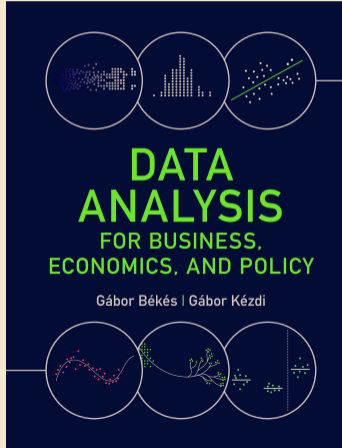


# Békés-Kézdi: Data Analysis, Chapter 13: Prediction setup



## Data Analysis for Business, Economics, and Policy

Gábor Békés (Central European University)  
Gábor Kézdi (University of Michigan)

Cambridge University Press, 2021

[gabors-data-analysis.com](http://gabors-data-analysis.com)

Central European University

Version: v3.1 License: CC BY-NC 4.0

Any comments or suggestions:  
[gabors.da.contact@gmail.com](mailto:gabors.da.contact@gmail.com)

## Prediction setup

- ▶ Original data (what we have)  $\rightarrow$  to build a model
- ▶ Live data (data we do not have yet)
- ▶ Target variable  $Y$  (=dependent variable, response, outcome)
- ▶ Predictor variables  $X$  (= inputs, covariates, features, independent variables)
- ▶ Need to predict value of  $Y$  for target observation  $j$  in live data
  - ▶ Actual value for  $Y_j$  unknown
  - ▶ Value for  $X_j$  known
  - ▶ Maybe more than one target observation
    - ▶ Need predicted value of  $Y$  for each

# Price cars (Case study 1)

## The situation

- ▶ You want to sell your car through online advertising
- ▶ Target is continuous (in dollars)
- ▶ Features are continuous or categorical
  
- ▶ The business question
  - ▶ What price should you put into the ad?

## Price apartments (Case study 2)

### The situation

- ▶ You are planning to run an AirBnB business
  - ▶ Maybe several rooms
- ▶ Target is continuous (in dollars)
- ▶ Features are varied from text to binary
- ▶ The business question
  - ▶ How should you price apartments/houses?

## Predict company's exit from business (Case study 3)

- ▶ Consulting company
- ▶ Predict which firms will go out of business (exit) from a pool of partners
- ▶ Target is binary: exit / stay
  
- ▶ Features of financial and management info
  
- ▶ Business decision
  - ▶ Which firms to give loan to?

## Predictive Analysis: what is new?

- ▶ Most of econometrics focused on the relationship between  $X$  and  $Y$ 
  - ▶ What is the relationship like
  - ▶ Is it a robust relationship – true in the population /general pattern?
- ▶ Now, we use  $x_1, x_2, \dots$  to predict  $y$

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

- ▶ How is this different?
- ▶ We care less about
  - ▶ Individual coefficient values, multicollinearity
  - ▶ We still care about the stability of our results.
  - ▶ Should we care about causality?

# Prediction setup

## Prediction setup

- ▶ Y is quantitative (e.g price)
- ▶ Quantitative prediction
  - ▶ „Regression” problem
- ▶ Y is binary (e.g. Default or nor)
- ▶ Probability prediction
- ▶ Classification problem
  - ▶ Broadly: Y takes values in a finite set of (unordered) classes (survived/died, sold/not sold, car model)
- ▶ Time series prediction (Forecasting)



## Our focus in this course

- ▶ Feature engineering (variable selection)
  - ▶ selecting variables
  - ▶ coding, functional form
- ▶ Model building and prediction
  - ▶ Estimate models
  - ▶ Regressions with a variety of interactions, non-linear functional forms
    - ▶ Remember splines, polynomials
  - ▶ Machine learning methods
    - ▶ Automated model selection under some conditions
- ▶ Model evaluation and selection
  - ▶ Compare models based on some measure of fit
- ▶ Key idea in prediction: systematically combine estimation and model selection

## Regression and prediction

- ▶ Linear regression produces a predicted value for the dependent variable.
  - ▶ Predictions: regressions tell the expected value of  $y$  if we know  $x$ .
- ▶ Linear regression with  $y$ ,  $x_1$ ,  $x_2$ , etc., 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  $\hat{y}$

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

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

# The Prediction Error

- ▶ 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

$$e_j = \hat{y}_j - y_j$$

- ▶ Error = predicted value – actual value

## Prediction Error

- ▶ The ideal prediction error, is zero: our predicted value is right on target. The prediction error is defined by the direction of miss and size.
- ▶ Direction of miss
  - ▶ 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.
- ▶ Size
  - ▶ 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.

