Predicting Disability Prejudice

Jenna Harder

## The Data

This is a dataset from a web study of prejudice against people with disabilities. Here, we’ll be playing with a subset of that data–45,605 people who indicated that they themselves had disabilities–to identify how best to predict explicit (conscious) prejudice against people with disabilities among that very population.

Our variables include demographic variables, measures of certain experiences (e.g. self-reported severity of own disability), and some measures of different kinds of prejudice.

diat <- read.csv("diatCleanPWD.csv")  
library(caret)

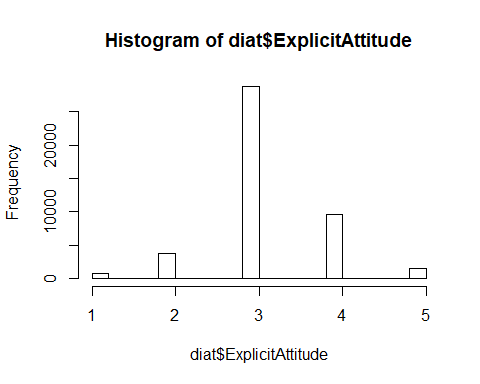
## Warning: package 'caret' was built under R version 3.4.4

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.4.4

Even though this is a question of people being prejudiced against themselves, there is still some variation in their prejudicial attitudes:



# Building a model

First, let’s split our data and train an initial model. We’ll use 10-fold cross validation with 5 repeats, which should give us a decent level of precision.

We’ll use linear regression; however, since we’re including both “age” and “age-squared,” the model is able to test for a u-shaped relationship with age (which has been observed in some past studies on disability prejudice among people who don’t have disabilities). If both age and age-squared are significant and have opposite signs, that will indicate a u-shaped relationship.

set.seed(112) #for reproducibility  
  
train\_control <- trainControl(method="repeatedcv", number=10, repeats=5)  
  
model <- train(ExplicitAttitude~., data=diat,   
 preProcess = c("scale", "center"),   
 trControl=train\_control,   
 method="lm",   
 na.action = na.omit) #this will shrink our sample  
  
summary(model)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.7877 -0.2629 -0.1260 0.3087 3.2647   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.161603 0.003925 805.523 < 2e-16 \*\*\*  
## numdate -0.016234 0.004125 -3.935 8.34e-05 \*\*\*  
## age2 -0.010654 0.007841 -1.359 0.174204   
## Gen -0.024801 0.004003 -6.196 5.88e-10 \*\*\*  
## eduC 0.016930 0.004621 3.663 0.000250 \*\*\*  
## contactPWD -0.023223 0.003961 -5.863 4.61e-09 \*\*\*  
## hide2 0.008999 0.004195 2.145 0.031955 \*   
## pervasive2 -0.004019 0.004884 -0.823 0.410601   
## duration2 -0.009154 0.004935 -1.855 0.063641 .   
## permEffects 0.004163 0.004829 0.862 0.388583   
## severenew2 -0.032621 0.004959 -6.579 4.85e-11 \*\*\*  
## ageSquared 0.024600 0.007073 3.478 0.000506 \*\*\*  
## Dscore 0.057387 0.004077 14.076 < 2e-16 \*\*\*  
## tabled2 0.318045 0.005082 62.584 < 2e-16 \*\*\*  
## tdisabled2 -0.368141 0.005140 -71.619 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5928 on 22794 degrees of freedom  
## Multiple R-squared: 0.2377, Adjusted R-squared: 0.2372   
## F-statistic: 507.6 on 14 and 22794 DF, p-value: < 2.2e-16

model$results[2] #RMSE

## RMSE  
## 1 0.5928558

Our model has an adjusted R-squared of 0.2372 and an RMSE of 0.5929 (note that our DV ranges from 1 to 5).

## Refining the model

Several variables were nonsignificant in our model, and we may want to get rid of those if we can do so without decreasing our predictive power.

The nonsignificant variables are:

Age of the participant

“Pervasiveness” of the disability (to what extent does it affect daily activities?)

Duration of the disability (how long has the participant had it?)

Perceived permanence of the disability (binary variable: does the participant expect it to continue over the long term?)

