MS&E 226: Fundamentals of Data Science

Lecture 6: Bias and variance

Ramesh Johari

Our plan today

In general, when creating predictive models, we are trading off two different goals:

- On one hand, we want models to fit the training data well.
- On the other hand, we want to avoid models that become so finely tuned to the training data that they perform poorly on new data.

In this lecture we develop a systematic vocabulary for talking about this kind of trade off, through the notions of *bias* and *variance*.

Conditional expectation

Conditional expectation

Given the population model for $\vec{X}$ and Y, suppose we are allowed to choose *any* predictive model $\hat{f}$ we want. What is the best one?

$$\text{minimize } \mathbb{E}[(Y - \hat{f}(\vec{X}))^2].$$

(Here expectation is over $(\vec{X}, Y)$.)

Theorem

The predictive model that minmizes squared error is $\hat{f}(\vec{X}) = \mathbb{E}[Y|\vec{X}]$.

Conditional expectation

Proof:

$$\begin{aligned}
&\mathbb{E}[(Y - \hat{f}(\vec{X}))^2] \\
&\quad = \mathbb{E}[(Y - \mathbb{E}[Y|\vec{X}] + \mathbb{E}[Y|\vec{X}] - \hat{f}(\vec{X}))^2] \\
&\quad = \mathbb{E}[(Y - \mathbb{E}[Y|\vec{X}])^2] + \mathbb{E}[(\mathbb{E}[Y|\vec{X}] - \hat{f}(\vec{X}))^2] \\
&\qquad + 2\mathbb{E}[(Y - \mathbb{E}[Y|\vec{X}])(\mathbb{E}[Y|\vec{X}] - \hat{f}(\vec{X}))].
\end{aligned}$$

The first two terms are positive, and minimized if $\hat{f}(\vec{X}) = \mathbb{E}[Y|\vec{X}]$. For the third term, using the tower property of conditional expectation:

$$\begin{aligned}
&\mathbb{E}[(Y - \mathbb{E}[Y|\vec{X}])(\mathbb{E}[Y|\vec{X}] - \hat{f}(\vec{X}))] \\
&\quad = \mathbb{E}\Big[\mathbb{E}[Y - \mathbb{E}[Y|\vec{X}]|\vec{X}]\ (\mathbb{E}[Y|\vec{X}] - \hat{f}(\vec{X}))\Big] = 0.
\end{aligned}$$

So the squared error minimizing solution is to choose $\hat{f}(\vec{X}) = \mathbb{E}[Y|\vec{X}]$.

Conditional expectation

Why don't we just choose $\mathbb{E}[Y|\vec{X}]$ as our predictive model?

Because we don't know the distribution of $(\vec{X}, Y)$*!*

In other words: *We don't know the population model.*

Nevertheless, the preceding result is a useful guide:

- It provides the benchmark that every squared-error-minimizing predictive model is striving for: approximate the conditional expectation.
- It provides intuition for why linear regression approximates the conditional mean.

Population model

For the rest of the lecture write:

$$Y = f(\vec{X}) + \varepsilon,$$

where $\mathbb{E}[\varepsilon|\vec{X}] = 0$. In other words, $f(\vec{X}) = \mathbb{E}[Y|\vec{X}]$.

We make the assumption that $\mathrm{Var}(\varepsilon|\vec{X}) = \sigma^2$ (i.e., it does not depend on $\vec{X}$).

We will make additional assumptions about the population model as we go along.

Prediction error revisited

A note on prediction error and conditioning

When you see "prediction error", it typically means:

$$\mathbb{E}_{\vec{X},Y}[(Y - \hat{f}(\vec{X}))^2 | (*)],$$

where $(*)$ can be one of many things:

- $\mathbf{X}, \mathbf{Y}, \vec{X}$;
- $\mathbf{X}, \mathbf{Y}$;
- $\mathbf{X}, \vec{X}$;
- $\mathbf{X}$;
- nothing.

As long we don't condition on both $\mathbf{X}$ and $\mathbf{Y}$, the model is random!

Models and conditional expectation

So now suppose we have data $\mathbf{X}, \mathbf{Y}$, and use it to build a model $\hat{f}$.

What is the prediction error if we see a new $\vec{X}$?

$$\begin{aligned}
&\mathbb{E}[(Y - \hat{f}(\vec{X}))^2 | \mathbf{X}, \mathbf{Y}, \vec{X}] \\
&= \mathbb{E}[(Y - f(\vec{X}))^2 | \vec{X}] \\
&\quad + (\hat{f}(\vec{X}) - f(\vec{X}))^2 \\
&= \sigma^2 + (\hat{f}(\vec{X}) - f(\vec{X}))^2.
\end{aligned}$$

$Y = f(\vec{X}) + \varepsilon$

$\mathbb{E}[\varepsilon | \vec{X}] = 0$

$\mathrm{Var}(\varepsilon | \vec{X}) = \sigma^2$

$\mathbb{E}[Y | \vec{X}] = \mathbb{E}[f(\vec{X}) + \varepsilon | \vec{X}] = f(\vec{X})$

I.e.: *When minimizing mean squared error, "good" models should behave like conditional expectation.*[1]

Our goal: understand the second term.

[1] This is just another way of deriving that the prediction-error-minimizing solution is the conditional expectation.