## Decomposing the prediction error

- ▶ The prediction error is the difference between the predicted value of the target variable and its actual (yet unknown) value for the target observation:

$$e_j = \hat{y}_j - y_j$$

- ▶ The prediction error can be decomposed into three parts:
  1. **estimation error**: the difference between the estimated value from the model and the true value from the model
  2. **model error**: the difference between the true value from the model and the best predictor value; ie we may not have the best model
  3. **genuine error** (idiosyncratic or irreducible 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.

## Interval prediction for quantitative target variables

- ▶ One advantage of regressions - easy quantify uncertainty of prediction
- ▶ Interval predictions produce ranges to capture the uncertainty of predicted values
- ▶ Interval predictions quantify two out of the three sources of prediction uncertainty: estimation error and genuine (or irreducible) error.
- ▶ They do not include the third source, model uncertainty!
- ▶ The 95% prediction interval (PI) tells where to expect the actual value for the target observation.
  - ▶ The PI for linear regression requires homoskedasticity.

# The Loss function

# Loss Functions

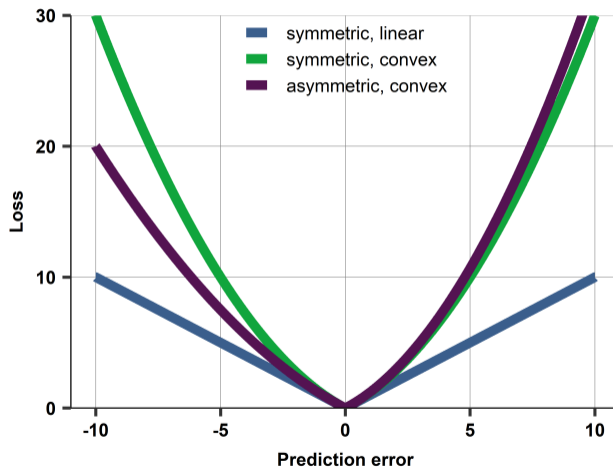
- ▶ Value attached to the prediction error
  - ▶ Specifying how bad it is
- ▶ Loss function determines best predictor
- ▶ Ideally derived from decision problem
  - ▶ Consequence of error is bad decision
  - ▶ Loss due to bad decision
- ▶ Difficult to quantify exact value of loss in practice
- ▶ But this could be super important in some business cases
  - ▶ Even if hard to adjust modeling



# Loss Functions

- ▶ Think about qualitative characteristics of loss function
- ▶ The most important qualitative characteristics of loss functions:
  - ▶ Symmetry
    - ▶ If losses due to errors in opposing direction are similar
  - ▶ Convexity
    - ▶ If twice as large errors generate more than twice as large losses

## Loss Functions of Various Shapes



## Examples 1 – used cars

- ▶ The loss function for predicting the value of our used car depends on how we value money and how we value how much time it takes to sell our car.
- ▶ A too low prediction may lead to selling our car cheap but fast;
- ▶ A too high prediction may make us wait a long time and, possibly, revising the sales price downwards before selling our car.
  
- ▶ What kind of loss function would make sense?

## Examples 2 - creditors

- ▶ Creditors decide whether to issue a loan only to potential debtors that are predicted to pay it back with high likelihood.
- ▶ Two kinds of errors are possible:
  - ▶ debtors that would pay back their loan don't get a loan
  - ▶ debtors that would not pay back their loan get one nevertheless.
- ▶ The costs of the first error are due to missed business opportunity; the costs of the second error are due to direct loss of money.
- ▶ These losses may be quantified in relatively straightforward ways.
- ▶ What kind of loss function would make sense?

