McCOMBS SCHOOL OF BUSINESS
Salem Center for Policy

Bias-variance tradeoff

David Puelz

November 9, 2021

Prediction

Let's go back to supervised learning aka prediction.

There was a lingering problem of which subset of variables I use for my regression model. It is closely related to model selection, and we will cover important ideas related to it here.

Remember our supervised learning goal

Predict a target variable Y with input variables X.

Remember our supervised learning goal

Predict a target variable Y with input variables X.

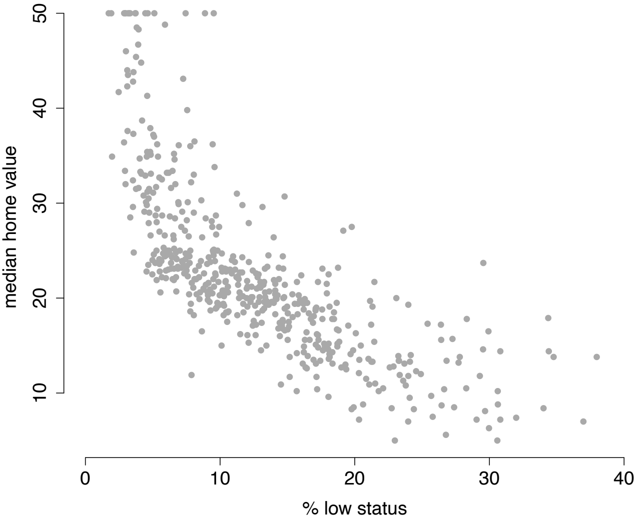
We can frame the problem by supposing Y and X are related in the following way:

$$Y_i = f(X_i) + \epsilon_i$$

To achieve our goal, we need to: *Learn or estimate* $f(\cdot)$ from data.

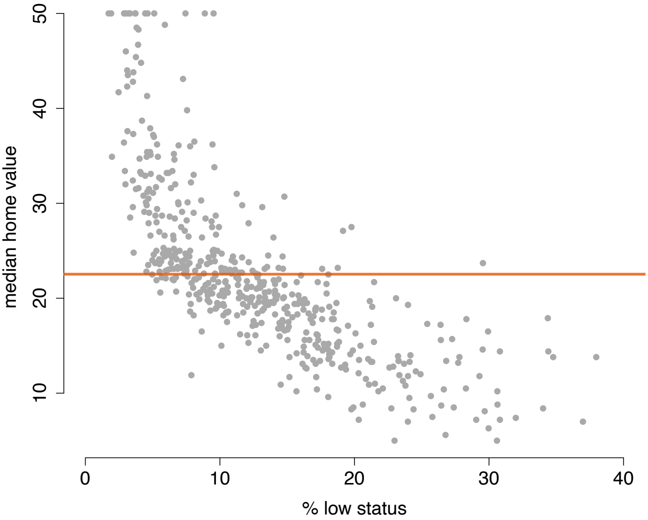
Boston housing data

Predict median home value with percent low economic status.



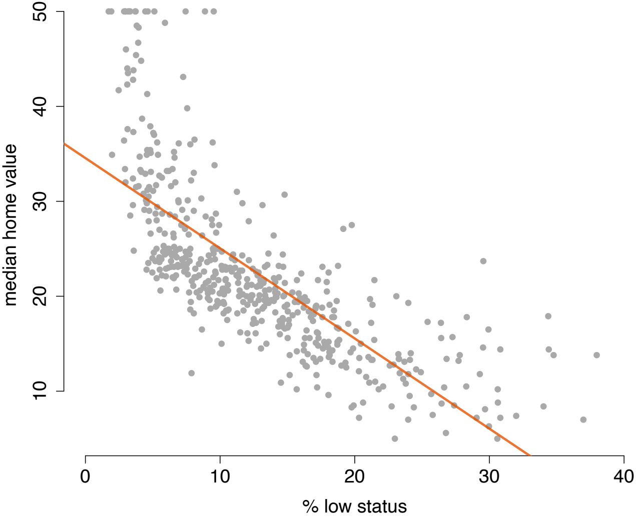
Boston housing data

Prediction at % low status = 30?



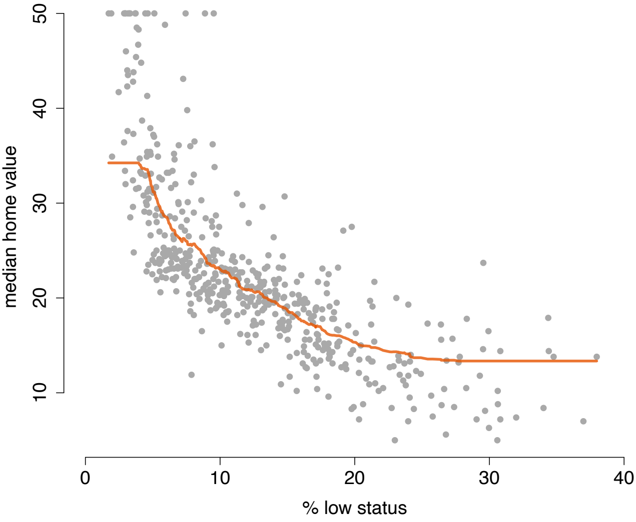
Boston housing data

Prediction at % low status = 30?



