Ch 2.1: What is Statistical Learning?

Lecture 2 - CMSE 381

Prof. Rongrong Wang

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

January 9, 2024

Announcements

Last time:

- Discussed where to find everything
  - Github
  - Slack
  - D2L
- Check out the syllabus!

B	C	D	E
Dates		**Content**	**Reading**
Jan 8	Mon	Intro/Python review	Sec 1
1/10	Wed	Statistical learning and model Accuracy	Sec 2.1, 2.2.1, 2.2.2
1/15	Mon	No-class Holiday	
1/17	Wed	Linear regression (I)	Sec 3.1, 3.2
1/22	Mon	Linear regression (II)	Sec 3.3
1/24	Wed	Llinear regression (III)	
1/29	Mon	Introd to classification, Bayes and KNN clas	Sec 2.2.3
1/31	Wed	Logistic Regression	Sec 4.1-4.3
2/5	Mon	Multiple Logistic Regression / Multinomial L	Sec 4.3-4.5
2/7	Wed	Leave one out CV and k-fold CV	Sec 5.1.1 - 5.1.3
2/12	Mon	k-fold CV for classification	Sec 5.1.5

Announcements:

- Get on slack!
- First homework due Wed Jan 17

To be covered in this class

- Input/output variables
- Prediction vs inference
- Reduceable vs irreduceable error
- Overfitting
- Classification vs regression
- Supervised vs Unsupervised learning
- Mean Square Error
- Variance bias trade-off

An example data set: `Advertising`

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1	4.8
10	199.8	2.6	21.2	10.6
11	66.1	5.8	24.2	8.6

- Sales of a product in 200 markets, along with amount spent on three differnt types of advertising
- Goal:
- Input variables:
- Output variable:

Data available at
https://github.com/nguyen-toan/ISLR/blob/master/dataset/Advertising.csv

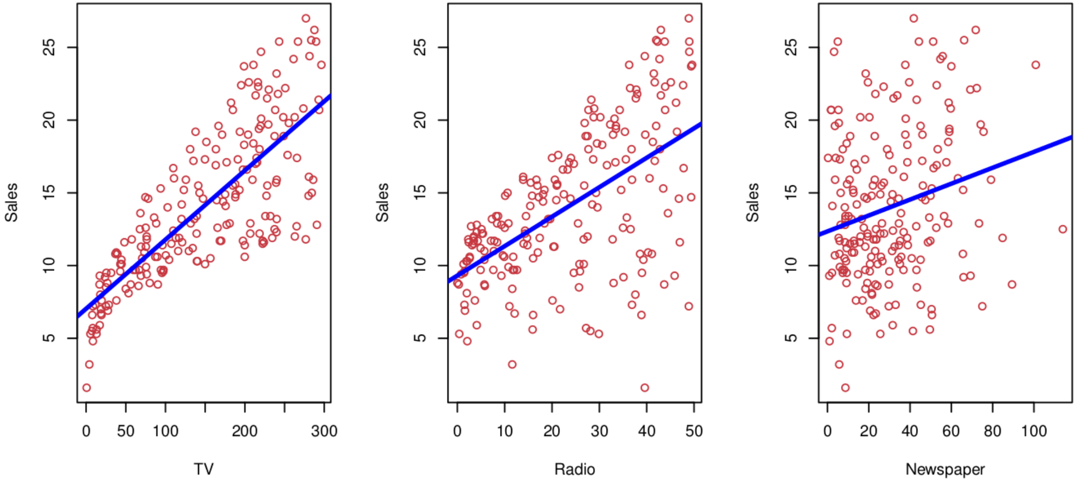
Notation and Big Assumption

Input variables: $X_1, X_2, \cdots, X_p$
Output variable: Y

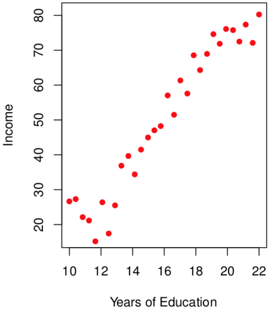
$$Y = f(X) + \varepsilon$$

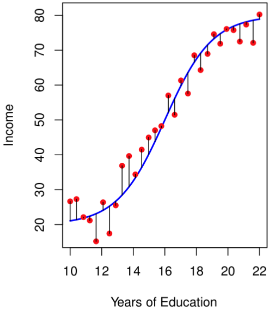
- f is the systemic information that X provides about Y.
- It is the ground truth that we want but can't access
- So our goal is to come up with an estimated model $\hat{f}$.
- ε is a random error term which is independent of X and has mean 0

