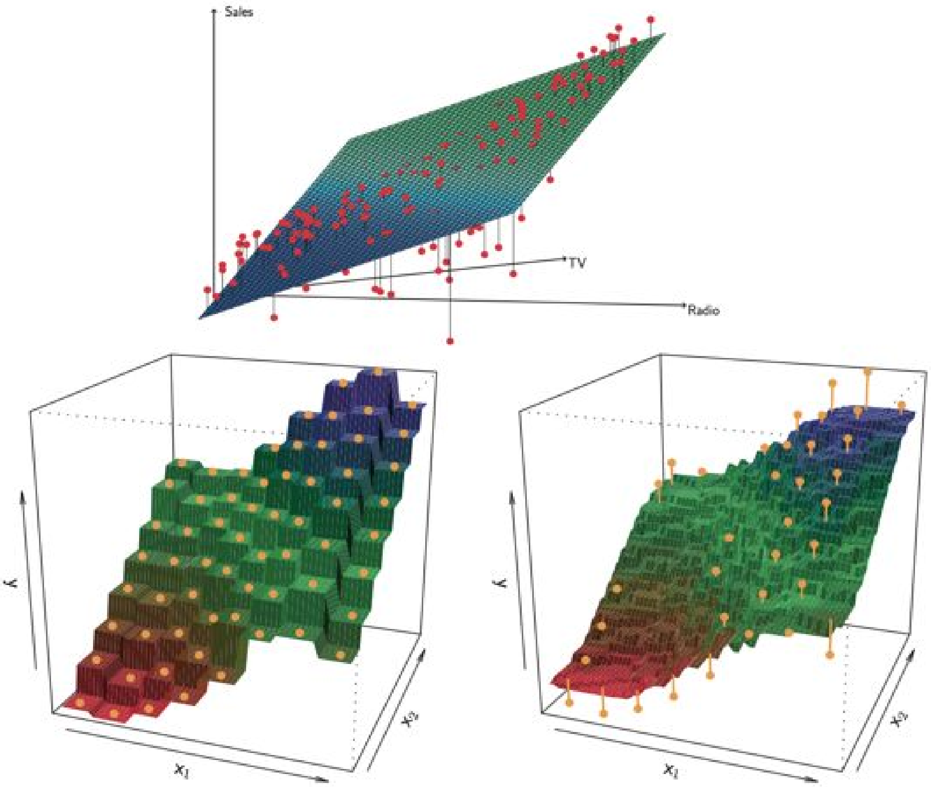
Foundations of Statistics, kNN, linear regression

1. Give an example when you would use the mode in practice.

2. You are tasked with creating an experiment to judge whether a new user interface causes users to purchase more widgets. From your single population of users, you randomly split the users into a control group, that uses the former layout, and a treatment group, that uses the new layout. What statistic would you use to determine if the new layout causes an increase in the number of widgets sold?

3. For what type of variable should you NOT use Pearson linear correlation?

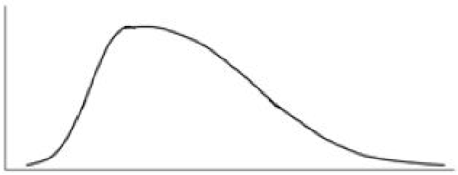
4. Y is our continuous response variable and x1 and x2 are our explanatory features. The 64 dots represent observations while the surface represents the predictions of 3 different models. Of the machine learning algorithms we learned, which is used in each graph? If you think it's knn, try to guess the value of k.

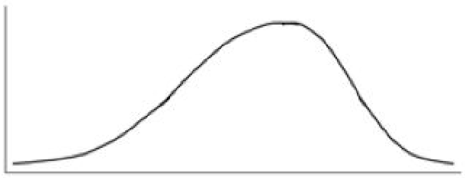


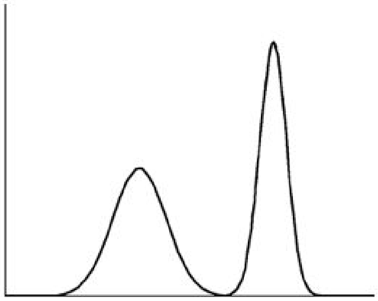
5. For knn, what is a heuristic for choosing a value of k?

6. What are the assumptions of simple linear regression?

7. What assumption of linear regression is heteroskedasticity a violation of?
8. How do you interpret a coefficient of determination (R^2) of .8?
9. For a linear regression with one explanatory variable, what do you do if the p-value for the intercept is insignificant?
10. What are the null and alternative hypotheses of a chi-square test of independence?
11. What is standard deviation and what is its equation for a sample of n observations?
12. What is the difference between a one-tail t-test and a two-tail t-test?
13. What are missing at random, missing completely at random, and missing not at random?
14. What are some pros of knn?
15. What are some cons of knn?
16. Interpret beta1 in the following regression model equation: Income = 20,000 + 3,000*(years schooling) + 30,000*(ability to use a search engine)
17. How can you mitigate a violation of constant variance?
18. How do you calculate the coefficient of determination (R^2)?
19. Describe the following variables distributions from their histograms:







20. Is the point (xbar, ybar) always on the linear regression line y = beta0 + beta1 * x?

21. What does an overall F-value tell you for a linear regression? What null hypothesis does it test?

22. What would you use a partial F-test for? What null hypothesis does it test?

23. How can you test for multicollinearity between explanatory variables? How can you eliminate or reduce multicollinearity?