Applied Logistic Regression – Assignment 4

Juho Ruohonen

DateComesHere, 2017

1. Residuals in Logistic Regression

We'll go back to the dataset on Sentiments Toward Racial Integration in Public Housing. That dataset is currently in so-called Bernoulli format, in which one row corresponds to one binary observation. To ease the following computation of residuals, we will convert it into binomial format, which is better for assessing the accuracy of the model predictions for each combination of predictor values.

```
w<-aggregate(PositiveSentiment ~ Contact + Norms, FUN = sum, data=contact)
n<-aggregate(PositiveSentiment ~ Contact + Norms, FUN = length, data=contact)
(Contact<-data.frame(Contact=w$Contact, Norms=w$Norms, Y=w$PositiveSentiment,
  Obs.=n[,3], p=w$PositiveSentiment/n[,3]))
```

```
##   Contact       Norms   Y Obs.         p
## 1       0   Favorable  41   96 0.4270833
## 2       1   Favorable 120  172 0.6976744
## 3       0 Unfavorable  56  201 0.2786070
## 4       1 Unfavorable  66  139 0.4748201
```

```
summary(Model<-glm(Y/Obs. ~ Contact + Norms, weights = Obs.,
  family=binomial(link=logit), data=Contact))
```

```
##
## Call:
## glm(formula = Y/Obs. ~ Contact + Norms, family = binomial(link = logit),
##     data = Contact, weights = Obs.)
##
## Deviance Residuals:
##       1        2        3        4
## -0.4652   0.3729   0.3592  -0.3842
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.1980     0.1664  -1.189    0.234
## Contact            0.9726     0.1742   5.583 2.37e-08 ***
## NormsUnfavorable  -0.8102     0.1747  -4.638 3.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 67.93867  on 3  degrees of freedom
## Residual deviance:  0.63214  on 1  degrees of freedom
## AIC: 27.992
##
## Number of Fisher Scoring iterations: 3
```

In logistic regression, the residual of a single observation is of limited interest due to its dichotomous nature – with the response limited to 0 or 1, its residual is also limited to theoretical maximum of 1. Therefore, rather than examining residuals of individual observations, it is more illuminating to group the data by each distinct *covariate pattern* – also known as explanatory variable pattern or EVP – i.e., each distinct combination of predictor values observed in the data. Each group sharing an EVP also necessarily shares the same predicted probability. Multiplying each group's n by the predicted probability yields the expected number of success outcomes for that group, which can then be compared to the observed success count. Assuming n is sufficiently large for a given covariate pattern, the difference between the observed and expected success counts is asymptotically standard normal.

Grouping by EVP requires that the explanatory variables be categorical. The presence of a truly continuous covariate in the model has the consequence that the number of EVPs approaches or equals the total number of observations in the dataset. In such scenarios, we need the Hosmer-Lemeshow procedure, which groups the data by deciles of predicted probability instead of covariate patterns. Fortunately, both predictors in our racial integration model are dichotomous. We will now calculate both the Pearson residuals and deviance residuals for each covariate pattern:

```
Contact$phat<-fitted.values(Model)
Contact$yhat<-with(Contact,Obs.*phat)
Contact$PearsonRes<-with(Contact,(Y-yhat)/sqrt(yhat*(1-phat)))
Contact$DevianceRes<-with(Contact,
  sign(Y-yhat)*sqrt(2)*sqrt(Y*log(Y/yhat)+(Obs.-Y)*log((Obs.-Y)/(Obs.-yhat))))
Contact[,c(1:2,5:6,3,7:9)]
```

```
##   Contact       Norms         p      phat   Y      yhat PearsonRes
## 1       0   Favorable 0.4270833 0.4506658  41  43.26392 -0.4643859
## 2       1   Favorable 0.6976744 0.6845121 120 117.73608  0.3714618
## 3       0 Unfavorable 0.2786070 0.2673437  56  53.73608  0.3608088
## 4       1 Unfavorable 0.4748201 0.4911073  66  68.26392 -0.3841064
##   DevianceRes
## 1  -0.4652041
## 2   0.3728929
## 3   0.3592364
## 4  -0.3842147
```

The group-specific residuals suggest a good fit. None of them is anywhere near 1.96 standard deviations in absolute value. We'll also calculate a summary statistic of the fit by adding up the squared Pearson residuals. This statistic follows a chi-squared distribution, with degrees of freedom equal to the number of covariate patterns (groups) minus the number of parameters (intercept plus predictors):

```
(OverallPearson <- sum(Contact$PearsonRes^2))
```

```
## [1] 0.6313589
```

```
1-pchisq(OverallPearson, df = 1)
```

```
## [1] 0.4268573
```

The residual degrees of freedom equal the number of binomial observations (4) minus the number of parameters in the model (3) = 1. The p-value is .43, indicating a statistically non-significant lack of fit. We can thereby conclude that the model fits quite well.

2. Effects of Variable Centering: Simulation

```
n <- 50
G <- c(rep(1,20),rep(0,30))
x <- vector(); for(i in 1:n) {xi <- i/n; x <- c(x,xi)}; rm(xi)
X <- cbind(rep(1,n),G,x)
```

```
B <- c(-0.9, 0.3, 1)
logodds <- X[,1]*B[1] + X[,2]*B[2] + X[,3]*B[3]
p <- exp(logodds)/(1+exp(logodds))
N <- 300

# x NOT centered:
MLEs <- data.frame()
for (i in 1:N){
  U <- runif(n)
  y <- as.numeric(p > U)
  model <- glm(y ~ G+x, family=binomial(link=logit))
  MLEs <- rbind(MLEs, coef(model))
};rm(model,y,U); names(MLEs)<-c("B0","B1","B2")
#Look at the MLEs:
sapply(MLEs, mean);cat("\n B2B0 correlation:\n");with(MLEs,cor(B2,B0))
```

```
##         B0         B1         B2
## -0.9927144  0.3579013  1.1147464
```

```
##
##  B2B0 correlation:
```

```
## [1] -0.9620566
```

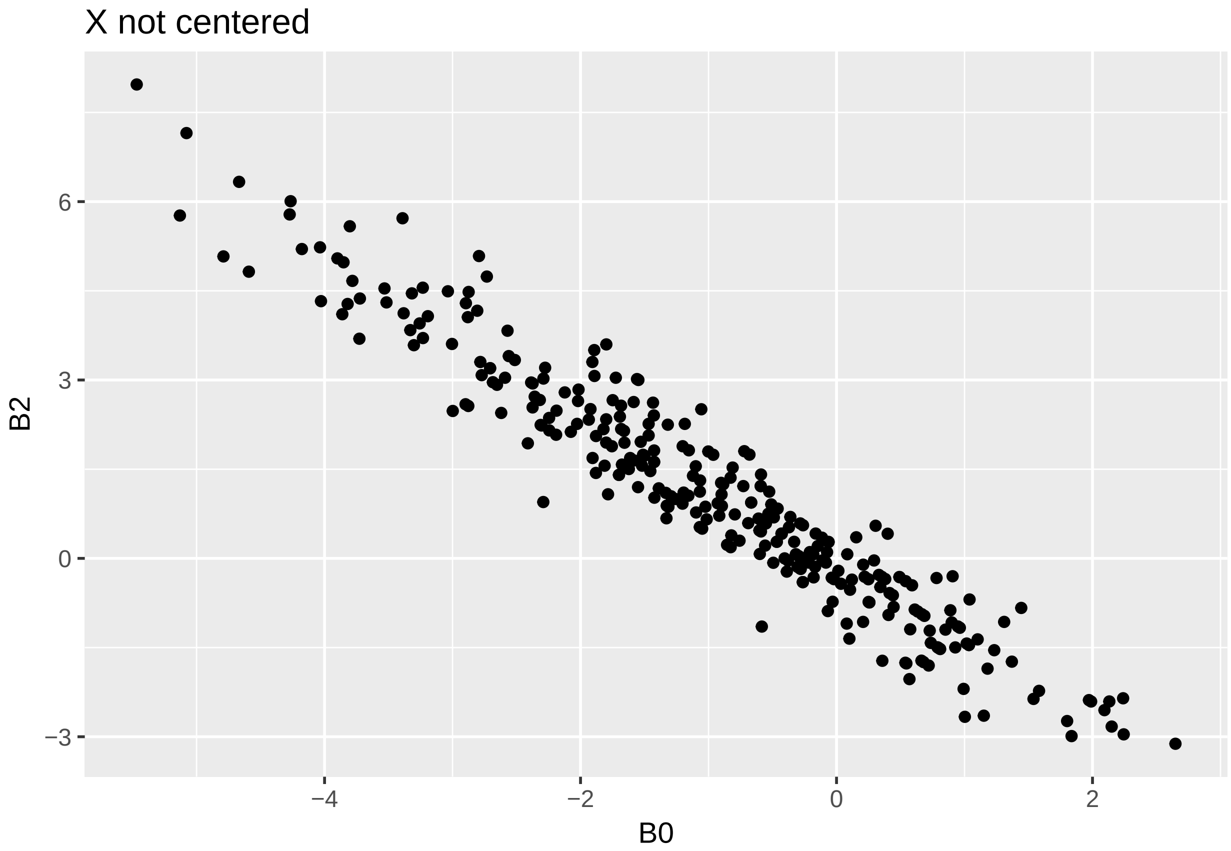
```
#x CENTERED:
x2 <- x-mean(x)
MLEs.2 <- data.frame()
for (i in 1:N){
  U <- runif(n)
  y <- as.numeric(p > U)
  model <- glm(y ~ G+x2, family=binomial(link=logit))
  MLEs.2 <- rbind(MLEs.2, coef(model))
};rm(model,y,U); names(MLEs.2)<-c("B0","B1","B2")
#Look at the MLEs:
sapply(MLEs.2, mean);cat("\n B2B0 correlation:\n");with(MLEs.2,cor(B2,B0))
```

```
##         B0         B1         B2
## -0.4298326  0.3495416  1.2124099
```