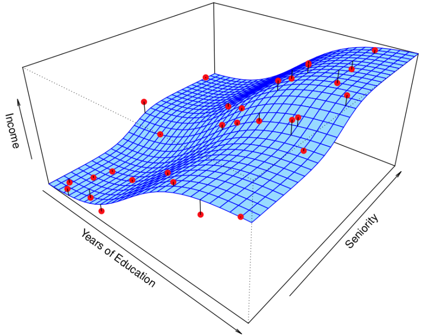
Advertising Example



More examples







Section 1

Prediction vs Inference

Prediction

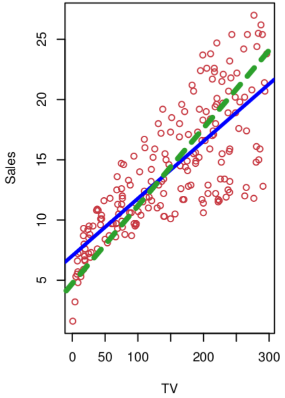
Given a value X, try to provide an estimate for $f(X)$.

Build a model:

$$\hat{Y} = \hat{f}(X)$$

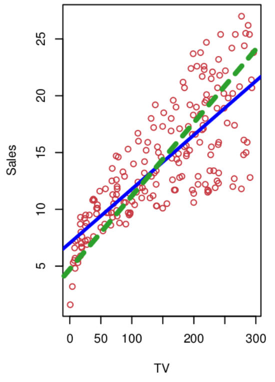
- Want to get a good guess for f, which is unknown blue
- Model is $\hat{f}$ is green dashed lines

Example: If we spend \$150 on TV advertising, what do we predict we will we make in sales?



Group question:

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3



The blue solid line is f. The green dashed line is $\hat{f}$.

- What is the predicted sales for the first three data points using the green dashed line $\hat{f}$ shown in the graph?
  - Note all values approximate
  - $\hat{f}(230.1) = 19$,
  - $\hat{f}(44.5) = 7$,
  - $\hat{f}(17.2) = 5$,
- Using the dashed green line as the predicted model $\hat{f}$, what is the error in each of the three predicions?

Reduceable vs irreducable error

All models are wrong, some are useful.

$$Y - \hat{Y}$$

Reducible Error

- $\hat{f}$ will not be a perfect estimate for f.
- We can potentially improve the irreducible accuracy of $\hat{f}$ by using the most appropriate statistical learning technique

Irreducible Error

- Model was $Y = f(X) + \varepsilon$,
- Variability of ε also affects predictions
- Not matter how well we estimate f, we can't get rid of this error.
- Would expect this in real life though:

Computing the error

- Given estimate $\hat{f}$ (fixed)
- Set of predictors X (fixed)
- Prediction $\hat{Y} = \hat{f}(X)$

$$E(Y - \hat{Y})^2 =$$

Inference

Want f, but not for prediction (or possibly combined with prediction)

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation? Is it more complicated?

Determine whether each scenario is prediction, inference, or both.

Application	**Prediction**	**Inference**
Predict effectiveness of vaccine		
Determine the address written on the image of an envelope.		
Identify risk factors for getting long covid.		
Predict stock prices.		

Section 2

How to estimate f?

Input: Training data

- n data points observed
- x_{ij} is the jth predictor for observation i
- y_i is the response variable for the ith observation
- Training data:
    - $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$
    - $x_i = (x_i, x_{i2}, \cdots, x_{ip})^T$

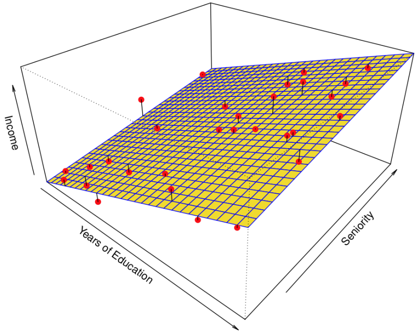
1		TV	Radio	Newspaper	Sales
2	1	230.1	37.8	69.2	22.1
3	2	44.5	39.3	45.1	10.4
4	3	17.2	45.9	69.3	9.3
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
10	9	8.6	2.1	1	4.8
11	10	199.8	2.6	21.2	10.6
12	11	66.1	5.8	24.2	8.6

