Machine Learning for Business Analytics

Lecture 02

Recap of lecture 01

- What is machine learning?
  - A program that improves at a given task with increased experience
- Different types of learning
  - Supervised vs unsupervised
  - Regression vs classification
  - Prediction vs inference
- Tradeoff between model flexibility and model interpretability
- Linear regression
  - $\mathrm{y} = \mathrm{f}(\mathrm{x}) = \beta x + e$
  - Residuals $\mathrm{r}_i = \widehat{y_i} - y_i$
  - Mean Squared Error: average of squared residuals
  - Linear regression minimizes MSE in the training sample
- Train-test paradigm: we want low MSE out of sample
  - To achieve this split the sample in two parts (train and test) and use one part for training and the other for evaluation

Today, lecture 02

- Overfitting and underfitting
- The bias-variance tradeoff in machine learning
  - What is bias?
  - What is variance?
- How do bias and variance relate MSE?
- Exploring the bias-variance tradeoff in practice
- Introduction to model selection

Overfitting and underfitting

- Suppose we do the following
  1. Take a dataset and split it randomly in train and test

		Y	X1	X2	X3	X4
Training set	1					
	2					
	3					
Test set	4					
	5					

  2. A linear regression model whose coefficients I have picked in random – in other words I have not used the training set to fit my model
     - The class of models I am interested in is:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

     - I set the betas randomly, for example: $\beta_1 = .3, \beta_2 = -.1, \beta_3 = .9, \beta_4 = -.2$
  3. Using the random model, compute MSE_{train} and MSE_{test}

- On average, if I repeat the process over many random models, what do you expect the relationship between MSE_{train} and MSE_{test} to be?

Overfitting and underfitting

- Suppose we do the following
  1. Take a dataset and split it randomly in train and test

		Y	X1	X2	X3	X4
Training set	1					
	2					
	3					
Test set	4					
	5					

  2. A linear regression model whose coefficients I have picked in random – in other words I have not used the training set to fit my model
     - The class of models I am interested in is:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

     - I set the betas randomly, for example: $\beta_1 = .3, \beta_2 = -.1, \beta_3 = .9, \beta_4 = -.2$
  3. Using the random model, compute MSE_{train} and MSE_{test}

- On average, if I repeat the process over many random models, what do you expect the relationship between MSE_{train} and MSE_{test} to be? Equal!

Overfitting and underfitting

- Instead of choosing betas randomly, I could select them instead using linear regression

- Linear regression works by minimizing MSE in the *train* set

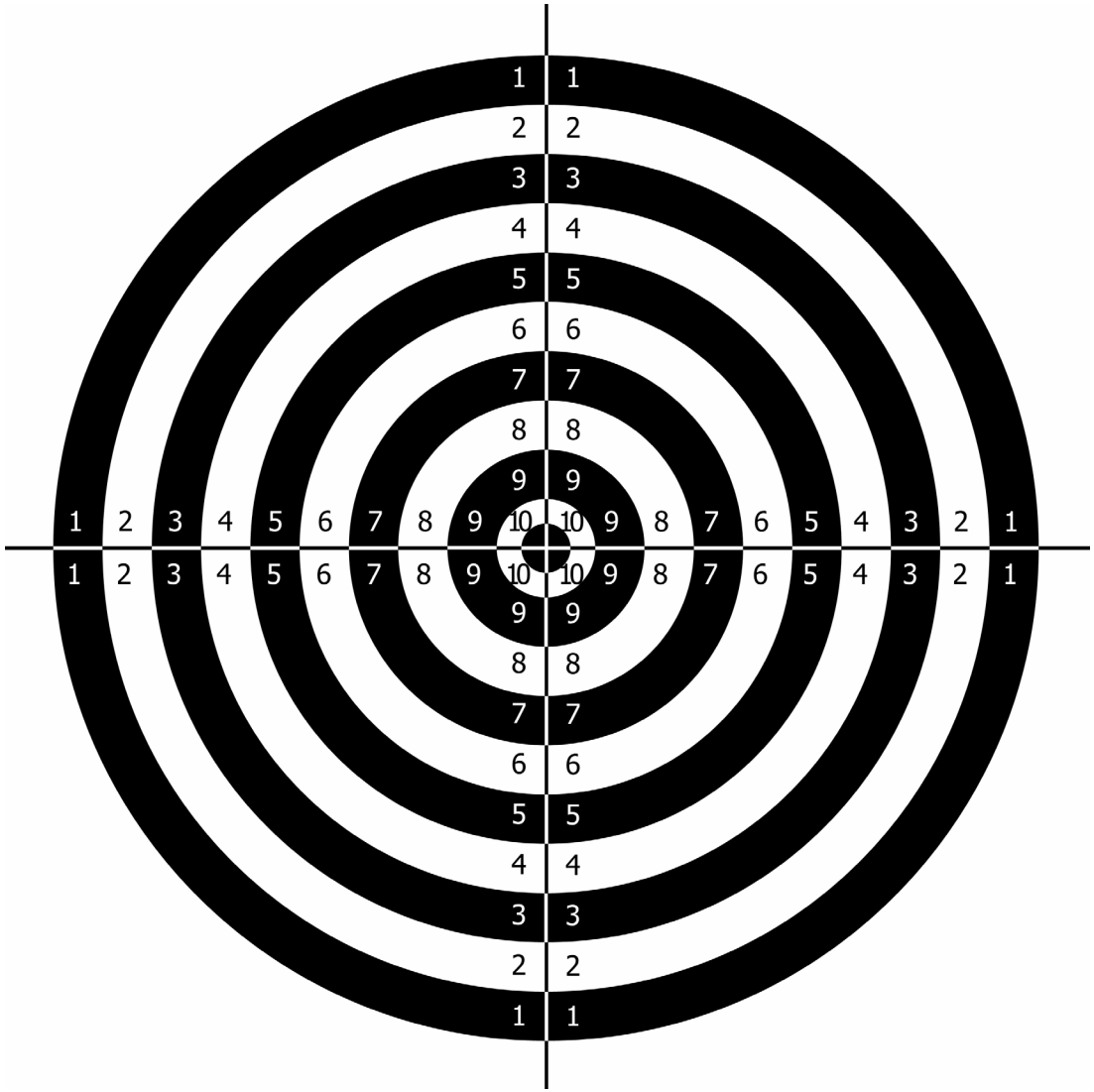
- In this case, what relationship do we expect between MSE_{train} and MSE_{test}?

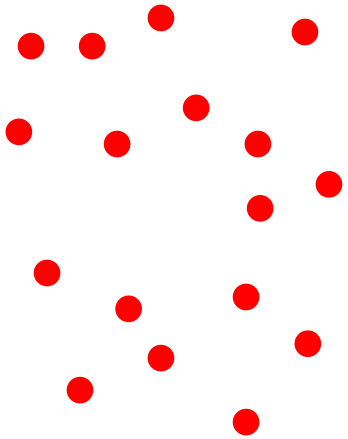
Overfitting and underfitting

- Instead of choosing betas randomly, I could select them instead using linear regression
- Linear regression works by minimizing MSE in the *train* set
- In this case, what relationship do we expect between MSE_{train} and MSE_{test}?
- We expect $MSE_{train} < MSE_{test}$
  - Why is that?
- **Underfitting**: both MSE_{train} and MSE_{test} are large
- **Overfitting**: MSE_{train} is small and MSE_{test} is large
- How do we close the gap between MSE_{train} and MSE_{test}?