# Squared Loss

- ▶  $L(e_j) = e_j^2 = (\hat{y}_j - y_j)^2$
- ▶ The most widely used loss function
  - ▶ Symmetric: Losses due to errors in opposing direction are same
  - ▶ Convex: Twice as large errors generate more than twice as large losses
- ▶ **Business sense ?**

## Adding up – MSE and MAE

- ▶ Many target observations in practice
- ▶ Or we can think about many situations with a single target observation
- ▶ Squared loss -> Mean Squared Error (MSE)

For  $k = 1 \dots K$  observations:

$$MSE = \frac{1}{K} \sum_{k=1}^K (\hat{y}_k - y_k)^2$$

(1)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{K} \sum_{k=1}^K (\hat{y}_k - y_k)^2} \quad (2)$$

## MSE implies expected value

$$MSE = \frac{1}{K} \sum_{k=1}^K (\hat{y}_k - y_k)^2$$

- ▶ MSE implies mean value for best predictor
  - ▶ Linear (least squares) regression
  - ▶ Why?
    - ▶ Because the solution to least squares minimization problem is the average.
    - ▶ Technically: first-order condition of minimization problem
- ▶ RMSE = square root of MSE
  - ▶ MSE is the numerator of the R-squared!

# MSE decomposition : Bias and Variance

May decompose MSE into Bias + Variance

- ▶ The bias of a prediction is the average of its prediction error.
  - ▶ An unbiased prediction produces zero error on average across multiple predictions.
  - ▶ A biased prediction produces nonzero error on average; the bias can be positive or negative
- ▶ The variance of a prediction describes how it varies around its average value when multiple predictions are made.
  - ▶ It's the variance of the prediction error around its average value.
  - ▶ The variance is zero if the prediction error is the same for all predictions.
  - ▶ The variance is higher the larger the spread of specific predictions around the average prediction



## MSE decomposition : Bias and Variance

- ▶ MSE is the sum of squared bias and the prediction variance.
- ▶ This decomposition helps appreciate a trade-off.

$$\begin{aligned}MSE &= \frac{1}{K} \sum_{k=1}^K (\hat{y}_k - y_k)^2 \\&= \left( \frac{1}{K} \sum_{k=1}^K (\hat{y}_k - \bar{y}) \right)^2 + \frac{1}{K} \sum_{k=1}^K (y_k - \bar{y})^2 \\&= \text{Bias}^2 + \text{PredictionVariance}\end{aligned}$$

- ▶ OLS is unbiased. Some other methods will allow for some bias in return for lower variance.

## Case study: used cars data

- ▶ Suppose you want to sell your car of a certain make, type, year, miles, condition and other features.
- ▶ The prediction analysis helps uncover the average advertised price of cars with these characteristics
  - ▶ That helps decide what price you may want to put on your ad.



## Case study: used cars data

- ▶ Scraped from a website
- ▶ Year of make (age), Odometer (miles)
- ▶ Tech specifications such as fuel and drive
- ▶ Dealer or private seller

## Case study: Loss function

- ▶ The loss function for predicting the value of our used car depends on how we value money and how we value how much time it takes to sell our car.
- ▶ A too low prediction may lead to selling our car cheap but fast;
- ▶ A too high prediction may make us wait a long time and, possibly, revising the sales price downwards before selling our car.
  
- ▶ Symmetric
- ▶ Sensitive to big deviations
- ▶ RMSE and OLS

## Case study - used cars: features

- ▶ Odometer, measuring miles the car traveled (**Continuous, linear**)
- ▶ More specific type of the car: LE, XLE, SE (missing in about 30% of the observations). (**Factor – set of dummies , incl N/A**)
- ▶ Good condition, excellent condition or it is like new (missing for about one third of the ads). (**Factor – set of dummies, incl N/A**)
- ▶ Car's engine has 6 cylinders (20% of ads say this; 43% says 4 cylinders, and the rest has no information on this). (**Binary for 6 cylinders**)