Parametric methods

Step 1: Select a model

Example:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$



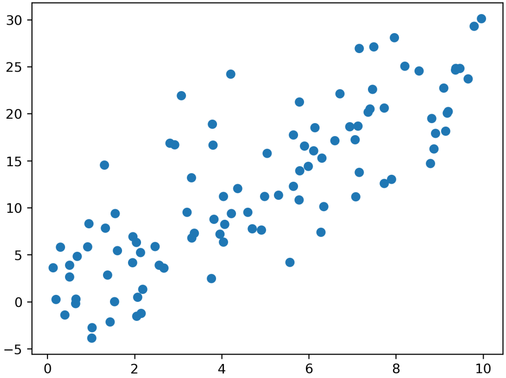
Step 2: Train the model

Example:
Find $\beta_i's$ so that

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

How do you decide on the coefficients?

$$Y \approx \beta_0 + \beta_1 X_1$$

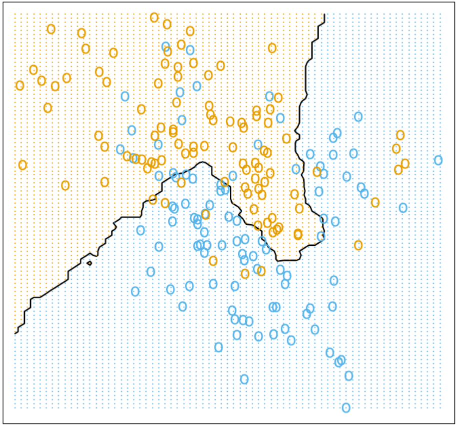


Desmos toy: https://www.desmos.com/calculator/skvt8c7317

Example Non-parametric method: Nearest Neighbors

$$N_k(x) = \text{Set of } k \text{ nearest neighbors of } x$$

$$\hat{f}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$



$k = 15$

Parametric methods: Pros and Cons

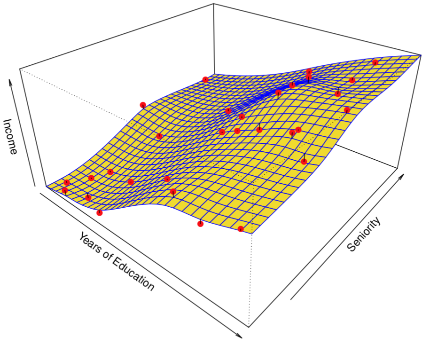
Pros

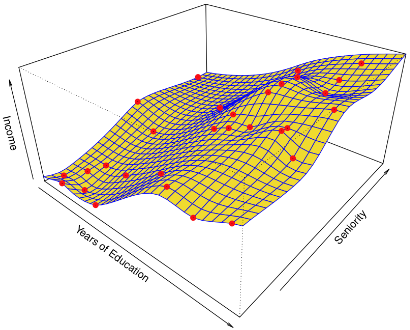
- Easier to estimate paramters than to figure out a completely arbitrary function

Cons

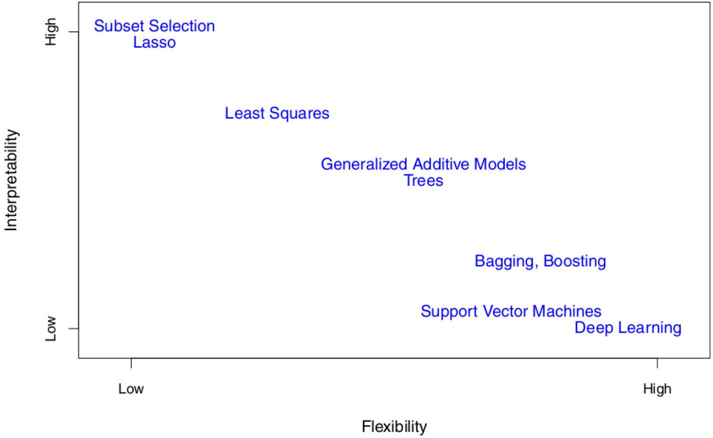
- You might have chosen the wrong function type

