Panther CLOGIT

Wenjing Xu

4/17/2020

dataframe prep

```
# reclassify vegetation class
data <- read.csv("Final_Variable_all.csv")
veg.code <- read.csv("VegCombineCode.csv")
# reshape data
data <- data %>% dplyr::left_join(veg.code, by = "CLC_STATE")
data.full <- data %>% filter (!is.na(Fire_Cat))
# reduce to 10 random points for each den
set.seed(17)
data.0 <- data.full %>% group_by (DenID) %>% filter( Used == 0 ) %>% sample_n(10)
data.1 <- data.full %>% filter( Used == 1 )
data <- union (data.1, data.0) %>% arrange (DenID)

# specify factors. Levels order will affect the baseline level in the modeling process
data$DenID <- as.factor(data$DenID)
data$Fire_Cat <- factor(data$Fire_Cat, levels = c("U", "A", "B", "C", "D"))
```

Examining variables

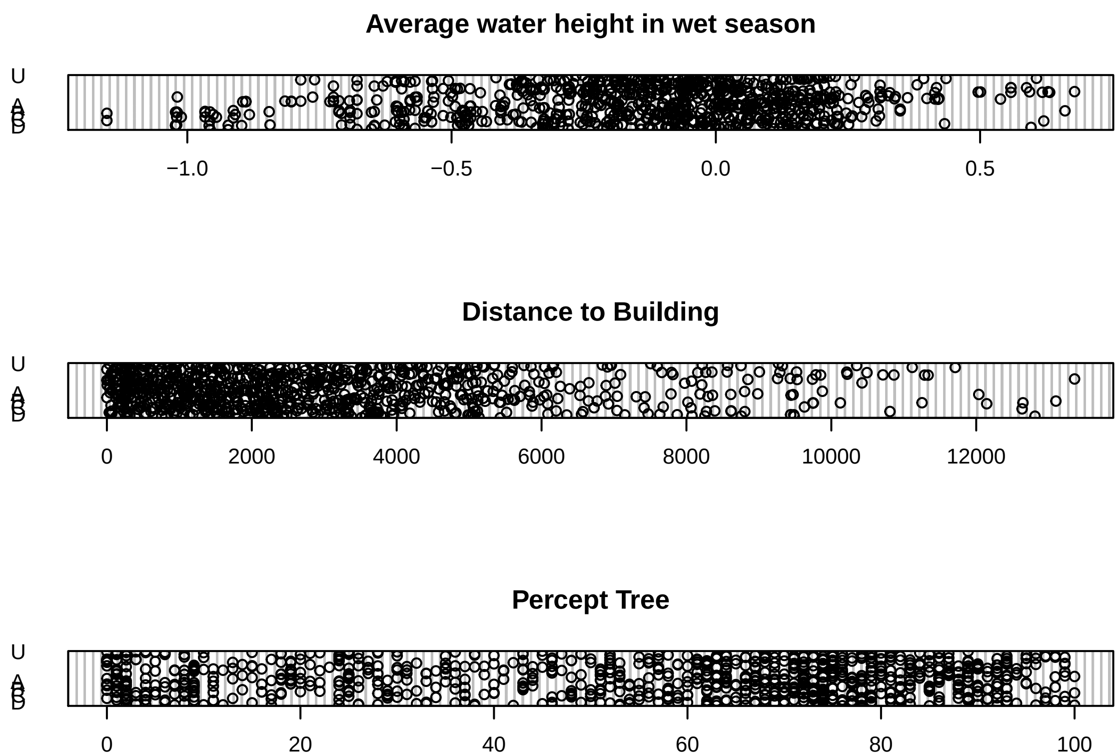
```
data %>% group_by(Combined) %>% summarise(used.count = n())
```

```
## # A tibble: 5 x 2
##   Combined          used.count
##   <fct>                  <int>
## 1 Marsh-Shrub-Swamp        286
## 2 Other                     13
## 3 Prairie-Grassland         91
## 4 Upland Forest             98
## 5 Wetland Forest           425
```

```
# get rid of the factors that are ecologically unlikely to be meaningful
# or that likely do not have much explanatory power (drop CatID because not enough replicates for each
# or are obviously correlated (e.g. ave_wet and ave_dry)
# also deleted age because does not really show importance but interfere with the modeling process from
data <- data %>% dplyr::select(Used, DenID, Fire_Cat, Ave_Wet, Prec_Tree, dist2build, Combined) %>% dply
data$Veg_Class <- factor(data$Veg_Class, levels = c("Upland Forest", "Other", "Marsh-Shrub-Swamp", "Prai
```

