Thesis: Causal Analysis

Laurens van der Tas

Initialize: Load R environment, data and packages

In this document we will investigate the causal relationships in our dataset using regression analysis.

```
#Set working directory
#setwd("C:/Users/Laurens/Dropbox/University/Year 4/Period 2/Applied Economics Research Course/Thesis/da

#Load required packages
library(foreign)
library(stargazer)
library(ggplot2)
library(aod)
library(gridExtra)
library(ggthemes)
library(dplyr)
library(mfx)
library(corrplot)
library(car)
library(LogisticDx)
library(rms)

#Enable anti-aliasing for rendered graphics
library(knitr)
#opts_chunk$set(out.width = '\\maxwidth')
#dev = "CairoPNG",

#Load dataset
data.dropout <- read.dta("DatasetTrimmed.dta")
  #Read name vector of dataset
  names(data.dropout)

#Factorize binaries
data.dropout$geslachtBin <- factor(data.dropout$geslachtBin, labels = c("Female", "Male"))
data.dropout$allochtoonBin <- factor(data.dropout$allochtoonBin, labels = c("No", "Yes"))
```

Regression Models

In this section all the investigated models, as well as the obsolete ones, are listed.

LPM Model

For testing purposes.

```
#TRIVIAL: LPM model of Big Five on dropout
lpm.1 <- lm(vrv_1 ~ zorgvuldig + stabiel + open + extravert + altrusme + allochtoonBin + geslachtBin,
            data = data.dropout)
#Display results
summary(lpm.1)
```

