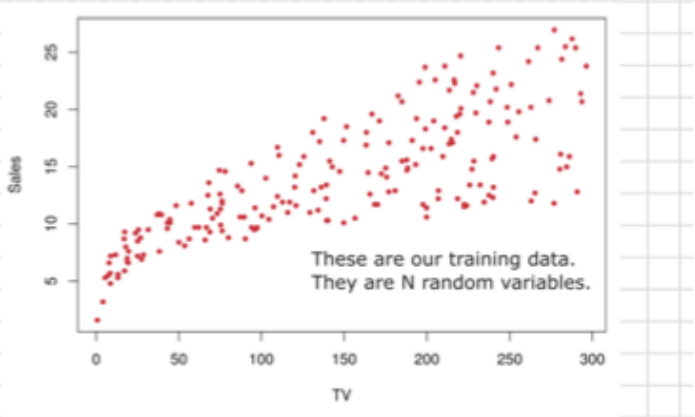
Lecture 3 - 03/09/24

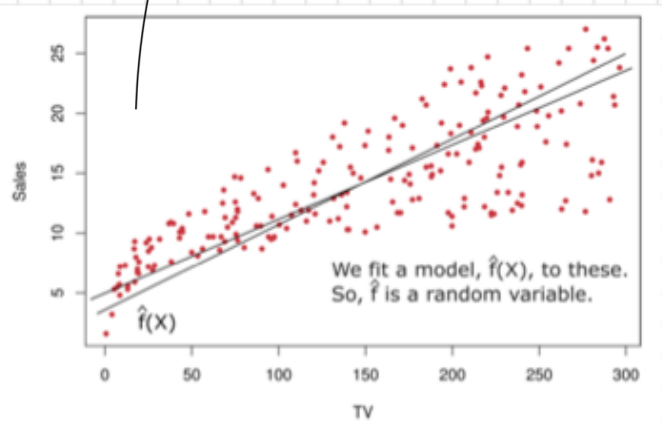
- **Quick notions**

  The N observations we refer to as **training data**

  The algorithm used for fitting the function is called **learner**

  Applying this learner to the training data, we call **training**





- **What is our goal?**

  Our goal is to predict unseen data points: (x_0, y_0)

  Imagine we have 2 models fitted to the data:

  $f_1(x)$

  $f_2(x)$

  We now give them an unseen data point (x_0, y_0)

  How do we know which of the two is best?

  Imagine now we give them n unseen data points (x_i, y_i) for $i = 1, \dots, n$

  Which of the two is best?

- **The answer to previous question...**

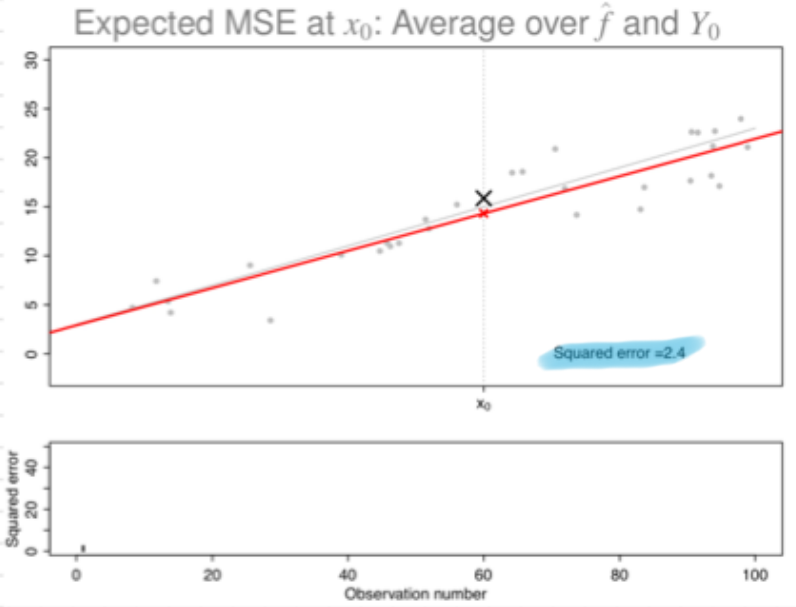
  Relies on the **Expected test MSE**: (mean squared error)

  $$E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right]$$ → we want this to be as small as possible.

