MS&E 226: Fundamentals of Data Science

Lecture 7: Further discussion of bias and variance

Ramesh Johari

Bias and variance in linear regression

Linear regression: Linear population model

In this part of the lecture, we're going to study how the bias and variance of linear regression behave.

We'll start with some specific (typically unrealistic!) assumptions on the population model, and conclude by discussing what happens as we relax these assumptions.

Notation

Throughout, we'll assume we have n observations, and p covariates observed with each outcome.

As usual Y denotes a generic outcome drawn from the population model, and $\vec{X} = (1, X_1, \ldots, X_p)$ denotes the associated covariate (row) vector, with an intercept term included. In addition, $\mathbf{X}$ and $\mathbf{Y}$ denote the data sample used to build a fitted model.

The linear population model

Three key assumptions:

(A1) The population model is *linear*: there are *parameters* $\beta_0, \ldots, \beta_p$ such that:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon,$$

where ε is an *error* random variable.

(A2) The sample data $\mathbf{X}, \mathbf{Y}$, consists of n independent draws from the population model.
In particular, this means errors ε_i are independent and identically distributed, and importantly, *all covariates in the population model are present in our sample*.

(A3) The errors have zero mean conditional on the sample: $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$. They are also *homoskedastic*: $\mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

The linear population model

This is a very strong set of assumptions!

Example: houses.

(A1) The price of a house is *linear* in a given set of features (living area, number of bedrooms, number of bathrooms), plus some (random) error.

(A2) The houses in our sample data set are independent draws from this model, *including all the features from (A1)*.

(A3) Errors have zero mean and constant variance.

OLS 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?

The Gauss-Markov Theorem

Recall: Irreducible error, bias, and variance allow us to decompose the generalization error we expect to see *as our training data varies*.

The *Gauss-Markov Theorem* (GMT) for OLS (informally) states that if (A1)-(A3) hold, then:

- The irreducible error is σ^2.
- The bias of OLS is **zero**.[1]
- In addition, among *all* linear modeling strategies using the covariates in $\mathbf{X}$ that have *zero bias*, OLS has the *lowest variance*.

[1] In fact, it can be shown that $\mathbb{E}_{\mathbf{Y}}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$, i.e., the OLS coefficients are an unbiased estimate of the true parameters.

The Gauss-Markov Theorem

Why is the GMT practically important?

- To lower prediction error, we have to lower bias, variance, or both.
- The GMT says we can't lower variance if we insist on a linear modeling strategy with zero bias.
- So: any linear modeling strategy with lower prediction error than OLS must have **positive bias**, and **lower variance** than OLS.

This observation is one of the important justifications for the value of linear modeling strategies such as lasso and ridge regression: these will generally have positive bias, but might offer lower variance in return.

Revisiting assumptions

Revisiting assumptions

The assumptions (A1)-(A3) play a critical role in our analysis.

We're now going to discuss implications for bias and variance if these assumptions are violated.

Our discussion will not be by any means exhaustive, but will overview some of the most important consequences to be aware of.

Revisiting assumptions: (A2)-(A3)

(A2)-(A3) says errors are homoskedastic and independent, and thus uncorrelated:

$\mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

What if this is violated?

Revisiting assumptions: (A2)-(A3)

What if these are violated?

- In general, error terms may be correlated with each other. *Example:* Observations were "grouped" when collected, e.g., houses from a common neighborhood have correlated sales prices.
- In general, error terms may have variance that *depends* on the covariates. *Example:* Houses with higher square footage likely have more variance in their sales price than houses with lower square footage.

Both violations of (A3) mean that there may be other unbiased techniques that yield *lower variance* than OLS (so GMT no longer holds). (See *weighted least squares* (WLS) on your problem set.)

Revisiting assumptions: (A1)

(A1) asserts that the population model relationship between covariates and outcome is *linear*; in particular, $f(\vec{X}) = \mathbb{E}[Y|\vec{X}]$ is linear in the covariates.

There are certainly situations where this is an appropriate assumption: for example, in many physical systems, there are laws that are known to be linear (e.g., Hooke's law for springs).

In most practical scenarios, it's a *simplifying* assumption. If the true population model relationship between covariates and the outcome is *nonlinear*, then any linear models will be *biased*.

Nonlinearity and covariate transformations

Previously, we talked about how we can model nonlinear relationships using linear models, by using data transformations (higher order terms, interactions, etc.).

This technique can help address bias when the true population model is nonlinear: by adding these additional features (covariates) to our $\mathbf{X}$, we can potentially reduce bias compared to the true population model.

However ... a potential concern is that introducing these additional features might increase variance, as we build more "complex" models.

