1)

a)

```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/
> log.fit1=glm(default~income+balance,family = binomial)
> summary(log.fit1)

Call:
glm(formula = default ~ income + balance, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4725  -0.1444  -0.0574  -0.0211   3.7245

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
balance      5.647e-03  2.274e-04  24.836  < 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: 1579.0  on 9997  degrees of freedom
AIC: 1585

Number of Fisher Scoring iterations: 8
```

b)

1)

```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/
> str(Default)
'data.frame':   10000 obs. of  4 variables:
 $ default: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ student: Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 2 1 1 ...
 $ balance: num  730 817 1074 529 786 ...
 $ income : num  44362 12106 31767 35704 38463 ...
>
> set.seed(1)
> train=sample(10000,5000)
```

2)

```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/
> log.fit=glm(default~income+balance,family = binomial,subset=train)
> summary(log.fit)

Call:
glm(formula = default ~ income + balance, family = binomial,
    subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3583  -0.1268  -0.0475  -0.0165   3.8116

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.208e+01  6.658e-01 -18.148   <2e-16 ***
income       1.858e-05  7.573e-06   2.454   0.0141 *
balance      6.053e-03  3.467e-04  17.457   <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: 1457.0  on 4999  degrees of freedom
Residual deviance:  734.4  on 4997  degrees of freedom
AIC: 740.4

Number of Fisher Scoring iterations: 8

>
```

3)

```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/

> log.prob=predict(log.fit,Default[-train,])
> log.pred=rep("No",length(log.prob))
> log.pred[log.prob>0.5]="Yes"
> default.test=default[-train]
> table(log.pred,default.test)
        default.test
log.pred   No  Yes
     No  4815  124
     Yes   18   43
```

4)

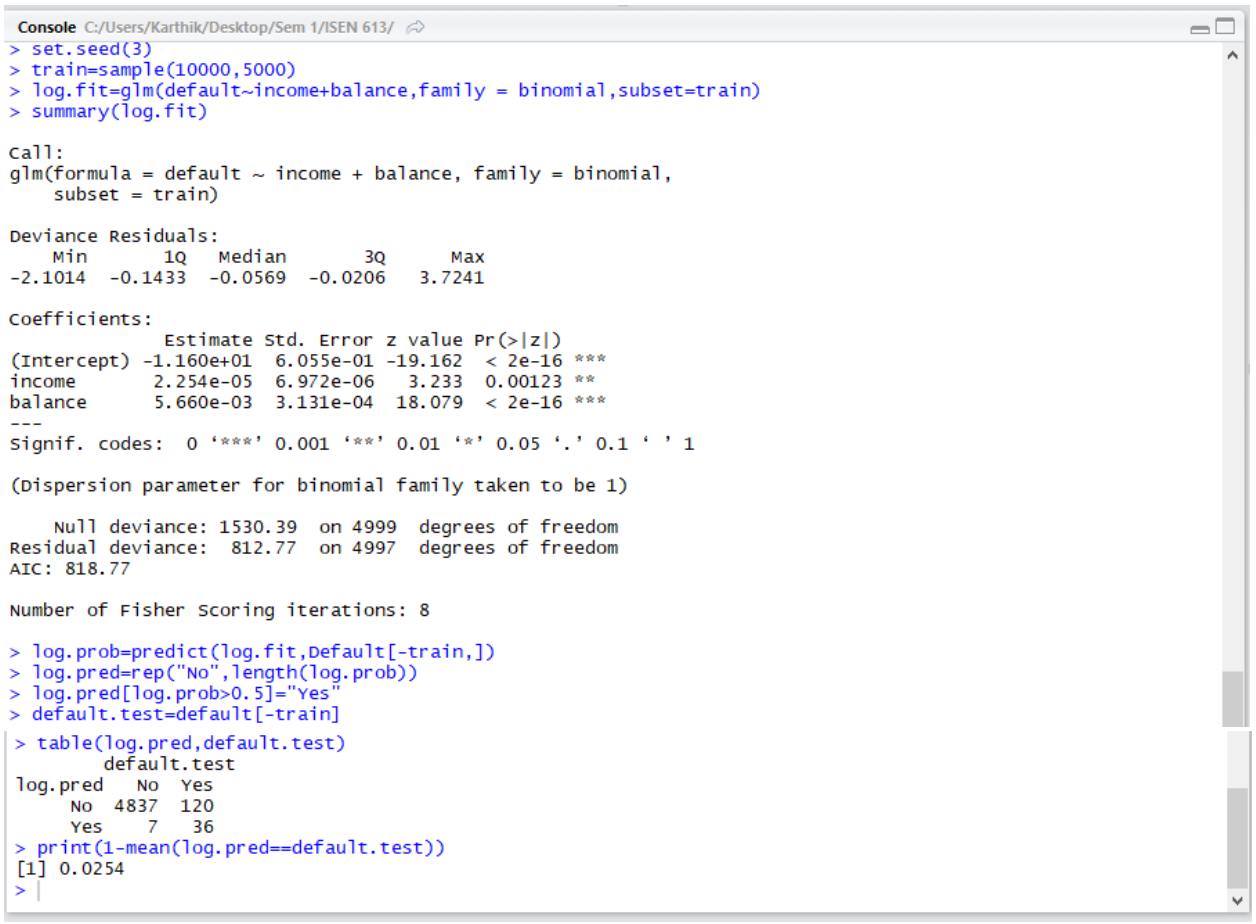
```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/

> print(1-mean(log.pred==default.test))
[1] 0.0284
>
```

