lab8 - Jin Kweon (3032235207)

Jin Kweon

10/23/2017

```r
default <- Default
names(default)
```

```
## [1] "default" "student" "balance" "income"
```

```r
summary(default)
```

```
##  default     student       balance           income     
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772  
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340  
##                        Median : 823.6   Median :34553  
##                        Mean   : 835.4   Mean   :33517  
##                        3rd Qu.:1166.3   3rd Qu.:43808  
##                        Max.   :2654.3   Max.   :73554
```

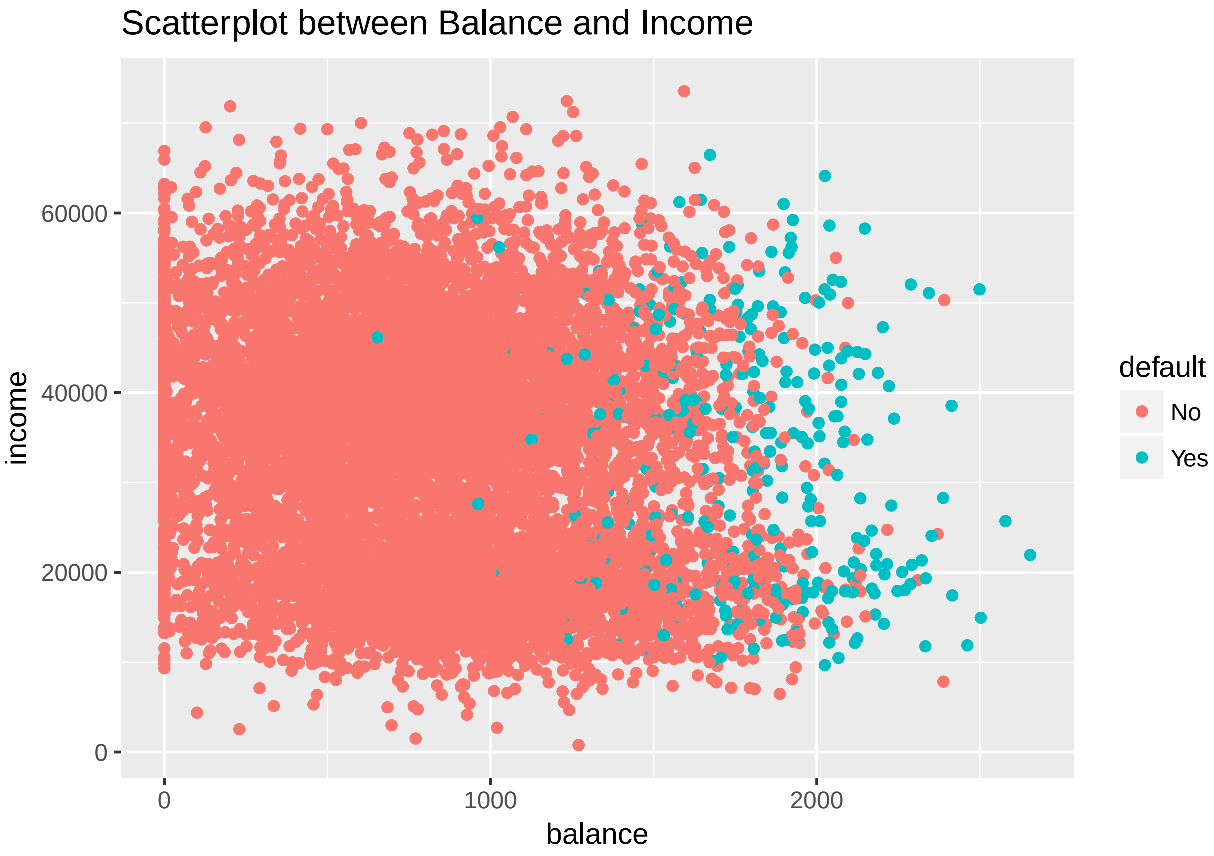
```r
summary(subset(default, default == 'Yes'))
```

```
##  default   student       balance           income     
##  No :  0   No :206   Min.   : 652.4   Min.   : 9664  
##  Yes:333   Yes:127   1st Qu.:1511.6   1st Qu.:19028  
##                      Median :1789.1   Median :31515  
##                      Mean   :1747.8   Mean   :32089  
##                      3rd Qu.:1988.9   3rd Qu.:43067  
##                      Max.   :2654.3   Max.   :66466
```

