Machine Learning and Programming in Python

Lecture for Master and PhD students

Chair of Data Science in Economics

Ruhr University Bochum

Summer semester 2024

Lecture 4

Model complexity

- Statistical Learning $\Rightarrow$ methods to understand data
  - Fitting a model y=f(x)
  - Prediction
  - Inference
- Measuring model performance – Mean squared error (MSE)
- Under vs. over-fitting
- Bias-variance trade-off – intuition and proof

- Example: a lending firm aims to improve quality of its applicant screening.
- For each new applicant:
  - Input (X): Credit score, income, wealth, occupation, demographics.
  - Output (Y): Default or not?
- Can build a model using existing clients to estimate a relationship between observed X and Y.
  $y = f(X) + \epsilon$
- f is an unknown fixed function of X; systematic information in X about y.
- ϵ is random noise error (of mean zero).

Prediction (focus in Machine Learning)

- Interested in the the most accurate prediction $\hat{y}$, $\hat{y} = \hat{f}(X)$
- $\hat{f}$ often treated as a 'black box'.
- Two sources of prediction error:
  - Reducible Error: $\hat{f}$, $\hat{f}$ is the error in modelling the true unknown f .
  - Irreducible error: ϵ combines unmeasurable variables and pure randomness.

(Causal) **Inference** (focus in Empirical Economics/ Econometrics):

- Understand how y changes as $x_1, x_2, \ldots$ change.
- What is the relation f between X and y
- Not a black box approach; need to unpack f .
- Which X variables matter and how do they affect y ?
- f should be easy to interpret; often prefer linear or simple models.

Prediction:

- Modeling: Consider several different models and different parameter settings.
- Model selection: Identify the model with the greatest predictive performance using validation/test sets; select the model with the highest performance on the test set.
- Prediction: Apply the selected model on new data with the expectation that the selected model also generalizes to the unseen data.

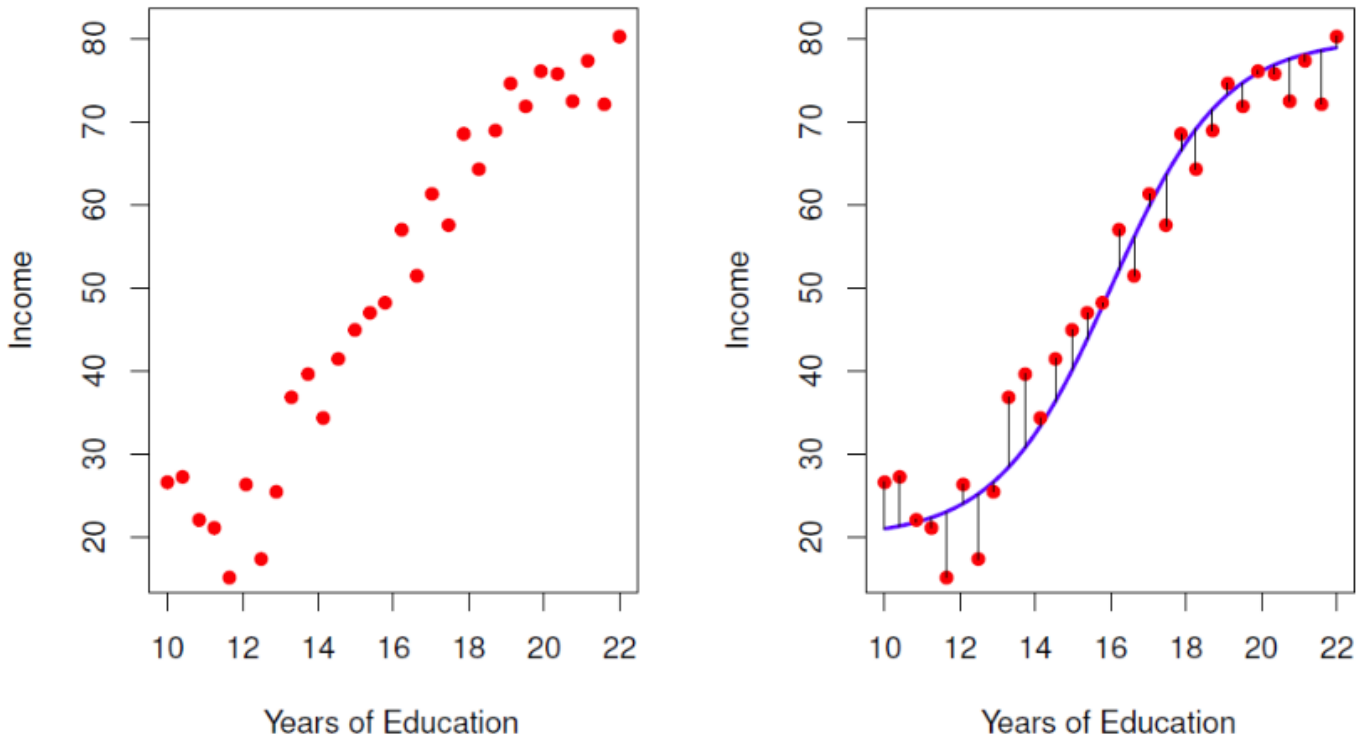
Inference:

- Modeling: Reason about the data generation process and choose the stochastic model that approximates the data generation process best.
- Model validation: Evaluate the validity of the stochastic model using residual analysis or goodness-of-fit tests.
- Inference: Use the stochastic model to understand the data generation process .

- Suppose we wish to estimate how demand for hotel rooms changes with price.
- A Machine Learning analyst has historical data on daily occupancy and prices.
  - A model to predict occupancy as a function of price can be trained.
  - Will find that high price implies high occupancy!
  - Reverse causality: Hotel yield management systems raise price as rooms fill up.
- An accurate prediction does not give us the required answer.
  - Example: What would happen to occupancy if price was raised by 5%?
- Need a natural experiment or an instrumental variable helping isolate "exogenous" changes in price.

Model fit

FIGURE 2.2. *The* `Income` *data set.* Left: *The red dots are the observed values of* `income` *(in tens of thousands of dollars) and* `years of education` *for* 30 *individuals.* Right: *The blue curve represents the true underlying relationship between* `income` *and* `years of education`*, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.*



