# Supervised Learning (Part I)

DS-6030 | Spring 2026

supervised_1.pdf

## Table of contents

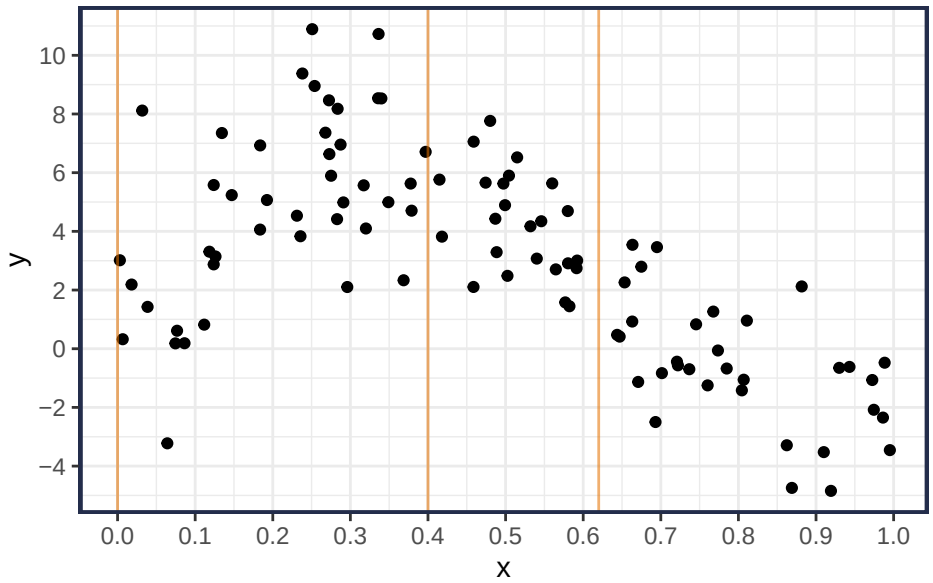
# 1 Supervised Learning Intro

## 1.1 Supervised Learning

- In *supervised learning*, each observation can be partitioned into two sets: the predictor variables and the outcome variable(s).
  - Predictor variables are sometimes called independent/feature variables
  - Outcome variables are sometimes called target/labels/response/dependent variables.
- Usually the predictor variables are represented by $X$ and the outcome variable(s) represented by $Y$
- The goal in supervised learning is to find the patterns and relationships between the predictors, $X$, and the outcome, $Y$.
  - Usually the goal is to *predict* the value of $Y$ given $X$.

# 2 Example Data

Consider some data $D = \{(X_i, Y_i)\}_{i=1}^{n}$ with $Y_i \in \mathbb{R}$, $X_i \in [0, 1]$ and $n = 100$.




**Your Turn #1**

The goal is to predict new $Y$ values if we are given the $X$'s.

- If $x = .40$, predict $Y$.
- If $x = 0$, predict $Y$.
- If $x = .62$, predict $Y$.
- How should we build a *model* that will automatically predict $Y$ for any given $X$?

# 3 Linear Models

- Linear regression refer to a class of models where the output (predicted value) is a linear combination (weighted sum) of the input variables

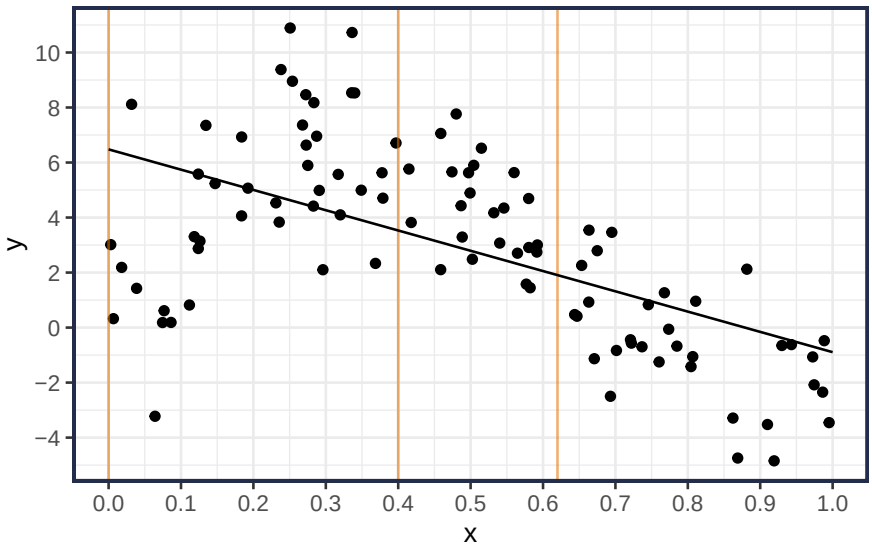
$$f(x;\beta) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

where $x = [x_1, \ldots, x_p]^\mathsf{T}$ is a vector of features/variables/attributes and $\hat{Y}|x = f(x;\hat{\beta}) = x^\mathsf{T}\hat{\beta}$ is the predicted outcome at $X = x$.

