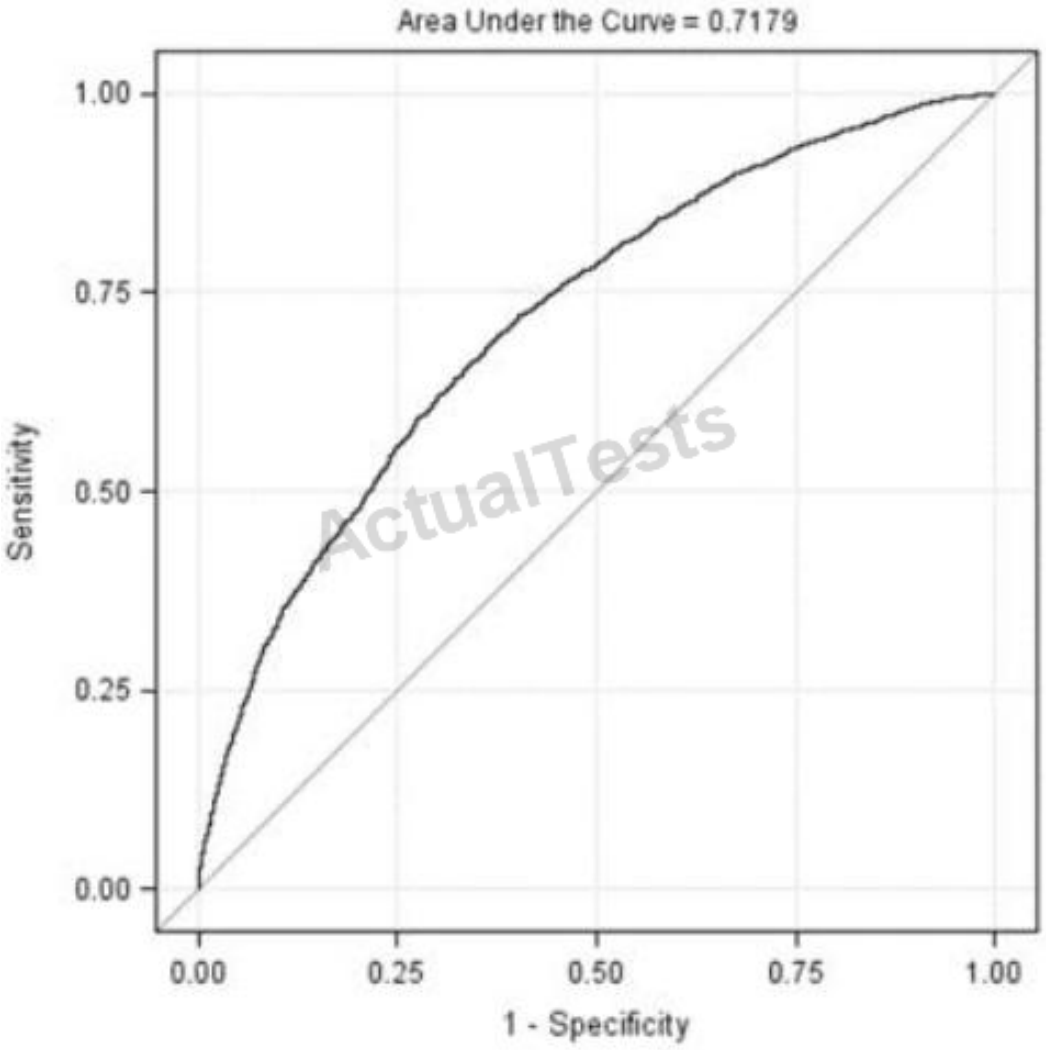
SAS Institute A00-240

SAS Statistical Business Analysis SAS9: Regression and Model

Version: 4.0

QUESTION NO: 1

Refer to the ROC curve:



As you move along the curve, what changes?

A. The priors in the population
B. The true negative rate in the population
C. The proportion of events in the training data
D. The probability cutoff for scoring

Answer: D
Explanation:

QUESTION NO: 2

When mean imputation is performed on data after the data is partitioned for honest assessment,

what is the most appropriate method for handling the mean imputation?

A. The sample means from the validation data set are applied to the training and test data sets.
B. The sample means from the training data set are applied to the validation and test data sets.
C. The sample means from the test data set are applied to the training and validation data sets.
D. The sample means from each partition of the data are applied to their own partition.

Answer: B
Explanation:

QUESTION NO: 3

An analyst generates a model using the LOGISTIC procedure. They are now interested in getting the sensitivity and specificity statistics on a validation data set for a variety of cutoff values.

Which statement and option combination will generate these statistics?

A. Scoredata=valid1 out=roc;
B. Scoredata=valid1 outroc=roc;
C. mode1resp(event= '1') = gender region/outroc=roc;
D. mode1resp(event"1") = gender region/ out=roc;

Answer: B
Explanation:

QUESTION NO: 4

In partitioning data for model assessment, which sampling methods are acceptable? (Choose two.)

A. Simple random sampling without replacement
B. Simple random sampling with replacement
C. Stratified random sampling without replacement
D. Sequential random sampling with replacement

Answer: A,C
Explanation:

QUESTION NO: 5

Which SAS program will divide the original data set into 60% training and 40% validation data sets, stratified by county?

A.
```
proc surveyselect data=SASUSER.DATABASE samprate=0.6 out=sample;
   strata county;
run;
```

B.
```
proc sort data=SASUSER.DATABASE;
   by county;
run;
proc surveyselect data=SASUSER.DATABASE samprate=0.6 out=sample outall;
run;
```

C.
```
proc sort data=SASUSER.DATABASE;
   by county;
run;
proc surveyselect data=SASUSER.DATABASE samprate =0.6 out=sample outall;
   strata county;
run;
```

D.
```
proc sort data=SASUSER.DATABASE;
   by county;
run;
proc surveyselect data=SASUSER.DATABASE samprate =0.6  out=sample;
   strata county;
run;
```

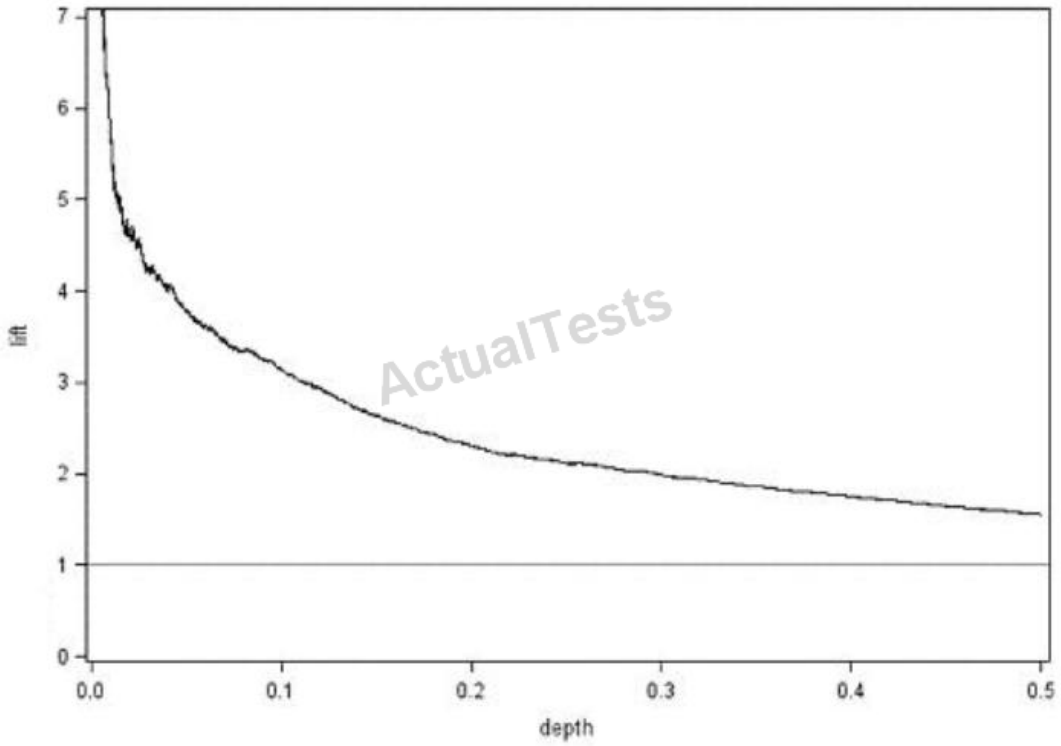

A. Option A
B. Option B
C. Option C
D. Option D

Answer: C

Explanation:

QUESTION NO: 6

Refer to the lift chart:



At a depth of 0.1, Lift = 3.14. What does this mean?