Which covariates?

Revisiting assumptions: (A2)

Our ability to add higher order terms draws attention to a critical part of assumption (A2):

All covariates in the population model are also in our data.

Revisiting assumptions: (A2)

Let's explore this assumption in greater detail,
especially the role of the number of covariates in affecting bias and variance.

(Unless otherwise stated, continue to assume that all other parts of (A1)-(A3) hold.)

Linear regression: Fewer covariates

What happens if we fit our model using only a subset of covariates $S \subset \{0, \ldots, p\}$ from the population model?

In general, the resulting OLS model will be *biased*.

(Note that this is what we observed in the previous lecture.)

Linear regression: Fewer covariates

A couple remarks:

- In general, the amount of bias introduced will depend on how correlated the *remaining* covariates are with the *omitted* covariates. (This is why it can be possible to make good predictions despite the omission of variables, as you saw on your problem set.)
- When 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; we will return to this in our next unit on inference.

Linear regression: More covariates

What happens if we introduce a new covariate into our modeling strategy?

As an extreme case, suppose our new covariate is *uncorrelated* with the existing covariates and the outcome. Then:

- The bias remains zero.
- However, the variance will *increase*.

(See the R Shiny dashboard for an example.)

A bias-variance "tradeoff"?

So we find that:

- If (A1)-(A3) hold, then GMT $\implies$ OLS has zero bias.
- If we fit with *fewer* covariates than the population model, but the other parts of (A1)-(A3) hold, then in general OLS has *nonzero* bias. (In our dashboard we see that variance can also decrease in this case.)
- If we fit with all the covariates in the population model, as well as *additional* uncorrelated covariates, and the other parts of (A1)-(A3) hold, then in general OLS still has zero bias but *higher* variance.

A bias-variance "tradeoff"?

If we measure "complexity" of a modeling strategy in terms of *the number of covariates we use to fit the model*, then our preceding discussion suggests that:

- More "complex" modeling strategies (i.e., more covariates) tend to have higher variance and lower bias.
- Less "complex" modeling strategies (i.e., fewer covariates) tend to have lower variance and higher bias.

This is usually what's meant by a "bias-variance tradeoff."

Is this all there is to the story?

A bias-variance "tradeoff"?

The "tradeoff" view can be misleading!

It's possible to have high bias and high variance at the same time, and bias and variance can move in unusual ways with model "complexity"...
In general, adding covariates can affect both bias and variance, depending on the correlation between the new covariates, existing covariates, and the outcome.

Let's dig into this a bit more...

A puzzle: Double descent

Recent developments in machine learning have challenged the "tradeoff": for some combinations of data context, modeling strategy, and model fitting procedure, the following relationship between train error, test error, and "complexity" is observed:

(From `openai.com/blog/deep-double-descent/`)

"Overparameterization"

The double descent phenomenon highlights the intriguing phenomenon that modeling strategies that appear to be "overparameterized" can in fact generalize very well:

Informally, models so "complex" that they can nearly perfectly fit the training data, are nevertheless not necessarily *overfitting* the data; they appear to be both *low bias* and *low variance*.

How?

