Questions 1-4 are based on the Conceptual question at the end of Chapter 4 in the JWHT textbook (6th edition). The scenario involves data collected about students in a graduate statistics class with variables X1 = hours studied, X2 = undergraduate GPA, and Y = 1 if the student received an A, 0 otherwise. A logistic regression is fitted, and the estimated coefficients are b0 = -6, b1 = 0.05, and b2 = 1.

Question 1

Estimate the probability that a student who studies for 30 hours and has an undergraduate GPA of 3.5 gets an A in the class.

0.2688

0.3775

0.2990

0.4750

Can’t be estimated

Question 2

Estimate the odds ratio in favor of student, who studies for 40 hours and has an undergraduate GPA of 3.5, getting an A in the class.

0.2869

0.6065

0.9040

0.3775

Can’t be estimated

Question 3

Estimate the probability that a student who studies for 40 hours and has an undergraduate GPA of 3.9 gets an A in the class.

0.2869

0.3775

0.2890

0.4750

Can’t be estimated

Question 4

How many hours should a student study with an undergraduate GPA of 3.5 to have a probability of getting an A in the class equal to 0.8?

77.72

90

62.04

53.82

Can’t be estimated

Question 5-8 are based on the Applied Problem 10 at the end of Chapter 4 in the JWHT textbook. Load the dataset named Weekly (in the ISLR package) into R. Create a new data frame (call is C1) that is a copy of Weekly for your analysis.  Assume “UP” as positive event and “DOWN” as negative event.

Question 5

Perform a logistic regression with Direction as the response variable and the five lag variables plus Volume as predictors. Use the summary function to review the results. Do any of the predictors appear to be statistically significant at alpha = 0.05? If so, which one(s)?

Volume

Lag2

Lag1, Lag2

Lag1, Lag3

Lag1, Volume

Question 6

Compute the confusion matrix with 0.5 as the cutoff value. What are the values of true negative, false positive, false negative, true positive, and accuracy?

TN=430, FP=48, FN=54, TP=557, Accuracy=46.1%.

TN=48, FP=557, FN=54, TP=430, Accuracy=43.9%.

TN=54, FP=557, FN=48, TP=430, Accuracy=44.4%.

TN=48, FP=430, FN=54, TP=557, Accuracy=55.5%.

TN=54, FP=430, FN=48, TP=557, Accuracy=56.1%.

Question 7

Fit a logistic regression model using a training data period from 1990 to 2008 with Direction as the response variable and with Lag2 as the only predictor. Compute the confusion matrix and the values of true positive, false negative, true negative, false positive for the holdout data (that is the data from 2009 to 2010). What are the values of true negative, false positive, false negative, true positive, and accuracy?

TN=34, FP=5, FN=9, TP=56, Accuracy=86.5%.

TN=5, FP=56, FN=9, TP=34, Accuracy=37.5%.

TN=9, FP=56, FN=5, TP=34, Accuracy=41.3%.

TN=5, FP=34, FN=9, TP=56, Accuracy=58.6%.

TN=9, FP=34, FN=5, TP=56, Accuracy=62.5%.

Question 8

Based on the regression model obtained in Question 7, estimate the Sensitivity and Specificity of the model.

Sensitivity = 0.918, Specificity = 0.209

Sensitivity = 0.459, Specificity = 0.209

Sensitivity = 0.918, Specificity = 0.104

Sensitivity = 0.209, Specificity = 0.918

None of these

Question 9

Which of the following is expected when cut-off probability for determining positive event in a classification problem is increased?

Specificity increases, Sensitivity decreases

Specificity decreases, Sensitivity decreases

Specificity decreases, Sensitivity increases

Specificity increases, Sensitivity increases

Sensitivity and Specificity can both increase and decrease

Question 10

Which of the following statements most closely represent the relationship between correlation and causation?

Causation always implies linear correlation

Correlation implies causation

Correlation and causation are unrelated

Correlation doesn’t imply causation

Both A and D

A study was done to understand the dependence of car prices on various car types (Sports car/ Sedan/ SUV). Researchers studied 30 different models of each car type. Assume only other factors that might affect car prices are brand of the car (Ferrari, Hyundai etc.) and how long ago the specific models were launched. This scenario applies to Questions 11 and 12.

Question 11

Which of the following is the explanatory variable in this study?

Price

Car type

Brand

Launch timeframe

B, C and D

Question 12

Which of the following is a confounding variable in this study?

Price

Car type

Brand

Launch timeframe

Both C and D

Question 13

Which of the following statements is correct about Difference-in-Difference estimation?

DID relies on assumption that the unobserved differences between treatment and control groups are the same overtime

DID makes use of longitudinal data from treatment and control groups

DID requires data from pre-/post-intervention

Used when randomization on the individual level is not possible

All of these

For the remaining questions (14 and 15) use the star dataset in the ECDAT package. Create a new dataset *mydata* that only has records for the small and regular.with.aide classes. We are not interested in regular sized classes. Create a dummy variable called *small* which is 1 if a record in*mydata* is of a student in a small size class, and is 0 otherwise. Create totalscore which is the sum of the math and the reading scores for each record.

Question 14

Please run a linear regression using all the data in *mydata* using *totalscore* as the response variable and small as the predictor. What is the estimated coefficient of small? What is its t-value? Is small statistically significant?

13.554, 5.601, Yes

13.554, 5.601, No

13.554, 5.601, Yes at 5%, No at 1%

13.554, 2.420. Yes at 5%, No at 1%

13.554, 558.211, Yes

Question 15

Please run a linear regression using all the data in *mydata* using *totalscore* as the response variable and small and teacher experience as the predictors. What are the estimated coefficient of small and teacher experience? What are their t-value? Are they statistically significant?

**Small**: coef=14.6502,t-value=6.072, Significant; **totexp**: coef=1.3429, t-value=6.460,Not Significant

**Small**: coef=14.6502,t-value=6.072, Significant; **totexp**: coef=1.3429, t-value=6.460, Significant

**Small**: coef=14.6502,t-value=6.072, Not Significant; **totexp**: coef=1.3429, t-value=6.460, Significant

**Small**: coef=14.6502,t-value=6.460, Significant; **totexp**: coef=1.3429, t-value=6.072, Significant

**Small**: coef=14.6502,t-value=6.072, Not Significant; **totexp**: coef=1.3429, t-value=6.460, Not Significant