Data Analytics in Business

Resampling Methods

Sridhar Narasimhan
Professor
Scheller College of Business

Introduction

Lessons

A. Introduction
B. The Validation Set Approach
C. Leave-One-Out Cross-Validation (LOOCV)
D. K-fold Cross Validation
E. Bias-Variance Tradeoff for K-fold Cross-Validation
F. Cross-Validation in Classification Problems
G. Bootstrap
H. CV and Auto Data Set Application

* This module is based on Chapter 5 in the ISLR textbook.

Resampling

- A **resampling method** is when data from a training set is drawn repeatedly. For each sample that is drawn, the model is refitted.
- One objective of resampling is to gauge if the fitted models differ, which could give us information about the model's performance.
- Two most commonly used methods of resampling
    - Cross-Validation (or CV)
    - Bootstrap

Resampling

- **Cross Validation** can be used to evaluate the test error of a model or to select the level of flexibility of a model.
    - Evaluating the model's performance is known as **model assessment**.
    - Selecting the proper level of flexibility is known as **model selection**.
- **Bootstrap** is used mostly to measure the accuracy of a parameter estimate of a learning method.

Training Error and Test Error

- We assume that we are interested in regressions with quantitative response variable (i.e., *Y*). The concept remains the same even if the response is a categorical variable.
- The **test error** is the average error (typically the Mean Square Error) obtained from using a model (or method) to predict the response on a new observation.
- Note that we would like to use a particular model if it performs well on test data assuming such a test data set is available. Often, a test data set may not be readily available.
- The **training error** is the average error (typically the Mean Square Error) obtained from using a model (or method) to predict the response on the data in the training set.

Summary

- Introduction to Resampling Methods

Next

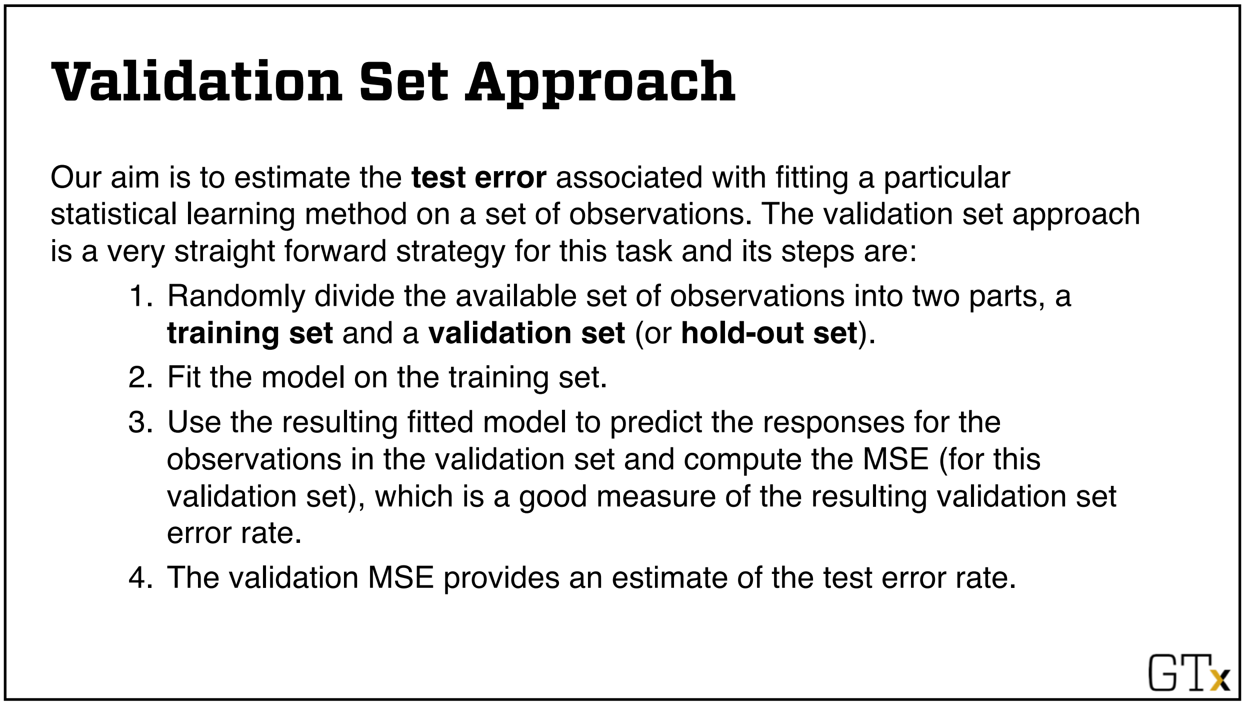
- The Validation Set Approach

Data Analytics in Business

Resampling Methods

Sridhar Narasimhan
Professor
Scheller College of Business

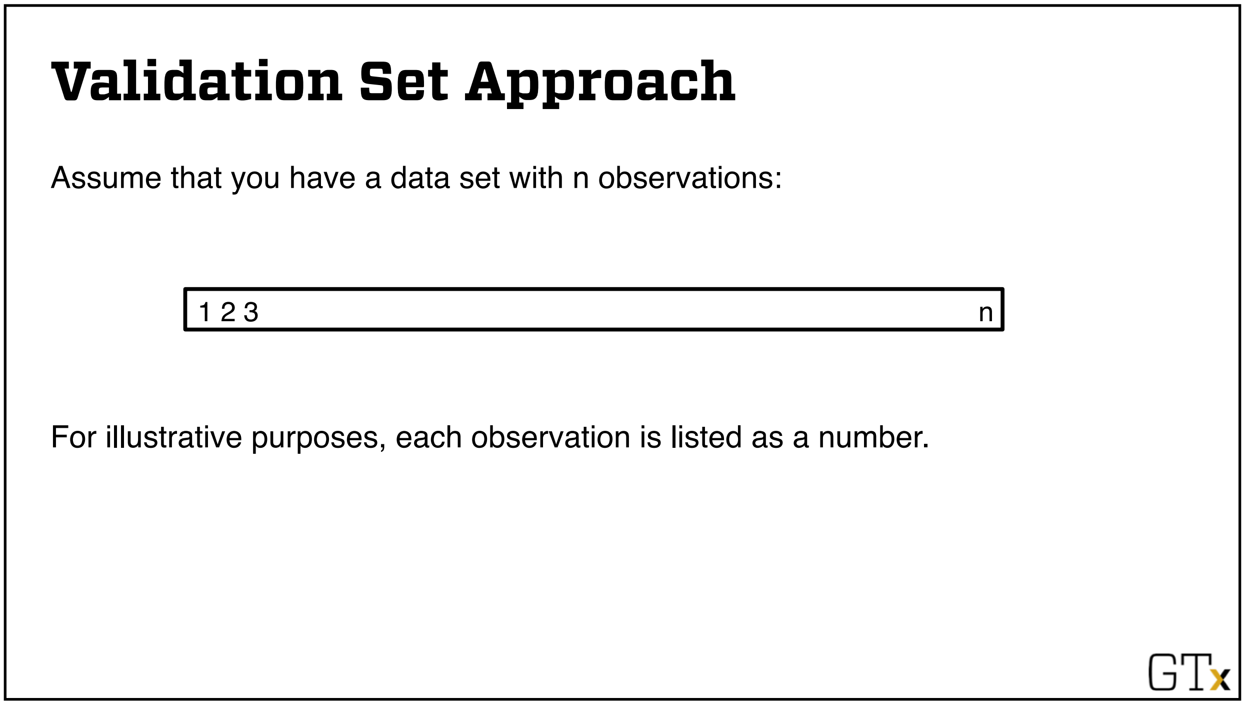
The Validation Set Approach



Validation Set Approach

Our aim is to estimate the **test error** associated with fitting a particular statistical learning method on a set of observations. The validation set approach is a very straight forward strategy for this task and its steps are:

1. Randomly divide the available set of observations into two parts, a **training set** and a **validation set** (or **hold-out set**).
2. Fit the model on the training set.
3. Use the resulting fitted model to predict the responses for the observations in the validation set and compute the MSE (for this validation set), which is a good measure of the resulting validation set error rate.
4. The validation MSE provides an estimate of the test error rate.

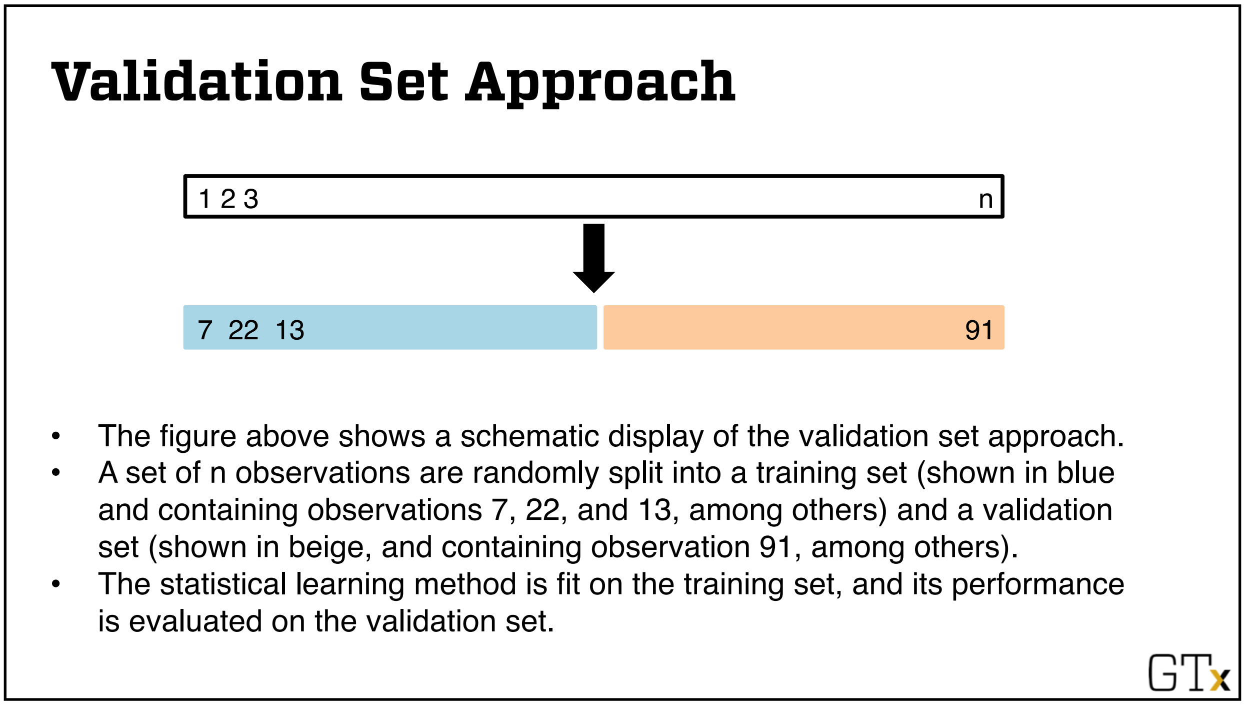


Validation Set Approach

Assume that you have a data set with n observations:

1 2 3 n

For illustrative purposes, each observation is listed as a number.



Validation Set Approach

1 2 3 n

7 22 13 91

- The figure above shows a schematic display of the validation set approach.
- A set of n observations are randomly split into a training set (shown in blue and containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others).
- The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.

