x - u)^2/(2*o)))/ sqrt(2*pi*o)}
fy = normaldist(4, 10 , 36)
fn = normaldist(4, 0 ,36)
piy = .8
pin = .2
prob = piy*fy /(pin*fn + piy*fy)
prob
```

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

Therefore, given that the percentage profit was four, the likelihood of getting a dividend is around 75%.