PassQueen
a.k.a dumpsqueen.com

SAS Statistical Business Analysis SAS9: Regression and Model

SAS Institute A00-240

Version Demo

Total Demo Questions: 16

Total Premium Questions: 161

Buy Premium PDF

https://passqueen.com

support@passqueen.com

Topic Break Down

Topic	No. of Questions
Topic 1, ANOVA and Regression	59
Topic 2, Logistic Regression	51
Topic 3, Model Selection and Validation	51
Total	161

QUESTION NO: 1

The SAS data set RESULT contains the following variables:

- Region (GrpA or GrpB)
- Sales (dollars per year)

Which SAS programs can be used to find the p-value for comparing GrpA sales with GrpB sales? (Choose two.)

A.
```
proc ttest data = RESULT;
   class Region;
   var Sales;
run;
```

B.
```
proc ttest data = RESULT;
   class Region;
   model Sales = Region;
run;
```

C.
```
proc glm data = RESULT;
   class Region;
   model Sales = Region;
run;
```

D.
```
proc glm data = RESULT;
   class Sales;
   model Sales = Region;
run;
```

A. Option A

B. Option B

C. Option C

D. Option D

ANSWER: A C

Explanation:

The programs represented by “Option A” and “Option C” can produce the p-value for a comparison of annual Sales between the two Region groups. This is a two-independent-group comparison: each observation belongs to either GrpA or GrpB, and Sales is a continuous response variable. SAS can test whether the group means differ by using a two-sample t test, where Region is specified as the classification variable and Sales as the analysis variable. The resulting output includes the relevant test statistic and associated p-value, with separate pooled-variance and unequal-variance forms available as appropriate.

The same comparison can also be fit as a one-factor linear model or analysis of variance with Sales as the dependent variable and Region as the categorical predictor. With only two Region levels, the overall model test for Region is mathematically equivalent to testing the difference between the two group means; consequently, its p-value answers the requested comparison. In SAS, this is commonly implemented using `PROC TTEST` or a modeling procedure such as `PROC`

GLM, with an appropriate CLASS statement for Region. SAS documentation for the two-sample analysis is available in the PROC TTEST syntax reference; model-based testing is described in the PROC GLM syntax reference.

QUESTION NO: 2

Refer to the confusion matrix:

Calculate the sensitivity. (0 - negative outcome, 1 - positive outcome)

Click the calculator button to display a calculator if needed.

A. 25/48

B. 58/102

C. 25/B9

D. 58/81

ANSWER: A

Explanation:

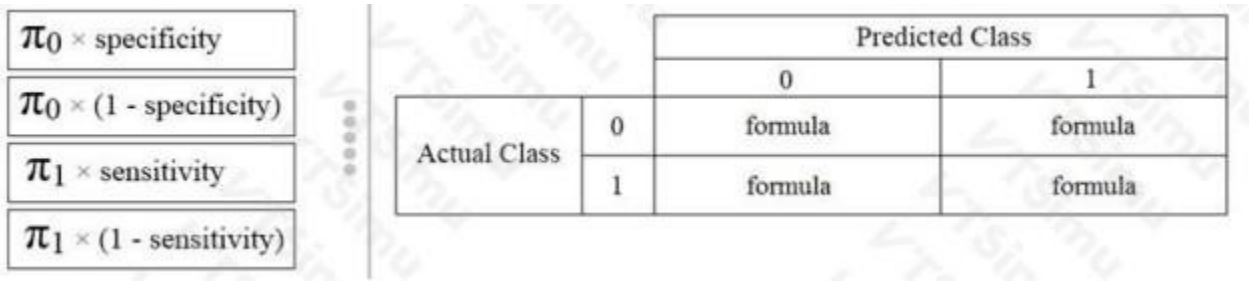
The correct answer is **25/48**. Sensitivity, also called the true-positive rate or recall for the positive class, measures how effectively a model identifies observations that truly have the positive outcome. Because the question defines 1 as the positive outcome, sensitivity is calculated using the counts for actual positive observations only: true positives divided by all actual positives. In the confusion matrix, 25 observations with actual outcome 1 were correctly classified as 1, giving the true-positive count. There were 48 actual positive observations in total, consisting of the 25 true positives plus the positive observations classified as 0. Therefore, the sensitivity is calculated as 25 / 48, or approximately 0.521 (52.1%). This statistic is particularly useful when the cost of failing to identify a positive event is important, because it focuses on the proportion of genuine positive cases detected by the model. SAS model-assessment output uses confusion-matrix cell counts to derive classification measures, including sensitivity, for a specified event level. See SAS documentation on classification tables in PROC LOGISTIC and SAS's overview of machine learning analytics.

