Problem Bank: Statistical Learning

September 10, 2019

1. Consider the general statistical model:

$$Y = f(X) + \epsilon, \tag{1}$$

where Y is the response variable, and $X = (X_1, \cdots, X_p)$ are the input variables.

Mark each statement below as TRUE or FALSE, and justify why in 1-2 sentences. If FALSE, reword the incorrect statement into a true one.

(a) The function f represents the known systematic information that X provides on Y.

(b) The error term ϵ is typically assumed to be independent of X.

(c) If we are missing important predictors in X, then the error term ϵ may have non-zero mean.

(d) If we include too many inert predictors in X (i.e., predictors which are non-influential for Y), then the error term ϵ may have inflated variance.

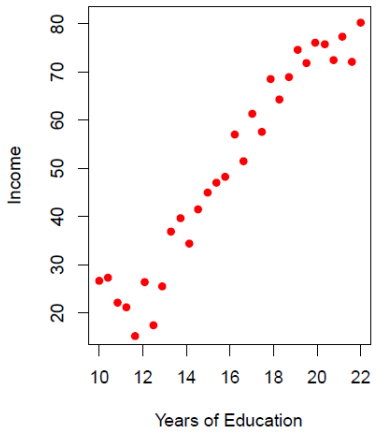
(e) Hypothetically, as we collect more and more data, the error term ϵ should become smaller and smaller.

(f) Hypothetically, as we collect more and more data, our prediction $\hat{f}$ should become closer and closer to f.

(g) Model (1) can be equivalently expressed as:

$$Y = f_1(X_1) + \cdots + f_p(X_p) + \epsilon.$$

2. Consider the following dataset on annual income (in \$10,000) vs. years of education for $n = 30$ individuals:



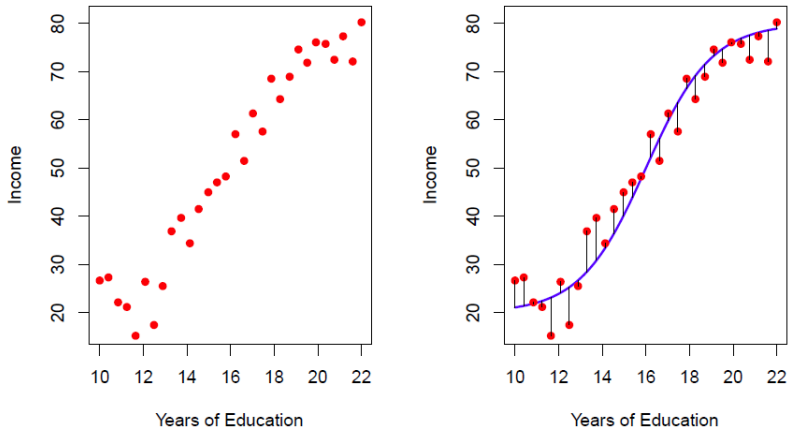
(a) Draw a linear regression fit on the above figure. Label the predictor $\hat{f}(\cdot)$ and the fitted error terms e_i for each data point $i = 1, \cdots, n$.

(b) Based on your plot, does a linear model fit the data well? Explain why or why not.

(c) Given the *context* of the problem (i.e., education vs. income), would you expect a linear model to fit well? Explain why or why not.

(d) Suppose we decide on using linear regression, and refit the model on $n = 1,000,000$ individuals using census data. Would you expect the resulting regression predictor $\hat{f}(\cdot)$ converge to the true regression function $f(\cdot)$?

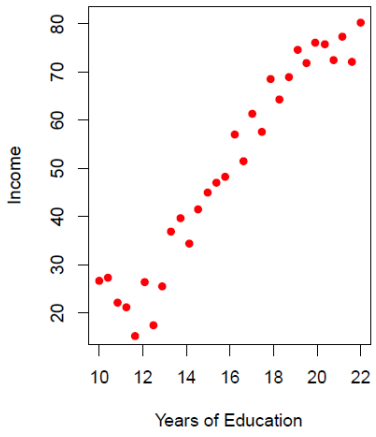
3. Consider the following non-linear regression fit on annual income (in \$10,000) vs. years of education for $n = 30$ individuals:



(a) Does the model fit the data well? Justify why / why not based on the fitted errors.

(b) Plot out what the bias, variance, test MSE and training MSE curves may look like as a function of model flexibility. Justify important features in these curves.