Drawbacks of the Validation Set Approach

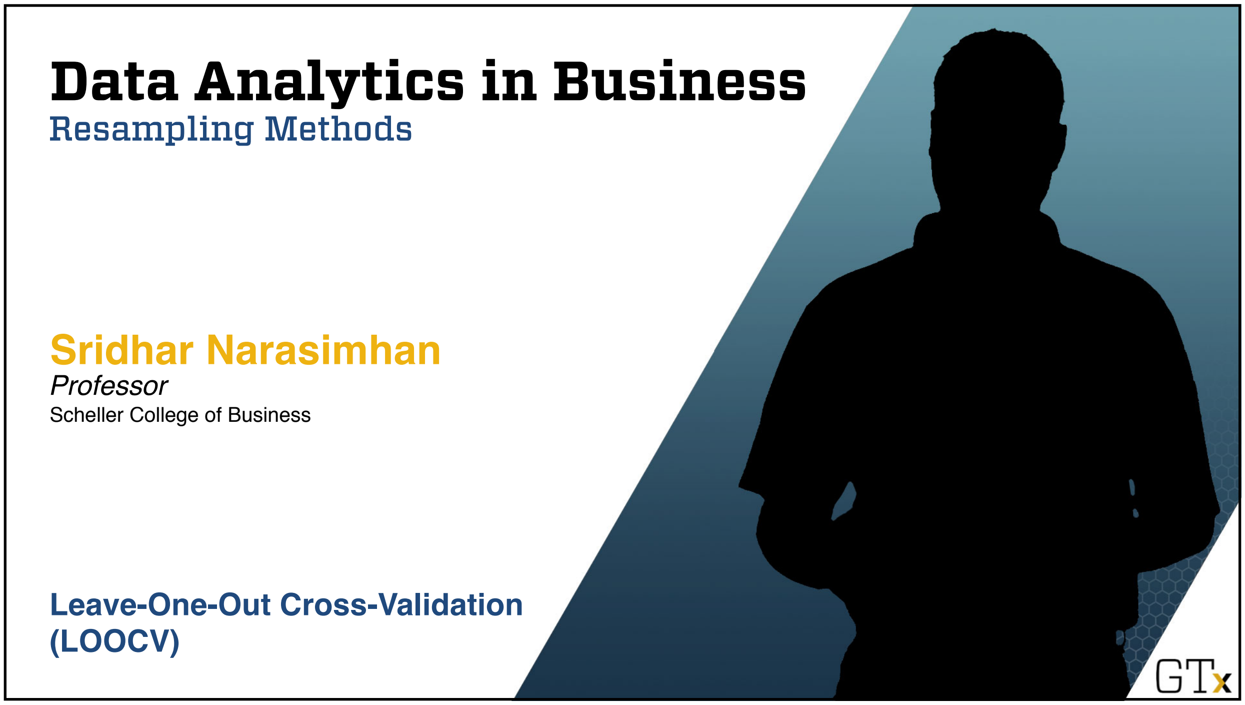
- While the validation set approach is easy to understand and implement, it has two potential shortcomings:
  1. The validation estimate of the test error rate could be highly variable, depending on which observations are included in the training sets and in the validation sets.
  2. Since only the training data is used to fit the model, the validation set error may tend to over-estimate the test error for the model fit on the **entire** data set.
- **Cross-Validation**, which we will discuss more in subsequent lessons, addresses these two shortcomings.

Summary

- The Validation Set Approach

Next

- Leave-One-Out Cross-Validation (LOOCV)

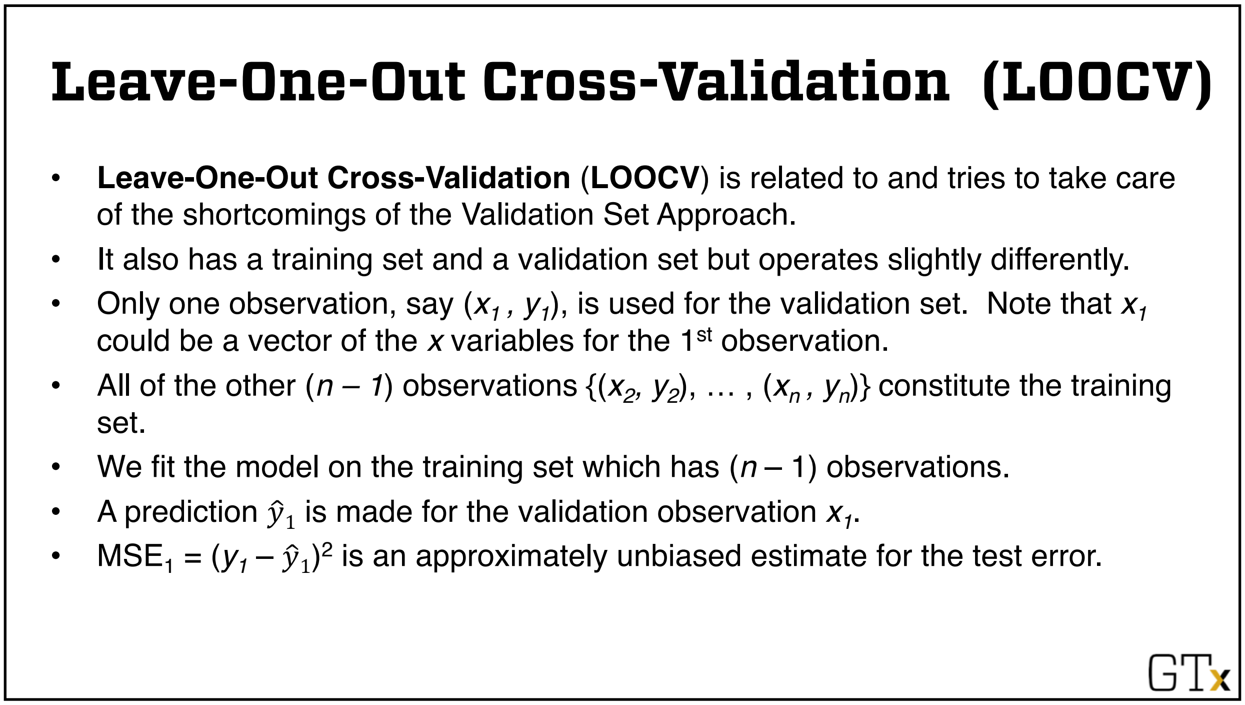


Data Analytics in Business

Resampling Methods

Sridhar Narasimhan
Professor
Scheller College of Business

Leave-One-Out Cross-Validation (LOOCV)



Leave-One-Out Cross-Validation (LOOCV)

- **Leave-One-Out Cross-Validation** (**LOOCV**) is related to and tries to take care of the shortcomings of the Validation Set Approach.
- It also has a training set and a validation set but operates slightly differently.
- Only one observation, say (x_1, y_1), is used for the validation set. Note that x_1 could be a vector of the x variables for the 1st observation.
- All of the other $(n – 1)$ observations $\{(x_2, y_2), \dots, (x_n, y_n)\}$ constitute the training set.
- We fit the model on the training set which has $(n – 1)$ observations.
- A prediction $\hat{y}_1$ is made for the validation observation x_1.
- $\text{MSE}_1 = (y_1 – \hat{y}_1)^2$ is an approximately unbiased estimate for the test error.

Leave-One-Out Cross-Validation (LOOCV)

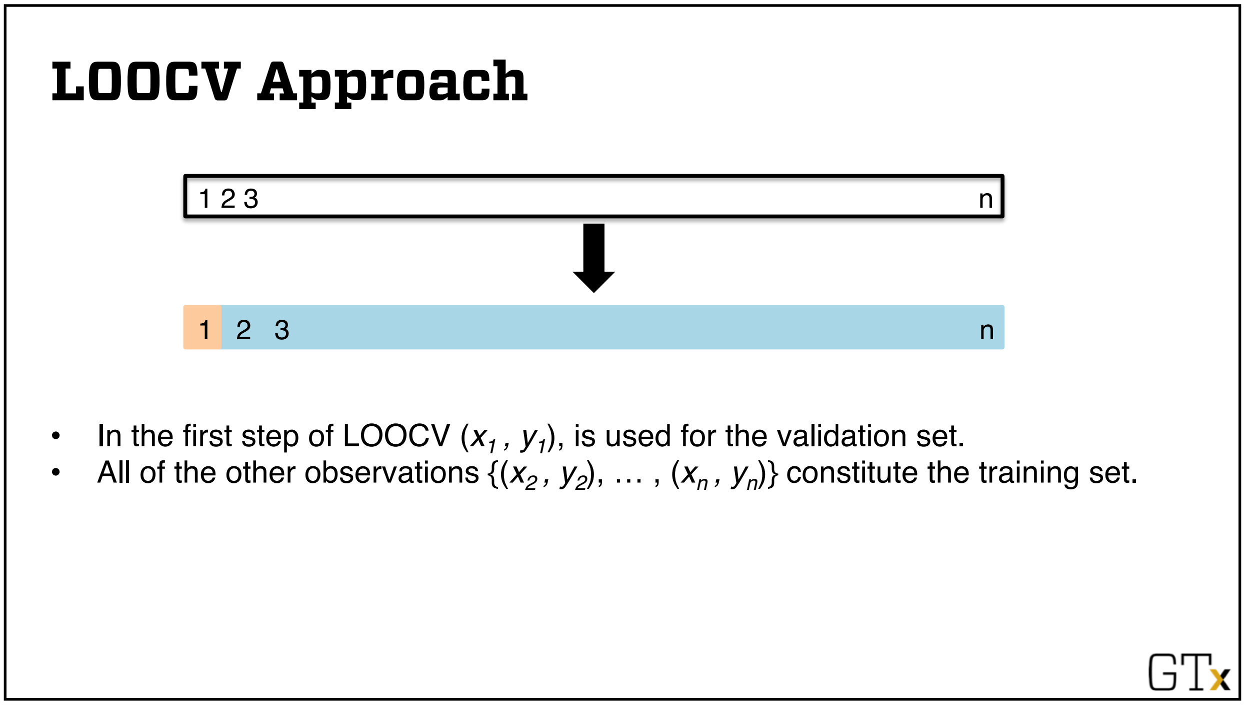
- Note that MSE_1 is a poor estimate because it is highly variable because it is only for one observation.
- We repeat the above procedure by selecting observation (x_2 , y_2) as the validation set and the other (n – 1) observations as the training set.
- Compute MSE_2 similar to the way MSE_1 was computed.
- LOOCV repeats this a total of n times to yield n mean squared errors, MSE_1, … , MSE_n
- The LOOCV estimate for the test MSE is the average of these n MSEs:

$$\boldsymbol{CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i}$$

LOOCV

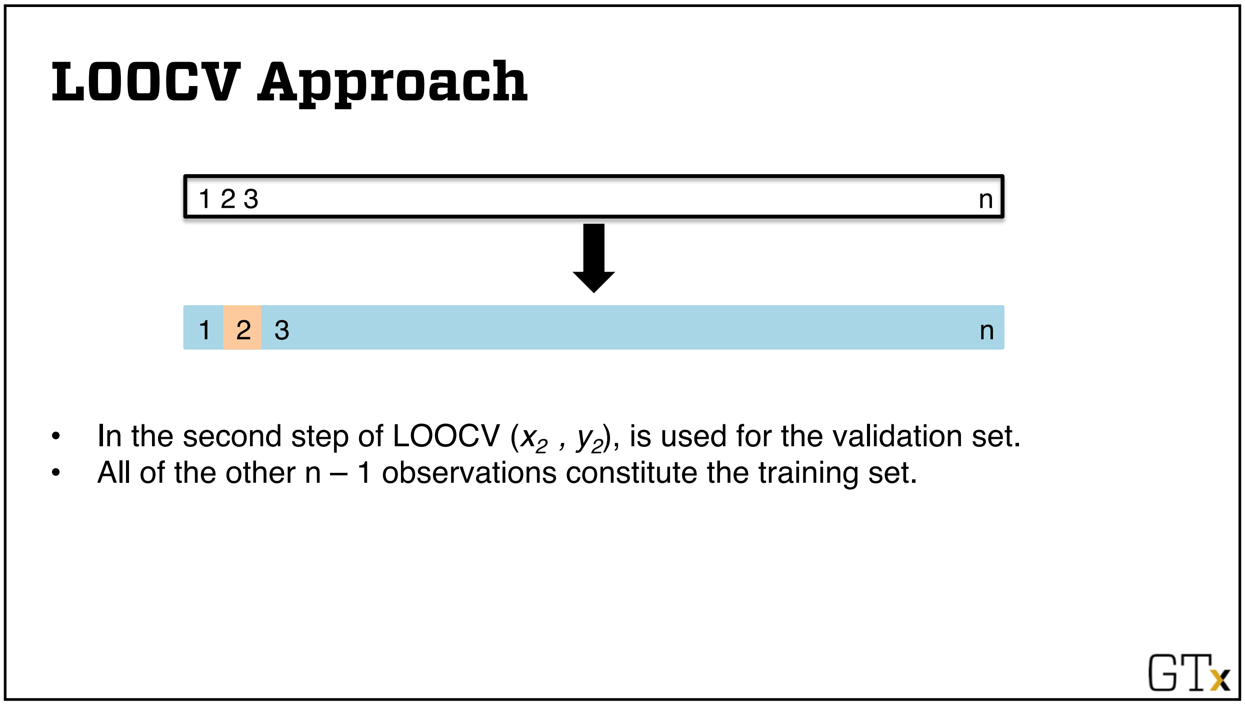
Assume that you have a data set with n observations:

For illustrative purposes, each observation is listed as a number.



LOOCV Approach

- In the first step of LOOCV (x_1 , y_1), is used for the validation set.
- All of the other observations {(x_2 , y_2), … , (x_n , y_n)} constitute the training set.

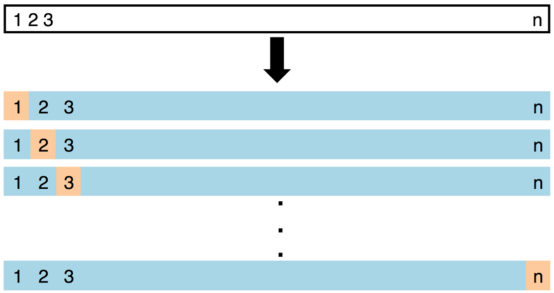


LOOCV Approach

- In the second step of LOOCV (x_2 , y_2), is used for the validation set.
- All of the other n – 1 observations constitute the training set.

LOOCV Approach

- The figure to the right is a schematic display of LOOCV.
- A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige).
- The test error is then estimated by averaging the n resulting MSE's.
- The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.



GTx

LOOCV Versus the Validation Set Approach

1. LOOCV has far less bias since it uses training sets with (n – 1) observations (i.e., almost the entire data set), as opposed to Validation Set Approach where the training set is usually 50% of the original data set. Hence, LOOCV tends not to overestimate the test error rate as much as the Validation Set Approach does.
2. Performing the LOOCV multiple times will always give the same results unlike the Validation Set Approach. This is because there is no randomness in the training/validation set splits.
3. LOOCV can be computationally expensive since the model has to be fitted n times. This could be a problem if n is large and if each individual model takes time to fit.

GTx

Summary

- Leave-One-Out Cross-Validation (LOOCV)

Next

- K-fold Cross-Validation

GTx

Data Analytics in Business

Resampling Methods

Sridhar Narasimhan

Professor

Scheller College of Business

K-fold Cross-Validation

GTx

K-fold CV

- K-fold CV approach involves dividing the set of observations into *k* groups, or **folds**, of about the same size.
- It starts with the first fold treated as the validation set and the method (e.g., regression) is fit on the remaining (*k* – 1) folds.
- The mean square error, MSE_1, is then computed on the observations in the held-out fold.
- This procedure is repeated *k* times and we get *k* estimates of the test error, $\text{MSE}_1, \dots , \text{MSE}_k$
- The k-fold CV estimate for the test MSE is the average of these *k* MSEs

$$\boldsymbol{CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i}$$

K-fold CV

- One can observe that LOOCV is a special case of k-fold when *k = n*
- Typically, k = 5 or 10, when performing k-fold CV; i.e., the method is fitted 5 or 10 times.
- This is an advantage over LOOCV which requires fitting the method *n* times.

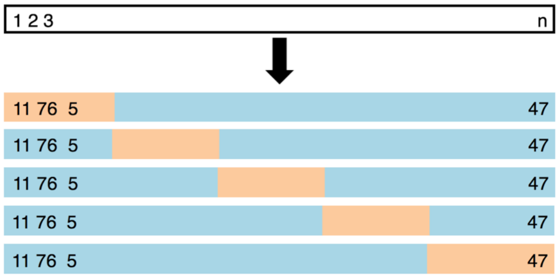
K-fold CV

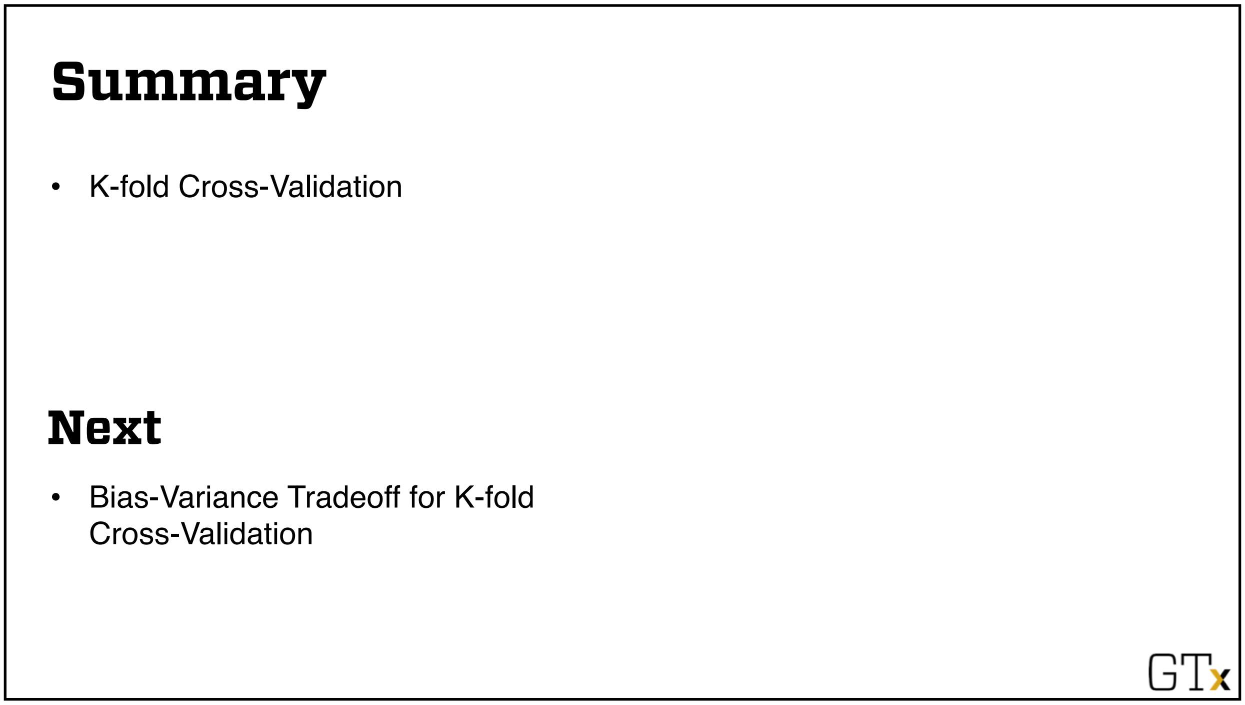
Assume that you have a data set with n observations:

For illustrative purposes, each observation is listed as a number.

K-fold CV

- The figure to the right is a schematic display of 5-fold CV.
- A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue).
- The test error is estimated by averaging the five resulting MSE estimates.



