Lab 9 – last one!

PSY V0500 Statistical Methods in Psychology

Kevin R Foster, Colin Powell School, the City College of New York, CUNY

For this lab we will look at tree models, which are useful for establishing a hierarchy of explanatory power.

We’ll use the Household Pulse Data.

```
require(plyr)
require(dplyr)
require(tidyverse)
require(ggplot2)
require(stargazer)

require(rpart)
require(rpart.plot)

load("Household_Pulse_data_ph4c2.RData")
```

We’ll look at incidence of depressive symptoms.

```
summary(Household_Pulse_data$DOWN)
```

```
##                                       NA no days in past 2 wks feeling depressed
##                                     5600                                   44300
##             several days over past 2 wks more than half the days over past 2 wks
##                                    15438                                    2965
##                         nearly every day
##                                     2849
```

```
select2 <- (Household_Pulse_data$DOWN != "NA")
d_down <- subset(Household_Pulse_data, select2)
d_down$DOWN <- fct_drop(d_down$DOWN) # drop unused levels
d_down$down_severe <- as.numeric(
    (d_down$DOWN == "more than half the days over past 2 wks") |
    (d_down$DOWN == "nearly every day") )

xtabs(~ d_down$DOWN + d_down$down_severe)
```

```
##                                           d_down$down_severe
## d_down$DOWN                                    0     1
##   no days in past 2 wks feeling depressed 44300     0
##   several days over past 2 wks            15438     0
##   more than half the days over past 2 wks     0  2965
##   nearly every day                            0  2849
```

The crosstab just verifies that everything looks the way we think it should. You can debate about what to do with the people who report “several days over past 2 weeks” when they were feeling depressed. You could do something like the following (which is not run),

```
d_down$down_recode <- as.numeric(
  (d_down$DOWN == "several days over past 2 wks") |
    (d_down$DOWN == "more than half the days over past 2 wks") |
    (d_down$DOWN == "nearly every day") )
```

Check some basic stats. What crosstabs might be useful? Perhaps try some graphs?

```
summary(d_down$down_severe)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.00000 0.00000 0.00000 0.08869 0.00000 1.00000
```

First note that tree models are sophisticated math but still kinda dumb. If we ask it, which are the most important variables in classifying which people report more severe depressive symptoms,

```
tree_mod1 <- rpart(down_severe ~ ., data = d_down, method = "class", cp = 0.01)
tree_mod1
```

```
## n= 65552
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 65552 5814 0 (0.91130705 0.08869295)
##   2) DOWN=no days in past 2 wks feeling depressed,several days over past 2 wks 597
##   3) DOWN=more than half the days over past 2 wks,nearly every day 5814    0 1 (0.
```

It will spit back the `DOWN` variable with which we created that `down_severe` variable. Not helpful. So we drop that variable from the data that we give the tree model,

```
d_down_for_tree <- select(d_down,-DOWN)
```

