# PREDICTION

## MODEL BUILDING FOR PREDICTION

Tamer Çetin

# A Framework for Prediction: Motivation

- Assume we have two prediction problems

1. You have a car and you want to sell it in the near future.
   a. You want to know what price you can expect from the sale.
   b. You may also want to know what you could expect if you were to wait one more year and sell it then.

- As long as you have data on used cars with their age and other features you can predict future price using several regression models in different functional forms.
- *** The problem is to select the right regression model that would give the best prediction!

# A Framework for Prediction: Motivation

2. You want to predict the next year sales for the company you are working.

* You have historical data on the sales with their right-hand side variables such as price, costumer profiles, and product features.

- *** The problem is how to use data to make sales predictions that will be as close to actual sales as possible!

# A Framework for Prediction: Motivation

- In the previous lectures, we introduced the logic of predictive data analysis and its most widely used methods:
  - Linear and logistic regressions.
- As of today, we focus on predicting a target variable $y$ with the help of predictor variables $x$.
- The main logic of prediction is:
  - estimating a model to capture the patterns of association between $y$ and $x$ in existing (original) data
  - and then using that model to predict $y$ for observations in the prediction situation (in the live data), in which we observe $x$ but not $y$.
- The ultimate goal of predictive analysis is to find the model that would give the best prediction in the live data by using the information from the original data.

# A Framework for Prediction: Flow

- We will discuss:
  - the distinction between various types of prediction such as quantitative predictions, probability predictions, and classification with a specific focus on quantitative predictions,
  - point prediction versus interval prediction,
  - the components of prediction error,
  - how to find the best prediction model that will likely produce the best fit/smallest prediction error in the live data using the observations in the original data.
- We will introduce:
  - loss functions in general,
  - mean squared error (MSE) and its square root (RMSE) in particular in order to evaluate predictions.

# A Framework for Prediction: Flow

- We will discuss three ways of finding the best predictor model:
  - using all data and the Bayesian Information Criterion (BIC) as the measure of fit,
  - using training-test splitting of data,
  - and using k-fold cross-validation, which is an improvement on the training-test split.
- We will:
  - discuss those assessment methods and try to improve the external validity of predictions, if possible,
  - and conclude the lecture by discussing what machine learning means.

# A Framework for Prediction: Learning Outcomes

- After going through this lecture, you should be able to:
  - identify situations where you want to and can predict $y$ with the help of $x$,
  - understand the concept of prediction error and its components,
  - carry out point predictions and interval predictions of quantitative outcomes with linear regression,
  - understand the role of model complexity in overfitting the original data,
  - use the BIC and k-fold cross-validation to find the model that best fits the population, or general patterns, behind the original data,
  - assess external validity with the help of domain knowledge, and, if possible, with additional analysis of the original data.

# Prediction Basics: Terminology

- In data science analysis, prediction means assigning a value to $y$, the target variable or outcome variable for a target observation or more target observations.
- The value of $y$ is not known for the target observations, but the value of one or more $x$ variables is known.
- In this context, those $x$ variables are called predictors.
- Data analysis with the aim of prediction is called predictive data analysis or predictive analytics.

# Prediction Basics: Terminology

- The basic logic of prediction is as follows.
- We have data with observations on both $y$ and $x$.
- We call this data the original data, which does not include the target observations for which we want to make the prediction.
- We use the original data to uncover the patterns of association between $y$ and $x$ that will be used in prediction.
- To uncover those patterns, we specify and estimate a model.
- With the help of that model, we then predict the value of $y$ for the target observations, for which we can observe $x$ but not $y$.
- We call data including target observations the live data.
- In short, we uncover patterns in the original data that we use for prediction in the live data.

# Prediction Basics: Mathematics

- Patterns of associations are expressed as a function giving a value for $y$ if we plug in values for $x$.

$$\hat{y}_j = \hat{f}(x_j)$$

- For a target observation $j$ in the live data, we observe the values of the predictor variables $x_j$.
- If we know the function all we need to do is feed those $x_j$ values into it.
- The abstract notation for the variable and observation we want to predict is $y_j$.
- The specific predicted value we call is $\hat{y}_j$.
- The abstract notation for the function we use for prediction is $f$.
- The notation for the specific function we use is $\hat{y}_j$.

# Prediction Basics: Mathematics

- $\hat{y}_j = \hat{f}(x_j)$ is called predictive model.
- There are many different predictive models that give some value of $y$ if we plug in values of $x$ but we need the model with the best prediction.
- For this aim, we use the original data to find a model that will give the best prediction in the live data.
- An example for a predictive model is the linear regression with a specific choice of functional forms, for instance, with or without interactions as follows:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

# Prediction Basics: Mathematics

- In the abstract, the linear regression $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is a model for the conditional expected value of $y$, and it has coefficients $\beta$.
- We need estimated coefficients ($\hat{\beta}$) and actual $x$ values ($x_j$) to predict an actual value $\hat{y}$ as in the following equilibrium.

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + \ldots$$

- The function or model is $f(x_1, x_2, \ldots) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$,
- And its estimated version with specific values for its coefficients is $\hat{f}(x_{1j}, x_{2j}, \ldots) = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + \cdots$.

