Print

Quiz Feedback: Preparing the Input Variables

Your Score: 100%

Congratulations! Your score of 100% indicates that you've mastered the topics in this lesson. If you'd like, you can review the feedback.

When you're finished, exit the lesson.

1. A continuous variable, **Household_Income**, has 85% missing values. In predictive modeling, which strategy would be the most reasonable to handle the missing values?

   - *a.* Create a missing indicator variable only and do not use **Household_Income** in the model.
   - *b.* Use median imputation to replace the missing values.
   - *c.* Use mean imputation to replace the missing values.
   - *d.* Use cluster imputation to replace the missing values.

   Your answer: a
   Correct answer: a

   If a very large percentage of values is missing, then the variable would be better handled by omitting it from the analysis and creating the missing indicator variable only.

---

2. The following table was created in PROC CLUSTER using Greenacre's method. Each of the six levels (B1 through B6) has 100 observations.

Number of Clusters	Clusters Joined	Clusters Joined	Freq	Semipartial R-Square	R-Square
5	B1	B2	200	0.0000	1.00
4	B3	B4	200	0.0005	.999
3	B5	B6	200	0.0150	.985
2	CL5	CL3	400	0.1500	.834
1	CL4	CL2	500	0.834	.000

   Which of the following statements correctly describes the two-cluster solution?

   - *a.* B1 and B2 were combined with B3. 83.4% of the variability is explained by the clusters.
   - *b.* B1 and B2 were combined with B5 and B6. The two-cluster solution (B1, B2, B5, B6 versus B3, B4) has a chi-square that is 83.4% of the original chi-square.
   - *c.* B1 and B2 were combined with B5 and B6. 15% of the total chi-square remains after the levels are collapsed.
   - *d.* B5 and B3 were combined, and the total chi-square changed 15%.

   Your answer: b

Correct answer: b

In the two-cluster solution, CL5 (the results of the five-cluster solution, in which B1 joins B2) combines with CL3 (the results of the three-cluster solution, in which B5 joins B6). The R-square is 0.834, which is the proportion of the original chi-square remaining after the clusters are collapsed.

---

3. The following cluster was created in PROC VARCLUS:

Cluster	Variable	1-R**2 Ratio
Cluster 3	MEDIAN_HOME_VALUE	0.2957
	MEDIAN_HOUSEHOLD_INCOME	0.2358
	PER_CAPITA_INCOME	0.1945
	NSES1	0.5374

Which variable should be selected based on the 1-R^2 ratio?

- ○ *a.* **Median_Home_Value**
- ○ *b.* **Median_Household_Income**
- ○ *c.* **Per_Capita_Income**
- ○ *d.* **NSES1**

Your answer: c
Correct answer: c

The variable **Per_Capita_Income** has the lowest 1-R^2 ratio. This indicates that the variable has the highest correlation with its own cluster and the lowest correlation with the other clusters.

---

4. Which of the following situations indicates a nonlinear and non-monotonic relationship with the target, where a high rank indicates a strong association?

- ○ *a.* high Spearman rank and high Hoeffding rank
- ○ *b.* low Spearman rank and high Hoeffding rank
- ○ *c.* high Spearman rank and low Hoeffding rank
- ○ *d.* low Spearman rank and low Hoeffding rank

Your answer: b
Correct answer: b

A variable with a low Spearman rank and a high Hoeffding rank indicates a nonlinear and non-monotonic relationship with the target.

5. Which statement is true regarding empirical logit plots?

- ○ *a.* Logit plots can be used with continuous target variables.
- ○ *b.* When there are no events in the bin, you set the logit to 0.
- ○ *c.* For continuous predictor variables, the plots should be fairly linear if the assumptions of the standard logistic regression model were met.

○ *d.* Adding a small constant to the numerator of the empirical logit formula eliminates the problem caused by zero counts.

