STATS 202: Data Mining and Analysis
Homework #2

Adam Kainikara

July 17, 2023

Problem 1 (5 points)

Chapter 4, Exercise 1 (p. 189).

4.2 $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$

4.3 $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$

Starting with 4.3

$p(X) = (1-p(X))e^{\beta_0+\beta_1 X}$

$p(X) = e^{\beta_0+\beta_1 X} - e^{\beta_0+\beta_1 X}p(X)$

$p(X) + e^{\beta_0+\beta_1 X}p(X) = e^{\beta_0+\beta_1 X}$

$p(X)(1+e^{\beta_0+\beta_1 X}) = e^{\beta_0+\beta_1 X}$

$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$

This is 4.2

Therefore 4.2 is equivalent to 4.3

Problem 2 (5 points)

Chapter 4, Exercise 4 (p. 189).

a) Because we are using 10% of observations in the range of X. This is the same as [100x +5)%. To find the average, we will have to integrate it.

$\int_{0.05}^{0.95} 10x dx + \int_{0}^{0.05} 10x + 5\, dx + \int_{0.95}^{1} 105 - 100x\, dx = 9.75$

b) X1 and X2 are independent. Under the same assumptions the fraction of available observations would be $9.75\%^2 = \sim 0.95\%$

c) This is similar to (b) just with 100 times. $9.75\%^{100} = \sim 0\%$

d) As the p increases, the fraction of amiable observations tends to 0.

e) p=100 $l = 0.1^{0.01}$

Problem 3 (5 points)

Chapter 4, Exercise 6 (p. 191).

a) $\hat{\beta_0} = -6\hat{\beta_1} = 0.05\hat{\beta_2} = 1$ are the coefficients for this question. These are coefficients for the constant, number of hours studied and undergrad GPA. Plugging in the values for this particular student gives $-6 + (0.05 \times 40) + (1 \times 3.5) = -0.5$. This value can be plugged into the logistic formula to calculate the probability.

$\hat{p}(x) = \frac{e^x}{1+e^x}$

$\hat{p}(x) = \frac{e^{-0.5}}{1+e^{-0.5}} = 0.378$

A student who studied for 40 hours and has an undergrad GPA of 3.5 has about a 37.8% chance of getting an A in the class.

b) How long does the student need to study for a 50% chance of an A.

Want $\hat{p}(x) = 0.5$

$0.5 = \frac{e^x}{1+e^x}$

$\frac{1}{2e^x} = \frac{1}{1+e^x}$

$2e^x = 1 + e^x$

$x = 0$

$0 = -6 + 0.05X_1 + 1X$

Because its the same student and only the hours studied is changing X_2(the GPA) is still the same

$0 = -6 + 0.05X_1 + 1(3.5)$

$X_1 = 50$

If this student wants a 50% of an A then they should study for 50 hours.

Problem 4 (5 points)

Chapter 4, Exercise 8 (p. 191).

Using KNN with K = 1, the training error rate can be calculated to be 0%. This means that the test error rate has to be 36% in order for the average to be 18%. Because the logistic regression fit had a test error rate of 20%, we should use the logistic fit because it

has a lower test error rate.

Problem 5 (5 points)

Chapter 4, Exercise 13 parts a-h (p. 193)

