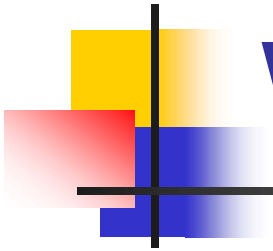
Session 6
Resampling Methods

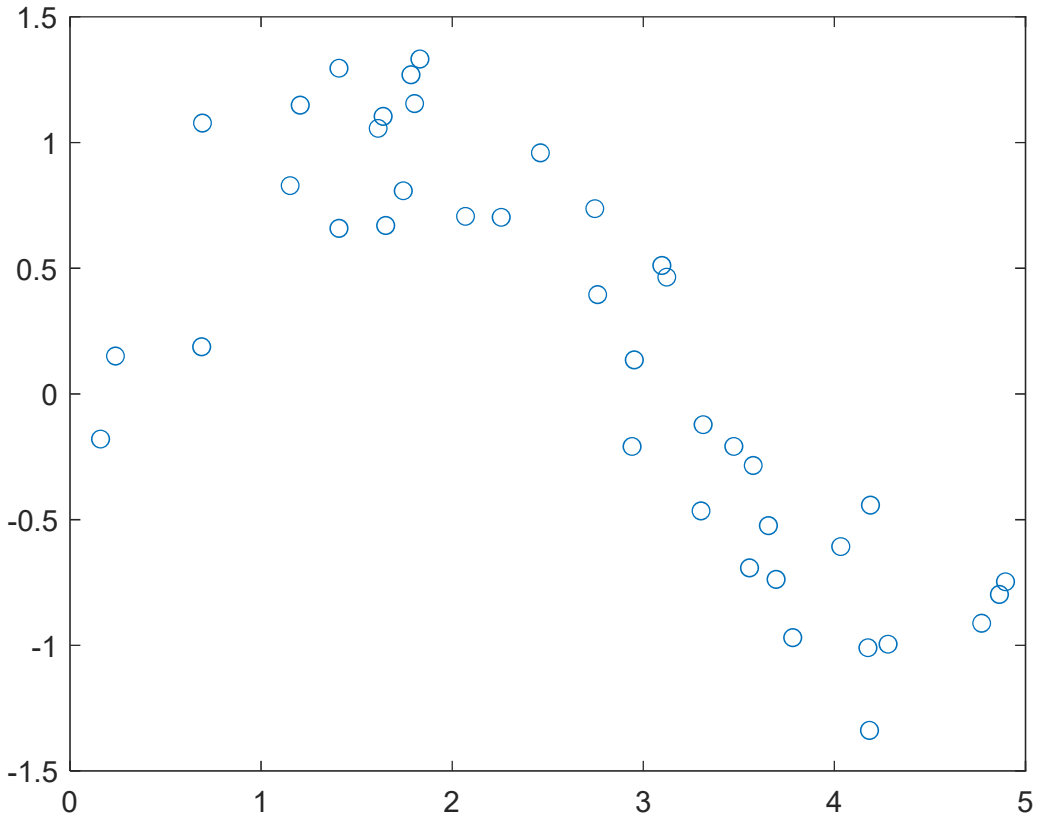
PSYC 560
Gyeongcheol Cho
Heungsun Hwang



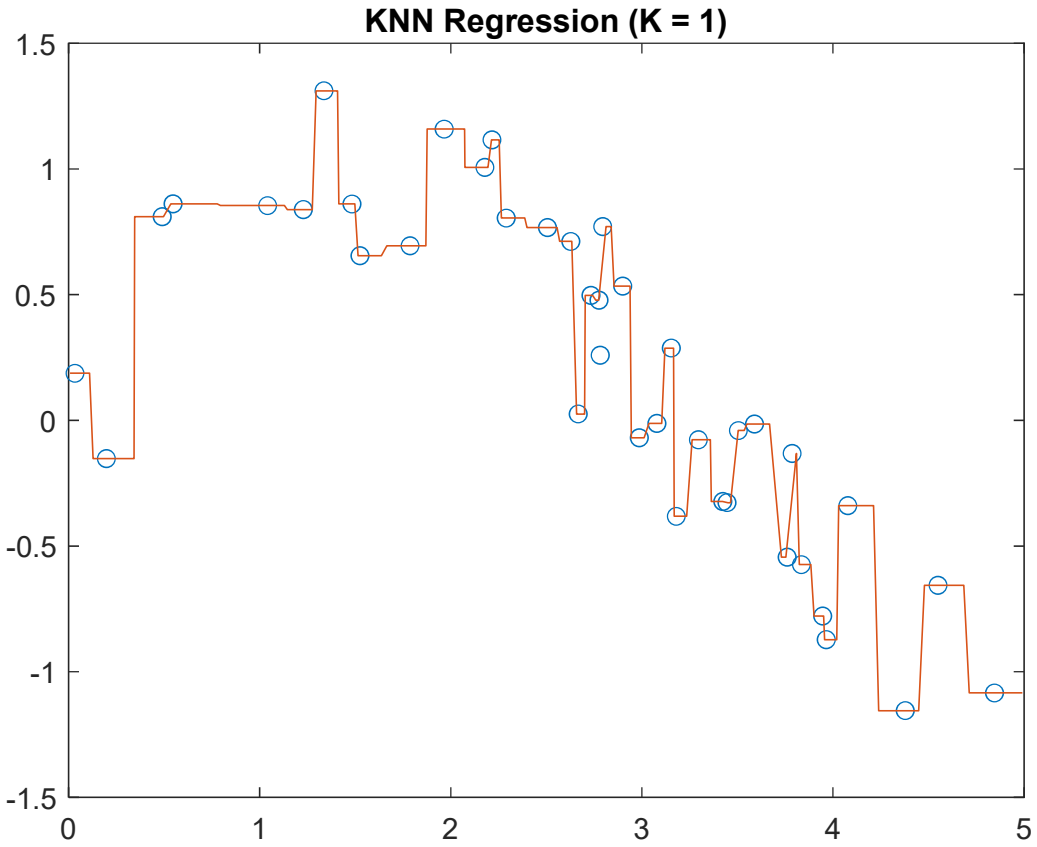
Why Resampling Methods?

- A researcher trained a linear regression model on a sample. Now he wants to obtain the test error of the model but does not have a designated test sample.
- A researcher wants to apply a KNN regression to her dataset but is not sure which value she should choose for *K*.
- A researcher trained a linear regression model and wants to conduct hypothesis tests for individual coefficients. However, the normality assumption seems to be violated.

