Machine Learning: Problem Set 2

Jingpeng Hong

February 6, 2022

Question 1

(a) Consider the true model:

$$\beta_0 = 0$$
$$\beta_1 = 1$$
$$\beta_2 = 2$$
$$f_1(x_{i,1}) = -x_{i,1}^2 - x_{i,1}$$
$$f_2(x_{i,2}) = x_{i,2}^2 - x_{i,2}$$

Hence, we have $\beta_2 > \beta_1$. We simulate x_1 and x_2 from the standard normal distribution, and the results from the estimation of true model and linear model are shown in the Column (1) and (2) of Table (1), respectively.

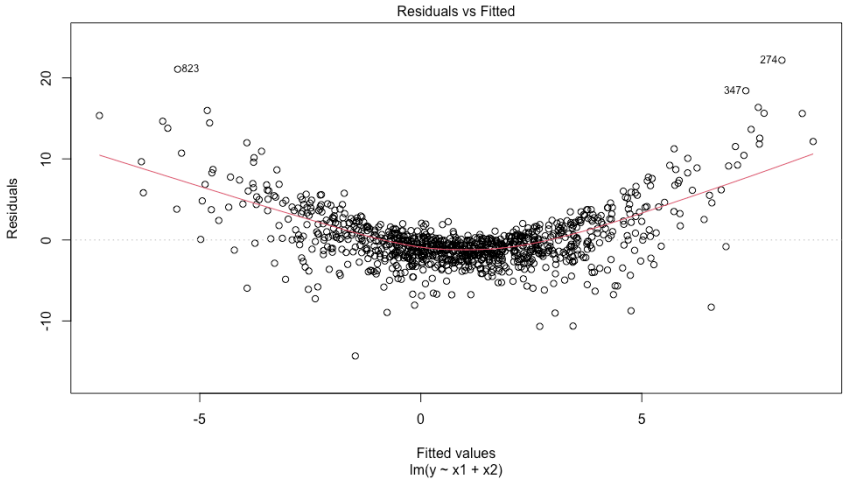
We can find that $\hat{\beta_2} > \hat{\beta_1}$ (1.986>1.001), while $\hat{b_2} < \hat{b_1}$ (-2.275<-0.900)

Table 1: OLS results from true model and linear model

	Dependent variable:	
	y	
	(1)	(2)
$f_1(x_1)$	1.001***	
	(0.018)	
$f_2(x_2)$	1.986***	
	(0.016)	
x_1		−0.900***
		(0.108)
x_2		−2.275***
		(0.113)
Constant	−0.016	0.844***
	(0.040)	(0.112)
Observations	1,000	1,000
R^2	0.949	0.321
Adjusted R^2	0.949	0.320
Note:	*p<0.1; **p<0.05; ***p<0.01	

(b) Consider the relationship between the residuals and fitted values. Ideally, if the linear model is the true model, the residual plot will show no fitted pattern and the red line should be approximately horizontal at zero. However, in Figure (1), we find that residuals and fitted values have a non-linear pattern, which may indicate a problem with the linear model.

Figure 1: Residuals vs Fitted for regression diagnostics



Question 2

Interpretation:

a) As shown in the Column (1) of Table (2), the constant is 50.887, which can be interpreted as the average reading test results of male (gender=0). The coefficient of gender indicates the difference between the female (gender=1) and the male (gender=0), i.e., female has a significantly (at 1% level) higher average reading test results than male, by 1.711.

b) As shown in the Column (2) of Table (2), the coefficient of male represents the average reading test result of male, 50.887; the coefficient of female represents the average reading test result of female, 52.598.

