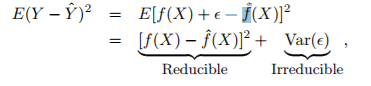
Ch 2

* 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

'