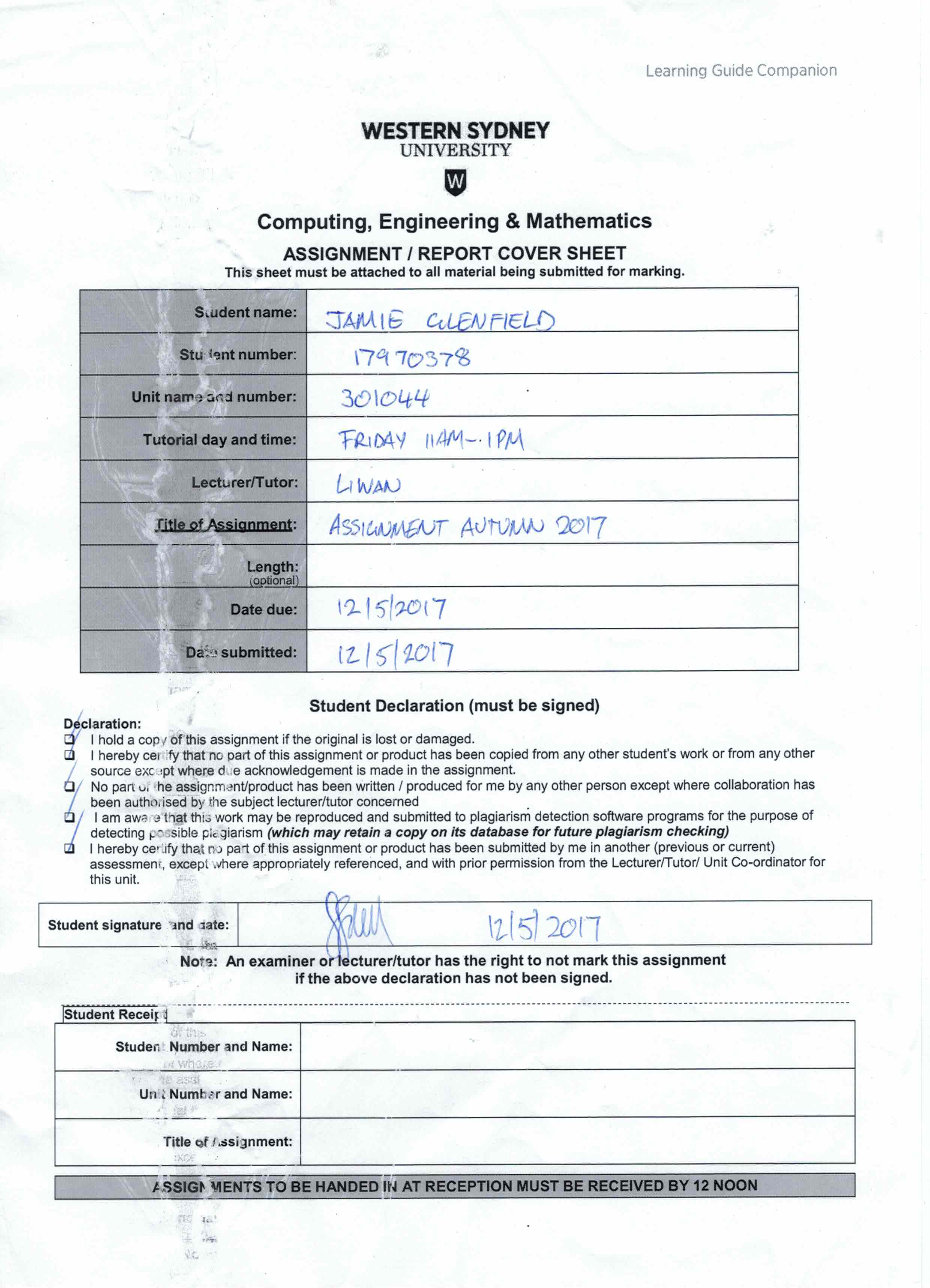
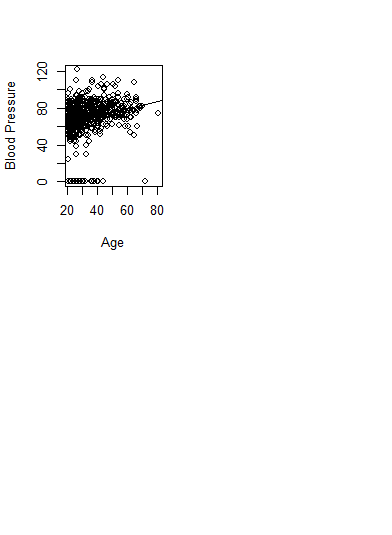
****