Boston housing data

Prediction at % low status = 30?



How do we estimate $f(\cdot)$?

1. Choose set of training data: $(Y_1, X_1), \ldots, (Y_N, X_N)$.

How do we estimate $f(\cdot)$?

1. Choose set of training data: $(Y_1, X_1), \ldots, (Y_N, X_N)$.

2. Fit $f(\cdot)$ to training data using:

   - Parametric model, or

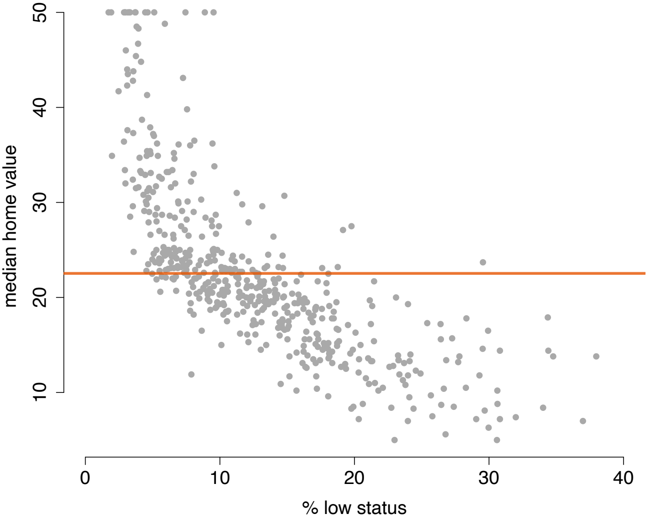
   - Nonparametric model

How do we estimate $f(\cdot)$?

1. Choose set of training data: $(Y_1, X_1), \ldots, (Y_N, X_N)$.
2. Fit $f(\cdot)$ to training data using:
   - Parametric model, or
   - Nonparametric model
3. Evaluate performance on testing data and *adjust*.

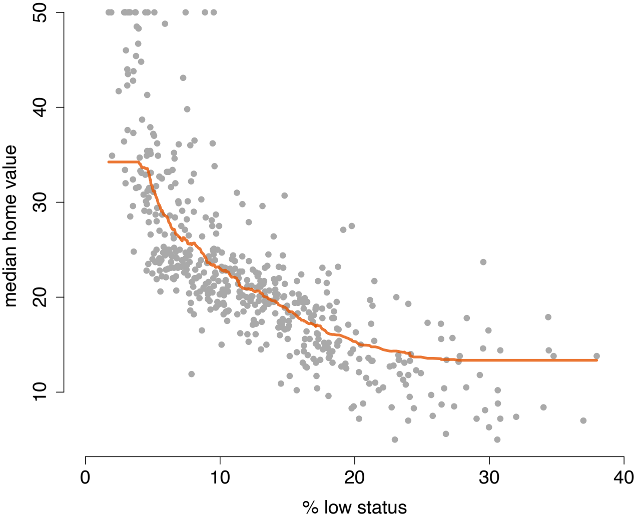
How do we estimate $f(\cdot)$?

Parametric: $Y = \mu + \epsilon$. restrictive assumptions, but simple interpretation.



How do we estimate $f(\cdot)$?

Nonparametric: "Knn" with $k = 100$. flexible assumptions, but complex interpretation.



The challenge when estimating predictions $\widehat{f(\cdot)}$

Balancing *restrictiveness* of assumptions with simplicity of *interpretation*.

Let's look at k-nearest-neighbors (knn)