First Logit Model

Neglects the theory in the sense that it disregards *gender* and *foreign origin* as control variables.

```
logit.1 <- glm(vrv_1 ~ stabiel + open + zorgvuldig + extravert + altrusme + age_opleiding,
               family = binomial(logit),
               data = data.dropout)

#Display results
summary(logit.1)
```

```
##
## Call:
## glm(formula = vrv_1 ~ stabiel + open + zorgvuldig + extravert +
##     altrusme + age_opleiding, family = binomial(logit), data = data.dropout)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8954  -1.2830   0.8695   0.9972   1.2225
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.02021    0.53107   0.038 0.969644
## stabiel       -0.03668    0.08695  -0.422 0.673153
## open          -0.01039    0.10280  -0.101 0.919523
## zorgvuldig    -0.20678    0.09600  -2.154 0.031244 *
## extravert      0.07932    0.10499   0.755 0.449957
## altrusme       0.01165    0.12152   0.096 0.923644
## age_opleiding  0.05571    0.01645   3.386 0.000708 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 659.08  on 498  degrees of freedom
## Residual deviance: 639.80  on 492  degrees of freedom
## AIC: 653.8
##
## Number of Fisher Scoring iterations: 4
```

```
#Calculate odds ratio vector
coef.vector.1 <- exp(logit.1$coef)

#R squared and Wald test significance
require(rms)
lrm(logit.1)
```

```
## Logistic Regression Model
##
##  lrm(formula = logit.1)
##
##
##                      Model Likelihood      Discrimination      Rank Discrim.
##                        Ratio Test              Indexes             Indexes
##  Obs          499    LR chi2      19.28    R2       0.052    C       0.599
##   0           186    d.f.             6    g        0.467    Dxy     0.199
##   1           313    Pr(> chi2) 0.0037    gr       1.595    gamma   0.200
```

```
##  max |deriv| 7e-06                          gp       0.098    tau-a   0.093
##                                             Brier    0.226
##
##              Coef    S.E.    Wald Z Pr(>|Z|)
##  Intercept    0.0202 0.5311   0.04  0.9696
##  stabiel     -0.0367 0.0870  -0.42  0.6732
##  open        -0.0104 0.1028  -0.10  0.9195
##  zorgvuldig  -0.2068 0.0960  -2.15  0.0312
##  extravert    0.0793 0.1050   0.76  0.4500
##  altrusme     0.0116 0.1215   0.10  0.9236
##  age_opleiding 0.0557 0.0165  3.39  0.0007
##
```

```r
#Marginal effects
require(mfx)
mfx.1 <- logitmfx(logit.1, data = data.dropout)

#Variance inflation factor: multicollinearity test
require(car)
vif(logit.1)
```

```
##      stabiel          open    zorgvuldig      extravert      altrusme
##     1.183144      1.257312      1.223593      1.428105      1.182988
## age_opleiding
##     1.036169
```

```r
sqrt(vif(logit.1)) > 2 #Bigger than 2 (or sometimes 2.5) signals relatively high multicollinearity
```

```
##      stabiel          open    zorgvuldig      extravert      altrusme
##        FALSE         FALSE         FALSE         FALSE         FALSE
## age_opleiding
##        FALSE
```

```r
#Test for outliers
outlierTest(logit.1)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferonni p
## 406 -1.929694          0.053645           NA
```

```r
#Cook's distance
#plot(logit.1, which = 4, main = "Model One")

ggplot(aes(x = seq_along(.cooksd), y = .cooksd), data = logit.1) +
  geom_bar(stat = "identity") +
  theme_bw(base_family = "serif") +
  labs(title = "Model One: Cook's Distance",
       x = "Observation Number",
       y = "Cook's Distance") +
  geom_hline(yintercept = 0.0081, colour = 2, alpha = 0.75, linetype = 5)
```
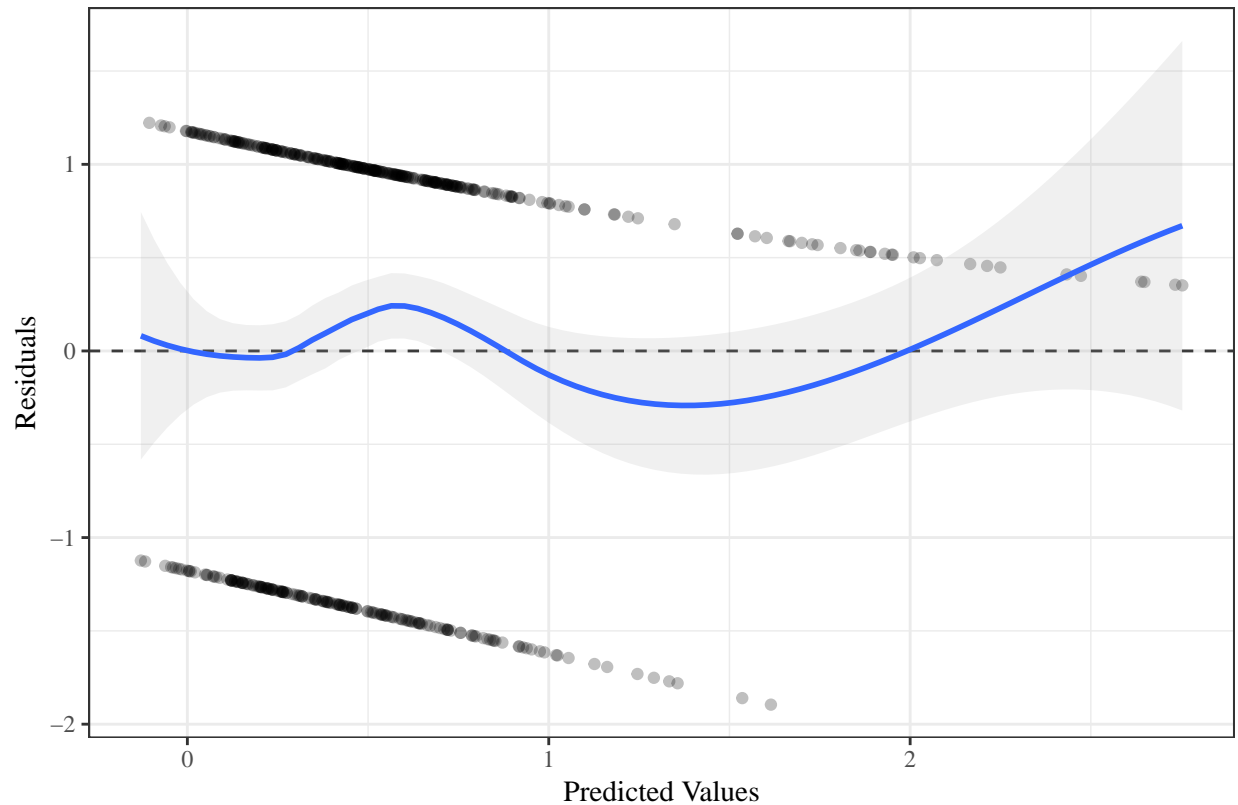
Model One: Cook's Distance
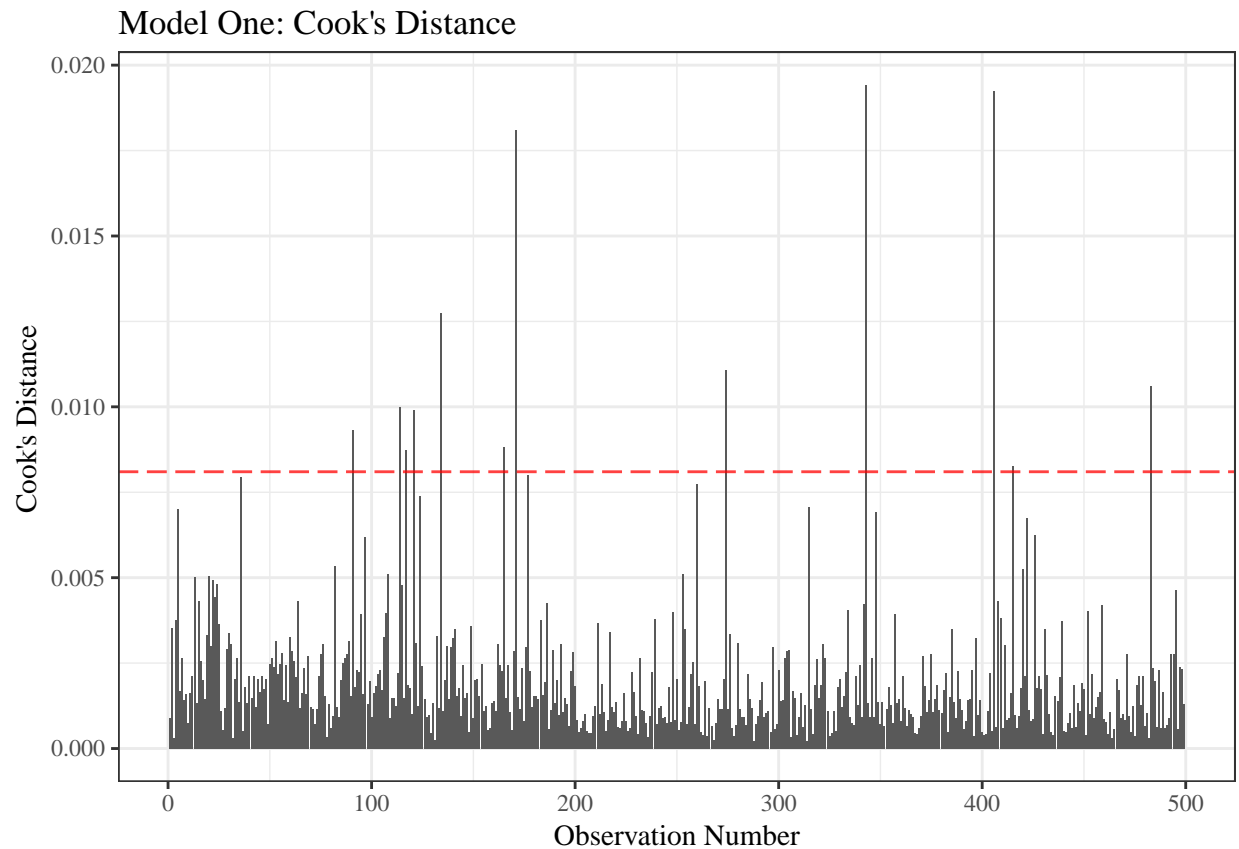

```
#Resid vs fitted
#plot(logit.1, which = 1, main = "Model One")

ggplot(aes(x = .fitted, y = .resid), data = logit.1) +
  geom_point(alpha = 1/4) +
  geom_hline(yintercept = 0, linetype = 2, alpha = 0.75) +
  geom_smooth(alpha = 0.15, method = "loess") +
  theme_bw(base_family = "serif") +
  labs(title = "Model One: Residuals vs. Fitted",
       x = "Predicted Values",
       y = "Residuals")
```

