CSCI 556 Data Analysis & Visualization

Regression and Prediction

Instructor: Dr. Jinoh Kim

Topics

- Linear regression and least squares regression
- Multiple linear regression
- Assessment metrics: RMSE, RSS, R-squared
- Cross validation
- Confidence and prediction intervals

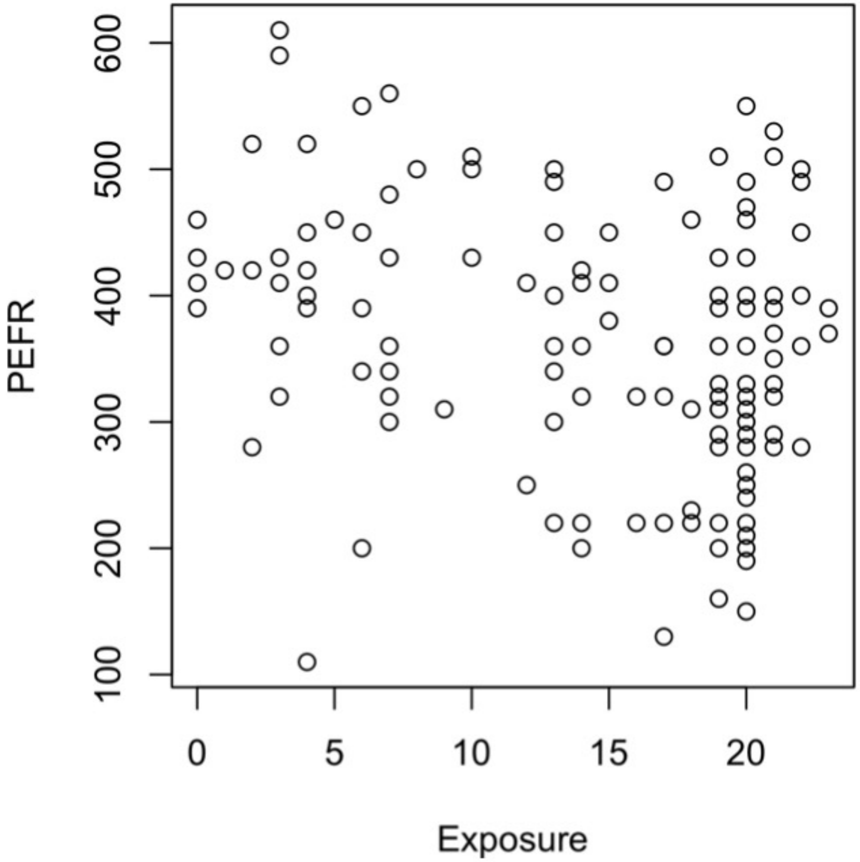
Regression and prediction

- Common goal in statistics:
  - Is the variable X associated with a variable Y?
  - If so, what is the relationship and can we use it to predict Y?
- Connection between statistics and data science:
  - Prediction of an outcome (target) variable based on the values of other "predictor" variables
- Another important connection is in the area of anomaly detection
  - Regression diagnostics originally intended for data analysis and improving the regression model can be used to detect unusual records

Simple linear regression

- Correlation: measures the strength of an association between two variables
- Regression: quantifies the nature of the relationship
- Simple linear regression: $Y = b_0 + b_1 X$
  - Y = response or dependent variable (or a target in ML)
  - X = predictor or independent variable (or a feature vector in ML)

Example:
how is PEFR related to Exposure?



$$\text{PEFR} = b_0 + b_1\,\text{Exposure}$$

Fitted values and residuals

- In general, the data does not fall exactly on a line, so the regression equation should include an explicit error term (e_i):