# Prediction Basics: Simple Example

- Predict ice cream sale with information day of the week, month of the year, highest temperature on that day.
- $y$ is sales, $j$ is the target day, the $x$ are binary variables for day of the week, month of the year, and temperature.
- We have access to daily data from the last three years with sales, temperature, day, and month variables.
- So this is our original data.
- One way to do prediction is:
  - estimate a linear regression using all observations in the original data to find $\hat{\beta}$ coefficient values
  - then plug in the $x_j$ values and multiply them by the estimated $\hat{\beta}$ coefficient values
  - The results are the $\hat{y}_j$.

# Prediction Basics: Simple Example

- In practice, instead of predicting one simple model, we carry out additional steps when making a prediction or use different prediction models since those other models can provide better results using the same data.
- Note that the fundamental task of predictive analysis is to find the best model.
- We have original data we can use to find the best prediction model.
- Prediction model is the best model in the sense that it gives the best prediction in the live data, but not the original data.

# Types of Prediction

- When we talk about prediction, we basically talk about three different prediction models in general: quantitative prediction, probability prediction, and classification.
- When we predict the value of a quantitative outcome $y$, this would be a quantitative prediction.
- When $y$ is binary;
  - If we predict the probability of $y = 1$ this prediction is called a probability prediction.
  - If we predict whether outcome is 0 or 1, this kind of prediction is called classification.
- In principle, probability predictions and classification models can be applied not only to binary $y$ variables but to qualitative $y$ variables with more than two sub-categories.

# Types of Prediction

- Predicting sales in an ice cream shop?
  - Quantitative prediction
- Predicting whether a loan applicant will repay?
  - Probability prediction

  or
  - Classification
- In both quantitative and probability predictions, because we basically predict a single value of $\hat{y}_j$ this is called point prediction.
- When we predict a probability interval this is called prediction interval (PI) showing the value of $y_j$ with a certain likelihood (e.g. a 95% prediction interval).

# Types of Prediction

- When the original data is a time series data and when we predict the future values of $y$, this is called time series forecasting.
- Note that most of the time the original data is from the past and the live data with the target observations is in the future.
- Basically, forecasts are predictions that use time series data!

# Prediction Error and its Components

- Predicted value $\hat{y}_j$ for target observation $j$
- Actual value $y_j$ for target observation $j$
- Unknown when we make the prediction
- Prediction error
  - $e_j = \hat{y}_j - y_j$
- The ideal prediction error, is zero: our predicted value is right on target.

# Prediction Error and its Components

- The prediction error is:
  - Positive if we overpredict the value: we predict a higher value than actual value.
  - Negative if we underpredict the value: our prediction is too low.
- Whether positive versus negative errors matter more, or they are equally bad, depends on the decision problem.
- Larger in absolute value the further away our prediction is from the actual value.
  - It is smaller the closer we are.
  - It is always better to have a prediction with as small an error as possible.

# Prediction Error and its Components

- Prediction error can be decomposed into three parts:
  - **estimation error**: the difference between the estimated value from the model and the true value from the model
  - **model error**: the difference between the true value from the model and the best predictor value; i.e. we may not have the best model
  - **irreducible error** (idiosyncratic or genuine error): error due to not being able to perfectly estimate all predicted values even if estimation error is zero, and we have the best possible model.

# Prediction Error and its Components

- Estimation error comes from the fact that, with a model $f$, we use original data to find $\hat{f}$.
- When using a linear regression for prediction, this error is due to the fact that we do not know the values of the coefficients of the regression($\beta$), only their estimated values ($\hat{\beta}$).
- The error is captured by the SE of the regression line (ch.9 section 9.3).
  - The smaller the SE of regression line, the smaller estimation error we can expect.
- This tends to happen
  - the larger the sample, the smaller the residual standard deviation (the better the fit),
  - the closer the target observation to the average of $x$ variables,
  - and the more spread the $x$ variables are.

# Prediction Error and its Components

- Model error reflets the fact that we may not selected the best model for our prediction.
- More specifically, instead of model $f$, there may be a better model $g$, which could use the **same** or **different** predictor variables $x$ from the **same** original data.
  - We should have included interactions that we did not, vice versa,
  - Or, maybe, the best model to predict $y$ is not a linear regression but something else.
- Irreducible error stems from the fact we are not able to make a perfect prediction for $y_j$ with the help of the $x_j$ variables even if we find the best model and we can estimate it without any estimation error.
- Maybe, if we had a data with more variables, we could have a better prediction using those additional variables.
- But, with the variables we have, we cannot do anything about this error.
- This is why it is called irreducible error.

# Prediction Error and its Components

- Note that we use prediction interval and PI depends on the standard prediction error and SPE include two of those three prediction errors:
  - Estimation error and irreducible error.
- It includes model error within the irreducible error component.
- So, it measures uncertainty in the context of specific model used to compute model error using the results from that specific model.
- The only way to reduce prediction error is to reduce estimation and model errors.
- There is always some, often large, irreducible part of prediction error.
  - To reduce estimation error, we should try to use as many observations as possible.
  - To reduce model error, we should find the best possible model.

# The Loss Function

- Recall that we want as small a prediction error as possible in prediction.
- Prediction error forms the basis of measuring the quality of prediction models.
  - Large error is always worse than small one.
  - Direction of error also matters.
    - Overpredicting
    - Underpredicting
- Ultimately, both magnitude and direction of prediction error matter because predictions lead to decision having consequences.

