

COMPSCI 689

Lecture 3: Properties of ERM for Linear Regression

Brendan O'Connor

College of Information and Computer Sciences
University of Massachusetts Amherst

Slides by Benjamin M. Marlin (marlin@cs.umass.edu).

Outline

1 Review

2 Assumptions

3 Empirical Analysis

4 Theoretical Analysis

5 Discussion

6 Appendix

Empirical Risk Minimization

Let Θ be the parameter space for a set \mathcal{F} of prediction functions f_θ mapping from \mathcal{X} to \mathcal{Y} . The principle of Empirical Risk Minimization (ERM) states that we should select the parameters $\theta \in \Theta$ that *minimize the average of the prediction loss* $L(\mathbf{y}_n, f_\theta(\mathbf{x}_n))$ computed over the data set \mathcal{D} , also known as the empirical risk $R(f_\theta, \mathcal{D})$:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} R(f_\theta, \mathcal{D})$$

$$R(f_\theta, \mathcal{D}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{y}_n - f_\theta(\mathbf{x}_n))^2$$

ERM for Linear Regression

Given the choice of the squared prediction loss $L_{sqr}(y, y') = (y - y')^2$ and the space of prediction functions $\mathcal{F} = \{f_{\theta}(\mathbf{x}) = \mathbf{x}\mathbf{w} + b \mid \theta \in \Theta\}$, we can apply ERM to define the optimal model parameters:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} R(f_{\theta}, \mathcal{D})$$

$$R(f_{\theta}, \mathcal{D}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{y}_n - (\mathbf{x}_n \mathbf{w} + b))^2$$

ERM for Linear Regression Solution

Using closed-form optimization methods, we derived the solution to the ERM learning problem for linear regression:

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- The formula for computing $\hat{\theta}$ is referred to as an *estimator*. This estimator is only valid if $\mathbf{X}^T \mathbf{X}$ is an invertible matrix.

ERM for Linear Regression Solution

Using closed-form optimization methods, we derived the solution to the ERM learning problem for linear regression:

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- The formula for computing $\hat{\theta}$ is referred to as an *estimator*. This estimator is only valid if $\mathbf{X}^T \mathbf{X}$ is an invertible matrix.
- The process of actually computing the value of $\hat{\theta}$ is referred to as *parameter estimation*, *model fitting*, or *learning*.

ERM for Linear Regression Solution

Using closed-form optimization methods, we derived the solution to the ERM learning problem for linear regression:

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- The formula for computing $\hat{\theta}$ is referred to as an *estimator*. This estimator is only valid if $\mathbf{X}^T \mathbf{X}$ is an invertible matrix.
- The process of actually computing the value of $\hat{\theta}$ is referred to as *parameter estimation, model fitting, or learning*.
- The computed value of $\hat{\theta}$ is referred to as the *parameter estimate, the fit parameters, the optimal parameters or the learned parameters*.

ERM for Linear Regression Solution

Using closed-form optimization methods, we derived the solution to the ERM learning problem for linear regression:

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- The formula for computing $\hat{\theta}$ is referred to as an *estimator*. This estimator is only valid if $\mathbf{X}^T \mathbf{X}$ is an invertible matrix.
- The process of actually computing the value of $\hat{\theta}$ is referred to as *parameter estimation, model fitting, or learning*.
- The computed value of $\hat{\theta}$ is referred to as the *parameter estimate, the fit parameters, the optimal parameters or the learned parameters*.
- The $\hat{\theta}$ estimator that we derived for linear regression is often referred to as the *ordinary least squares* (OLS) estimator.

ERM for Linear Regression Solution

Using closed-form optimization methods, we derived the solution to the ERM learning problem for linear regression. This solution is often referred to as the Ordinary Least Squares (OLS) *estimator* of θ .

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- Given a new input vector \mathbf{x} we predict the corresponding output using the estimated parameters as follows: $\hat{y} = \underbrace{f_{\hat{\theta}}(\mathbf{x})}_{\text{prediction}} = \underbrace{\mathbf{x}\hat{\theta}}_{\text{prediction}}$

ERM for Linear Regression Solution

Using closed-form optimization methods, we derived the solution to the ERM learning problem for linear regression. This solution is often referred to as the Ordinary Least Squares (OLS) *estimator* of θ .