Overfitting



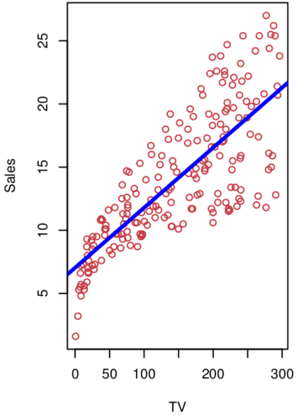


Prediction Accuracy vs Model Interpretability

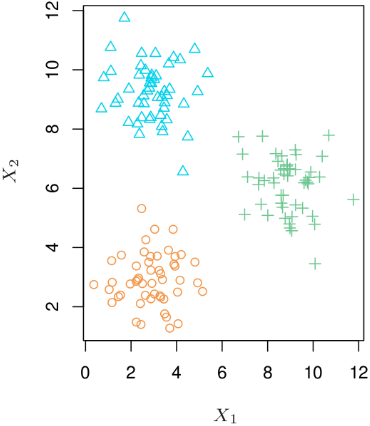


- More flexible allows for greater accuracy, but potential for overfitting
- Also more restrictive makes it easier to understand and interpret the results

Supervised learning:
Training data has response variable y for every input x



Unsupervised Learning:
Training data has response variable y for every input x



Regression vs Classification

Types of variables:

- Quantitative

- Qualitative / Categorical

Section 3

Group work

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

- Is this classification or regression?
- Do we want inference or prediction?
- What is n, the number of data points?
- What is p, the number of variables?

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

- Is this classification or regression?
- Do we want inference or prediction?
- What is n, the number of data points?
- What is p, the number of variables?

Quick review of notation

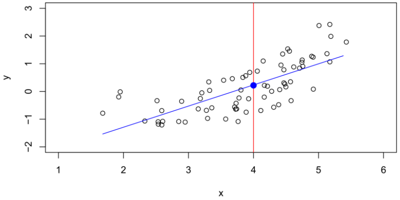
- $X = (X_1, \cdots, X_p)$ number of variables
- Ground truth $Y = f(X) + \varepsilon$
- Approximation $\hat{Y} = \hat{f}(X)$
- Number of data points n
- X_{ij} is *jth* predictor for observation i

Section 4

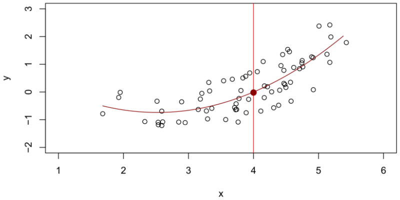
Mean Squared Error

Which is better?

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



A quadratic model $\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ fits slightly better.



No free lunch

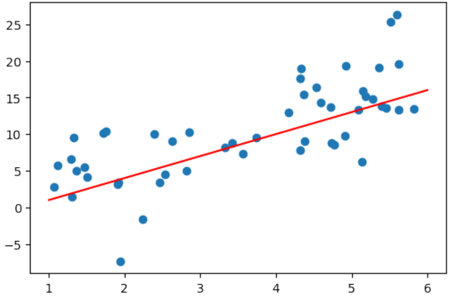
No best model for all data sets

So, need to measure quality of a model on a given data set

Mean Squared Error-Measuring the Quality of Fit

Error in the regression setting

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



Group Work

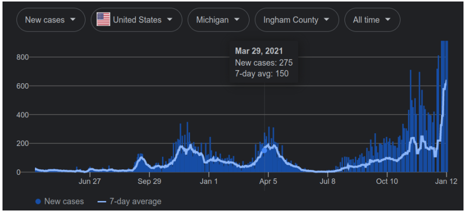
Given the following data, you decide to use the model

$$\hat{f}(X_1, X_2) = 1 - 3X_1 + 2X_2.$$

What is the MSE?

X_1	X_2	Y
0	7	14
1	-3	-6
5	2	-10
-1	1	7

Training MSE



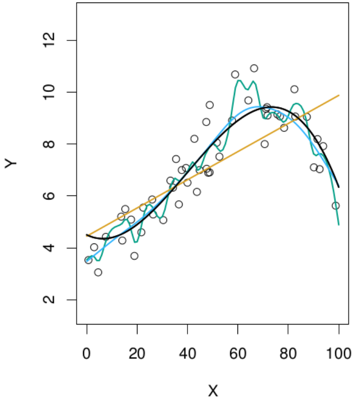
Train vs test

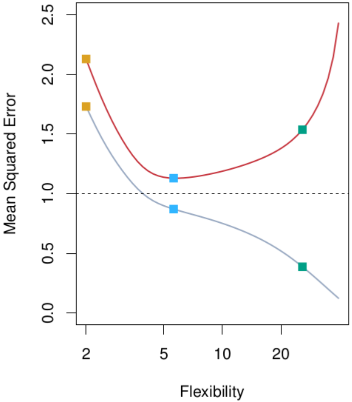
Training set:
The observations
$\{(x_1, y_1), \cdots, (x_n, y_n)\}$ used to get
the estimate $\hat{f}$