# The Loss Function

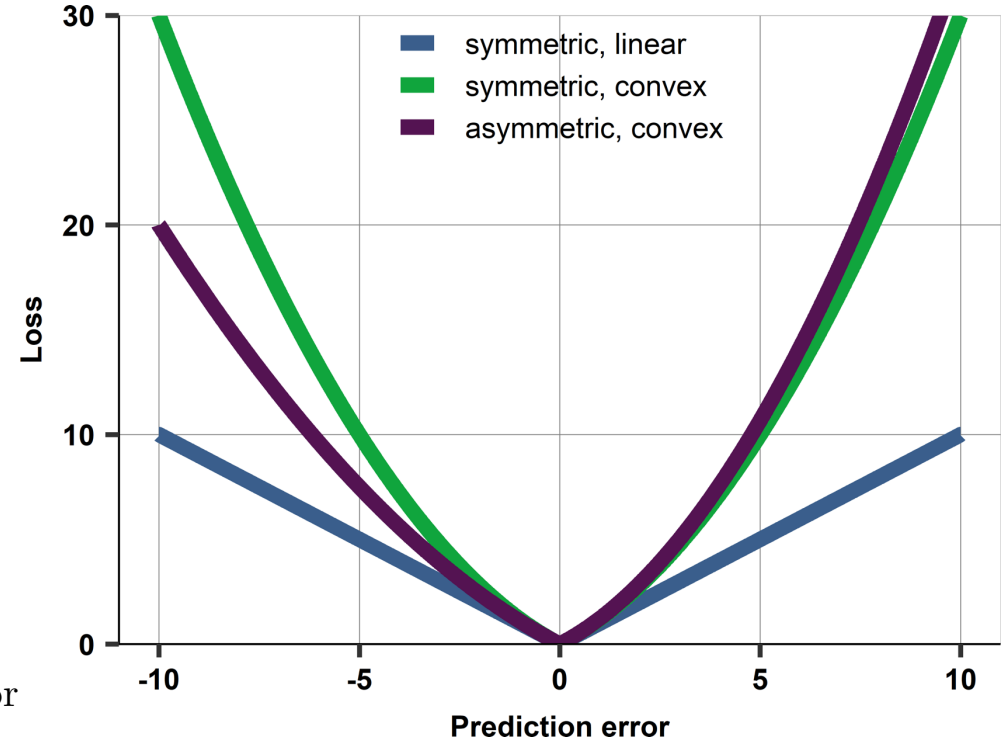
- For instance;
  - If we underpredict ice cream sales, we will produce less ice cream.
    - Less profit
    - Bad reputation
  - If we overpredict ice cream sales, we will produce more than needed.
    - Less profit
- Loss function is defined as a loss value that we can attach to the consequences of prediction error that we incur due to decisions based on bad prediction.
- This idea is formulated in a loss function translating prediction error into a number that makes more sense for business decisions and their consequences.

# The Loss Function

- Typically, even though we have more than one target observation loss function takes prediction error for each of those target observations and generates a single number.
- Loss function helps rank predictions.
- Of course, we could rank predictions using prediction errors but this would be only linear loss function.
- Actually, a loss function may be a non-linear function of the errors.
- If we take into account all the values for each target observation, it would be very hard to quantify loss function or all the decisions and their results.

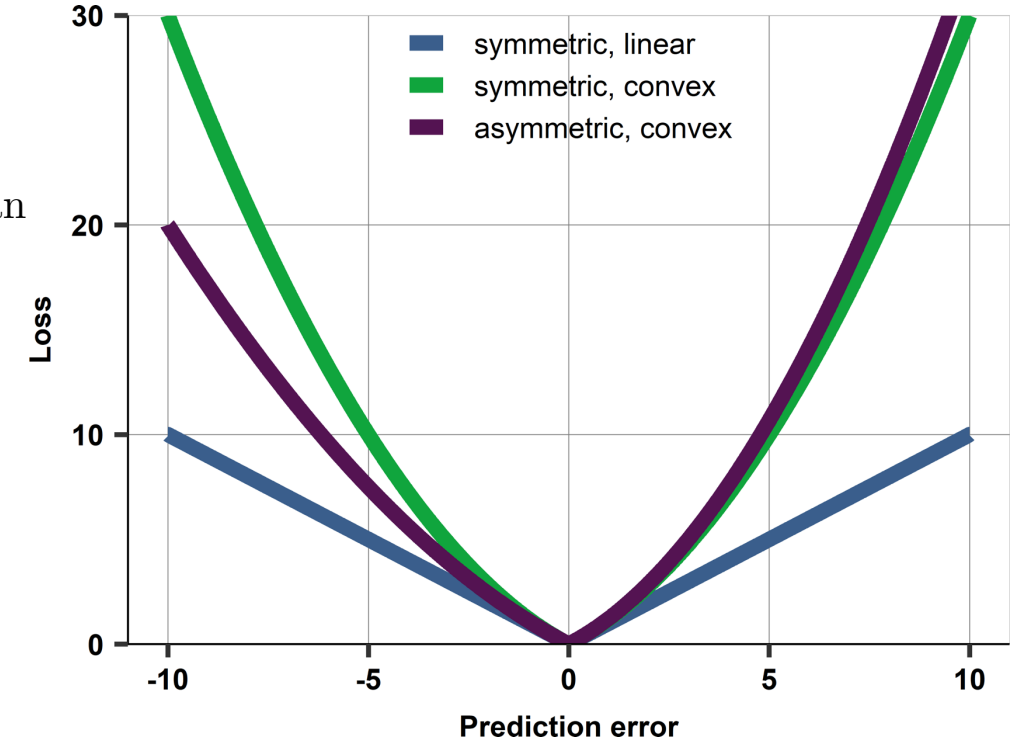
# The Loss Function

- For that reason, we use some general properties of a loss function:
  - symmetry vs asymmetry
  - linearity vs convexity
- Symmetric loss functions add the same value to positive and negative errors of the same magnitude.
- Asymmetric loss functions attach different values to errors of the same magnitude on different sides.
- Linear loss functions add proportionally larger loss to an error, regardless of its magnitude.
- Conversely, convex loss functions attach a disproportionally larger loss to an error that is larger.
- Linearity vs convexity is about the magnitudes of error in quantitative or probability prediction.