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- Given a new input vector \mathbf{x} we predict the corresponding output using the estimated parameters as follows: $\hat{y} = f_{\hat{\theta}}(\mathbf{x}) = \mathbf{x}\hat{\theta}$
- **Questions:** How well should we expect the estimated parameters to perform on *future data*? What can we say about the *accuracy* of the fit parameters?

Outline

1 Review

2 Assumptions

3 Empirical Analysis

4 Theoretical Analysis

5 Discussion

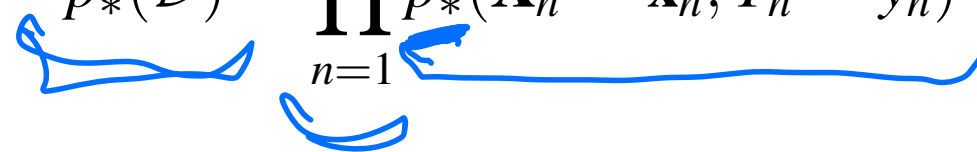
6 Appendix

Data Generating Distributions

- We introduced the concept of a data set \mathcal{D} , but we have not made any assumptions about the process that generates the data set.

Data Generating Distributions

- We introduced the concept of a data set \mathcal{D} , but we have not made any assumptions about the process that generates the data set.
- The most common assumption is that data case are generated independently and identically (IID) from an underlying joint probability distribution over (\mathbf{x}, y) pairs:

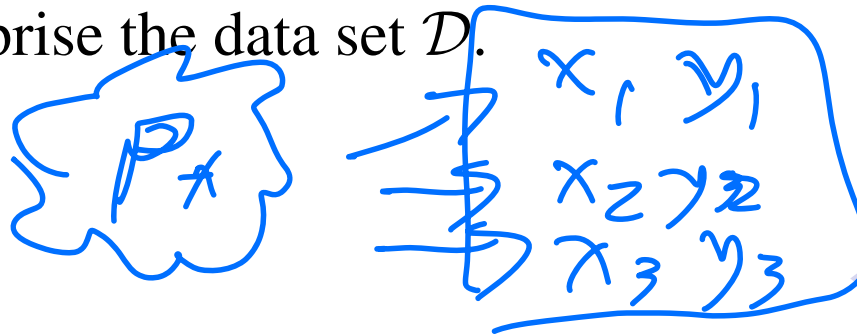
$$p_*(\mathcal{D}) = \prod_{n=1}^N p_*(\mathbf{X}_n = \mathbf{x}_n, Y_n = y_n)$$


Data Generating Distributions

- We introduced the concept of a data set \mathcal{D} , but we have not made any assumptions about the process that generates the data set.
- The most common assumption is that data case are generated independently and identically (IID) from an underlying joint probability distribution over (\mathbf{x}, y) pairs:

$$p_*(\mathcal{D}) = \prod_{n=1}^N p_*(\mathbf{X}_n = \mathbf{x}_n, Y_n = y_n)$$

- We refer to p_* as the true data generating distribution. Our only information about p_* is provided through the (\mathbf{x}_n, y_n) *samples* that comprise the data set \mathcal{D} .



Data Generating Distributions

- Using the chain rule of probability, we can write:

$$p_*(\mathbf{X}_n = \mathbf{x}_n, Y_n = y_n) = \underbrace{p_*(\mathbf{X}_n = \mathbf{x}_n)}_{\text{joint distribution of } \mathbf{X}_n} \underbrace{p_*(Y_n = y_n | \mathbf{X}_n = \mathbf{x}_n)}_{\text{conditional distribution of } Y_n \text{ given } \mathbf{X}_n}$$

Data Generating Distributions

- Using the chain rule of probability, we can write:

$$p_*(\mathbf{X}_n = \mathbf{x}_n, Y_n = y_n) = p_*(\mathbf{X}_n = \mathbf{x}_n)p_*(Y_n = y_n|\mathbf{X}_n = \mathbf{x}_n)$$

- We can then define $p_*(\mathbf{X}) = \prod_{n=1}^N p_*(\mathbf{X}_n = \mathbf{x}_n)$