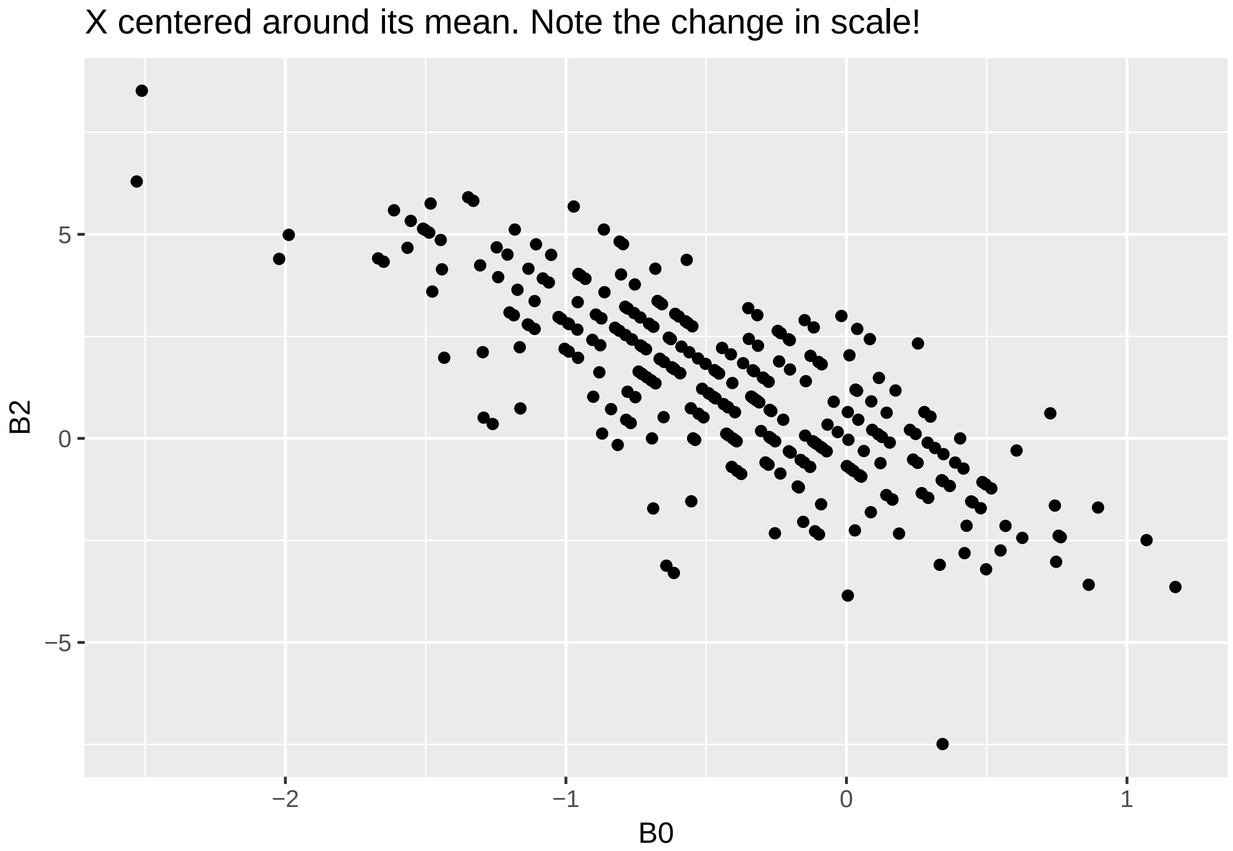
```
##
##  B2B0 correlation:
```

```
## [1] -0.7827631
```

```
library(ggplot2)
(scatter1<-ggplot(data=MLEs, aes(x=B0, y=B2)) +
  geom_point(aes(x=B0,y=B2)) +
  ggtitle("X not centered"))
```



```
(scatter2<-ggplot(data=MLEs.2, aes(x=B0, y=B2)) +
  geom_point(aes(x=B0,y=B2)) +
  ggtitle("X centered around its mean. Note the change in scale!"))
```



The constant term, or intercept, is the estimated logit of the outcome when all continuous covariates equal 0 and all categorical covariates are at their reference level. This means that the coding of continuous covariates ought to be given careful thought – else the MLE for the intercept becomes uninteresting. In general, centering continuous covariates around their means and contrasting qualitative covariates with their most common category simplifies the interpretation of any regression model – in particular, it makes the constant term immediately informative. This is well illustrated by continuous covariates such as a subject's height and weight, where zero values are impossible in real data. Without centering, the intercept will be the logit-scale probability of the outcome for an individual who weighs 0 pounds and stands 0 feet tall, which is much less useful than an intercept representing the logit-scale probability for a person of average weight and height.

In our simulation data above, the true mean probability of the outcome is .43. The uncentered continuous covariate x takes on positive values from 0 to 1. True $B2$ is 1, while "true" $B0$ (intercept) is -0.9. If we center x around a smaller value after the outcomes are generated, its MLE remains the same because it is not based on the x values themselves but on the relationship between the observed variation in x and the observed variation in y. $B0$, by contrast, represents not the relationship between some variable and the outcome but the estimated logit-scale probability of the outcome before the covariates have had an effect – in other words, when they equal zero. Therefore, when the values of x are reduced, i.e., brought closer to zero by centering it while the data is kept fixed, $B0$ must necessarily increase, because the effect of x at "zero" now represents the effect of x at its mean.

All of the above entails that it is possible to manipulate the intercept of a model in almost any way we want by adjusting the origins and/or scales of the continuous covariates. Such manipulation is sometimes advisable if it can make the interpretation of the model coefficients simpler and more intuitive.

3. Conditional Logistic Regression for Matched-Pairs Case Control Data

We'll now investigate what factors might be associated with low birth weight in newborns. The data comes in matched-pairs case-control format, so we'll use conditional logistic regression to remove the unmeasured, pair-specific effects from the equation. The measured covariates available are:

1. Race (nominal: 1=white, 2=black, 3=other)
2. Smoking during pregnancy (dichotomous)
3. PTD: history of premature labor (dichotomous)
4. HT: History of hypertension (dichotomous)
5. UI: Intrauterine Irritability (dichotomous)

I began the analysis by entering the aforementioned 5 variables plus every conceivable two-way interaction term into the model. It was messy and therefore isn't shown here. Many of the interactions could not be analyzed because the data was too sparse, or there were singularities. Out of the interactions that could be analyzed, however, none was statistically significant. We can thus proceed to fitting a model with main effects only:

```r
library(survival)
lowbw <- read.table(file="lowbwtm11CSV.csv",sep=";",header=T)
names(lowbw)
```

```
## [1] "ï..PAIR" "LOW"     "AGE"     "LWT"     "RACE"    "SMOKE"   "PTD"    
## [8] "HT"      "UI"
```

```r
names(lowbw)[1]<-"PAIR"
lowbw$RACE <- as.factor(lowbw$RACE)
levels(lowbw$RACE) <- c("WHITE","BLACK","OTHER")
summary(bw1 <- clogit(LOW ~ RACE + SMOKE + PTD + HT + UI + strata(PAIR), data = lowbw))
```

```
## Call:
## coxph(formula = Surv(rep(1, 112L), LOW) ~ RACE + SMOKE + PTD + 
##     HT + UI + strata(PAIR), data = lowbw, method = "exact")
## 
##   n= 112, number of events= 56 
## 
##              coef exp(coef) se(coef)     z Pr(>|z|)   
## RACEBLACK 0.4340    1.5434   0.6788 0.639  0.52260   
## RACEOTHER 0.5308    1.7003   0.5803 0.915  0.36031   
## SMOKE     1.5939    4.9230   0.5947 2.680  0.00736 **
## PTD       1.6826    5.3795   0.7209 2.334  0.01960 * 
## HT        1.5011    4.4865   0.8322 1.804  0.07126 . 
## UI        1.3320    3.7887   0.7034 1.894  0.05826 . 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##           exp(coef) exp(-coef) lower .95 upper .95
## RACEBLACK     1.543     0.6479    0.4080     5.838
## RACEOTHER     1.700     0.5881    0.5453     5.302
## SMOKE         4.923     0.2031    1.5347    15.792
## PTD           5.379     0.1859    1.3095    22.099
## HT            4.487     0.2229    0.8782    22.921
## UI            3.789     0.2639    0.9545    15.039
## 
## Rsquare= 0.18   (max possible= 0.5 )
## Likelihood ratio test= 22.25  on 6 df,   p=0.001089
## Wald test            = 12.42  on 6 df,   p=0.05325
## Score (logrank) test = 17.83  on 6 df,   p=0.00666
```

