Print

Quiz: Model Building for Scoring and Prediction

*Select the best answer for each question. When you are finished, click **Submit Quiz**.*

1. Predictive models can be based on which of the following?

   - *a.* parametric models only
   - *b.* non-parametric models only
   - *c.* both parametric and non-parametric models

2. In predictive modeling, your goal is to create the best possible model to score new data. The model you choose should be which of the following?

   - *a.* It should overfit the training data in order to accommodate random noise in the sample.
   - *b.* It should underfit the training data to avoid modeling chance relationships.
   - *c.* It should be flexible enough to fit the training data well but generalize to new data sets.

3. Honest assessment might generate multiple candidate models that have the same (or nearly the same) validation assessment values. In this situation, which model should be selected?

   - *a.* the model that has the highest variance when it is applied to the population
   - *b.* the model that has the most terms
   - *c.* the most parsimonious model
   - *d.* the most biased model

4. With a large enough data set, observations can be divided into three subset data sets for use in honest assessment. Which of the following is not the name of one of these three subset data sets?

   - *a.* training
   - *b.* validation
   - *c.* score
   - *d.* test

5. Which of the following PROC GLMSELECT steps splits the original data set into a training data set that contains 80% of the original data and a validation data set that contains 20% of the original data?

   - *a.*
   ```
   proc glmselect data=housing;
      class fireplace lot_shape;
      model Sale_price = fireplace lot_shape / fraction(test=0 validate=.20);
   run;
   ```
   - *b.*
   ```
   proc glmselect data=housing;
      class fireplace lot_shape;
      model Sale_price = fireplace lot_shape / partition(test=0 validate=.20);
   run;
   ```
   - *c.*
   ```
   proc glmselect data=housing;
      class fireplace lot_shape;
      model Sale_price = fireplace lot_shape;
      fraction(test=0 validate=.20);
   run;
   ```
   - *d.*
   ```
   proc glmselect data=housing;
      class fireplace lot_shape;
      model Sale_price = fireplace lot_shape;
      partition fraction(test=0 validate=.20);
   run;
   ```

---

6. Which of the following statements is true about the SEED= option in PROC GLMSELECT?

   - *a.* You can reproduce your results if you specify an integer that is greater than zero in the SEED= option and then rerun the code using the same SEED= value.
   - *b.* The SEED= option offers an alternative way to specify the proportion of observations to allocate to the validation data set.
   - *c.* If a valid value is not specified for the SEED= option, the code does not run.
   - *d.* You can use the SEED= option only when you already partitioned the data before model building.

---

7. Which of the following does PROC GLMSELECT use to select a model from the candidate models when a validation data set is provided?

   - *a.* the smallest number of predictors
   - *b.* the largest adjusted R-square value
   - *c.* the smallest overall validation average squared error
   - *d.* none of the above

---

8. Which of the following statements about scoring is true?

- *a.* When you score data, you must rerun the algorithm that was used to build the model.
- *b.* When you score data, you apply the score code (the equations obtained from the final model) to the scoring data.
- *c.* If you made any modifications to the training or validation data, it is not necessary to make the same modifications to the scoring data.
- *d.* The scoring data set cannot be larger than either the training data set or the validation data set.

---

9. A department store is deploying a chosen model to make predictions for an upcoming sales period. They have the necessary data and are ready to proceed. Which of the following methods can be used for scoring?

   - *a.* a PROC GLMSELECT step that contains the SCORE statement
   - *b.* a PROC PLM step that contains the SCORE statement and references an item store that was created in PROC GLMSELECT
   - *c.* a PROC PLM step with the CODE statement that writes the score code based on an item store created in PROC GLMSELECT, and a DATA step that scores the data
   - *d.* any of the above

---

10. Suppose you ran a PROC GLMSELECT step that saved the context and results of the statistical analysis in an item store named **homestore**. Which of the following programs scores new observations in a data set named **new** and saves the predictions in a data set named **new_out**?

    - *a.*
      ```
      proc plm restore=homestore;
         score data=new out=new_out;
      run;
      ```
    - *b.*
      ```
      proc plm restore=new;
         score data=homestore out=new_out;
      run;
      ```
    - *c.*
      ```
      proc plm data=homestore;
         score data=new out=new_out;
      run;
      ```
    - *d.*
      ```
      proc plm restore=homestore;
         model data=new out=new_out;
      run;
      ```

Submit Quiz