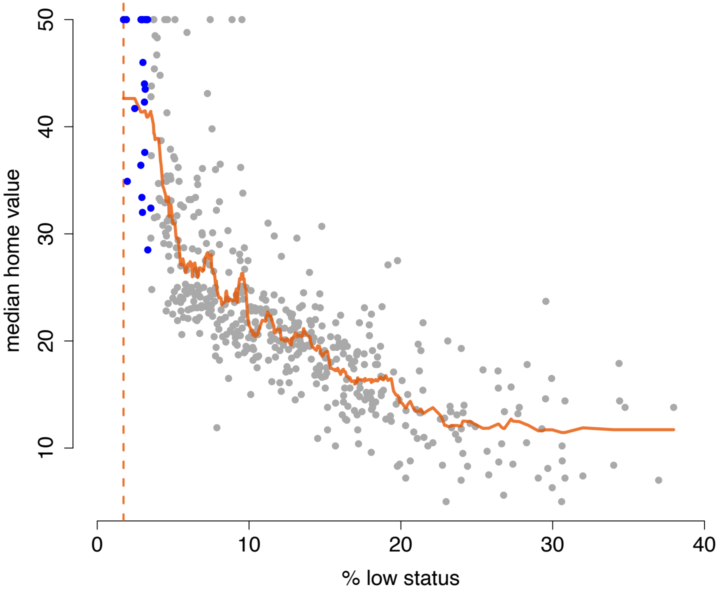
Prediction at point x, $\widehat{f(x)}$ = average of k nearest points around x.

Let's look at k-nearest-neighbors (knn)

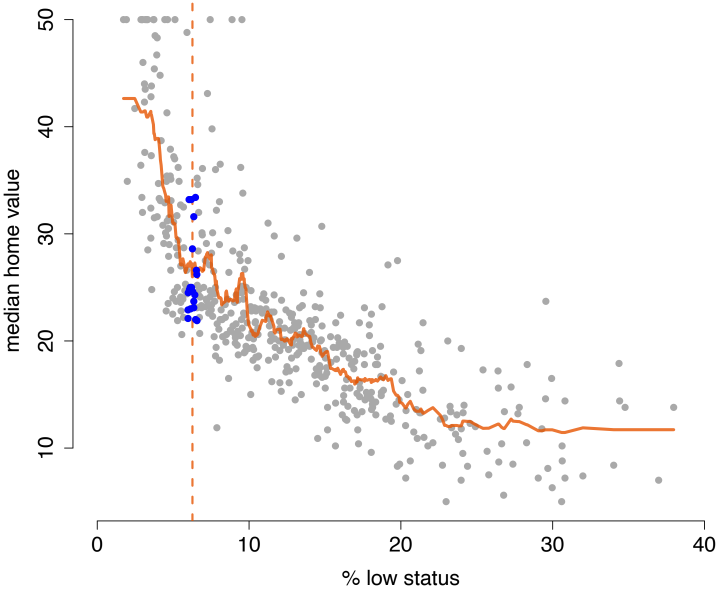
Prediction at point x, $\widehat{f(x)}$ = average of k nearest points around x.

Let's look at $k = 20$ ...

