Microeconometrics, Empirical project, Group 8

Atanasov Georgi* Fitter Jonathan† Hochholzer Matthias‡ Woharcik Verena§

17th February 2021

Contents

*student ID 11776393
†student ID 11709902
‡student ID 11724853
§student ID 11701581

1 Data

1.1 Importing the dataset

from Wooldridge, his source: J. Grogger (1991), “Certainty vs. Severity of Punishment,” Economic Inquiry 29, 297-309.

```
df<-read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/crime1.dta")
attach(df)
head(df)
```

```
##   narr86 nfarr86 nparr86 pcnv avgsen tottime ptime86 qemp86 inc86 durat black
## 1      0       0       0 0.38   17.6    35.2      12      0   0.0     0     0
## 2      2       2       0 0.44    0.0     0.0       0      1   0.8     0     0
## 3      1       1       0 0.33   22.8    22.8       0      0   0.0    11     1
## 4      2       2       1 0.25    0.0     0.0       5      2   8.8     0     0
## 5      1       1       0 0.00    0.0     0.0       0      2   8.1     1     0
## 6      0       0       0 1.00    0.0     0.0       0      4  97.6     0     0
##   hispan born60 pcnvsq pt86sq     inc86sq
## 1      0      1 0.1444    144     0.00000
## 2      1      0 0.1936      0     0.64000
## 3      0      1 0.1089      0     0.00000
## 4      1      1 0.0625     25    77.44000
## 5      0      0 0.0000      0    65.61001
## 6      0      1 1.0000      0  9525.75977
```

```
str(df)
```

```
## 'data.frame':    2725 obs. of  16 variables:
##  $ narr86 : num  0 2 1 2 1 0 2 5 0 0 ...
##  $ nfarr86: num  0 2 1 2 1 0 2 3 0 0 ...
##  $ nparr86: num  0 0 0 1 0 0 1 5 0 0 ...
##  $ pcnv   : num  0.38 0.44 0.33 0.25 0 ...
##  $ avgsen : num  17.6 0 22.8 0 0 ...
##  $ tottime: num  35.2 0 22.8 0 0 ...
##  $ ptime86: num  12 0 0 5 0 0 0 0 9 0 ...
##  $ qemp86 : num  0 1 0 2 2 4 0 0 0 3 ...
##  $ inc86  : num  0 0.8 0 8.8 8.1 ...
##  $ durat  : num  0 0 11 0 1 ...
##  $ black  : num  0 0 1 0 0 0 1 0 1 0 ...
##  $ hispan : num  0 1 0 1 0 0 0 0 0 1 ...
##  $ born60 : num  1 0 1 1 0 1 1 1 1 1 ...
##  $ pcnvsq : num  0.1444 0.1936 0.1089 0.0625 0 ...
##  $ pt86sq : num  144 0 0 25 0 0 0 0 81 0 ...
##  $ inc86sq: num  0 0.64 0 77.44 65.61 ...
##  - attr(*, "datalabel")= chr ""
##  - attr(*, "time.stamp")= chr "10 Jan 2000 16:54"
##  - attr(*, "formats")= chr  "%9.0g" "%9.0g" "%9.0g" "%9.0g" ...
##  - attr(*, "types")= int  102 102 102 102 102 102 102 102 102 102 ...
##  - attr(*, "val.labels")= chr  "" "" "" "" ...
##  - attr(*, "var.labels")= chr  "" "" "" "" ...
##  - attr(*, "version")= int 6
```

```r
summary(df)
```

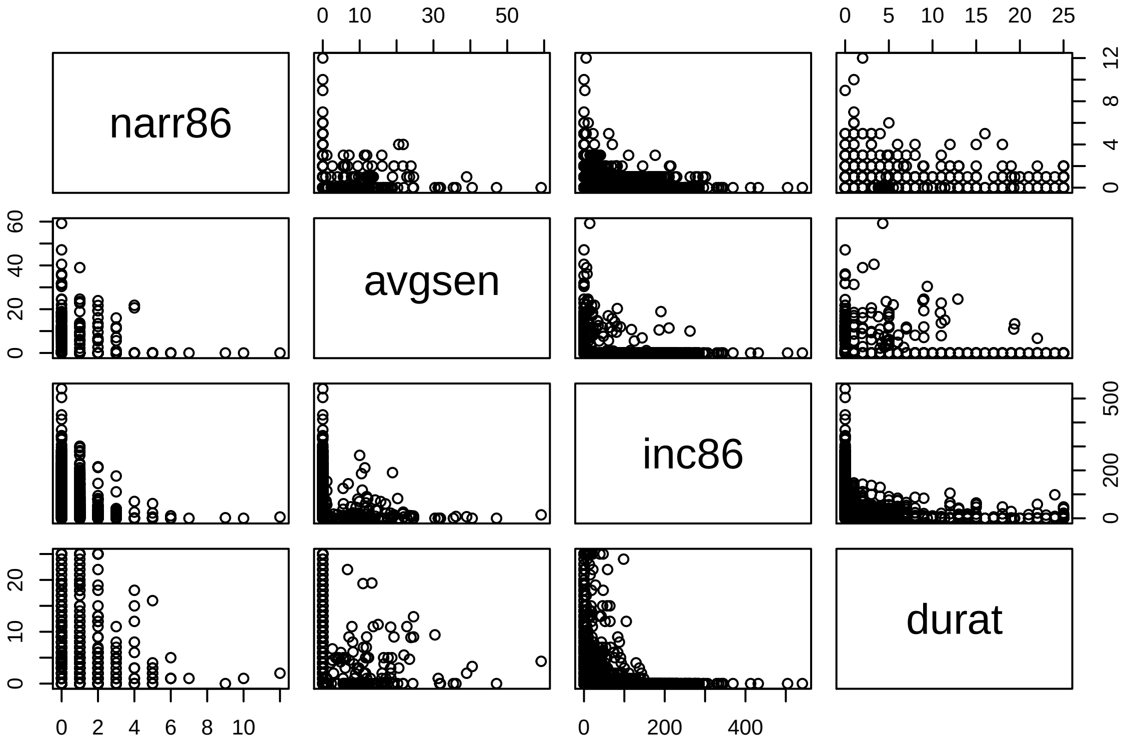
```
##      narr86          nfarr86          nparr86           pcnv       
##  Min.   : 0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median : 0.0000   Median :0.0000   Median :0.0000   Median :0.2500  
##  Mean   : 0.4044   Mean   :0.2334   Mean   :0.1255   Mean   :0.3578  
##  3rd Qu.: 1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.6700  
##  Max.   :12.0000   Max.   :6.0000   Max.   :8.0000   Max.   :1.0000  
##      avgsen           tottime            ptime86            qemp86     
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   :0.000  
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:1.000  
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median :3.000  
##  Mean   : 0.6323   Mean   : 0.8387   Mean   : 0.3872   Mean   :2.309  
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:4.000  
##  Max.   :59.2000   Max.   :63.4000   Max.   :12.0000   Max.   :4.000  
##      inc86           durat            black            hispan      
##  Min.   :  0.00   Min.   : 0.000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:  0.40   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median : 29.00   Median : 0.000   Median :0.0000   Median :0.0000  
##  Mean   : 54.97   Mean   : 2.251   Mean   :0.1611   Mean   :0.2176  
##  3rd Qu.: 90.10   3rd Qu.: 2.000   3rd Qu.:0.0000   3rd Qu.:0.0000  
##  Max.   :541.00   Max.   :25.000   Max.   :1.0000   Max.   :1.0000  
##      born60           pcnvsq           pt86sq            inc86sq         
##  Min.   :0.0000   Min.   :0.0000   Min.   :  0.000   Min.   :     0.00  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  0.000   1st Qu.:     0.16  
##  Median :0.0000   Median :0.0625   Median :  0.000   Median :   841.00  
##  Mean   :0.3626   Mean   :0.2841   Mean   :  3.951   Mean   :  7458.93  
##  3rd Qu.:1.0000   3rd Qu.:0.4489   3rd Qu.:  0.000   3rd Qu.:  8118.01  
##  Max.   :1.0000   Max.   :1.0000   Max.   :144.000   Max.   :292681.00
```