## Case study: models by hand

- ▶ Model 1: age, age squared
- ▶ Model 2: age, age squared, odometer, odometer squared
- ▶ Model 3: age, age squared, odometer, odometer squared, LE, excellent condition, good condition, dealer
- ▶ Model 4: age, age squared, odometer, odometer squared, LE, excellent condition, good condition, dealer, LE, XLE, cylinder
- ▶ Model 5: same as Model 4 but with all variables interacted with age (won't show in next table)

## Case study: Car price model results

	(1)	(2)	(3)	(4)
Variables	Model 1	Model 2	Model 3	Model 4
age	-1,530.09	-1,149.22	-873.47	-836.64
agesq	35.05	27.65	18.21	17.63
odometer		-303.84	-779.90	-788.70
odometersq			18.81	19.20
LE			28.11	-20.48
XLE				301.69
SE				1,338.79
cond_likenew				558.67
cond_excellent			176.49	190.40
cond_good			293.36	321.56
cylind6				-370.27
dealer			572.98	822.65
Constant	18,365.45	18,860.20	19,431.89	18,963.35

## Case study: Results

- ▶ When doing prediction, coefficients are less important.
- ▶ But we shall use them for sanity check: age negative, convex (flattens out)
- ▶ SE may not be even displayed. It is helpful for model selection, but only along with other measures
- ▶ and values of the predictor variables for our car: age = 10 (years), odometer= 12 (10 thousand miles), type= LE, excellent condition=1.



## Case study: Results

- ▶ When doing prediction, coefficients are less important.
- ▶ But we shall use them for sanity check: age negative, convex (flattens out)
- ▶ SE may not be even displayed. It is helpful for model selection, but only along with other measures
- ▶ and values of the predictor variables for our car: age = 10 (years), odometer= 12 (10 thousand miles), type= LE, excellent condition=1.
- ▶ A point prediction, Model 3: age: -873.47, age squared=18.21, odometer -799.90, odometer sq = 18.81, LE=28.11, cond excellent: 176.49+ C=19.431.89
- ▶ Predicted price is 6073.

## Case study: Prediction Interval

- ▶ Calculating prediction intervals for the baseline models
- ▶ Very wide interval despite high  $R^2$
- ▶ Prediction is hard!
- ▶ Even with a good model, you'll make plenty of errors
- ▶ Should be aware
- ▶ Let your clients know in advance...

## Case study: Prediction Interval

- ▶ Based on the third model, we have a point prediction of \$6073
- ▶ Have a 80% prediction intervals (PI) – Ads for cars just like ours may ask a price ranging from \$4,317 to \$7,829 with a 80% chance.

Table: Car price model

	Model 1	Model 3
Point prediction	6,569	6,073
Prediction Interval (80%)	[4,296-8,843]	[4,317-7,829]
Prediction Interval (95%)	[3,085-10,053]	[3,382-8,763]

Note: *Chicago cars. Prices in dollars.*  
 Source: *used-cars dataset.*

# Model selection

## Data for prediction

- ▶ We have a dataset
- ▶ We wanna make some prediction

# Model selection

Model selection is finding the best fit while avoiding overfitting and aiming for high external validity

## External validity, avoiding overfitting and model selection

- ▶ Have a dataset and a target variable. Compare various models of prediction.
- ▶ How to choose a model?
- ▶ Pick a model that can predict well....
  - ▶ Best prediction - best model that would produce the smallest prediction error.
  - ▶ Context of squared loss function → finding the regression that would produce the smallest RMSE for the target observations.

## External validity, avoiding overfitting and model selection

- ▶ Have a dataset and a target variable. Compare various models of prediction.
- ▶ How to choose a model?
- ▶ Pick a model that can predict well....
  - ▶ Best prediction - best model that would produce the smallest prediction error.
  - ▶ Context of squared loss function → finding the regression that would produce the smallest RMSE for the target observations.
- ▶ Pick a model that can predict well on the **live data**



## Underfit, overfit

- ▶ Comparing two models (model 1 and model 2)
- ▶ Model 1 can give a worse fit in the live data than model 2 in two ways.
- ▶ Model 1 may give a worse fit both in the original data and the **live** data. In this case, we say that model 1 underfits the original data.
  - ▶ Simple: we should build a better model.
- ▶ Model 1 may actually give a better fit in the original, but a worse fit in the **live** data. In this case, we say that model 1 overfits the original data.