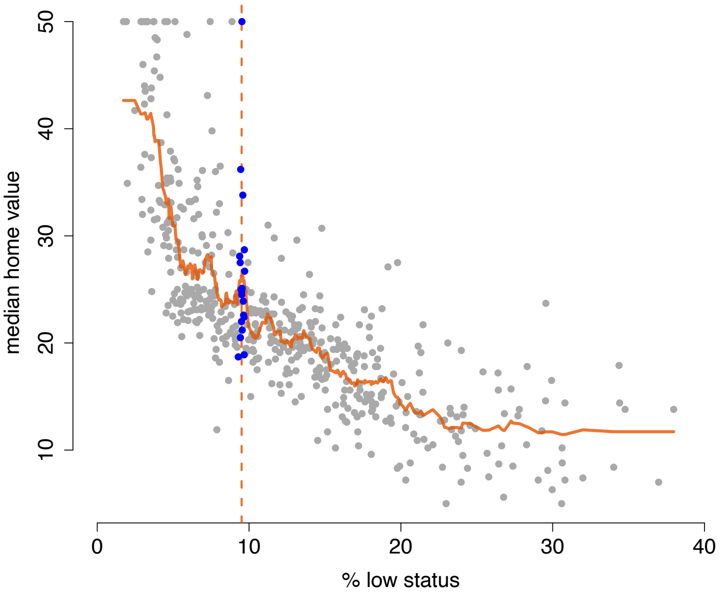
knn with $k = 20$



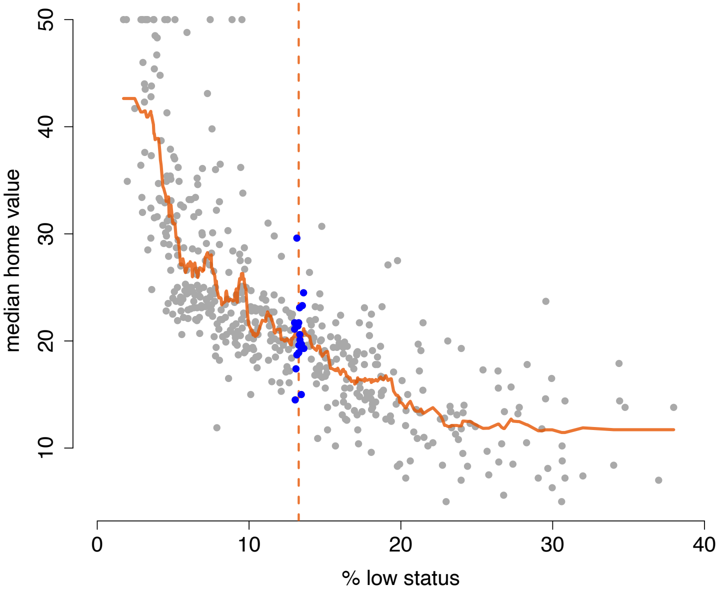
knn with $k = 20$



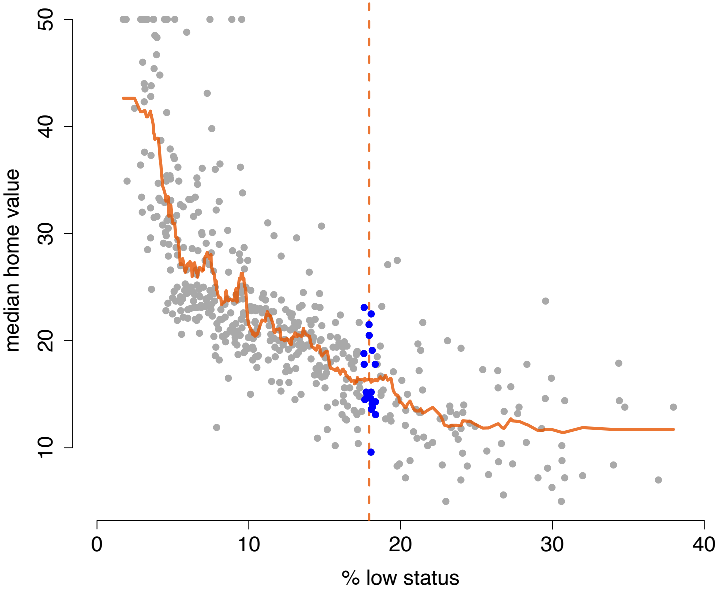
knn with $k = 20$



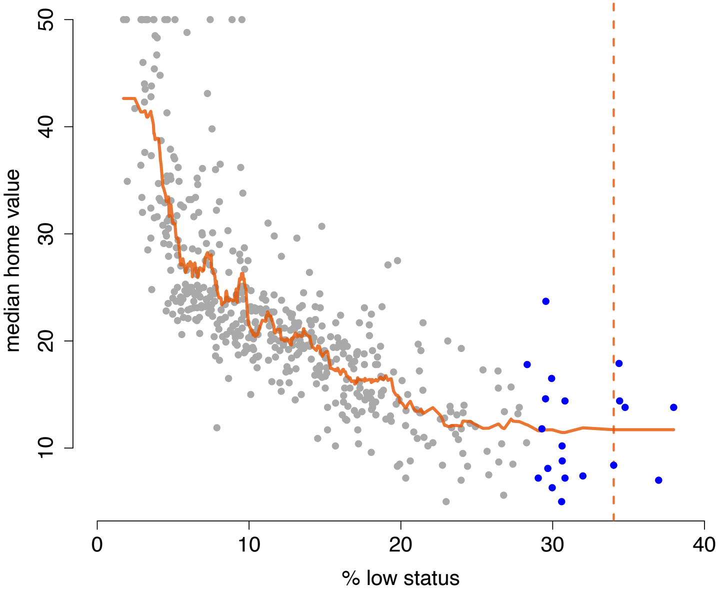
knn with $k = 20$