- the *model parameters* for linear (or coefficients or weights), $\hat{\beta}$ determine how much influence each feature has on the predicted output.

**Model Structure**

- Outcome variable: $y \in \mathbb{R}$
- Predictor variables $\mathbf{x} \in \mathbb{R}^p$
- Model parameters: $\beta = (\beta_0, \beta_1)$
- Prediction function: $f(\mathbf{x};\hat{\beta}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 = \mathbf{x}^\mathsf{T}\beta$




**Your Turn #2**

1. How did we do? If $X_{\text{new}}$ is close to 0, or close to 0.4, or close to .62?
2. How to make it better?

# 4 Polynomial inputs

- In the *simple* linear regression model, we had 2 parameters that we needed to estimation, $\beta_0$ and $\beta_1$. Thus, the *model flexibility/complexity* is minimal.
  - The only thing simpler is an intercept only model.
- But the data appears to have a more *complex* structure than linear.
- A *parametric approach* to add flexibility is to incorporate *polynomial terms* into the model.
  - A quadratic model is $f(x;\beta) = \beta_0 + \beta_1 x + \beta_2 x^2$

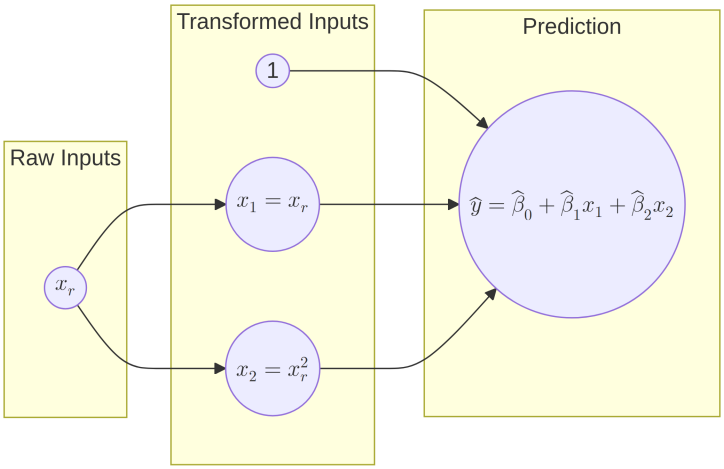
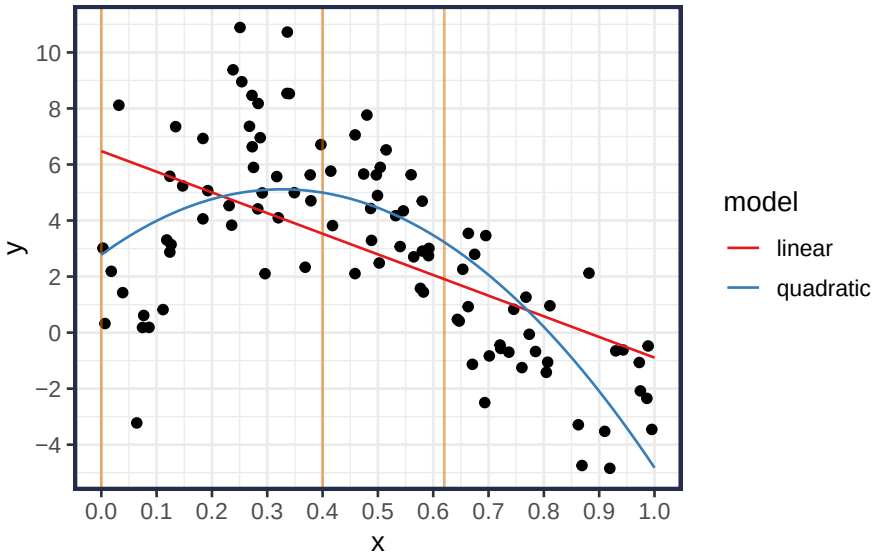