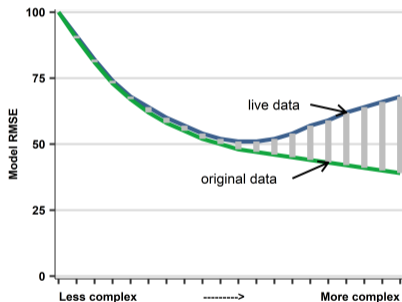
# Overfitting

- ▶ Overfitting is a key aspect of external validity
  - ▶ finding a model that fits the data better than alternative models
  - ▶ but makes worse actual prediction.
- ▶ Thus, the problem of overfitting the original data is best split into two problems:
- ▶ fitting patterns in the original data that are not there in the population, or general pattern, it represents;
- ▶ fitting patterns in the world of the original data that will not be there in the world of the live data.

## Reason for overfitting

- ▶ The typical reason for overfitting is fitting a model that is too complex on the dataset.
  - ▶ Complexity: number of estimated coefficients
- ▶ Often: fitting a model with too many predictor variables.
  - ▶ Including too many variables from the dataset that do not really add to the predictive power of the regression,
    - ▶ often because they are strongly correlated with other predictor variables.
- ▶ Specifying too many interactions,
- ▶ Too detailed nonlinear patterns
  - ▶ as piecewise linear splines with many knots
  - ▶ polynomials of high degree.

# Increasing model complexity



- ▶ As we increase model complexity
- ▶ Such as number of features (variables)
  - ▶ By adding interactions, etc.
- ▶ We will see
  - ▶ RMSE within dataset to fall monotonously
  - ▶ RMSE for target observations (ie. not in our dataset) to fall and then rise as we overfit
    - ▶ example to come in class 2

## Finding the best model by best fit and penalty: The BIC

- ▶ *Approach 1: Indirectly*
- ▶ Estimate it by an adjustment
  - ▶ Use a method based on some distributional assumptions
  - ▶ Need to pick an evaluation criterion
- ▶ =In-sample evaluation with penalty
  - ▶ Specify and estimate model using all data
  - ▶ Use a measure of fit that helps avoid overfitting
- ▶ Such as
  - ▶ adjusted  $R^2$
  - ▶ BIC = Bayesian Information Criterion, or Schwarz criterion

## Indirect evaluation criteria

- ▶ Main methods: AIC, BIC and adjusted  $R^2$ 
  - ▶ Advantage: easy to compute
  - ▶ Disadvantage: assumptions
- ▶ Adjusted  $R^2$  – just add a penalty for having many RHS vars
  - ▶ corrects with  $(n - 1)/(n - p - 1)$
- ▶ Akaike Information Criterion
  - ▶  $AIC = -2 \times \ln(\text{likelihood}) + 2 \times k$
- ▶ Schwarz – Bayesian Information Criterion
  - ▶  $BIC = -2 \times \ln(\text{likelihood}) + \ln(N) \times k$ 
    - ▶ Both quantities that take the log likelihood and apply a penalty for the number of parameters being estimated. Both are based on information loss theory from the fifties.
    - ▶ BIC puts heavier penalty on models with many RHS variables, than AIC.

## Model fit evaluation

- ▶ Use a good measure of fit to compare models.
- ▶ Don't
  - ▶ Don't use MSE or R-squared (the two very closely related).
  - ▶ They choose best fit in data and don't care about overfitting.
- ▶ In practice, use BIC.
  - ▶ BIC good approximation of what more sophisticated methods would pick. Or even more conservative...
  - ▶ That introduces a "penalty term"
    - ▶ More predictor variables leads to worse value
    - ▶ Even more so in large samples.

