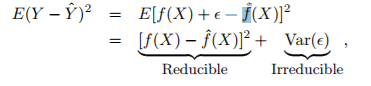
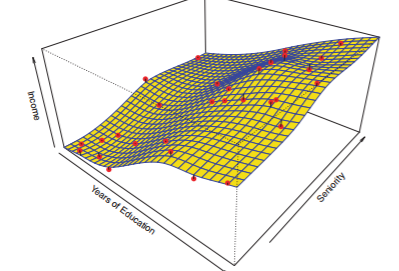
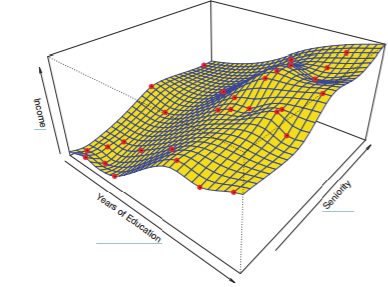
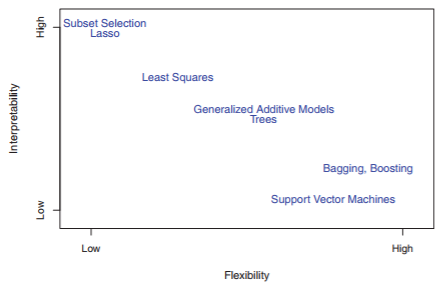
Ch 2

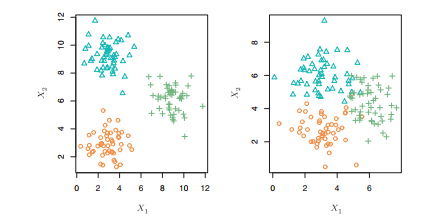
* Suppose we are statistical consultants hired by a client to provide advice on how to improve sales of a product.
* The dataset contains sales (in thousands of units) of a product over 200 different markets w/ advertising budgets for different channels/media.
* It is not possible for our client to directly increase sales of the product. On the other hand, they can control the advertising expenditure in each of the 3 channels.
* Therefore, if we determine that there is an association between advertising and sales, we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales.
* In other words, our goal is to develop an accurate model that can be used to predict sales on the basis of the 3 channels advertising budgets.
* Quantitative response variable Y = Sales, then with 3 predictors (channels)
* We assume some relationship written as **Y = f(X) + ε** where f(X) is a function of the values of the channels and **ε** = epsilon/error w/ mean = 0 and is *independent of X*
* We are estimating Y based on given X values/points.
* The errors in our predictions vs. actual values should have a mean = 0.
* **Statistical learning** refers to a set of approaches for estimating f
* 2 main reasons to estimate f:
* Prediction
* may have X values readily available, but cannot easily obtain Y values
* Since **ε** averages to 0, we can predict with w/ Ŷ = ˆf (X)
* ˆf is usually a black box, provided it yields accurate predictions for Y
* Accuracy of Ŷ in relation to Y depends on **reducible error** and **irreducible error**
* ˆf will generally not be a perfect estimate of f, and the error in this estimate is the reducible one
* We can potentially improve the accuracy of ˆf via the most appropriate statistical learning technique
* Even if we found the "perfect" estimate of f, we would still have error because b/c Y is also
* a function of **ε**, which cannot be predicted with X, by definition
* Therefore, variability in **ε** also affects prediction accuracy, and is the irreducible error
* **ε** may contain unmeasured variables useful in predicting Y, and since we do not measure them, we cannot use them in f to predict Y
* **ε** may also carry unmeasurable variation (variation in drug manufacturing or in how a patient is feeling may vary the risk of an adverse reaction)



* Inference
* Often interested in understanding *how* Y is affected by how X values change
* Estimating f, but now w/ the goal of making predictions of Y
* Want to understand the *relationship* (how Y changes as a function of/with respect to X)
* Here, f^ cannot be treated as a black box, b/c we need to know its exact form
* Possible interesting questions:
* *Which predictors are associated w/ the response?*
* Often only small fractions of predictors are substantially associated w/ Y
* Must ID a few *important* predictors among a possible large set of them
* *What is the relationship between each predictor + the response?*
* Positive, negative, how strong? Do relationships between the response + a predictor depend on values of other predictors?
