Statistical Learning Fall 2019
Homework 1

Name Student: Yizhen Dai

Student Number: S2395479

Introduction

In this homework, we will apply linear regression on the housing data set, which concerns housing values in the suburbs of Boston. We will consider three ways of improving prediction accuracy: feature selection, ridge regression and the lasso. We will split the data set into two parts (training and test set), then fit the models on the training set (including choice of the best parameter using cross-validation) and test their performance on the separate test set.

Introductory analyses

1. The data set is downloaded from the website loaded to Jupyter Notebook. It is then divided into two parts, with the first 350 examples as a training set and the rest (156 examples) as a test set.

2. After adding the dummy variable into training set as the intercept, we fit the least squares model to the training set.

   Figure 1 shows a boat.

Cross-validation and regularization

Argue that the expressions on the left hand side of the equality evaluate to the real number, ∞ or $-\infty$ on the right hand side.

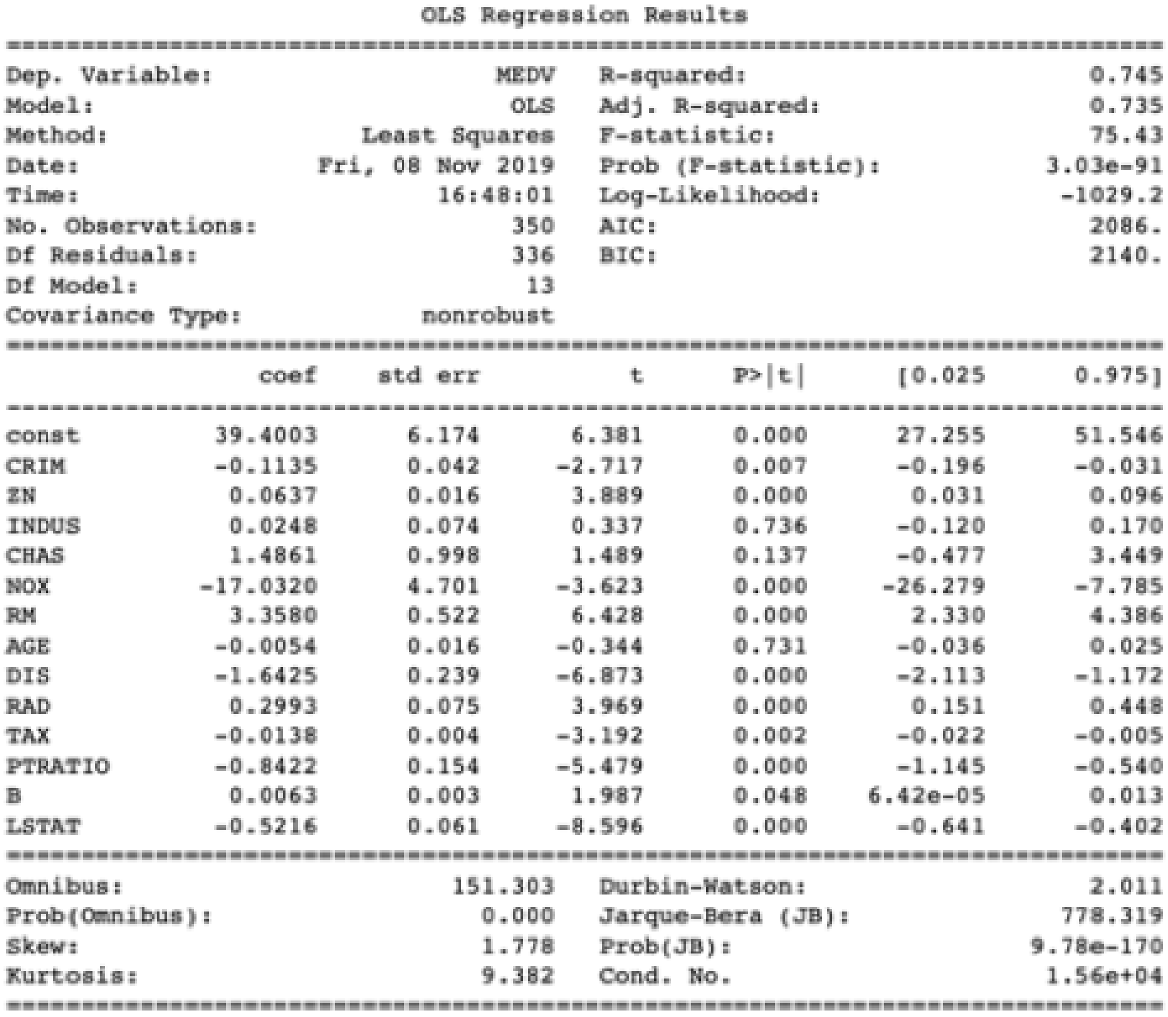
a) **2 points:** $\lim_{x \to -2} \frac{x^2-4}{x+2} = -4$.

