Department of Statistics

STATS 762 : Topics in Regression Modelling
Extra Term Test: October 2006

Instructions: Attempt both questions. Time: 60 minutes.

Question 1. Answer each of the following parts. You should write no mote than a sentence or two for each part. Each part is worth 2 marks.

a) What is the relationship between the probability of success and the covariates in the logistic regression model?

b) Under what circumstances can we use the residual deviance as a goodness-of-fit measure in logistic regression?

c) Calculate the odds ratio for the following 2x2 table which gives the results of car accidents in Florida, classified by seat belt use and type of injury.

	Injury type	
Seat belts	Fatal	Non-fatal
Not used	1,601	162,527
used	510	412,368

d) Does the table in (c) provide evidence that seat belts reduce fatal injury?

e) Suppose we fit a Poisson regression model with a single continuous covariate x . The estimated coefficients are $\hat{\beta}_0$ = -2 and $\hat{\beta}_1$ =1. What is the estimated mean count when x = 2?

f) What problem with linear regression can be remedied by using weighted least squares?

g) Consider a simple linear regression model with a single continuous covariate x and response y. Draw a scatterplot that shows an observation with high leverage but low influence.

h) What do we use gam plots for?

CONTINUED

i) Name 4 criteria that can be used to select a good model in the "all possible regressions" approach to model selection.

j) Suppose we have a contingency table with 3 factors A, B and C. How can you test if the odds ratios between A and B conditional on C depend on the levels of C?

Question 2. Minor faults occur in an industrial process, and it is thought that they are related to the purity of the raw materials used. We did an experiment to explore a possible solution to the process. Twenty-two batches of raw material were made, and divided into sub-batches. One of the sub-batches was use in the standard process, and the other in a modified process. For each of these 44 process runs, we recorded if there was a fault in the process or not. The purity of each batch was also recorded. (Each sub-batch made from the same batch had the same purity.)

The following data were obtained.

Purity Index	Standard process	Modified process	Purity Index	Standard process	Modified process
7.2	NF	NF	6.5	NF	F
6.3	F	NF	4.9	F	F
8.5	F	NF	5.3	F	NF
7.1	NF	F	7.1	NF	F
8.2	F	NF	8.4	F	NF
4.6	F	NF	8.5	NF	F
8.5	NF	NF	6.6	F	NF
6.9	F	F	9.1	NF	NF
8.0	NF	NF	7.1	F	NF
8.0	F	NF	7.5	NF	F
9.1	NF	NF	8.3	NF	NF

NB: NF = no fault, F = fault.

(a) Explain how you would enter the data into a data frame **`batches.df`** for analysis in R, having variables **`y`** (having value 1 if there is no fault, 0 if there is a fault), **`purity`** and **`process`** (a factor having two values, "modified" and "standard"). Write down the first few lines of the resulting data frame. [7 marks]

CONTINUED

(b) The regression output on the next page was obtained.

```
> summary(glm(y~process+purity, family=binomial,
data=batches.df))

Call:
glm(formula = y ~ process + purity, family = binomial, data
= batches.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8622  -1.0687   0.6321   0.9187   1.5381

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.5963     2.0778  -1.731   0.0835 .
processstandard  -0.8645     0.6717  -1.287   0.1981
purity            0.6042     0.2838   2.129   0.0333 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.534  on 43  degrees of freedom
Residual deviance: 52.812  on 41  degrees of freedom
AIC: 58.812

Number of Fisher Scoring iterations: 4
```

Do you think that the modified method has a higher probability of success for a given purity than the standard method? Give a reason for your answer. [6 marks]

(c) Use this model to predict the probability of getting a fault for the standard method when the purity is 8 units. [7 marks]