Race is about to be dropped from the model. We'll verify this through a comparison of nested models:

```
bw2 <- clogit(LOW ~ SMOKE + PTD + HT + UI + strata(PAIR), data = lowbw)
anova(bw1,bw2,test = "Chisq")
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(rep(1, 112L), LOW)
##  Model 1: ~ RACE + SMOKE + PTD + HT + UI + strata(PAIR)
##  Model 2: ~ SMOKE + PTD + HT + UI + strata(PAIR)
##    loglik  Chisq Df P(>|Chi|)
## 1 -27.689
## 2 -28.149 0.9197  2    0.6314
```

Dropping RACE did not significantly worsen the fit, and it made the model more parsimonious:

```
summary(bw2)
```

```
## Call:
## coxph(formula = Surv(rep(1, 112L), LOW) ~ SMOKE + PTD + HT +
##     UI + strata(PAIR), data = lowbw, method = "exact")
##
##   n= 112, number of events= 56
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## SMOKE 1.3808    3.9779   0.5256 2.627  0.00861 **
## PTD   1.6411    5.1611   0.6891 2.382  0.01723 *
## HT    1.6042    4.9739   0.7918 2.026  0.04275 *
## UI    1.3592    3.8929   0.7017 1.937  0.05277 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## SMOKE     3.978     0.2514    1.4199     11.14
## PTD       5.161     0.1938    1.3372     19.92
## HT        4.974     0.2010    1.0538     23.48
## UI        3.893     0.2569    0.9839     15.40
##
## Rsquare= 0.173   (max possible= 0.5 )
## Likelihood ratio test= 21.33  on 4 df,   p=0.0002719
## Wald test            = 12.15  on 4 df,   p=0.01629
## Score (logrank) test = 17.15  on 4 df,   p=0.001804
```

All the remaining variables are significant. Intrauterine irritability has a p-value on the borderline of significance, but we'll keep it in the model. Newborn babies' health is at stake here, so it's prudent to be cautious and attentive of any suspicious factors, even if they fall 2.7 per mille short of the conventional significance threshold. The model output shows the logit-scale MLEs converted into odds ratios as well as the 95% confidence intervals of these odds ratios: smoking during pregnancy increases the odds of the baby having a low birth weight by a factor of 1.4 to 11. For those having a history of premature labor, the odds of delivering a baby with a low birth weight are from 1.3 to 20 times as high as the odds for those without such a history. Hypertension increases the odds by a factor ranging from 1.05 to 23. Intrauterine irritability multiplies the odds by factor of 1 to 15.

4. Predicting Membership in the Hilmo Drug User Database

We'll now use logistic regression to model the probability of double registration for criminal drug users. Unfortunately the data file is not in a format immediately usable by R, so we have to edit it a little before importing:

```r
orig.file<-scan(what="char",file="drug-users.txt",sep="\n")
head(orig.file,10)
```

```
##  [1] "Data on Drug Users"
##  [2] "hilmo = 1, if registered in hilmo,otherwise = 0"
##  [3] "riki = 1, if registered in riki, otherwise = 0"
##  [4] "age (in years)"
##  [5] "male = 1, if male, = 0 if female"
##  [6] "sta = family status: 1 = married or cohabiting, 2 = single, 3 = widowed, 4 = divorced, 5 = not
##  [7] "hilmo   riki   age male   sta\t"
##  [8] "  1     0     41     0     4"
##  [9] "  1     1     32     1     2"
## [10] "  0     1     23     0     2"
```

```r
library(stringr)
drug_orig<-str_replace_all(orig.file[7:length(orig.file)],"^ +","")
drug_orig<-str_replace_all(drug_orig," +","\t")
drug_orig<-str_replace_all(drug_orig,"\t(?=$)",""); cat(drug_orig,file="drug_orig.txt",sep="\n")
drug_orig<-read.table(file="drug_orig.txt",sep="\t",header=T)
```

We also have to convert the relationship status variable from numeric to nominal:

```r
drug_orig$sta<-as.factor(drug_orig$sta); levels(drug_orig$sta)
```

```
## [1] "1" "2" "3" "4" "5"
```

```r
(levels(drug_orig$sta)<-c("HasPartner","Single","Widowed","Divorced",NA))
```

```
## [1] "HasPartner" "Single"     "Widowed"    "Divorced"   NA
```

```r
drug_orig[which(is.na(drug_orig$sta)),]
```

```
##    hilmo riki age male  sta
## 98     0    1  46    0 <NA>
```

Unfortunately, this lone NA value in the *sta* varuabke **will cause problems** with our probability calculations later. A NA value will prevent the automated calculation of a fitted probability for that observation, which will consequently prevent the calculation of means and other statistical parameters for the fitted probabilities. We will therefore impute a likely value for the missing field. And since the *sta* variable is categorical, what better way to impute it than logistic regression? We'll just need to use the polytomous variety:

```r
library(nnet)
drug <- drug_orig
drug$sta[98] <- predict(
  multinom(sta ~ riki + hilmo + age + male, data=drug), newdata=drug[98,], type="class")
```

```
## # weights:  24 (15 variable)
## initial  value 3593.274984
## iter  10 value 1774.095050
## iter  20 value 1479.097664
## iter  30 value 1478.872792
## iter  30 value 1478.872779
## iter  30 value 1478.872779
## final  value 1478.872779
## converged
```

```
drug[98,]
```

```
##    hilmo riki age male      sta
## 98     0    1  46    0 Divorced
```

Much better. Now we can proceed.

There are two databases that keep track of drug users in Finland – the Hospital Discharge Register *hilmo* and the Criminal Report Register *riki*. Our aggregate dataset *drug* lists everyone who was registered as a drug user in one or both of these databases around the year 2000. The ultimate goal of these exercises is to estimate the total number of drug users in the Uusimaa province of Southern Finland at the turn of the millennium. First we will model the probability that subjects registered in *riki* are also found in hilmo. Thecovariates at our disposal are age, sex, and relationship status. We'll start by analyzing the main effects plus all potential two-way interactions. Also, having learned from exercise 2, we'll center the age variable around its mean:

```
d0<-drug[drug$riki==1,]
d0$ageCentered<-d0$age-mean(d0$age)
summary(drugmodel0<-glm(hilmo ~ ageCentered*male + male*sta + ageCentered*sta,
  family=binomial(link=logit), data=d0))
```

```
## 
## Call:
## glm(formula = hilmo ~ ageCentered * male + male * sta + ageCentered * 
##     sta, family = binomial(link = logit), data = d0)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.6927  -0.4091  -0.3752  -0.3265   2.5018  
## 
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)    
## (Intercept)             -2.443e+00  6.346e-01  -3.849 0.000119 ***
## ageCentered             -1.261e-01  7.705e-02  -1.636 0.101809    
## male                    -6.202e-01  8.659e-01  -0.716 0.473838    
## staSingle                1.837e-01  6.936e-01   0.265 0.791151    
## staWidowed               1.803e+01  1.145e+03   0.016 0.987439    
## staDivorced             -1.228e-01  9.436e-01  -0.130 0.896430    
## ageCentered:male        -3.288e-03  4.302e-02  -0.076 0.939080    
## male:staSingle           1.883e-01  9.152e-01   0.206 0.836987    
## male:staWidowed         -3.053e+01  1.913e+03  -0.016 0.987269    
## male:staDivorced         6.291e-01  1.163e+00   0.541 0.588636    
## ageCentered:staSingle    9.346e-02  7.607e-02   1.229 0.219208    
## ageCentered:staWidowed   1.293e-01  9.077e+01   0.001 0.998863    
## ageCentered:staDivorced  1.024e-01  8.361e-02   1.225 0.220499    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 813.71  on 1578  degrees of freedom
## Residual deviance: 787.62  on 1566  degrees of freedom
## AIC: 813.62
##
## Number of Fisher Scoring iterations: 14
```

None of the interactions is significant, so we'll drop them. Next, we'll analyze main effects only:

```
summary(drugmodel1<-glm(hilmo ~ ageCentered + male + sta, family=binomial(link=logit), data=d0))
```

```
##
## Call:
## glm(formula = hilmo ~ ageCentered + male + sta, family = binomial(link = logit),
##     data = d0)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9888  -0.4137  -0.3731  -0.3157   2.6095
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.58483    0.44110  -5.860 4.63e-09 ***
## ageCentered -0.04297    0.01677  -2.562   0.0104 *
## male        -0.42248    0.23168  -1.824   0.0682 .
## staSingle    0.29022    0.44789   0.648   0.5170
## staWidowed   2.54505    1.03910   2.449   0.0143 *
## staDivorced  0.45167    0.52100   0.867   0.3860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 813.71  on 1578  degrees of freedom
## Residual deviance: 794.79  on 1573  degrees of freedom
## AIC: 806.79
##
## Number of Fisher Scoring iterations: 5
```

Thanks to our centering of age, the intercept is immediately relevant. It tells us that, according to this model, a drug user who is aged 28, female, and in a relationship, has a probability of

$$e^{-}2.58388/(1+e^{-}2.58388) = .07$$

of inclusion in *hilmo*. Age is negatively associated with inclusion in the database, while being Widowed has a very strong positive association. Maleness is negatively associated – a phenomenon which is intuitively believable since common sense suggests that men are less likely than women to seek medical help in distress. The effect falls just short of statistical significance, however.