**301044 Data Science — Assignment Autumn 2017**

1. **A linear regression analysis of diastolic blood pressure against age including; (marks 5)**



Firstly after plotting the data, it was noted that blood pressure readings of 0 were present. After cleaning the data to replace the zero’s with the median blood pressure, I performed 10 fold cross-validation. This indicated a drop in the estimates test MSE between the linear and quadratic, but no further improvement for higher order polynomials.

The first model I investigated (fit1) was linear, but after further investigation with 10fold cross validation it was found that fit2 with polynomial of the second power produced a more constant variance with residuals vs fitted plot and also returned a higher R2 value, indicating more data is covered by the fit2 model.

setwd("J:/test")

> mydata <- read.csv("diabetes.csv")

> attach(mydata)

> View(mydata)

> mydata$BloodPressure[is.na(mydata$BloodPressure)] <- median(mydata$BloodPressure,na.rm = TRUE)

> View(mydata)

> mydata$BloodPressure[mydata$BloodPressure==0] <- median(mydata$BloodPressure,na.rm = TRUE)

> View(mydata)

> install.packages("boot")

library(boot)

> set.seed (17)

> cv.error.10= rep (0,10)

> for (i in 1:10) {

+ glm.fit=glm(BloodPressure~poly(Age ,i),data=mydata)

+ cv.error.10[i]=cv.glm (mydata ,glm.fit ,K=10) $delta [1]

+ }

> cv.error.10

[1] 131.3337 129.3500 129.5618 129.6266 131.5937

[6] 132.5849 508.6457 610.4992 614.6097 4554.3449

> fit1=lm(BloodPressure~Age,data = mydata)

> summary(fit1)

Call:

lm(formula = BloodPressure ~ Age, data = mydata)

Residuals:

Min 1Q Median 3Q Max

-44.296 -6.972 -0.479 6.362 51.699

Coefficients:

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

(Intercept) 61.27729 1.23925 49.447 <2e-16 \*\*\*

Age 0.33421 0.03515 9.508 <2e-16 \*\*\*

---

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

Residual standard error: 11.45 on 766 degrees of freedom

Multiple R-squared: 0.1056, Adjusted R-squared: 0.1044

F-statistic: 90.41 on 1 and 766 DF, p-value: < 2.2e-16

> fit2<-lm(BloodPressure~Age+I(Age^2),data = mydata)

> summary(fit2)

Call:

lm(formula = BloodPressure ~ Age + I(Age^2), data = mydata)

Residuals:

Min 1Q Median 3Q Max

-43.605 -6.784 -0.627 6.214 51.540

Coefficients:

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

(Intercept) 47.721618 3.782745 12.616 < 2e-16 \*\*\*

Age 1.102234 0.205674 5.359 1.11e-07 \*\*\*

I(Age^2) -0.009633 0.002542 -3.789 0.000163 \*\*\*

---

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

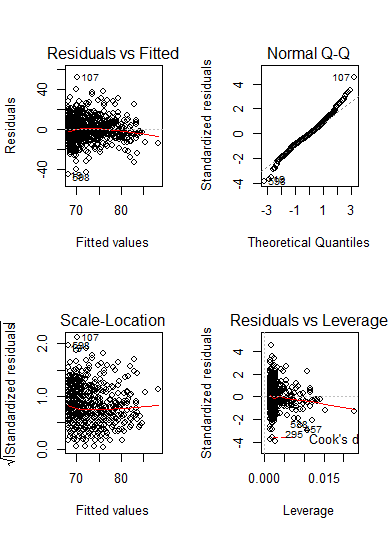
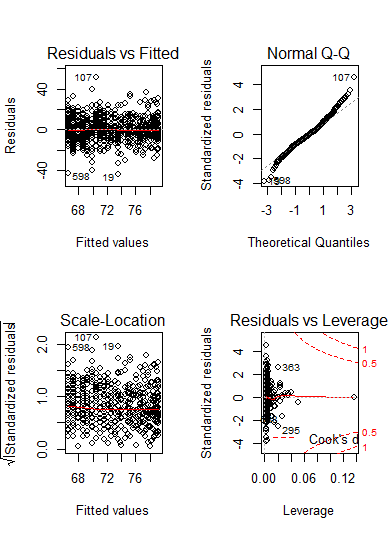
Residual standard error: 11.35 on 765 degrees of freedom

