**CS 5565, HW2 FS19 (Linear Regression) 65 pts.**

Name _____

1. (10 points) Describe the null hypotheses to which the $p$-values given in the following table correspond. The table depicts sales for widgets based on the amount of advertising spent on ads sold on TV, radio, newspaper, and the Internet.

   Explain what conclusions you can draw based on these $p$-values. Your explanation should be phrased in terms of TV, radio, newspaper, and the Internet, rather than in terms of the coefficients of the linear model.

| | Coefficient | Std. Error | $t$-statistic | $p$-value |
|---|---|---|---|---|
| Intercept | 2.225 | 0.3126 | 9.21 | $< 0.0001$ |
| TV | 3.525 | 0.4126 | 39.01 | $< 0.0001$ |
| Radio | 0.025 | 0.0032 | 1.6 | 0.089 |
| Newspaper | -0.033 | 0.0052 | -0.62 | 0.232 |
| Internet | 0.325 | 0.0026 | 32.53 | $< 0.0001$ |

2. (10 points) Carefully explain the differences between the KNN classifier and KNN regression methods.

3. (20 points total) Suppose we have a data set with five predictors, $X_A$ = Age, $X_W$ = Weight, $X_G$ = Gender (1 for Female and 0 for Male), $X_{AW}$ = Interaction between Age and Weight, and $X_{AG}$ = Interaction between Age and Gender. The response is the level of an agent in the blood which may indicate a higher risk of heart attack. Suppose we use least squares to fit the model, and get $\widehat{\beta}_0 = 50$, $\widehat{\beta}_A = 0.5$, $\widehat{\beta}_W = 1.0$, $\widehat{\beta}_G = -40$, $\widehat{\beta}_{AW} = 0.01$, $\widehat{\beta}_{AG} = 1.0$.

   (a) (4 points) What is the blood level given you have a male patient who is 30 years old and 150 pounds?

   (b) (4 points) What is the blood level given you have a female patient who is 30 years old and 150 pounds?

   (c) (4 points) What is the blood level given you have a male patient who is 60 years old and 150 pounds?

   (d) (4 points) What is the blood level given you have a female patient who is 60 years old and 150 pounds?

   (e) (4 points) At what age are the blood levels the same given both patients are 150 pounds?

4. (10 points) I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \widehat{\beta}_0 + \widehat{\beta}_1 X + \widehat{\beta}_2 X^2 + \widehat{\beta}_3 X^3 + \epsilon$

   (a) (3 points) Suppose that the true relationship between $X$ and $Y$ is linear, i.e. $Y = \widehat{\beta}_0 + \widehat{\beta}_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) (3 points) Answer the previous question using the test rather than the training RSS.

(c) (2 points) Suppose that the true relationship between $X$ and $Y$ is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) (2 points) Answer the previous question using test rather than training RSS.

5. (15 points) Fill in an ANOVA table and compute the F-statistic and $p$-value for the following data.

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|
| 7 | 4 | 11 | 5 | 6 |
| 6 | 3 | 10 | 4 | 5 |
| 8 | 2 | 10 | 5 | 7 |
| 9 | 4 | 10 | 4 | 6 |