$$Y_i = b_0 + b_1 X_i + e_i$$

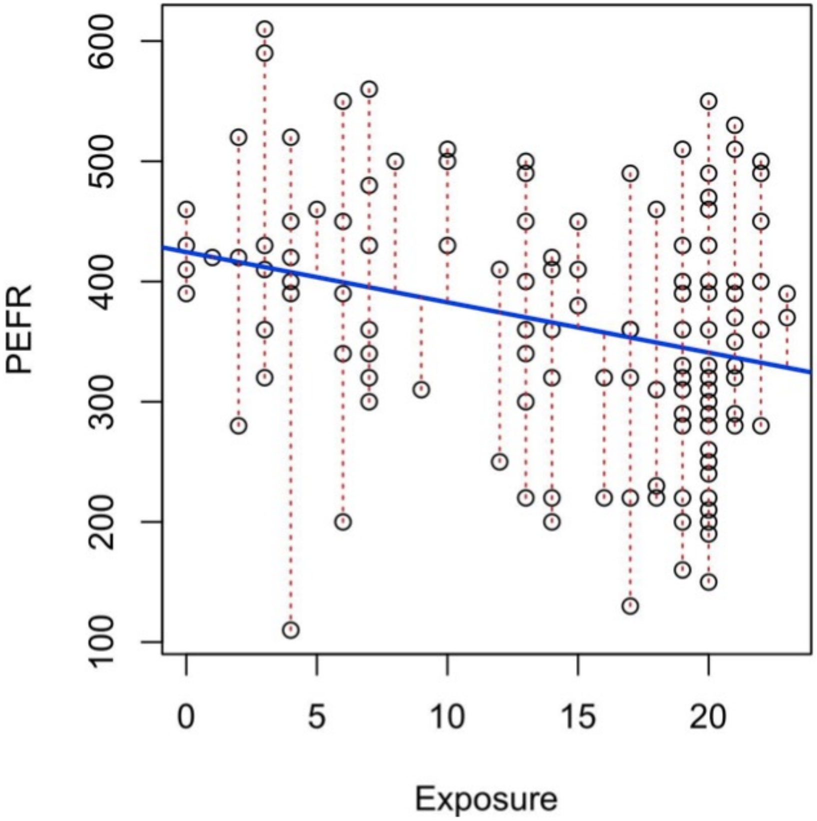
- Predicted values (fitted values):

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

- Residuals:

$$\hat{e}_i = Y_i - \hat{Y}_i$$

Example:
Residuals from a regression line



- Regression line: PEFR = -4.185*Exposure + 424.583

Least squares regression

- Minimizing the sum of squared residual values, also called *residual sum of squares* (*RSS*):

$$
\begin{aligned}
RSS &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\
&= \sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2
\end{aligned}
$$

- Computing coefficients:

$$
\begin{aligned}
\hat{b}_1 &= \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \\
\hat{b}_0 &= \overline{Y} - \hat{b}_1 \overline{X}
\end{aligned}
$$

Multiple linear regression

- For multiple predictors, the equation is simply extended to:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ \ldots + bpXp + e$$

- The relationship between each coefficient and its variable (feature) is linear (instead of a single line)

- The fitted values are given by:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1,i} + \hat{b}_2 X_{2,i} + \ \ldots + \hat{b}_p X_{p,i}$$

Assessing the model

- The most important performance metric from a data science perspective is *root mean squared error* (*RMSE*), defined as:

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

- A variant: *Residual standard error* (*RSE*), with a difference of denominator: (*n-p-1*) instead of *n*

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

Other metrics

- Another useful metric: *Coefficient of determination*, also called the *R-squared* statistic (R^2):

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

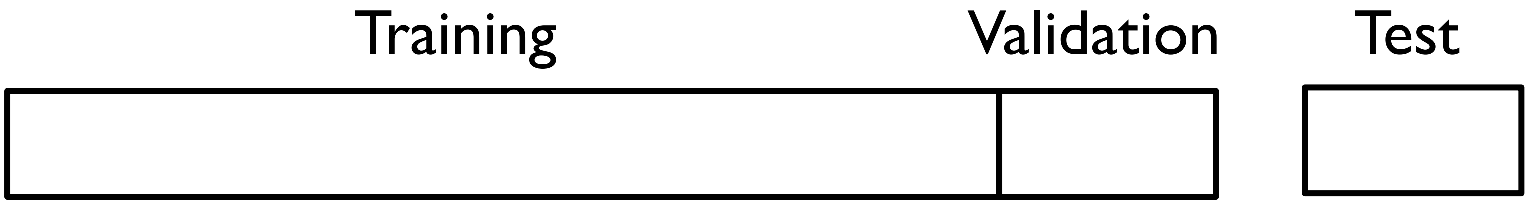
- R-squared ranges from 0 to 1 and measures the proportion of variation in the data that is accounted for in the model.
- Higher R-squared values represent smaller differences between the observed data and the fitted values