Validation set Error: 2.84%

c)

```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/

> set.seed(2)
> train=sample(10000,5000)
> log.fit=glm(default~income+balance,family = binomial,subset=train)
> summary(log.fit)

Call:
glm(formula = default ~ income + balance, family = binomial,
    subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2043  -0.1385  -0.0552  -0.0203   3.7058

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.184e+01  6.403e-01 -18.492  < 2e-16 ***
income       2.717e-05  7.183e-06   3.783 0.000155 ***
balance      5.703e-03  3.266e-04  17.460  < 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: 1443.44  on 4999  degrees of freedom
Residual deviance:  776.64  on 4997  degrees of freedom
AIC: 782.64

Number of Fisher Scoring iterations: 8

> log.prob=predict(log.fit,Default[-train,])
> log.pred=rep("No",length(log.prob))
> log.pred[log.prob>0.5]="Yes"
> default.test=default[-train]
> table(log.pred,default.test)
        default.test
log.pred   No  Yes
     No  4822  130
     Yes    9   39
> print(1-mean(log.pred==default.test))
[1] 0.0278
>
```

Validation error: 2.78%



Validation error: 2.54%

```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/
> set.seed(4)
> train=sample(10000,5000)
> log.fit=glm(default~income+balance,family = binomial,subset=train)
> summary(log.fit)

Call:
glm(formula = default ~ income + balance, family = binomial,
    subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4799  -0.1411  -0.0559  -0.0208   3.7223

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.156e+01  6.146e-01 -18.803  < 2e-16 ***
income       2.004e-05  6.997e-06   2.864  0.00418 **
balance      5.678e-03  3.244e-04  17.502  < 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: 1450.21  on 4999  degrees of freedom
Residual deviance:  780.49  on 4997  degrees of freedom
AIC: 786.49

Number of Fisher Scoring iterations: 8

> log.prob=predict(log.fit,Default[-train,])
> log.pred=rep("No",length(log.prob))
> log.pred[log.prob>0.5]="Yes"
> default.test=default[-train]
> table(log.pred,default.test)
        default.test
log.pred   No  Yes
     No  4823  129
     Yes    9   39
> print(1-mean(log.pred==default.test))
[1] 0.0276
```

Validation error: 2.76%

Each of the 3 times, the results are different as the seed or the randomness of the split of the data for validating the data is different each time.

d)

```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/
> log.fit2=glm(default~income+balance+student,family = binomial)
> summary(log.fit2)

Call:
glm(formula = default ~ income + balance + student, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
---
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: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/
> set.seed(1)
> train=sample(10000,5000)
> log.fit=glm(default~income+balance+student,family = binomial,subset=train)
> summary(log.fit)

Call:
glm(formula = default ~ income + balance + student, family = binomial,
    subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2905  -0.1260  -0.0465  -0.0161   3.7715

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.147e+01  7.562e-01 -15.164   <2e-16 ***
income       2.433e-06  1.256e-05   0.194    0.846
balance      6.124e-03  3.525e-04  17.373   <2e-16 ***
studentYes  -5.608e-01  3.473e-01  -1.615    0.106
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1456.95  on 4999  degrees of freedom
Residual deviance:  731.81  on 4996  degrees of freedom
AIC: 739.81

Number of Fisher Scoring iterations: 8

> log.prob=predict(log.fit,Default[-train,])
> log.pred=rep("No",length(log.prob))
> log.pred[log.prob>0.5]="Yes"
> default.test=default[-train]
> table(log.pred,default.test)
        default.test
log.pred   No  Yes
     No  4817  126
     Yes   16   41
> print(1-mean(log.pred==default.test))
[1] 0.0284
>
```

Validation error:2.84%

Comparing the results of seed(1) of the logistic model, before and after adding the 'student' variable, the test error rate does not change much.