Multiple R-squared: 0.122, Adjusted R-squared: 0.1198

F-statistic: 53.17 on 2 and 765 DF, p-value: < 2.2e-16

1. **a plot of the data,**

**FIT1 FIT2**

** **

The plot for fit1 shows variance decreasing for residuals vs fitted. This is much improved with the fit2 model; Variance for Residuals vs Fitted and Std residuals vs Fitted values has improved constancy in this model.

**b. a discussion of the significance of the slope,**

For the chosen model **fit2** the slope is has a high significance with a very small (almost zero) p-value of 1.11e-7 as seen in the summary below. The three asterisk at the end of the line give a visual indication of high significance. The intercept of the line is 47.741618 and the slope of the line is Age(1.102234) + I(Age2)(-0.009633). This indicates that for every increase of one unit in year plus one unit in (year)2 blood pressure rises by a quadratic factor.

For fit2

> summary(fit2)

Call:

lm(formula = BloodPressure ~ Age + I(Age^2), data = mydata)

Residuals:

Min 1Q Median 3Q Max

-43.605 -6.784 -0.627 6.214 51.540

Coefficients:

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

(Intercept) 47.721618 3.782745 12.616 < 2e-16 \*\*\*

Age 1.102234 0.205674 5.359 1.11e-07 \*\*\*

I(Age^2) -0.009633 0.002542 -3.789 0.000163 \*\*\*

---

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

Residual standard error: 11.35 on 765 degrees of freedom

Multiple R-squared: 0.122, Adjusted R-squared: 0.1198

F-statistic: 53.17 on 2 and 765 DF, p-value: < 2.2e-16

**c. the R2 statistic and**

The first model has a R2value of 0.1056, so for only 10.56% of the data.

Although this can be problematic for making precise predictions, the significant values of the coefficient offer valuable information on the mean change in the response for each unit increase in the predictor.

The second model increases the R2 value to 0.122 encompassing an extra 1.66% of the data

**d. the leave-one-out cross-validation estimate of the mean square error.**

> glm.fit=glm(fit2 ,data=mydata)

> cv.err =cv.glm(mydata ,glm.fit)

> cv.err$delta

[1] 129.2451 129.2445

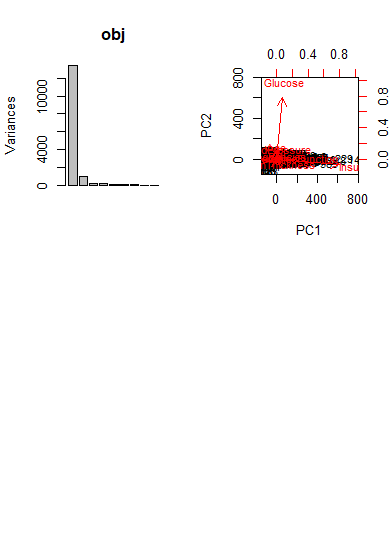
**2. A principal components biplot of the 5 numeric variables without scaling. (marks 5)**

*Rcode*

obj = prcomp(mydata[,1:9])

screeplot(obj)

biplot(obj,scale = 0)



We can see from the Screeplot above that we should use the first two components (PC1 & PC2) for the biplot. Through dimension reduction using PCA we are able to find the dominant dimensions of the dataset, reducing to only 2 dimensions by ignoring all eigenvectors with insignificant eigenvalues.

