> setwd("D:/Documents (Louis Booth)/R/Term Project/train.csv")

> train <- read.table("train.csv", sep=",", header=TRUE)

>

> setwd("D:/Documents (Louis Booth)/R")

>

> var.names <- c("Response", "Product\_Info\_4", "Ins\_Age", "Ht", "Wt", "BMI", "Employment\_Info\_1", "Employment\_Info\_4", "Employment\_Info\_6", "Insurance\_History\_5", "Family\_Hist\_2", "Family\_Hist\_3", "Family\_Hist\_4", "Family\_Hist\_5", "Medical\_History\_2", "Medical\_History\_10")

> var.cols <- which(names(train) %in% var.names)

> train <- train[,var.cols]

>

> dim(train)

[1] 59381 16

> head(train)

Product\_Info\_4 Ins\_Age Ht Wt BMI Employment\_Info\_1 Employment\_Info\_4 Employment\_Info\_6

1 0.07692308 0.64179104 0.5818182 0.1485356 0.3230080 0.028 0 NA

2 0.07692308 0.05970149 0.6000000 0.1317992 0.2722877 0.000 0 0.0018

3 0.07692308 0.02985075 0.7454545 0.2887029 0.4287804 0.030 0 0.0300

4 0.48717949 0.16417910 0.6727273 0.2050209 0.3524377 0.042 0 0.2000

5 0.23076923 0.41791045 0.6545455 0.2343096 0.4240456 0.027 0 0.0500

6 0.23076923 0.50746269 0.8363636 0.2991632 0.3648867 0.325 0 1.0000

Insurance\_History\_5 Family\_Hist\_2 Family\_Hist\_3 Family\_Hist\_4 Family\_Hist\_5 Medical\_History\_2

1 0.000666667 NA 0.5980392 NA 0.5267857 112

2 0.000133333 0.1884058 NA 0.08450704 NA 412

3 NA 0.3043478 NA 0.22535211 NA 3

4 NA 0.4202899 NA 0.35211268 NA 350

5 NA 0.4637681 NA 0.40845070 NA 162

6 0.005000000 NA 0.2941176 0.50704225 NA 491

Medical\_History\_10 Response

1 NA 8

2 NA 4

3 NA 8

4 NA 8

5 NA 8

6 NA 8

>

> pdf(file="Lab Week 11-1.pdf")

>

> for (j in 1:ncol(train)) {

+ hist(train[,j], xlab=colnames(train)[j],

+ main=paste("Histogram of", colnames(train)[j]),

+ col="red")

+ }

>

>

>

> cor.train <- cor(train[,c(1:5, 16, 14)])

> round(cor.train,3)

Product\_Info\_4 Ins\_Age Ht Wt BMI Response Medical\_History\_2

Product\_Info\_4 1.000 -0.261 0.142 -0.028 -0.129 0.202 -0.006

Ins\_Age -0.261 1.000 0.008 0.110 0.137 -0.210 -0.016

Ht 0.142 0.008 1.000 0.610 0.123 -0.094 -0.014

Wt -0.028 0.110 0.610 1.000 0.854 -0.351 -0.009

BMI -0.129 0.137 0.123 0.854 1.000 -0.382 -0.001

Response 0.202 -0.210 -0.094 -0.351 -0.382 1.000 0.016

Medical\_History\_2 -0.006 -0.016 -0.014 -0.009 -0.001 0.016 1.000

>

> cor.train[lower.tri(cor.train,diag=TRUE)] = 0

> cor.train

Product\_Info\_4 Ins\_Age Ht Wt BMI Response Medical\_History\_2

Product\_Info\_4 0 -0.261356 0.141779803 -0.02804119 -0.1293686 0.20243403 -0.0063673422

Ins\_Age 0 0.000000 0.008419251 0.11036572 0.1370757 -0.20960984 -0.0160104337

Ht 0 0.000000 0.000000000 0.61042482 0.1231248 -0.09357624 -0.0144035827

Wt 0 0.000000 0.000000000 0.00000000 0.8540833 -0.35139480 -0.0094073865

BMI 0 0.000000 0.000000000 0.00000000 0.0000000 -0.38160075 -0.0006528926

Response 0 0.000000 0.000000000 0.00000000 0.0000000 0.00000000 0.0160355919

Medical\_History\_2 0 0.000000 0.000000000 0.00000000 0.0000000 0.00000000 0.0000000000