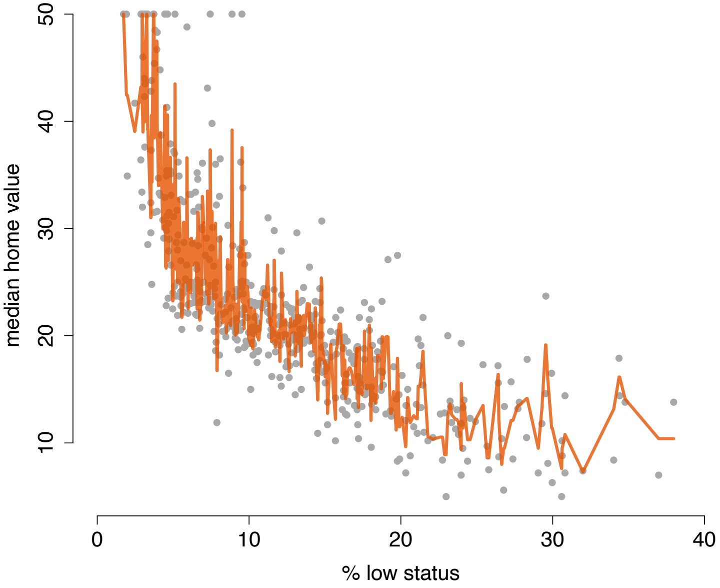
knn with $k = 20$



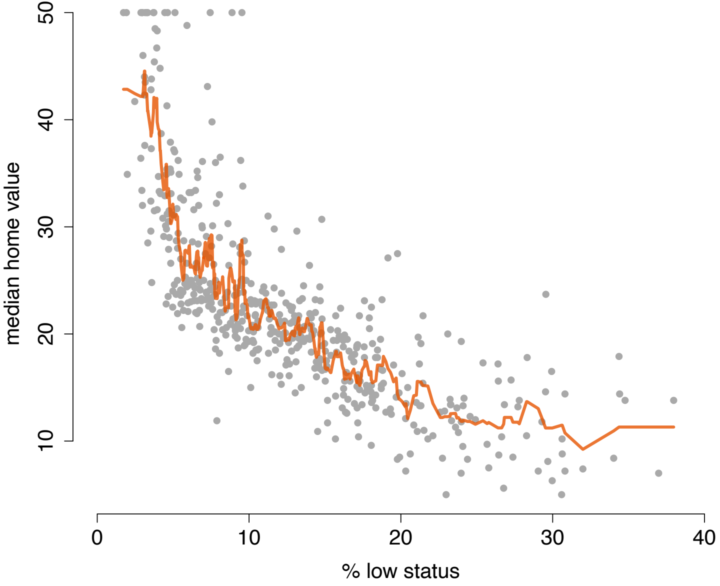
knn with $k = 20$



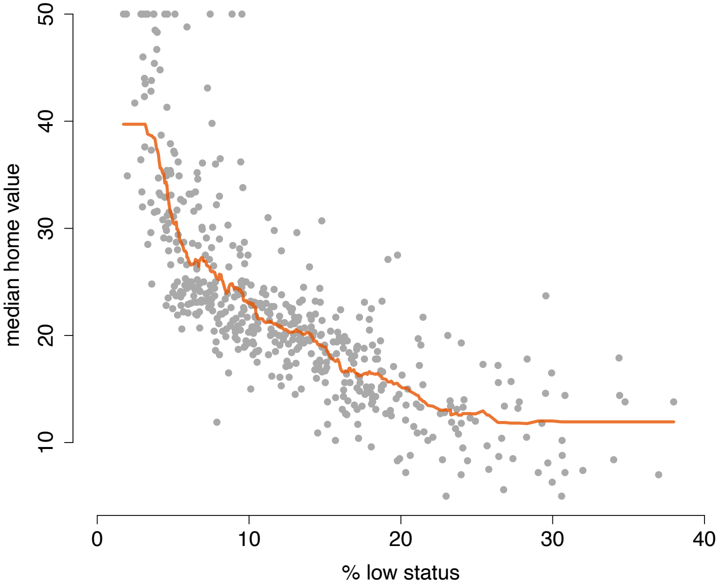
Why don't I choose $k = 2$ instead?



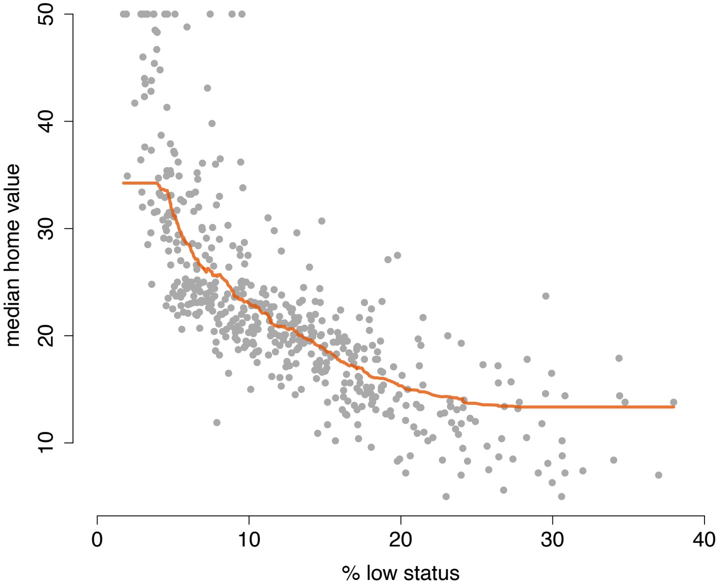
or $k = 10$ ...



