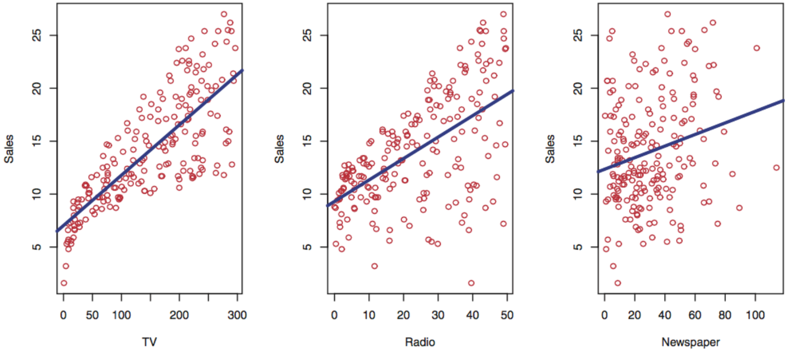
Lecture 2: Statistical Learning (Textbook 2.1)

Nayel Bettache

Department of Statistical Science, Cornell University

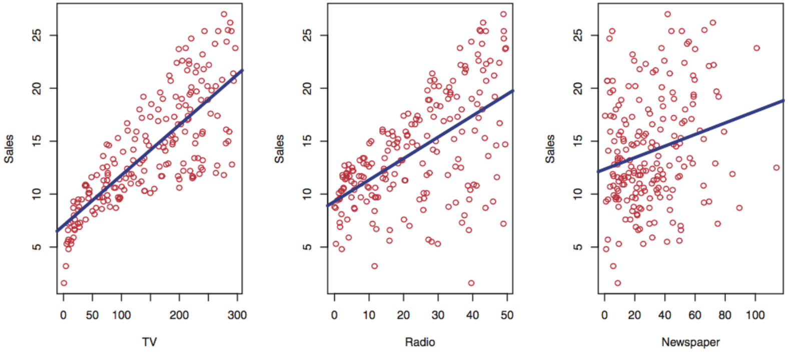
A Simple Example

- **Advertising data set**: sales of a product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.



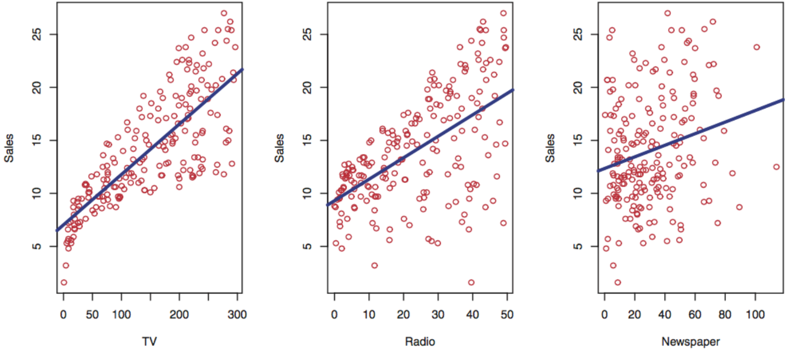
A Simple Example

- **Advertising data set**: sales of a product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.
- Suppose that we are statistical consultants hired to investigate the association between advertising and sales of this product.



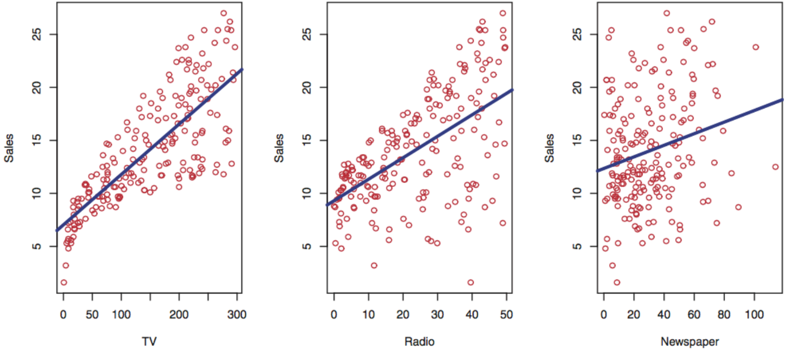
A Simple Example

- It is not possible for our client to directly increase sales of the product.