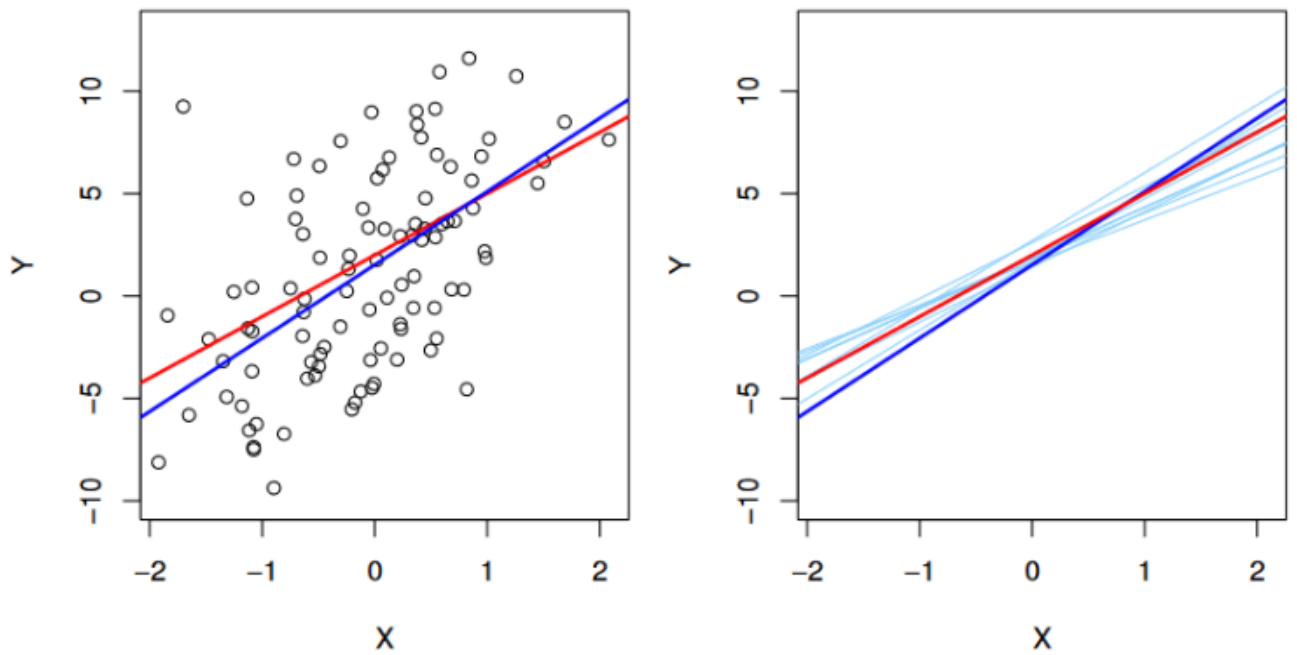
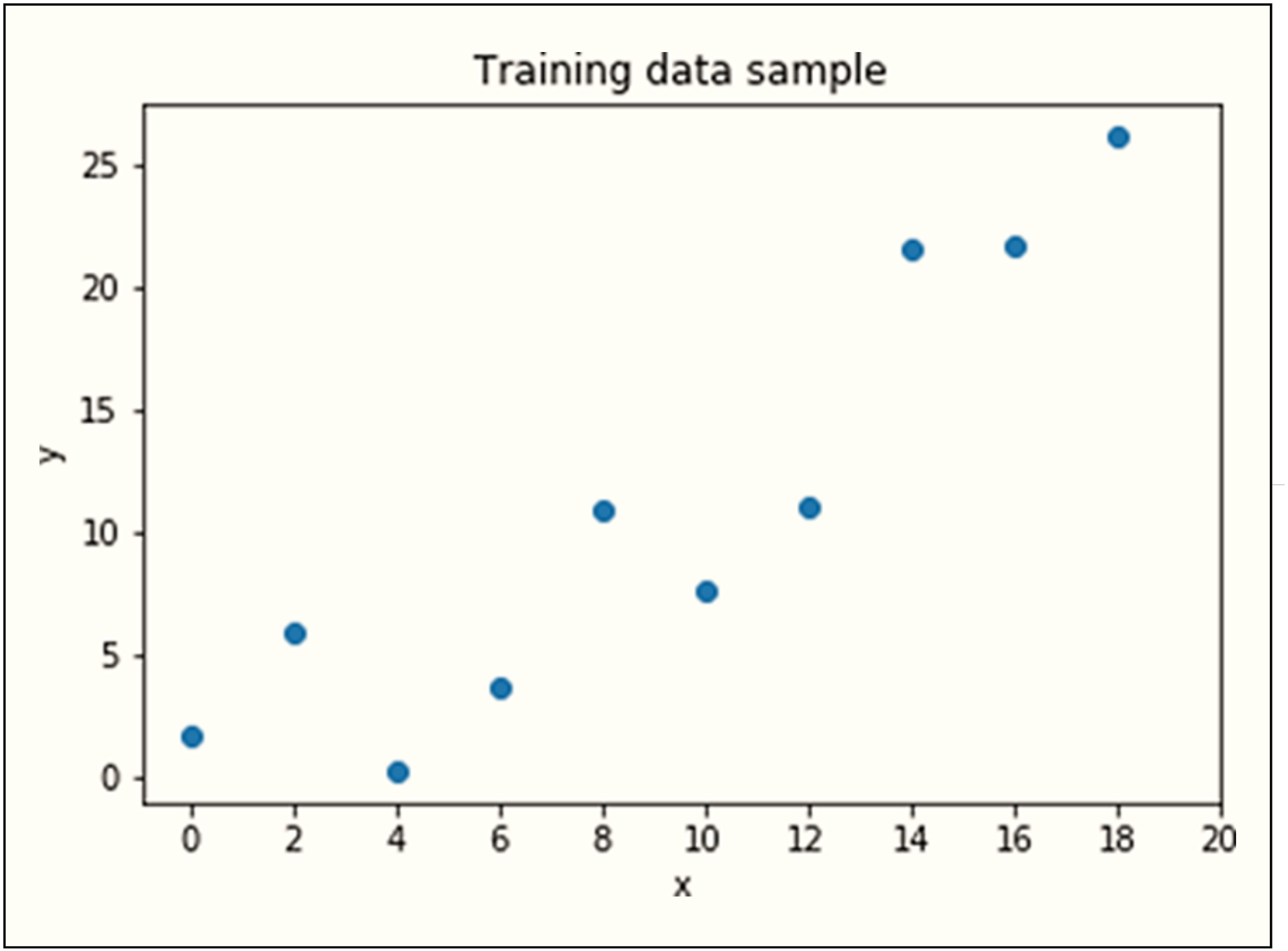
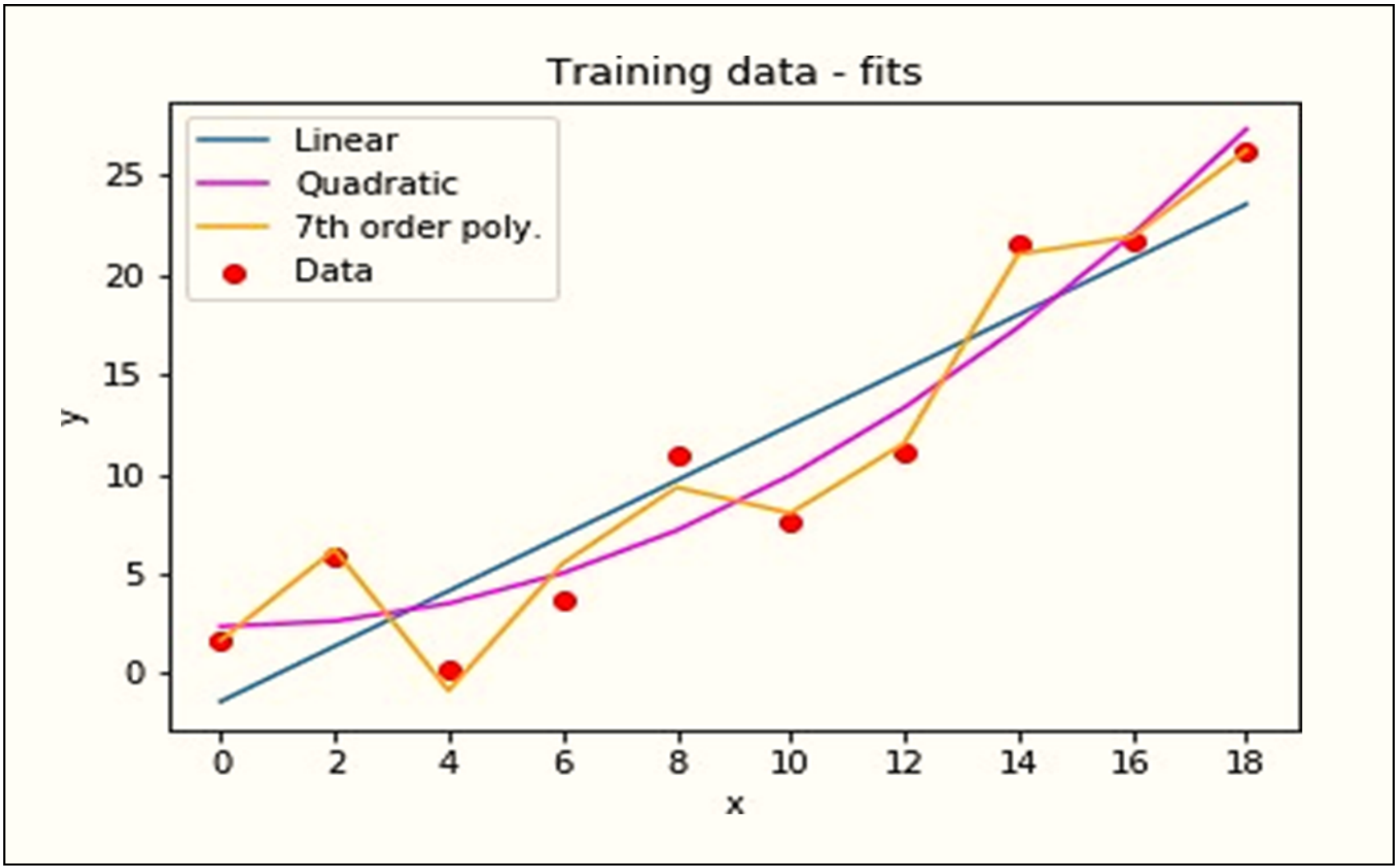


FIGURE 3.3. *A simulated data set.* Left: *The red line represents the true relationship,* $f(X) = 2 + 3X$, *which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for* $f(X)$ *based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*


