# Dog Breed Ordinal Logistic Regression Analysis

Xinyu Xie

2022/4/10

## Introduction

This project tries to address how different traits of different dogs breeds influence their popularity. In that case I can get what traits is usually liked by people, and what is not. It is can also be a good reference for me to choose my next dog as pet. Since our popularity is based on the rank in AKC registration statistics in 2020, and a ordinal value may fit the rank more, I perform an ordinal logistic regression in the analysis.

## Data Description

The dog breeds dataset comes from the American Kennel Club[1]. There are three important worksheet will be used in this project. “trait_description.csv” shows us the meaning of each traits and meaning of the value 1 to 5 represent for each traits. “breed_rank_all.csv” shows us the breeds popularity rank from the year 2013 to 2020, we will just focus on the rank in 2020 in the project. Finally, “breed_traits.csv” shows values for each trait and breed.

```
dplyr::select(descrip,Trait:Trait_5)
```

```
##                         Trait                   Trait_1
## 1    Affectionate With Family                Independent
## 2    Good With Young Children            Not Recommended
## 3        Good With Other Dogs            Not Recommended
## 4              Shedding Level                No Shedding
## 5     Coat Grooming Frequency                    Monthly
## 6              Drooling Level       Less Likely to Drool
## 7                  Coat Type                           -
## 8                Coat Length                           -
## 9       Openness To Strangers                   Reserved
## 10          Playfulness Level Only When You Want To Play
## 11 Watchdog/Protective Nature       What's Mine Is Yours
## 12         Adaptability Level          Lives For Routine
## 13         Trainability Level                Self-Willed
## 14               Energy Level              Couch Potato
## 15              Barking Level             Only To Alert
## 16   Mental Stimulation Needs            Happy to Lounge
##                       Trait_5
## 1                 Lovey-Dovey
```

```
## 2          Good With Children
## 3        Good With Other Dogs
## 4             Hair Everywhere
## 5                       Daily
## 6         Always Have a Towel
## 7                           -
## 8                           -
## 9  Everyone Is My Best Friend
## 10                  Non-Stop
## 11                  Vigilant
## 12           Highly Adaptable
## 13            Eager to Please
## 14                High Energy
## 15                 Very Vocal
## 16    Needs a Job or Activity
```

## Missing value and outliers

By simply implementing summary() function, I found some missing values in ranks of the year 2020, they can be removed by function na.omit(). There are some traits have values that is zero, which is not acceptable, as the traits value must be integer from 1 to 5.

```
summary(rank)
summary(traits)

traits <- traits %>%
  filter_at(vars(-
c("Coat.Type","Coat.Length","Breed")),any_vars(.>=1&.<=5))

rank <- na.omit(rank)
traits <- na.omit(traits)
```

## Transformation

First, we need to combine two rank and traits table by dog breeds into a dataframe. So, inner_join() helps us join all value in traits table into rank table by matching the variable breed. After remiving some useless columnns, I get all the data I need in a single dataframe.

As we mentioned above, the rank is more like a ordinal value, but we cannot have one value for each class. As a result, I create new variable called “Level”, we divided rank into 10 levels. Level 1 represent the least popular dog breed and level 10 represent the most popular breed. Here, is how Level distributed below.

```
##
##  1  2  3  4  5  6  7  8  9 10
## 13 16 14 18 20 16 20 20 19 19
```

# Methods and Results

## Simple full model

Since it is an ordinal logistic regression analysis, polr() function is used to fit the model. First we use all variable to fit a first-order model.

Before interpret the full first-order model, multicollinearity test is needed as it will influence the interpretation of parameters coefficient. Variance inflation factor shows us all the predict variables have normal VIF value (<10). However I decided to remove the variable Coat.Type for two reasons.