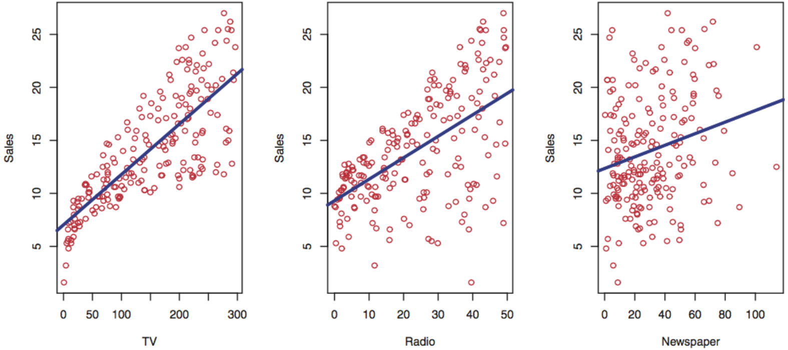
A Simple Example

- It is not possible for our client to directly increase sales of the product.
- They can control the advertising expenditure in each of the three media.



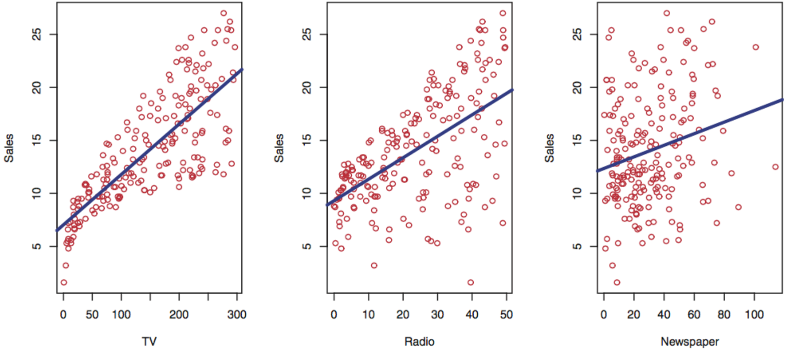
A Simple Example

- It is not possible for our client to directly increase sales of the product.
- They can control the advertising expenditure in each of the three media.
- 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.



A Simple Example

- It is not possible for our client to directly increase sales of the product.
- They can control the advertising expenditure in each of the three media.
- 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.
- Goal: Develop a model that can be used to predict sales on the basis of the three media budgets: $Sales \approx f(TV, Radio, Newspaper)$.'



Notation

- Sales is a target we want to predict. It will be denoted Y.

Notation

- Sales is a target we want to predict. It will be denoted Y.
- TV budget, Radio budget and Newspaper budget are predictors. They will be denoted X_1, X_2, X_3 respectively.

Notation

- Sales is a target we want to predict. It will be denoted Y.
- TV budget, Radio budget and Newspaper budget are predictors. They will be denoted X_1, X_2, X_3 respectively.
- The predictor vector is denoted $X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$.

Notation

- Sales is a target we want to predict. It will be denoted Y.
- TV budget, Radio budget and Newspaper budget are predictors. They will be denoted X_1, X_2, X_3 respectively.
- The predictor vector is denoted $X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$.
- We consider the regression model $Y = f(X) + \epsilon$, where ϵ is a random error term, independent of X with zero mean, capturing measurement errors.

Notation

- Sales is a target we want to predict. It will be denoted Y.
- TV budget, Radio budget and Newspaper budget are predictors. They will be denoted X_1, X_2, X_3 respectively.
- The predictor vector is denoted $X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$.
- We consider the regression model $Y = f(X) + \epsilon$, where ϵ is a random error term, independent of X with zero mean, capturing measurement errors.
- f is some fixed but unknown function. It represents the systematic information that X provides about Y.

Notation

- Sales is a target we want to predict. It will be denoted Y.
- TV budget, Radio budget and Newspaper budget are predictors. They will be denoted X_1, X_2, X_3 respectively.
- The predictor vector is denoted $X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$.
- We consider the regression model $Y = f(X) + \epsilon$, where ϵ is a random error term, independent of X with zero mean, capturing measurement errors.
- f is some fixed but unknown function. It represents the systematic information that X provides about Y.
- Objective: Estimate f based on the observed samples.

Why Estimate f ?

We denote $\hat{f}$ the estimate of f based on the observed samples.

Why Estimate f ?