Figure 1



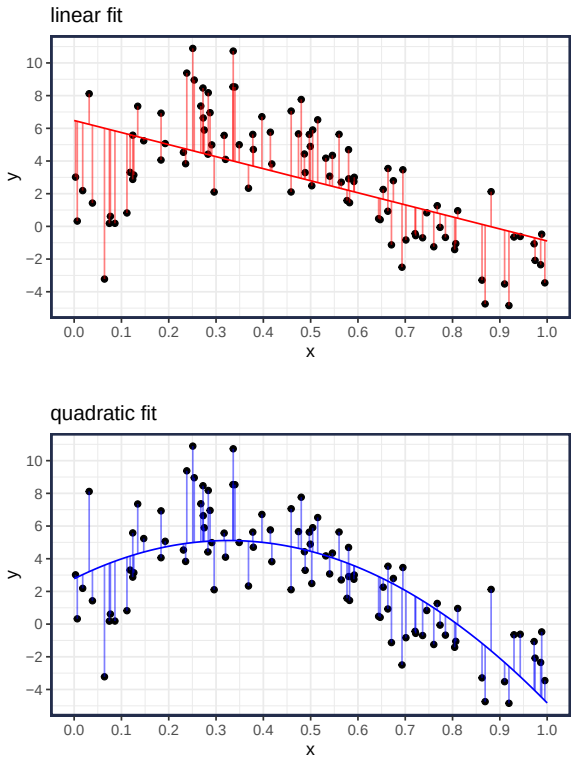

**Your Turn #3**

1. How did we do? If $X_{\text{new}}$ is close to 0, or close to 0.4, or close to .62?
2. But does the quadratic model fit better *overall*?
3. What is the *complexity/flexibility* of the quadratic model?

## 4.1 Performance Comparison (on Training Data)

Comparing the two models (according to MSE), the quadratic model does much better!

| degree | MSE | # pars |
|---|---|---|
| 1 | 8.29 | 2 |
| 2 | 5.58 | 3 |




As my kids always reason, "if a little is good, than a lot must be better". So why not try more complex models by increasing the polynomial degree.

- Polynomial of degree $d$

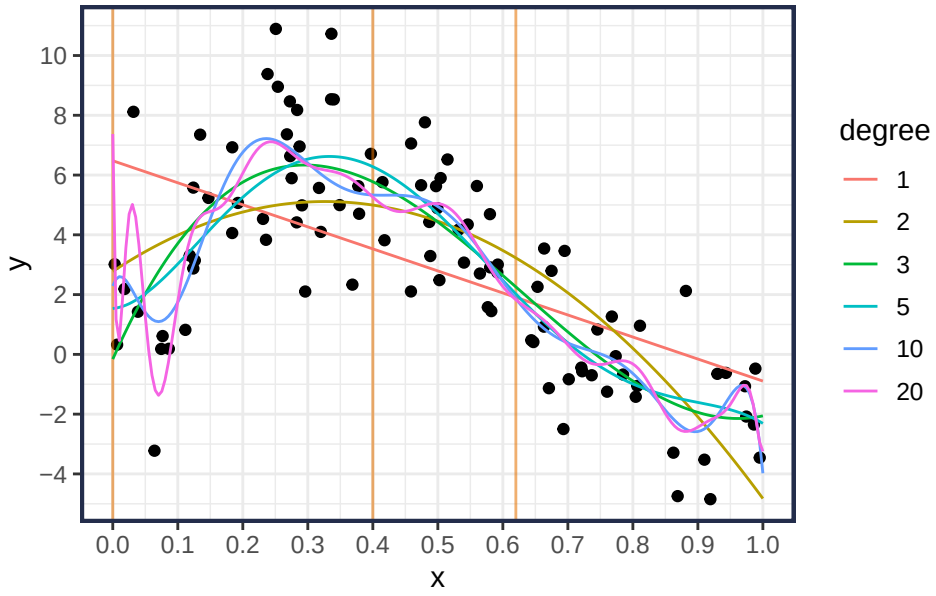
$$f_{\text{poly}}(x;\beta,d) = \beta_0 + \sum_{j=1}^{d} \beta_j x^j$$

**Model and Tuning parameters**

- The $\beta = (\beta_0, \beta_1, \ldots, \beta_d)$ are the *model parameters*.
- The polynomial degree $d$ is the *tuning parameter*.

| degree | MSE | # pars |
|---|---|---|
| 1 | 8.29 | 2 |
| 2 | 5.58 | 3 |
| 3 | 4.28 | 4 |
| 5 | 4.10 | 6 |
| 10 | 3.65 | 11 |
| 20 | 3.16 | 21 |

And its always good to observe the plot




- For `degree=20`, the behavior at the end points are a bit erratic.
- Using a higher degree would further reduce the MSE, but the fitted curve would be more "complex" and may not be as good for new data.