Within the categorical variable describing relationship status, only the Widowed category has a significant effect. It is therefore worthwhile experimenting with a dummy variable that indicates simply whether the person is Widowed (1) or not (0). This effectively collapses the three remaining levels together into one large reference category representing the vast majority of the subjects in the dataset:

```
d0$Widowed<-as.numeric(d0$sta=="Widowed")
summary(drugmodel2<-glm(hilmo~ageCentered+male+Widowed,family=binomial(link=logit),data=d0))
```

```
##
## Call:
## glm(formula = hilmo ~ ageCentered + male + Widowed, family = binomial(link = logit),
##     data = d0)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9886  -0.4122  -0.3730  -0.3174   2.6582
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.29755    0.20219 -11.364  < 2e-16 ***
## ageCentered -0.04157    0.01444  -2.879  0.00399 **
## male        -0.41708    0.22765  -1.832  0.06694 .
## Widowed      2.25190    0.95705   2.353  0.01863 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 813.71  on 1578  degrees of freedom
## Residual deviance: 795.58  on 1575  degrees of freedom
## AIC: 803.58
##
## Number of Fisher Scoring iterations: 5
```

```r
anova(drugmodel1,drugmodel2,test="LRT")
```

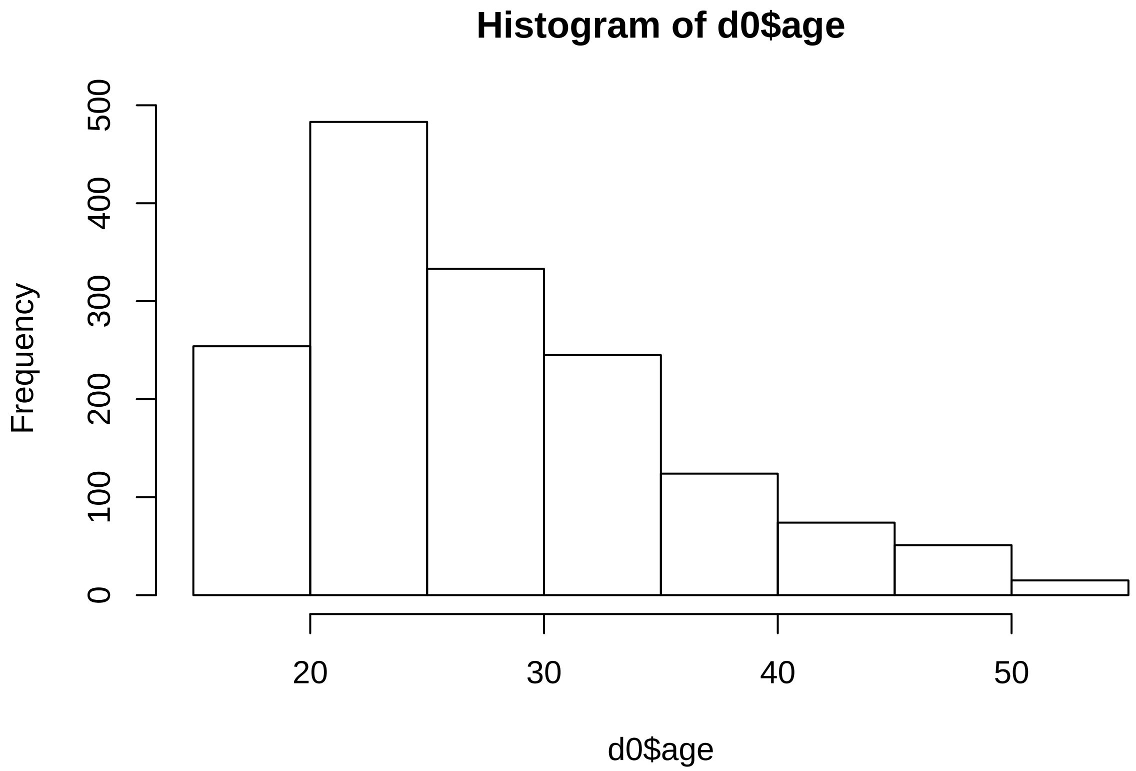
```
## Analysis of Deviance Table
##
## Model 1: hilmo ~ ageCentered + male + sta
## Model 2: hilmo ~ ageCentered + male + Widowed
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1573     794.79
## 2      1575     795.58 -2 -0.78917    0.674
```

This is a clear improvement. Residual deviance increased only negligibly (p = .67), while df improved by 2 points and AIC by 5 points. We'll keep this change. Next, let's try breaking age into cohorts in order to examine the exact nature of its effect. We'll contrast the other age brackets with the one containing the mean.

```r
summary(d0$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   22.00   26.00   28.02   33.00   55.00
```

```r
hist(d0$age)
```



```r
d0$agegroup<-cut(d0$age,breaks=c(min(d0$age),19,24,29,34,39,44,49,max(d0$age)),include.lowest = TRUE)
addmargins(table(d0$agegroup)); nrow(d0)
```

```
##
## [15,19] (19,24] (24,29] (29,34] (34,39] (39,44] (44,49] (49,55]     Sum
##     162     483     365     272     138      83      55      21    1579

## [1] 1579
```

```r
d0$agegroup<-relevel(d0$agegroup,ref="(24,29]")
summary(drugmodel3<-glm(hilmo~agegroup+male+Widowed,family=binomial(link=logit),data=d0))
```

```
##
## Call:
## glm(formula = hilmo ~ agegroup + male + Widowed, family = binomial(link = logit),
##     data = d0)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2114  -0.4368  -0.3808  -0.3340   2.8437
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.8041     0.2594  -6.956 3.51e-12 ***
## agegroup[15,19]  -0.4075     0.3531  -1.154   0.2485
## agegroup(19,24]  -0.2858     0.2457  -1.163   0.2448
## agegroup(29,34]  -0.5567     0.3127  -1.781   0.0750 .
## agegroup(34,39]  -0.7636     0.4296  -1.778   0.0755 .
## agegroup(39,44]  -2.2214     1.0220  -2.174   0.0297 *
## agegroup(44,49]  -1.0527     0.7421  -1.419   0.1560
## agegroup(49,55] -14.5978   505.4343  -0.029   0.9770
## male             -0.4978     0.2304  -2.160   0.0308 *
## Widowed           2.3814     1.0301   2.312   0.0208 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 813.71  on 1578  degrees of freedom
## Residual deviance: 787.33  on 1569  degrees of freedom
## AIC: 807.33
##
## Number of Fisher Scoring iterations: 15
```

The effect of age not linear. Instead, it seems to have a similar effect as it does on athletic performance, which is generally high in youth, peaks in the late 20s and then starts to deteriorate as a function of time. The reference category (25-to-29-year-olds) has by far the highest risk of inclusion.

One unexpected consequence of dividing age into cohorts is that maleness becomes statisically significant, which it wasn't in the model where age was continuous. How can this be explained? One possibility is that one or more of the cohorts may, purely by chance, have a high proportion of males avoiding hospitalization on the top end of the age bracket. With age as categorical, the model would attribute such within-bracket associations between age and the outcome solely to maleness rather than dividing it between age and maleness.

Which is the better model between the one with continuous age and the one where it's bracketed? We cannot compare them directly using a LRT because the models are not nested – one treats age as continuous, the other as nominal, so this is a matter of different variables rather than extra variables.

Judging by the smaller AIC, we should perhaps prefer the simpler model with continuous age to the one with age brackets. Though the effect of age is not exactly linear, the overall trend is still that old people are less likely to be registered than young people. In fact, a binary age division between the old and the young might further improve the model. The output of the bracketed-age model already gives us clues as to what the best threshold value might be. We'll try to find one now.

After experimenting with different age cutoffs (not shown), I learned that a binary age distinction between those under 33 and the rest results in a better fit than having age as continuous, and the model is equally parsimonious.

```
d0$AgeUnder33<-as.numeric(d0$age<33)
summary(drugmodel4<-glm(hilmo~AgeUnder33+Widowed+male,family=binomial(link=logit),data=d0))
```

```
##
## Call:
## glm(formula = hilmo ~ AgeUnder33 + Widowed + male, family = binomial(link = logit),
##     data = d0)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0666  -0.3955  -0.3955  -0.2527   2.6307
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.9655     0.3274  -9.058  < 2e-16 ***
## AgeUnder33    0.9193     0.2935   3.132  0.00173 **
## Widowed       2.2429     0.9618   2.332  0.01970 *
## male         -0.4630     0.2263  -2.046  0.04075 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 813.71  on 1578  degrees of freedom
## Residual deviance: 792.83  on 1575  degrees of freedom
## AIC: 800.83
##
## Number of Fisher Scoring iterations: 6
```

```
anova(drugmodel2, drugmodel4)
```

```
## Analysis of Deviance Table
##
## Model 1: hilmo ~ ageCentered + male + Widowed
## Model 2: hilmo ~ AgeUnder33 + Widowed + male
##   Resid. Df Resid. Dev Df Deviance
## 1      1575     795.58
## 2      1575     792.83  0   2.7572
```