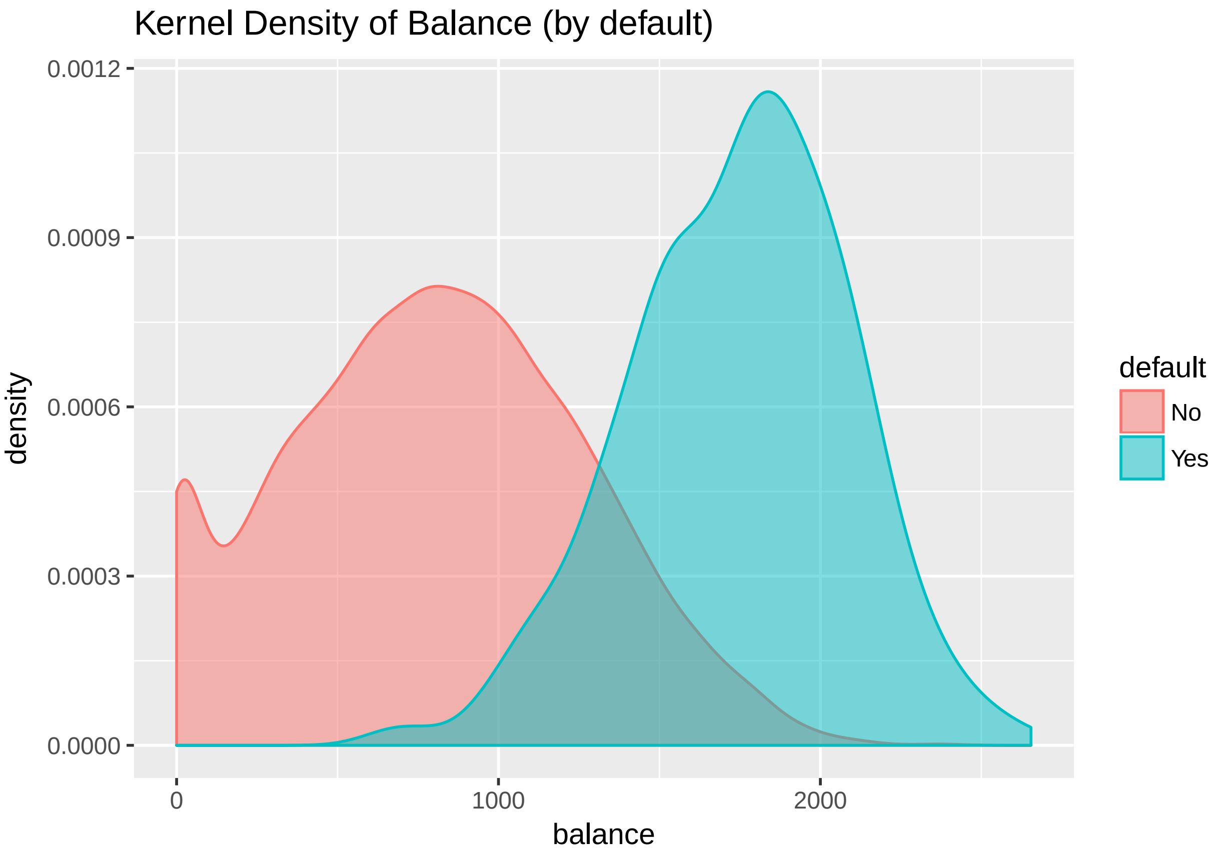
```r
summary(subset(default, default == 'No'))
```

```
##  default     student       balance           income     
##  No :9667   No :6850   Min.   :   0.0   Min.   :  772  
##  Yes:   0   Yes:2817   1st Qu.: 465.7   1st Qu.:21405  
##                        Median : 802.9   Median :34589  
##                        Mean   : 803.9   Mean   :33566  
##                        3rd Qu.:1128.2   3rd Qu.:43824  
##                        Max.   :2391.0   Max.   :73554
```

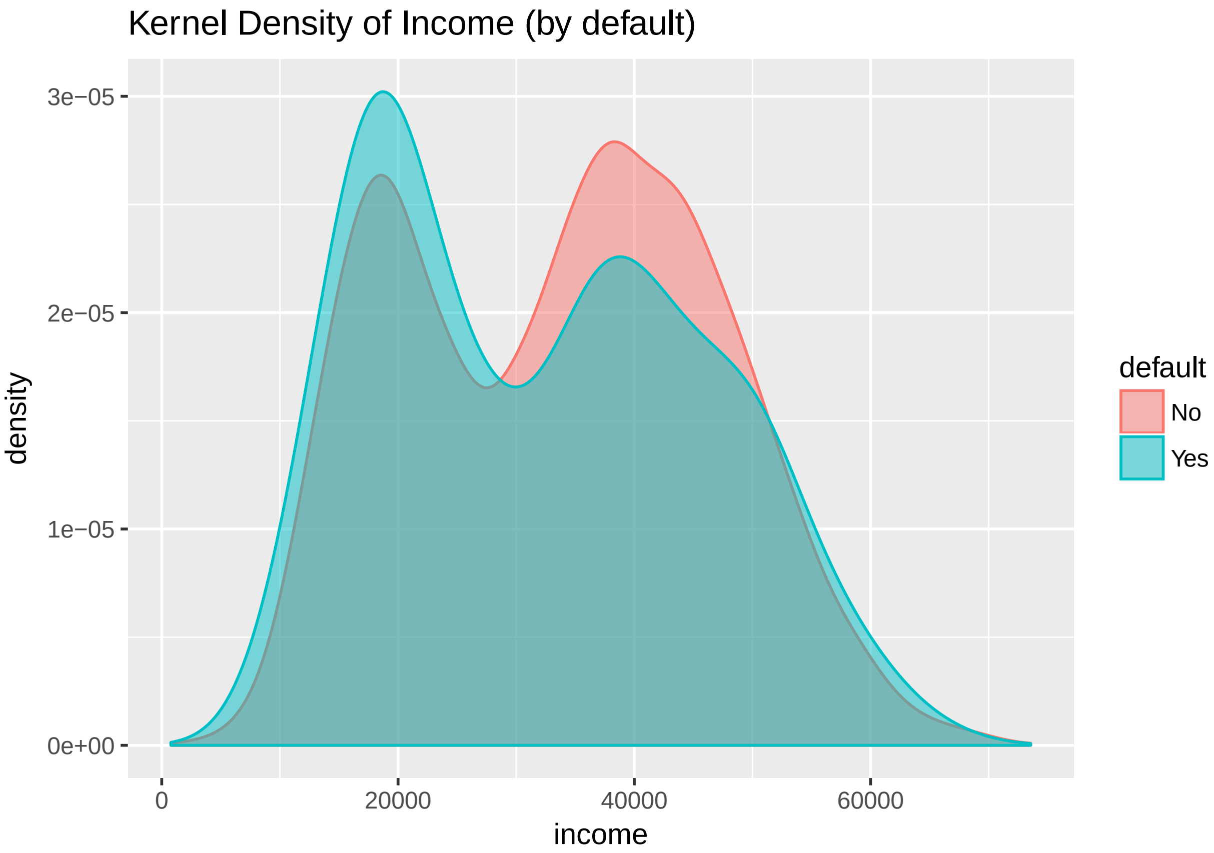
```r
ggplot(default, aes(x = balance, y = income, fill = default, col = default)) + geom_point() +
  ggtitle("Scatterplot between Balance and Income")
```



```
ggplot(default, aes(x = balance, fill = default, col = default)) + geom_density(alpha = 0.5) +
  ggtitle("Kernel Density of Balance (by default)")
```



```
ggplot(default, aes(x = income, fill = default, col = default)) + geom_density(alpha = 0.5) +
  ggtitle("Kernel Density of Income (by default)")
```