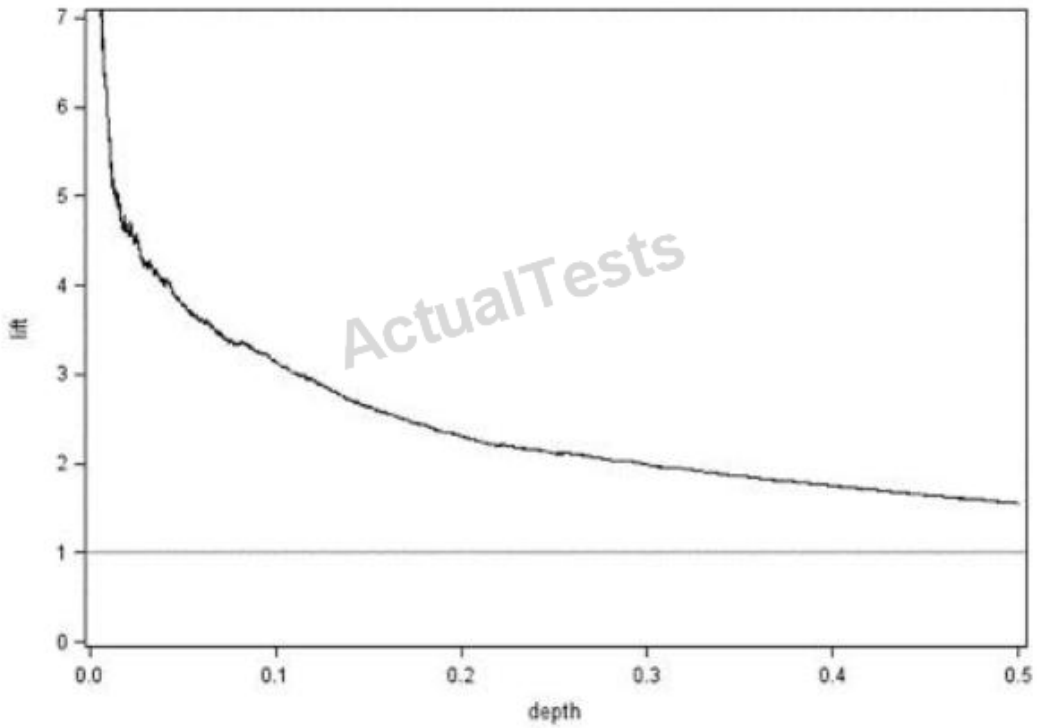
A. Selecting the top 10% of the population scored by the model should result in 3.14 times more events than a random draw of 10%.
B. Selecting the observations with a response probability of at least 10% should result in 3.14 times more events than a random draw of 10%.
C. Selecting the top 10% of the population scored by the model should result in 3.14 timesgreater accuracy than a random draw of 10%.
D. Selecting the observations with a response probability of atleast 10% should result in 3.14times greater accuracy than a random draw of 10%.

Answer: A
Explanation:

QUESTION NO: 7

Refer to the lift chart:

Refer to the lift chart:



What does the reference line at lift = 1 corresponds to?

A. The predicted lift for the best 50% of validation data cases
B. The predicted lift if the entire population is scored as event cases
C. The predicted lift if none of the population are scored as event cases
D. The predicted lift if 50% of the population are randomly scored as event cases

Answer: B

Explanation:

QUESTION NO: 8

Suppose training data are oversampled in the event group to make the number of events and non-events roughly equal. A logistic regression is run and the probabilities are output to a data set NEW and given the variable name PE. A decision rule considered is, "Classify data as an event if probability is greater than 0.5." Also the data set NEW contains a variable TG that indicates whether there is an event (1=Event, 0= No event).

The following SAS program was used.

```
data NEW;
   set NEW;
   Solicit = PE > .5;
run;
proc means data=NEW(where = (TG=1)) mean;
   var Solicit;
run;
```


What does this program calculate?

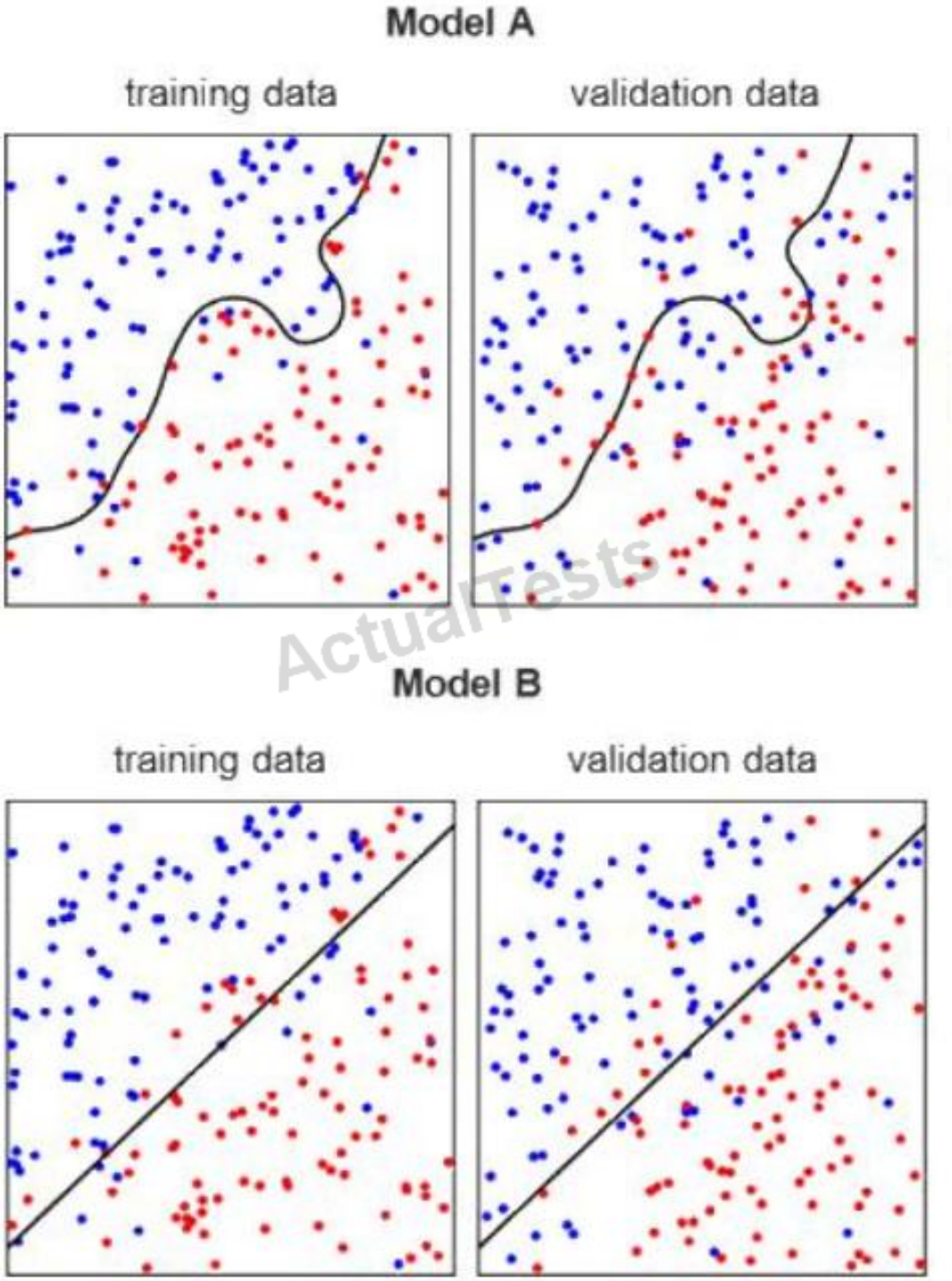
A. Depth
B. Sensitivity
C. Specificity
D. Positive predictive value

Answer: B

Explanation:

QUESTION NO: 9

Refer to the exhibit:



The plots represent two models, A and B, being fit to the same two data sets, training and validation.