Test set:
The observations
$\{(x'_1, y'_1), \cdots, (x'_{n'}, y'_{n'})\}$ used test the
model. We care the average squared
test error more

$$\frac{1}{n'} \sum_i (y'_i - \hat{f}(x'_i))^2$$

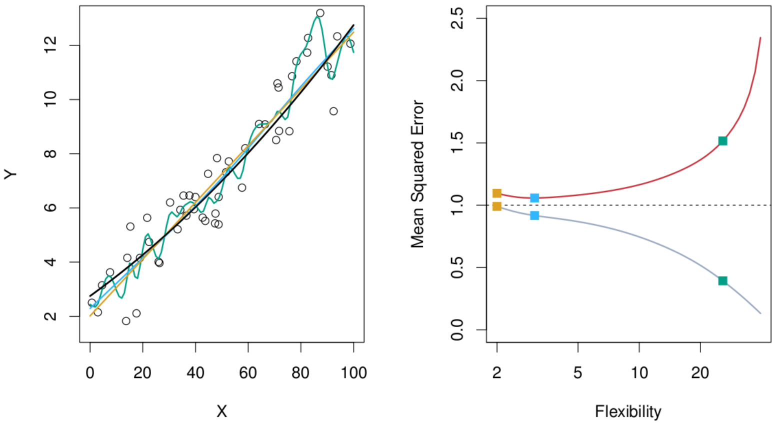
Why not just get the best model for the training data?





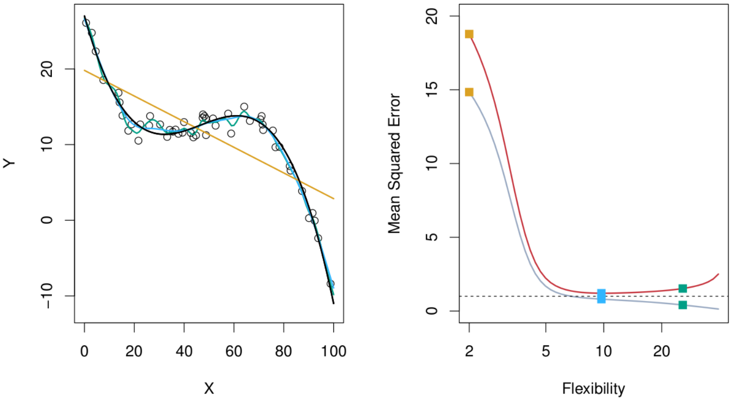
- Left: Black line is model for simulation of data
- Right side: training in blue/grey, testing in red
- Variance of ε is dashed line
- Point out that training error goes down but test goes up (overfitting)
- Flexibility = degrees of freedom

A more linear example



- Truth is linear, so test MSE down only a bit before going up

A more non-linear example



- Similar structure to previous but needs more degrees of freedom before getting a good test

Section 5

Bias-Variance Trade-Off

Bias-variance

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

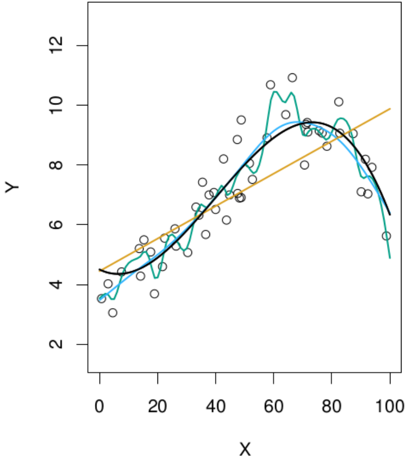
- Proof are not in the textbook
- $E(y_0 - \hat{f}(x_0))^2$ is expected test MSE at x_0; avg test MSE if repeatedly estimated f with lots of training sets and tested each at x_0
- Computed by averaging $E(y_0 - \hat{f}(x_0))^2$ over all values in the test set

- Eqn says we need an $\hat{f}$ with both low variance and low bias
- Also says error is bounded below by irreducible error $\mathrm{Var}(\varepsilon)$

Variance

Variance: the amount by which $\hat{f}$ would change if we estimated it using a different training data set.

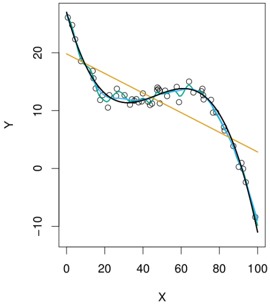
- High variance: small changes in training data result in large changes in $\hat{f}$.
- Example right: green curve more flexible, but also small changes in data set could cause large changes in the computed model.

