CS 5565, HW2 FS19 (Linear Regression) 150 pts.

Name _____

1. (15 points) Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

   (a) (5 points) We collect a set of data on 10,000 home sales in a particular city. For each house sold, we record the price, number of floors (ranch, 1 1/2, reverse 1 1/2, 2 story), age, total number of rooms, number of bedrooms, number of bathrooms, basement or no basement, lot size, and school district. We are interested in understanding which factors affect the house sales price.

   (b) (5 points) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2016. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

   (c) (5 points) We are considering an analysis on group 100 employees in a petrochemical plant to determine whether they have either a high or low risk of being diagnosed with a certain disease. We collect data on 30 different chemical compounds that they may be exposed to and record the amount of time they are exposed to each chemical each day. For each employee, we want to determine if they have high risk or a low risk of getting the disease.

2. (10 points) Consider the bias-variance decomposition.

   (a) (5 points) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

   (b) (5 points) explain why each of the five curves has the shape displayed in part (a).

3. (20 points) For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

   (a) (5 points)
   The sample size n is extremely large, and the number of predictors p is small.

   (b) (5 points)
   The number of predictors p is extremely large, and the number of observations n is small.

   (c) (5 points)
   The sample size n is extremely large, and the number of predictors p is small.

   (d) (5 points)
   The relationship between the predictors and response is highly non-linear.

   (e) (5 points)
   The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

4. (20 points) Suppose you are given the following data.

X	Y
1.00	0.8
2.00	2.1
3.00	3.8
4.00	3.7
5.00	6.1
6.00	7.0
7.00	9.2
8.00	9.3
9.00	10.1
10.00	12.2

(a) (5 points) Use linear regression to find the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$

(b) (5 points) What is R^2

(c) (5 points) What is the predicted value of Y_{15}?

(d) (5 points) What is the predicted value of Y_{18}?

You may use the expression in the book or the following:

$$\begin{aligned} \beta_1 &= \frac{n\sum_{t=1}^{n} X_t Y_t - \left(\sum_{t=1}^{n} X_t\right)\left(\sum_{t=1}^{n} Y_t\right)}{n\sum_{t=1}^{n} X_t^2 - \left(\sum_{t=1}^{n} X_t\right)^2} \\ \beta_0 &= \bar{Y} - \beta_1 \bar{X}, \end{aligned}$$

where $\bar{Y} = \left(\sum_{t=1}^{n} Y_t\right)/n$ and $\bar{X} = \left(\sum_{t=1}^{n} X_t\right)/n$

5. (10 points) In your own words, explain the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification. Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

6. (10 points) Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

7. (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

8. (10 points) Carefully explain the differences between the KNN classifier and KNN regression methods.

9. (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?

10. (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.

11. (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	6
9	4	10	6	7