- The VIF = 8.67, which is means there is moderate correlation and not good enough.
- We have only 175 observations. If one predict variable have to have 10 observations at least, we have to much predict variable. Coat.Type exactly generate 7 more dummy variables for our model.

```
##                    GVIF Df GVIF^(1/(2*Df))
## Family         2.043244  1        1.429421
## Children       1.528535  1        1.236339
## Other.Dogs     1.563867  1        1.250547
## Shedding       1.603315  1        1.266221
## Coat.Grooming 2.293754  1        1.514514
## Drooling       1.633755  1        1.278184
## Coat.Type      8.668698  8        1.144517
## Coat.Length    5.337450  2        1.519965
## Strangers      1.633297  1        1.278005
## Playfulness    2.221219  1        1.490375
## Protective     1.415371  1        1.189694
## Adaptability   2.166907  1        1.472042
## Trainability   1.537553  1        1.239981
## Energy         1.884538  1        1.372785
## Barking        1.298426  1        1.139485
## Mental         2.044408  1        1.429828
```

We can get summary of a model by summary function directly. Here we get intercepts for each logit odd. We can see the intercept is getting higher from logit(1|2) to logit(9|10). That is because for a breed of dog that have level x, p(x<=1)/p(x>2) is clearly smaller than p(x<=9)/p(x>10), as the difference in the number of observations.

Take two coefficients for Interpreting:

1. Playfulness Value=1.01 exp(Value)=2.74
- Every one unit increase in Playfulness, the odds of being higher level is 2.74 times odds of be lower level
2. Barking Value=-0.47 exp(Value)=0.63

- Every one unit increase in Playfulness, the odds of being higher level is 0.64 times odds of be lower level

```
model_1 <-polr(Level~.,data=data_2020[,-7],Hess = TRUE)
model_1$zeta

##      1|2      2|3      3|4      4|5      5|6      6|7      7|8
8|9
## 3.434321 4.472383 5.070358 5.695303 6.307480 6.760771 7.340332
8.031607
##     9|10
## 8.988374

exp(coef(model_1))

##            Family          Children        Other.Dogs
Shedding
##         0.7244263         1.0750756         0.8864936
1.1481088
##     Coat.Grooming          Drooling Coat.LengthMedium
Coat.LengthShort
##         1.5524752         1.3384854         0.8498331
2.0352343
##         Strangers       Playfulness        Protective
Adaptability
##         1.2927175         2.7431366         1.3511351
1.1169774
##      Trainability            Energy           Barking
Mental
##         1.1042867         1.0023548         0.6267548
1.2537557
```

Same as other logistic regression, we test model utility by global F test. We can get a very low p value by pchisq() function, which means at least one beta is not 0.

I use MASS::Anova() function to calculate likelihood ratio-test, there are two variable are more important in the model. The fact can also be detect by confidence interval of each coefficient. However, from confidence interval, we can also found only three variables have the interval on one side of 1, which means we cannot reject the hypothesis that other variable is 0.

We can also detect if remove Coat.Type have us get much better model by anova() function. The p-value show us it is not a significant improvement, but an acceptable better model.

```
pchisq(deviance(model_0),df=df.residual(model_1),lower.tail = F)

## [1] 2.988424e-78

confint(model_1)

## Waiting for profiling to be done...
```

```
##                          2.5 %     97.5 %
## Family             -0.76397071  0.1141970
## Children           -0.25318754  0.3989441
## Other.Dogs         -0.46507392  0.2257875
## Shedding           -0.23131165  0.5085251
## Coat.Grooming       0.00333490  0.8768242
## Drooling           -0.02773531  0.6086375
## Coat.LengthMedium -1.17387296  0.8294127
## Coat.LengthShort  -0.46759330  1.8745033
## Strangers          -0.10962929  0.6337835
## Playfulness         0.43585974  1.5904408
## Protective         -0.03324031  0.6409592
## Adaptability       -0.51838625  0.7406680
## Trainability       -0.25969326  0.4594934
## Energy             -0.46816112  0.4779563
## Barking            -0.73712778 -0.2026347
## Mental             -0.31295706  0.7644764

Anova(model_1)

## Analysis of Deviance Table (Type II tests)
##
## Response: Level
##                LR Chisq Df Pr(>Chisq)
## Family           2.0951  1  0.1477734
## Children         0.1902  1  0.6627556
## Other.Dogs       0.4683  1  0.4937550
## Shedding         0.5385  1  0.4630605
## Coat.Grooming    3.9002  1  0.0482806 *
## Drooling         3.2068  1  0.0733315 .
## Coat.Length      6.4195  2  0.0403672 *
## Strangers        1.8771  1  0.1706595
## Playfulness     11.9407  1  0.0005492 ***
## Protective       3.1148  1  0.0775854 .
## Adaptability     0.1195  1  0.7295446
## Trainability     0.2943  1  0.5874749
## Energy           0.0001  1  0.9916729
## Barking         12.0510  1  0.0005177 ***
## Mental           0.6797  1  0.4096965
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(model_0,model_1)

## Likelihood ratio tests of ordinal regression models
##
## Response: Level
##
Model
## 1             Family + Children + Other.Dogs + Shedding +
Coat.Grooming + Drooling + Coat.Length + Strangers + Playfulness +
```

```
Protective + Adaptability + Trainability + Energy + Barking + Mental
## 2 Family + Children + Other.Dogs + Shedding + Coat.Grooming +
Drooling + Coat.Type + Coat.Length + Strangers + Playfulness +
Protective + Adaptability + Trainability + Energy + Barking + Mental
##   Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
## 1       150   744.5771
## 2       142   736.8897 1 vs 2     8 7.687364 0.4645895
```