Training and testing

- Training set: a set of data instances used to fit the parameters
- Test set: independent instances that have played no part in formation of classifier
  - Assumption: both training data and test data are representative samples of the underlying problem
- Test and training data may differ in nature
  - Example: classifiers built using customer data from two different towns *A* and *B*
  - To estimate performance of classifier from town *A* in completely new town, test it on data from *B*

Note on parameter tuning

- It is important that the test data is not used *in any way* to create the classifier (model)
- Some learning schemes operate in two stages:
  - Stage 1: build the basic structure
  - Stage 2: optimize parameter settings
- The test data cannot be used for parameter tuning!
- Proper procedure uses *three* sets: training data, validation data, and test data
  - Validation data is used to optimize parameters

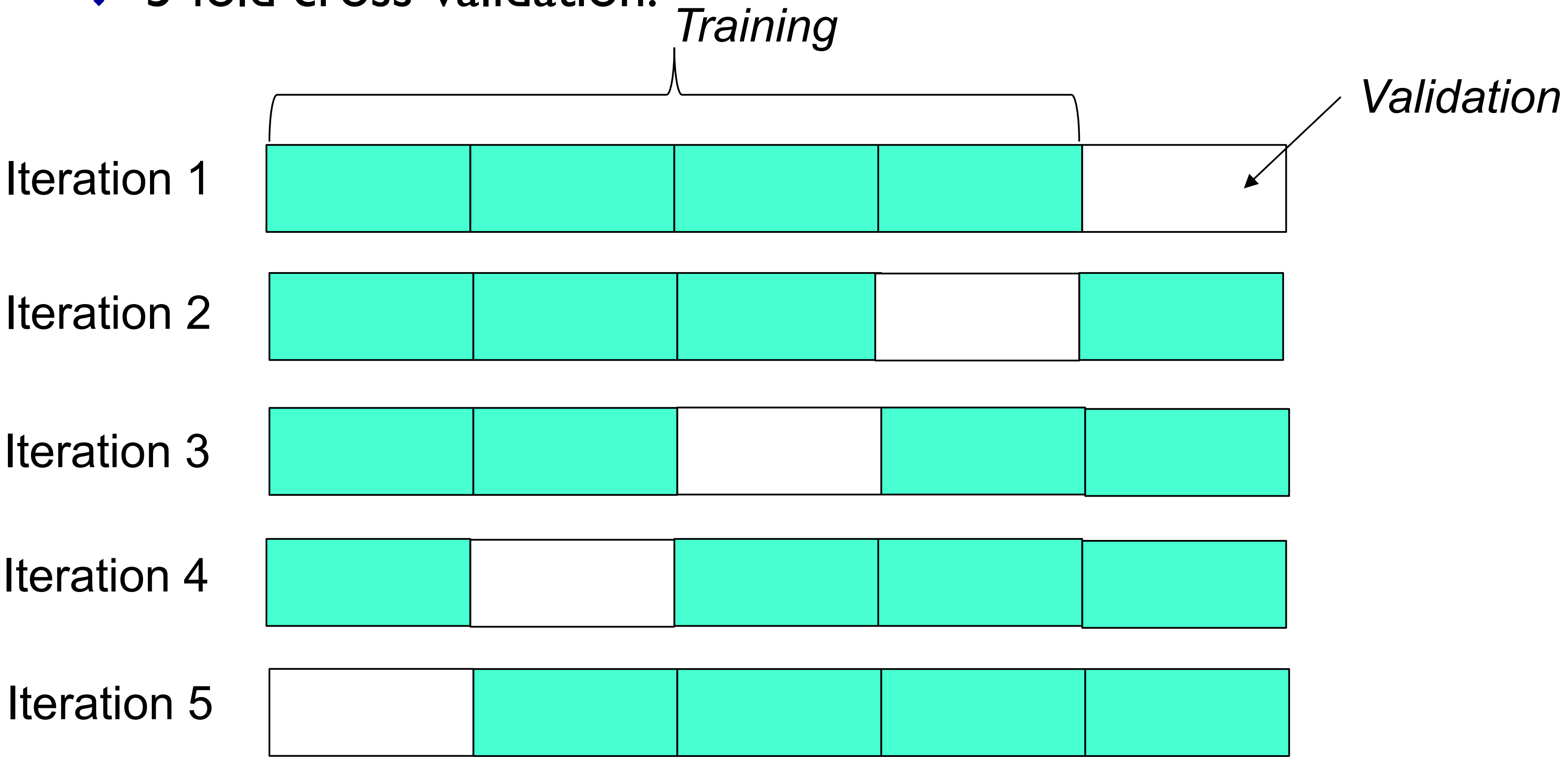


Cross validation

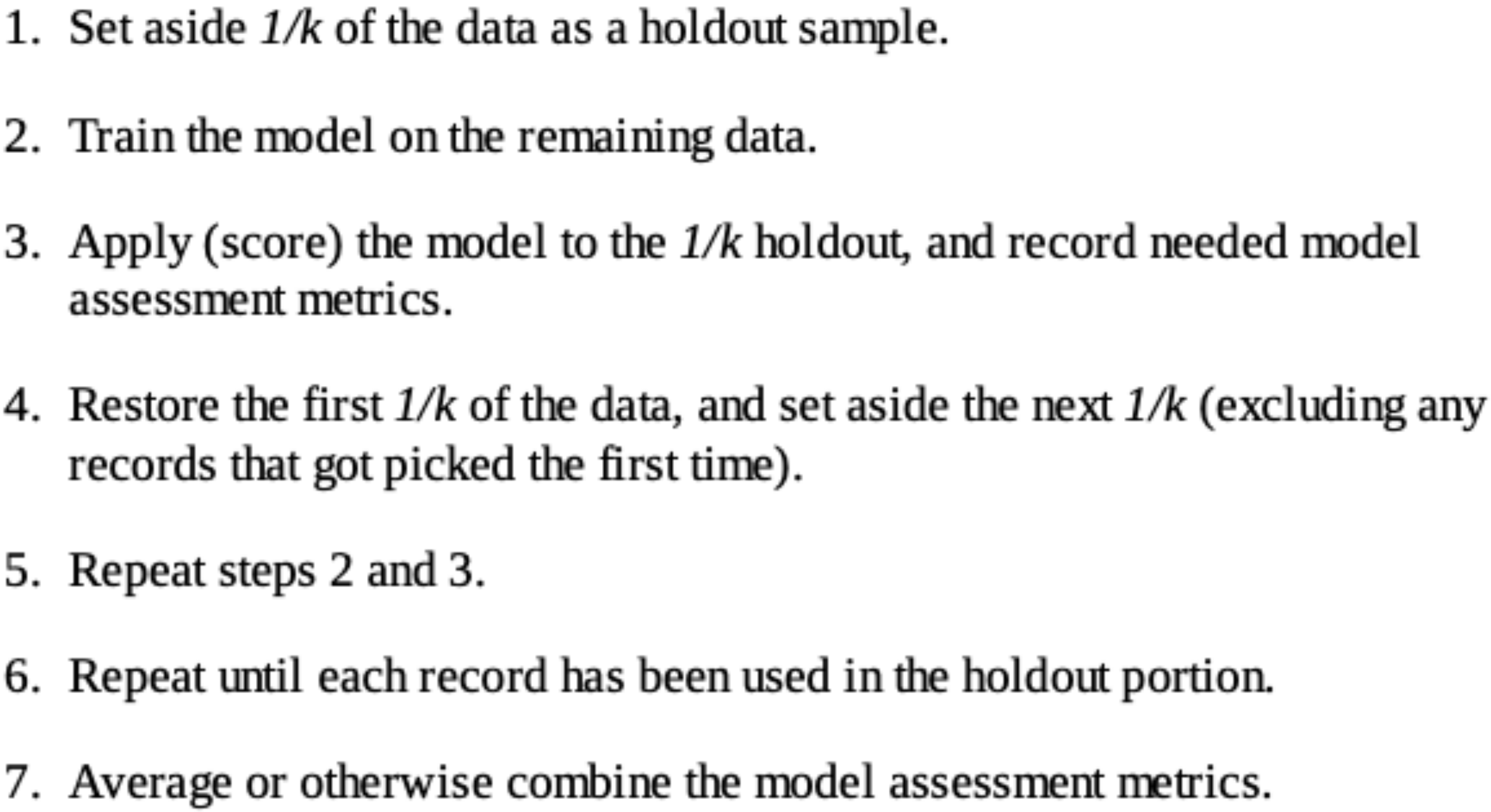
- In-sample metrics (e.g., R^2): applied to the same data that was used to fit the model
- Out-of-sample validation: use a majority of the data to fit the model, and use a smaller portion to test the model
- Hold-out sample: a random sample from a data set that is withheld and not used in the model fitting process
- How different would the assessment be if you selected a different holdout sample?

