Homework #3

Zhijian Liu

Multiple Linear Regression

1. Page 123, chap. 3, ≈ #10. Consider the following variables from the *Carseats* data set.

Sales	Unit sales (in thousands) at each location
Price	Price company charges for car seats at each site
Urban	A factor with levels No and Yes to indicate whether the store is in an urban or rural location
US	A factor with levels No and Yes to indicate whether the store is in the US or not

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
##
## Call:
## lm(formula = Sales ~ Price + as.factor(Urban) + as.factor(US))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         13.043469   0.651012  20.036  < 2e-16 ***
## Price               -0.054459   0.005242 -10.389  < 2e-16 ***
## as.factor(Urban)Yes -0.021916   0.271650  -0.081    0.936
## as.factor(US)Yes     1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393,    Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b) Write out the model in equation form, being careful to handle the qualitative variables properly. That is, write a separate model for each category.

$$\begin{cases} Sales = 13.04 - 0.05 \cdot Price & \text{Urban} = \text{No, US} = \text{No} \\ Sales = 13.02 - 0.05 \cdot Price & \text{Urban} = \text{Yes, US} = \text{No} \\ Sales = 14.24 - 0.05 \cdot Price & \text{Urban} = \text{No, US} = \text{Yes} \\ Sales = 14.22 - 0.05 \cdot Price & \text{Urban} = \text{Yes, US} = \text{Yes} \end{cases}$$

(c) For which of the predictors the null hypothesis $H_0 : \beta_j = 0$ is <u>not</u> rejected? Verify your conclusion with the appropriate partial F-test (you can find examples in HW2 and the regression handout). Compare p-values of this test and the corresponding t-test. Also, compare the partial F-statistic with the squared t-statistic.

From (a), we know that for Urban, the null hypothesis $H_0 : \beta_j = 0$ is not rejected. p-values of this test is the same as the corresponding t-test. The partial F-statistic is the same as the squared t-statistic as well.

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + as.factor(Urban) + as.factor(US)
## Model 2: Sales ~ Price + as.factor(US)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    396 2420.8
## 2    397 2420.9 -1  -0.03979 0.0065 0.9357
```

	t-test	F-test
p-value	0.936	0.9357

	t^2	F
statistics	0.006561	0.0065

(d) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```
## lm(formula = Sales ~ Price + as.factor(US))
##                   coef.est coef.se
## (Intercept)       13.03     0.63
## Price             -0.05     0.01
## as.factor(US)Yes   1.20     0.26
## ---
## n = 400, k = 3
## residual sd = 2.47, R-Squared = 0.24
```

(e) Test linearity of Sales as a function of Price, according to our regression model. Is there a significant lack of fit?

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price
## Model 2: Sales ~ as.factor(Price)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    398 2552.2
## 2    299 1932.3 99    619.92 0.9689 0.5652
```

The lack of fit is not significant, so the linearity of Sales over Price is not violated.

Logistic Regression

2. Page 170, chap. 4, #9. This problem has to do with odds.

(a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

$$\frac{\pi}{1-\pi} = 0.37$$
$$\pi = 0.270073$$

The fraction of people is 0.270073 on average.

(b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

$$\pi = 0.16$$
$$\frac{\pi}{1-\pi} = 0.1904762$$

The odds that she will default is 0.1904762.

3. Page 170, chap. 4, #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 coefficients, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

(a) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 + \hat{\beta}_2 \cdot X_2$$
$$log\frac{\hat{\pi}}{1-\hat{\pi}} = -6 + 0.05 \cdot 40 + 1 \cdot 3.5$$
$$log\frac{\hat{\pi}}{1-\hat{\pi}} = -0.5$$
$$\hat{\pi} = 0.3775407$$

The probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class is 0.3775407.

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

$$log\frac{0.5}{1-0.5} = -6 + 0.05 \cdot (X_1) + 1 \cdot 3.5$$
$$0 = -2.5 + 0.05 \cdot (X_1)$$
$$X_1 = 50$$

The student need to study 50 hours to have a 50% chance of getting an A in the class.

KNN

4. Pages 53-54, chap. 2, #7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

Obs	X1	X2	X3	Y	distance
1	0	3	0	Red	3.00
2	2	0	0	Red	2.00
3	0	1	3	Red	3.16
4	0	1	2	Green	2.24
5	-1	0	1	Green	1.41
6	1	1	1	Red	1.73

(b) What is our prediction with $K = 1$? Why?
Y would be Green, because the 5^{th} observation is the nearest neighbour with Y = Green.

(c) What is our prediction with K = 3? Why?
Y would be rend, because the 2^{th}, 5^{th} and 6^{th} observations are the nearest neighbours with the mode of Y = Green.

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?
We expect a small K, since a small K stands for a flexible method that would produce a highly nonlinear boundary.

5. When the number of features p is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the curse of dimensionality, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigate this curse.

(a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly (evenly) distributed on [0, 1]. Associated with each observation is a response value. Suppose that we wish to predict a test observations response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.6$, we will use observations in the range [0.55, 0.65]. On average, what fraction of the available observations will we use to make the prediction?

```
X <- runif(10000)
predict.X <- X[which(X >= 0.55 & X <= 0.65)] # where x in [0.55, 0.65]
fraction <- length(predict.X)/length(X)
fraction
```

```
## [1] 0.1028
```

$\frac{0.65-0.55}{1} = 0.1$. So 10% of the available observations will be used.

(b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2. We assume that (X_1 , X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observations response using only observations that are within 10% of the range of X_1 and within 10% of the range of X_2 closest to that test observation. For instance, in order to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$, we will use observations in the range [0.55, 0.65] for X_1 and in the range [0.3, 0.4] for X_2. On average, what fraction of the available observations will we use to make the prediction?

```
X1 <- X2 <- runif(10000)
predict.X1 <- X1[which(X1 >= 0.55 & X1 <= 0.65)] # where x in [0.55, 0.65]
predict.X2 <- X2[which(X2 >= 0.3 & X2 <= 0.4)] # where x in [0.3, 0.4]
fraction <- (length(predict.X1) *length(predict.X2))/(length(X1)*length(X2))
fraction
```

```
## [1] 0.01030081
```

$\frac{(0.65-0.55)\times(0.4-0.3)}{1\times1} = 0.1$. So 1% of the available observations will be used.

(c) Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observations response using observations within the 10% of each features range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

$fraction = \frac{(0.1)\times(0.1)\times(0.1)\times(0.1)\times\cdots\times(0.1)}{1\times1\times1\times1\times\cdots\times1} = 0.1^{100} = 10^{-100}$. So 10^{-100} of the available observations will be used.

(d) Using your answers to parts (a)(c), argue that a drawback of KNN when p is large is that there are very few training observations "near" any given test observation.

Using the same percentage of range within each feature, when p is large, the number of "near" training observations will be very small.

(e) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For p = 1, 2, and 100, what is the length of each side of the hypercube? Comment on your answer.

The length of each side of the hypercube would be 10% of the range of each feature, if each feature is uniformly distributed. In this example, the length would be 0.1.