Lab 8: Logistic Regression

Donggyun Kim
27008257

3/16/2018

The Default Data Set

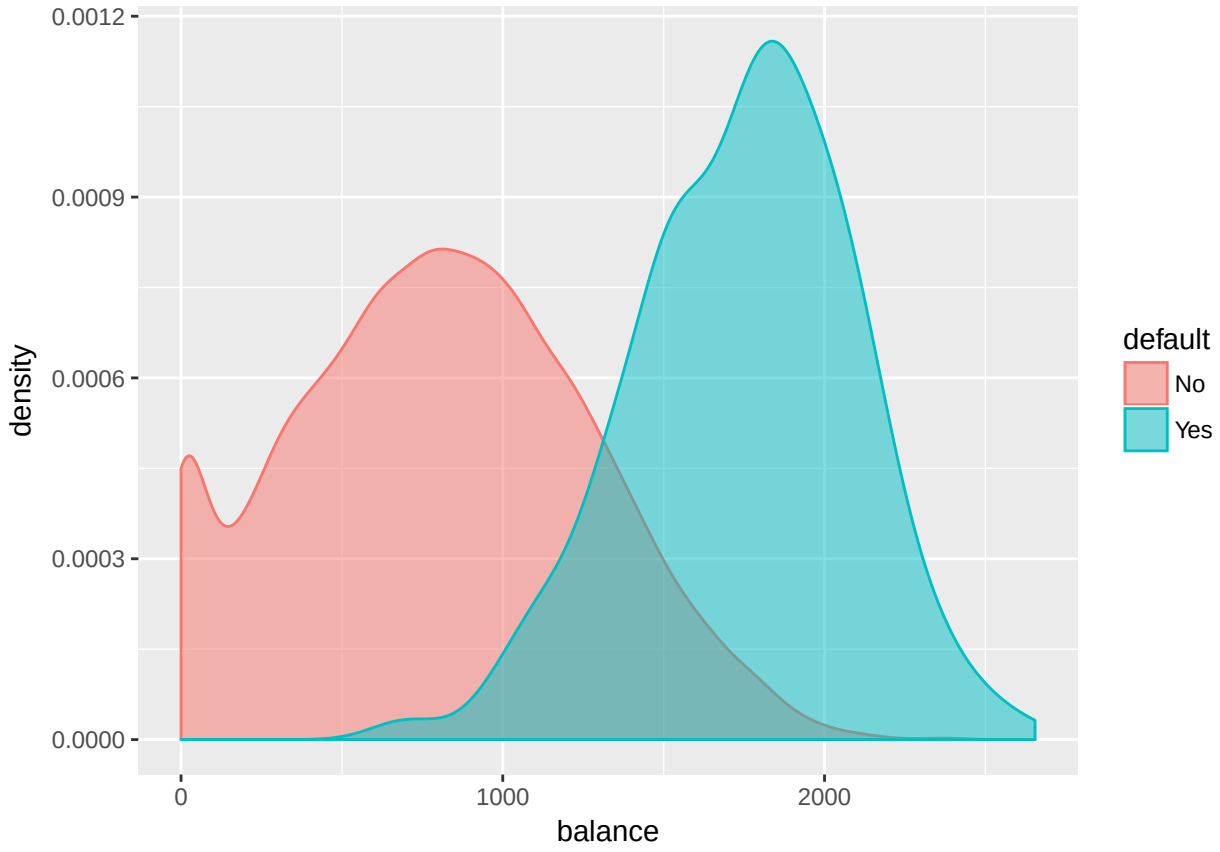
```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.4.2
```