Models and conditional expectation [$*$]

Proof of preceding statement:
The proof is essentially identical to the earlier proof for conditional expectation:

$$\begin{aligned}
&\mathbb{E}[(Y - \hat{f}(\vec{X}))^2 | \mathbf{X}, \mathbf{Y}, \vec{X}] \\
&\quad = \mathbb{E}[(Y - f(\vec{X}) + f(\vec{X}) - \hat{f}(\vec{X}))^2 | \mathbf{X}, \mathbf{Y}, \vec{X}] \\
&\quad = \mathbb{E}[(Y - f(\vec{X}))^2 | \vec{X}] + (\hat{f}(\vec{X}) - f(\vec{X}))^2 \\
&\qquad + 2\mathbb{E}[Y - f(\vec{X}) | \vec{X}](f(\vec{X}) - \hat{f}(\vec{X})) \\
&\quad = \mathbb{E}[(Y - f(\vec{X}))^2 | \vec{X}] + (\hat{f}(\vec{X}) - f(\vec{X}))^2,
\end{aligned}$$

because $\mathbb{E}[Y - f(\vec{X}) | \vec{X}] = \mathbb{E}[Y - \mathbb{E}[Y | \vec{X}] | \vec{X}] = 0$.

A thought experiment

Our goal is to understand:

$$(\hat{f}(\vec{X}) - f(\vec{X}))^2. \qquad (**)$$

Here is one way we might think about the quality of our modeling approach:

- Fix the design matrix $\mathbf{X}$.
- Generate data $\mathbf{Y}$ many times (parallel "universes").
- In each universe, create a $\hat{f}$.
- In each universe, evaluate $(**)$.

A thought experiment

Our goal is to understand:

$$(\hat{f}(\vec{X}) - f(\vec{X}))^2. \qquad (**)$$

What kinds of things can we evaluate?

- If $\hat{f}(\vec{X})$ is "close" to the conditional expectation, then on average in our universes, it should look like $f(\vec{X})$.
- $\hat{f}(\vec{X})$ might be close to $f(\vec{X})$ on average, but still vary wildly across our universes.

The first is *bias*. The second is *variance*.

Example

Let's carry out some simulations with a synthetic model.

Population model:

- We generate X_1, X_2 as i.i.d. $N(0,1)$ random variables.
- Given X_1, X_2, the distribution of Y is given by:

$$Y = 1 + X_1 + 2X_2 + \varepsilon,$$

where ε is an independent $N(0,1)$ random variable.

Thus $f(X_1, X_2) = 1 + X_1 + 2X_2$.

Example: Parallel universes

Generate a design matrix $\mathbf{X}$ by sampling 100 i.i.d. values of (X_1, X_2).

Now we run $m = 500$ simulations. These are our "universes." In each simulation, generate data $\mathbf{Y}$ according to:

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i,$$

where ε_i are i.i.d. $N(0, 1)$ random variables.

In each simulation, what changes is the specific values of the ε_i*.* This is what it means to condition on $\mathbf{X}$.

Example: OLS in parallel universes

In each simulation, given the design matrix $\mathbf{X}$ and $\mathbf{Y}$, we build a fitted model $\hat{f}$ using ordinary least squares.

Finally, we use our models to make predicitions at a new covariate vector: $\vec{X} = (1, 1)$.
In each universe we evaluate:

$$\text{error} = f(\vec{X}) - \hat{f}(\vec{X}).$$

We then create a density plot of these errors across the 500 universes.

Example: Results

We go through the process with three models: A, B, C.

The three we try are:

Red
- `Y ~ 1 + X1.`

Green
- `Y ~ 1 + X1 + X2.`

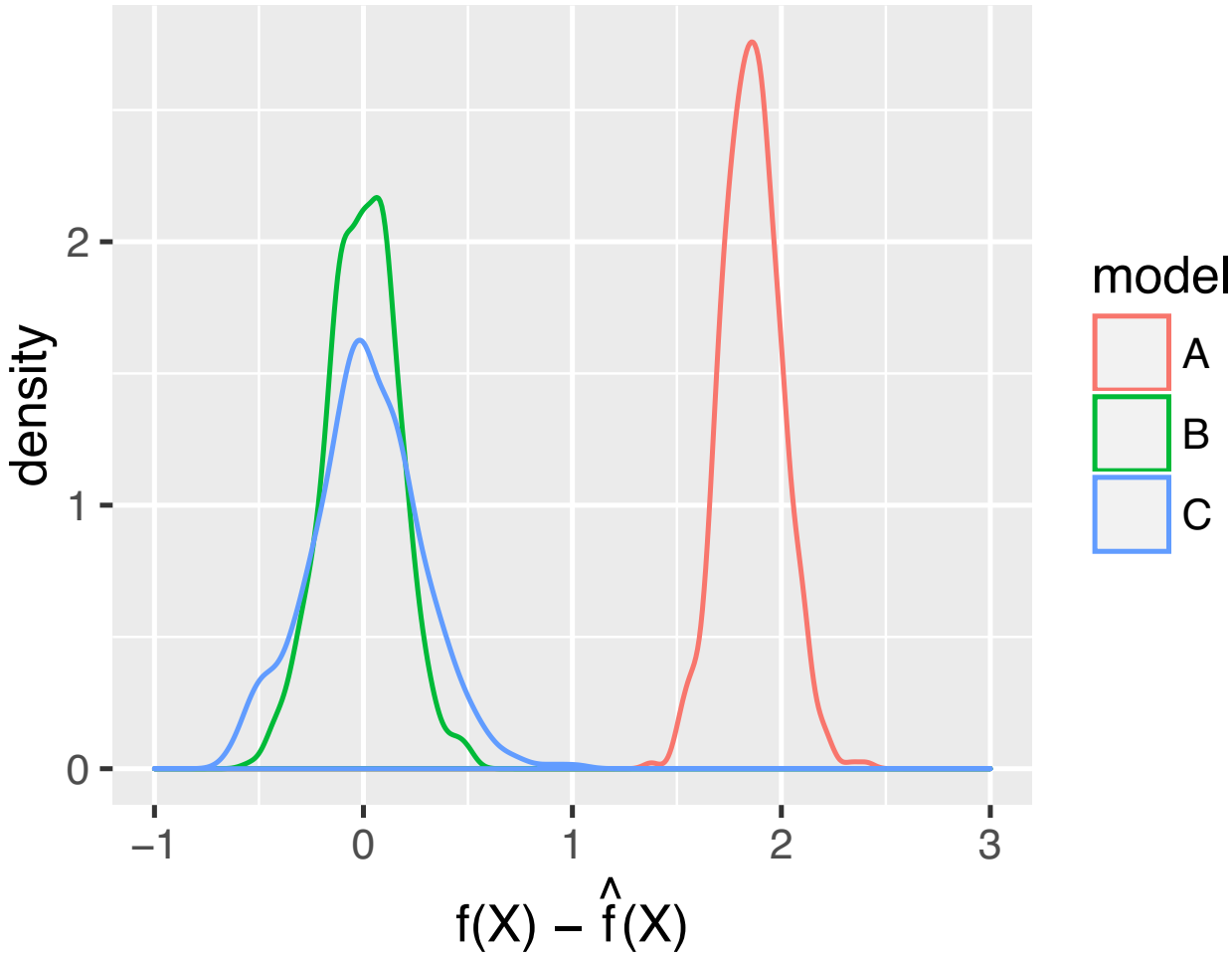
Blue
- `Y ~ 1 + X1 + X2 + ... + I(X1^4) + I(X2^4).`