Data Generating Distributions

- Using the chain rule of probability, we can write:

$$p_*(\mathbf{X}_n = \mathbf{x}_n, Y_n = y_n) = p_*(\mathbf{X}_n = \mathbf{x}_n)p_*(Y_n = y_n|\mathbf{X}_n = \mathbf{x}_n)$$

- We can then define $p_*(\mathbf{X}) = \prod_{n=1}^N p_*(\mathbf{X}_n = \mathbf{x}_n)$
- As well as $p_*(\mathbf{Y}|\mathbf{X}) = \prod_{n=1}^N p_*(Y_n = y_n|\mathbf{X}_n = \mathbf{x}_n)$

Data Generating Distributions

- Using the chain rule of probability, we can write:

$$p_*(\mathbf{X}_n = \mathbf{x}_n, Y_n = y_n) = p_*(\mathbf{X}_n = \mathbf{x}_n)p_*(Y_n = y_n|\mathbf{X}_n = \mathbf{x}_n)$$

- We can then define $p_*(\mathbf{X}) = \prod_{n=1}^N p_*(\mathbf{X}_n = \mathbf{x}_n)$
- As well as $p_*(\mathbf{Y}|\mathbf{X}) = \prod_{n=1}^N p_*(Y_n = y_n|\mathbf{X}_n = \mathbf{x}_n)$
- We thus have $p_*(\mathcal{D}) = p_*(\mathbf{X})p_*(\mathbf{Y}|\mathbf{X})$

Future Data

- For parameter estimates derived from \mathcal{D} to generalize to future data cases from a new data set \mathcal{D}' sampled from a distribution p'_* , the distributions p_* and p'_* must be related in some way.

Future Data

- For parameter estimates derived from \mathcal{D} to generalize to future data cases from a new data set \mathcal{D}' sampled from a distribution p'_* , the distributions p_* and p'_* must be related in some way.
- The most common assumption is simply that $p'_* = p_*$.

Future Data

- For parameter estimates derived from \mathcal{D} to generalize to future data cases from a new data set \mathcal{D}' sampled from a distribution p'_* , the distributions p_* and p'_* must be related in some way.
- The most common assumption is simply that $p'_* = p_*$.
- When $p'_* \neq p_*$, the data in \mathcal{D}' are referred to as *out of distribution* or OOD.

Future Data

- For parameter estimates derived from \mathcal{D} to generalize to future data cases from a new data set \mathcal{D}' sampled from a distribution p'_* , the distributions p_* and p'_* must be related in some way.
- The most common assumption is simply that $p'_* = p_*$.
- When $p'_* \neq p_*$, the data in \mathcal{D}' are referred to as *out of distribution* or OOD.
- Detection and/or robustness to OOD instances is a critical issue in many current machine learning applications where systems are deployed “in the wild.”

Future Data

- For parameter estimates derived from \mathcal{D} to generalize to future data cases from a new data set \mathcal{D}' sampled from a distribution p'_* , the distributions p_* and p'_* must be related in some way.
- The most common assumption is simply that $p'_* = p_*$.
- When $p'_* \neq p_*$, the data in \mathcal{D}' are referred to as *out of distribution* or OOD.
- Detection and/or robustness to OOD instances is a critical issue in many current machine learning applications where systems are deployed “in the wild.”
- In many cases, p'_* may also not be static over time. In particular, p'_* may drift away from p_* over time resulting in the need for *continual learning*.

Outline

1 Review

2 Assumptions

3 Empirical Analysis

4 Theoretical Analysis

5 Discussion

6 Appendix

Expected Prediction Loss

- **Question:** How well should we expect estimated parameters to perform on *future data*?

Expected Prediction Loss

- **Question:** How well should we expect estimated parameters to perform on *future data*?
- If we had access to the true generating distribution of the future data, the statistic we are interested in is the expected prediction loss of our estimated model under this distribution.


