STAT 526 - Spring 2011 Olga Vitek

Homework 5 - Solution

Each part of the problems 5 points

1. Agresti 10.1 (a) and (b).

	Let Patient Die		
Suicide	Yes	No	sum
Yes	1097	90	1187
No	203	435	638
sum	1300	525	1825

(a) Test of marginal homogeneity : $H_0 : \pi_{1+} = \pi_{+1}$, $H_a : \pi_{1+} \neq \pi_{+1}$
$\delta = \pi_{+1} - \pi_{1+}$
$\hat{\delta} = p_{+1} - p_{1+} = \frac{1300}{1825} - \frac{1187}{1825} = 0.712 - 0.65 = 0.062$
$\hat{\sigma}^2\{\hat{\delta}\} = \frac{(p_{12}+p_{21})-(p_{12}-p_{21})^2}{n} = \frac{1}{1825}\{(\frac{90}{1825} + \frac{203}{1825}) - (\frac{90}{1825} - \frac{203}{1825})^2\} = 0.000086$
95% Confidence interval for δ :

$$\hat{\delta} \pm z_{\alpha/2} s\{\hat{\delta}\} = 0.062 \pm (1.96)(\sqrt{0.00086}) = 0.062 \pm 0.018 = (0.044, 0.08)$$

which does not include zero. Therefore we reject H_0, and conclude that the marginal proportions are different.

(b) McNemar Test for $H_0 : \pi_{1+} = \pi_{+1}$, $H_a : \pi_{1+} \neq \pi_{+1}$
$z_0^2 = \frac{(n_{21}-n_{12})^2}{n_{21}+n_{12}} = \frac{(203-90)^2}{203+90} = 43.6 \sim \chi_1^2 = 3.841459$, p-value$\approx 0$
Therefore we reject H_0, and conclude that the marginal proportions are different. In other words, there is strong evidence of a higher proportion of "yes" response for "let patient die".

2. Agresti 10.3

	Drug B		
Drug A	Success(1)	Failure(0)	sum
Success(1)	16	45	61
Failure(0)	22	17	39
sum	38	62	100

(a) Ignoring order, McNemar Test for $H_0 : \pi_{1+} = \pi_{+1}$, $H_a : \pi_{1+} \neq \pi_{+1}$
$z_0^2 = \frac{(n_{21}-n_{12})^2}{n_{21}+n_{12}} = \frac{(22-45)^2}{22+45} = 7.895522 \sim \chi_1^2 = 3.841459$, p-value= 0.004955733
Therefore, we reject H_0, and conclude that the marginal proportions are different. This provides evidence that the response rate of successes is higher for drug A.

(b) Pearson χ^2 statistic= $\sum_{ij} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(25-19.33)^2}{19.33} + \frac{(10-15.67)^2}{15.67} + \frac{(12-17.67)^2}{17.67} + \frac{(20-14.33)^2}{14.33} = 7.8 \sim \chi_1^2 = 3.841459$, p-value= 0.005224623
Therefore we conclude that success rate differ for the two treatments.

	Treatment that is better		
Order	First	Second	sum
A, then B	25(19.33)	10(15.67)	35
B, then A	12(17.67)	20(14.33)	32
sum	37	30	67

Table 1: the values in the parentheses are expected values

3. *[Methods qualifying exam, ???: use paper and pencil.]* Results from a Copenhagen housing condition survey are compiled in an R data frame `housing1` consisting of the following components:

   Infl Influence of renters on management: Low, Medium, High.
   Type Type of rental property: Tower, Atrium, Apartment, Terrace.
   Cont Contact between renters: Low, High.
   Sat Highly satisfied or not: two columns of counts.

   A model is fitted to the data using the following commands.

```
fit <- glm(Sat~Infl+Type+Cont,family=binomial,data=housing1)
```

   Part of the results are summarized below (`summary(fit)`).

