CMDA 4654: Intermed Data Analytics & ML
Homework 2
Christopher J. Mobley
Due: Monday, 21MAR16

1. (4.7 – 1)Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and logit representation for the logistic regression model are equivalent.

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

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

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

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

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

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

2. (4.7 – 2)It was stated in the text that classifying an observation to the class for which (4.12) is largest is equivalent to classifying an observation to the class for which (4.13) is largest. Prove that this is the case. In other words, under the assumption that the observations in the kth class are drawn from a $N(\mu_k, \sigma^2)$ distribution, the Bayes' classifier assigns an observation to the class for which the discriminant function is maximized.

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-u_k)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-u_l)^2}} \tag{4.12}$$

Given x, the denominator of $p_k(x)$ becomes constant for all k. Thus maximizing the function does not depend on the denominator.

$$p_k(x) = \pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-u_k)^2}$$

Given $\frac{1}{\sqrt{2\pi}\sigma}$ is a constant for all k, maximizing the function does not depend on it.

$$p_k(x) = \pi_k e^{-\frac{1}{2\sigma^2}(x-u_k)^2}$$

Taking the logarithm of the function does not change the maximum k value since a logarithm is a monotonically increasing function.

$$p_k(x) = \log(\pi_k) - \frac{1}{2\sigma^2}(x - u_k)^2$$

$$p_k(x) = \log(\pi_k) - \frac{1}{2\sigma^2}(x^2 - 2u_k x + {u_k}^2)$$

$$p_k(x) = \log(\pi_k) - \frac{1}{2\sigma^2}x^2 + \frac{2u_k x}{2\sigma^2} - \frac{{u_k}^2}{2\sigma^2})$$

Given x, $\frac{1}{2\sigma^2}x^2$ becomes constant for all k. Thus maximizing the function does not depend on it.

$$p_k(x) = \log(\pi_k) + \frac{2u_k x}{2\sigma^2} - \frac{{u_k}^2}{2\sigma^2})$$

After a little rearranging equation 4.12 become 4.13, which prove that they are equivalent when assigning a classification via maximization.

$$p_k(x) = x\frac{u_k}{\sigma^2} - \frac{{u_k}^2}{2\sigma^2} + \log(\pi_k) = \delta_k(x) \qquad (4.13)$$

3. (4.7 – 6)Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\beta_0 = -6$, $\beta_1 = 0.05$, $\beta_2 = 1$.
    a. Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

$$p(X) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2}}$$
$$p(Get\ A) = \frac{e^{-6+0.05*40+1*3.5}}{1+e^{-6+0.05*40+1*3.5}}$$
$$p(Get\ A) = .377541$$

    b. How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?

$$p(X) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2}}$$
$$0.5 = \frac{e^{-6+0.05*x+1*3.5}}{1+e^{-6+0.05*x+1*3.5}}$$

Used TI-89 to solve for x

$$x = 50\ hrs$$

So, the student in part a would have to study 50 hours or more in order to have a 50% chance of getting an A in the class.

4. (4.7 – 10)This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.
    a. Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

Table 1: Summary of the Weekly Dataset

Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
Min. :1990	Min. :-18.1950	Min. :-18.1950	Min. :-18.1950	Min. :-18.1950	Min. :-18.1950	Min. :0.08747	Min. :-18.1950	Down:484
1st Qu.:1995	1st Qu.: -1.1540	1st Qu.: -1.1540	1st Qu.: -1.1580	1st Qu.: -1.1580	1st Qu.: -1.1660	1st Qu.:0.33202	1st Qu.: -1.1540	Up :605
Median :2000	Median : 0.2410	Median : 0.2410	Median : 0.2410	Median : 0.2380	Median : 0.2340	Median :1.00268	Median : 0.2410	
Mean :2000	Mean : 0.1506	Mean : 0.1511	Mean : 0.1472	Mean : 0.1458	Mean : 0.1399	Mean :1.57462	Mean : 0.1499	
3rd Qu.:2005	3rd Qu.: 1.4050	3rd Qu.: 1.4090	3rd Qu.: 1.4090	3rd Qu.: 1.4090	3rd Qu.: 1.4050	3rd Qu.:2.05373	3rd Qu.: 1.4050	
Max. :2010	Max. : 12.0260	Max. : 12.0260	Max. : 12.0260	Max. : 12.0260	Max. : 12.0260	Max. :9.32821	Max. : 12.0260	