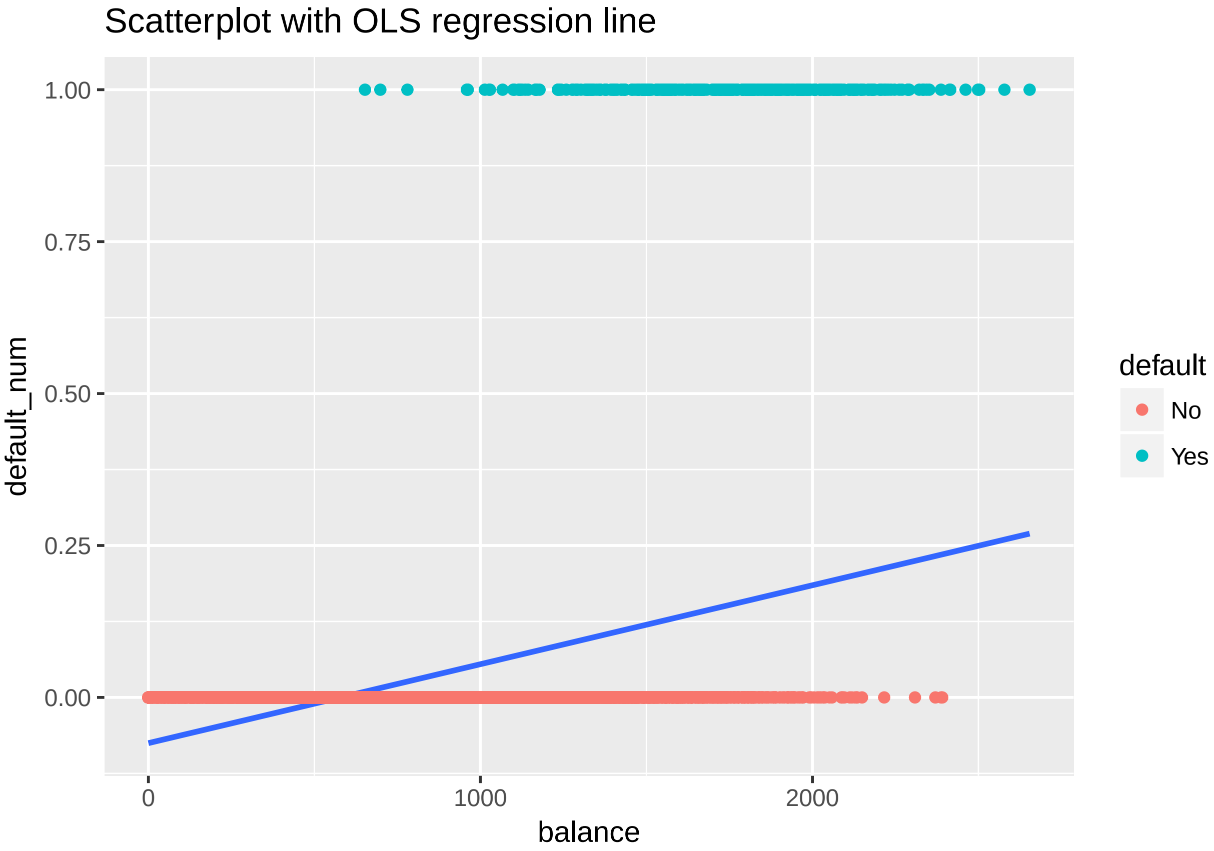
Balance matters a lot on how ppl decide to go for either default yes or no.

```
default_numeric <- rep(0, nrow(default))
default_numeric[default$default == 'Yes'] <- 1

default$default_num <- default_numeric
ols_reg <- lm(default_num ~ balance, data = default)
summary(ols_reg)
```

```
##
## Call:
## lm(formula = default_num ~ balance, data = default)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23533 -0.06939 -0.02628  0.02004  0.99046
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.519e-02  3.354e-03  -22.42   <2e-16 ***
## balance      1.299e-04  3.475e-06   37.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1681 on 9998 degrees of freedom
## Multiple R-squared:  0.1226, Adjusted R-squared:  0.1225
## F-statistic:  1397 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
ggplot(default, aes(x = balance, y = default_num)) +
  geom_smooth(method = "lm", se = FALSE) + geom_point(aes(col = default)) +
  ggtitle("Scatterplot with OLS regression line")
```