# 5 Nearest Neighbor models

- The $k$-NN method is a non-parametric *local* method, meaning that to make a prediction $\hat{y}|x$, it only uses the training data in the *vicinity* of $x$.
  - contrast with OLS linear regression, which uses all $x$'s to get prediction.
- The model is simple to describe. It finds the $k$ most similar/closest points in the training data and uses their average.

$$\begin{aligned} f_{\text{knn}}(x; k) &= \frac{1}{k} \sum_{i:x_i \in N_k(x)} y_i \\ &= \text{Avg}(y_i \mid x_i \in N_k(x)) \end{aligned}$$

  - $N_k(x)$ are the set of $k$ nearest neighbors to $x$
  - only the $k$ closest $y$'s are used to generate a prediction
  - it is a *simple mean* of the $k$ nearest observations

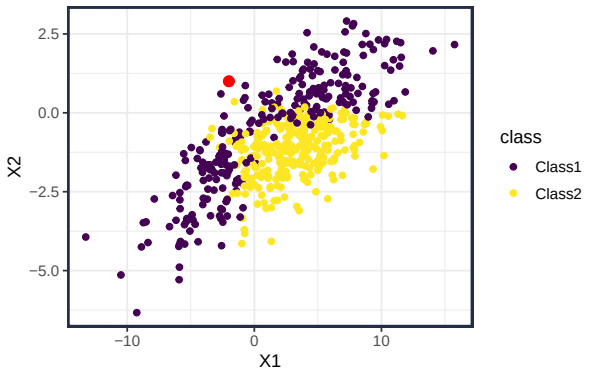
**Your Turn #4**

What is the estimate $f_{\text{knn}}(x; k = n)$?

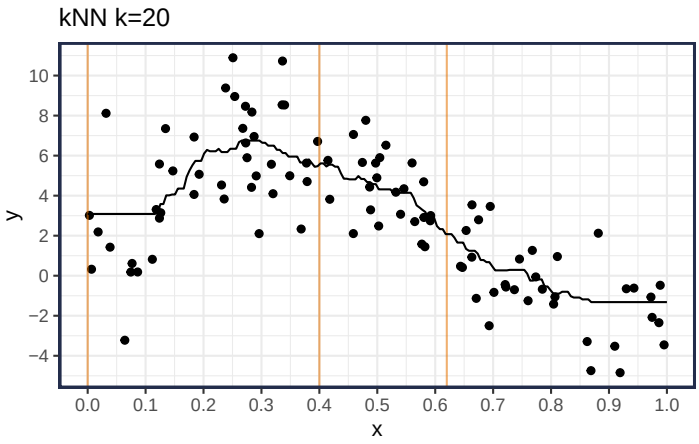
## 5.1 Notes about knn

- One computational drawback of knn methods is that all the training data must be stored in order to make predictions.
  - For very large training data, you may need to sample (or use prototypes/clusters)
  - At prediction time, the nearest neighbors for the new data needs to be computed. This can be computationally costly.
- The *flexibility* of a knn model increases as $k$ decreases.
- The least complex model, which is a constant, occurs when $k = n$
- The most complex model when $k = 1$
- The effective degrees of freedom or *edf* for a knn model is $n/k$
  - this is a measure of the model *flexibility/complexity*. It is approximately the number of parameters that are estimated in the model (to allow comparison with parametric models)

There are some additional considerations for the usual case when the feature space is multidimensional:




- A suitable *distance* measure (e.g. Euclidean) must be chosen.
  - And predictors are often *scaled* (same sd or range) so one variable doesn't dominate the distance calculation
- Because the distance to neighbors grows exponentially with increased dimensionality/features, the *curse of dimensionality* is often referenced with respect to knn.
  - This means that in high dimensions most *neighbors* are not very close and the method becomes less *local*