data exploration

```
op <- par(mfrow = c(3, 1))
dotchart(data$Ave_Wet, main = "Average water height in wet season", group = data$Fire_Cat)
dotchart(data$dist2build, main = "Distance to Building", group = data$Fire_Cat)
dotchart(data$Prec_Tree, main = "Percept Tree", group = data$Fire_Cat)
```




```
par(op)
```

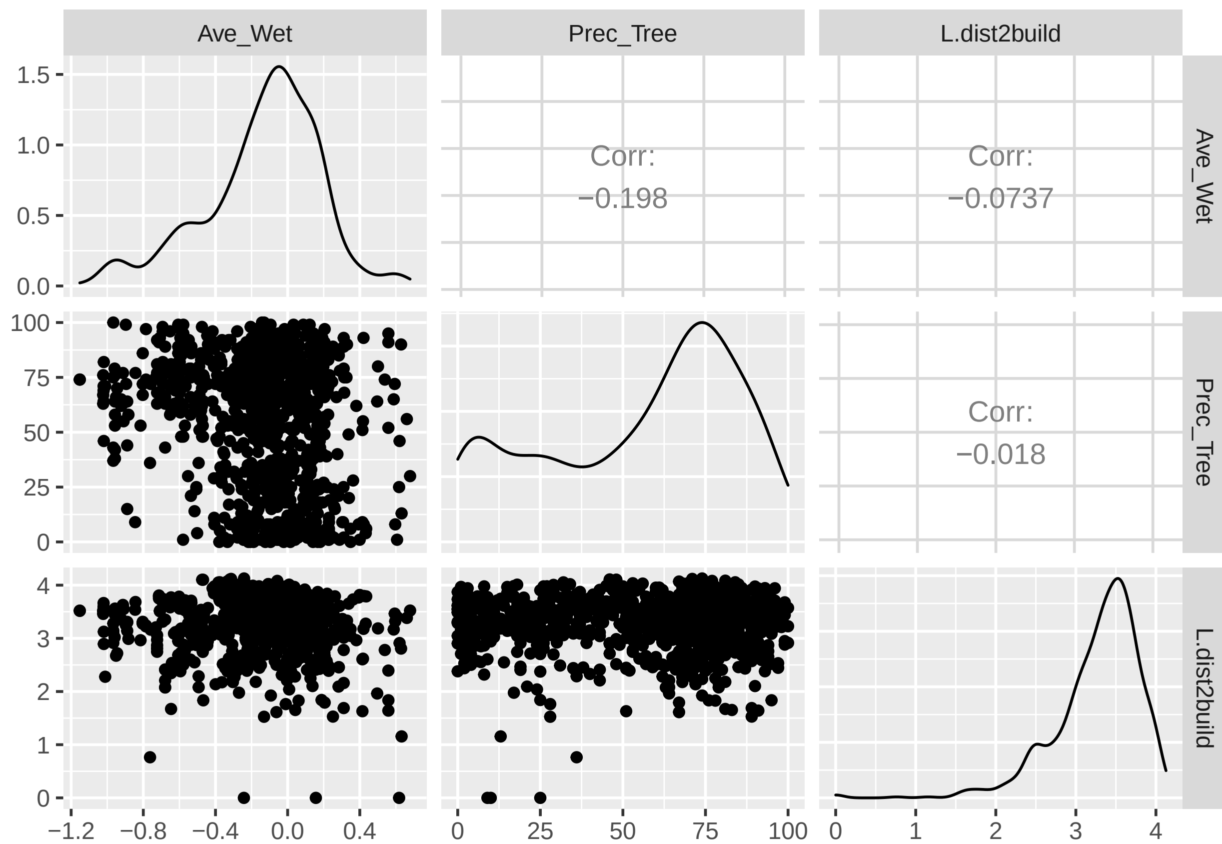
Age and dist2build has some outliers at the far end. So do transformation on these two.

```
data$L.dist2build <- log10(data$dist2build+1) #since some of the dist is 0
data.L <- data %>% dplyr::select(Used, DenID, Fire_Cat, Ave_Wet, Prec_Tree, L.dist2build, Veg_Class)

cor(data.L[,c(4:6)]) #does not seem to have strong linear relations
```

```
##                 Ave_Wet  Prec_Tree L.dist2build
## Ave_Wet       1.00000000 -0.1981786  -0.07366867
## Prec_Tree    -0.19817865  1.0000000  -0.01799120
## L.dist2build -0.07366867 -0.0179912   1.00000000
```