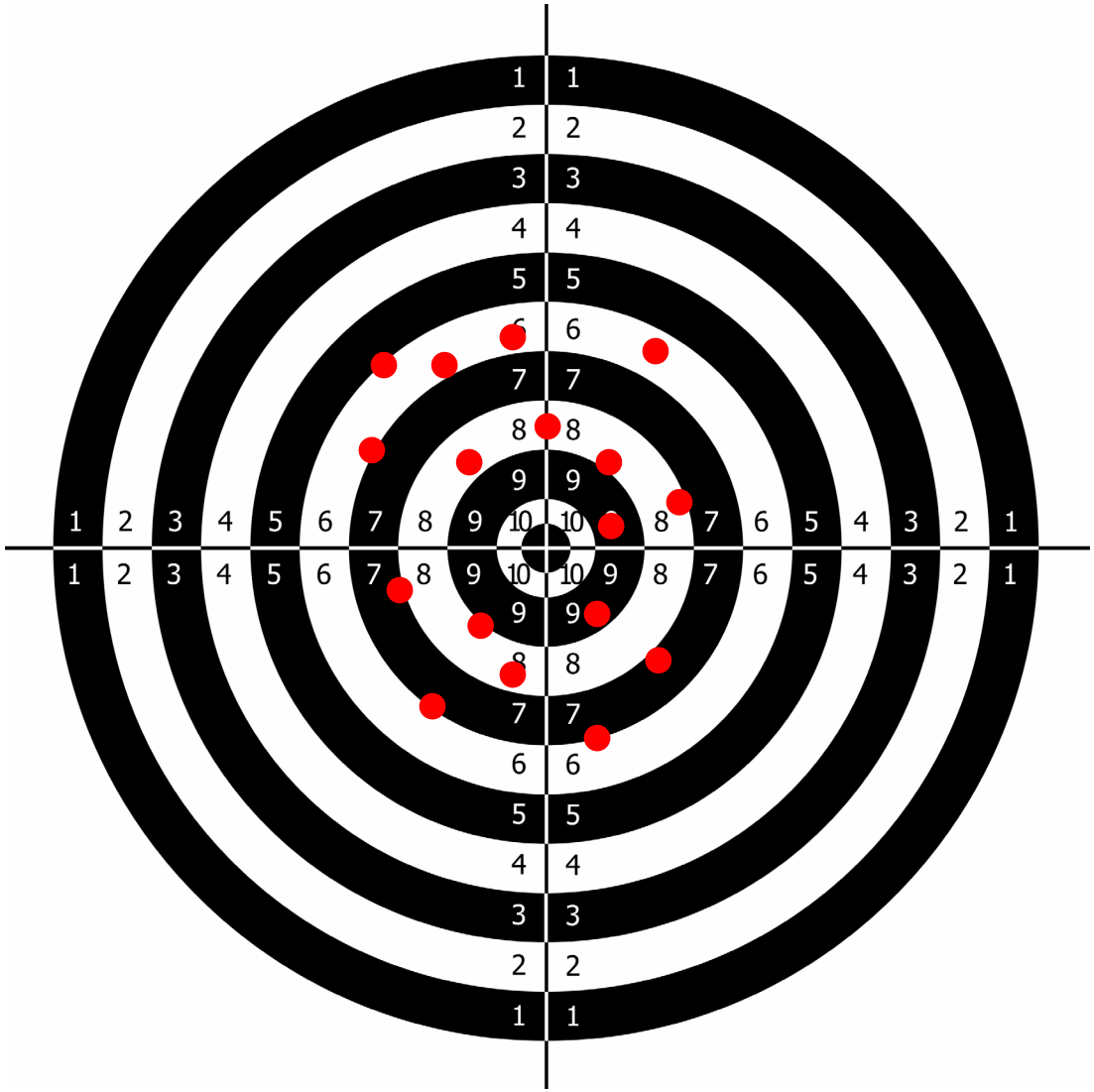
The bias-variance tradeoff

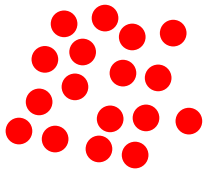
- Reducing MSE on your training data does not necessarily reduce MSE on data you have not trained on
- In other words, MSE_{Test} could be going up as MSE_{Train} is going down!
- This is a fundamental trade-off in machine learning, called the bias-variance trade-off
- What is bias? What is variance? Let's get some intuition…


1 1
2 2
3 3
4 4
5 5
6 6
7 7
8 8
9 9
1 2 3 4 5 6 7 8 9 10 10 9 8 7 6 5 4 3 2 1
1 2 3 4 5 6 7 8 9 10 10 9 8 7 6 5 4 3 2 1
9 9
8 8
7 7
6 6
5 5
4 4
3 3
2 2
1 1