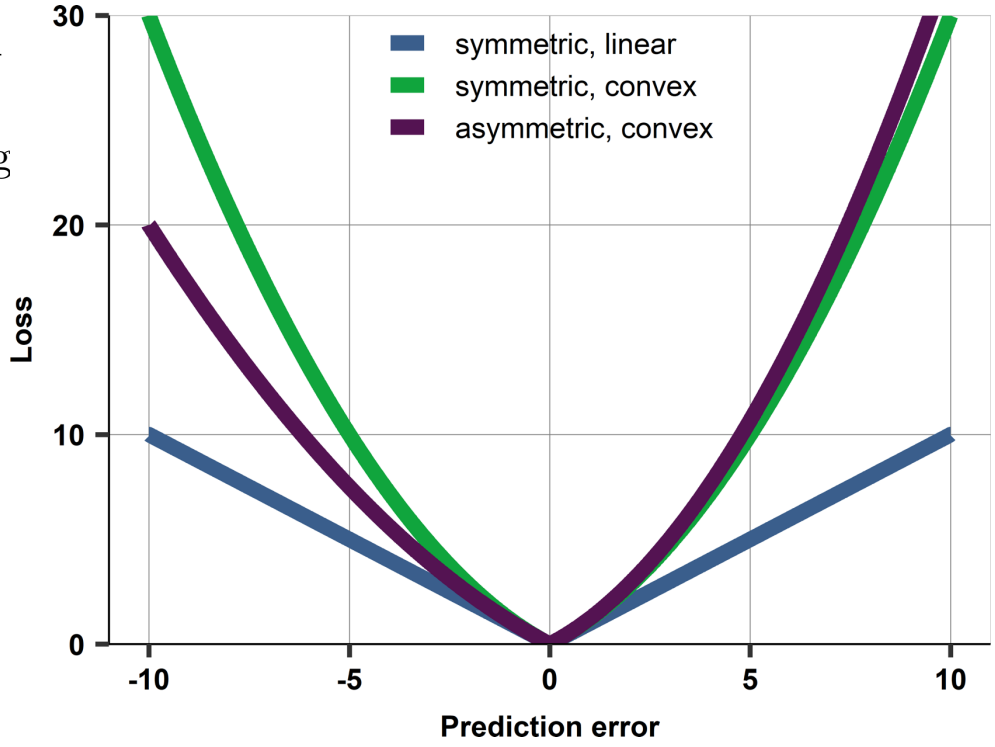
# The Loss Function

- Prediction for Ice Cream Sales under Severe Competition in Negative Error
  - Loss function will be asymmetric with larger loss attached to a negative prediction error than to a positive prediction error of the same size.
  - So, loss function will be convex in both directions:
    - Larger prediction error would have disproportionally larger consequences than a smaller prediction error.
  - Reputation is more important than production costs
  -



# The Loss Function

- Prediction (forecasting) inflation by central banks to make decision on monetary policy
  - The larger the prediction error, the larger wrong policies, and thus the larger the social costs.
  - Arguably, larger errors may induce costs that are disproportionally larger
  - Small prediction errors may not divert monetary policy from the best choice at all.
  - So, loss function would be either symmetric or asymmetric but most likely convex.
  - In practice, it is very difficult to attach a loss function to the prediction error based on the actual consequences of that error.
  - Instead, we use a generic loss function that we think represents the most important properties of prediction case.



# Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

- Most widely used generic loss function is the squared loss defining the loss as the square of prediction error.

$$L(e_j) = e_j^2 = (\hat{y}_j - y_j)^2$$

- It is a symmetric and convex loss function.
- When we have a prediction for multiple observations, we aggregate those squared losses across them to calculate prediction error.

$$MSE = \frac{1}{j}\sum_{j=1}^{j}(\hat{y}_j - y_j)^2$$

- The most widely used aggregation is the mean
- MSE is the average squared loss across several target observations.

# Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

- Often, we use its square root (RMSE) because RMSE is easier to interpret.

$$RMSE = \sqrt{\overline{MSE}} = \sqrt{\frac{1}{j}\sum_{j=1}^{j}(\hat{y}_j - yj)2}$$

- MSE is actually like the numerator of R-Squared formula for linear regression.
- The difference is;
  - that R-Squared is calculated using predicted values in the original data, which is used for linear regression
  - while MSE as a loss function is defined for the live data, although it can be computed for the original data as well.

# Bias and Variance of Predictions

- Earlier, we talked about three components of prediction error: estimation, model, and irreducible errors.
- Analogously, MSE can be decomposed into these three components as estimation, model, and irreducible MSEs.
- This decomposition is useful because it helps guide our thinking of how to find the best model.
- This model should give us as precise estimates as possible and it should be as close to the theoretically best model as possible.
- Note that with the data we have, we cannot do anything with the third component.
- Moreover, the first two can traded off to find the best model.