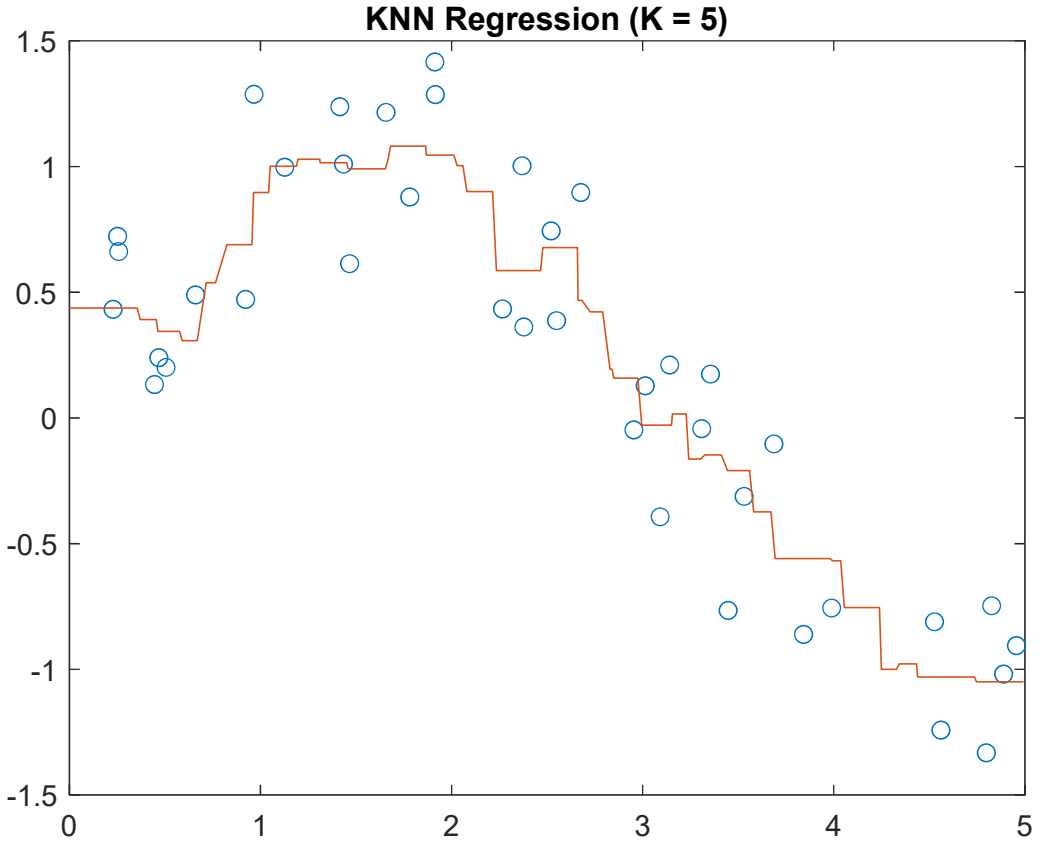
K-Nearest Neighbors Regression



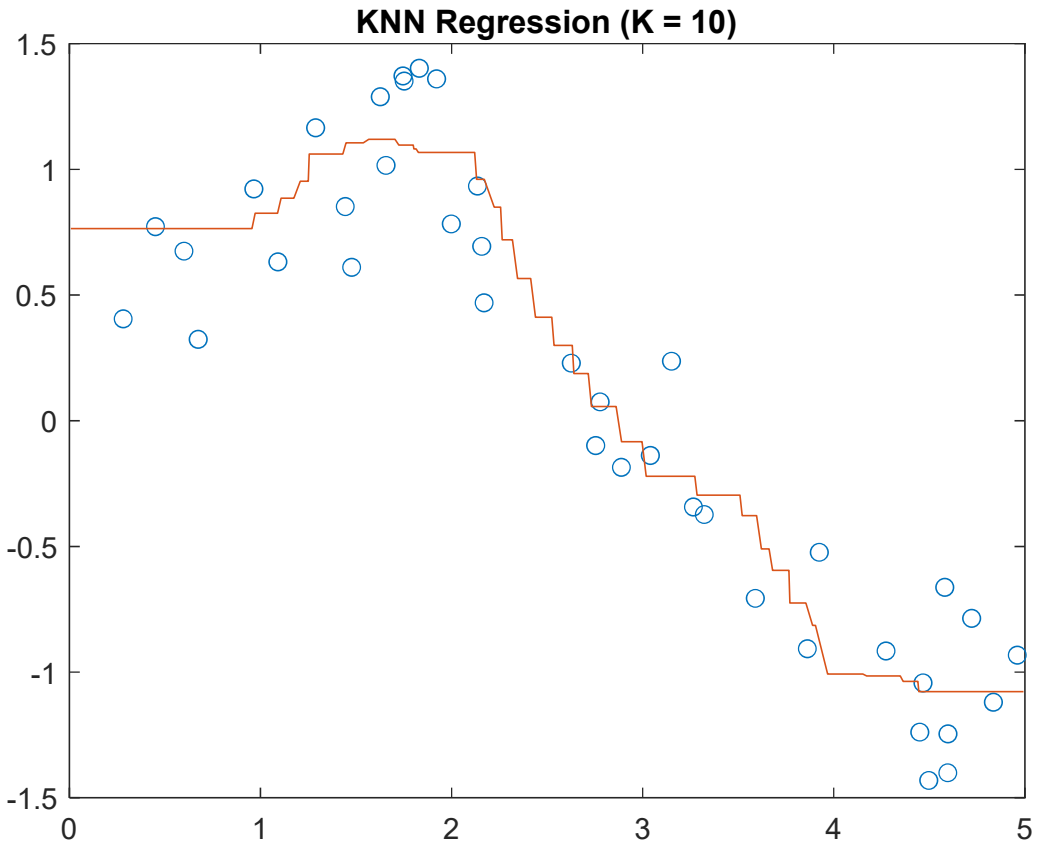
K-Nearest Neighbors Regression

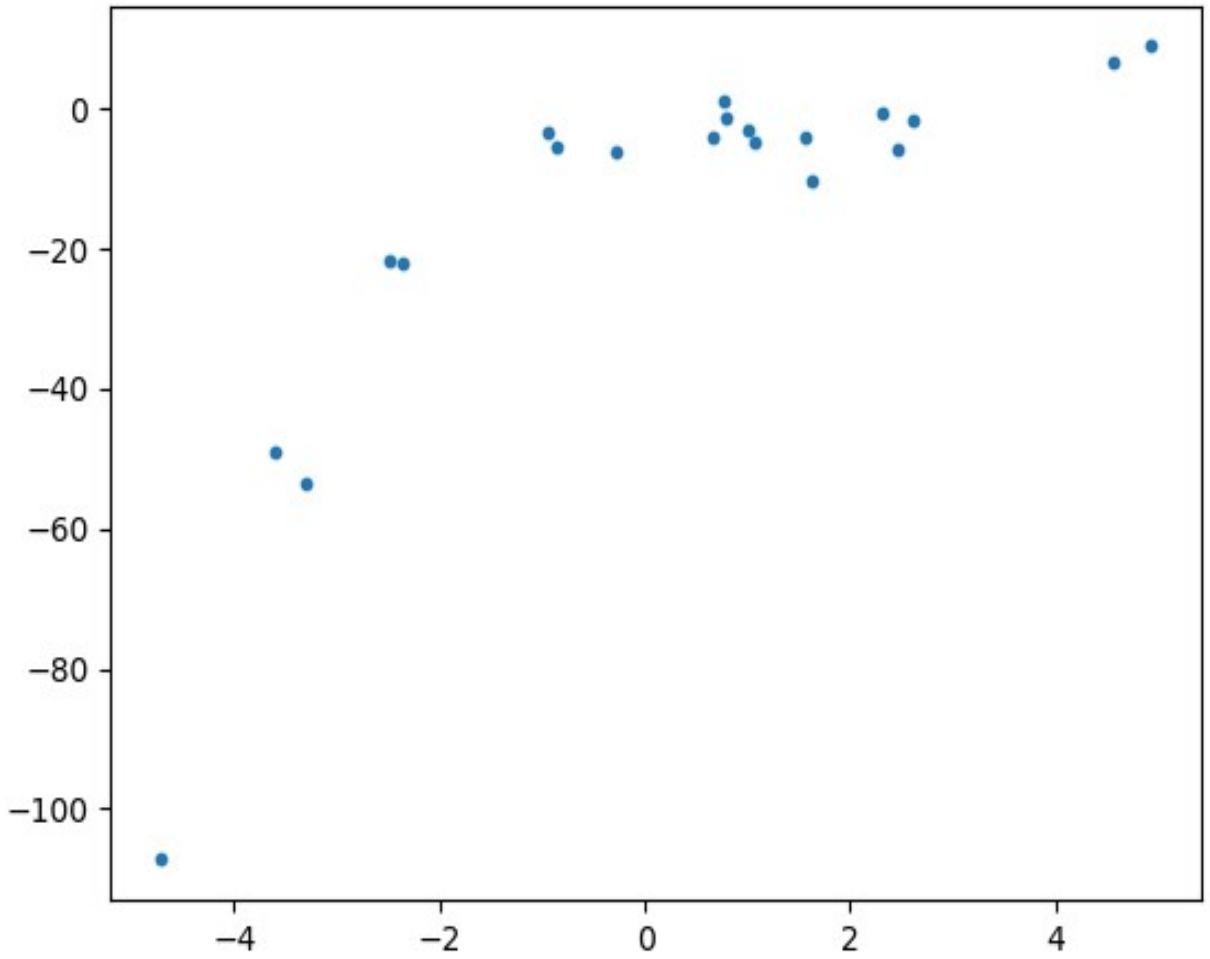


K-Nearest Neighbors Regression



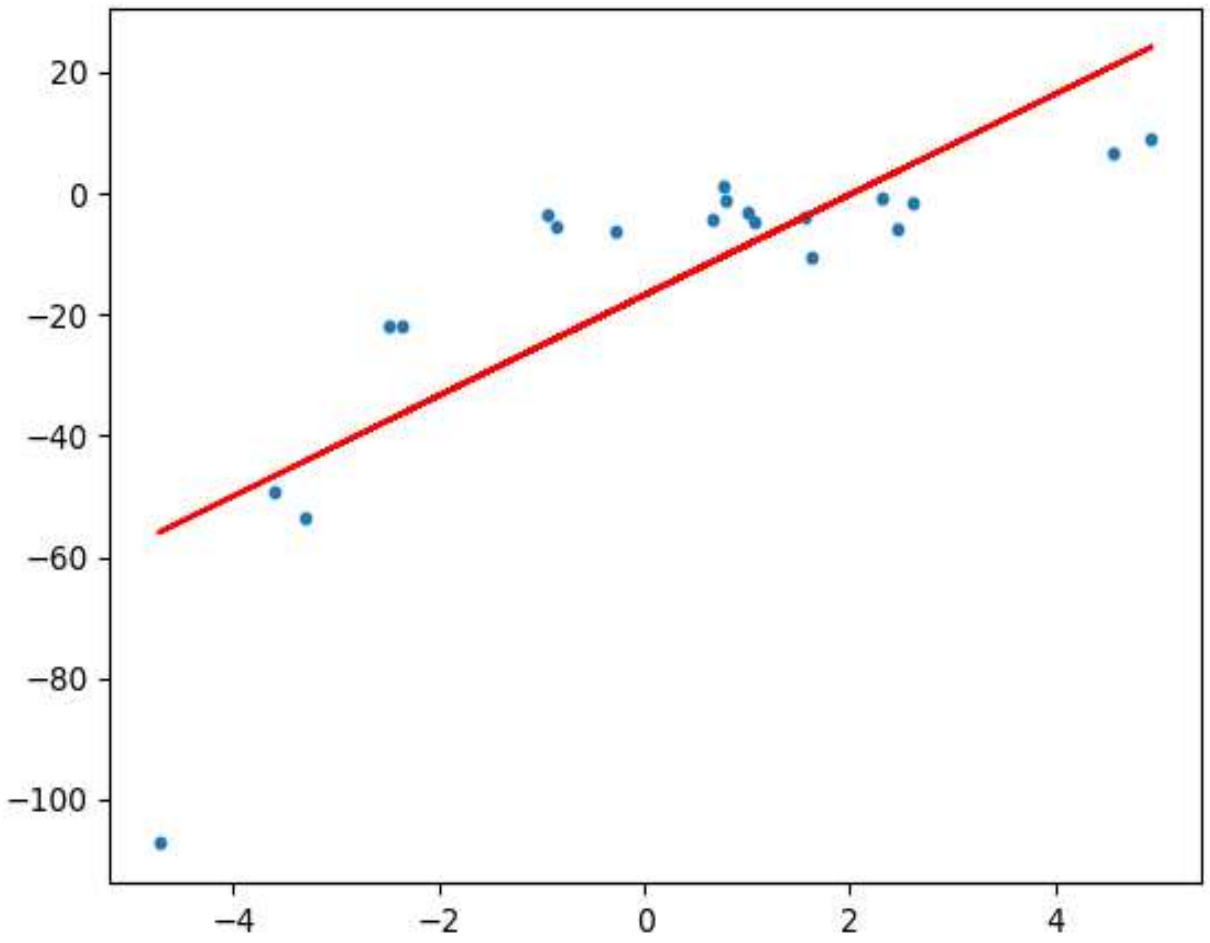
