**Q1.** Explain whether each scenario is a classification or regression problem and indicate whether we are most interested in inference or prediction. Finally, provide $n$n and $p$p.

a. 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.

*Regression and inference with* $n=10000$n=10000 *and* $p=9$p=9

*b. 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.*

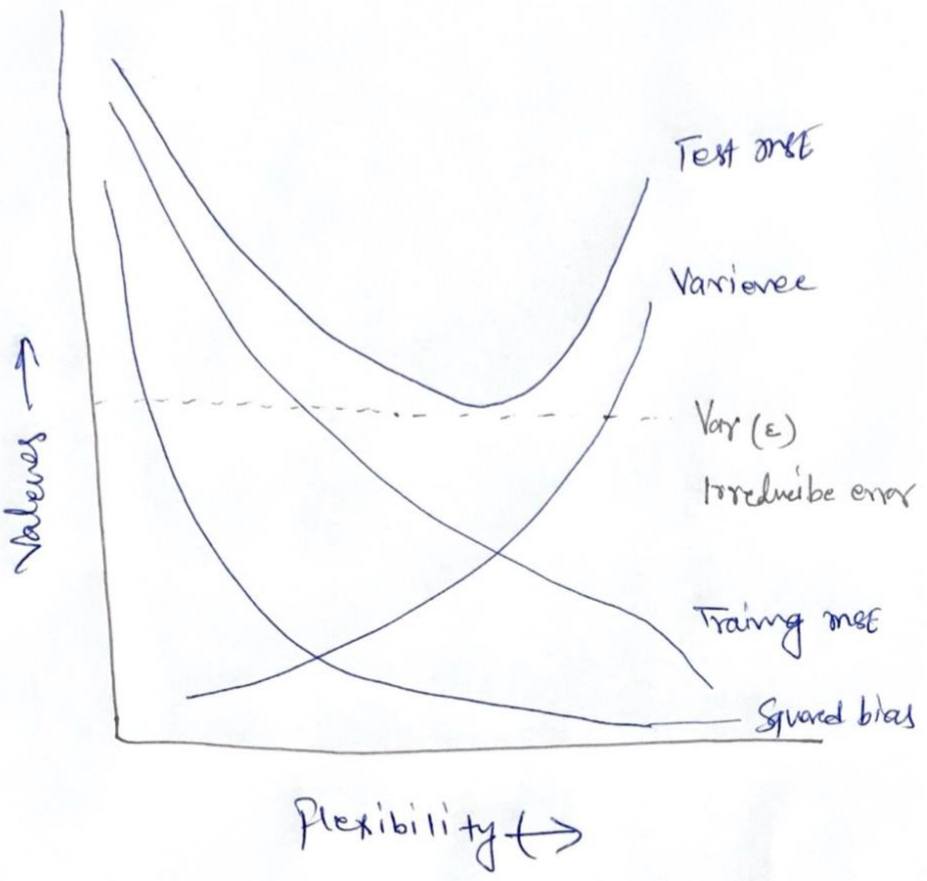
*Classification and prediction with* $n=52$n=52 *and* $p=4$p=4

c. 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.

*Regression and prediction with* $n=100$n=100 *and* $p=30$

**Q2.** We now revisit the bias-variance decomposition.

a. 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. Explain why each of the five curves has the shape displayed in part (a).

*The training MSE declines monotonically as flexibility increases, this is because as flexibility increases the* $f$ *curve fits the observed data more closely. The test MSE initially declines as flexibility increases but at some point it levels off and then starts to increase again (U-shape), this is because when a* $f$ *curve yields a small training MSE but a large test MSE we are actually overfitting the data The squared bias decreases monotonically and the variance increases monotonically; as a general rule, as we use more flexible methods, the variance will increase and the bias will decrease. Variance refers to the amount by wich* $\hat{f}$ *would change if we estimated it using a different training data set, so if the curve fits the observations very closely, changing any point may cause* $\hat{f}$ *to change considerably, and so will result in some variance. Bias refers to the error that is introduced by approximating a real-life problem by a much simpler model, so if we use a very simple model (linear regression) it is unlikely that any real-life problem has such a simple linear relationship, and so performing linear regression will result in some bias in the estimate of* $f$. *The irreducible error is a constant so it is a parallel line, this curve lies below the test MSE curve because the expected test MSE will always be greater the* $\text{Var}(\varepsilon)$.

**Q3.** For each of parts (a) through (d), indicate whether i. or ii. is correct and explain your answer. In general, do we expect the performance of a flexible statistical learning method to perform better or worse than an inflexible method when:

a. The sample size $n$ is extremely large, and the number of predictors $p$ is small?