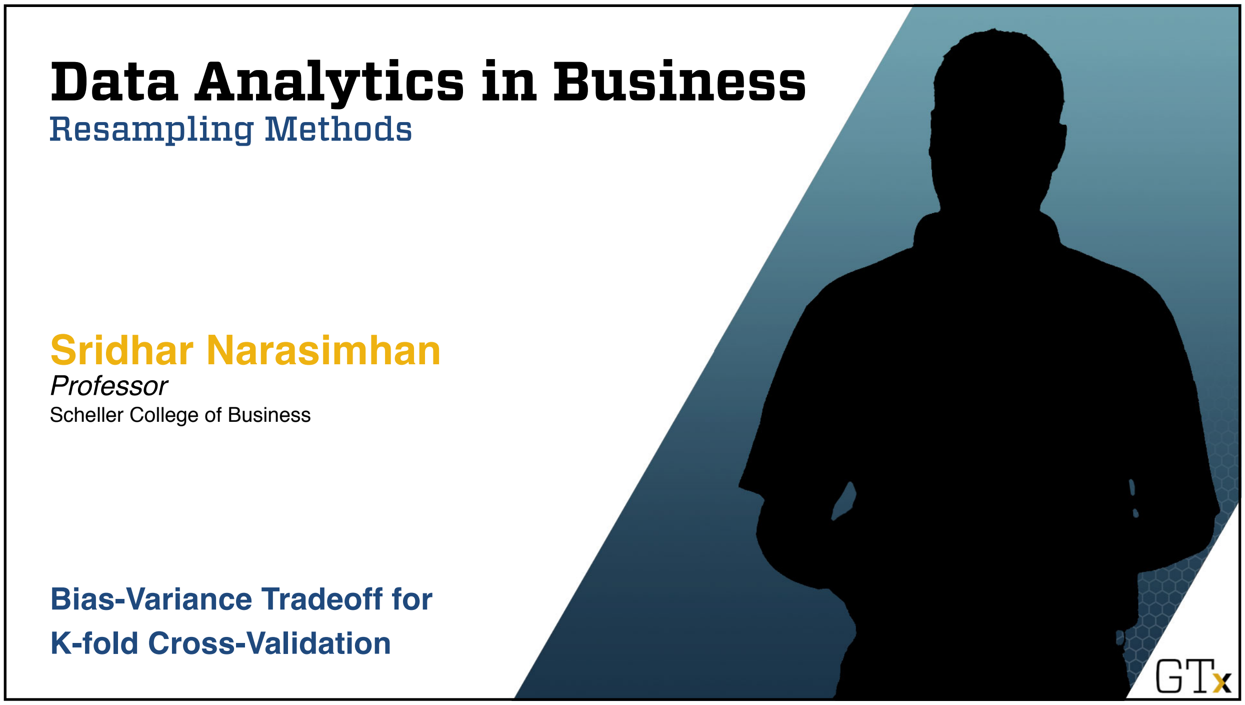


Summary

- K-fold Cross-Validation

Next

- Bias-Variance Tradeoff for K-fold Cross-Validation



Data Analytics in Business

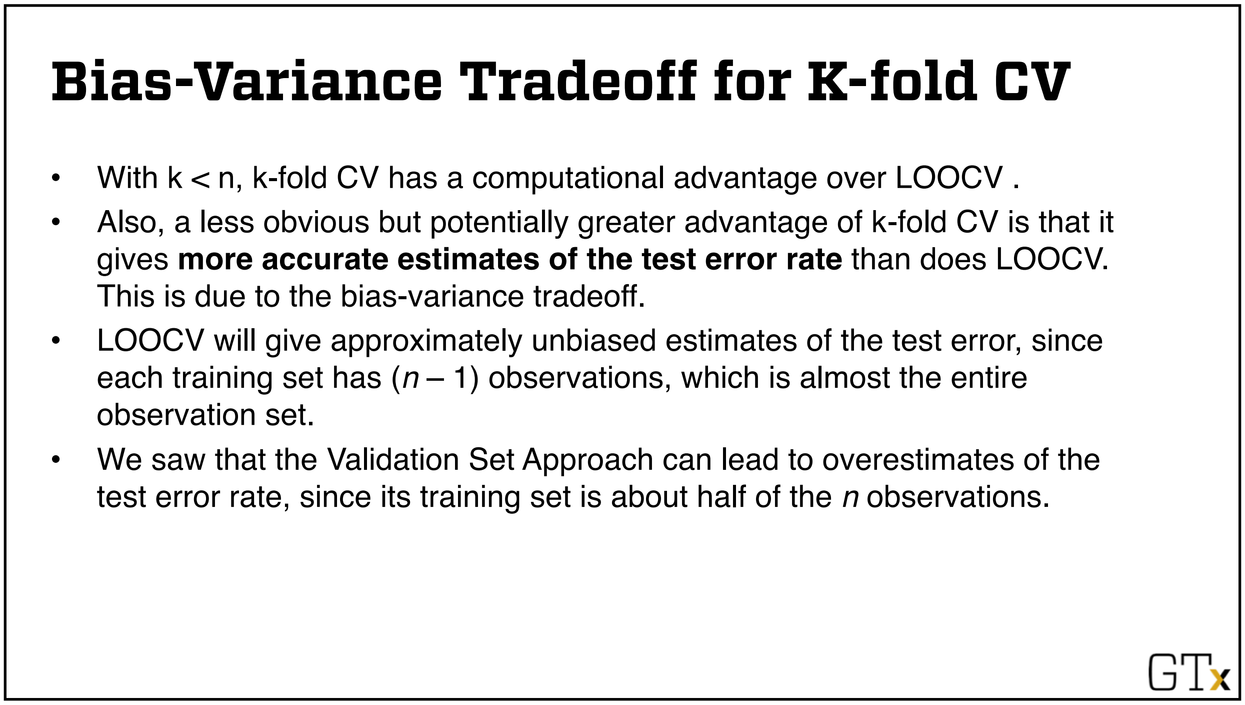
Resampling Methods

Sridhar Narasimhan

Professor

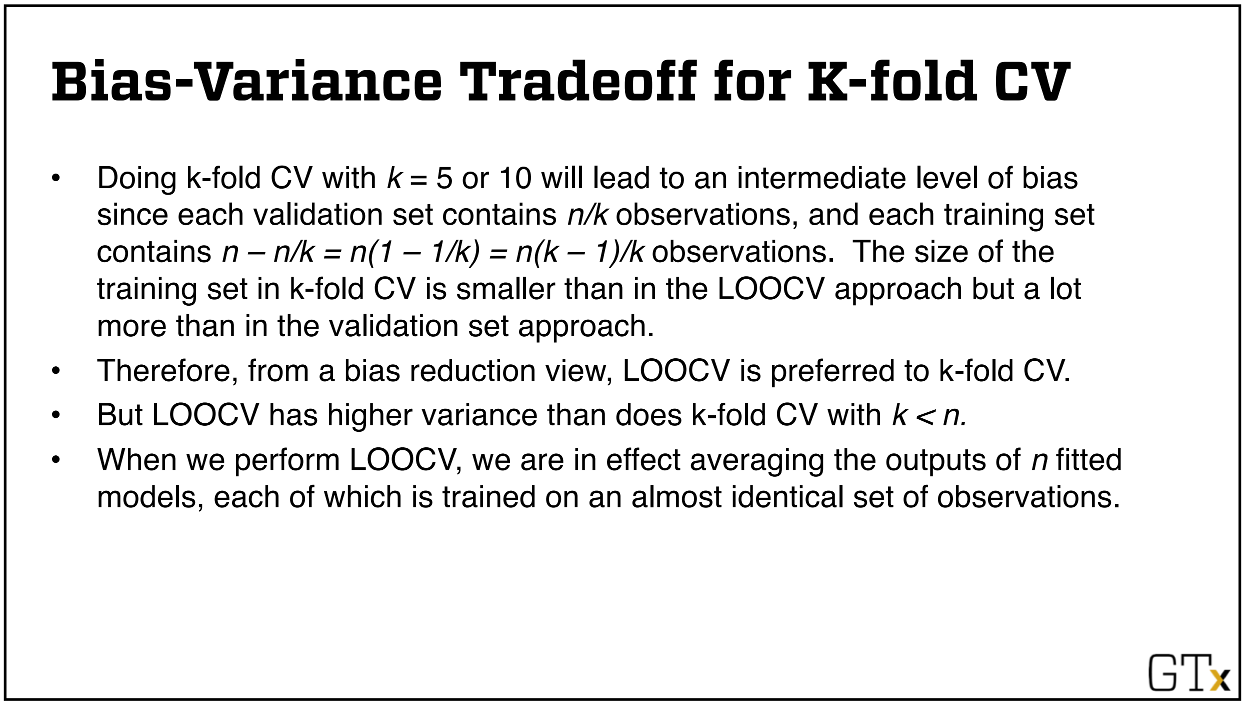
Scheller College of Business

Bias-Variance Tradeoff for K-fold Cross-Validation



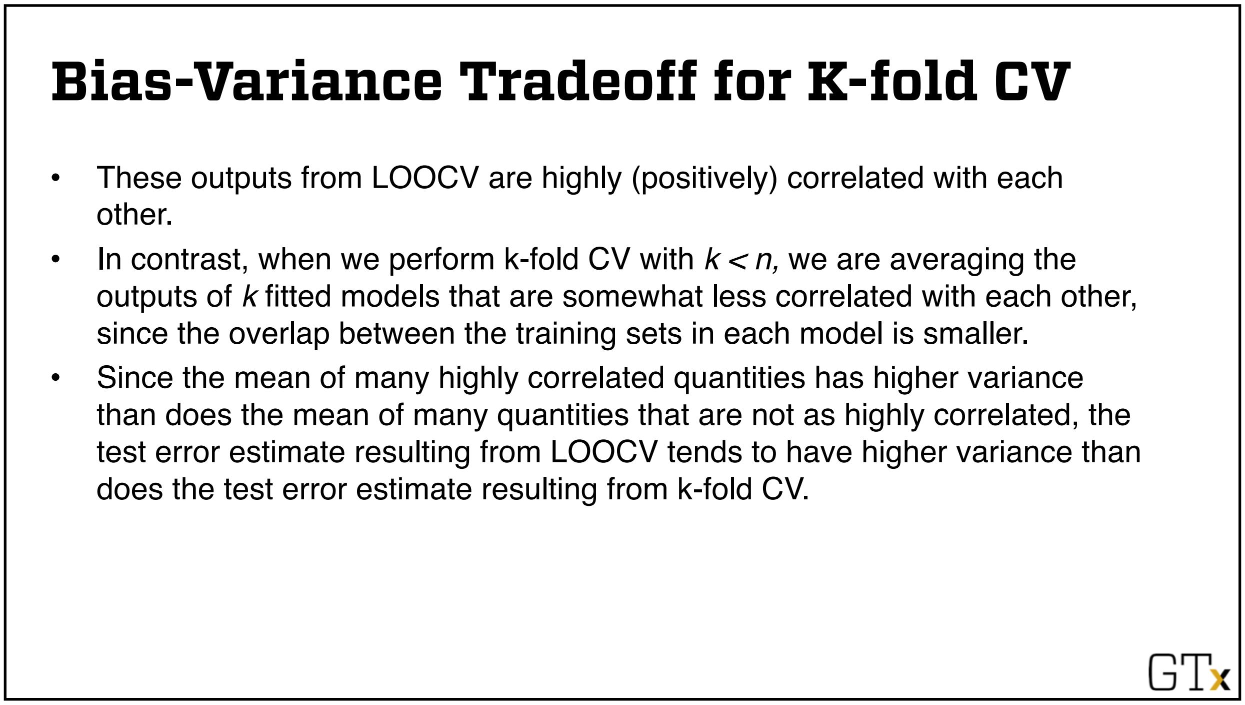
Bias-Variance Tradeoff for K-fold CV

- With k < n, k-fold CV has a computational advantage over LOOCV .
- Also, a less obvious but potentially greater advantage of k-fold CV is that it gives **more accurate estimates of the test error rate** than does LOOCV. This is due to the bias-variance tradeoff.
- LOOCV will give approximately unbiased estimates of the test error, since each training set has (*n* – 1) observations, which is almost the entire observation set.
- We saw that the Validation Set Approach can lead to overestimates of the test error rate, since its training set is about half of the *n* observations.