Model A is 90.5% accurate at distinguishing blue from red on the training data and 75.5% accurate at doing the same on validation data. Model B is 83% accurate at distinguishing blue from red on the training data and 78.3% accurate at doing the same on the validation data.

Which of the two models should be selected and why?

A. Model A. It is more complex with a higher accuracy than model B on training data.
B. Model A. It performs better on the boundary for the training data.
C. Model B. It is more complex with a higher accuracy than model A on validation data.
D. Model B. It is simpler with a higher accuracy than model A on validation data.

Answer: D

Explanation:

QUESTION NO: 10

Assume a $10 cost for soliciting a non-responder and a $200 profit for soliciting a responder. The logistic regression model gives a probability score named P_R on a SAS data set called VALID. The VALID data set contains the responder variable Pinch, a 1/0 variable coded as 1 for responder. Customers will be solicited when their probability score is more than 0.05.

Which SAS program computes the profit for each customer in the data set VALID?

A.
```
data VALID;
   set VALID;
   Profit = (P_R > .05)*Purch*200  - (P_R > .05)*(1 - Purch)*10;
run;
```

B.
```
data VALID;
   set VALID;
   Profit = (P_R <= .05)*Purch*200  - (P_R > .05)*(1 - Purch)*10;
run;
```

C.
```
data VALID;
   set VALID;
   if P_R > .05;
   Profit = (P_R > .05)*Purch*200  - (P_R > .05)*(1 - Purch)*10;
run;
```

D.
```
data VALID;
   set VALID;
    if P_R >.05;
   Profit = (P_R > .05)*Purch*200  + (P_R <= .05)*(1 - Purch)*10;
run;
```


A. Option A
B. Option B
C. Option C
D. Option D

Answer: A

Explanation:

QUESTION NO: 11

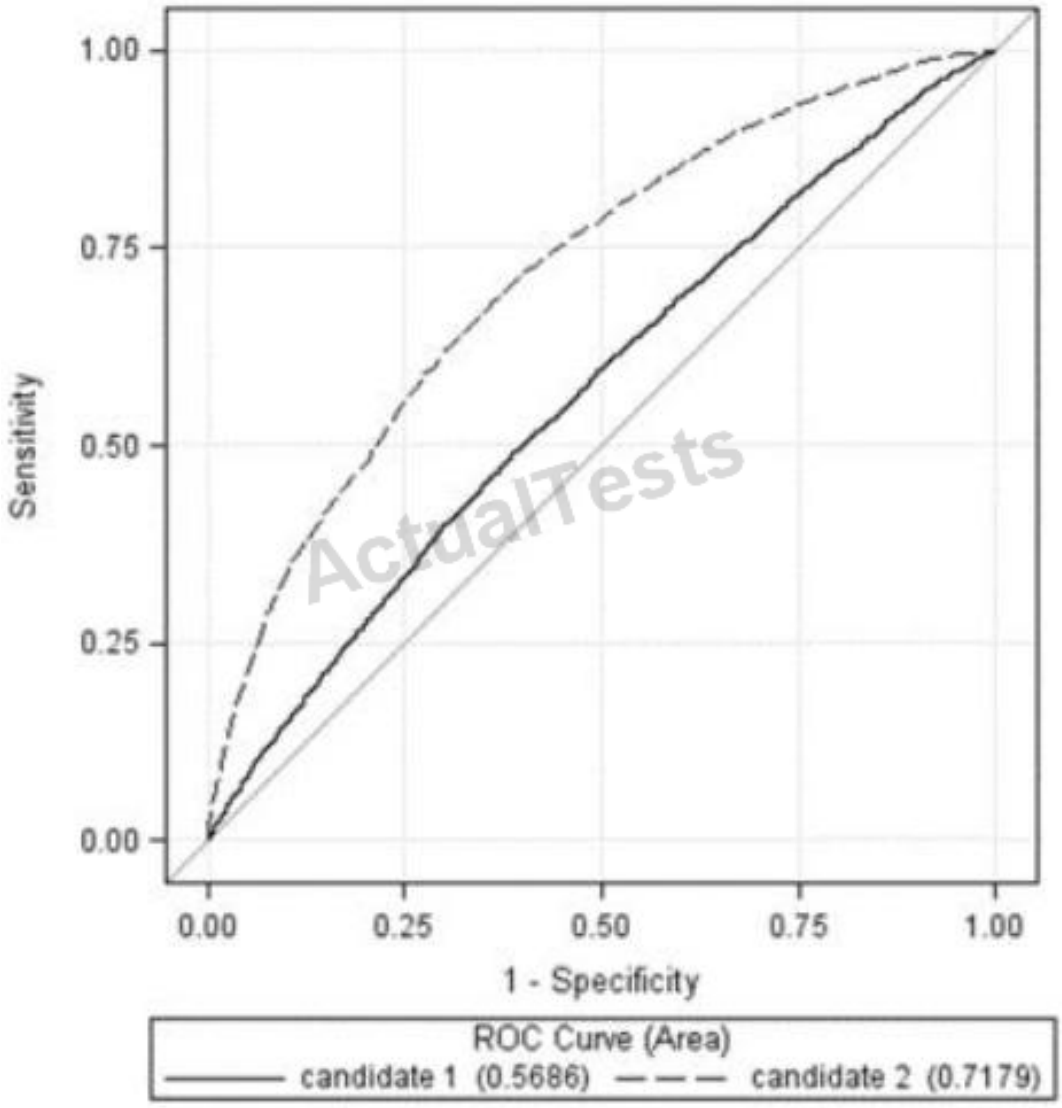
In order to perform honest assessment on a predictive model, what is an acceptable division between training, validation, and testing data?

A. Training: 50% Validation: 0% Testing: 50%
B. Training: 100% Validation: 0% Testing: 0%
C. Training: 0% Validation: 100% Testing: 0%
D. Training: 50% Validation: 50% Testing: 0%

Answer: D
Explanation:

QUESTION NO: 12

Refer to the exhibit:



Based upon the comparative ROC plot for two competing models, which is the champion model and why?

A. Candidate 1, because the area outside the curve is greater
B. Candidate 2, because the area under the curve is greater
C. Candidate 1, because it is closer to the diagonal reference curve
D. Candidate 2, because it shows less over fit than Candidate 1

Answer: B
Explanation:

QUESTION NO: 13

A marketing campaign will send brochures describing an expensive product to a set of customers. The cost for mailing and production per customer is $50. The company makes $500 revenue for

each sale.

What is the profit matrix for a typical person in the population?

A.

	Purchase	
Solicit	No	Yes
No	-50	0
Yes	0	450

B.

	Purchase	
Solicit	No	Yes
No	0	0
Yes	-50	500

C.

	Purchase	
Solicit	No	Yes
No	0	0
Yes	-50	450

D.

	Purchase	
Solicit	No	Yes
No	-50	0
Yes	0	500

A. Option A
B. Option B
C. Option C
D. Option D

Answer: C

Explanation:

QUESTION NO: 14

A confusion matrix is created for data that were oversampled due to a rare target.

What values are not affected by this oversampling?

A. Sensitivity and PV+
B. Specificity and PV-
C. PV+ and PV-
D. Sensitivity and Specificity

Answer: D
Explanation:

QUESTION NO: 15

This question will ask you to provide missing code segments.

A logistic regression model was fit on a data set where 40% of the outcomes were events (TARGET=1) and 60% were non-events (TARGET=0). The analyst knows that the population where the model will be deployed has 5% events and 95% non-events. The analyst also knows that the company's profit margin for correctly targeted events is nine times higher than the company's loss for incorrectly targeted non-event.

