Assignment 3

Zach Proux

1/30/2018

Univariate Assignment

Read in tree data, metadata can be found in: `./data/tree_metadata.txt`

```
library(knitr)
opts_knit$set(root.dir='../')
trees = read.csv('../Data/treedata.csv')
```

1 Carry out an exploratory analysis using the tree dataset. Develop and compare models for species cover for a habitat generalist Acer rubrum (Red maple) and a habitat specialist [Abies fraseri (Frasier fir)](https://upload.wikimedia.org/wikipedia/commons/d/d0/Abies_fraseri_Mitchell.jpg).

Because this dataset includes both continuous and discrete explanatory variables use the function `Anova` in the packages `car`

```
library(car)
```

```
tree_rm = trees[trees$spcode=="ACERRUB",]
tree_ff = trees[trees$spcode=="ABIEFRA",]
```

```
#generalist model
rm_mod = lm(tree_rm$cover ~ tree_rm$plotsize + tree_rm$utme + tree_rm$utmn +
            tree_rm$elev + tree_rm$tci + tree_rm$streamdist + tree_rm$disturb +
            tree_rm$beers)
summary(rm_mod)
#significant explanatory variables: plotsize, utmn, elev, streamdist, and beers.
```

```
rm_mod = update(rm_mod, ~ . - utme - tci - disturb, data = tree_rm)
summary(rm_mod)
Anova(rm_mod, type = 3)
```

```
#specialist model
ff_mod = lm(tree_ff$cover ~ tree_ff$plotsize + tree_ff$utme + tree_ff$utmn +
            tree_ff$elev + tree_ff$tci + tree_ff$streamdist + tree_ff$disturb +
            tree_ff$beers)
summary(ff_mod)
#significant explanatory variables: utmn, elevation, tci, and beers.
```

```
ff_mod = update(ff_mod, ~ . - tree_ff$streamdist - tree_ff$disturb - tree_ff$utme
                - tree_ff$plotsize)
summary(ff_mod)
Anova(ff_mod, type = 3)
```

Compare the p-values you observe using the function `Anova` to those generated using `summary`.

```
The p-values are the same in summary and in the 'Anova' table.
```

For each species address the following additional questions:

How well does the exploratory model appear to explain cover?

```
Our generalist model explains 4.1% of variation in cover and our
specialist model explains 41.2% of variation in cover.
```

Which explanatory variables are the most important?

> Elevation is the most important in both models. In the generalist model, beers and streamdist were also important. In the specialist model, tci and beers were important.

Do model diagnostics indicate any problems with violations of OLS assumptions?

```
#From the residual plots below, it doesn't look like any of the assumptions are violated.
plot(rm_mod)
plot(ff_mod)
```

Are you able to explain variance in one species better than another, why might this be the case?

> Yes, we can explain variance in the specialist much better which makes perfect sense because specialists tend to be tolerant of a much narrower range of conditions. This also means their characteristics might be heavily dependent on just a few variables as opposed to generalists that can live in a variety of habitats and be effected by a variety of different variables.

2 You may have noticed that the variable cover is defined as positive integers between 1 and 10. and is therefore better treated as a discrete rather than continuous variable.

Re-examine your solutions to the question above but from the perspective of a General Linear Model (GLM) with a Poisson error term (rather than a Gaussian one as in OLS).

The Poisson distribution generates integers 0 to positive infinity so this may provide a good first approximation.

Your new model calls will look as follows:

```
rm_glm = glm(cover ~ plotsize + utmn + elev + tci + streamdist + beers,
             data = tree_rm, family='poisson')
summary (rm_glm)
```

```
##
## Call:
## glm(formula = cover ~ plotsize + utmn + elev + tci + streamdist +
##     beers, family = "poisson", data = tree_rm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5729  -0.6135   0.1548   0.5955   2.0716
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.667e+01  1.235e+01   2.159 0.030844 *
## plotsize    -7.749e-05  3.707e-05  -2.090 0.036578 *
## utmn        -6.275e-06  3.133e-06  -2.003 0.045198 *
## elev        -1.890e-04  5.552e-05  -3.404 0.000663 ***
## tci         -9.498e-03  7.707e-03  -1.232 0.217778
## streamdist   2.345e-04  9.698e-05   2.418 0.015595 *
## beers       -5.734e-02  2.260e-02  -2.537 0.011191 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 721.46  on 793  degrees of freedom
## Residual deviance: 687.86  on 787  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 3411
##
## Number of Fisher Scoring iterations: 4
```

```
ff_glm = glm(cover ~ tci + utmn + elev + beers, data = tree_ff, family='poisson')
summary (ff_glm)
```

```
##
## Call:
## glm(formula = cover ~ tci + utmn + elev + beers, family = "poisson",
##     data = tree_ff)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.82725  -0.60859   0.07017   0.48130   2.07161
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.488e+01  4.994e+01   1.099  0.27179
## tci          1.142e-01  3.642e-02   3.137  0.00171 **
## utmn        -1.450e-05  1.278e-05  -1.135  0.25644
## elev         1.724e-03  3.771e-04   4.572 4.82e-06 ***
## beers        1.911e-01  7.109e-02   2.688  0.00720 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 98.822  on 101  degrees of freedom
## Residual deviance: 53.715  on  97  degrees of freedom
## AIC: 402.24
##
## Number of Fisher Scoring iterations: 4
```

For assessing the degree of variation explained you can use a pseudo-R-squared statistic (note this is just one of many possible)

```
    pseudo_r2 = function(glm_mod) {
        1 -  glm_mod$deviance / glm_mod$null.deviance
    }
pseudo_r2(rm_glm)
```

```
## [1] 0.04657819
```

```
pseudo_r2(ff_glm)
```

```
## [1] 0.4564418
```

Compare the residual sums of squares between the traditional OLS and glm models using `anova` (Note: not `Anova`)

```
The change in residual sum of swaures was negligible, although I can't
tell if my anova() ran correctly because it returned this statement:
"Models with response cover removed because response differs from model 1"
```

```
anova(rm_mod, rm_glm)
anova(ff_mod, ff_glm)
```

Does it appear that changing the error distribution changed the results much? In what ways?

```
The pseudo_r2 values in the GLM were slightly higher than the adjusted
r2 values from the OLS model, but only by 4% in the specialist model
and 0.4% in the generalist model.
```

3 Provide a plain English summary (i.e., no statistics) of what you have found and what conclusions we can take away from your analysis?

```
We learned that can explain about 45% of the variation in cover of a
specialist tree (Abies fraseri), but only about 4% in a generalist
(Acer rubrum).  We learned the residuals more closely follow a poisson
distribution than a gaussian one, but only slightly.  We also learned
elevation is the most important variable in explaining variation in cover
in both species, but only because we don't have the variables that
explain the majority of the variation in our models.
```

4 (optional) Examine the behavior of the function `step()` using the exploratory models developed above. This is a very simple and not very robust machine learning stepwise algorithm that uses AIC to select a best model. By default it does a backward selection routine.

5 (optional) Develop a model for the number of species in each site (i.e., unique plotID). This variable will also be discrete so the Poisson may be a good starting approximation. Side note: the Poisson distribution converges asymptotically on the Gaussian distribution as the mean of the distribution increases. Thus Poisson regression does not differ much from traditional OLS when means are large.

Ran into a problem when I was knitting my pdf. All of the chunks ran correctly

but for some reason it kept getting stuck at any lines with anova functions so

I coded 'eval = FALSE' and that got the knit to work