# Bias and Variance of Predictions

- The best model should be with the smallest estimation MSE.
- But, the best model should be fine with a little bit more estimation error if this will lead to a smaller model error.
- This is the trade off for the best model.
- We call that the bias-variance tradeoff, which is technically another way of the decomposition of MSE.
- The bias of prediction is the average of its prediction error.
- An unbiased prediction produces zero error on average across multiple predictions.
- Bias can be positive and negative when prediction is biased.

# Bias and Variance of Predictions

- The variance of prediction describes how prediction varies around its average value when multiple predictions are made.
- It is the variance of prediction error around its average value.
- That average can be zero or not, so prediction may be unbiased or biased.
- The variance is zero if prediction error is the same for all predictions.
- The variance is higher the larger the spread of specific predictions around the average prediction.

# Bias and Variance of Predictions

- Accordingly, MSE is the sum of the squared bias and prediction variance if we decompose MSE as follows.

$$MSE = \frac{1}{j}\sum_{j=1}^{j}(\hat{y}_j - y_j)^2$$

$$= \left(\frac{1}{j}\sum_{j=1}^{j}(\hat{y}_j - y_j)\right)^2 + \frac{1}{j}\sum_{j=1}^{j}(y_j - \bar{\hat{y}})^2$$

$$= Bias^2 + Variance$$

- Decomposition shows that a biased prediction with small variance may be better than an unbiased prediction with a large variance.

# Bias and Variance of Predictions

- This decomposition is commonly known as the bias-variance trade-off decomposition for mean squared error (MSE).
- It explains the relationship between bias, variance, and the overall error in prediction models.
- We will use the decomposition to explain why a biased prediction with small variance may sometimes be preferable to an unbiased prediction with large variance.

# Bias and Variance of Predictions

- First Term: $\left(\frac{1}{j}\sum_{j=1}^{j}(\hat{y}_j - y_j)\right)^2$
  - Represents the bias component of the MSE,
  - Measures the squared difference between the predicted values ($\hat{y}_j$) and the true values ($y_j$) for each sample, averaged over all samples, and
  - Captures the systematic error or deviation of the model's predictions from the true values.
- Second Term: $\frac{1}{j}\sum_{j=1}^{j}(y_j - \overline{\hat{y}})^2$
  - Represents the variance component of the MSE,
  - Measures the squared difference between the true values ($y_j$) and the average of the predicted values ($\overline{\hat{y}}$), averaged over all samples, and
  - Captures the variability or spread of the model's predictions around their mean.

# Bias and Variance of Predictions

- Bias:
  - A biased prediction implies that the model consistently overestimates or underestimates the true values.
  - A small bias suggests that the model's predictions are on average relatively close to the true values, even though they may consistently deviate from them.
- Variance:
  - High variance indicates that the model's predictions exhibit significant variability or sensitivity to the training data.
  - A large variance suggests that the model is heavily influenced by the specific instances in the training set, leading to potential overfitting.
  - This means the model may perform well on the training data but poorly on new, unseen data.

# Bias and Variance of Predictions

- In some cases, it can be preferable to have a slightly biased prediction with small variance rather than an unbiased prediction with large variance.
- This is because a small bias implies that the model's predictions are consistently close to the true values on average, while a small variance indicates that the predictions are relatively stable and less sensitive to the training data.
- This combination can result in a model that generalizes well to unseen data, providing more reliable and robust predictions.
- However, it's important to note that the optimal trade-off between bias and variance depends on the specific problem, dataset, and the desired characteristics of the prediction model.
- Different scenarios may require different levels of bias and variance trade-off.
- The goal is to strike a balance that minimizes the overall error and achieves good generalization performance on unseen data.

# The Task of Finding the Best Model

- Note that the main difficulty is the difference between original data and live data.
- All we have is original data, but the best prediction is actually about live data we do not have.
- So, we will never know what the best prediction model is during prediction time.
- The only option is to compare the predictions of various models using only original data.
- But, again, we want to know the best model in terms of live data, but not original data.

# The Task of Finding the Best Model

- An important dimension is model complexity.
- Model in terms of live data, but not original data.
- In a linear regression model, more complex model has more coefficients because:
  - this model maybe has more $x$ variables
  - or more complicated functional form with the same variables
  - or more interactions.
- Finding the best model involves finding the right degree of complexity.

# The Task of Finding the Best Model

- See the results from used car value prediction.
- Five models we considered were five linear regressions of increasing level of complexity.

| Variables | (1) Model 1 | (2) Model 2 | (3) Model 3 | (4) Model 4 |
|---|---|---|---|---|
| age | −1,530.09 | −1,149.22 | −873.47 | −836.64 |
| age squared | 35.05 | 27.65 | 18.21 | 17.63 |
| odometer | | −303.84 | −779.90 | −788.70 |
| odometer squared | | | 18.81 | 19.20 |
| type: LE | | | 28.11 | −20.48 |
| type: XLE | | | | 301.69 |
| type: SE | | | | 1,338.79 |
| condition: like new | | | | 558.67 |
| condition: excellent | | | 176.49 | 190.40 |
| condition: good | | | 293.36 | 321.56 |
| cylinder_6 | | | | −370.27 |
| dealer | | | 572.98 | 822.65 |
| Constant | 18365.45 | 18860.20 | 19431.89 | 18963.35 |
| Observations | 281 | 281 | 281 | 281 |
| R-squared | 0.847 | 0.898 | 0.913 | 0.919 |