QUESTION NO: 3 - (DRAG DROP)

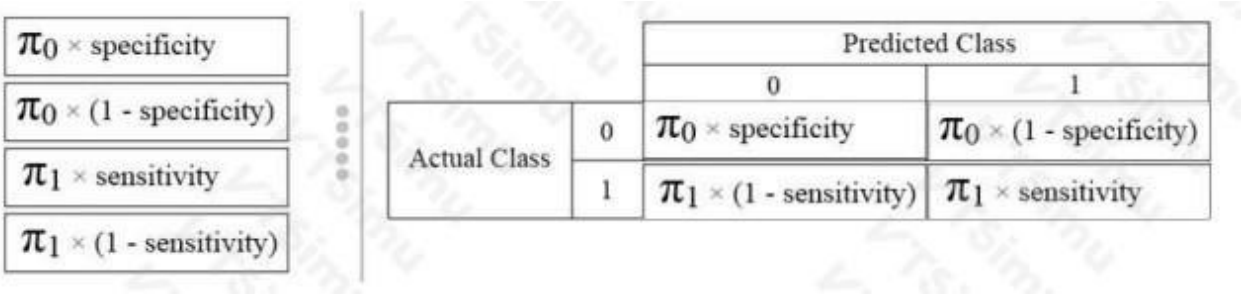
DRAG DROP

Drag the adjustment formulas for oversamping from the left and place them into the correct location in the confusion matrix shown on the right.

Select and Place:



ANSWER:



Explanation:

A confusion matrix is formed by combining the prevalence of the actual class with the classifier's conditional probability of producing each prediction for that class. The prevalence π_0 applies to every row whose actual class is 0, while π_1 applies to every row whose actual class is 1. Specificity is the probability that an observation from actual class 0 is predicted as 0. Therefore, the upper-left cell, where Actual Class is 0 and Predicted Class is 0, is π_0 × specificity. The remaining probability within the actual-class-0 row is the false-positive probability, 1 - specificity, so the upper-right cell is π_0 × (1 - specificity).

Sensitivity is the probability that an observation from actual class 1 is predicted as 1. Thus, the lower-right cell is π_1 × sensitivity. The remaining probability in that row is the false-negative probability, 1 - sensitivity, making the lower-left cell π_1 × (1 - sensitivity). These four joint probabilities preserve the row totals π_0 and π_1 and connect the confusion matrix to the population class distribution, which is the essential adjustment needed when a modeling sample has been oversampled. SAS documents classification assessment measures, including sensitivity and specificity, in its assessment details documentation. The corresponding definitions of sensitivity and specificity are also summarized by the SAS machine learning resources.

QUESTION NO: 4

An analyst has predicted event probabilities from a binary logistic regression model and considers several classification cutoff values.

Which assessment measures do not depend on the selected classification cutoff? (Choose two.)

A. Area under the ROC curve

B. Average Squared Error

C. Sensitivity

D. Misclassification rate

ANSWER: A B

Explanation:

The area under the ROC curve and average squared error are calculated from the predicted probabilities and observed target values without converting probabilities to event or nonevent classifications at one selected cutoff. The area under the ROC curve measures ranking discrimination, and average squared error measures the average squared difference between the observed binary response and the predicted probability.

