Homework 10 Solution

Mohsen Nabian

8/1/2015

1. Using the anes_2008tr.csv dataset in Course Resources, model vote-rep (whether the respondent voted Republican in the last election) as a function of age, race, income, and ideology.

```
anes_2008tr <- read.csv("C:/Users/monabiyan/SkyDrive/Summer 2015/Statistics/HW8/anes_2008tr.csv",sep=",
```

a. What's the probability of voting Republican for a white person of average age, income, and ideology?

```
age_mean=mean(anes_2008tr$age)
income_mean=mean(anes_2008tr$income)
ideology_mean=mean(anes_2008tr$ideology_con)
RaceWhite<-1
lr1 <- glm(vote_rep ~ age + income + ideology_con+race_white,data=anes_2008tr,family="binomial")

summary(lr1)
```

```
##
## Call:
## glm(formula = vote_rep ~ age + income + ideology_con + race_white,
##     family = "binomial", data = anes_2008tr)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3038  -0.4875  -0.2900   0.4763   3.3359
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.076556   0.462932 -17.447  < 2e-16 ***
## age           0.004738   0.004190   1.131    0.258
## income        0.404587   0.074955   5.398 6.75e-08 ***
## ideology_con  1.041977   0.066277  15.721  < 2e-16 ***
## race_white    2.436605   0.167899  14.512  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1960.6  on 1538  degrees of freedom
## Residual deviance: 1170.3  on 1534  degrees of freedom
##   (783 observations deleted due to missingness)
## AIC: 1180.3
##
## Number of Fisher Scoring iterations: 5
```

```
# The nice thing about glm is that it automotically delete the rows with NA values. Here. 783 rows were
B0=as.numeric((lr1$coef)[1])
B1=as.numeric((lr1$coef)[2])
```

```r
B2=as.numeric((lr1$coef)[3])
B3=as.numeric((lr1$coef)[4])
B4=as.numeric((lr1$coef)[5])
H=exp(B0+B1*age_mean+B2*income_mean+B3*ideology_mean+B4*RaceWhite)
prob_vote_rep_white<-(H)/(1+H)
prob_vote_rep_white
```

```
## [1] 0.4860377
```

b. What's the change in probability of voting Republican for a person of average age, income, and ideology who switches from black to white?

```r
RaceBlack=0

RaceWhite=1

odd_ratio_white=exp(B0+B1*age_mean+B2*income_mean+B3*ideology_mean+B4*RaceWhite)

odd_ratio_black=exp(B0+B1*age_mean+B2*income_mean+B3*ideology_mean+B4*RaceBlack)

prob_vote_rep_white<-(odd_ratio_white)/(1+odd_ratio_white)

prob_vote_rep_black<-(odd_ratio_black)/(1+odd_ratio_black)

diff<-prob_vote_rep_white-prob_vote_rep_black

diff
```

```
## [1] 0.4096499
```

c. Using the e formula from the lesson, what's the effect on the odds ratio of shifting from black to white?

```r
diff<-odd_ratio_white-odd_ratio_black
diff
```

```
## [1] 0.8629626
```

d. What has a greater effect on voting Republican: an age increase of 50 years, or an incease of one income bracket?

lets caclculate based on the odd ratio which make the calculations simpler. As odd ratio increases, The P(y=1) increases.

one unit increase in x_i is equivalent to multiplying the odd with exp(Bi),two units increase in xi is equivalent to multiplying the odd with exp(2*Bi)… So

```r
age_increase_factor<-exp(50*B1)
age_increase_factor
```

```
## [1] 1.267313
```

```
income_increase_factor<-exp(1*B2)
income_increase_factor
```

```
## [1] 1.498683
```

So one unit increase in income has higher effect that 50 years of age increase in the probability of a person vote for a republican