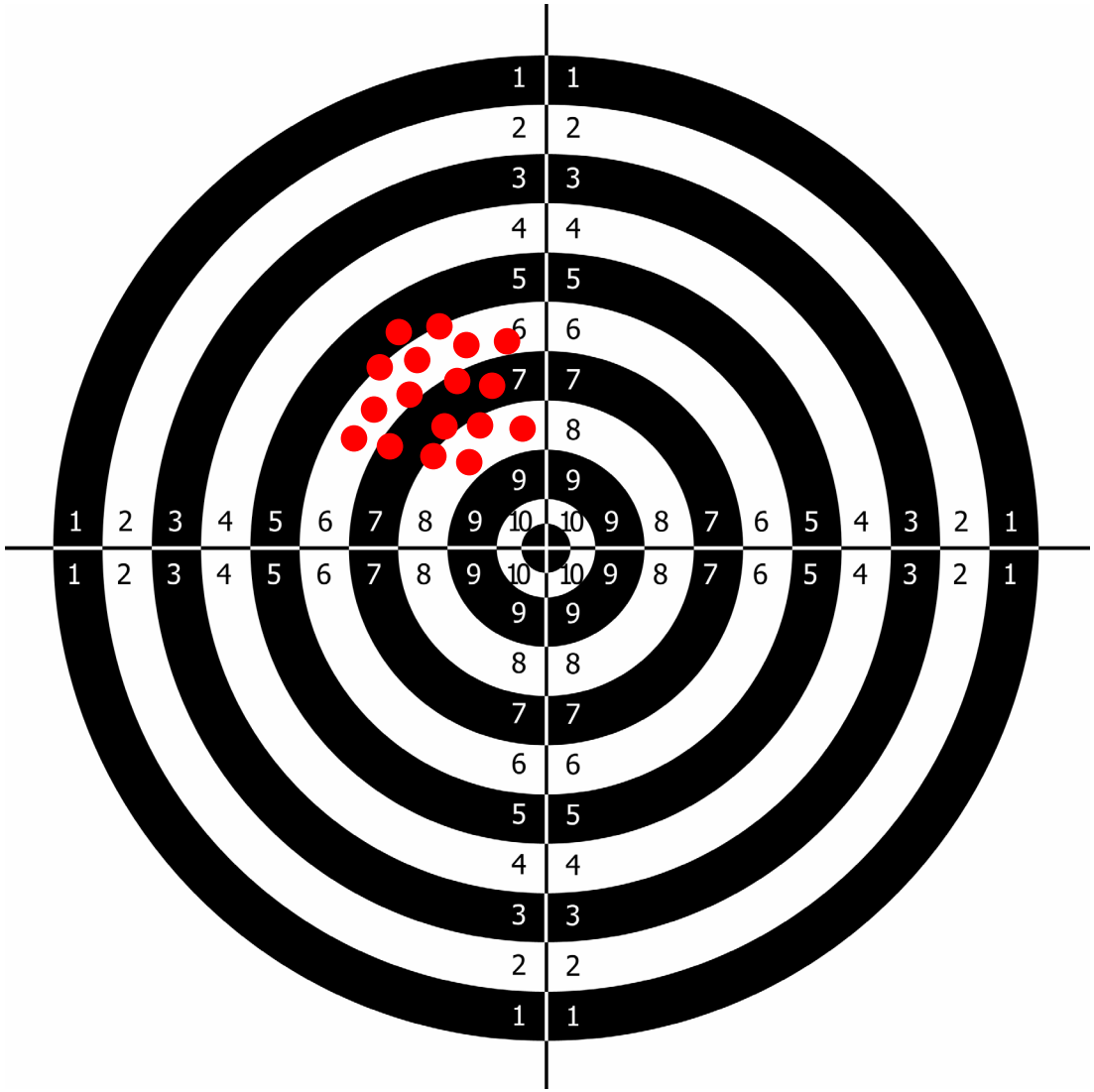


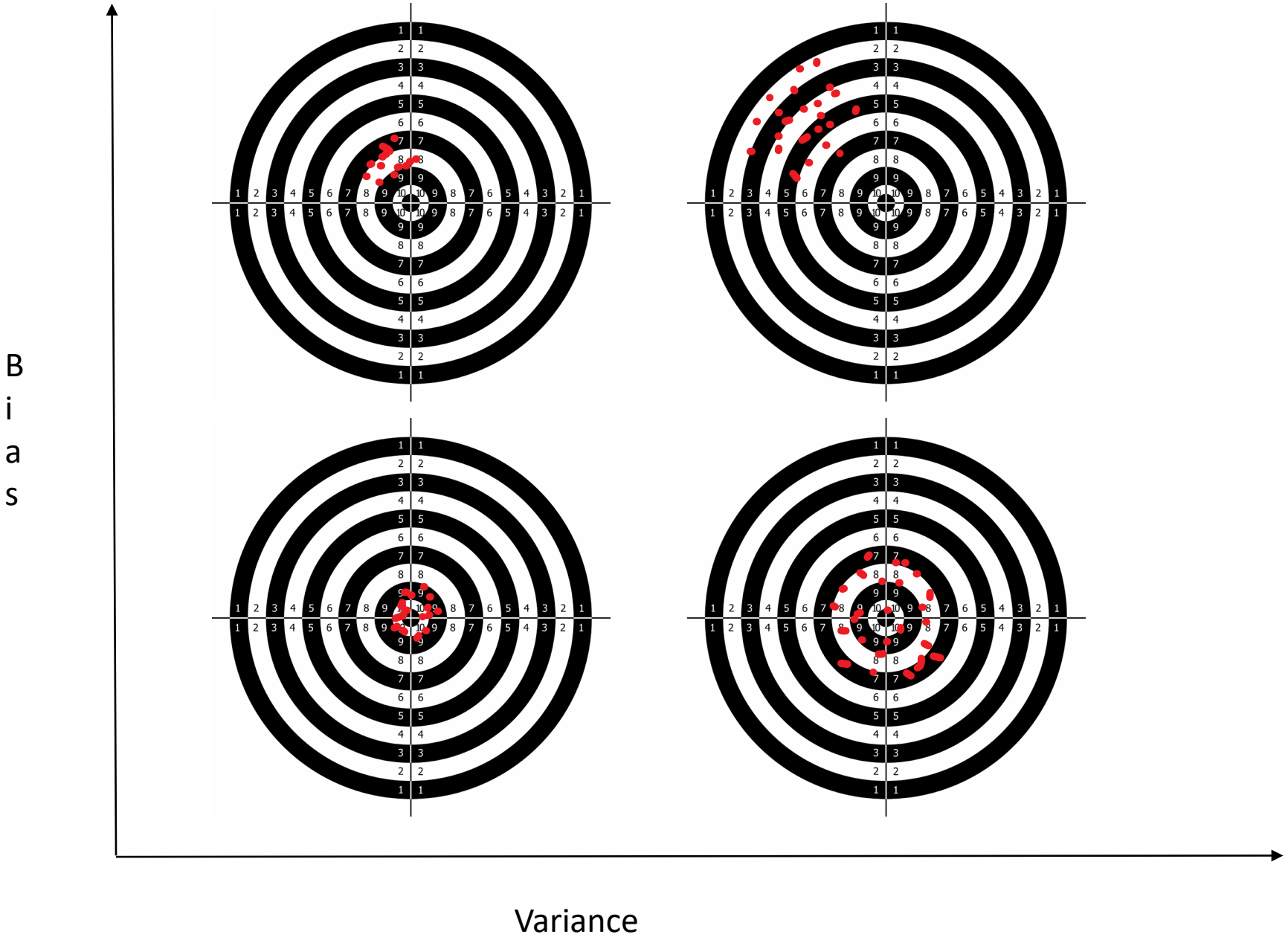
High variance

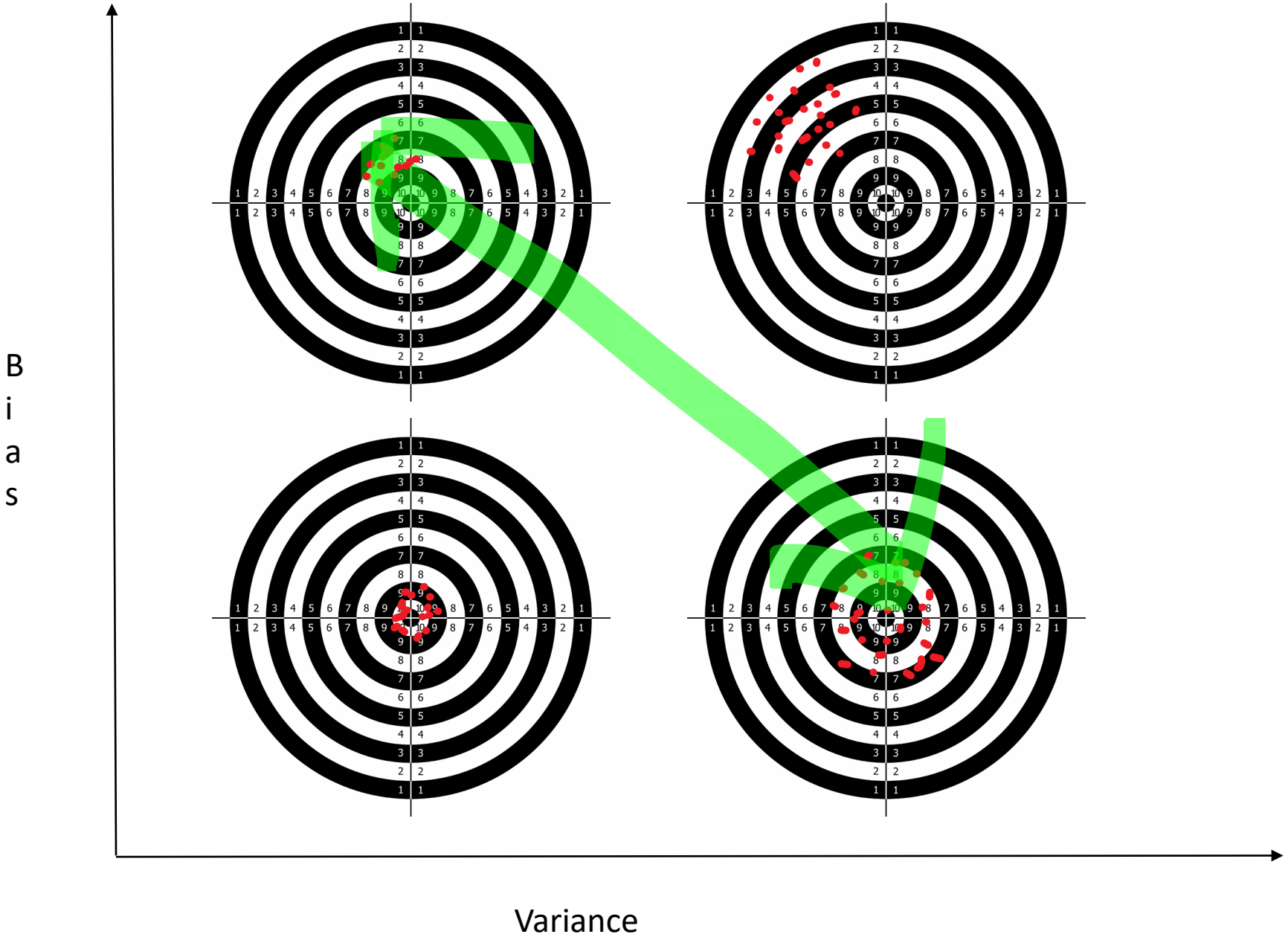




High bias


1 1
2 2
3 3
4 4
5 5
6 6
7 7
8
9 9
1 2 3 4 5 6 7 8 9 10 10 9 8 7 6 5 4 3 2 1
1 2 3 4 5 6 7 8 9 10 10 9 8 7 6 5 4 3 2 1
9 9
8 8
7 7
6 6
5 5
4 4
3 3
2 2
1 1


Bias
Variance


Bias
Variance

Intuitive definitions

- Suppose we create many training sets from the same data
  - Eg, by re-spitting randomly into train and test
  - In this case some rows that were in the train set might end up in the test set and vice versa

- For each training set (n=1 … N) we estimate a model $\hat{f}^{(n)}$ with parameters ($\beta^{(n)}$)

- Then, for a point x_0 we estimate $f^{(n)}(x_0) = \hat{y}^{(n)}$

- **Bias:** the difference between our average prediction $\frac{1}{n}\sum\hat{y}^{(n)}$ (over many training sets) and the true value y.