The difference between Column (1) and Column (2) is that the first column indicates the gender gap, while the second column shows the mean of male and female separately.

c) As shown in the Column (3) of Table (2), the sample can be divided into 4 groups:

- no high school diploma male, with the average 44.285
- high school diploma male, with the average 44.285+6.858
- no high school diploma female, with the average 44.285+0.279
- high school diploma female, with the average 44.285+6.858+0.279+1.505

The intuitive interpretation is that the effect of acquiring a high school diploma depends on the individual's gender. If the individual is male, the effect of acquiring a high school diploma is 6.858; but if the individual is female, the

effect is 6.858+1.505. The coefficient of the interaction term Female × High School is the difference in the effect of acquiring a college degree for women versus men.

d) As shown in the Column (4) of Table 2, the coefficient of ln(Family Income) is the effect of family income on reading test results for male. For male, 1% increase in family income will lead to reading test results increased by 0.02815; the effect of family income on reading test results for female is the sum of coefficients of ln(Family Income) and ln(Family Income) × Female. For female, 1% increase in family income will lead to reading test results increased by 0.02815+0.00534. The coefficient of ln(Family Income) × Female is the difference in the effect of a 1% increase in family income for female versus male.

Bonus Question:

e) If factor1 has m levels and factor2 has n levels, then there are $m \times n$ groups needed to be identified. Given we have the constant term represent the baseline group, there remains $m \times n - 1$ to be identified. Hence, 1 interaction term has to be set to zero and the remaining $m \times n - 1$ ones are identified.

Table 2: OLS Regression

	Dependent variable:			
	Reading Test Results			
	(1)	(2)	(3)	(4)
Male		50.887***		
		(0.156)		
Female	1.711***	52.598***	0.279	−3.605
	(0.218)	(0.151)	(0.902)	(2.766)
High School			6.858***	
			(0.693)	
Female × High School			1.505	
			(0.935)	
ln(Family Income)				2.815***
				(0.200)
Female × ln(Family Income)				0.534**
				(0.266)
Constant	50.887***		44.285***	21.643***
	(0.156)		(0.669)	(2.079)
Observations	7,430	7,430	5,970	7,430
R^2	0.008	0.968	0.051	0.078
Adjusted R^2	0.008	0.968	0.051	0.077
Note:				*p<0.1; **p<0.05; ***p<0.01

Question 3

ISL Chapter 5 Exercise 2

We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

(a) What is the probability that the first bootstrap observation is not the jth observation from the original sample? Justify your answer.

For the first bootstrap observation, we randomly select an observation from the data set. The probability that the jth observation from the original sample is the first bootstrap observation is $\frac{1}{n}$. Hence, the probability that the first bootstrap observation is not the jth observation from the original sample is $1 - \frac{1}{n}$.

(b) What is the probability that the second bootstrap observation is not the jth observation from the original sample?

Since in the bootstrap approach, the sampling is performed with replacement. The probability that the second bootstrap observation is not the jth observation from the original sample will be the same as the probability in (a), which is $1 - \frac{1}{n}$.

(c) Argue that the probability that the jth observation is not in the bootstrap sample is $(1\text{-}\frac{1}{n})^n$.

For each selection, the probability that the jth observation is not selected is $1 - \frac{1}{n}$. For n times selections with replacement, the probability of not being selected will be the same. Hence, the probability that the jth observation is not in the bootstrap sample (the jth observation is not selected for n times) is $(1\text{-}\frac{1}{n})^n$

(d) When $n = 5$, what is the probability that the jth observation is in the bootstrap sample?

The probability that the jth observation is in the bootstrap sample will be

$$p = 1 - (1\text{-}\frac{1}{n})^n = 1 - (1\text{-}\frac{1}{5})^5 = 0.67232$$

(e) When $n = 100$, what is the probability that the jth observation is in the bootstrap sample?

The probability that the jth observation is in the bootstrap sample will be

$$p = 1 - (1\text{-}\frac{1}{n})^n = 1 - (1\text{-}\frac{1}{100})^100 = 0.6339677$$

(f) When $n = 10,000$, what is the probability that the jth observation is in the bootstrap sample?