The response default falls into one of two categories: “Yes” or “No”. Rather than modeling default directly, logistic regression models the probability that the response Y belongs to a particular category.

Q. In “How do we interpret the coefficients? A one-unit increase in balance is associated with an increase in the log odds of default by 0.005 units,” what is log odds? ===> log (p / (1-p)).

Q. On pg 136 of ISL, it says “The negative coefficient for student in the multiple logistic regression indicates that for a fixed value of balance and income, a student is less likely to default than a non-student. Indeed, we observe from the left-hand panel of Figure 4.3 that the student default rate is at or below that of the non-student default rate for every value of balance. But the horizontal broken lines near the base of the plot, which show the default rates for students and non-students averaged over all values of balance and income, suggest the opposite effect: the overall student default rate is higher than the non-student default rate. Consequently, there is a positive coefficient for student in the single variable logistic regression output shown in Table 4.2.” And, I dont understand the part “But the horizontal broken lines near the base of the plot, which show the default rates for students and non-students averaged over all values of balance and income, suggest the opposite effect: the overall student default rate is higher than the non-student default rate.”… ==> So, at the fixed points of balance and income, student is less likely to default than a non-student, which makes sense. Cuz, when students have the same balance and income with non-student, they do not want to take risk and default less. However, when we do not fix balance and income, in average/in overall, students default rate is higher. So, at fixed point student is less likely to default, but overall student default more often.

Q. But how does the above reasoning makes sense??? If we add up all fixed points of balance and income, then it will be overall…??? Is it because the number of students and nonstudents are different in the data?

==> cuz it is only true for some fixed points. . . .

```r
logreg_default <- glm(default ~ balance, family = binomial, data = default)

summary(logreg_default)$coefficients
```

```
##                  Estimate    Std. Error   z value      Pr(>|z|)
## (Intercept) -10.651330614 0.3611573721 -29.49221 3.623124e-191
## balance       0.005498917 0.0002203702  24.95309 1.976602e-137
```

```r
predictor <- as.data.frame(seq(100, 2000, by = 100))
colnames(predictor) <- "balance"


predict(logreg_default, predictor, type = "response")
```

```
##            1            2            3            4            5
## 4.101880e-05 7.108613e-05 1.231905e-04 2.134779e-04 3.699132e-04
##            6            7            8            9           10
## 6.409100e-04 1.110217e-03 1.922514e-03 3.327154e-03 5.752145e-03
##           11           12           13           14           15
## 9.926984e-03 1.707982e-02 2.923441e-02 4.960213e-02 8.294762e-02
##           16           17           18           19           20
## 1.355136e-01 2.136317e-01 3.201070e-01 4.493274e-01 5.857694e-01
```

```r
logreg_default2 <- glm(default ~ student, family = binomial, data = default)
summary(logreg_default2)$coefficients
```

```
##               Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) -3.5041278 0.07071301 -49.554219 0.0000000000
## studentYes   0.4048871 0.11501883   3.520181 0.0004312529
```

```r
#A one-unit increase in studentyes is is associated with an increase in the log odds of defaults by 0.4
logreg_default3 <- glm(default ~ student + balance + income, family = binomial, data = default)
summary(logreg_default3)$coefficients
```

```
##                  Estimate   Std. Error    z value      Pr(>|z|)
## (Intercept) -1.086905e+01 4.922555e-01 -22.080088 4.911280e-108
## studentYes  -6.467758e-01 2.362525e-01  -2.737646  6.188063e-03
## balance      5.736505e-03 2.318945e-04  24.737563 4.219578e-135
## income       3.033450e-06 8.202615e-06   0.369815  7.115203e-01
```

```r
#income is not that significant => it makes sense when I saw the graph at the beginning!!
```

How do we interpret the coefficients? A one-unit increase in balance is associated with an increase in the log odds of default by 0.005 units. Or, odds of being default with every one unit increase in balance are e^0.005 times higher, keeping everything else fixed

This data consists of percentage returns fro the S&P 500 stock index over 1,250 days, from the beginning of 2001 until the end of 2005. For each date, the percentage returns for each of the five previous tradings has been records, Lag1 through Lag5. Other variables are: • Volume = the number of shares traded on the prevous day, in billions • Today = the percentage return on the data in question • Direction = whether the market was Up or Down on this date

Q. How to answer "Are previous day's returns highly correlated with today's returns?" ==> lag1 and today correlation. . .

