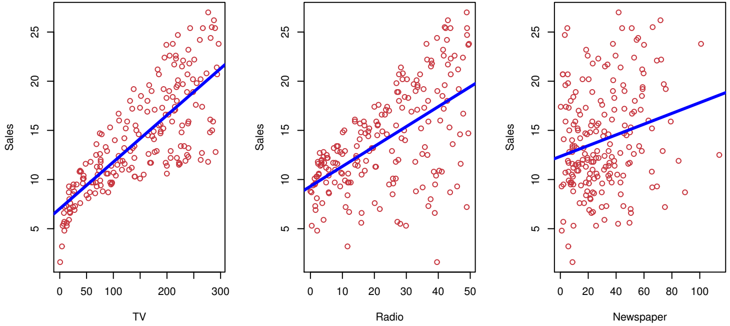
What is Statistical Learning?



- Shown are $\texttt{Sales}$ vs $\texttt{TV}$, $\texttt{Radio}$ and $\texttt{Newspaper}$, with a blue linear-regression line fit separately to each.
- Can we predict $\texttt{Sales}$ using these three?
- Perhaps we can do better using a model

$$\texttt{Sales} \approx f(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$

Notations

- Here `Sales` is a *response* or *target* that we wish to predict. We generically refer to the response as Y.
- `TV` is a *feature*, or *input*, or *predictor*. We name it X_1.
- Likewise, name `Radio` as X_2, and `Newspaper` as X_3.
- We can refer to the *input vector* collectively as

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}.$$

- Now we write our model as

$$Y = f(X) + \epsilon$$

  where ϵ captures measurement errors and other discrepancies, and is independent of X and has mean zero.

Why learn $f(X)$?

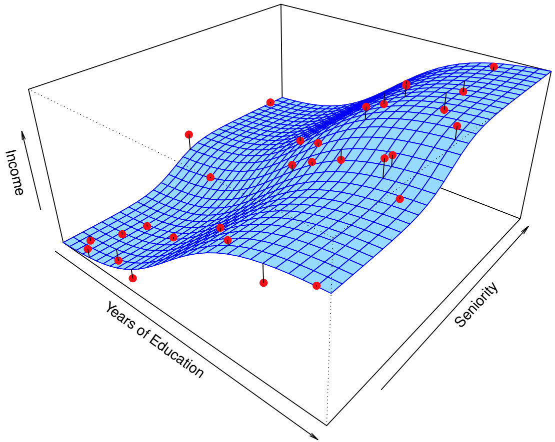
- With a good f we can make predictions of Y at new points $X = x$.
- We can understand which components of $X = (X_1, X_2, ..., X_p)$ are important in explaining Y, and which are irrelevant.
    - e.g. features such as `Seniority` and `Years of Education` have a big impact on a response `Income`, but `Marital Status` feature may not.
- Depending on the complexity of f, we may be able to understand how each component X_j of X affects Y.

Parametric and structured models

The linear model is an important example of a parametric model:

$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

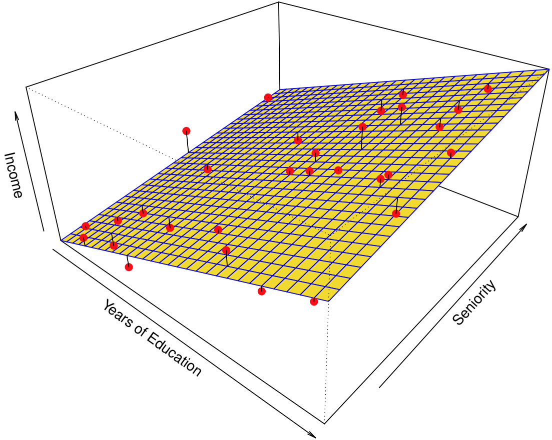
- A linear model is specified in terms of $p + 1$ parameters $\beta_0, \beta_1, ..., \beta_p$.
- We estimate the parameters by fitting the model to training data.
- Although the linear model assumption is almost never exactly correct, a linear model often serves as a good and interpretable approximation to the unknown true function $f(X)$.



Simulated example. Red points are simulated values for income from the model

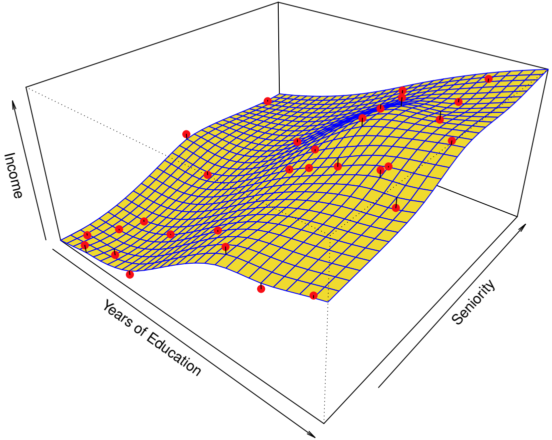
$$\texttt{income} = f(\texttt{education}, \texttt{seniority}) + \epsilon$$

f is the blue surface.