The probability that the jth observation is in the bootstrap sample will be

$$p = 1 - (1\text{-}\frac{1}{n})^n = 1 - (1\text{-}\frac{1}{10000})^10000 = 0.632139$$

(g) Create a plot that displays, for each integer value of n from 1 to 100,000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

As shown in Figure (2), the probability that the jth observation is in the bootstrap sample will converge to a constant when n is large enough, because

$$p = \lim_{n\to\infty} 1 - (1\text{-}\frac{1}{n})^n = 1 - \frac{1}{e} = 0.6321$$

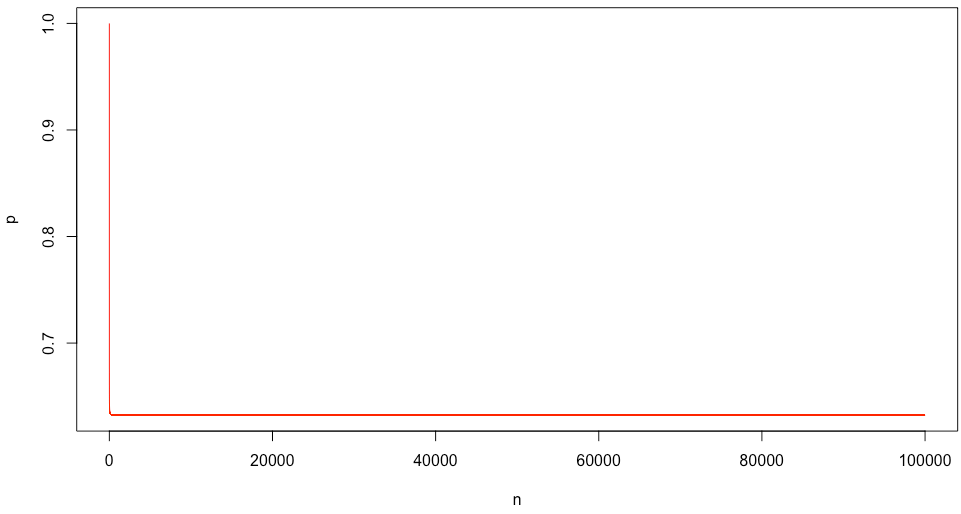


Figure 2: The probability that the jth observation is in the bootstrap sample

(h) We will now investigate numerically the probability that a bootstrap sample of size $n = 100$ contains the jth observation. Here j = 4. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

From the result in R code, we find that mean(store) is 0.6314, which is similar to the result we calculated above.

ISL Chapter 5 Exercise 3

We now review k-fold cross-validation.

(a) Explain how k-fold cross-validation is implemented.

We randomly divide the observation sets into k folds, of approximately equal size n/k. We treat the first fold as the test set, and use the remaining $k-1$ folds as training set to fit the model. Then, we compute the MSE_1 based on the observations in the first folder, i.e., the held-out fold. We repeat this procedure for k times, and each time treat a different fold as validation set with estimates of test error, $\text{MSE}_1, \text{MSE}_2, ..., \text{MSE}_k$, then the cross-validation estimate is calulated by averaging the MSE:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \text{MSE}_i$$

(b) What are the advantages and disadvantages of k-fold cross- validation relative to:

i. The validation set approach?

Advantages:

- The k-fold cross-validation approach will have less bias compared with the validation set approach. The training set in the validation set approach is typically around half the size of the original data set, while in k-fold CV, we repeatedly fit the model using training sets with $\frac{k-1}{k}n$ observations. Therefore, from the perspective of bias reduction, the k-fold CV is to be preferred to the validation set approach.
- In contrast to the validation approach which will yield different results when applied repeatedly due to randomness in the training/validation set splits, performing k-fold cross-validation approach multiple times will always yield similar results: there is less randomness in the training/validation set splits.

Disadvantages:

- In terms of the computational complexity, the validation set approach will only fit the learning procedure once, while k-fold cross-validation approach will perform k times. s. This computational disadvantage will be more significant when n is very large, or when the statistical learning methods have computationally intensive fitting procedures.

ii. LOOCV?