Which is which?

Example: Results

$X = (1, 1)$

Results:



A thought experiment: Aside

This is our first example of a *frequentist* approach:

- The population model is *fixed*.
- The data is *random*.
- We reason about a particular modeling procedure by considering what happens if we carry out the same procedure over and over again (in this case, fitting a model from data).

Examples: Bias and variance

Suppose you are predicting, e.g., wealth based on a collection of demographic covariates.

- Suppose we make a constant prediction: $\hat{f}(\mathbf{X}_i) = c$ for all i. Is this biased? Does it have low variance?
- Suppose that every time you get your data, you use enough parameters to fit $\mathbf{Y}$ *exactly*: $\hat{f}(\mathbf{X}_i) = Y_i$ for all i. Is this biased? Does it have low variance?

The bias-variance decomposition

We can be more precise about our discussion.

$$\begin{aligned}
&\mathbb{E}[(Y - \hat{f}(\vec{X}))^2 | \mathbf{X}, \vec{X}] \\
&\quad = \sigma^2 + \mathbb{E}[(\hat{f}(\vec{X}) - f(\vec{X}))^2 | \mathbf{X}, \vec{X}] \\
&\quad = \sigma^2 + \left(f(\vec{X}) - \mathbb{E}[\hat{f}(\vec{X}) | \mathbf{X}, \vec{X}] \right)^2 \\
&\qquad + \mathbb{E}\left[\left(\hat{f}(\vec{X}) - \mathbb{E}[\hat{f}(\vec{X}) | \mathbf{X}, \vec{X}] \right)^2 \Big| \mathbf{X}, \vec{X} \right].
\end{aligned}$$

The first term is *irreducible error*.

The second term is BIAS^2.

The third term is VARIANCE.

The bias-variance decomposition

The bias-variance decomposition measures how sensitive prediction error is to *changes in the training data* (in this case, $\mathbf{Y}$.

- If there are systematic errors in prediction made regardless of the training data, then there is high *bias*.
- If the fitted model is very sensitive to the choice of training data, then there is high *variance*.

The bias-variance decomposition: Proof [$*$]

Proof: We already showed that:

$$\mathbb{E}[(Y - \hat{f}(\vec{X}))^2 | \mathbf{X}, \mathbf{Y}, \vec{X}] = \sigma^2 + (\hat{f}(\vec{X}) - f(\vec{X}))^2.$$

Take expectations over $\mathbf{Y}$:

$$\mathbb{E}[(Y - \hat{f}(\vec{X}))^2 | \mathbf{X}, \vec{X}] = \sigma^2 + \mathbb{E}[(\hat{f}(\vec{X}) - f(\vec{X}))^2 | \mathbf{X}, \vec{X}].$$

Add and subtract $\mathbb{E}[\hat{f}(\vec{X}) | \mathbf{X}, \vec{X}]$ in the second term:

$$\begin{aligned}
&\mathbb{E}[(\hat{f}(\vec{X}) - f(\vec{X}))^2 | \mathbf{X}, \vec{X}] \\
&\quad = \mathbb{E}\Big[\Big(\hat{f}(\vec{X}) - \mathbb{E}[\hat{f}(\vec{X}) | \mathbf{X}, \vec{X}] \\
&\qquad + \mathbb{E}[\hat{f}(\vec{X}) | \mathbf{X}, \vec{X}] - f(\vec{X})\Big)^2 \Big| \mathbf{X}, \vec{X}\Big]
\end{aligned}$$

The bias-variance decomposition: Proof [$*$]