A data.frame with 2725 observations on 16 variables: - narr86: times arrested, 1986 - nfarr86: felony arrests, 1986 - nparr86: property crme arr., 1986 - pcnv: proportion of prior convictions - avgsen: avg sentence length, mos. - tottime: time in prison since 18 (mos.) - ptime86: mos. in prison during 1986 - qemp86: quarters employed, 1986 - inc86: legal income, 1986, \$100s - durat: recent unemp duration - black: =1 if black - hispan: =1 if Hispanic - born60: =1 if born in 1960 - pcnvsq: pcnvˆ2 - pt86sq: ptime86ˆ2 - inc86sq: inc86ˆ2

1.2 Descriptive Statistics

1.2.1 Correlation Plots

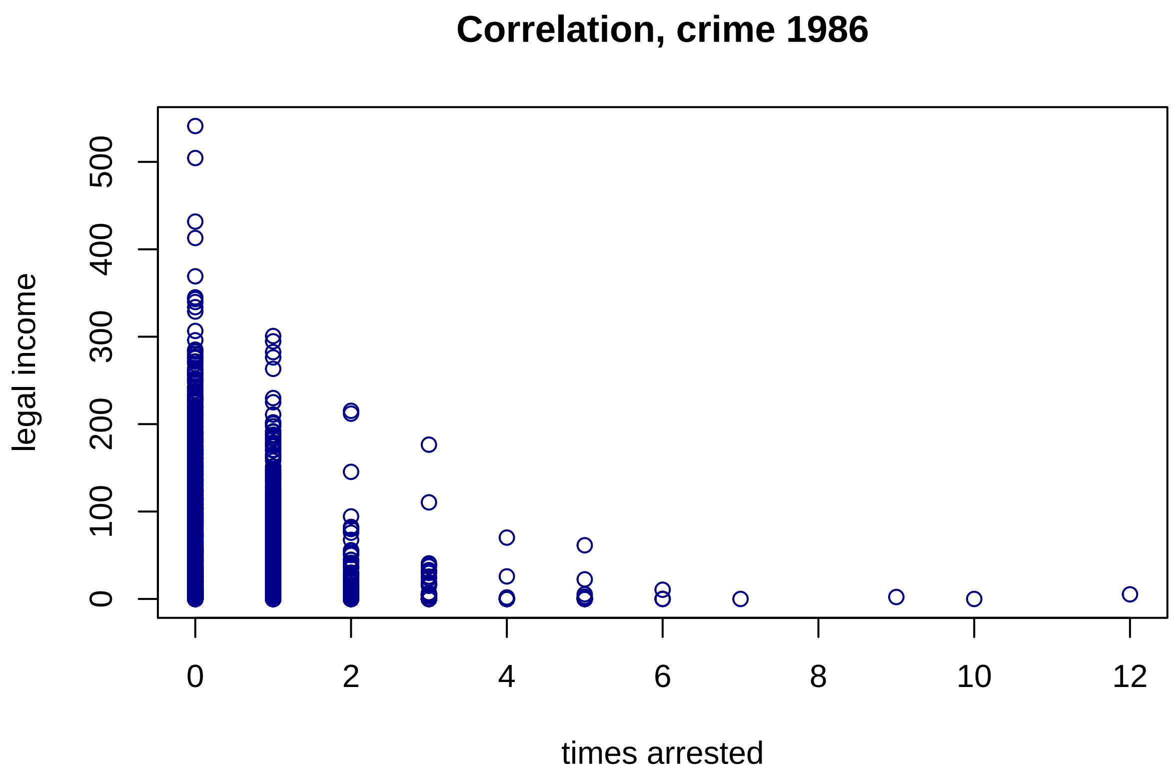
```r
plot(df[,c("narr86", "avgsen", "inc86", "durat")])
```

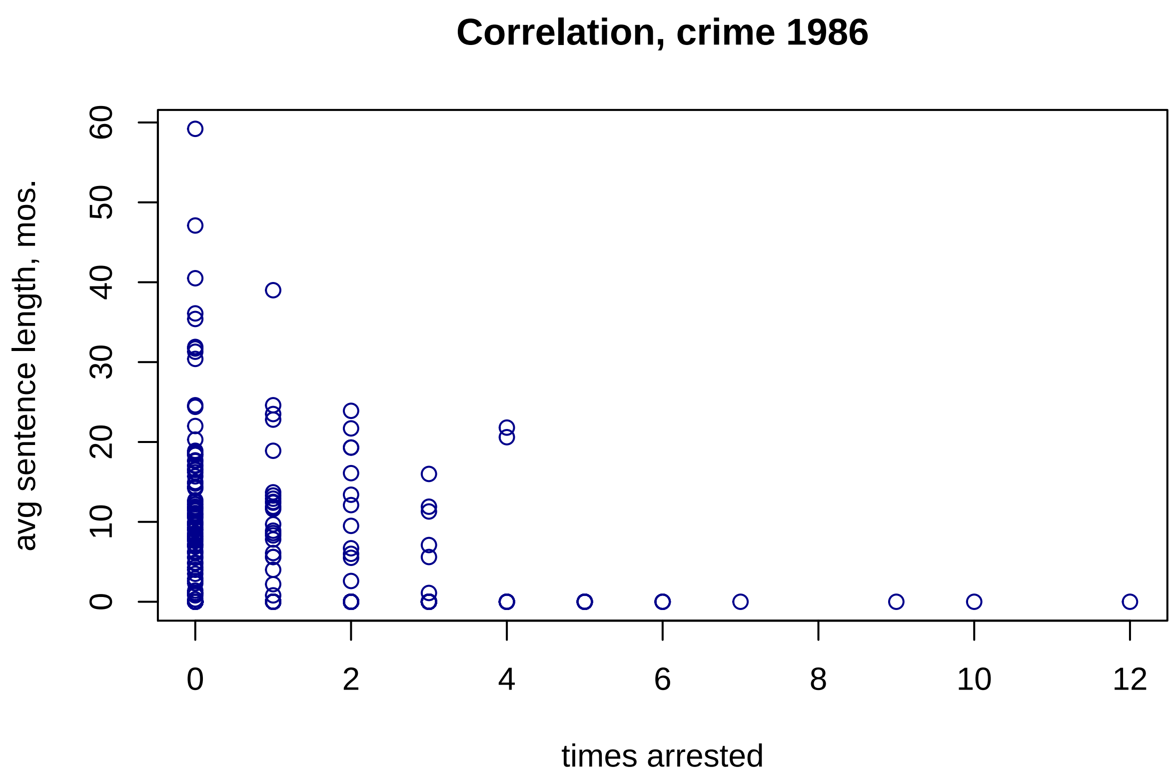



```r
cor(df[,c("narr86", "avgsen", "inc86", "durat")])
```

```
##             narr86      avgsen       inc86       durat
## narr86  1.00000000  0.02929780 -0.18997653  0.08232769
## avgsen  0.02929780  1.00000000 -0.09580596  0.02843162
## inc86  -0.18997653 -0.09580596  1.00000000 -0.34292954
## durat   0.08232769  0.02843162 -0.34292954  1.00000000
```