Model One: Residuals vs. Fitted


```r
#Model fitted on covariates
ggplot(aes(x = data.dropout$zorgvuldig, y = .resid), data = logit.1) +
  theme_bw(base_family = "serif") +
  geom_smooth(aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 1/6) +
  geom_point(aes(x = data.dropout$open+0.1, y = .resid), colour = 2, alpha = 1/6) +
  geom_point(aes(x = data.dropout$stabiel+0.2, y = .resid), colour = 3, alpha = 1/6) +
  geom_point(aes(x = data.dropout$extravert+0.3, y = .resid), colour = 4, alpha = 1/6) +
  geom_point(aes(x = data.dropout$altrusme+0.4, y = .resid), colour = 5, alpha = 1/6)
```

```
## `geom_smooth()` using method = 'loess'
```
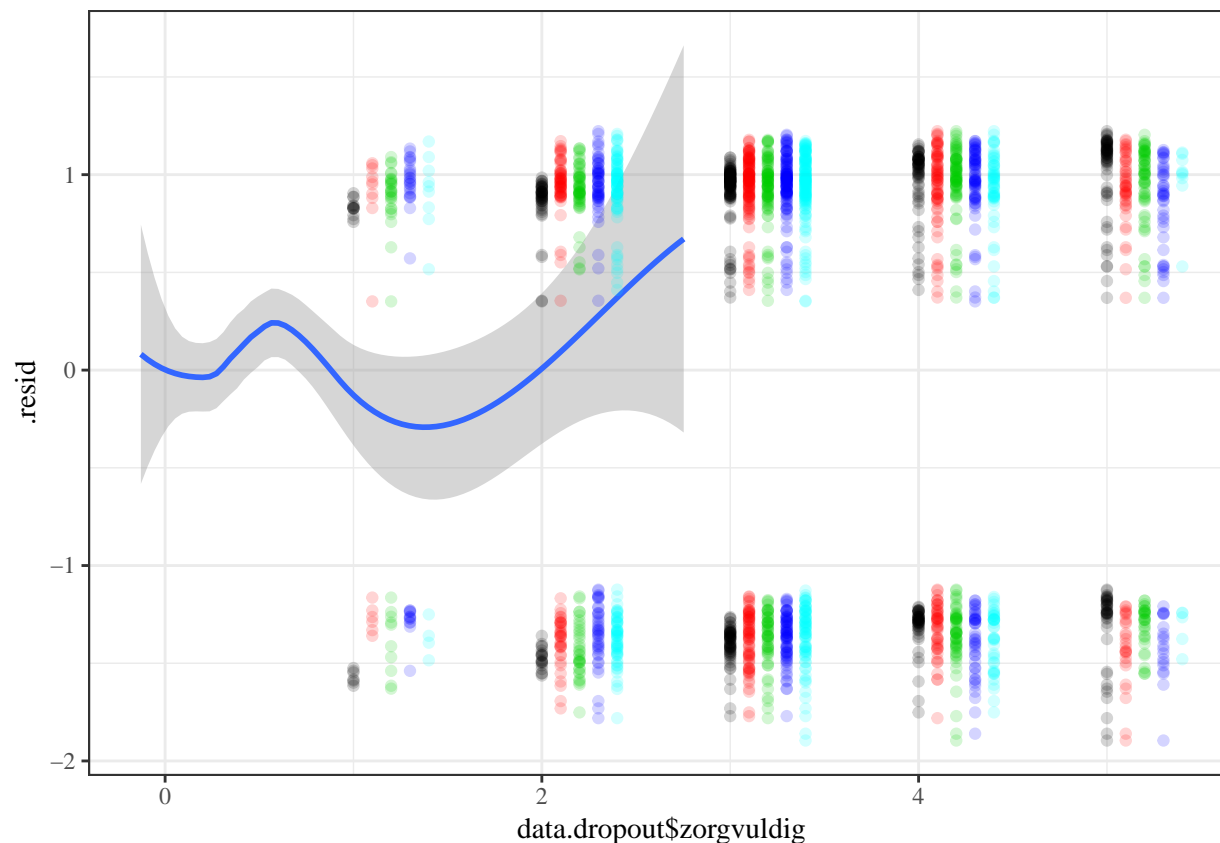
Second Logit Model