(c) Draw out below what the fitted model $\hat{f}(\cdot)$ may look like if we assumed high model flexibility. Use this to justify the test and training MSEs in part (b).



(d) In the first plot, the fitted model $\hat{f}(\cdot)$ suggests significant slope changes at $x = 12$ and $x = 18$. Interpret what this means in terms of the problem. Based on purely income considerations, what advice would you give a graduating high-school student?

4. Assume the general learning model:

$$Y(x) = f(x) + \epsilon,$$

where $Y(x)$ is the response at input predictors x, and ϵ is a zero-mean random error term. We discussed in-class the reducible-irreducible error decomposition of the MSPE:

$$\text{MSPE}(x) := \mathbb{E}[(Y(x) - \hat{f}(x))^2] = [f(x) - \hat{f}(x)]^2 + \text{Var}(\epsilon),$$

where $\hat{f}(\cdot)$ is a chosen predictor function.

(a) Of the variables $Y(x)$, x, ϵ and $\hat{f}(x)$, which are random? Which are not?

(b) Prove this decomposition, justifying each step.

(c) Use this decomposition to show that, if $Y(x)$ is known for each x, the optimal predictor minimizing $\text{MSPE}(x)$ is $\hat{f}(x) = \mathbb{E}[Y(x)]$.

(d) Explain the intuition behind this predictor in layman's terms (i.e., to someone who is not well-versed in statistics).

(e) Why can't one apply this predictor in practice?

5. Assume the general learning model:

$$Y(x) = f(x) + \epsilon,$$

where $Y(x)$ is the response at input predictors x, and ϵ is a random error term. Instead of the MSPE discussed in class, suppose we use a different error measure – the mean absolute predictive error (MAPE):

$$\text{MAPE}(x) := \mathbb{E}[|Y(x) - \hat{f}(x)|].$$

We wish to find the optimal predictor under this new MAPE error measure.

(a) Of the variables $Y(x)$, x, ϵ and $\hat{f}(x)$, which are random? Which are not?

(b) Let Z be a continuous random variable with distribution function $F(\cdot)$. For any number m, show that:

$$\mathbb{E}[|Z - m|] = \int_{-\infty}^{m} (m - z)\, dF(z) + \int_{m}^{\infty} (z - m)\, dF(z).$$

(c) Define $m^* = \text{med}(Z)$ as the *median* of Z, satisfying $F(m^*) = 0.5$. Using (b), show that for any number m greater than m^*, we have:

$$\mathbb{E}[|Z-m|]-\mathbb{E}[|Z-m^*|] = (m-m^*)\,[P(Z \le m^*) - P(Z > m)]+2\int_{m^*}^{m}(m-z)\, dF(z).$$

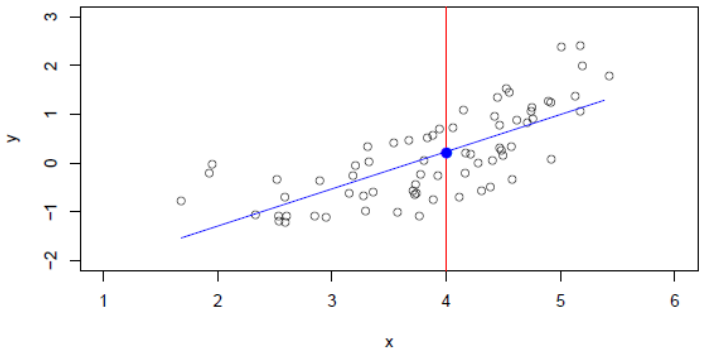
(d) Using (c), argue that $\mathbb{E}[|Z - m|] - \mathbb{E}[|Z - m^*|] \geq 0$ for any $m > m^*$.

(e) Using (d), show that if $Y(x)$ is known for each x, the optimal predictor minimizing $\text{MAPE}(x)$ is $\hat{f}(x) = \text{med}[Y(x)]$.

(f) Explain the intuition behind this predictor in layman's terms (i.e., to someone who is not well-versed in statistics).

(g) From (e), the optimal predictor minimizing MSPE (i.e., $\hat{f}(x) = \mathbb{E}[Y(x)]$) is different from the optimal predictor minimizing MAPE (i.e., $\hat{f}(x) = \text{med}[Y(x)]$). Give a real-world scenario where the latter predictor may be more preferable to the former.