Training data sample
y
x
0
5
10
15
20
25
0
2
4
6
8
10
12
14
16
18
20


Training data - fits
Linear
Quadratic
7th order poly.
Data
y
x
0
5
10
15
20
25
0
2
4
6
8
10
12
14
16
18

Model performance

- How close are predictions to observed data?
- Prediction error: $y_i - \hat{f}(x_i)$
- Popular measure of fit is the mean squared error (MSE):

  $MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

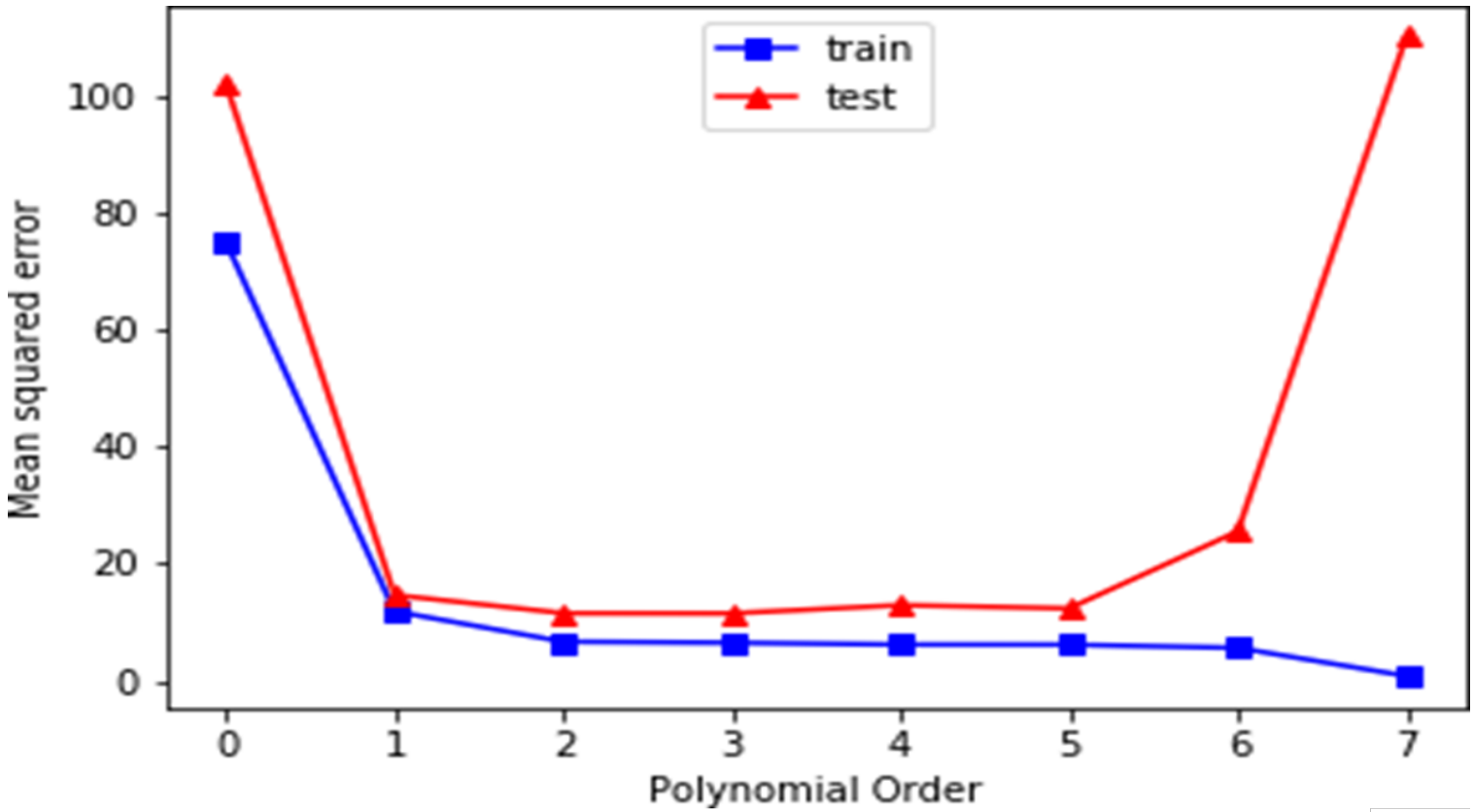
- What matters here is the test MSE. Why?
- Quality of prediction for a new observation.
- Select the model that gives the lowest test MSE.

Underfitting versus Overfitting

- Underfitting: the model is too simple, several patterns in the data cannot be modelled; for example: we fit a linear model, but the true relation is quadratic. This leads to a high Bias
- Overfitting: the model is too complex, it fits the model very closely to the training data and even uses the noise-information. This leads to high Variance

- Bias-Variance trade-off:
  - Simple models have high Bias, but low Variance (for example, take the mean of the training data as prediction)
  - Complex models have low Bias, but high Variance

- Maximising in-sample fit does not usually lead to accurate predictions out-of-sample.
- More flexible functional forms improve in-sample fit; training mean squared error (MSE) is decreasing in degrees of freedom.
- But test MSE is U-shaped (it first decreases but then increases).
- Models with low training MSE but large test MSE are overfitting the data.


train
test
Mean squared error
100
80
60
40
20
0
0
1
2
3
4
5
6
7
Polynomial Order

Bias-Variance trade-off

- The expected test MSE for a given x_0 can always be decomposed as:

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

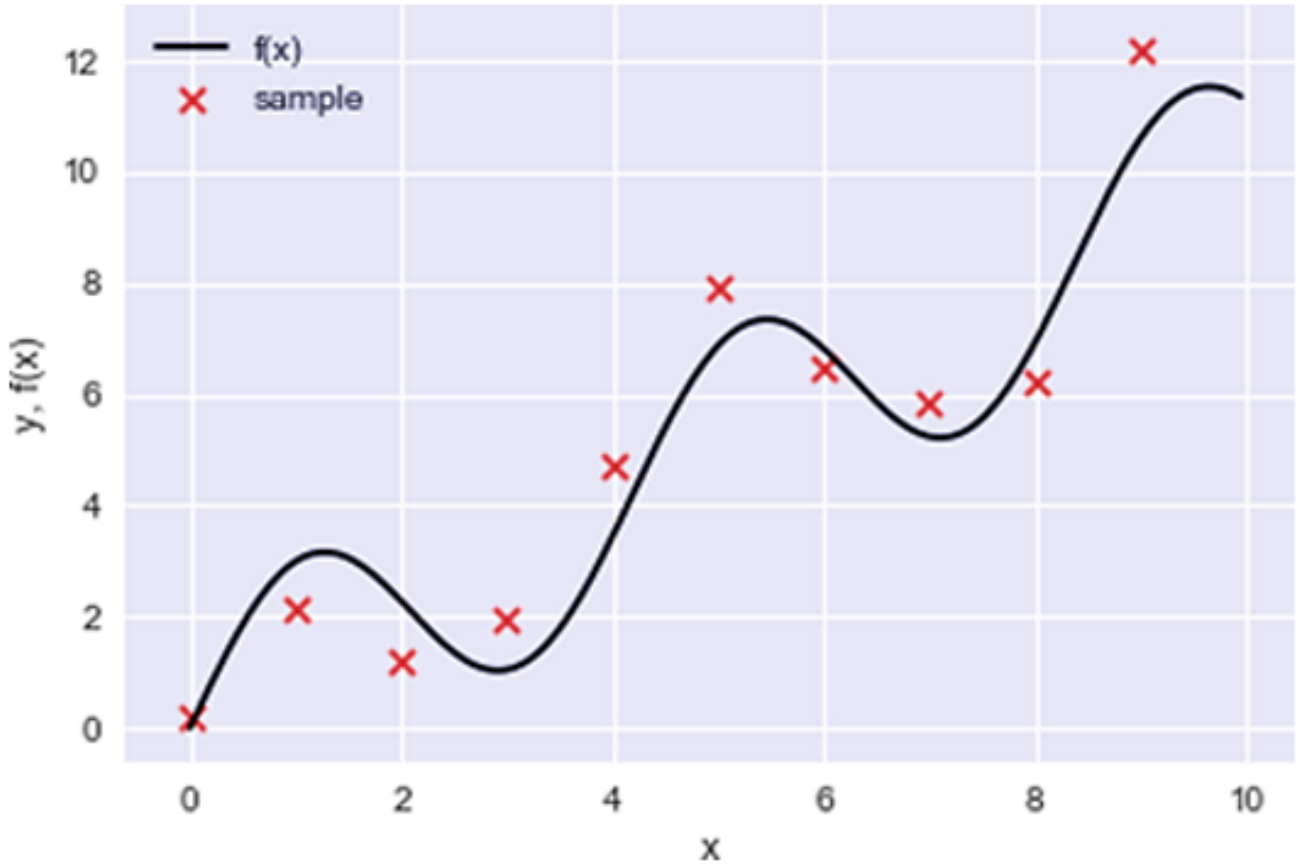
- Var($\hat{f}$): $\hat{f}$ depends on the training dataset, and changes if we use a different training dataset. More flexible models have higher variance.