Added *gender* and *foreign origin* as control variables to the first logit model.

```
logit.2 <- glm(vrv_1 ~ stabiel + open + zorgvuldig + extravert + altrusme + age_opleiding + geslachtBin
               allochtoonBin,
               family = binomial(logit),
               data = data.dropout)

#Display results
summary(logit.2)
```

```
##
## Call:
## glm(formula = vrv_1 ~ stabiel + open + zorgvuldig + extravert +
##     altrusme + age_opleiding + geslachtBin + allochtoonBin, family = binomial(logit),
##     data = data.dropout)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8665  -1.2884   0.8566   0.9997   1.2473
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.04660    0.53913  -0.086 0.931121
## stabiel         -0.05222    0.08969  -0.582 0.560368
```

```
## open              -0.02572      0.10521   -0.244 0.806885
## zorgvuldig        -0.19280      0.09915   -1.944 0.051836 .
## extravert          0.08508      0.10532    0.808 0.419199
## altrusme           0.01471      0.12168    0.121 0.903769
## age_opleiding      0.05646      0.01652    3.418 0.000631 ***
## geslachtBinMale    0.15153      0.20361    0.744 0.456721
## allochtoonBinYes   0.06659      0.22302    0.299 0.765271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 659.08  on 498  degrees of freedom
## Residual deviance: 639.19  on 490  degrees of freedom
## AIC: 657.19
##
## Number of Fisher Scoring iterations: 4
```

```
#R squared and Wald test significance
require(rms)
lrm(logit.2)
```

```
## Logistic Regression Model
##
##  lrm(formula = logit.2)
##
##                       Model Likelihood      Discrimination    Rank Discrim.
##                          Ratio Test              Indexes          Indexes
##  Obs          499      LR chi2      19.89    R2     0.053    C      0.601
##   0           186      d.f.             8    g      0.476    Dxy    0.201
##   1           313      Pr(> chi2) 0.0108    gr     1.610    gamma  0.203
##  max |deriv| 8e-06                          gp     0.100    tau-a  0.094
##                                             Brier  0.226
##
##
##                   Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept       -0.0466 0.5391 -0.09  0.9311
##  stabiel         -0.0522 0.0897 -0.58  0.5604
##  open            -0.0257 0.1052 -0.24  0.8069
##  zorgvuldig      -0.1928 0.0992 -1.94  0.0518
##  extravert        0.0851 0.1053  0.81  0.4192
##  altrusme         0.0147 0.1217  0.12  0.9038
##  age_opleiding    0.0565 0.0165  3.42  0.0006
##  geslachtBin=Male 0.1515 0.2036  0.74  0.4567
##  allochtoonBin=Yes 0.0666 0.2230  0.30  0.7653
##
```

```
#Marginal effects
require(mfx)
logitmfx(logit.2, data = data.dropout)
```

```
## Call:
## logitmfx(formula = logit.2, data = data.dropout)
##
## Marginal Effects:
##                      dF/dx   Std. Err.        z      P>|z|
## stabiel         -0.0120952  0.0207698 -0.5823 0.5603351
```

```
## open            -0.0059565  0.0243658 -0.2445 0.8068734
## zorgvuldig      -0.0446532  0.0229390 -1.9466 0.0515822 .
## extravert        0.0197039  0.0243864  0.8080 0.4190991
## altrusme         0.0034071  0.0281810  0.1209 0.9037686
## age_opleiding    0.0130772  0.0037752  3.4640 0.0005322 ***
## geslachtBinMale  0.0349592  0.0467581  0.7477 0.4546648
## allochtoonBinYes 0.0153493  0.0511618  0.3000 0.7641653
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "geslachtBinMale"  "allochtoonBinYes"
```

```r
#Variance inflation factor: multicollinearity test
require(car)
vif(logit.2)
```

```
##         stabiel             open       zorgvuldig         extravert
##        1.257415         1.315879         1.303901          1.436032
##        altrusme    age_opleiding  geslachtBinMale allochtoonBinYes
##        1.183412         1.040891         1.124696          1.054038
```

```r
sqrt(vif(logit.2)) > 2
```

```
##         stabiel             open       zorgvuldig         extravert
##           FALSE            FALSE            FALSE             FALSE
##        altrusme    age_opleiding  geslachtBinMale allochtoonBinYes
##           FALSE            FALSE            FALSE             FALSE
```

```r
#Test for outliers
outlierTest(logit.2)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 406 -1.902455           0.057112           NA
```