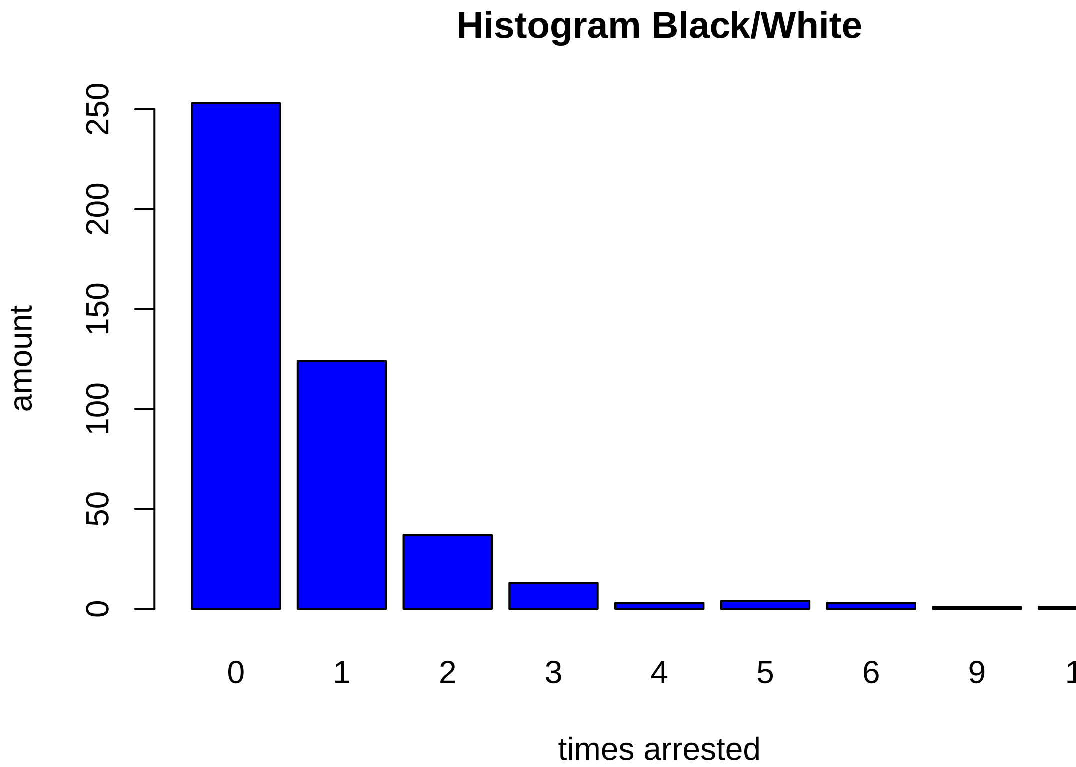
1.2.2 Specific Plots:



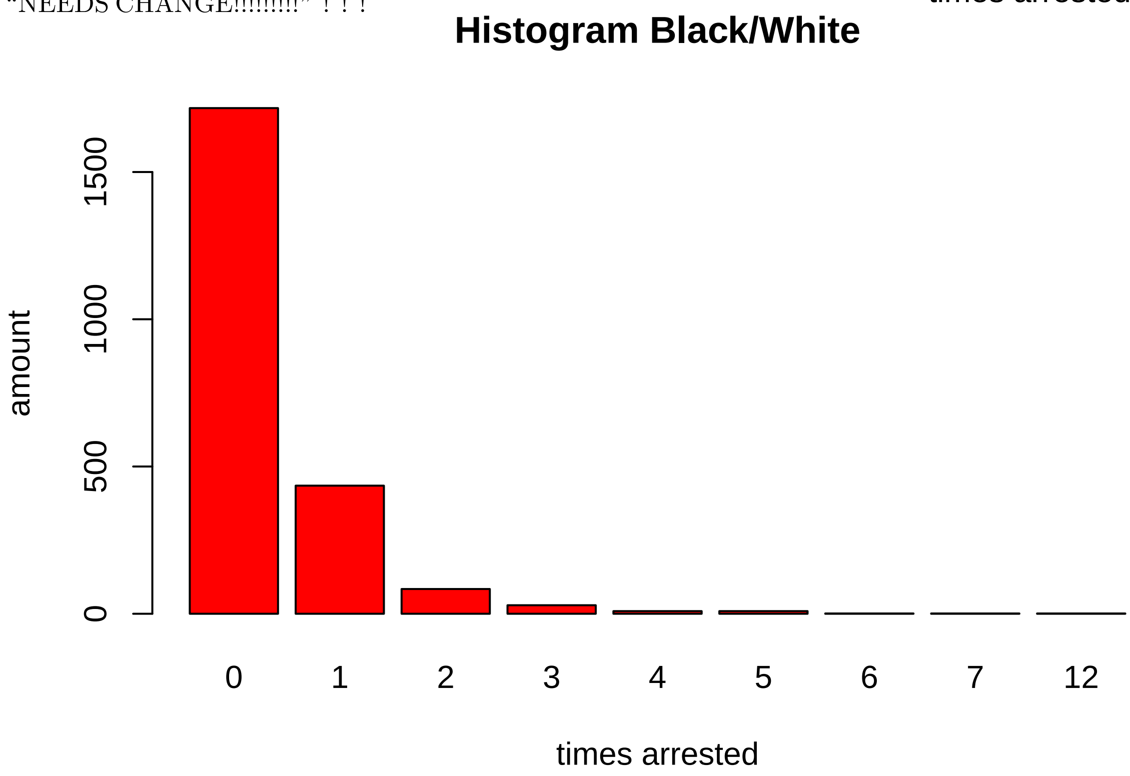



```
"HISTOGRAMME !"
```

```
## [1] "HISTOGRAMME !"
```



“NEEDS CHANGE!!!!!!!!!” ! ! !



2 PART 1

** Modeling “avgsen” ** Building model estimating expected severity of conviction when arrested in 1986 using level of income, employment, total time spend in prison and color (black ad non-black) of the arrested

Our hypothesis is, that the mentioned variables have a significant effect on the average sentence length.

$$avgsen = \beta_0 + \beta_1\ inc86 + \beta_2\ black + \beta_3\ tottime + \beta_4\ qemp86$$

2.1 Simple OLS-Estimation

A General OLS estimation including all potential regressors:

```
lm_all<-lm(avgsen~. -nfarr86 - nparr86 , data = df)
summary(lm_all)
```

```
##
## Call:
## lm(formula = avgsen ~ . - nfarr86 - nparr86, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4560  -0.0948  -0.0346   0.0093  16.7462
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.743e-02  7.345e-02   0.237   0.8124
## narr86      -4.337e-02  3.060e-02  -1.418   0.1564
## pcnv         3.163e-01  2.467e-01   1.282   0.1998
## tottime      7.103e-01  5.712e-03 124.346   <2e-16 ***
## ptime86      9.820e-02  7.117e-02   1.380   0.1678
## qemp86       3.911e-02  2.932e-02   1.334   0.1824
## inc86       -1.652e-03  1.286e-03  -1.284   0.1991
## durat       -3.538e-04  6.318e-03  -0.056   0.9553
## black        1.361e-01  7.193e-02   1.893   0.0585 .
## hispan      -2.537e-02  6.304e-02  -0.402   0.6875
## born60      -1.248e-02  5.225e-02  -0.239   0.8113
## pcnvsq      -3.393e-01  2.500e-01  -1.357   0.1749
## pt86sq      -1.364e-02  6.247e-03  -2.183   0.0291 *
## inc86sq      3.373e-06  4.077e-06   0.827   0.4081
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.298 on 2711 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8631
## F-statistic:  1322 on 13 and 2711 DF,  p-value: < 2.2e-16
```

Interpretation: A high R-squared is observable. Only few variables are significant for 0.05 and 0.1 significance level. Also the p-Value of the F-statistic is low, which implies that there are some variables which can be used to explain the average sentence length.

We have proceeded our further estimation of avsen after excluding variables which have considerably high p-values.

The average severity is regressed on the income in 1986, employment in 1986, color (black and non-black) and total time spend in prison.

```r
lm_sev<-lm(avgsen~ tottime+ black+ qemp86+ inc86, data = df)
summary(lm_sev)
```

```
##
## Call:
## lm(formula = avgsen ~ tottime + black + qemp86 + inc86, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2801  -0.0774  -0.0329   0.0213  17.2152
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0374053  0.0476873  -0.784   0.4329
## tottime      0.7064354  0.0054793 128.928   <2e-16 ***
## black        0.1402641  0.0690914   2.030   0.0424 *
## qemp86       0.0425101  0.0221607   1.918   0.0552 .
## inc86       -0.0007928  0.0005335  -1.486   0.1374
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.301 on 2720 degrees of freedom
## Multiple R-squared:  0.8626, Adjusted R-squared:  0.8624
## F-statistic:  4268 on 4 and 2720 DF,  p-value: < 2.2e-16
```

Output of an OLS-Estimation is given:

```r
summary((lm_sev))
```


```
##
## Call:
## lm(formula = avgsen ~ tottime + black + qemp86 + inc86, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2801  -0.0774  -0.0329   0.0213  17.2152
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0374053  0.0476873  -0.784   0.4329
## tottime      0.7064354  0.0054793 128.928   <2e-16 ***
## black        0.1402641  0.0690914   2.030   0.0424 *
## qemp86       0.0425101  0.0221607   1.918   0.0552 .
## inc86       -0.0007928  0.0005335  -1.486   0.1374
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.301 on 2720 degrees of freedom
## Multiple R-squared:  0.8626, Adjusted R-squared:  0.8624
## F-statistic:  4268 on 4 and 2720 DF,  p-value: < 2.2e-16
```


Interpretation: We see almost the same R-squared as from the previous OLS-Estimation. The significant variables for 0.05 significance level are the total time spend in prison and the color. No significance of the other variables is proven.