* *Can the relationship between Y + each predictor be adequately summarized using a linear equation, or is it more complicated?*
* Most methods for estimating f have been linear, and sometimes this assumption is reasonable/desirable
* But, often, a true relationship is more complicated
* Reasons for estimating f can be a combo of both prediction and inference
* Prediction Ex: Company + a direct-marketing campaign (response) w/ goal of IDing units who will respond positively to a mailing, based on observations of demographic variables (predictors) measures on each unit
* Company doesn’t need deep understanding of the relationships between each predictor + the outcome, just want an accurate model to predict the response w/ the predictors
* Inference Ex: Ask questions:
* Which media channel contributes to sales? Which media generates the largest boost in sales? How much increase in sales is associated w/a given increase in TV ads? What affect will changing price of a product have on sales?
* Combo Ex: Real-estate agent estimating values of homes to inputs like crime rate, zoning, distance from rivers, air quality, schools, community income level, house size, etc.
* May be interested in how individual inputs affect prices, or in predicting house price given its characteristics and if it’s over or under valued
* Linear models are fit for simple + interpretable inference, but may not yield as accurate predictions as some other approaches
* Some highly non-linear approaches can potentially provide very accurate predictions for Y, but at the expense of a less interpretable model, for which inference is more challenging
* Many linear and non-linear approaches for estimating generally share certain characteristics.
* x(i,j) is the value of the jth predictor for the ith observation, and y(i) is the outcome value for i
* Training data therefore consists of { (x1,y1),…..(x(n),y(n) } where x(i) = { xi1, xi2, ….. xij)
* Want to apply a statistical learning method to estimtate an unkonw function f such that Y = f^(X) for any observation (X,Y)
* Most statistical learning methods for this task can be characterized as **parametric** or **non-parametric**
* **Parametric methods** involve a 2-step model-based approach.
* 1. Make an assumption about the **functional form/**shape of f.
* Ex: 1 very simple assumption is that f is linear in X: f(X) = β0 + β1X1 + β2X2 + ... + βpXp.
* Once we have assumed that f is linear, the problem of estimating f is greatly simplified.
* Instead of having to estimate an entirely arbitrary p-dimensional function f(X), one only needs to estimate the *p + 1* coefficients β0-βp.
* 2. After selecting a model, need a procedure that uses training data to fit/train it.
* In the case of the linear model fit train: Estimate the parameters β0-βp = find values of these parameters such that Y ≈ β0 + β1X1 + β2X2 + βpXp.
* The most common approach to fitting the model is **(ordinary) least squares,** but is only 1 of many possible ways to fit a linear model.
* A parametric model-based approach reduces the problem of estimating f down to just estimating a set of parameters.
* Assuming a parametric form for f simplifies the problem of estimating f b/c it is generally much easier to estimate a set of parameters than to fit an entirely arbitrary function
* Potential disadvantage of a parametric approach = model we choose will usually not match the *true* unknown form of f.
* If a chosen model is too far from the true f, our estimate will be poor.
* Can try to address this problem by choosing flexible models that can fit many different possible functional forms flexible for f.
* But in general, fitting a more flexible model requires estimating a greater number of parameter, + more complex models can lead to **overfitting**
* Ex: Linear model applied to Income data 🡪 **income ≈ β0 + β1 × education + β2 × seniority**
* Since we have assumed a linear relationship between the response + 2 predictors, the entire fitting problem reduces to estimating β0, β1, + β2, which we do using **least squares linear regression**.
* The linear fit may not quite be right if the true f has some curvature not captured in the linear fit
* However, the linear fit can still appear to do a reasonable job of capturing a positive relationship between years of education and income, as well as a slightly less positive relationship between seniority and income.
* It may be that with such a small number of observations, this is the best we can do
* **Non-parametric Methods** do NOT make explicit assumptions about the functional form of f.
* Instead they seek an *estimate of f* that gets *as close to the data points as possible* w/out being too rough/wiggly.
* Major advantage over parametric approaches 🡺 *by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f.*
* Any parametric approach brings w/ it the possibility the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well.
* In contrast, non-parametric approaches completely avoid this danger, since essentially no assumption about the form of f is made.
* Non-parametric major disadvantage: They don’t reduce the problem of estimating f to a small number of parameters, so a very large number of observations (far more than typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

* Non-parametric thin-plate spline approach is used to estimate f 🡪 does not impose any pre-specified model on f + instead attempts produce an estimate for f as close as possible to the observed data, subject to the fit being *smooth* (left)
* This non-parametric fit has produced a remarkably accurate estimate of the true f 🡪 no errors seen
* In order to fit a thin-plate spline, a data analyst must select a *level of smoothness*.
* Figure on right shows same thin-plate spline fit using a lower level of smoothness, allowing for a rougher fit 🡪 far more variable than the true function f
* Left = example of overfitting data, an undesirable situation b/c model fit will not yield accurate estimates of the response on new observations that were not part of the original training data set.