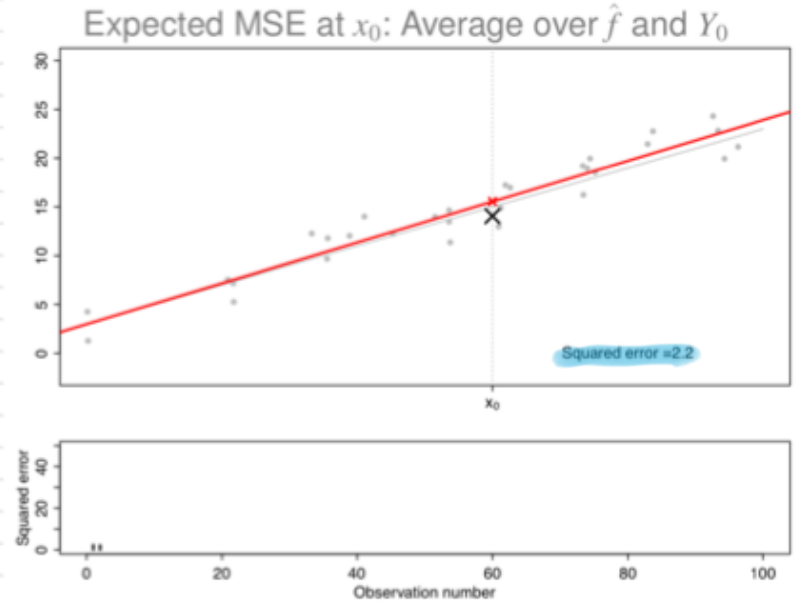
  This is basically the positive difference between the real y_0 & our predicted y outputted by our model $f(x_0)$

  Remember

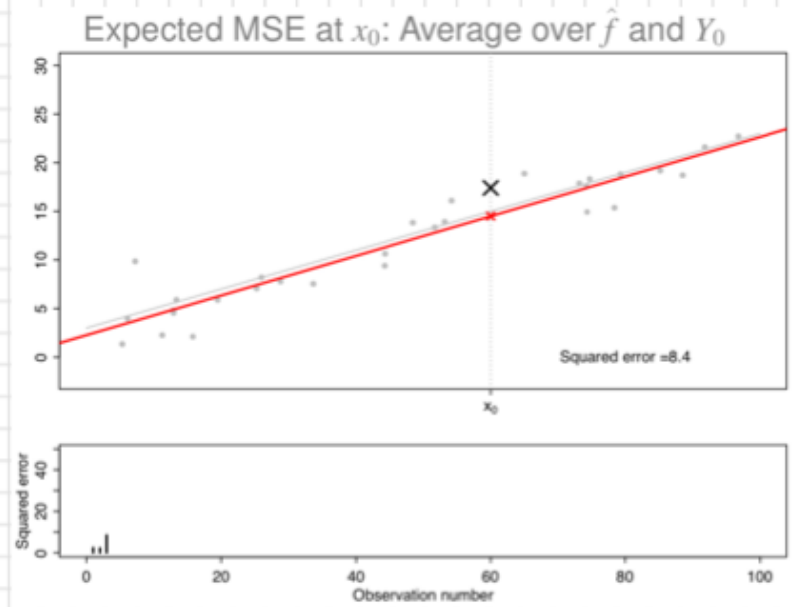
  The unseen data is what we care about



- In this case our f(x) predicted ŷ with a 1.2 unit difference from the real y
  - this leads to our squared error = 2.4