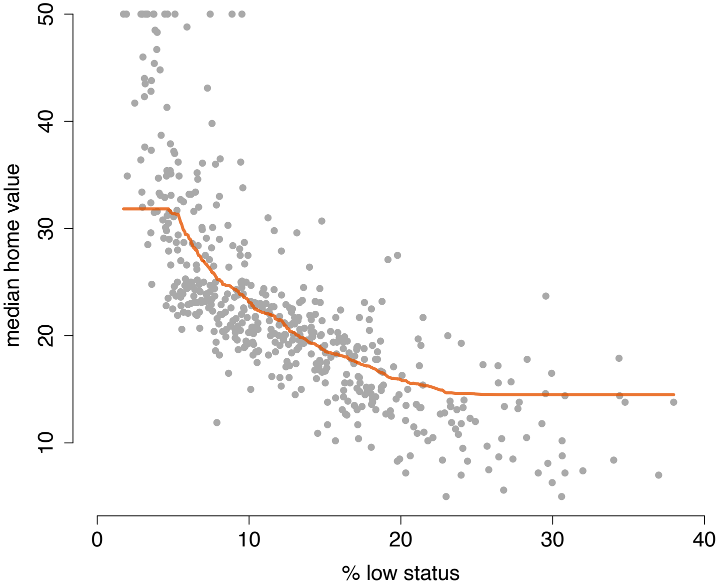
or $k = 50$ …



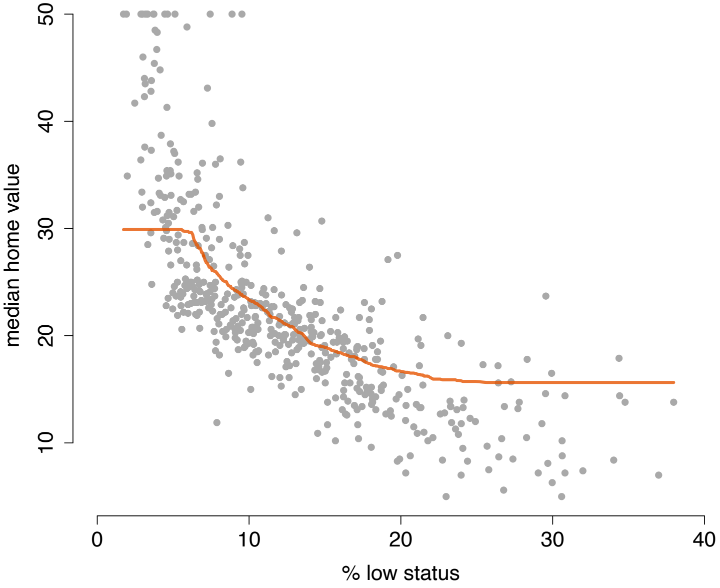
or $k = 100$ ...



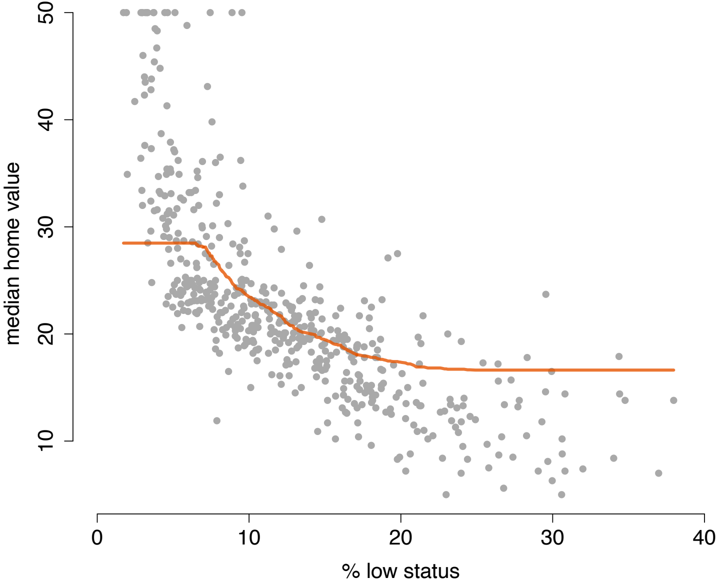
or $k = 150$ ...