Table 2: Covariance Matrix of Weekly Data Set

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
Year	1.00000000	-0.032289274	-0.03339001	-0.03000649	-0.031127923	-0.030519101	0.84194162	-0.032459894	-0.02220025
Lag1	-0.03228927	1.000000000	-0.07485305	0.05863568	-0.071273876	-0.008183096	-0.06495131	-0.075031842	-0.05000380
Lag2	-0.03339001	-0.074853051	1.00000000	-0.07572091	0.058381535	-0.072499482	-0.08551314	0.059166717	0.07269634
Lag3	-0.03000649	0.058635682	-0.07572091	1.00000000	-0.075395865	0.060657175	-0.06928771	-0.071243639	-0.02291281
Lag4	-0.03112792	-0.071273876	0.05838153	-0.07539587	1.000000000	-0.075675027	-0.06107462	-0.007825873	-0.02054946
Lag5	-0.03051910	-0.008183096	-0.07249948	0.06065717	-0.075675027	1.000000000	-0.05851741	0.011012698	-0.01816827
Volume	0.84194162	-0.064951313	-0.08551314	-0.06928771	-0.061074617	-0.058517414	1.00000000	-0.033077783	-0.01799521
Today	-0.03245989	-0.075031842	0.05916672	-0.07124364	-0.007825873	0.011012698	-0.03307778	1.000000000	0.72002470
Direction	-0.02220025	-0.050003804	0.07269634	-0.02291281	-0.020549456	-0.018168272	-0.01799521	0.720024704	1.00000000