K-Nearest Neighbors Regression





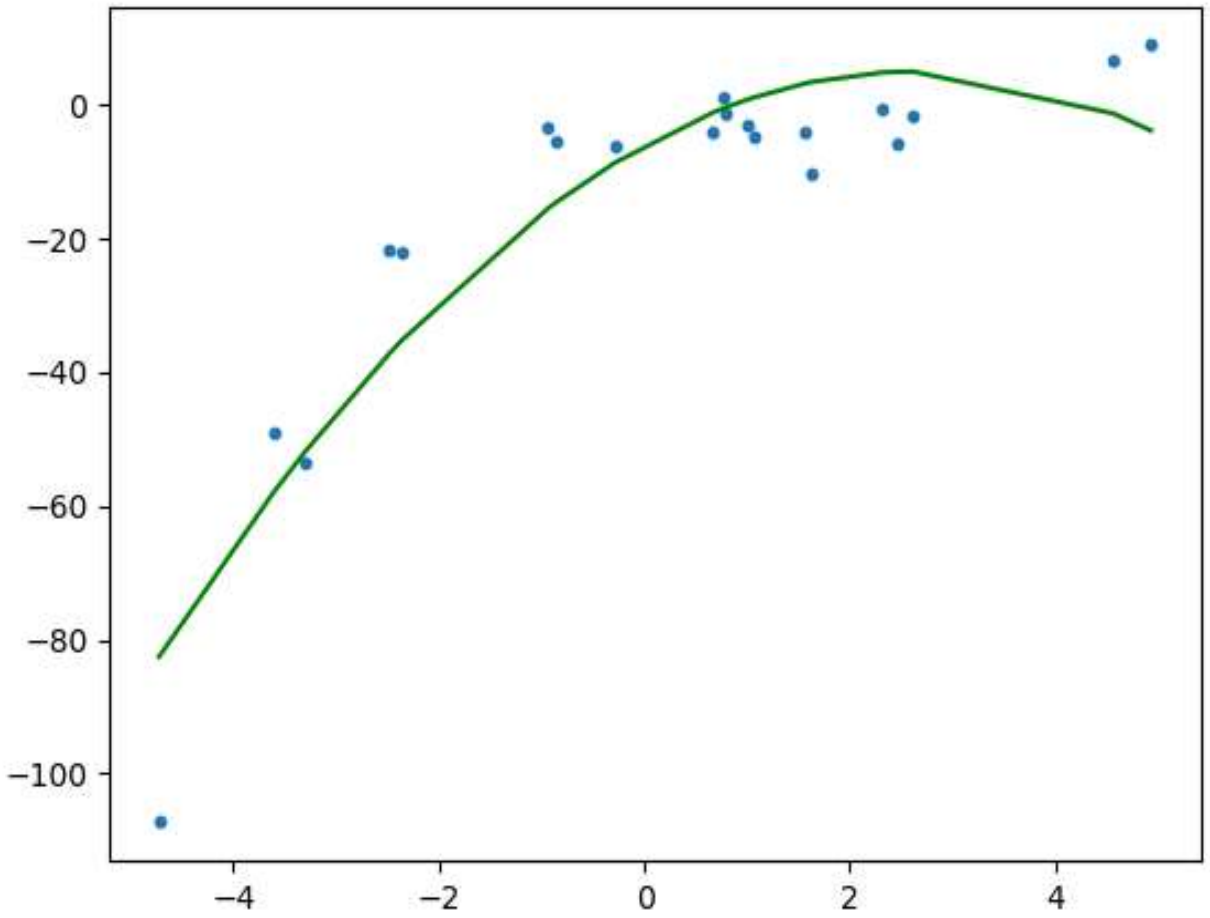
https://towardsdatascience.com/polynomial-regression-bbe8b9d97491

Linear regression



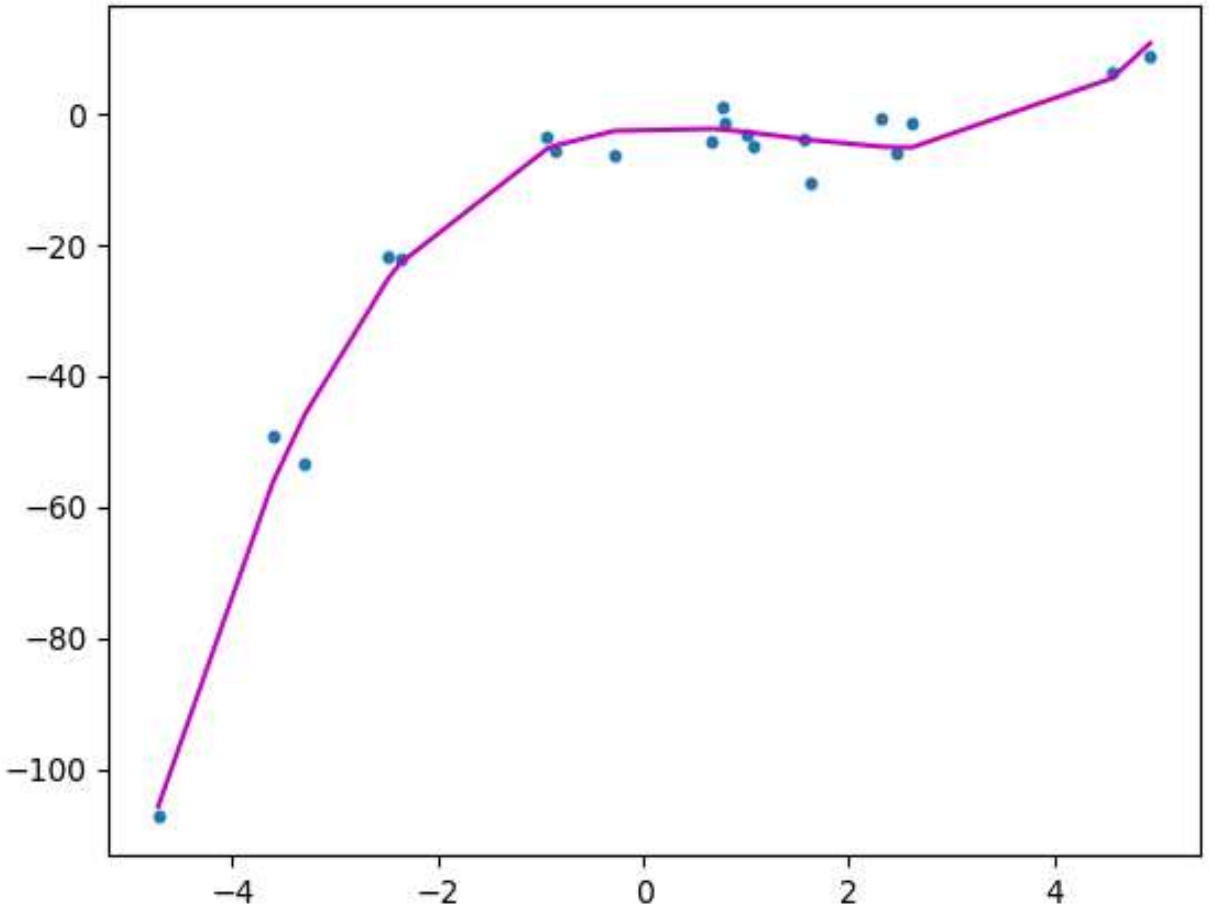
https://towardsdatascience.com/polynomial-regression-bbe8b9d97491