The age-bracketed model above has lower residual deviance, but its higher number of parameters results in diminished parsimony and a higher AIC (807.33 as opposed to the current model's 800.83). Therefore, drugmodel4 is our choice. It's got three binary predictors, each of which has high predictive power and is simple to interpret. We'll finish by predicting the probabilities of belonging to *hilmo* for every subject in the entire drug dataset, and viewing 10 of them at random:

```
drug$Widowed<-as.numeric(drug$sta=="Widowed")
drug$AgeUnder33<-as.numeric(drug$age<33)
drug$FittedHilmoProbs <- predict(drugmodel4,newdata=drug,type="response")
sample(drug$FittedHilmoProbs, size=10)
```

```
##  [1] 0.07521518 0.11443921 0.07521518 0.07521518 0.07521518 0.11443921
##  [7] 0.07521518 0.03141681 0.07521518 0.03141681
```

5. Predicting Membership in Criminal Report Rekister Riki

Now we'll perform a similar analysis in the other direction, trying to predict whether those registered as drug users in *hilmo* have been caught committing crimes and listed in *riki* as drug users. We'll begin with a model containing all the main effects and potential two-variable interactions. We will also center age around its mean again:

```
d1<-drug[drug$hilmo==1,]; row.names(d1)<-1:nrow(d1)
d1$ageCentered<-d1$age-mean(d1$age)
summary(DrugModel0<-glm(riki ~ ageCentered*male + male*sta +ageCentered*sta,
  family=binomial(link=logit), data=d1))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Call:
## glm(formula = riki ~ ageCentered * male + male * sta + ageCentered *
##     sta, family = binomial(link = logit), data = d1)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -0.6647  -0.5146  -0.4518  -0.3492   2.6432
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -3.002e+00  6.513e-01  -4.609 4.05e-06 ***
## ageCentered             -1.359e-01  7.922e-02  -1.716   0.0862 .
## male                     3.222e-01  8.764e-01   0.368   0.7132
## staSingle                5.382e-01  7.049e-01   0.764   0.4452
## staWidowed               2.429e+00  5.770e+02   0.004   0.9966
## staDivorced              3.587e-01  9.379e-01   0.382   0.7022
## ageCentered:male         1.141e-03  3.935e-02   0.029   0.9769
## male:staSingle          -2.495e-02  9.214e-01  -0.027   0.9784
## male:staWidowed         -2.769e+01  2.059e+03  -0.013   0.9893
## male:staDivorced         6.150e-01  1.166e+00   0.527   0.5979
## ageCentered:staSingle    8.848e-02  7.974e-02   1.110   0.2672
## ageCentered:staWidowed  -3.142e+00  1.285e+02  -0.024   0.9805
## ageCentered:staDivorced  8.817e-02  8.733e-02   1.010   0.3127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 734.05  on 1126  degrees of freedom
## Residual deviance: 702.04  on 1114  degrees of freedom
## AIC: 728.04
##
## Number of Fisher Scoring iterations: 16
```

None of the interaction terms is significant, and the whole model is quite confusing to interpret. We'll drop the interactions:

```
summary(DrugModel1<-glm(riki ~ ageCentered + male + sta, family=binomial(link=logit), data=d1))
```

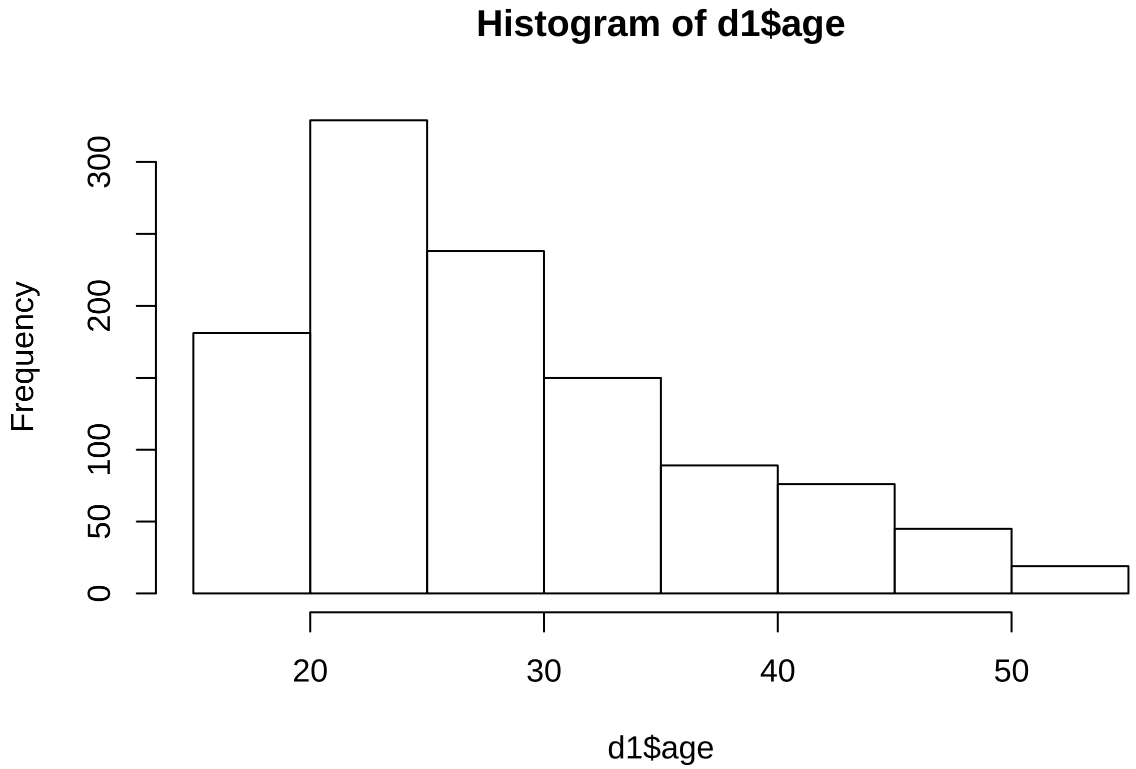
```
##
## Call:
## glm(formula = riki ~ ageCentered + male + sta, family = binomial(link = logit),
##     data = d1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9205  -0.5166  -0.4427  -0.3387   2.5389
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.98793    0.43928  -6.802 1.03e-11 ***
## ageCentered -0.05742    0.01659  -3.460  0.00054 ***
## male         0.36228    0.23368   1.550  0.12107
## staSingle    0.43188    0.44917   0.962  0.33630
## staWidowed   1.77874    0.91552   1.943  0.05203 .
## staDivorced  0.81244    0.52622   1.544  0.12260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 734.05  on 1126  degrees of freedom
## Residual deviance: 712.68  on 1121  degrees of freedom
## AIC: 724.68
##
## Number of Fisher Scoring iterations: 5
```

Here, age has an even stronger (disfavoring) effect than in the previous dataset. This makes intuitive sense – the old are less hotheaded than the young and therefore less prone to rash acts of violence and crime. Much like the risk of inclusion in *hilmo*, being Widowed also increases the risk of inclusion in *riki*. Maleness increases the risk of inclusion, but at p = .12 the effect falls somewhat short of statistical significance. Let's study the effect of age more closely by dividing the variable into cohorts:

```
summary(d1$age); hist(d1$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   22.00   26.00   28.61   33.00   55.00
```



```
d1$agegroup<-cut(d1$age,breaks=c(min(d1$age),19,24,29,34,39,44,49,max(d1$age)),include.lowest = TRUE)
addmargins(table(d1$agegroup)); nrow(d1)
```

```
##
## [15,19] (19,24] (24,29] (29,34] (34,39] (39,44] (44,49] (49,55]     Sum
##     111     339     265     157      98      77      57      23    1127
```

```
## [1] 1127
```

```
d1$agegroup<-relevel(d1$agegroup,ref="(24,29]") #Make 25-30 the reference level
summary(DrugModel2<-glm(riki ~ agegroup + male + sta, family=binomial(link=logit), data=d1))
```

```
##
## Call:
## glm(formula = riki ~ agegroup + male + sta, family = binomial(link = logit),
##     data = d1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9877  -0.5161  -0.4604  -0.3203   2.8992
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.5467     0.4646  -5.482 4.22e-08 ***
## agegroup[15,19]  -0.2224     0.3617  -0.615   0.5387
## agegroup(19,24]  -0.1635     0.2515  -0.650   0.5155
## agegroup(29,34]  -0.3977     0.3266  -1.218   0.2233
## agegroup(34,39]  -0.8433     0.4475  -1.885   0.0595 .
## agegroup(39,44]  -2.5417     1.0384  -2.448   0.0144 *
## agegroup(44,49]  -1.5648     0.7662  -2.042   0.0411 *
## agegroup(49,55] -14.8648   487.9613  -0.030   0.9757
## male              0.2421     0.2367   1.023   0.3065
## staSingle         0.5193     0.4468   1.162   0.2452
## staWidowed        1.8404     0.9395   1.959   0.0501 .
## staDivorced       0.9008     0.5329   1.691   0.0909 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 734.05  on 1126  degrees of freedom
## Residual deviance: 704.57  on 1115  degrees of freedom
## AIC: 728.57
##
## Number of Fisher Scoring iterations: 15
```

The effect of age seems similarly shaped as in the previous dataset – its favoring effect peaks in the late twenties, then reverses. I'll now see if I can find a threshold value at which age could be conveniently dichotomized to improve the model (experimentation not shown).