- Bias($\hat{f}$): on average how close is the $\hat{f}$ to the true (unknown) f. More flexible models have lower bias.

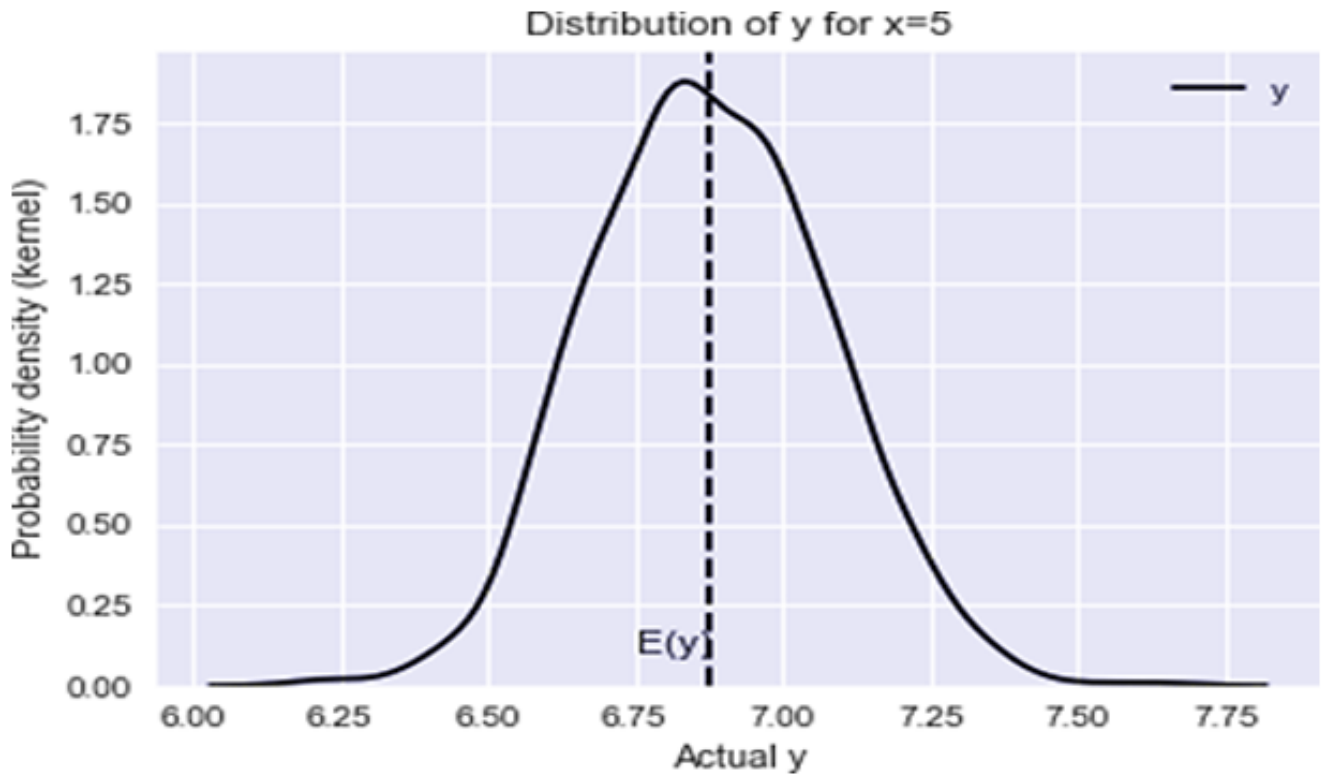
- Var(ϵ): intrinsic noise in the data. Even if we figure out the true f , we can't reduce error below this level.

- In statistics, variance decomposition is used for multivariate analysis.
- Widely used in macroeconomics, for instance when using VAR methodologies. In finance, researchers use it to understand the percentage of the variance driving changes in a particular stock.
- The idea is that, in a multivariate environment, shocks in one variable may have immediate effects over the rest of the variables; this changes can be permanent or seasonal, depends on the data.
- Impulse-Response functions (interest rate, inflation and money supply), maximum likelihood (probits and logits), or Bayesian statistics, among others.

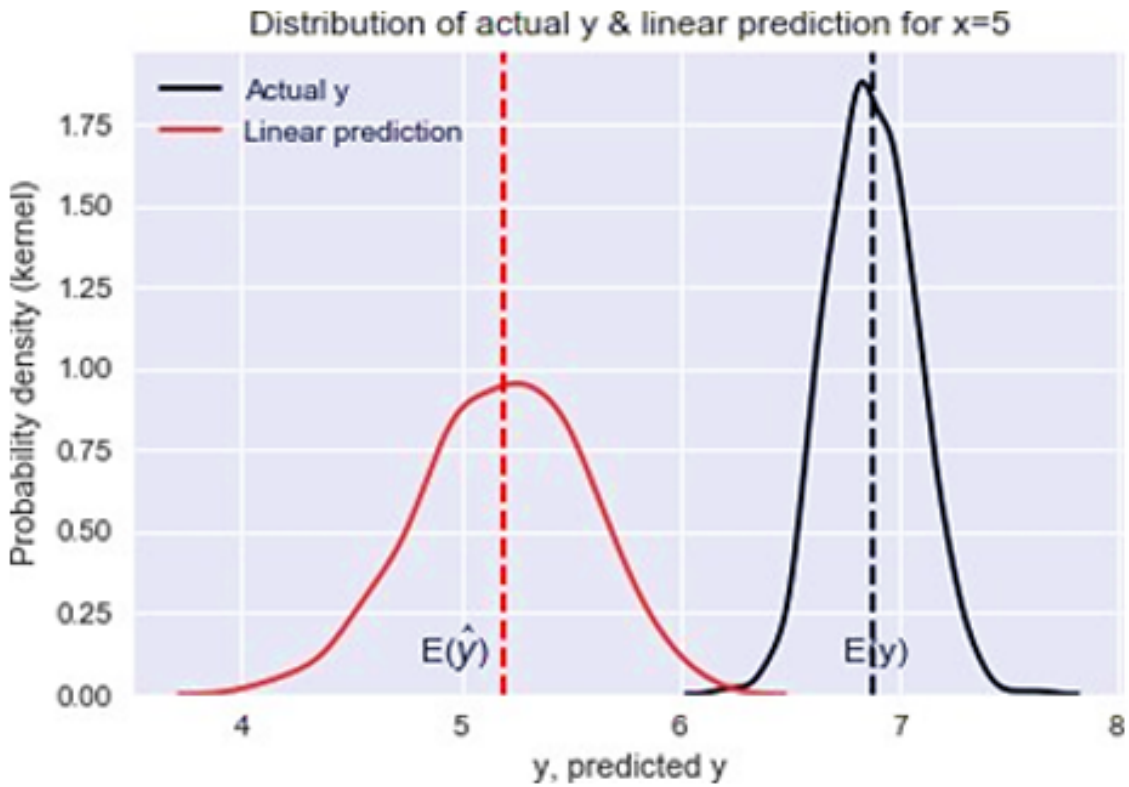
Let y=f(X) + $\epsilon = x + 2sin(1.5x) + \epsilon$ and ϵ $N(0, 0.2)$. But we only observe samples (with intrinsic noise).