>

> cor.train.sorted = sort(abs(cor.train), decreasing=T)

> cor.train.sorted

[1] 0.8540833446 0.6104248206 0.3816007472 0.3513948010 0.2613559844 0.2096098417 0.2024340337 0.1417798026

[9] 0.1370757035 0.1293686411 0.1231247518 0.1103657195 0.0935762445 0.0280411946 0.0160355919 0.0160104337

[17] 0.0144035827 0.0094073865 0.0084192507 0.0063673422 0.0006528926 0.0000000000 0.0000000000 0.0000000000

[25] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000

[33] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000

[41] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000

[49] 0.0000000000

> cor.train.sorted[cor.train.sorted>0]

[1] 0.8540833446 0.6104248206 0.3816007472 0.3513948010 0.2613559844 0.2096098417 0.2024340337 0.1417798026

[9] 0.1370757035 0.1293686411 0.1231247518 0.1103657195 0.0935762445 0.0280411946 0.0160355919 0.0160104337

[17] 0.0144035827 0.0094073865 0.0084192507 0.0063673422 0.0006528926

>

> a <- which(abs(cor.train)==cor.train.sorted[1])

> a

[1] 32

> var.big.cor <- arrayInd(a, dim(cor.train))

> var.big.cor

[,1] [,2]

[1,] 4 5

>

> var.big.cor1 <- c(a%%nrow(cor.train), (a%/%nrow(cor.train))+1)

> var.big.cor1

[1] 4 5

> colnames(cor.train)[var.big.cor]

[1] "Wt" "BMI"

>

> cor.train.sorted[2]

[1] 0.6104248

> a2 <- which(abs(cor.train)==cor.train.sorted[2])

> var.big.cor2 <- arrayInd(a2,dim(cor.train))

> colnames(cor.train)[var.big.cor2]

[1] "Ht" "Wt"

>

> cor.train.sorted[3]

[1] 0.3816007

> a3 <- which(abs(cor.train)==cor.train.sorted[3])

> var.big.cor3 <- arrayInd(a3,dim(cor.train))

> colnames(cor.train)[var.big.cor3]

[1] "BMI" "Response"

>

> var1 <- rep(NA,10)

> var2 <- rep(NA,10)

> high.cor <- rep(NA,10)

>

> high.cor.data <- data.frame(var1, var2, high.cor, row.names=1:10)

> for (i in 1:10) {

+ which(abs(cor.train)==cor.train.sorted[i])

+ a <- which(abs(cor.train)==cor.train.sorted[i])

+ var.big.cor <- arrayInd(a,dim(cor.train))

+ high.cor.data[i,1:2] <- colnames(cor.train)[var.big.cor]

+ high.cor.data[i,3] <- cor.train[var.big.cor]

+ }

>

> high.cor.data

var1 var2 high.cor

1 Wt BMI 0.8540833

2 Ht Wt 0.6104248

3 BMI Response -0.3816007

4 Wt Response -0.3513948

5 Product\_Info\_4 Ins\_Age -0.2613560

6 Ins\_Age Response -0.2096098

7 Product\_Info\_4 Response 0.2024340

8 Product\_Info\_4 Ht 0.1417798

9 Ins\_Age BMI 0.1370757

10 Product\_Info\_4 BMI -0.1293686

>

>

>

> png(file="Lab Week 11-1 pairs1.png")

>

> pairs(train[,c(1:5, 16, 14)])

>

> dev.off()

null device

1

>

> png(file="Lab Week 11-1 pairs2.png")

>

> pairs(~Response+Product\_Info\_4+Ins\_Age+Ht+Wt+BMI+Medical\_History\_2, data=train)

>

> dev.off()

null device

1

>

> pc.train <- prcomp(train[,c(1:5, 16, 14)], scale.=TRUE)

>

> summary(pc.train)

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6 PC7

Standard deviation 1.5379 1.1848 1.0010 0.9244 0.83643 0.81792 0.07727

Proportion of Variance 0.3379 0.2005 0.1432 0.1221 0.09994 0.09557 0.00085

Cumulative Proportion 0.3379 0.5384 0.6816 0.8036 0.90358 0.99915 1.00000

> par(mfrow=c(1,1))

