MATH 523: EXAMPLES - LOGISTIC REGRESSION

LOGISTIC REGRESSION WITH A SINGLE COVARIATE

EXAMPLE: Crabs Data:- 173 observations of the shell widths of female horseshoe crabs, with response value being the number of male "satellite" crabs discovered within a given distance.

- Response Y: Indicator of number of satellites being greater than zero.
- Covariate x: Shell width

These data are available from

www.math.mcgill.ca/dstephens/523/Data/Crabs.txt

The following code performs the analysis of two competing models: let

$$\pi(x) = \Pr[Y = 1|x]$$

- `m1`: Linear

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

- `m2`: Quadratic

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x + \beta_2 x^2$$

The code also produces a plot using categorized data: the deciles of width are used to define ten categories, and the proportion of crabs that have response 0 and 1 in each category is recorded. This proportion is then plotted against the mean width for the category.

```
Crabs<-read.table("Crabs.txt",header=T)

m1<-glm(I(satell>0)~width,data=Crabs,family=binomial)
m2<-glm(I(satell>0)~width+I(width^2),data=Crabs,family=binomial)

ncat<-10
wcat<-cut(Crabs$width,quantile(Crabs$width,prob=c(0:ncat)/ncat),include.lowest=T)
wi<-as.numeric(wcat)

wm<-rep(0,ncat)
for(k in 1:ncat){
        wm[k]<-mean(Crabs$width[wi == k])
}

wtab<-table(wcat,I(Crabs$satell>0))

prop<-wtab/apply(wtab,1,sum)

plot(wm,prop[,2],ylim=range(0,1),
xlab="x=Width",ylab=expression(pi(x)),pch=19,xlim=range(22,32))
xv<-seq(22,32,by=0.01)
lp1<-m1$coef[1]+xv*m1$coef[2]
p1<-exp(lp1)/(1+exp(lp1))
lines(xv,p1)
lp2<-m2$coef[1]+xv*m2$coef[2]+xv^2*m2$coef[3]
p2<-exp(lp2)/(1+exp(lp2))
lines(xv,p2,lty=2)
legend(28,0.4,c("Linear","Quadratic"),lty=c(1,2))
```

The categorized data are as follows

	Y	
Category	0	1
[21.0,23.7]	14	5
(23.7,24.5]	10	8
(24.5,25.1]	5	10
(25.1,25.7]	10	9
(25.7,26.1]	5	11
(26.1,26.7]	7	11
(26.7,27.4]	4	12
(27.4,28.2]	4	12
(28.2,29.0]	3	19
(29.0,33.5]	0	14

Results:

```
> summary(m1)

Call: glm(formula = I(satell > 0) ~ width, family = binomial, data = Crabs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0281  -1.0458   0.5480   0.9066   1.6942

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
width         0.4972     0.1017   4.887 1.02e-06 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom
AIC: 198.45


> summary(m2)

Call: glm(formula = I(satell > 0) ~ width + I(width^2), family = binomial,
    data = Crabs)

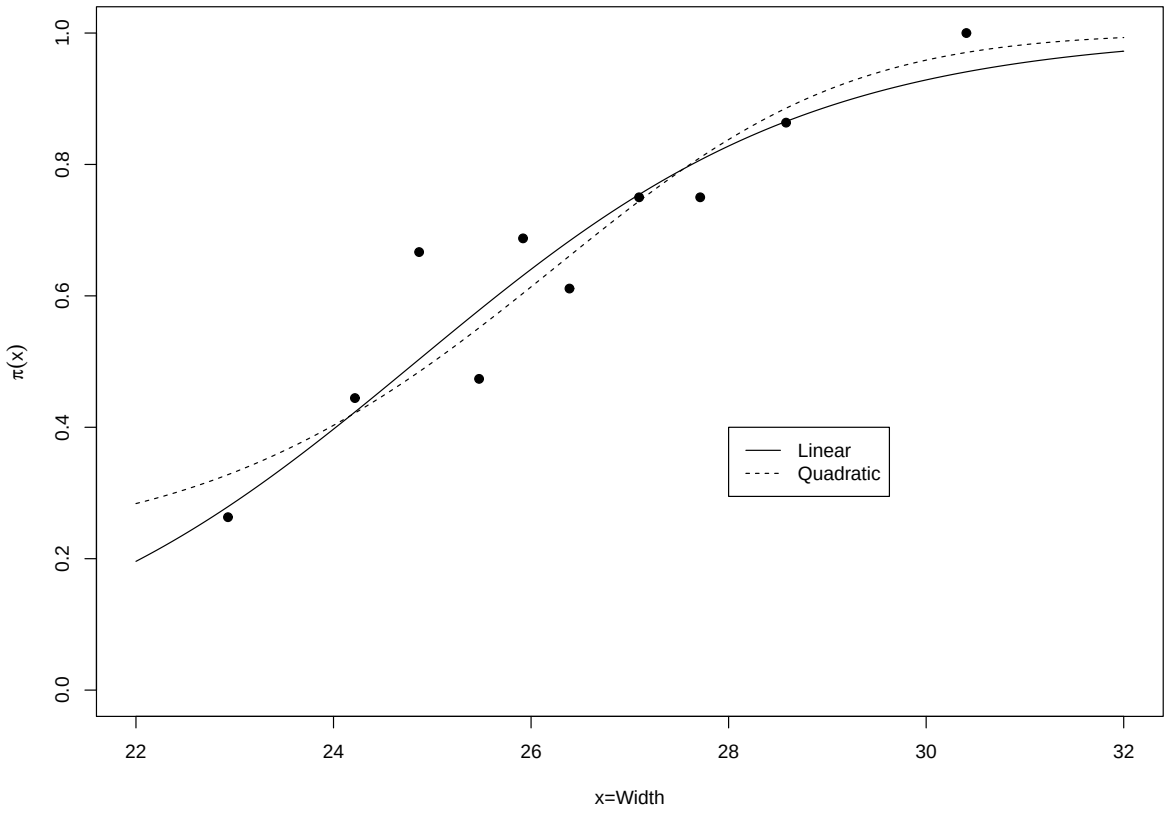
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1185  -1.0441   0.5067   0.9483   1.5406

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 14.59156   30.22371   0.483    0.629
width       -1.59572    2.35195  -0.678    0.497
I(width^2)   0.04047    0.04566   0.886    0.375

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 193.63  on 170  degrees of freedom
AIC: 199.63

Number of Fisher Scoring iterations: 5
```



