Lab 6: Model specification

Not graded, just practice

Katie Schuler

2024-10-10

Practice your new *modeling* skills with these practice exam questions! Best to open a fresh Google Colab notebook and test things out! Refer to the study guide to find answers as well.

1 Types of models

(a) Which of the following best describes the goal of a regression model?

- (A) To classify observations into distinct categories
- (B) To predict continuous outcomes
- (C) To find the median of the dataset
- (D) To determine the probability of each class

(b) In classification tasks, the output variable (label) is typically:

- (A) Continuous
- (B) Discrete
- (C) Ordinal
- (D) A linear function

(c) Which of the following is an example of a regression problem?

- (A) Predicting whether an email is spam or not

- (B) Predicting the price of a house based on its features
- (C) Identifying the species of a flower
- (D) Grouping customers into clusters based on purchasing behavior

(d) What is the primary difference between regression and classification?

- (A) Predicting whether an email is spam or not
- (B) Predicting the price of a house based on its features
- (C) Identifying the species of a flower
- (D) Grouping customers into clusters based on purchasing behavior

(e) Which of the following tasks is a classification problem?

- (A) Estimating a person's height based on their age
- (B) Predicting if a student will pass or fail a course
- (C) Predicting the temperature next week
- (D) Estimating the number of sales for the next quarter

(f) True or false, supervised learning requires labeled data to train the model.

- (A) True
- (B) False

(g) True or false, in unsupervised learning, the model attempts to identify patterns or structures in data without any specific target variable.

- (A) True
- (B) False

2 Model specification

(a) Which of the following is the first step in model specification?

- (A) Fitting the model
- (B) Defining the response variable
- (C) Calculating residuals
- (D) Transforming variables

(b) What does model specification involve?

- (A) Estimating the parameters
- (B) Defining the functional form of the model
- (C) Calculating prediction accuracy
- (D) Testing the model's reliability

(c) Which of the following is NOT part of model specification?

- (A) Choosing which variables to include
- (B) Defining the relationship between predictors and response
- (C) Assessing the goodness-of-fit
- (D) Determining if interaction terms are necessary

(d) Which of the following describes a correctly specified model?

- (A) A model that includes irrelevant variables
- (B) A model that excludes important variables
- (C) A model that represents the true relationship between predictors and response
- (D) A model that overfits the training data

(e) True or false, Adding interaction terms between predictors is part of the model specification process.

- (A) True
- (B) False

(f) Model specification is the final step in the model-building process.

- (A) True
- (B) False

3 Functional form of linear models

4 Question

(a) Write the equation that expresses the response variable as a weighted sum of regressors (our favorite).

5 Answer

$y = \sum_{i=1}^{n} w_i x_i$

(b) In the linear regression equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$, what do the β's represent?

- (A) The predicted values
- (B) The error terms
- (C) The weights for each regressor
- (D) The intercept

6 Question

(c) Write the linear model equation in matrix notation.

7 Answer

$y = X\beta + \varepsilon$

or similar

(d) In matrix notation, what is $\mathbf{X}$?

- (A) A vector of error terms
- (B) A matrix of predictors (explanatory variables)
- (C) A vector of residuals
- (D) The coefficients of the model

(e) Suppose our `SwimRecords` data includes the year, sex, record time, swimsuit type, and swim cap type. Which of the following variables is most likely to be irrelevant for predicting swim times?

- (A) suit type
- (B) year
- (C) sex
- (D) swim cap type

(f) What is the potential issue of including too many irrelevant variables in your model?

- (A) It will improve model accuracy.
- (B) It can lead to overfitting and increased model complexity.
- (C) It will simplify the interpretation of results.
- (D) It has no effect on the model.

8 Finished?

Work together with support from TAs on **problem set 3, questions 5**:

Suppose your roommate is keeping a bunch of plants in your apartment. You notice that the plants exposed to more light seem to be taller, and — as an emerging data scientist — you record these data in a csv file: polynomial_plants.csv. Explore the relationship between light_exposure and plant_height across different plant species by plotting the data using an appropriate geom. Then, specify, fit and compare polynomial models of increasing degrees (linear, quadratic, and cubic) to the data. Start by specifying and fitting a simple linear model. Next, specify and fit second- and third-degree polynomial models, and visualize each using geom_smooth(). Which best captures the relationship between light_exposure and

plant_height? For each model, make sure you specify as a mathematical expression first in LaTex, then use `infer` to specify and fit the model.