Advantages:

- The k-fold cross- validation approach has the computational advantage. The k-fold cross- validation approach requires fitting the learning procedure for k times, while LOOCV requires n times. Hence, when n is very large, or when the statistical learning methods have computationally intensive fitting procedures, k-fold cross-validation approach will be more feasible in terms of the computational burden.
- In terms of the variance, the k-fold cross- validation approach will have smaller variance than the LOOCV. For LOOCV, since each training set has an almost identical set of observations, our MSE will be highly correlated; In contrast, since the overlap between the training sets in each model is smaller in k-fold cross- validation approach, our MSE will be somewhat less correlated with each other. Since the mean of many highly correlated quantities has higher variance than does the mean of many quantities that are not as highly correlated, the test error estimate resulting from LOOCV tends to have higher variance than does the test error estimate resulting from k-fold CV.

Disadvantages:

- For each training set, the k-fold cross-validation approach contains $\frac{k-1}{k}n$ observations, while LOOCV approach contains $n-1$ observations. The LOOCV approach has a larger training set to fit the statistical learning model. Hence, compared with LOOCV, the k-fold cross-validation approach will have a larger bias for the estiamtes of the test error.

ISL Chapter 5 Exercise 8

(a) In this data set, what is n and what is p? Write out the model used to generate the data in equation form.

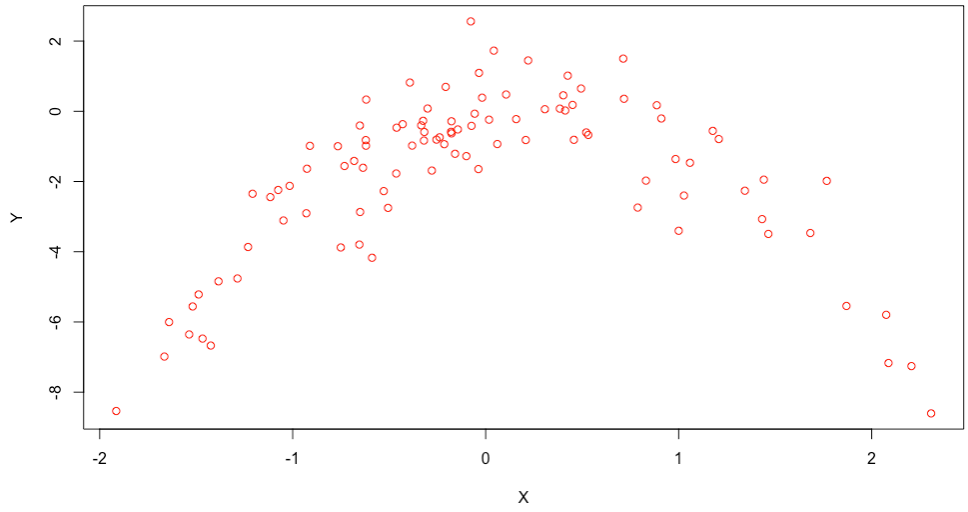
In this data set, $n = 100$ and $p = 2$. The model can be written as:

$$Y = X - X^2 + \epsilon \tag{1}$$

(b) Create a scatterplot of X against Y. Comment on what you find.

We can see that X and Y roughly follow the equation $y = x - x^2$, while x has a standard normal distribution.

Figure 3: Scatter Plot of X against Y



(c) The LOOCV errors that result from fitting the following four models using least squares are listed below:

Table 3: Cross-validation error and p

p	1	2	3	4
Cross-validation error	5.891	1.087	1.103	1.115

(d) We change random seed from 1 to 100. The LOOCV errors that result from fitting the following four models using least squares are listed below:

Table 4: Cross-validation error and p with different random seed

p	1	2	3	4
Cross-validation error	5.891	1.087	1.103	1.115

The results with other random seeds are the same as what we got in (c). Since each time we choose one observation as validation set and the remaining $n-1$ as training set, there is no randomness in the training/validation set splits. In other words, for all training/validation set splits, we have the same n validation sets with one observation in each of them. Therefore, we will have the same estimation of the test error.