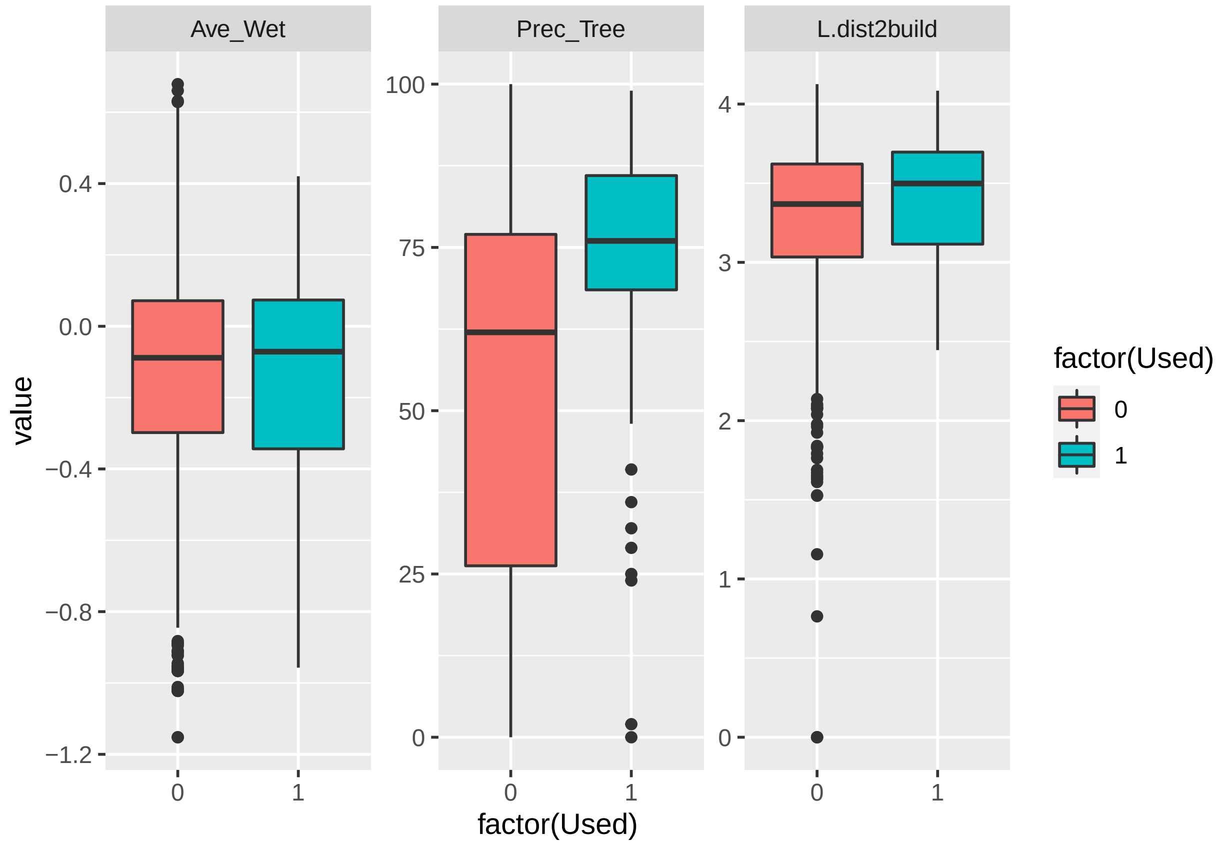
```r
ggpairs(data.L[, c( "Ave_Wet", "Prec_Tree", "L.dist2build")])
```



This plot further confirmed that there are no obvious linear relations.