2.1.0.1 Problems with the OLS

Some of the variables may be endogenous E.g assumptions may be violated. => Testing this way may not be correct.

2.2 IV-Regression (using 2SLS-Estimation)

Use instrumental variables in the estimation of the expected severity. Define: endogenous var: income86, qemp86, tottime exogenuos var: black instruments: durat, nparr, nfarr, narr, ptime86

$$avgsen = \beta_0 + \beta_1\, in\hat{c}86 + \beta_2\, black + \beta_3\, tott\hat{i}me + \beta_4\, qem\hat{p}86$$

,with $in\hat{c}86 = \beta_0 + \beta_1\; durat + \beta_2\, nparr + \beta_3\, nfarr + \beta_4\, narr + \beta_5\, ptime86$ $tott\hat{i}me = \beta_0 + \beta_1\; durat + \beta_2\, nparr + \beta_3\, nfarr + \beta_4\, narr + \beta_5\, ptime86$ $qem\hat{p}86 = \beta_0 + \beta_1\; durat + \beta_2\, nparr + \beta_3\, nfarr + \beta_4\, narr + \beta_5\, ptime86$

The regression code is given by:

```
IV_sev1<-ivreg(avgsen~ tottime+ black+ qemp86+ inc86 | black+ durat+ narr86+ nfarr86+ nparr86+ ptime86

summary(IV_sev1, diagnostics=TRUE)
```

```
##
## Call:
## ivreg(formula = avgsen ~ tottime + black + qemp86 + inc86 | black +
##     durat + narr86 + nfarr86 + nparr86 + ptime86, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -11.72068  -0.17947  -0.09283   0.02787  21.82055
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.122938   0.132308   0.929  0.35288
## tottime      0.627074   0.026174  23.958  < 2e-16 ***
## black        0.258763   0.090411   2.862  0.00424 **
## qemp86      -0.099968   0.185116  -0.540  0.58922
## inc86        0.003139   0.006011   0.522  0.60157
##
## Diagnostic tests:
##                             df1  df2 statistic p-value
## Weak instruments (tottime)    5 2718    46.827 < 2e-16 ***
## Weak instruments (qemp86)     5 2718   258.873 < 2e-16 ***
## Weak instruments (inc86)      5 2718   105.504 < 2e-16 ***
## Wu-Hausman                    3 2717     4.503 0.00371 **
## Sargan                        2   NA     2.277 0.32028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.364 on 2720 degrees of freedom
## Multiple R-Squared: 0.8491,  Adjusted R-squared: 0.8488
## Wald test:   309 on 4 and 2720 DF,  p-value: < 2.2e-16
```

Interpretation: Here a high R-squared is observed. Tottime and black are the only significant variables for 0.05 significance level. Furthermore, diagnostics of the instruments are provided. We observe small p-values, which means that instruments are not weak e.g they are appropriate. The value of the Hausmans-test is smaller than than the significance level of 0.05. Thus, meaning that instruments and residuals can be considered as uncorrelated.

2.2.0.1 Manual Check if Instuments are adequate

1. Check if regressors and instruments are correlated

```
i1lm_sev1<- lm(tottime~ black+ durat+ narr86+ nfarr86+ nparr86+ ptime86, data=df)
summary(i1lm_sev1)
```

```
##
## Call:
## lm(formula = tottime ~ black + durat + narr86 + nfarr86 + nparr86 +
##     ptime86, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.254 -0.662 -0.306 -0.281 55.743
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.28080    0.10601   2.649  0.00812 **
## black        1.09852    0.23369   4.701 2.72e-06 ***
## durat        0.02531    0.01845   1.372  0.17013
## narr86       0.38140    0.17599   2.167  0.03031 *
## nfarr86     -0.11213    0.25363  -0.442  0.65845
## nparr86     -0.46308    0.23883  -1.939  0.05261 .
## ptime86      0.65619    0.04338  15.127  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.397 on 2718 degrees of freedom
## Multiple R-squared:  0.09124,    Adjusted R-squared:  0.08923
## F-statistic: 45.48 on 6 and 2718 DF,  p-value: < 2.2e-16
```

```
i2lm_sev1<- lm(qemp86~ black+ durat+ narr86+ nfarr86+ nparr86+ ptime86, data=df)
summary(i2lm_sev1)
```

```
##
## Call:
## lm(formula = qemp86 ~ black + durat + narr86 + nfarr86 + nparr86 +
##     ptime86, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7583 -0.9233  0.2340  1.0767  4.6960
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.923252   0.031584  92.554  < 2e-16 ***
## black       -0.328207   0.069629  -4.714 2.56e-06 ***
```

```
## durat       -0.164969   0.005496 -30.016  < 2e-16 ***
## narr86      -0.157216   0.052437  -2.998  0.00274 **
## nfarr86     -0.159808   0.075569  -2.115  0.03454 *
## nparr86     -0.009608   0.071160  -0.135  0.89260
## ptime86     -0.226935   0.012925 -17.558  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.31 on 2718 degrees of freedom
## Multiple R-squared:  0.3398, Adjusted R-squared:  0.3383
## F-statistic: 233.1 on 6 and 2718 DF,  p-value: < 2.2e-16
```

```
i3lm_sev1<- lm( inc86~ black+ durat+ narr86+ nfarr86+ nparr86+ ptime86, data=df)
summary(i3lm_sev1)
```

```
##
## Call:
## lm(formula = inc86 ~ black + durat + narr86 + nfarr86 + nparr86 +
##     ptime86, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -74.93 -44.37 -16.33  30.17 465.97
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  75.0256     1.4557  51.539  < 2e-16 ***
## black       -14.8285     3.2092  -4.621 4.00e-06 ***
## durat        -4.7147     0.2533 -18.612  < 2e-16 ***
## narr86      -10.3329     2.4168  -4.275 1.97e-05 ***
## nfarr86      -1.7791     3.4829  -0.511    0.610
## nparr86      -2.1158     3.2797  -0.645    0.519
## ptime86      -5.6714     0.5957  -9.520  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.38 on 2718 degrees of freedom
## Multiple R-squared:  0.1806, Adjusted R-squared:  0.1788
## F-statistic: 99.86 on 6 and 2718 DF,  p-value: < 2.2e-16
```

R-squared >> 0 is observed in every regression => first criterion is met.

2. Check if errors and instruments are uncorrelated.

```
resid_sev1<-resid(IV_sev1)
lm_resid_sev1<-lm(resid_sev1~black+ durat+ narr86+ nfarr86+ nparr86+ ptime86, data=df)
summary(lm_resid_sev1)
```

```
##
## Call:
## lm(formula = resid_sev1 ~ black + durat + narr86 + nfarr86 +
##     nparr86 + ptime86, data = df)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7167  -0.1842  -0.0872   0.0257  21.8264
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.003974   0.032882  -0.121    0.904
## black        0.001268   0.072489   0.017    0.986
## durat        0.001107   0.005722   0.193    0.847
## narr86       0.047213   0.054592   0.865    0.387
## nfarr86     -0.017087   0.078673  -0.217    0.828
## nparr86     -0.102002   0.074083  -1.377    0.169
## ptime86     -0.002650   0.013456  -0.197    0.844
##
## Residual standard error: 1.364 on 2718 degrees of freedom
## Multiple R-squared:  0.0008356,  Adjusted R-squared:  -0.00137
## F-statistic: 0.3789 on 6 and 2718 DF,  p-value: 0.8929
```

A really small R-squared is observed. The p-values of variables are considerably higher than 0.05 significance level.

What can be done in addition is a test on \$ n*R^{2} \$, where \$ R^{2} \$ is the non-centered \$ R^{2} \$ (\$ R^{2} \$ used)

```r
summary(lm_resid_sev1)$r.squared*length(resid_sev1)
```

```
## [1] 2.2771
```

Value is smaller than the Chi-square value on 2 df and 0.05 significance level=> also the second criterion is met.