## 5.2 Nearest Neighbor Models are Kernel Smoothers

$$\begin{aligned} f_{\text{knn}}(x;k) &= \frac{1}{k} \sum_{i:x_i \in N_k(x)} y_i \\ &= \text{Avg}(y_i \mid x_i \in N_k(x)) \end{aligned}$$

- $N_k(x)$ are the set of $k$ nearest neighbors to $x$
- only the $k$ closest $y$'s are used to generate a prediction
- it is a *simple mean* of the $k$ nearest observations

Kernel smoother (ks) can be written

$$f_{\text{ks}}(x;k) = \frac{\sum_{i=1}^{n} K(x,x_i)\, y_i}{\sum_{i=1}^{n} K(x,x_i)}$$

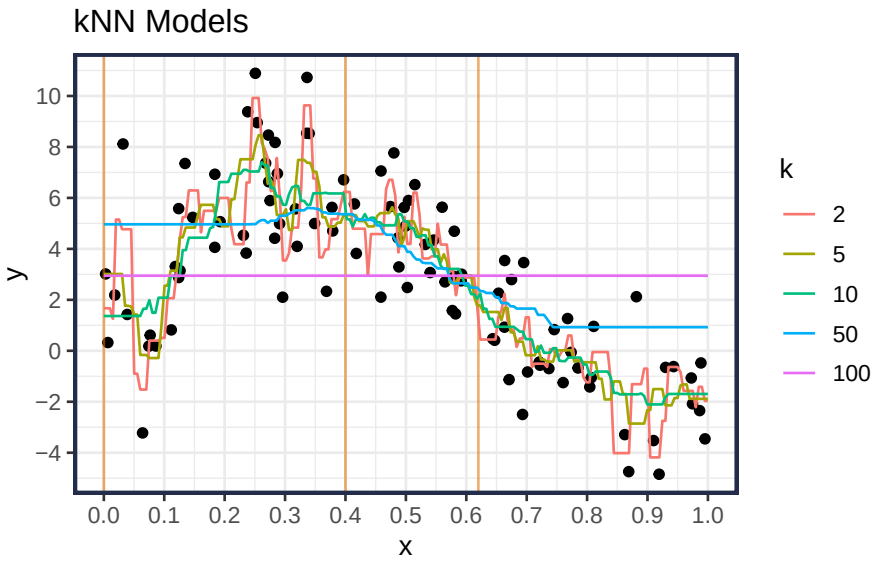
where

- $K(x, x_i)$ is the kernel that describes the similarity between the input $x$ and the training data observation $x_i$.
  - e.g., for a uniform kernel: $K(x, x_i) = \mathbb{1}(d(x, x_i) \leq h)$ where $d(x, x_i)$ is a distance measure like Euclidean or Manhattan and $h$ is the bandwidth.

The k-nearest neighbor model is equivalent to a kernel smoother with an *adaptive bandwidth*. Let $h_k(x)$ be the distance between $x$ and the $k$th nearest point in the training data. And define the kernel as $K(x, x_i) = \mathbb{1}(d(x, x_i) \leq h_k(x)) = \mathbb{1}(x_i \in N_k(x))$.

$$\begin{aligned} f_{\text{ks}}(x; k) &= \frac{\sum_{i=1}^n K(x, x_i)\, y_i}{\sum_{i=1}^n K(x, x_i)} \\ &= \frac{\sum_{i=1}^n \mathbb{1}(d(x, x_i) \leq h_k(x))\, y_i}{\sum_{i=1}^n \mathbb{1}(d(x, x_i) \leq h_k(x))} \\ &= \frac{\sum_{i=1}^n \mathbb{1}(x_i \in N_k(x))\, y_i}{k} \end{aligned}$$

## 5.3 Performance of the knn models (on training data)




| k | MSE | edf |
|---|---|---|
| 100 | 12.40 | 1 |
| 50 | 6.87 | 2 |
| 20 | 4.18 | 5 |
| 10 | 3.86 | 10 |
| 5 | 3.16 | 20 |
| 2 | 1.84 | 50 |