k-fold cross-validation

- Use multiple sequential holdout samples
- Fold: division of the data into the training sample and the holdout sample
- 5-fold cross-validation:



k-fold cross-validation (algorithm)



1. Set aside *1/k* of the data as a holdout sample.
2. Train the model on the remaining data.
3. Apply (score) the model to the *1/k* holdout, and record needed model assessment metrics.
4. Restore the first *1/k* of the data, and set aside the next *1/k* (excluding any records that got picked the first time).
5. Repeat steps 2 and 3.
6. Repeat until each record has been used in the holdout portion.
7. Average or otherwise combine the model assessment metrics.

More on cross-validation

- Standard method for evaluation: stratified ten-fold cross-validation
- Why ten?
  - Extensive experiments have shown that this is the best choice to get an accurate estimate
  - There is also some theoretical evidence for this
- Stratification reduces the estimate's variance
- Even better: repeated stratified cross-validation
  - E.g., ten-fold cross-validation is repeated ten times and results are averaged (reduces the variance)

Model selection

- In some problems, many variables could be used as predictors in a regression
- Adding more variables, however, does not necessarily mean we have a better model
- Occam's razor: all things being equal, a simpler model should be used in preference to a more complicated model
- Measures to determine:
  - AIC (Akaike's Information Criteria)
  - BIC (Bayesian Information Criteria)

AIC and stepwise regression

- AIC penalizes adding terms to a model using:

$$\text{AIC} = 2p + n\log(\frac{RSS}{n})$$

  - Here, p is the number of variables and n is the number of records
  - Find the model that minimizes AIC
- Stepwise regression: successively adds and drops predictors to find a model that lowers AIC (rather than searching through all possible models)
- BIC is similar to AIC with a stronger penalty for including additional variables to the model

Prediction using regression

- Primary purpose of regression in data science
- Regression models should not be used to extrapolate beyond the range of the data
  - Extrapolation: predicting hypothetical values that fall outside a particular data set

- Prediction vs. explanation
  - Explanation: understanding a relationship between predictor variables and an outcome variable
  - Prediction: predicting individual outcomes for new data, rather than explain data in hand

