1. 假设我们希望判断一个水果是苹果，还是橙子。基于某个数据集，我们建立了一个逻辑回归模型，模型估算结果如下（x是一个未知的自变量）：

$P(Y=\text{橙子}|X=x)=\dfrac{\exp(\hat{\beta}_0+\hat{\beta}_1*x)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1*x)}$利用同样的数据集，你的朋友也建立了一个逻辑回归模型，但是他采用的是 softmax 的多分类建模格式（参考教科书"ISLR" 第 141 页，或者我们的课件 Multinomial Logistic Regression 部分），模型估算结果如下：

$$P(Y=\text{橙子}|X=x)=\frac{\exp(\hat{\alpha}_{¿ange0}+\hat{\alpha}_{orange1}*x)}{\exp(\hat{\alpha}_{¿ange0}+\hat{\alpha}_{orange1}*x)+\exp(\hat{\alpha}_{apple0}+\hat{\alpha}_{apple1}*x)}$$

请回答一下问题。（10 分）

(a) 你的模型中，橙子相比于苹果的对数几率(log odds)是多少？（2 分）

Answer: B0+B1*X

(b) 你朋友的模型中，橙子相比于苹果的对数几率(log odds)是多少？（2 分）

Answer: a orange0 -a apple0 +（a orange1 – a apple1）* X

(c) 假设你的模型中，$\hat{\beta}_0=2$，$\hat{\beta}_1=-1$。那么，你朋友的模型中回归系数之间的关系是怎么样的。（2 分）

Answer :

$P(Y=\text{橙子}|X=x)=\dfrac{\exp(\hat{\beta}_0+\hat{\beta}_1*x)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1*x)} = \dfrac{1}{1+e^{-\beta_0-\beta_1 x}}$

$P(Y=\text{橙子}|X=x)=\dfrac{\exp(\hat{\alpha}_{¿ange0}+\hat{\alpha}_{orange1}*x)}{\exp(\hat{\alpha}_{¿ange0}+\hat{\alpha}_{orange1}*x)+\exp(\hat{\alpha}_{apple0}+\hat{\alpha}_{apple1}*x)} = \dfrac{1}{1+e^{\alpha_{ap0}-\alpha_{or0}+(\alpha_{ap1}-\alpha_{or1})\cdot x}}$

朋友回归系数有 a orange0 -a apple0 =2， a orange1 – a apple1 =-1 的关系

(d) 假设针对另外一个数据集，你和你的朋友建立了两个同样的模型。这一次，你朋友模型的各回归系数估计结果分别为$\hat{\alpha}_{¿ange0}=1.2$，$\hat{\alpha}_{¿ange1}=-2$，$\hat{\alpha}_{app\leq 0}=3$，$\hat{\alpha}_{apple1}=0.6$。那么，你的模型中各回归系数的估计结果是多少？（2 分）

B0=-1.8，B1=-2.6

2、

(a) 利用所有数据，用 5 个 Lag 变量和 Volume 作为预测变量，以 Direction 作为响应变量，建立一个逻辑回归模型(Logistic Regression Model)。利用 summary 函数打出模型估计结果。有变量展示出显著影响吗？是哪些？（3 分）

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

选取 0.05 为显著水平门槛时，Intercept 和 Lag2 表现显著影响

(b) 基于上述模型，给出预测结果的混淆矩阵(confusion matrix)，以及准确率(Accuracy)。这个混淆矩阵显示逻辑回归模型犯了哪种错误？ （2 分）

```
      Down  Up
FALSE   54  48
TRUE   430 557
```

准确度 0.561

(Accuracy)。这个混淆矩阵显示逻辑回归模型犯了哪种错误？ （2 分）

(c) 对数据进行分割，选取 1990-2008 年的数据作为训练(train)数据集，2009-2010 年的数据作为检验(test)数据集。对训练数据集，只用 ~~Lag2~~ 作为影响因素，建立逻辑回归模型(Logistic Regression Model)来预测 Direction，给出所建模型估计结果。（5 分）

Volume

```
predict_res
FALSE  TRUE
   12    92
```

```
table(test_dat$Direction,test_dat$predict_res)
...

      FALSE TRUE
Down      8   35
Up        4   57
```

（d） 利用线性判别分析(LDA)方法重复问题(d)。（5 分）

```
prect_lda
Down    Up
  45    59
```

```
table(test_dat$Direction,prect_lda)
```

```
      prect_lda
       Down Up
Down     20 23
Up       25 36
```

（e） 利用 K-最近邻(KNN)方法重复问题(d)，其中 K=1。（5 分）

```
knn1
Down    Up
  55    49
```

（f） 利用朴素贝叶斯(Naïve Bayes)方法重复问题(d)。（5 分）

```
       Down  Up
Down     43   0
Up       57   4
```

```
Down    Up
 100     4
```

table(test_dat$prectit_nb)

