Biostats - Inference and Diagnostics

Rob Leonard (robleonard@tamu.edu (mailto:robleonard@tamu.edu))

1) Pilot Study Patient Recruitment

1a) Construct two-sided CI's for pi using AC, J, W and CP methods.

```
library(DescTools)
BinomCI(8,12, conf.level = 0.95, method = c("agresti-coul","jeffreys","wilson","clopper-pearson"
))
```

```
##                      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
```

1b) Find the required sample size. $n = (\frac{Z_{(1-alpha)}\sqrt{p_0(1-p_0)}+Z_{(1-beta)}\sqrt{p_1(1-p_1)}}{\delta})^2$ $n = (\frac{Z_{0.95}\sqrt{.6(1-.6)}+Z_{.1}\sqrt{.7(1-.7)}}{.1})^2$

```
(n1 = ((((qnorm(.95))*sqrt(.6*.4))+((-1*qnorm(.1))*sqrt(.7*.3)))/.1)^2)
```

```
## [1] 194.0703
```

We would need a sample size of 195 (so 1 more than 194 since we needed slightly over 194).

1c Recalculate n for a potential 35% dropout rate

```
(n1/(1-.35))
```

```
## [1] 298.5697
```

We would need n=300 samples if there is a potential drop out rate of 35%. (299 might be enough depending on rounding, but a 35% dropout rate for 299 samples is 104.65 which is essentially 105 people with a net of 194 observed. But we need 195 observed so n=300 is preferable).

2) PTSD Observational Study

2) Test for an association between PTSD and Gender using Chi Square Test of Independence and Odds Ratio

1. Pearson Chi Squared Test of Independence Null Hypothesis is H_0 : there is no association between the variables PTSD and Gender (either having PTSD or not and being either Male or Female). Alternative Hypothesis is H_a : there is an association between the two variables PTSD and Gender.

```
ptsd = matrix(c(40,60,280,156),byrow=TRUE, ncol=2)
colnames(ptsd) = c('Male', 'Female')
rownames(ptsd) = c('PTSD','No PTSD')
ptsd.tb = as.table(ptsd)
chisq.test(ptsd.tb, correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ptsd.tb
## X-squared = 19.834, df = 1, p-value = 8.448e-06
```

Check to make sure expected frequencies greater than 5 for all cells.

```
n = 40+60+280+156
sum(ptsd[1,1]+ptsd[1,2])*sum(ptsd[1,1]+ptsd[2,1])/n
```

```
## [1] 59.70149
```

```
sum(ptsd[1,1]+ptsd[1,2])*sum(ptsd[1,2]+ptsd[2,2])/n
```

```
## [1] 40.29851
```

```
sum(ptsd[2,1]+ptsd[2,2])*sum(ptsd[1,1]+ptsd[2,1])/n
```

```
## [1] 260.2985
```

```
sum(ptsd[2,1]+ptsd[2,2])*sum(ptsd[1,2]+ptsd[2,2])/n
```

```
## [1] 175.7015
```

The high Chi Squared test statistic and very low p-value (< .05%) means that we have significant evidence to reject the null hypothesis that there is not an association between PTSD and Gender. The condition that the expected cell values all exceed 5 are met, so using Pearson's Chi Squared Test of independence is valid.

2. Now repeat test of no association using the Odds Ratio Method

```
tausq = 1/40+1/60+1/280+1/156   # get variance using large sample approximation which is sum 1/f
requencies
sdtau = sqrt(tausq)   # get std dev for CI
or = (40*156)/(60*280)
lor = log(or)
(lor-qnorm(.975)*sdtau)  # lower bound
```

```
## [1] -1.435825
```

```
(lor+qnorm(.975)*sdtau)   # upper bound
```

```
## [1] -0.5449719
```

The confidence interval for the log odds ratio is (-1.435,-0.545). Since this interval does not include 0 which is the null hypothesis value for the log(OR), we have significant evidence to reject the null hypothesis that there is not an association between PTSD and Gender. So both the Chi Squared and Odds Ratio methods reach the same conclusion, to reject the null hypotheses. There is evidence of an association between PTSD and Gender.