e. Now run the regression with all the other variables in anes_2008tr (except for voted). How do your coefficients change? What do you think explains any coefficient that became or lost significance?

```
reg <- glm(vote_rep ~ age + income + ideology_con+ race_white+ gender_male+partyid_rep,            data
summary(lr1)
```

```
##
## Call:
## glm(formula = vote_rep ~ age + income + ideology_con + race_white,
##     family = "binomial", data = anes_2008tr)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3038  -0.4875  -0.2900   0.4763   3.3359
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.076556   0.462932 -17.447  < 2e-16 ***
## age           0.004738   0.004190   1.131    0.258
## income        0.404587   0.074955   5.398 6.75e-08 ***
## ideology_con  1.041977   0.066277  15.721  < 2e-16 ***
## race_white    2.436605   0.167899  14.512  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1960.6  on 1538  degrees of freedom
## Residual deviance: 1170.3  on 1534  degrees of freedom
##   (783 observations deleted due to missingness)
## AIC: 1180.3
##
## Number of Fisher Scoring iterations: 5
```

```
summary(reg)
```

```
##
## Call:
## glm(formula = vote_rep ~ age + income + ideology_con + race_white +
##     gender_male + partyid_rep, family = "binomial", data = anes_2008tr)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6959  -0.3689  -0.1715   0.2522   3.2794
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.582552   0.565275 -15.183  < 2e-16 ***
## age           0.014868   0.005328   2.791  0.00526 **
## income        0.258349   0.092098   2.805  0.00503 **
## ideology_con  0.509896   0.085593   5.957 2.57e-09 ***
## race_white    1.634050   0.201764   8.099 5.55e-16 ***
## gender_male  -0.138592   0.187991  -0.737  0.46098
## partyid_rep   0.894828   0.057354  15.602  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1960.57  on 1538  degrees of freedom
## Residual deviance:  814.69  on 1532  degrees of freedom
##   (783 observations deleted due to missingness)
## AIC: 828.69
##
## Number of Fisher Scoring iterations: 6
```

In the full regression, 'age' regains the significance, and gender_male is not significant. That means gender in fact has no effect on the vote. 'income','ideology' and 'race' still have their significance but have lower coefficients. Party-id has also significant effect. 'age' and 'income' got less significant. 'income' could be a function of other variables like, age and race and that is why it loses its significance.

2. Construct a simulated y variable with 100 observations where each observation (year) is a function of the previous observation: specifically, yt is 80% yt???1 + 20% random noise with mean 0 and sd 1 (and y1 = 1). Estimate an ARIMA model using auto.arima() from the forecast package and interpret the results, in particular the ARIMA(p,d,q) numbers and the coefficients reported, if any. What do you think is going on here?

```
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 3.1.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 3.1.3
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
## Loading required package: timeDate
```

```
## Warning: package 'timeDate' was built under R version 3.1.3
## This is forecast 6.1
```

```
yy=0
yy[1]=1
for (i in 2:100)
{
 yy[i]<-0.8*yy[i-1]+0.2*rnorm(n=1,mean=0,sd=1)
}
auto.arima(yy)
```

```
## Series: yy
## ARIMA(1,0,0) with zero mean
##
## Coefficients:
##          ar1
##       0.7029
## s.e.  0.0786
##
## sigma^2 estimated as 0.04545:  log likelihood=12.32
## AIC=-20.64   AICc=-20.52   BIC=-15.43
```

The auto arima sayas AR=1, Differencing=0, MA=0 that means y(t) is only dependent to y(t-1) and not more lags. which corresponds to our expectations. Moreover, 0 differencing is that y is not a time series variable which also make sense since there is no time dependency for y.

Finally 0 for MA means there is no moving average. That is true beacuase every element depends on the preivious element as well as external noises.

The results also suggests 0.8627 for the coefficient for y(t-1) which is close to 0.8 that we chose in our formula.

3. Find some existing data that either has temporal or bindary dependent variable data and run a ARIMA or logit model on it and interpret the results in detail. You can use data from the previous assignment, and you can construct a binary variable out of some existing continuous variable if you like.

```
mydata <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")
## view the first few rows of the data
head(mydata)
```

```
##   admit gre  gpa rank
## 1     0 380 3.61    3
## 2     1 660 3.67    3
## 3     1 800 4.00    1
## 4     1 640 3.19    4
## 5     0 520 2.93    4
## 6     1 760 3.00    2
```

```
admission_reg<- glm(admit ~ gre + gpa + rank ,data=mydata ,family="binomial")
summary(admission_reg)
```

```
##
## Call:
## glm(formula = admit ~ gre + gpa + rank, family = "binomial",
##     data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5802  -0.8848  -0.6382   1.1575   2.1732
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.449548   1.132846  -3.045  0.00233 **
## gre          0.002294   0.001092   2.101  0.03564 *
## gpa          0.777014   0.327484   2.373  0.01766 *
## rank        -0.560031   0.127137  -4.405 1.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 459.44  on 396  degrees of freedom
## AIC: 467.44
##
## Number of Fisher Scoring iterations: 4
```

accordig to the result, 'rank' plays an important role on acceptance of the applicants.Of course lower rank means more qualified applicant and thats why we see negetive coefficient. However, GRE and GPA are important but less significant.1 reason that Gre coefficient is very low is that the gre scale is 100-800 and much higher to rank and gpa.