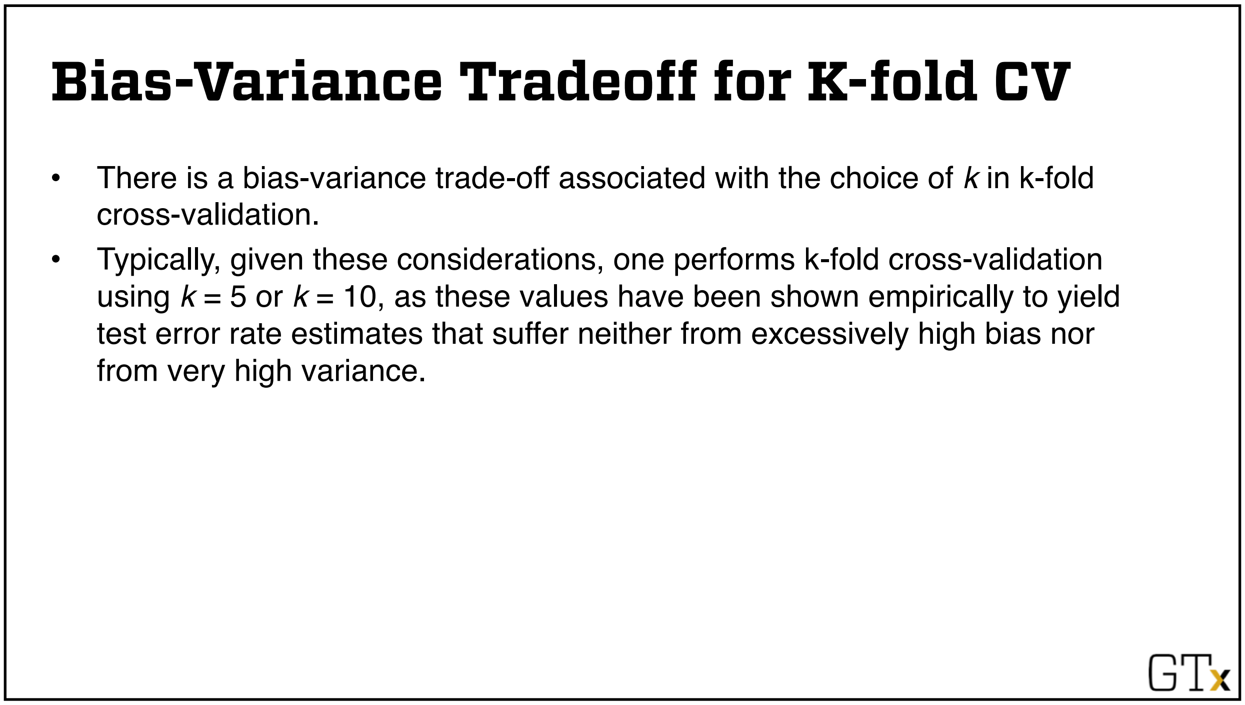
Bias-Variance Tradeoff for K-fold CV

- Doing k-fold CV with *k* = 5 or 10 will lead to an intermediate level of bias since each validation set contains *n/k* observations, and each training set contains *n – n/k = n(1 – 1/k) = n(k – 1)/k* observations. The size of the training set in k-fold CV is smaller than in the LOOCV approach but a lot more than in the validation set approach.
- Therefore, from a bias reduction view, LOOCV is preferred to k-fold CV.
- But LOOCV has higher variance than does k-fold CV with *k < n.*
- When we perform LOOCV, we are in effect averaging the outputs of *n* fitted models, each of which is trained on an almost identical set of observations.



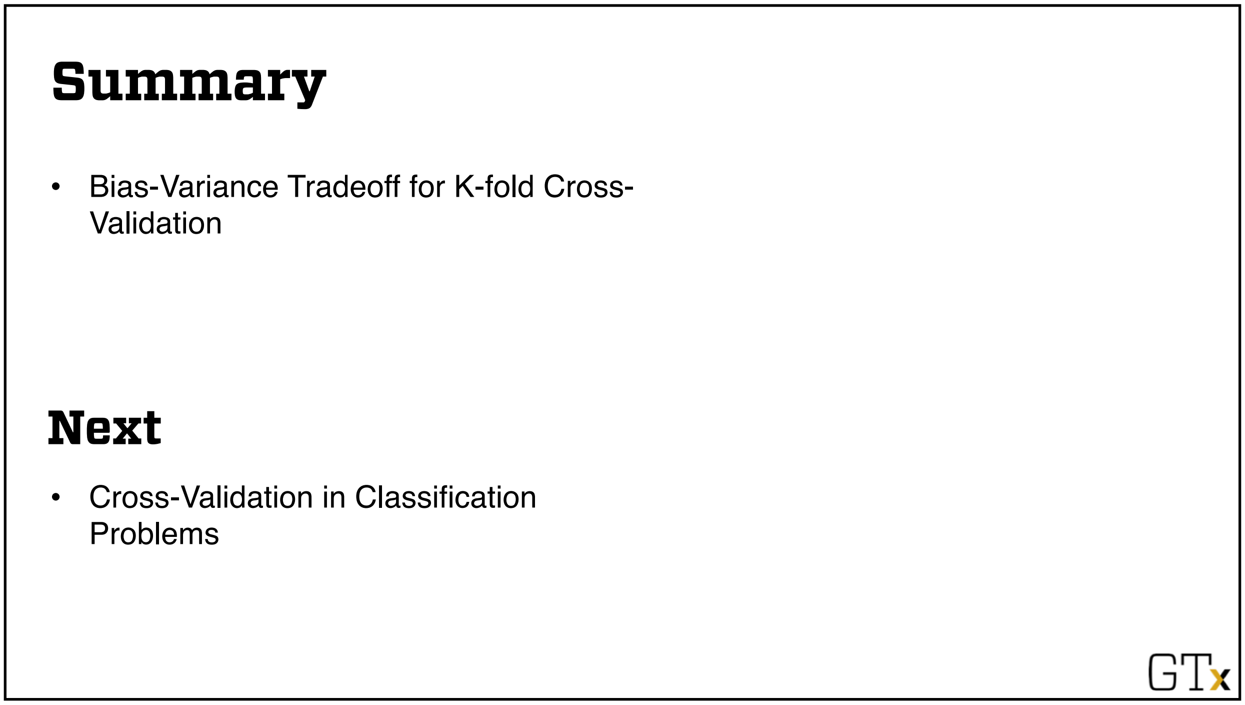
Bias-Variance Tradeoff for K-fold CV

- These outputs from LOOCV are highly (positively) correlated with each other.
- In contrast, when we perform k-fold CV with *k* < *n,* we are averaging the outputs of *k* fitted models that are somewhat less correlated with each other, since the overlap between the training sets in each model is smaller.
- Since the mean of many highly correlated quantities has higher variance than does the mean of many quantities that are not as highly correlated, the test error estimate resulting from LOOCV tends to have higher variance than does the test error estimate resulting from k-fold CV.



Bias-Variance Tradeoff for K-fold CV

- There is a bias-variance trade-off associated with the choice of *k* in k-fold cross-validation.
- Typically, given these considerations, one performs k-fold cross-validation using *k* = 5 or *k* = 10, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance.

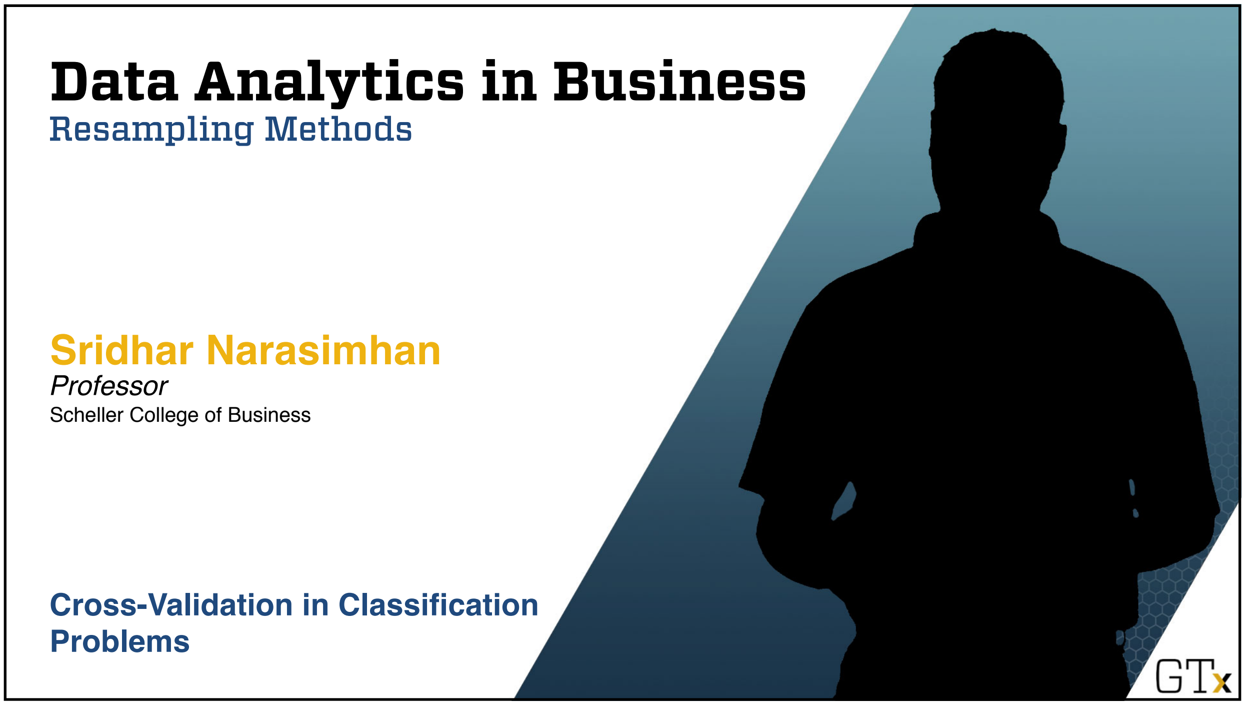


Summary

- Bias-Variance Tradeoff for K-fold Cross-Validation

Next

- Cross-Validation in Classification Problems



Data Analytics in Business

Resampling Methods

Sridhar Narasimhan
Professor
Scheller College of Business

Cross-Validation in Classification Problems

Cross-Validation in Classification

- Cross-validation can also be a very useful approach in the classification setting when Y is qualitative.
- Instead of using MSE to quantify test error, we instead use the number of misclassified observations.
- For instance, in the classification setting, the LOOCV error rate takes the form

$$CV_{(n)}= \frac{1}{n} \sum_{i=1}^{n} Err_i$$

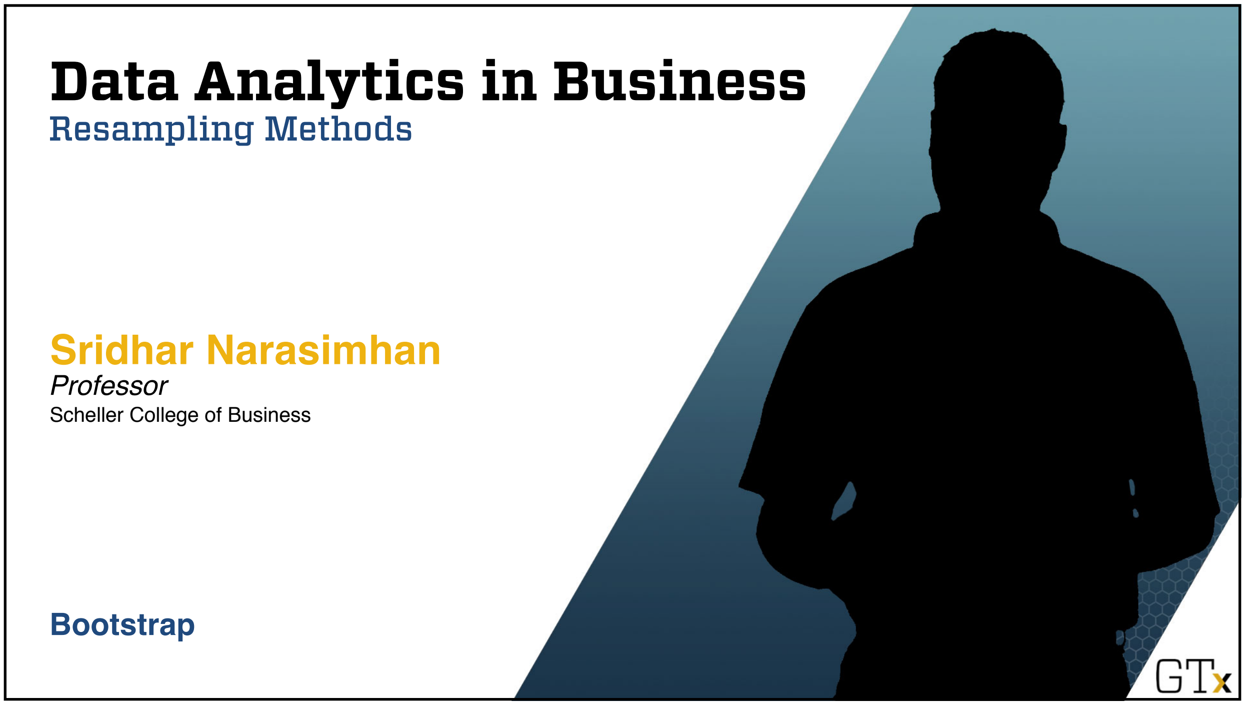
where $ERR_i = I(y_i \neq \hat{y}_i), \text{i.e.}, ERR_i = 1 \; if (y_i \neq \hat{y}_i) \text{ and 0 otherwise.}$
- The k-fold CV error rate and validation set error rates are defined analogously.
- See section 5.1.5 of ISLR for more details.