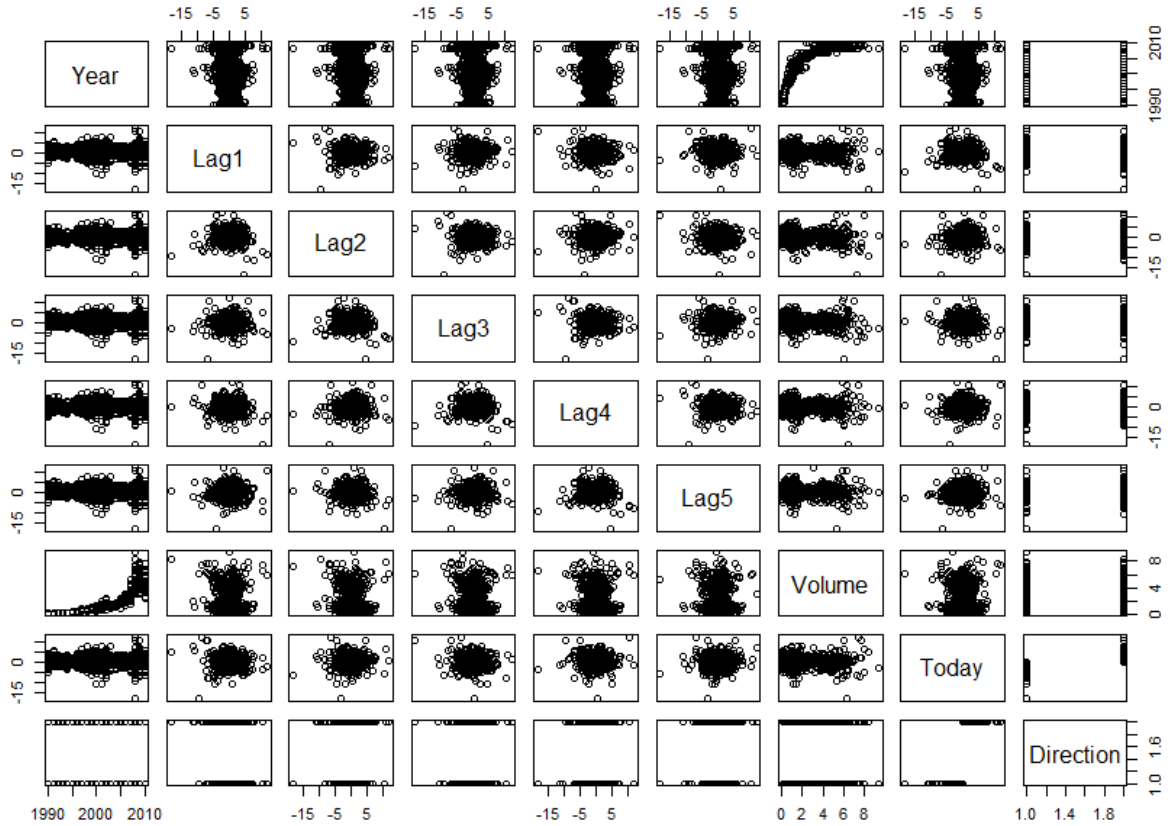


Figure 1: Scatterplot matrix of the Weekly dataset

Table 2 and Figure 3 shows that there is a strong positive correlation between Year and Volume. In addition, there is a strong positive correlation between Direction and Today; however, Direction is based on whether today was positive or negative. So, a correlation is expected. The remaining data appears uncorrelated with no discernable pattern.

b. Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

```
Call:
glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
    Volume, family = binomial, data = Weekly)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6949  -1.2565   0.9913   1.0849   1.4579

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.26686    0.08593   3.106   0.0019 **
Lag1        -0.04127    0.02641  -1.563   0.1181
Lag2         0.05844    0.02686   2.175   0.0296 *
Lag3        -0.01606    0.02666  -0.602   0.5469
Lag4        -0.02779    0.02646  -1.050   0.2937
Lag5        -0.01447    0.02638  -0.549   0.5833
Volume      -0.02274    0.03690  -0.616   0.5377
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1496.2  on 1088  degrees of freedom
Residual deviance: 1486.4  on 1082  degrees of freedom
AIC: 1500.4

Number of Fisher Scoring iterations: 4
```

Figure 2: Summary of logistic regression with response Direction

Figure 2 shows that Lag 2 appears to be statistically significant given by it low p-value.

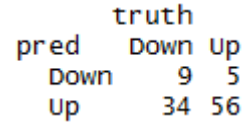
c. Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

```
      truth
pred   Down  Up
  Down   54  48
  Up    430 557
```

Figure 3: Confusion matrix of logistic regression model

Figure 3 shows that the logistic model has an accuracy of ~56.1%. In addition, the model performs significantly better predicting up days, ~92.1% accurate, versus down days, ~11.2% accurate.

d. Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

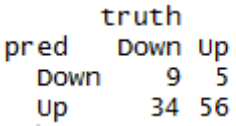


	truth	
pred	Down	Up
Down	9	5
Up	34	56

Figure 4: Confusion matrix of 2nd logistic regression model

Figure 4 shows that the 2nd logistic model has an accuracy of ~62.5%. In addition, the model performs significantly better predicting up days, ~91.8% accurate, versus down days, ~20.9% accurate.

e. Repeat (d) using LDA.

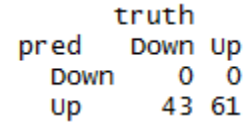


	truth	
pred	Down	Up
Down	9	5
Up	34	56

Figure 5: Confusion matrix of lda model

Figure 5 shows that the lda model has an accuracy of ~62.5%. In addition, the model performs significantly better predicting up days, ~91.8% accurate, versus down days, ~20.9% accurate.

f. Repeat (d) using QDA.

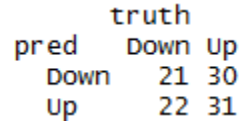


	truth	
pred	Down	Up
Down	0	0
Up	43	61

Figure 6: Confusion matrix of qda model

Figure 6 shows that the qda model has an accuracy of ~58.7%. In addition, the model performs significantly better predicting up days, ~100.0% accurate, versus down days, ~0.0% accurate.

g. Repeat (d) using KNN with K = 1.



	truth	
pred	Down	Up
Down	21	30
Up	22	31

Figure 7: Confusion matrix of knn model

Figure 6 shows that the knn model has an accuracy of ~50.0%. In addition, the model performs approximately the same predicting up days, ~50.8% accurate, as down days, ~51.2% accurate.

h. Which of these methods appears to provide the best results on this data?