Expected Prediction Loss

- **Question:** How well should we expect estimated parameters to perform on *future data*?
- If we had access to the true generating distribution of the future data, the statistic we are interested in is the expected prediction loss of our estimated model under this distribution.
- Under the assumption that the future data are generated by p'_* , the expected prediction loss of a model with estimated parameters $\hat{\theta}$ is given by:

$$E_{p'_*(y, \mathbf{x})}[L(y, f_{\hat{\theta}}(\mathbf{x}))] = \int_{\mathcal{X}} \int_{\mathcal{Y}} p'_*(y, \mathbf{x}) L(y, f_{\hat{\theta}}(\mathbf{x})) d\mathbf{x} dy$$

Estimating Expected Prediction Loss

- We can use a *newly sampled* data set \mathcal{D}' of size N' sampled from p'_* to estimate the expected prediction loss using an empirical average over \mathcal{D}' :


$$\underbrace{E_{p'_*(y, \mathbf{x})} [L(y, f_{\hat{\theta}}(\mathbf{x}))]}_{\text{Expected Prediction Loss}} \approx \frac{1}{N'} \sum_{n=1}^{N'} L(\underbrace{y'_n}_{\text{sampled } y}, \underbrace{f_{\hat{\theta}}(\mathbf{x}'_n)}_{\text{sampled } f_{\hat{\theta}}(\mathbf{x})})$$

Estimating Expected Prediction Loss

- We can use a *newly sampled* data set \mathcal{D}' of size N' sampled from p'_* to estimate the expected prediction loss using an empirical average over \mathcal{D}' :

$$E_{p'_*(y, \mathbf{x})} [L(y, f_{\hat{\theta}}(\mathbf{x}))] \approx \frac{1}{N'} \sum_{n=1}^{N'} L(y'_n, f_{\hat{\theta}}(\mathbf{x}'_n))$$

- By the law of large numbers, this average will converge to the expected prediction loss as N' goes to infinity.

Train/Test Experiment

- In machine learning, we refer to the data set \mathcal{D} used to estimate the model parameters $\hat{\theta}$ as the *training set* and the data set \mathcal{D}' used to estimate the expected prediction loss as the *test set*.

Train/Test Experiment

- In machine learning, we refer to the data set \mathcal{D} used to estimate the model parameters $\hat{\theta}$ as the *training set* and the data set \mathcal{D}' used to estimate the expected prediction loss as the *test set*.
- The training data set is often denoted \mathcal{D}_{Tr} while the test set is often denoted \mathcal{D}_{Te} .

Train/Test Experiment

- In machine learning, we refer to the data set \mathcal{D} used to estimate the model parameters $\hat{\theta}$ as the *training set* and the data set \mathcal{D}' used to estimate the expected prediction loss as the *test set*.
- The training data set is often denoted \mathcal{D}_{Tr} while the test set is often denoted \mathcal{D}_{Te} .
- A train/test experiment design is the most basic and most common machine learning experiment.

Train/Test Experiment

- In machine learning, we refer to the data set \mathcal{D} used to estimate the model parameters $\hat{\theta}$ as the *training set* and the data set \mathcal{D}' used to estimate the expected prediction loss as the *test set*.
- The training data set is often denoted \mathcal{D}_{Tr} while the test set is often denoted \mathcal{D}_{Te} .
- A train/test experiment design is the most basic and most common machine learning experiment.
- If we only have a single data set \mathcal{D} , we can split it into a training set \mathcal{D}_{Tr} and a test set \mathcal{D}_{Te} at random. However, this is equivalent to assuming that $p_* = p'_*$, which may or may not be true.



Outline

1 Review

2 Assumptions

3 Empirical Analysis

4 Theoretical Analysis

5 Discussion

6 Appendix

Theoretical Analysis

- **Question:** What can we say about the *accuracy* of the fit parameters?

Theoretical Analysis

- **Question:** What can we say about the *accuracy* of the fit parameters?
- We can consider the *theoretical* properties of estimators under different analysis frameworks.

Theoretical Analysis

- **Question:** What can we say about the *accuracy* of the fit parameters?
- We can consider the *theoretical* properties of estimators under different analysis frameworks.
- This is the purview of estimation theory and learning theory. In this lecture, we'll look at some basic estimation theory.