Given the following SAS program:

```
proc logistic data = LOANS descending;
  model Purch =  Inc Edu;
  score data = LOANS_V out = LOANS_VS priorevent = (insert X here);
  run;
data LOANS_VS;  set LOANS_VS;
  Solicit = P_1 > (insert Y here);
run;
```

What X and Y values should be added to the program to correctly score the data?

A. X=40, Y=10
B. X=.05, Y=10
C. X=.05, Y=.40
D. X=.10.Y=05

Answer: B
Explanation:

QUESTION NO: 16

An analyst has a sufficient volume of data to perform a 3-way partition of the data into training, validation, and test sets to perform honest assessment during the model building process.

What is the purpose of the test data set?

A. To provide a unbiased measure of assessment for the final model.
B. To compare models and select and fine-tune the final model.
C. To reduce total sample size to make computations more efficient.
D. To build the predictive models.

Answer: A
Explanation:

QUESTION NO: 17

Refer to the confusion matrix:

		Predicted Outcome	
		0	1
Actual Outcome	0	58	44
	1	23	25

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:

QUESTION NO: 18

The total modeling data has been split into training, validation, and test data. What is the best data to use for model assessment?

A. Training data
B. Total data
C. Test data
D. Validation data

Answer: D

Explanation:

QUESTION NO: 19

What is a drawback to performing data cleansing (imputation, transformations, etc.) on raw data prior to partitioning the data for honest assessment as opposed to performing the data cleansing after partitioning the data?

A. It violates assumptions of the model.
B. It requires extra computational effort and time.
C. It omits the training (and test) data sets from the benefits of the cleansing methods.
D. There is no ability to compare the effectiveness of different cleansing methods.

Answer: D

Explanation:

QUESTION NO: 20

A company has branch offices in eight regions. Customers within each region are classified as either "High Value" or "Medium Value" and are coded using the variable name VALUE. In the last year, the total amount of purchases per customer is used as the response variable.

Suppose there is a significant interaction between REGION and VALUE. What can you conclude?

A. More high value customers are found in some regions than others.
B. The difference between average purchases for medium and high value customers depends on the region.
C. Regions with higher average purchases have more high value customers.
D. Regions with higher average purchases have more medium value customers.

Answer: B

Explanation:

QUESTION NO: 21

This question will ask you to provide a missing option.

Complete the following syntax to test the homogeneity of variance assumption in the GLM procedure:

Means Region / <insert option here> =levene;

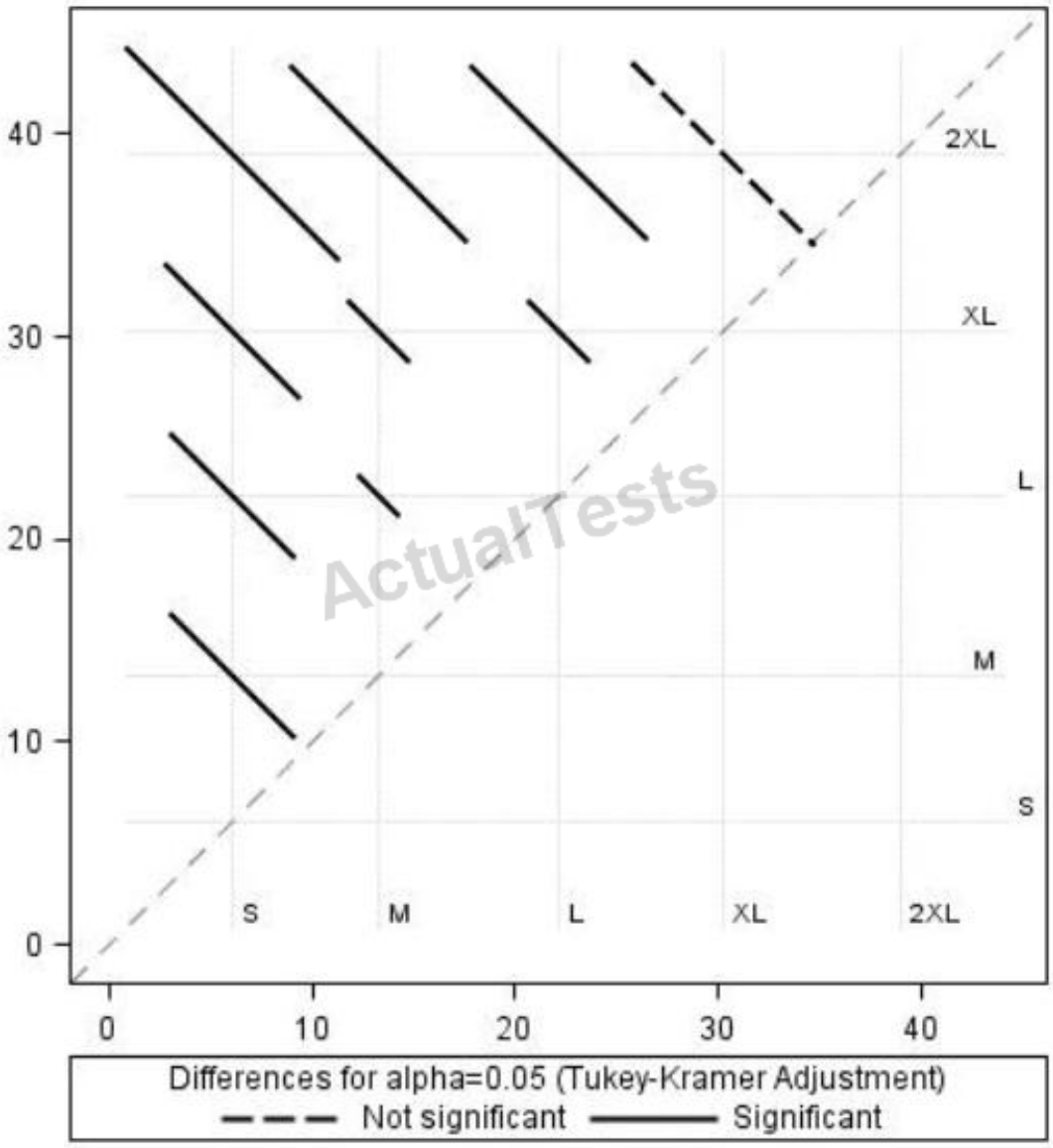
A. test
B. adjust
C. var
D. hovtest

Answer: D

Explanation:

QUESTION NO: 22

Refer to the exhibit.



Based on the control plot, which conclusion is justified regarding the means of the response?

A. All groups are significantly different from each other.
B. 2XL is significantly different from all other groups.
C. Only XL and 2XL are not significantly different from each other.
D. No groups are significantly different from each other.

Answer: C
Explanation:

QUESTION NO: 23

Customers were surveyed to assess their intent to purchase a product. An analyst divided the customers into groups defined by the company's pre-assigned market segments and tested for

difference in the customers' average intent to purchase. The following is the output from the GLM procedure:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	15716.87902	2245.26843	64.98	<.0001
Error	146	5044.56579	34.55182		
Corrected Total	153	20761.44481			

What percentage of customers' intent to purchase is explained by market segment?