It appears that 33 is the best threshold for a binary age classification is this dataset too:

```
d1$AgeUnder33<-as.numeric(d1$age<33)
summary(DrugModel3<-glm(riki~AgeUnder33+male+sta,family=binomial(link=logit),data=d1))
```

```
##
## Call:
## glm(formula = riki ~ AgeUnder33 + male + sta, family = binomial(link = logit),
##     data = d1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0006  -0.5302  -0.4623  -0.3023   2.6032
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.8801     0.5085  -7.630 2.35e-14 ***
## AgeUnder33    1.1715     0.3295   3.556 0.000377 ***
## male          0.2913     0.2306   1.263 0.206486
## staSingle     0.5261     0.4429   1.188 0.234881
## staWidowed    1.9862     0.9269   2.143 0.032120 *
## staDivorced   0.8484     0.5280   1.607 0.108067
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 734.05  on 1126  degrees of freedom
## Residual deviance: 710.95  on 1121  degrees of freedom
## AIC: 722.95
##
## Number of Fisher Scoring iterations: 5
```

```
anova(DrugModel1,DrugModel3)
```

```
## Analysis of Deviance Table
##
## Model 1: riki ~ ageCentered + male + sta
## Model 2: riki ~ AgeUnder33 + male + sta
##   Resid. Df Resid. Dev Df Deviance
## 1      1121     712.68
## 2      1121     710.95  0   1.7256
```

This model fits the data better, on the same degrees of freedom, as the one where age was continuous. It seems safe to prefer the binary classification.

A further improvement can be accomplished by removing the non-significant distinction between single and divorced people within the *sta* variable. We will collapse these two categories together into a new reference category that will be contrasted with the more statistically significant classes. Being Widowed is clearly significant, but having a partner might be significant too – we have no direct measure of its effect because it is the variable's current baseline, and it is being contrasted with three different categories. By changing the baseline to Single/Divorced, we will get an overt estimate of the effect of having a partner. The merging the two non-significant cateogries will gain us a degree of freedom at little cost in residual deviance. We will call our new, trimmed version of the variable *sta.3way*:

```
levels(d1$sta)
```

```
## [1] "HasPartner" "Single"     "Widowed"    "Divorced"
```

```
d1$sta.3way <- d1$sta
levels(d1$sta.3way) <- c("HasPartner","Single/Divorced", "Widowed", "Single/Divorced")
d1$sta.3way <- relevel(d1$sta.3way, "Single/Divorced")
summary(DrugModel4<-glm(riki~AgeUnder33+male+sta.3way,family=binomial(link=logit),data=d1))
```

```
##
## Call:
## glm(formula = riki ~ AgeUnder33 + male + sta.3way, family = binomial(link = logit),
##     data = d1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9687  -0.5337  -0.4708  -0.3220   2.5458
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.2006     0.3340  -9.584  < 2e-16 ***
## AgeUnder33           1.0569     0.3001   3.522 0.000429 ***
## male                 0.2668     0.2287   1.167 0.243398    
## sta.3wayHasPartner  -0.5809     0.4384  -1.325 0.185164    
## sta.3wayWidowed      1.3639     0.8369   1.630 0.103156    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 734.05  on 1126  degrees of freedom
## Residual deviance: 711.72  on 1122  degrees of freedom
## AIC: 721.72
## 
## Number of Fisher Scoring iterations: 5
```

```
anova(DrugModel3,DrugModel4,test="LRT")
```

```
## Analysis of Deviance Table
## 
## Model 1: riki ~ AgeUnder33 + male + sta
## Model 2: riki ~ AgeUnder33 + male + sta.3way
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1121     710.95                     
## 2      1122     711.72 -1 -0.76874   0.3806
```

The chi-square statistic for the increase in deviance is only .77 at 1 degree of freedom, which supports our decision to drop the variable.

We now face a decision regarding the gender variable. It is not having the same effect as in the *riki* dataset. Let's try dropping it:

```
summary(DrugModel5<-glm(riki~AgeUnder33+sta.3way,family=binomial(link=logit),data=d1))
```

```
## 
## Call:
## glm(formula = riki ~ AgeUnder33 + sta.3way, family = binomial(link = logit), 
##     data = d1)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.8909  -0.5158  -0.5158  -0.3114   2.4702  
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         -3.0024     0.2844 -10.556  < 2e-16 ***
## AgeUnder33           1.0525     0.3001   3.507 0.000453 ***
## sta.3wayHasPartner  -0.6396     0.4355  -1.469 0.141909    
## sta.3wayWidowed      1.2307     0.8261   1.490 0.136312    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 734.05  on 1126  degrees of freedom
## Residual deviance: 713.12  on 1123  degrees of freedom
## AIC: 721.12
##
## Number of Fisher Scoring iterations: 5
```

```
anova(DrugModel4,DrugModel5,test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: riki ~ AgeUnder33 + male + sta.3way
## Model 2: riki ~ AgeUnder33 + sta.3way
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1122     711.72
## 2      1123     713.12 -1  -1.4042    0.236
```

At p = 0.24, there is relatively robust statistical evidence for removing gender from the model. We will drop the variable. However, since gender is always of scientific interest, we will state for the record that its effect on inclusion in *riki* generally ranged between 1 and 1.6 on the z-scale, varying according to what other predictors were in the model.

The choices get harder from this point on. We will make use of two more statistics in our endeavor to select the best model. In addition to the familiar LRT, ANOVA and AIC, we will be measuring each model candidate's testing error in leave-one-out crossvalidation. Secondly, we will calculate BIC for each model. AIC and BIC are both designed to provide a measure of the model's fit while taking its complexity into account. They accomplish this by adding a penalty to the deviance residual for each additional parameter, and their difference lies in the magnitude of that penalty. No consensus exists on which criterion is better, but AIC is considered relatively lenient towards added complexity, while BIC penalizes it more severely.

The two relationship statuses contrasted with the new baseline have opposite effects of almost identical magnitude. Both also fall somewhat short of the statistical significance threshold. Let's see what happens if we remove the ostensibly least significant remaining variable, i.e., the indicator for having a romantic partner.

```
summary(DrugModel6<-glm(riki~AgeUnder33+Widowed,family=binomial(link=logit),data=d1))
```

```
##
## Call:
## glm(formula = riki ~ AgeUnder33 + Widowed, family = binomial(link = logit),
##     data = d1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9031  -0.5062  -0.5062  -0.2967   2.5080
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.1011     0.2793 -11.104  < 2e-16 ***
## AgeUnder33    1.1109     0.2981   3.727 0.000194 ***
## Widowed       1.3039     0.8270   1.577 0.114859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 734.05  on 1126  degrees of freedom
## Residual deviance: 715.66  on 1124  degrees of freedom
## AIC: 721.66
##
## Number of Fisher Scoring iterations: 5
```

```r
anova(DrugModel5,DrugModel6,test = "LRT")
```

```
## Analysis of Deviance Table
## 
## Model 1: riki ~ AgeUnder33 + sta.3way
## Model 2: riki ~ AgeUnder33 + Widowed
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1123     713.12                     
## 2      1124     715.66 -1  -2.5351   0.1113
```

```r
if(exists("survival",mode="function")){detach("package:survival",unload=TRUE)}
library(boot)
```

```
## 
## Attaching package: 'boot'

## The following object is masked from 'package:survival':
## 
##     aml
```

```r
ErrorRate<-function(Y,prob){mean(abs(Y-prob)>0.5)} #Define how Error Rate is calculated
ModelCandidates<-data.frame(TestingErrorCV=cv.glm(data=d1,glmfit=DrugModel5,
  cost=ErrorRate)$delta[1],AIC=DrugModel5$aic,BIC=BIC(DrugModel5),DevianceRes=DrugModel5$deviance,
  row.names="AgeUnder33+sta.3way")
ModelCandidates<-rbind(ModelCandidates,
  "AgeUnder33+Widowed"=c(TestingErrorCV=cv.glm(data=d1,glmfit=DrugModel6,
  cost=ErrorRate)$delta[1],AIC=DrugModel6$aic,BIC=BIC(DrugModel6),DevianceRes=DrugModel6$deviance))
ModelCandidates
```

```
##                       TestingErrorCV      AIC      BIC DevianceRes
## AgeUnder33+sta.3way        0.1002662 721.1245 741.2338    713.1245
## AgeUnder33+Widowed         0.1002662 721.6596 736.7416    715.6596
```