Proof (continued): Cross-multiply:

$$\begin{aligned}
&\mathbb{E}[(\hat{f}(\vec{X}) - f(\vec{X}))^2 | \mathbf{X}, \vec{X}] \\
&\quad = \mathbb{E}\Big[\Big(\hat{f}(\vec{X}) - \mathbb{E}[\hat{f}(\vec{X})|\mathbf{X}, \vec{X}]\Big)^2 | \mathbf{X}, \vec{X}\Big] \\
&\qquad + \Big(\mathbb{E}[\hat{f}(\vec{X})|\mathbf{X}, \vec{X}] - f(\vec{X})\Big)^2 \\
&\qquad + 2\mathbb{E}\Big[\Big(\hat{f}(\vec{X}) - \mathbb{E}[\hat{f}(\vec{X})|\mathbf{X}, \vec{X}]\Big)\Big(\mathbb{E}[\hat{f}(\vec{X})|\mathbf{X}, \vec{X}] - f(\vec{X})\Big)\Big|\mathbf{X}, \vec{X}\Big].
\end{aligned}$$

(The conditional expectation drops out on the middle term, because given $\mathbf{X}$ and $\vec{X}$, it is no longer random.)

The first term is the VARIANCE. The second term is the BIAS^2. We have to show the third term is zero.

The bias-variance decomposition: Proof [$*$]

Proof (continued). We have:

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\hat{f}(\vec{X}) - \mathbb{E}[\hat{f}(\vec{X})|\mathbf{X}, \vec{X}]\Big)\Big(\mathbb{E}[\hat{f}(\vec{X})|\mathbf{X}, \vec{X}] - f(\vec{X})\Big)\Big|\mathbf{X}, \vec{X}\Big] \\
&\quad = \Big(\mathbb{E}[\hat{f}(\vec{X})|\mathbf{X}, \vec{X}] - f(\vec{X})\Big)\mathbb{E}\Big[\Big(\hat{f}(\vec{X}) - \mathbb{E}[\hat{f}(\vec{X})|\mathbf{X}, \vec{X}]\Big)\Big|\mathbf{X}, \vec{X}\Big] \\
&\quad = 0,
\end{aligned}$$

using the tower property of conditional expectation.

Other forms of the bias-variance decomposition

The decomposition varies depending *on what you condition on*, but always takes a similar form.

For example, suppose that you also wish to average over the particular covariate vector $\vec{X}$ at which you make predictions; in particular suppose that $\vec{X}$ is also drawn from the population model.

Then the bias-variance decomposition can be shown to be:

$$\begin{aligned}
&\mathbb{E}[(Y - \hat{f}(\vec{X}))^2 | \mathbf{X}] \\
&\qquad = \mathrm{Var}(Y) + \Big(\mathbb{E}[f(\vec{X}) - \hat{f}(\vec{X}) | \mathbf{X}]\Big)^2 \\
&\qquad\quad + \mathbb{E}\Big[\Big(\hat{f}(\vec{X}) - \mathbb{E}[\hat{f}(\vec{X}) | \mathbf{X}]\Big)^2 \Big| \mathbf{X}\Big].
\end{aligned}$$

(In this case the first term is the irreducible error.)

Example: k-nearest-neighbor fit

Generate data the same way as before:

We generate 1000 X_1, X_2 as i.i.d. $N(0,1)$ random variables.

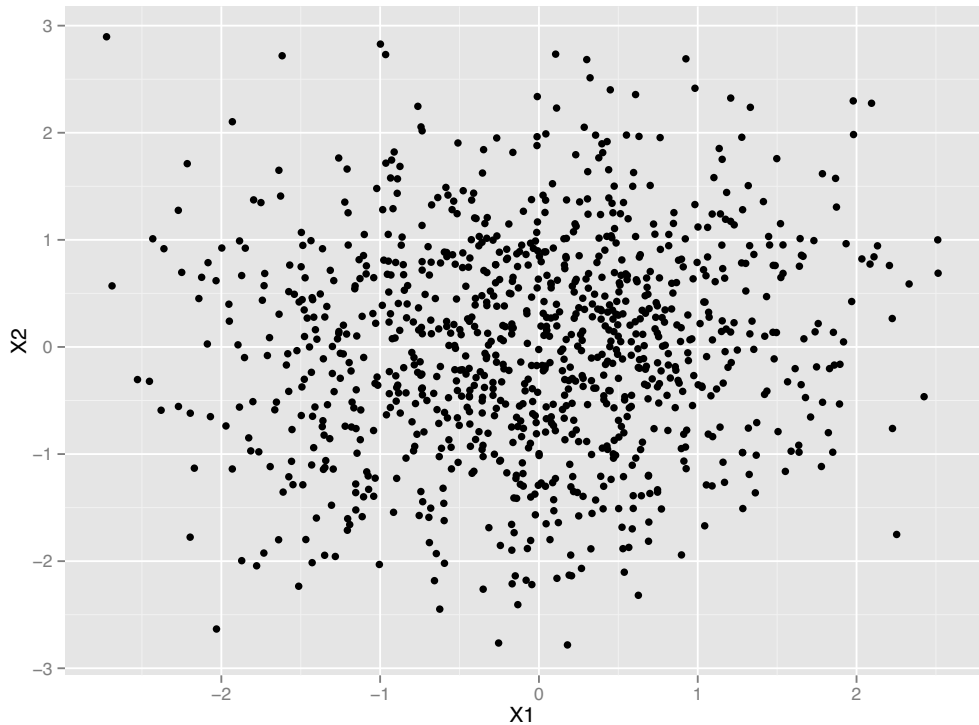
We then generate 1000 Y random variables as:

$$Y_i = \overbrace{1 + 2X_{i1} + 3X_{i2}}^{f(\vec{x})} + \varepsilon_i,$$

where ε_i are i.i.d. $N(0,5)$ random variables.