>

> png(file="Lab Week 11-1 PC plot.png")

>

> plot(pc.train, type="l")

>

> dev.off()

RStudioGD

2

>

> png(file="Lab Week 11-1 PC biplot.png")

>

> biplot(pc.train)

>

> dev.off()

RStudioGD

2

>

>

> reg <- lm(Response~Product\_Info\_4+Ins\_Age+Ht+Wt+BMI+Medical\_History\_2, data=train)

> summary(reg)

Call:

lm(formula = Response ~ Product\_Info\_4 + Ins\_Age + Ht + Wt +

BMI + Medical\_History\_2, data = train)

Residuals:

Min 1Q Median 3Q Max

-7.4962 -1.2289 0.7395 1.5965 5.0850

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.595e+00 4.061e-01 21.162 < 2e-16 \*\*\*

Product\_Info\_4 1.180e+00 3.388e-02 34.816 < 2e-16 \*\*\*

Ins\_Age -1.588e+00 4.799e-02 -33.087 < 2e-16 \*\*\*

Ht 7.804e-01 6.025e-01 1.295 0.1952

Wt -5.030e+00 9.583e-01 -5.249 1.54e-07 \*\*\*

BMI -3.896e+00 5.577e-01 -6.985 2.87e-12 \*\*\*

Medical\_History\_2 1.837e-04 5.082e-05 3.615 0.0003 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.211 on 59374 degrees of freedom

Multiple R-squared: 0.1902, Adjusted R-squared: 0.1901

F-statistic: 2324 on 6 and 59374 DF, p-value: < 2.2e-16

> round(reg$coefficients,5)

(Intercept) Product\_Info\_4 Ins\_Age Ht Wt BMI

8.59469 1.17975 -1.58794 0.78037 -5.02971 -3.89572

Medical\_History\_2

0.00018

>

> install.packages("MASS")

Error in install.packages : Updating loaded packages

> library(MASS)

> train$Response <- as.factor(train$Response)

> str(train)

'data.frame': 59381 obs. of 16 variables:

$ Product\_Info\_4 : num 0.0769 0.0769 0.0769 0.4872 0.2308 ...

$ Ins\_Age : num 0.6418 0.0597 0.0299 0.1642 0.4179 ...

$ Ht : num 0.582 0.6 0.745 0.673 0.655 ...

$ Wt : num 0.149 0.132 0.289 0.205 0.234 ...

$ BMI : num 0.323 0.272 0.429 0.352 0.424 ...

$ Employment\_Info\_1 : num 0.028 0 0.03 0.042 0.027 0.325 0.11 0.12 0.165 0.025 ...

$ Employment\_Info\_4 : num 0 0 0 0 0 0 NA 0 0 0 ...

$ Employment\_Info\_6 : num NA 0.0018 0.03 0.2 0.05 1 0.8 1 1 0.05 ...

$ Insurance\_History\_5: num 0.000667 0.000133 NA NA NA ...

$ Family\_Hist\_2 : num NA 0.188 0.304 0.42 0.464 ...

$ Family\_Hist\_3 : num 0.598 NA NA NA NA ...

$ Family\_Hist\_4 : num NA 0.0845 0.2254 0.3521 0.4085 ...

$ Family\_Hist\_5 : num 0.527 NA NA NA NA ...

$ Medical\_History\_2 : int 112 412 3 350 162 491 600 145 16 162 ...

$ Medical\_History\_10 : int NA NA NA NA NA NA NA NA NA NA ...

$ Response : Factor w/ 8 levels "1","2","3","4",..: 8 4 8 8 8 8 8 1 8 1 ...

> logreg <- polr(Response~Product\_Info\_4+Ins\_Age+Ht+Wt+BMI+Medical\_History\_2, data=train, Hess = TRUE)

> summary(logreg)

Call:

polr(formula = Response ~ Product\_Info\_4 + Ins\_Age + Ht + Wt +

BMI + Medical\_History\_2, data = train, Hess = TRUE)

Coefficients:

Value Std. Error t value

Product\_Info\_4 1.0177677 2.887e-02 35.258

Ins\_Age -1.4266012 4.027e-02 -35.422

Ht -3.8187611 7.439e-02 -51.336

Wt 1.7225294 5.613e-02 30.691

BMI -7.3734802 5.917e-02 -124.607

Medical\_History\_2 0.0001697 4.203e-05 4.037

Intercepts:

Value Std. Error t value

1|2 -8.4457 0.0597 -141.4626

2|3 -7.4987 0.0586 -128.0062

3|4 -7.3848 0.0585 -126.3236

4|5 -7.2364 0.0583 -124.0790

5|6 -6.7246 0.0579 -116.1579

6|7 -5.7645 0.0570 -101.1552

7|8 -5.0488 0.0563 -89.7475

Residual Deviance: 198661.21

AIC: 198687.21

> logLik(logreg)

'log Lik.' -99330.61 (df=13)

>

> summary(update(logreg, method = "probit", Hess = TRUE), digits = 3)

Call:

polr(formula = Response ~ Product\_Info\_4 + Ins\_Age + Ht + Wt +

BMI + Medical\_History\_2, data = train, Hess = TRUE, method = "probit")

Coefficients:

Value Std. Error t value

Product\_Info\_4 6.00e-01 1.70e-02 35.35

Ins\_Age -8.76e-01 2.35e-02 -37.24

Ht -1.64e+00 4.47e-02 -36.79

Wt 2.00e-01 3.24e-02 6.16

BMI -3.74e+00 3.53e-02 -105.83

Medical\_History\_2 9.71e-05 2.49e-05 3.90

Intercepts:

Value Std. Error t value

1|2 -4.427 0.034 -128.730

2|3 -3.920 0.034 -115.085

3|4 -3.856 0.034 -113.330

4|5 -3.772 0.034 -110.973

5|6 -3.475 0.034 -102.578

6|7 -2.905 0.034 -86.497

7|8 -2.478 0.033 -74.425

Residual Deviance: 199658.10

AIC: 199684.10

> logLik(update(logreg, method = "probit", Hess = TRUE), digits = 3)

'log Lik.' -99829.05 (df=13)

> summary(update(logreg, method = "loglog", Hess = TRUE), digits = 3)

Call:

polr(formula = Response ~ Product\_Info\_4 + Ins\_Age + Ht + Wt +

BMI + Medical\_History\_2, data = train, Hess = TRUE, method = "loglog")

Coefficients:

Value Std. Error t value

Product\_Info\_4 5.42e-01 0.015440 35.13

Ins\_Age -8.66e-01 0.022483 -38.53

Ht -9.15e-01 0.044686 -20.49

Wt -4.92e-01 0.031047 -15.86

BMI -2.14e+00 0.034444 -62.14

Medical\_History\_2 8.04e-05 0.000024 3.35

Intercepts:

Value Std. Error t value

1|2 -2.867 0.033 -86.671

2|3 -2.461 0.033 -74.643

3|4 -2.406 0.033 -73.008

4|5 -2.333 0.033 -70.779

5|6 -2.063 0.033 -62.558

6|7 -1.491 0.033 -45.203

7|8 -1.008 0.033 -30.536

Residual Deviance: 205314.35

AIC: 205340.35

> logLik(update(logreg, method = "loglog", Hess = TRUE), digits = 3)

'log Lik.' -102657.2 (df=13)

> summary(update(logreg, method = "cloglog", Hess = TRUE), digits = 3)

Call:

polr(formula = Response ~ Product\_Info\_4 + Ins\_Age + Ht + Wt +

BMI + Medical\_History\_2, data = train, Hess = TRUE, method = "cloglog")

Coefficients:

Value Std. Error t value

Product\_Info\_4 0.654630 2.13e-02 30.77

Ins\_Age -0.771910 2.76e-02 -28.02

Ht -2.233053 4.88e-02 -45.77

Wt 0.539531 3.86e-02 13.98

BMI -5.833552 3.97e-02 -147.12

Medical\_History\_2 0.000107 2.77e-05 3.89

Intercepts:

Value Std. Error t value

1|2 -6.761 0.041 -164.545

2|3 -5.915 0.040 -147.988

3|4 -5.817 0.040 -145.932

4|5 -5.690 0.040 -143.203

5|6 -5.265 0.039 -133.844

6|7 -4.529 0.039 -117.457

7|8 -4.036 0.038 -106.369

Residual Deviance: 194727.08

AIC: 194753.08

> logLik(update(logreg, method = "cloglog", Hess = TRUE), digits = 3)

'log Lik.' -97363.54 (df=13)