snewt
ISYE6501x HW#4
Due June 14 2018

Question 9.1
PCA of uscrime.txt

```
> library(stats)
> pca1<- prcomp(crime, scale. = TRUE)
> head(pca1)
$`sdev`
 [1] 2.49443367 1.71114001 1.42083523 1.19585483 1.06341246 0.75086767 0.60237227
 [8] 0.55502694 0.49243978 0.47036049 0.43856093 0.41777035 0.29147362 0.26063133
[15] 0.21812568 0.06584351

$center
           M           So           Ed          Po1          Po2           LF 
1.385745e+01 3.404255e-01 1.056383e+01 8.500000e+00 8.023404e+00 5.611915e-01 
         M.F          Pop           NW           U1           U2       Wealth 
9.830213e+01 3.661702e+01 1.011277e+01 9.546809e-02 3.397872e+00 5.253830e+03 
        Ineq         Prob         Time        Crime 
1.940000e+01 4.709138e-02 2.659792e+01 9.050851e+02 

$scale
           M           So           Ed          Po1          Po2           LF 
  1.25676339   0.47897516   1.11869985   2.97189736   2.79613186   0.04041181 
         M.F          Pop           NW           U1           U2       Wealth 
  2.94673654  38.07118801  10.28288187   0.01802878   0.84454499 964.90944200 
        Ineq         Prob         Time        Crime 
  3.98960606   0.02273697   7.08689519 386.76269715 
```

snewt
ISYE6501x HW#4
Due June 14 2018

Build regression from this

```
> lmpca1 <- lm(crime$Crime~pca1[,1]+pca1[,2])
Error in pca1[, 1] : incorrect number of dimensions
> lmpca1 <- lm(crime$Crime~pca1$x[,1]+pca1$x[,2])
> summary(lmpca1)

Call:
lm(formula = crime$Crime ~ pca1$x[, 1] + pca1$x[, 2])

Residuals:
    Min      1Q  Median      3Q     Max 
-602.51 -157.14   14.85  141.47  792.87 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   905.09      44.40  20.386  < 2e-16 ***
pca1$x[, 1]    75.89      17.99   4.218 0.000121 ***
pca1$x[, 2]   -92.65      26.23  -3.533 0.000980 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 304.4 on 44 degrees of freedom
Multiple R-squared:  0.4076,	Adjusted R-squared:  0.3807 
F-statistic: 15.14 on 2 and 44 DF,  p-value: 9.945e-06
```

That's not a very good R2 value… Let's try a couple more PC's

```
> lmpca4 <- lm(crime$Crime~pca1$x[,1]+pca1$x[,2]+pca1$x[,3]+pca1$x[,4]+pca1$x[,5])
> summary(lmpca4)

Call:
lm(formula = crime$Crime ~ pca1$x[, 1] + pca1$x[, 2] + pca1$x[,
    3] + pca1$x[, 4] + pca1$x[, 5])

Residuals:
     Min       1Q   Median       3Q      Max
-305.496  -89.435    6.064   73.323  281.078

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  905.085     20.610  43.916  < 2e-16 ***
pca1$x[, 1]   75.891      8.352   9.087 2.25e-11 ***
pca1$x[, 2]  -92.650     12.175  -7.610 2.30e-09 ***
pca1$x[, 3]   40.535     14.662   2.765   0.0085 **
pca1$x[, 4] -212.374     17.420 -12.191 3.22e-15 ***
pca1$x[, 5]   51.545     19.590   2.631   0.0119 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 141.3 on 41 degrees of freedom
Multiple R-squared:  0.881,	Adjusted R-squared:  0.8665
F-statistic: 60.74 on 5 and 41 DF,  p-value: < 2.2e-16
```

But we actually requested that the regression be in terms of the original variables, not of the principle components. I'm really not sure how to do that, currently, so we're going to leave it, and come back if we have time.

Question 10.1.a
Regression tree model of uscrime.txt
Following the lead of: https://www.statmethods.net/advstats/cart.html

```
> library(tree)
> library(rpart)
> list.files(getwd())
[1] "10.3germancreditSummer2018.txt" "9.1uscrimeSummer2018.txt"
> crime<-read.delim("9.1uscrimeSummer2018.txt",header = TRUE)

> crimetree <- rpart(crime$Crime~., data = crime[1:15],method="anova")
```