## Finding the best model by training and test samples

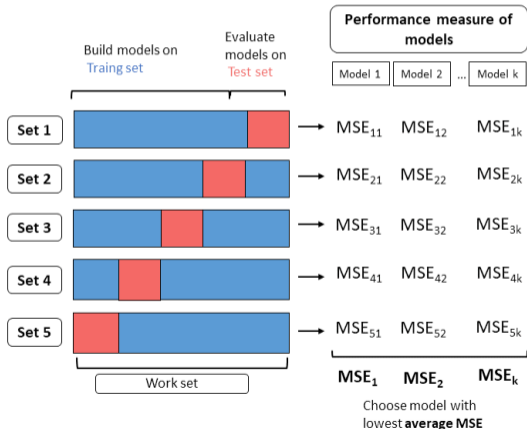
- ▶ *Approach Nr.2: Directly*
- ▶ Estimate it using a test (validation) set approach.
  - ▶ Needs cutting the dataset into training and test sample
  - ▶ No assumption
  - ▶ Need to pick evaluation criterion (loss function) = RMSE (root mean squared error)
- ▶ Estimate the model in part of the data (say, 80%).
  - ▶ *Training sample*
- ▶ Evaluate predictive performance on the rest of the data.
  - ▶ *Test sample*
- ▶ Avoid overfitting in training data by evaluating on test data.



## Training and Test Samples

- ▶ Creating two sub-samples
  - ▶ Randomly! (ie. not 1—80 and 81—100)
- ▶ Randomly generate an ID, sort and create two sub-samples.
- ▶ Training sample 80%
  - ▶ Regressions will be on run on this sample
  - ▶ Coefficients estimated
- ▶ Test (validation) sample 20%
  - ▶ Using estimated coefficients, we predict values for flats in the validation sample
  - ▶ Calculate residual, RMSE in the test sample
    - ▶ RMSE rather than MSE – smaller numbers....

## 5-fold cross-validation



- ▶ Split sample  $k=5$  times to train and test
- ▶ For each folds:
  - ▶ Estimate model on training.
  - ▶ Get coefficients.
  - ▶ Use them to estimate on Test
  - ▶ Calculate test MSE
- ▶ Average and take Sqrt
- ▶ Repeat for models
- ▶ Pick model w lowest avg RMSE

## BIC vs test RMSE

- ▶ In our experience, in practice, BIC is the best indirect criterion – closest to test sample.
- ▶ The advantage of BIC is that it needs no sample splitting which may be a problem in small samples.
- ▶ The advantage of test MSE is that it makes no assumption.
- ▶ BIC is a good first run, quick, is often not very wrong.
- ▶ Ultimately, you want to do a test MSE.

## Case study: Model selection

- ▶ We have the ingredients, we need to pick a model.
- ▶ This process involves variable selection and a decision rule of choosing the model based on some loss function.
- ▶ BIC on the actual data
- ▶ Test-sample RMSE
- ▶ Cross-validated (CV) RMSE
- ▶ If enough data / computer power, use CV RMSE
- ▶ With larger dataset, overfit becomes less of an issue.

## Case study: Model selection

Table: Car price models -BIC and in-sample RMSE

	Model	N vars	N coeff	R-squared	RMSE	BIC
1	Model 1	1	3	0.85	1,755	5,018
2	Model 2	2	5	0.90	1,433	4,910
3	Model 3	5	9	0.91	1,322	<b>4,893</b>
4	Model 4	6	12	0.92	1,273	<b>4,894</b>
5	Model 5	6	22	0.92	<b>1,239</b>	4,935

Note: *In sample values. Model 1: age, age squared, Model 2= Model 1 +odometer, odometer squared, Model 3= Model2 + SE, excellent condition, good condition, dealer, Model 4= Model 3 + LE, XLE, like new condition, 6cylinder, Model 5 = Model 4 + many interactions.*

Source:

used-cars

dataset.

## Case study: Model selection

- ▶ Cross-validate using 4-fold cross validation.
- ▶ Run the regression on 3/4 of the sample, predicting on the remaining 1/4 of the sample, get RMSE on test sample.
- ▶ We then average out RMSE values over the 4 test samples

Table: Car price models -CV RMSE

	Fold No.	Model 1	Model 2	Model 3	Model 4	Model 5
1	Fold1	1,734	1,428	1,331	1,395	1,391
2	Fold2	2,010	1,781	1,692	1,638	1,693
3	Fold3	1,465	1,251	1,256	1,253	1,436
4	Fold4	1,823	1,325	1,250	1,246	1,307
5	Average	1,769	1,460	<b>1,394</b>	<b>1,392</b>	1,464

Source:

used-cars

dataset.