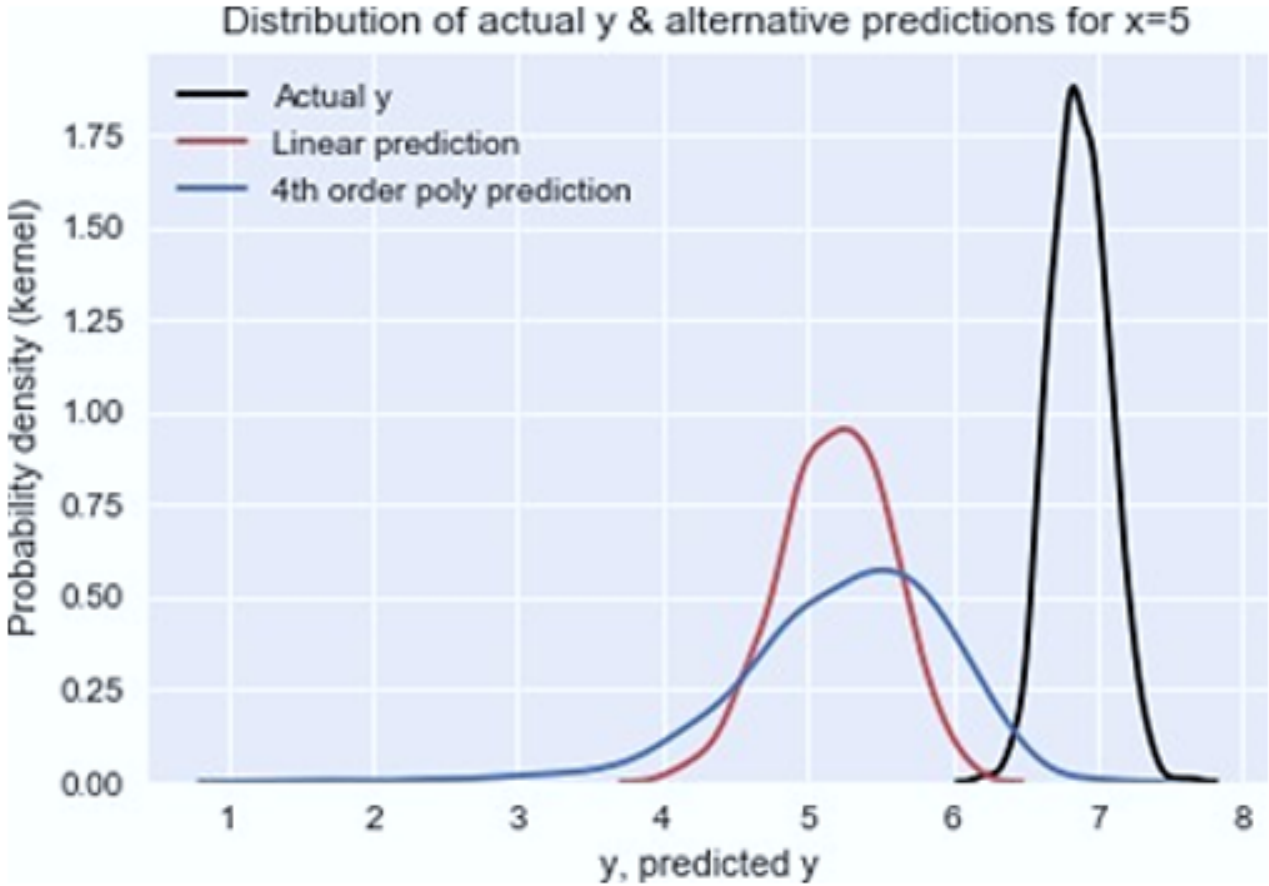
The probability distribution (pdf) of y at x = 5 shows the irreducible error we would face even if we knew the exact f (x).



The pdf of $\hat{f}$ at x = 5 from (linear fit) is the red curve.
The difference $E(\hat{y} - y)$ in the mean of the red and the black curve is the bias.
It is high because we use too simple a model.
The spread of the red curve is the variance of the prediction.



Red and blue curves are pdfs of a linear and a 4th order polynomial fit.
In general, more complex models have smaller bias but higher variance.



Bias-Variance decomposition - Proof:

Assume independent random variables, and remember:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbb{E}[(y - \hat{f}(x))^2] \qquad y = f(x) + \varepsilon \qquad \hat{y} = \hat{f}(x)$$

$$(a+b)^2 = a^2 + 2ab + b^2 \quad \text{und} \quad (a-b)^2 = a^2 - 2ab + b^2$$

We want to show that:

$$\mathbb{E}[(y - \hat{f}(x))^2] = Bias(\hat{f}(x))^2 + Var(\hat{f}(x)) + Var(\varepsilon)$$

$$\mathbb{E}[(y-\widehat{f}(x))^2] = \mathbb{E}[(\underbrace{f(x)-\widehat{f}(x)}_{a} + \underbrace{\varepsilon}_{b})^2]$$

$$= \mathbb{E}[(f(x)-\widehat{f}(x))^2] + \underbrace{\mathbb{E}[\varepsilon^2]}_{Var(\varepsilon)} + \underbrace{2\,\mathbb{E}[f(x)-\widehat{f}(x))\cdot\varepsilon]}_{\mathbb{E}(\varepsilon)=0}$$

$$\Longrightarrow 2\,\mathbb{E}[f(x)-\widehat{f}(x)]\cdot\mathbb{E}(\varepsilon)=0$$

$$= \mathbb{E}[(f(x)-\widehat{f}(x))^2] + Var(\varepsilon)$$

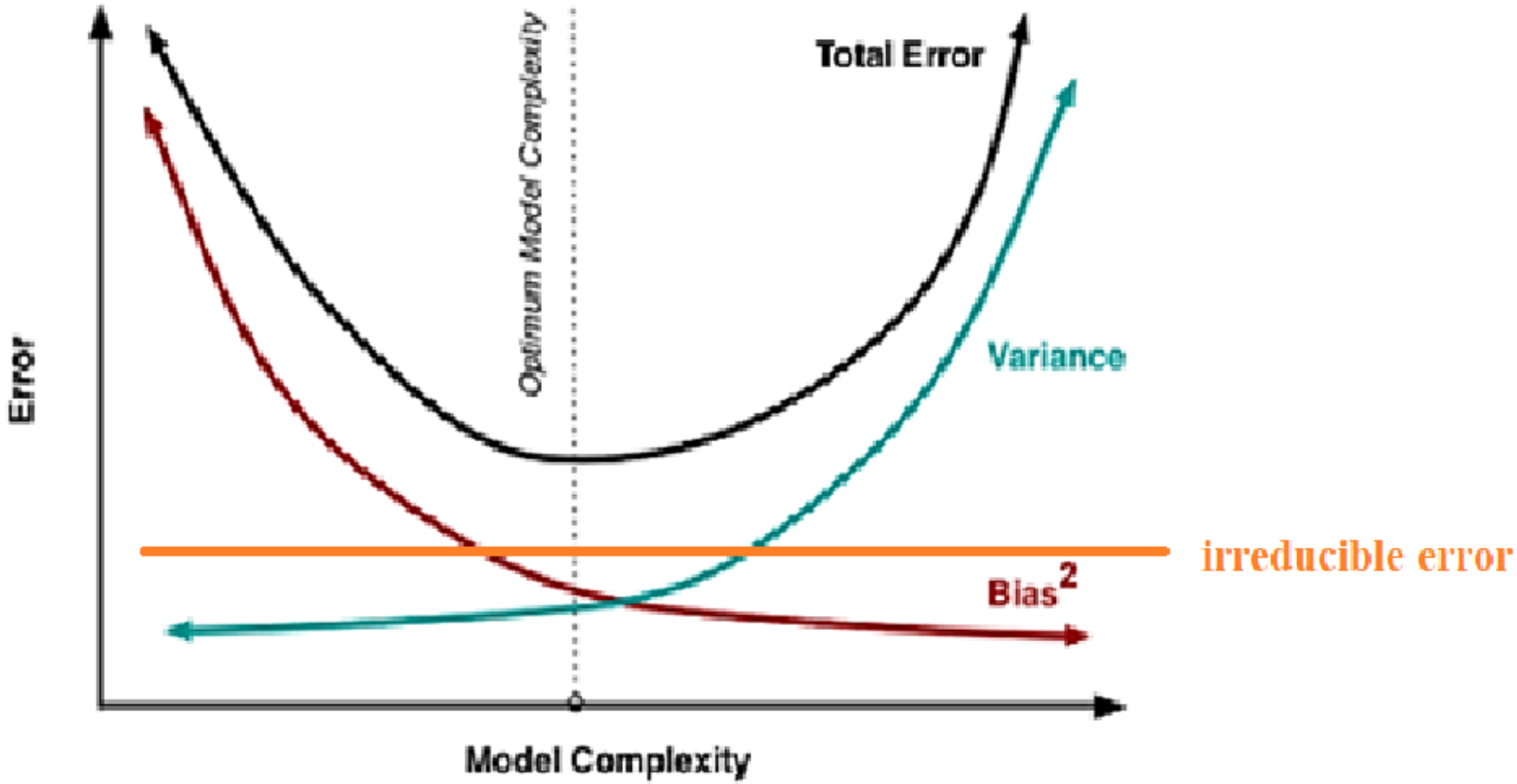
$$\mathbb{E}[(f(x)-\widehat{f}(x))^2] = \mathbb{E}\left[\left(\underbrace{(f(x)-\mathbb{E}[\widehat{f}(x)])}_{a} - \underbrace{(\widehat{f}(x)-\mathbb{E}[\widehat{f}(x)])}_{b}\right)^2\right]$$