Then you need to summarize the crimetree you just built:

snewt
ISYE6501x HW#4
Due June 14 2018

```
> summary(crimetree)
Call:
rpart(formula = crime$Crime ~ ., data = crime[1:15], method = "anova")
  n= 47

          CP nsplit rel error    xerror      xstd
1 0.36296293      0 1.0000000 1.0271566 0.2580187
2 0.14814320      1 0.6370371 0.8766895 0.2076665
3 0.05173165      2 0.4888939 1.1602041 0.2589807
4 0.01000000      3 0.4371622 1.1158500 0.2612314

Variable importance
   Po1    Po2 Wealth   Ineq   Prob      M     NW    Pop   Time     Ed     LF     So
    17     17     11     11     10     10      9      5      4      4      1      1
```

Let's describe the first node first:

```
Node number 1: 47 observations,    complexity param=0.3629629
  mean=905.0851, MSE=146402.7
  left son=2 (23 obs) right son=3 (24 obs)
  Primary splits:
      Po1    < 7.65      to the left,  improve=0.3629629, (0 missing)
      Po2    < 7.2       to the left,  improve=0.3629629, (0 missing)
      Prob   < 0.0418485 to the right, improve=0.3217700, (0 missing)
      NW     < 7.65      to the left,  improve=0.2356621, (0 missing)
      Wealth < 6240      to the left,  improve=0.2002403, (0 missing)
  Surrogate splits:
      Po2    < 7.2       to the left,  agree=1.000, adj=1.000, (0 split)
      Wealth < 5330      to the left,  agree=0.830, adj=0.652, (0 split)
      Prob   < 0.043598  to the right, agree=0.809, adj=0.609, (0 split)
      M      < 13.25     to the right, agree=0.745, adj=0.478, (0 split)
      Ineq   < 17.15     to the right, agree=0.745, adj=0.478, (0 split)
```

Then the second:

```
Node number 2: 23 observations,    complexity param=0.05173165
  mean=669.6087, MSE=33880.15
  left son=4 (12 obs) right son=5 (11 obs)
  Primary splits:
      Pop < 22.5      to the left,  improve=0.4568043, (0 missing)
      M   < 14.5      to the left,  improve=0.3931567, (0 missing)
      NW  < 5.4       to the left,  improve=0.3184074, (0 missing)
      Po1 < 5.75      to the left,  improve=0.2310098, (0 missing)
      U1  < 0.093     to the right, improve=0.2119062, (0 missing)
  Surrogate splits:
      NW   < 5.4       to the left,  agree=0.826, adj=0.636, (0 split)
      M    < 14.5      to the left,  agree=0.783, adj=0.545, (0 split)
      Time < 22.30055  to the left,  agree=0.783, adj=0.545, (0 split)
      So   < 0.5       to the left,  agree=0.739, adj=0.455, (0 split)
      Ed   < 10.85     to the right, agree=0.739, adj=0.455, (0 split)
```

Then the third:

```
Node number 3: 24 observations,    complexity param=0.1481432
  mean=1130.75, MSE=150173.4
  left son=6 (10 obs) right son=7 (14 obs)
  Primary splits:
      NW   < 7.65      to the left,  improve=0.2828293, (0 missing)
      M    < 13.05     to the left,  improve=0.2714159, (0 missing)
      Time < 21.9001   to the left,  improve=0.2060170, (0 missing)
      M.F  < 99.2      to the left,  improve=0.1703438, (0 missing)
      Po1  < 10.75     to the left,  improve=0.1659433, (0 missing)
  Surrogate splits:
      Ed   < 11.45     to the right, agree=0.750, adj=0.4, (0 split)
      Ineq < 16.25     to the left,  agree=0.750, adj=0.4, (0 split)
      Time < 21.9001   to the left,  agree=0.750, adj=0.4, (0 split)
      Pop  < 30        to the left,  agree=0.708, adj=0.3, (0 split)
      LF   < 0.5885    to the right, agree=0.667, adj=0.2, (0 split)
```

Subsequent nodes have less information as they're really trees:

```
Node number 4: 12 observations
  mean=550.5, MSE=20317.58

Node number 5: 11 observations
  mean=799.5455, MSE=16315.52

Node number 6: 10 observations
  mean=886.9, MSE=55757.49

Node number 7: 14 observations
  mean=1304.929, MSE=144801.8
```