（g） 分别利用 4 个所建模型预测检验数据集，给出预测结果的混淆矩阵(confusion matrix)，以及准确率(Accuracy)。对比分析评价上述 4 个模型的表现，哪个模型表现最优？（5 分）

```
table(test_dat$Direction,test_dat$predict_res)

       FALSE  TRUE
Down       8    35
Up         4    57
```

Logitstic 准确度0.625

```
table(test_dat$Direction,prect_lda)

      prect_lda
       Down Up
Down     20 23
Up       25 36
```

Lda 准确率 0.538

```
       Down  Up
Down     43   0
Up       57   4
```

Naïve bayes 准确率 0.452

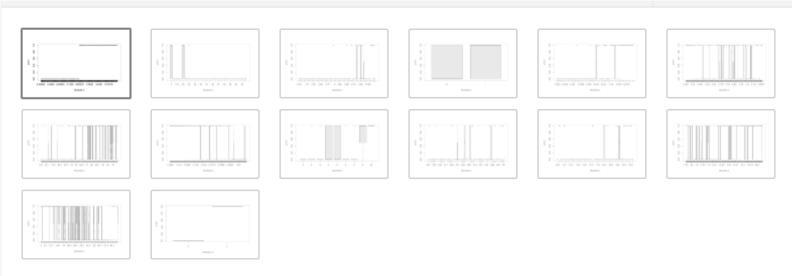
```
      knn1
       Down  Up
Down     26  17
Up       29  32
```

Knn 准确率 0.556

四个中逻辑回归的准确率最高，使用在检验数据集时，逻辑回归模型表现最优

3 、

（a） 创造 1 个新的 binary 变量 crim1，也就是新建 1 列数据。如果一个普查区的犯罪率(crim)大于所有人口普查区犯罪率的中位数(median)，那么 crim1 等于 1；否则，crim1 等于 0。画出 crim1 与其他变量的箱型图。观察箱型图，分析哪些变量可用于有效预测 crim1。注：计算犯罪率的中位数用 median() 函数。（5 分）



其中 crim，nox,age,rad,medv 可用于有效预测 crim1

（b） 对数据进行分割，选取前 400 个数据作为训练(train)数据集，而后 106 个数据作为检验(test)数据集。对训练数据集，选择合适的变量建立逻辑回归模型(Logistic Regression Model)来预测 crim1。给出所建模型估计结果，并描述你的发现。（5 分）

```
warning: glm.fit
predict_lot
FALSE  TRUE
   14    92
```

所用参数的 pr 都大于 0.05，都对 crim1 影响不显著

（c） 利用线性判别分析(LDA)方法重复问题(d)。（5 分）

```
predict_lda2
  0   1
  3 103
```

（d） 利用朴素贝叶斯(Naïve Bayes)方法重复问题(d)。（5 分）

```
test_dat
 0  1
15 91
```

（e） 利用 K-最近邻(KNN)方法重复问题(d)。你需要自己确定合适的 K 值，并做出解释。（10 分）

逻辑回归准确率最高

4、

```
[1] "k取为"
[1] 1
   knn1
     0  1
  0  9  6
  1  7 84
[1] "k取为"
[1] 2
   knn1
     0  1
  0  9  6
  1  6 85
[1] "k取为"
[1] 3
   knn1
     0  1
  0  9  6
  1 13 78
[1] "k取为"
[1] 4
   knn1
     0  1
  0  8  7
  1 14 77
[1] "k取为"
[1] 5
   knn1
     0  1
  0 10  5
  1 17 74
[1] "k取为"
[1] 6
   knn1
     0  1
  0 10  5
  1 15 76
[1] "k取为"
[1] 7
   knn1
     0  1
  0  9  6
  1 19 72
```

选取 k 为 2，因为此时准确率最高

(f) 分别利用 4 个所建模型预测检验数据集，给出预测结果的混淆矩阵(confusion matrix)，以及准确率(Accuracy)。对比分析评价上述四个模型的表现，哪个模型表现最优？可能需要用到 predict()函数。（5 分）

```
   knn1
     0  1
  0  9  6
  1  6 85
```