6. Consider the following simple linear regression fit $\hat{f}$ (in blue), with data points in black:



Suppose a modeler is interested in predicting the response value Y at $x = 4$.

Mark each statement below as TRUE or FALSE, and justify why in 1-2 sentences. If FALSE, reword the incorrect statement into a true one.

(a) The "reducible" error $[\mathbb{E}(Y) - \hat{f}(4)]^2$ reduces to zero as we collect more data.

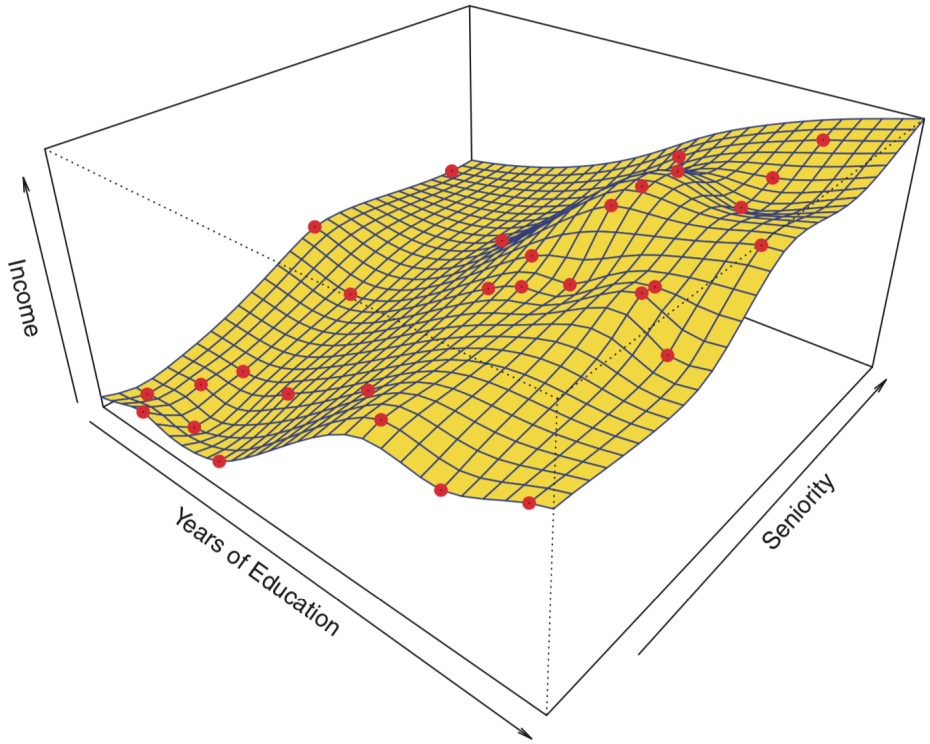
(b) The "irreducible" error $\text{Var}(Y)$ goes to zero as we collect more data.

(c) The prediction error $\mathbb{E}[(Y - \hat{f}(4))^2]$ decreases as we collect more data.

(d) There is strong evidence for a non-parametric model on f over a parametric model.

(e) The linear regression fit $\hat{f}$ is underfitting the data.

7. Consider the following fitted surface $\hat{f}$ (in yellow) on income as a function of education and seniority. Data points are marked in red.



Mark each statement below as TRUE or FALSE, and justify why in 1-2 sentences. If FALSE, reword the incorrect statement into a true one.

(a) The fitted model $\hat{f}$ is likely overfitting the data, and therefore has high bias with respect to the true function f.

(b) The fitted model $\hat{f}$ should not be used if our goal is to understand and interpret how education and seniority affects income.

(c) If we know the data was collected without noise, then $\hat{f}$ may be a good model to use for predictive purposes.

(d) The fitted model $\hat{f}$ requires a large amount of data to fit well.

8. Mark each statement below as TRUE or FALSE, and justify why in 1-2 sentences. If FALSE, reword the incorrect statement into a true one.

   (a) In problems where inference is preferred over prediction, a simpler statistical model is oftentimes preferred.

   (b) In problems where prediction is preferred over inference, a more complex statistical model always yields better performance.