snewt
ISYE6501x HW#4
Due June 14 2018

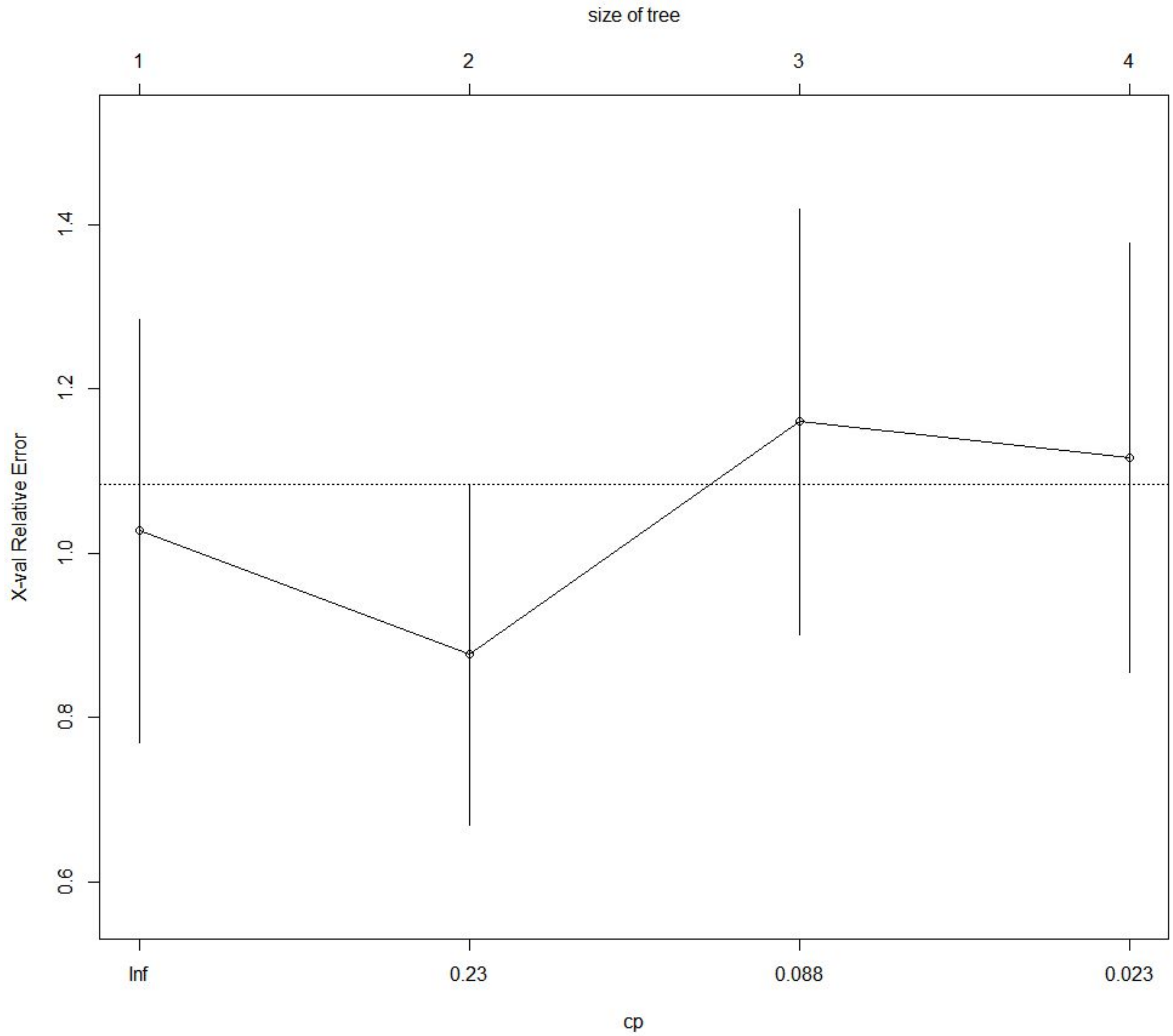
Display the actual results:

```
> printcp(crimetree)

Regression tree:
rpart(formula = crimes$Crime ~ ., data = crimes[1:15], method = "anova")

Variables actually used in tree construction:
[1] NW  Po1 Pop

Root node error: 6880928/47 = 146403

n= 47

        CP nsplit rel error  xerror    xstd
1 0.362963      0   1.00000 1.02716 0.25802
2 0.148143      1   0.63704 0.87669 0.20767
3 0.051732      2   0.48889 1.16020 0.25898
4 0.010000      3   0.43716 1.11585 0.26123
```