We have a lot of points in U. Unbalanced dataset. Later for modeling we will create one dataset that combines A and B, and C and D.

```r
data.2 <- melt(data.L[, c("Used", "Ave_Wet", "Prec_Tree", "L.dist2build")], id.vars="Used")
ggplot(data.2, aes(factor(Used), y = value, fill=factor(Used))) +
    geom_boxplot() +
    facet_wrap(~variable, scales="free_y")
```



Super clear that they pick high tree precentage area. slightly far away from building, almost no difference in age and ave wet. We expect precentage tree will play an important role.

fitting conditional logistic regression model and model selection

The model selection process can be described as evaluating candidate models (with Fire_cat always be one variable). We considered relevant interactions into the most parsimonious models to see if they would increase fit, but models with interactions always ranked poorer (higher AICc) than models excluding them.

```
# full model
m.0 <- clogit(Used ~ Fire_Cat + Prec_Tree + Veg_Class + Ave_Wet + L.dist2build + strata(DenID), data =
summary(m.0)
```

```
## Call:
## coxph(formula = Surv(rep(1, 913L), Used) ~ Fire_Cat + Prec_Tree +
##     Veg_Class + Ave_Wet + L.dist2build + strata(DenID), data = data.L,
##     method = "efron")
##
##   n= 913, number of events= 83
##
##                               coef exp(coef)  se(coef)     z Pr(>|z|)
## Fire_CatA                 0.262255  1.299858  0.423529  0.619   0.5358
## Fire_CatB                 0.451709  1.570994  0.448292  1.008   0.3136
## Fire_CatC                 0.728005  2.070945  0.389094  1.871   0.0613 .
## Fire_CatD                 0.730217  2.075531  0.484937  1.506   0.1321
## Prec_Tree                 0.038032  1.038764  0.007038  5.404 6.51e-08 ***
```

```
## Veg_ClassOther              1.173883  3.234529  0.962944  1.219   0.2228
## Veg_ClassMarsh-Shrub-Swamp -0.540321  0.582561  0.449621 -1.202   0.2295
## Veg_ClassPrairie-Grassland -1.018116  0.361275  0.577134 -1.764   0.0777 .
## Veg_ClassWetland Forest    -0.665926  0.513798  0.348579 -1.910   0.0561 .
## Ave_Wet                     0.005177  1.005191  0.746553  0.007   0.9945
## L.dist2build                0.961724  2.616203  0.351396  2.737   0.0062 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                            exp(coef) exp(-coef) lower .95 upper .95
## Fire_CatA                     1.2999     0.7693    0.5667     2.981
## Fire_CatB                     1.5710     0.6365    0.6525     3.782
## Fire_CatC                     2.0709     0.4829    0.9660     4.440
## Fire_CatD                     2.0755     0.4818    0.8023     5.369
## Prec_Tree                     1.0388     0.9627    1.0245     1.053
## Veg_ClassOther                3.2345     0.3092    0.4900    21.354
## Veg_ClassMarsh-Shrub-Swamp    0.5826     1.7166    0.2413     1.406
## Veg_ClassPrairie-Grassland    0.3613     2.7680    0.1166     1.120
## Veg_ClassWetland Forest       0.5138     1.9463    0.2595     1.017
## Ave_Wet                       1.0052     0.9948    0.2327     4.342
## L.dist2build                  2.6162     0.3822    1.3139     5.209
##
## Concordance= 0.745  (se = 0.032 )
## Likelihood ratio test= 66  on 11 df,   p=7e-10
## Wald test            = 49.25  on 11 df,   p=9e-07
## Score (logrank) test = 60.62  on 11 df,   p=7e-09
```

```
# # if combine fire cat
# data.LC <- data.L %>% mutate(Fire_Cat_C = ifelse (Fire_Cat == "A" | Fire_Cat == "B", "AB",
#                              ifelse (Fire_Cat == "U", "U", "CD")))
# data.LC$Fire_Cat_C <- factor(data.LC$Fire_Cat_C, levels = c("AB", "CD", "U"))
# m.0c <- clogit(Used ~ Fire_Cat_C + Prec_Tree + Veg_Class + Ave_Wet + L.dist2build + strata(DenID), da
# summary(m.0c)
# # combining class attenuate impacts of fire class C... decide to go back to not combined classes

m.1 <- clogit(Used ~ Fire_Cat + Prec_Tree + Veg_Class + L.dist2build + strata(DenID), data = data.L, me
m.2 <- clogit(Used ~ Fire_Cat + Prec_Tree  + Veg_Class+ Ave_Wet + strata(DenID), data = data.L, method=
m.3 <- clogit(Used ~ Fire_Cat + Veg_Class + Ave_Wet + L.dist2build + strata(DenID), data = data.L, meth
m.4 <- clogit(Used ~ Fire_Cat + Prec_Tree + Ave_Wet + L.dist2build + strata(DenID), data = data.L, meth
AICc(m.0, m.1,m.2,m.3,m.4)
```

```
##     df     AICc
## m.0 11 357.7658
## m.1 10 355.1031
## m.2 10 363.7658
## m.3 10 388.5831
## m.4  7 355.0989
```