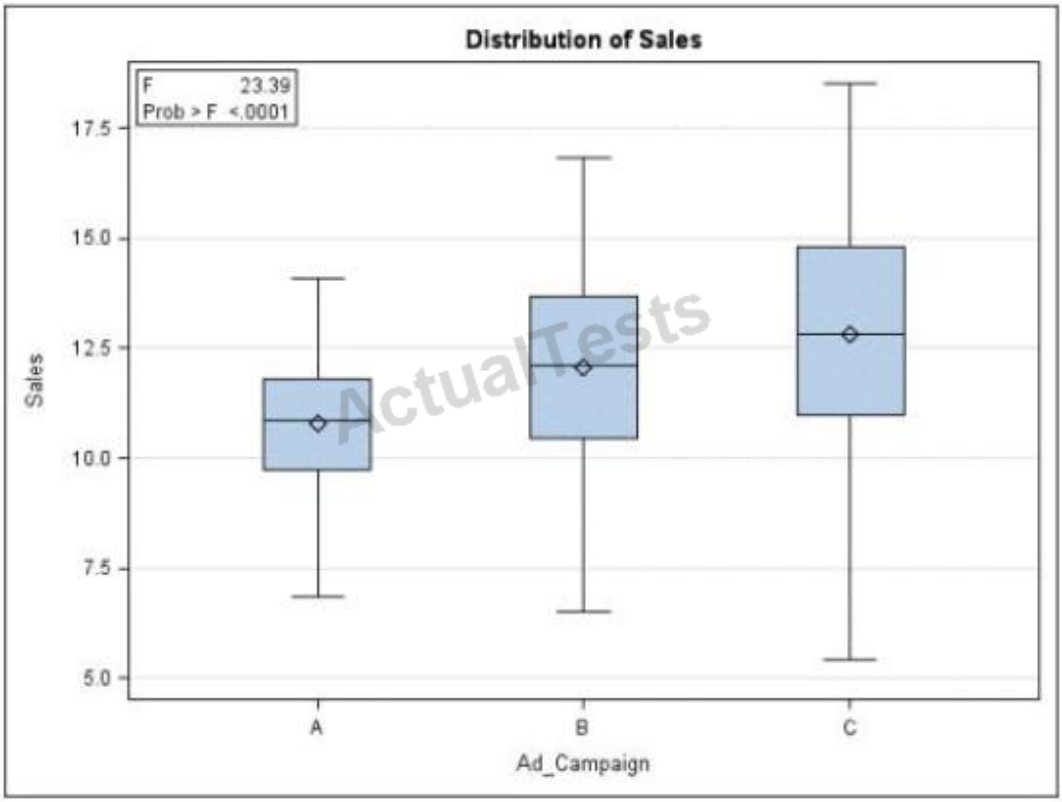
Click the calculator button to display a calculator if needed.

A. <0.01%
B. 35%
C. 65%
D. 76%

Answer: D
Explanation:

QUESTION NO: 24

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.

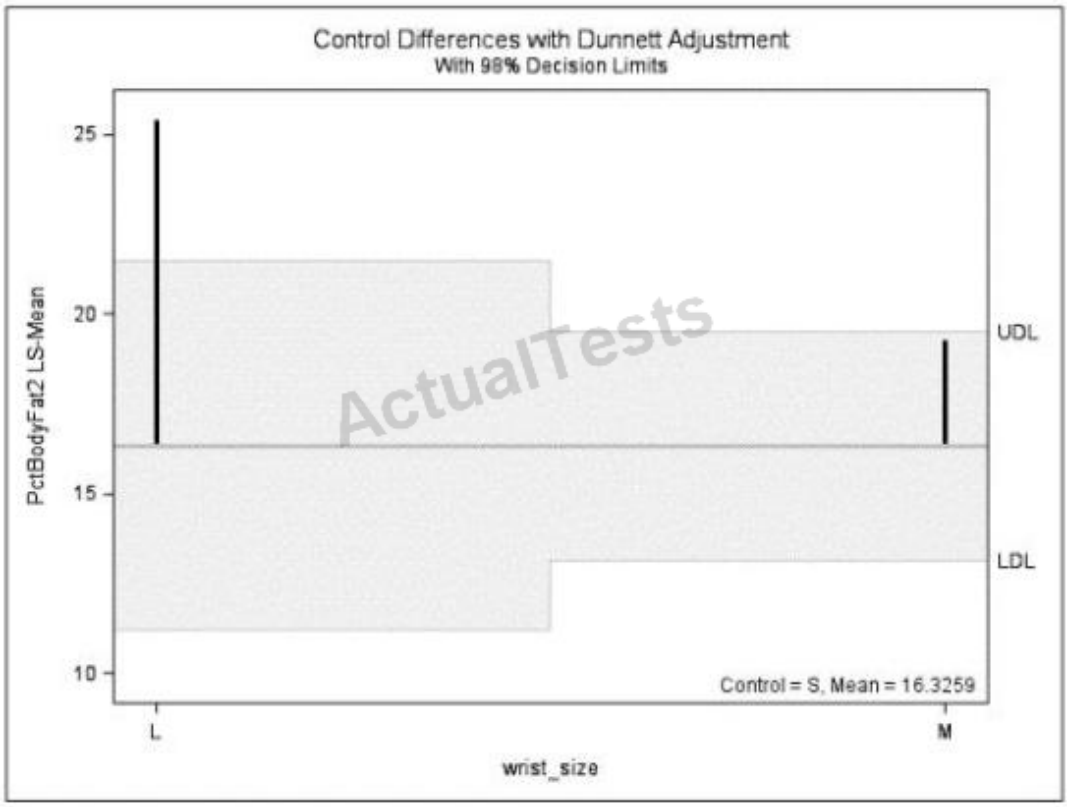
Which assumption has been violated and why?

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:

QUESTION NO: 25

Refer to the exhibit.



Given alpha=0.02, which conclusion is justified regarding percentage of body fat, comparing small (S), medium (M), and large (L) wrist sizes?

A. Medium wrist size is significantly different than small wrist size.
B. Large wrist size is significantly different than medium wrist size.
C. Large wrist size is significantly different than small wrist size.
D. There is no significant difference due to wrist size.

Answer: C
Explanation:

QUESTION NO: 26

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

B.
```
proc ttest data = SALARY;
   class Gender;
   var Pay;
run;
```

C.
```
proc glm data = SALARY;
   class Pay;
   model Pay = Gender;
run;
```

D.
```
proc ttest data = SALARY;
   class Gender;
   model Pay = Gender;
run;
```


A. Option A
B. Option B
C. Option C
D. Option D

Answer: A,B

Explanation:

QUESTION NO: 27

Given the following GLM procedure output:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
School	3	17905.24929	5968.41643	4.14	0.0073
Gender	1	1578.63006	1578.63006	1.09	0.2971
School*Gender	3	17205.36689	5735.12230	3.97	0.0091

Which statement is correct at an alpha level of 0.05?

A. School*Gender should be removed because it is non-significant.
B. Gender should be removed because it is non-significant.
C. School should be removed because it is significant.
D. Gender should not be removed due to its involvement in the significant interaction.

Answer: D
Explanation:

QUESTION NO: 28

There are missing values in the input variables for a regression application.

Which SAS procedure provides a viable solution?

A. GLM
B. VARCLUS
C. STDI2E
D. CLUSTER

Answer: C
Explanation:

QUESTION NO: 29

Screening for non-linearity in binary logistic regression can be achieved by visualizing:

A. A scatter plot of binary response versus a predictor variable.
B. A trend plot of empirical logit versus a predictor variable.

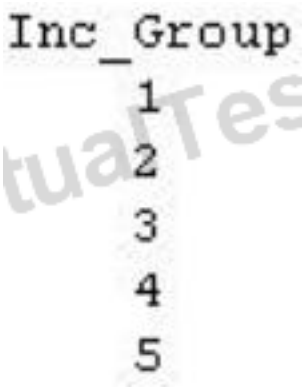
C. A logistic regression plot of predicted probability values versus a predictor variable.
D. A box plot of the odds ratio values versus a predictor variable.

Answer: B

Explanation:

QUESTION NO: 30

Given the following SAS data set TEST:



Which SAS program is NOT a correct way to create dummy variables?

