Bernhard Nordemann
CS 4342

1. The null hypothesis for the below table is that TV, radio and newspaper do not affect sales. Since the p value of TV and radio are very low, and the t-statistic is relatively high, then we can reject these two null hypotheses and conclude that TV and radio have an effect on sales. Since the p value of newspaper is very high, and the t statistic is relatively close to zero, we do not reject the null hypothesis for newspaper, and conclude that newspaper has no significant effect on sales.
2. The KNN classifier takes the k neighboring values of a data point, and classifies the data point based on the highest occurring type of the surrounding k neighboring data points. KNN classifiers are used to classify qualitative data. The KNN regression method is used to solve regression problems with a quantitative response, by identifying the k nearest neighbors of a data point, x, and then estimating f(x) as the average of this neighborhood.
3. .
    a. For a fixed value of IQ and GPA, males earn more on average than females, provided that the GPA is high enough
        i. If we plug the parameters into a model, we get :
           male = 50 + 20GPA + .07IQ + 0.01GPAxIQ
           female = 50 + 20GPA + 0.07IQ + 35(female) + 0.01GPAxIQ -10GPA
           Or, female = 85 + 10GPA + 0.07IQ + 0.01GPAxIQ
           So, males make more, given that their GPA is high enough (given GPA > 3.5) and IQ is the same.
    b. female = 85 + 10GPA + 0.07IQ + 0.01GPAxIQ
       female = 85 + 10(4) + 0.07(110) + 0.01(110*4) = 137.1
       So, we predict a starting salary of $137,100
    c. False. To verify the interaction between GPA/ IQ and its relevance, we need to view the p value, or the t statistic to judge the impact this parameter has on the model.
4. .
    a. We would expect that the training RSS for the linear regression would be lower than the training RSS for cubic regression. Because the true relationship between X and Y is linear, it is more easily modeled by a linear regression, and is less prone to having errors.
    b. We would also expect that the test RSS for linear regression would be lower than the test RSS for cubic regression. We would expect the cubic regression to overfit more than the linear regression which would give us a general fit, closer to the true relationship.
    c. We would expect the training RSS to be lower for the cubic regression. The reason for this is that the cubic regression is able to fit more to the training data than the linear regression.
    d. There is not enough information on which RSS would be lower for test data. Since we do not know how far the relationship is from linear, we cannot say. If it

is more cubic than linear, then RSS for cubic could be smaller. If it is more linear than cubic, then RSS for linear could be smaller. It also depends on the variance of the training model.

5. Substituting beta into the equation, we get:

$$\hat{y}_i = \frac{x_i}{\sum_{i'=1}^{n} x_{i'}^2} \sum_{i=1}^{n} x_i y_i$$

$$\hat{y} = \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{i'=1}^{n} x_{i'}^2} \cdot x_i \right) \cdot y_i$$

$$\boxed{a_{i'} = \frac{x_i \cdot x_i}{\sum_{i'=1}^{n} x_{i'}^2}}$$

6. The least squares line equation always passes through the point ($\bar{x}$, $\bar{y}$), because the beta value chosen to minimize the RSS is based on the average values of X and Y. the full equation would be f($\bar{x}$) = $\bar{y} - \hat{\beta}_1\bar{x}$ + $\hat{\beta}_1\bar{x}$ when x = $\bar{x}$. We can cancel out the 2nd and 3rd term, which leaves us with f($\bar{x}$) = $\bar{y}$.

APPLIED QUESTIONS

- See applied questions code, answers, and plots in separate document: "Assignment2 applied pdf"