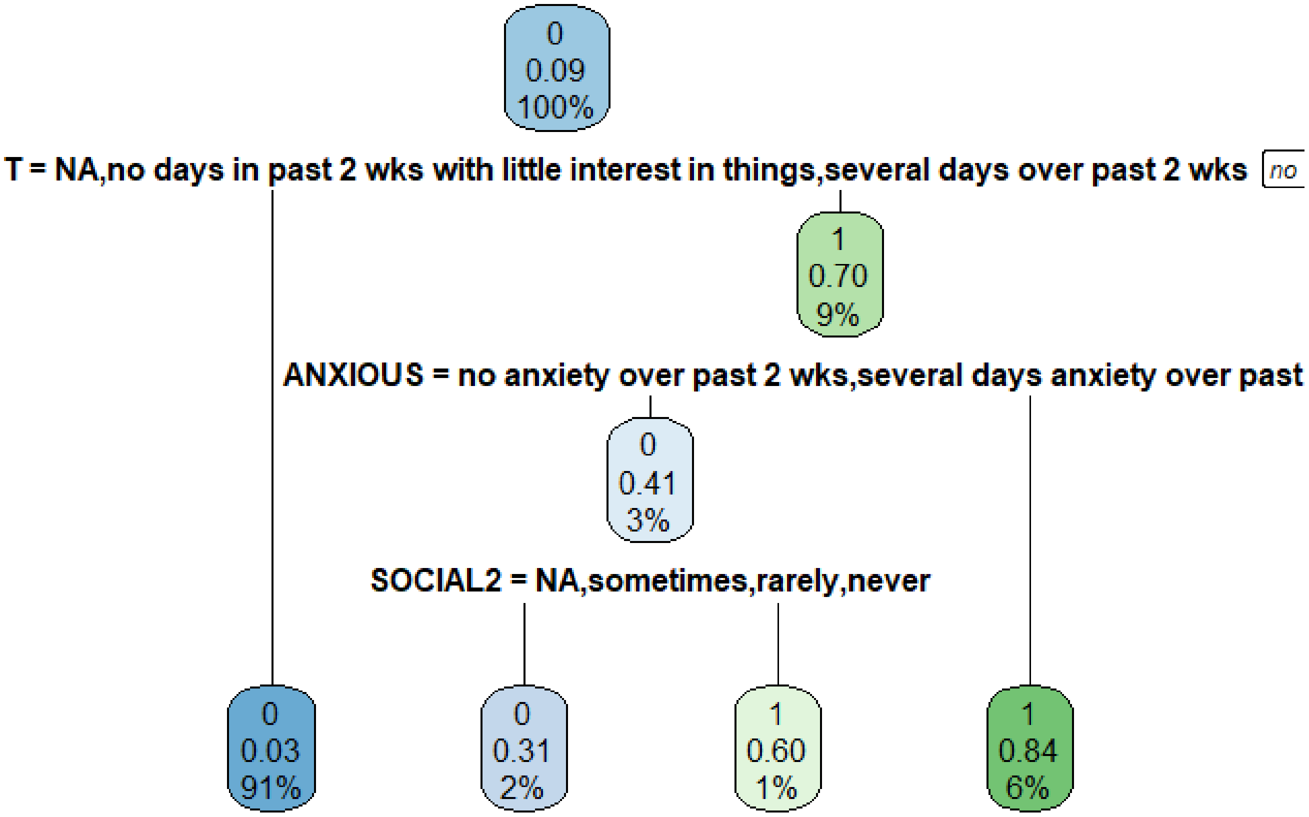
How is this?

```
tree_mod2 <- rpart(down_severe ~ ., data = d_down_for_tree, method = "class", cp = 0.
tree_mod2
```

```
## n= 65552
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 65552 5814 0 (0.91130705 0.08869295)
##    2) INTEREST=NA,no days in past 2 wks with little interest in things,several day
##    3) INTEREST=more than half the days over past 2 wks,nearly every day 6036 1826
##      6) ANXIOUS=no anxiety over past 2 wks,several days anxiety over past 2 wks 19
##       12) SOCIAL2=NA,sometimes,rarely,never 1325  409 0 (0.69132075 0.30867925) *
##       13) SOCIAL2=always lonely,usually 663  264 1 (0.39819005 0.60180995) *
##      7) ANXIOUS=NA,more than half the days anxiety over past 2 wks,nearly every da
```

The text output is tough to read, so try this,

```
rpart.plot(tree_mod2)
```


0
0.09
100%
T = NA,no days in past 2 wks with little interest in things,several days over past 2 wks
no
1
0.70
9%
ANXIOUS = no anxiety over past 2 wks,several days anxiety over past
0
0.41
3%
SOCIAL2 = NA,sometimes,rarely,never
0
0.03
91%
0
0.31
2%
1
0.60
1%
1
0.84
6%

That shows the fraction in each ‘leaf’ that are 1, where greener are more accurate and blue get less.

You can mess with the `cp` hyperparameter, the ‘complexity parameter’. A lower value allows more complexity and more branches to the tree. But takes more time to calculate!

This tree model is basically showing how symptoms tend to correlate. Maybe we wanted to know something about what populations had greater or lesser incidence, but that’s not revealed here. For a clinician, however, this is useful to validate that lack of interest in things and anxiety do correlate.

You might want to ask if the correlations change for certain population groups – perhaps anxiety or loneliness are different reported?

Or snip off some of those columns from the data, to see if the tree model can reveal other information.

```
d_down_for_tree2 <- select(d_down_for_tree, - ANXIOUS, - WORRY, - INTEREST,
                           - SOCIAL1, - SOCIAL2, - SUPPORT1, - SUPPORT2,
                           - SUPPORT3, - SUPPORT4, - SUPPORT1EXP)
```

```
tree_mod3 <- rpart(down_severe ~ ., data = d_down_for_tree2, method = "class", cp = 0
tree_mod3
```