3) Pima Indian Dataset

Load data and scale variables

```
library(MASS)
data("Pima.tr")
ScaledData=(Pima.tr)
ScaledData$glu=scale(ScaledData$glu)
ScaledData$bp=scale(ScaledData$bp)
ScaledData$bmi=scale(ScaledData$bmi)
ScaledData$ped=scale(ScaledData$ped)
ScaledData$age=scale(ScaledData$age)
```

3a) Test if age is positively associated with the disease

```
model3 = glm(type~npreg+glu+bp+skin+bmi+ped+age, data=ScaledData, family="binomial")
summary(model3)
```

```
##
## Call:
## glm(formula = type ~ npreg + glu + bp + skin + bmi + ped + age,
##     family = "binomial", data = ScaledData)
##
## 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
## glu          1.017051   0.214935   4.732 2.22e-06 ***
## bp          -0.054729   0.212840  -0.257  0.79707
## skin        -0.001917   0.022500  -0.085  0.93211
## bmi          0.512632   0.262538   1.953  0.05087 .
## ped          0.559275   0.204462   2.735  0.00623 **
## age          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
```

```
# test if age is positively associated with type, so a 1-sided test
qt(.95,192)
```

```
## [1] 1.652829
```

Yes, there is a positive association between age and type. The null hypotheses is $H_0 : \beta_7 <= 0$ No positive association between Age and Type (chances of having the disease) controlling for the other variables versus the alternative hypothesis of $H_a : \beta_7 > 0$ a positive association between age and type. The t-test statistic is shown in the summary table as 1.864 and the t-critical value for a 1-sided hypothesis test at the usual α=.05 level is 1.65. The test statistic is greater than the critical value which indicates that we have sufficient evidence to reject the null hypothesis that there is not any non positive association between age and type.

3b) Test that the coefficients for skin, bp and bmi are all 0 using likelihood and Wald at alpha=0.05.
Likelihood Ratio Test First $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ versus H_a : at least one non-zero.

```
library(lmtest)
```

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

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
model3alt = glm(type~npreg+glu+ped+age, data=ScaledData, family="binomial")
lrtest(model3,model3alt) # using R functions
```

```
## Likelihood ratio test
##
## Model 1: type ~ npreg + glu + bp + skin + bmi + ped + age
## Model 2: type ~ npreg + glu + ped + age
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -89.195
## 2   5 -92.370 -3 6.3487    0.09582 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# or calculate by hand
diff01 = as.numeric(logLik(model3alt))-as.numeric(logLik(model3)) #loglik rest-unrest
teststat1 = -2*diff01
df1 = df.residual(model3alt)-df.residual(model3)
pval1 = 1-pchisq(teststat1, df1)
cat('P-value calcualted by hand', pval1) # p-value calculated by hand
```

```
## P-value calcualted by hand 0.09582443
```

The p-value of 0.096 is more than .05, indicates we do not have significant evidence to reject the null hypothesis that all three coefficients are equal to zero.

Recheck using Wald.