Third Logit Model

Added interaction effect between emotional *stability* and *conscientiousness*, as suggested by the psychologist referred to by Yolanda.

```r
logit.3 <- glm(vrv_1 ~ stabiel + open + zorgvuldig + extravert + altrusme + age_opleiding + stabiel:zorg
               family = binomial(logit),
               data = data.dropout)
```

```r
#Display results
summary(logit.3)
```

```
##
## Call:
## glm(formula = vrv_1 ~ stabiel + open + zorgvuldig + extravert +
##     altrusme + age_opleiding + stabiel:zorgvuldig, family = binomial(logit),
##     data = data.dropout)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9140  -1.2881   0.8516   1.0022   1.2020
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.859716   0.989461   0.869 0.384917
## stabiel           -0.290425   0.265614  -1.093 0.274213
## open              -0.020262   0.103352  -0.196 0.844575
## zorgvuldig        -0.441306   0.251862  -1.752 0.079744 .
## extravert          0.076597   0.105109   0.729 0.466160
## altrusme           0.006072   0.121729   0.050 0.960217
## age_opleiding      0.055556   0.016508   3.365 0.000764 ***
## stabiel:zorgvuldig 0.073843   0.072905   1.013 0.311126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 659.08  on 498  degrees of freedom
## Residual deviance: 638.77  on 491  degrees of freedom
## AIC: 654.77
##
## Number of Fisher Scoring iterations: 4
```

```r
#R squared and Wald test significance
require(rms)
lrm(logit.3)
```

```
## Logistic Regression Model
##
##  lrm(formula = logit.3)
##
##                    Model Likelihood      Discrimination    Rank Discrim.
##                       Ratio Test            Indexes            Indexes
##  Obs         499    LR chi2      20.31  R2       0.054   C      0.605
##   0          186    d.f.             7  g        0.481   Dxy    0.210
##   1          313    Pr(> chi2) 0.0049  gr       1.618   gamma  0.211
##  max |deriv| 8e-06                      gp       0.100   tau-a  0.098
##                                         Brier    0.226
##
##
##                       Coef     S.E.    Wald Z Pr(>|Z|)
##  Intercept            0.8597  0.9895   0.87   0.3849
##  stabiel             -0.2904  0.2656  -1.09   0.2742
##  open                -0.0203  0.1034  -0.20   0.8446
##  zorgvuldig          -0.4413  0.2519  -1.75   0.0797
##  extravert            0.0766  0.1051   0.73   0.4662
##  altrusme             0.0061  0.1217   0.05   0.9602
##  age_opleiding        0.0556  0.0165   3.37   0.0008
##  stabiel * zorgvuldig 0.0738  0.0729   1.01   0.3111
##
```

```r
#Interpret marginal effects
require(mfx)
logitmfx(logit.3, data = data.dropout)
```

```
## Call:
## logitmfx(formula = logit.3, data = data.dropout)
##
## Marginal Effects:
##                         dF/dx   Std. Err.        z   P>|z|
## stabiel             -0.0672418  0.0614541  -1.0942 0.273876
## open                -0.0046911  0.0239275  -0.1961 0.844566
## zorgvuldig          -0.1021752  0.0582265  -1.7548 0.079296 .
## extravert            0.0177345  0.0243313   0.7289 0.466077
## altrusme             0.0014058  0.0281838   0.0499 0.960217
## age_opleiding        0.0128629  0.0037718   3.4103 0.000649 ***
## stabiel:zorgvuldig   0.0170968  0.0168691   1.0135 0.310823
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Variance inflation factor: multicollinearity test
require(car)
vif(logit.3)
```

```
##            stabiel               open          zorgvuldig
##          10.899160           1.270075            8.289497
##           extravert           altrusme        age_opleiding
##           1.424272           1.184704            1.035412
## stabiel:zorgvuldig
##          22.659327
```

```r
sqrt(vif(logit.3)) > 2
```

```
##            stabiel               open          zorgvuldig
##               TRUE              FALSE                TRUE
##           extravert           altrusme        age_opleiding
##              FALSE              FALSE               FALSE
## stabiel:zorgvuldig
##               TRUE
```

```r
#Test for outliers
outlierTest(logit.3)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferonni p
## 406 -1.949069           0.051287           NA
```