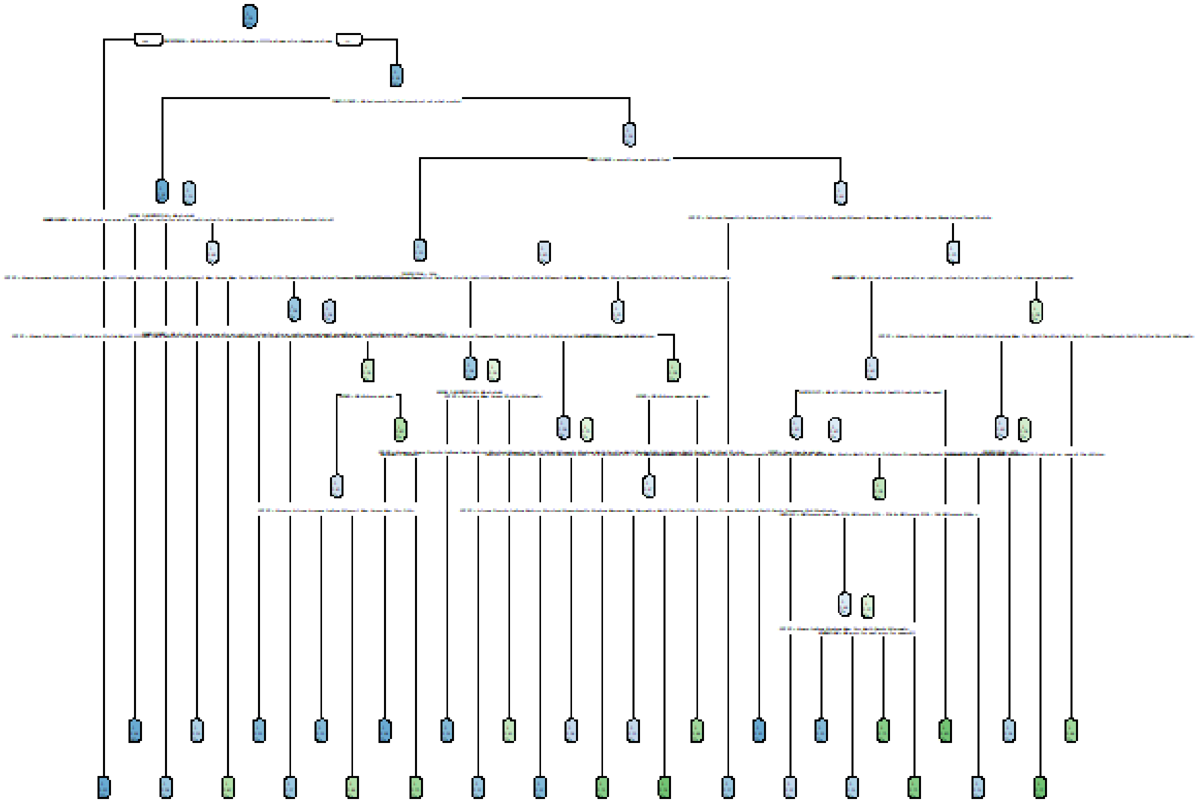
```
## n= 65552
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##    1) root 65552 5814 0 (0.91130705 0.08869295)
##      2) PRICESTRESS=NA,Moderate stress price changes,a little stress price changes
##      3) PRICESTRESS=very stressed about price changes 18092 3512 0 (0.80588105 0.1
##        6) CURFOODSUF=NA,had enough food,had enough but not what wanted 14540 2299
##         12) SEXUAL_ORIENTATION=NA,straight 12775 1753 0 (0.86277886 0.13722114) *
##         13) SEXUAL_ORIENTATION=gay or lesbian,bisexual,something else,dont know 17
##           26) RSNNOWRKRV=NA,did not want,am/was sick w covid or caring for sick w
##           27) RSNNOWRKRV=caring for elderly,retired,employer closed because covid
##             54) EST_ST=Alaska,Arkansas,Colorado,Florida,Georgia,Hawaii,Illinois,Ke
##             55) EST_ST=Alabama,Arizona,California,Connecticut,Delaware,District of
##        7) CURFOODSUF=sometimes not enough food,often not enough food 3552 1213 0 (
##         14) CURFOODSUF=sometimes not enough food 2562  780 0 (0.69555035 0.3044496
##           28) TBIRTH_YEAR< 1983.5 1742  448 0 (0.74282434 0.25717566)
##             56) RSNNOWRKRV=NA,did not want,am/was sick w covid or caring for sick
##             57) RSNNOWRKRV=caring for kids,caring for elderly,retired,laid off 446
##              114) EST_ST=Alaska,Colorado,Connecticut,Delaware,Florida,Hawaii,Illin
##              115) EST_ST=Alabama,Arizona,Arkansas,California,District of Columbia,
##                230) EEDUC=HS diploma,adv deg 57   23 0 (0.59649123 0.40350877)
##                  460) EST_ST=Alabama,Arizona,Arkansas,Indiana,Missouri,New Jersey,
##                  461) EST_ST=California,Georgia,Idaho,Kansas,Minnesota,New Hampshi
##                231) EEDUC=less than hs,some hs,some coll,assoc deg,bach deg 110
##                  462) RHISPANIC=Hispanic 11    2 0 (0.81818182 0.18181818) *
##                  463) RHISPANIC=Not Hispanic 99   30 1 (0.30303030 0.69696970) *
##           29) TBIRTH_YEAR>=1983.5 820  332 0 (0.59512195 0.40487805)
##             58) EST_ST=California,Colorado,Connecticut,Delaware,Florida,Idaho,Illi
##              116) SEXUAL_ORIENTATION=NA,straight 272   56 0 (0.79411765 0.20588235
##              117) SEXUAL_ORIENTATION=gay or lesbian,bisexual,something else,dont k
##                234) EST_ST=Delaware,New Jersey,Virginia,Wisconsin 20    5 0 (0.750
##                235) EST_ST=California,Colorado,Connecticut,Florida,Idaho,Illinois,
##             59) EST_ST=Alabama,Alaska,Arizona,Arkansas,District of Columbia,Georgi
##              118) MHLTH_NEED=no, none of the children 174   62 0 (0.64367816 0.356
##                