Polynomial regression (degree = 2)

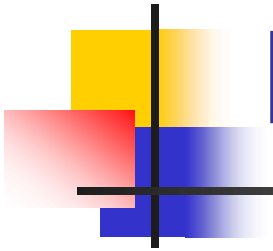


https://towardsdatascience.com/polynomial-regression-bbe8b9d97491

Polynomial regression (degree = 3)

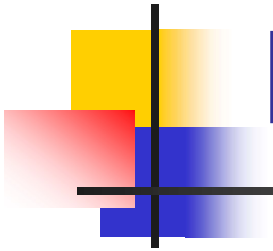


https://towardsdatascience.com/polynomial-regression-bbe8b9d97491



Resampling Methods

- Resampling methods involve repeatedly drawing samples from a training set and refitting the same model to each sample to obtain additional information about the fitted model, such as:
  - Test error estimate
  - Standard errors or confidence intervals



Resampling Methods

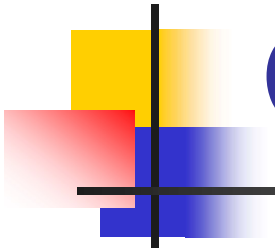
- Cross validation is used to estimate the test error associated with a given statistical model/method to evaluate its performance (model assessment) or to select the appropriate level of model flexibility (model selection).

- The bootstrap is used in several contexts, most commonly to provide a measure of the accuracy of a parameter estimate or of a statistical method.

Why Resampling Methods?

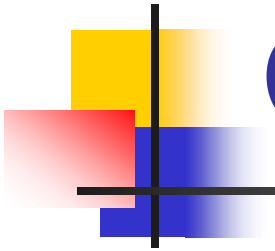
- A researcher trained a linear regression model on a sample. Now he wants to obtain the test error of the model but does not have a designated test sample.
- A researcher wants to apply a KNN regression to her dataset but is not sure which value she should choose for *K*.

- A researcher trained a linear regression model and wants to conduct hypothesis tests for individual coefficients. However, the normality assumption seems to be violated.



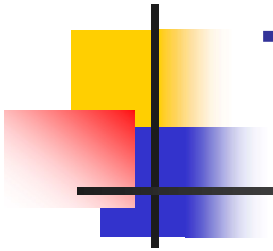
Cross Validation

- Cross validation refers to a class of resampling methods that estimate the test error by holding out a subset of the training observations from the fitting process and then applying a statistical method to those hold-out observations.
  - The test error is the average error that results from using a statistical method/model to predict the response on a new observation, i.e., a measurement that was not used in training the method.



Cross Validation Methods

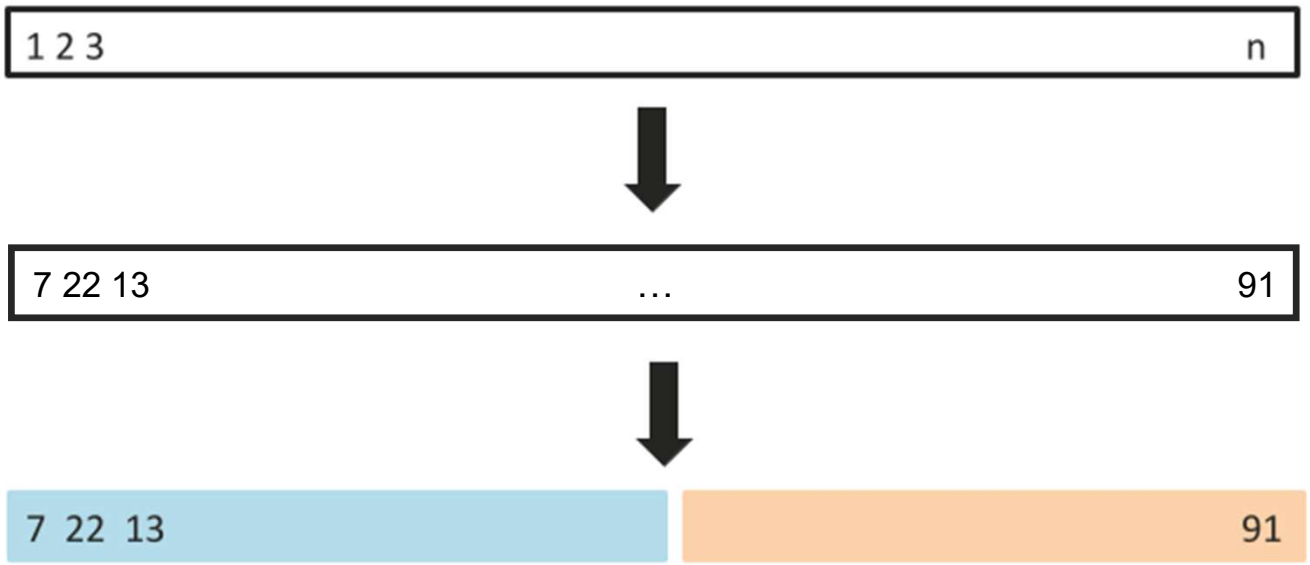
- The validation set approach
- Leave-one-out cross validation
- K-fold cross validation



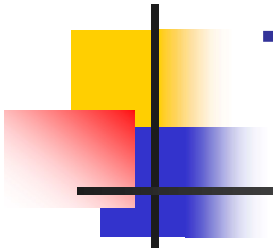
The Validation Set Approach

- This approach involves randomly dividing a set of observations into two subsets of comparable size, a training set and a validation (hold-out) set.

The Validation Set Approach



Revised from Figure 5.1 in James et al., (2021, p. 199)



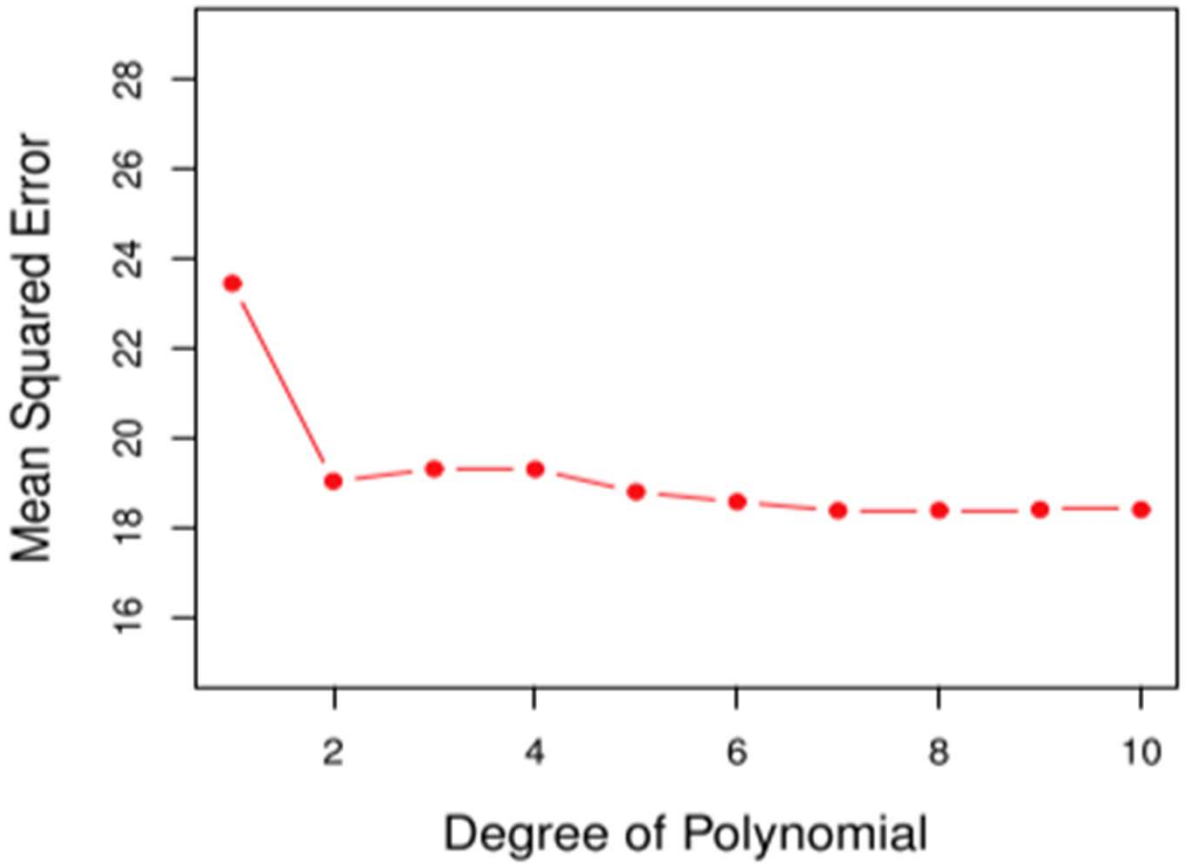
The Validation Set Approach

- The model is fit to the training set, and the fitted model is used to predict the responses for the observations in the validation set. The resulting validation set error rate (e.g., MSE in the case of quantitative responses) provides an estimate of the test error rate.