Comparing the two models via their deviances, or by inspecting the coefficient of the quadratic term in model `m2`, it seems that the linear model does not provide an inferior fit compared to the quadratic model.

```
> anova(m1,m2)
Analysis of Deviance Table

Model 1: I(satell > 0) ~ width
Model 2: I(satell > 0) ~ width + I(width^2)
  Resid. Df Resid. Dev Df Deviance
1       171     194.45
2       170     193.63  1  0.82542
```

Recall, however, that comparing models via their deviance must be done with care for binary response data. Therefore the more reliable test is based on the Wald statistic

$$W_n = \frac{\widehat{\beta}_2}{\text{s.e.}(\widehat{\beta}_2)}$$

where $W_n^2 \div \chi_1^2$ for large n.

ANALYSIS BASED ON FACTOR PREDICTORS

EXAMPLE: AIDS and AZT:- The following data are from a study on the effects of drug AZT in slowing the development of AIDS symptoms. In the study, 338 veterans with compromised immune systems after infection with the AIDS virus were randomly assigned either to receive AZT immediately, or to wait until their T cells showed severe immune weakness. The table below cross-classifies the veterans race, whether they received AZT immediately, and whether they developed AIDS symptoms during the 3-year study.

- Response Y: Indicator of number of developing AIDS symptoms.
- Factor `azt`: Recipient of AZT therapy (Yes, No)
- Factor `race`: Race of subject (White, Black)

		Symptoms	
Race	AZT	Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

We have a $2 \times 2 \times 2$ table of counts; via the randomization, the sampling design effectively fixes the number of subjects receiving/not receiving AZT (we have 107+63=170 receiving AZT, 113+55=168 not receiving AZT). However, we can fit logistic regression models to the individual response data, taking `azt` and `race` as factor predictors taking two levels.

We may fit five models to the binary response data

```
m0 Null
m1 azt
m2 race
m3 azt+race
m4 azt*race = azt+race+azt.race
```

Code to do this in `R` is as follows:

```
AZT.counts<-c(14,93,32,81,11,52,12,43)
AZT.acounts<-c(sum(AZT.counts[1:2]),sum(AZT.counts[3:4]),
               sum(AZT.counts[5:6]),sum(AZT.counts[7:8]))
AZT.aids<-as.factor(rep(c("Y","N","Y","N","Y","N","Y","N"),AZT.counts))
AZT.azt<-as.factor(rep(c("Y","N","Y","N"),AZT.acounts))
AZT.race<-as.factor(rep(c("W","B"),c(sum(AZT.counts[1:4]),sum(AZT.counts[5:8]))))

table(AZT.azt,AZT.aids,AZT.race)

AZT<-data.frame(AZT.race,AZT.azt,AZT.aids)
names(AZT)<-c("race","azt","aids")

m0<-glm(aids~1,data=AZT,family=binomial)
m1<-glm(aids~azt,data=AZT,family=binomial)
m2<-glm(aids~race,data=AZT,family=binomial)
m3<-glm(aids~azt+race,data=AZT,family=binomial)
m4<-glm(aids~azt*race,data=AZT,family=binomial)
```