- **Variance:** the variance of predictions at x_0– are they similar or they vary a lot?

Bias-variance trade off

The MSE (average prediction error) can be written as:

$$MSE = Variance + Bias^2 + Irreducible\ error$$

As flexibility increases, variance goes up, and bias goes down

Therefore, we are facing a trade-off between bias and variance

Bias-variance trade off

$$MSE = E\left(\left(Y - \hat{f}(x)\right)^2\right) = Var\left(\hat{f}(x)\right) + \left\{Bias(\hat{f}(x))\right\}^2 + Var(e)$$

…where $Bias\left(\hat{f}(x)\right) = E\left(\hat{f}(x)\right) - f(x)$

As flexibility increases, variance goes up, and bias goes down

Therefore, we are facing a trade-off between bias and variance

Bias-variance trade off

$$MSE = E\left(\left(Y - \hat{f}(x)\right)^2\right) = \underbrace{Var\left(\hat{f}(x)\right) + \left\{Bias(\hat{f}(x))\right\}^2}_{\text{Reducible}} + \underbrace{Var(e)}_{\text{Irreducible}}$$

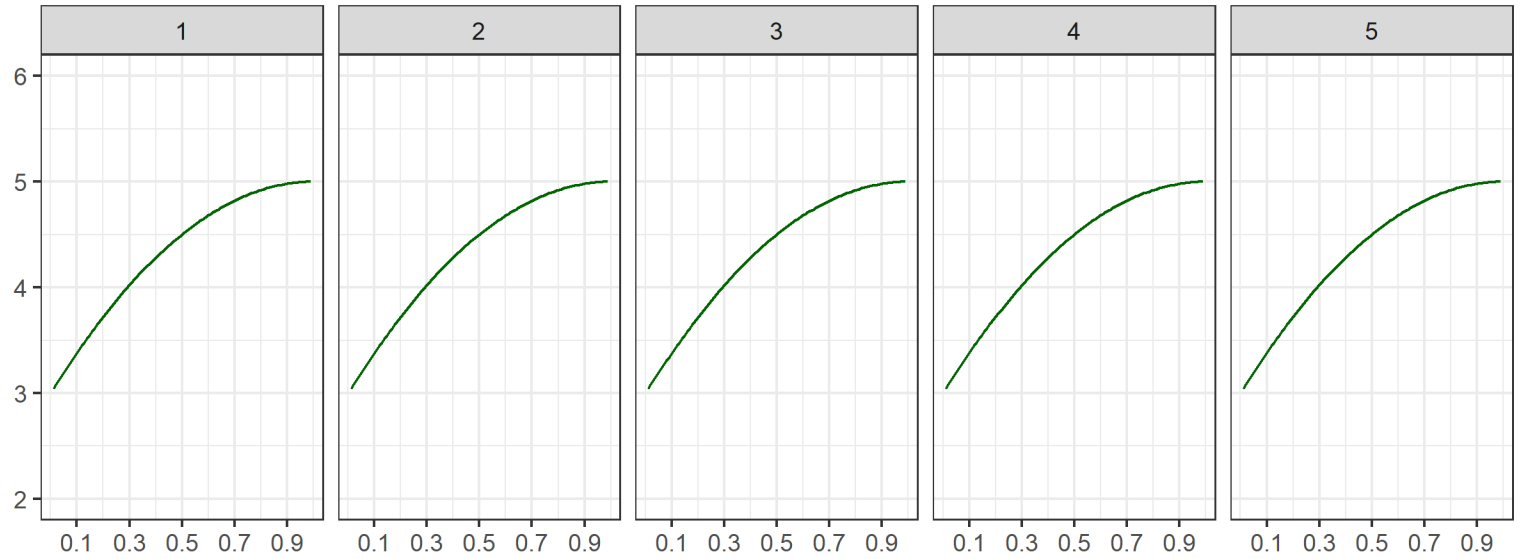
…where $Bias\left(\hat{f}(x)\right) = E\left(\hat{f}(x)\right) - f(x)$

As flexibility increases, variance goes up, and bias goes down

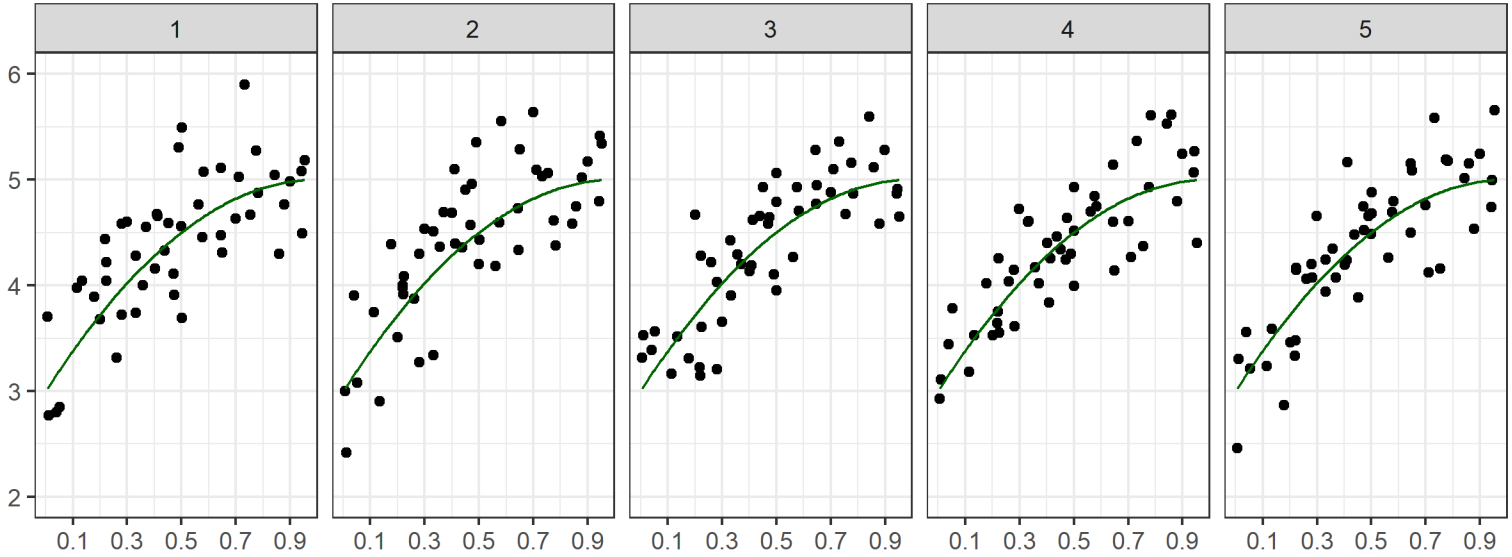
Therefore, we are facing a trade-off between bias and variance