We denote $\hat{f}$ the estimate of f based on the observed samples.
In machine learning, $\hat{f}$ is often a *black box*, in the sense that one is not typically concerned with the exact form of $\hat{f}$.

Why Estimate f ?

We denote $\hat{f}$ the estimate of f based on the observed samples.
In machine learning, $\hat{f}$ is often a *black box*, in the sense that one is not typically concerned with the exact form of $\hat{f}$.
There are two main reasons that we may wish to estimate f:

Why Estimate f ?

We denote $\hat{f}$ the estimate of f based on the observed samples.
In machine learning, $\hat{f}$ is often a *black box*, in the sense that one is not typically concerned with the exact form of $\hat{f}$.
There are two main reasons that we may wish to estimate f:

- **Prediction**: With a good $\hat{f}$ we can make predictions of Y at new unobserved points X. We then would have $\hat{Y} = \hat{f}(X)$.

Why Estimate f ?

We denote $\hat{f}$ the estimate of f based on the observed samples.
In machine learning, $\hat{f}$ is often a *black box*, in the sense that one is not typically concerned with the exact form of $\hat{f}$.
There are two main reasons that we may wish to estimate f:

- **Prediction**: With a good $\hat{f}$ we can make predictions of Y at new unobserved points X. We then would have $\hat{Y} = \hat{f}(X)$.
- **Inference**: We can understand which components of $X = (X_1, X_2, \ldots, X_p)$ are important in explaining Y, and which are irrelevant.

Prediction

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

Prediction

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

- **Reducible error**: $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error.

Prediction

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

- **Reducible error**: $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error.
  This error is reducible because we can potentially improve the accuracy of $\hat{f}$ by using a more efficient statistical learning technique to estimate f.

Prediction

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

- **Reducible error**: $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error.
  This error is reducible because we can potentially improve the accuracy of $\hat{f}$ by using a more efficient statistical learning technique to estimate f.
- **Irreducible error**: Even if we retrieve the exact f that generated the target, our prediction would still have some error in it!

Prediction

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

- **Reducible error**: $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error.
  This error is reducible because we can potentially improve the accuracy of $\hat{f}$ by using a more efficient statistical learning technique to estimate f.
- **Irreducible error**: Even if we retrieve the exact f that generated the target, our prediction would still have some error in it!
  This is because Y is also a function of ϵ which, by definition, cannot be predicted using X.

Prediction

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

- **Reducible error**: $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error.
  This error is reducible because we can potentially improve the accuracy of $\hat{f}$ by using a more efficient statistical learning technique to estimate f.
- **Irreducible error**: Even if we retrieve the exact f that generated the target, our prediction would still have some error in it!
  This is because Y is also a function of ϵ which, by definition, cannot be predicted using X.

Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$.

Prediction

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

- **Reducible error**: $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error.
  This error is reducible because we can potentially improve the accuracy of $\hat{f}$ by using a more efficient statistical learning technique to estimate f.
- **Irreducible error**: Even if we retrieve the exact f that generated the target, our prediction would still have some error in it!
  This is because Y is also a function of ϵ which, by definition, cannot be predicted using X.

Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$.
Assume here that both $\hat{f}$ and X are fixed, so that the only variability comes from ϵ. Then we have:

Prediction

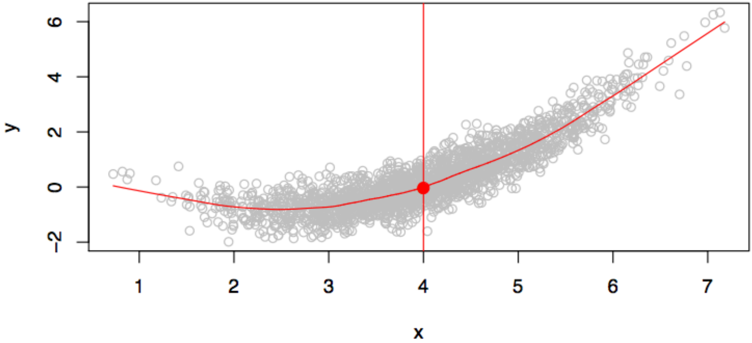
The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

- **Reducible error**: $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error.
  This error is reducible because we can potentially improve the accuracy of $\hat{f}$ by using a more efficient statistical learning technique to estimate f.
- **Irreducible error**: Even if we retrieve the exact f that generated the target, our prediction would still have some error in it!
  This is because Y is also a function of ϵ which, by definition, cannot be predicted using X.

Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$.
Assume here that both $\hat{f}$ and X are fixed, so that the only variability comes from ϵ. Then we have:

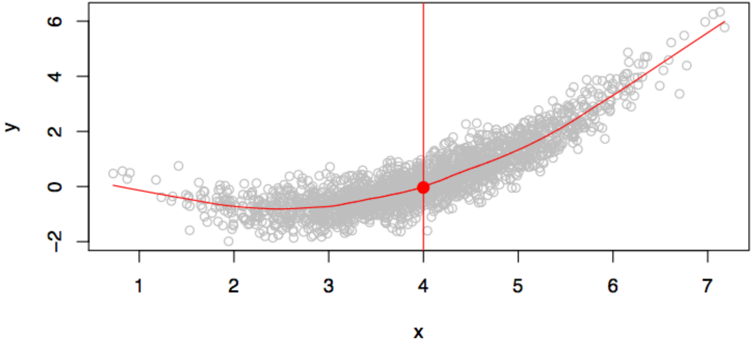
$$\begin{aligned}\mathbb{E}\left[\left(\hat{Y} - Y\right)^2\right] &= \mathbb{E}\left[\left(\hat{f}(X) + \epsilon - f(X)\right)^2\right] \\ &= \underbrace{\left(\hat{f}(X) - f(X)\right)^2}_{\textit{reducible}} + \underbrace{\mathbb{V}(\epsilon)}_{\textit{irreducible}}\end{aligned}$$

Irreducible error



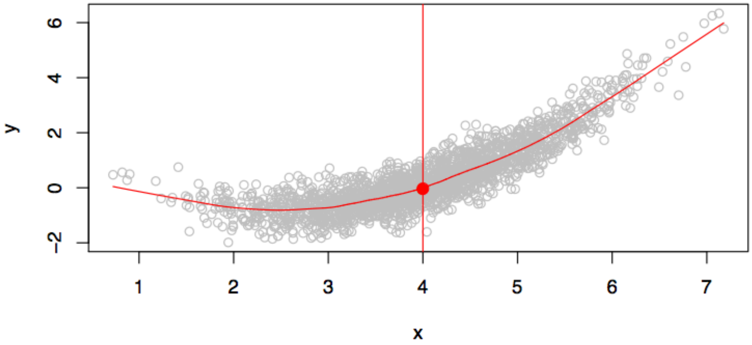
Given this dataset, the red curve seems to be a good estimate of f.

Irreducible error



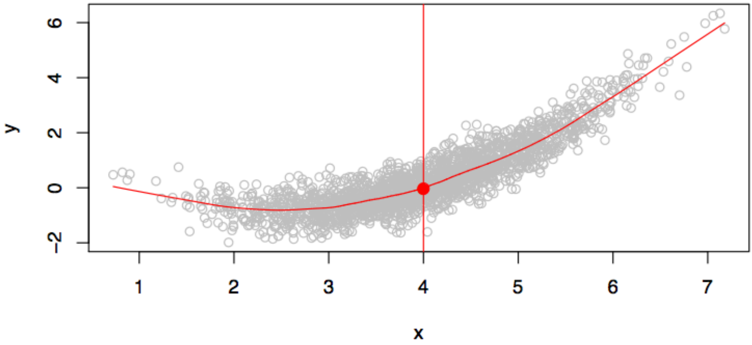
Given this dataset, the red curve seems to be a good estimate of f.
This predictor $\hat{f}$ is defined as follows:

Irreducible error



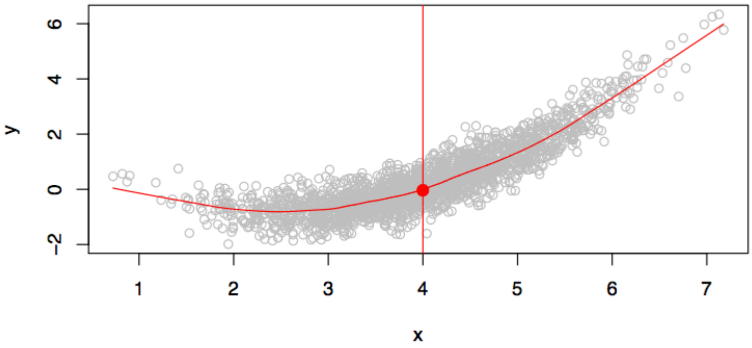
Given this dataset, the red curve seems to be a good estimate of f.
This predictor $\hat{f}$ is defined as follows:
For each value x taken by the predictor X, we consider $\hat{f}(x) = E(Y|X = x)$.