snewt
ISYE6501x HW#4
Due June 14 2018

View the cross validation results:

```
> plotcp(crimetree)
```

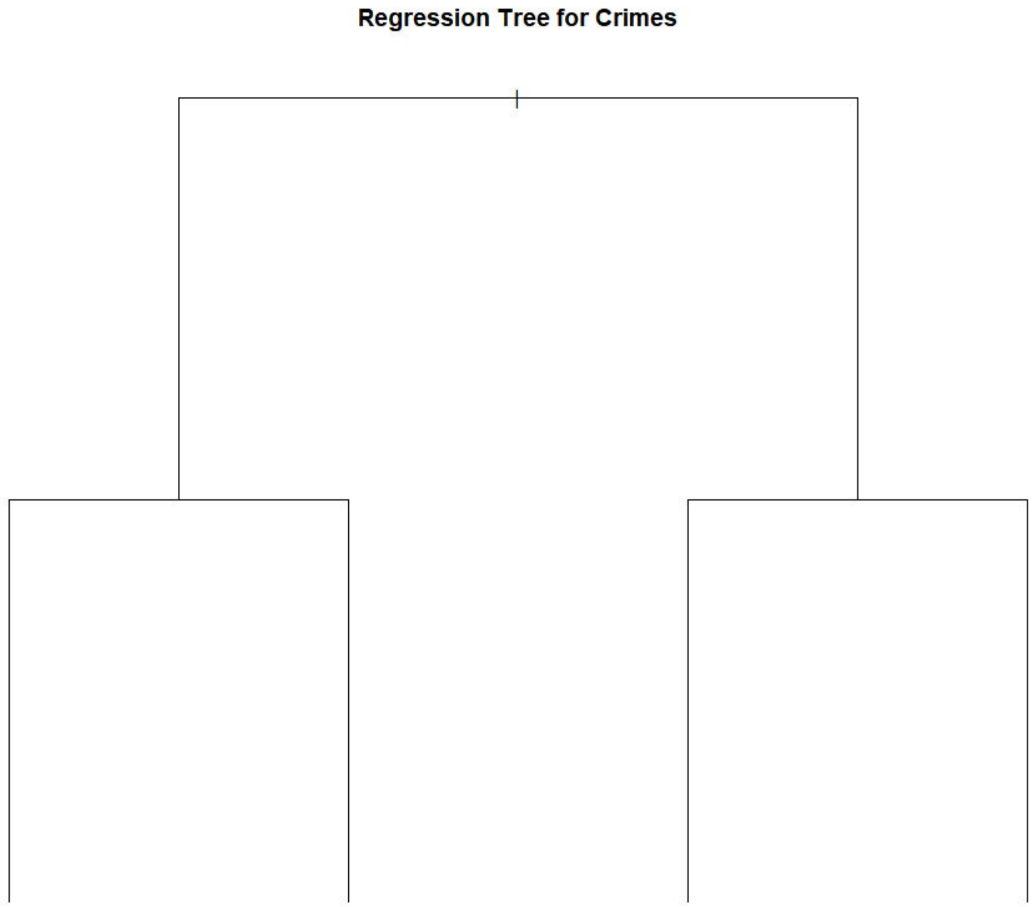


Translation: least relative x-value error, is with a tree of size 4.

snewt
