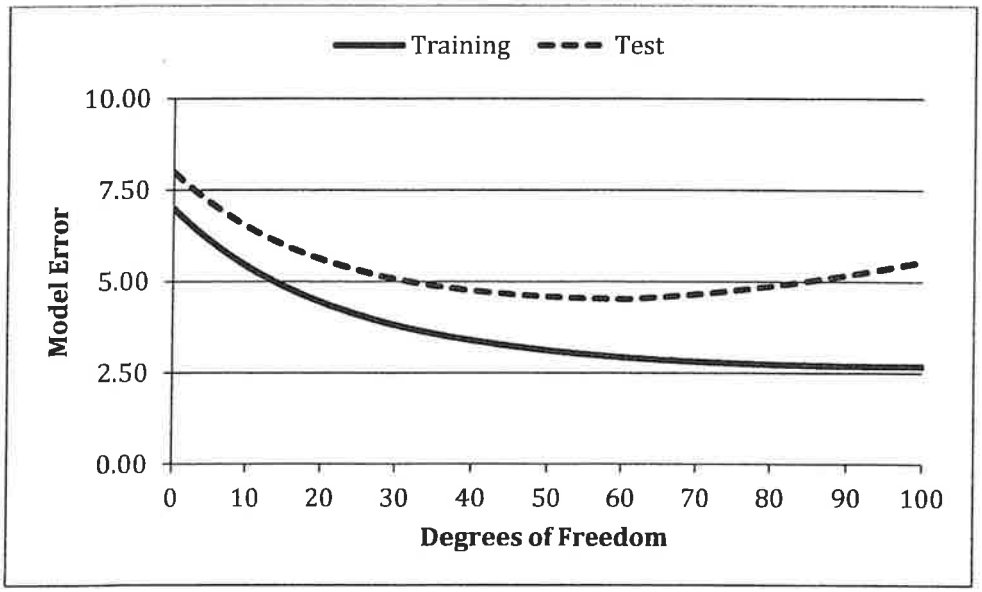
4. (1.75 points)

An actuary has split data into training and test groups for a model. The chart below shows the relationship between model performance and model complexity. Model performance is represented by model error and model complexity is represented by degrees of freedom.



a. (0.5 point)

Briefly describe two reasons for splitting modeling data into training and test groups.

b. (0.75 point)

Briefly describe whether each of the following model iterations has an optimal balance of complexity and performance.

i. Model iteration 1: 10 degrees of freedom
ii. Model iteration 2: 60 degrees of freedom
iii. Model iteration 3: 100 degrees of freedom

c. (0.5 points)

Identify and briefly describe one situation where it is an advantage to split the data by time rather than by random assignment.

CONTINUED ON NEXT PAGE

SAMPLE ANSWERS AND EXAMINER'S REPORT

QUESTION 4	
TOTAL POINT VALUE: 1.75	**LEARNING OBJECTIVE(S): A3b, A2d**
SAMPLE ANSWERS	
Part a: 0.5 point	
Sample 1 • Splitting data into training and test groups prevents overfitting of the model because the model will always fit the training better with more parameters but it also could pick up random variation as a predictive variable. • It also allows for the testing of the predictive power of the model, if it doesn't fit the test data well it likely won't predict future outcomes well either. *Sample 2* To test the predictive power of the model & to avoid overfitting to the noise of the training set. We use the training set to fit the model but then we test this on the test set (which is data that is "unseen" by the model) to make sure we have not overfit & that the model is predictive.	
Part b: 0.75 point	
Sample 1 • Model iteration 1: The model does not have an optimal balance here as the model error for both training and test data is relatively high and can be decreased. • Model iteration 2: This model has an optimal balance as the training error is lower which is expected with more variables. However the test error is also at it lowest point, meaning adding more parameters has not lead to overfitting. • Model iteration 3: This model does not have an optimal balance. Adding more parameters has lowered the training data errors but increased the error for the test data. This model has been overfit. *Sample 2* i. not optimal because we can still improve model performance on the test set by adding additional degrees of freedom ii. This appears to be the optimal balance because this is right around where the test set has lowest model error iii. not optimal, too much complexity (ie degrees of freedom) in the model, which has caused the model error on the test set to actually increase (ie we have overfit to the training set).	
Part c: 0.5 point	
Sample 1 Splitting data by time would be advantageous when weather events occur. This keeps the entire event in one section of data and prevents overfitting. *Sample 2* When data is affected by a large weather event. In which case, this event will only be included in one of the two sets & we won't get overly optimistic results that would occur if the event affected both sets.	