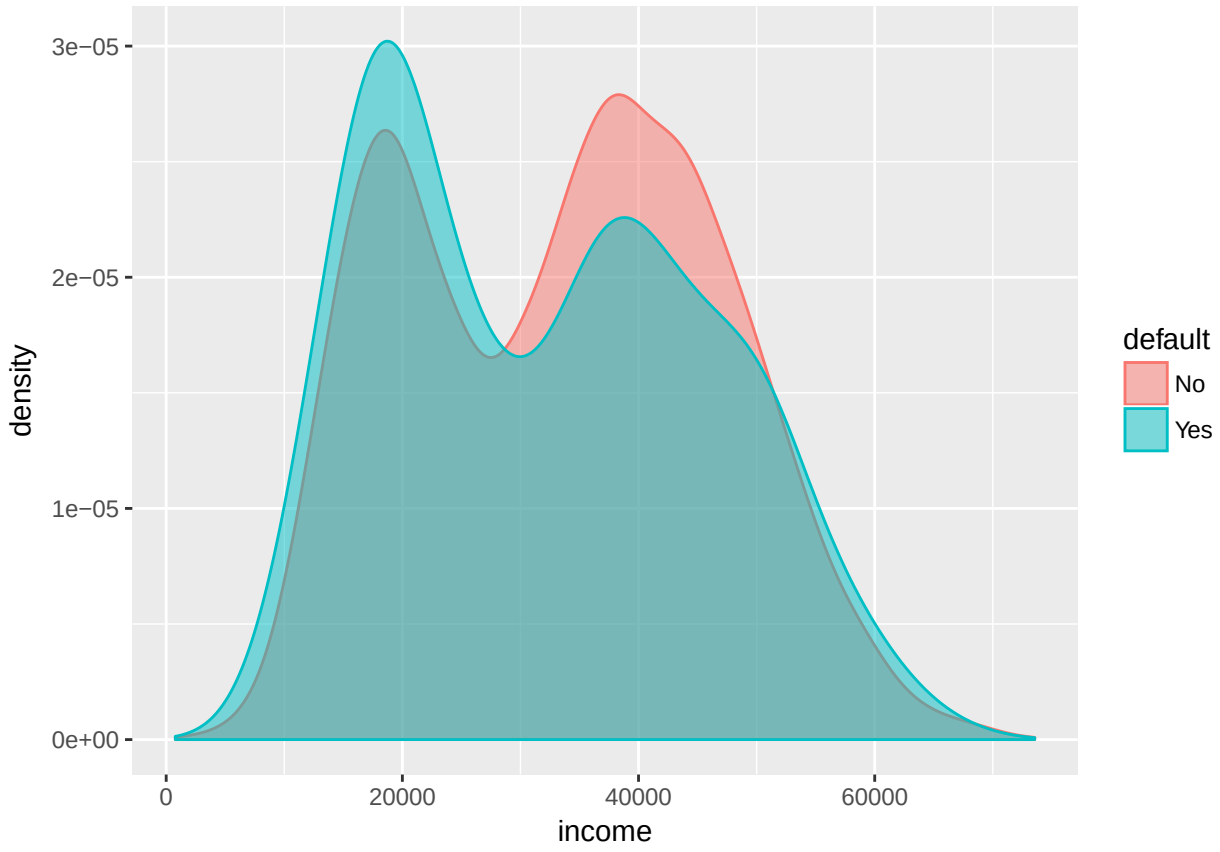
```
library(ggplot2)

ggplot(Default) +
  geom_point(aes(x = balance, y = income, col = default), alpha = 0.5)
```


```
ggplot(Default) +
  geom_density(aes(x = balance, fill = default, col = default), alpha = 0.5)
```




```
ggplot(Default) +
  geom_density(aes(x = income, col = default, fill = default), alpha = 0.5)
```