Fourth Logit Model

Dropped all the outliers for *age*: bigger than 35. Resulting in a sample size of 466 (before: 499).

```r
logit.4 <- glm(vrv_1 ~ stabiel + open + zorgvuldig + extravert + altrusme + age_opleiding,
               family = binomial(logit),
               data = subset(data.dropout, !age_opleiding > 35))

#Display results
summary(logit.4)
```

```
##
## Call:
```

```
## glm(formula = vrv_1 ~ stabiel + open + zorgvuldig + extravert +
##     altrusme + age_opleiding, family = binomial(logit), data = subset(data.dropout,
##     !age_opleiding > 35))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5952  -1.3122   0.9136   1.0067   1.1694
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.457961   0.638476   0.717   0.4732
## stabiel       -0.045877   0.087652  -0.523   0.6007
## open           0.023058   0.103870   0.222   0.8243
## zorgvuldig    -0.171712   0.096692  -1.776   0.0758 .
## extravert      0.069962   0.106084   0.659   0.5096
## altrusme      -0.003223   0.123037  -0.026   0.9791
## age_opleiding  0.023985   0.026921   0.891   0.3730
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 624.39  on 465  degrees of freedom
## Residual deviance: 619.72  on 459  degrees of freedom
## AIC: 633.72
##
## Number of Fisher Scoring iterations: 4
```

```r
#R squared and Wald test significance
require(rms)
lrm(logit.4)
```

```
## Logistic Regression Model
##
##  lrm(formula = logit.4)
##
##
##                      Model Likelihood     Discrimination    Rank Discrim.
##                        Ratio Test            Indexes          Indexes
##  Obs          466    LR chi2      4.67   R2      0.014   C      0.559
##   0           183    d.f.            6   g       0.236   Dxy    0.117
##   1           283    Pr(> chi2) 0.5869   gr      1.267   gamma  0.118
##  max |deriv| 2e-08                       gp      0.056   tau-a  0.056
##                                          Brier   0.236
##
##
##              Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept    0.4580 0.6385   0.72  0.4732
##  stabiel     -0.0459 0.0877  -0.52  0.6007
##  open         0.0231 0.1039   0.22  0.8243
##  zorgvuldig  -0.1717 0.0967  -1.78  0.0758
##  extravert    0.0700 0.1061   0.66  0.5096
##  altrusme    -0.0032 0.1230  -0.03  0.9791
##  age_opleiding 0.0240 0.0269   0.89  0.3730
##
```

```
#Marginal effects
require(mfx)
logitmfx(logit.4, data = data.dropout)
```

```
## Call:
## logitmfx(formula = logit.4, data = data.dropout)
##
## Marginal Effects:
##                   dF/dx   Std. Err.        z     P>|z|
## stabiel       -0.0084973  0.0201438 -0.4218 0.6731477
## open          -0.0024060  0.0238133 -0.1010 0.9195210
## zorgvuldig    -0.0479037  0.0222079 -2.1571 0.0310010 *
## extravert      0.0183749  0.0243173  0.7556 0.4498710
## altrusme       0.0026981  0.0281509  0.0958 0.9236436
## age_opleiding  0.0129064  0.0037613  3.4314 0.0006005 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Variance inflation factor: multicollinearity test
require(car)
vif(logit.4)
```

```
##        stabiel           open      zorgvuldig       extravert       altrusme
##       1.175143       1.249847        1.218655        1.411745       1.185207
## age_opleiding
##       1.037346
```

```
sqrt(vif(logit.4)) > 2
```

```
##        stabiel           open      zorgvuldig       extravert       altrusme
##          FALSE          FALSE           FALSE           FALSE          FALSE
## age_opleiding
##          FALSE
```

```
#Test for outliers
outlierTest(logit.4)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferonni p
## 483 -1.617687            0.10573           NA
```