Assumptions About Models

- To relate the parameters of a model to a data generating distribution, we need to make additional assumptions.
- Strong assumption: Linear Gaussian generating distribution implies the least-squares estimator is the maximum likelihood estimate (MLPP ch. 7).

$$p_*(y|\mathbf{x}) = \mathcal{N}(y; \mathbf{x}\theta_*, \sigma_*^2)$$

$$p(y_i|x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i - x_i\theta)^2\right)$$

$$p(Y/X) = \prod_i p(y_i|x_i) = \left[\frac{1}{\sqrt{2\pi}\sigma}\right]^N \prod_i e^{-\frac{1}{2\sigma^2}(y_i - x_i\theta)^2}$$

$$\underset{\theta}{\operatorname{argmax}} p(Y/X) = \underset{\theta}{\operatorname{argmax}} \log p(Y/X) = \underset{\theta}{\operatorname{argmax}} \sum_i -\frac{1}{2\sigma^2}(y_i - x_i\theta)^2$$

Assumptions About Models

- To relate the parameters of a model to a data generating distribution, we need to make assumptions, but can be weaker than the last slide.

Assumptions About Models

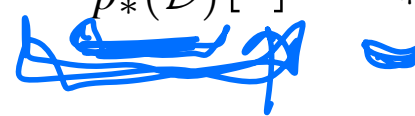
- To relate the parameters of a model to a data generating distribution, we need to make assumptions, but can be weaker than the last slide.
- In the case of the linear regression model, one common assumption is that $\mathbb{E}_{p_*(y|\mathbf{x})}[y] = \mathbf{x}\theta_*$.

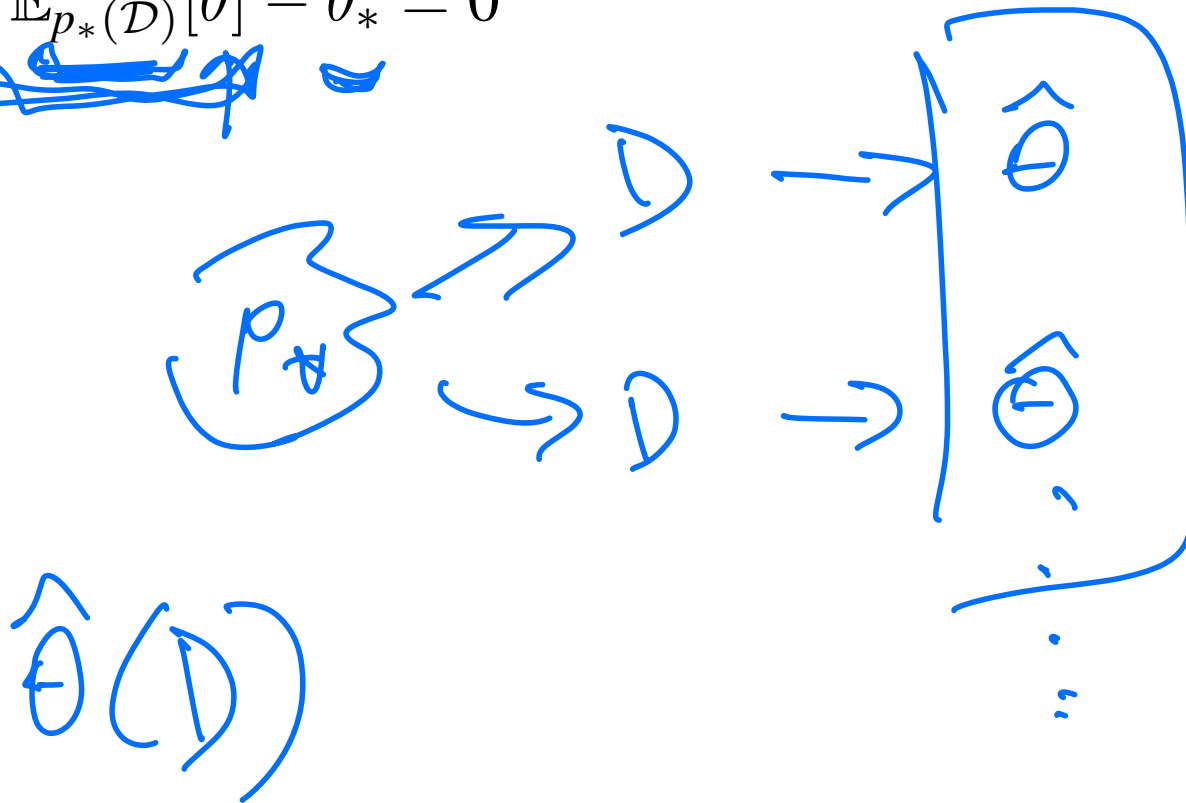
Assumptions About Models

- To relate the parameters of a model to a data generating distribution, we need to make assumptions, but can be weaker than the last slide.
- In the case of the linear regression model, one common assumption is that $\mathbb{E}_{p_*(y|\mathbf{x})}[y] = \mathbf{x}\theta_*$.
- This assumption states that the expected value of the output given the input is a linear function of the input and an unknown parameter θ_* . This means that $E_{p_*(y|\mathbf{x})}[y] \in \mathcal{F}$.

