MANIPAL
ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

Linear Regression Models

Segment 4 – Model Diagnostics

Topic 1 – In-Sample Estimation of Prediction Error

Sudarsan N.S. Acharya (sudarsan.acharya@manipal.edu)

Topics

1. Prediction error
2. Train-Test-Validation Split
3. Validation
4. Testing
5. Linear Regression Example

Prediction error

Prediction error

- Recall the prediction problem and the associated error:

Prediction error

- Recall the prediction problem and the associated error:

true population relationship $Y = f(X) + \varepsilon,$

Prediction error

- Recall the prediction problem and the associated error:

$$\text{true population relationship } Y = f(X) + \varepsilon,$$
$$\text{using data build approximation } Y \approx \hat{f}(X).$$

Prediction error

- Recall the prediction problem and the associated error:

$$\text{true population relationship } Y = f(X) + \varepsilon,$$

$$\text{using data build approximation } Y \approx \hat{f}(X).$$

- Prediction error:

Prediction error

- Recall the prediction problem and the associated error:

$$\text{true population relationship } Y = f(X) + \varepsilon,$$

$$\text{using data build approximation } Y \approx \hat{f}(X).$$

- Prediction error:

$$\text{squared error } \left(Y - \hat{f}(X)\right)^2,$$

Prediction error

- Recall the prediction problem and the associated error:

$$\text{true population relationship } Y = f(X) + \varepsilon,$$

$$\text{using data build approximation } Y \approx \hat{f}(X).$$

- Prediction error:

$$\text{squared error } \left(Y - \hat{f}(X)\right)^2,$$

$$\text{absolute deviation } \left|Y - \hat{f}(X)\right|.$$

Prediction error

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Recall the prediction problem and the associated error:

$$\text{true population relationship } Y = f(X) + \varepsilon,$$
$$\text{using data build approximation } Y \approx \hat{f}(X).$$

- Prediction error:

$$\text{squared error } \left(Y - \hat{f}(X)\right)^2,$$
$$\text{absolute deviation } \left|Y - \hat{f}(X)\right|.$$

- The fitted model $\hat{f}$ should minimize the *expected* prediction error on new unseen data.

Prediction error

- Recall the prediction problem and the associated error:

$$\text{true population relationship } Y = f(X) + \varepsilon,$$
$$\text{using data build approximation } Y \approx \hat{f}(X).$$

- Prediction error:

$$\text{squared error } \left(Y - \hat{f}(X)\right)^2,$$
$$\text{absolute deviation } \left|Y - \hat{f}(X)\right|.$$

- The fitted model $\hat{f}$ should minimize the *expected* prediction error on new unseen data.
- How do we calculate the expected prediction error?

Train-Test-Validation Split

Train-Test-Validation Split

- Split data into three groups:

Train-Test-Validation Split

- Split data into three groups: (1) train

Train-Test-Validation Split

- Split data into three groups: (1) train (2) validation

Train-Test-Validation Split

- Split data into three groups: (1) train (2) validation (3) test:

Train-Test-Validation Split

- Split data into three groups: (1) train (2) validation (3) test:

$$\underbrace{\Big[\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \Big]}_{\text{data}}.$$

Train-Test-Validation Split

- Split data into three groups: (1) train (2) validation (3) test:

$$\underbrace{\left[\underbrace{\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(k)}, y^{(k)}\right)}_{\text{train}} \qquad\qquad\qquad\qquad\qquad\qquad \right]}_{\text{data}}.$$

Train-Test-Validation Split

- Split data into three groups: (1) train (2) validation (3) test:

$$\underbrace{\left[\underbrace{\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(k)}, y^{(k)}\right)}_{\text{train}} \quad \underbrace{\left(\mathbf{x}^{(k+1)}, y^{(k+1)}\right), \ldots, \left(\mathbf{x}^{(l)}, y^{(l)}\right)}_{\text{validation}} \qquad \qquad \right]}_{\text{data}}.$$

Train-Test-Validation Split

- Split data into three groups: (1) train (2) validation (3) test:

$$\underbrace{\left[\underbrace{\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(k)}, y^{(k)}\right)}_{\text{train}} \quad \underbrace{\left(\mathbf{x}^{(k+1)}, y^{(k+1)}\right), \ldots, \left(\mathbf{x}^{(l)}, y^{(l)}\right)}_{\text{validation}} \quad \underbrace{\left(\mathbf{x}^{(l+1)}, y^{(l+1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)}_{\text{test}}\right]}_{\text{data}}.$$

Train-Test-Validation Split

- Split data into three groups: (1) train (2) validation (3) test:

$$\underbrace{\Bigg[\underbrace{\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(k)}, y^{(k)}\right)}_{\text{train}} \quad \underbrace{\left(\mathbf{x}^{(k+1)}, y^{(k+1)}\right), \ldots, \left(\mathbf{x}^{(l)}, y^{(l)}\right)}_{\text{validation}} \quad \underbrace{\left(\mathbf{x}^{(l+1)}, y^{(l+1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)}_{\text{test}}\Bigg]}_{\text{data}}.$$

- Build different models ($\hat{f}$) using training data.

