STAT645 - Homework 5

Salih Kilicli

10/1/2019

```r
library(MASS)
library(DescTools)
setwd("~/Desktop/STAT645/Data")
```

Problem 1: Suppose that a pilot study was conducted to assess the feasibility of patient recruitment, the ability of patients and clinicians to comply with study protocols, and the use of data collection instruments to collect cost-effective data, and to obtain variability estimates for sample-size calculations for a full-scale trial. Suppose that twenty patients were randomized into the study with treatment and control group. Out of twelve patients in the treatment, 8 showed substantial improvements in the main patient-rated outcomes at the end of the 12-week intervention phase. Let π be the proportion patients who showed substantial improvements in the main patient-rated outcomes at the end of the 12-week intervention phase in the treatment group.

(a) Construct two-sided 95% confidence interval for π using Agresti-Coull, Jeffreys, Wilson, Clopper-Pearson methods.

```r
library(DescTools)
cinterval=c("agresti-coul","jeffreys", "wilson", "clopper-pearson")
BinomCI(8, 12, conf.level=0.95, sides="two.sided", method=cinterval) # 8 no of succes, 12 sample size
```

```
##                       est    lwr.ci    upr.ci
## agresti-coul    0.6262510 0.3880110 0.8644910
## jeffreys        0.6666667 0.3875639 0.8754627
## wilson          0.6666667 0.3906221 0.8618799
## clopper-pearson 0.6666667 0.3488755 0.9007539
```

(b) What would be the required sample size for the actual study if we want to test $H_0 : \pi = 0.6 \text{ versus } H_a : \pi > 0.6$ at the 5% level, and we desire to have 90% power to reject H_0 when in fact $\pi = 0.7$?

```r
pi0=0.6; a=0.05;   # 1-sided z_crit=qnorm(1-a) whereas 2-sided z_crit=qnorm(1-a/2) (quantile)
pi1=0.7; b=0.10;   # to find p value for a given z use pnorm(z)
n=((qnorm(1-a)*sqrt(pi0*(1-pi0))+qnorm(1-b)*sqrt(pi1*(1-pi1)))/(pi1-pi0))^2
cat('The required sample size for one sided alternative is n =', n)
```

```
## The required sample size for one sided alternative is n = 194.0703
```

(c) Recalculate the needed sample size for the above scenario considering that there is a possibility of 35% drop-out or study non-compliance.

```r
n1=n/(1-0.35)
cat('The required sample size considering drop-out possibility is n* =', n1)
```

```
## The required sample size considering drop-out possibility is n* = 298.5697
```

Problem 2: Suppose in an observational study on PTSD we have obtained the following data. Test at the 5% level if there is any association between PTSD and gender. Write the hypothesis, do the analysis, and write your conclusion. Use the both methods, the chi-square test of independence and the odds ratio approach.

PTSD	Gender(M)	Gender(F)
Y	40	60
N	280	156

H_0 : There is no association between two variables, PTSD and Gender H_a : There exists an association between two variables, PTSD and Gender

```r
library(MASS)
PTSD=matrix(c(40,60,280,156),ncol=2,byrow=TRUE)
rownames(PTSD)=c("Y","N")
colnames(PTSD)=c("Gender(M)","Gender(F)")
PTSD=as.table(PTSD)
PTSD
```

```
##   Gender(M) Gender(F)
## Y        40        60
## N       280       156
```

```r
chsq=chisq.test(PTSD, correct=F)
cat('p-value using chisquared test is p =', chsq$p.value, 'which is much smaller than alpha=0.05')
```

```
## p-value using chisquared test is p = 8.448273e-06 which is much smaller than alpha=0.05
```

```r
orhat=(156*40)/(60*280)# ORhat=[pr(A=1|B=1)pr(A=0|B=0)]/[pr(A=0|B=1)pr(A=1|B=0)]=[n_11xn_00]/[n_01xn_10]
lorhat=log(orhat)      # log(ORhat)=log([n_11xn_00]/[n_01xn_10])
stder=sqrt(1/40+1/60+1/280+1/156) # Tao=sqrt(1/n_11+1/n_00+1/n_01+1/n_10)
CI=lorhat+c(-1,1)*qnorm(1-0.05/2)*stder
cat('95% CI for log(ORhat) is given by CI =', CI, 'which clearly doesnt contain 0')
```

```
## 95% CI for log(ORhat) is given by CI = -1.435825 -0.5449719 which clearly doesnt contain 0
```

Both of the methods implies that the data provide a strong evidence that the two variables are associated. Because in chisq test we found that p << 0.05 and 0 is not in the %95 confidence interval for log().