Bias of an Estimator

- An estimator $\hat{\theta}$ for a true parameter θ_* is said to be unbiased if:

$$\mathbb{E}_{p_*(\mathcal{D})}[\hat{\theta}] - \theta_* = 0$$




Bias of an Estimator

- An estimator $\hat{\theta}$ for a true parameter θ_* is said to be unbiased if:
$$\mathbb{E}_{p_*(\mathcal{D})}[\hat{\theta}] - \theta_* = 0$$
- We will show that the OLS estimator $\hat{\theta}$ is unbiased under the assumption that $\mathbb{E}_{p_*(y|\mathbf{x})}[y] = \mathbf{x}\theta_*$.

Bias of an Estimator

- An estimator $\hat{\theta}$ for a true parameter θ_* is said to be unbiased if:
$$\mathbb{E}_{p_*(\mathcal{D})}[\hat{\theta}] - \theta_* = 0$$
- We will show that the OLS estimator $\hat{\theta}$ is unbiased under the assumption that $\mathbb{E}_{p_*(y|\mathbf{x})}[y] = \mathbf{x}\theta_*$.
- When an estimator is unbiased, the estimate $\hat{\theta}$ only differs from θ_* due to randomness derived from the finite data set \mathcal{D} . A biased estimator differs from θ_* systematically.

Bias of OLS

$\hat{\theta}(\mathcal{D})$
↓

$$\underbrace{\mathbb{E}_{p_*(\mathcal{D})}[\hat{\theta}]} = \mathbb{E}_{p_*(\mathcal{D})}[\underbrace{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}}]$$

Bias of OLS

$$\begin{aligned}\mathbb{E}_{p_*(\mathcal{D})}[\hat{\theta}] &= \mathbb{E}_{p_*(\mathcal{D})}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}] \\ &= \mathbb{E}_{p_*(\mathbf{X})} \left[\mathbb{E}_{p_*(\mathbf{Y}|\mathbf{X})} \left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \right] \right]\end{aligned}$$

Bias of OLS

$$\begin{aligned}\mathbb{E}_{p_*(\mathcal{D})}[\hat{\theta}] &= \mathbb{E}_{p_*(\mathcal{D})}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}] \\ &= \mathbb{E}_{p_*(\mathbf{X})} \left[\mathbb{E}_{p_*(\mathbf{Y}|\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}] \right] \\ &= \mathbb{E}_{p_*(\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}_{p_*(\mathbf{Y}|\mathbf{X})} [\mathbf{Y}]]\end{aligned}$$

Bias of OLS

$$\begin{aligned}\mathbb{E}_{p_*(\mathcal{D})}[\hat{\theta}] &= \mathbb{E}_{p_*(\mathcal{D})}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}] \\ &= \mathbb{E}_{p_*(\mathbf{X})} [\mathbb{E}_{p_*(\mathbf{Y}|\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}]] \\ &= \mathbb{E}_{p_*(\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \underbrace{\mathbb{E}_{p_*(\mathbf{Y}|\mathbf{X})} [\mathbf{Y}]}_{\mathbf{X}\theta_*}] \\ &= \mathbb{E}_{p_*(\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \underbrace{(\mathbf{X}\theta_*)}_{\mathbf{X}\theta_*}]\end{aligned}$$

Bias of OLS

$$\begin{aligned}\mathbb{E}_{p_*(\mathcal{D})}[\hat{\theta}] &= \mathbb{E}_{p_*(\mathcal{D})}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}] \\&= \mathbb{E}_{p_*(\mathbf{X})} [\mathbb{E}_{p_*(\mathbf{Y}|\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}]] \\&= \mathbb{E}_{p_*(\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}_{p_*(\mathbf{Y}|\mathbf{X})} [\mathbf{Y}]] \\&= \mathbb{E}_{p_*(\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X} \theta_*)] \\&= \mathbb{E}_{p_*(\mathbf{X})} [\underbrace{(\mathbf{X}^T \mathbf{X})^{-1}}_{\text{red bracket}} \underbrace{(\mathbf{X}^T \mathbf{X}) \theta_*}_{\text{red bracket}}]\end{aligned}$$