## Diagnostics

The diagnostics of ordinal logistic is very different, many methods are given online, like sign-based statistic(SBS)[3], and Surrogate residuals[4]. Considered we cannot make diagnostics directly on ordinal logistic directly, surrogate residuals transform that into a continuous value R(s), R(s)=S − E (S|X)[4] If our model fit the true model, the Rs should have three properties as below.

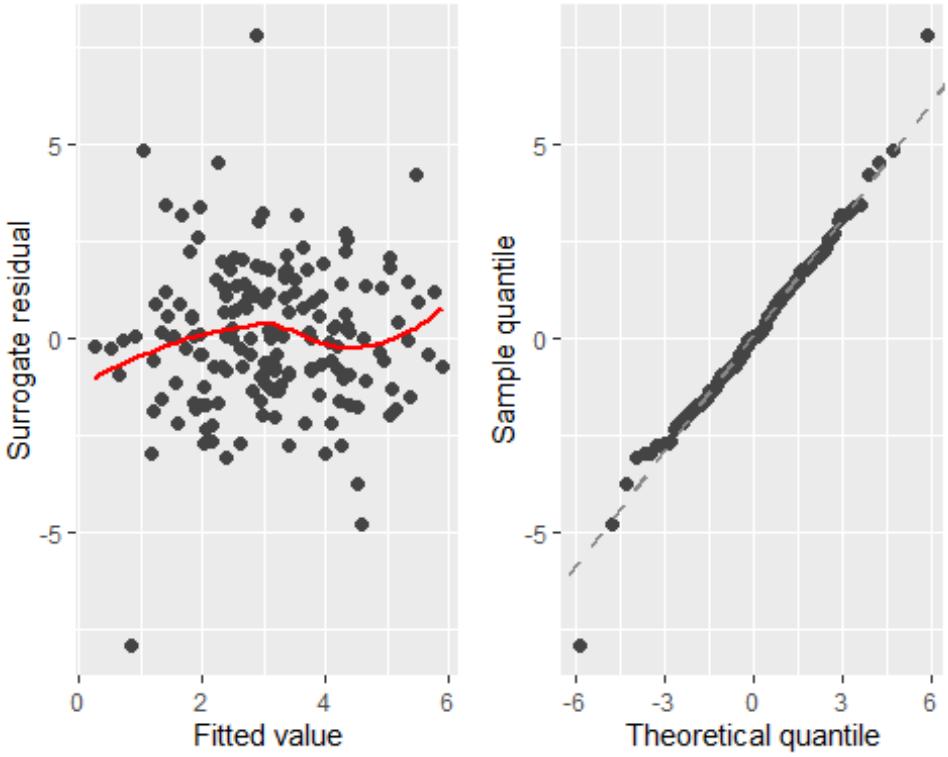
a. *(Symmetry around zero)* $E\{R \mid \boldsymbol{X}\} = 0$.