Train-Test-Validation Split

- Split data into three groups: (1) train (2) validation (3) test:

$$\underbrace{\Big[\underbrace{\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(k)}, y^{(k)}\right)}_{\text{train}} \quad \underbrace{\left(\mathbf{x}^{(k+1)}, y^{(k+1)}\right), \ldots, \left(\mathbf{x}^{(l)}, y^{(l)}\right)}_{\text{validation}} \quad \underbrace{\left(\mathbf{x}^{(l+1)}, y^{(l+1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)}_{\text{test}}\Big]}_{\text{data}}.$$

- Build different models $(\hat{f})$ using training data.
- Estimate prediction error for different models and pick the best one using validation data.

Train-Test-Validation Split

- Split data into three groups: (1) train (2) validation (3) test:

$$\underbrace{\left[\underbrace{\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(k)}, y^{(k)}\right)}_{\text{train}} \quad \underbrace{\left(\mathbf{x}^{(k+1)}, y^{(k+1)}\right), \ldots, \left(\mathbf{x}^{(l)}, y^{(l)}\right)}_{\text{validation}} \quad \underbrace{\left(\mathbf{x}^{(l+1)}, y^{(l+1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)}_{\text{test}}\right]}_{\text{data}}.$$

- Build different models $(\hat{f})$ using training data.
- Estimate prediction error for different models and pick the best one using validation data.
- Finally, assess the best model's performance using test data.

Validation

Validation

- The validation data $\left(\mathbf{x}^{(k+1)}, y^{(k+1)}\right), \ldots, \left(\mathbf{x}^{(l)}, y^{(l)}\right)$ is used to estimate the generalization error of the different models and choose the best one.

Validation

- The validation data $\left(\mathbf{x}^{(k+1)}, y^{(k+1)}\right), \ldots, \left(\mathbf{x}^{(l)}, y^{(l)}\right)$ is used to estimate the generalization error of the different models and choose the best one.
- Calculate estimated generalization error for each fitted model $\hat{f}$ as

Validation

- The validation data $\left(\mathbf{x}^{(k+1)}, y^{(k+1)}\right), \ldots, \left(\mathbf{x}^{(l)}, y^{(l)}\right)$ is used to estimate the generalization error of the different models and choose the best one.
- Calculate estimated generalization error for each fitted model $\hat{f}$ as

$$\frac{1}{l-k} \sum_{j=k+1}^{l} \left[y^{(j)} - \hat{f}\left(\mathbf{x}^{(j)}\right) \right]^2 .$$

Validation

- The validation data $\left(\mathbf{x}^{(k+1)}, y^{(k+1)}\right), \ldots, \left(\mathbf{x}^{(l)}, y^{(l)}\right)$ is used to estimate the generalization error of the different models and choose the best one.
- Calculate estimated generalization error for each fitted model $\hat{f}$ as

$$\frac{1}{l-k} \sum_{j=k+1}^{l} \left[y^{(j)} - \hat{f}\left(\mathbf{x}^{(j)}\right)\right]^2 .$$

- Choose the model $\hat{f}_{\text{best}}$ with the smallest estimated generalization error.

Validation

- The validation data $\left(\mathbf{x}^{(k+1)}, y^{(k+1)}\right), \ldots, \left(\mathbf{x}^{(l)}, y^{(l)}\right)$ is used to estimate the generalization error of the different models and choose the best one.
- Calculate estimated generalization error for each fitted model $\hat{f}$ as

$$\frac{1}{l-k} \sum_{j=k+1}^{l} \left[y^{(j)} - \hat{f}\left(\mathbf{x}^{(j)}\right)\right]^2 .$$

- Choose the model $\hat{f}_{\text{best}}$ with the smallest estimated generalization error.
- Once we have the best model $\hat{f}_{\text{best}}$, what is the need for the test data?

Testing

Testing

- The error corresponding to the best model $\hat{f}_{\text{best}}$ is typically an *underestimate* of the true generalization error.

Testing

- The error corresponding to the best model $\hat{f}_{\text{best}}$ is typically an *underestimate* of the true generalization error.
- We use the test data $\left(\mathbf{x}^{(l+1)}, y^{(l+1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)$ to calculate an *unbiased* estimate of the true generalization error of the best model:

Testing

- The error corresponding to the best model $\hat{f}_{\text{best}}$ is typically an *underestimate* of the true generalization error.
- We use the test data $\left(\mathbf{x}^{(l+1)}, y^{(l+1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)$ to calculate an *unbiased* estimate of the true generalization error of the best model:

$$\frac{1}{n-l} \sum_{j=l+1}^{n} \left[y^{(j)} - \hat{f}_{\text{best}}\left(\mathbf{x}^{(j)}\right)\right]^2 .$$

Linear Regression Example

Linear Regression Example

- Using the train, validation, and test split of the data, we can come up with a linear regression model with a small prediction error.