Figures 3 through 7 shows that the Logistic Regression and the LDA model preforms the best on the test data.

i. Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

	truth	
pred	Down	Up
Down	17	20
Up	26	41

Figure 8: Confusion matrix of knn model, k = 9

Figure 8 shows that the knn model has an accuracy of ~55.8%.

	truth	
pred	Down	Up
Down	7	8
Up	36	53

Figure 9: Confusion matrix of logistic regression model, Lag1+Lag2

Figure 9 shows that the logistic regression model has an accuracy of ~57.6%.

	truth	
pred	Down	Up
Down	7	8
Up	36	53

Figure 10: Confusion matrix of lda model, Lag1+Lag2

Figure 10 shows that the lda model has an accuracy of ~57.6%.

	truth	
pred	Down	Up
Down	7	10
Up	36	51

Figure 11: Confusion matrix of qda model, Lag1+Lag2

Figure 11 shows that the qda model has an accuracy of ~55.8%.

Figures 8 through 11 shows that logistic regression and lda are the most appropriate model for this dataset.

5. (5.4 – 1) Using basic statistical properties of the variance, as well as single-variable calculus, derive (5.6). In other words, prove that α given by (5.6) does indeed minimize Var(αX + (1 − α)Y).

$$Var(\alpha X) = \alpha^2 Var(X)$$

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

Given the properties of variance above $Var(\alpha X + (1 - \alpha)Y)$ becomes

$$Var(\alpha X + (1 - \alpha)Y) = \alpha^2 Var(X) + (1-\alpha)^2 Var(Y) + 2\alpha(1-\alpha)Cov(X,Y)$$

In order to minimize the $Var(\alpha X + (1 - \alpha)Y)$ with respect to α, we take the derivative, set it equal to 0 and solve for α.

$$\frac{d}{d\alpha} Var(\alpha X + (1 - \alpha)Y) = 0$$

$$2\alpha Var(X) + 2(1-\alpha)(-1)Var(Y) + 2(1-2\alpha)Cov(X,Y) = 0$$

$$2\alpha Var(X) + 2\alpha Var(Y) - 4\alpha Cov(X,Y) = 2Var(Y) - 2Cov(X,Y)$$

$$\alpha = \frac{Var(Y) - Cov(X,Y)}{Var(X) + Var(Y) - 2Cov(X,Y)}$$

$\sigma_x^2 = Var(X), \sigma_y^2 = Var(Y)$, and $\sigma_{xy} = Cov(X,Y)$. So, α becomes which is equation 5.6

$$\alpha = \frac{\sigma_y^2 - \sigma_{xy}}{\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}} \qquad (5.6)$$

6. (5.4 – 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.
        i. $\Pr = \frac{n-1}{n}$
    b. What is the probability that the second bootstrap observation is *not* the jth observation from the original sample?
        i. $\Pr = \frac{n-1}{n}$
    c. Argue that the probability that the jth observation is *not* in the bootstrap sample is $(1 - 1/n)^n$.

The probability of that a bootstrap sample is not the jth observation is $\frac{n-1}{n}$ due to the fact that there a *n* total number of sample and *n-1* samples in the original dataset that are not the jth observations. So, the probability that the jth observation is not in the bootstrap is $\left(\frac{n-1}{n}\right)^n$, which is computed via $\left(\frac{n-1}{n}\right)_1 * \left(\frac{n-1}{n}\right)_2 * \ldots \left(\frac{n-1}{n}\right)_n$ or the probability of it not being in the first instance multiplied by the probability of it not being in the second and so on.

    d. When *n* = 5, what is the probability that the jth observation is in the bootstrap sample?

$$\Pr(\text{in}) = 1 - \Pr(out)$$

$$\Pr(\text{in}) = 1 - \left(\frac{n-1}{n}\right)^n$$

$$\Pr(\text{in}) = 1 - \left(\frac{5-1}{5}\right)^5$$

$$\Pr(\text{in}) = .672$$

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

$$\Pr(\text{in}) = .634$$

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

$$\Pr(\text{in}) = .632$$

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.