$$= \mathbb{E}\left[\underbrace{\left(f(x)-\mathbb{E}[\widehat{f}(x)]\right)^2}_{Bias}\right] + \underbrace{\mathbb{E}\left[\left(\widehat{f}(x)-\mathbb{E}[\widehat{f}(x)]\right)^2\right]}_{Var}$$

$$-2\,\mathbb{E}\left[\left(f(x)-\mathbb{E}[\widehat{f}(x)]\right)\left(\widehat{f}(x)-\mathbb{E}[\widehat{f}(x)]\right)\right]$$

$$= Bias(\widehat{f}(x))^2 + Var(\widehat{f}(x)) - 2\left(f(x)-\mathbb{E}[\widehat{f}(x)]\right)\underbrace{\left(\mathbb{E}[\widehat{f}(x)]-\mathbb{E}[\widehat{f}(x)]\right)}_{=0}$$

$$= Bias(\widehat{f}(x))^2 + Var(\widehat{f}(x))$$


Error
Optimum Model Complexity
Total Error
Variance
irreducible error
Bias^2
Model Complexity

Bias-Variance trade-off - Summary:

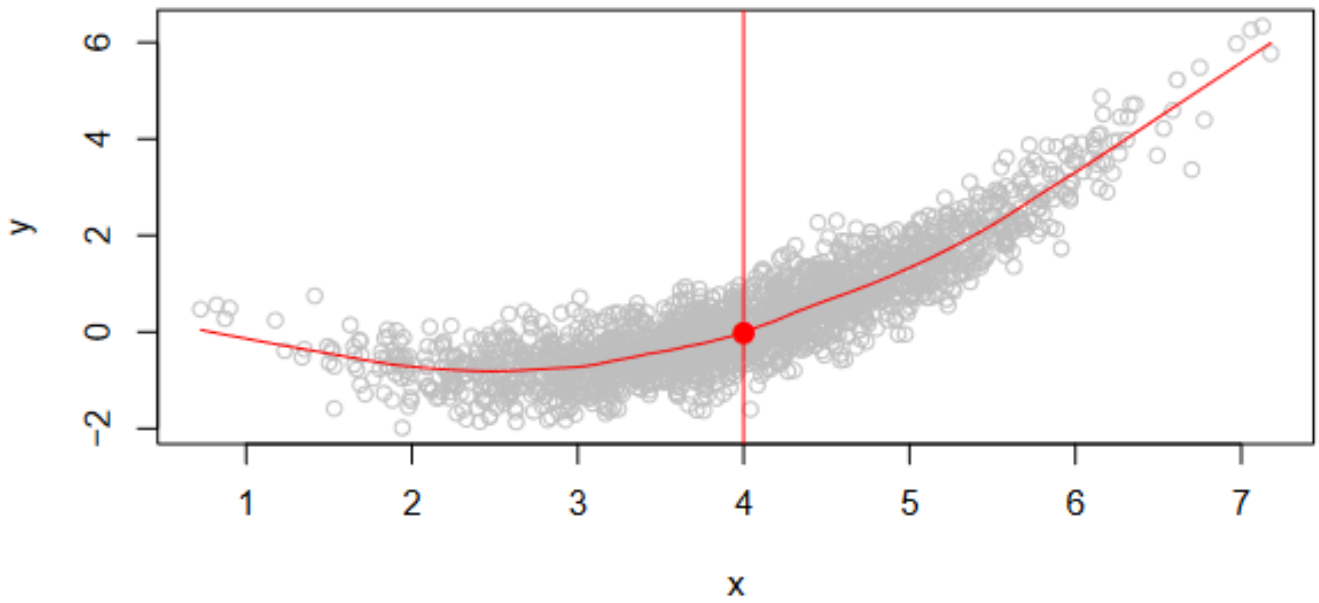
- In practice, we don't know the true data generating process i.e., f or the distribution D of x, ϵ.
- We only observe a sample of x, y .
- We split the sample into training and test data.
- Then we experiment with different prediction models.
- Simple models will tend to have low prediction variance but high bias.
- Complex models will usually be the converse.
- We try to strike a balance between bias and variance.

Regression

Modeling the relationship between Y and X

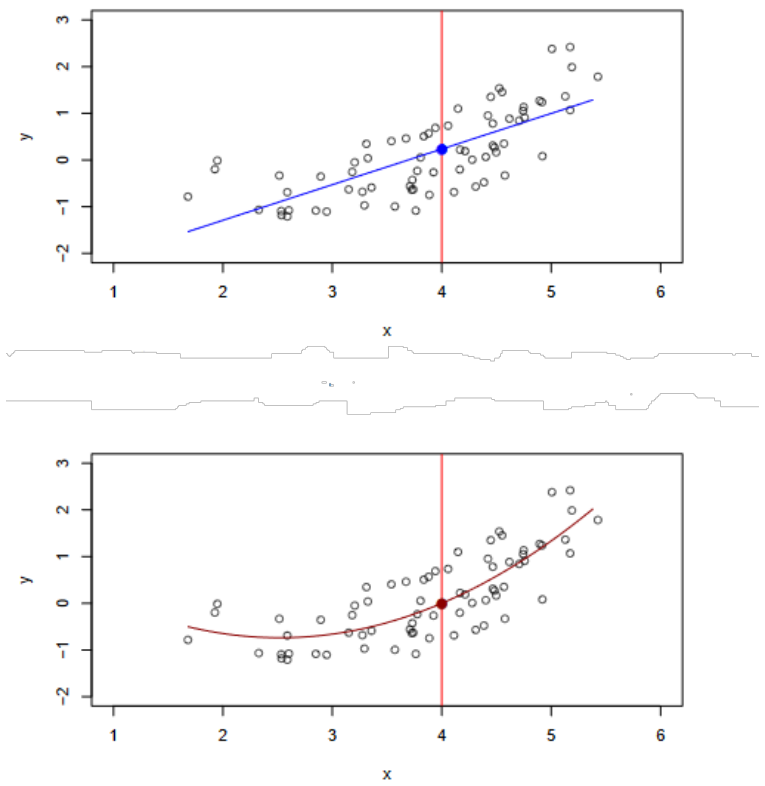
- Model $Y = f(X) + \epsilon$
- Y: income
- X: education, job experience, ...
- ϵ captures measurement errors and other discrepancies

- 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
- Depending on the complexity of f , we may be able to understand how each component X_j of X affects Y