b) **2 points:** $\lim_{x \to -2} \frac{1}{x+2} = \infty$.

c) **2 points:** $\lim_{x \to 0} \frac{1}{x+2} = \frac{1}{2}$.

d) **4 points:** $\lim_{x \to 0} \log_2(x) = -\infty$. **Hint:** Recall that the base 2 logarithm of a real number $x \in (0, \infty)$, denoted by $\log_2(x)$, is defined as the (unique) real number $y = \log_2(x)$ satisfying $2^y = x$.

Solution:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   MEDV   R-squared:                       0.745
Model:                            OLS   Adj. R-squared:                  0.735
Method:                 Least Squares   F-statistic:                     75.43
Date:                Fri, 08 Nov 2019   Prob (F-statistic):           3.03e-91
Time:                        16:48:01   Log-Likelihood:                -1029.2
No. Observations:                 350   AIC:                             2086.
Df Residuals:                     336   BIC:                             2140.
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         39.4003      6.174      6.381      0.000      27.255      51.546
CRIM          -0.1135      0.042     -2.717      0.007      -0.196      -0.031
ZN             0.0637      0.016      3.889      0.000       0.031       0.096
INDUS          0.0248      0.074      0.337      0.736      -0.120       0.170
CHAS           1.4861      0.998      1.489      0.137      -0.477       3.449
NOX          -17.0320      4.701     -3.623      0.000     -26.279      -7.785
RM             3.3580      0.522      6.428      0.000       2.330       4.386
AGE           -0.0054      0.016     -0.344      0.731      -0.036       0.025
DIS           -1.6425      0.239     -6.873      0.000      -2.113      -1.172
RAD            0.2993      0.075      3.969      0.000       0.151       0.448
TAX           -0.0138      0.004     -3.192      0.002      -0.022      -0.005
PTRATIO       -0.8422      0.154     -5.479      0.000      -1.145      -0.540
B              0.0063      0.003      1.987      0.048    6.42e-05       0.013
LSTAT         -0.5216      0.061     -8.596      0.000      -0.641      -0.402
==============================================================================
Omnibus:                      151.303   Durbin-Watson:                   2.011
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              778.319
Skew:                           1.778   Prob(JB):                    9.78e-170
Kurtosis:                       9.382   Cond. No.                     1.56e+04
==============================================================================
```

Figure 1: A boat.

a) First, we consider the behaviour of the denominator. As x approaches -2 from either the left or right side, $x + 2$ goes to 0. The numerator $x^2 - 4$ also approaches 0 as x goes to -2. Rewriting the expression yields that for any $x \in \mathbb{R} \setminus \{2\}$,

$$\frac{x^2 - 4}{x + 2} = \frac{(x - 2)(x + 2)}{x + 2} = x - 2,$$

which tends to -4 as x tends to -2 from either the left or the right side, which verifies what is asked of us.