Fifth Logit Model

Our definitive, kick-ass model.

```
logit.5 <- glm(vrv_1 ~ zorgvuldig + open + stabiel + extravert + altrusme + zorgvuldig:allochtoonBin + 
               family = binomial(logit),
               data = data.dropout)
```

```
#Display results
summary(logit.5)
```

```
##
## Call:
## glm(formula = vrv_1 ~ zorgvuldig + open + stabiel + extravert +
```

```
##       altrusme + zorgvuldig:allochtoonBin + open:allochtoonBin +
##       stabiel:allochtoonBin + allochtoonBin + age_opleiding + leerintelligentie,
##       family = binomial(logit), data = data.dropout)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1111  -1.2039   0.7125   0.9641   1.6623
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.22638    0.68232  -0.332 0.740052
## zorgvuldig                  -0.35826    0.11199  -3.199 0.001379 **
## open                         0.09935    0.12336   0.805 0.420595
## stabiel                     -0.15678    0.10259  -1.528 0.126463
## extravert                    0.05850    0.10827   0.540 0.588996
## altrusme                     0.05489    0.12679   0.433 0.665057
## allochtoonBinYes            -1.56605    1.04383  -1.500 0.133540
## age_opleiding                0.05777    0.01711   3.377 0.000732 ***
## leerintelligentie            0.20454    0.11260   1.817 0.069283 .
## zorgvuldig:allochtoonBinYes  0.73863    0.25009   2.953 0.003143 **
## open:allochtoonBinYes       -0.63185    0.24154  -2.616 0.008898 **
## stabiel:allochtoonBinYes     0.40362    0.21439   1.883 0.059748 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 659.08  on 498  degrees of freedom
## Residual deviance: 613.36  on 487  degrees of freedom
## AIC: 637.36
##
## Number of Fisher Scoring iterations: 4
```

```r
#Calculate odds ratio vector
coef.vector.5 <- exp(logit.5$coef)

#R squared and Wald test significance
require(rms)
lrm(logit.5)
```

```
## Logistic Regression Model
##
##  lrm(formula = logit.5)
##
##                         Model Likelihood    Discrimination    Rank Discrim.
##                           Ratio Test           Indexes           Indexes
## Obs           499    LR chi2      45.72    R2       0.119    C       0.669
##   0           186    d.f.            11    g        0.774    Dxy     0.337
##   1           313    Pr(> chi2) <0.0001    gr       2.169    gamma   0.339
## max |deriv| 3e-05                          gp       0.162    tau-a   0.158
##                                            Brier    0.214
##
##
##                          Coef    S.E.   Wald Z Pr(>|Z|)
## Intercept              -0.2264 0.6823 -0.33   0.7401
## zorgvuldig             -0.3583 0.1120 -3.20   0.0014
```

```
##   open                             0.0994 0.1234  0.81  0.4206
##   stabiel                         -0.1568 0.1026 -1.53  0.1265
##   extravert                        0.0585 0.1083  0.54  0.5890
##   altrusme                         0.0549 0.1268  0.43  0.6651
##   allochtoonBin=Yes               -1.5660 1.0438 -1.50  0.1335
##   age_opleiding                    0.0578 0.0171  3.38  0.0007
##   leerintelligentie                0.2045 0.1126  1.82  0.0693
##   zorgvuldig * allochtoonBin=Yes   0.7386 0.2501  2.95  0.0031
##   open * allochtoonBin=Yes        -0.6319 0.2415 -2.62  0.0089
##   stabiel * allochtoonBin=Yes      0.4036 0.2144  1.88  0.0597
##
```

```r
#Marginal effects
require(mfx)
mfx.5 <- logitmfx(logit.5, data = data.dropout)

#Variance inflation factor: multicollinearity test
require(car)
vif(logit.5)
```

```
##              zorgvuldig                         open
##                1.517166                     1.703091
##                 stabiel                     extravert
##                1.518127                     1.454943
##                altrusme              allochtoonBinYes
##                1.214333                    21.299938
##           age_opleiding             leerintelligentie
##                1.055990                     1.260560
## zorgvuldig:allochtoonBinYes     open:allochtoonBinYes
##               18.119427                    16.813727
##    stabiel:allochtoonBinYes
##               11.301315
```

```r
sqrt(vif(logit.5)) > 2
```

```
##              zorgvuldig                         open
##                   FALSE                        FALSE
##                 stabiel                     extravert
##                   FALSE                        FALSE
##                altrusme              allochtoonBinYes
##                   FALSE                         TRUE
##           age_opleiding             leerintelligentie
##                   FALSE                        FALSE
## zorgvuldig:allochtoonBinYes     open:allochtoonBinYes
##                    TRUE                         TRUE
##    stabiel:allochtoonBinYes
##                    TRUE
```

```r
#Test for outliers
outlierTest(logit.5)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 274 -2.167205           0.030219           NA
```

```
#Cook's distance
#plot(logit.5, which = 4, main = "Model Two")

ggplot(aes(x = seq_along(.cooksd), y = .cooksd), data = logit.5) +
  geom_bar(stat = "identity") +
  theme_bw(base_family = "serif") +
  labs(title = "Model Two: Cook's Distance",
       x = "Observation Number",
       y = "Cook's Distance") +
  geom_hline(yintercept = 0.0082, colour = 2, alpha = 0.75, linetype = 5)
```