Double descent: An example

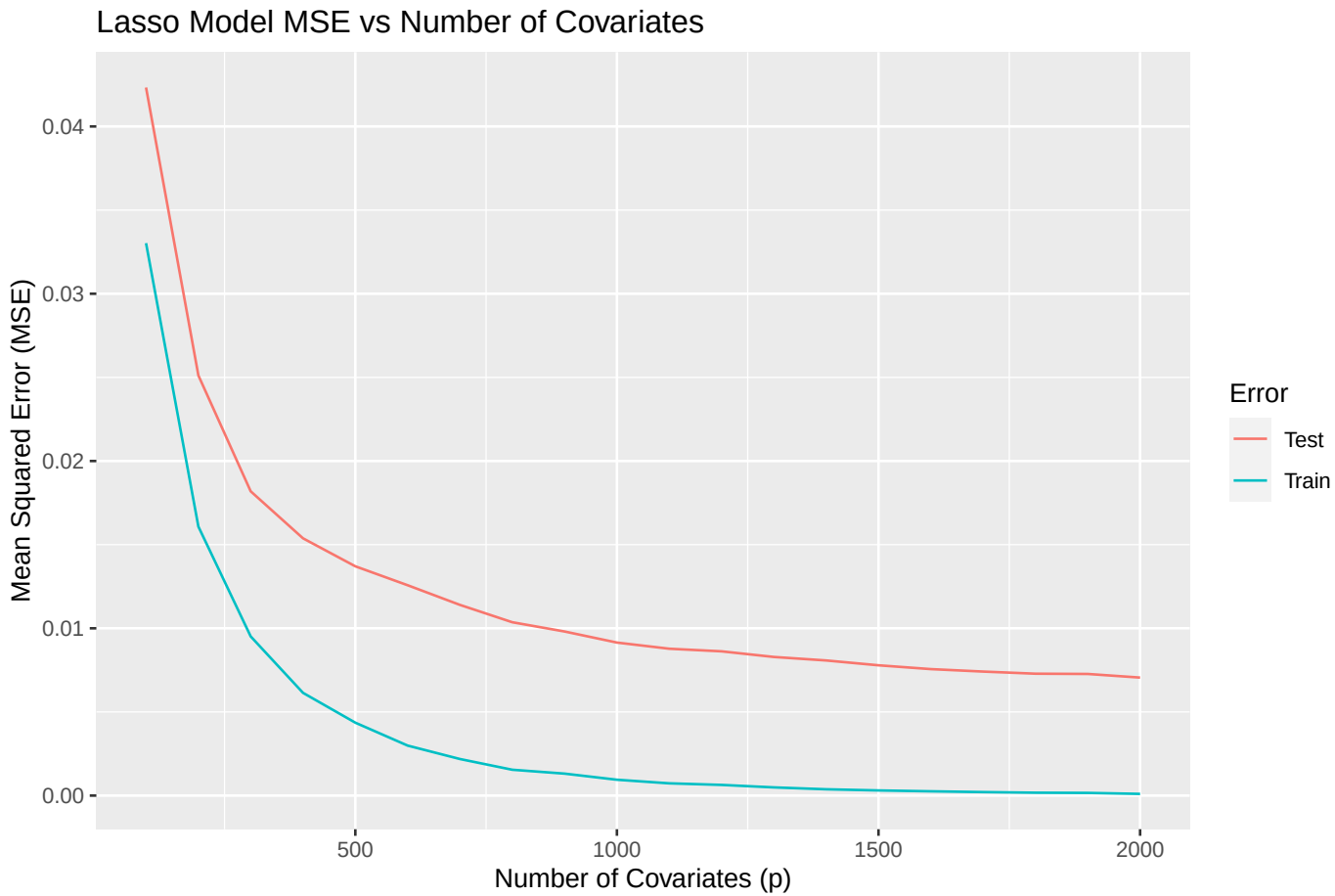
Let's explore double descent with lasso. We'll use synthetic data generated as follows:

1. *First*, generate $Y_1, \ldots, Y_{1000}$ as $N(0,1)$ random variables.
2. *Then*, for each $i = 1, \ldots, 1000$, generate 2000 *covariates* $X_{i1}, \ldots, X_{i,2000}$ as $X_{ij} = Y_i + s_{ij}$, where s_{ij} is $N(0, \tau^2)$.
3. Finally, we fit lasso models using *only* the first p covariates, for $p = 100, 200, \ldots, 2000$ (with a fixed value of λ), and test them on a test set with 5000 data points,

What's unusual about this data generating procedure?