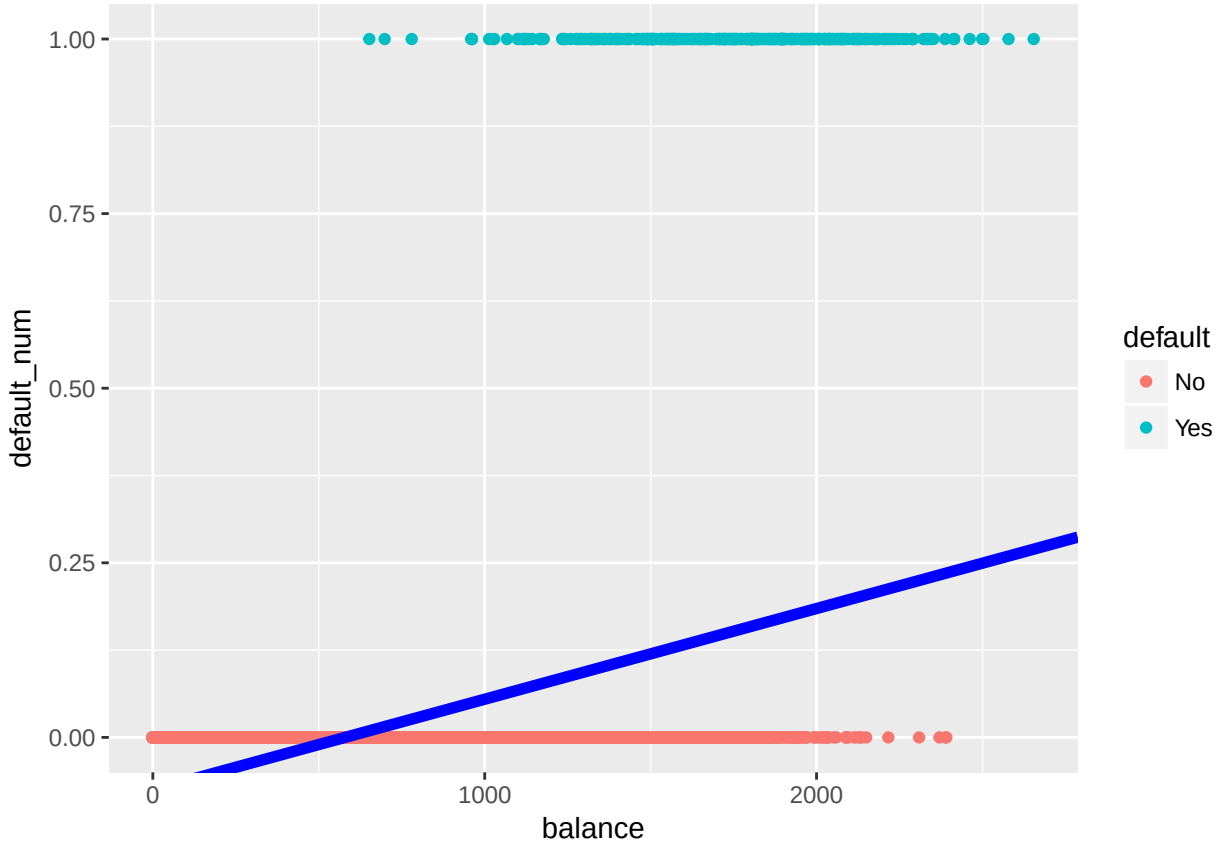



```
default_numeric <- rep(0, nrow(Default))
default_numeric[Default$default == "Yes"] <- 1
Default$default_num <- default_numeric

ols_reg <- lm(default_num ~ balance, 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) +
  geom_point(aes(x = balance, y = default_num, col = default)) +
  geom_abline(slope = ols_reg$coefficients[2],
              intercept = ols_reg$coefficients[1],
              col = "blue", lwd = 2)
```




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

```
##
## Call:
## glm(formula = default ~ balance, family = binomial, data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <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: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

```
new.default <- data.frame(balance = seq(100, 2000, 100))
predict(logreg_default, new.default)
```

```
##           1           2           3           4           5           6
## -10.1014389  -9.5515472  -9.0016555  -8.4517638  -7.9018721  -7.3519805
##           7           8           9          10          11          12
##  -6.8020888  -6.2521971  -5.7023054  -5.1524137  -4.6025220  -4.0526303
##          13          14          15          16          17          18
##  -3.5027386  -2.9528469  -2.4029552  -1.8530635  -1.3031718  -0.7532801
##          19          20
##  -0.2033884   0.3465032
```

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

```
##
## Call:
## glm(formula = default ~ student, family = binomial, data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2970  -0.2970  -0.2434  -0.2434   2.6585
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
## studentYes   0.40489    0.11502    3.52 0.000431 ***
## ---
## 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: 2908.7  on 9998  degrees of freedom
## AIC: 2912.7
##
## Number of Fisher Scoring iterations: 6
```

The Stock Market Smarket Data

```
str(Smarket)
```

```
## 'data.frame':    1250 obs. of  9 variables:
##  $ Year     : num  2001 2001 2001 2001 2001 ...
##  $ Lag1     : num  0.381 0.959 1.032 -0.623 0.614 ...
##  $ Lag2     : num  -0.192 0.381 0.959 1.032 -0.623 ...
##  $ Lag3     : num  -2.624 -0.192 0.381 0.959 1.032 ...
##  $ Lag4     : num  -1.055 -2.624 -0.192 0.381 0.959 ...
##  $ Lag5     : num  5.01 -1.055 -2.624 -0.192 0.381 ...
```