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.6551     0.1374  -4.768 1.86e-06 ***
InflMedium     0.5362     0.1213   4.421 9.81e-06 ***
InflHigh       1.3039     0.1387   9.401  < 2e-16 ***
TypeApartment -0.5285     0.1295  -4.081 4.49e-05 ***
TypeAtrium    -0.4872     0.1728  -2.820  0.00480 **
TypeTerrace   -1.1107     0.1765  -6.294 3.10e-10 ***
ContHigh       0.3130     0.1077   2.905  0.00367 **

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 166.179  on 23  degrees of freedom
Residual deviance:  27.294  on 17  degrees of freedom
AIC: 146.55
```

   (a) Write the assumptions of the model, and the expression of the log-likelihood.

   Answer:
   Assumptions include:
   Y_i are independent independent Bernoulli random variables, $i = 1, \ldots, 24$.
   $P(Y_i = 1) = 1 - P(Y_i) = \pi_i$;
   $E\{Y_i\} = \pi_i$, where $\pi_i = \frac{\exp(X_i'\beta)}{\exp(X_i'\beta)+1}$.

   The log-likelihood function is

$$l(\pi|y_1, \ldots, y_{24}) = \ln \prod_{i=1}^{24} \pi_i^{y_i} (1 - {\pi_i}^{1-y_i}) = \sum_{i=1}^{24} \ln \frac{\pi_i}{1-\pi_i} + \sum_{i=1}^{24} \ln(1-\pi_i)$$

(b) According to the fitted model, what percentage of renters, who have low influence on management, live in apartment, and have high contact between neighbors, are highly satisfied?

Answer:
The logit for these values of the covariates is

$$\hat{\eta} = -0.6551 - 0.5285 + 0.3130 = -0.8706$$

Therefore, $\hat{p} = e^{-0.8706}/(1 + e^{-0.8706}) = 0.2951$.

(c) Do people who live in apartments have a significantly different probability of satisfaction than people who live in atriums? The correlation between the respective coefficients is 0.494.

Answer:
We test
H_0 : Pr{ `Satisfaction` | `Type=Apartment` } = Pr{`Satisfaction` | `Type=Atrium`}
against
H_a : Pr{ `Satisfaction` | `Type=Apartment` } $\neq$ Pr{`Satisfaction` | `Type=Atrium`}
when all other predictors are help fixed.

This translates into testing
H_0 : $\beta_{\text{TypeApartment}} = \beta_{\text{TypeAtrium}}$, or equivalently,
H_0 : $\beta_{\text{TypeApartment}} - \beta_{\text{TypeAtrium}} = 0$

The test statistic is

$$z = \frac{-0.5285 - (-0.4872) - 0}{\sqrt{0.1295^2 + 0.1728^2 - 2(0.1295)(0.1728)(0.494)}} = -0.264$$

Since $|z| < z^{1-0.05/2} = 1.96$, we fail to reject H_0.

(d) Estimate the odds ratio of high satisfaction for groups with high contact among neighbors over groups with low contact among neighbors, using a 95% confidence interval.

Answer:
The odds ratio $OR = e^{0.3130} = 1.367522$.
The 95% CI for the log-odds ratio is $0.313 \pm 1.96(0.1077) = (0.1019, 0.5241)$.
Therefore, the CI for the odds ratio is $(e^{0.1019}, e^{0.5241}) = (1.1073, 1.6889)$.

4. *[Methods qualifying exam, August 2005: use paper and pencil.]* A sample of elderly people was given a psychiatric examination to determine whether symptoms of senility were present. Other measurements taken at the same time included the score on a subset of the Wechsler Adult Intelligence Scale (WAIS). The data are shown below.

x	9	13	6	8	10	4	14	8	11	7	9	7	5	14	13	16	10	12
s	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
x	11	14	15	18	7	16	9	9	11	13	15	13	10	11	6	17	14	19
s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x	9	11	14	10	16	10	16	14	13	13	9	15	10	11	12	4	14	20
s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

A linear logistic regression model is tted to the data,

$$log\frac{p}{1-p} = \alpha + \beta x$$

where $p = P(y = 1)$, with $\hat{\alpha} = 2.404$, $\hat{\beta} = -.3235$, and a deviance of 51.017.

(a) For a person with WAIS score $x = 10$, what is the estimated probability that the person has symptoms of senility.

Answer:
$log\frac{p}{1-p} = 2.404 - 0.3235(10)$
$\hat{p}_{x=10} = (1 + e^{-2.404+0.3235(10)})^{-1} = 0.3034$

(b) The standard errors of $\hat{\alpha}, \hat{\beta}$ are given by $s\{\hat{\alpha}\} = 1.1918$, $s\{\hat{\beta}\} = .1140$, and the correlation between $\hat{\alpha}, \hat{\beta}$ is estimated to be -.96. Obtain a approximate 95% condence interval for the sensility probability of a person with a WAIS score $x = 10$.