Summary

- Cross-Validation in Classification Problems

Next

- Bootstrap



Data Analytics in Business

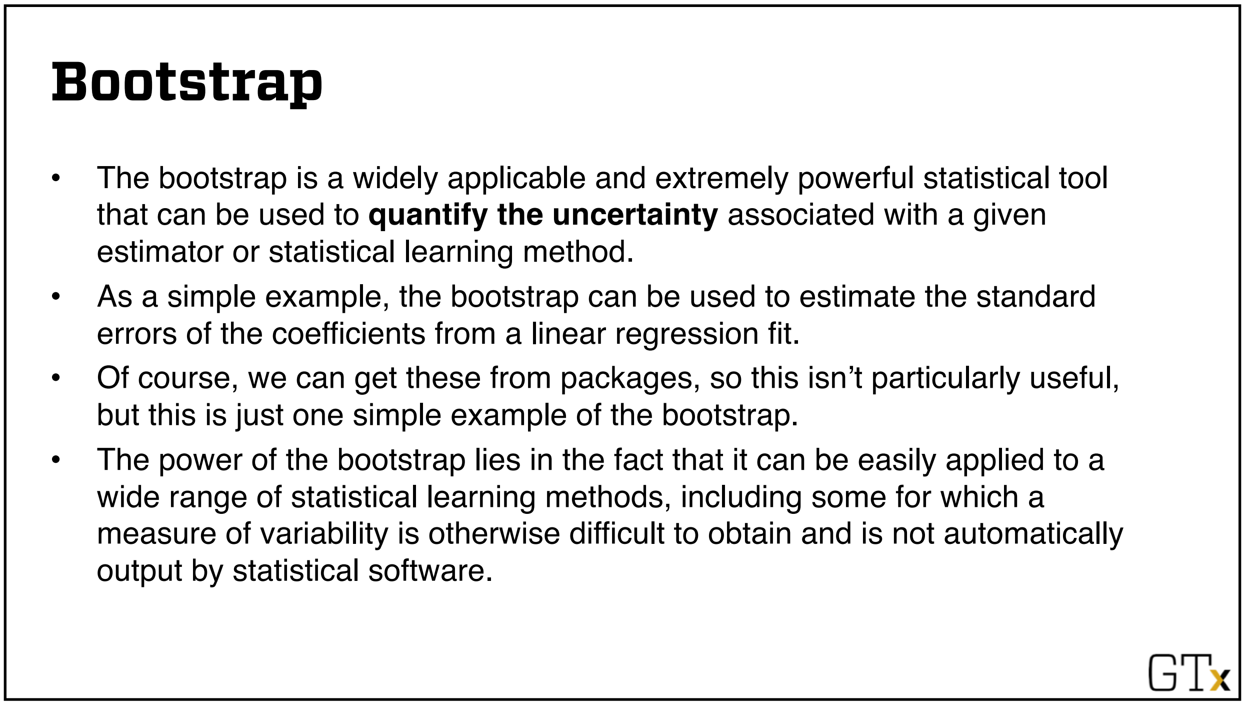
Resampling Methods

Sridhar Narasimhan

Professor

Scheller College of Business

Bootstrap



Bootstrap

- The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to **quantify the uncertainty** associated with a given estimator or statistical learning method.
- As a simple example, the bootstrap can be used to estimate the standard errors of the coefficients from a linear regression fit.
- Of course, we can get these from packages, so this isn't particularly useful, but this is just one simple example of the bootstrap.
- The power of the bootstrap lies in the fact that it can be easily applied to a wide range of statistical learning methods, including some for which a measure of variability is otherwise difficult to obtain and is not automatically output by statistical software.

Bootstrap

- For real data we cannot generate new samples from the original population.
- However, the bootstrap approach allows us to use a computer to emulate the process of obtaining new sample sets so that we can estimate the variability of an estimator without generating additional samples.
- Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set.
- See ISLR, page 195 – 197 for an excellent lab on applying the bootstrap to linear regression. Please work through this on your own.

Summary

- Bootstrap

Next

- CV and Auto Data Set Application

Data Analytics in Business

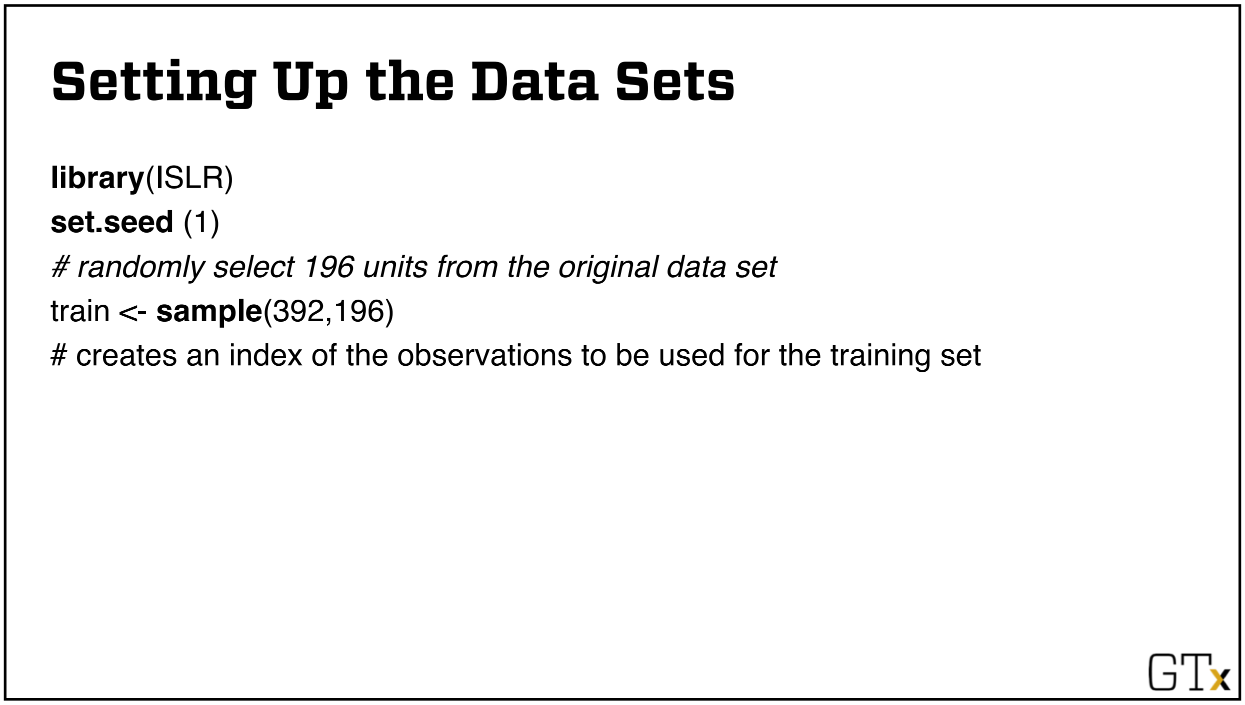
Resampling Methods

Sridhar Narasimhan
Professor
Scheller College of Business

CV and Auto Data Set Application

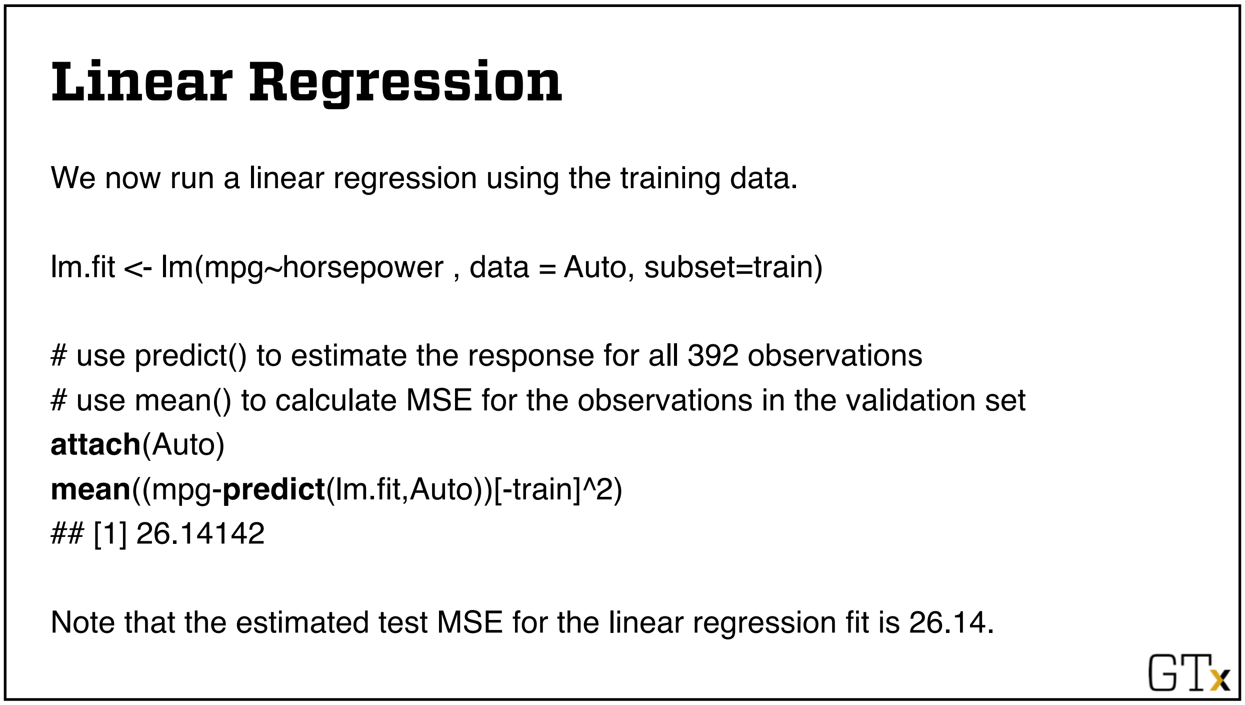
CV and Auto Data Set Application

- We explore the use of the Validation Set Approach in order to estimate the test error rates that result from fitting various linear models on the **Auto** data set.
- We first do some data setup and pre-processing before we begin.
- We set a seed to make our methods reproducible.
- Use the sample() function to split the set of observations into two equal parts.
- We randomly select a random subset of 196 observations out of the original data set of 392 observations.
- We refer to this new sample as the **training set**.