Let’s explore the bias-variance trade-off with an example

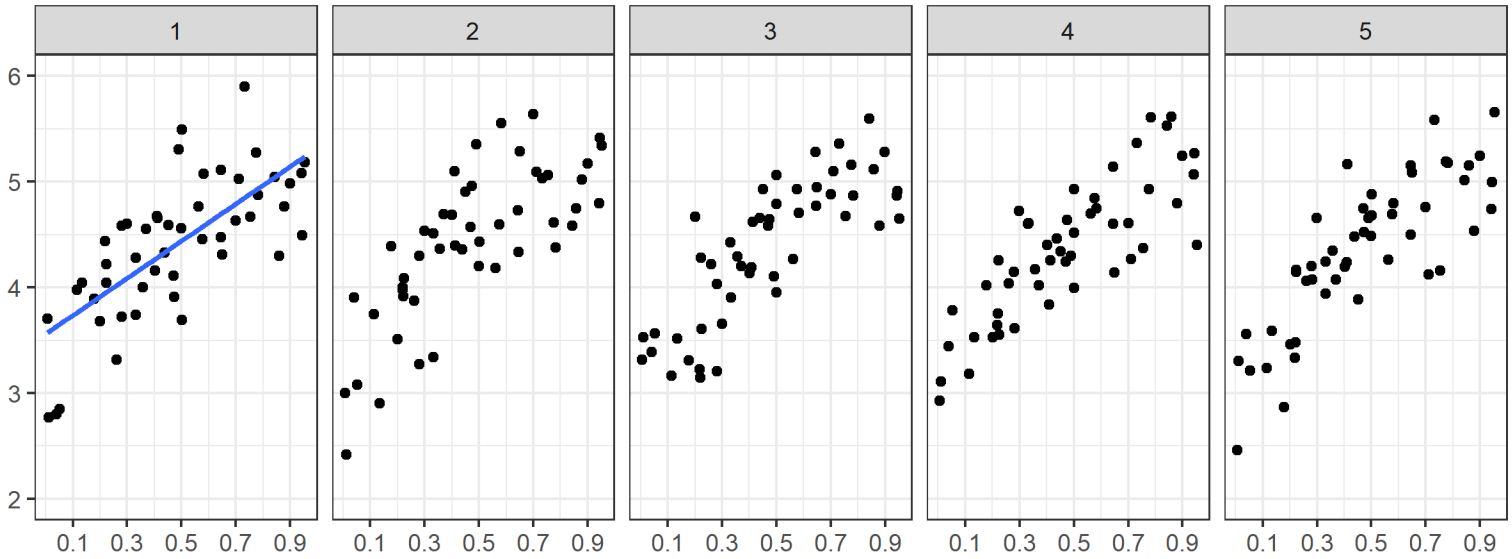
1. True relationship is $Sales(AdSpend) = 3 + 4 \times AdSpend - 2 \times AdSpend^2$
2. Five different analysts are given data, which has some random noise
3. In other words, data is of the form $x = AdSpend, \mathrm{y} = Sales(AdSpend) + e$



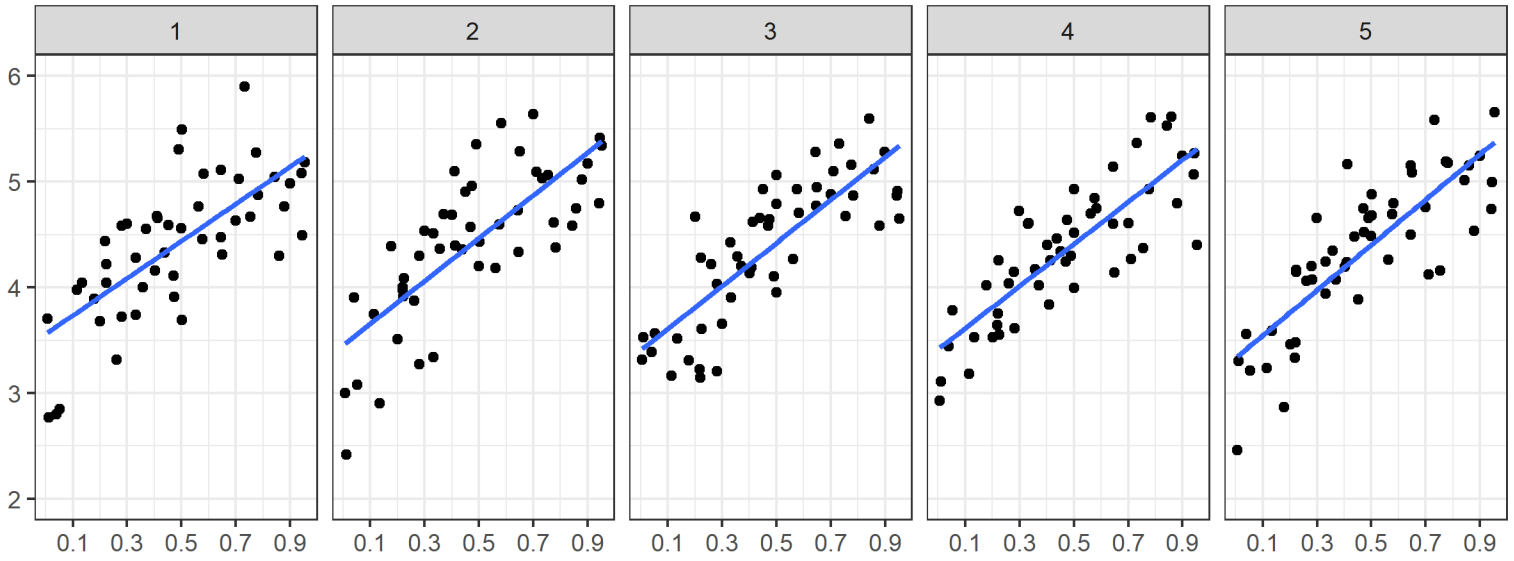
1. True relationship is $Sales(AdSpend) = 3 + 4 \times AdSpend - 2 \times AdSpend^2$
2. Five different analysts are given data, which has some random noise
3. In other words, data is of the form $x = AdSpend, \mathrm{y} = Sales(AdSpend) + e$



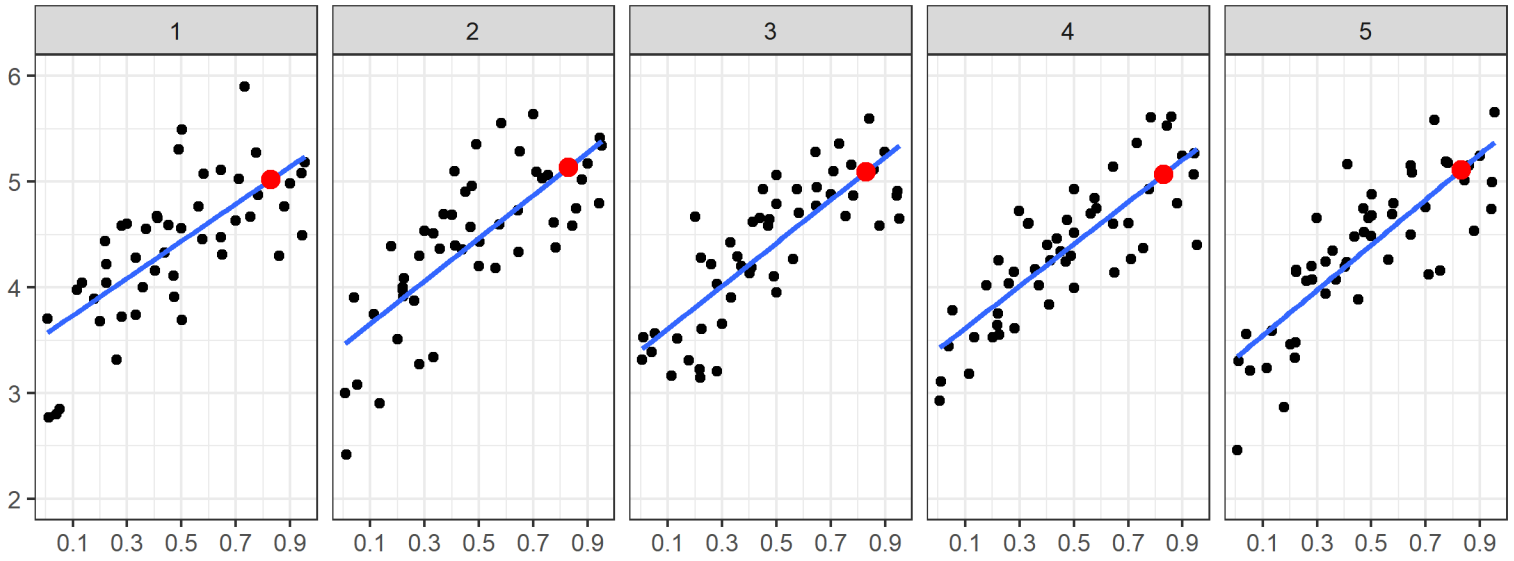
$$Sales(AdSpend) = 3 + 4 \times AdSpend - 2 \times AdSpend^2$$