Answer:
$\eta = log\frac{p}{1-p} = \hat{\alpha} + \hat{\beta}x = 2.404 - 0.3235(10) = -0.831$
$\widehat{Var}(\eta) = \begin{bmatrix} 1 & 10 \end{bmatrix} \begin{bmatrix} 1.1918^2 & -0.96 \cdot 1.1918 \cdot 0.114 \\ -0.96 \cdot 1.1918 \cdot 0.114 & 0.114^2 \end{bmatrix} \begin{bmatrix} 1 \\ 10 \end{bmatrix} = 0.1114$
95% Confidence interval for η : $\eta \pm z_{0.975}\sqrt{\widehat{Var}(\eta)} = -0.831 \pm (1.96)\sqrt{0.1114} = (-1.485, -0.1768)$
95% Confidence interval for $p_{x=10}$: $((1 + e^{1.485})^{-1}, (1 + e^{0.1768})^{-1}) = (0.1846, 0.456)$

(c) Assuming $\beta = 0$, t the constant model by estimating α, and obtain the deviance of the t.

Answer:
If $\beta = 0$, $log\frac{p}{1-p} = \alpha$, $\hat{p} = \frac{14}{54} = 0.2593$, So, $\hat{\alpha} = log\frac{\hat{p}}{1-\hat{p}} = -1.049822$
Deviance $= -2logL(\hat{p}) = -2log[\hat{p}^{14}(1-\hat{p})^{54-14}] = -2[14log(0.2593) + 40log(0.7407)] = 61.806$

(d) Test the hypothesis that $\beta = 0$ using the likelihood ratio test.

Answer:
$H_0 : \beta = 0$, $H_a : \beta \neq 0$
$G^2 = 2[logL(full) - logL(reduced)] = Deviance(reduced) - Deviance(full) = 61.806 - 51.017 = 10.789 \sim \chi^2_1 = 3.841459$, p-value= 0.00102
Therefore reject H_0 and conclude that β is not zero, which means that constant model is not true and accept the two-variable model with x and the intercept.

5. *[Methods qualifying exam, August 2010: use paper and pencil.]* When modeling a binary response in logistic regression, input data can have two forms: (1) individual observations, where the response records the status of "failure" or "success", and (2) grouped data, where the response records the number of experimental units with the same covariate pattern, and the corresponding number of "successes".

For the same data set, would a logistic regression analysis of each type of input data yield the same parameter estimates? Would they yield the same deviance and test of goodness-of-fit? Why or why not?

Answer:

Both types of input data will yield the same parameter estimates, because they involve the same likelihood function. ($\pi_i = (1 + e^{-X\beta})^{-1}$) However they will yield different deviances, because deviance compares the model fit to the saturated model, and the saturated model is different for the two types of data.

6. *Data analysis: Adapted from Faraway, p. 52, Problem 2.* The dataset `wbca` from the library `Faraway` comes from a study of breast cancer in Wisconsin. There are 681 cases of potentially cancerous tumors, of which 238 are actually malignant. Malignant tumors are traditionally determined using a surgically invasive procedure. Our goal is to determine whether a new procedure called the needle aspiration, which draws only a small sample of tissue, could be effective at determining tumor status.

   (a) Split the data into two parts: assign every third observation to a test set, and the remaining two thirds to the training set. Parts (a) -(i) below will only use the training set. Use the training set to fit a binomial regression with `class` as response, and the other nine variables as predictors.

   Answer:

   The binomial model is fitted with all the 9 predictors:

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_{\texttt{Adhes}}\texttt{Adhes} + \ldots + \beta_{\texttt{USize}}\texttt{USize}$$

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  12.0244     2.0462   5.876 4.19e-09 ***
Adhes        -0.4859     0.1555  -3.126  0.00177 **
BNucl        -0.3732     0.1292  -2.888  0.00388 **
Chrom        -0.6655     0.2536  -2.625  0.00868 **
Epith         0.1779     0.2148   0.828  0.40744
Mitos        -0.6075     0.5103  -1.190  0.23388
NNucl        -0.5168     0.1828  -2.828  0.00469 **
Thick        -0.6533     0.2044  -3.197  0.00139 **
UShap        -0.5291     0.2612  -2.026  0.04280 *
USize         0.2672     0.2320   1.152  0.24947
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 592.796  on 453  degrees of freedom
Residual deviance:  57.651  on 444  degrees of freedom
AIC: 77.65
```