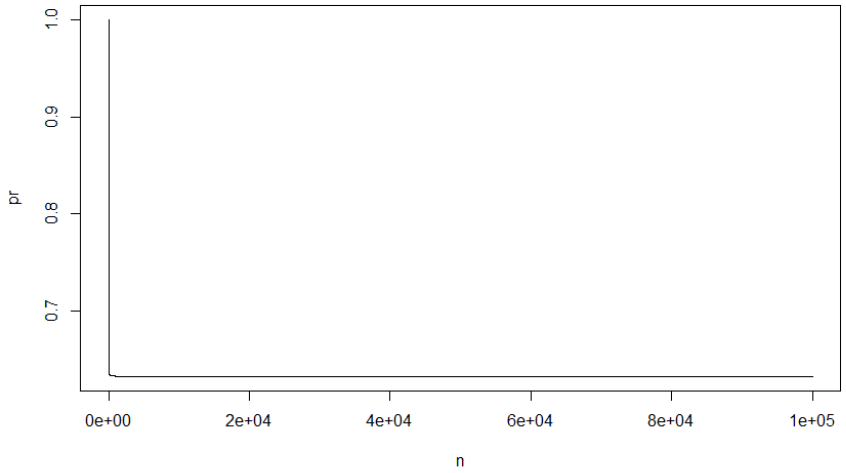


Figure 12: Probability of the jth observation in the bootstrap sample

Figure 12 shows that the probability of the jth observation being in the bootstrap sample quickly reaches an asymptote of approximately ~.632.

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.

```
 > store=rep(NA, 10000)
> for(i in 1:10000) {
        store[i]=sum(sample (1:100, rep=TRUE)==4) >0
} > mean(store)
```

Comment on the results obtained.

i. The result obtained was .6407 which is close to our calculated value of .634.

7. (5.4 – 6)We continue to consider the use of a logistic regression model to predict the probability of default using income and balance on the Default data set. In particular, we will now compute estimates for the standard errors of the income and balance logistic regression coefficients in two different ways: (1) using the bootstrap, and (2) using the standard formula for computing the standard errors in the glm() function. Do not forget to set a random seed before beginning your analysis.
   a. Using the summary() and glm() functions, determine the estimated standard errors for the coefficients associated with income and balance in a multiple logistic regression model that uses both predictors.

```
Call:
glm(formula = default ~ income + balance, family = binomial,
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4725  -0.1444  -0.0574  -0.0211   3.7245

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
balance      5.647e-03  2.274e-04  24.836  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1579.0  on 9997  degrees of freedom
AIC: 1585

Number of Fisher Scoring iterations: 8
```

Figure 13: Summary of logistic regression with response Default

Figure 13 shows the standard error for the coefficients associated with income and balance.

   b. Write a function, boot.fn(), that takes as input the Default data set as well as an index of the observations, and that outputs the coefficient estimates for income and balance in the multiple logistic regression model.

```
boot.fn = function(data,index){
  logreg.fit4 = glm(default ~ income + balance, data=data, family="binomial", subset=index)
  return (coefficients(logreg.fit4))
}
```

Figure 14: Boot.fn function

   c. Use the boot() function together with your boot.fn() function to estimate the standard errors of the logistic regression coefficients for income and balance.