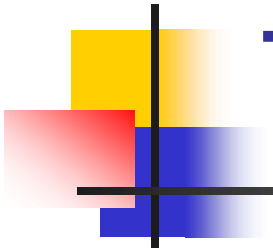
Example: The Validation Set Approach

- **Auto MPG data set (auto_mpg.csv)**
  - It includes eight attributes of cars that can be used to predict their mile per gallon (mpg)
    1. mpg: continuous
    2. cylinders: multi-valued discrete
    3. displacement: continuous
    4. horsepower: continuous
    5. weight: continuous
    6. acceleration: continuous
    7. model year: multi-valued discrete
    8. origin: multi-valued discrete
    9. car name: string (unique for each instance)
  - N = 392
  - The original dataset can be downloaded from https://archive.ics.uci.edu/ml/datasets/auto+mpg

Example: The Validation Set Approach



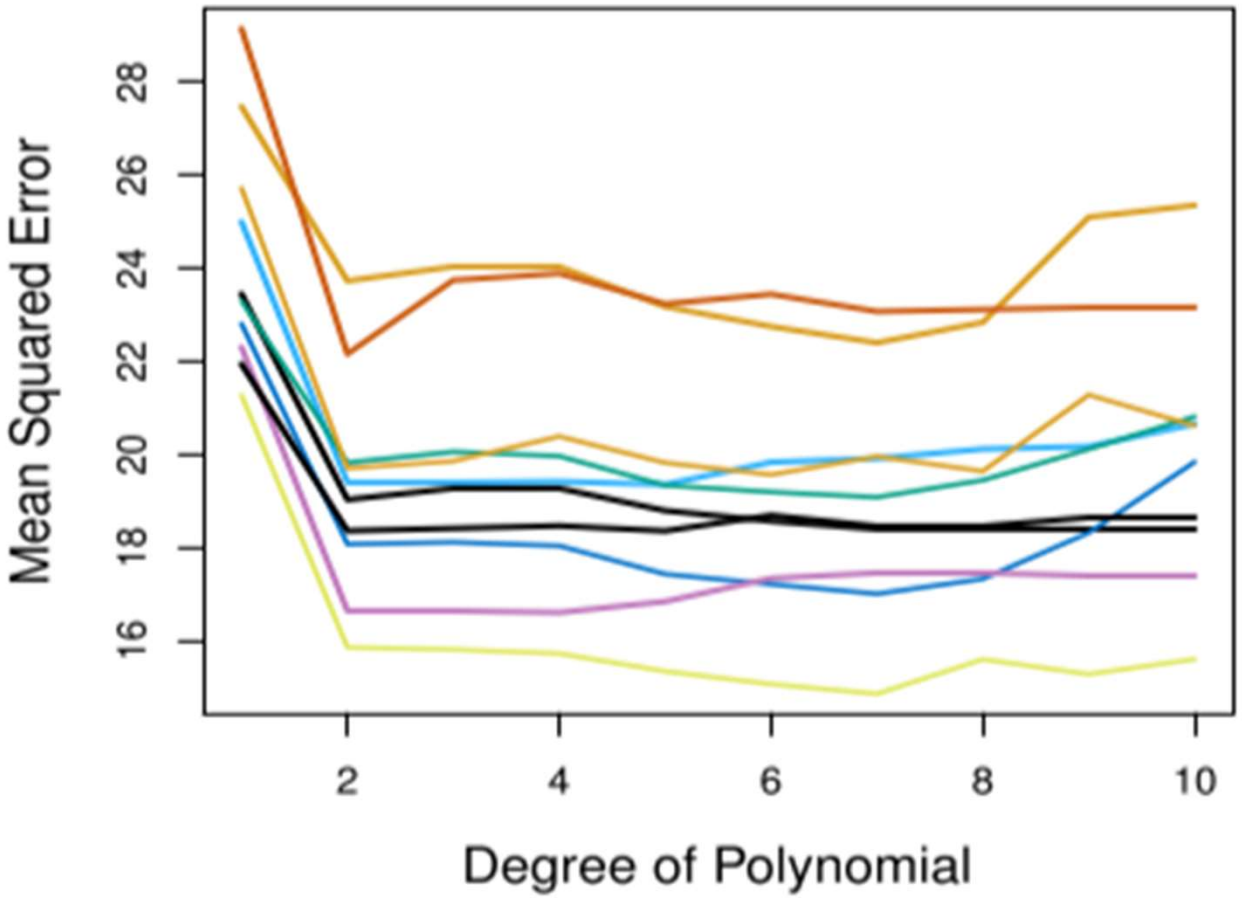
James et al. (2021, p. 200)



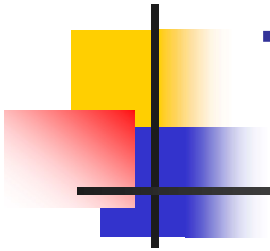
The Validation Set Approach

- The validation set approach is conceptually simple and easy to implement.
- However, the validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are in the validation set.

Example: The Validation Set Approach

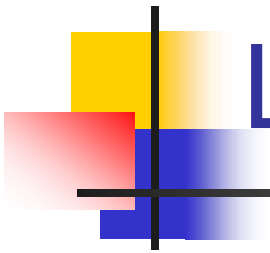


James et al. (2021, p. 200)



The Validation Set Approach

- Moreover, only a subset of all observations (in the training set) is used to fit the model. As statistical methods tend to perform worse when trained on fewer observations, this suggests that the validation set error rate may tend to be a biased estimate of the test error for the model fit on the entire data set.

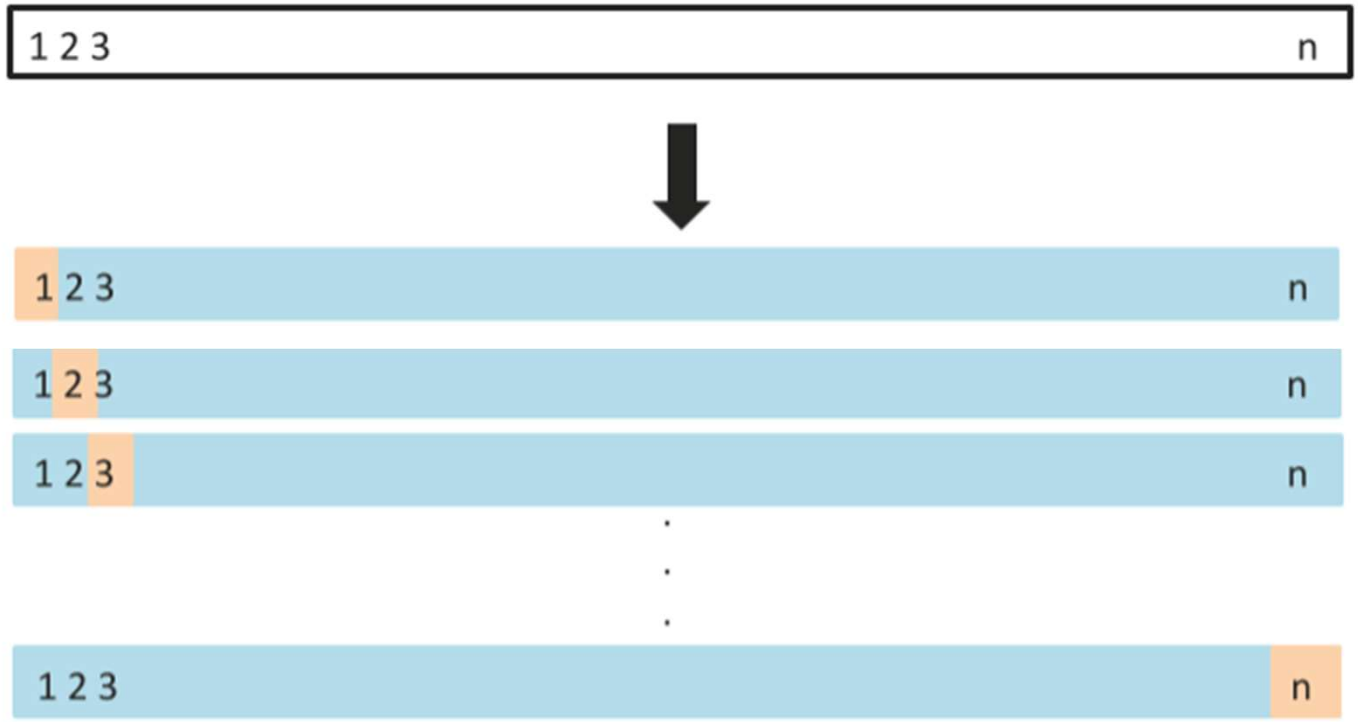


Leave-One-Out Cross Validation

- Leave-one-out cross validation (LOOCV) is closely related to the validation set approach, but it attempts to address that method's drawbacks.