Sensitivity and misclassification rate are calculated from a confusion matrix. Their values depend on which predicted probabilities are classified as events, so they can change when the cutoff changes.

QUESTION NO: 5

A financial services manager wants to assess the probability that certain clients will default on their Home Equity Line of Credit (HELOC). A former employee left the code listed below.

```
proc logistic data = MYDIR.HELOC des outest=MSG;
    model DEFAULT = amount job_code years_at_residence;
run;

proc score data = MYDIR.RECENT_HELOC
                          out = SCORED_HELOC
                          score = MSG
                          type = parms;
   var Amount Job_code Years_at_residence;
run;
```

The training data set is named HELOC, while a similar data set of more recent clients is named RECENT_HELOC.

Which SAS data steps will calculate the predicted probability of default on recent clients? (Choose two.)

A.
```
data NEW_PROB;
   set SCORED_HELOC;
   p=1/(1+exp(-DEFAULT));
run;
```

B.
```
data NEW_PROB;
   set SCORED_HELOC;
   ODDS = exp(DEFAULT);
   p = ODDS /(1+ODDS);
run;
```

C.
```
data NEW_PROB;
   set SCORED_HELOC;
   p=(1+exp(DEFAULT))/exp(DEFAULT);
run;
```

D.
```
data NEW_PROB;
   set SCORED_HELOC;
   p = DEFAULT /(1+DEFAULT);
run;
```

A. Option A

B. Option B

C. Option C

D. Option D

ANSWER: A B

Explanation:

The two selected data steps correctly apply the fitted logistic-regression model to each observation in `RECENT_HELOC`. Logistic regression first computes a linear predictor, often called the logit: an intercept plus the estimated coefficient for each model input multiplied by that client's corresponding value. This score is not itself a probability. To obtain the predicted probability of default, the data step must transform the linear predictor with the inverse-logit function, `1/(1+exp(-logit))`.

SAS also provides the equivalent `LOGISTIC` function, so either expression produces the same probability when applied to the same linear predictor.

The calculation must use the parameter estimates from the model trained on `HELOC`, while reading the predictor values from `RECENT_HELOC`. The resulting probability is defined for the event that was modeled as the default event; therefore, the event coding and sign of the linear predictor must remain consistent with the original fitted model. The output data set can retain the recent-client variables and add the calculated probability for downstream risk ranking, reporting, or decision rules. SAS documents the logistic link and the inverse transformation from a linear predictor to an event probability in the PROC LOGISTIC details documentation. The LOGISTIC function reference describes the equivalent SAS function form.

QUESTION NO: 6 - (SIMULATION)

Refer to the confusion matrix:

		Predicted Outcome	
		0	1
Actual Outcome	0	345	155
	1	188	312

An analyst determines that loan defaults occur at the rate of 3% in the overall population. The above confusion matrix is from an oversampled test set (1 = default).

What is the sensitivity adjusted for the population event probability?

Enter your answer in the space below. Round to three decimals (example: n.nnn).

ANSWER: See the explanation for the answer

Explanation:

Answer: `0.624`

Sensitivity is the proportion of actual defaults that were predicted as defaults:

```
Sensitivity = True Positives / (True Positives + False Negatives)
            = 312 / (312 + 188)
            = 312 / 500
            = 0.624
```

Although the test data were oversampled, both true positives and false negatives are from the actual-default group and receive the same population-event adjustment. Therefore, the 3% population event probability does not change sensitivity.

QUESTION NO: 7

An analyst adds a predictor to a multiple linear regression model. The R-Square increases slightly, but the adjusted R-Square decreases.

What does the decrease in adjusted R-Square indicate?