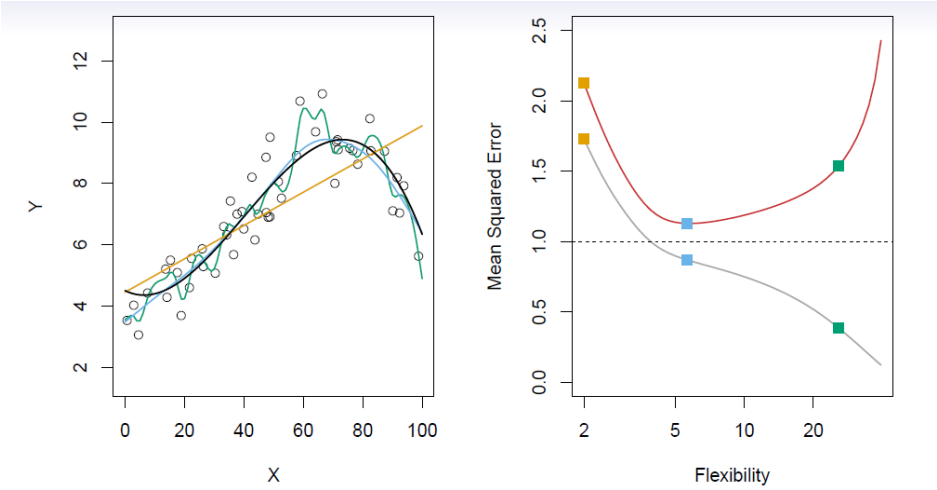
   (c) When choosing an appropriate statistical model, the optimal trade-off between prediction and inference depends solely on the data at hand.

   (d) In statistical modeling, inference involves finding which predictors are not associated with the response.

   (e) Linear regression is appealing in many applications, since the fitted models are both interpretable and flexible.

   (f) Ignoring model interpretability, a more complex model often yields better predictive power when there is little noise.

9. Plotted below are the fitted predictors for three statistical models. The left plot shows the training data (black dots), the true regression function (black curve), and the predictors from the three models (orange, green and blue curves). The right plot shows the test error (in red) and training error (in grey) for these models.



Mark each statement below as TRUE or FALSE, and justify why in 1-2 sentences. If FALSE, reword the incorrect statement into a true one.

(a) The fitted model in green is picking up erroneous signals from noise.

(b) The test errors are larger than the training errors because the true regression function is nonlinear.

(c) The fitted model in green is optimal, since it minimizes training error.

(d) The test error forms a "U-shaped" curve, since greater model flexibility often results in overfitting the data.

(e) The training error curve may form a "U-shaped" curve when the true model is highly nonlinear.

(f) In practice, one often resorts to minimizing training error, since test errors are difficult to estimate from observed data.

10. Assume the general learning model:

$$Y(x) = f(x) + \epsilon,$$

where $Y(x)$ is the response at input predictors x, and ϵ is a random error term. We discussed in-class the following bias-variance decomposition:

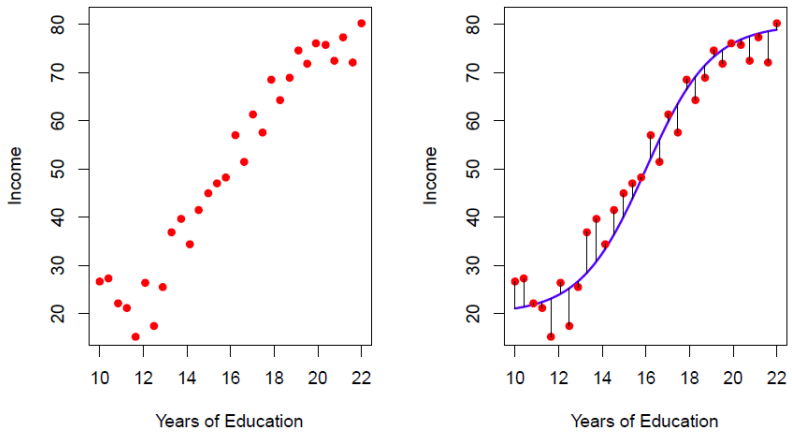
$$\mathbb{E}[(Y(x) - \hat{f}(x))^2] = \text{Var}[\hat{f}(x)] + \text{Bias}^2[\hat{f}(x)] + \text{Var}(\epsilon),$$

where $\text{Bias}[\hat{f}(x)] := \mathbb{E}[\hat{f}(x)] - f(x)$ is the *bias* of $\hat{f}(x)$ in estimating the true regression function $f(x)$. Let's investigate this further.