Problem 3: . Consider the Pima.tr dataset in library(MASS). This dataset contains information on some 200 Pima Indian women who were all at least 21 years old. Please look at https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Pima.tr.html (https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Pima.tr.html) for details of the data. Suppose that interest is finding association between type and npreg, glu, bp, skin, bmi, ped, age. Before the analysis, tranform glu, bp, bmi, ped, age into variables that have zero mean and standard deviation one.

(a) Test if age is positively associated with the disease (chances of the disease).

```
library(MASS)
attach(Pima.tr)
glu0=as.vector((scale(glu, center=T, scale=T)))
bp0=as.vector((scale(bp, center=T, scale=T)))
bmi0=as.vector((scale(bmi, center=T, scale=T)))
ped0=as.vector((scale(ped, center=T, scale=T)))
age0=as.vector((scale(age, center=T, scale=T)))

logit=glm(type~npreg+glu0+bp0+skin+bmi0+ped0+age0, family="binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = type ~ npreg + glu0 + bp0 + skin + bmi0 + ped0 +
##     age0, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9830  -0.6773  -0.3681   0.6439   2.3154
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.268201   0.724581  -1.750  0.08007 .
## npreg        0.103183   0.064694   1.595  0.11073
## glu0         1.017051   0.214935   4.732 2.22e-06 ***
## bp0         -0.054729   0.212840  -0.257  0.79707
## skin        -0.001917   0.022500  -0.085  0.93211
## bmi0         0.512632   0.262538   1.953  0.05087 .
## ped0         0.559275   0.204462   2.735  0.00623 **
## age0         0.452007   0.242458   1.864  0.06228 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 256.41  on 199  degrees of freedom
## Residual deviance: 178.39  on 192  degrees of freedom
## AIC: 194.39
##
## Number of Fisher Scoring iterations: 5
```

```
T=(0.452007-0)/0.242458  # T= (\hat{\beta_age}-0)/se(\hat{\beta_age}) Test statistic for H_0=\beta_age=0
p=length(logit$coefficients)
n=length(age0)
CI=0.452007+qt(1-0.05,n-p)*0.242458
OCI=c(CI,Inf)
cat("One sided interval for %95 confidence is",OCI)
```

```
## One sided interval for %95 confidence is 0.8527485 Inf
```

```
cat("One sided p-value of age0 is the half of the two-sided p-value, where p=", 0.06228/2)
```

```
## One sided p-value of age0 is the half of the two-sided p-value, where p= 0.03114
```

For this problem $H_0 : \hat{\beta}_{age0} = 0$, whereas the alternative hypothesis is given by $H_a : \hat{\beta}_{age0} > 0$. At $\%5$ level since one sided CI doesn't include 0 or one-sided $p_{value} < 0.05$, we reject the Null hypothesis. Thus, there is enough evidence that age is positively related with the disease.

(b) Test $H_0 : \beta_{skin} = \beta_{bp} = \beta_{bmi} = 0$ at the 5% level. Use both the likelihood ratio and Wald test approaches.

```
library(aod)
logit.ha=glm(type~npreg+glu0+bp0+skin+bmi0+ped0+age0, family="binomial")
logit.h0=glm(type~npreg+glu0+ped0+age0, family="binomial")
anova(logit.h0,logit.ha, test="LRT") # LRH - Likelihood Ratio Test
```

```
## Analysis of Deviance Table
##
## Model 1: type ~ npreg + glu0 + ped0 + age0
## Model 2: type ~ npreg + glu0 + bp0 + skin + bmi0 + ped0 + age0
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       195     184.74
## 2       192     178.39  3   6.3487  0.09582 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
wald.test(b=coef(logit), Sigma=vcov(logit), Terms=4:6)
```

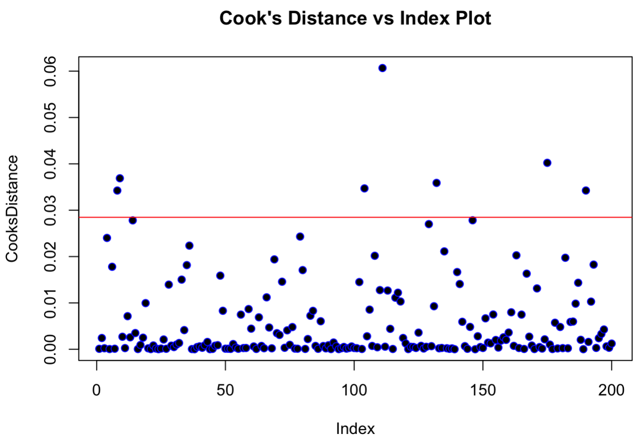
```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 6.0, df = 3, P(> X2) = 0.11
```