(e) Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.

The model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ had the smallest LOOCV error. This is the result we expected because our simulated data was generated from the function $Y = X - X^2 + \epsilon$. We should expect that our estimated model has the same flexibility as the function used for simulation.

(f) As shown in table (5), these results agree with the conclusions drawn based on the cross-validation results. We can see that the coefficients of x and x^2 is statistically significant, while the coefficients of x^3 and x^4 are not. Hence, a cubic or higher-order fit will not provide better fit, i.e., there is not much benefit in including cubic or higher-order polynomial terms in the model.

Table 5: OLS coefficient estimates from fitting models in (c)

	Dependent variable:			
	y			
	(1)	(2)	(3)	(4)
x	2.316	2.316**	2.316**	2.316**
	(2.362)	(1.032)	(1.037)	(1.041)
x^2		−21.059***	−21.059***	−21.059***
		(1.032)	(1.037)	(1.041)
x^3			−0.305	−0.305
			(1.037)	(1.041)
x^4				−0.493
				(1.041)
Constant	−1.828***	−1.828***	−1.828***	−1.828***
	(0.236)	(0.103)	(0.104)	(0.104)
Observations	100	100	100	100
Note:				*p<0.1; **p<0.05; ***p<0.01

ISL Chapter 5 Exercise 9

We will now consider the Boston housing data set, from the MASS library.

(a) Based on this data set, provide an estimate for the population mean of medv.

$$\hat{\mu} = 22.53281$$

(b) Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result.

$$SE(\hat{\mu}) = \frac{\text{Std.Dev}(\hat{\mu})}{\sqrt{n}} = \frac{9.197104}{\sqrt{506}} = 0.4088611$$

The standard error of $\hat{\mu}$ is the estimate of variance of sample means in the sampling distribution. Hence, we have the estimate of the variance of the sample mean is 0.4088611. This result assesses how far $\hat{\mu}$ is likely to fall from the population mean of medv.

(c) Now estimate the standard error of $\hat{\mu}$ using the bootstrap.

The bootstrap estimate for SE($\hat{\mu}$) is 0.4106622, which is quite similar to our answers from (b).

(d) The 95% confidence interval for the mean of medv is

$$[\hat{\mu}\text{-}2\text{SE}(\hat{\mu}), \hat{\mu} + 2\text{SE}(\hat{\mu})] = [22.533 - 2 \times 0.411, 22.533 + 2 \times 0.411] = [21.711, 23.354]$$

The results obtained using t.test(Boston$medv) is [21.730, 23.336], which is also quite close to our Bootstrap results.

(e) Based on this data set, provide an estimate, $\hat{\mu}_{med}$, for the median value of medv in the population.

We use the median the sample as our estimate for the median value of medv in the population, $\hat{\mu}_{med} = 21.2$

(f)Using the bootstrap, we estimate that SE($\hat{\mu}_{med}$) is 0.3778075. We can find that the standard error of ($\hat{\mu}_{med}$) is a little smaller than the standard error of ($\hat{\mu}$), which indicates that the median can be a more stable estimator than the mean.

(g) Based on this data set, provide an estimate, $\hat{\mu}_{0.1}$, for the tenth percentile of medv in the population.

We use the tenth percentile the sample as our estimate for the tenth percentile of medv in the population, $\hat{\mu}_{0.1} = 12.75$

(h)Using the bootstrap, we estimate that SE($\hat{\mu}_{0.1}$) is 0.4767526. We can find that the standard error of $\hat{\mu}_{0.1}$ is a little larger than the standard error of $\hat{\mu}$ and $\hat{\mu}_{med}$, which indicates that the tenth percentile can be a more unstable estimator than the mean and the median. The tenth percentile is more likely to be influenced by the extreme values or outliers in the tail.