Setting Up the Data Sets

```
library(ISLR)
set.seed (1)
# randomly select 196 units from the original data set
train <- sample(392,196)
# creates an index of the observations to be used for the training set
```



Linear Regression

We now run a linear regression using the training data.

```
lm.fit <- lm(mpg~horsepower , data = Auto, subset=train)

# use predict() to estimate the response for all 392 observations
# use mean() to calculate MSE for the observations in the validation set
attach(Auto)
mean((mpg-predict(lm.fit,Auto))[-train]^2)
## [1] 26.14142
```

Note that the estimated test MSE for the linear regression fit is 26.14.

Linear Regression

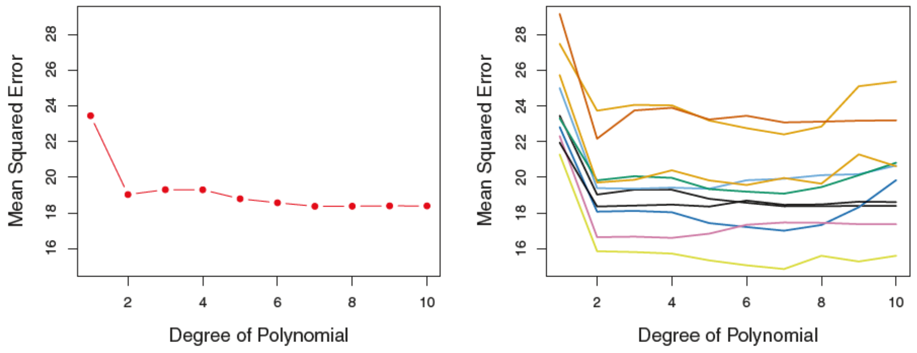
We can use the poly() function to estimate the test error for the quadratic and cubic regressions.

```
lm.fit2=lm(mpg ~ poly(horsepower ,2),data=Auto, subset=train)
mean((mpg-predict(lm.fit2,Auto))[-train]^2)
## [1] 19.82259
lm.fit3=lm(mpg ~ poly(horsepower ,3),data=Auto, subset=train)
mean((mpg — predict(lm.fit3,Auto))[-train]^2)
## [1] 19.78252
```

These error rates are 19.82 and 19.78, respectively. If we choose a different training set instead, then we will obtain somewhat different errors on the validation set. (Verify this on your own).

Summary of Validation Method

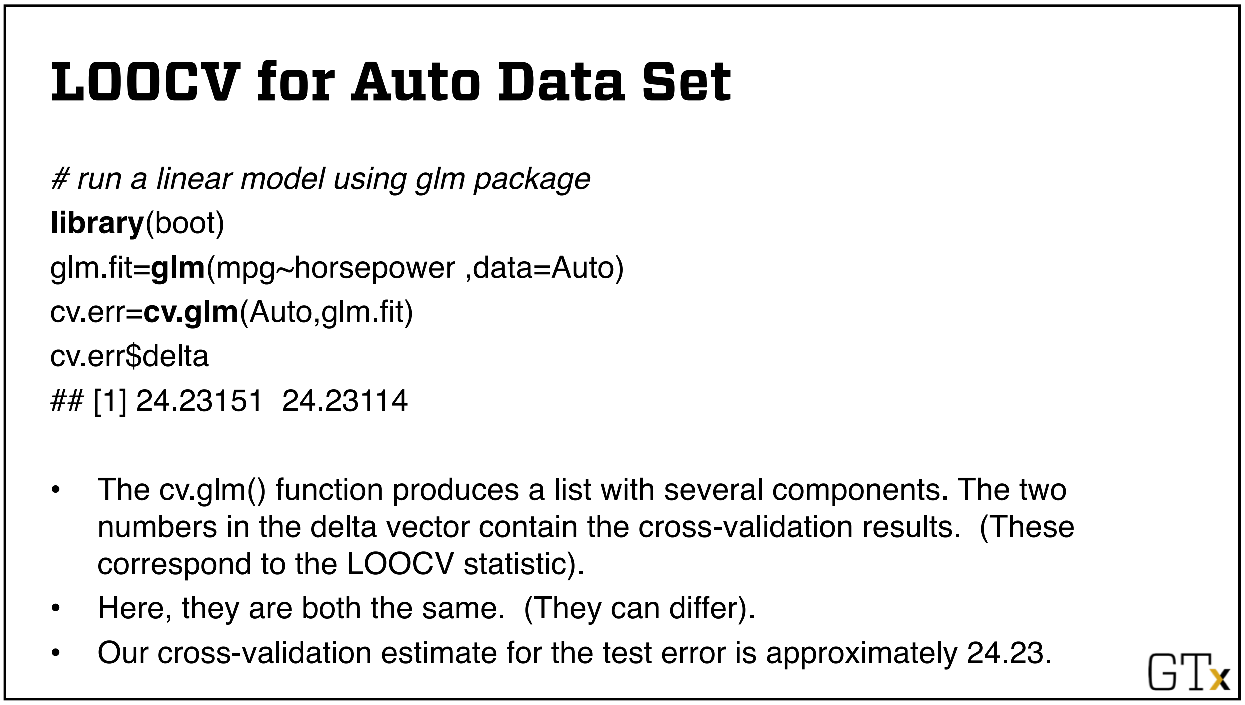
- Using this split of the observations into a training set and a validation set, we find that the validation set error rates for the models with linear, quadratic, and cubic terms are 26.14142, 19.82259, and 19.78252, respectively.
- These results suggest that a model that predicts mpg using a quadratic function of horsepower performs better than a model that involves only a linear function of horsepower, and there is little evidence in favor of a model that uses a cubic function of horsepower.



- The figures above show the Validation Set Approach used on the **Auto** data set in order to estimate the test error that results from prediction **mpg** using polynomial functions of **horsepower**.
- **Left**: Validation error estimates for a single split into training and validation data sets.
- **Right**: The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set.
- This illustrates the variability in the estimated test MSE that results from this approach.

LOOCV for Auto Data Set

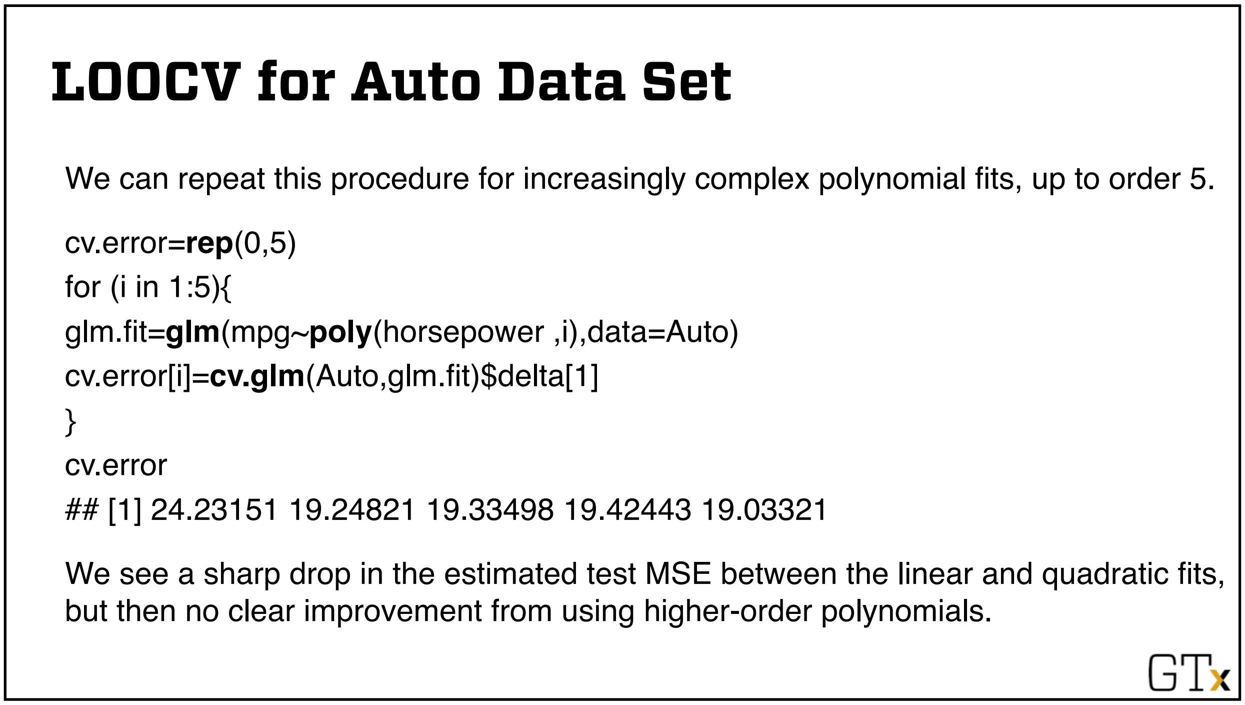
- The LOOCV estimate can be automatically computed for any generalized linear model using the glm() and cv.glm() functions.
- If we do not give the glm() a family, then it will perform linear regression.
- The cv.glm() function is part of the boot library.



LOOCV for Auto Data Set

```
# run a linear model using glm package
library(boot)
glm.fit=glm(mpg~horsepower ,data=Auto)
cv.err=cv.glm(Auto,glm.fit)
cv.err$delta
## [1] 24.23151  24.23114
```

- The cv.glm() function produces a list with several components. The two numbers in the delta vector contain the cross-validation results. (These correspond to the LOOCV statistic).
- Here, they are both the same. (They can differ).
- Our cross-validation estimate for the test error is approximately 24.23.



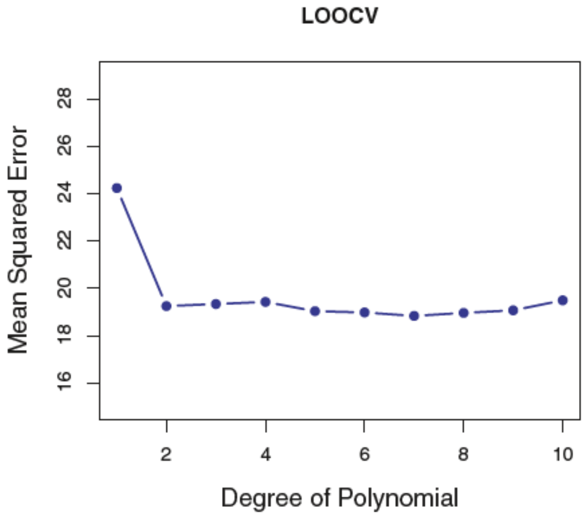
LOOCV for Auto Data Set

We can repeat this procedure for increasingly complex polynomial fits, up to order 5.

```
cv.error=rep(0,5)
for (i in 1:5){
glm.fit=glm(mpg~poly(horsepower ,i),data=Auto)
cv.error[i]=cv.glm(Auto,glm.fit)$delta[1]
}
cv.error
## [1] 24.23151 19.24821 19.33498 19.42443 19.03321
```