   R code

```
library(faraway)
data(wbca)
select<-seq(from=3, to=nrow(wbca), by=3)
test<-wbca[select,]
```

```
train<-wbca[-select,]
reg.binomial<-glm(Class~., family=binomial, data=train)
summary(reg.binomial)
```

(b) For both null and residual deviance tests, specify the null and the alternative hypotheses, and report the conclusions if possible. (*Hint:* can we use the residual deviance test in this case?). Use the Hosmer-Lemeshow test, and compare the results.

Answer:

- the Null deviance test,
  $H_0: \log \frac{\pi_i}{1-\pi_i} = \beta_0$ i.e. $\beta_1 = \ldots = \beta_9 = 0$
  H_a : at least one $\beta_j \neq 0$, $j = 1, \cdots, 9$
  $G_0^2 = 592.796 \sim \chi^2_{453}$ The p-value $= 1.025 \times 10^{-5}$ Therefore we reject H_0, and conclude that at least some of the predictors are necessary.

  R code

  ```
  pchisq(592.796, 453, lower.tail=FALSE)
  ```

- residual deviance test,
  $H_0: \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$
  $H_a: \log \frac{\pi_i}{1-\pi_i} \neq \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$ (i.e. the saturated model).
  We cannot use the residual deviance test directly, due to the lack of replication for covariate patterns.

- The Hosmer-Lemeshow is an approximate test that can be used in this case.

  ```
        X^2         Df   P(>Chi)
  2.3534984 8.0000000 0.9682128
  ```

  Therefore fail to reject H_0 and conclude that logistic regression model fits well.

  R code

  ```
  hosmerlem(train$Class,reg.binomial$fitted.values,g=10)
  ```

  The P-value of the test $= 0.95 > 0.05$. We failed to reject the null hypothesis, therefore the model fit is appropriate.

(c) Use the full model to check for outliers, and for influential observations. Report the results.

Answer:

```
par(mfrow=c(2,2))
for (i in 1:4)
    plot(reg.binomial, which=i)
```

- Graph 1: "Predicted values" are predictions on the logit scale (i.e. $\hat{\eta}$), "Residuals" are deviance residuals. The "smoothed" pattern of the residuals is roughly horizontal, no substantial systematic deviations in predicted values is detected. Observation 244 is a potential outlier or an influential observation.
- Graph 2: Normal q-q plot of standardized deviance residuals. Observation 244 again appears as an outlier.
- Graph 3: "Predicted values" are predictions on the logit scale (i.e. $\hat{\eta}$), plotted against the (absolute value of) $\sqrt{\text{standardized deviance residuals}}$. This is essentially the same plot as Graph 1. Since these are not signed residuals, we do not expect a horizontal smoothed line here. Observation 244 is a potential outlier or an influential observation.