Irreducible error



Given this dataset, the red curve seems to be a good estimate of f.
This predictor $\hat{f}$ is defined as follows:
For each value x taken by the predictor X, we consider $\hat{f}(x) = E(Y|X = x)$.
$E(Y|X = x)$ is the expected value of Y given $X = x$. Basically $\hat{f}$ returns the average of all the observed values of Y when predictors take the value x.

Irreducible error



Given this dataset, the red curve seems to be a good estimate of f.
This predictor $\hat{f}$ is defined as follows:
For each value x taken by the predictor X, we consider $\hat{f}(x) = E(Y|X = x)$.
$E(Y|X = x)$ is the expected value of Y given $X = x$. Basically $\hat{f}$ returns the average of all the observed values of Y when predictors take the value x.
$\hat{f}(x) = E(Y|X = x)$ is called the *regression function*.

Inference

We would like to estimate f based on the training data.

Inference

We would like to estimate f based on the training data.

- We can understand which components of $X = (X_1, X_2, ..., X_p)$ are important in explaining Y, and which are irrelevant.

Inference

We would like to estimate f based on the training data.

- We can understand which components of $X = (X_1, X_2, ..., X_p)$ are important in explaining Y, and which are irrelevant.
- Depending on the complexity of f, we may be able to understand how each component X_j of X affects Y.

Inference

We would like to estimate f based on the training data.

- We can understand which components of $X = (X_1, X_2, ..., X_p)$ are important in explaining Y, and which are irrelevant.
- Depending on the complexity of f, we may be able to understand how each component X_j of X affects Y.
- Now, $\hat{f}$ cannot be treated as a black box, because we need to know its form to know the relationship between X and Y.

Inference

We would like to estimate f based on the training data.

- We can understand which components of $X = (X_1, X_2, ..., X_p)$ are important in explaining Y, and which are irrelevant.
- Depending on the complexity of f, we may be able to understand how each component X_j of X affects Y.
- Now, $\hat{f}$ cannot be treated as a black box, because we need to know its form to know the relationship between X and Y.
- Trade-off between prediction and inference: Linear models allow for simple and interpretable inference, but may not yield good predictions; non-linear models (introduced later) may have better prediction but is less interpretable and inference is challenging.

Inference

We would like to estimate f based on the training data.

- We can understand which components of $X = (X_1, X_2, ..., X_p)$ are important in explaining Y, and which are irrelevant.
- Depending on the complexity of f, we may be able to understand how each component X_j of X affects Y.
- Now, $\hat{f}$ cannot be treated as a black box, because we need to know its form to know the relationship between X and Y.
- Trade-off between prediction and inference: Linear models allow for simple and interpretable inference, but may not yield good predictions; non-linear models (introduced later) may have better prediction but is less interpretable and inference is challenging.

Two different approaches to estimate f: **Parametric methods** and **Non-parametric methods.**

Parametric methods: presentation

Parametric methods involve a two-step model-based approach.

Parametric methods: presentation

Parametric methods involve a two-step model-based approach.

1. Make an assumption about the functional form of f.

Parametric methods: presentation

Parametric methods involve a two-step model-based approach.

1. Make an assumption about the functional form of f.
   - For example, one very simple assumption is that f is linear in X.

Parametric methods: presentation

Parametric methods involve a two-step model-based approach.

1. Make an assumption about the functional form of f.
    - For example, one very simple assumption is that f is linear in X.
    - $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_p X_p.$$

Parametric methods: presentation

Parametric methods involve a two-step model-based approach.

1. Make an assumption about the functional form of f.
    - For example, one very simple assumption is that f is linear in X.
    - $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_p X_p.$$
    - Once we have assumed that f is linear, the problem of estimating f is greatly simplified.

Parametric methods: presentation

Parametric methods involve a two-step model-based approach.