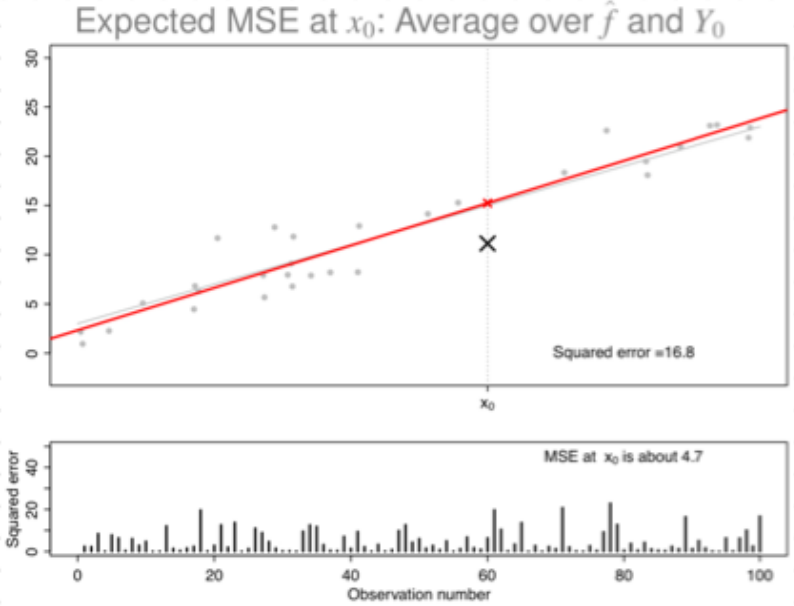


- In this case our f(x) predicted ŷ with a 1.1 unit difference from the real y
  - this leads to our squared error = 2.2



- In this case our f(x) predicted ŷ with a 4.2 unit difference from the real y
  - this leads to our squared error = 8.4

..... We make lots of them (watch slides)



WE END UP WITH THIS.

- WE CAN NOW THINK ABOUT THE FOLLOWING QUESTION:

> *What is the expected squared error that we get from the entire process of first training the learner on some training data and then using it for predicting the response Y_0 for a new observation X_0?*

Remember:

- The function $\hat{f}$ is a random variable, because it is obtained from applying a learner to a dataset of N random variables (X, Y).
- A new observation (X_0, Y_0) is a random variable.

BIAS-VARIANCE DECOMPOSITION

- THE EXPECTED TEST MSE AT X_0 CAN BE DECOMPOSED AS:

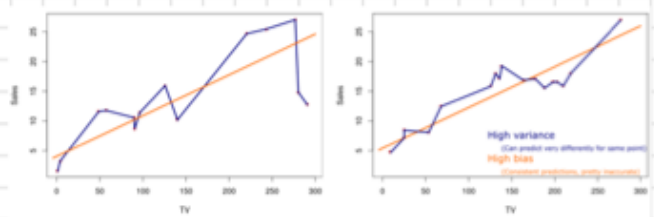
$$\mathbb{E}\left(Y_0 - \hat{f}(x_0)\right)^2 = \underbrace{\mathbb{E}\left(\hat{f}(x_0) - \mathbb{E}\hat{f}(x_0)\right)^2}_{\text{Variance of } \hat{f}(x_0)} + \underbrace{\left(\mathbb{E}\hat{f}(x_0) - f(x_0)\right)^2}_{\text{Bias of } \hat{f}(X_0)} + \text{Var}(\epsilon)$$

FIRST TERM IS THE VARIANCE

RECALL: MEASURE OF SPREAD IN DATA FROM IT'S MEAN. AMOUNT OF CHANGE DUE TO NEW SUBSETS OF DATA

SECOND TERM IS THE EXPECTED DEVIATION OF OUR MODEL PREDICTION FROM THE TRUE VALUE (BIAS)

RECALL: DIFF BETWEEN MODEL PREDICTED VALUE & REAL VALUE.



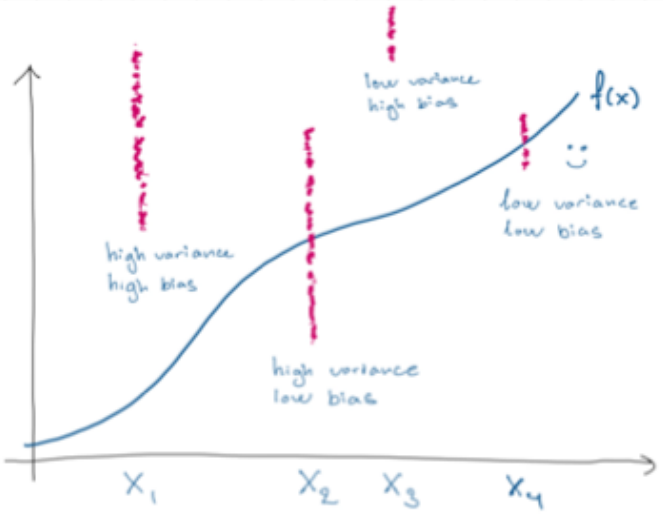
From ISL:

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

"*Bias* refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model."