* Some methods are less flexible/more restrictive + produce a relatively small range of shapes to estimate f (linear regression is relatively inflexible b/c it can only generate linear functions)
* Other methods (thin plate spline) are considerably more flexible b/c they can generate a much wider range of possible shapes to estimate f.
* *Why would we ever choose to use a more restrictive method instead of a very flexible approach?*
* If mainly interested in *inference*, **restrictive models are much more interpretable**.
* When inference is the goal, a linear model may be a good choice since it will be quite easy to understand the relationship between Y and X1-Xp.
* Very flexible approaches, such as splines + **boosting methods** can lead to such complicated estimates of f where it’s difficult to understand how any individual X is associated w/ Y response
* Trade-off between flexibility + interpretability for some methods:
* 
* Least squares linear regression is relatively inflexible but quite interpretable
* The lasso relies upon the **lasso linear model** but uses an alternative fitting procedure for estimating coefficients β0-βp.
* The new procedure is more restrictive in estimating coefficients, + sets a number of them to exactly 0.
* Hence in this sense the lasso is a less flexible approach than linear regression, but is also more interpretable, b/c in the final model the response variable will only be related to a small subset of predictors those w/ nonzero coefficient estimates)
* **Generalized additive models (GAMs** instead *extend the linear model* to allow for certain non-linear relationships.
* Consequently, GAMs are more flexible than linear regression but also somewhat less interpretable, b/c the relationship between each X + the Y is now modeled using a curve.
* Finally, fully non-linear methods such as **bagging, boosting**, + **support vector machines (SVM)** w/ **non-linear kernels** are highly flexible approaches that are harder to interpret.
* We have established that when inference is the goal, there are clear advantages to using simple + relatively inflexible statistical learning methods.
* When interested in prediction, interpretability of the predictive model is simply not of interest.
* If we seek to develop an algorithm to predict price of stock, our sole requirement for the algorithm is that it predict accurately + interpretability is not a concern 🡪 expect it will be best to use the most flexible model available.
* *Surprisingly, this is not always the case!*
* We will often obtain *more accurate predictions* using a *less flexible method*.
* This phenomenon, which may seem counterintuitive at first glance, has to do with the potential for overfitting in highly flexible methods.
* **Supervised Learning**
* For each observation of the predictor measurement(s) x(i)-x(n), there is an associated response measurement y(i) + we want to fit a model that relates the response to the predictors w/ the aim of:
* accurately predicting the response for future observations (*prediction*)
* Or better understanding the relationship between the response + predictors (*inference*).
* Linear + logistic regression, GAM, boosting, SVM
* **Unsupervised Learnin**g
* Somewhat more challenging situation in which for every observation x(i)-x(n), we observe a vector of measurements X(i) but no associated response y(i).
* It is not possible to fit a linear regression model, since there is no response variable to predict
* working blind 🡪 lack a response variable that can supervise our analysis
* Seek to understand the relationships between variables or between observations.
* **Cluster analysis** to ascertain, on the basis of x(i)-x(n), whether an observations fall into relatively distinct groups
* Ex: Market Segmentation Study 🡪 observe multiple characteristics/variables for potential customers, such as ZIP, family income, shopping habits.
* Might believe that customers fall into different groups, such as big vs spenders.
* If the info about each customer’s spending patterns were available, a supervised analysis would be possible.
* However, this info is NOT available 🡪 we do not know whether each potential customer is a big spender or not
* So try to cluster customers on the basis of variables measured in order to ID distinct groups of potential customers.
* IDing such groups can be of interest b/c it might be that groups differ w/ respect to some property of interest, such as spending habits.



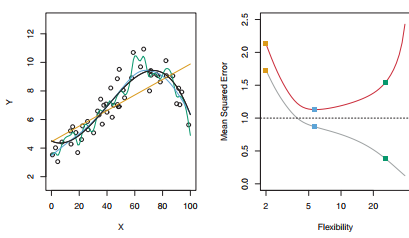
* Left 🡪 Each observation corresponds to one of 3 distinct groups + groups are well-separated + clustering is easier
* Right 🡪 more challenging problem w/ some overlap between groups + clustering could not be expected to assign all overlapping points to their correct group
* W/ 2 variables, one can simply visually inspect scatterplots of observations in order to ID clusters
* However, in practice, data sets contain many more than 2 variables + we cannot easily plot observations
* For this reason, **automated clustering methods** are important
* Many problems fall naturally into supervised or unsupervised learning paradigms but, sometimes the question of whether an analysis should be considered supervised or unsupervised is less clear-cut
* Suppose we have a set of n observations + for m of the observations, we have both predictor + response measurements.