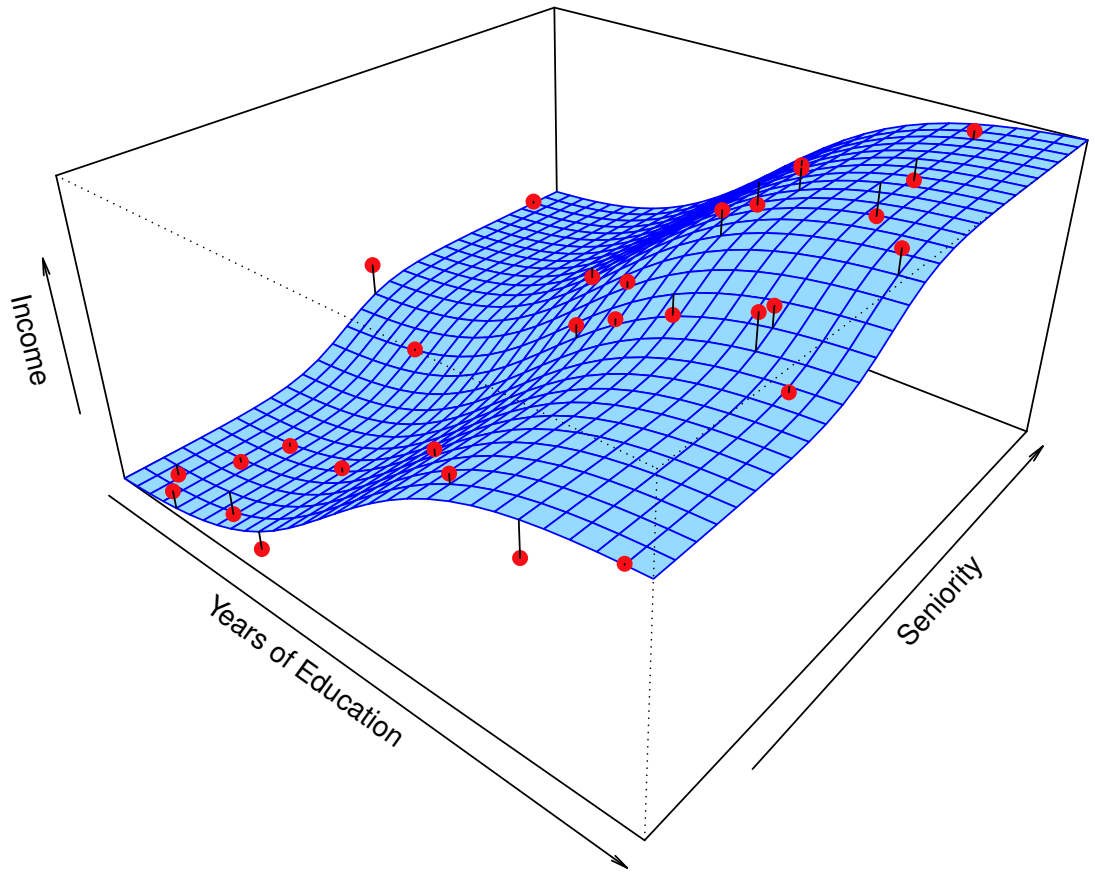
- Is there an ideal f(X)? In particular, what is a good value for f(X) at any selected value of X, say X = 4? There can be many Y values at X = 4. A good value is
  f(4) = $E(Y|X = 4)$
  $E(Y|X = 4)$ means expected value (average) of Y given X = 4. This ideal f(x) = $E(Y|X = x)$ is called the regression function

- The linear model is an important example of a parametric model: $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$
- 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 it is almost never correct, a linear model often serves as a good and interpretable approximation to the unknown true function f(X).



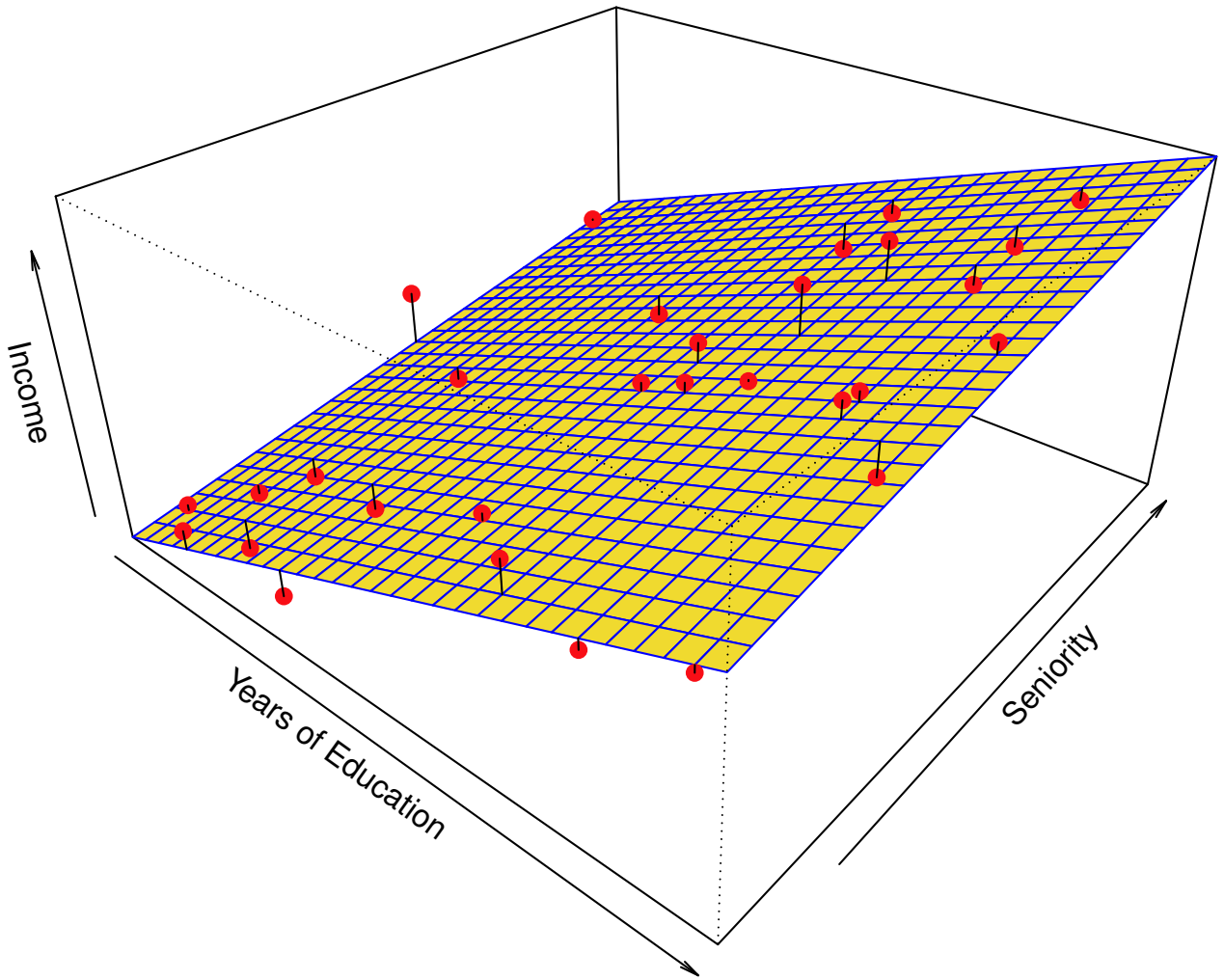
- A linear model $\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here
- A quadratic model $\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_1 X^2$ fits slightly better.

Colored in red are the data points for the model $income = f(education, jobexperience) + \epsilon$. f is the blue surface for the true relation between Y (income) and X (education; and job experience == seniority)

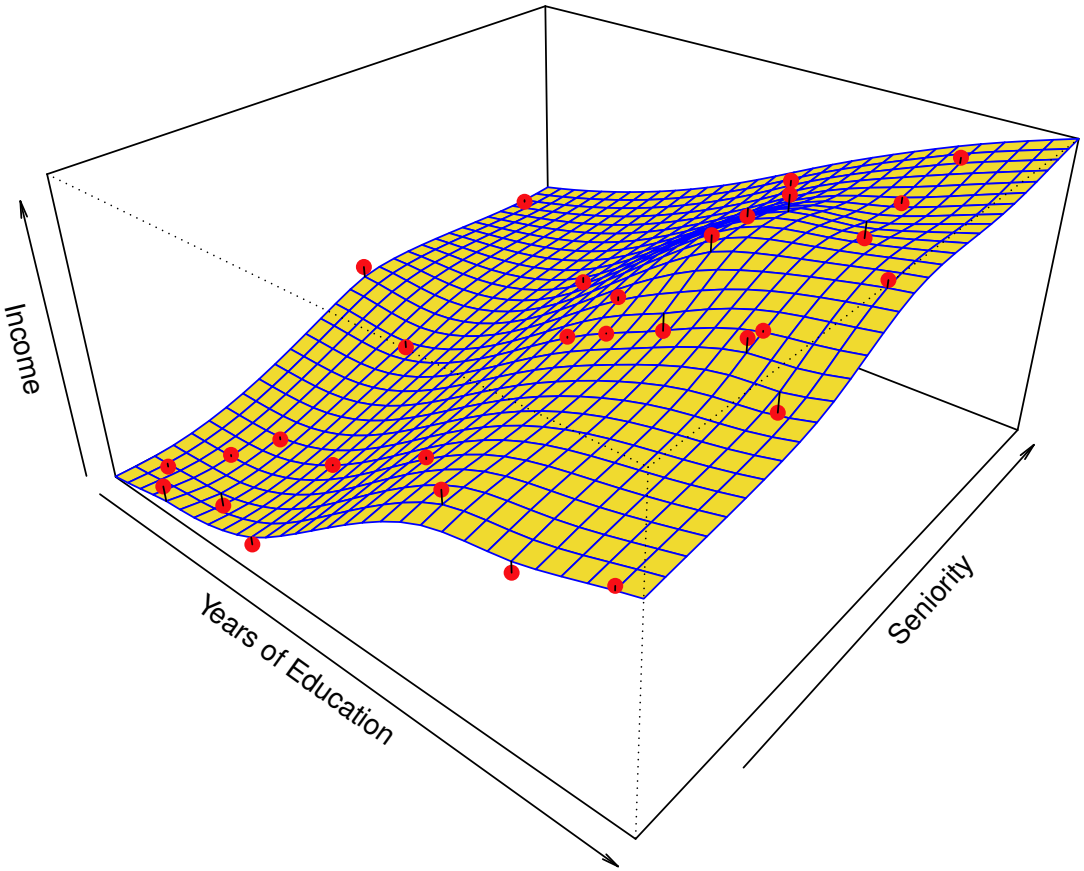


Linear regression model fit to the data.