b) As x approaches -2 from the left side, $x + 2$ approach 0 from the left side (negative) and $\frac{1}{x+2}$ approaches $-\infty$. As x approaches -2 from the right side, $x + 2$ approach 0 from the right side (positive) and $\frac{1}{x+2}$ approaches ∞. Therefore,

$$\lim_{x \searrow -2} \frac{1}{x + 2} = +\infty,$$
$$\lim_{x \nearrow -2} \frac{1}{x + 2} = -\infty$$

which verifies the two-sided limit does not exist.

c) The factor $x+2$ approaches 2 as x approaches 0. Therefore, $\frac{1}{x+2}$ tends to $\frac{1}{2}$ as x approaches 0 from either the left or the right side, which verifies what is asked of us.

d) Since $x \in (0, \infty)$, $\log_2(x)$ only has right-hand limit at $x = 0$. To meet $2^y = x$, y becomes larger and larger negative numbers as the values of x approaches 0 from the right. Give x any natural number $c \in (0, 1)$, we get $\log_2(c)$. There will always be another postive natural number $d < c$ gives a larger negative number $\log_2(d)$. Therefore, we can say $\log_2(x)$ has right-hand limit $-\infty$ at $x = 0$.

Exercise 3 (10 points)

Consider the functions $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 1 \text{ if } x \geq 2 \\ 0 \text{ if } x < 2. \end{cases}$$

and $g : \mathbb{R} \setminus \{2\} \to \mathbb{R}$ defined by

$$g(x) = \frac{x^2 + 2x}{x^2 - 4} \text{ for } x \in \mathbb{R} \setminus \{2\}.$$

Evaluate the specified (one or two sided) limits for the functions defined above. If their exist no limit, is the limit $-\infty$, ∞ or neither?

a) **2 points:** Evaluate the (the right-side) limit $\lim_{x \searrow 2} f(x)$. Give an argument based on the informal definition of a limit.

b) **2 points:** Evaluate the (the left-side) limit $\lim_{x \nearrow 2} f(x)$. Give an argument of why the based on the informal definition of a limit.

c) **2 points:** Evaluate the limit $\lim_{x \to 2} f(x)$. **Hint:** Consider your answer to part (a) and (b).

d) **2 points:** $\lim_{x\to 2} g(x)$.

e) **2 points:** $\lim_{x\to\infty} g(x)$.

Solution:

a) As x approaches 2 from the right side, $f(x)$ goes to 1. Therefore $\lim_{x\searrow 2} f(x) = 1$

b) As x approaches 2 from the left side, $f(x)$ goes to 0. Therefore $\lim_{x\nearrow 2} f(x) = 0$

c) Based on part (a) and (b), $f(x)$ goes to different values as x approaches 2 from either side. Since the one-sided limits are not equal even, the two-sided limit does not exist.

d) Rewrite the expression for $f(x)$ as follows:

$$g(x) = \frac{x^2+2x}{x^2-4} = \frac{x*(x+2)}{(x-2)*(x+2)} = \frac{x}{x-2} = \frac{1}{1-\frac{2}{x}} \text{ for } x \in \mathbb{R}\setminus\{-2,2\}.$$

The factor $1-\frac{2}{x}$ approaches 0 as x approaches 2. Therefore, $g(x)$ has right-hand ∞ and left-hand $-\infty$ at $x=0$. We can put it as:

$$\lim_{x\searrow 2} g(x) = +\infty,$$
$$\lim_{x\nearrow 2} g(x) = -\infty$$

Since the one-sided limits are not equal even, the two-sided limit does not exist.

e) As shown in part (d), $g(x) = \frac{1}{1-\frac{2}{x}}$ for $x \in \mathbb{R}\setminus\{-2,2\}$. The factor $1-\frac{2}{x}$ approaches 1 as x approaches ∞. Therefore, $\lim_{x\to\infty} g(x) = 1$.

Exercise 4 (10 points)

Consider the function $f : (0,\infty)\setminus\{4\} \to \mathbb{R}$ defined by the equation

$$f(x) = \frac{\sqrt{x}-2}{x^2-16} \text{ for } x \in (0,\infty)\setminus\{4\}.$$