ISYE6501x HW#4
Due June 14 2018

Print the naked plot:

```
> plot(crimetree,uniform=TRUE,main="Regression Tree for Crimes")
```



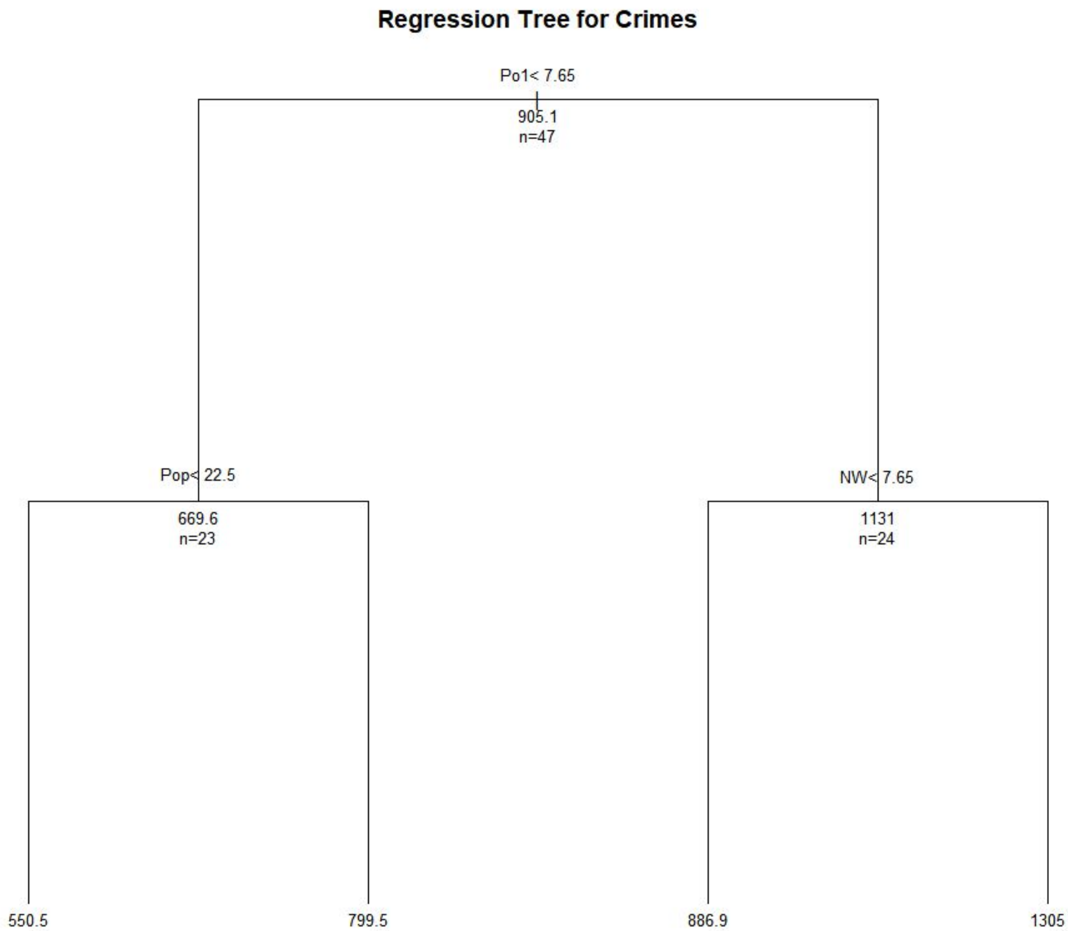
Unsurprisingly, the naked tree shows a 4 branched tree.

