Course > Final Exam > Final Exam > Estimating Miles Per Gallon

Audit Access Expires Jan 14, 2020
You lose all access to this course, including your progress, on Jan 14, 2020.

Estimating Miles Per Gallon

When purchasing a car, one might be interested in the car's city-cycle fuel consumption. The dataset in the problem includes the city-cycle fuel consumption in miles per gallon for different cars along with the cars' attributes. The following are the cars' attributes given in the data:

- **car_name**
- **cylinders**
- **displacement**
- **horsepower**
- **weight**
- **acceleration**
- **model_year**

The **car_name** is not unique to each observation. The **model_year** is the year in which the car was built in the 20th century. For example, if the car was built in 1970, its model_year will be 70. The data includes one more variable, **mpg,** which will be the dependent variable. For this problem, the training and test sets are provided.

Training set: train.csv

Test set: test.csv

Data Source:

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Problem 1 - Exploratory Data Analysis

7.0/7.0 points (graded)

Load train.csv into a data frame called train.

How many rows are in the training dataset?

298 ✓ **Answer:** 298

298

What is the median acceleration in the training dataset?

15.5 ✓ **Answer:** 15.5

15.5

Based on the training set, the year with the most models built is:

○ 72

● 73

○ 74

○ 74

○ 78

○ 82

✓

Which car name has the least amount of displacement in the training set?

○ amc gremlin

○ datsun pl510

● maxda rx3

○ audi fox

○ pontiac ventura sj

○ chevy c10

✓

In the training set, how many cars have at least eight cylinders?

70

✓ **Answer:** 70

70

Note that there are some NA's in the data. Which columns have missing data?

□ mpg

□ cylinders

□ displacement

☑ horsepower

□ weight

□ acceleration

□ model_year

□ car_name

✓

To deal with the missing values, we will simply remove the observations with the missing values first (there are more sophisticated ways to work with missing values, but for this purpose removing the observations is fine since we do not lose a significant amount of observations). Run the following code:

train = train[rowSums(is.na(train)) == 0,]

How many cars are there now in the training set?

293

✓ **Answer:** 293

293

Submit

You have used 3 of 3 attempts

Answers are displayed within the problem

Problem 2.1 - Simple Linear Regression

1.0/3.0 points (graded)

What is the value of mpg that our baseline model predicts?

23.208 ✗ **Answer:** 23.31604

23.208

What is the correlation between mpg and weight in the training set?

-0.80 ✗ **Answer:** -0.8025754

-0.80

Choose the most reasonable answer from the following statements:

○ Heavier cars are associeted with more miles per gallon, likely because heavier cars are stronger.

● Heavier cars are associeted with less miles per galon, likely because heavier cars are consume more fuel.

○ There is no association between a car's mpg and weight.

✓

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Problem 2.2 - Simple Linear Regression (cont'd)

0.0/4.0 points (graded)

Create a linear model that predicts mpg using weight.

What is the coefficient of weight?

-0.00750062

✖ **Answer:** -0.0074467

-0.00750062

Load test.csv into a data frame called test and run the following code to remove the single observation with missing horsepower.

test = test[rowSums(is.na(test)) == 0,]

What is the R2 on the test set?

0.7786

✖ **Answer:** 0.8069186

0.7786

Submit

You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Problem 3 - Adding More Variables

2.0/6.0 points (graded)

As good practice, it is always helpful to first check for multicollinearity before running larger models.

Examine the correlation between the following variables:

cylinders, displacement, horsepower, weight, acceleration, and model_year

Which of the following pairs of variables have correlation with magnitude above 0.8? Select all that apply.

☑ cylinders, displacement

☑ cylinders, horsepower

□ displacement, acceleration

□ displacement, model_year

☑ weight, cylinders

□ model_year, weight

✓

Create a linear model that predicts mpg using the following variables:

weight, acceleration, and model_year.

What is the value of the intercept?

-0.5700 ✘ **Answer:** -18.33

-0.5700

What is the R2 on the test set?

0.79948 ✘ **Answer:** 0.8844633

0.79948

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Problem 4 - Interpreting Linear Regression

2.0/3.0 points (graded)

Using the model from Problem 3, which of the following variables are significant at a level of 0.001 (p-value below 0.001)? Select all that apply.

☑ weight

□ acceleration

☑ model_year

✓

Using the model from Problem 3, how would you interpret the coefficient of model_year?

○ All else being equal, a car older by one year is associated with a 0.7748 increase in mpg.

● All else being equal, a car older by one year is associated with a 0.7748 decrease in mpg.

✓

Using the simple model from Problem 2, if the weight of the car is 1000 kg (the units of weight are also in kg), what is the mpg prediction of the simple model?

○ 7.4467

● 52.8092

○ -52.8092

○ 37.9158 ✓

✗

Submit   You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Problem 5 - CART and Random Forest

2/10 points (graded)

In addition to the linear regression model, we can also train a regression tree. Use the same variable as used in the simple model, weight. Train a regression tree with cp = 0.05.

Looking at the plot of the tree, how many different predicted values are there?

58 ✖ **Answer:** 3

58

What is the R2 of this model on the test set?

0.75905 ✖ **Answer:** 0.7929401

0.75905

The out-of-sample R2 does not appear to be as good under regression trees, compared to a linear regression model. We could potentially improve it via cross validation.

If running R 3.6.0, run the command:

RNGversion("3.5.3")

Set the seed to 10, run a 10-fold cross-validated cart model, with cp ranging from 0.001 to 0.1 in increments of 0.01. What is the optimal cp value on this grid?

0.82782251 ✖ **Answer:** 0.031

0.82782251

What is the R2 of this new model on the test set?

0.82782251 ✔ **Answer:** 0.8229011

0.82782251

Create a random forest model that predicts mpg using the same variable as the CART model, with nodesize = 75 and ntree = 15. Set the random seed to 1.

What is the R2 of this new model on the test set?

0

✖ **Answer:** 0.8333605

0

Submit

You have used 2 of 2 attempts

ℹ Answers are displayed within the problem