# 6 Predictive Model Comparison (or how to choose the best model)

## 6.1 Predictive Model Evaluation

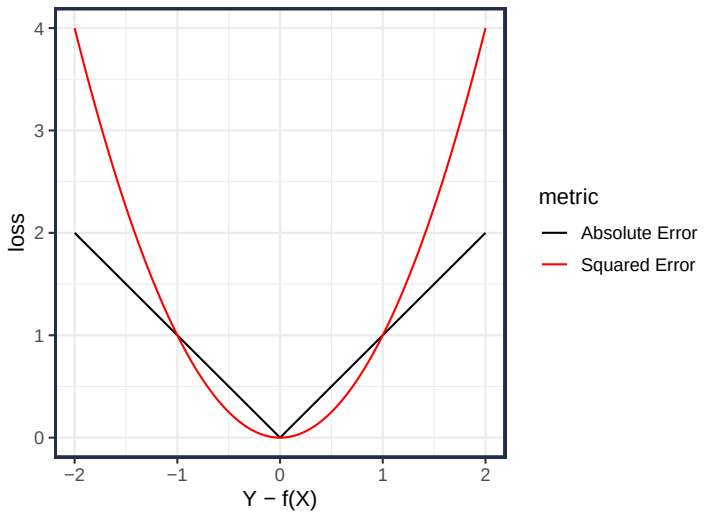
Our goal is prediction, so we should evaluate the models on their *predictive performance*.

- We need to use hold-out data (i.e., data not used to fit the model) to evaluate how well our models do in prediction
- Call these data *test data* $D_{\text{test}} = \{(X_i, Y_i)\}_{i=1}^{M}$
  - Important: the test data must come from the same distribution as the training data
  - Or $P_{\text{test}}(X, Y) = P_{\text{train}}(X, Y)$
  - **both** $Y$ and $X$ from same distribution
- Later in the course we will cover ways to do this when we only have training data (e.g., cross-validation)
- But for today, we have an unlimited amount of *test data* at our disposal (since we know how the data were generated)

## 6.2 Statistical Decision Theory

- In a prediction context, we want a *point estimate* for the value of an unobserved r.v. $Y \in \mathbb{R}$ given an input feature $X \in \mathbb{R}$.
- Let $f(X)$ be the prediction of $Y$ given $X$.
- Define a *loss function* $L(Y, f(X))$ that indicates how bad it is if we estimate the value $Y$ by $f(X)$
  - E.g. $Y$ is the number of customers complaints in a call center and $X$ is the day of week
  - If we guess $f(X) = 500$, but there are really $Y = 2000$, how bad would that be?
- Two common loss functions are *squared error* and *absolute error*

$$L_{\text{sq}}(Y, f(X)) = (Y - f(X))^2$$
$$L_{\text{abs}}(Y, f(X)) = |Y - f(X)|$$



- The best model is the one that minimizes the *expected loss* or Risk or Expected Prediction Error (EPE)

$$\text{Risk} = \text{EPE} = E[\text{loss}]$$

- For *squared error*, the *risk* for using the model $f$ is:

$$\begin{aligned} R(f) &= E_{XY}[L(Y, f(X))] \\ &= E_{XY}[(Y - f(X))^2] \end{aligned}$$

where the expectation is w.r.t. the *test values* of $X, Y$.

**Mean Squared Error (MSE)**

Under squared error loss, the risk/EPE is also known as the *mean squared error* (MSE)

- To simplify a bit, let's examine the EPE of model $f$ at a given fixed input $X = x$. This removes the uncertainty in $X$, so we only have uncertainty coming from $Y$.

$$\begin{aligned} \text{EPE}_x(f) &= E[L(Y, f(x)) \mid X = x] \\ &= \text{E}[(Y - f(x))^2 \mid X = x] \qquad \text{for squared error loss} \end{aligned}$$

where the expectation is taken with respect to $Y|X = x$

- The best prediction $f^*(x)$, given $X = x$, is the value that minimizes the risk

$$\begin{aligned} f^*(x) &= \arg\min_c \text{EPE}_x(c) \\ &= \arg\min_c \text{E}[(Y - c)^2 \mid X = x] \end{aligned}$$

**Your Turn #5**

What is the optimal prediction at $X = x$ under the squared error loss?

- I.e., find $f^*(x)$.