Q. how to include the similar smooth lines on the scatter plot? ==> get average and connect them!

```r
smarket <- Smarket
names(smarket)
```

```
## [1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"
## [7] "Volume"    "Today"     "Direction"
```

```r
dim(smarket)
```

```
## [1] 1250    9
```

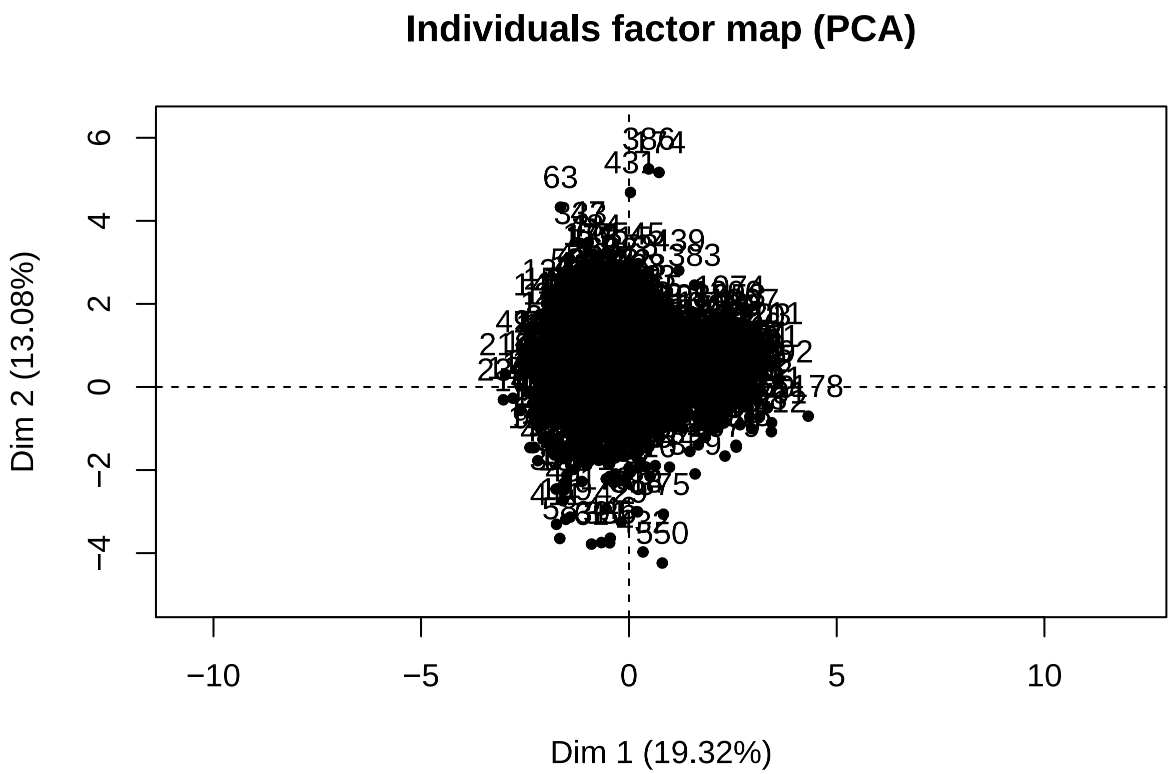
```r
summary(smarket)
```

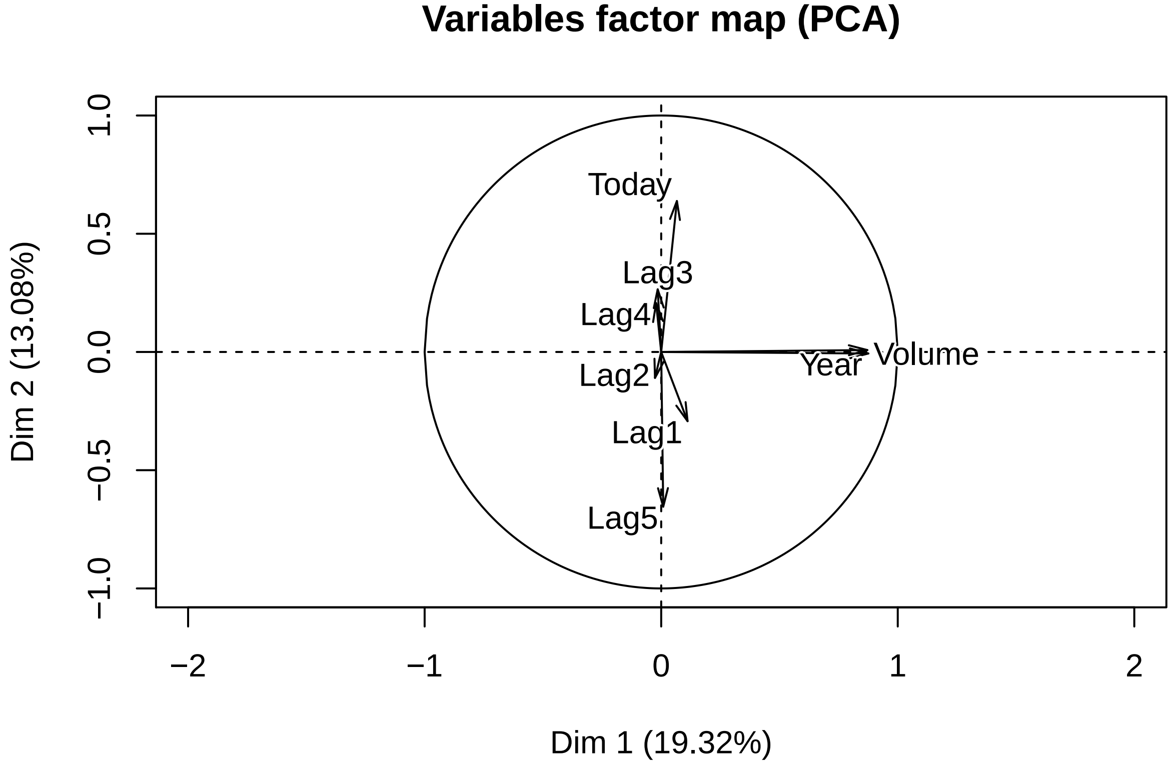
```
##       Year           Lag1                Lag2
##  Min.   :2001   Min.   :-4.922000   Min.   :-4.922000
##  1st Qu.:2002   1st Qu.:-0.639500   1st Qu.:-0.639500
##  Median :2003   Median : 0.039000   Median : 0.039000
##  Mean   :2003   Mean   : 0.003834   Mean   : 0.003919
##  3rd Qu.:2004   3rd Qu.: 0.596750   3rd Qu.: 0.596750
##  Max.   :2005   Max.   : 5.733000   Max.   : 5.733000
##       Lag3                Lag4                Lag5
##  Min.   :-4.922000   Min.   :-4.922000   Min.   :-4.92200
##  1st Qu.:-0.640000   1st Qu.:-0.640000   1st Qu.:-0.64000
##  Median : 0.038500   Median : 0.038500   Median : 0.03850
##  Mean   : 0.001716   Mean   : 0.001636   Mean   : 0.00561
##  3rd Qu.: 0.596750   3rd Qu.: 0.596750   3rd Qu.: 0.59700
##  Max.   : 5.733000   Max.   : 5.733000   Max.   : 5.73300
##      Volume           Today           Direction
##  Min.   :0.3561   Min.   :-4.922000   Down:602
##  1st Qu.:1.2574   1st Qu.:-0.639500   Up  :648
##  Median :1.4229   Median : 0.038500
##  Mean   :1.4783   Mean   : 0.003138
##  3rd Qu.:1.6417   3rd Qu.: 0.596750
##  Max.   :3.1525   Max.   : 5.733000
```

```r
cor(smarket[,-9])
```

```
##              Year         Lag1         Lag2         Lag3         Lag4
## Year   1.00000000  0.029699649  0.030596422  0.033194581  0.035688718
## Lag1   0.02969965  1.000000000 -0.026294328 -0.010803402 -0.002985911
## Lag2   0.03059642 -0.026294328  1.000000000 -0.025896670 -0.010853533
## Lag3   0.03319458 -0.010803402 -0.025896670  1.000000000 -0.024051036
## Lag4   0.03568872 -0.002985911 -0.010853533 -0.024051036  1.000000000
## Lag5   0.02978799 -0.005674606 -0.003557949 -0.018808338 -0.027083641
## Volume 0.53900647  0.040909908 -0.043383215 -0.041823686 -0.048414246
## Today  0.03009523 -0.026155045 -0.010250033 -0.002447647 -0.006899527
##                Lag5      Volume        Today
## Year    0.029787995  0.53900647  0.030095229
## Lag1   -0.005674606  0.04090991 -0.026155045
## Lag2   -0.003557949 -0.04338321 -0.010250033
## Lag3   -0.018808338 -0.04182369 -0.002447647
## Lag4   -0.027083641 -0.04841425 -0.006899527
## Lag5    1.000000000 -0.02200231 -0.034860083
## Volume -0.022002315  1.00000000  0.014591823
## Today  -0.034860083  0.01459182  1.000000000
```