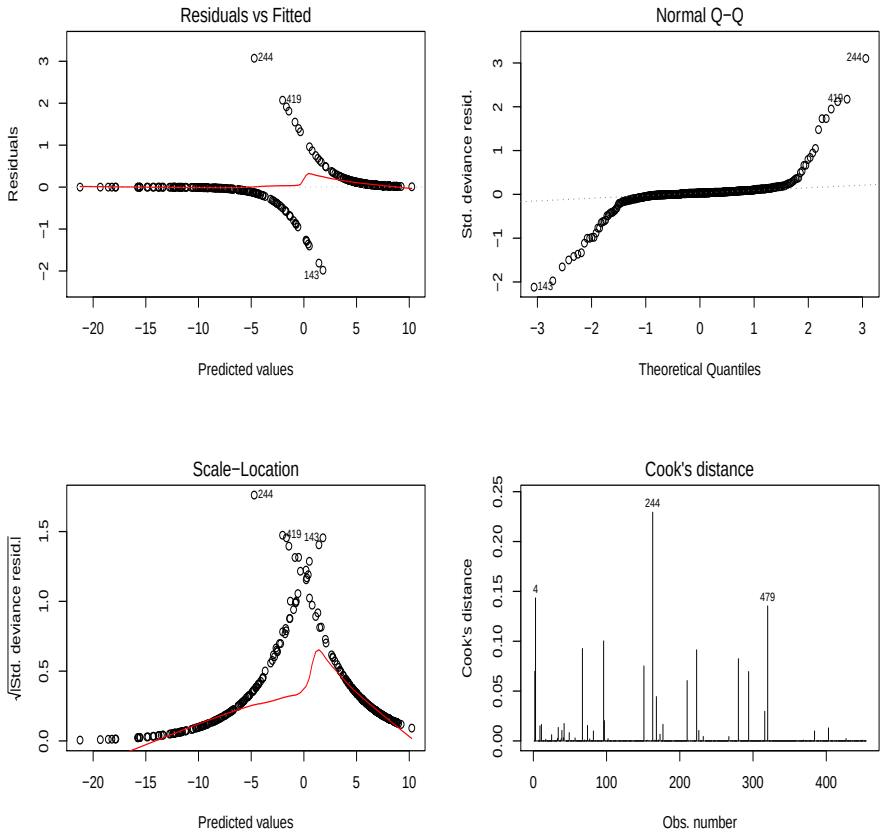


Figure 1: Check Residuals

- Cook's distance measures (an approximation of) the summary change in predicted values $\hat{\pi}$ after removing each individual observation. Observation 244 again stands out, since removal of this observation results in a large change of predicted values.

(d) Use stepwise variable selection combined with the AIC criterion to determine the best subset of predictors. (*Hint:* use `step(fullModelFit, direction="both")`).

Answer:

The stepwise variable selection procedure starts from the full model, and sequentially removes or adds predictors in the attempt to minimize the AIC criterion. In this case, the procedure first removed `Epith` and then `USize`. The resulting model is

```
Call:  glm(formula = Class ~ Adhes + BNucl + Chrom + Mitos + NNucl +
Thick + UShap, family = binomial, data = train)

Coefficients:
(Intercept)        Adhes       BNucl       Chrom       Mitos       NNucl
    11.5571      -0.4249     -0.3341     -0.5963     -0.5822     -0.4192
      Thick       UShap
    -0.6037      -0.2943
```

```
Degrees of Freedom: 453 Total (i.e. Null);  446 Residual
Null Deviance:      592.8
Residual Deviance: 59.54        AIC: 75.54
```

R code

```
reg.stepwise<-step(reg.binomial, direction="both")
```

(e) Provide the definition of overdispersion. Check for evidence of over- or under-dispersion in this dataset. Compare the model selected in (d) to the full model, while accounting for over- or under-dispersion (*Hint:* use approximate F test).

Answer :

The basic assumption of the model is $Y_i \overset{ind}{\sim} Binomial(n_i, \pi_i)$, which implies $Var(Y_i) > n_i\pi_i(1-\pi_i)$. If in the dataset $Var(Y_i) >> n_i\pi_i(1-\pi_i)$, this phenomenon is called overdispersion.

To check for evidence of overdispersion, one approach is to use the full model with family `quasibinomial`. We got

```
...
(Dispersion parameter for quasibinomial family taken to be 0.3587316)
....
```

The estimated parameter $\hat{\phi} = 0.3587316 < 1$ which indicates underdispersion (not accounting for underdispersion is less problematic than in case of overdispersion, since it will yield more conservative conclusions).

In presence of over- or underdispersion, likelihood-based tests such as χ^2 test and analysis of deviance cannot be used to compare models. An approximate F test is employed instead.

We test $H_0 : \beta_{\texttt{Epith}} = \beta_{\texttt{USize}} = 0$ against $H_a : \beta_{\texttt{Epith}} \neq 0$ or $\beta_{\texttt{USize}} \neq 0$

The F-test is constructed as

$$F = \frac{D_{reduced} - D_{full}}{df_{reduced} - df_{full}} / \hat{\phi} \overset{ass.,H_0}{\sim} F_{df_reduced - df_{full}, df_{full}}$$