Model Two: Cook's Distance
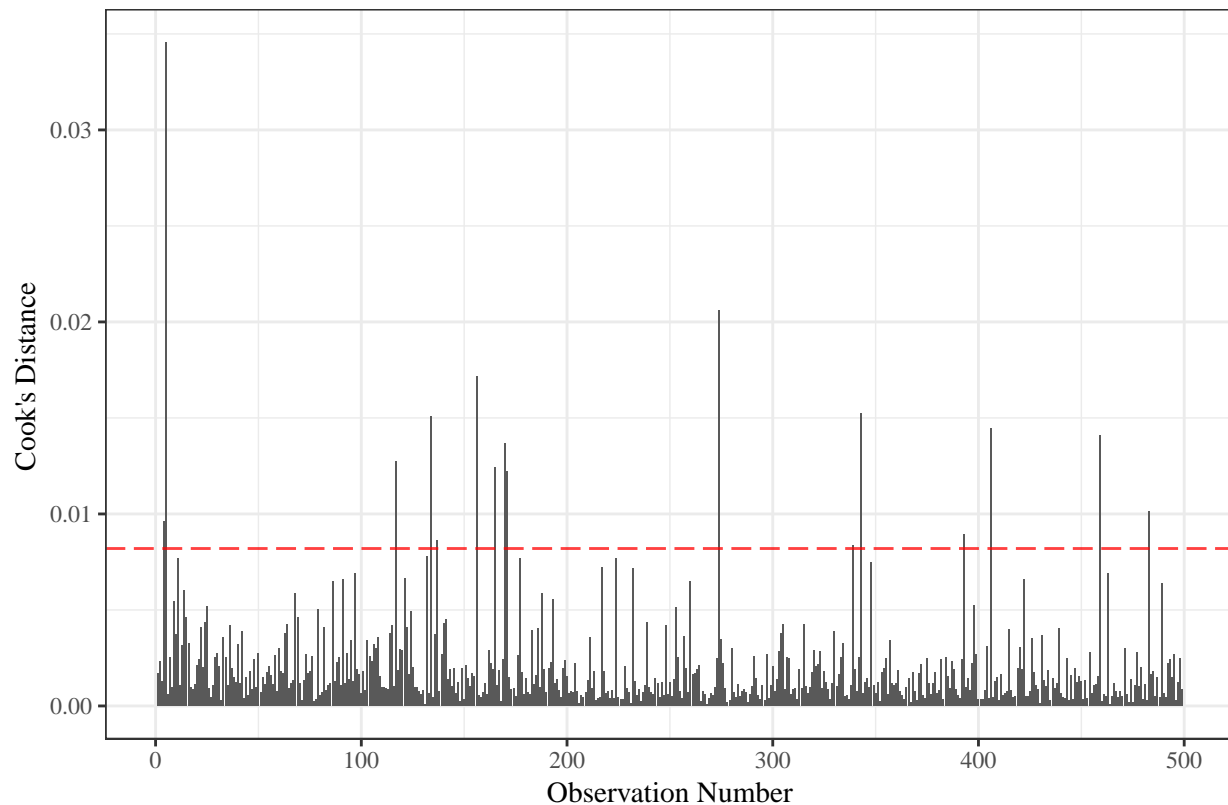

```
#Resid vs fitted
#plot(logit.5, which = 1, main = "Model Two")

ggplot(aes(x = .fitted, y = .resid), data = logit.5) +
  geom_point(alpha = 1/4) +
  geom_hline(yintercept = 0, linetype = 2, alpha = 0.75) +
  geom_smooth(alpha = 0.15, method = "loess") +
  theme_bw(base_family = "serif") +
  labs(title = "Model Two: Residuals vs. Fitted",
       x = "Predicted Values",
       y = "Residuals")
```

Model Two: Residuals vs. Fitted