Example: k-nearest-neighbor fit

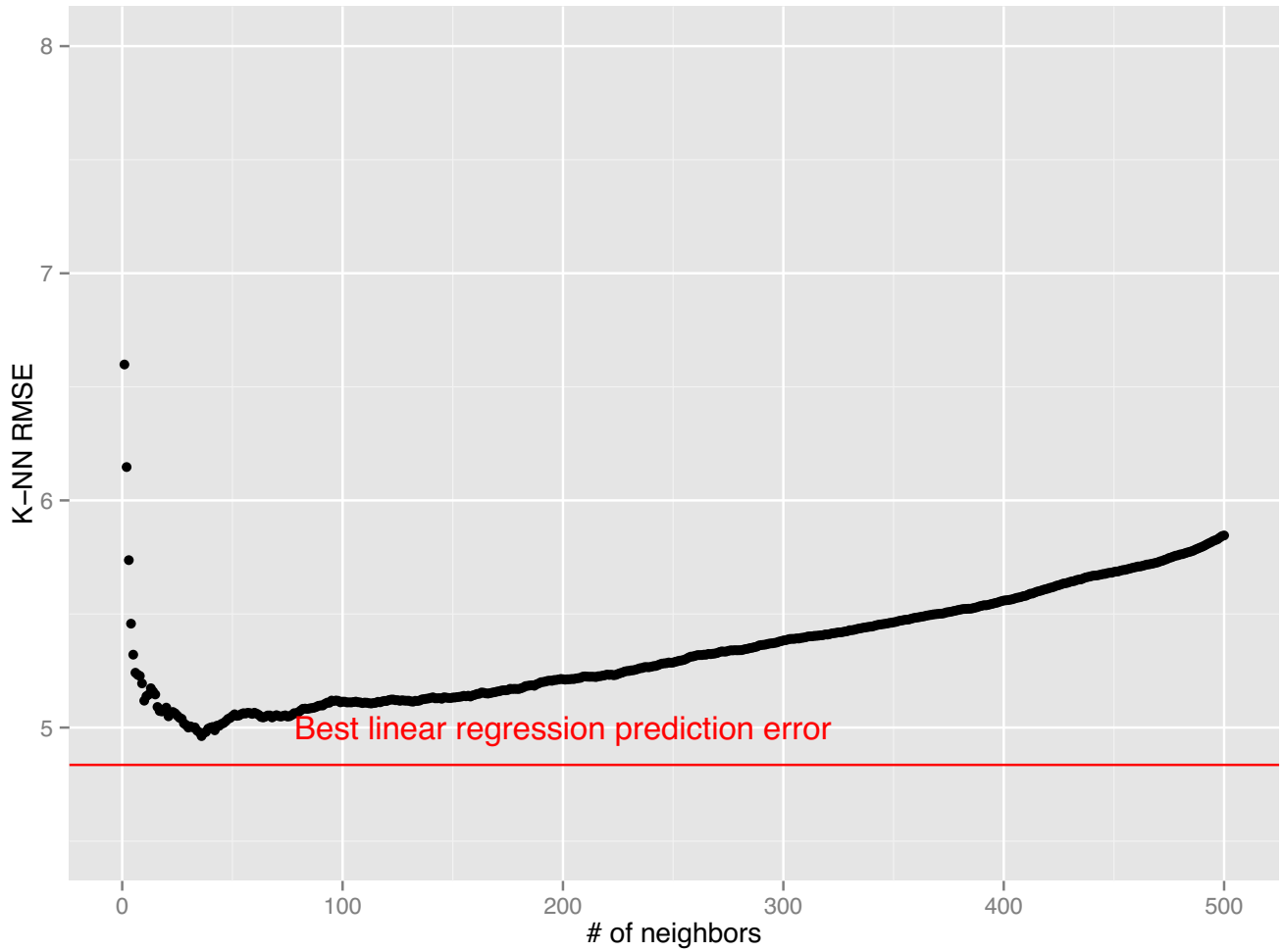
Using the first 500 points, we create a *k-nearest-neighbor* (k-NN) model: For any $\vec{X}$, let $\hat{f}(\vec{X})$ be the average value of Y_i over the k nearest neighbors $\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(k)}$ to $\vec{X}$ in the training set.



How does this fit behave as a function of k?

Example: k-nearest-neighbor fit

The graph shows root mean squared error (RMSE) over the remaining 500 samples (test set) as a function of k:



Example: k-nearest-neighbor fit

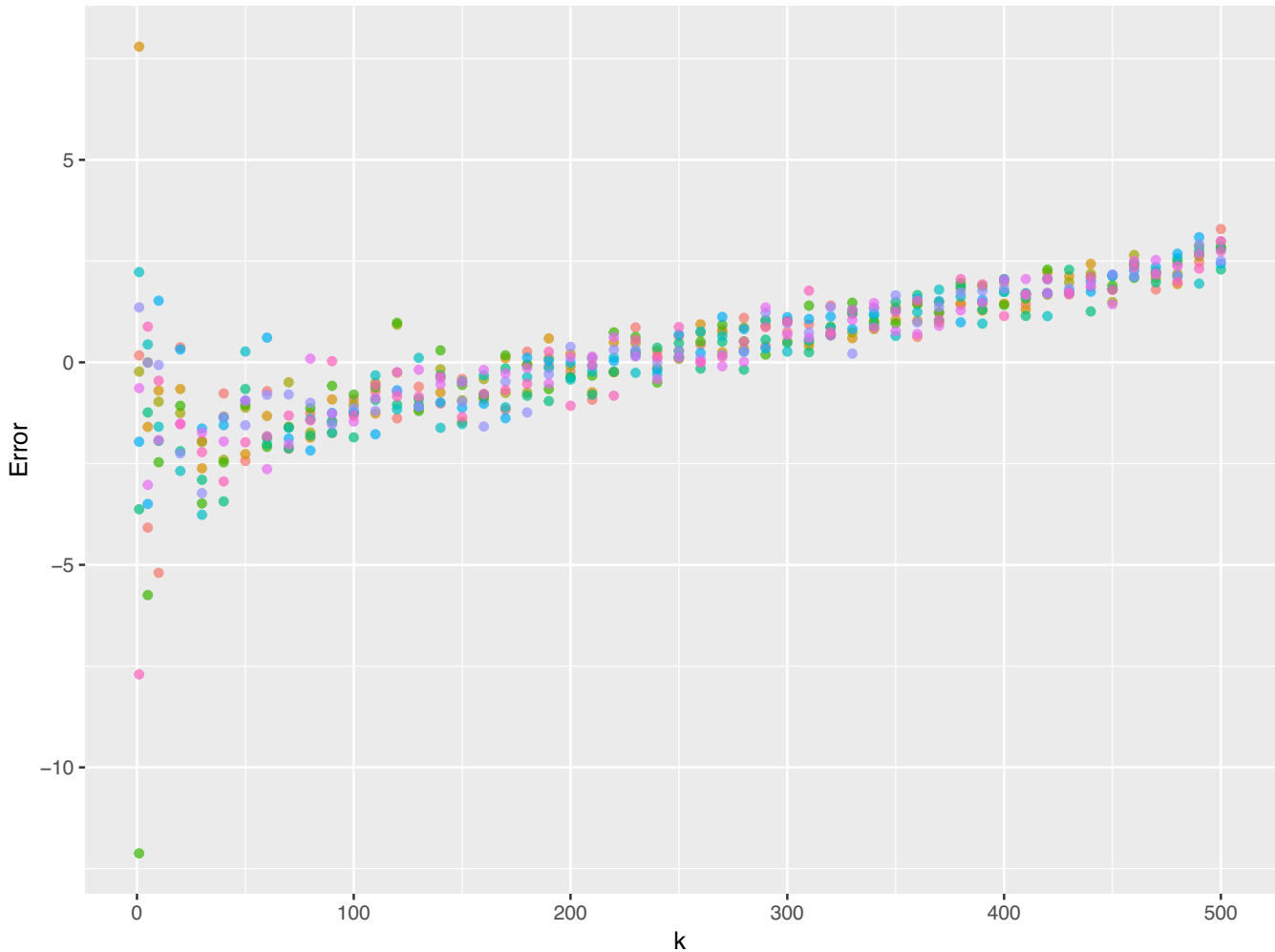
We can get more insight into why RMSE behaves this way if we pick a specific point and look at how our prediction error varies with k.

In particular, repeat the following process 10 times:

- Each time, we start with a new training dataset of 500 samples.
- We use this training dataset to make predictions at the specific point $X_1 = 1$, $X_2 = 1$.
- We plot the error $f(\vec{X})$ minus our prediction, where $f(\vec{X}) = 1 + 2X_1 + 3X_2 = 6$ is the true conditional expectation of Y given $\vec{X}$ in the population model.

Example: k-nearest neighbor fit

Results (each color is one repetition):



In other words, as k increases, *variance* goes down while *bias* goes up.

k-nearest-neighbor fit

Given $\vec{X}$ and $\mathbf{X}$, let $X_{(1)}, \ldots, X_{(k)}$ be the k closest points to $\vec{X}$ in the data. You will show that:

$$\text{BIAS} = f(\vec{X}) - \frac{1}{k} \sum_{i=1}^{k} f(X_{(i)}),$$

and

$$\text{VARIANCE} = \frac{\sigma^2}{k}.$$

This type of result is why we often refer to a "tradeoff" between bias and variance.

Linear regression: Linear population model

Now suppose the population model itself is *linear* with p covariates:[2]

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon, \quad \text{i.e. } Y = \vec{X}\boldsymbol{\beta} + \varepsilon,$$

for some fixed parameter vector $\boldsymbol{\beta}$. Note that $f(\vec{X}) = \vec{X}\boldsymbol{\beta}$.

The errors ε are i.i.d. with $\mathbb{E}[\varepsilon_i|\mathbf{X}] = 0$ and $\mathrm{Var}(\varepsilon_i|\mathbf{X}) = \sigma^2$.

Also assume errors ε are *uncorrelated* across observations. So for our given data $\mathbf{X}$, $\mathbf{Y}$, we have $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where:

$$\mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}.$$

[2]Note: Here $\vec{X}$ is viewed as a row vector $(1, X_1, \ldots, X_p)$.

Linear regression

Suppose we are given data $\mathbf{X}, \mathbf{Y}$ and fit the resulting model by ordinary least squares. Let $\hat{\boldsymbol{\beta}}$ denote the resulting fit:

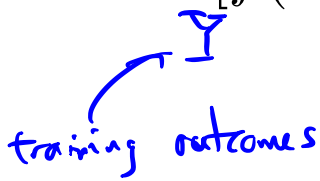
$$\hat{f}(\vec{X}) = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_j$$

with $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$.
What can we say about bias and variance?