Bias of OLS

$$\begin{aligned}\mathbb{E}_{p_*(\mathcal{D})}[\hat{\theta}] &= \mathbb{E}_{p_*(\mathcal{D})}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}] \\&= \mathbb{E}_{p_*(\mathbf{X})} \left[\mathbb{E}_{p_*(\mathbf{Y}|\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}] \right] \\&= \mathbb{E}_{p_*(\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}_{p_*(\mathbf{Y}|\mathbf{X})} [\mathbf{Y}]] \\&= \mathbb{E}_{p_*(\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X} \theta_*)] \\&= \mathbb{E}_{p_*(\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{X}) \theta_*] \\&= \theta_*\end{aligned}$$

Bias of OLS

$$\begin{aligned}\mathbb{E}_{p_*(\mathcal{D})}[\hat{\theta}] &= \mathbb{E}_{p_*(\mathcal{D})}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}] \\&= \mathbb{E}_{p_*(\mathbf{X})} [\mathbb{E}_{p_*(\mathbf{Y}|\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}]] \\&= \mathbb{E}_{p_*(\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}_{p_*(\mathbf{Y}|\mathbf{X})} [\mathbf{Y}]] \\&= \mathbb{E}_{p_*(\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X} \theta_*)] \\&= \mathbb{E}_{p_*(\mathbf{X})} [(\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{X}) \theta_*] \\&= \theta_*\end{aligned}$$

Thus, $\mathbb{E}_{p_*(\mathcal{D})}[\hat{\theta}] - \theta_* = \theta_* - \theta_* = 0$ and the OLS estimator is unbiased under the stated assumptions.

More Properties of OLS

- Consider the setting where $p'_* = p_*$

More Properties of OLS

- Consider the setting where $p'_* = p_*$
- If we make the assumption that $p_*(y|\mathbf{x}) = \mathcal{N}(y; \mathbf{x}\theta_*, \sigma_*^2)$, it's also possible to derive the expected squared loss of the OLS estimated model computed on a data set of size N :

$$E_{p_*(y, \mathbf{x})} [E_{p_*(\mathcal{D})} [(y - f_{\hat{\theta}}(\mathbf{x}))^2]] = \sigma_*^2 \frac{N-1}{N-D-1}$$

Best case loss: σ_*^2

Bayes error

$\uparrow D \rightarrow$ lower loss

$D > N$? No!!!

More Properties of OLS

- Consider the setting where $p'_* = p_*$
- If we make the assumption that $p_*(y|\mathbf{x}) = \mathcal{N}(y; \mathbf{x}\theta_*, \sigma_*^2)$, it's also possible to derive the expected squared loss of the OLS estimated model computed on a data set of size N :

$$E_{p_*(y, \mathbf{x})}[E_{p_*(\mathcal{D})}[(y - f_{\hat{\theta}}(\mathbf{x}))^2]] = \sigma_*^2 \frac{N-1}{N-D-1}$$

- Under some additional assumptions, one can prove that the OLS estimator is *asymptotically consistent*, which means $p_*(\|\theta_* - \hat{\theta}\|_1 \geq \epsilon) = 0$ for any $\epsilon > 0$ as N goes to infinity.

$$\|\theta_* - \hat{\theta}\|_1 = \sum_d |\theta_{*d} - \hat{\theta}_d|$$

Outline

1 Review

2 Assumptions

3 Empirical Analysis

4 Theoretical Analysis

5 Discussion

6 Appendix

Discussion

- We have presented two different ways to think about an estimator.

Discussion

- We have presented two different ways to think about an estimator.
- Under the assumption that we can obtain a test data set \mathcal{D}_{te} , we can obtain an unbiased estimate of the generalization loss of an estimated model. This holds regardless of the relationship between p_* and p'_* .