236) EST_ST=Alabama,Alaska,Georgia,Indiana,Iowa,Kentucky,Maryland,M
##                237) EST_ST=Arizona,Arkansas,Hawaii,Mississippi,Nebraska,New Hampsh
##                  474) INCOME=HH income less than $25k,HH income $50k - 74.9,HH inc
##                  475) INCOME=NA,HH income $25k - $34.9k,HH income $35k - 49.9 28
##              119) MHLTH_NEED=NA,all children need mental health treatment,some but
##                238) EEDUC=HS diploma,assoc deg,adv deg 116   54 0 (0.53448276 0.46
##                  476) EST_ST=Arizona,Georgia,Indiana,Kentucky,Maryland,Massachuset
##                  477) EST_ST=Alabama,Alaska,Arkansas,Iowa,Michigan,Minnesota,Missi
##                239) EEDUC=less than hs,some hs,some coll,bach deg 176   60 1 (0.34
##         15) CURFOODSUF=often not enough food 990  433 0 (0.56262626 0.43737374)
##           30) EST_ST=Colorado,Connecticut,Delaware,Florida,Hawaii,Illinois,Maine,M
##          31) EST_ST=Alabama,Alaska,Arizona,Arkansas,California,District of Columb
##            62) RSNNOWRKRV=NA,did not want,am/was sick w covid or caring for sick 
##             124) MHLTH_GET=NA,all children get the mental health treatment they n
##               248) EEDUC=less than hs,adv deg 40    4 0 (0.90000000 0.10000000) *
##               249) EEDUC=some hs,HS diploma,some coll,assoc deg,bach deg 370  168
##                 498) EST_ST=Alabama,Arizona,Arkansas,Georgia,Iowa,Massachusetts,M
##                 499) EST_ST=Alaska,California,District of Columbia,Idaho,Indiana,
##                   998) INCOME=HH income less than $25k,HH income $25k - $34.9k,HH
##                    1996) EST_ST=Alaska,Indiana,Montana,New York,North Dakota,Wisc
##                    1997) EST_ST=California,Idaho,Kansas,Kentucky,Louisiana,Minnes
##                      3994) KINDWORK=NA,work for govt,work for nonprofit 26    8 0
##                      3995) KINDWORK=work for private co,self employed,work in fam
##                   999) INCOME=NA,HH income $35k - 49.9,HH income $50k - 74.9,HH i
##             125) MHLTH_GET=some but not all children,no, none of the children 30 
##            63) RSNNOWRKRV=caring for elderly,sick or disabled,retired,laid off,em
##             126) EST_ST=Alaska,Georgia,Indiana,Kansas,Louisiana,Michigan,Montana,
##               252) TBIRTH_YEAR< 1977.5 59   18 0 (0.69491525 0.30508475) *
##               253) TBIRTH_YEAR>=1977.5 41   18 1 (0.43902439 0.56097561)  
##                 506) MHLTH_NEED=all children need mental health treatment,no, non
##                 507) MHLTH_NEED=NA,some but not all children 24    6 1 (0.2500000
##             127) EST_ST=Alabama,Arizona,Arkansas,California,District of Columbia,
```