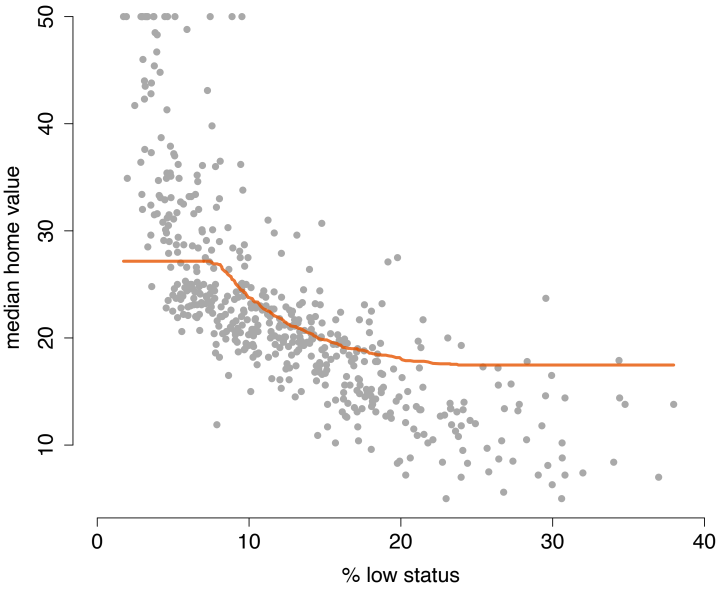
or $k = 200$ …



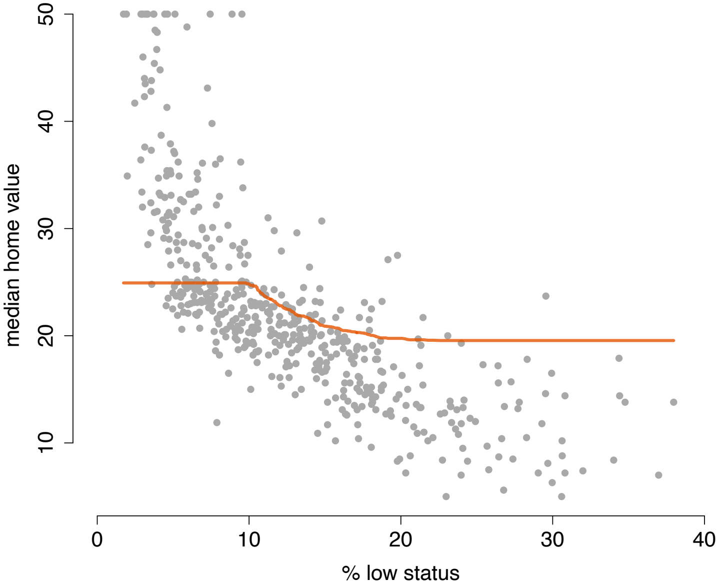
or $k = 250$ ...



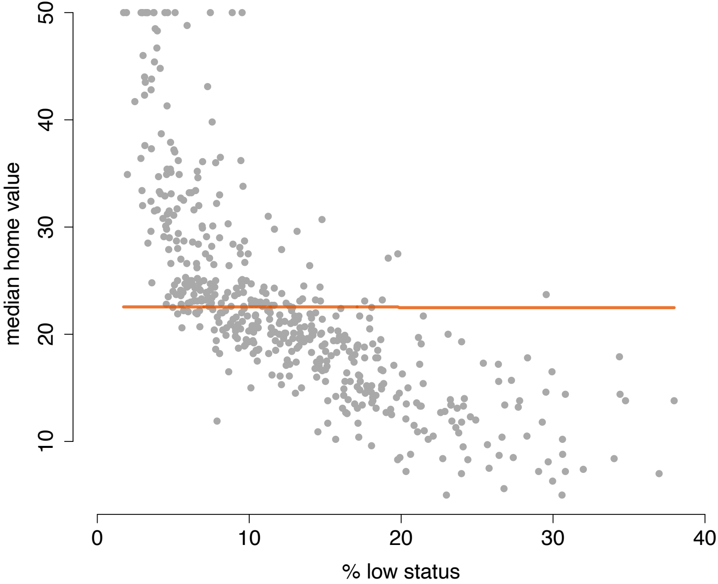
Or $k = 300$ ...



or $k = 400$ ...



or $k = 505$ …



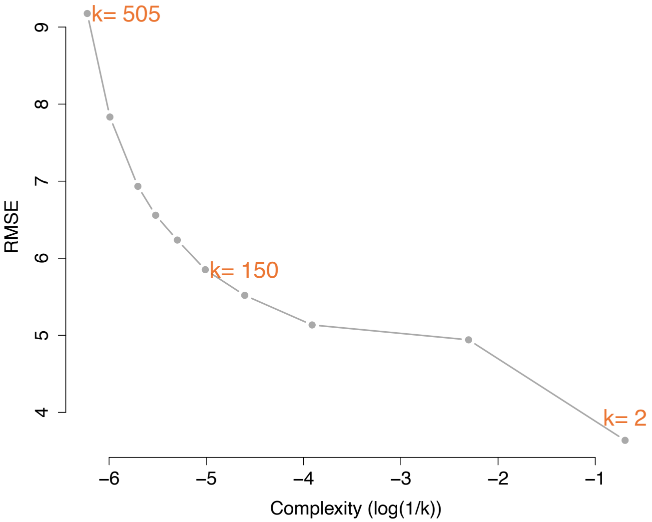
A rigorous way to select

– The root mean squared error measures how accurate my predictions are, on average.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \widehat{f(X_i)}\right]^2}$$

In sample RMSE

It looks like $k = 2$ is the best. Should we choose this model?