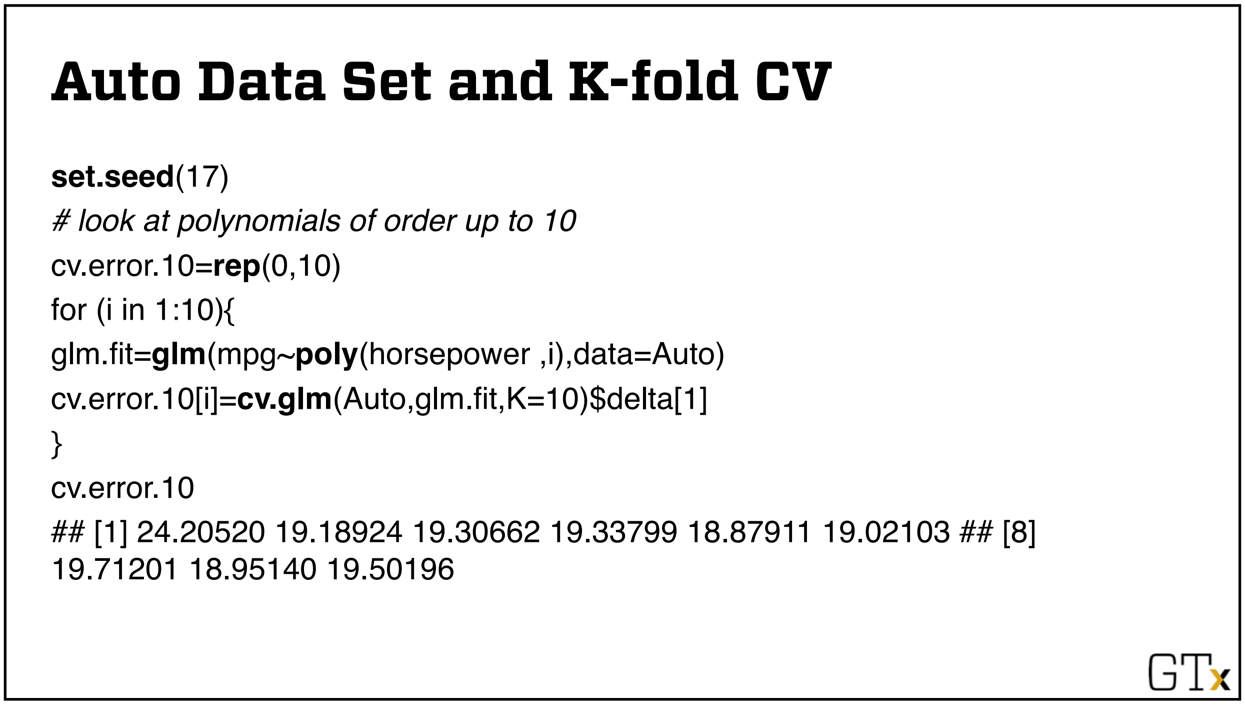
We see a sharp drop in the estimated test MSE between the linear and quadratic fits, but then no clear improvement from using higher-order polynomials.

LOOCV Error Curve



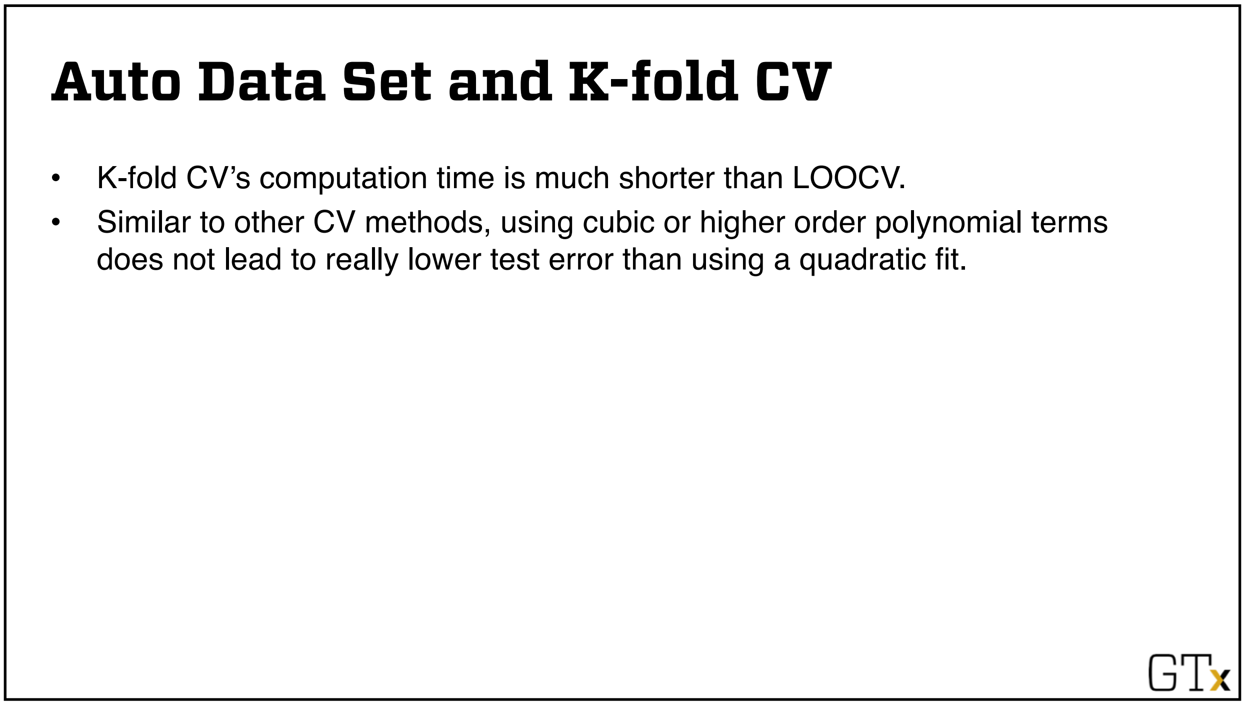
Auto Data Set and K-fold CV

- The cv.glm() function can also be used to implement k-fold CV.
- We will use k = 10 on the Auto data set.
- We once again set a random seed and initialize a vector in which we will store the CV errors corresponding to the polynomial fits of orders one to ten.



Auto Data Set and K-fold CV

```
set.seed(17)
# look at polynomials of order up to 10
cv.error.10=rep(0,10)
for (i in 1:10){
glm.fit=glm(mpg~poly(horsepower ,i),data=Auto)
cv.error.10[i]=cv.glm(Auto,glm.fit,K=10)$delta[1]
}
cv.error.10
## [1] 24.20520 19.18924 19.30662 19.33799 18.87911 19.02103 ## [8]
19.71201 18.95140 19.50196
```



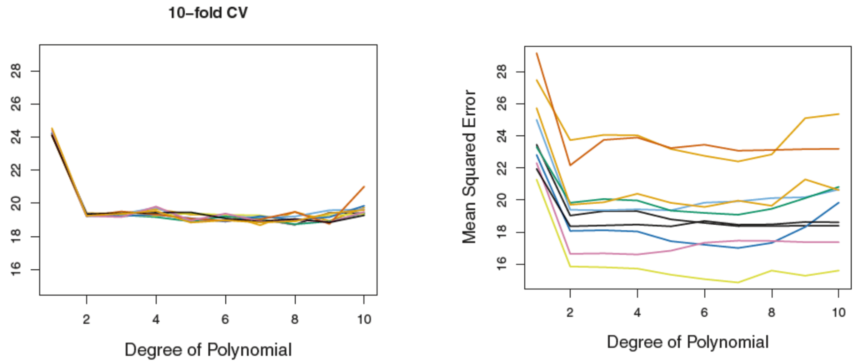
Auto Data Set and K-fold CV

- K-fold CV's computation time is much shorter than LOOCV.
- Similar to other CV methods, using cubic or higher order polynomial terms does not lead to really lower test error than using a quadratic fit.

Delta for CV

- We saw that the two numbers associated with delta are essentially the same when LOOCV is performed.
- When we instead perform k-fold CV, then the two numbers associated with delta differ slightly.
  - The first is the standard k-fold CV estimate.
  - The second is a bias-corrected version.
- On this data set, the two estimates are very similar to each other (but not shown here).

CV Error Curves



- The figure on the left shows the CV error curves for 10-fold CV run 9 times, each with a different random split.
- Contrast this with the figure on the right that we saw earlier for the Validation Method.

Summary

- CV and Auto Data Set Application

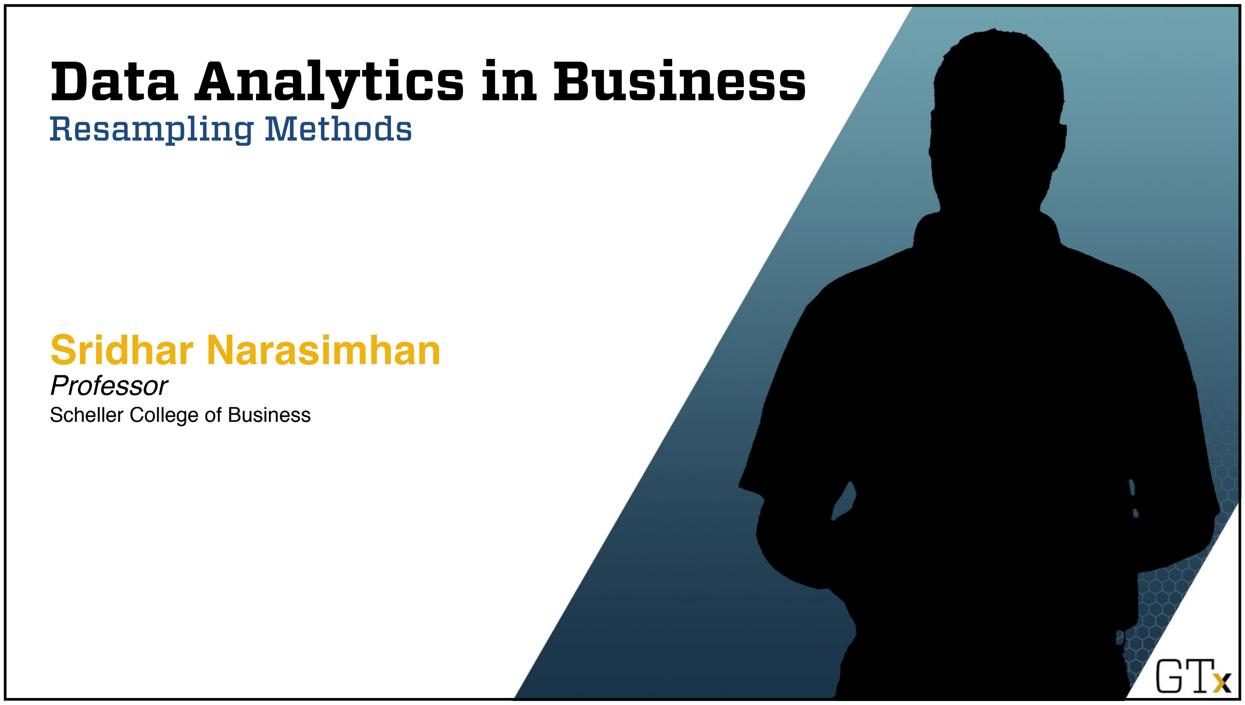
Next

- End of Module

Recap of the Lessons

A. Introduction
B. The Validation Set Approach
C. Leave-One-Out Cross-Validation (LOOCV)
D. K-fold Cross Validation
E. Bias-Variance Tradeoff for K-fold Cross-Validation
F. Cross-Validation in Classification Problems
G. Bootstrap
H. CV and Auto Data Set Application

- This module is based on Chapter 5 in the ISLR textbook.



Data Analytics in Business

Resampling Methods

Sridhar Narasimhan

Professor

Scheller College of Business