I had to reduce the complexity parameter penalty. This reveals other features, such as answers about whether the person is stressed about price changes or having enough food. I don’t know if that tells us that economic precarity correlates or if those are just another symptom. Although it never surfaces household income as important so perhaps the latter.

With all of the branches, the plots are a mess. I’ll offer a few options – the `prp` shows less information; both offer a `tweak` parameter to increase size of text even though it overwrites. You could do better.

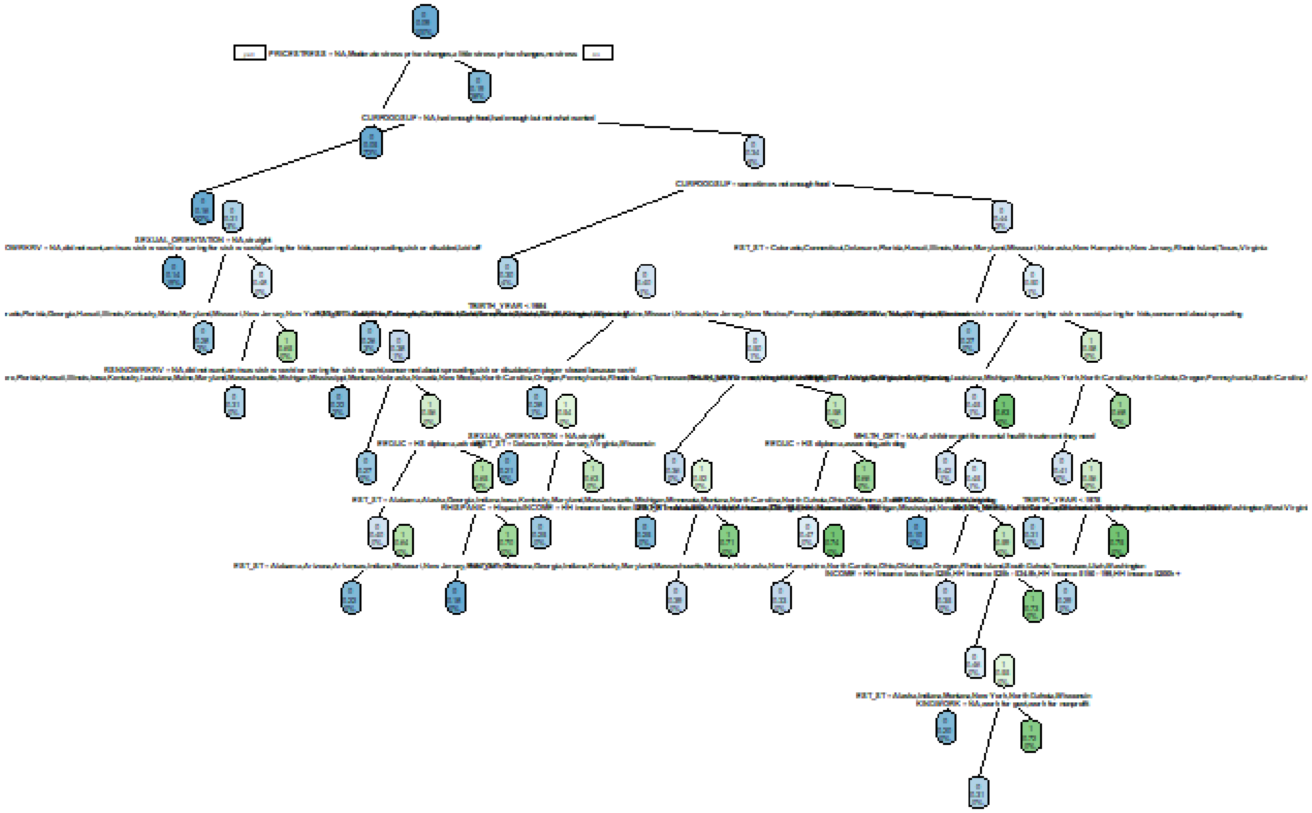
```
rpart.plot(tree_mod3)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



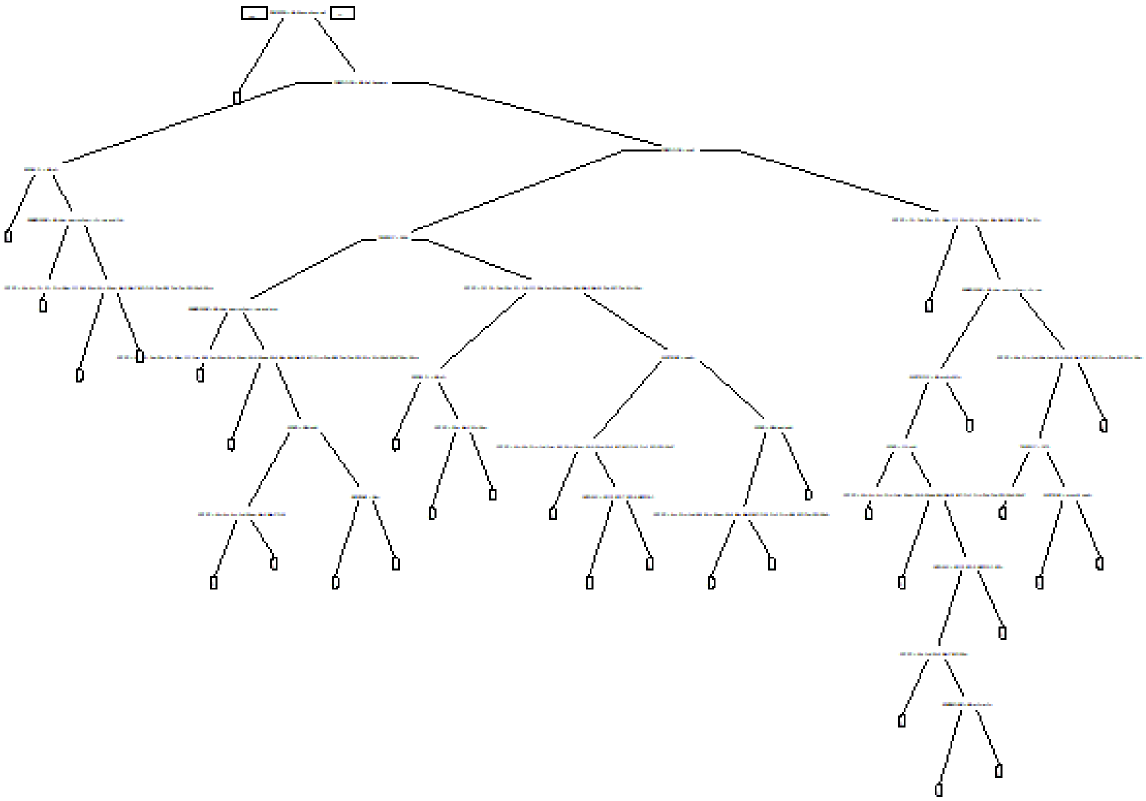
```
rpart.plot(tree_mod3, fallen.leaves = FALSE, tweak = 1.5)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```