A. The added predictor explains no variation in the response.

B. The added predictor does not improve fit enough to offset the adjusted R-Square complexity penalty.

C. The original model has a larger ordinary R-Square than the expanded model.

D. The regression error sum of squares must be unchanged.

ANSWER: B

Explanation:

The added predictor does not improve the model sufficiently to justify its additional complexity according to adjusted R-Square. Unlike ordinary R-Square, adjusted R-Square includes a penalty for adding model parameters. It can decrease when the reduction in error variation from a new predictor is too small relative to the loss of degrees of freedom.

Ordinary R-Square cannot decrease when an effect is added to a least-squares regression model fit to the same observations. Therefore, the reported increase in R-Square and decrease in adjusted R-Square are compatible results.

QUESTION NO: 8

An analyst compares the mean salaries of men and women working at a company.

The SAS data set SALARY contains variables:
Gender (M or F)
Pay (dollars per year)

Which SAS programs can be used to find the p-value for comparing men's salaries with women's salaries? (Choose two.)

A. Option A

B. Option B

C. Option C

D. Option D

E. `proc ttest data=salary; class Gender; var Pay; run;`

F. `proc glm data=salary; class Gender; model Pay=Gender; run;`

ANSWER: E F

Explanation:

`proc ttest data=salary; class Gender; var Pay; run;` is appropriate because PROC TTEST directly performs a two-sample t test when the CLASS statement identifies the two groups and the VAR statement identifies the numeric response. Its output includes hypothesis-test p-values for the difference between the mean Pay values for the M and F Gender groups. SAS also provides both pooled and unequal-variance (Satterthwaite) forms of the two-sample comparison, allowing the analyst to use the reported p-value suitable for the variance assumption selected.

`proc glm data=salary; class Gender; model Pay=Gender; run;` is also appropriate. With a single two-level classification effect, the GLM model tests whether mean Pay differs by Gender. The Type III test for Gender is an analysis-of-variance F test; for a two-group comparison, that test is mathematically equivalent to the usual equal-variance two-sample t test, with the F statistic equal to the squared t statistic. Therefore, the model output supplies a p-value for the same null hypothesis that the mean salaries are equal. SAS documents the required CLASS and VAR statements for PROC TTEST and the CLASS and MODEL statements for PROC GLM in the PROC TTEST syntax documentation and PROC GLM syntax documentation.

QUESTION NO: 9

Refer to the REG procedure output:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	31848	15924	13.42	<.0001
Error	97	115082	1186.40833		
Corrected Total	99	146930			

Root MSE	34.44428	R-Square	0.2168
Dependent Mean	606.38715	Adj R-Sq	0.2006
Coeff Var	5.68025		

An analyst has selected this model as a champion because it shows better model fit than a competing model with more predictors.

Which statistic justifies this rationale?

A. R-Square

B. Coeff Var

C. Adj R-Sq

D. Error DF

ANSWER: C

Explanation:

Adj R-Sq justifies selecting a model with fewer predictors when it has better fit than a competing model with more predictors. Ordinary R-Square measures the proportion of variation in the response explained by the fitted model, but it cannot decrease merely because additional predictors are included. Therefore, R-Square alone does not provide an appropriate basis for preferring a more parsimonious model. Adjusted R-Square modifies R-Square using the model and error degrees of freedom, incorporating a penalty for the number of predictors estimated. A newly added predictor improves Adjusted R-Square only when its contribution to explaining variation is sufficient to offset that complexity penalty. Consequently, when comparing models fit to the same response using the same observations, the model with the higher Adjusted R-Square has the preferable balance of explanatory fit and number of predictors. The REG procedure reports this statistic in its summary output as Adj R-Sq, making it directly useful for the analyst's champion-model rationale. SAS documents R-Square and Adjusted R-Square among the fit statistics produced by PROC REG. See the SAS PROC REG model fit and summary statistics documentation and the SAS PROC REG documentation.

QUESTION NO: 10

An analyst has fitted a binary logistic regression model and uses its predicted event probabilities to classify observations.

Which statements about the c statistic are correct? (Choose two.)