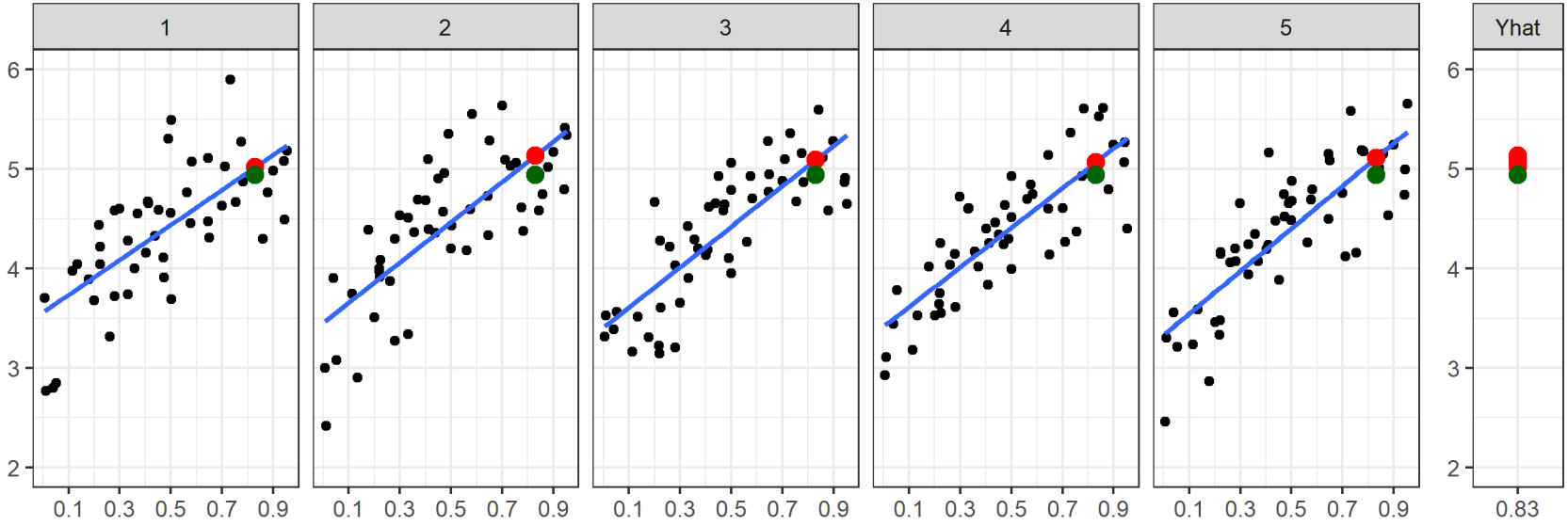
$$Sales(AdSpend) = 3 + 4 \times AdSpend - 2 \times AdSpend^2$$



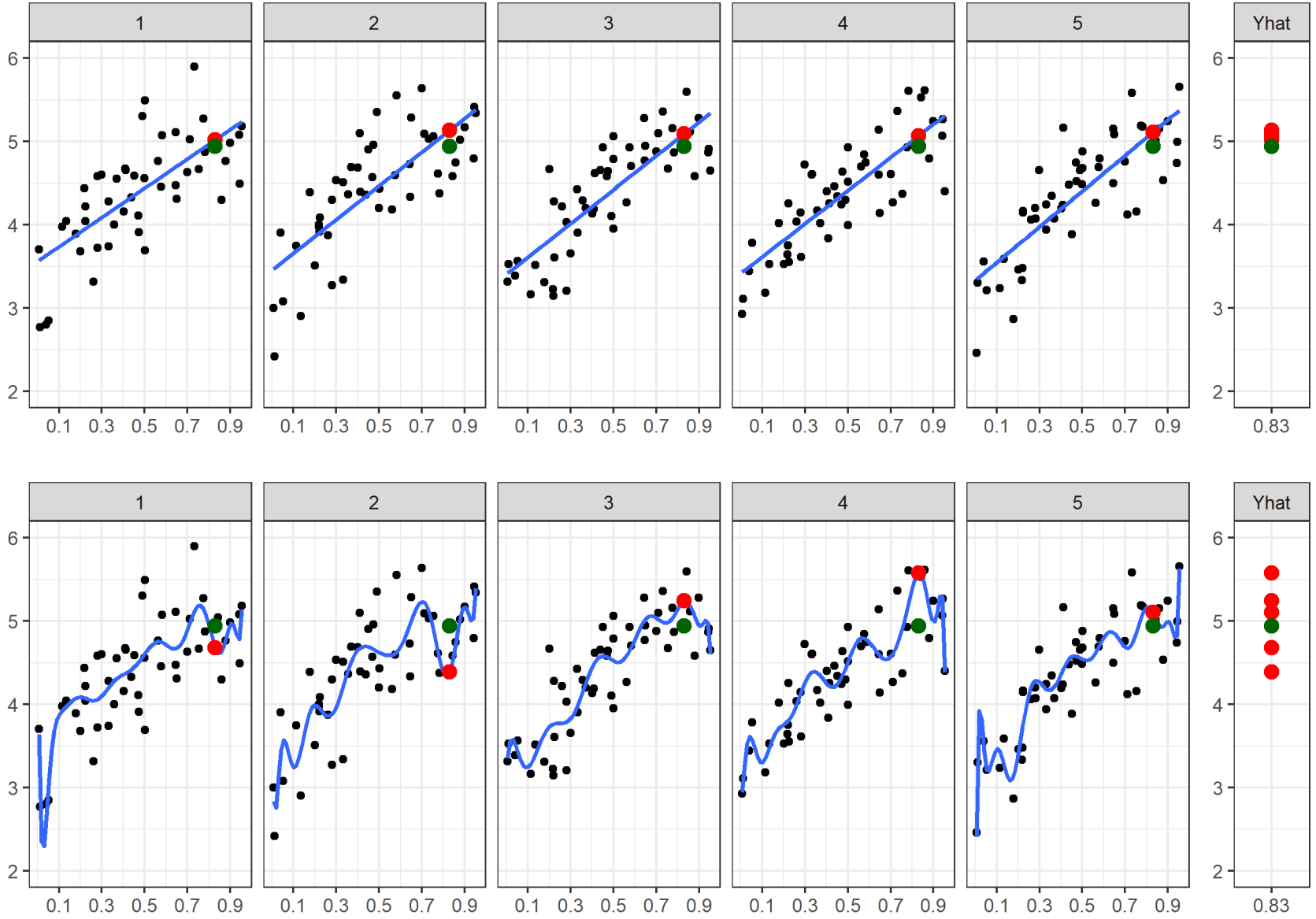
$$Sales(AdSpend) = 3 + 4 \times AdSpend - 2 \times AdSpend^2$$