yes
PRICESTRESS = NA,Moderate stress price changes,a little stress price changes,no stress
no
CURFOODSUF = NA,had enough food,had enough but not what wanted
CURFOODSUF = sometimes not enough food
SEXUAL_ORIENTATION = NA,straight
BIRTH_YEAR < 1984
BIRTH_YEAR < 1978

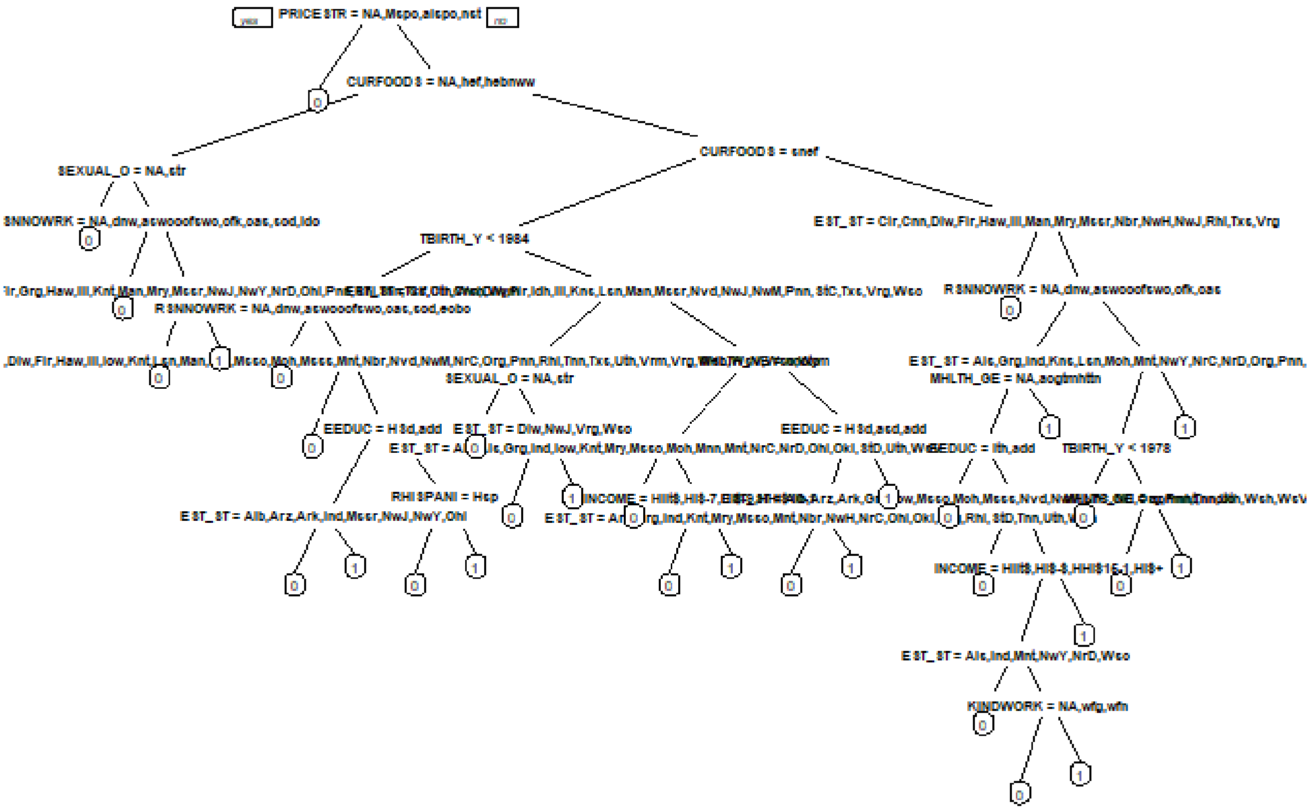
```
prp(tree_mod3)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