Results:

```
> anova(m0,m1,m2,m3,m4)
Analysis of Deviance Table

Model 1: aids ~ 1
Model 2: aids ~ azt
Model 3: aids ~ race
Model 4: aids ~ azt + race
Model 5: aids ~ azt * race
  Resid. Df Resid. Dev Df Deviance
1       337     342.12
2       336     335.19  1   6.9293
3       336     342.02  0  -6.8338
4       335     335.15  1   6.8709
5       334     333.77  1   1.3835
```

The immediate conclusion of this analysis is that the interaction term `azt.race` can be omitted. The results for the additive model `azt+race` are

```
> summary(m3)

Call: glm(formula = aids ~ azt + race, family = binomial, data = AZT)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7854  -0.7668  -0.5694  -0.5549   1.9733

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.07357    0.26294  -4.083 4.45e-05 ***
aztY        -0.71946    0.27898  -2.579  0.00991 **
raceW        0.05548    0.28861   0.192  0.84755
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 342.12  on 337  degrees of freedom
Residual deviance: 335.15  on 335  degrees of freedom
AIC: 341.15

Number of Fisher Scoring iterations: 4
```

This indicates a strong effect of AZT therapy (significant coefficient for `aztY`, estimate (s.e.) is -0.71946 (0.27898)). Comparing the models `m1` and `m3` yields an assessment of whether there is a need for the factor `race` to be included:

```
> anova(m1,m3)
Analysis of Deviance Table

Model 1: aids ~ azt
Model 2: aids ~ azt + race
  Resid. Df Resid. Dev Df Deviance
1       336     335.19
2       335     335.15  1 0.037084
```

which, in conjunction with the magnitude of the coefficient of `raceW` in the `m3` analysis leads us to conclude that `race` does not need to be included in the model.

Finally, the models `m1` yields an assessment of the magnitude of the AZT effect ignoring `race`, that is, pooling the two race groups.

```
> summary(m1)

Call:glm(formula = aids ~ azt, family = binomial, data = AZT)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7793  -0.7793  -0.5640  -0.5640   1.9580

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0361     0.1755  -5.904 3.54e-09 ***
aztY         -0.7218     0.2787  -2.590  0.00961 **
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 342.12  on 337  degrees of freedom
Residual deviance: 335.19  on 336  degrees of freedom
AIC: 339.19

Number of Fisher Scoring iterations: 4
```

Interpretation: The additive main effects model `azt+race` models the log odds of developing AIDS symptoms as follows: coding `aids=Y` as 1, `aids=N` as 0, the models seek to explain

$$\log\left(\frac{\pi_{1jk}}{\pi_{0jk}}\right) = \log\left(\frac{\pi_{1jk}}{1-\pi_{1jk}}\right)$$

and we have

$$\log\left(\frac{\pi_{1jk}}{\pi_{0jk}}\right) = \beta + \beta_j^{(\texttt{azt})} + \beta_k^{(\texttt{race})} = \begin{cases} \beta & j=1,k=1 \quad (\texttt{azt=N}, \texttt{race=black}) \\ \beta + \beta_2^{(\texttt{azt})} & j=2,k=1 \quad (\texttt{azt=Y}, \texttt{race=black}) \\ \beta + \beta_2^{(\texttt{race})} & j=1,k=2 \quad (\texttt{azt=N}, \texttt{race=white}) \\ \beta + \beta_2^{(\texttt{azt})} + \beta_2^{(\texttt{race})} & j=2,k=2 \quad (\texttt{azt=Y}, \texttt{race=white}) \end{cases}$$

with estimates (from the `summary(m3)` table from the previous page),

$$\widehat{\beta} = -1.07357 \qquad \widehat{\beta}_2^{(\texttt{azt})} = -0.71946 \qquad \widehat{\beta}_2^{(\texttt{race})} = 0.05548$$

Rearranging the formulae above, we have that

$$\log\left(\frac{\pi_{121}/\pi_{021}}{\pi_{111}/\pi_{011}}\right) = \beta_2^{(\texttt{azt})} \qquad \log\left(\frac{\pi_{122}/\pi_{022}}{\pi_{111}/\pi_{011}}\right) = \beta_2^{(\texttt{azt})} + \beta_2^{(\texttt{race})} \qquad \log\left(\frac{\pi_{122}/\pi_{022}}{\pi_{121}/\pi_{021}}\right) = \beta_2^{(\texttt{race})}$$

Thus, we conclude that as the coefficient $\beta_2^{(\texttt{race})}$ has estimate (standard error) 0.05548 (0.28861), with z value 0.192 and p-value 0.84755, there is no evidence to support that the odds-ratio comparing the `azt=Y` and `azt=N` levels differs across the races.