→ PROBLEMS:

OVERFITTING
UNDERFITTING



(In this illustration, red dots are predictions made by different instances of $\hat{f}$)

- RECALL PROFESSOR EXAMPLE ABOUT THE DIFFERENT SCENARIO
- IT IS VERY DIFFICULT TO GET LOW BIAS & LOW VARIANCE
- BY PROVIDING LOW BIAS & LOW VARIANCE WE PREVENT THE TWO PROBLEMS:
  - OVERFITING (STUDY TOO MUCH TO TRAIN)
  - UNDERFITING (NOT ENOUGH TRAIN)

AS SAID, PROVIDING BOTH LOW & LOW VARIANCE IS VERY DIFFICULT THAT'S WHY WE TALK ABOUT **BIAS - VARIANCE TRADEOFF**

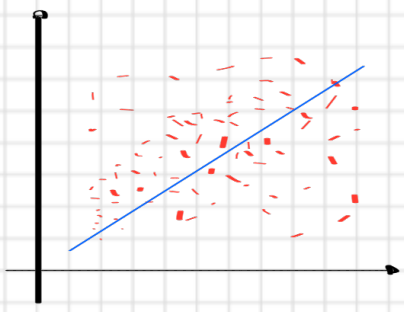
$$\mathbb{E}\left(Y_0 - \hat{f}(x_0)\right)^2 = \underbrace{\mathbb{E}\left(\hat{f}(x_0) - \mathbb{E}\hat{f}(x_0)\right)^2 + \left(\mathbb{E}\hat{f}(x_0) - f(x_0)\right)^2}_{\text{Reducible error}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible error}}$$

All three terms are non-negative, so if any is large, the MSE is large.

- THE REDUCIBLE ERROR CAN BE LOWERED BY USING AN ESTIMATOR $\hat{f}$ THAT HAS BOTH LOW VARIANCE & LOW BIAS
- THE IRREDUCIBLE ERROR IS A LOWER BOUND ON THE ACCURACY OF OUR PREDICTION Y

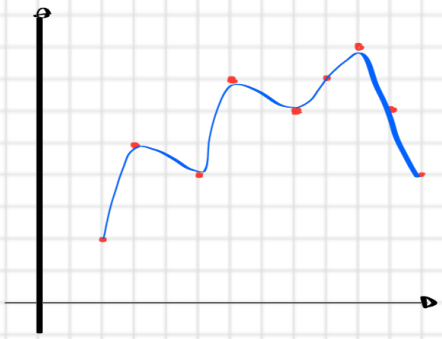
WE CAN THINK ABOUT TWO EXAMPLES:

FIT A CONSTANT LINE



LOW VARIANCE
HIGH BIAS

FIT A CURVE THAT INTERPOLATES ALL POINTS

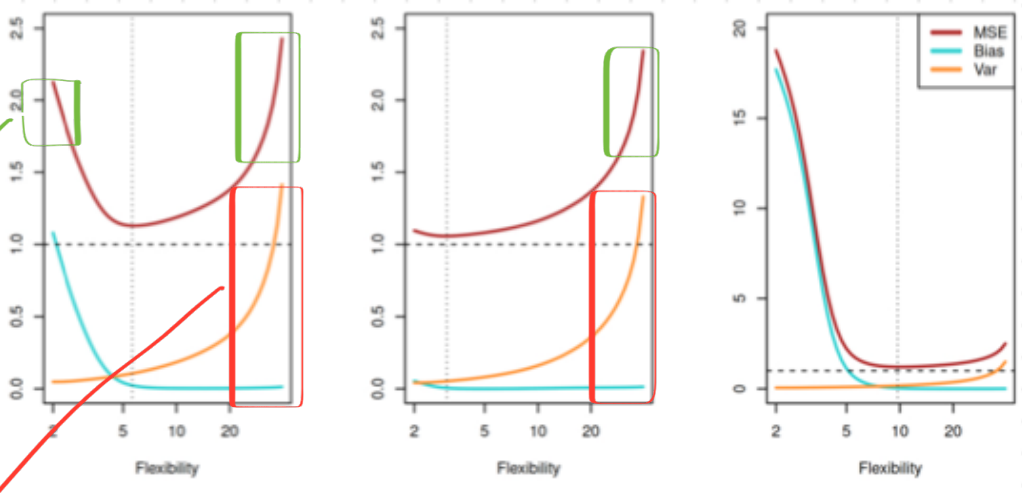


HIGH VARIANCE
LOW BIAS

* NOTE

IT IS VERY EASY TO FIND THEN AN $\hat{f}$ WITH EITHER VERY LOW BIAS OR VERY LOW VARIANCE. BUT LESS STRAIGHTFORWARD TO FIND ONE WITH BOTH

- We can further more analyze Bias, Variance & MSE with respect to complexity



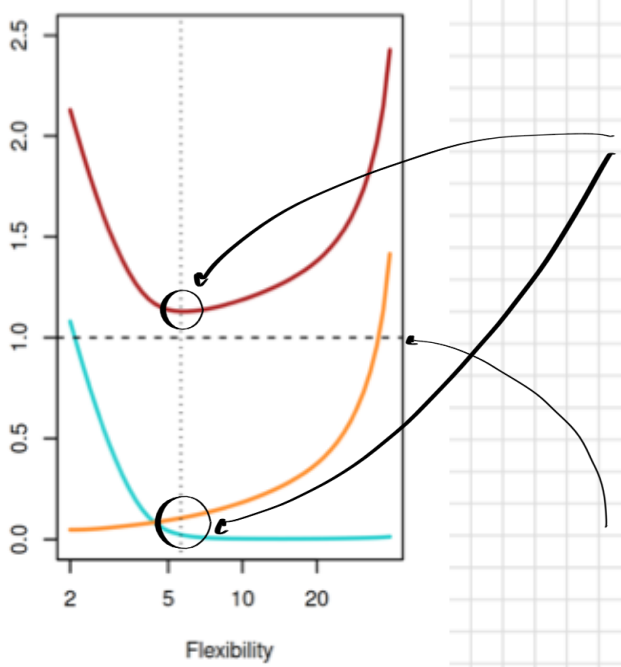
"Flexibility" is model complexity.

Panels are 3 different data sets.

Q: What do these 3 panels have in common?

- From the three graphs we were multiple things:
  - MSE is very HIGH when model is too simple or too complex
    - Remember!!! → UNDERFITHING & OVERFITHING
  - Variance grows as complexity arises & it is the inverse for Bias that decreases as model gets more complex.
    - Remember!!! → Easy to find inverse match, very difficult to find both low

- So where do we want our MSE to be?