(a) Using the reducible-irreducible error decomposition earlier, prove the above bias-variance decomposition.
Hint: Show, for a random variable Z with mean μ and finite variance, that:

$$\mathbb{E}[(Z - a)^2] = \text{Var}(Z) + (\mu - a)^2.$$

(b) Draw out the predictor $\hat{f}(x)$ from a linear regression fit of the following data (left plot).



Which of the three terms in the bias-variance trade-off will decrease to zero as we collect more data?

(c) For the above data, let us fit the following nonlinear model (right plot). Suppose the variance term $\text{Var}[\hat{f}(x)] = 0.5$ and the bias term $\text{Bias}[\hat{f}(x)] = 0.5$. Hypothetically, if the same data were observed with less noise, what would happen to both the variance and bias terms? Justify your answer.

(d) Using part (c), make a general statement on how the bias-variance trade-off changes as observation noise changes.

11. Consider the four possible learning scenarios, obtained by varying the regression function smoothness (smooth / wiggly) and the noise variance (low / high).

    (a) For *smooth* regression functions with *low* noise, what is likely the optimal model flexibility in terms of the bias-variance trade-off? Draw two plausible curves for the test and training error.

    (b) For *smooth* regression functions with *high* noise, what is likely the optimal model flexibility in terms of the bias-variance trade-off? Draw two plausible curves for the test and training error.

    (c) For *wiggly* regression functions with *low* noise, what is likely the optimal model flexibility in terms of the bias-variance trade-off? Draw two plausible curves for the test and training error.

    (d) For *wiggly* regression functions with *high* noise, what is likely the optimal model flexibility in terms of the bias-variance trade-off? Draw two plausible curves for the test and training error.

12. For the classification problem (with K classes), we typically assume the general model:

$$p_k(x) = P[Y(x) = k], \quad k = 1, \cdots, K,$$

where $Y(x)$ is the discrete response at input predictors x. We discussed in-class the misclassification error measure:

$$\text{MCE}(x) := P[Y(x) \neq \hat{C}(x)]$$

where $\hat{C}(\cdot)$ is a chosen classifier function.

(a) Of the variables $Y(x)$, x and $\hat{C}(x)$, which are random? Which are not?

(b) In class, it was claimed that if $p_k(x)$ is known for each k and x, then the optimal (or "Bayes-optimal") classifier which minimizes $\text{MCE}(x)$ is:

$$\hat{C}(x) = k^*, \quad k^* := \underset{k=1,\cdots,K}{\text{argmax}} \; p_k(x).$$

Argue why this is true mathematically, justifying each step.

(c) Explain the intuition behind this classifier in layman's terms (i.e., to someone who is not well-versed in statistics).

(d) Why is this predictor not that useful in practice?

Ch.2 Statistical Learning - Problem Bank Questions

August 28, 2019

1. Suppose you are interested in predicting how many points the Duke basketball team will score in each game this season. Your data set includes the following variables about last year's season:

1. Date of Game
2. Home or Away Game?
3. Opponent
4. Attendance at the game
5. Points scored
6. Starting Duke players

(a) What is the response variable here? What are the input variables?

(b) What are the data types of each variable? Which variables are continuous and which are categorical?

(c) Do you think these variables are sufficient to have a good predictive model? What additional input variables might be helpful for predicting points scored?

(d) Which of the following scenarios related to this data would be prediction problems? Which scenarios would be inference problems?

i. Coach K wants to understand how the starting lineup impacts how many points are scored.

ii. Duke Athletics wants to know how the attendance at games, especially at home games, impacts how many games Duke wins over the season.

iii. Your friend only wants to attend games where Duke is likely to score more than 100 points and you want to help them determine which games to attend.

2. Suppose you are conducting a science experiment. You measure a predictor variable X and a response variable Y. Which of the following is an example of reducible error? Which of the following is an example of irreducible error?

(a) X and Y appear to have a quadratic relationship, but you fit a linear model to the data.

(b) Your measurements of Y are made with a stopwatch that is only accurate to within a tenth of a second.

3. (a) Why is it important to report error on the test data rather than the training data?

(b) What do you think are some important assumptions about the test data as related to the training data? What circumstances might make the training error a very bad estimate of the test error?

4. There were two methods presented this chapter to evaluate supervised methods; one each for regression and classification. However, these evaluation measures cannot be used for unsupervised models. How might we evaluate unsupervised methods?