A.
```
data DUMMY_TEST1;
  set TEST;
  Inc_Group1=(Inc_Group=1);
  Inc_Group2=(Inc_Group=2);
  Inc_Group3=(Inc_Group=3);
  Inc_Group4=(Inc_Group=4);
  Inc_Group5=(Inc_Group=5);
run;
```

B.
```
data DUMMY_TEST1;
  set TEST;
  if Inc_Group=1 then Inc_Group1=1;
    else Inc_Group1=0;
  if Inc_Group=2 then Inc_Group2=1;
    else Inc_Group2=0;
  if Inc_Group=3 then Inc_Group3=1;
    else Inc_Group3=0;
  if Inc_Group=4 then Inc_Group4=1;
    else Inc_Group4=0;
  if Inc_Group=5 then Inc_Group5=1;
    else Inc_Group5=0;
run;
```

C.
```
data DUMMY_TEST1 (drop=i);
  set TEST;
  array inc(*) Inc_Group1 -  Inc_Group5;
  do i = 1 to 5;
    inc(i) = ( Inc_Group = i );
  end;
run;
```

D.
```
data DUMMY_TEST1 (drop=i);
  set TEST;
  array inc(*) Inc_Group1 Inc_Group2 Inc_Group3
               Inc_Group4 Inc_Group5;
  do i = 1 to 5;
    ( Inc_Group = i );
  end;
run;
```

A. Option A
B. Option B
C. Option C
D. Option D

Answer: D

Explanation:

QUESTION NO: 31

An analyst fits a logistic regression model to predict whether or not a client will default on a loan. One of the predictors in the model is agent, and each agent serves 15-20 clients each. The model fails to converge. The analyst prints the summarized data, showing the number of defaulted loans per agent. See the partial output below:

Obs	agent	clients	defaults
1	1	17	12
2	2	19	0
3	3	16	7
4	4	15	5
5	5	19	13
6	6	17	8
7	7	16	9
8	8	17	10
9	9	17	11
10	10	16	8

What is the most likely reason that the model fails to converge?

A. There is quasi-complete separation in the data.
B. There is collinearity among the predictors.
C. There are missing values in the data.

D. There are too many observations in the data.

Answer: A

Explanation:

QUESTION NO: 32

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:

QUESTION NO: 33

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:

QUESTION NO: 34

An analyst investigates Region (A, B, or C) as an input variable in a logistic regression model.

The analyst discovers that the probability of purchasing a certain item when Region = A is 1.

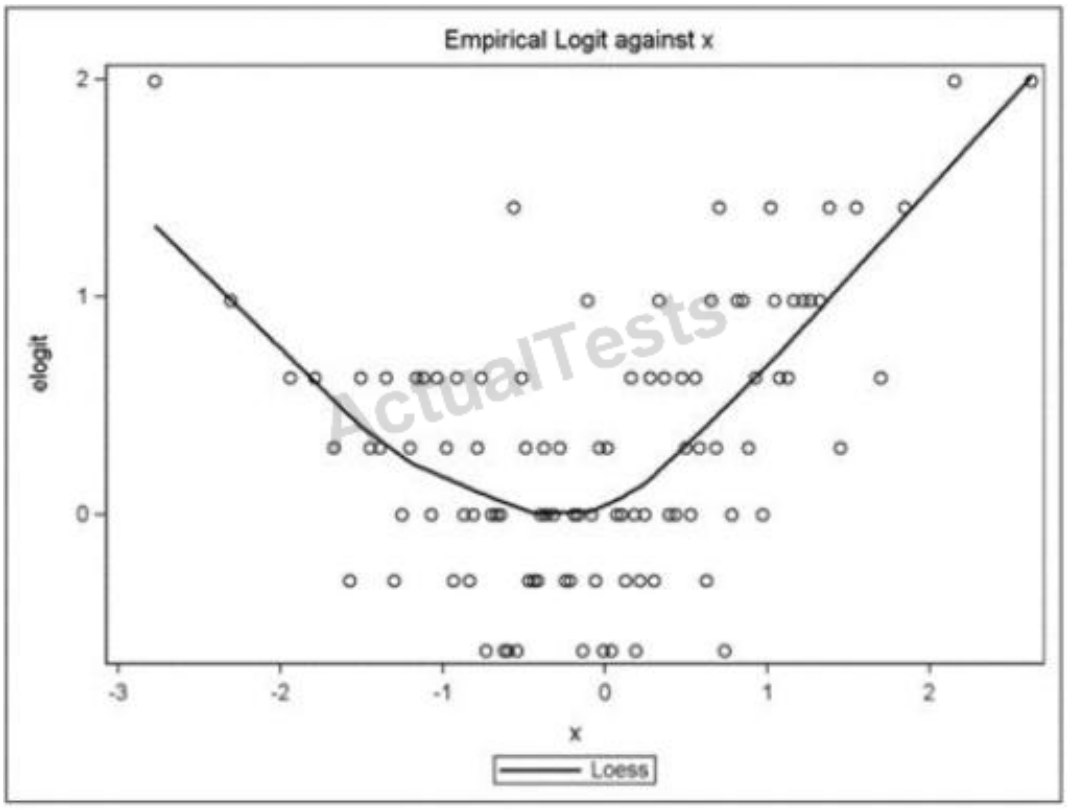
What problem does this illustrate?

A. Collinearity
B. Influential observations
C. Quasi-complete separation
D. Problems that arise due to missing values

Answer: C
Explanation:

QUESTION NO: 35

Refer to the following exhibit:



What is a correct interpretation of this graph?

A. The association between the continuous predictor and the binary response is quadratic.
B. The association between the continuous predictor and the log-odds is quadratic.
C. The association between the continuous predictor and the continuous response is quadratic.

D. The association between the binary predictor and the log-odds is quadratic.

Answer: B

Explanation:

QUESTION NO: 36

This question will ask you to provide a missing option. Given the following SAS program:

```
proc corr data = MYDATA <insert option here> ;
  var x1 x2 x3 x4 x5;
  with Target;
run;
```

What option must be added to the program to obtain a data set containing Pearson statistics?

A. OUTPUT=estimates
B. OUTP=estimates
C. OUTSTAT=estimates
D. OUTCORR=estimates

Answer: B

Explanation:

QUESTION NO: 37

A predictive model uses a data set that has several variables with missing values.

What two problems can arise with this model? (Choose two.)

A. The model will likely be overfit.
B. There will be a high rate of collinearity among input variables.
C. Complete case analysis means that fewer observations will be used in the model building process.
D. New cases with missing values on input variables cannot be scored without extra data processing.

Answer: C,D

Explanation:

QUESTION NO: 38

Spearman statistics in the CORR procedure are useful for screening for irrelevant variables by investigating the association between which function of the input variables?

A. Concordant and discordant pairs of ranked observations
B. Logit link (log(p/1-p))
C. Rank-ordered values of the variables
D. Weighted sum of chi-square statistics for 2x2 tables

Answer: C
Explanation:

QUESTION NO: 39

A non-contributing predictor variable (Pr > |t| =0.658) is added to an existing multiple linear regression model.

What will be the result?

A. An increase in R-Square
B. A decrease in R-Square
C. A decrease in Mean Square Error
D. No change in R-Square

Answer: A
Explanation:

QUESTION NO: 40

The standard form of a linear regression model is:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Which statement best summarizes the assumptions placed on the errors?

A. The errors are correlated, normally distributed with constant mean and zero variance.
B. The errors are correlated, normally distributed with zero mean and constant variance.
C. The errors are independent, normally distributed with constant mean and zero variance.
D. The errors are independent, normally distributed with zero mean and constant variance.

Answer: D
Explanation:

QUESTION NO: 41

Refer to the REG procedure output:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	33033	11011	115.63	<.0001
Error	496	47231	95.22454		
Corrected Total	499	80265			

Click on the calculator button to display a calculator if needed.

A. 0.4115
B. 0.6994
C. 0.5884
D. 0.1372

Answer: A
Explanation:

QUESTION NO: 42

Identify the correct SAS program for fitting a multiple linear regression model with dependent

variable (y) and four predictor variables (x1-x4).

A.
```
proc reg data=SASUSER.MLR;
   var y x1 x2 x3 x4;
   model y = x1-x4;
run;
```

B.
```
proc reg data=SASUSER.MLR;
   model y = x1-x4;
run;
```

C.
```
proc reg data=SASUSER.MLR;
   model y = x1;
   model y = x2;
   model y = x3;
   model y = x4;
run;
```

D.
```
proc reg data=SASUSER.MLR;
   model y = x1 x2 x3 x4 /solution;
run;
```


A. Option A
B. Option B
C. Option C
D. Option D

Answer: B

Explanation:

QUESTION NO: 43

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:

QUESTION NO: 44

The selection criterion used in the forward selection method in the REG procedure is:

A. Adjusted R-Square
B. SLE
C. Mallows' Cp
D. AIC

Answer: B

Explanation:

QUESTION NO: 45

Which SAS program will correctly use backward elimination selection criterion within the REG procedure?

