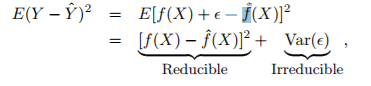
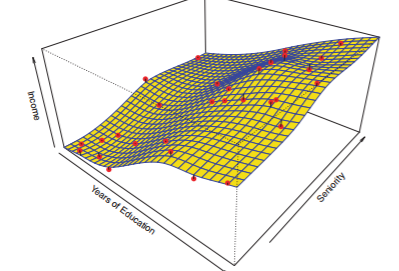
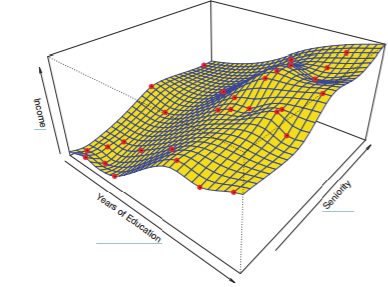
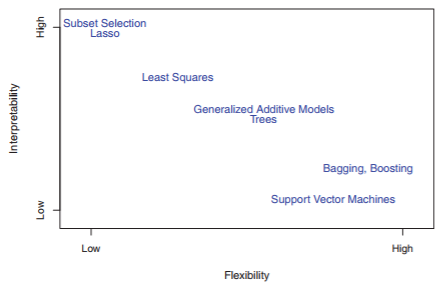
Ch 2

* 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.