Linear regression

Let's look at bias:

$$\begin{aligned}\mathbb{E}[\hat{f}(\vec{X})|\vec{X},\mathbf{X}] &= \mathbb{E}[\vec{X}\hat{\boldsymbol{\beta}}|\vec{X},\mathbf{X}] \\ &= \mathbb{E}[\vec{X}(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{Y})|\vec{X},\mathbf{X}] \\ &= \vec{X}(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\big(\mathbb{E}[\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}|\vec{X},\mathbf{X}]\big) \\ &= \vec{X}(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X})\boldsymbol{\beta} = \vec{X}\boldsymbol{\beta} = f(\vec{X}).\end{aligned}$$



In other words: the ordinary least squares solution is *unbiased*!

Linear regression: The Gauss-Markov theorem

In fact: among all *unbiased linear models*, the OLS solution has *minimum variance*.

This famous result in statistics is called the *Gauss-Markov* theorem.

Theorem

Assume a linear population model with uncorrelated errors. Fix a (row) covariate vector $\vec{X}$*, and let* $\gamma = \vec{X}\boldsymbol{\beta} = \sum_j \beta_j X_j$*.*

Given data $\mathbf{X}, \mathbf{Y}$*, let* $\hat{\boldsymbol{\beta}}$ *be the OLS solution. Let* $\hat{\gamma} = \vec{X}\hat{\boldsymbol{\beta}} = \vec{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$*.*

Let $\hat{\delta} = \mathbf{g}(\mathbf{X}, \vec{X})\mathbf{Y}$ *be any other estimator for* γ *that is linear in* $\mathbf{Y}$ *and* unbiased *for all* $\mathbf{X}$*:* $\mathbb{E}[\hat{\delta}|\mathbf{X}, \vec{X}] = \gamma$*.*

Then $\mathrm{Var}(\hat{\delta}|\mathbf{X}, \vec{X}) \geq \mathrm{Var}(\hat{\gamma}|\mathbf{X}, \vec{X})$*, with equality if and only if* $\hat{\delta} = \hat{\gamma}$*.*

The Gauss-Markov theorem: Proof [$*$]

Proof: We compute the variance of $\hat{\delta}$.

$$\begin{aligned}
&\mathbb{E}[(\hat{\delta} - \mathbb{E}[\hat{\delta}|\vec{X}, \mathbf{X}])^2|\vec{X}, \mathbf{X}] \\
&\quad = \mathbb{E}[(\hat{\delta} - \vec{X}\boldsymbol{\beta})^2|\vec{X}, \mathbf{X}] \\
&\quad = \mathbb{E}[(\hat{\delta} - \hat{\gamma} + \hat{\gamma} - \vec{X}\boldsymbol{\beta})^2|\vec{X}, \mathbf{X}] \\
&\quad = \mathbb{E}[(\hat{\delta} - \hat{\gamma})^2|\vec{X}, \mathbf{X}] \\
&\qquad + \mathbb{E}[(\hat{\gamma} - \vec{X}\boldsymbol{\beta})^2|\vec{X}, \mathbf{X}] \\
&\qquad + 2\mathbb{E}[(\hat{\delta} - \hat{\gamma})(\hat{\gamma} - \vec{X}\boldsymbol{\beta})|\vec{X}, \mathbf{X}].
\end{aligned}$$

Look at the last equality: If we can show the last term is zero, then we would be done, because the first two terms are uniquely minimized if $\hat{\delta} = \hat{\gamma}$.

The Gauss-Markov theorem: Proof [$*$]

Proof continued: For notational simplicity let $\mathbf{c} = \mathbf{g}(\mathbf{X}, \vec{X})$. We have:

$$\begin{aligned}
&\mathbb{E}[(\hat{\delta} - \hat{\gamma})(\hat{\gamma} - \vec{X}\boldsymbol{\beta})|\vec{X}, \mathbf{X}] \\
&\quad = \mathbb{E}[(\mathbf{c}\mathbf{Y} - \vec{X}\hat{\boldsymbol{\beta}})^2(\vec{X}\hat{\boldsymbol{\beta}} - \vec{X}\boldsymbol{\beta})|\vec{X}, \mathbf{X}].
\end{aligned}$$

Now using the fact that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$, that $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and the fact that $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top|\vec{X}, \mathbf{X}] = \sigma^2\mathbf{I}$, the last quantity reduces (after some tedious algebra) to:

$$\sigma^2\vec{X}(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{c}^\top - \vec{X}^\top).$$

The Gauss-Markov theorem: Proof [$*$]

Proof continued: To finish the proof, notice that from unbiasedness we have:

$$\mathbb{E}[\mathbf{c}\mathbf{Y}|\mathbf{X}, \vec{X}] = \vec{X}\boldsymbol{\beta}.$$

But since $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}, \vec{X}] = 0$, we have:

$$\mathbf{c}\mathbf{X}\boldsymbol{\beta} = \vec{X}\boldsymbol{\beta}.$$

Since this has to hold true for every $\mathbf{X}$, we must have $\mathbf{c}\mathbf{X} = \vec{X}$, i.e., that:

$$\mathbf{X}^\top \mathbf{c}^\top - \vec{X}^\top = 0.$$

This concludes the proof.