```
##  $ Volume   : num  1.19 1.3 1.41 1.28 1.21 ...
##  $ Today    : num  0.959 1.032 -0.623 0.614 0.213 ...
##  $ Direction: Factor w/ 2 levels "Down","Up": 2 2 1 2 2 2 1 2 2 2 ...
```

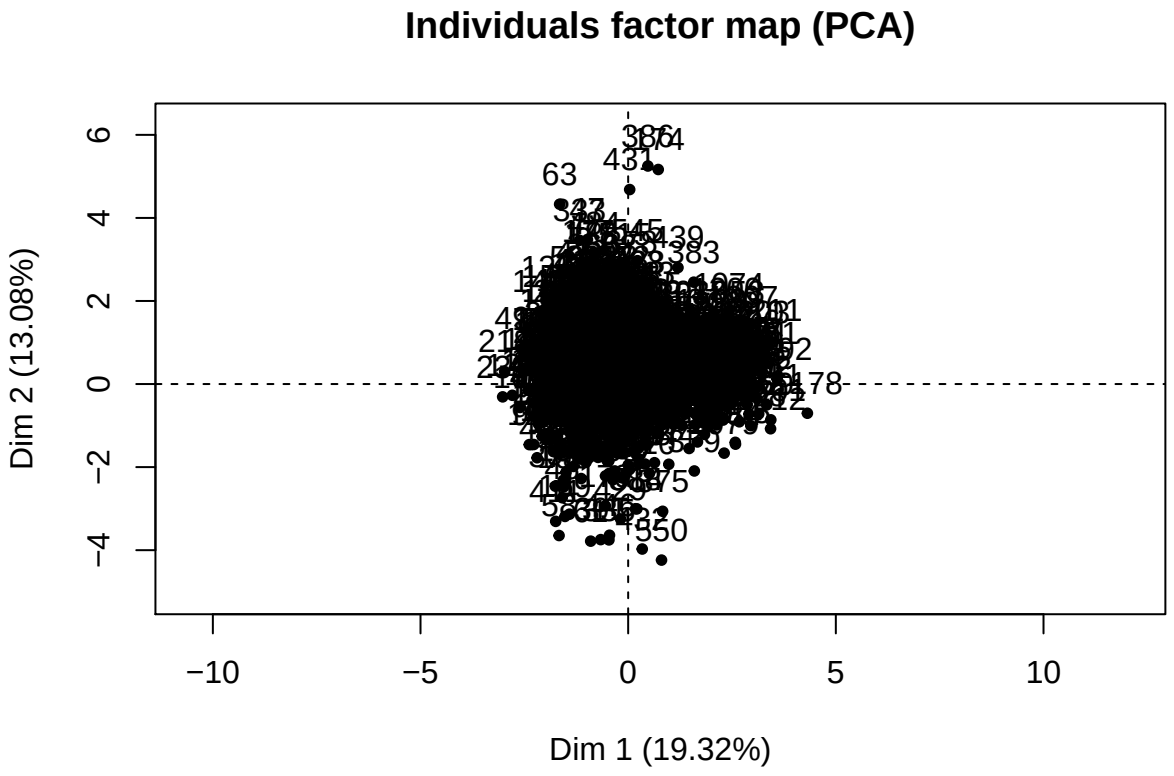
```
X <- as.matrix(Smarket[, -9])
cor(X)
```

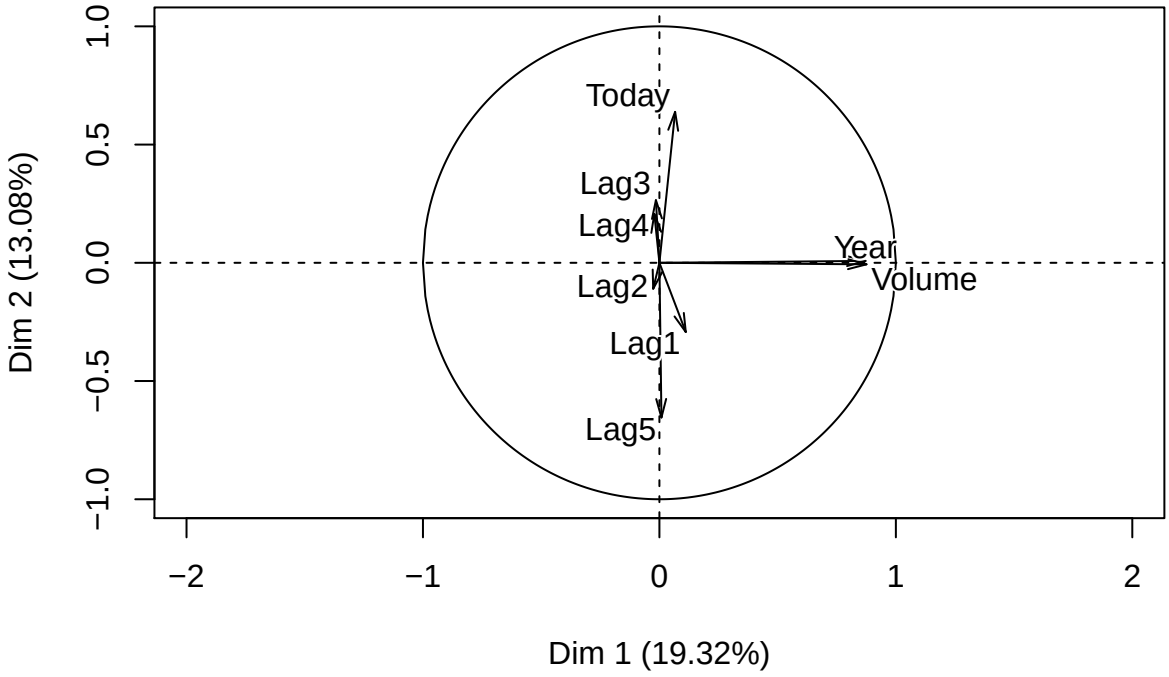
```
##              Year         Lag1         Lag2         Lag3         Lag4
## Year   1.00000000  0.029699649  0.030596422  0.033194581  0.035688718
## Lag1   0.02969965  1.000000000 -0.026294328 -0.010803402 -0.002985911
## Lag2   0.03059642 -0.026294328  1.000000000 -0.025896670 -0.010853533
## Lag3   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
```