```r
prp(tree_mod3, fallen.leaves = FALSE, tweak = 3)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```


yes
PRICESTR = NA,Mspo,alspo,nst
no
CURFOODS = NA,hef,hebnww
SEXUAL_O = NA,str
CURFOODS = snef
SNNOWRK = NA,dnw,aswooofswo,ofk,oas,sod,ldo
TBIRTH_Y < 1984
EST_ST = Clr,Cnn,Dlw,Flr,Haw,Ill,Man,Mry,Mssr,Nbr,NwH,NwJ,Rhl,Txs,Vrg
lr,Grg,Haw,Ill,Knt,Man,Mry,Mssr,NwJ,NwY,NrD,Ohl,Pnn
R SNNOWRK = NA,dnw,aswooofswo,oas,sod,eobo
Man,Mssr,Nvd,NwJ,NwM,Pnn,StC,Txs,Vrg,Wso
R SNNOWRK = NA,dnw,aswooofswo,ofk,oas
,Dlw,Flr,Haw,Ill,low,Knt,Lsn,Man,
,Mssо,Moh,Msss,Mnt,Nbr,Nvd,NwM,NrC,Org,Pnn,Rhl,Tnn,Txs,Uth,Vrm,Vrg,
SEXUAL_O = NA,str
EST_ST = Als,Grg,Ind,Kns,Lsn,Moh,Mnt,NwY,NrC,NrD,Org,Pnn,
MHLTH_GE = NA,aogtmhttn
EEDUC = H Sd,add
EST_ST = Dlw,NwJ,Vrg,Wso
EEDUC = H Sd,asd,add
EST_ST = Al
ls,Grg,Ind,low,Knt,Mry,Mssо,Moh,Mnn,Mnt,NrC,NrD,Ohl,Okl,StD,Uth,W
EEDUC = lth,add
TBIRTH_Y < 1978
RHISPANI = Hsp
INCOME = HIlt$,HI$-7,
,Arz,Ark,Gr
ow,Mssо,Moh,Msss,Nvd,N
,Wsh,WsV
EST_ST = Alb,Arz,Ark,Ind,Mssr,NwJ,NwY,Ohl
EST_ST = Ar
rg,Ind,Knt,Mry,Mssо,Mnt,Nbr,NwH,NrC,Ohl,Okl,
,Rhl,StD,Tnn,Uth,W
INCOME = HIlt$,HI$-$,HHI$15,1,HI$+
EST_ST = Als,Ind,Mnt,NwY,NrD,Wso
KINDWORK = NA,wfg,wfn

If you worry about overfitting tree models, there's random forest, which estimates trees on random subsets of the data.

You can go back to OLS and logit models to confirm that there's not much predictive power in all of these. (not run here, you can do for yourself)

```
ols_down1 <- lm(down_severe ~ ANXIOUS + WORRY + INTEREST +
                  SOCIAL1 + SOCIAL2 +
                  SUPPORT1 + SUPPORT2 + SUPPORT3 + SUPPORT4 + SUPPORT1EXP +
                  TBIRTH_YEAR + RHISPANIC*RRACE + EEDUC + MS +
                  EGENID_BIRTH + GENID_DESCRIBE + SEXUAL_ORIENTATION +
                  REGION, data = d_down)

stargazer(ols_down1, type = "text")
logit_down1 <- glm(down_severe ~ ANXIOUS + WORRY + INTEREST +
                     SOCIAL1 + SOCIAL2 +
                     SUPPORT1 + SUPPORT2 + SUPPORT3 + SUPPORT4 + SUPPORT1EXP +
                     TBIRTH_YEAR + RHISPANIC*RRACE + EEDUC + MS +
                     EGENID_BIRTH + GENID_DESCRIBE + SEXUAL_ORIENTATION +
                     REGION, data = d_down,
                   family = binomial)
stargazer(logit_down1, type = "text")
stargazer(ols_down1,logit_down1, type = "text")
```