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = Default, statistic = boot.fn, R = 1000)


Bootstrap Statistics :
        original        bias     std. error
t1* -1.154047e+01 -3.566749e-02 4.408921e-01
t2*  2.080898e-05  2.589821e-07 4.968550e-06
t3*  5.647103e-03  1.518670e-05 2.332941e-04
```

Figure 14: Estimated standard error of the logistic regression coefficients for income and balance

d. Comment on the estimated standard errors obtained using the glm() function and using your bootstrap function.
    i. The estimated standard error obtained using the glm() function and my bootstrap function are similar.

8. (5.4 – 8) We will now perform cross-validation on a simulated data set.
    a. Generate a simulated data set as follows:
    > set.seed(1)
    > y=rnorm(100)
    > x=rnorm(100)
    > y=x-2*x^2+rnorm (100)
    In this data set, what is n and what is p? Write out the model used to generate the data in equation form.
        i. $n = 100, p = 2; Y = X - 2X^2 + \epsilon$

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

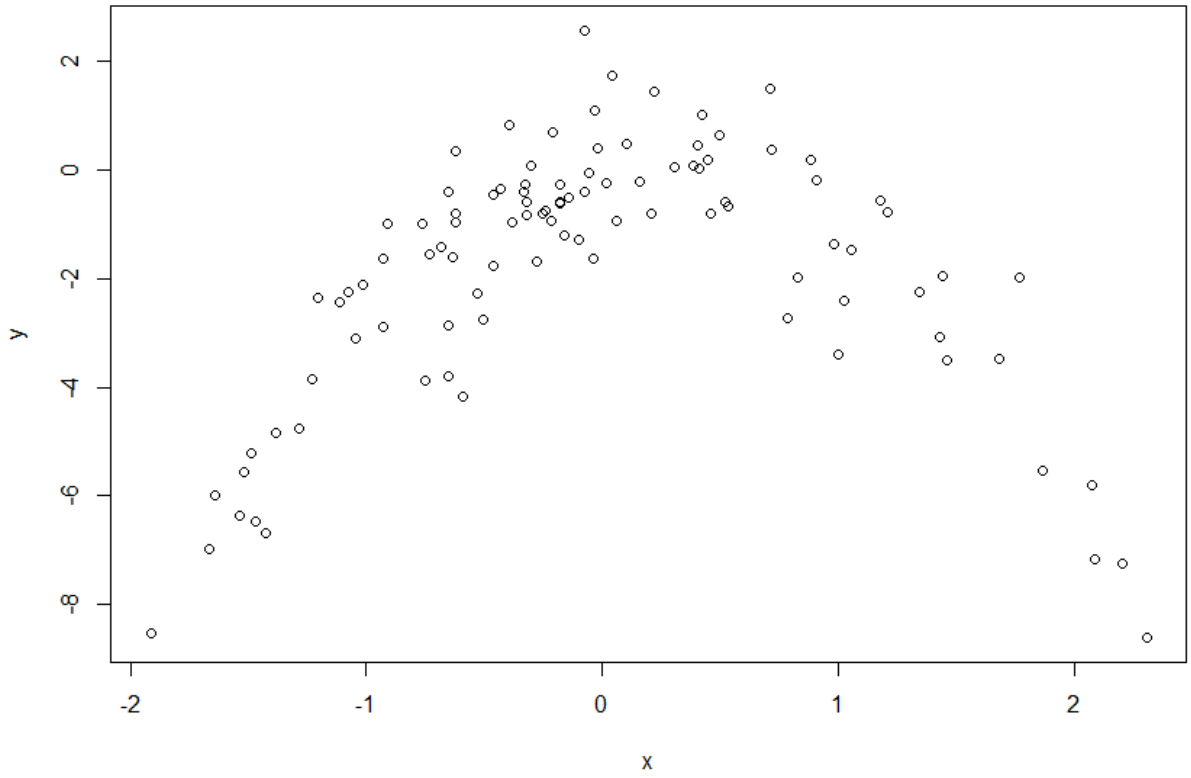


Figure 15: Scatterplot of X and Y

Figure 15 shows a point that appears to follow a quadratic ranging in the x from -2 to 2 and y from -8 to 2, which makes sense given the function used to create it.

c. Set a random seed, and then compute the LOOCV errors that result from fitting the following four models using least squares:
   i. $Y = \beta0 + \beta1X + \varepsilon$
      1. 5.89

   ii. $Y = \beta0 + \beta1X + \beta2X2 + \varepsilon$
      1. 1.09

   iii. $Y = \beta0 + \beta1X + \beta2X2 + \beta3X3 + \varepsilon$
      1. 1.1

   iv. $Y = \beta0 + \beta1X + \beta2X2 + \beta3X3 + \beta4X4 + \varepsilon$
      1. 1.11

   Note you may find it helpful to use the data.frame() function to create a single data set containing both X and Y .

d. Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?
    i. $Y = \beta0 + \beta1X + \varepsilon$
        1. 5.89

    ii. $Y = \beta0 + \beta1X + \beta2X2 + \varepsilon$
        1. 1.09

    iii. $Y = \beta0 + \beta1X + \beta2X2 + \beta3X3 + \varepsilon$
        1. 1.1

    iv. $Y = \beta0 + \beta1X + \beta2X2 + \beta3X3 + \beta4X4 + \varepsilon$
        1. 1.11

    Yes, the results were the exact same for part c and d. Changing the seed had no effect on the LOOCV because no randomness is involved in the procedure. LOOCV merely fits the model n time leaving out each observation once and takes the average of the MSE error.

e. Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.
    i. Model ii or the quadratic model has the smallest LOOCV error. This is as expected since the model that generated the data was a quadratic.

f. Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