```
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 3.4.2
```

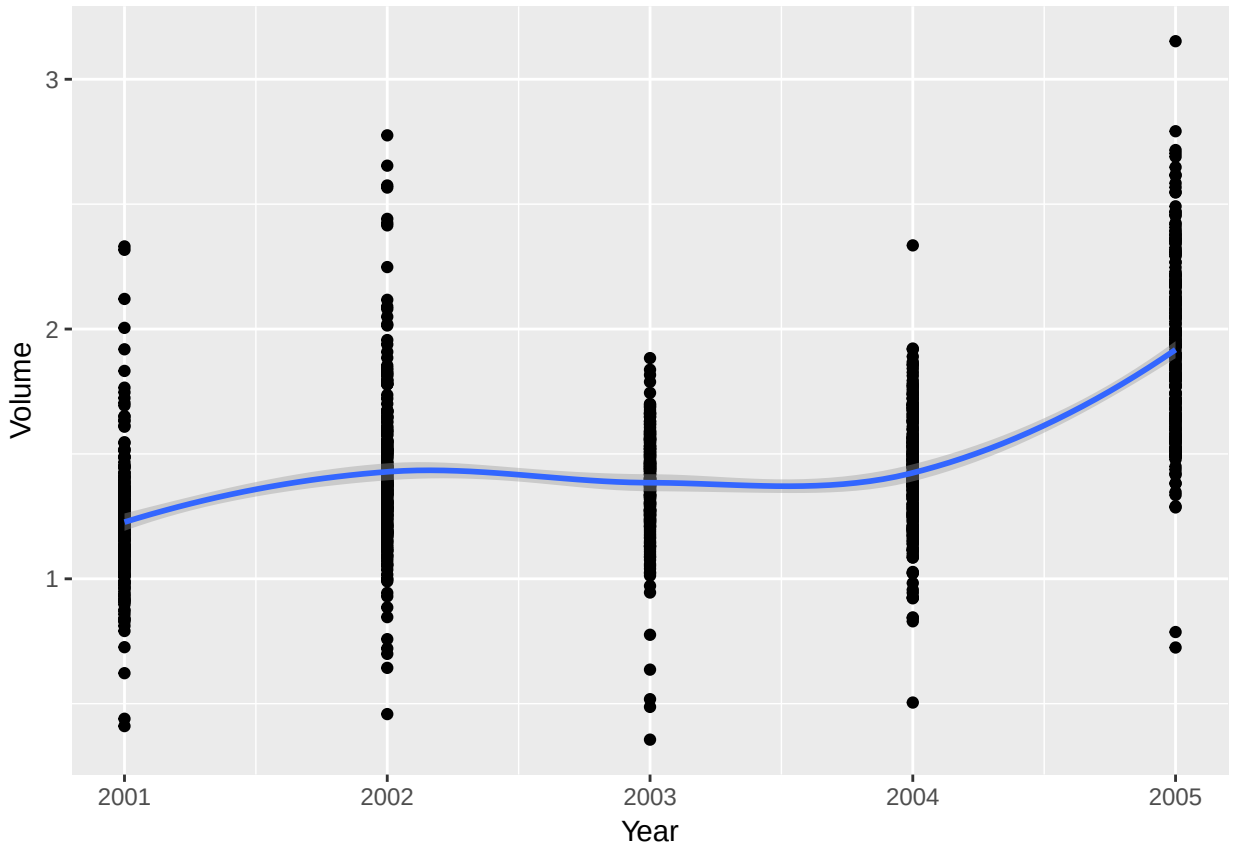
```
PCA(Smarket[, -9])
```