a) **5 points:** Compute the limit $\lim_{x\to 4} f(x)$.

b) **5 points:** Consider now for some number $L \in \mathbb{R}$ the function $g : \mathbb{R} \to \mathbb{R}$ defined by

$$g(x) = \begin{cases} \frac{\sqrt{x}-2}{x^2-16} & \text{for } x \in (0,\infty)\setminus\{4\} \\ L & \text{for } x = 4. \end{cases}$$

For what value of L is g continuous on its entire domain?

Solution:

a) Rewriting the expression yields that for any $x \in (0, \infty) \setminus \{4\}$,

$$f(x) = \frac{\sqrt{x} - 2}{(x+4)(x-4)} = \frac{\sqrt{x} - 2}{(x+4)(\sqrt{x}+2)(\sqrt{x}-2)} = \frac{1}{(x+4)(\sqrt{x}+2)}$$

which tends to $\frac{1}{32}$ as x tends to 4 from either the left or the right side. Therefore,

$$\lim_{x \to 4} f(x) = \frac{1}{32}$$

b) We say that a function f is continuous at an interior point c of its domain if $\lim_{x \to c} f(x) = f(c)$. To make g continous on its entire domain, $\lim_{x \to 4} g(x) = g(4)$. Therefore,

$$L = g(4) = \lim_{x \to 4} g(x) = \frac{1}{32}$$

Exercise 5 (10 points)

We call a function $f : \mathbb{R} \to \mathbb{R}$ a *polynomial* if it satisfies

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \text{ for all } x \in \mathbb{R}$$

for some natural number n and $a_n, a_{n-1}, \ldots, a_0 \in \mathbb{R}$.

In this exercise we will show that any polynomial is a continuous function.

a) **3 points:** Argue that the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x$ is continuous at every point of its domain. **Hint:** see the examples in 1.5 and recall the definition of a continuous function.

b) **3 points:** Argue, using Theorem 2 of section 1.2 in the book, that $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is continuous. **Hint:** can we write f as the product of two continuous functions?

c) **2 points:** Argue in a similar way to part (b) that actually for any $n \in \mathbb{N}$ the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^n$ is continuous.

d) **2 points:** Now, using Theorem 2 twice more, conclude that any polynomial is a continuous function.

Solution:

a) For any number $c \in \mathbb{R}$, we want to show that there exists a $\delta > 0$ for each number $\epsilon > 0$ that for all $x \in \mathbb{N} \setminus \{c\}$ with $|x - c| < \delta$, it should hold that $|f(x) - f(c)| < \epsilon$. Set $\delta = \epsilon$, $|f(x) - f(c)| = |x - c| < \delta = \epsilon$. It verifies $\lim_{x \to c} f(x) = f(c)$.
We say that a function f is continuous at an interior point c of its domain if $\lim_{x \to c} f(x) = f(c)$. Therefore, $f(x) = x$ is continuous at every point of its domain.

b) Define $g : \mathbb{R} \to \mathbb{R}$ by $g(x) = x$. Based on part (a), $g(x) = x$ is continuous at every point $c \in \mathbb{R}$ of its domain and $\lim_{x \to c} g(x) = c$. Since $f(x) = x^2 = g(x) * g(x)$, based on Theorem 2.3(Limit of a product), $\lim_{x \to c} f(x) = \lim_{x \to c} g(x) * \lim_{x \to c} g(x) = c^2$. It verifies $\lim_{x \to c} f(x) = f(c)$.
We say that a function f is continuous at an interior point c of its domain if $\lim_{x \to c} f(x) = f(c)$.Therefore, $f(x) = x^2$ is continuous at every point of its domain.