A.
```
proc reg data=SASUSER.MLR;
   model y = x1-x10 /selection=backward sls=aic;
run;
```

B.
```
proc reg data=SASUSER.MLR;
   model y = x1-x10 /selection=backward sls=0.15;
run;
```

C.
```
proc reg data=SASUSER.MLR;
   model y = x1-x10 /selection=backward sle=cp;
run;
```

D.
```
proc reg data=SASUSER.MLR;
   model y = x1-x10 /selection=backward sle=all;
run;
```


A. Option A
B. Option B
C. Option C
D. Option D

Answer: B

Explanation:

QUESTION NO: 46

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:

QUESTION NO: 47

Refer to the exhibit:

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	SBC	Variables in Model
1	0.7434	0.7345	13.6988	64.5341	2.74478	67.40210	RunTime
1	0.1595	0.1305	106.3021	101.3131	4.96748	104.18108	RestPulse
2	0.7642	0.7474	12.3894	63.9050	2.67739	68.20695	Age RunTime
2	0.7614	0.7444	12.8372	64.2740	2.69337	68.57597	RunTime RunPulse
3	0.8111	0.7901	6.9596	59.0373	2.44063	64.77326	Age RunTime RunPulse
3	0.8100	0.7889	7.1350	59.2183	2.44777	64.95424	RunTime RunPulse MaxPulse
4	0.8368	0.8117	4.8800	56.4995	2.31159	63.66941	Age RunTime RunPulse MaxPulse
4	0.8165	0.7883	8.1035	60.1386	2.45133	67.30850	Age Weight RunTime RunPulse
5	0.8480	0.8176	5.1063	56.2986	2.27516	64.90250	Age Weight RunTime RunPulse MaxPulse
5	0.8370	0.8044	6.8461	58.4590	2.35583	67.06288	Age RunTime RunPulse RestPulse MaxPulse
6	0.8487	0.8108	7.0000	58.1616	2.31695	68.19952	Age Weight RunTime RunPulse RestPulse MaxPulse

SAS output from the RSOUARE selection method, within the REG procedure, is shown. The top two models in each subset are given.

Based on the AIC statistic, which model is the champion model?

A. Age Weight RunTime RunPulse MaxPulse
B. Age Weight RunTime RunPulse RestPulse MaxPulse
C. RestPulse
D. RunTime

Answer: A

Explanation:

QUESTION NO: 48 CORRECT TEXT

A linear model has the following characteristics:

*A dependent variable (y)

*One continuous variable (xl), including a quadratic term (x12)

*One categorical (d with 3 levels) predictor variable and an interaction term (d by x1)

How many parameters, including the intercept, are associated with this model?

Enter your numeric answer in the space below. Do not add leading or trailing spaces to your answer.

Answer: 7

QUESTION NO: 49

A linear model has the following characteristics:

- A dependent variable (y)
- Three continuous predictor variables (x1-x3)
- One categorical predictor variable (c1with 3 levels)

Which SAS program fits this model?

A.
```
proc glm data=SASUSER.MLR;
   class c1 x1 x2 x3;
   model y = c1 x1-x3 /solution;
run;
```

B.
```
proc reg data=SASUSER.MLR;
   model y = c1 x1-x3 /solution;
run;
```

C.
```
proc reg data=SASUSER.MLR;
   class c1;
   model y = c1 x1-x3;
run;
```

D.
```
proc glm data=SASUSER.MLR;
   class c1;
   model y = c1 x1-x3 /solution;
run;
```


A. Option A
B. Option B
C. Option C
D. Option D

Answer: D

Explanation:

QUESTION NO: 50

Which SAS program will detect collinearity in a multiple regression application?

A.
```
proc reg data = SASUSER.RETAIL;
   model Purchase = Gender Age Income / lackfit;
run;
```

B.
```
proc reg data = SASUSER.RETAIL;
   model Purchase = Gender Age Income / vif;
run;
```

C.
```
proc reg data=SASUSER.RETAIL plots(only)=(COOKSD);
   model Purchase = Gender Age Income;
run;
```

D.
```
proc reg data=sasuser.retail plots(only)=(RSTUDENTBYPREDICTED);
   model Purchase = Gender Age Income;
run;
```

A. Option A
B. Option B
C. Option C
D. Option D

Answer: B

Explanation:

QUESTION NO: 51

Refer to the following odds ratio table:

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
salary	1.0000	1.142	1.083	1.220

What is a correct interpretation of the estimate?

A. The odds of the event are 1.142 greater for each one dollar increase in salary.
B. The odds of the event are 1.142 greater for each one thousand dollar increase in salary.
C. The probability of the event is 1.142 greater for each one dollar increase in salary.
D. The probability of the event is 1.142 greater for each one thousand dollar increase in salary.

Answer: B

Explanation:

QUESTION NO: 52

Which method is NOT an appropriate way to score new observations with a known target in a logistic regression model?

A. Use the SCORE statement in the LOGISTIC procedure.
B. Augment the training data set with new observations and set their responses to missing.
C. Augment the training data set with new observations and rerun the LOGISTIC procedure.
D. Use the saved parameter estimates from the LOGISTIC procedure and score new observations in the SCORE procedure.

Answer: C

Explanation:

QUESTION NO: 53

Consider scoring new observations in the SCORE procedure versus the SCORE statement in the LOGISTIC procedure.

Which statement is true?

A. The SCORE statement in the LOGISTIC procedure returns only predicted probabilities, whereas the SCORE procedure returns only predicted logits.
B. The SCORE statement in the LOGISTIC procedure returns only predicted logits, whereas the SCORE procedure returns only predicted probabilities.
C. Unlike the SCORE procedure, the SCORE statement in the LOGISTIC procedure produces both predicted probabilities and predicted logits.
D. The SCORE procedure and the SCORE statement in the LOGISTIC procedure produce the same output.

Answer: A

Explanation:

QUESTION NO: 54

Select the equivalent LOGISTIC procedure model statements. (Choose two.)

A. Mode1 Purchase * Gender Age Region;
B. Mode1 Purchase * Gender | Age | Region;
C. Mode1 Purchase * Gender|Age|Region @1;
D. Mode1 Purchase * Gender|Age|Region @2;

Answer: A,C

Explanation:

QUESTION NO: 55

Given the following LOGISTIC procedure:

```
proc logistic data = MYDIR.CONVERT des outest=OUTFILE_1;
    model  Attrite = Calls Plan Billing_code;
    score data=MYDIR.NEW_ATTRITE_DATA out=OUTFILE_2;
run;
```

What is the difference between the datasets OUTFILEJ and OUTFILE_2?

A. OUTFILE_1 contains the final parameter estimates while OUTFILE_2 contains the newly scored probabilities.
B. OUTFILE_1 contains the model goodness of fit statistics while OUTFILE_2 contains the newly scored probabilities
C. OUTFILE_1 contains the model goodness of fit statistics while OUTFILE_2 contains the newly scored logits.
D. OUTFILEJ contains the final parameter estimates and Wald Chi-Square values while OUTFILE_2 contains the newly scored probabilities.

Answer: A

Explanation:

QUESTION NO: 56

The following LOGISTIC procedure output analyzes the relationship between a binary response and an ordinal predictor variable, wrist_size Using reference cell coding, the analyst selects Large (L) as the reference level.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.0415	0.4749	4.8101	0.0283
wrist_size	M	1	1.1234	0.4989	5.0697	0.0243
wrist_size	S	1	1.6078	0.5478	8.6133	0.0033

What is the estimated logit for a person with large wrist size?

Click the calculator button to display a calculator if needed.