```
## **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"
```

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




```
logreg_smarket <- glm(Direction ~ . - Year - Today,
                      family = binomial, data = Smarket)

summary(logreg_smarket)
```

```
##
## Call:
## glm(formula = Direction ~ . - Year - Today, family = binomial,
##     data = Smarket)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.446  -1.203   1.065   1.145   1.326
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.126000   0.240736  -0.523    0.601
## Lag1        -0.073074   0.050167  -1.457    0.145
## Lag2        -0.042301   0.050086  -0.845    0.398
## Lag3         0.011085   0.049939   0.222    0.824
## Lag4         0.009359   0.049974   0.187    0.851
## Lag5         0.010313   0.049511   0.208    0.835
## Volume       0.135441   0.158360   0.855    0.392
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1731.2  on 1249  degrees of freedom
```

```
## Residual deviance: 1727.6  on 1243  degrees of freedom
## AIC: 1741.6
##
## Number of Fisher Scoring iterations: 3
```

```
head(predict(logreg_smarket, type = "response"), 10)
```

```
##         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
```

Estimation of Parameters

```
y <- numeric(length(Smarket$Direction))
y[Smarket$Direction == "Up"] <- 1
X <- as.matrix(Smarket[c(paste0("Lag", 1:5), "Volume")])
X <- cbind(1, X)
b_old <- rep(0, ncol(X))
N <- 100
j <- 1

while(j <= N) {
  p <- numeric(nrow(X))
  for (i in 1:nrow(X)) {
    p[i] <- exp(crossprod(X[i, ], b_old)) / (1 + exp(crossprod(X[i, ], b_old)))
  }

  W <- diag(nrow(X))
  for (i in 1:nrow(X)) {
    W[i, i] <- p[i] * (1 - p[i])
  }

  z <- X %*% b_old + solve(W) %*% (y - p)

  b_new <- solve(crossprod(X, W) %*% X) %*% crossprod(X, W) %*% z

  if (crossprod(b_new - b_old, b_new - b_old) < 10^-20) {
    break
    } else {
      b_old <- b_new
      j <- j + 1
    }
}

# number of iteration
j
```

```
## [1] 4
```

```
# coefficients
b_new
```

```
##                [,1]
```

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

Simplified Algorithm

```
y <- numeric(length(Smarket$Direction))
y[Smarket$Direction == "Up"] <- 1
X <- as.matrix(Smarket[c(paste0("Lag", 1:5), "Volume")])
X <- cbind(1, X)
b_old <- rep(0, ncol(X))
N <- 100
j <- 1

while (j < N) {
  p <- numeric(nrow(X))
  X_tilda <- matrix(0, nrow = nrow(X), ncol = ncol(X))

  for (i in 1:nrow(X)) {
    p[i] <- exp(crossprod(X[i, ], b_old)) / (1 + exp(crossprod(X[i, ], b_old)))
    X_tilda[i, ] <- p[i] * X[i, ]
  }

  b_new <- b_old + solve(crossprod(X, X_tilda)) %*% crossprod(X, (y - p))

  if (crossprod(b_new - b_old, b_new - b_old) < 10^-20) {
    break
  } else {
    b_old <- b_new
    j <- j + 1
  }
}

# number of iteration
j
```

```
## [1] 36
```

```
# coefficients
b_new
```

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