The term exp(coef) is giving the odds ratio for an increase of 1 unit in the independent variable. We drop vegclass here.

```
m.10 <- clogit(Used ~ Fire_Cat + Prec_Tree + Ave_Wet + L.dist2build + strata(DenID), data = data.L, met
m.11 <- clogit(Used ~ Fire_Cat + Prec_Tree + L.dist2build + strata(DenID), data = data.L, method='efron
```

```r
m.12 <- clogit(Used ~ Fire_Cat + Prec_Tree + Ave_Wet + strata(DenID), data = data.L, method='efron')
m.13 <- clogit(Used ~ Fire_Cat  + Ave_Wet + L.dist2build + strata(DenID), data = data.L, method='efron')
AICc(m.10, m.11, m.12, m.13)
```

```
##      df     AICc
## m.10  7 355.0989
## m.11  6 352.7164
## m.12  6 360.4667
## m.13  6 402.7659
```

Drop ave_wet.

```r
m.100 <- clogit(Used ~ Fire_Cat + Prec_Tree + L.dist2build + strata(DenID), data = data.L, method='efron
m.101 <-  clogit(Used ~ Fire_Cat + Prec_Tree + strata(DenID), data = data.L, method='efron')
m.102 <- clogit(Used ~ Fire_Cat + L.dist2build + strata(DenID), data = data.L, method='efron')

# test nonlinear effect of road
data.L2 <- data.L
data.L2$L.dist2build2 <- data.L2$L.dist2build*data.L2$L.dist2build
m.10001 <- clogit(Used ~ Fire_Cat + Prec_Tree + L.dist2build +  L.dist2build2 + strata(DenID), data = da

#test interaction terms
m.10002 <- clogit(Used ~ Fire_Cat + Fire_Cat:Prec_Tree + L.dist2build + strata(DenID), data = data.L, me
m.10003 <- clogit(Used ~ Prec_Tree + Fire_Cat:Prec_Tree + L.dist2build + strata(DenID), data = data.L, m

AICc(m.100, m.101, m.102, m.10001, m.10002, m.10003)
```

```
##         df     AICc
## m.100    6 352.7164
## m.101    5 358.1627
## m.102    5 400.4413
## m.10001  7 353.7842
## m.10002 10 358.3962
## m.10003  6 355.0943
```