5. (a) Calculate the training MSE for the two models presented below.

x_i	y_i	$\hat{f}_1(x_i)$	$\hat{f}_2(x_i)$
1	10	10.2	9.8
2	20	19.4	18.1
3	30	35.5	30.5
4	40	40.1	42.4
5	50	56.8	51.7

(b) Calculate the classification error rate for the two classifiers presented below.

x_i	y_i	$\hat{f}_1(x_i)$	$\hat{f}_2(x_i)$
1	1	0	1
2	1	1	1
3	0	1	1
4	0	0	0
5	0	0	0

6. Assume that you have a data set with 20 points in the training set and you are comparing two different models. For model 1, there are 5 predicted points that are off by a small amount from their true value, and the 15 other points are predicted with 0 error. For model 2, 1 point is predicted to be much larger than its true value, while the 19 other points are all predicted with 0 error. Which model will have the lower training MSE? What does this mean about the MSE as a measure of accuracy? Under which settings might this not be a good assumption?

7. Which of the following is a scenario where we might expect the test MSE to be much larger than the training MSE?

(a) The model has overfit to the training data.
(b) The model was trained on a different population than it is tested on.
(c) Much time has passed between when the training data and the test data was collected.
(d) All of the above.

8. Consider the U-shaped curve in Figure 2.9 (b). Which model (orange, blue or green) has the highest bias? Which model has the highest variance?

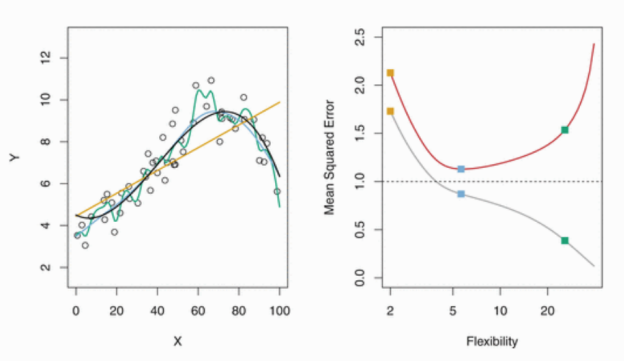


Figure 2.9

9. Which of the following is generally true?

(a) More flexible models tend to have higher bias
(b) More flexible models tend to have higher variance
(c) More flexible models tend to have higher bias and variance
(d) None of the above

10. Suppose you are interested in predicting Y from X. When you plot the data, there appears to be a highly non-linear relationship between X and Y. Which of the following models is likely to have the largest bias?

(a) Linear regression model
(b) Random Forest
(c) Predicting the mean of X for all values of X
(d) Flexible spline

11. The training MSE is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i)),$$

for training data points $x_1, \ldots, x_n$ and a given model $\hat{f}$. We want to show that the training MSE is minimized when

$$\hat{f}(x_i) = \frac{1}{n}\sum y_i,$$

that is, the training MSE is minimized by the mean.

(a) Let a be an arbitrary model. We want to find a such that the training MSE is minimized. First, expand the training MSE and distribute the sum.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - a)^2.$$

(b) Next, take the derivative of this expanded expression with respect to a.

(c) Finally, set the derivate of the MSE with respect to a equal to 0 and solve for the value of a that minimizes the training MSE. Use the second derivative test to check that this value is indeed a minimum.

12. For a K-nearest neighbor classifier, as K increases, what is likely to happen to the model fit?
(a) The bias of the classifier decreases.
(b) The variance of the classifier increases.
(c) The classifier is likely to overfit.
(d) All of the above.

13. Consider Figure 2.7 in the text, showing flexibility vs. interpretability. Suppose you have data that appears to have a linear relationship. Of the models in Figure 2.7, which model is likely to:

(a) Provide the smallest bias and variance at the same time.
(b) Overfit the data.
(c) Provide the best fit in terms of answering an inference problem.
(d) Have a high variance.

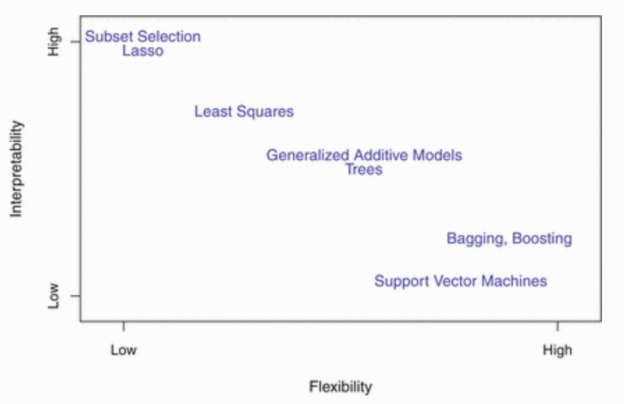


Figure 2.7

14. Consider Figure 2.9 in the text.
(a) Which model (orange, blue or green) appears to be overfitting the data?
(b) Why might this model be overfitting the data?
(c) Which model appears to be the most biased?
(d) Why might this model be biased?