a) See bar graph

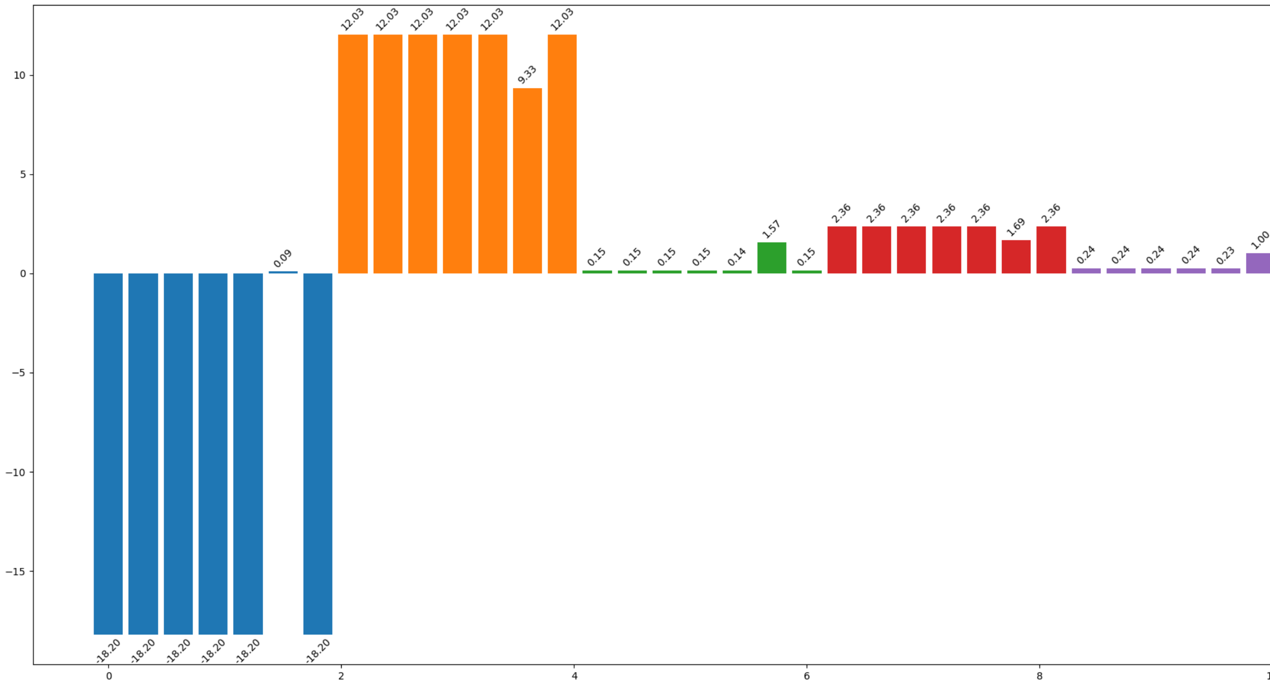

Figure 1: Min, Max, Mean Stand Dev and Median in the Colors Blue, Orange, Green, Red and Purple Respectively

The bar graph shows the min, max, mean stand dev and median in the colors blue, orange, green, red and purple respectively. These values are for Lag 1, 2, 3, 4, 5, Volume and Today. The bar graph shows similar results for each parameter. Only volume seems to change. The numerical summary for this data is below.

min: [-18.195 -18.195 -18.195 -18.195 -18.195 0.087465 -18.195]

max: [12.026 12.026 12.026 12.026 12.026 9.328214 12.026]

mean: [0.15058494 0.15107897 0.14720478 0.14581818 0.13989256 1.57461763 0.14989899]

stand dev: [2.35593008 2.35617168 2.35941794 2.35919491 2.36020027 1.68586184 2.35584499]

median: [0.241 0.241 0.241 0.238 0.234 1.00268 0.241]

b) The coefficients for the logistic regression are below.

[[5.37367327e-02 5.34098644e-03 -1.50782005e-02 5.70986281e-02 9.18621394e-02 7.21885742e+00]]

The logistic regression was performed with direction as the response and the five lag variables plus volume as predictors.

Used stats models API to obtain p values. Lag 2 had a p value of 0.0296 which is lower than 0.05 so the null hypothesis can be rejected. Lag 2 does not influence direction.

c) I first did logistic regression on the whole data set. Then I computed the confusion matrix. The confusion matrix resulted the following:

[[482 2]

[0 605]]

The model was very successful in predicting whether the data would go up or down. The model correctly predicted down every single time. When predicting up, there were only two instances the model predicted wrong of the 607 times. Of all 1089 data points, just two where predicted wrong.

d) Training data is 1990 to 2008 which is until row 986.

I fitted a logistic regression using the training data. I then computed the confusion matrix for the left out data (test data) which was the data from 2009 and 2010. The confusion matrix resulted the following:

Confusion Matrix of the test data with Lag2 as the only predictor

[[9 34]

[5 56]]

Of the 104 data points from 2009 and 2010, 39 (37.5%) where incorrect. 37.5% is the test error rate. In this time period there were 61 ups and 43 downs. Of the 61 ups, 5 were incorrect (8.2%). Of the 43 downs, 34 where incorrect (79.1%).

e, f, g, h, i) Doing d again but with LDA, QDA, KNN.