Discussion

- We have presented two different ways to think about an estimator.
- Under the assumption that we can obtain a test data set \mathcal{D}_{te} , we can obtain an unbiased estimate of the generalization loss of an estimated model. This holds regardless of the relationship between p_* and p'_* .
- If we make additional assumptions about the structure of p_* and the relationship between p_* and p'_* , we can prove various properties of an estimator such as unbiasedness, consistency and expected prediction loss.

Outline

1 Review

2 Assumptions

3 Empirical Analysis

4 Theoretical Analysis

5 Discussion

6 Appendix

Background

■ Joint Probability Distribution:

$$p(\mathbf{A} = \mathbf{a}, \mathbf{B} = \mathbf{b})$$

■ Marginal Probability Distribution:

$$p(\mathbf{A} = \mathbf{a}) = \int_{\mathcal{B}} p(\mathbf{A} = \mathbf{a}, \mathbf{B} = \mathbf{b}) d\mathbf{b}$$

■ Conditional Probability Distribution:

$$p(\mathbf{B} = \mathbf{b} | \mathbf{A} = \mathbf{a}) = \frac{p(\mathbf{A} = \mathbf{a}, \mathbf{B} = \mathbf{b})}{p(\mathbf{A} = \mathbf{a})}$$

Background

■ Chain Rule of Probability:

$$p(\mathbf{A} = \mathbf{a}, \mathbf{B} = \mathbf{b}) = p(\mathbf{A} = \mathbf{a})p(\mathbf{B} = \mathbf{b}|\mathbf{A} = \mathbf{a})$$

■ Expectation:

$$\mathbb{E}_{p(\mathbf{A}=\mathbf{a})}[f(\mathbf{a})] = \int_{\mathcal{A}} f(\mathbf{a})p(\mathbf{A} = \mathbf{a})d\mathbf{a}$$

■ Conditional Expectation:

$$E_{p(\mathbf{B}=\mathbf{b}|\mathbf{A}=\mathbf{a})}[h(\mathbf{b})] = \int_{\mathcal{B}} h(\mathbf{b})p(\mathbf{B} = \mathbf{b}|\mathbf{A} = \mathbf{a})d\mathbf{b}$$

Background

■ Joint Expectation:

$$\begin{aligned} & E_{p(\mathbf{A}=\mathbf{a}, \mathbf{B}=\mathbf{b})} [g(\mathbf{a}, \mathbf{b})] \\ &= \int_{\mathcal{A}} \int_{\mathcal{B}} g(\mathbf{a}, \mathbf{b}) p(\mathbf{A} = \mathbf{a}, \mathbf{B} = \mathbf{b}) d\mathbf{a} d\mathbf{b} \\ &= \int_{\mathcal{A}} \int_{\mathcal{B}} g(\mathbf{a}, \mathbf{b}) p(\mathbf{B} = \mathbf{b} | \mathbf{A} = \mathbf{a}) p(\mathbf{A} = \mathbf{a}) d\mathbf{a} d\mathbf{b} \\ &= \int_{\mathcal{A}} \left(\int_{\mathcal{B}} g(\mathbf{a}, \mathbf{b}) p(\mathbf{B} = \mathbf{b} | \mathbf{A} = \mathbf{a}) d\mathbf{b} \right) p(\mathbf{A} = \mathbf{a}) d\mathbf{a} \\ &= \int_{\mathcal{A}} E_{p(\mathbf{B}=\mathbf{b}|\mathbf{A}=\mathbf{a})} [g(\mathbf{a}, \mathbf{b})] p(\mathbf{A} = \mathbf{a}) d\mathbf{a} \\ &= \mathbb{E}_{p(\mathbf{A}=\mathbf{a})} [E_{p(\mathbf{B}=\mathbf{b}|\mathbf{A}=\mathbf{a})} [g(\mathbf{a}, \mathbf{b})]] \end{aligned}$$

Background

■ Normal/Gaussian Distribution:

$$\begin{aligned} p(X = x | \mu, \sigma^2) &= \mathcal{N}(x; \mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \end{aligned}$$

■ Multivariate Normal/Multivariate Gaussian Distribution:

$$\begin{aligned} p(\mathbf{X} = \mathbf{x} | \mu, \Sigma) &= \mathcal{N}(\mathbf{x}; \mu, \Sigma) \\ &= \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right) \end{aligned}$$