(c) Provide a Cook's distance plot and check if there is any influential observation.

```
CooksDistance=cooks.distance(logit)
summary(CooksDistance)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 6.660e-06 2.376e-04 1.066e-03 5.646e-03 7.295e-03 6.067e-02
```

```
Q3R=7.295e-03
Q1R=2.376e-04
IQR=Q3R-Q1R
plot(CooksDistance, main="Cook's Distance vs Index Plot", pch=21, col="blue", bg=1)
abline(h=Q3R+3*IQR, col='red')
```



```
index=CooksDistance[which(CooksDistance>Q3R+3*IQR)]
cat("Indexes and values of influential observations are given by \n")
```

```
## Indexes and values of influential observations are given by
```

```
index
```

```
##          8          9        104        111        132        175 
## 0.03424452 0.03689080 0.03470689 0.06067031 0.03589055 0.04022298 
##        190 
## 0.03424811
```

```
cat("Number of influential points is", length(index))
```

```
## Number of influential points is 7
```

(d) Consider the model containing, npreg, glu, ped, age, age^2, ped×age, glu×age, glu×ped as explanatory variables. Do a stepwise regression to find the the best fitted model for this data based on the above specified explanatory variables [Hint use the step(obj) function].

```
logit1=glm(type~npreg+glu0+ped0+age0+I(age0^2)+I(ped0*age0)+I(glu0*age0)+I(glu0*ped0), family="binomial")
step(logit1)
```

```
## Start:  AIC=191.64
## type ~ npreg + glu0 + ped0 + age0 + I(age0^2) + I(ped0 * age0) + 
##     I(glu0 * age0) + I(glu0 * ped0)
## 
##                  Df Deviance    AIC
## - I(glu0 * ped0)  1   173.73 189.73
## - npreg           1   173.98 189.98
## - I(glu0 * age0)  1   174.10 190.10
## <none>                173.64 191.64
## - I(age0^2)       1   175.84 191.84
## - I(ped0 * age0)  1   178.57 194.57
## - age0            1   183.46 199.46
## - ped0            1   183.68 199.68
## - glu0            1   205.32 221.32
## 
## Step:  AIC=189.73
## type ~ npreg + glu0 + ped0 + age0 + I(age0^2) + I(ped0 * age0) + 
##     I(glu0 * age0)
## 
##                  Df Deviance    AIC
## - npreg           1   174.08 188.08
## - I(glu0 * age0)  1   174.48 188.48
## <none>                173.73 189.73
## - I(age0^2)       1   175.85 189.85
## - I(ped0 * age0)  1   178.82 192.82
## - age0            1   183.46 197.46
## - ped0            1   184.58 198.58
## - glu0            1   207.17 221.17
## 
## Step:  AIC=188.08
## type ~ glu0 + ped0 + age0 + I(age0^2) + I(ped0 * age0) + I(glu0 * 
##     age0)
## 
##                  Df Deviance    AIC
## - I(glu0 * age0)  1   174.97 186.97
## <none>                174.08 188.08
## - I(age0^2)       1   176.87 188.87
## - I(ped0 * age0)  1   179.27 191.27
## - ped0            1   184.66 196.66
## - age0            1   193.47 205.47
## - glu0            1   207.53 219.53
## 
## Step:  AIC=186.97
## type ~ glu0 + ped0 + age0 + I(age0^2) + I(ped0 * age0)
## 
##                  Df Deviance    AIC
## <none>                174.97 186.97
## - I(age0^2)       1   178.56 188.56
## - I(ped0 * age0)  1   179.97 189.97
## - ped0            1   186.15 196.15
## - age0            1   193.60 203.60
## - glu0            1   207.55 217.55
```

```
##
## Call:  glm(formula = type ~ glu0 + ped0 + age0 + I(age0^2) + I(ped0 *
##     age0), family = "binomial")
##
## Coefficients:
##    (Intercept)            glu0            ped0            age0
##        -0.4746          1.1084          0.6801          1.2032
##      I(age0^2)  I(ped0 * age0)
##        -0.3655          0.6335
##
## Degrees of Freedom: 199 Total (i.e. Null);  194 Residual
## Null Deviance:       256.4
## Residual Deviance: 175   AIC: 187
```