Linear regression: Variance of OLS

We can explicitly work out the variance of OLS in the linear population model.

$$\begin{aligned}\mathrm{Var}(\hat{f}(\vec{X})|\mathbf{X},\vec{X}) &= \mathrm{Var}(\vec{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}|\mathbf{X},\vec{X}) \\ &= \vec{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathrm{Var}(\mathbf{Y}|\mathbf{X})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\vec{X}^\top.\end{aligned}$$

Now note that $\mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}$ where $\mathbf{I}$ is the $n \times n$ identity matrix.

Therefore:

$$\mathrm{Var}(\hat{f}(\vec{X})|\mathbf{X},\vec{X}) = \sigma^2\vec{X}(\mathbf{X}^\top\mathbf{X})^{-1}\vec{X}^\top.$$

Linear regression: In-sample prediction error

The preceding formula is not particularly intuitive. To get more intuition, evaluate *in-sample prediction error*:

$$\mathsf{Err}_{\mathsf{in}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[(Y - \hat{f}(\vec{X}))^2 | \mathbf{X}, \mathbf{Y}, \vec{X} = \mathbf{X}_i].$$

This is the prediction error if we received new samples of Y corresponding to each covariate vector in our existing data. Note that the only randomness in the preceding expression is in the new observations Y.

(We saw in-sample prediction error in our discussion of model scores.)

Linear regression: In-sample prediction error [$*$]

Taking expectations over $\mathbf{Y}$ and using the bias-variance decomposition on each term of Err_{in}, we have:

$$\mathbb{E}[(Y - \hat{f}(\vec{X}))^2 | \mathbf{X}, \vec{X} = \mathbf{X}_i] = \sigma^2 + 0^2 + \sigma^2 \mathbf{X}_i (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i^\top.$$

First term is irreducible error; second term is zero because OLS is unbiased; and third term is variance when $\vec{X} = \mathbf{X}_i$.

Now note that:

$$\sum_{i=1}^{n} \mathbf{X}_i (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i = \text{Trace}(\mathbf{H}),$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is the *hat matrix*.

It can be shown that the $\text{Trace}(\mathbf{H}) = p$ (see appendix).

Linear regression: In-sample prediction error

Therefore we have:

$$\mathbb{E}[\mathsf{Err}_{\mathsf{in}}|\mathbf{X}] = \sigma^2 + 0^2 + \frac{p}{n}\sigma^2.$$

This is the bias-variance decomposition for linear regression:

- As before σ^2 is the *irreducible error*.
- 0^2 is the BIAS^2: OLS is unbiased.
- The last term, $(p/n)\sigma^2$, is the VARIANCE. It increases with p.

Linear regression: Model specification

It was very important in the preceding analysis that the covariates we used were the same as in the population model!
This is why:

- The bias is zero.
- The variance is $(p/n)\sigma^2$.

On the other hand, with a large number of potential covariates, will we want to use all of them?

Linear regression: Model specification

What happens if we use a subset of covariates $S \subset \{0, \ldots, p\}$?

- In general, the resulting model will be *biased*.
- It can be shown that the same analysis holds:

$$\mathbb{E}[\mathsf{Err}_{\mathsf{in}}|\mathbf{X}] = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i\boldsymbol{\beta} - \mathbb{E}[\hat{f}(\mathbf{X}_i)|\mathbf{X}])^2 + \frac{|S|}{n}\sigma^2.$$

(Second term is BIAS^2.)

So in general, bias increases while variance decreases.[3]

[3] As shown on Problem Set 2, it's possible to make good predictions despite the omission of variables, provided that the covariates that are included have sufficient correlation with the omitted variables. This is what is captured by the bias expression on the slide. Whenever covariates are omitted, another concern is that the estimates of the coefficients of the included covariates may be incorrect. This phenomenon often appears as the *omitted variable bias* in econometrics.

Linear regression: Model specification

What happens if we introduce a new covariate that is uncorrelated with the existing covariates and the outcome?

- The bias remains zero.
- However, now the variance increases, since VARIANCE $= (p/n)\sigma^2$.

Bias-variance decomposition: Summary

The bias-variance decomposition is a conceptual guide to understanding what influences test error:

- Frames the question by asking: *What if we were to repeatedly use the same modeling approach, but on different training sets?*
- On average (across models built from different training sets), test error looks like:

$$\text{irreducible error} + \text{bias}^2 + \text{variance}.$$

- *Bias*: systematic mistakes in predictions on test data, regardless of the training set
- *Variance*: variation in predictions on test data, as training data changes

So key point is: models can be "bad" (high test error) because of high bias or high variance (or both).

Bias-variance decomposition: Summary

In practice, e.g., for linear regression models:

- More "complex" models tend to have higher variance and lower bias.
- Less "complex" models tend to have lower variance and higher bias.

But this is far from ironclad...in practice, you want to understand reasons why bias may be high (or low) as well as why variance might be high (or low).

Appendix: Trace of $\mathbf{H}$ [$*$]

Why does $\mathbf{H}$ have trace p?

- The trace of a (square) matrix is the sum of its eigenvalues.
- Recall that $\mathbf{H}$ is the matrix that projects $\mathbf{Y}$ into the column space of $\mathbf{X}$. (See Lecture 2.)
- Since it is a projection matrix, for any $\mathbf{v}$ in the column space of $\mathbf{X}$, we will have $\mathbf{H}\mathbf{v} = \mathbf{v}$.
- This means that 1 is an eigenvalue of $\mathbf{H}$ with multiplicity p.
- On the other hand, the remaining eigenvalues of $\mathbf{H}$ must be zero, since the column space of $\mathbf{X}$ has rank p.

So we conclude that $\mathbf{H}$ has p eigenvalues that are 1, and the rest are zero.