Results

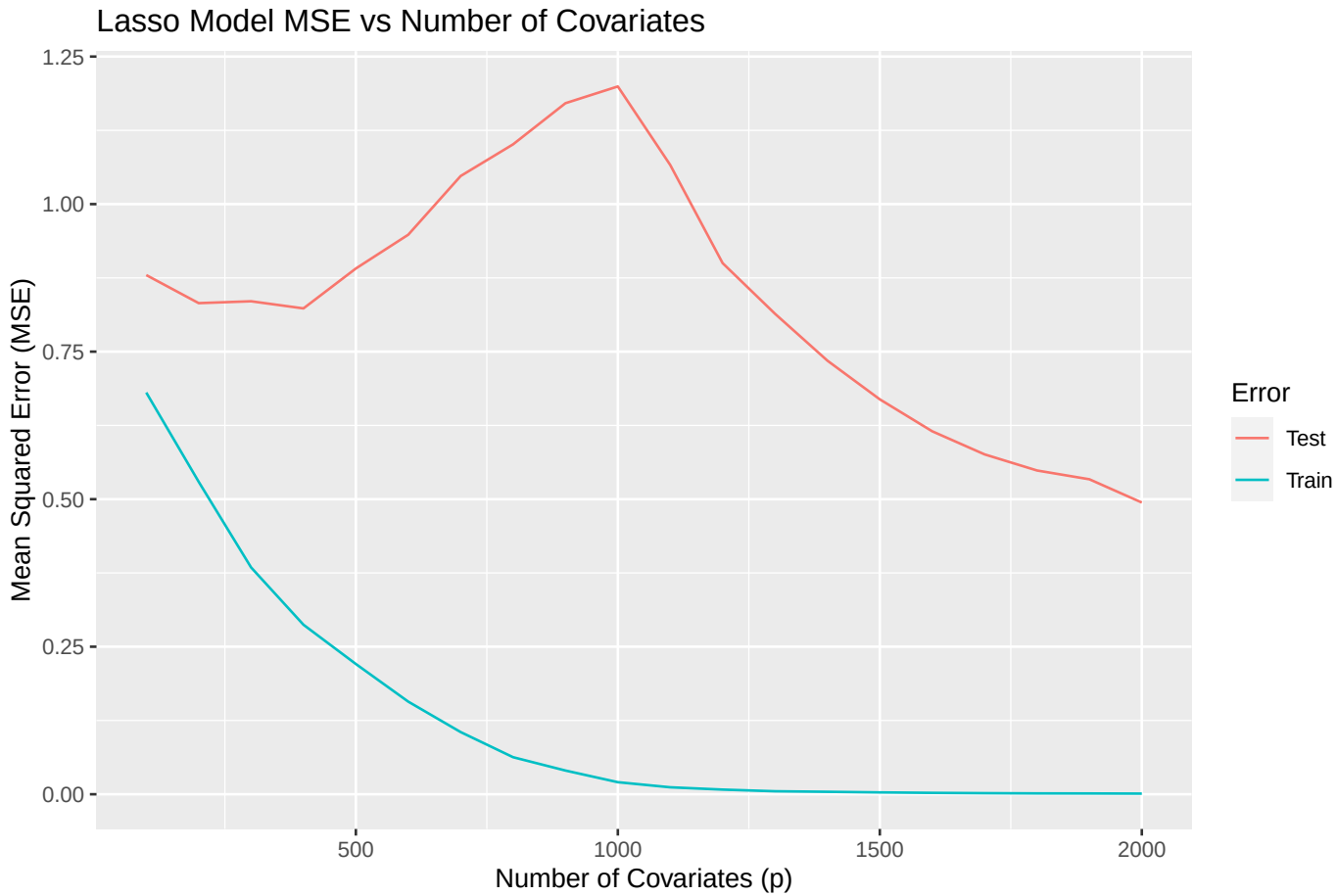
If $\tau = 2$ and $\lambda = 0.001$, our test error continues to drop with increasing p:



Why? Because *every* additional covariate brings new information about the outcome.

Results

What happened to "double descent"? Now suppose $\tau = 20$ and $\lambda = 0.001$:



Why? As $p \to 1000$, lasso gets "confused": τ^2 is high, and λ isn't very large, so it fits the training data well ... but then generalizes poorly. But with more covariates, this "confusion" is overcome.

Moral: "Big data" and highly parameterized models

Usually, we think of "big data" as having more samples in our data (larger n).

But an equally consequential change is the availability (and relevance to the population model!) of *many more covariates* (larger p).

Moral: "Big data" and highly parameterized models

In many ways, "double descent" is not the central story here: it is a practical artifact of a particular modeling strategy and model fitting procedure passing through "confusion" as p grows, so that variance rises then falls.

Instead a key takeaway is:

With the right data context, right modeling strategy, and right model fitting procedure, *if* those new covariates all bring novel insight about the outcome, bias *and* variance can continue to decrease as you use them. (See R Shiny dashboard for an example.)

Summary

Bias-variance decomposition: Summary

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

- Frames the question by asking: *What if we were to repeatedly use the same modeling strategy, but on different training sets?*
- *Generalization error has only three components!* On average (across models built from different training sets):

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

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

Modeling strategies can be "bad" (high test error) because of high bias or high variance (or both).

Bias and variance in linear regression: Technical details

Technical details for linear regression [$*$]

In this section of slides, we provide technical details for the claims made about linear regression under assumptions (A1)-(A3).

Linear regression: bias [$*$]

This derivation shows bias of OLS is zero:

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

The Gauss-Markov theorem: Precise statement [$*$]

The Gauss-Markov theorem shows that among all unbiased linear modeling strategies, OLS has minimum variance.

Theorem (Gauss-Markov)

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* $\mathbf{X}$*:* $\mathbb{E}_{\mathbf{Y}}[\hat{\delta}|\mathbf{X}, \vec{X}] = \gamma$.

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:[2] 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 $\hat{\delta} = \hat{\gamma}$.

[2] Throughout this proof we suppress subscripts on the expectations for ease of exposition.

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.

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:

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