2.3 Censored Tobit

```r
summary(tobit(avgsen~ tottime+ black+ qemp86+ inc86, left=-Inf, right = 12, data=df))
```

```
##
## Call:
## tobit(formula = avgsen ~ tottime + black + qemp86 + inc86, left = -Inf,
##     right = 12, data = df)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##           2725              0           2669             56
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.0313374  0.0257231   1.218    0.223
## tottime      0.6000381  0.0051706 116.049   <2e-16 ***
## black        0.0090406  0.0375819   0.241    0.810
## qemp86       0.0146955  0.0119743   1.227    0.220
```

```
## inc86       -0.0004182  0.0002874  -1.455    0.146
## Log(scale)  -0.3582712  0.0137308 -26.093   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 0.6989
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 10
## Log-likelihood: -2874 on 6 Df
## Wald-statistic: 1.367e+04 on 4 Df, p-value: < 2.22e-16
```

3 PART 2

Building a model, which aims at estimating probability of arrest during 1986. A dependend binory variable, describing the states: arrested and not arrested, is to be regressed.

In this part we test the hypothesis that every single regressor has a significant impact on the dependend variable.

3.1 Simple OLS Regression, LPM

3.1.1 OLS estimation of the variable narr86

Regressing the variable narr86 on almost all variables

```
##
## Call:
## lm(formula = narr86 ~ pcnv + avgsen + tottime + ptime86 + qemp86 +
##     inc86 + durat + black + hispan + born60 + pcnvsq + pt86sq +
##     inc86sq, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5542 -0.4622 -0.2097  0.2374 11.3955
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.618e-01  4.481e-02  12.537  < 2e-16 ***
## pcnv         5.710e-01  1.544e-01   3.697 0.000222 ***
## avgsen      -1.708e-02  1.205e-02  -1.418 0.156417
## tottime      1.203e-02  9.277e-03   1.297 0.194806
## ptime86      2.936e-01  4.432e-02   6.624 4.19e-11 ***
## qemp86      -2.706e-02  1.840e-02  -1.471 0.141512
## inc86       -3.348e-03  8.048e-04  -4.160 3.28e-05 ***
## durat       -7.652e-03  3.962e-03  -1.931 0.053535 .
## black        2.936e-01  4.481e-02   6.551 6.80e-11 ***
## hispan       1.616e-01  3.944e-02   4.098 4.29e-05 ***
## born60      -3.767e-02  3.278e-02  -1.149 0.250623
## pcnvsq      -7.488e-01  1.563e-01  -4.792 1.74e-06 ***
## pt86sq      -3.044e-02  3.879e-03  -7.846 6.12e-15 ***
## inc86sq      7.148e-06  2.555e-06   2.798 0.005178 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8146 on 2711 degrees of freedom
## Multiple R-squared:  0.1051, Adjusted R-squared:  0.1008
## F-statistic:  24.5 on 13 and 2711 DF,  p-value: < 2.2e-16
```

We will proceed our estimations ommitting insignificant variables from this estimation.

3.1.2 The Chosen Model:

After omitting the insignificant variables, we create the following model:

$$narr86 = \beta_0 + \beta_1\, pcnv + \beta_2\, ptime86 + \beta_3\, inc86 + \beta_4\, black + \beta_5\, hispan + \beta_6\, pcnvsq + \beta_7\, pt86sq + \beta_8\, inc86sq$$

```
##
## Call:
## lm(formula = narr86 ~ pcnv + ptime86 + inc86 + black + hispan +
##     pcnvsq + pt86sq + inc86sq, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5498 -0.4692 -0.2159  0.2309 11.4326
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.896e-01  3.227e-02  15.173  < 2e-16 ***
## pcnv         5.500e-01  1.533e-01   3.587  0.00034 ***
## ptime86      2.880e-01  4.388e-02   6.563 6.30e-11 ***
## inc86       -3.906e-03  5.257e-04  -7.430 1.45e-13 ***
## black        2.908e-01  4.464e-02   6.514 8.71e-11 ***
## hispan       1.623e-01  3.938e-02   4.120 3.89e-05 ***
## pcnvsq      -7.286e-01  1.552e-01  -4.695 2.80e-06 ***
## pt86sq      -2.946e-02  3.850e-03  -7.652 2.72e-14 ***
## inc86sq      8.377e-06  2.096e-06   3.996 6.60e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.815 on 2716 degrees of freedom
## Multiple R-squared:  0.1026, Adjusted R-squared:  0.09991
## F-statistic:  38.8 on 8 and 2716 DF,  p-value: < 2.2e-16
```

```
## [1] "Robust Standard Errors"
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  4.8963e-01  3.1484e-02 15.5517 < 2.2e-16 ***
## pcnv         5.4998e-01  1.6713e-01  3.2908  0.001012 **
## ptime86      2.8797e-01  6.9228e-02  4.1597 3.286e-05 ***
```

```
## inc86       -3.9062e-03  4.6991e-04 -8.3126 < 2.2e-16 ***
## black        2.9076e-01  5.7624e-02  5.0457 4.816e-07 ***
## hispan       1.6227e-01  3.9962e-02  4.0606 5.034e-05 ***
## pcnvsq      -7.2855e-01  1.6900e-01 -4.3109 1.684e-05 ***
## pt86sq      -2.9464e-02  5.8454e-03 -5.0405 4.948e-07 ***
## inc86sq      8.3771e-06  1.7314e-06  4.8384 1.382e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation: First to notice is the neglection of parameter restrictions: E.g. negative values cannot easily be interpreted in this scenario.

Although OLS yields unbiased estimators, heteroskedasticity among other things leads to inefficient ones.

Additionally: Errors also not normal

3.2 LOGIT model

We are creating a binary variable arr86, when a person gets arrested at least once. Define: arr86 = 1 if arrested in 1986 arr86 = 0 if not arrested in 1986

```
df$arr86 <- ifelse(df$narr86>0 ,1 ,0)
```

We create a Logit-Model with all variables

```
log_all <- glm(arr86 ~ pcnv + avgsen + tottime + ptime86 + qemp86 + inc86 + durat + black + hispan + bo

summary(log_all)
```

```
##
## Call:
## glm(formula = arr86 ~ pcnv + avgsen + tottime + ptime86 + qemp86 +
##     inc86 + durat + black + hispan + born60 + pcnvsq + pt86sq +
##     inc86sq, family = binomial(link = "logit"), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1656  -0.8658  -0.5644   1.1201   2.6271
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.302e-01  1.225e-01  -5.960 2.53e-09 ***
## pcnv         4.390e-01  4.348e-01   1.010 0.312619
## avgsen       2.614e-02  4.384e-02   0.596 0.550956
## tottime     -3.245e-02  3.562e-02  -0.911 0.362387
## ptime86      1.263e+00  2.523e-01   5.007 5.52e-07 ***
## qemp86       1.373e-01  5.144e-02   2.669 0.007607 **
## inc86       -1.448e-02  2.471e-03  -5.860 4.63e-09 ***
## durat        1.235e-02  1.039e-02   1.189 0.234550
## black        7.322e-01  1.209e-01   6.058 1.38e-09 ***
## hispan       4.386e-01  1.129e-01   3.886 0.000102 ***
## born60      -1.587e-02  9.635e-02  -0.165 0.869192
## pcnvsq      -1.552e+00  4.618e-01  -3.361 0.000776 ***
```

```
## pt86sq      -1.742e-01  3.911e-02  -4.453 8.48e-06 ***
## inc86sq      2.468e-05  8.186e-06   3.015 0.002570 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3216.4  on 2724  degrees of freedom
## Residual deviance: 2871.9  on 2711  degrees of freedom
## AIC: 2899.9
##
## Number of Fisher Scoring iterations: 8
```

Further Logit-Model with reduced number of variables: ################################qepmpl not included)(pcnv included) in the Latex############### ################################
##

$$Pr(arr86 = 1|X) = \frac{exp(\beta_0 + \beta_1\ pcnv + \beta_2\ ptime86 + \beta_3\ inc86 + \beta_4\ black + \beta_5\ hispan + \beta_6\ pcnvsq + \beta_7\ pt86sq + \beta_8\ inc8}{1 + exp(\beta_0 + \beta_1\ pcnv + \beta_2\ ptime86 + \beta_3\ inc86 + \beta_4\ black + \beta_5\ hispan + \beta_6\ pcnvsq + \beta_7\ pt86sq + \beta_8\ in}$$