Best model is m.100.

```r
summary(m.100)
```

```
## Call:
## coxph(formula = Surv(rep(1, 913L), Used) ~ Fire_Cat + Prec_Tree +
##     L.dist2build + strata(DenID), data = data.L, method = "efron")
##
##   n= 913, number of events= 83
##
##                  coef exp(coef) se(coef)     z Pr(>|z|)
## Fire_CatA    0.337913  1.402019 0.413952 0.816  0.41432
## Fire_CatB    0.549577  1.732521 0.433504 1.268  0.20488
## Fire_CatC    0.724843  2.064408 0.382156 1.897  0.05786 .
## Fire_CatD    0.808655  2.244888 0.472747 1.711  0.08716 .
## Prec_Tree    0.038191  1.038929 0.006241 6.119  9.4e-10 ***
```

```
## L.dist2build 0.900075  2.459789 0.345537 2.605  0.00919 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## Fire_CatA        1.402     0.7133    0.6229     3.156
## Fire_CatB        1.733     0.5772    0.7408     4.052
## Fire_CatC        2.064     0.4844    0.9761     4.366
## Fire_CatD        2.245     0.4455    0.8888     5.670
## Prec_Tree        1.039     0.9625    1.0263     1.052
## L.dist2build     2.460     0.4065    1.2496     4.842
##
## Concordance= 0.741  (se = 0.032 )
## Likelihood ratio test= 58.44  on 6 df,   p=9e-11
## Wald test            = 43.51  on 6 df,   p=9e-08
## Score (logrank) test = 50.5  on 6 df,   p=4e-09
```

model validation

```
#source code https://github.com/basille/hab/blob/master/R/kfold.r
# in the output, cor is Spearman rank correlations r_s
k_fold <- kfold(m.100, k=5, nrepet = 5, jitter = FALSE, reproducible = TRUE, details = FALSE)
rs.mean <- k_fold %>% dplyr::filter (type == "obs") %>% summarise(mean(cor))
rs.p <- t.test(k_fold %>% dplyr::filter (type == "obs") %>% dplyr::select(cor))
```

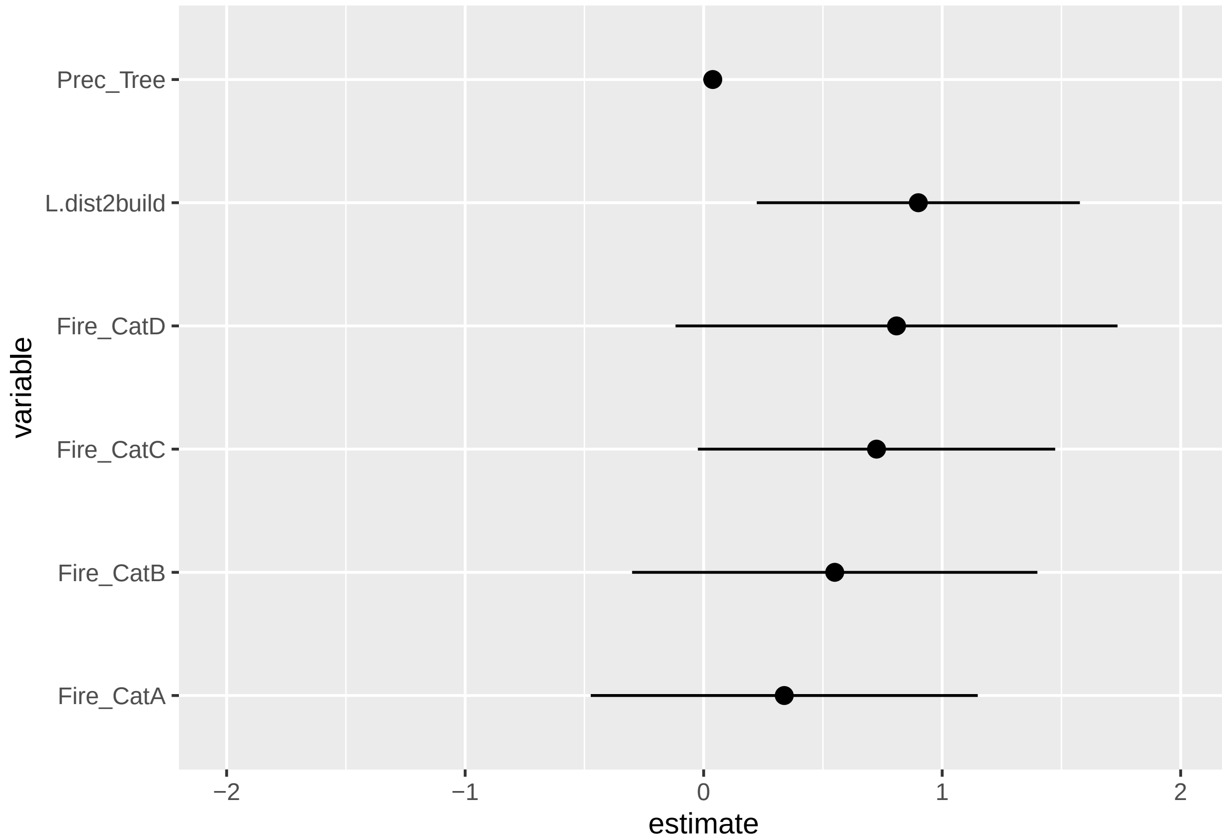
Residual deviance : 339.6111595 on 905 degree of freedom. Over dispersion is not a problem.