1. Make an assumption about the functional form of f.
    - For example, one very simple assumption is that f is linear in X.
    - $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_p X_p.$$
    - 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, one only needs to estimate the $p + 1$ coefficients $\beta_0, ..., \beta_p$.

Parametric methods: presentation

Parametric methods involve a two-step model-based approach.

1. Make an assumption about the functional form of f.
    - For example, one very simple assumption is that f is linear in X.
    - $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... \beta_p X_p.$$
    - 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, one only needs to estimate the $p + 1$ coefficients $\beta_0, ..., \beta_p$.
2. After a model has been selected, we need a procedure that uses the training data to fit or train the model.

Parametric methods: presentation

Parametric methods involve a two-step model-based approach.

1. Make an assumption about the functional form of f.
   - For example, one very simple assumption is that f is linear in X.
   - $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p.$$
   - 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, one only needs to estimate the $p+1$ coefficients $\beta_0, \ldots., \beta_p$.
2. After a model has been selected, we need a procedure that uses the training data to fit or train the model.
   - The most common approach to fitting the model is the ordinary least squares.

Parametric methods: presentation

Parametric methods involve a two-step model-based approach.

1. Make an assumption about the functional form of f.
   - For example, one very simple assumption is that f is linear in X.
   - $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... \beta_p X_p.$$
   - 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, one only needs to estimate the $p + 1$ coefficients $\beta_0, ..., \beta_p$.
2. After a model has been selected, we need a procedure that uses the training data to fit or train the model.
   - The most common approach to fitting the model is the ordinary least squares.

The model-based approach just described is referred to as parametric; it reduces the problem of estimating f down to one of estimating a set of parameters.

Parametric methods: pros and cons

Assuming a parametric form for f has pros and cons.

Parametric methods: pros and cons

Assuming a parametric form for f has pros and cons.

- **Pros**: It simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters than it is to fit an entirely arbitrary function f.

Parametric methods: pros and cons

Assuming a parametric form for f has pros and cons.

- **Pros**: It simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters than it is to fit an entirely arbitrary function f.
- **Cons**: The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f.

Parametric methods: pros and cons

Assuming a parametric form for f has pros and cons.

- **Pros**: It simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters than it is to fit an entirely arbitrary function f.
- **Cons**: The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f.
  - If the chosen model is too far from the true f, then our estimate will be poor.

Parametric methods: pros and cons

Assuming a parametric form for f has pros and cons.

- **Pros**: It simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters than it is to fit an entirely arbitrary function f.
- **Cons**: The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f.
  - If the chosen model is too far from the true f, then our estimate will be poor.
  - We can try to address this problem by choosing flexible models that can ft many different possible functional forms flexible for f.

Parametric methods: pros and cons

Assuming a parametric form for f has pros and cons.

- **Pros**: It simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters than it is to fit an entirely arbitrary function f.
- **Cons**: The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f.
  - If the chosen model is too far from the true f, then our estimate will be poor.
  - We can try to address this problem by choosing flexible models that can ft many different possible functional forms flexible for f.
  - In general, fitting a more flexible model requires estimating a greater number of parameters.

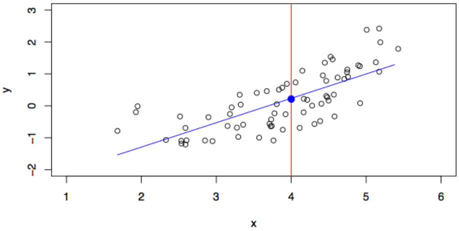
Parametric methods: pros and cons

Assuming a parametric form for f has pros and cons.

- **Pros**: It simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters than it is to fit an entirely arbitrary function f.
- **Cons**: The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f.
  - If the chosen model is too far from the true f, then our estimate will be poor.
  - We can try to address this problem by choosing flexible models that can ft many different possible functional forms flexible for f.
  - In general, fitting a more flexible model requires estimating a greater number of parameters.
  - These more complex models can lead to a phenomenon known as overfitting the data, which essentially means they follow the errors, or noise, too closely.