### 6.2.1 Squared Error Loss Functions

- **Conclusion:** If quality of prediction is measured by squared error, then the best predictor is the (conditional) expected value $f^*(x) = \text{E}[Y|X = x]$.
  - And the minimum Risk/MSE is $\text{EPE}_x(f^*) = \text{V}[Y|X = x]$
- **Summary:** Under *squared error loss* the Risk (at input $x$) is

$$\begin{aligned} \text{EPE}_x(f) &= \text{E}_Y[L(Y, f(X)) \mid X = x] \\ &= \text{E}_Y[(Y - f(x))^2 \mid X = x] \qquad \text{using squared error loss} \\ &= \text{V}[Y \mid X = x] + (E_Y[Y \mid X = x] - f(x))^2 \\ &= \text{Irreducible Variance} + \text{model squared error} \end{aligned}$$

### 6.2.2 kNN and Polynomial Regression

- The kNN model estimates the conditional expectation by using the data in a *local region* around $x$

$$\hat{f}_{\text{knn}}(x; k) = \text{Ave}(y_i \mid x_i \in N_k(x))$$

  This assumes that the true $f(x)$ can be well approximated by a *locally constant* function

- Polynomial (linear) regression, on the other hand, assumes that the true $f(x)$ is well approximated by a *globally polynomial* function

$$\hat{f}_{\text{poly}}(x; d) = \beta_0 + \sum_{j=1}^{d} \beta_j x^j$$

### 6.2.3 Empirical Risk

- The actual Risk/EPE is based on the expected error from *test data* (out-of-sample), or data that was not used to estimate $\hat{f}$

$$\begin{aligned} \text{EPE}(f) &= E_{XY}[L(Y, f(X))] \\ &= E_{XY}[(Y - f(X))^2] \qquad \text{for squared error loss} \end{aligned}$$

  where $X, Y$ are from $\Pr(X, Y)$ (i.e., test data)

- But is it a bad idea to choose the best model according to *empirical risk* or *training error*?

$$\begin{aligned} \text{Train error}(f) &= \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \\ &= \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \qquad \text{for squared error loss} \end{aligned}$$

### 6.3 Choose the best *predictive* model

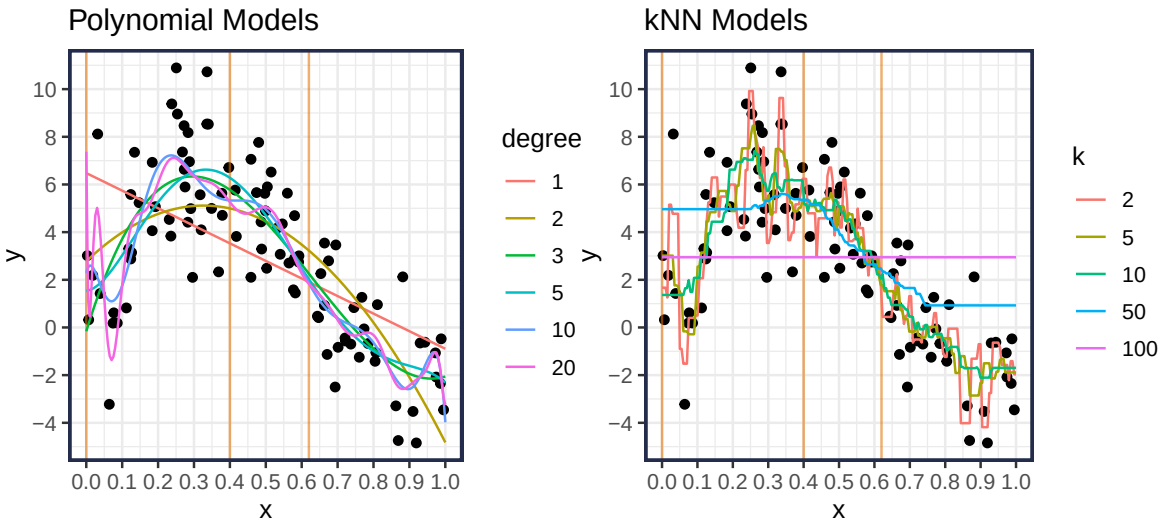



Figure 2: Polynomial

Figure 3: kNN

| degree | MSE | # pars |
|---|---|---|
| 1 | 8.29 | 2 |
| 2 | 5.58 | 3 |
| 3 | 4.28 | 4 |
| 5 | 4.10 | 6 |
| 10 | 3.65 | 11 |
| 20 | 3.16 | 21 |

| k | MSE | edf |
|---|---|---|
| 50 | 6.87 | 2.00 |
| 30 | 5.06 | 3.33 |
| 20 | 4.18 | 5.00 |
| 15 | 4.13 | 6.67 |
| 10 | 3.86 | 10.00 |
| 5 | 3.16 | 20.00 |

# 7 Evaluate Models on Simuated Test Data

I simulated 50,000 test observations, evaluated the predictions from each model, and recorded the estimated MSE/Risk.