snewt
ISYE6501x HW#4
Due June 14 2018

Add labels to the tree:

```
> text(crimetree,use.n=TRUE,all=TRUE,cex=.8)
```

We W

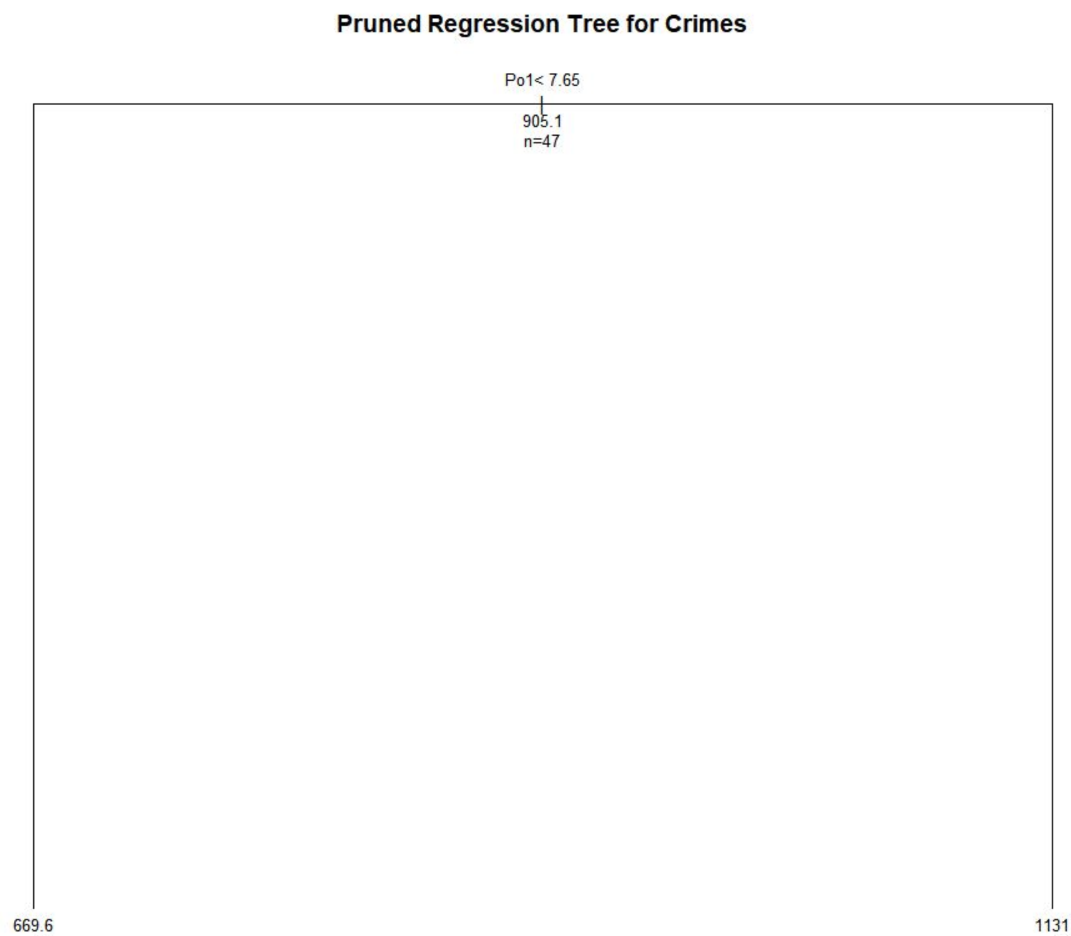


We have splits of Po < 7.65, then subsequent splits of Pop < 22.5, NW<7.65.

snewt
ISYE6501x HW#4
Due June 14 2018

Prune the tree, then plot, label:

```
> pcrimetree<-prune(crimetree,
cp=crimetree$cptable[which.min(crimetree$cptable[,"xerror"]),"CP"])
> plot(pcrimetree,uniform = TRUE,main="Pruned Regression Tree for Crimes")
> text(pcrimetree,use.n=TRUE,all=TRUE,cex=.8)
```



Now we’ve only got the first split.

snewt
ISYE6501x HW#4
Due June 14 2018

Question 10.1.b

Random Forest model

```
> library(randomForest)
randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.
> rfcrime <- randomForest(crimes$Crime~.,data=crimes)
> print(rfcrime)

Call:
 randomForest(formula = crimes$Crime ~ ., data = crimes)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 5

          Mean of squared residuals: 88932.7
                    % Var explained: 39.25
```

Summary summarizes

```
> importance(rfcrime)
       IncNodePurity
M          216029.72
So          21331.46
Ed         261648.92
Po1       1200658.12
Po2       1090817.22
LF         264774.42
M.F        296961.70
Pop        327208.27
NW         565322.65
U1         144781.07
U2         139341.07
Wealth     664870.39
Ineq       196863.70
Prob       727230.58
Time       227307.06
```

Question 10.2

Where would I use a logistic regression? What predictors would I use? These are questions. According to http://logisticregressionanalysis.com/33-when-to-use-logistic-regression/ We need 3 things:

1. We have a binary or dichotomous *Y* variable.
2. We have explanatory *X*-variables that we think are related to the *Y*-variable.

3. It is reasonable to think that the value the *Y*-variable takes on is like a coin flip where the probability of getting a 1 (“heads”) depends on the explanatory variables.

I’ve said before, I think, that I work in healthcare software. Specifically, I work in incentives reporting and regulation of that software. A lot of those incentives measures deal with people with cancers. They either have cancer, or they don’t have cancer. That’s a binary variable. What can explain this? Could be sex - it’s a lot less likely that a male patient born male will have breast cancer (but not impossible). Could be age - we don’t normally even screen patients under 18 for breast cancer (that’s not crazy either). Could be percentage of family members with history of cancer. So if we have that percentage of family members with a history as a fraction of 1, and age as an integer, and sex as a binary (0 for male, 1 for female) and cancer status as 0/1 for not/yes cancer. That seems like a reasonable situation for a logistic regression to me.

Question 10.3

So let’s split into a test, training set?

```
> gdata <- german
> set.seed(9)
> rowindices <- sample(1:nrow(gdata),round(.8*nrow(gdata)),replace = FALSE)
> rowindices <- sample(1:nrow(gdata),round(.8*nrow(gdata)),replace = FALSE)
```