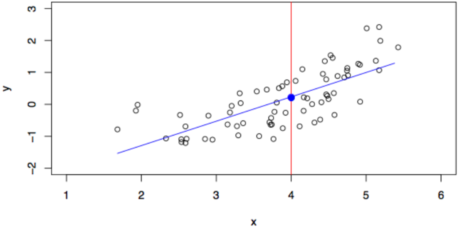
Parametric methods

A linear model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here

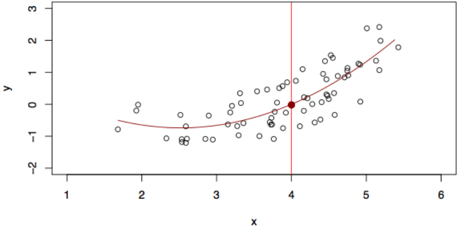


Parametric methods

A linear model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here



A more flexible model $\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ gives a slightly better fit



Non-parametric methods

Non-parametric methods do not make explicit assumptions about the functional form of f.

Non-parametric methods

Non-parametric methods do not make explicit assumptions about the functional form of f.
They seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.

Non-parametric methods

Non-parametric methods do not make explicit assumptions about the functional form of f.

They seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.

- **Pros**: Potential to accurately fit a wider range of possible shapes for f.

Non-parametric methods

Non-parametric methods do not make explicit assumptions about the functional form of f.
They seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.

- **Pros**: Potential to accurately fit a wider range of possible shapes for f.
  - Any parametric approach brings with it the possibility that 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.

Non-parametric methods

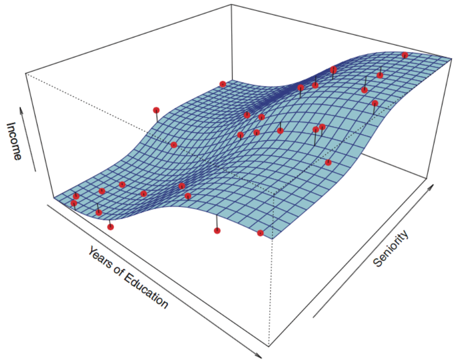
Non-parametric methods do not make explicit assumptions about the functional form of f.

They seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.

- **Pros**: Potential to accurately fit a wider range of possible shapes for f.
    - Any parametric approach brings with it the possibility that 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.
- **Cons**: Since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations is required in order to obtain an accurate estimate for f.

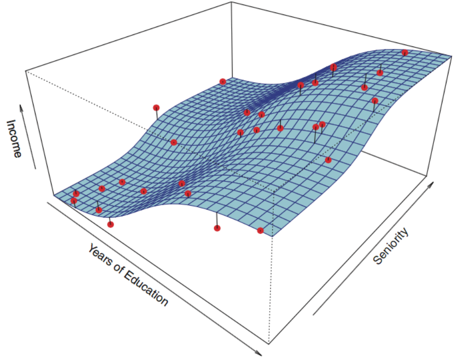
Example: Simulated data points

- Consider the following simulated example.



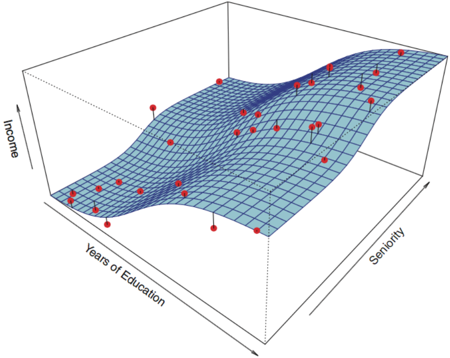
Example: Simulated data points

- Consider the following simulated example.
- Red points are simulated values for income from the model *income* $= f(\textit{education}, \textit{seniority}) + \epsilon$ where f is the blue surface and ϵ a random noise.



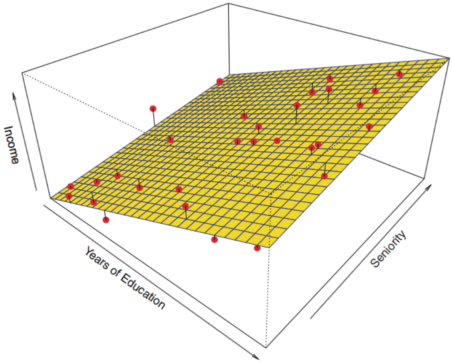
Example: Simulated data points

- Consider the following simulated example.
- Red points are simulated values for income from the model $income = f(education, seniority) + \epsilon$ where f is the blue surface and ϵ a random noise.
- If we are only given the red points, how can we estimate the blue surface ?