P-value for removing the indicator is .11. Though not statistically significant, I find this uncomfortably low. AIC also deteriorates. BIC improves, but deviance jumps by 2.5 points. Testing error rates are identical. This evidence is very inconclusive. Before deciding, let's check what happens if we remove the widowhood indicator instead, and keep the dummy for having a partner.

```r
d1$HasPartner<-as.numeric(d1$sta.3way=="HasPartner")
summary(DrugModel6b<-glm(riki~AgeUnder33+HasPartner,family=binomial(link=logit),data=d1))
```

```
## 
## Call:
## glm(formula = riki ~ AgeUnder33 + HasPartner, family = binomial(link = logit), 
##     data = d1)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.5178  -0.5178  -0.5178  -0.3198   2.4489  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -2.9473     0.2783 -10.589  < 2e-16 ***
## AgeUnder33    1.0055     0.2955   3.402 0.000669 ***
## HasPartner   -0.6603     0.4351  -1.518 0.129123    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 734.05  on 1126  degrees of freedom
## Residual deviance: 714.90  on 1124  degrees of freedom
## AIC: 720.9
##
## Number of Fisher Scoring iterations: 5
```

```
anova(DrugModel5,DrugModel6b,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: riki ~ AgeUnder33 + sta.3way
## Model 2: riki ~ AgeUnder33 + HasPartner
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1123     713.12
## 2      1124     714.90 -1  -1.7769   0.1825
```

```
ModelCandidates<-rbind(ModelCandidates,
  "AgeUnder33+HasPartner"=c(TestingErrorCV=cv.glm(data=d1,glmfit=DrugModel6b,
  cost=ErrorRate)$delta[1],AIC=DrugModel6b$aic,BIC=BIC(DrugModel6b),DevianceRes=DrugModel6b$deviance))
```

Paradoxically, removing the widowhood indicator has less of an effect on deviance than removing the indicator for having a romantic partner, even though the former variable was estimated to be a more significant predictor. How can this be? The answer lies in different vastly different n's.

```
table(d1$sta.3way)
```

```
##
## Single/Divorced      HasPartner         Widowed
##             993             124              10
```

Widowhood is the better predictor in the few cases where it applies, but it is applicable to less than 1% of the subjects in the dataset. HasPartner is applicable to over 10% of the data. Thus, removing HasPartner worsens the fit considerably more than does removing Widowed, even though the former's predictive power on a single observation is slightly lower. Put another way, the widowhood dummy looks more impressive in the model output but is overall less useful than HasPartner. CV Testing Error remains identical between models 5 and 6b. AIC and BIC, as well as LRT, seem to support the removal of the widowhood dummy, so out it goes. Two predictors remain. Can we remove one?

```
DrugModel7<-glm(riki~AgeUnder33,family=binomial(link=logit),data=d1)
anova(DrugModel6b,DrugModel7,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: riki ~ AgeUnder33 + HasPartner
## Model 2: riki ~ AgeUnder33
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1124     714.90
## 2      1125     717.62 -1  -2.7228  0.09892 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(ModelCandidates<-rbind(ModelCandidates,
  "AgeUnder33"=c(TestingErrorCV=cv.glm(data=d1,glmfit=DrugModel7,
  cost=ErrorRate)$delta[1],AIC=DrugModel7$aic,BIC=BIC(DrugModel7),DevianceRes=DrugModel7$deviance)))
```

```
##                           TestingErrorCV      AIC      BIC DevianceRes
## AgeUnder33+sta.3way          0.1002662 721.1245 741.2338    713.1245
## AgeUnder33+Widowed           0.1002662 721.6596 736.7416    715.6596
## AgeUnder33+HasPartner        0.1002662 720.9015 735.9834    714.9015
## AgeUnder33                   0.1002662 721.6243 731.6789    717.6243
```

No, I don't think this removal is a good idea. The p-value for dropping HasPartner is below .10, and the deviance increase is unsavory. We have a pretty good model now. It is parsimonious with only two predictors, both of which make intuitive sense are simple to interpret. However, we perhaps do still a little better in terms of predictive power. Let's look more closely at the effect of relationship status:

```
table(d1$sta.3way,d1$riki)
```

```
##
##                      0   1
##   Single/Divorced 888 105
##   HasPartner      118   6
##   Widowed           8   2
```

Widowed people seem at an elevated risk of inclusion in *riki*. We knew this. Now let's condition the effect on sex:

```
cat("men\n");with(d1[d1$male==1,],table(sta.3way,riki));cat("\n\n")
```

```
## men
```

```
##                  riki
## sta.3way            0   1
##   Single/Divorced 625  80
##   HasPartner       58   3
##   Widowed           2   0
```

```
cat("women\n");with(d1[d1$male==0,],table(sta.3way,riki))
```

```
## women
```

```
##                  riki
## sta.3way            0   1
##   Single/Divorced 263  25
##   HasPartner       60   3
##   Widowed           6   2
```

The odds ratio of being in *riki* between widowed men and widowed women is infinite. For the bigger picture, let's view the same conditional crosstables from the *riki* dataset from the previous exercise:

```
cat("men\n");with(d0[d0$male==1,],table(sta,hilmo));cat("\n\n")
```

```
## men
```

```
##              hilmo
## sta             0   1
##   HasPartner   87   3
##   Single      968  71
##   Widowed       3   0
##   Divorced    148   9
```

```
cat("women\n");with(d0[d0$male==0,],table(sta,hilmo))
```

```
## women
```

```
##              hilmo
## sta            0   1
##   HasPartner  35   3
##   Single     179  22
##   Widowed      0   2
##   Divorced    46   3
```

Again the odds ratio between women and men is infinite. Based on this (very limited) data, widowhood seems much more dangerous to women than to men. The implication for our model selection exercise is that we should consider making a dummy variable for being a widow in the word's exact sense, i.e., a **woman** who has lost her partner to death and remains alone.

```
d1$fem.widow<-as.numeric(d1$male==0 & d1$Widowed)
summary(DrugModel8<-glm(riki~AgeUnder33+HasPartner+fem.widow,family=binomial,data=d1))
```

```
## 
## Call:
## glm(formula = riki ~ AgeUnder33 + HasPartner + fem.widow, family = binomial, 
##     data = d1)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.6592  -0.5159  -0.5159  -0.3088   2.4767  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -3.0192     0.2865 -10.537  < 2e-16 ***
## AgeUnder33    1.0695     0.3022   3.539 0.000402 ***
## HasPartner   -0.6353     0.4355  -1.459 0.144655    
## fem.widow     1.6031     0.8591   1.866 0.062027 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 734.05  on 1126  degrees of freedom
## Residual deviance: 712.20  on 1123  degrees of freedom
## AIC: 720.2
## 
## Number of Fisher Scoring iterations: 5
```

```
ModelCandidates<-rbind(ModelCandidates,
  "AgeUnder33+HasPartner+fem.widow"=c(TestingErrorCV=cv.glm(data=d1,glmfit=DrugModel8,
  cost=ErrorRate)$delta[1],AIC=DrugModel8$aic,BIC=BIC(DrugModel8),DevianceRes=DrugModel8$deviance))
ModelCandidates[,2:4]
```

```
##                                      AIC      BIC DevianceRes
## AgeUnder33+sta.3way             721.1245 741.2338    713.1245
## AgeUnder33+Widowed              721.6596 736.7416    715.6596
## AgeUnder33+HasPartner           720.9015 735.9834    714.9015
## AgeUnder33                      721.6243 731.6789    717.6243
## AgeUnder33+HasPartner+fem.widow 720.2001 740.3093    712.2001
```

```
anova(DrugModel6b,DrugModel8,test="LRT")
```

```
## Analysis of Deviance Table
## 
## Model 1: riki ~ AgeUnder33 + HasPartner
## Model 2: riki ~ AgeUnder33 + HasPartner + fem.widow
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1124      714.9
## 2      1123      712.2  1   2.7014   0.1003
```

(Testing Error in crossvalidation remains identical for each model, so the column was left out to make the table fit within the page margins.)

The improvement in fit that is achieved by adding the dummy for being a widow has a p-value of .10, so there's no clear-cut answer to whether its inclusion is justified. It is risky to include covariates whose predictive power is based on such tiny samples – the reported effect could very well be due to mere chance. To illustrate, let's compare these three predictors by the extent to which they are influenced by single observations:

```
leverage <- dfbetas(DrugModel8); leverage <- as.data.frame(leverage)
(maxlev<-sapply(abs(leverage), max))
```

```
## (Intercept)  AgeUnder33  HasPartner   fem.widow
##   0.1892240   0.1999408   0.3110654   0.8549638
```

The *dfbeta* estimates how much the MLE of the covariate would change if the observation was deleted. We can see that the maximum leverage of an individual observation on the widow dummy is about 3 to 4 times as high as on the other two covariates. We see this difference leverage difference at work when we delete a single observation exerting maximum influence for each predictor and compare the changes in the MLEs:

```
leverage[which(abs(leverage$fem.widow)==maxlev["fem.widow"]),]
```

```
##     (Intercept) AgeUnder33 HasPartner fem.widow
## 393  -0.1863602  0.1999408 0.02011602 0.8549638
```

```
leverage[which(abs(leverage$HasPartner)==maxlev["HasPartner"]),]
```

```
##     (Intercept) AgeUnder33 HasPartner  fem.widow
## 128 -0.05525641 0.05928311  0.3110654 0.01134464
```

```
summary(glm(riki~AgeUnder33+HasPartner+fem.widow,family=binomial,data=d1[c(-128,-393),]))
```

```
##
## Call:
## glm(formula = riki ~ AgeUnder33 + HasPartner + fem.widow, family = binomial,
##     data = d1[c(-128, -393), ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5140  -0.5140  -0.5140  -0.3175   2.4547
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.9624     0.2835 -10.450  < 2e-16 ***
## AgeUnder33    1.0049     0.3006   3.343 0.000827 ***
## HasPartner   -0.8252     0.4721  -1.748 0.080475 .
## fem.widow     0.9803     1.1145   0.880 0.379114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 724.82  on 1124  degrees of freedom
## Residual deviance: 704.59  on 1121  degrees of freedom
## AIC: 712.59
##
## Number of Fisher Scoring iterations: 5
```

The MLE for AgeUnder33 stays virtually unchanged. HasPartner becomes more significant as its p-value decreases from .14 to .8 – this is because the most influential observation is one that deviates from the general trend, so getting rid of it makes the trend more clear. The effect of the widow dummy, on the other hand, is based on a mere 2 success outcomes, thus it vanishes completely when we remove one of them.