```r
log <- glm(arr86 ~   ptime86 + qemp86 + inc86  + black + hispan + pcnvsq + pt86sq + inc86sq , data = df

summary(log)
```

```
##
## Call:
## glm(formula = arr86 ~ ptime86 + qemp86 + inc86 + black + hispan +
##     pcnvsq + pt86sq + inc86sq, family = binomial(link = "logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1653  -0.8654  -0.5673   1.1359   2.6267
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.312e-01  9.372e-02  -6.735 1.64e-11 ***
## ptime86      1.251e+00  2.467e-01   5.070 3.97e-07 ***
## qemp86       1.175e-01  4.857e-02   2.420   0.0155 *
## inc86       -1.458e-02  2.459e-03  -5.929 3.05e-09 ***
## black        7.297e-01  1.202e-01   6.073 1.26e-09 ***
## hispan       4.471e-01  1.116e-01   4.008 6.13e-05 ***
## pcnvsq      -1.114e+00  1.379e-01  -8.079 6.55e-16 ***
## pt86sq      -1.733e-01  3.847e-02  -4.504 6.67e-06 ***
## inc86sq      2.480e-05  8.170e-06   3.036   0.0024 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3216.4  on 2724  degrees of freedom
## Residual deviance: 2875.7  on 2716  degrees of freedom
```

```
## AIC: 2893.7
##
## Number of Fisher Scoring iterations: 8
```

For comparison a Probit-Model with same regressors is given:

```
prob <- glm(arr86 ~   ptime86 + qemp86 + inc86  + black + hispan + pcnvsq + pt86sq + inc86sq , data = d

summary(prob)
```

```
##
## Call:
## glm(formula = arr86 ~ ptime86 + qemp86 + inc86 + black + hispan +
##     pcnvsq + pt86sq + inc86sq, family = binomial(link = "probit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1724  -0.8682  -0.5697   1.1467   2.7138
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.917e-01  5.648e-02  -6.936 4.04e-12 ***
## ptime86      7.387e-01  1.400e-01   5.278 1.31e-07 ***
## qemp86       6.771e-02  2.898e-02   2.337  0.01944 *
## inc86       -8.503e-03  1.417e-03  -6.001 1.96e-09 ***
## black        4.373e-01  7.299e-02   5.992 2.08e-09 ***
## hispan       2.615e-01  6.643e-02   3.936 8.28e-05 ***
## pcnvsq      -6.503e-01  7.687e-02  -8.461  < 2e-16 ***
## pt86sq      -1.021e-01  2.183e-02  -4.676 2.93e-06 ***
## inc86sq      1.520e-05  4.623e-06   3.287  0.00101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3216.4  on 2724  degrees of freedom
## Residual deviance: 2876.2  on 2716  degrees of freedom
## AIC: 2894.2
##
## Number of Fisher Scoring iterations: 8
```

3.3 Models diagnostics

3.3.1 Calculation of MC Faddens pseudo R^2

```
r_log<- 1-(log$deviance/log$null.deviance)

r_prob<- 1-(prob$deviance/prob$null.deviance)
```

MC Faddens pseudo R^2 for Logit is `r_log` and for Probit it is `r_prob`.

3.3.2 Scaling of probit to logit (ptime86)

The factor between our Probit and Logit is `factor_log_prob`. And it is close to 1.6

3.3.3 Interpretation of Coefficients: Odds and Average-Marginal-Effects

```
# for logit
odds<- exp(log$coefficients)
odds
```

```
## (Intercept)     ptime86      qemp86       inc86       black      hispan
##   0.5319334   3.4933867   1.1247170   0.9855262   2.0743757   1.5637392
##      pcnvsq      pt86sq     inc86sq
##   0.3282620   0.8409137   1.0000248
```

```
fav <- mean(dnorm(predict(log,type="link")))
fav*coef(log)
```

```
##   (Intercept)       ptime86        qemp86         inc86         black
## -1.391845e-01  2.758107e-01  2.591507e-02 -3.214709e-03  1.608863e-01
##        hispan        pcnvsq        pt86sq       inc86sq
##  9.857880e-02 -2.456187e-01 -3.820432e-02  5.468947e-06
```

3.3.4 Classification table

```
tab <- table(true= df$arr86, pred= ifelse(fitted(log)>0.5,1,0))
tab
```

```
##     pred
## true    0    1
##    0 1883   87
##    1  625  130
```

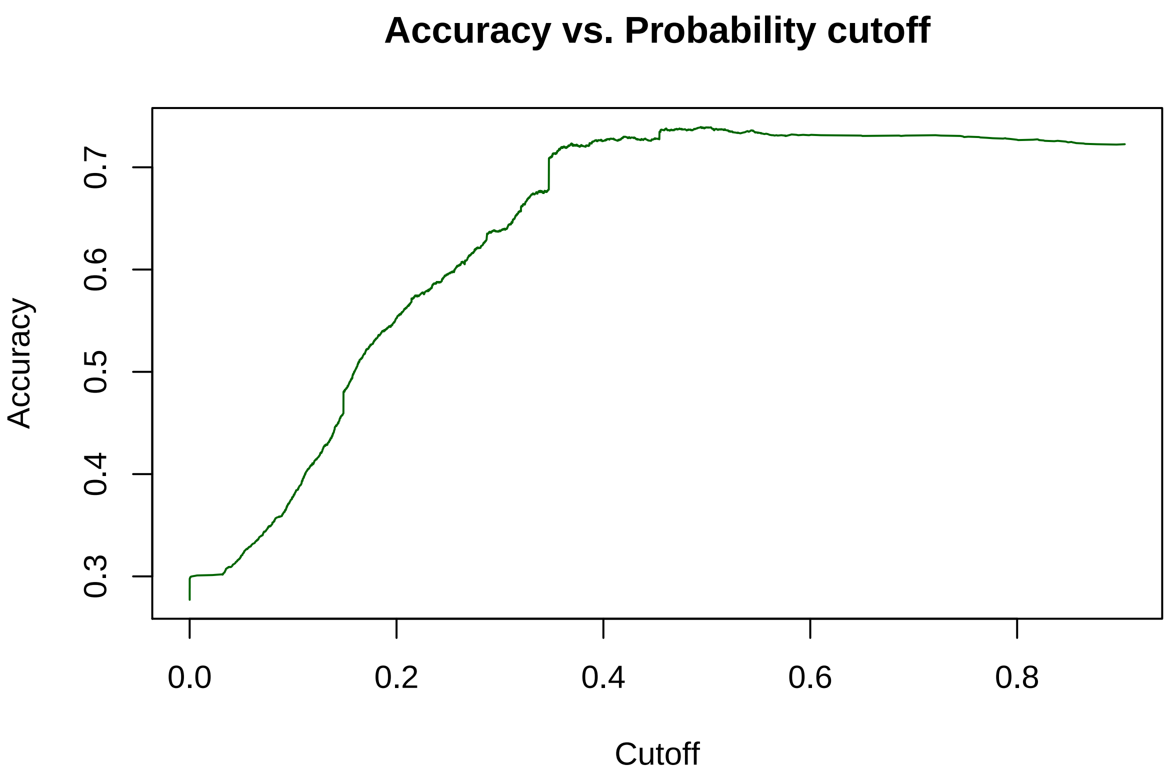
```
TP<-tab[2,2]
FP<-tab[2,1]
FN<-tab[1,2]
TN<-tab[1,1]