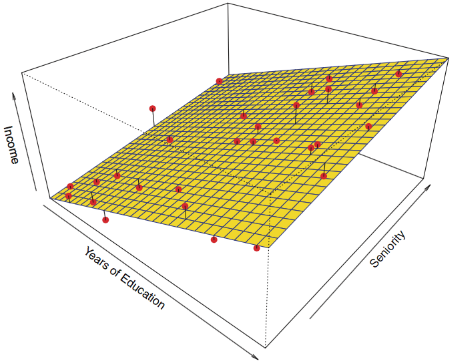
Example 1: Linear regression

- We estimate the blue surface with linear regression.



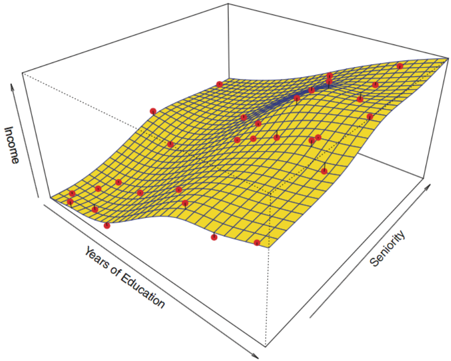
Example 1: Linear regression

- We estimate the blue surface with linear regression.
- $\hat{f}_L(education, seniority) = \hat{\beta}_0 + \hat{\beta}_1 \times education + \hat{\beta}_2 \times seniority.$



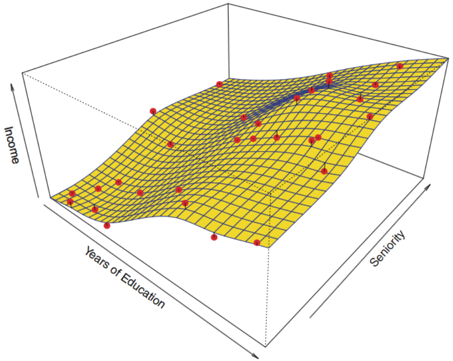
Example2: Non-parametric method

- We estimate the blue surface with a non parametric method.



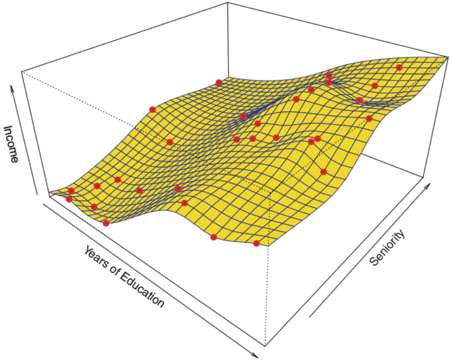
Example2: Non-parametric method

- We estimate the blue surface with a non parametric method.
- Looks closer to the target blue surface !



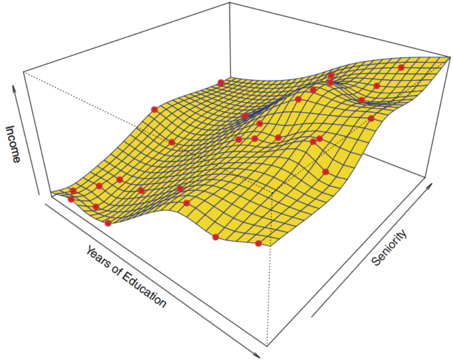
Example3: Overfitting

- We estimate the blue surface with a too flexible non parametric method.



Example3: Overfitting

- We estimate the blue surface with a too flexible non parametric method.
- This fit makes zero errors on the training data!



Some Trade-off

- **Prediction accuracy (flexibility) versus interpretability.**
  Linear models are easy to interpret; thin-plate splines are not.

Some Trade-off

- **Prediction accuracy (flexibility) versus interpretability.**
  Linear models are easy to interpret; thin-plate splines are not.
- **Good fit versus over-fit or under-fit.**
  How do we know when the fit is just right?

Some Trade-off

- **Prediction accuracy (flexibility) versus interpretability.**
  Linear models are easy to interpret; thin-plate splines are not.
- **Good fit versus over-fit or under-fit.**
  How do we know when the fit is just right?
- **Parsimony versus black-box.**
  We often prefer a simpler model involving fewer variables over a black-box predictor involving them all.

Flexibility versus interpretability