c) Define $g : \mathbb{R} \to \mathbb{R}$ by $g(x) = x$. Based on part (a), $g(x) = x$ is continuous at every point $c \in \mathbb{R}$ of its domain and $\lim\limits_{x \to c} g(x) = c$. Since $f(x) = x^n = g(x)^n$, based on Theorem 2.6(Limit of a power), $\lim\limits_{x \to c} f(x) = \lim\limits_{x \to c} g(x)^n = c^n$ for any number $c \in \mathbb{R}$. It verifies $\lim\limits_{x \to c} f(x) = f(c)$.
We say that a function f is continuous at an interior point c of its domain if $\lim\limits_{x \to c} f(x) = f(c)$.Therefore, $f(x) = x^n$ is continuous at every point of its domain.

d) Based on part(3), $f(x) = x^n$ is continuous at every point of its domain and $\lim\limits_{x \to c} f(x) = c^n$. Set $g(x) = a * x^n$, $a \in \mathbb{R}$. Based on Theorem 2.4(Limit of a multiple), $\lim\limits_{x \to c} g(x) = a * c^n$ for any number $c \in \mathbb{R}$. It verifies $\lim\limits_{x \to c} g(x) = g(c)$.
We say that a function g is continuous at an interior point c of its domain if $\lim\limits_{x \to c} g(x) = g(c)$.Therefore, $g(x) = a * x^n$ is continuous at every point of its domain.

For $f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ for all $x \in \mathbb{R}$, based on Theorem 2.1(Limit of a sum), $\lim\limits_{x \to c} f(x) = a_n c^n + a_{n-1} c^{n-1} + \cdots + a_1 c + a_0$ for any number $c \in \mathbb{R}$. It verifies $\lim\limits_{x \to c} f(x) = f(c)$.
We say that a function f is continuous at an interior point c of its domain if $\lim\limits_{x \to c} f(x) = f(c)$. Therefore, $f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ is continuous at every point of its domain.

Bonus exercise (+5 points)

a) **+3 points:** Does the limit $\lim\limits_{x \to 0} \sin(\frac{1}{x})$ exist? If not, is it $-\infty$, ∞ or neither? Base your reasoning on the behaviour of the function on the interval $(0, \delta)$, for $\delta > 0$ a small (and shrinking) number.

b) **+2 points:** Prove using the formal definition of a limit that $\lim\limits_{x \to 0} [x \cdot \sin(\frac{1}{x})]$ exists. **Hint:** Argue that $|x \cdot \sin(\frac{1}{x})| \leq |x|$ based on the properties of the sine function. Then argue based on the formal definition as in Example 2 of section 1.5 in the book.

Solution:

a) To prove the limit $\lim\limits_{x \to 0} \sin(\frac{1}{x})$ does not exist, we need to show no $L \in \mathbb{R}$ exists as x goes to 0 that for each number ϵ there exists a number δ such that for all $x \in \mathbb{R}$ with $|x - 0| < \delta$ it holds that $|sin(\frac{1}{x}) - L| < \epsilon$.

As x approaches 0 from the right, $\frac{1}{x}$ approaches approaches ∞. Therefore, $\sin(\frac{1}{x})$ will oscillate between -1 and 1. Therefore. no L can hold $|sin(\frac{1}{x}) - L| < \epsilon$ for a small ϵ.

b) We want to show that as x goes to 0 that for each number ϵ there exists a number δ such that for all $x \in \mathbb{N} \setminus \{0\}$ with $|x - 0| < \delta$ it holds that $|x \cdot sin(\frac{1}{x}) - 0| < \epsilon$.

Since $0 \leq |sin(\frac{1}{x})| \leq 1$, we can get $|x \cdot \sin(\frac{1}{x})| = |x| \cdot |\sin(\frac{1}{x})| \leq |x|$.

Set $\delta = \epsilon$, $|x \cdot sin(\frac{1}{x}) - 0| < |x| < \delta = \epsilon$. Hence it proves what is asked of us.