We may not want to get rid of age, because we have a wide range of ages in this sample and the education variable (which was significant) will mean something very different if we are not controlling for age. “Duration” is also nearly significant, so I’m tempted to leave that in. But let’s find out what happens if we get rid of Pervasiveness and Permanence:

model2 <- train(ExplicitAttitude~. -permEffects -pervasive2,   
 data=diat,   
 preProcess = c("scale", "center"),   
 trControl=train\_control,   
 method="lm",   
 na.action = na.omit)  
  
summary(model2)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.7822 -0.2625 -0.1267 0.3094 3.2656   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.161603 0.003925 805.533 < 2e-16 \*\*\*  
## numdate -0.015348 0.003968 -3.868 0.000110 \*\*\*  
## age2 -0.011708 0.007782 -1.505 0.132459   
## Gen -0.024898 0.003999 -6.227 4.84e-10 \*\*\*  
## eduC 0.017088 0.004619 3.700 0.000216 \*\*\*  
## contactPWD -0.023298 0.003958 -5.886 4.01e-09 \*\*\*  
## hide2 0.009319 0.004178 2.231 0.025721 \*   
## duration2 -0.006610 0.004151 -1.592 0.111370   
## severenew2 -0.034814 0.004173 -8.342 < 2e-16 \*\*\*  
## ageSquared 0.025004 0.007064 3.539 0.000402 \*\*\*  
## Dscore 0.057493 0.004070 14.125 < 2e-16 \*\*\*  
## tabled2 0.318222 0.005076 62.691 < 2e-16 \*\*\*  
## tdisabled2 -0.368257 0.005139 -71.662 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5928 on 22796 degrees of freedom  
## Multiple R-squared: 0.2376, Adjusted R-squared: 0.2372   
## F-statistic: 592.1 on 12 and 22796 DF, p-value: < 2.2e-16

model2$results[2]

## RMSE  
## 1 0.5928022

The adjusted R-squared is still .2372, and the RMSE is 0.5928: negligibly lower than the previous one. So we probably don’t need those variables.

Let’s see what happens when we get rid of age and duration, just to be thorough.

model3 <- train(ExplicitAttitude~. -permEffects -pervasive2 -age2 -duration2,   
 data=diat,   
 preProcess = c("scale", "center"),   
 trControl=train\_control,   
 method="lm",   
 na.action = na.omit)  
  
summary(model3)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.7813 -0.2626 -0.1263 0.3079 3.2616   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.161603 0.003925 805.474 < 2e-16 \*\*\*  
## numdate -0.014576 0.003947 -3.693 0.000222 \*\*\*  
## Gen -0.024364 0.003991 -6.105 1.05e-09 \*\*\*  
## eduC 0.012427 0.004004 3.104 0.001912 \*\*   
## contactPWD -0.023633 0.003955 -5.975 2.33e-09 \*\*\*  
## hide2 0.010316 0.004153 2.484 0.012987 \*   
## severenew2 -0.034970 0.004140 -8.447 < 2e-16 \*\*\*  
## ageSquared 0.015157 0.004041 3.751 0.000176 \*\*\*  
## Dscore 0.056993 0.004042 14.101 < 2e-16 \*\*\*  
## tabled2 0.318929 0.005065 62.966 < 2e-16 \*\*\*  
## tdisabled2 -0.368942 0.005127 -71.959 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5928 on 22798 degrees of freedom  
## Multiple R-squared: 0.2375, Adjusted R-squared: 0.2371   
## F-statistic: 709.9 on 10 and 22798 DF, p-value: < 2.2e-16

model3$results[2]

## RMSE  
## 1 0.5928783

Excluding age and duration has changed very little: our adjusted R-squared has gone down by 0.0001, and our RMSE remains about the same. Given this, there probably isn’t much justification to leave age and duration in the model.

This model’s predictive ability isn’t dramatic: it explains just shy of 24% of the variance. It would have been helpful if we’d had more data about these people, such as the specific type of disability each of them had. Our strongest predictors were “tabled” and “tdisabled,” both of which are other measures of prejudice (those are measures of a more emotion-based form of prejudice, whereas the measure we’re predicting here is tapping into a more cognitive/opinion-based form of prejudice.)

Some of the other variables are only slightly contributing to the prediction of explicit prejudice. We could probably streamline the model further by getting rid of some of those without decreasing the model’s predictive ability too much. However, efficiency isn’t really a concern here with only 10 predictors, and all these predictors are significant: so we’ll call this our final model.