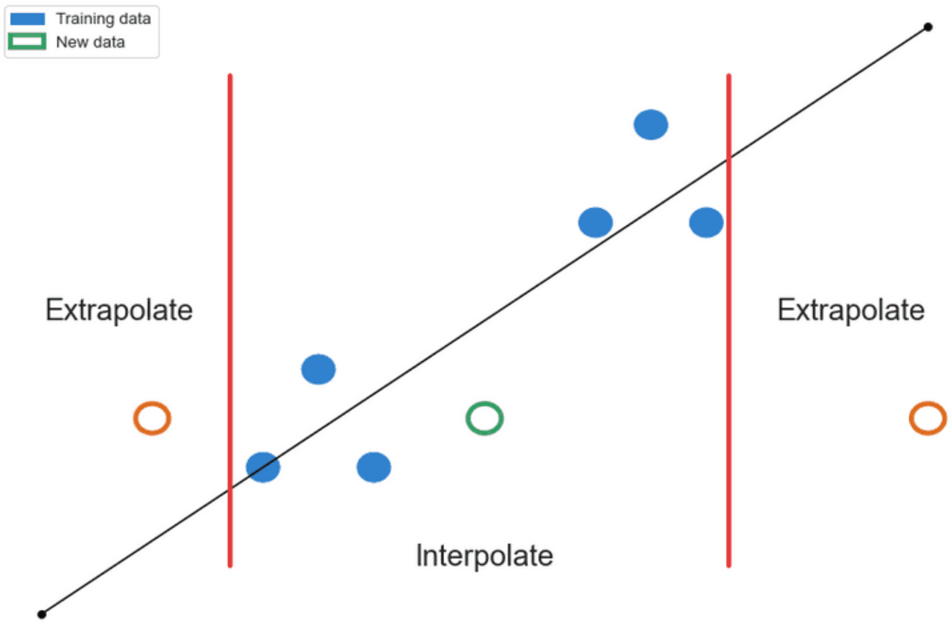
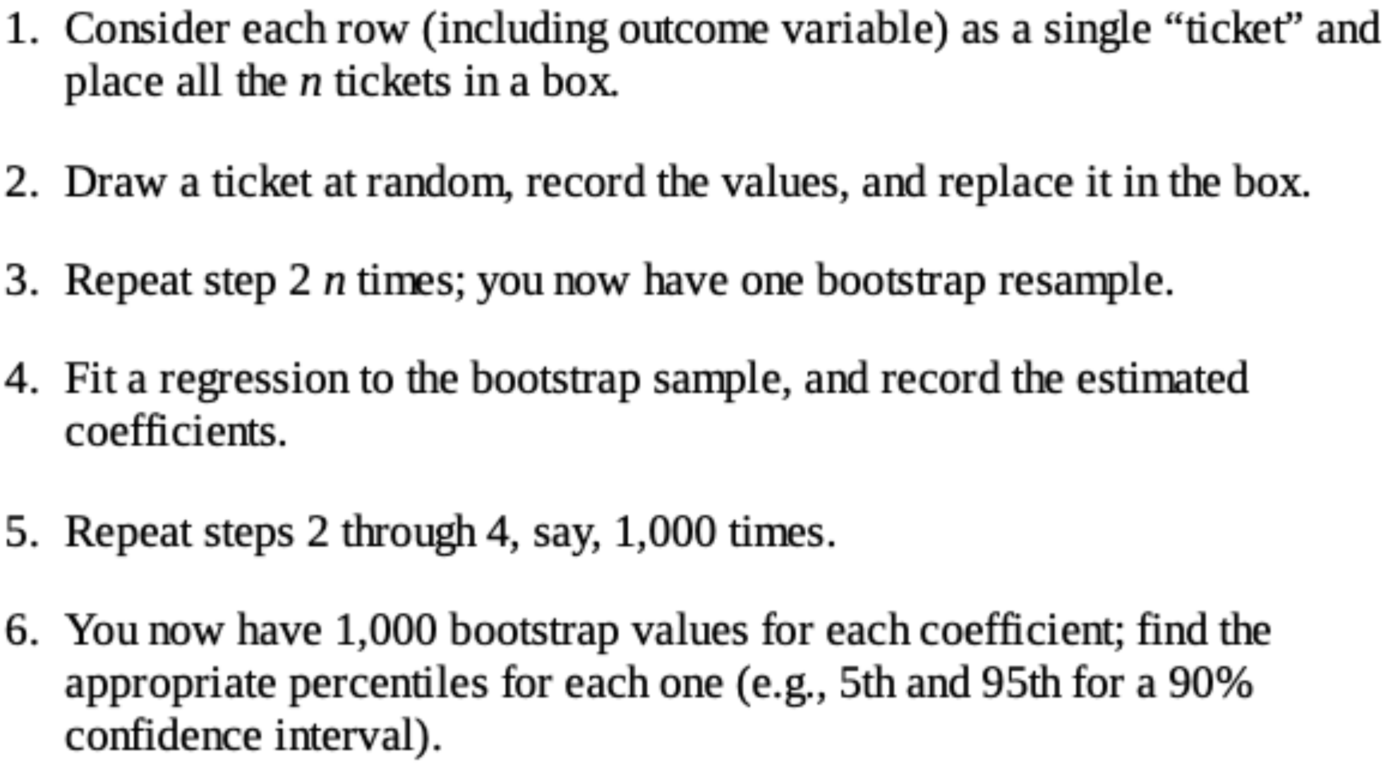