```
library(aod)
wald.test(b=coef(model3), Sigma = vcov(model3), Terms= 4:6)
```

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

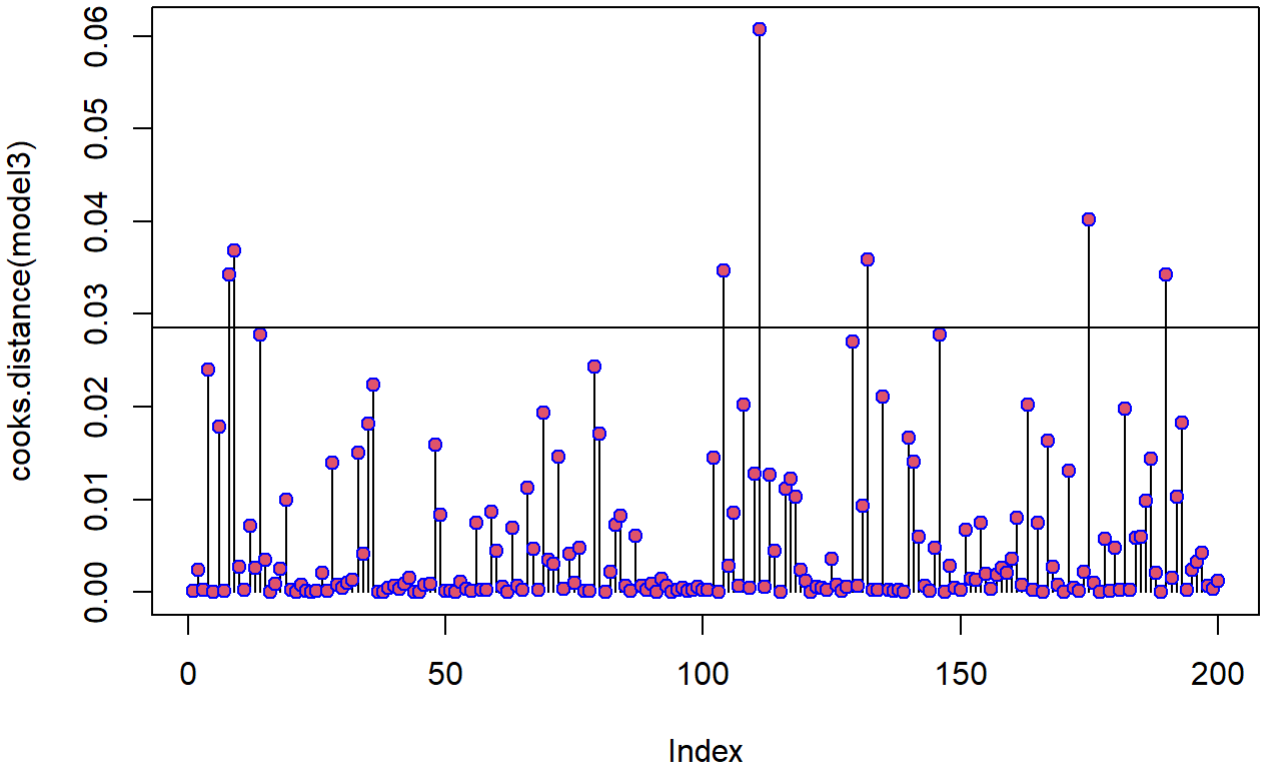
The Wald test results concur with the likelihood ratio test. At the 5% level, the p-value from the Wald test of 0.11 means we cannot reject the null hypothesis.

3c) Provide a Cook's distance plot and check if there are any influential observations.

```
plot(cooks.distance(model3), type="h")
points(cooks.distance(model3), pch=21, col="blue", bg=2)
summary(cooks.distance(model3))
```

```
##      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
```

```
abline(h=.0285)  #3*iqr +3rd quartile
```




```
which(cooks.distance(model3)>.0285)
```

```
##   8   9 104 111 132 175 190
##   8   9 104 111 132 175 190
```

Observations (8,9,104,111,132,175,190) are possible influential observations.

3d) New model and employ stepwise regression.

```
model3d = glm(type~npreg+glu+ped+age+I(age*age)+I(ped*age)+I(glu*age)+I(glu*ped), data=ScaledDat
a, family="binomial")
summary(model3d)
```

```
##
## Call:
## glm(formula = type ~ npreg + glu + ped + age + I(age * age) +
##     I(ped * age) + I(glu * age) + I(glu * ped), family = "binomial",
##     data = ScaledData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4261  -0.6662  -0.3774   0.5783   2.4307
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.66212    0.38910  -1.702  0.08882 .
## npreg         0.04019    0.06870   0.585  0.55857
## glu           1.12532    0.22353   5.034  4.8e-07 ***
## ped           0.68062    0.22412   3.037  0.00239 **
## age           1.12694    0.36797   3.063  0.00219 **
## I(age * age) -0.30021    0.20433  -1.469  0.14176
## I(ped * age)  0.63669    0.30812   2.066  0.03879 *
## I(glu * age) -0.15739    0.22894  -0.687  0.49178
## I(glu * ped)  0.06322    0.21286   0.297  0.76648
## ---
## 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: 173.64  on 191  degrees of freedom
## AIC: 191.64
##
## Number of Fisher Scoring iterations: 5
```