* For the remaining n - m observations, we have predictor but no response measurement.
* Such a scenario can arise if predictors can be measured relatively cheaply but corresponding responses are much more expensive to collect.
* This setting = a **semi-supervised learning problem**🡪 need a statistical learning method that can incorporate the m observations for which response measurements are available as well as the n − m observations for which they are not
* Variables can be characterized as either **quantitative/numerical** or **qualitative/categorical**
* Problems w/ a quantitative response = **regression** problems + qualitative response = **classification**
* However, the distinction is not always that crisp.
* **Least squares linear regression** is used w/ a *quantitative* response, whereas **logistic regression** is typically used w/ a *qualitative* (2-class, or binary) response + as such is often used as a classification
* But since it estimates class probabilities, it can be thought of as a regression
* Some statistical methods, such as **K-nearest neighbors** + **boosting** can be used in the either quantitative or qualitative responses.
* Select statistical learning methods on the basis of whether a response is quantitative or qualitative;
* i.e. might use linear regression when quantitative + logistic regression when qualitative.
* However, whether predictors are qualitative or quantitative is generally considered *less important*.
* Most statistical learning methods can be applied regardless of the predictor variable type, *provided that any qualitative predictors are properly coded before the analysis is performed.*
* Why is it necessary to introduce so many different statistical learning approaches, rather than just a single best method?
* There is **no free lunch** in statistics 🡪 no 1 method dominates all others over all possible data sets.
* On a particular data set, 1 specific method may work best, but some other method may work better on a similar, *but different*, data set.
* It is an important task to decide, for any given set of data, which method produces the best results.
* Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice.
* In order to evaluate performance of a statistical learning method on a given data set, we need to measure how well its predictions actually match observed data, or quantify the extent to which a predicted response for a given observation is close to the true response value for that observation
* In the regression setting, the most commonly-used measure is the **mean squared error (MSE)**



* ˆf^(xi) = the prediction f^ gives for the ith observation
* MSE will be small if predicted responses are very close to true responses + large if for some observations, the predicted + true responses differ substantially.
* MSE computed using the training data is used to fit the model = **training MSE**.
* *But in general, we do not really care how well the method works training on the training data.*
* Rather, we are interested in the accuracy of the predictions we obtain when we apply our method to previously *unseen test data*.
* Suppose we’re interested in developing an algorithm to predict a stock’s price based on previous returns
* Can train the method using stock returns from the past 6 months, but we don’t really care how well our method predicts last week’s stock price.
* Care about how well it will predict tomorrow’s or next month’s price.
* Suppose we have clinical measurements (e.g. weight, BP, height, age, family history of disease) for a number of patients, as well as info about whether each patient has diabetes.
* We can use *these* patients to train a statistical learning method to predict risk of diabetes based on clinical measurements.
* In practice, we want to accurately predict diabetes risk for *future* patients based on *their* clinical measurements
* We are not very interested in whether or not the method accurately predicts diabetes risk for patients used to train the model, since we already know which of those patients have diabetes.
* Suppose we fit our statistical learning method on our training observations {(x1, y1),(x2, y2), ...,(x(n), y(n))} + we obtain the estimate f^ (model)
* We can then compute f^(x1), f^f(x2), ..., f^(x(n)).
* If these are approximately equal to y1, y2,...,y(n), then the training MSE is small.
* However, we’re really not interested in whether f^(x(i)) ≈ y(i) + want to know whether f^(x0) ≈ y0, where (x0, y0) is *a previously unseen test observation not used to train a statistical learning method.*
* We want to choose a method that gives the lowest test MSE, as opposed to the lowest training MSE
* In other words, if we had a large number of test observations, we could compute



* This is the **average squared prediction error** for these test observations (x0, y0).
* We’d like to select a model for which the average of *this* quantity (**test MSE**) is as small as possible.
* How can we go about trying to select a method that minimizes the test MSE?
* In some settings, we may have a test data set available/access to a set of observations not used to train the statistical learning method.
* We can then simply evaluate the MSPE/test MSE on the test observations + select the learning method for which it’s is smallest.
* *But what if no test observations are available?*
* In that case, one might imagine selecting a statistical learning method that minimizes training MSE
* Seems sensible since training + test MSE appear to be closely related.
* Unfortunately, there is a fundamental problem w/ this strategy: **there is no guarantee the method w/ lowest training MSE will also have the lowest test MSE.**
* Many statistical methods specifically estimate coefficients so as to minimize training set MSE.