It is calculated using

```
>anova(reg.binomial, reg.stepwise, test="F", dispersion=0.3588)
Analysis of Deviance Table

Model 1: Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
    UShap + USize

Model 2: Class ~ Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap
  Resid. Df Resid. Dev  Df Deviance      F  Pr(>F)
1       444     57.651
2       446     59.536  -2   -1.885 2.6271 0.07341 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The p-value of the test is 0.0734. We fail to reject the null hypothesis, while accounting for underdispersion.

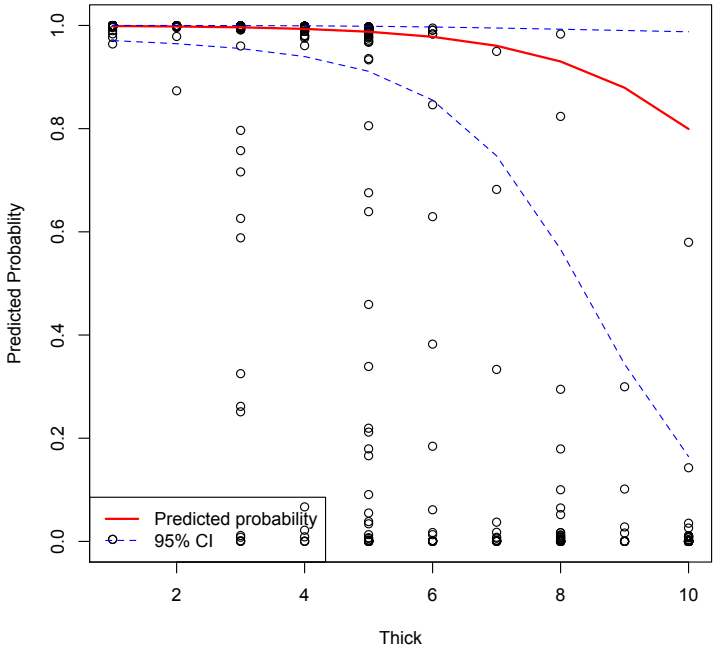
R code

```
reg.quasi<-glm(Class~., family=quasibinomial, data=train)
summary(reg.quasi)

anova(reg.binomial, reg.stepwise, test="F", dispersion=0.3587)
```

(f) Based on the logistic regression fit in (d) without overdispersion, plot the predicted probability of receiving a flu shot as a function of `Thick`, while setting the values of the remaining predictors at the median value observed in the dataset. Overlay the corresponding confidence interval and interpret the results.

Answer:



Since the coefficient of `Thick` is negative, the predicted probability $\hat{\pi}_i$ decreases with increasing values of `Thick`. The confidence intervals are wide due to the conservative Bonferroni correction for multiple comparisons. The confidence intervals are approximate, and are not guaranteed to be between 0 and 1.

R code

```
sapply(train, median) # get median
range(train$Thick) #1-10

# make a new arfiticial dataset
newdata<-data.frame(Adhes=rep(1, length=10), BNucl=rep(1, length=10), Chrom=rep(3, length=10),
  Mitos=rep(1, length=10),NNucl=rep(1, length=10), UShap=rep(2, length=10),
  Thick=seq(from=1, to=10, length=10))

# get prediction
newprediction<-predict(reg.stepwise, newdata=newdata, type="response", se.fit=TRUE)
```

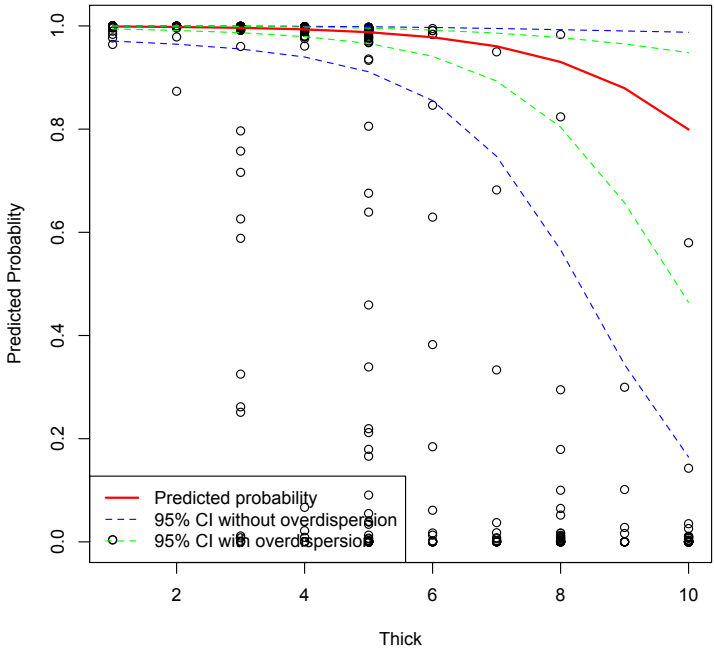
```
#plot
plot(train$Thick,reg.stepwise$fitted.values, xlab="Thick",ylab="Predicted Probablity", ylim=c(0,1))
lines(newdata$Thick, newprediction$fit, lwd=2, col="red")