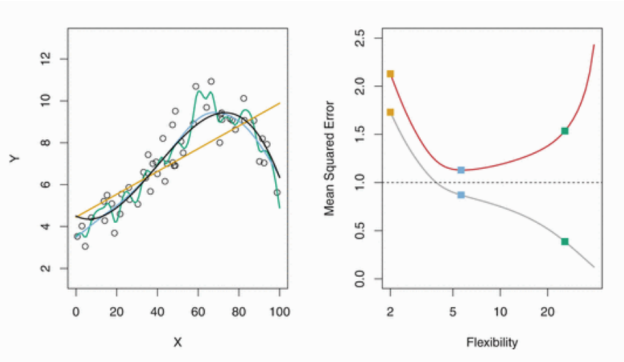


Figure 2.9

15. Suppose you are considering a linear regression problem and have found an unbiased estimator. Will this model have the best expected test MSE of any possible model? Why or why not?

Problem Bank: Statistical Learning

August 31, 2019

1. For each of the following scenarios, identify whether the model should be used for prediction or inference.

   (a) A sociology researcher is conducting a study to identify which environmental factors are commonly associated with high standardized test scores.

   (b) A sports bettor is planning to bet on every single NBA game during the upcoming season. She must win at least 52 percent of her bets in order to turn a profit.

   (c) A meteorologist wants to make an app that will predict whether it will be raining in the next 1 minute. His biggest concern is the accuracy, because he doesn't want the users to get mad at him if his predictions are wrong.

   (d) The head of Human Resources for a large company wants to figure out why some employees quit their jobs within the first year of joining.

2. Which of the following terms can be used interchangeably with the term "input variable"?

   (a) Response variable

   (b) Feature

   (c) Independent variable

   (d) Output variable

   (e) Dependent variable

   (f) Predictor

3. ISL identifies two broad reasons for using statistical models: prediction and inference. In one of these settings, the model is sometimes treated as a "black box." Which setting is this, and what does it mean to treat a model as a black box?

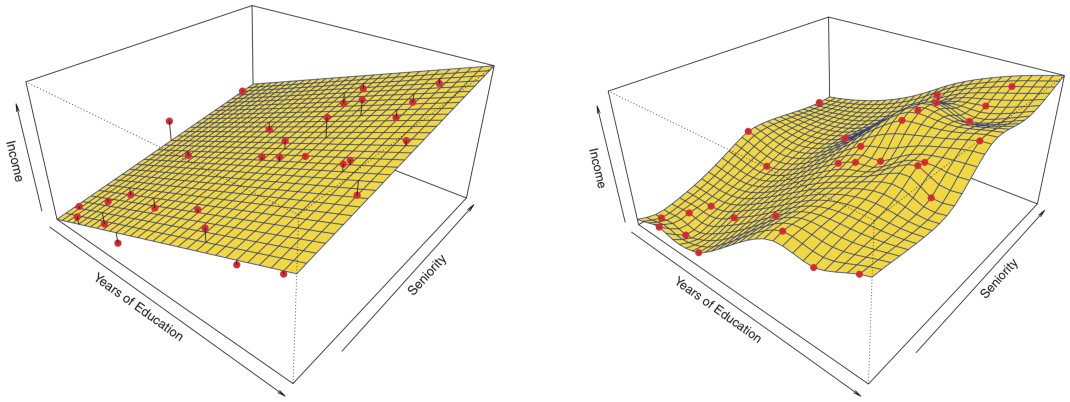
4. Say whether each of the following statements is true or false.

   (a) Parametric methods generally require more data than non-parametric methods in order to accurately estimate f.

   (b) Linear regression is an example of a non-parametric model.

   (c) Non-parametric methods are more flexible than parametric models. Thus, non-parametric models are always more accurate than parametric models.