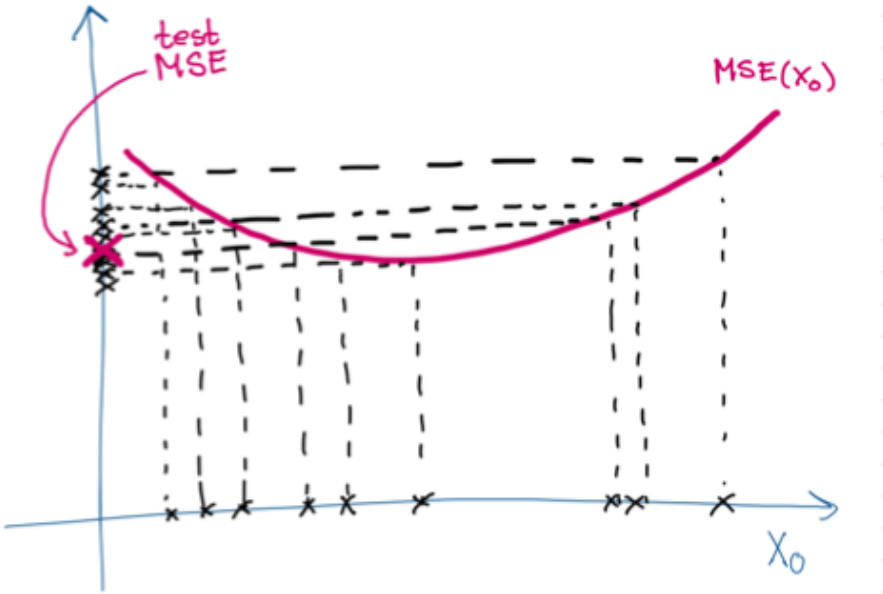


Sweet Spot

This is called sweet spot, we see that we have a good tradeoff between variance & bias & the MSE is at it's minimum.

This is the irreducible variance, an error that we can't minimize

- So far we have considered the test MSE at an individual point X_0
- To have thought a great model fit, we want to do well in ALL points



Expected test MSE

- What we want to do is take expectation over new values $X_0, y_0, \hat{f}$
- This can be done sequentially by:
  1. First finding the MSE at X_0, which is then a function of X_0
  2. After find the expectation of the MSE at X_0 when it varies:

$$\mathbb{E}(MSE(\hat{f}, Y_0, X_0)) = \mathbb{E}_{X_0}\left(\mathbb{E}_{\hat{f},Y_0}(MSE(\hat{f}, Y_0, X_0)\,|\,X_0 = x_0)\right)$$

- By doing this we take expectations over X_0 in the bias-variance decomposition that result into:

$$\text{MSE} = \underbrace{\mathbb{E}\left(\text{Var}(\hat{f}(X_0))\right)}_{\substack{\text{“Average” variance of} \\ \hat{f}(x_0) \text{ across the range} \\ \text{of } X_0.}} + \underbrace{\mathbb{E}\left(\text{Bias}(\hat{f}(X_0))\right)^2}_{\substack{\text{“Average” squared bias} \\ \text{of } \hat{f}(x_0) \text{ across the} \\ \text{range of } X_0.}} + \underbrace{\text{Var}(\epsilon)}_{\substack{\text{Residual} \\ \text{variance}}}$$

Test MSE vs Training MSE

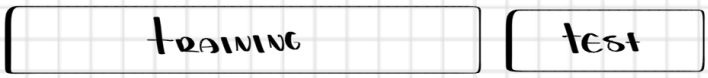
- The MSE is a theoretical quantity that we can estimate from data
- What we do is we use data points that were NOT used during training (train set)
  - This is always due to the underfitting & overfitting problems

Always keep in mind this distinction

- Model Selection: Estimating the prediction model aiming to choose best one.
- Model Assessment: Having chosen a final model, estimating its prediction error.

Holding Out Data for Testing

- In linear regression model we used a single dataset to train the models & selected between models using p-values or AIC.



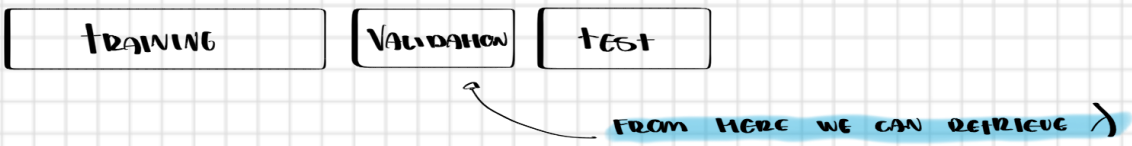
  - This is very good because we left out data that we can now give to our model & that it has never seen
  - Sometimes we have Hyperparameters to specify in our model:

    i.e $\lambda \in [0,1]$ where a specific val gives best-performing model

Question

How can we train model for different λ's & choose the best?

- The answer is to change the dataset division into:



  - The process now sees training, tuning hyperparams, test.
  - A typical split is:

    Training: 60%

    Validation: 20%

    test: 20%