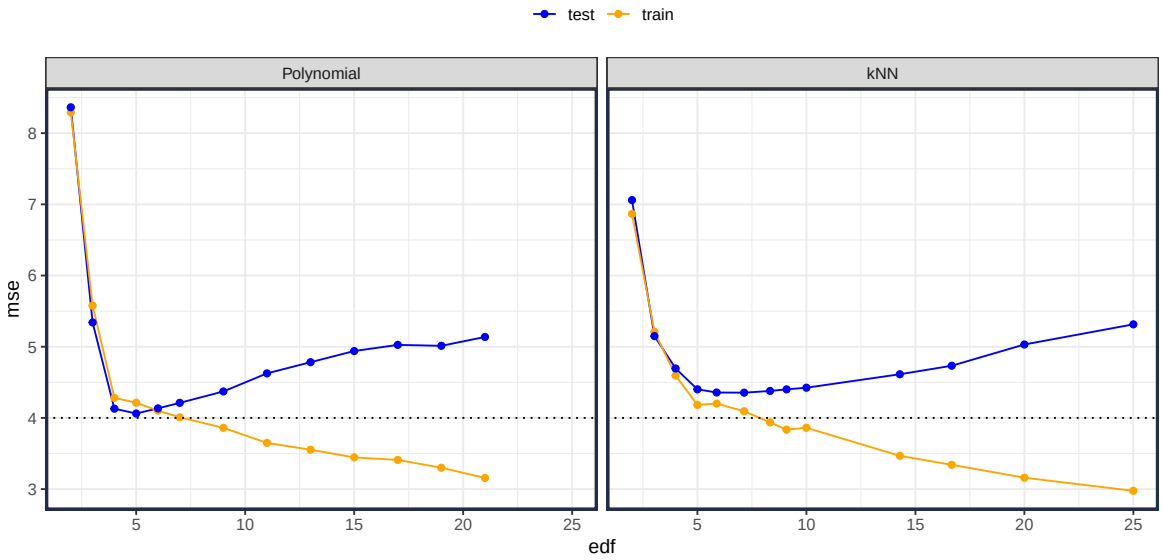



Figure 4: Polynomial

Figure 5: kNN

| degree | edf | mse_train | mse_test |
|---|---|---|---|
| 1 | 2 | 8.29 | 8.36 |
| 2 | 3 | 5.58 | 5.34 |
| 3 | 4 | 4.28 | 4.13 |
| 4 | 5 | 4.21 | 4.06 |
| 5 | 6 | 4.10 | 4.13 |
| 6 | 7 | 4.01 | 4.21 |
| 8 | 9 | 3.86 | 4.37 |
| 10 | 11 | 3.65 | 4.63 |
| 12 | 13 | 3.55 | 4.78 |
| 14 | 15 | 3.44 | 4.94 |
| 16 | 17 | 3.41 | 5.03 |
| 18 | 19 | 3.30 | 5.01 |
| 20 | 21 | 3.16 | 5.14 |

| k | edf | mse_train | mse_test |
|---|---|---|---|
| 50 | 2.00 | 6.87 | 7.06 |
| 33 | 3.03 | 5.21 | 5.15 |
| 25 | 4.00 | 4.59 | 4.69 |
| 20 | 5.00 | 4.18 | 4.40 |
| 17 | 5.88 | 4.20 | 4.36 |
| 14 | 7.14 | 4.09 | 4.35 |
| 12 | 8.33 | 3.94 | 4.38 |
| 11 | 9.09 | 3.84 | 4.40 |
| 10 | 10.00 | 3.86 | 4.42 |
| 7 | 14.29 | 3.47 | 4.61 |
| 6 | 16.67 | 3.34 | 4.73 |
| 5 | 20.00 | 3.16 | 5.03 |
| 4 | 25.00 | 2.98 | 5.31 |

**Observations:**

- as the flexibility increases, both classes of models *overfit*.
  - overfit means model too flexible
  - underfit means model is not flexible enough
  - see discrepancy between training and test performance
- The polynomial with degree = 4 has the best test performance with an approximate MSE = 4.06.

- The optimal MSE $= 4$.
  - I only know this because I know how the data was generated