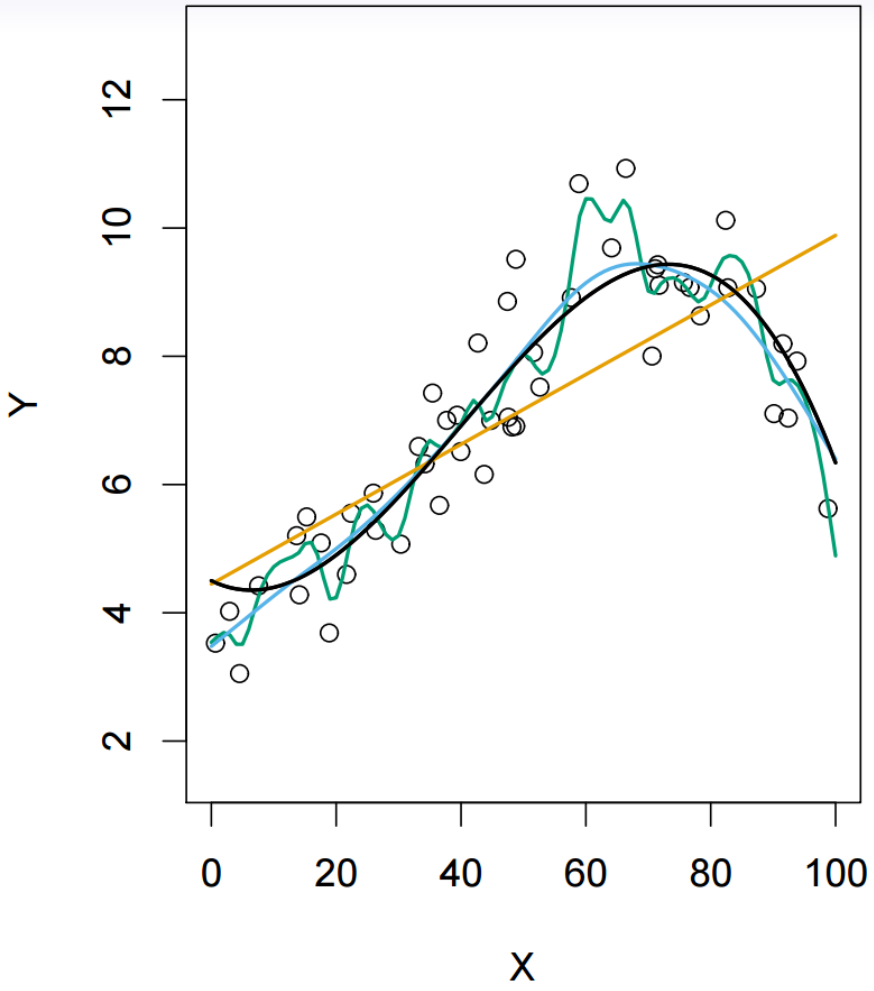


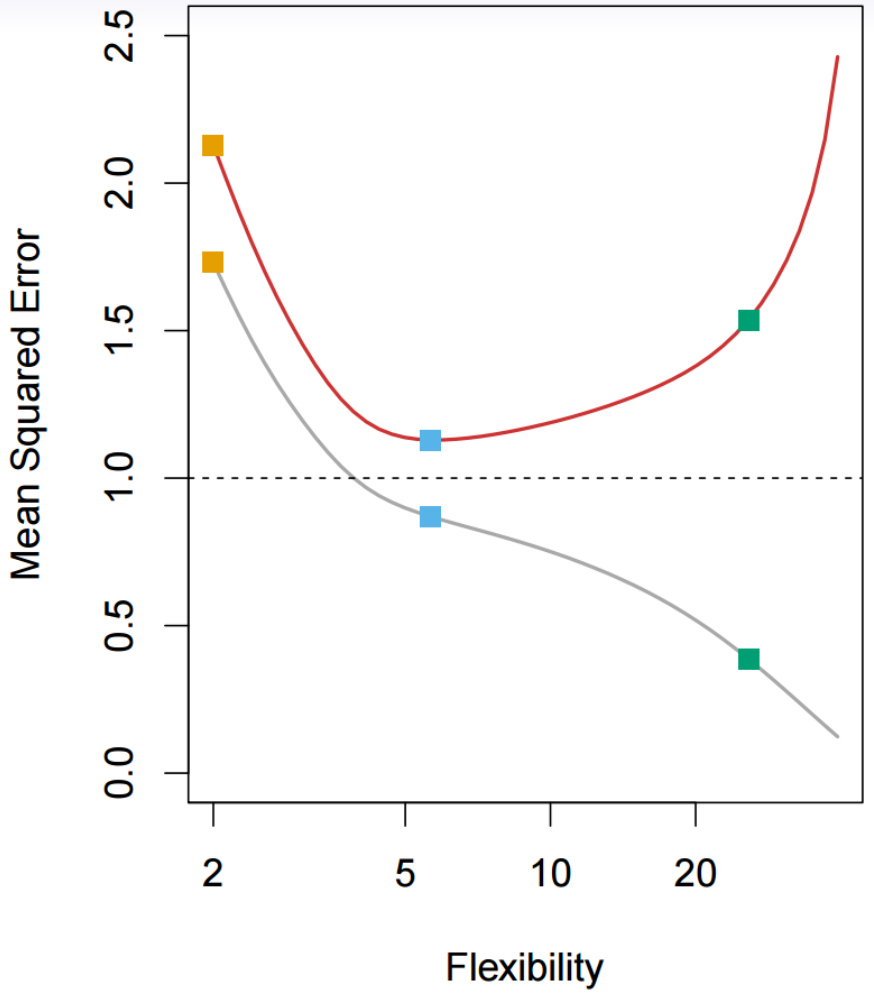
$$Sales(AdSpend) = 3 + 4 \times AdSpend - 2 \times AdSpend^2$$