# The Task of Finding the Best Model

- Recall the best model is the one that fits the live data best.
- But there is no way to tell how predictions would fit the live data.
- We need to use the original data to find the best model.
- So, one measure could be R-Squared to find the best model using the original data.
- Remember R-Squared is just like MSE, except it is computed using predictions within the original data and it is divided by the variance of $y$ and subtracted from one:

$$= 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

# The Task of Finding the Best Model

- So we can rank the models by their R-squared.
- This ranking would be the same as ranking by their MSE.
- But, the best R-squared in the original data does not give the best MSE in the live data.
- This means that more complexity can reflect the patterns in the original data but not the live data.
- Consider the results from Model 1 and Model 2 in Table.

# The Task of Finding the Best Model

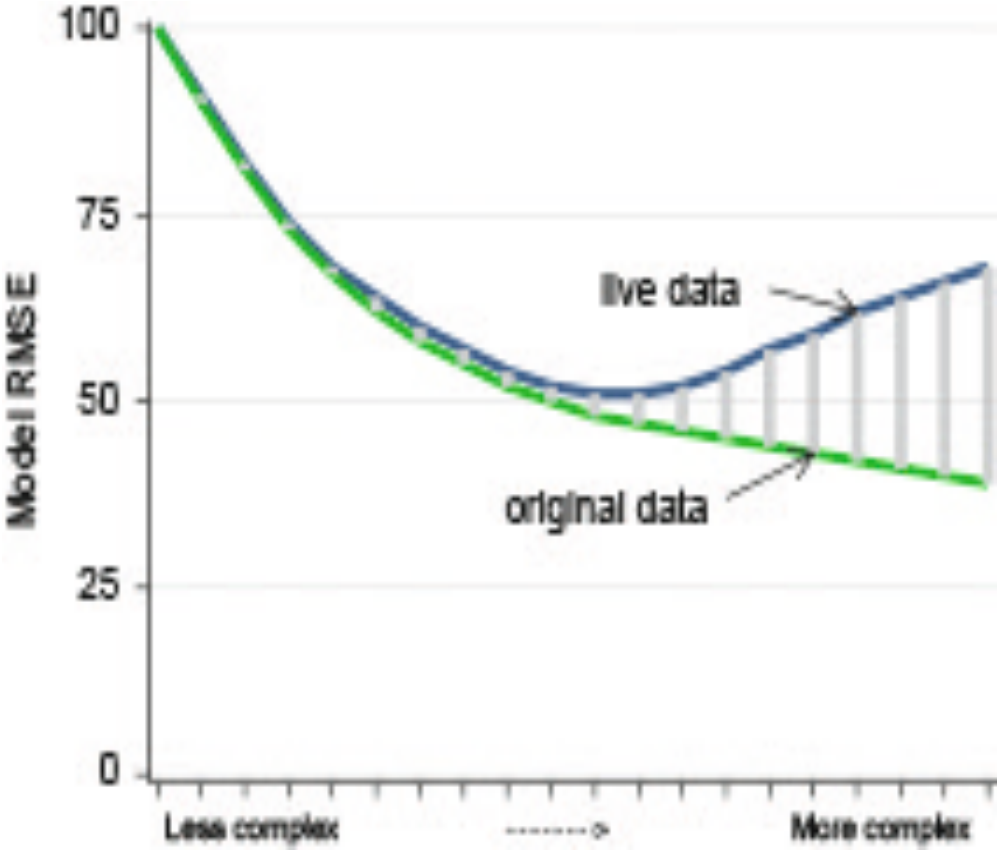
- Model 1 can give a worse fit in the live data than Model 2 in two ways.
  - First, Model 1 might give a worse fit in both original and live data.
    - In this case, we infer that Model 1 underfits the original data.
    - This is quite straightforward.
  - Second, Model 1 might give a better fit in the original data, but a worse fit in the live data.
    - In this case, Model 1 overfits the original data.
    - This is less straightforward, but more important.
  - Overfitting is serious threat because all we have is the original data to find the best model.

# The Task of Finding the Best Model

- Overfitting means that the prediction model gives the best fit in the original data but actually it is not best fit or model for the live data.
- How does that work?
- Note that more complex models always give higher R-squared in the original data.
- This is because a more complex model tends to find more detailed patterns in the original data.
- The problem is that some of those patterns represent the patterns only in original data but not live data.
- This means that a very complex model can lead to a very good fit in original data, but with a high likelihood, this would overfit original data.

# The Task of Finding the Best Model

- Figure shows underfitting and overfitting based on RMSE and its relation to model complexity.
- Blue curve represents RMSE in live data green curve shows RMSE in original data.
- Less complex models cause bad fit in both original and live data.
- More complex models generate better fit for original data but worse fit for live data.
- So more complex models overfit original data.
- In other words, they underfit live data.



# The Task of Finding the Best Model

- Overfitting:
  - In terms of original data and population original data represents, fitting patterns in original data are not there in population.
  - In terms of original data and live data that we will see in the future, fitting patterns in original data are not there in live data.
- What we have learned:
  - The best model is the best prediction for live data.
  - The only data we can use to find the best model is original data.
  - The model that gives the best prediction (best fit) in original data may not give the best prediction (best fit) in live data.
  - Overfitting meant achieving a good fit original data by finding patterns that are specific to original data only and would not achieve a good fit in live data.