Linear Regression Example

- Using the train, validation, and test split of the data, we can come up with a linear regression model with a small prediction error.
- The train data is used to build different linear regression models:

Linear Regression Example

- Using the train, validation, and test split of the data, we can come up with a linear regression model with a small prediction error.
- The train data is used to build different linear regression models: different choice of covariates,

Linear Regression Example

- Using the train, validation, and test split of the data, we can come up with a linear regression model with a small prediction error.
- The train data is used to build different linear regression models: different choice of covariates, interaction between covariates,

Linear Regression Example

- Using the train, validation, and test split of the data, we can come up with a linear regression model with a small prediction error.
- The train data is used to build different linear regression models: different choice of covariates, interaction between covariates, transformation of variables,

Linear Regression Example

- Using the train, validation, and test split of the data, we can come up with a linear regression model with a small prediction error.
- The train data is used to build different linear regression models: different choice of covariates, interaction between covariates, transformation of variables, regularization etc.

Linear Regression Example

- Using the train, validation, and test split of the data, we can come up with a linear regression model with a small prediction error.
- The train data is used to build different linear regression models: different choice of covariates, interaction between covariates, transformation of variables, regularization etc.
- The validation data is used to estimate the generalization error of the different models leading to the selection of the best model.

Linear Regression Example

MANIPAL ACADEMY of HIGHER EDUCATION (Institution of Eminence Deemed to be University)

- Using the train, validation, and test split of the data, we can come up with a linear regression model with a small prediction error.
- The train data is used to build different linear regression models: different choice of covariates, interaction between covariates, transformation of variables, regularization etc.
- The validation data is used to estimate the generalization error of the different models leading to the selection of the best model.
- The test data is used to obtain an unbiased estimate of the generalization error of the best model and assess its performance.

Linear Regression Example–Continued

Linear Regression Example–Continued

- We will simulate a dataset with 1000 samples:

Linear Regression Example–Continued

- We will simulate a dataset with 1000 samples: 750 train,

Linear Regression Example–Continued

- We will simulate a dataset with 1000 samples: 750 train, 150 validation,

Linear Regression Example–Continued

- We will simulate a dataset with 1000 samples: 750 train, 150 validation, and 100 test.

Linear Regression Example–Continued

- We will simulate a dataset with 1000 samples: 750 train, 150 validation, and 100 test.
- The population model that we will simulate is $Y = 1 + 2X_1 + 9X_2 + \varepsilon$, where

Linear Regression Example–Continued

- We will simulate a dataset with 1000 samples: 750 train, 150 validation, and 100 test.
- The population model that we will simulate is $Y = 1 + 2X_1 + 9X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and

Linear Regression Example–Continued

- We will simulate a dataset with 1000 samples: 750 train, 150 validation, and 100 test.
- The population model that we will simulate is $Y = 1 + 2X_1 + 9X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.

Linear Regression Example–Continued

- We will simulate a dataset with 1000 samples: 750 train, 150 validation, and 100 test.
- The population model that we will simulate is $Y = 1 + 2X_1 + 9X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.

Model	Train Error	Validation Error	Test Error

Linear Regression Example–Continued

- We will simulate a dataset with 1000 samples: 750 train, 150 validation, and 100 test.
- The population model that we will simulate is $Y = 1 + 2X_1 + 9X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.

Model	Train Error	Validation Error	Test Error
Y~X_1	10.9	10.6	11.0

Linear Regression Example–Continued

- We will simulate a dataset with 1000 samples: 750 train, 150 validation, and 100 test.
- The population model that we will simulate is $Y = 1 + 2X_1 + 9X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.

Model	**Train Error**	**Validation Error**	**Test Error**
Y~X_1	10.9	10.6	11.0
Y~X_2	5.7	5.2	5.7

Linear Regression Example–Continued

- We will simulate a dataset with 1000 samples: 750 train, 150 validation, and 100 test.
- The population model that we will simulate is $Y = 1 + 2X_1 + 9X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.

Model	**Train Error**	**Validation Error**	**Test Error**
`Y~X_1`	10.9	10.6	11.0
`Y~X_2`	5.7	5.2	5.7
`Y~X_1+X_2`	5.2	4.8	5.3

Linear Regression Example–Continued

- We will simulate a dataset with 1000 samples: 750 train, 150 validation, and 100 test.
- The population model that we will simulate is $Y = 1 + 2X_1 + 9X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.

Model	Train Error	Validation Error	Test Error
`Y~X_1`	10.9	10.6	11.0
`Y~X_2`	5.7	5.2	5.7
`Y~X_1+X_2`	5.2	4.8	5.3
`Y~X_1+X_2+I(X_1^2)+I(X_2^2)`	5.2	4.8	5.3

Summary

Summary

- Describe how model performance can be studied using prediction error.

Summary

- Describe how model performance can be studied using prediction error.
- Differentiate between train, test, and validation parts of the data for assessing model performance