Image from mlexpert.io

Confidence intervals

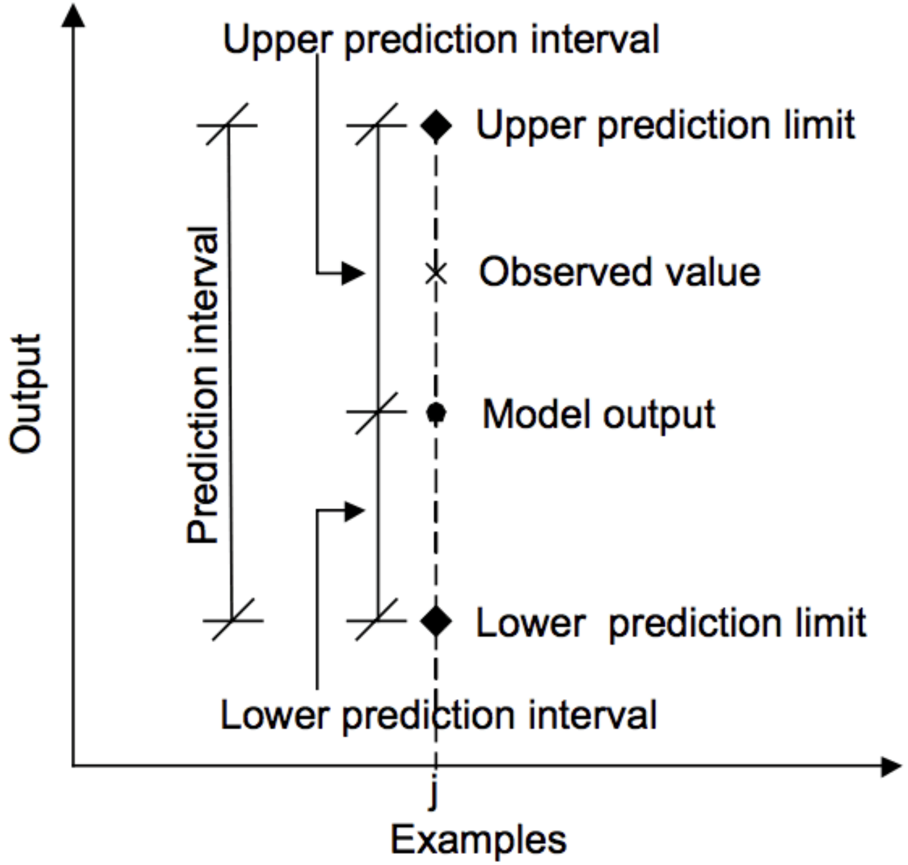
- How to calculate confidence intervals for regression parameters (coefficients) for a data set with *P* predictors and *n* records (rows)?
- Algorithm (based on bootstrapping):



1. Consider each row (including outcome variable) as a single “ticket” and place all the *n* tickets in a box.
2. Draw a ticket at random, record the values, and replace it in the box.
3. Repeat step 2 *n* times; you now have one bootstrap resample.
4. Fit a regression to the bootstrap sample, and record the estimated coefficients.
5. Repeat steps 2 through 4, say, 1,000 times.
6. You now have 1,000 bootstrap values for each coefficient; find the appropriate percentiles for each one (e.g., 5th and 95th for a 90% confidence interval).

Prediction intervals

- Prediction interval quantifies the uncertainty on a single observation estimated from the population
- Confidence interval quantifies the uncertainty on an estimated population variable (e.g., mean or standard deviation)



Relationship between prediction, actual value and prediction interval.
Taken from "Machine learning approaches for estimation of prediction interval for the model output", 2006.

Summary

- Linear regression and least squares regression
- Multiple linear regression
- Assessment metrics: RMSE, RSS, R-squared
- Cross validation
- Confidence and prediction intervals