The 5-fold cross validation (sensu Boyce et al 2002) provided a mean Spearman's Rank corerltion of 0.75216932214027, P $= 4.4551423 \times 10^{-4}$, < 0.01. The model is pretty good.

result plots

```
co.table <- cbind(OR = coef(m.100), confint(m.100))
co.df <- data.frame(variable = row.names(co.table), estimate = co.table[1:nrow(co.table), 1], LL = co.ta
co.df$variable <- as.character(co.df$variable)

ggplot(co.df, aes(y=estimate, x=variable, ymin=LL, ymax=UL)) + geom_pointrange() + coord_flip() + ylim(
```



The selection coefficients estimated by the conditional logistic regression are the log odds ratio for a habitat type being chosen relative to a reference habitat type (beta = 0) (there is no intercept in clogit model). As such, selection for the reference habitat occurs when the other habitat types have beta < 0. ABCD are relatively more selected than U. But only C is slightly significant, others CI overlap with 0. CI of prec_tree is high.

```
newdata <- data.frame(DenID = factor(data.L$DenID[6]),
                      Fire_Cat = factor(rep(c("A", "B", "C", "D", "U"), each = 100)),
                      Prec_Tree = rep(seq(from = 0, to = 100, length.out = 100), 5),
                      L.dist2build = mean(data.L$L.dist2build))

newdata <- cbind(newdata, predict(m.100, newdata = newdata, type = "lp", se = TRUE))
newdata <- within(newdata, {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
  UL <- plogis(fit + (1.96 * se.fit))
})

p.1 <- ggplot(newdata, aes(x = Prec_Tree, y = PredictedProb)) + geom_ribbon(aes(ymin = LL,
    ymax = UL, fill = Fire_Cat), alpha = 0.2) + geom_line(aes(colour = Fire_Cat),
    size = 1)  + theme(legend.position = "none")


newdata <- data.frame(DenID = factor(data.L$DenID[8]),
                      Fire_Cat = factor(rep(c("A", "B", "C", "D", "U"), each = 100)),
                      Prec_Tree = mean(data.L$Prec_Tree),
```

```
                        L.dist2build = rep(seq(from = min(data.L$L.dist2build), to = max(data.L$L.dist2bu

newdata <- cbind(newdata, predict(m.100, newdata = newdata, type = "lp", se = TRUE))
newdata <- within(newdata, {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
  UL <- plogis(fit + (1.96 * se.fit))
})
p.2 <- ggplot(newdata, aes(x = (10^(L.dist2build)-1), y = PredictedProb)) + geom_ribbon(aes(ymin = LL,
    ymax = UL, fill = Fire_Cat), alpha = 0.2) + geom_line(aes(colour = Fire_Cat),
    size = 1)  + xlab("dist2build") + theme(legend.position = c(0.9, 0.2))

plot_grid(p.1, p.2)
```