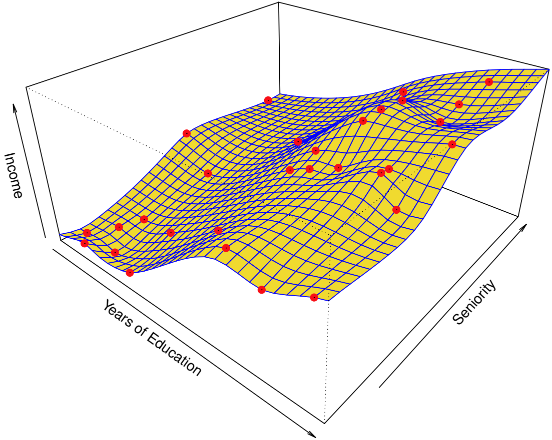


Linear regression model fit to the simulated data.

$$\hat{f}(\texttt{education}, \texttt{seniority}) = \hat{\beta}_0 + \hat{\beta}_1 \times \texttt{education} + \hat{\beta}_2 \times \texttt{seniority}$$



More flexible regression model $\hat{f}_S(\texttt{education}, \texttt{seniority})$ fit to the simulated data. Here we use a technique called a thin-plate spline to fit a flexible surface. We control the roughness of the fit (Ch.7 in the textbook).

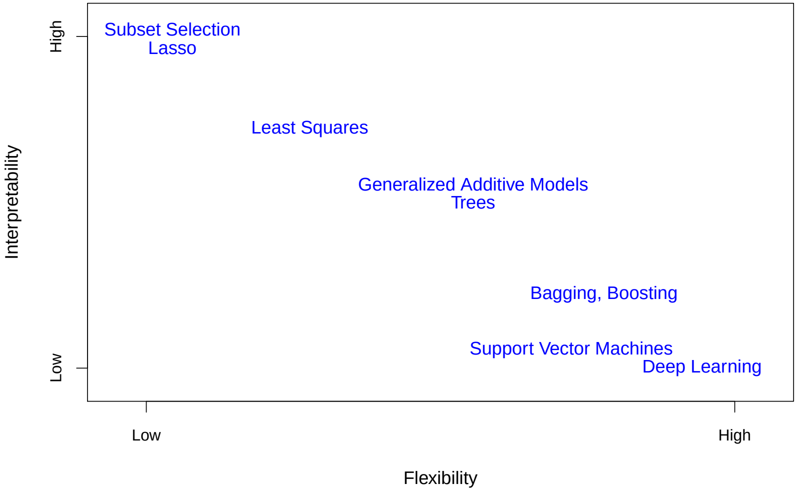


Even more flexible spline regression model $\hat{f}_S($education, seniority$)$ fit to the simulated data. Here the fitted model makes no errors on the training data! Also known as overfitting.

Some trade-offs

- Prediction accuracy 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.

Interpretability vs. flexibility



Assessing model accuracy

Suppose we fit a model $\hat{f}(x)$ to some training data $\mathsf{Tr} = \{x_i, y_i\}_{i=1}^N$, and we wish to see how well it performs.

- We could compute the average squared prediction error over Tr:

$$\text{MSE}_{\mathsf{Tr}} = \text{Ave}_{i \in \mathsf{Tr}} \left[y_i - \hat{f}(x_i) \right]^2 = \frac{1}{N} \sum_{i \in \mathsf{Tr}} \left[y_i - \hat{f}(x_i) \right]^2$$

- This may be biased toward more overfit models!

