report

Jingning Yang, Yanwen Liu, He Guo

2/21/2020

Introduction

Our client, Reshma Satish, a PhD student at Boston University's Department of Forensic Anthropology,administered a project involves fracturing 135 long deer bones using a standardized fracturing apparatus over a period of 540 days. The bones are broken into 9 groups with 15 bones in each group and each group is fractured at different times(0 days,30 days,60 days,90 days,180 days,270 days, etc.) Her goal was to investigate how time of exposure of long bones affects variables such as the type of fracture angle, the type of fracture surface morphology,etc.that is created on the bones. Currently, our client only has data until day 270 since this experiment still keep going.

This report will start with a description of the EDA, as well as Data analysis/Data structure. Then we will describe the methods we will use on modeling part, followed by our analysis.

Data Cleaning and EDA

Data structure

We are provided the data in an excel document with 7 groups of drying days and a total of 105 bones. There are 10 variables in the dataset, but we will only focus on 8 of them, with Bisect Fracture and Fracture-like as binary responses, Number of fragments as a numerical response, and Fracture Angle, Fracture Type, Fracture Surface, Fracture Outline, and Fractographic Features as unordered categorical responses.

Data Cleaning

```
## [1] "integer"
```

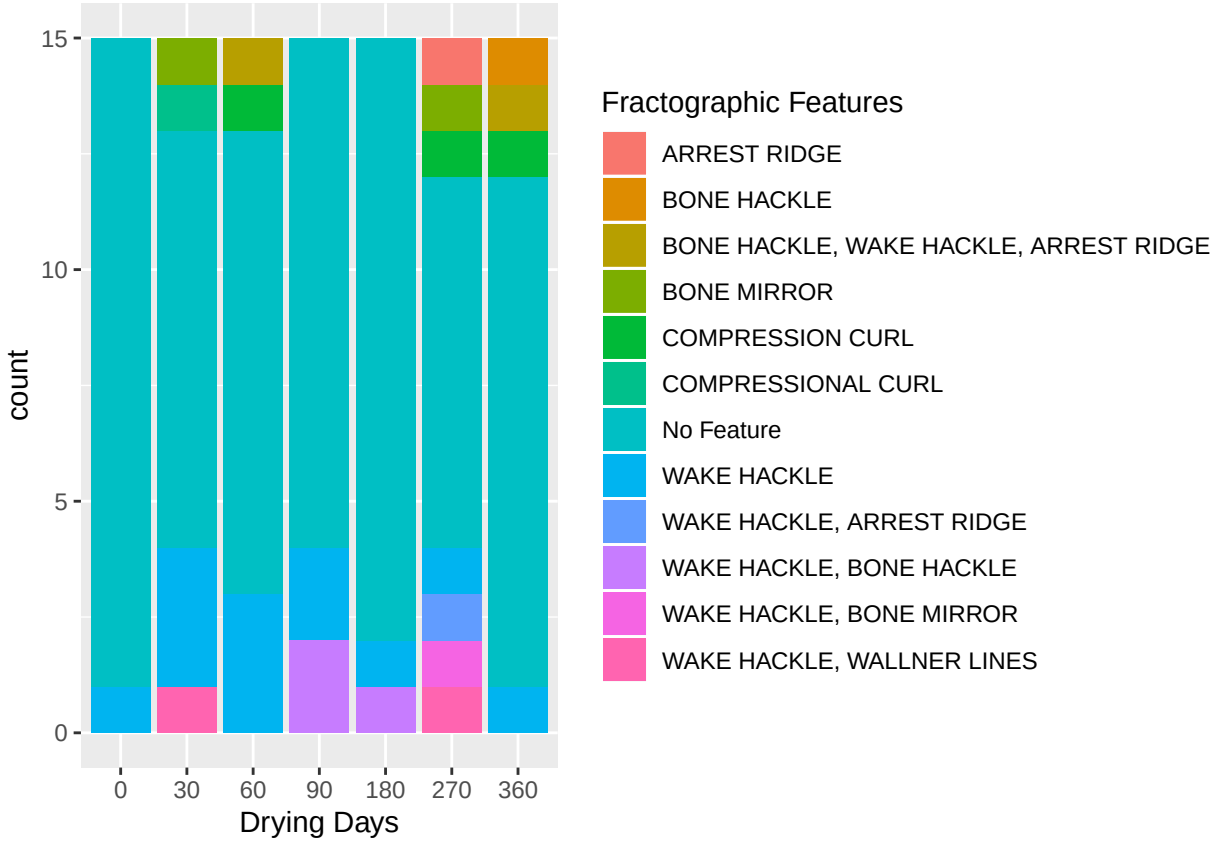
At a first glance of the raw dataset, there are many NA inputs. After discussed with our client, we decide to change NA in Fracture Angle, Fracture Type, Fracture Surface, and Fracture Outline to non applicable and change NA in Fractographic Features to No Feature. Moreover, we create numerical variable drying days from variable Bone Specimen for future analysis.