At the same time, the effect of being a widowed female has the same direction in both datasets – a strong positive association. Our choice is whether to ignore this effect due to the small sample or to assume that the association seen in this meager sample is an indicator of a wider phenomenon of prematurely widowed women having an elevated risk of antisocial and self-destructive behavior.

I'll take the conservative road here, and leave fem.widow out of the model. While the p-value is impressive, it hinges on a single observation. There might well be a significant effect, but making confident claims based on this kind of sample size would be sloppy science. We'll choose DrugModel6b, with AgeUnder33 and HasPartner as the only predictors. It is a parsimonious model with two simple predictors that make intuitive sense and aren't based on an embarrassingly small sample. We will now use this model to calculate the probability of inclusion in *riki* for everyone in the combined *drug* dataset:

```
drug$HasPartner<-as.numeric(drug$sta=="HasPartner")
drug$FittedRikiProbs<-predict(DrugModel6b, newdata=drug, type="response")
```

6. The Horvitz-Thompson Estimator of Population Total

If we take a random sample from a population of unknown size, mark each subject in the sample, release them back into the population, and then collect a new random sample of the same population, the degree of overlap between the two random samples is going to be approximately inversely proportional to the total population size. This is the capture-recapture method, which is used to estimate population sizes in fields such as biology. In our example, we are estimating the total population of drug users in Uusimaa, and our two random samples are the *hilmo* and *riki* registries, both of which keep independent track of drug users in the area. As per the instructions, we'll rename the *hilmo* and *riki* inclusion probabilities calculated for each subject in the *drug* dataset as $p1$ and $p2$, respectively:

```
names(drug)[which(names(drug)=="FittedHilmoProbs")]<-"p1"
names(drug)[which(names(drug)=="FittedRikiProbs")]<-"p2"
drug<-drug[,c(1:7,9,8,10)]
```

For each of the 2593 subjects, the probability of being registered in either *hilmo* or *riki* is:

```
drug$theta <- with(drug, p1+p2 - p2*p2)
```

Here's a summary of *theta*:

```
summary(drug$theta)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05712 0.09639 0.18490 0.16280 0.18490 0.65870
```

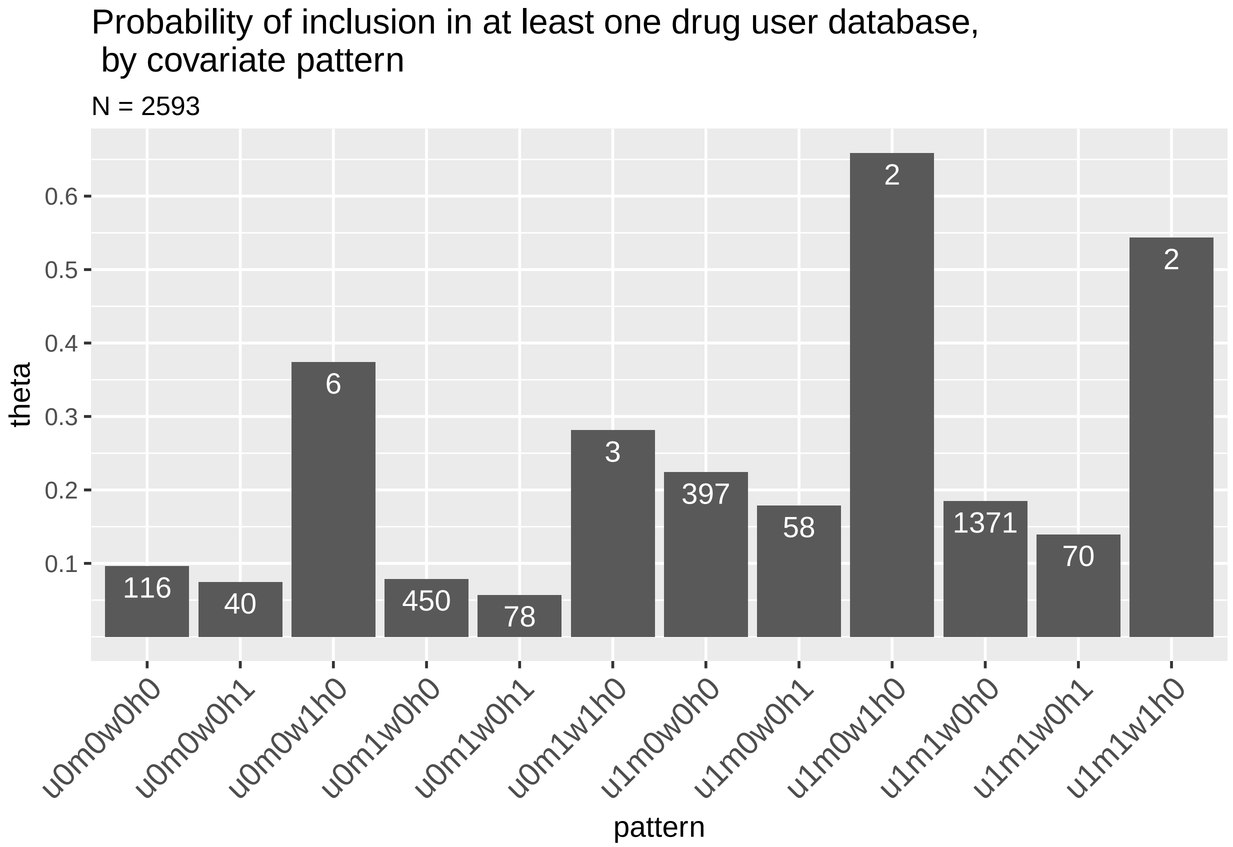
Our two models rely on a combined total of 4 explanatory variables, all of which are dichotomous. In the *drug* dataset, 12 different combinations of these variables occur. The following table lists the estimated $p1$, $p2$, and *theta* for each combination:

```
EVPs.Bernoulli<-split(drug, list(drug$AgeUnder33,drug$Widowed,drug$male,drug$HasPartner),drop=TRUE)
EVPs <- data.frame(); n<-vector()
for(i in 1:length(EVPs.Bernoulli)){
    EVPs <- rbind(EVPs,EVPs.Bernoulli[[i]][1,3:11]); n<-c(n,nrow(EVPs.Bernoulli[[i]]))}
EVPs$n<-n; row.names(EVPs)<-1:nrow(EVPs)
EVPs<-EVPs[,c(5,2,4,6,7:10)]
round(EVPs,digits=6)
```

```
##    AgeUnder33 male Widowed HasPartner       p1       p2    theta    n
## 1           0    0       0          0 0.049011 0.049863 0.096388  116
## 2           1    0       0          0 0.114439 0.125443 0.224146  397
## 3           0    0       1          0 0.326838 0.049863 0.374215    6
## 4           1    0       1          0 0.549033 0.125443 0.658740    2
## 5           0    1       0          0 0.031417 0.049863 0.078794  450
## 6           1    1       0          0 0.075215 0.125443 0.184922 1371
## 7           0    1       1          0 0.234056 0.049863 0.281433    3
## 8           1    1       1          0 0.433824 0.125443 0.543531    2
## 9           0    0       0          1 0.049011 0.026401 0.074714   40
## 10          1    0       0          1 0.114439 0.068999 0.178678   58
## 11          0    1       0          1 0.031417 0.026401 0.057120   78
## 12          1    1       0          1 0.075215 0.068999 0.139454   70
```

Let **u1 = age under 33, m1 = male, w1 = widowed, and h1 = has a partner**. The following plot illustrates the variation in *theta* across covariate patterns:

```
EVPs<-cbind(pattern=c("u0m0w0h0","u1m0w0h0","u0m0w1h0","u1m0w1h0","u0m1w0h0",
  "u1m1w0h0","u0m1w1h0","u1m1w1h0","u0m0w0h1","u1m0w0h1","u0m1w0h1","u1m1w0h1"),EVPs)
ggplot(EVPs, aes(x=pattern)) + geom_col(aes(y=theta)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 12)) +
  geom_text(aes(y=theta,label=n), vjust=1.5, colour="white") +
  scale_y_continuous(breaks = 1:6/10) +
  ggtitle("Probability of inclusion in at least one drug user database,\n by covariate pattern",
  subtitle="N = 2593")
```



The right side of the plot represents those under 33. The left side are the older people. In both age groups, it is the widowed women followed by the widowed men who have the highest probability of inclusion. We can also see how low *n* is in these four groups.

Finally, we'll calculate the Horvitz-Thompson estimate of the population total of drug users in Uusimaa around 2000. It is the sum over the inverse inclusion probabilities (thetas) of each subject in the full dataset:

```
(Horvitz.Thompson <- sum(1/drug$theta))
```

```
## [1] 18860.56
```

Based on the probability estimates yielded by our models, there were a little under 19,000 drug users in Uusimaa at the turn of the century.