2)

a)

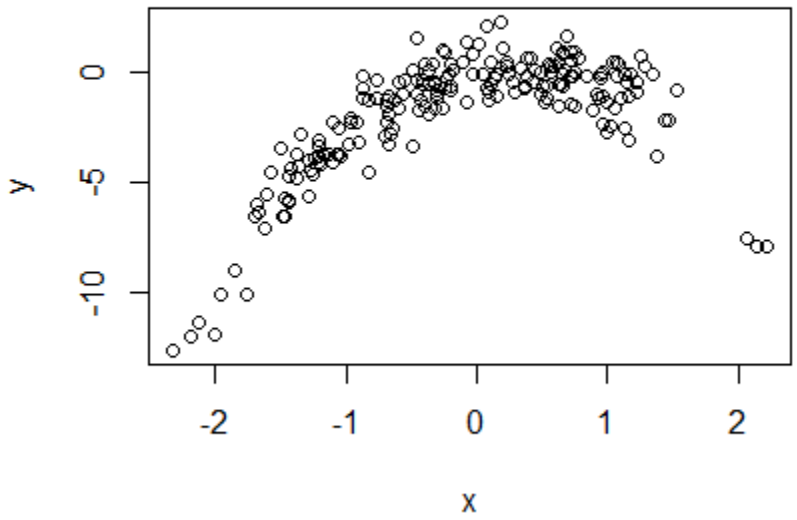
```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/
> set.seed(10)
> x=rnorm(200)
> y=x-2*x^2+rnorm(200)
```

N=200, p=1

True model: Y= X-2*X^2 + E

E:error term

b)



Y seems to have a quadratic relationship with a variation around the fitted curve with an error (which belongs to the one generated from rnorm function.) Also the x axis has a value that has a mean of 0 and +/-3 standard deviations with many corresponding y values towards the centre as many x values will be generated as near to 0 as possible and lesser y values as one moves away from the mean 0.

c)

```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/
> reg.fit=lm(y~x,data=df)
> summary(reg.fit)

Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8679 -0.6873  0.3621  1.3118  3.9028

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6485     0.1585 -10.399   <2e-16 ***
x             1.6295     0.1655   9.846   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.226 on 198 degrees of freedom
Multiple R-squared:  0.3287,    Adjusted R-squared:  0.3253
F-statistic: 96.94 on 1 and 198 DF,  p-value: < 2.2e-16

> set.seed(12)
>
> cv.error=rep(0,4)
> for (i in 1:4)
+ {
+   reg.fit=glm(y~poly(x,i),data=df)
+   summary(reg.fit)
+   cv.error[i]=cv.glm(df,reg.fit)$delta[1]
+ }
> cv.error
[1] 5.115157 1.045824 1.059718 1.019268
```

Error for order 1: 5.1151
Error for order 2: 1.0458
Error for order 3: 1.0597
Error for order 4: 1.0193

d)

```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/
> set.seed(13)
> cv.err=cv.glm(df,reg.fit)
> cv.err$delta
[1] 1.019268 1.019108
>
> cv.error=rep(0,4)
> for (i in 1:4)
+ {
+   reg.fit=glm(y~poly(x,i),data=df)
+   cv.error[i]=cv.glm(df,reg.fit)$delta[1]
+ }
> cv.error
[1] 5.115157 1.045824 1.059718 1.019268
```

Error for order 1: 5.1151
Error for order 2: 1.0458
Error for order 3: 1.0597
Error for order 4: 1.0193

The results obtained in LOOCV using a different seed is same as the previous result with a different seed as there is no randomness as the data is split into 'n-1' and '1' and all possibilities are covered any way the data is split.

e)

Model 2 ($Y = \beta 0 + \beta 1X + \beta 2X^2 + \epsilon$) has the smallest test error which is expected as the relation between y and x is quadratic.

f)

```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/
> set.seed(12)
> cv.error.5=rep(0,4)
> for(i in 1:4)
+ {glm.fit=glm(y~poly(x,i),data=df)
+   cv.error.5[i]=cv.glm(df,glm.fit,K=5)$delta[1]
+ }
> cv.error.5
[1] 5.232012 1.044780 1.068223 1.045473
```

Error for order 1: 5.232
Error for order 2: 1.044
Error for order 3: 1.0682
Error for order 4: 1.045

Model 2 ($Y = \beta 0 + \beta 1X + \beta 2X^2 + \epsilon$) has the smallest test error. It is consistent with the LOOCV method, in the sense that even in LOOCV model, Model 2 had the least test error, but the individual test error values, vary a bit for the corresponding models in LOOCV and K-Fold method.

g)

```
Console C:/Users/Karthik/Desktop/Sem 1/ISEN 613/
> set.seed(12)
> cv.error.10=rep(0,4)
> for(i in 1:4)
+ {glm.fit=glm(y~poly(x,i),data=df)
+ cv.error.10[i]=cv.glm(df,glm.fit,K=10)$delta[1]
+ }
> cv.error.10
[1] 5.086932 1.040679 1.056471 1.050927
>
```

Error for order 1: 5.0869
Error for order 2: 1.0407
Error for order 3: 1.0564
Error for order 4: 1.0509

The individual test errors are different for 5 Fold and 10 Fold Methods, but the conclusion that Model 2 ($Y = \beta 0 + \beta 1X + \beta 2X^2 + \epsilon$) has the smallest test error is common and same for both methods.