$$Sales(AdSpend) = 3 + 4 \times AdSpend - 2 \times AdSpend^2$$

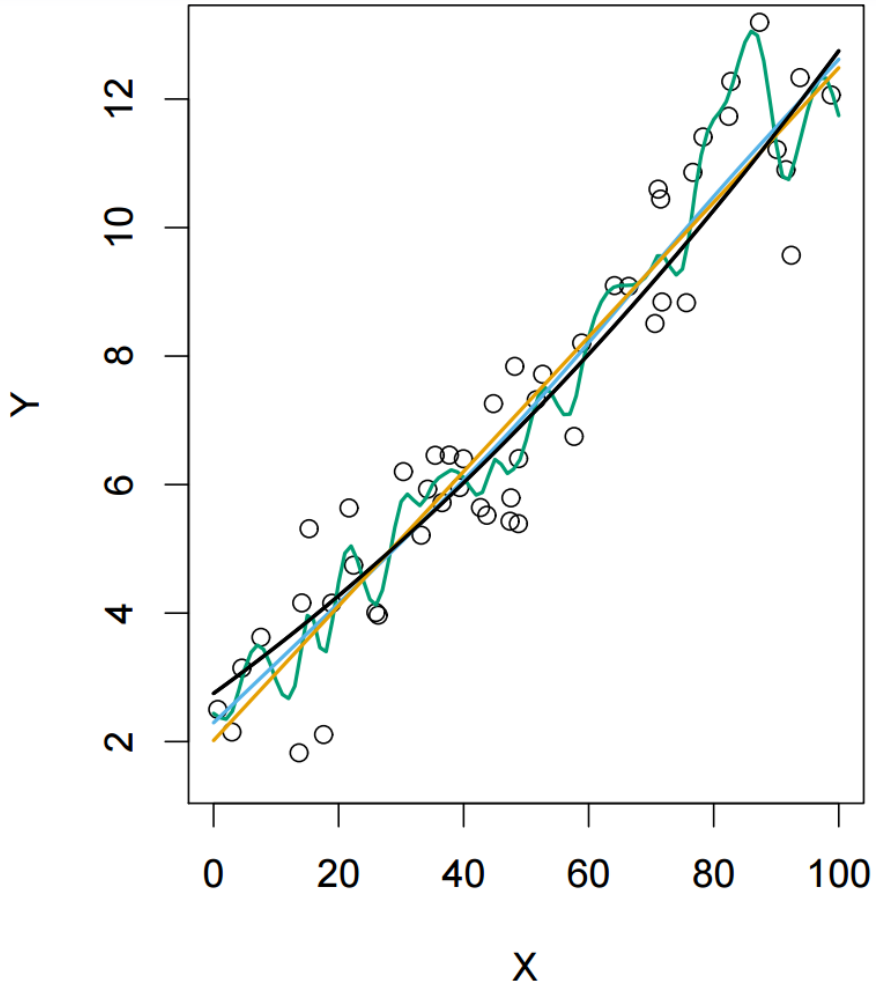


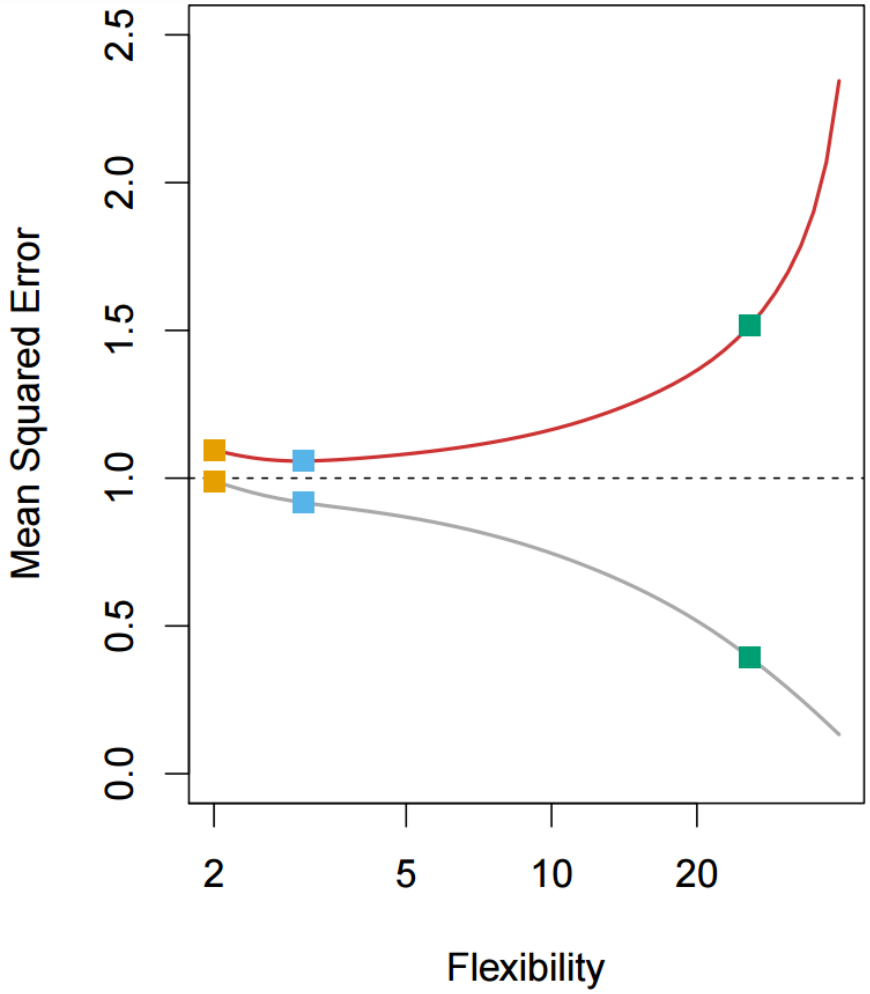




Black curve is truth. Red curve on right is MSE_{Te}, grey curve is MSE_{Tr}. Orange, blue and green curves/squares correspond to fits of different flexibility.

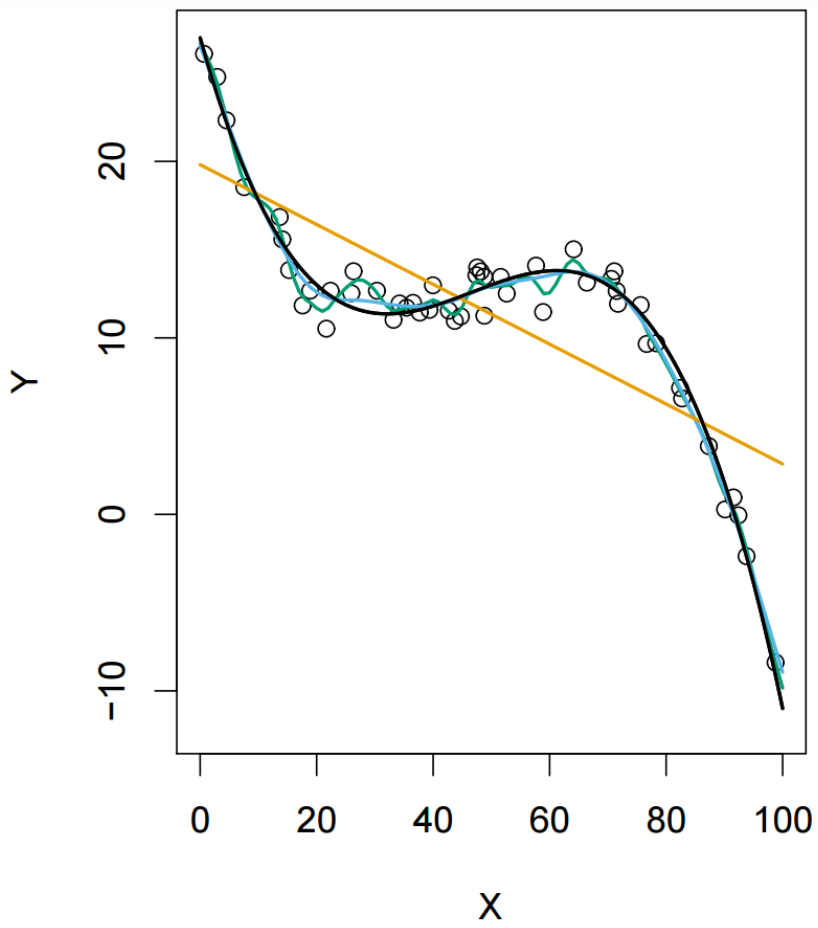
Source: Tibshirani et al.

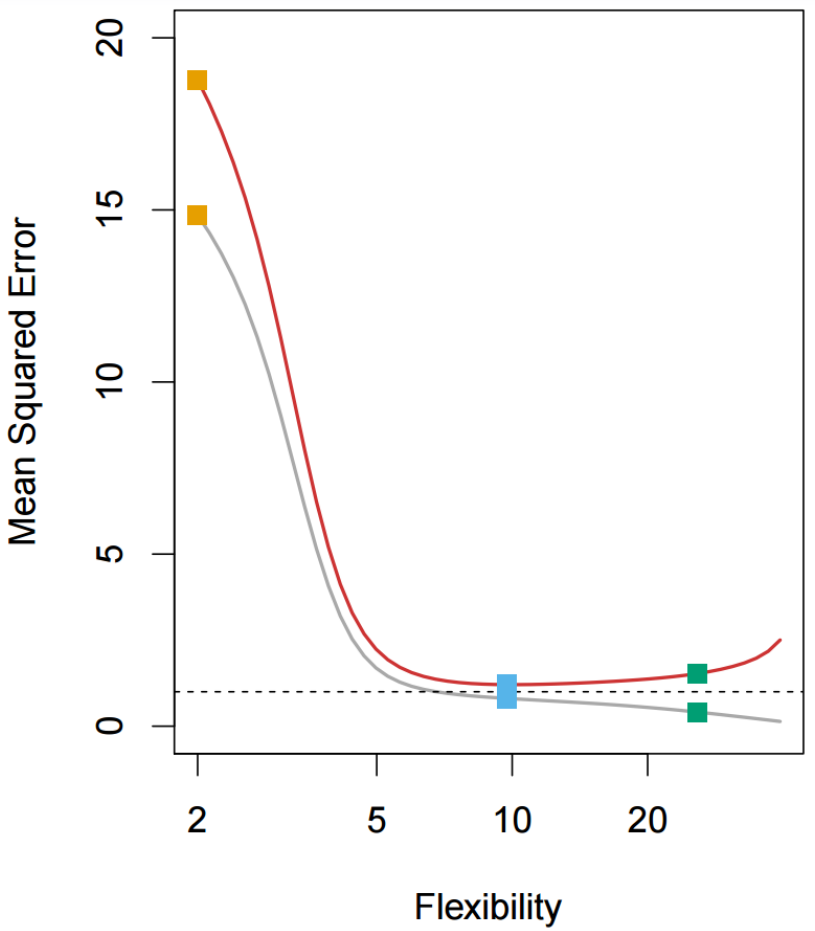




Here the truth is smoother, so the smoother fit and linear model do really well.

Source: Tibshirani et al.





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

Source: Tibshirani et al.

Navigating the bias-variance trade-off

- How do we balance model complexity and prediction accuracy?

- How about this:
  - For every possible combination of predictors, fit a linear regression
  - Compute test MSE
  - Pick the best model

- Would this work?

- Is it practical?

Forward selection

Simple idea

1. Start with a model that only contains only an intercept
2. Fit p linear regressions, one for each predictor X_p, and compute $MSE(p)$
3. Add the predictor X_p with the smallest $MSE(p)$ to the model
4. Repeat

Backward selection

Similar idea

1. Start with a model that only contains all variables
2. Fit p linear regressions, removing one predictor X_p each time, and compute $MSE(p)$
3. Choose the model with smallest $MSE(p)$
4. Repeat

END

- End