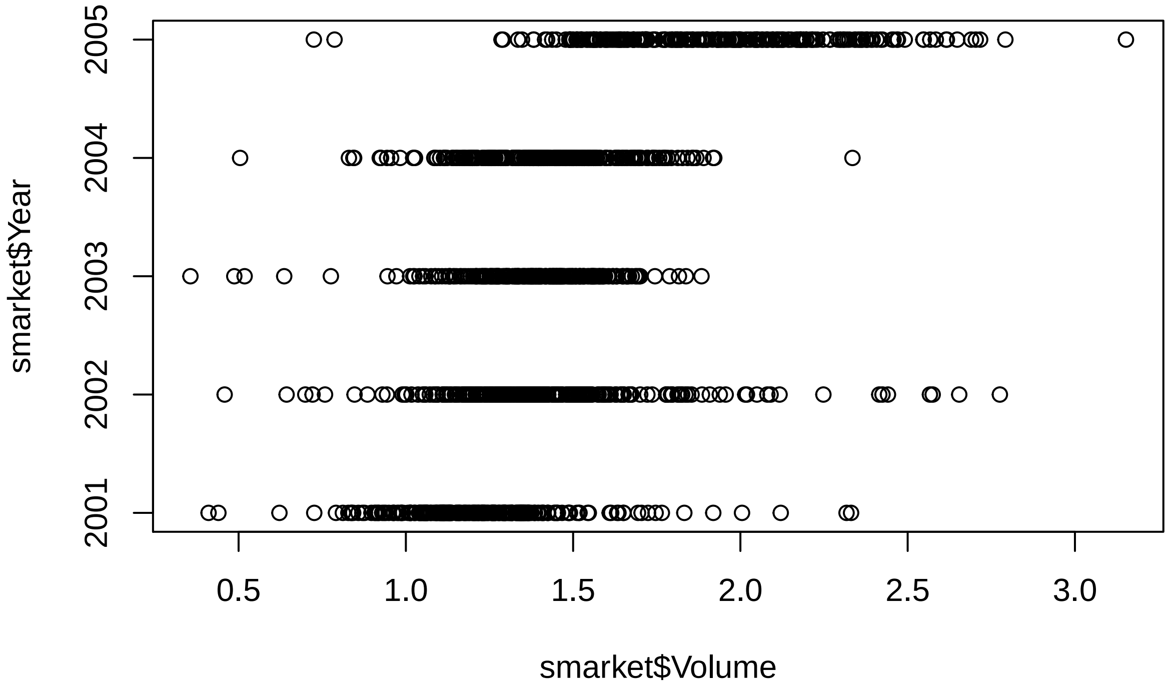
```r
#chart.Correlation(smarket[,-9])
PCA(smarket[,-9])
```


Individuals factor map (PCA)

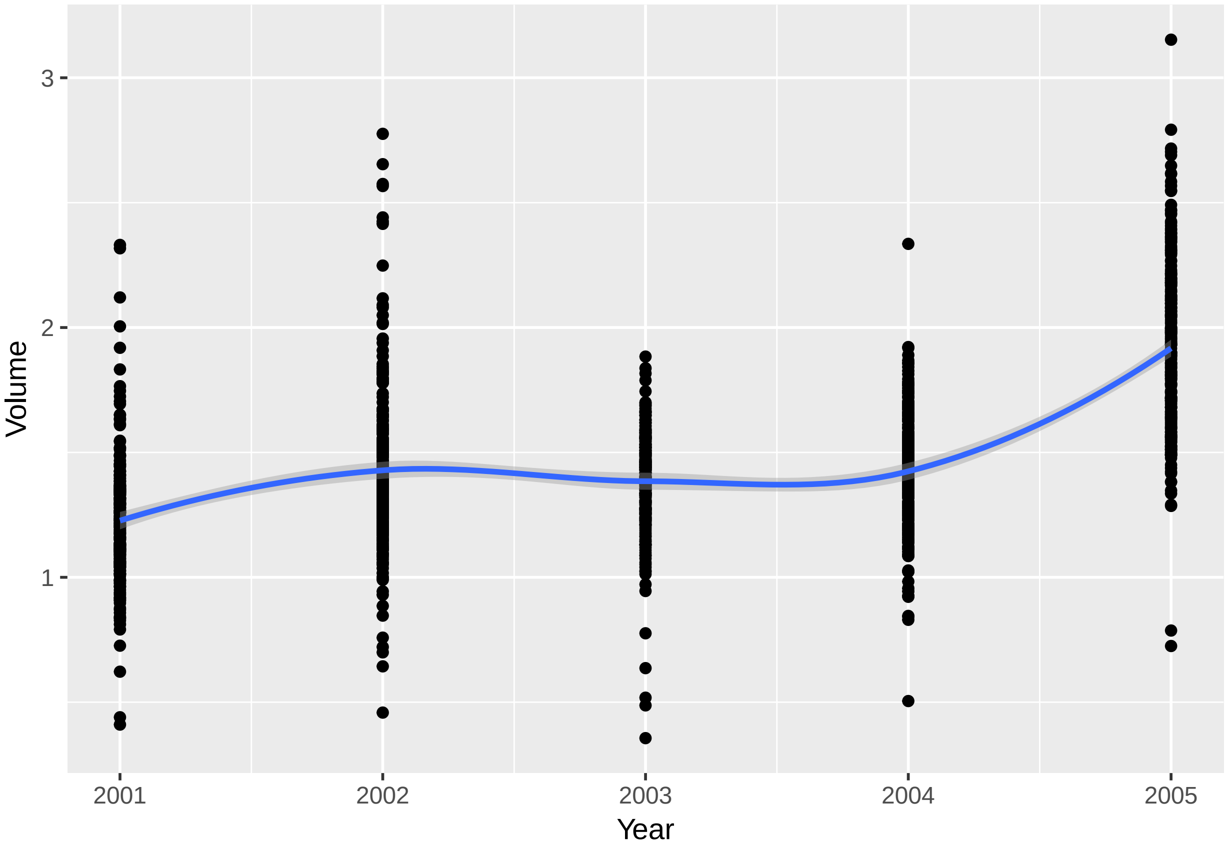

Variables factor map (PCA)