```
Call:
lm(formula = y ~ x, data = dataframe)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3469 -0.9275  0.8028  1.5608  4.3974

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.8185     0.2364  -7.692 1.14e-11 ***
x             0.2430     0.2479   0.981    0.329
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.362 on 98 degrees of freedom
Multiple R-squared:  0.009717,  Adjusted R-squared:  -0.0003881
F-statistic: 0.9616 on 1 and 98 DF,  p-value: 0.3292
```

Figure 16: Least Squares Regression model of i

```
Call:
lm(formula = y ~ x + I(x^2), data = dataframe)

Residuals:
     Min       1Q   Median       3Q      Max
-2.89884 -0.53765  0.04135  0.61490  2.73607

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.09544    0.13345  -0.715    0.476
x            0.89961    0.11300   7.961 3.24e-12 ***
I(x^2)      -1.86665    0.09151 -20.399  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.032 on 97 degrees of freedom
Multiple R-squared:  0.8128,    Adjusted R-squared:  0.8089
F-statistic: 210.6 on 2 and 97 DF,  p-value: < 2.2e-16
```

Figure 17: Least Squares Regression model of ii

```
Call:
lm(formula = y ~ x + I(x^2) + I(x^3), data = dataframe)

Residuals:
     Min       1Q   Median       3Q      Max
-2.87250 -0.53881  0.02862  0.59383  2.74350

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.09865    0.13453  -0.733    0.465
x            0.95551    0.22150   4.314  3.9e-05 ***
I(x^2)      -1.85303    0.10296 -17.998  < 2e-16 ***
I(x^3)      -0.02479    0.08435  -0.294    0.769
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.037 on 96 degrees of freedom
Multiple R-squared:  0.813,     Adjusted R-squared:  0.8071
F-statistic: 139.1 on 3 and 96 DF,  p-value: < 2.2e-16
```

Figure 18: Least Squares Regression model of iii

```
Call:
lm(formula = y ~ x + I(x^2) + I(x^3) + I(x^4), data = dataframe)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8914 -0.5244  0.0749  0.5932  2.7796

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.13897    0.15973  -0.870 0.386455
x            0.90980    0.24249   3.752 0.000302 ***
I(x^2)      -1.72802    0.28379  -6.089  2.4e-08 ***
I(x^3)       0.00715    0.10832   0.066 0.947510
I(x^4)      -0.03807    0.08049  -0.473 0.637291
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.041 on 95 degrees of freedom
Multiple R-squared:  0.8134,    Adjusted R-squared:  0.8055
F-statistic: 103.5 on 4 and 95 DF,  p-value: < 2.2e-16
```

Figure 19: Least Squares Regression model of iv

Figure 16 through 18 show least squares regression models of i through iv. In the models that include x^2, x and x^2 are the only significant coefficient as determined by the p-value. This agrees with our cross-validation results and makes since the model that created the data was quadratic.