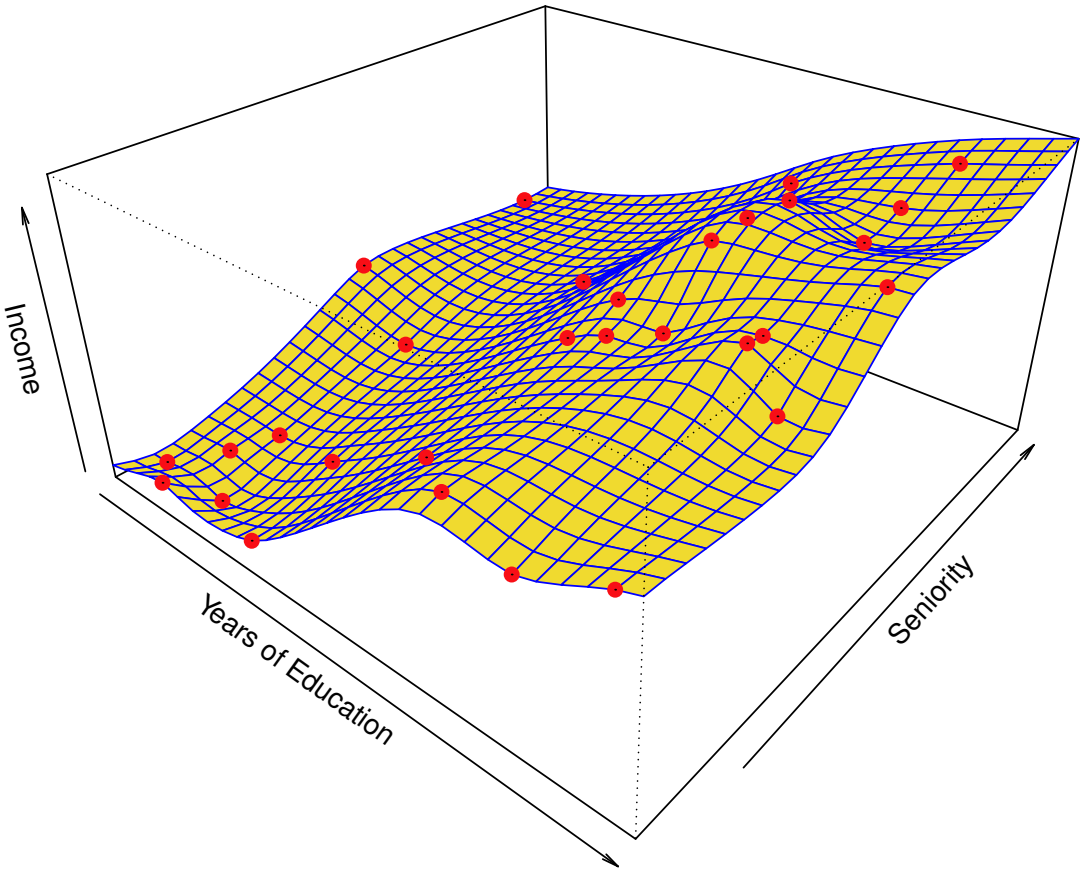
$\hat{f}(education, jobexperience) = \hat{\beta}_0 + \hat{\beta}_1 * education + \hat{\beta}_2 * jobexperience$



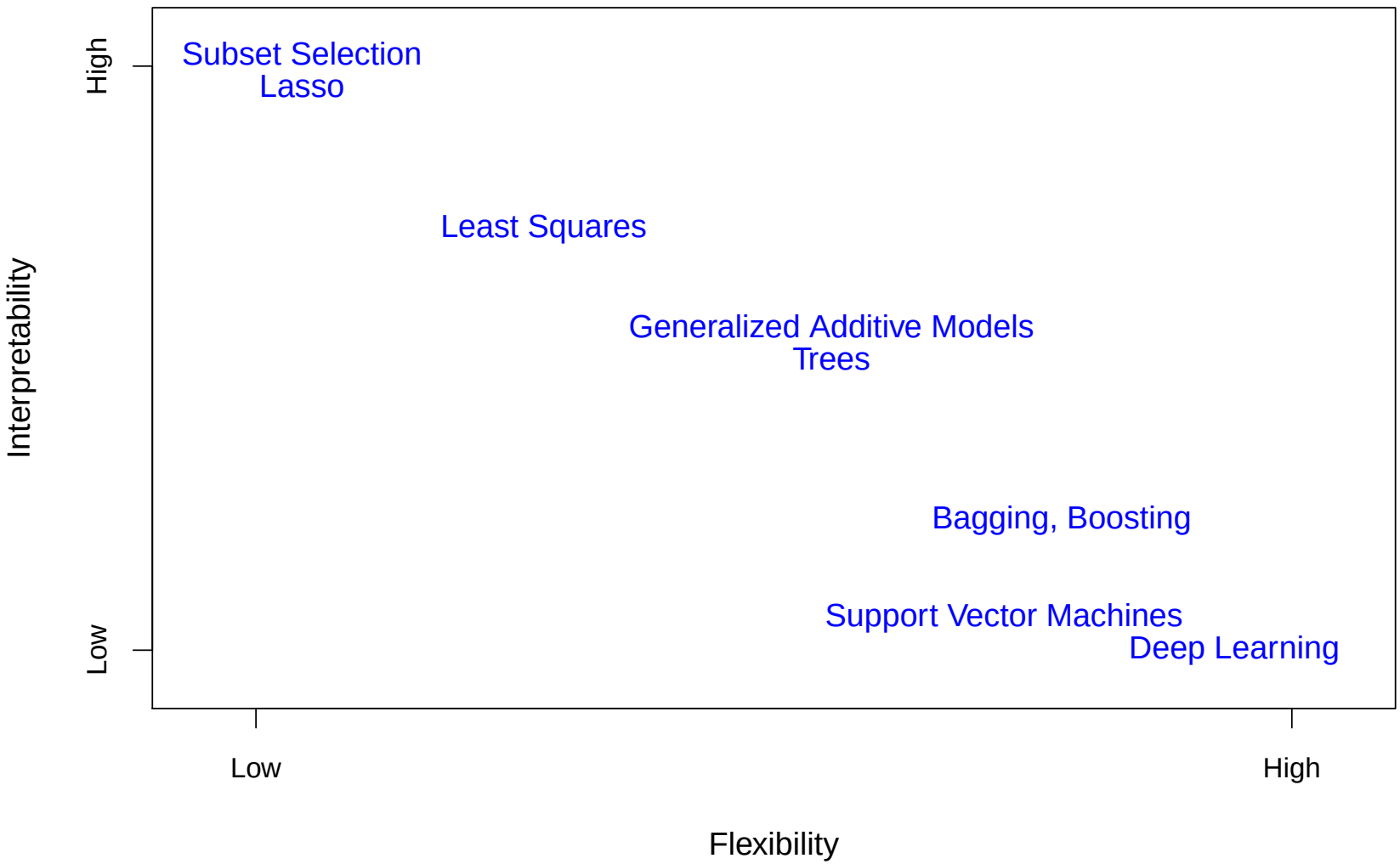
More flexible regression model $\hat{f}$ (education, jobexperience) fit to the data. Here we use a technique called a thin-plate spline to fit a flexible surface. We control the roughness of the fit



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



- 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 prediction model involving them all


Subset Selection
Lasso
Least Squares
Generalized Additive Models
Trees
Bagging, Boosting
Support Vector Machines
Deep Learning
High
Low
Interpretability
Low
High
Flexibility

Literature

Literature:

James, Witten, Hastie, Tibshirani, Taylor (2023), An Introduction to Statistical Learning, Springer, Chapter 2, pp. 15-39