Comments on Assignment 4

Q1

Contingency Tables Descriptive examination of the relation between categorical variables (e.g. diagnostic group and discharge status) is best done in terms of proportions or percents, "conditioned" on the explanatory variable. Thus, if you create a cross-tabulation with an explanatory variable as the row variable and outcome as the column variable, you should provide row proportions (or percents).

Goodness of Fit Test The residual deviance is a only valid as goodness of fit test if the data has been recorded in replicate sets i.e. where each row in the data table records the number of successes and failures for a group of replicate subjects.

Q2

d. The estimated probabilities are all functions of the same estimated coefficients, and so are not independent. Thus the delta method should be applied to the vector of estimated probabilities to obtain it's asymptotic covariance matrix. If we let V donote this matrix and let $\underset{\sim}{1}$ represent a column of 1's, the variance of the sum of probabilities takes the form $\underset{\sim}{1}^t V \underset{\sim}{1}$, which when divided by n^2 provides the variance of the mean.

Q3

Part a.

This question was intended to get you to think "outside the box" about prediction for binary outcomes. The simplest way to examine the accuracy of the predicted risk is to look at observed and expected deaths (obtained taking the mean of the risk estimates) in risk categories.

	observed	expected
[0,10]	51/556(9.2%)	4.4%
(10,20]	51/320(15.9%)	15%
(20,30]	47/247(19.0%)	25%
(30,40]	44/175(25.1%)	34.5%
(40,50]	62/171(36.3%)	45.3%
(50,60]	47/105(44.8%)	56.1%
(60,70]	66/118(55.9%)	65.7%
(70,80]	59/96(61.5%)	75.8%
(80,90]	80/109(73.4%)	85.6%
(90,100]	79/96(82.3%)	94.8%

From the above table we see that the observed death percentages are consistently lower than the expected, indicating that risk is over-estimated.

Though the question doesn't ask for this, it is informative to recalibrate the given risk values to better match the data using a "simple linear" logistic regression. Here are the results.

	observed	expected
[0,10]	26/354(7.3%)	8.9%
(10,20]	95/663(14.3%)	13.8%
(20,30]	72/279(25.8%)	24.3%
(30,40]	62/173(35.8%)	35.4%
(40,50]	47/105(44.8%)	45.9%
(50,60]	66/118(55.9%)	55.6%
(60,70]	64/106(60.4%)	65.8%
(70,80]	111/143(77.6%)	75.5%
(80,90]	43/52(82.7%)	81.5%
(90,100]	0/0(NaN%)	NA%

Part b.

If we fit a spline we see the same distinct improvement over the "raw" risk score (but a minimal improvement over the linear fit).

	observed	expected
[0,10]	28/409(6.8%)	8.3%
(10,20]	86/563(15.3%)	14.4%
(20,30]	79/324(24.4%)	24.5%
(30,40]	65/183(35.5%)	34.9%
(40,50]	46/107(43.0%)	45.5%
(50,60]	70/114(61.4%)	55.3%
(60,70]	58/98(59.2%)	65.8%
(70,80]	94/123(76.4%)	75.5%
(80,90]	60/72(83.3%)	82.1%
(90,100]	0/0(NaN%)	NA%

Part c.

The last part indicates an essential weakness of the AUC as an indicator of predictive strength. Mathematically, the construction of the ROC curve doesn't change if we transform the underlying score - it's numerical values don't matter, only the ordering it induces in the observations.