# Finding the Best Model by the Best Fit and Penalty: The BIC

- Now, we will introduce three methods to select the best model that is the best in terms of prediction for the general pattern represented by original data and its external validity.
- The solution to finding the best model is using all the original data for estimation and finding the model that fits original data best but measuring fit in a way that discourages overfitting.
- A measure of fit includes a penalty for model complexity.
- A measure of fit has two components:
  - One part is measuring fit.
  - Other component is measuring complexity.

# Finding the Best Model by the Best Fit and Penalty: The BIC

- Comparing models with different fits and different complexities at the same time, those two components are traded off to rank the models.
- There are several measures of fit that includes a penalty term for model complexity.
- They differ in how they trade off fit with complexity.
- Adjusted R-squared, AIC, and BIC or Schwarz Criterion.
- They are included in the regression results of most statistical software packages.

# Finding the Best Model by the Best Fit and Penalty: The BIC

- Like others, BIC measures how far the predicted values are from actual values:
  - the lower the BIC value, the better the expected fit!
- So, a lower BIC implies a better fit, a less complex model, or both!
- In many cases, BIC can help avoid severe overfitting and pick the best model, or at least one of the best models.
- It is easy to calculate and good enough to answer to fitting problem.

# Finding the Best Model by the Best Fit and Penalty: The BIC

- However, BIC is based on strict assumptions.
- Also, it is calculated using original data so if external validity is low, BIC is not a good measure of fit.
- It is tuned to measure fit in original data and penalize complexity.
- This would create problems for live data due to the lack of external validity.
- Thus, it is an ideal way to select the best model.

# Finding the Best Model by Training and Test Samples

- Splitting data into training and test sets is a popular method for measure of fit and.
- We still use original data but initiate original versus live data problem in a more direct way using training and test datasets to measure fit by penalizing complexity.
- We use training set as if it were all original data we had.
- We use all data for EDA that we use to specify regression models and then we estimate the regression model(s) using the training set only.
- Then we use test set to evaluate the predictions.

# Finding the Best Model by Training and Test Samples

- We pretend that we can observe $x$ in this data but not $y$, and we use the model estimates from training set to predict $\hat{y}$ for each observation in test set.
- Then we compare those predictions or predicted values with actual $y$ values to calculate the prediction error.
- After finding prediction errors, we can calculate measure of fit using MSE, RMSE, and/or R-squared.
- The model giving the best fit to test set is the best model.
- Finally, once the best model is found, we re-estimate that model using the entire original data and we use this last estimated model for prediction.

# Finding the Best Model by Training and Test Samples

- Basically, training a model means all the steps of data analysis leading to prediction:
  - Selection the $x$ variables, transforming the $y$ and/or some of the $x$ variables, specifying the model, and estimating it.
- Testing means:
  - predicting $y$ values using the model that was previously trained and evaluating that prediction.
- This testing is different from hypothesis testing!
- Mostly, the size of training sample is 75, 80, or 90% of original data and testing set is the remaining.
- The guiding principle is to have test set as similar to live data as possible.

# Finding the Best Model by Training and Test Samples

- Basically, training a model means all the steps of data analysis leading to prediction:
  - Selection the $x$ variables, transforming the $y$ and/or some of the $x$ variables, specifying the model, and estimating it.
- Testing means:
  - predicting $y$ values using the model that was previously trained and evaluating that prediction.
- This testing is different from hypothesis testing!
- Mostly, the size of training sample is 75, 80, or 90% of original data and testing set is the remaining.

# Finding the Best Model by Training and Test Samples

- The selection of observations in train-test split is not easy.
- The guiding principle is to have test set as similar to live data as possible.
- However, we know little about what live data will look like in the future.
- More importantly, we do not know whether and how the patterns of association between predictors and outcome will differ from original data and the prediction results from it.
- For that reason, we usually do random sampling.
- Training sample is a random sub-sample of original data; test sample is the remaining sub-sample, which is still random.

# Finding the Best Model by Training and Test Samples

- Another issue with train-test split method is that we use less observations than the ones in original data to find the best model for prediction during the training set analysis.
- Note that when fewer observations are used estimation error is larger and when estimation error is larger prediction error tends to be larger.
- Even though we go back and use all the observations in original data in prediction after finding the best estimation model using training data and this avoids the issue for prediction, higher estimation error may still matter to find the best model.
- This is why we use more observations in training set: to minimize the issue.

# Finding the Best Model by Training and Test Samples

- The training-test method does not really avoid overfitting original data.
- True, it avoids overfitting training sample as the predictions are evaluated without using observations from that sample.
- But it may very well overfit test sample.
- In fact, test sample may have patterns in it that are specific to test set and not present in live data just as the training sample would have such patterns, which are not present in test sample.
- So, overfitting test sample is a likely problem.

# Finding the Best Model by Training and Test Samples

- Another issue is more fundamental: external validity!
- All we have is original data.
- With a clever method, we might be able to find the best model for the population, or general pattern, represented by original data
- That way we could avoid overfitting patterns that are present in original data but would not not be there in population represented by original data.
- Whether the model that we find this way is also the best model for live data requires external validity that we will talk about later.