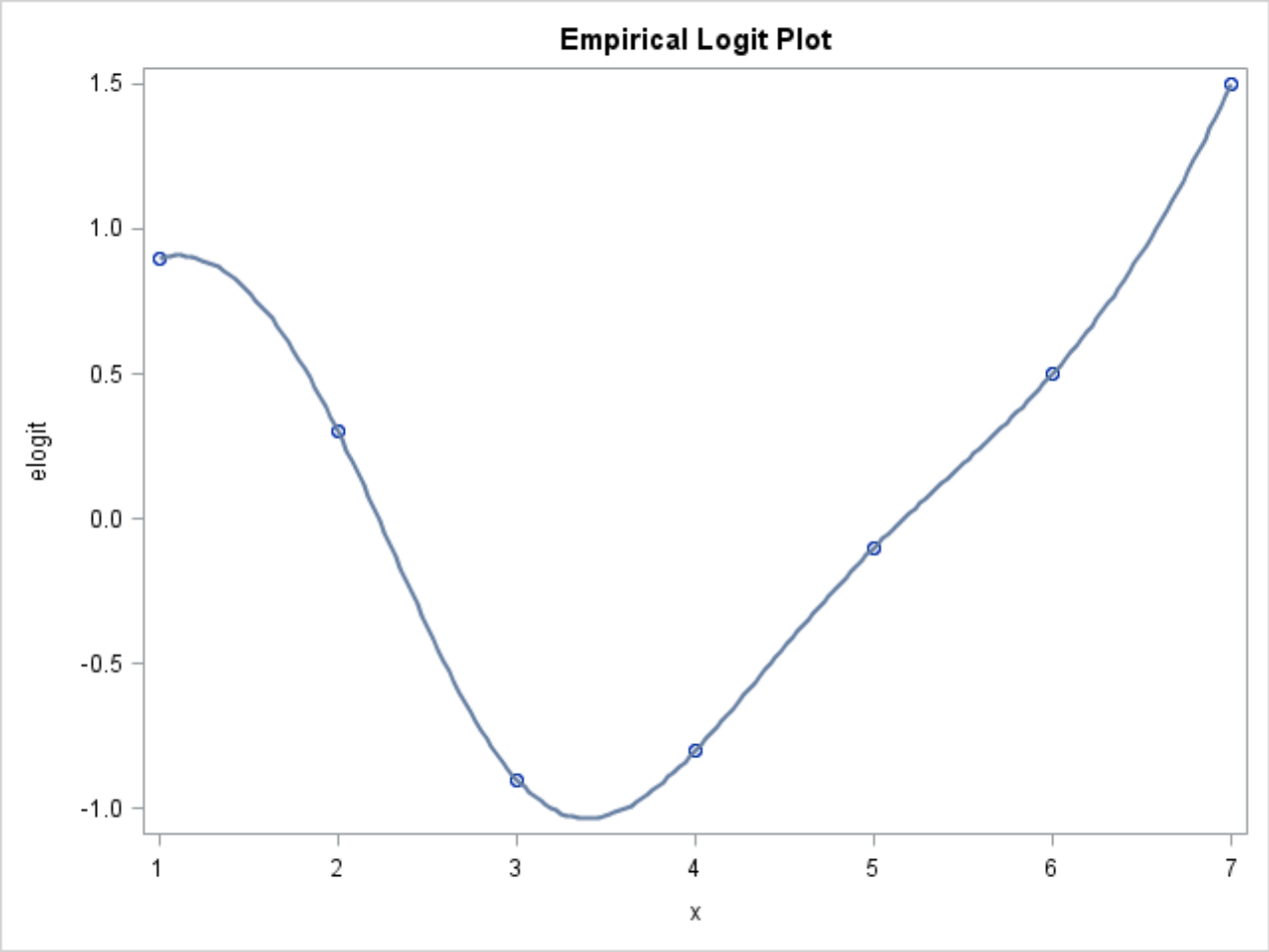
Your answer: c
Correct answer: c

Answer *a* is incorrect because logit plots are used only with categorical target variables.

Answer *b* is incorrect because the logit is the log(p/1-p) and would be negative infinity if there were no events in the bin. To eliminate the problem caused by zero counts, a small nonzero constant is added to the numerator and denominator of the empirical logit formula.