A. 0.0819
B. 0.5663
C. -3.7727
D. -1.0415

Answer: D
Explanation:

QUESTION NO: 57

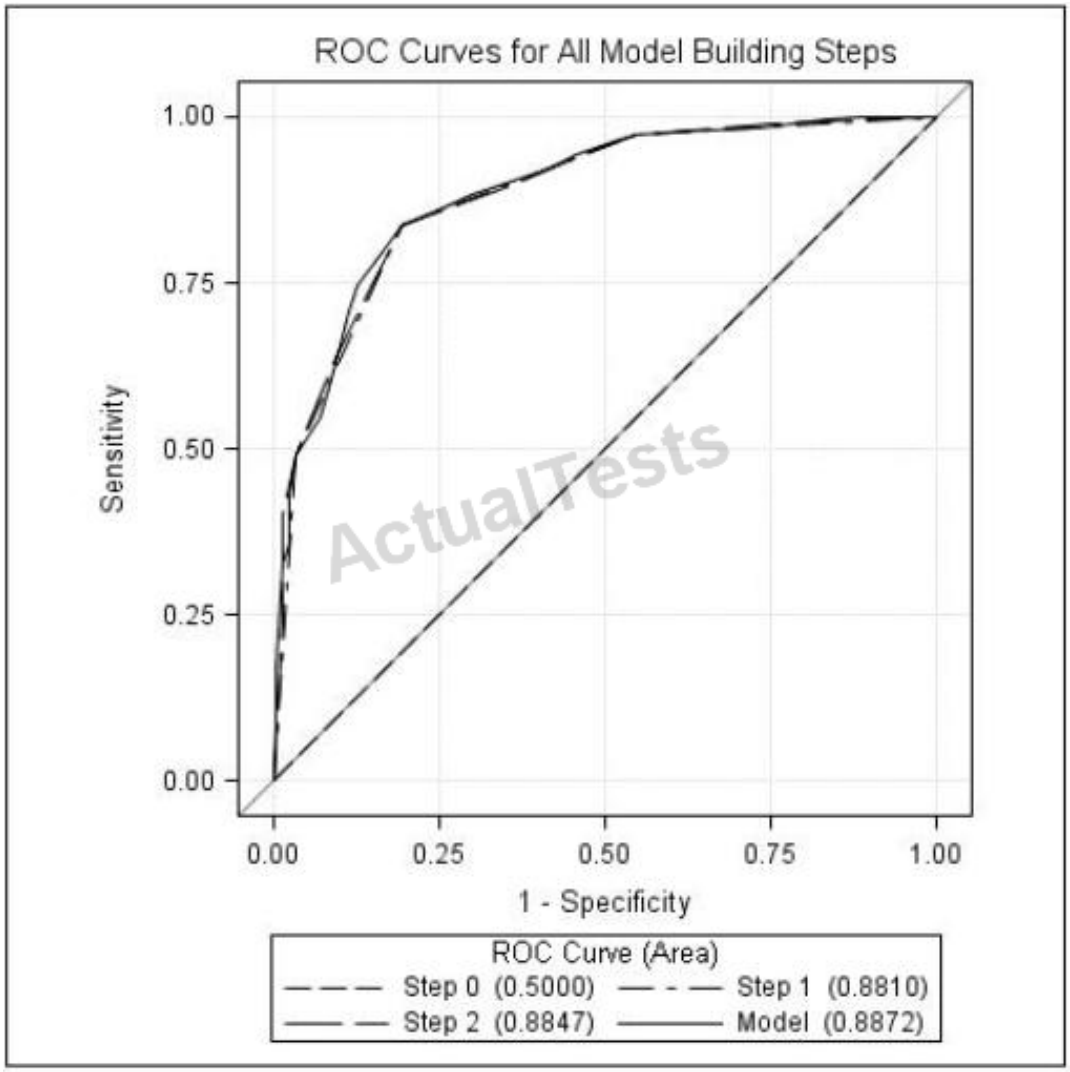
Which of the following describes a concordant pair of observations in the LOGISTIC procedure?

A. An observation with the event has an equal probability as another observation with the event.
B. An observation with the event has a lower predicted probability than the observation without the event.
C. An observation with the event has an equal predicted probability as the observation without the event.
D. An observation with the event has a higher predicted probability than the observation without the event

Answer: D
Explanation:

QUESTION NO: 58

Refer to the exhibit:



An analyst examined logistic regression models for predicting whether a customer would make a purchase. The ROC curve displayed summarizes the models. Using the selected model and the analyst's decision rule, 25% of the customers who did not make a purchase are incorrectly classified as purchasers.

What can be concluded from the graph?

A. About 25% of the customers who did make a purchase are correctly classified as making a purchase.
B. About 50% of the customers who did make a purchase are correctly classified as making a purchase.
C. About 85% of the customers who did make a purchase are correctly classified as making a purchase.
D. About 95% of the customers who did make a purchase are correctly classified as making a purchase.

Answer: C

Explanation:

QUESTION NO: 59

One common approach for predicting rare events in the LOGISTIC procedure is to build a model that disproportionately over-re presents those cases with an event occurring (e.g. a 50-50 event/non-event split).

What problem does this present?

A. All parameter estimates are biased.
B. Only the intercept estimate is biased.
C. Only the non-intercept parameter estimates are biased.
D. Sensitivity estimates are biased.

Answer: B
Explanation:

QUESTION NO: 60

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:

QUESTION NO: 61

Which statistic, calculated from a validation sample, can help decide which model to use for prediction of a binary target variable?

A. Adjusted R Square
B. Mallow's Cp
C. Chi Square

D. Average Squared Error

Answer: D

Explanation:

QUESTION NO: 62

The question will ask you to provide a missing statement. Given the following SAS program:

```
proc logistic data = MYDIR.DEFAULT_DATA des;
   model Purchase = Money Acct_type Debt Employment;
   <insert statement here>
run;
```

Which SAS statement will complete the program to correctly score the data set NEW_DATA?

A. Scoredata data=MYDIR.NEW_DATA out=scores;
B. Scoredata data=MYDIR.NEW_DATA output=scores;
C. Scoredata=HYDIR.NEU_DATA output=scores;
D. Scoredata=MYDIR,NEW DATA out=scores;

Answer: D

Explanation:

QUESTION NO: 63

A marketing manager attempts to determine those customers most likely to purchase additional products as the result of a nation-wide marketing campaign.

The manager possesses a historical dataset (CAMPAIGN) of a similar campaign from last year.

It has the following characteristics:

- Target variable Respond (0,1)
- Continuous predictor Income
- Categorical predictor Homeowner(Y,N)

Which SAS program performs this analysis?

A.
```
proc logistic data=MYDIR.CAMPAIGN descending;
   class Homeowner;
   model Respond = Income Homeowner;
run;
```

B.
```
proc logistic data = MYDIR.CAMPAIGN descending;
   by Homeowner;
   model Respond = Income Homeowner;
run;
```

C.
```
proc logistic data = MYDIR.CAMPAIGN descending;
   model Respond = Income Homeowner;
run;
```

D.
```
proc logistic data = MYDIR.CAMPAIGN descending;
   class Income Homeowner;
   model Respond = Income Homeowner;
run;
```

A. Option A
B. Option B
C. Option C
D. Option D

Answer: A

Explanation:

QUESTION NO: 64

Given the following output from the LOGISTIC procedure:

Parameter	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Estimate
Intercept	1	0.1119	0.0304	24.8969	<.0001	
MBA	1	0.9699	0.0385	633.6092	<.0001	0.2074
DOWN_AMT	1	0.000072	3.39E-6	448.5988	<.0001	-0.2883
CASH	1	-0.5629	0.1145	24.1615	<.0001	-0.0408
HOME	1	-0.00402	0.00317	1.6168	0.2035	-0.0114

Which variables, among those that are statistically significant at an alpha of 0.05, have the greatest and least relative importance on the fitted model?

A. Greatest: MBA

Least: DOWN_AMT
B. Greatest: MBA
Least: CASH
C. Greatest: DOWN_AMT
Least: CASH
D. Greatest: DOWN_AMT
Least: HOME

Answer: C

Explanation:

QUESTION NO: 65

What is the default method in the LOGISTIC procedure to handle observations with missing data?

A. Missing values are imputed.
B. Parameters are estimated accounting for the missing values.
C. Parameter estimates are made on all available data.
D. Only cases with variables that are fully populated are used.

Answer: D