We care about out of sample performance

- Suppose we have m additional observations (X_i^o, Y_i^o), for $i = 1, \ldots, m$, that we did not use to fit the model. Let's call this dataset the *validation set* (a.k.a *hold-out set* or *test set*)

We care about out of sample performance

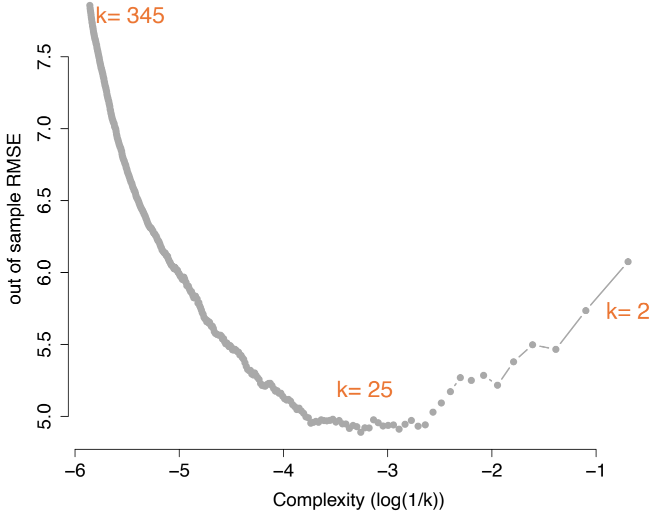
- Suppose we have m additional observations (X_i^o, Y_i^o), for $i = 1, \ldots, m$, that we did not use to fit the model. Let's call this dataset the *validation set* (a.k.a *hold-out set* or *test set*)

- We evaluate the fit with out of sample RMSE:

$$RMSE^o = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left[Y_i^o - \widehat{f(X_i^o)} \right]^2}$$

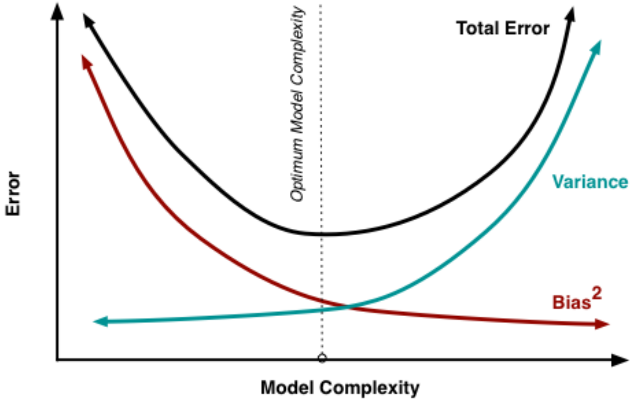
Out of sample RMSE

Fit each model on training set of size 400. Test each model (*out of sample*) on testing set of size 106. Here, we plot the out of sample performance.



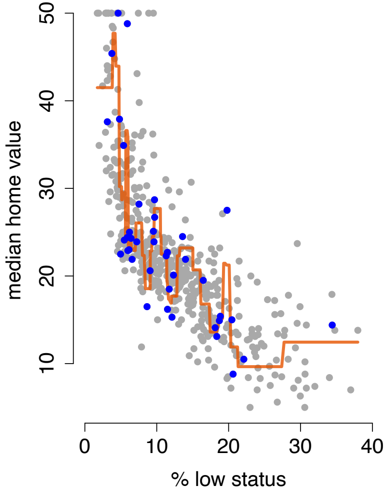
The Bias-variance tradeoff!

When fitting a predictive model, there is a tradeoff between bias and variance of predictions.



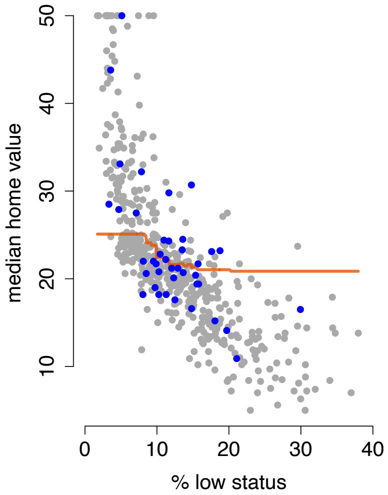
$k = 2$: low bias, high variance

Training set of size 40.



$k = 25$: high bias, low variance

Training set of size 40.



Relationship to linear regression

Selecting k is Knn is the same as selecting which variables to include in your regression model!

In both cases, you are trying to build the best model for your outcome Y.

Questions that remain unanswered:

$\rightarrow$ How does model selection relate to causal inference?

$\rightarrow$ More directly, how can we use the best ideas from machine learning to help us automatically **control for** the variables we need in our model?