- Like the validation set approach, LOOCV involves splitting the set of observations into two parts. However, instead of creating two subsets of comparable size, a single observation is used for the validation set and the remaining observations make up the training set.

Leave-One-Out Cross Validation

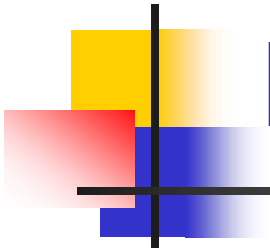


James et al. (2021, Figure 5.3, p. 201)

Leave-One-Out Cross Validation

- The statistical learning method is fit on the N-1 training observations, and a prediction of a single response is made for the excluded observation.
- We can repeat the procedure of training and prediction N times, using each observation as the validation set.
- The LOOCV estimate for the test error is obtained as follows.

$$CV = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{N}\sum_{i=1}^{N} MSE_i$$



Leave-One-Out Cross Validation

- LOOCV has two major advantages over the validation set approach.
  - It has far less bias because a statistical method is fit on training sets that contain N-1 observations, almost as many as are in the entire dataset.
  - Performing LOOCV multiple times will always yield the same results.

Leave-One-Out Cross Validation

- But LOOCV has the potential to be expensive to implement because the model must be fit N times. This can be very time-consuming if N is large, and if each individual model is slow to fit.
  - In (least squares) linear regression, there exists an amazing shortcut that can obtain the LOOCV estimate based on a single model fit.
    - e.g., refer to Equation 5.2 in James et al., 2021, p. 202.

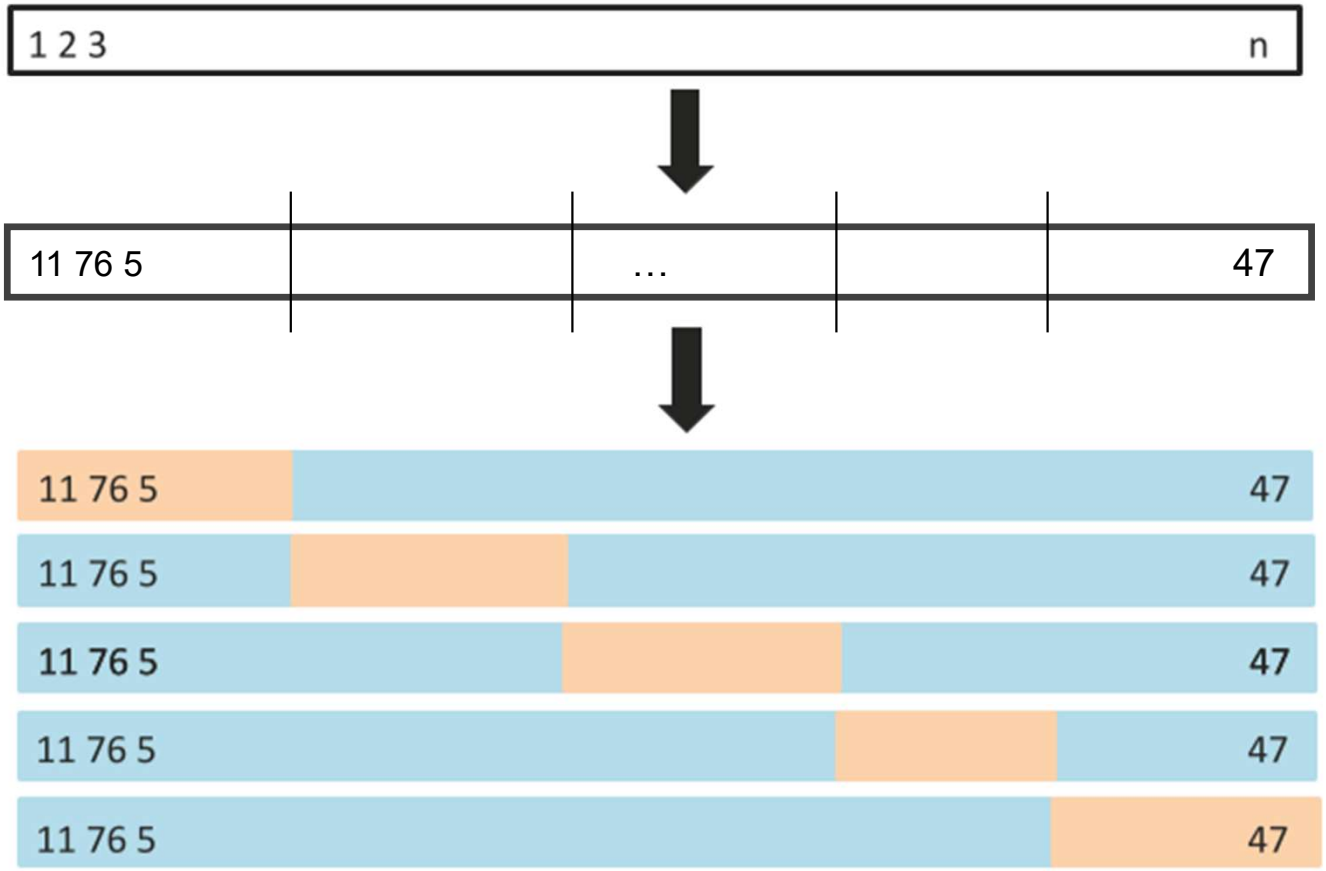
K-Fold Cross Validation

- K-fold CV is an alternative to LOOCV. This approach involves randomly dividing a set of observations into K groups (or folds) of approximately equal size. The first set is treated as a validation set and the method is fit on the remaining K-1 folds. The MSE is then computed on the observations in the validation set. The procedure is repeated K times; each time, the *k*th fold is treated as a validation set.

- The K-fold CV estimate of the test error is computed as:

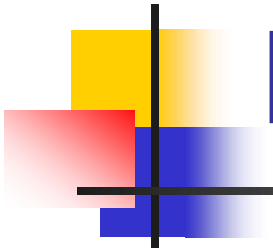
$$CV = \frac{1}{K}\sum_{i=1}^{K} MSE_i$$

K-Fold Cross Validation



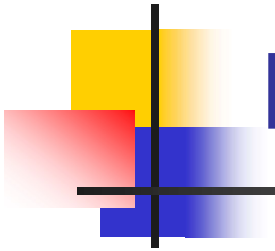
Revised from Figure 5.5 in James et al. (2021, p. 203)

K-Fold Cross Validation



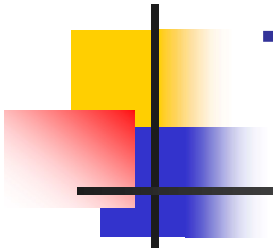
- In practice, K = 5 or 10 is used.
  - If K = N, LOOCV = K-fold CV
- K-fold CV is computationally more efficient than LOOCV especially if N is very large.
- K-fold CV can often give a more accurate estimate of the test error than LOOCV.
  - The LOOCV estimate tends to be less biased yet more highly variable than the K-fold CV estimate. (James et al., 2021, pp. 205-206).

Cross Validation on Classification Problems



- We have thus far illustrated the use of CV in the regression setting where the outcome is quantitative, and so have used MSE to quantify test error.

- CV can also be used in the classification setting where the outcome is qualitative. In this setting, CV works in the same way as in the regression setting, except that we replace MSE with the number of misclassified observations.