snewt
ISYE6501x HW#4
Due June 14 2018

```
> summary(myglm)

Call:
glm(formula = gdata$bgdata ~ ., family = binomial(link = "logit"),
    data = gdata2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3410  -0.6994  -0.3752   0.7095   2.6116

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.005e-01  1.084e+00   0.369 0.711869
V1A12       -3.749e-01  2.179e-01  -1.720 0.085400 .
V1A13       -9.657e-01  3.692e-01  -2.616 0.008905 **
V1A14       -1.712e+00  2.322e-01  -7.373 1.66e-13 ***
V2           2.786e-02  9.296e-03   2.997 0.002724 **
V3A31        1.434e-01  5.489e-01   0.261 0.793921
V3A32       -5.861e-01  4.305e-01  -1.362 0.173348
V3A33       -8.532e-01  4.717e-01  -1.809 0.070470 .
V3A34       -1.436e+00  4.399e-01  -3.264 0.001099 **
V4A41       -1.666e+00  3.743e-01  -4.452 8.51e-06 ***
V4A410      -1.489e+00  7.764e-01  -1.918 0.055163 .
V4A42       -7.916e-01  2.610e-01  -3.033 0.002421 **
V4A43       -8.916e-01  2.471e-01  -3.609 0.000308 ***
V4A44       -5.228e-01  7.623e-01  -0.686 0.492831
V4A45       -2.164e-01  5.500e-01  -0.393 0.694000
V4A46        3.628e-02  3.965e-01   0.092 0.927082
V4A48       -2.059e+00  1.212e+00  -1.699 0.089297 .
V4A49       -7.401e-01  3.339e-01  -2.216 0.026668 *
V5           1.283e-04  4.444e-05   2.887 0.003894 **
V6A62       -3.577e-01  2.861e-01  -1.250 0.211130
V6A63       -3.761e-01  4.011e-01  -0.938 0.348476
V6A64       -1.339e+00  5.249e-01  -2.551 0.010729 *
V6A65       -9.467e-01  2.625e-01  -3.607 0.000310 ***
V7A72       -6.691e-02  4.270e-01  -0.157 0.875475
V7A73       -1.828e-01  4.105e-01  -0.445 0.656049
V7A74       -8.310e-01  4.455e-01  -1.866 0.062110 .
V7A75       -2.766e-01  4.134e-01  -0.669 0.503410
V8           3.301e-01  8.828e-02   3.739 0.000185 ***
V9A92       -2.755e-01  3.865e-01  -0.713 0.476040
V9A93       -8.161e-01  3.799e-01  -2.148 0.031718 *
V9A94       -3.671e-01  4.537e-01  -0.809 0.418448
```

```
V10A102      4.360e-01  4.101e-01   1.063 0.287700
V10A103     -9.786e-01  4.243e-01  -2.307 0.021072 *
V11          4.776e-03  8.641e-02   0.055 0.955920
V12A122      2.814e-01  2.534e-01   1.111 0.266630
V12A123      1.945e-01  2.360e-01   0.824 0.409743
V12A124      7.304e-01  4.245e-01   1.721 0.085308 .
V13         -1.454e-02  9.222e-03  -1.576 0.114982
V14A142     -1.232e-01  4.119e-01  -0.299 0.764878
V14A143     -6.463e-01  2.391e-01  -2.703 0.006871 **
V15A152     -4.436e-01  2.347e-01  -1.890 0.058715 .
V15A153     -6.839e-01  4.770e-01  -1.434 0.151657
V16          2.721e-01  1.895e-01   1.436 0.151109
V17A172      5.361e-01  6.796e-01   0.789 0.430160
V17A173      5.547e-01  6.549e-01   0.847 0.397015
V17A174      4.795e-01  6.623e-01   0.724 0.469086
V18          2.647e-01  2.492e-01   1.062 0.288249
V19A192     -3.000e-01  2.013e-01  -1.491 0.136060
V20A202     -1.392e+00  6.258e-01  -2.225 0.026095 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.73  on 999  degrees of freedom
Residual deviance:  895.82  on 951  degrees of freedom
AIC: 993.82

Number of Fisher Scoring iterations: 5
```