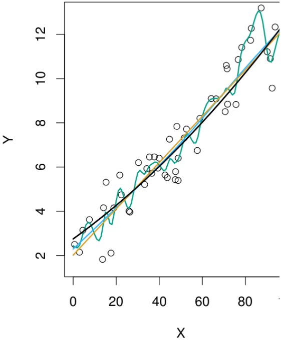


Bias

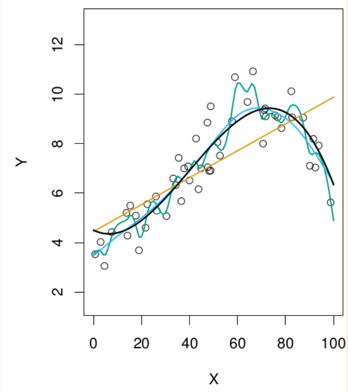
Bias: the error that is introduced by approximating a (complicated) real-life problem by a much simpler model.

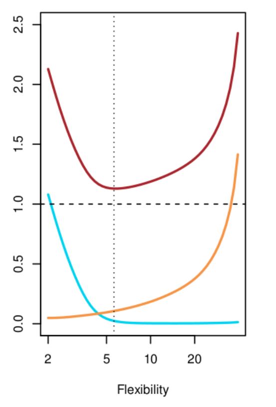
- Example: Linear regression, likely too simple for any real-life problem.
- Figure at left: True f is non-linear, so no good estimate possible with linear regression. (linear regression = high bias)
- Figure at right: True f is linear, so linear regression should do a good job (linear regression = low bias)





Group work



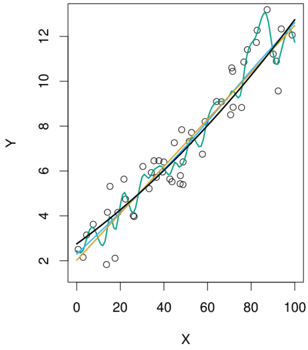


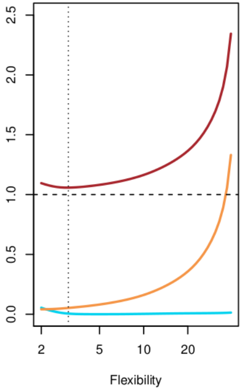
Label the line corresponding to each of the following:

- MSE
- Bias
- Variance of $\hat{f}(x_0)$
- Variance of ε

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

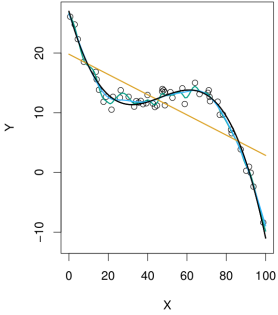
Another example

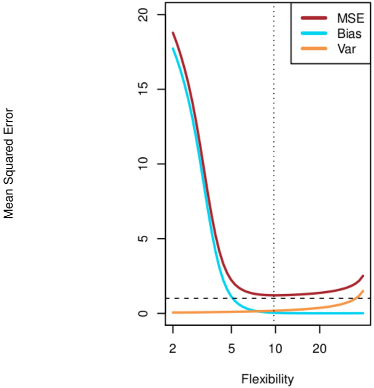




$$E(y_0 - \hat{f}(x_0))^2 = \mathrm{Var}(\hat{f}(x_0)) + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + \mathrm{Var}(\varepsilon)$$

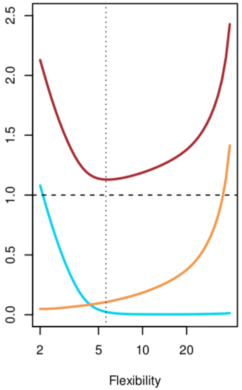
Yet another example





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

Bias-variance trade off



- More flexible: variance up, bias down.
- Relative rate of change between bias and variance determines whether MSE goes up or down
- Initially, bias goes down faster, so MSE declines
- Eventually little effect on bias but lots of increased variance, so MSE increases

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

Group work: coding

See jupyter notebook

Wrap up

Next time:

- Friday:
  - Bring Laptop!
  - There will be no quiz this week
- Monday:
  - No class!
- Wednesday: Linear regression

Announcements:

- Get on slack!
- Office hours!
  - Dr. Wang: Weds 5-7pm
  - Maryclare: Tues and Thur 12:30 - 2:30pm