Assessing model accuracy

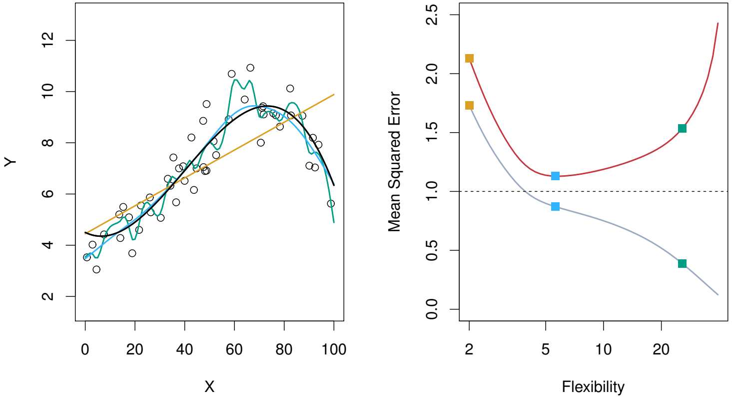
Suppose we fit a model $\hat{f}(x)$ to some training data $\mathsf{Tr} = \{x_i, y_i\}_{i=1}^N$, and we wish to see how well it performs.

- We could compute the average squared prediction error over Tr:

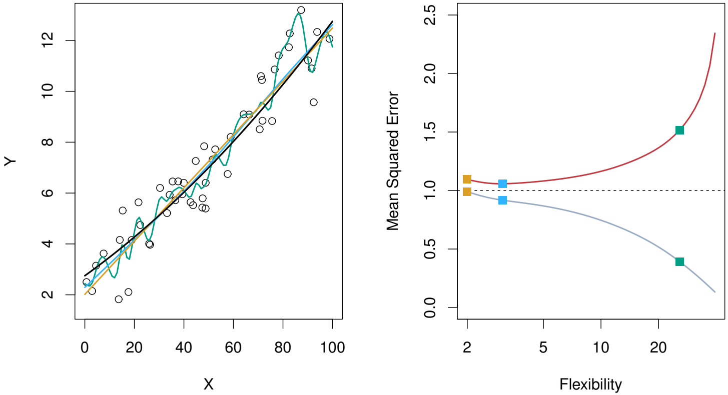
$$\text{MSE}_{\mathsf{Tr}} = \text{Ave}_{i \in \mathsf{Tr}} \left[y_i - \hat{f}(x_i) \right]^2 = \frac{1}{N} \sum_{i \in \mathsf{Tr}} \left[y_i - \hat{f}(x_i) \right]^2$$

- This may be biased toward more overfit models!
- Instead we should, if possible, compute it using fresh test data $\mathsf{Te} = \{x_i, y_i\}_{i=1}^M$ diffrent from the training data

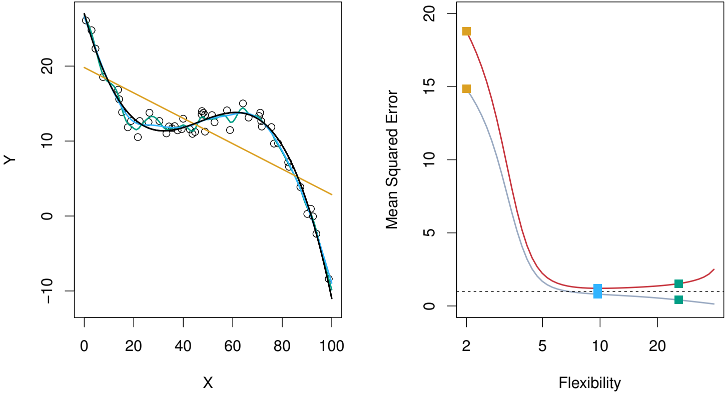
$$\text{MSE}_{\mathsf{Te}} = \text{Ave}_{i \in \mathsf{Te}} \left[y_i - \hat{f}(x_i) \right]^2 = \frac{1}{M} \sum_{i \in \mathsf{Te}} \left[y_i - \hat{f}(x_i) \right]^2$$



- Black curve on the left is (unknown) truth.
- Red curve on right is $\mathrm{MSE}_{\mathrm{Te}}$, grey curve is $\mathrm{MSE}_{\mathrm{Tr}}$.
- Orange curve/square: linear model
- Blue and green curves/squares: two different smoothing splines



- Here the truth is smoother, so the smoother fit and linear model do well.



- Here the truth is wiggly and the noise is low, so the more flexible fits do better.

Bias-Variance Trade-off

- Suppose we have fit a model $\hat{f}(x)$ to some training data `Tr`, and let (x_0, y_0) be a test observation drawn from the population. If the true model is $Y = f(X) + \epsilon$ (with $f(x) = \mathbb{E}[Y|X = x]$), then

$$\mathbb{E}\left[y_0 - \hat{f}(x_0)\right]^2 = \mathrm{Var}\big(\hat{f}(x_0)\big) + \left[\mathrm{Bias}\big(\hat{f}(x_0)\big)\right]^2 + \mathrm{Var}(\epsilon)$$

- What is the expectation over?

Bias-variance trade-off for the three examples