Whoops forgot to separate the test/train set

```
> myglm <- glm(train[,21]~., data=train[,1:20], family=binomial(link="logit"))
> summary(myglm)

Call:
glm(formula = train[, 21] ~ ., family = binomial(link = "logit"),
    data = train[, 1:20])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2887  -0.6931  -0.3706   0.7119   2.7160
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.955e-01  1.258e+00  -0.155 0.876488
V1A12       -5.158e-01  2.441e-01  -2.113 0.034579 *
V1A13       -1.229e+00  4.150e-01  -2.962 0.003054 **
V1A14       -1.910e+00  2.610e-01  -7.318 2.52e-13 ***
V2           1.843e-02  1.045e-02   1.765 0.077636 .
V3A31        1.094e-01  6.150e-01   0.178 0.858844
V3A32       -2.468e-01  4.819e-01  -0.512 0.608510
V3A33       -4.037e-01  5.291e-01  -0.763 0.445471
V3A34       -9.601e-01  4.933e-01  -1.946 0.051640 .
V4A41       -1.614e+00  4.222e-01  -3.824 0.000131 ***
V4A410      -1.867e+00  8.656e-01  -2.157 0.031038 *
V4A42       -7.119e-01  2.949e-01  -2.414 0.015759 *
V4A43       -9.678e-01  2.807e-01  -3.447 0.000566 ***
V4A44       -6.074e-01  7.847e-01  -0.774 0.438921
V4A45       -3.768e-01  6.238e-01  -0.604 0.545814
V4A46        7.630e-02  4.295e-01   0.178 0.859004
V4A48       -2.153e+00  1.249e+00  -1.723 0.084826 .
V4A49       -7.657e-01  3.747e-01  -2.043 0.041011 *
V5           1.625e-04  5.013e-05   3.241 0.001189 **
V6A62       -2.486e-01  3.164e-01  -0.786 0.432042
V6A63       -3.374e-01  4.426e-01  -0.762 0.445905
V6A64       -1.245e+00  5.915e-01  -2.106 0.035220 *
V6A65       -8.516e-01  2.913e-01  -2.923 0.003463 **
V7A72        4.279e-02  4.928e-01   0.087 0.930808
V7A73        1.253e-01  4.719e-01   0.265 0.790675
V7A74       -5.570e-01  5.069e-01  -1.099 0.271835
V7A75       -1.126e-01  4.756e-01  -0.237 0.812910
V8           4.438e-01  1.019e-01   4.357 1.32e-05 ***
V9A92       -2.121e-01  4.236e-01  -0.501 0.616631
V9A93       -7.935e-01  4.196e-01  -1.891 0.058596 .
V9A94       -2.720e-01  5.015e-01  -0.542 0.587611
V10A102      3.833e-01  4.882e-01   0.785 0.432337
V10A103     -9.098e-01  4.942e-01  -1.841 0.065608 .
V11          3.621e-02  9.900e-02   0.366 0.714532
V12A122      3.882e-01  2.893e-01   1.342 0.179611
V12A123      2.908e-01  2.661e-01   1.093 0.274480
V12A124      7.136e-01  4.652e-01   1.534 0.125045
V13         -8.447e-03  1.026e-02  -0.824 0.410114
V14A142      6.211e-02  4.751e-01   0.131 0.895980
V14A143     -6.511e-01  2.706e-01  -2.406 0.016119 *
V15A152     -5.321e-01  2.621e-01  -2.030 0.042333 *
```

```
V15A153     -5.399e-01  5.204e-01  -1.037 0.299507
V16          1.442e-02  2.073e-01   0.070 0.944556
V17A172      6.536e-01  8.102e-01   0.807 0.419875
V17A173      5.292e-01  7.830e-01   0.676 0.499103
V17A174      6.161e-01  7.880e-01   0.782 0.434331
V18          1.896e-01  2.864e-01   0.662 0.507938
V19A192     -4.137e-01  2.326e-01  -1.779 0.075247 .
V20A202     -1.419e+00  6.488e-01  -2.186 0.028789 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 975.68  on 799  degrees of freedom
Residual deviance: 714.59  on 751  degrees of freedom
AIC: 812.59

Number of Fisher Scoring iterations: 5
```

So what’s got a $p < .001$? No checking account, seeking a loan for a used car, seeking a loan for a radio/television (all (-)), and payment as a percentage of disposable income (+). The negative values mean that all other variables being equal, seeking a loan for those items was significantly less likely to be approved. The positive value means that all other factors being equal, seeking a loan that is a smaller portion of your disposable income is more likely to be approved.

Let’s now try to take a look at $.001 < p < .01$. (-): unknown or no savings account, critical account or other credits with a different bank, or more than or equal to 200 German Marks beyond salary for one year; (+) smaller credit amount or longer loans.

Well now let's look at an ANOVA of those results for the model:

```
> anova(myglm,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: train[, 21]

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                   799     975.68
V1    3  116.524       796     859.16 < 2.2e-16 ***
V2    1   24.062       795     835.10 9.326e-07 ***
V3    4   16.031       791     819.07  0.002978 **
V4    9   25.694       782     793.37  0.002292 **
V5    1    1.570       781     791.80  0.210173
V6    4   12.858       777     778.94  0.011988 *
V7    4    8.189       773     770.75  0.084903 .
V8    1   18.134       772     752.62 2.059e-05 ***
V9    3    8.417       769     744.20  0.038136 *
V10   2    4.410       767     739.79  0.110232
V11   1    0.977       766     738.82  0.323052
V12   3    3.001       763     735.81  0.391481
V13   1    1.263       762     734.55  0.261058
V14   2    6.351       760     728.20  0.041764 *
V15   2    3.781       758     724.42  0.151034
V16   1    0.013       757     724.41  0.910433
V17   3    1.071       754     723.34  0.783984
V18   1    0.323       753     723.01  0.569653
V19   1    2.777       752     720.24  0.095649 .
V20   1    5.650       751     714.59  0.017458 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Translation: the model with only variables 1, 2, 8 explain the most. The model with variables 1,2,3,4,8 explains enough to still be worth the trade off. The other variables don't substantially help the model enough to be bothered with.