Answer *d* is incorrect because a small nonzero constant is added to the numerator and denominator of the empirical logit formula, not only the numerator.

✔ 6. Consider the following empirical logit plot for the variable **X**:



Which of the following logistic regression models is most likely to represent the relationship between the target and the input variable?

○ *a.* Logit(p) = $\beta_0+\beta_1*X$
○ *b.* Logit(p) = $\beta_0+\beta_1*X+\beta_2*X^2$
○ *c.* Logit(p) = $\beta_0+\beta_1*\log(X)$
○ *d.* Logit(p) = $\beta_0+\beta_1*(1/X)$

Your answer: b
Correct answer: b

The plot indicates a quadratic relationship in which the input variable and its squared term would be the best fit.

---

7. Which statement is true regarding the subset selection methods in PROC LOGISTIC?

- ○ *a.* For the final model, the backwards elimination method yields exactly the same parameter estimates as the fast backwards elimination method requested by the FAST option.
- ○ *b.* The models selected in the best-subsets selection method are ranked by the score chi-square statistic.
- ○ *c.* The best-subsets selection method provides parameter estimates for each model.
- ○ *d.* The stepwise selection method is the most efficient algorithm when you have a large number of predictor variables.

Your answer: b
Correct answer: b

Answer *a* is incorrect because the FAST option provides only approximations to the regression coefficients. If you compare models selected by backward elimination with and without the FAST option, you see different regression coefficients.

Answer *c* is incorrect because the output in PROC LOGISTIC for the best-subsets selection method gives only the number of variables in the model, the score chi-square, and the variable names in the model.

Answer *d* is incorrect because the stepwise selection method is very inefficient when there are a large number of predictor variables. Simulations conducted by the authors showed that backward elimination with the FAST option was more efficient than stepwise for any number of inputs.

Before answering questions 8-10, perform the following steps in SAS:

Note: Make sure you set up the course files before you continue.

a. Copy and paste the following code into the editor:

```
data work.pva(drop=CONTROL_NUMBER MONTHS_SINCE_LAST_PROM_RESP
              FILE_AVG_GIFT FILE_CARD_GIFT);
   set pmlr.pva_raw_data;
run;

title "Models Selected by Backward Selection";
proc logistic data=work.pva;
   model target_b(event='1')=recent_response_prop
         recent_avg_gift_amt lifetime_card_prom
         per_capita_income pct_male_veterans
         lifetime_gift_range pct_male_military
         / selection=backward;
run;
title;
```

b. Submit the code and review the results.

8. Which predictor variable was eliminated first?

   - *a.* **Pct_Male_Military**
   - *b.* **Lifetime_Gift_Range**
   - *c.* **Pct_Male_Veterans**
   - *d.* **Per_Capita_Income**

Your answer: a
Correct answer: a

The output shows that, in step 1, the variable **Pct_Male_Military** was eliminated.

9. In Step 3 of the backward elimination method, the residual chi-square had a *p*-value of 0.2949. You can interpret this as which of the following?

   a. the significance of the variable that was removed in step 3
   b. the joint significance of the variables remaining in the model
   c. the joint significance of the following three variables that were excluded from the model: **Pct_Male_Military**, **Lifetime_Gift_Range**, and **Pct_Male_Veterans**
   d. the significance of **Pct_Male_Military**

Your answer: c
Correct answer: c

The residual chi-square tests the significance of the variables that are not in the model. By step 3, the variables **Pct_Male_Military**, **Lifetime_Gift_Range**, and **Pct_Male_Veterans** were removed from the model.

10. Which of the following is the default significance level for the backward elimination method, as shown in the output?

   - *a.* 0.01
   - *b.* 0.05
   - *c.* 0.10
   - *d.* 0.15

Your answer: b
Correct answer: b

The program does not use the SLSTAY= option. In the output, a note states that no (additional) effects met the 0.05 significance level for removal from the model, which indicates the default significance level.

Close