Confusion Matrix of the training data with Lag2 as the only predictor but instead used LDA

[[22 419]

[20 524]]

Confusion Matrix of the test data with Lag2 as the only predictor but instead used LDA

[[9 34]

[5 56]]

Confusion Matrix of the training data with Lag2 as the only predictor but instead used QDA

[[0 441]

[0 544]]

Confusion Matrix of the test data with Lag2 as the only predictor but instead used QDA

[[0 43]

[0 61]]

Confusion Matrix of the training data with Lag2 as the only predictor but instead used NAIVE BAYES

[[0 441]

[0 544]]

Confusion Matrix of the test data with Lag2 as the only predictor but instead used NAIVE BAYES

[[0 43]

[0 61]]

Confusion Matrix of the training data with Lag2 as the only predictor but instead used KNN

[[298 143]

[106 438]]

Confusion Matrix of the test data with Lag2 as the only predictor but instead used KNN

[[16 27]

[19 42]]

Error Rates (test data)	Lag 2 Logi	LDA	QDA	N
Test Error Rate	37.5%	37.5%	41.3%	
Incorrectly Predicted Down	79.1% (34/43 wrong)	79.1% (34/43 wrong)	100% (43/43 wrong)	100
Incorrectly Predicted Up	8.2 % (5/61 wrong)	8.2 % (5/61 wrong)	0% (0/61 wrong)	0%

Lag 2 Logi only had the lowest test error rate. KNN was the best at predicting down. QDA and NAIVE BAYES was best at picking up.

j) I did some transformations. These where using Lag 2 and 3 and changing KNN to K = 7.

Confusion Matrix of the training data with Lag2 and Lag3 as the only predictors

[[23 418]

[23 521]]

Confusion Matrix of the test data with Lag2 and Lag 3 as the only predictor

[[8 35]

[4 57]]

Confusion Matrix of the training data with Lag2 and Lag3 as the only predictors but instead used LDA

[[22 419]

[22 522]]

Confusion Matrix of the test data with Lag2 and Lag3 as the only predictors but instead used LDA

[[8 35]

[4 57]]

Confusion Matrix of the training data with Lag2 and Lag 3 as the only predictors but instead used QDA

[[12 429]

[13 531]]

Confusion Matrix of the test data with Lag2 and Lag3 as the only predictors but instead used QDA

[[4 39]

[2 59]]

Confusion Matrix of the training data with Lag2 as the only predictor but instead used NAIVE BAYES

[[0 441]

[0 544]]

Confusion Matrix of the test data with Lag2 as the only predictor but instead used NAIVE BAYES

[[0 43]

[0 61]]

Confusion Matrix of the training data with Lag2 and Lag 3 as the only predictor but instead used KNN

[[254 187]

[120 424]]

Confusion Matrix of the test data with Lag2 and Lag 3 as the only predictor but instead used KNN [

[11 32]

[17 44]]

Error Rates (test data)	Lag 2 and 3 Logi	LDA	QDA	N
Test Error Rate	37.5	37.5	39.4	
Incorrectly Predicted Down	81.4% (35/43 wrong)	81.4% (35/43 wrong)	90.7% (39/43 wrong)	100
Incorrectly Predicted Up	6.3 % (4/61 wrong)	6.3 % (4/61 wrong)	3.3% (2/61 wrong)	0

Problem 6 (5 points)

Chapter 5, Exercise 2 (p. 219).

a) Let n be the number of observations. The probability that the jth observation is in the bootstrap is $\frac{1}{n}$so the probability the jth observation is not in the probability is $1 - \frac{1}{n}$.

b) Each bootstrap is independent. So the probability that the jth observation is not in the second is the same: $1 - \frac{1}{n}$.

c) Bootstrapping has sample with replacement and is independent. The probability that the jth observation is not in a observation is $1 - \frac{1}{n}$. The probability that the jth observation is not in the bootstrap sample is the product of this. So it becomes $(1 - \frac{1}{n}) \times (1 - \frac{1}{n}) \times \ldots = (1 - \frac{1}{n})^n$

d) n =5, finding if the jth observation is in the bootstrap.