* For these methods, training set MSE can be quite small, but test MSE is often much larger.



* Left: Data simulated from f in black w/ 3 estimates of f:
* linear regression line (orange) + 2 smoothing spline fits (blue + green)
* The orange, blue + green curves illustrate 3 possible estimates for f obtained using methods w/ increasing levels of flexibility.
* Orange linear regression fit = relatively inflexible
* Blue + green curves produced using smoothing splines w/ different levels of smoothness
* It’s clear that as level of flexibility increases, curves fit the observed data more closely.
* Green curve is the *most* flexible + matches the data very well
* However, it fits the true f (black) poorly because it is *too* wiggly.
* By adjusting level of flexibility of the smoothing spline fit, we can produce many different fits to data.
* Right: *Training* + (grey) *test* MSE (red) + minimum possible test MSE over all methods (dash)
* Grey curve displays the **average training MSE** *as* *a function of flexibility/***degrees of freedom** for a number of smoothing splines.
* **Degrees of freedom** is a quantity that summarizes flexibility of a curve
* Orange, blue + green squares indicate the MSEs associated w/ the corresponding curves on the left
* A more *restricted* + hence *smoother* curve has *fewer degrees of freedom* than a wiggly curve
* Linear regression = most restrictive, w/ 2 dF
* Training MSE declines as flexibility increases.
* In this example, the true f is non-linear, so the orange linear fit is not flexible enough to estimate f well.
* The green curve has the lowest *training* MSE of all 3 methods, since it corresponds to the most flexible of the 3 curves
* We know the true f here, so we can also compute the test MSE over a very large test set as a function of flexibility. (Of course, in general, f is unknown, so this will usually not be possible.)
* Test MSE (red) + as w/ the training MSE, test MSE initially declines as level of flexibility increases.
* However, at some point the test MSE levels off + starts to increase again.
* Consequently, orange + green both have high test MSE + blue minimizes the test MSE, which should not be surprising, given, visually, it appears to estimate f the best in the left
* The horizontal dashed line indicates **Var(),** the *irreducible error*, which corresponds to the lowest achievable test MSE among all possible methods.
* Hence, the blue smoothing spline curve is close to optimal.
* On the right, **as flexibility of the statistical learning method increases, we observe a monotone decrease in training MSE + a U-shape in the test MSE.**
* *This is a fundamental property of statistical learning that holds regardless of the particular data set at hand + regardless of the statistical method being used*.
* **As model flexibility increases, training MSE will decrease, but test MSE may not**.
* When a given method yields a small training MSE but a large test MSE, we are **overfitting** data, b/c our statistical learning procedure is working too hard to find patterns in training + may be picking up some patterns that are just caused by random chance
* rather than by true properties of the unknown function f. When we overfit
* the training data, the test MSE will be very large because the supposed
* patterns that the method found in the training data simply don’t exist
* in the test data. Note that regardless of whether or not overfitting has
* occurred, we almost always expect the training MSE to be smaller than
* the test MSE because most statistical learning methods either directly or
* indirectly seek to minimize the training MSE. Overfitting refers specifically
* to the case in which a less flexible model would have yielded a smaller
* test MSE.
* 2.2 Assessing Model Accuracy 33
* 0 20 40 60 80 100
* 2 4 6 8 10 12
* X
* Y
* 2 5 10 20
* 0.0 0.5 1.0 1.5 2.0 2.5
* Flexibility
* Mean Squared Error
* FIGURE 2.10. Details are as in Figure 2.9, using a different true f that is
* much closer to linear. In this setting, linear regression provides a very good fit to
* the data.
* Figure 2.10 provides another example in which the true f is approximately
* linear. Again we observe that the training MSE decreases monotonically
* as the model flexibility increases, and that there is a U-shape in
* the test MSE. However, because the truth is close to linear, the test MSE
* only decreases slightly before increasing again, so that the orange least
* squares fit is substantially better than the highly flexible green curve. Finally,
* Figure 2.11 displays an example in which f is highly non-linear. The
* training and test MSE curves still exhibit the same general patterns, but
* now there is a rapid decrease in both curves before the test MSE starts to
* increase slowly.
* In practice, one can usually compute the training MSE with relative
* ease, but estimating test MSE is considerably more difficult because usually
* no test data are available. As the previous three examples illustrate, the
* flexibility level corresponding to the model with the minimal test MSE can
* vary considerably among data sets. Throughout this book, we discuss a
* variety of approaches that can be used in practice to estimate this minimum
* point. One important method is cross-validation (Chapter 5), which is a crossmethod
* for estimating test MSE using the training data.