b. *(Homogeneous variance)* $Var\{R \mid \boldsymbol{X}\}$ *is a constant, not depending on* $\boldsymbol{X}$.

c. *(Explicit reference distribution)* $\sup_{c\in\mathbb{R}} \mid Q_n(c; R_1, ..., R_n) - G(c+ \int u\, dG(u)) \mid \to 0$ *almost surely as* $n \to \infty$, *where* $Q_n(c; R_1, \ldots, R_n) = \frac{1}{n}\sum_{i=1}^{n} I(R_i \le c)$ *is the empirical cumulative distribution function of* $\{R_1, ..., R_n\}$.

```
set.seed(99)
p1 <- autoplot.polr(model_0, what = "fitted")
set.seed(99)
p2 <- autoplot.polr(model_0, what = "qq")
grid.arrange(p1, p2, ncol = 2)

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




## Model building

Since I found I can only reject three variables' coefficient from 0, I would to fit a better model for this task. I implement two times of Stepwise regression by following steps. AIC is the only criteria to check whether model is better or worse.

1. I get a 10 variable first-order model from base model that has only intercept by Stepwise algorithm. The reason why the number is 10 is, for in plor(), it can only interact 10 variables,
2. Then I implement the second Stepwise regression from previous 10 variable model to their full interaction model.
3. Since we only have 175 observations, the model can have 17 variables at most.

Finally we build a model that have formula: Level ~ Family + Shedding + Coat.Grooming + Drooling + Coat.Length + Strangers + Playfulness + Protective + Barking + Mental + Coat.Grooming:Barking + Shedding:Mental + Family:Playfulness + Drooling:Coat.Length + Drooling:Mental + Strangers:Protective

We can also compare it with previous first-order simple full model, the p-value shows a significant improvement on our new model.

```
anova(model_1,model_2)
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: Level
##
Model
## 1
Family + Children + Other.Dogs + Shedding + Coat.Grooming + Drooling +
Coat.Length + Strangers + Playfulness + Protective + Adaptability +
Trainability + Energy + Barking + Mental
## 2 Family + Shedding + Coat.Grooming + Drooling + Coat.Length +
Strangers + Playfulness + Protective + Barking + Mental +
Coat.Grooming:Barking + Shedding:Mental + Family:Playfulness +
Drooling:Coat.Length + Drooling:Mental + Strangers:Protective
##   Resid. df Resid. Dev   Test    Df LR stat.      Pr(Chi)
## 1       150   744.5771
## 2       148   702.2457 1 vs 2     2 42.33141 6.424674e-10
```

## Discussion

From the interpretation of simple full model, we can get Barking and Playfulness are two most important traits for a dog that will influence their popularity. Dog owners usually prefer the dog that never stopping running, while they usually do not like bods that very vocal. Dog breed that usually need to be cleaned also get more popularity.

### Limitation

1. Influence points, leverage points, and outliers, I am not sure if there is a methods.
2. Diagnostics is not understood well. I am not sure if I can understand it, if I have enough time.

### Future work

1. Some prediction task can be implemented in the future, if I find better model.
2. Level 1 ~ Level 10 is the best way I can come up with to divide the rank, but I think there can be some distribution on rank value, which can help me get better ordinal value.

## Reference


1. kkakey, “I love dogs,” GitHub, Mar. 30, 2022.
2. D. Liu and H. Zhang, “Residuals and Diagnostics for Ordinal Regression Models: A Surrogate Approach,” Journal of the American Statistical Association, vol. 113, no. 522, pp. 845–854, Apr. 2018.

3. C. Li and B. E. Shepherd. A new residual for ordinal outcomes. Biometrika, 99(2):473–480, 2012. URL http://dx.doi.org/10.1093/biomet/asr073. [p382]
4. D. Liu and H. Zhang. Residuals and diagnostics for ordinal regression models: A surrogate approach. Journal of the American Statistical Association, 0(ja):0–0, 2017. URL http://dx.doi.org/10.1080/01621459.2017.1292915. [p382, 383, 389, 392]