#confidence interval
newprediction.link<-predict(reg.stepwise, newdata=newdata, type="link", se.fit=TRUE)
L <- newprediction.link$fit - qnorm(1-0.05/(2*10))*newprediction.link$se
U <- newprediction.link$fit + qnorm(1-0.05/(2*10))*newprediction.link$se

lines(newdata$Thick, 1/(1+exp(-L)), lty=2, col="blue")
lines(newdata$Thick, 1/(1+exp(-U)), lty=2, col="blue")
legend("bottomleft", c("Predicted probability","95\% CI"),lty=c(1,2), lwd=c(2,1), col=c("red","blue"))
```

(g) Overlay the confidence interval obtained in presence of overdispersion, and compare it to the confidence interval in (f).

Answer:



The confidence intervals adjusted for underdispersion are shown in green. Since $\hat{\phi} = 0.257132 < 1$, the confidence intervals are narrower.

R code

```
#allow overdispersion
reg.step.quasi<-glm(Class~Adhes+BNucl+Chrom+Mitos+NNucl+UShap+Thick, family=quasibinomial, data=train)

newprediction2<-predict(reg.step.quasi, newdata=newdata, type="response", se.fit=TRUE)
```

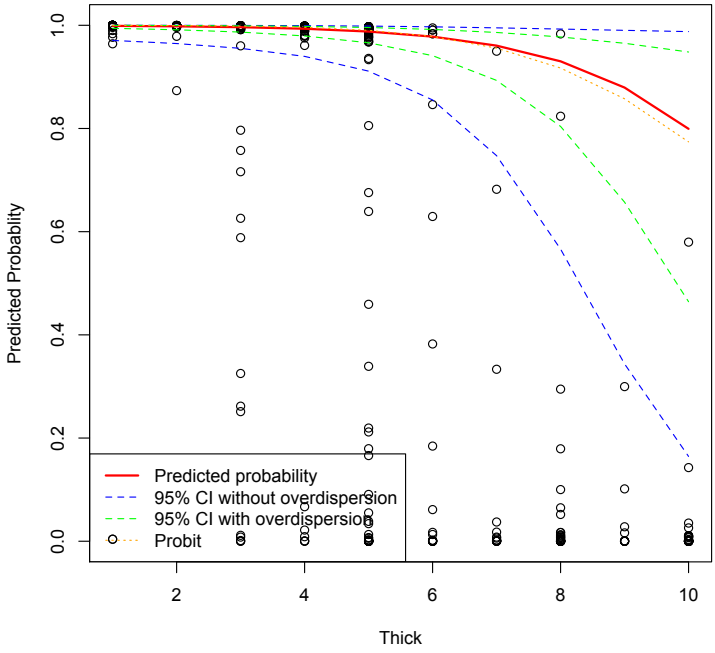
```
#add lines
L <- newprediction2.link$fit - qnorm(1-0.05/(2*10))*newprediction2.link$se
U <- newprediction2.link$fit + qnorm(1-0.05/(2*10))*newprediction2.link$se

lines(newdata$Thick, 1/(1+exp(-L)), lty=2, col="green")
lines(newdata$Thick, 1/(1+exp(-U)), lty=2, col="green")

legend("bottomleft", c("Predicted probability","95\% CI without overdispersion",
  "95\% CI with overdispersion"),lty=c(1,2,2), lwd=c(2,1,1), col=c("red","blue","green"))
```

(h) Fit the probit regression model using the same predictors as in parts (f)-(g). On the plot, overlay the predicted probabilities based on the probit regression (you do not need to plot confidence intervals this time). Discuss the differences between the two sets of curves obtained with the two link functions (if any).

Answer:



The probit link results in prediction quite similar to those with the logit link.

R code

```
reg.step.probit<-glm(Class~Adhes+BNucl+Chrom+Mitos+NNucl+UShap+Thick, family=binomial(link="probit"), dat

