5. (1.75 points)

An analyst has fit several different variations of a logistic GLM to a dataset containing 1,000 records of fraudulent claims and 9,000 records of legitimate claims. For each model variation listed below, draw a quintile plot based on the training data. Label the axes and identify each data series.

i. A saturated model

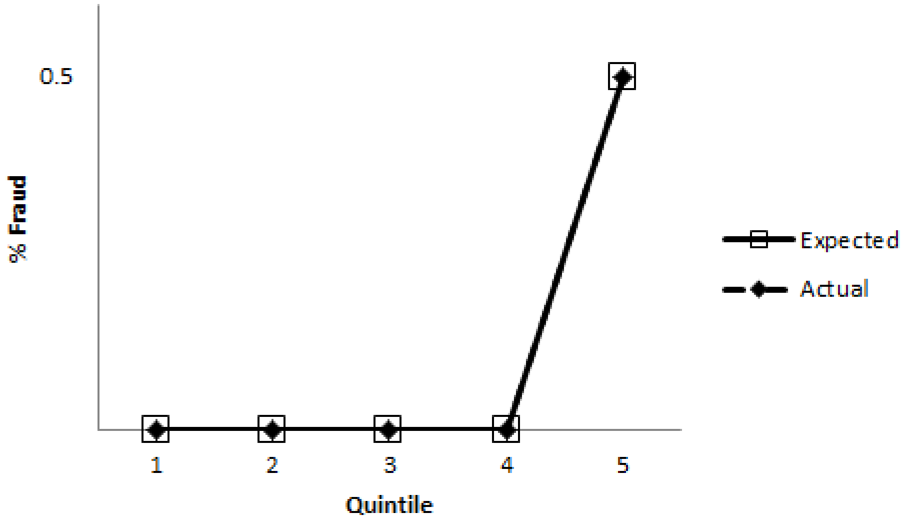
ii. A null model

iii. A model that could be used in practice

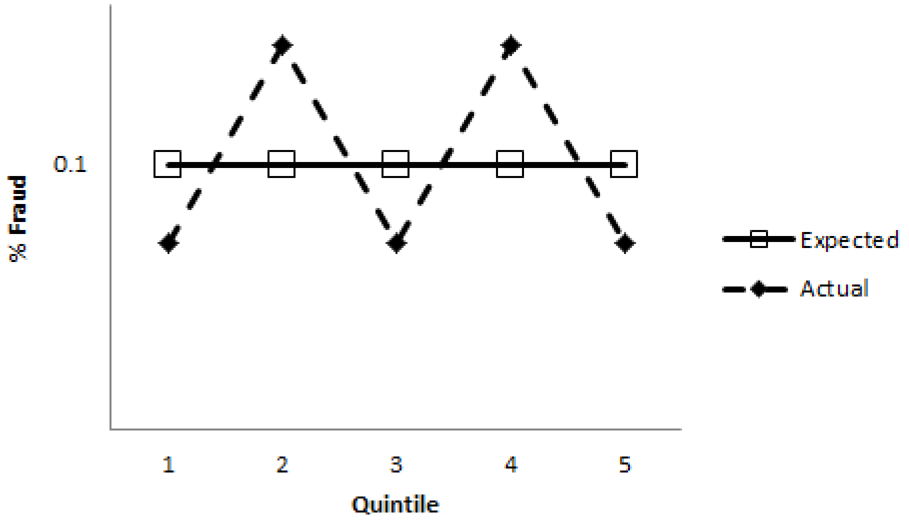
CONTINUED ON NEXT PAGE

| QUESTION 5 | |
|---|---|
| TOTAL POINT VALUE: 1.75 | LEARNING OBJECTIVE(S): A4c |
| SAMPLE ANSWERS | |

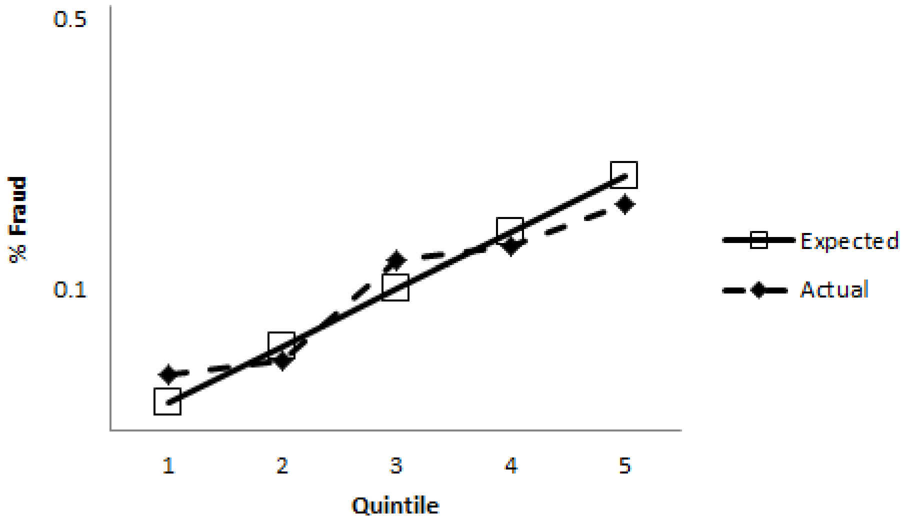
i. Saturated Model



ii. Null Model



iii. Model in Practice



## EXAMINER'S REPORT

Candidates were expected to create 3 separate quintile plots, which demonstrated how saturated, null and practical models compared to actuals for a logistic model. The response of each model was expected to identify the percentage of fraudulent (or alternatively percentage of non-fraudulent) records that were identified with the model.

Common mistakes included:

- Not plotting actuals on the graph. The purpose of a quintile plot is to compare how well predicted values compare to actual values
- Plotting the same actuals on each graph. Since the records are ordered by predicted values, the records in each bucket change for each graph. Thus, actuals are not the same for each graph.
- Plotting all models on one graph. Actuals are not the same for each model.
- Sorting records by Actuals. Quintile plots are sorted by predicted values from smallest to largest value. Thus the predicted values must be monotonically increasing. Actuals need not be.
- Plotting loss ratios, pure premiums, loss costs, or losses. The response for this logistic model is either % fraudulent or % non-fraudulent claims.
- Not labeling the axes or not including scale. The average % fraud of all graphs is 0.1.
- A null model plots the grand mean of the data. The mean of this dataset is 1,000/10,000 = 0.1, not 1/9, 0.5 or 0. A null model does not mean there is no prediction (i.e. predicted value of 0), nor does it mean that half the records are predicted to be fraud (i.e. predicted value of 0.5).
- A saturated model means there is an equal number of predictors as there are records in the dataset, not that all variables are used. Thus a saturated model would perfectly predict every historical outcome in the training data.
- Not plotting the practical model between the saturated and null models. Most common was for quintile 5 of the practical model to have a larger value than the saturated model.

- Plotting something other than a quintile plot. Partial credit was given for other graphs, including, but not limited to, ROC, Actual vs. Predicted, and QQ plots, as long as an understanding of saturated, null and practical models was demonstrated.