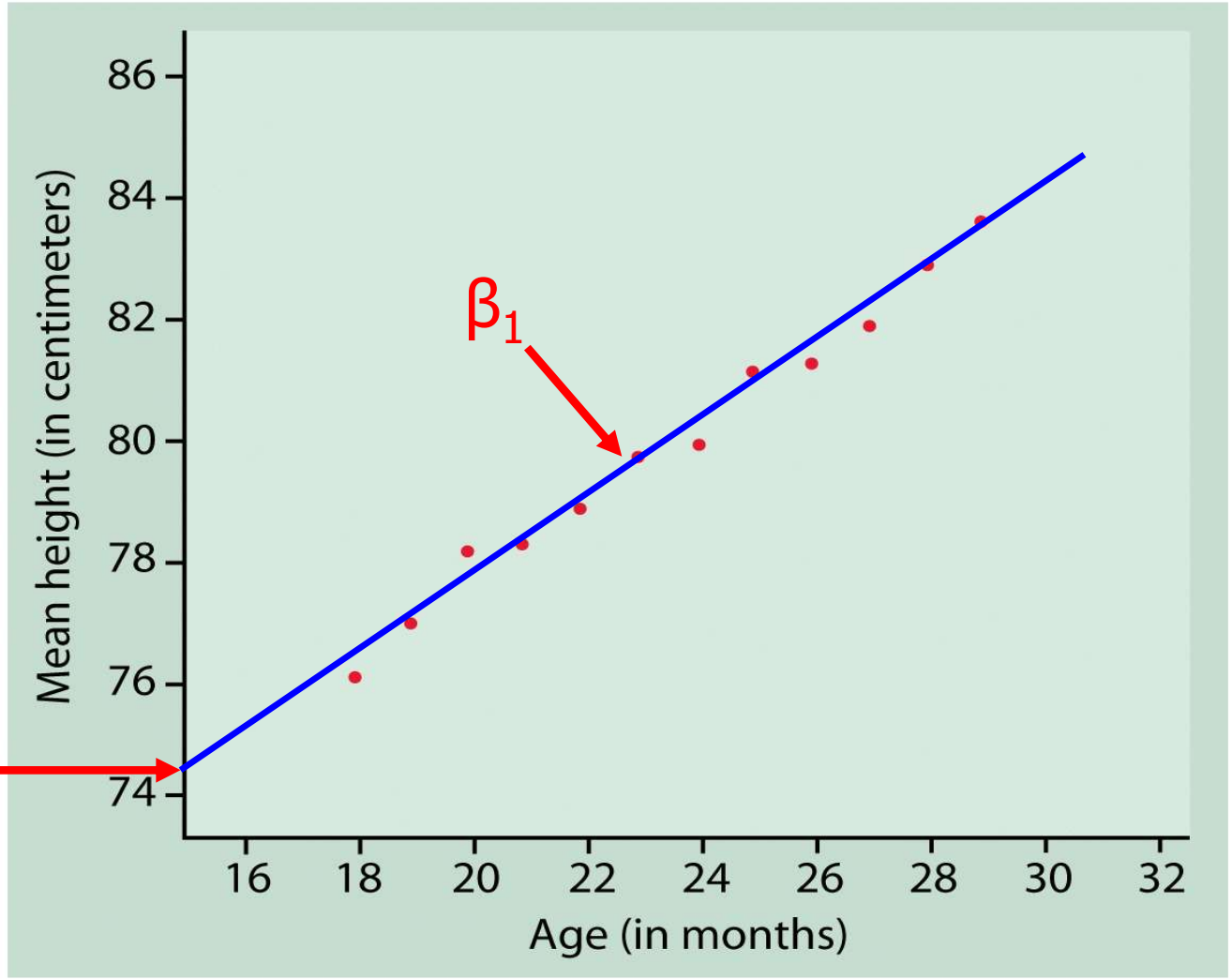


The Bootstrap

- The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given statistical method or estimator.

- For example, it can be used to estimate the standard errors of parameter estimates or their confidence intervals.

Simple Linear Regression



Moore & McCabe (1998, p.136)

Example: LS Coefficient Estimates

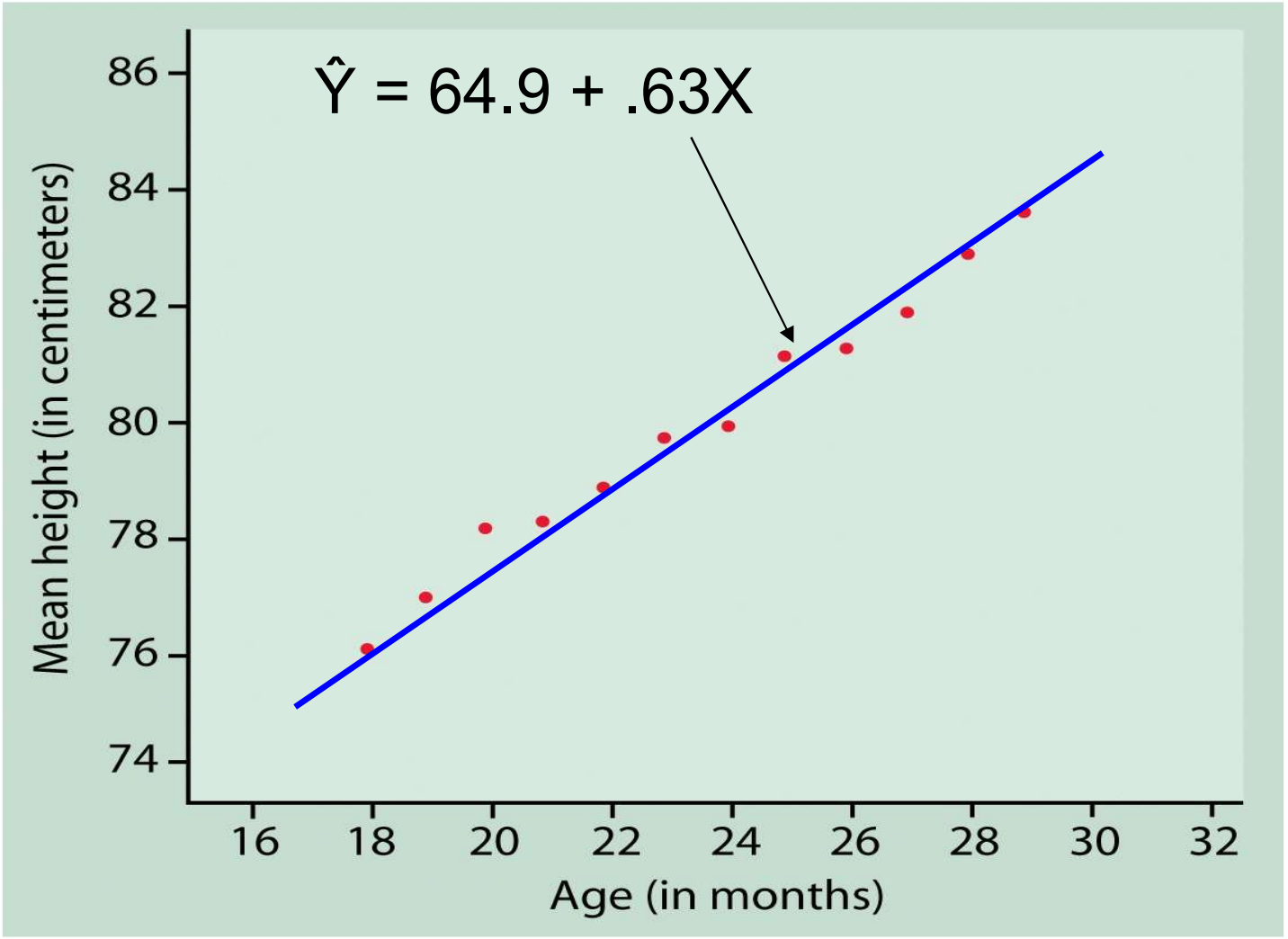
Age X (month)	Height Y (cm)
18	76.1
19	77.0
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5

Moore & McCabe (1998, p.136)

- N = 12
- $\bar{X}$ = 23.5
- $\bar{Y}$ = 79.85

- $\hat{\beta}_1$=.6348
- $\hat{\beta}_0$= 64.93

Example: LS Coefficient Estimates



Moore & McCabe (1998, p.136)

Statistical test for the significance of slope

- We can perform a hypothesis test on the relationship between X and Y (the effect of X on Y)
  - $H_0 : \beta_1 = 0$
    - There is no linear relationship between X and Y (no effect of X on Y)
  - $H_1 : \beta_1 \neq 0$
    - There is a linear relationship (an effect of X on Y)

Statistical test for the significance of slope

- We compute a t-statistic, given by

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$$

where $SE(\hat{\beta}_1)$ is the standard error of the estimate.

Statistical test for the significance of slope

- $SE(\hat{\beta}_1)$ is computed as follows:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N-2}} \Bigg/ \sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

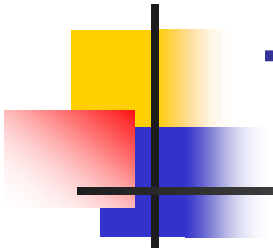
Statistical test for the significance of slope

- If the p-value of the t statistic is small enough, e.g., $p < \alpha = .05$ (or .01), we may reject the null hypothesis.