Knn 准确率为0.887

```
  test_dat
     0  1
  0 13  2
  1  2 89
```

Nab 准确率为0.981

```
test_dat
  0   1
  3 103
```

lda准确率为0.981

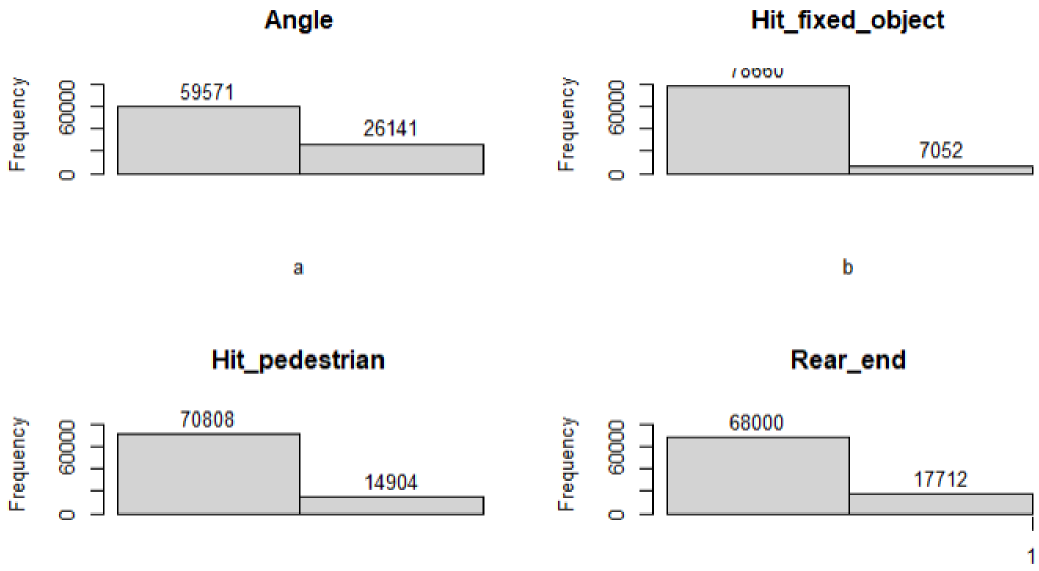
```
predict_lot  0  1
      FALSE 14  0
      TRUE   1 91
```

逻辑回归准确率 0.99

逻辑回归准确率最高

4、

(a)　创造 1 个新的变量 casualty（伤亡人数），定义为事故死亡人数与受伤人数之和。分别画出每种事故类型(Collision)对应的伤亡人数的直方图（可参考 Lecture5 第 6 页），并求出相应的平均值与方差（可使用 doBy 工具包的 summaryBy 函数）。（5 分）




```
          Angle Hit_fixed_object
      0.3913338        0.1482640
 Hit_pedestrian         Rear_end
      0.1741413        0.2862610
```

(b)　以 casualty 为因变量，以 Daytime，Weather 和 Collision 为自变量，建立一个泊松回归模型(Poisson Regression)模型。展示模型结果，并对除 Intercept 之外的所有其他回归系数估值(estimate)进行解释（是否显著影响伤亡人数，以及如何影响）。可参考 Lecture5 第 10 页。（10 分）

```
Call:
glm(formula = casualty ~ Daytime + Weather + Collision, family = poisson(link = "log"),
    data = philly)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5922  -0.2468  -0.1388   0.2840  15.1024

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               0.214098   0.007003  30.571  < 2e-16 ***
Daytime                   0.022937   0.007010   3.272 0.001067 **
WeatherRain              -0.036865   0.009502  -3.880 0.000104 ***
WeatherSnow              -0.137893   0.026044  -5.295 1.19e-07 ***
CollisionHit_fixed_object -0.535170  0.011737 -45.596  < 2e-16 ***
CollisionHit_pedestrian  -0.177213   0.009374 -18.905  < 2e-16 ***
CollisionRear_end        -0.101328   0.007742 -13.088  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 78370  on 85711  degrees of freedom
Residual deviance: 75720  on 85705  degrees of freedom
AIC: 221573

Number of Fisher Scoring iterations: 5
```

选取 0.05 为显著水平门槛时,**全部回归系数估值都对 casualty 产生显著影响，Daytime 的回归系数大于 0，与应变量正相关，其余的回归系数均小于 0，与应变量负相关**

(c) 利用线性回归模型重复问题(b)。（5 分）

```
Call:
lm(formula = casualty ~ Daytime + Weather + Collision, data = philly)

Residuals:
   Min     1Q Median     3Q    Max
-1.265 -0.265 -0.145  0.248 41.855

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.239936   0.007741 160.179  < 2e-16 ***
Daytime                    0.025147   0.007503   3.352 0.000803 ***
WeatherRain               -0.039369   0.009981  -3.944 8.01e-05 ***
WeatherSnow               -0.133354   0.025351  -5.260 1.44e-07 ***
CollisionHit_fixed_object -0.513221   0.010953 -46.856  < 2e-16 ***
CollisionHit_pedestrian   -0.202976   0.010101 -20.094  < 2e-16 ***
CollisionRear_end         -0.120416   0.008624 -13.963  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

选取 0.05 为显著水平门槛时，**全部回归系数估值都对 casualty 产生显著影响，Daytime 的回归系数大于 0，与应变量正相关，其余的回归系数均小于 0，与应变量负相关**

（d） 分别计算两个模型预测结果的均方根误差(RMSE, root mean squared error)。$RMSE=\sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2}{n}}$，其中$\hat{y}_i$为预测值，$y_i$为真实值。对比 RMSE，评价泊松模型是否表现更优，并讨论为什么我们倾向于对计数数据采用泊松回归模型。（5 分）

r_g	1.44735068598841
r_l	1.02644093655846

没有表现更优，计数变量一般只能取有限范围内的非负整数，虽然可以使用线性回归模型进行最小二乘法估计，但是会带来严重的异方差问题。泊松回归的特殊性在于，**它的因变量，是记录某个特定事件出现的次数（有序的非负整数），它们被称之为“计数数据”。**普通的线性回归模型是无法对计数数据建模的.