P(observation in the bootstrap) = 1 - P(observation is not in the bootstrap)

$1 - (1 - \frac{1}{5})^5 = 0.67$

e) Same as above but now n = 100

$1 - (1 - \frac{1}{100})^{100} = 0.63$

f) Same as above but now n = 10000

$1 - (1 - \frac{1}{10000})^{10000} = 0.63$

g) Note: For the graph I did n = 20,000 because when I did n = 100,000 my computer crashed and above 30,000 it was taking a very long time for the graph to load.

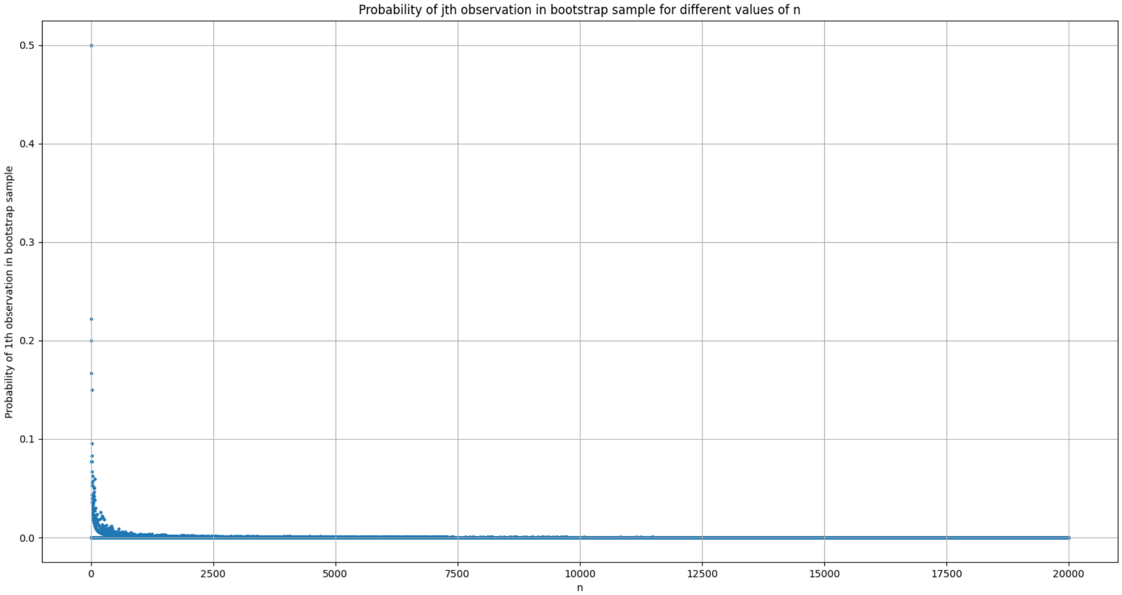

Figure 2: Probability of Jth Observation

The graph asymptotes around 0.63

h) As $n- > \infty, p = 0.632$

Problem 7 (5 points)

Chapter 5, Exercise 5 (p. 220).

a) The coefficients for income and balance fit using logistic regression to predict default are:

[[5.64710797e-03,

2.08091984e-05]]

b) 1) Did split into test and validation set. The data was split with 5000 observations in each set.

2) Fitted a multi logistic model. Coefficients are:

[[0.00041108

-0.00012325]]

3) Classified using posterior probability

4) Computed the validation set error. The confusion matrix was:

[[4996 1]

[3 0]]

The error for the validation set is 0.08%

c) 1) Did split into test and validation set. The data was split with 7500 observations in the training set and 2500 observations in the validation set.

2) Fitted a multi logistic model.Coefficients are:

```
[[5.79458398e-03
2.30839912e-05]]
```

3) Classified using posterior probability

4) Computed the validation set error. The confusion matrix was:

```
[[7275 107]
[ 112 6]]
```

The error for the validation set is 2.2%

5) Did split into test and validation set. The data was split with 2500 observations in the training set and 7500 observations in the validation set.

6) Fitted a multi logistic model. Coefficients are:

```
[[ 0.00043135
-0.00012144]]
```

7) Classified using posterior probability

8) Computed the validation set error. The confusion matrix was:

```
[[2499 0]
[ 1 0]]
```

The error for the validation set is 0.04%

9) Did split into test and validation set. The data was split with 6000 observations in the training set and 4000 observations in the validation set.