A. The c statistic is equivalent to the area under the ROC curve.

B. Changing the classification cutoff changes the c statistic.

C. The c statistic evaluates the ranking of event and nonevent observations.

D. A high c statistic guarantees that predicted probabilities are well calibrated.

ANSWER: A C

Explanation:

The c statistic measures the model's ability to rank event observations above nonevent observations. For a binary logistic regression model, it is equivalent to the area under the ROC curve.

The c statistic is based on the ordering of predicted probabilities rather than on a selected classification cutoff. Consequently, changing a cutoff used to classify observations as events or nonevents does not change the c statistic. The c statistic measures discrimination and does not, by itself, assess calibration of predicted probabilities.

QUESTION NO: 11

An analyst fits a multiple linear regression model with PROC REG and requests the Durbin-Watson statistic. The reported Durbin-Watson value is 0.82.

What does this result suggest?

A. The residuals exhibit positive autocorrelation.

B. The residuals exhibit negative autocorrelation.

C. The residuals have constant variance.

D. The model has no collinearity.

ANSWER: A

Explanation:

A Durbin-Watson statistic substantially below 2 suggests positive autocorrelation among the model residuals. The statistic ranges approximately from 0 to 4, with a value near 2 indicating no first-order autocorrelation. Values approaching 0 are consistent with positive autocorrelation, whereas values approaching 4 are consistent with negative autocorrelation.

Positive residual autocorrelation violates the ordinary least-squares assumption that errors are independent. This can cause standard errors and significance tests to be unreliable, even when the fitted regression coefficients appear reasonable. PROC REG can request this diagnostic by using the `DW` option in the MODEL statement. See the SAS PROC REG documentation.

QUESTION NO: 12

An analyst fits a binary logistic regression model and wants to evaluate whether the fitted probabilities are consistent with the observed event frequencies.

Which two statements are correct? (Choose two.)

A. The `LACKFIT` option requests the Hosmer-Lemeshow goodness-of-fit test.

B. A small Hosmer-Lemeshow test p-value can indicate inadequate model fit.

C. A nonsignificant Hosmer-Lemeshow test proves that every model effect is significant.

D. The Hosmer-Lemeshow test measures collinearity among the predictors.

ANSWER: A B

Explanation:

The Hosmer-Lemeshow goodness-of-fit test compares observed and expected event frequencies across groups based on predicted probabilities. In PROC LOGISTIC, the `LACKFIT` option in the MODEL statement requests this test for a binary response model. A small p-value provides evidence that the fitted probabilities do not adequately agree with the observed outcomes.

A nonsignificant Hosmer-Lemeshow test does not prove that a model is correct. Its result can depend on sample size and grouping, so it should be considered with other evidence, such as residual diagnostics, calibration assessment, discrimination statistics, and validation-sample performance. See the SAS PROC LOGISTIC documentation.

QUESTION NO: 13

Refer to the REG procedure output:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	618.44051	40.03665	15.45	<.0001	0
overhead	1	4.99845	0.00157	3181.24	<.0001	0.99993
scrap	1	2.82667	0.71581	3.95	<.0001	0.00124
training	1	-50.95436	2.82069	-18.06	<.0001	-0.00568

The Intercept estimate is interpreted as:

A. The predicted value of the response when all the predictors are at their current values.

B. The predicted value of the response when all predictors are at their means.

C. The predicted value of the response when all predictors = 0.

D. The predicted value of the response when all predictors are at their minimum values.

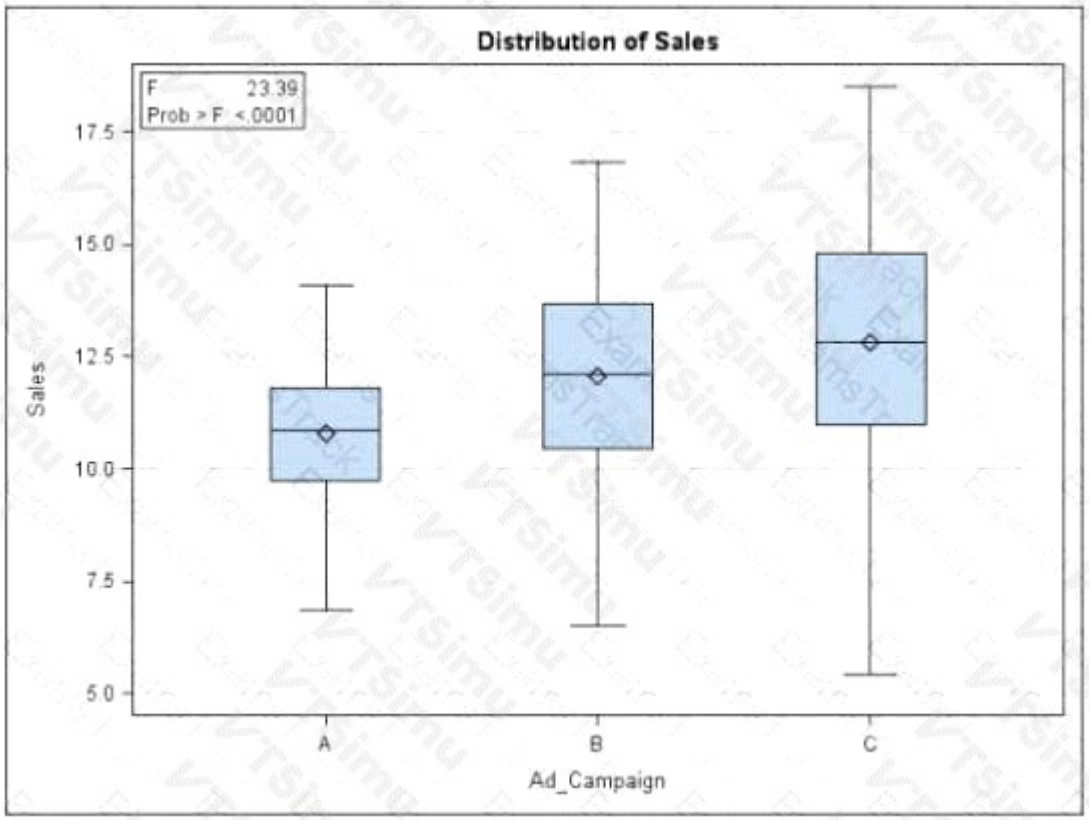
ANSWER: C

Explanation:

The predicted value of the response when all predictors = 0 is the correct interpretation. In a linear regression model fit with SAS PROC REG, the fitted mean response is represented as an intercept plus the sum of each predictor multiplied by its estimated regression coefficient. Setting every predictor variable to zero removes all of the predictor-coefficient terms from that equation. The remaining value is therefore the estimated intercept, which is the model's predicted response at the zero point for every explanatory variable. For example, for a model written as $Y = b_0 + b_1X_1 + b_2X_2$, when both X_1 and X_2 equal zero, the predicted value of Y is b_0. This interpretation is mathematical and applies regardless of whether zero is an observed, practical, or meaningful value for every predictor. If predictors have been centered before fitting the model, zero may correspond to their centering values, but the intercept still represents the prediction when the variables supplied to the model equal zero. SAS documents the regression model as an intercept term together with explanatory-variable effects in the PROC REG MODEL statement documentation; related parameter-estimate information is available in the SAS PROC REG details.

QUESTION NO: 14

Refer to the exhibit:



The box plot was used to analyze daily sales data following three different ad campaigns.

The business analyst concludes that one of the assumptions of ANOVA was violated.

Which assumption has been violated and why?