SAMPLE ANSWERS AND EXAMINER'S REPORT

EXAMINER'S REPORT

Candidates were expected to have a high-level understanding of how to segment a dataset for construction and evaluation of a predictive model, how to assess the quality of modeled output, and to comment on strengths and weaknesses of alternative data segmentation strategies.

Part a

Candidates were expected to clearly identify and describe two reasons for using a holdout dataset. In particular, acceptable reasons fell into two distinct groups: to avoid overfitting, and to assess the predictive power of the model.

Common variations that received full credit for the first of these categories (avoid overfitting) were:

- Discussion of the fact that the addition of variables will always improve the fit on train data, but that it may not improve the fit on test data
- Discussion of k-fold cross-validation methods (or similar techniques) when referencing parameters, quality of fit, etc.

Common variations that received full credit for the second of these categories (assess predictive power) were:

- Discussion of the fact that "attempting to test the performance of any model on the same set of data on which the model was built will produce overoptimistic results" (from the GLM paper)
- Selecting a best model from multiple alternative models
- Discussion of validating a model on an "unseen" dataset
- Discussion of assessing model stability
- Discussion of k-fold cross-validation methods (or similar techniques) when referencing model stability, lift, etc.

Common mistakes included:

- Discussing two variations of the same thing: either preventing overfitting or assessing predictive power
- Essentially re-stating the question: for example, simply stating that a test group is used to test a model (without being specific regarding what is being tested)

Part b

Candidates were expected to assess the balance of complexity and performance based on a graph of model error as related to model degrees of freedom (number of free parameters). The question specifically referred to three candidate model iterations. While full-credit responses varied significantly in length, in general to receive full credit, candidates needed to recognize and convey an understanding of each of the following:

- A training data curve will always decrease monotonically with addition of degrees of freedom
- Model performance on test data will improve until the model has been overfit
- An optimal balance of complexity and performance occurs at/near the minimum of the test data curve

Common mistakes included:

- Drawing conclusions based primarily on training set error rather than test set error
- Thinking that the gap between the training data curve and the test data curve is meaningful in itself, without addressing the actual magnitude of these curves
- Not addressing the trade-off between complexity and performance (most candidates implicitly did this by selecting an optimal model iteration)

Part c

Candidates were expected to know that it is an advantage to have an out-of-time validation dataset when many records are influenced by a single event. The most common examples were a catastrophe or other weather event. To receive full credit, candidates were expected to identify such a situation, to note that splitting by time would result in claims from the same event being assigned to either the training dataset or the testing dataset (not both), and would therefore lessen the chance for overfitting or producing overly optimistic validation results.

Common mistakes included:

- Correctly identifying the situation (CAT or other large event), but not explaining why a time-based split is preferred
- Identifying "weather" as a situation, rather than specifying "weather events" (which implies correlated claims, and which is a fundamental reason to prefer an out-of-time split)
- Arguing that because a random split ***might*** result in an imbalanced test or train dataset (most candidates identified "seasonal effects" as the cause, but other reasons were given), therefore splitting by time (even/odd years, for example) is more appropriate than randomly splitting (note: credit was given in this case to candidates who demonstrated specific cases in which this might be reasonable)
- Claiming that splitting by time would alleviate the problem of correlation due to insureds having more than one policy period included in the dataset
- Arguing that time-dependent signals in the data (for example, underlying trends in the data, shifts in the mix of business, or shifts of fitted parameters over time) could be either identified or adjusted for by using a ***single*** time-based train/test split of the data (note: using multiple train/test splits may help in these situations – for example, using a k-fold out-of-sample out-of-time validation – and credit was given if the candidate made this point)