*Better. A flexible method will fit the data closer and with the large sample size, would perform better than an inflexible approach.*

b. The number of predictors $p$ is extremely large, and the number of observations $n$ is small?

*Worse. A flexible method would overfit the small number of observations.*

c. The relationship between the predictors and response is highly non-linear?

*Better. With more degrees of freedom, a flexible method would fit better than an inflexible one.*

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

*Worse. A flexible method would fit to the noise in the error terms and increase variance.*

8.4

## Regression table

| N | X | Y | $\hat{y}$ | $y-\hat{y}$ | $(y-\hat{y})^2$ | XY | $x^2$ | $Y^2$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.0 | 0.8 | 0.88 | -0.08 | .0064 | 0.8 | 1 | 0.64 |
| 2 | 2.0 | 2.1 | 2.1133 | -0.01 | .0001 | 4.2 | 4 | 4.41 |
| 3 | 3.0 | 3.8 | 3.3466 | 0.45 | 0.2 | 11.4 | 9 | 14.44 |
| 4 | 4.0 | 3.7 | 4.58 | -0.88 | 0.77 | 14.8 | 16 | 13.69 |
| 5 | 5.0 | 6.1 | 5.81 | 0.29 | 0.08 | 30.5 | 25 | 37.21 |
| 6 | 6.0 | 7.0 | 7.04 | -0.05 | .0021 | 42 | 36 | 49 |
| 7 | 7.0 | 9.20 | 8.28 | 0.92 | 0.84 | 64.4 | 49 | 84.64 |
| 8 | 8.0 | 9.30 | 9.51 | -0.21 | 0.045 | 74.4 | 64 | 86.49 |
| 9 | 9.0 | 10.10 | 10.74 | -0.65 | 0.41 | 90.9 | 81 | 102.01 |
| 10 | 10.0 | 12.20 | 11.98 | 0.22 | 0.048 | 122 | 100 | 148.84 |
| Σ | 10 | 55 | 64.30 | 64.3 | 0.00 | 242 | 4554 | 385 | 541.37 |

$\hat{x} = 5.5$ $\quad SS_{xy} = 101.75$

$\hat{y} = 6.43$ $\quad SS_{yy} = 127.921$

$\beta_1 = 1.2333$ $\quad SS_{xx} = 82.5$

$\beta_0 = -0.3533$ $\quad S_{EE} = 2.42$

$S^2 = 0.303666$

$r^2 = 0.9810091$

$r = 0.99045$ $\quad S = 0.5510$

**Q5.** What are 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?

*The advantages of a very flexible approach are that it may give a better fit for non-linear models and it decreases the bias.*

*The disadvantages of a very flexible approach are that it requires estimating a greater number of parameters, it follows the noise too closely (overfit) and it increases the variance.*

*A more flexible approach would be preferred to a less flexible approach when we are interested in prediction and not the interpretability of the results.*

*A less flexible approach would be preferred to a more flexible approach when we are interested in inference and the interpretability of the results.*

**Q6.** 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?

**Q7.** Describe the null hypotheses to which the p-values given in following table correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

*The null hypotheses associated with following are that advertising budgets of "TV", "radio" or "newspaper" do not have an effect on sales. More precisely* $H_{(1)0}$: $\beta_1=0 H0(1)$: $\beta_1=0$, $H_{(2)0}$: $\beta_2=0 H0(2)$: $\beta_2=0$ *and* $H_{(3)0}$: $\beta_3=0 H0(3)$: $\beta_3=0$. *The corresponding p-values are highly significant for "TV" and "radio" and not significant for "newspaper"; so, we reject* $H(1)0H0(1)$ *and* $H(2)0H0(2)$ *and we do not reject* $H(3)0H0(3)$. *We may conclude that newspaper advertising budget do not affect sales.*

*A parametric approach reduces the problem of estimating* ff *down to one of estimating a set of parameters because it assumes a form for* ff.
*A non-parametric approach does not assume a particular form of* ff *and so requires a very large sample to accurately estimate* ff.
*The advantages of a parametric approach to regression or classification are the simplifying of modeling* ff *to a few parameters and not as many observations are required compared to a non-parametric approach.*
*The disadvantages of a parametric approach to regression or classification are a potentially inaccurate estimate* ff *if the form of* ff *assumed is wrong or to overfit the observations if more flexible models are used.*

**Q8.** Carefully explain the differences between the KNN classifier and KNN regression methods.

*The KNN classifier is typically used to solve classification problems (those with a qualitative response) by identifying the neighborhood of* $x0x0$ *and then estimating the conditional probability* $P(Y=j|X=x0)P(Y=j|X=x0)$ *for class* jj *as the fraction of points in the neighborhood whose*