accuracy=(TP+TN)/length(narr86)
specificity<-TN/(FP+TN)
sensitivity<-TP/(TP+FN)
```

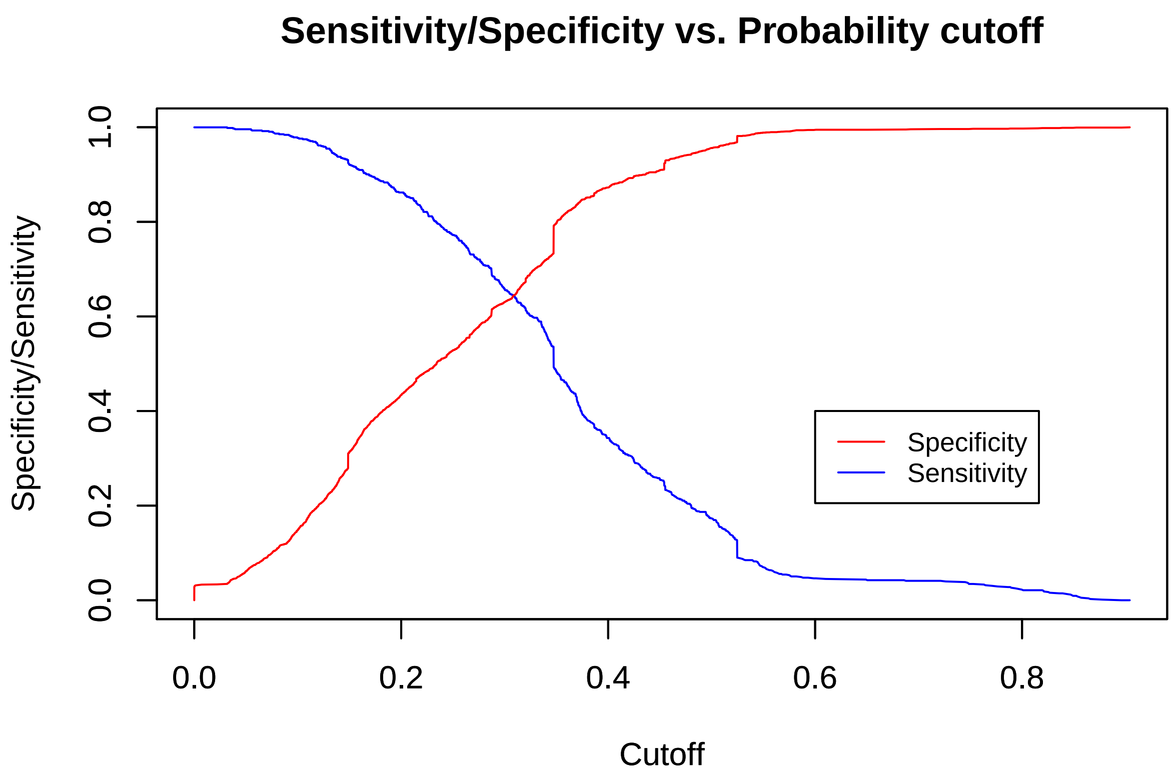
h accuracy = 0.7387156, h_0 specificity = 0.7507974 and h_1 sensitivity = 0.5990783

3.3.5 Finding Probability cutoff

```r
pred <- prediction(fitted(log),df$arr86)
plot(performance(pred, "acc"),col="darkgreen",main="Accuracy vs. Probability cutoff")
```




```r
plot(performance(pred, "sens"),col="blue",ylab="", main="Sensitivity/Specificity vs. Probability cutoff
par(new=TRUE)
plot(performance(pred, "spec"),col="red",ylab="Specificity/Sensitivity")
legend(0.6, 0.4, legend=c("Specificity", "Sensitivity"),
       col=c("red", "blue"), lty=1:1, cex=0.8)
```



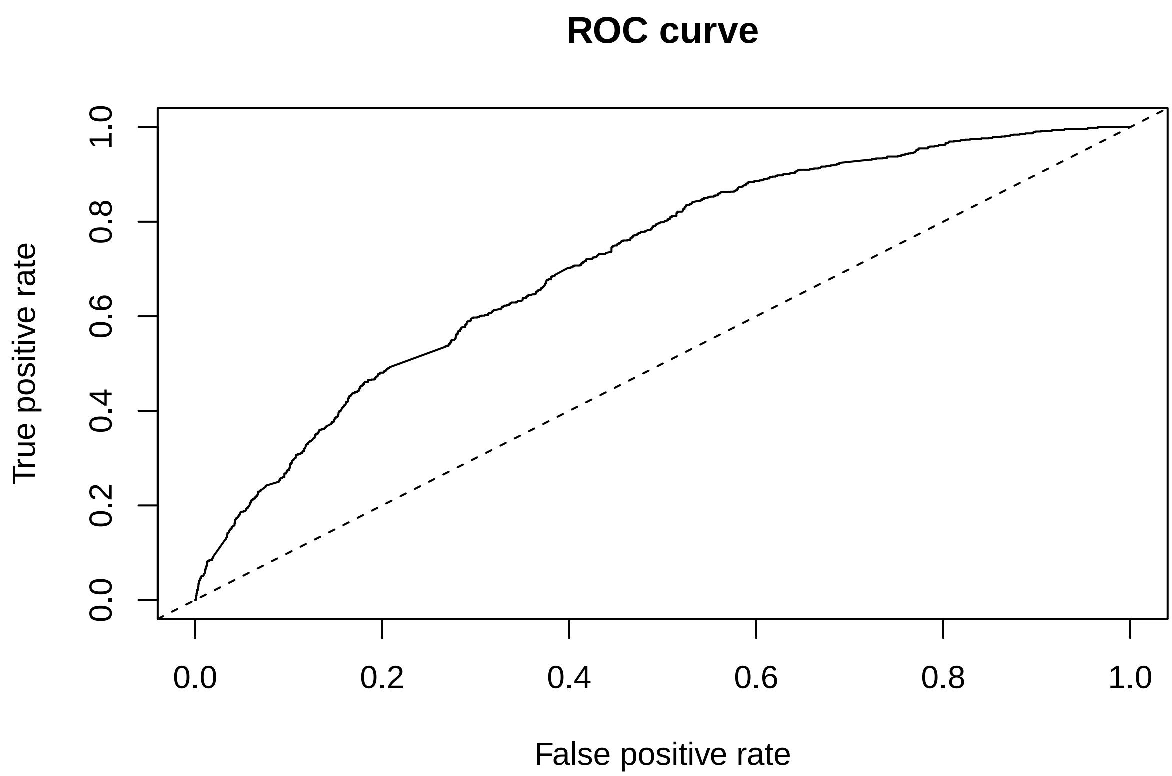

```
# -->adjusted cutoff value ... 0.3
tab_cut <- table(true= df$arr86, pred= ifelse(fitted(log)>0.3,1,0))
tab_cut
```

```
##     pred
## true    0    1
##    0 1242  728
##    1  258  497
```

3.3.6 ROC

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
## Area under the curve: 0.6885
```

The area under the ROC curve (AUC) amounts to `auc_number`.

3.4 Ordered Logit Model

Excluding narr86 > 4

```
dfn<- df %>%
  subset(df$narr<4)
head(dfn)
```

```
##   narr86 nfarr86 nparr86 pcnv avgsen tottime ptime86 qemp86 inc86 durat black
## 1      0       0       0 0.38   17.6    35.2      12      0   0.0     0     0
## 2      2       2       0 0.44    0.0     0.0       0      1   0.8     0     0
## 3      1       1       0 0.33   22.8    22.8       0      0   0.0    11     1
## 4      2       2       1 0.25    0.0     0.0       5      2   8.8     0     0
## 5      1       1       0 0.00    0.0     0.0       0      2   8.1     1     0
## 6      0       0       0 1.00    0.0     0.0       0      4  97.6     0     0
##   hispan born60 pcnvsq pt86sq     inc86sq arr86
## 1      0      1 0.1444    144    0.00000     0
## 2      1      0 0.1936      0    0.64000     1
## 3      0      1 0.1089      0    0.00000     1
## 4      1      1 0.0625     25   77.44000     1
## 5      0      0 0.0000      0   65.61001     1
## 6      0      1 1.0000      0 9525.75977     0
```

```r
results.olog<-oglmx(narr86 ~  0 + ptime86 + qemp86 + inc86  + black + hispan + pcnvsq + pt86sq + inc86s
                     delta=0,threshparam = NULL)
summary(results.olog)
```

```
## Ordered Logit Regression
## Log-Likelihood: -1879.718
## No. Iterations: 7
## McFadden's R2: 0.08034072
## AIC: 3781.436
##            Estimate  Std. error t value  Pr(>|t|)
## ptime86  1.2034e+00  2.1007e-01  5.7286 1.013e-08 ***
## qemp86   1.1647e-01  4.7887e-02  2.4323  0.015004 *
## inc86   -1.4482e-02  2.4354e-03 -5.9463 2.742e-09 ***
## black    7.1837e-01  1.1827e-01  6.0740 1.248e-09 ***
## hispan   4.6009e-01  1.1109e-01  4.1415 3.450e-05 ***
## pcnvsq  -1.0727e+00  1.3805e-01 -7.7702 7.837e-15 ***
## pt86sq  -1.6925e-01  3.4249e-02 -4.9418 7.742e-07 ***
## inc86sq  2.4796e-05  8.1111e-06  3.0570  0.002236 **
## ----- Threshold Parameters -----
##                  Estimate Std. error t value  Pr(>|t|)
## Threshold (0->1) 0.672784   0.094188   7.143 9.131e-13 ***
## Threshold (1->2) 2.552963   0.117131  21.796 < 2.2e-16 ***
## Threshold (2->3) 3.999906   0.178077  22.462 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.4.1 Marginal effects

