Homework 7

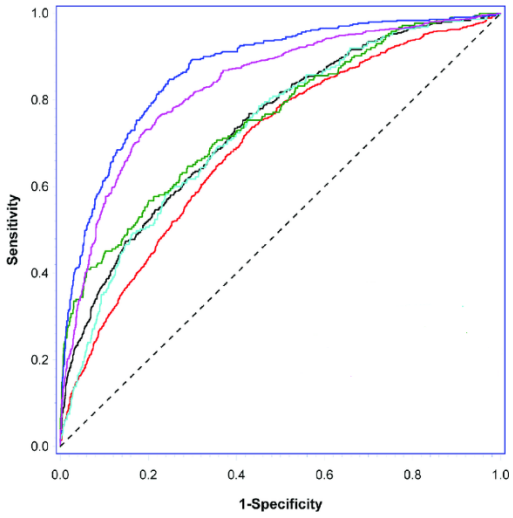
Statistical Inference, Spring 1400

1- Short answers and True/False Problems (explain/correct if False):

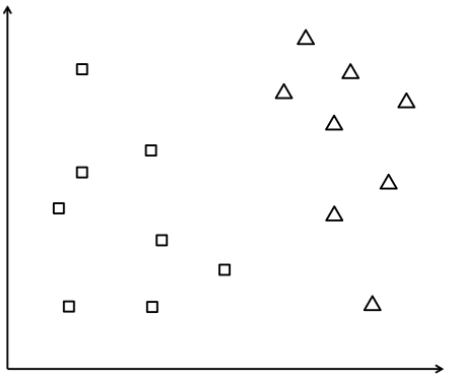
a. Can we apply logistic regression on a 3-class classification problem? (Explain your answer in less than two lines.)

b. Suppose you've been given a fair coin, write down the odds of getting heads, and describe your answer.

c. The below figure shows us AUC-ROC curves for different models. Which one of these models will give us the best result? Describe your answer in less than two lines.



2- Given the following samples (each shape belongs to different classes), could logistic regression produce zero training error? Explain your answer.



Homework 7

Statistical Inference, Spring 1400

3- We have measured the cracking power of two different types of bricks (Tolid-e-melli vs Chinese version). Test the claim that the probability distributions associated with the Bricks are equivalent (Use 5% significant level).

Tolid-e-melli	315	317	316	316	295	318	317	316	269	314
Chinese	321	319	267	242	324	323	284	258	257	322

4- ADHD is a mental health disorder that can cause above-normal levels of hyperactive and impulsive behaviors. A new method to detect ADHD is being developed here at UT. To assess this method, 1000 patients known to be ADHD positive were selected and 800 patients know to be ADHD negative were selected. Each of the 1800 patients was tested with the new method. As a result, 867 out of 1000 truly positive patients tested positive; and 85 out of 800 truly negative patients tested positive.

a. What is the sensitivity of the new method?

b. What is the specificity of this new method?

c. Using the fact that the prevalence of ADHD is 0.02, calculate the positive predictive value of the test?

5- The below information is based on records of car crashes in 1399 compiled by Iranian Traffic Police in Shiraz.

Safety Equipment in Shiraz	**Injury**	
	Fatal	**Nonfatal**
None	1601	162,527
Seat Belt	510	412,368

a. Identify the response variable, and find and interpret the difference of proportions, relative risk, and odds ratio.

b. Why is the relative risk and odds ratio almost equal?

6- The median age of developing BC (breast cancer) is thought to be 45 years. The random sample of 30 people with BC are:

35.5, 44.5, 39.8, 33.3, 51.4, 51.3, 30.5, 48.9, 42.1, 40.3,
46.8, 38.0, 40.1, 36.8, 39.3, 65.4, 42.6, 42.8, 59.8, 52.4,
26.2, 60.9, 45.6, 27.1, 47.3, 36.6, 55.6, 45.1, 52.2, 43.5

Does the median age in BC differ significantly from 45 years?

Homework 7

Statistical Inference, Spring 1400

7- (R) Simpson's paradox is "*the reversal of the direction of a comparison or an association when data from several groups are combined to form a single group*". The data concerned two hospitals, A and B, and whether or not patients undergoing surgery died or survived. Here are the data for all patients:

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800

And here are the more detailed data where the patients are categorized as being in good condition or poor condition:

Good condition			Poor condition		
	Hospital A	Hospital B		Hospital A	Hospital B
Died	6	8	Died	57	8
Survived	594	592	Survived	1443	192
Total	600	600	Total	1500	200

a. Use logistic regression to model the odds of death within hospital as the explanatory variable. Summarize the results of your analysis and give a 95% confidence interval for the odds ratio of hospital A relative to hospital B.

b. Rerun your analysis in (a) using the hospital and patient's condition as explanatory variables. Summarize the results of your analysis and give a 95% confidence interval for the odds ratio of hospital A relative to hospital B.

c. Explain Simpson's paradox in terms of your results in part (a) and part (b).

8- (R) Suppose we want to validate a new method for curing the illness. Answer the questions below using the attached dataset (Data.csv).

a. Divide the data into the test and the train parts. Use the "Response" column as the response variable and others as predictors. Use 1/3 of data as the test data. Fit logistic regression and write out the formula

b. Form a hypothesis to validate if predictors are significant. Specify the null and the alternative hypotheses.

c. Use stepwise variable selection to get the best model from the dataset using adjusted R^2 (forward method).

d. Plot the ROC and get the AUC threshold.

e. Plot the logistic regression assumptions using ggplot2 and specify the outliers.