A. Normality, because Prob > F < .0001.

B. Normality, because the interquartile ranges are different in different ad campaigns.

C. Constant variance, because Prob > F < .0001.

D. Constant variance, because the interquartile ranges are different in different ad campaigns.

ANSWER: D

Explanation:

Constant variance, because the interquartile ranges are different in different ad campaigns. ANOVA assumes homogeneity of variance: the variability of the response should be approximately the same in every treatment or group. In a box plot, the height of each box represents the interquartile range (IQR), the spread of the middle 50% of observations. The substantially different box heights across the ad-campaign groups indicate that the within-campaign spread of daily sales is not similar. This is visual evidence that the population variances may differ by campaign, violating the constant-variance assumption required for the usual ANOVA inference. A box plot is particularly useful as an initial graphical diagnostic because it permits comparison of center, spread, skewness, and potential outliers among groups. The reported ANOVA F-test probability concerns evidence for differences among group means; it is not itself a test of the equal-variance assumption. When unequal variances are suspected, an analyst should investigate further with residual diagnostics and, where appropriate, use a method designed for heterogeneous variances or consider a suitable transformation. SAS describes ANOVA model assumptions and diagnostic assessment in the PROC GLM documentation; its discussion of graphical summaries, including box plots, is available in the SAS box plot documentation.

QUESTION NO: 15

An analyst knows that the categorical predictor, storeId, is an important predictor of the target.

However, store_Id has too many levels to be a feasible predictor in the model. The analyst wants to combine stores and treat them as members of the same class level.

What are the two most effective ways to address the problem? (Choose two.)

A. Eliminate store_id as a predictor in the model because it has too many levels to be feasible.

B. Cluster by using Greenacre's method to combine stores that are similar.

C. Use subject matter expertise to combine stores that are similar.

D. Randomly combine the stores into five groups to keep the stochastic variation among the observations intact.

ANSWER: B C

Explanation:

Clustering by using Greenacre's method to combine stores that are similar is appropriate because it provides a data-driven way to reduce a high-cardinality categorical predictor to a manageable number of groups. The method evaluates category-level relationships using the target-related profile of each store, allowing stores with similar behavior to be assigned to the same consolidated class level. This retains useful predictive information while reducing the number of parameters, degrees of freedom, and sparse category effects that can make a regression model impractical or unstable.

Using subject matter expertise to combine stores that are similar is also effective. Business knowledge can identify meaningful, stable store groupings based on characteristics such as geography, format, size, customer segment, operating model, or merchandising strategy. Such groupings can preserve an interpretable relationship between store membership and the target, while avoiding an unwieldy number of indicator variables. In practice, analysts can use domain-defined groupings alone or compare them with target-based clustering results during model assessment. SAS documentation resources for modeling categorical effects and related statistical procedures are available through the SAS Documentation portal and the SAS Support documentation library.

QUESTION NO: 16

Including redundant input variables in a regression model can:

A. Stabilize parameter estimates and increase the risk of overfitting.

B. Destabilize parameter estimates and increase the risk of overfitting.

C. Stabilize parameter estimates and decrease the risk of overfitting.

D. Destabilize parameter estimates and decrease the risk of overfitting.

ANSWER: B

Explanation:

“Destabilize parameter estimates and increase the risk of overfitting.” is correct. Redundant input variables provide overlapping information about the response, often creating substantial correlation among predictor columns. In a regression model, this multicollinearity makes it difficult to separate the individual contribution of each correlated predictor. As a result, estimated regression coefficients can become unstable: relatively small changes in the data may produce large changes in coefficient values, signs, or standard errors. Although predictions may sometimes remain reasonably similar within the training data, the individual parameter estimates are less reliable for interpretation.

Adding redundant predictors also increases model complexity without adding commensurate independent information. A more complex model has more opportunity to fit sampling noise rather than the underlying relationship, which raises the risk of overfitting and can reduce performance on new observations. SAS model-selection procedures are designed to help identify a parsimonious set of effects and assess models using validation or selection criteria, rather than retaining unnecessary predictors. SAS discusses regression model selection in the GLMSELECT procedure documentation and provides regression modeling guidance in the SAS learning resources.