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 1250 individuals, described by 8 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"
```

```
plot(smarket$Volume, smarket$Year)
```



```
ggplot(smarket, aes(x = Year, y = Volume)) + geom_point() +
  geom_smooth(method = loess)
```



From circle of correlations of variables, I can see that lag 3 and 4 are highly correlated with Today’s returns!!!

Q. For question 2 and 3, “No variable is significant. . . .” and “Coef of lag1 is -0.073, and if fixed other variables, lag1 is likely to have different sign as direction. . . .” ??? ==> yes!

Q. So, if predict(smarketlog, type = “response”) gives >0.5, it means it is more likely to be up? ==> right!

```r
smarketlog <- glm(Direction ~., family = binomial, data = smarket[,-c(1, 8)])
summary(smarketlog)$coefficients
```

```
##                 Estimate Std. Error    z value  Pr(>|z|)
## (Intercept) -0.126000257 0.24073574 -0.5233966 0.6006983
## Lag1        -0.073073746 0.05016739 -1.4565986 0.1452272
## Lag2        -0.042301344 0.05008605 -0.8445733 0.3983491
## Lag3         0.011085108 0.04993854  0.2219750 0.8243333
## Lag4         0.009358938 0.04997413  0.1872757 0.8514445
## Lag5         0.010313068 0.04951146  0.2082966 0.8349974
## Volume       0.135440659 0.15835970  0.8552723 0.3924004
```

```r
head(predict(smarketlog, type = "response"), 10) # tells R to output probabilities of the form P(Y = 1|
```

```
##         1         2         3         4         5         6         7
## 0.5070841 0.4814679 0.4811388 0.5152224 0.5107812 0.5069565 0.4926509
##         8         9        10
## 0.5092292 0.5176135 0.4888378
```

Q. What does it mean by “no analytical solution”?? ==> no specific formula, like lambda in ridge and lasso, cuz we get lambda based on CV.

```r
smarket2 <- Smarket
default_numeric <- rep(0, nrow(smarket2))
```

```r
default_numeric[smarket2$Direction == 'Up'] <- 1
smarket2$default_num <- default_numeric

smarket2 <- smarket2[,-c(1, 8, 9)]

# dist(rbind(v1, v2))

# l2 <- function(v1, v2){
#   sqrt(sum((v1 - v2)^2))
# }
# l2(v1, v2)
b <- as.matrix(rep(0, ncol(smarket2)))
bnew <- as.matrix(rep(0, ncol(smarket2)))
k <- 1000
while (k > 0.00001) {
  p <- c()
  w <- diag(nrow(smarket2))
  z <- rep(0, nrow(smarket2))
  x <- as.matrix(cbind(1, smarket2[,-7]))
  b <- bnew
  for(i in 1:nrow(smarket2)){
    exponential <- exp(as.numeric(as.matrix(x[i, ] %*% b)))
    p[i] <- exponential / (1 + exponential)
    w[i, i] <- p[i] * (1 - p[i])
  }
  z <- ((x %*% b) + (solve(w) %*% (smarket2[,7] - p)))
  bnew <- solve(t(x) %*% w %*% x) %*% t(x) %*% w %*% z
  k <- abs( dist(rbind(as.vector(b), as.vector(bnew) ) ) )
}

bnew
```

```
##                 [,1]
## 1      -0.126000259
## Lag1   -0.073073747
## Lag2   -0.042301345
## Lag3    0.011085108
## Lag4    0.009358938
## Lag5    0.010313069
## Volume  0.135440661
```

```r
b2 <- as.matrix(rep(0, ncol(smarket2)))
xtilda <- matrix(0, nrow(smarket2), ncol(smarket2))
x <- as.matrix(cbind(1, smarket2[,-7]))
bnew2 <- as.matrix(rep(0, ncol(smarket2)))

k <- 1000
while(k > 0.00001) {
  b2 <- bnew2
  p2 <- c()
  for (i in 1: nrow(smarket2)){
    exponential <- exp(as.numeric(as.matrix(x[i, ] %*% b2)))
    p2[i] <- exponential / (1 + exponential)
    xtilda[i, ] <- x[i, ] * (p2[i] * (1 - p2[i]))
  }
  bnew2 <- b2 + (solve(t(x) %*% xtilda) %*% t(x) %*% (smarket2[,7] - p2))
  k <- abs(dist(rbind(as.vector(b2), as.vector(bnew2))))
}

bnew2
```

```
##              [,1]
## [1,] -0.126000259
## [2,] -0.073073747
## [3,] -0.042301345
## [4,]  0.011085108
## [5,]  0.009358938
## [6,]  0.010313069
## [7,]  0.135440661
```

```
print("The same!!!:")
```

```
## [1] "The same!!!:"
```

```
summary(glm(default_num ~., data = smarket2, family = binomial))$coefficients[,1]
```

```
##  (Intercept)         Lag1         Lag2         Lag3         Lag4
## -0.126000257 -0.073073746 -0.042301344  0.011085108  0.009358938
##         Lag5       Volume
##  0.010313068  0.135440659
```