*response values equal* jj. *The KNN regression method is used to solve regression problems (those with a quantitative response) by again identifying the neighborhood of* x0x0 *and then estimating* f(x0)f(x0) *as the average of all the training responses in the neighborhood.*

**Q9.** Suppose we have a data set with five predictors, X1=GPAX1=GPA, X2=IQX2=IQ, X3=GenderX3=Gender (1 for Female and 0 for Male), X4=Interaction between GPA and IQX4=Interaction between GPA and IQ, and X5=Interaction between GPA and GenderX5=Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta_0}=50$, $\hat{\beta_1}=20$, $\hat{\beta_2}=0.07$, $\hat{\beta_3}=35$, $\hat{\beta_4}=0.01$, $\hat{\beta_5}=-10$.

a. Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.
ii. For a fixed value of IQ and GPA, females earn more on average than males.
iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

*The least square line is given by*

$$\hat{y}=50+20GPA+0.07IQ+35Gender+0.01GPA\times IQ-10GPA\times Gender$$

*which becomes for the males*

$$\hat{y}=50+20GPA+0.07IQ+0.01GPA\times IQ,$$

*and for the females*

$$\hat{y}=85+10GPA+0.07IQ+0.01GPA\times IQ.$$

*So, the starting salary for males is higher than for females on average iff* $50+20GPA\geq 85+10GPA$ *which is equivalent to* $GPA\geq 3.5$. *Therefore iii. is the right answer.*

b. Predict the salary of a female with IQ of 110 and a GPA of 4.0.

*It suffices to plug in the given values in the least square line for females given above and we obtain*

$$\hat{y}=85+40+7.7+4.4=137.1,$$

*which gives us a starting salary of* 137100$137100$.

c. True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

*False. To verify if the GPA/IQ has an impact on the quality of the model we need to test the hypothesis* $H_0: \hat{\beta_4}=0$ *and look at the p-value associated with the* tt *or the* FF *statistic to draw a conclusion.*

**Q10.** 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. $Y=\beta_0+\beta_1X+\beta_2X_2+\beta_3X_3+\varepsilon$ .

a. Suppose that the true relationship between $X$ and $Y$ is linear, i.e. $Y=\beta_0+\beta_1X+\varepsilon$. 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.

*Without knowing more details about the training data, it is difficult to know which training RSS is lower between linear or cubic. However, as the true relationship between $X$ and $Y$ is linear, we may expect the least squares line to be close to the true regression line, and consequently the RSS for the linear regression may be lower than for the cubic regression.*

b. Answer (a) using test rather than training RSS.

*In this case the test RSS depends upon the test data, so we have not enough information to conclude. However, we may assume that polynomial regression will have a higher test RSS as the overfit from training would have more error than the linear regression.*

c. 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.

*Polynomial regression has lower train RSS than the linear fit because of higher flexibility: no matter what the underlying true relationshop is the more flexible model will closer follow points and reduce train RSS. An example of this behavior is shown on Figure 2.9 from Chapter 2.*

d. Answer (c) using test rather than training RSS.

*There is not enough information to tell which test RSS would be lower for either regression given the problem statement is defined as not knowing "how far it is from linear". If it is closer to linear than cubic, the linear regression test RSS could be lower than the cubic regression test RSS. Or, if it is closer to cubic than linear, the cubic regression test RSS could be lower than the linear regression test RSS. It is dues to bias-variance tradeoff: it is not clear what level of flexibility will fit data better.*

(11) ANOVA

$a = 4$
$b = 5$

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_1^2$ | $x_2^2$ | $x_3^2$ | $x_4^2$ | $x_5^2$ |
|---|---|---|---|---|---|---|---|---|---|
| 7 | 4 | 11 | 5 | 6 | 49 | 16 | 121 | 25 | 36 |
| 6 | 3 | 10 | 4 | 5 | 36 | 9 | 100 | 16 | 25 |
| 8 | 2 | 10 | 5 | 6 | 64 | 4 | 100 | 25 | 36 |
| 9 | 4 | 10 | 6 | 7 | 81 | 16 | 100 | 36 | 49 |

$|T| = 819.2$

$|B| = 931.5$
$|W| = 944$

| Source | expression | DF | SS | MS | F | p-Va |
|---|---|---|---|---|---|---|
| $\|B\|$ | 931.5 | 4 | 112.3 | 28.07 | 33.69 | f-to |
| $\|W\|$ | 944 | 15 | 12.5 | 0.833 | | |
| $\|T\|$ | 819.2 | 19 | 124.8 | | | |