```r
margins.oglmx(results.olog,ascontinuous = TRUE) #treating discrete variables like continuous ones, give
```

```
## Marginal Effects on Pr(Outcome==0)
##           Marg. Eff  Std. error t value  Pr(>|t|)
## ptime86 -2.0013e-01  2.9566e-02 -6.7690 1.297e-11 ***
## qemp86  -1.9371e-02  7.9601e-03 -2.4335  0.014954 *
## inc86    2.4084e-03  4.1095e-04  5.8606 4.611e-09 ***
## black   -1.1947e-01  2.0067e-02 -5.9538 2.620e-09 ***
## hispan  -7.6518e-02  1.8633e-02 -4.1065 4.016e-05 ***
## pcnvsq   1.7840e-01  2.3231e-02  7.6791 1.602e-14 ***
## pt86sq   2.8148e-02  4.7822e-03  5.8858 3.960e-09 ***
## inc86sq -4.1237e-06  1.3576e-06 -3.0375  0.002385 **
## ------------------------------------
## Marginal Effects on Pr(Outcome==1)
##           Marg. Eff  Std. error t value  Pr(>|t|)
## ptime86  1.5488e-01  2.4154e-02  6.4123 1.433e-10 ***
## qemp86   1.4991e-02  6.1684e-03  2.4303  0.015086 *
## inc86   -1.8639e-03  3.1974e-04 -5.8294 5.561e-09 ***
## black    9.2459e-02  1.5651e-02  5.9077 3.468e-09 ***
## hispan   5.9217e-02  1.4462e-02  4.0948 4.226e-05 ***
## pcnvsq  -1.3806e-01  1.8168e-02 -7.5989 2.986e-14 ***
## pt86sq  -2.1783e-02  3.8930e-03 -5.5955 2.200e-08 ***
## inc86sq  3.1914e-06  1.0519e-06  3.0340  0.002413 **
```

```
## ------------------------------------
## Marginal Effects on Pr(Outcome==2)
##           Marg. Eff  Std. error t value  Pr(>|t|)
## ptime86  3.3936e-02  5.2208e-03  6.5002 8.021e-11 ***
## qemp86   3.2847e-03  1.3806e-03  2.3792 0.0173512 *
## inc86   -4.0839e-04  7.9020e-05 -5.1682 2.364e-07 ***
## black    2.0258e-02  3.8561e-03  5.2536 1.492e-07 ***
## hispan   1.2975e-02  3.3740e-03  3.8456 0.0001203 ***
## pcnvsq  -3.0250e-02  4.7919e-03 -6.3128 2.740e-10 ***
## pt86sq  -4.7729e-03  8.1075e-04 -5.8870 3.933e-09 ***
## inc86sq  6.9925e-07  2.3911e-07  2.9244 0.0034511 **
## ------------------------------------
## Marginal Effects on Pr(Outcome==3)
##           Marg. Eff  Std. error t value  Pr(>|t|)
## ptime86  1.1314e-02  2.1670e-03  5.2209 1.781e-07 ***
## qemp86   1.0951e-03  4.8226e-04  2.2707 0.0231675 *
## inc86   -1.3615e-04  3.1761e-05 -4.2867 1.813e-05 ***
## black    6.7539e-03  1.5499e-03  4.3575 1.316e-05 ***
## hispan   4.3256e-03  1.2554e-03  3.4457 0.0005696 ***
## pcnvsq  -1.0085e-02  2.0714e-03 -4.8687 1.124e-06 ***
## pt86sq  -1.5912e-03  3.2368e-04 -4.9160 8.833e-07 ***
## inc86sq  2.3312e-07  8.5305e-08  2.7328 0.0062807 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.4.2 Alternative model with fixed thresholds (restrictions)

```
results.ologalt<-oglmx(narr86 ~  0 + ptime86 + qemp86 + inc86  + black + hispan + pcnvsq + pt86sq + inc8

"Unrestricted model"
```

```
## [1] "Unrestricted model"
```

```
summary(results.olog)
```

```
## Ordered Logit Regression
## Log-Likelihood: -1879.718
## No. Iterations: 7
## McFadden's R2: 0.08034072
## AIC: 3781.436
##            Estimate  Std. error t value  Pr(>|t|)
## ptime86  1.2034e+00  2.1007e-01  5.7286 1.013e-08 ***
## qemp86   1.1647e-01  4.7887e-02  2.4323  0.015004 *
## inc86   -1.4482e-02  2.4354e-03 -5.9463 2.742e-09 ***
## black    7.1837e-01  1.1827e-01  6.0740 1.248e-09 ***
## hispan   4.6009e-01  1.1109e-01  4.1415 3.450e-05 ***
## pcnvsq  -1.0727e+00  1.3805e-01 -7.7702 7.837e-15 ***
## pt86sq  -1.6925e-01  3.4249e-02 -4.9418 7.742e-07 ***
## inc86sq  2.4796e-05  8.1111e-06  3.0570  0.002236 **
## ----- Threshold Parameters -----
##                  Estimate Std. error t value  Pr(>|t|)
```

```
## Threshold (0->1) 0.672784   0.094188   7.143 9.131e-13 ***
## Threshold (1->2) 2.552963   0.117131  21.796 < 2.2e-16 ***
## Threshold (2->3) 3.999906   0.178077  22.462 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
"Alternative model with fixed thresholds"
```

```
## [1] "Alternative model with fixed thresholds"
```

```r
summary(results.ologalt)
```

```
## Ordered Logit Regression
## Log-Likelihood: -1926.135
## No. Iterations: 8
## McFadden's R2: 0.05763094
## AIC: 3870.27
## ----- Mean Equation ------
##            Estimate  Std. error t value  Pr(>|t|)
## ptime86  7.8028e-01  1.2531e-01  6.2266 4.766e-10 ***
## qemp86   1.8333e-01  2.5754e-02  7.1185 1.091e-12 ***
## inc86   -9.4246e-03  1.5021e-03 -6.2744 3.509e-10 ***
## black    6.4795e-01  6.5476e-02  9.8959 < 2.2e-16 ***
## hispan   4.4269e-01  6.4611e-02  6.8515 7.305e-12 ***
## pcnvsq  -4.5389e-01  8.4327e-02 -5.3825 7.344e-08 ***
## pt86sq  -1.0628e-01  2.0740e-02 -5.1242 2.989e-07 ***
## inc86sq  1.5782e-05  4.9918e-06  3.1615  0.001569 **
## ----- SD Equation ------
##     Estimate Std. error t value  Pr(>|t|)
## NA -0.488715   0.032473  -15.05 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.4.3 Likelihoodratio-Test to compare unrestricted and restricted model

```r
library("lmtest")

lrtest(results.olog,results.ologalt)
```

```
## Likelihood ratio test
##
## Model 1: narr86 ~ 0 + ptime86 + qemp86 + inc86 + black + hispan + pcnvsq +
##     pt86sq + inc86sq
## Model 2: narr86 ~ 0 + ptime86 + qemp86 + inc86 + black + hispan + pcnvsq +
##     pt86sq + inc86sq
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  11 -1879.7
## 2   9 -1926.1 -2 92.834  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```