10) Fitted a multi logistic model. Coefficients are:

```
[[5.95395704e-03
3.26622332e-05]]
```

11) Classified using posterior probability

12) Computed the validation set error. The confusion matrix was:

```
[[3872 65]
[ 61 2]]
```

The error for the validation set is 3.2%

d) I used a list comprehension to switch the “yes” and “no” to 1 and 0. Then added it to the array. Then computed the logistic regression and made a confusion matrix of the validation set. The results are below:

```
[[4996 1]
[ 3 0]]
```

The error for the validation set is 0.08%

Adding a dummy variable, student, did not result in an increase or decrease in the error rate.

Problem 8 (5 points)

Chapter 5, Exercise 6 (p. 221).

a) The standard error for the coefficients balance, income and intercept are as follows.

$Name$	$Standard\,Error\,for\,the\,coefficients$
$balance$	0
$income$	$4.99e-06$
$intercept$	0.435

b) See code

c,d) The standard error was 3.804649256670676.

Problem 9 (5 points)

Chapter 5, Exercise 8 (p. 222).

a) In this problem n = 100 and p = 2. The model in this problem is $Y = X - 2X^2 + constant$

b) See scatter plot. The scatter plot has a upside down parabolic shape. Most of the points occur at the top of the curve.

c, d ,e)

I first fitted these 4 models using linear algebra and coding it.

i $\quad Y = \beta_0 + \beta_1 X + \epsilon$

ii $\quad Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$

iii $\quad Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

iv $\quad Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$

The seed was set to 1:

$rng = np.random.default_rng(1)$

The resulting coefficients for the models are as follows:

	β_0	β_1	β_2	β_3	β_4
i	-1.46496301	1.94936857			
ii	-0.07275529	0.96627276	-2.00470902		
iii	-0.05719669	1.1145842	-2.04709357	-0.06430033	
iv	0.10084766	0.90499786	-2.50592308	0.03376837	0.10421699

I then used LOOCV to help fit the models. The coefficients for each of the models that produced the lowest MSE are as follows:

	β_0	β_1	β_2	β_3	β_4	$Error$
i	-1.46518099	1.94913361				5.9226536056081205
ii	-0.0726598	0.96615516	-2.00472843			0.9834354527808048
iii	-0.05737607	1.11461298	-2.04701671	-0.06429704		0.9718403920130584
iv	0.10079418	0.90506572	-2.50587159	0.03374964	0.10420779	0.9201948751940691

Changed the seed to 100.

$rng = np.random.default_rng(100)$

The resulting coefficients for the models are as follows:

	β_0	β_1	β_2	β_3	β_4
i	-1.69029473	0.50898235			
ii	0.13096828	0.77172222	-1.94030595		
iii	0.15961224	0.46259444	0.46259444	0.13689711	
iv	0.07938098	0.41879764	-1.75905017	0.16392833	-0.06110937

I then used LOOCV to help fit the models. The coefficients for each of the models that produced the lowest MSE are as follows:

	β_0	β_1	β_2	β_3	β_4	$Error$
i	-1.69066004	0.50920579				5.790791169159457
ii	0.13115837	0.7717761	-1.94038458			1.0288896723158567
iii	0.15971326	0.46265244	-1.98187348	0.13688155		0.994999233705984
iv	0.07952053	0.41882023	-1.75921296	0.16391697	-0.0610765	0.9864312854873168

In both seeds model $iv : Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$ had the lowest LOOCV. I had mostly expected it as I thought the error would decrease but *iv* would over fit because we knew what the true function was. I had expected that *ii* would have the lower error because it has the same polynomial type as the true function but *iv* does have a lower error it just over fits the data.

f) The coefficients for $\beta_0, \beta_1, \beta_2$ all have p values less than 0.05. This means that these coefficients significantly help predict y. This is what should happen because the original function is a quadratic.

Problem 10 (5 points)

Chapter 5, Exercise 9 (p. 223).

a) The population mean: $\hat{\mu} = 22.53$

b) The standard error of the population mean 0.408. On average the mean of the population will be off from the population mean by 0.408.

c) Using bootstrap I got a standard error of 0.4018559

d) Con Int: [21.752012851895014, 23.31359979632634

e) $\hat{\mu_{med}} = 21.2$

f) Standard Error of Median 0.36652890745478695

g,h) Tenth Percentile (µ̂0.1) of medv: 12.75 with an error of 0.49. This shows how much the 10 percentile median would be off by on average.