**3. A hierarchical cluster analysis of the 5 numeric variables using Euclidean distance**

**Cluster analysis, and defining 2 clusters. (include dendrogram) (marks 5)**

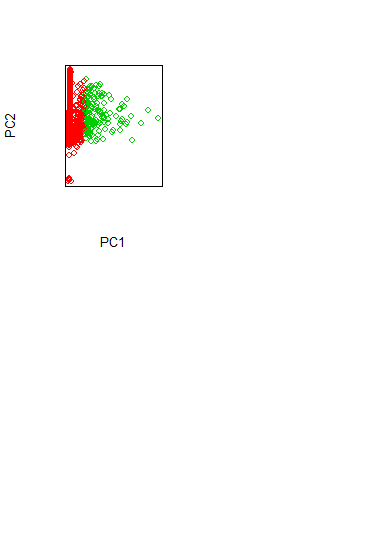
> X = mydata[,1:9]

> km = kmeans(X, centers = 2)

> fitted(km, "classes")

> pp = prcomp(X)

> plot(pp$x[,1:2],col=fitted(km,"classes")+1, xaxt="n", yaxt="n")



table(Outcome=mydata$Outcome, cluster=fitted(km,"classes"))

cluster

Outcome 1 2

0 421 79

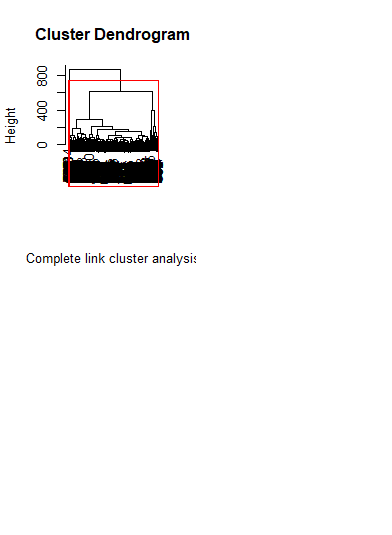
1 182 86

hh = hclust(dist(X),method = "complete")

cutree(hh, k=2)

plot(hh, xlab = "", sub = "Complete link cluster analysis")

rect.hclust(hh, k=2)



4. A logistic regression analysis of the outcome (diabetes) against the other 5 numeric variables

including;(marks 10)

mydata <- read.csv("diabetes.csv")

attach(mydata)

> dim(mydata)

[1] 768 9

#A logistic regression analysis of the outcome against the other 5 numeric variables

> fit3 = glm(Outcome~Pregnancies+Glucose+BloodPressure+BMI+Age,family = "binomial",data = mydata)

a. a table of regression coefficients and p-values

> summary(fit3)

Call:

glm(formula = Outcome ~ Pregnancies + Glucose + BloodPressure +

BMI + Age, family = "binomial", data = mydata)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.1658 -0.7268 -0.4316 0.7620 2.8357

Coefficients:

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

(Intercept) -8.112871 0.757507 -10.710 < 2e-16 \*\*\*

Pregnancies 0.114951 0.031388 3.662 0.00025 \*\*\*

Glucose 0.033465 0.003392 9.866 < 2e-16 \*\*\*

BloodPressure -0.006573 0.008270 -0.795 0.42675

BMI 0.085827 0.014287 6.007 1.89e-09 \*\*\*

Age 0.014660 0.009316 1.574 0.11557

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 993.48 on 767 degrees of freedom

Residual deviance: 741.47 on 762 degrees of freedom

AIC: 753.47

Number of Fisher Scoring iterations: 5

b. the training error rate

> glm.probs = predict(fit3,type = "response")

> glm.probs[1:10]

1 2 3 4 5

0.66222283 0.05469481 0.72692674 0.06101176 0.59673322

6 7 8 9 10

0.18159402 0.07983562 0.46672853 0.83811473 0.05470746

glm.pred = rep("0",768)

> glm.pred[glm.probs>.5]="1"

> table(glm.pred, Outcome)

Outcome

glm.pred 0 1

0 439 117

1 61 151

178/724

[1] 0.2458564

c. the leave-one-out cross-validation estimate of the error rate

> require(boot)

> glm.fit3=fit3

> cv.glm(mydata,glm.fit)$delta

[1] 129.2451 129.2445

d. Does removing the variable diastolic change the error rate (as a percentage to 1 decimal place).

The error rate changes from 129.2451 to 0.1607948, which is 99.9% difference (129.2451-0.1607948)/129.2451)

fit4=glm(Outcome~Pregnancies+Glucose+BMI+Age,family = "binomial",data = mydata)

> glm.fit4=fit4

> cv.glm(mydata,glm.fit4)$delta

[1] 0.1607948 0.1607930