5. In the graphs below, two models are shown. The models are fit to the *Income* dataset from ISL.



(a) *Model A* **(b)** *Model B*

Which of these models is most likely to be parametric, and which is most likely to be non-parametric? Also, which model is more likely to be "overfitting" to the data? Select one answer.

(a) Model A is parametric. Model A is more likely to be overfitting.

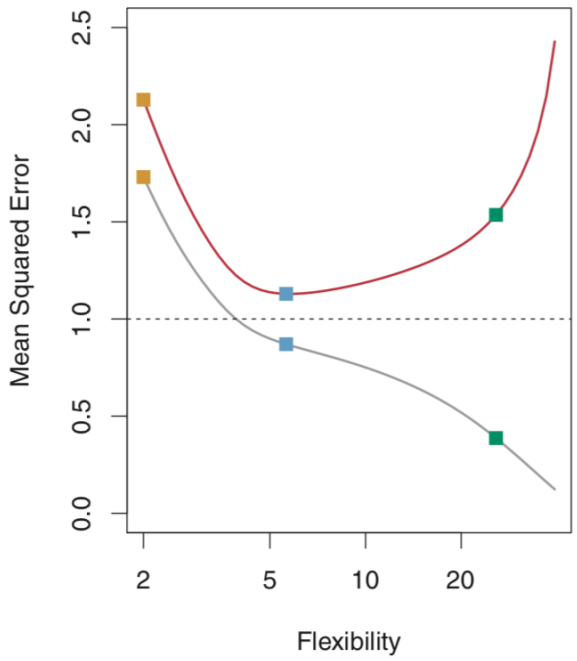
(b) Model A is parametric. Model B is more likely to be overfitting.

(c) Model B is parametric. Model A is more likely to be overfitting.

(d) Model B is parametric. Model B is more likely to be overfitting.

6. Let's imagine that we have two regression models that are fit to the same dataset. We compute the MSE of each model, using the training data. Model A has an MSE of 8, and Model B has an MSE of 12. A week later, we collect some more data. Do we know which model will perform better on this new dataset? If so, which one? Please explain the reasoning behind your answer.

7. Plotted below are the fitted predictors for three statistical models. The plot shows the testing error (in red) and training error (in grey) for these models.



The curve for the testing error is a "U-curve," which is a pattern that is commonly observed in test error plots. Explain why this U-curve is present, and why it is so common.

8. Fill in the blanks: as we increase model flexibility, the test MSE will increase or decrease depending on the relative rate of change of the _____ and the _____.

9. Which one of these methods will generally reduce the bias of a model?

   (a) Using a more flexible model.

   (b) Reducing the variance, since that will force the bias to decrease as well.

   (c) Using a less flexible model, since the lower flexibility makes it more stable across datasets.

   (d) Using a model that has fewer parameters.

10. Say whether each of the following statements is true or false.

    (a) Reducing bias is always a good thing.

    (b) The expected test MSE can never lie below $\text{Var}(\epsilon)$, which is the irreducible error.

    (c) Linear regression models tend to have low bias but high variance.

11. The Bayes classifier assigns a test observation with predictor vector x_0 to the class j for which

$$\Pr(Y = j \mid X = x_0)$$

is highest.

This classifier produces the lowest possible test error rate. However, it's essentially never used in practice. Why is that?

(a) Because calculating $\Pr(Y = j \mid X = x_0)$ is often computationally efficient.

(b) Because many people prefer frequentist methods rather than Bayesian methods.

(c) Because $\Pr(Y = j \mid X = x_0)$ is impossible to calculate in practice, unless you know the true underlying conditional distribution of Y given X.

(d) There isn't a good reason; the Bayes classifier really should be used more often!

12. In the K-Nearest Neighbors classifier, the choice of K is important. Imagine you have two models, one with $K = 3$ and one with $K = 15$. Which one of the following is true?

    (a) The model with $K = 3$ is less flexible, so it probably has lower bias and higher variance than the model with $K = 15$.

    (b) The model with $K = 3$ is less flexible, so it probably has higher bias and lower variance than the model with $K = 15$.

    (c) The model with $K = 3$ is more flexible, so it probably has lower bias and higher variance than the model with $K = 15$.

    (d) The model with $K = 3$ is more flexible, so it probably has higher bias and lower variance than the model with $K = 15$.

13. Often, when training a model, you may notice that your training error continues to drop lower and lower. You shouldn't let it go to zero, because that almost certainly means it is overfitting. So, how do you know when you should stop training?