# Finding the Best Model by Cross-Validation

- Using a test set to evaluate predictions:
  - rewards models fitting patterns that are in the test set and
  - penalizes models fitting patterns that are not in the test set.
- This way, we can avoid overfitting in the training data.
- At same time, we may fit a model on patterns that are specific to the test set.
- An improved version of train-test splitting is called k-fold cross-validation (CV).
- CV splits data into four or five samples (k-folds) mostly and looks at the average fit across all the test sets.

# Finding the Best Model by Cross-Validation

- Each observation appears exactly once in a test set and $k-1$ times in the training tests as the remaining sub-samples or other k-folds.
- A training set-test set combination is called fold.
- A k-fold CV works with $k$ folds.
- For example, with a 10-fold CV, we split data into ten subsets in a random way, so each subset consists of 10% of the observations.
- Then we take each of the ten 10% samples as a test set and define the corresponding training set as the other 90%.

# Finding the Best Model by Cross-Validation

- Each observation appears exactly once in a test set and $k-1$ times in the training tests as the remaining sub-samples or other k-folds.
- A training set-test set combination is called fold.
- A k-fold CV works with $k$ folds.
- For example, with a 10-fold CV, we split data into ten subsets in a random way, so each subset consists of 10% of the observations.
- Then we take each of the ten 10% samples as a test set and define the corresponding training set as the other 90%.
- There are ten folds: ten training-test combinations.

# Finding the Best Model by Cross-Validation

- The training-test method explained earlier is carried out ten times to calculate the loss function as MSE or RMSE ten times.
- The overall fit of the model is measured by the average of the ten MSE/RMSE values.

$$MSE_{CV(_k)} = \frac{1}{k}\sum_{i=1}^{k} MSE$$

$$RMSE_{CV(_k)} = \sqrt{\frac{1}{k}\sum_{i=1}^{k} MSE}$$

# Finding the Best Model by Cross-Validation

- CV approach decreases the role of a particular test set in determining the MSE and thus the rusk of overfitting the patterns of a particular test set.
- A further improvements would repeat CV many times, each for a different random partition of the dataset, and compute the overall average MSE.
- This approach decreases the problem of overfitting a particular test set by selecting a larger number of different test sets.
- Thus, CV helps with the second issue we described above: overfitting the test sample.
- As a result, CV is the most widely used method to find the best model in predictive analytics.

# External Validity and Stable Patterns

- Using CV, BIC, and/or R-Squared is the first step to find the model that is the best in predicting $y$ in the population or general pattern represented by original data.
- The second step is the hardest part to find the best model that will give the best prediction in live data.
- External validity is about whether and to what extent we can generalize findings or patterns in the population or live data.
- Or, external validity means to what extent prediction errors are similar in live data and in population.
- Accordingly, external validity is high when the model that produces the best prediction in population or live data.
- Most importantly, external validity of a prediction is high if the fit (e.g., RMSE) is similar in the two.

# External Validity and Stable Patterns

- For high external validity, we require that the patterns of association between $y$ and $x$ are similar in the two worlds:
  - The population, or general patterns, behind the original data,
  - and the population, or general patterns, behind the live data.
- All models make use of those patterns of association to make a prediction.
- Thus, for a model to be best in both worlds, those patterns of association must be similar in the two worlds.
- This requirement is one of stability:
  - The patterns of association between $y$ and $x$ should be stable across the two worlds.

# External Validity and Stable Patterns

- Because original data is from the past and prediction is usually about the future, this also requires stationary over time.
- Another component of external validity is domain knowledge.
- Domain knowledge means knowledge of the mechanisms behind the patterns of association between $y$ and $x$.
- When external validity is not perfect, there will be uncertainty in prediction, and it is very difficult to quantify external validity or lack of it.

# ML and the Role of Algorithms

- We will conclude this lecture with o note on terminology.
- The term predictive analytics is often used for data analysis whose goal is prediction or machine learning in a more popular term.
- In this sense, ML is an umbrella concept for methods using algorithms to find patterns in data and use them for prediction purposes.
- Algorithm is a set of rules and steps defining how to generate an output (predicted value) using various inputs (variables and observations in data).
- A formula is an example of an algorithm.
  - But not all algorithms can be translated to a formula (bootstrap).

# ML and the Role of Algorithms

- The other important component of ML is computers as machines, which are used to plug data into formula.
- We basically use computers and an algorithm (formula) to learn something about the future predicted value from data.
- In ML literature, understating the patterns of associations between $y$ and $x$ is secondary.
  - This is primary in econometrics!
- ML is interested in prediction of $y$, which should be applicable to the future live data.

# Main Takeaways

- Prediction uses original data with $y$ and $x$ to predict the value of $y$ for observations in live data, in which $x$ is observed but $y$ is not.
- Prediction uses a model that describes the patterns of associations between $y$ and $x$ in original data.
- CV can help find the best model in the population, or general pattern, represented by original data.
- Stability of the patterns of association is needed for a prediction with high external validity.

# Main Takeaways

- The RMSE is the square root of the variance of the residuals.
- It indicates the absolute fit of the model to the data—how close the observed data points are to the model's predicted values.
- Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit.
- As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable.
- Lower values of RMSE indicate better fit.
- RMSE is a good measure of how accurately the model predicts the response, and is the most important criterion for fit if the main purpose of the model is prediction.
- The best measure of model fit depends on the researcher's objectives, and more than one are often useful.
- The above statistics were described for the case of ordinary least squares regression.
- Other regression models, such as mixed or generalized linear models, have alternative statistics or diagnostics for assessing model fit.