- This indicates that the slope is different from zero, suggesting a statistically significant effect of X on Y.

Statistical test for the significance of slope

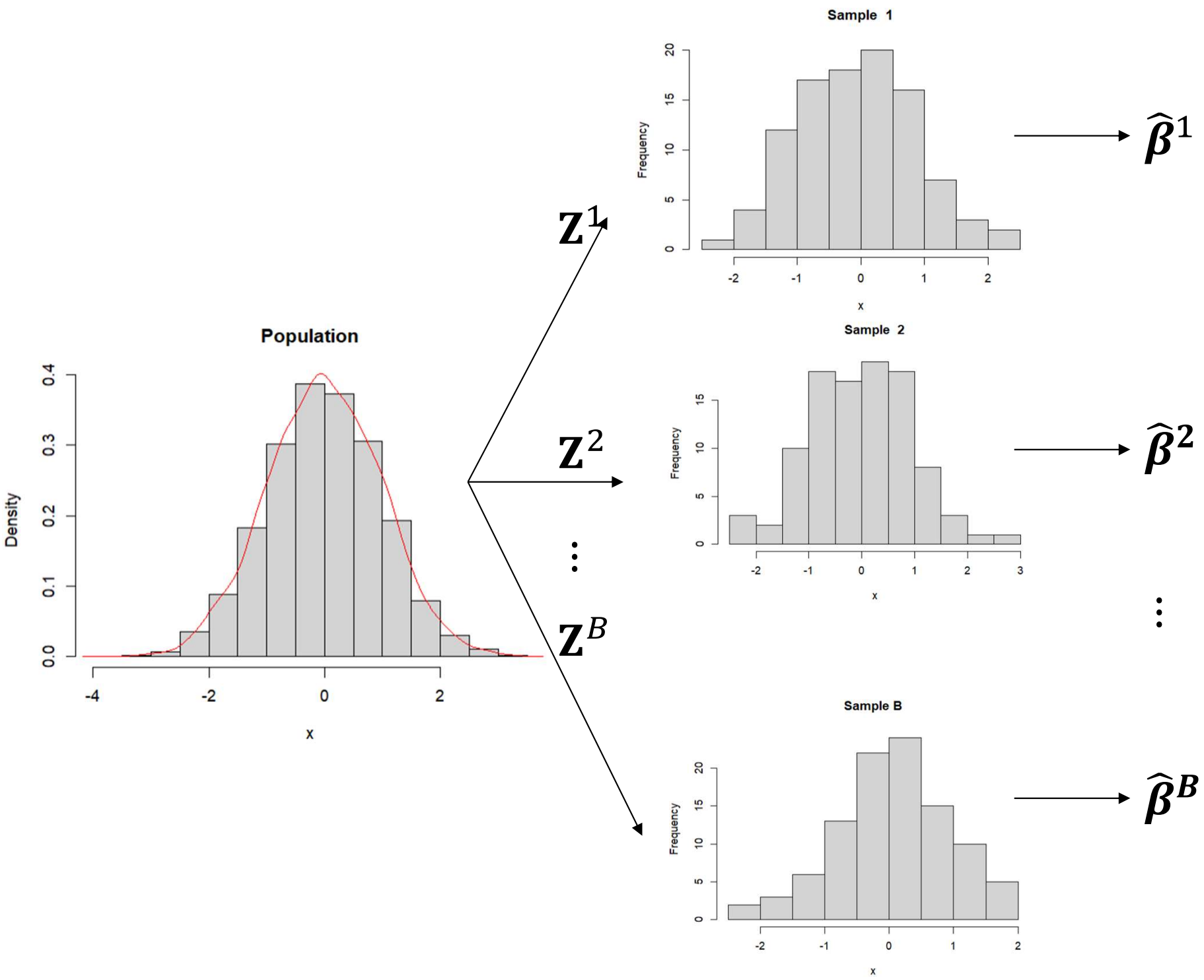
- To apply the *t* test for the slope, the following assumptions are required:
  - Normal distribution
  - Independent observations

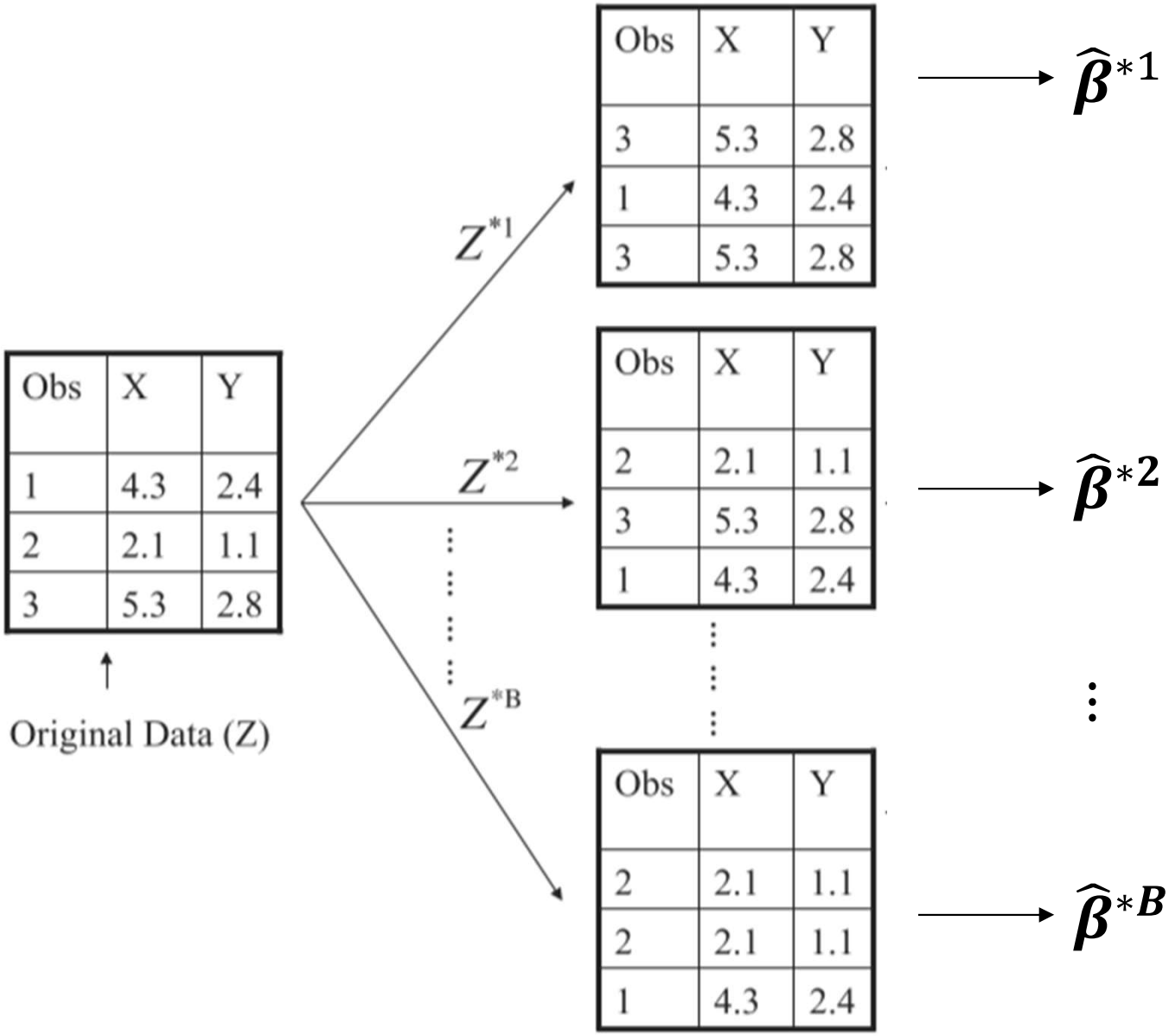


The Bootstrap

- The bootstrap allows us to use a computer to emulate the process of obtaining new samples so that we can estimate the variability of a parameter estimate without generating additional samples.

- Rather than repeatedly obtaining independent data sets from the population, the bootstrap obtains distinct datasets by repeatedly sampling observations from the original dataset with replacement.
  - Resampling with replacement


Population
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
x
$\mathbf{Z}^1$
$\mathbf{Z}^2$
$\mathbf{Z}^B$
Sample 1
Frequency
0
5
10
15
20
-2
-1
0
1
2
x
$\widehat{\boldsymbol{\beta}}^1$
Sample 2
Frequency
0
5
10
15
-2
-1
0
1
2
3
x
$\widehat{\boldsymbol{\beta}}^2$
Sample B
Frequency
0
5
10
15
20
-2
-1
0
1
2
x
$\widehat{\boldsymbol{\beta}}^B$



James et al. (2021, Figure 5.11, p. 212)