```
# stepwise
library(MASS)
model3d.min = glm(type~1, data=ScaledData, family="binomial")  # null model
out3d.both = step(model3d.min, scope=list(lower=model3d.min, upper=model3d))
```

```
## Start:  AIC=258.41
## type ~ 1
##
##                Df Deviance    AIC
## + glu           1   207.37 211.37
## + age           1   229.94 233.94
## + npreg         1   242.03 246.03
## + ped           1   248.11 252.11
## + I(age * age)  1   251.34 255.34
## + I(glu * age)  1   254.28 258.28
## <none>              256.41 258.41
## + I(ped * age)  1   254.90 258.90
## + I(glu * ped)  1   254.94 258.94
##
## Step:  AIC=211.37
## type ~ glu
##
##                Df Deviance    AIC
## + age           1   197.11 203.11
## + npreg         1   199.08 205.08
## + ped           1   199.26 205.26
## + I(ped * age)  1   204.27 210.27
## <none>              207.37 211.37
## + I(glu * ped)  1   206.77 212.77
## + I(age * age)  1   207.14 213.14
## + I(glu * age)  1   207.37 213.37
## - glu           1   256.41 258.41
##
## Step:  AIC=203.11
## type ~ glu + age
##
##                Df Deviance    AIC
## + ped           1   187.10 195.10
## + I(age * age)  1   189.80 197.80
## + I(ped * age)  1   191.38 199.38
## + I(glu * age)  1   194.86 202.86
## <none>              197.11 203.11
## + npreg         1   195.55 203.55
## + I(glu * ped)  1   196.14 204.14
## - age           1   207.37 211.37
## - glu           1   229.94 233.94
##
## Step:  AIC=195.1
## type ~ glu + age + ped
##
##                Df Deviance    AIC
## + I(ped * age)  1   178.56 188.56
## + I(age * age)  1   179.97 189.97
## + I(glu * age)  1   184.68 194.68
## + npreg         1   184.74 194.74
## <none>              187.10 195.10
## + I(glu * ped)  1   186.63 196.63
## - ped           1   197.11 203.11
```

```
## - age           1   199.26 205.26
## - glu           1   217.91 223.91
##
## Step:  AIC=188.56
## type ~ glu + age + ped + I(ped * age)
##
##                Df Deviance    AIC
## + I(age * age)  1   174.97 186.97
## <none>              178.56 188.56
## + I(glu * age)  1   176.87 188.87
## + npreg         1   177.16 189.16
## + I(glu * ped)  1   178.23 190.23
## - I(ped * age)  1   187.10 195.10
## - ped           1   191.38 199.38
## - age           1   195.37 203.37
## - glu           1   211.14 219.14
##
## Step:  AIC=186.97
## type ~ glu + age + ped + I(ped * age) + I(age * age)
##
##                Df Deviance    AIC
## <none>              174.97 186.97
## + I(glu * age)  1   174.08 188.08
## + npreg         1   174.48 188.48
## + I(glu * ped)  1   174.53 188.53
## - I(age * age)  1   178.56 188.56
## - I(ped * age)  1   179.97 189.97
## - ped           1   186.15 196.15
## - age           1   193.60 203.60
## - glu           1   207.55 217.55
```

The best fitting model is: $logit(P(Type = 1|glu, age, ped, ped * age, age^2))$
$= \beta_0 + \beta_1(Glu) + \beta_2(Age) + \beta_3(Ped) + \beta_4(Ped * Age) + \beta_5(Age^2) + e$