prediction.probit<-predict(reg.step.probit, newdata=newdata, type="response", se.fit=TRUE)
lines(newdata$Thick, prediction.probit$fit, lty=3, col="orange")
legend("bottomleft", c("Predicted probability","95\% CI without overdispersion",
  "95\% CI with overdispersion","Probit"),lty=c(1,2,2,3),
   lwd=c(2,1,1,1), col=c("red","blue","green","orange"))
```

(i) Use the model in (d) to predict the tumor status of the patients in the training set. Report confusion matrices for probability cutoffs $p = 0.5$ and $p = 0.9$. Discuss the predictive ability of the model, and the role of the cutoff.

Answer :
Probability cutoffs, $p = 0.5$

	Obs	
Pred	0	1
FALSE	156	7
TRUE	7	284

Table 2: Probability cutoffs, $p = 0.5$

Probability cutoffs, $p = 0.9$

	Obs	
	0	1
FALSE	163	14
TRUE	0	277

Table 3: Probability cutoffs, $p = 0.9$

On the training set, the model shows a relatively good predictive ability. A higher probability cutoff results in a better specificity (the number of false predictions of 'Class=1' decreased from 7 to 0), but is a lower sensitivity (the number of correct predictions of 'Class=1' decreased from 284 to 277). Therefore the choice of cutoff is a trade-off between sensitivity and specificity.

R code

```
pred1 <- predict (stepwise, train, type="response") > 0.5
table ("Pred" = pred1, "Obs" = train$Class)

pred2 <- predict (stepwise, train, type="response") > 0.9
table ("Pred" = pred2, "Obs" = train$Class)
```

(j) Plot two ROC curves: the ROC curve based on the training set, and the ROC curve based on the validation set. Report the areas under the curves. Discuss the difference between the in-sample results and the results on the validation set. Discuss the ability of the needle aspiration method to determine tumor status.

Answer:

While the area under train dataset is 0.9973, the area under test dataset is 0.9976. In this case, the areas under the ROC curves are nearly perfect (close to 1), and show a good predictive ability of the model on both training and validation dataset. In general, the area under the ROC curve on the validation set tends to be smaller than on the training set, and is a more accurate indicator of predictive ability of the model.

R code

```
library(ROCR)

# training set
score <- predict(reg.stepwise, data=train, type = "response")
pred <- prediction(score, labels = train$Class)
```

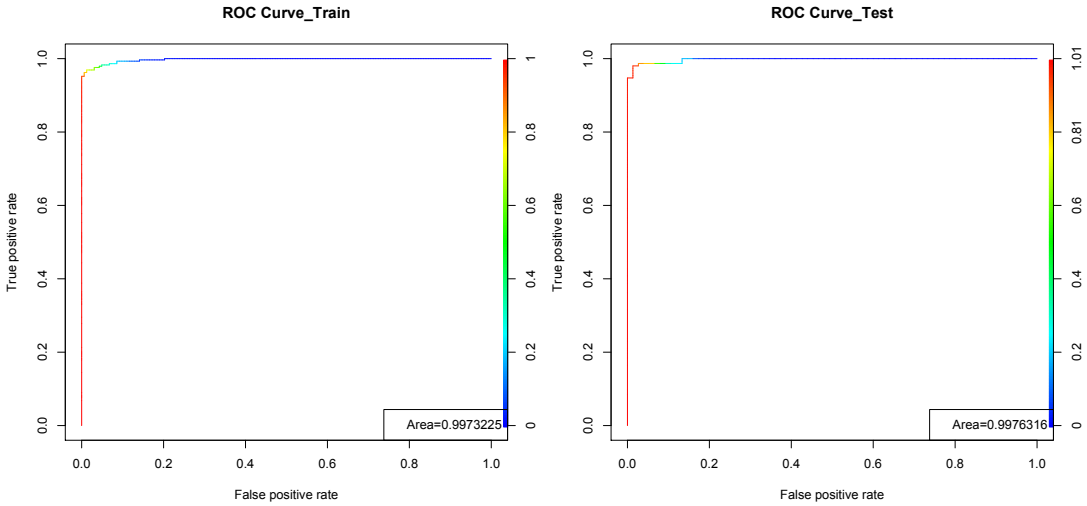


Figure 2: ROC curves for both training and validation datasets

```
perf <- performance (pred, 'tpr','fpr')
plot (perf, colorize =T, main = "ROC Curve_Train")

# area under the ROC curve
unlist(attributes(performance(pred,"auc"))$y.values)
legend("bottomright","Area=0.9973225")

# independent validation set
score2 <- predict(reg.stepwise, newdata=test, type = "response")
pred2 <- prediction(score2, labels = test$Class)
perf2 <- performance (pred2, 'tpr','fpr')
plot (perf2, colorize =T, main = "ROC Curve_Test")

# area under the ROC curve
unlist(attributes(performance(pred2,"auc"))$y.values)
legend("bottomright","Area=0.9976316")
```