Finally I’ll show some code for multilevel models, although it doesn’t help much in this case. There are many other places where it would be more useful. Note the regular `summary` command gives additional useful information that `stargazer` doesn’t.

```
require(lme4)

model_mm1 <- lmer(down_severe ~ (1 | EST_ST) +
                    TBIRTH_YEAR + RHISPANIC*RRACE + EEDUC + MS, data = d_down)

stargazer(model_mm1, type = "text")
```

```
##
## ========================================================
##                                  Dependent variable:
##                              ---------------------------
##                                      down_severe
## --------------------------------------------------------
## TBIRTH_YEAR                           0.002***
##                                       (0.0001)
##
## RHISPANICHispanic                     -0.015***
##                                        (0.004)
##
## RRACEBlack                            -0.015***
##                                        (0.004)
##
## RRACEAsian                            -0.032***
##                                        (0.005)
##
## RRACEOther                            0.032***
##                                        (0.005)
##
## EEDUCsome hs                          -0.056***
##                                        (0.015)
##
## EEDUCHS diploma                       -0.060***
##                                        (0.013)
##
## EEDUCsome coll                        -0.067***
##                                        (0.013)
##
## EEDUCassoc deg                        -0.082***
##                                        (0.013)
##
```

```
## EEDUCbach deg                         -0.109***
##                                        (0.013)
##
## EEDUCadv deg                          -0.121***
##                                        (0.013)
##
## MSmarried                             -0.048**
##                                        (0.020)
##
## MSwidowed                              -0.013
##                                        (0.021)
##
## MSdivorced                              0.007
##                                        (0.020)
##
## MSseparated                            0.052**
##                                        (0.022)
##
## MSnever                                 0.018
##                                        (0.020)
##
## RHISPANICHispanic:RRACEBlack            0.026
##                                        (0.017)
##
## RHISPANICHispanic:RRACEAsian            0.011
##                                        (0.024)
##
## RHISPANICHispanic:RRACEOther          0.040***
##                                        (0.012)
##
## Constant                              -4.475***
##                                        (0.154)
##
## --------------------------------------------------------
## Observations                           65,552
## Log Likelihood                       -9,135.082
## Akaike Inf. Crit.                    18,314.160
## Bayesian Inf. Crit.                  18,514.160
## ========================================================
## Note:                        *p<0.1; **p<0.05; ***p<0.01
```

```
summary(model_mm1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: down_severe ~ (1 | EST_ST) + TBIRTH_YEAR + RHISPANIC * RRACE +
##     EEDUC + MS
##    Data: d_down
## 
## REML criterion at convergence: 18270.2
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -1.3113 -0.4259 -0.2592 -0.1062  3.8296 
## 
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  EST_ST   (Intercept) 5.902e-05 0.007682
##  Residual             7.714e-02 0.277746
## Number of obs: 65552, groups:  EST_ST, 51
## 
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                  -4.475e+00  1.537e-01 -29.107
## TBIRTH_YEAR                   2.375e-03  7.751e-05  30.647
## RHISPANICHispanic            -1.474e-02  4.440e-03  -3.321
## RRACEBlack                   -1.460e-02  4.116e-03  -3.547
## RRACEAsian                   -3.180e-02  5.382e-03  -5.909
## RRACEOther                    3.167e-02  5.494e-03   5.764
## EEDUCsome hs                 -5.572e-02  1.485e-02  -3.752
## EEDUCHS diploma              -6.012e-02  1.269e-02  -4.738
## EEDUCsome coll               -6.668e-02  1.257e-02  -5.303
## EEDUCassoc deg               -8.240e-02  1.280e-02  -6.439
## EEDUCbach deg                -1.085e-01  1.255e-02  -8.648
## EEDUCadv deg                 -1.211e-01  1.257e-02  -9.633
## MSmarried                    -4.785e-02  2.033e-02  -2.353
## MSwidowed                    -1.317e-02  2.072e-02  -0.635
## MSdivorced                    7.183e-03  2.045e-02   0.351
## MSseparated                   5.167e-02  2.183e-02   2.367
## MSnever                       1.834e-02  2.047e-02   0.896
## RHISPANICHispanic:RRACEBlack  2.555e-02  1.714e-02   1.491
## RHISPANICHispanic:RRACEAsian  1.070e-02  2.375e-02   0.451
## RHISPANICHispanic:RRACEOther  3.998e-02  1.193e-02   3.353
```

For the lab, I’d like each group to pick a different dependent variable to explore. Maybe look at anxiety, worry, loss of interest, loneliness or social and emotional suppport; maybe look at different ways people get support though phone, text, getting together in person, at religious ceremonies or other organizations. Pick one of those that seems interesting and tune a tree model to see what are important correlates and what you might learn from those.