## Case study: Model selection

- ▶ Model 3 has lowest BIC, lowest average RMSE on test samples. Model 4 is close.
- ▶ Interestingly, both approaches suggests that Model 3 is the one that has the best prediction properties
- ▶ Small sample, simple model.

## External validity and stable patterns

- ▶ BIC, Training-test, k-fold cross-validation. . .
- ▶ All very nice
- ▶ But, in the end, they all use the information in the data.
- ▶ How would things look for the target observation(s)?
- ▶ The issue of stationarity – how our data is related to other datasets we may use our model
  - ▶ We may have some ideas
  - ▶ We may use non-random test samples that may mimic the difference in our data and the target observations
- ▶ In the end we can't know but need to think about it.
- ▶ Plus be aware, that some difference is likely, so your model fit in an outside data source is likely to be worse...



## External validity and stable patterns

- ▶ Most predictions will be on future data
- ▶ High external validity requires that the environment is **stationary**.
- ▶ Stationarity means that the way variables are distributed remains the same over time.
  - ▶ Here that distribution is to be understood in a general way: the joint distribution of predictor variables and target variable are required to remain the same throughout the time covered in the data and the time of the forecast.
- ▶ Stationarity ensures that the relationship between predictors and the target variable is the same in the data and the forecasted future.
  - ▶ If the relationship breaks down whatever we establish in our data won't be true in the future, leading to wrong forecasts.

## External validity and stable patterns

- ▶ External validity and stable patterns - Very broad concept
- ▶ It's about representativeness of actual data → to live data
- ▶ Often hard to know.
  - ▶ Remember hotels (other dates, other cities).
- ▶ Domain knowledge can help.
- ▶ Study if patterns were stable in the past / other locations were stable can help.

# Algorithms

# Machine Learning and the Role of Algorithms

- ▶ **Predictive analytics** is often used for data analysis whose goal is prediction. But a more popular, and related, term is machine learning.
- ▶ **Machine learning** is an umbrella concept for methods that use algorithms to find patterns in data and use them for prediction purposes.
- ▶ An **algorithm** is a set of rules and steps that defines how to generate an output (predicted values) using various inputs (variables, observations in the original data).
- ▶ A **formula** is an example of an algorithm – one that can be formulated in terms of an equation.
  - ▶ OLS formula for estimating the coefficients of a linear regression is an algorithm.

# Machine Learning Algorithms

- ▶ Machine learning is about algorithms, machines and learning
- ▶ Algorithms specify each and every step to follow in a clear way.
- ▶ Not all algorithms can be translated into a formula.
  - ▶ The bootstrap estimation of a standard error (Chapter 5, Section 5.6) is an example.
  - ▶ K-fold cross-validation.
- ▶ Heavy use of **machines** = computers. Steps of algorithm translated into computer code and make the computer follow those steps. Fast.
- ▶ Learning - learn something from the data with data and an algorithms.
  - ▶ Predicted value of  $y=?$  If combine  $x$  variables using a particular model.
  - ▶ learning which model is best for predicting  $y$  as well as what that predicted value is.

## What is, machine learning?

- ▶ Many definitions, discussions.
- ▶ Here: Machine learning is an approach to predictive data analysis – achieving the best possible prediction from available data.
- ▶ Consequence 1: understanding the patterns of associations between  $y$  and  $x$  is of secondary importance.
  - ▶ We need stable patterns for good prediction in live data, but that is it.
- ▶ The machine learning *attitude* - a preference for evaluating methods based on data as opposed to abstract principles.
  - ▶ Original data to live data
  - ▶ Not a general rule or philosophy
- ▶ Machine learning broadly: all prediction models including OLS
- ▶ Machine learning narrowly: prediction models with no formula, ie not OLS

## Main takeaways

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