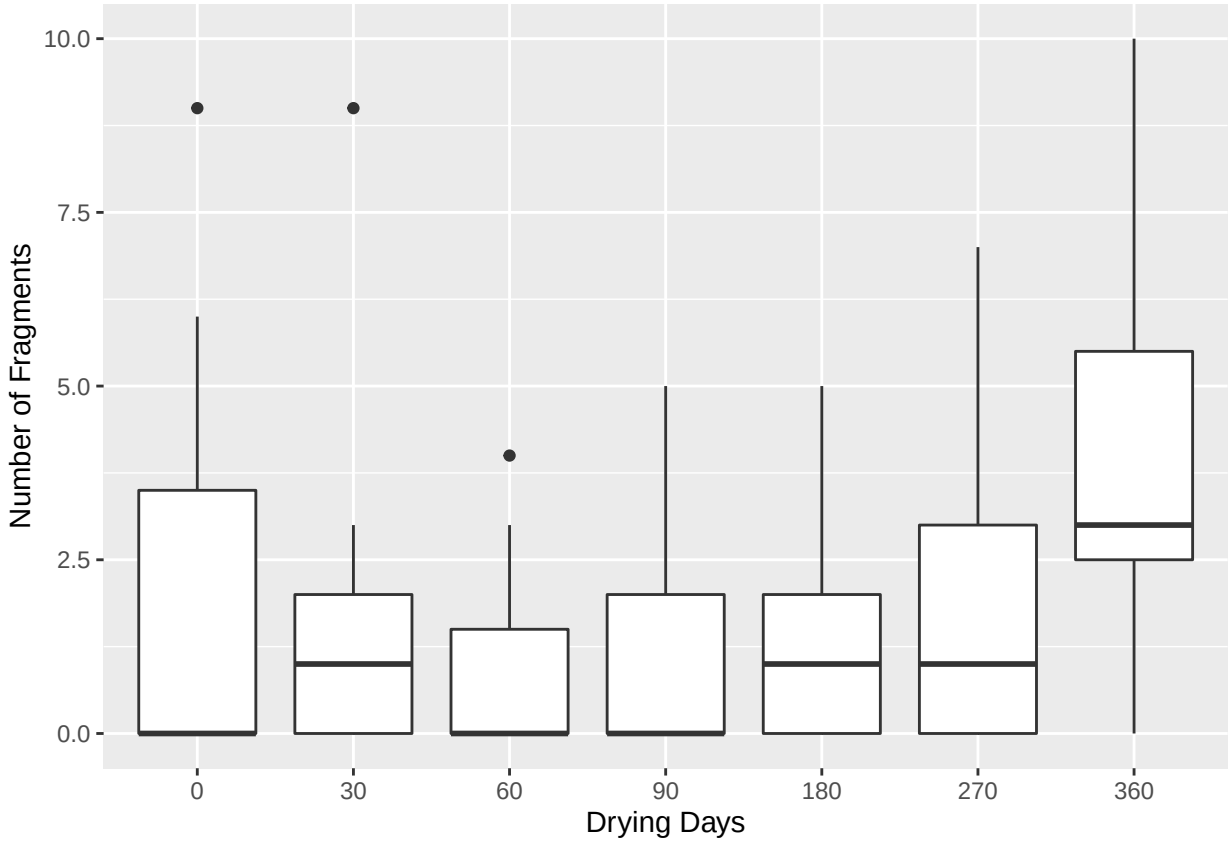
EDA

We conduct a basic exploratory data analysis (EDA) to visualize the data and find some patterns between variables. We focus on the two binary responses, Number of fragments, and the Fractographic Features as well.

Table 1: Bisecting Fracture and Fracture-like percentage for different drying-day groups

	0	30	60	90	180	270	360
Bisecting percentage	0.47	0.6	0.67	0.47	0.67	0.60	0.87
Fracture-like percentage	0.47	0.8	0.73	0.53	0.73	0.93	0.93


count
15
10
5
0
0
30
60
90
180
270
360
Drying Days
Fractographic Features
ARREST RIDGE
BONE HACKLE
BONE HACKLE, WAKE HACKLE, ARREST RIDGE
BONE MIRROR
COMPRESSION CURL
COMPRESSIONAL CURL
No Feature
WAKE HACKLE
WAKE HACKLE, ARREST RIDGE
WAKE HACKLE, BONE HACKLE
WAKE HACKLE, BONE MIRROR
WAKE HACKLE, WALLNER LINES


Number of Fragments
10.0
7.5
5.0
2.5
0.0
0
30
60
90
180
270
360
Drying Days

Modeling & Interpretation (YW-poisson) (HG-fracture type & surface)

We used logistic regression, poisson regression, and multinomial logit model based on different type of predictors in our dataset.

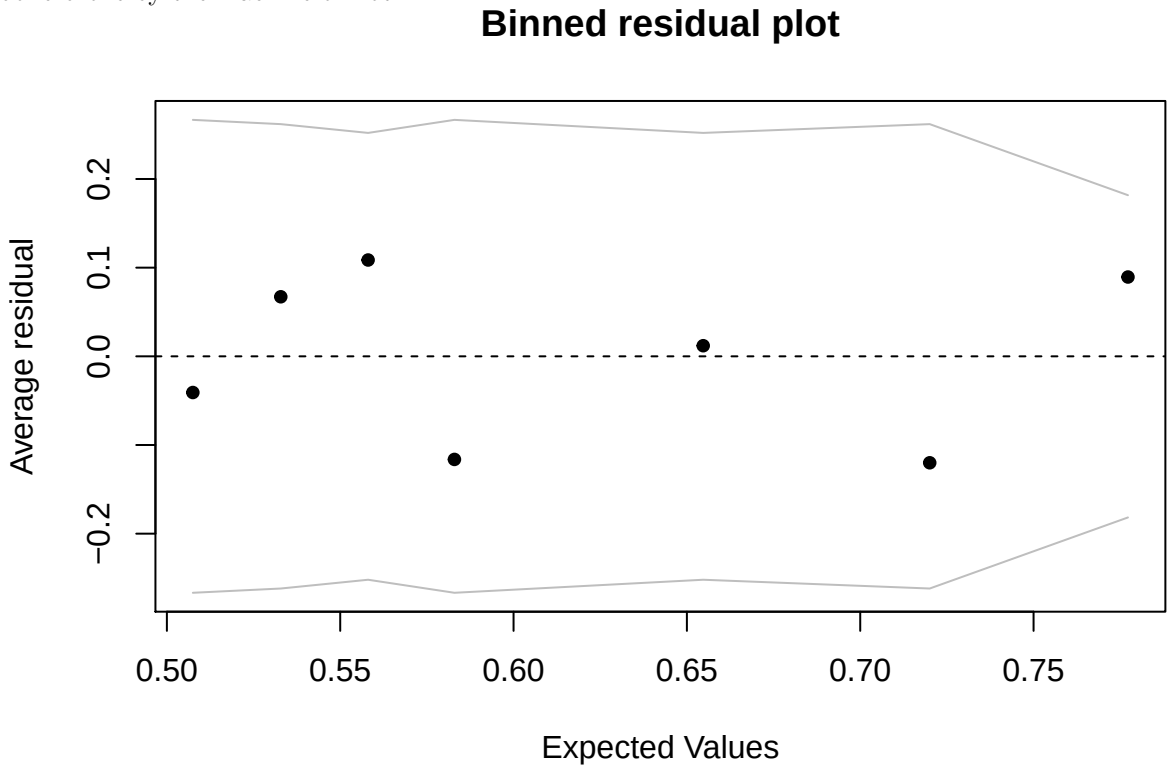
Bisect Fracture part

Since we assume response variable "bisect fracture" is a binary outcomes(0 or 1) and predictor "days" is continuous, we are using logistic regression to figure out relationship between the probability of "bisect fracture" and value of "days".
In addition, when "bisect fracture" is 0, it means the bone did not break by the standard fracturing machine, otherwise, it means the bone broke by the machine.

```
## (Intercept)   bone$days
##      0.0300      0.0034
```

```
## (Intercept)   bone$days
##       0.921       0.052
```

From the output, the estimated intercept coefficent 0.03 correspond to the log odds of broke the bone with average days, in other words, on average days, expected probability of broke the bone is expit(0.03)=0.4925=49%.
And one-unit increase in the predictor "days" increases the log odds of broke the bone(versus not broke the bone) by 0.0034.
Further, p-value of days is 0.052 indicates that it is somewhat not significant in determinging whether the bone broke by the machine or not.



Since the discrete residual plot does not tell us much about the model fit, therefore, we are using Binned residual plot for this model, due to every points fall within the bounds, our model fit well in this case.

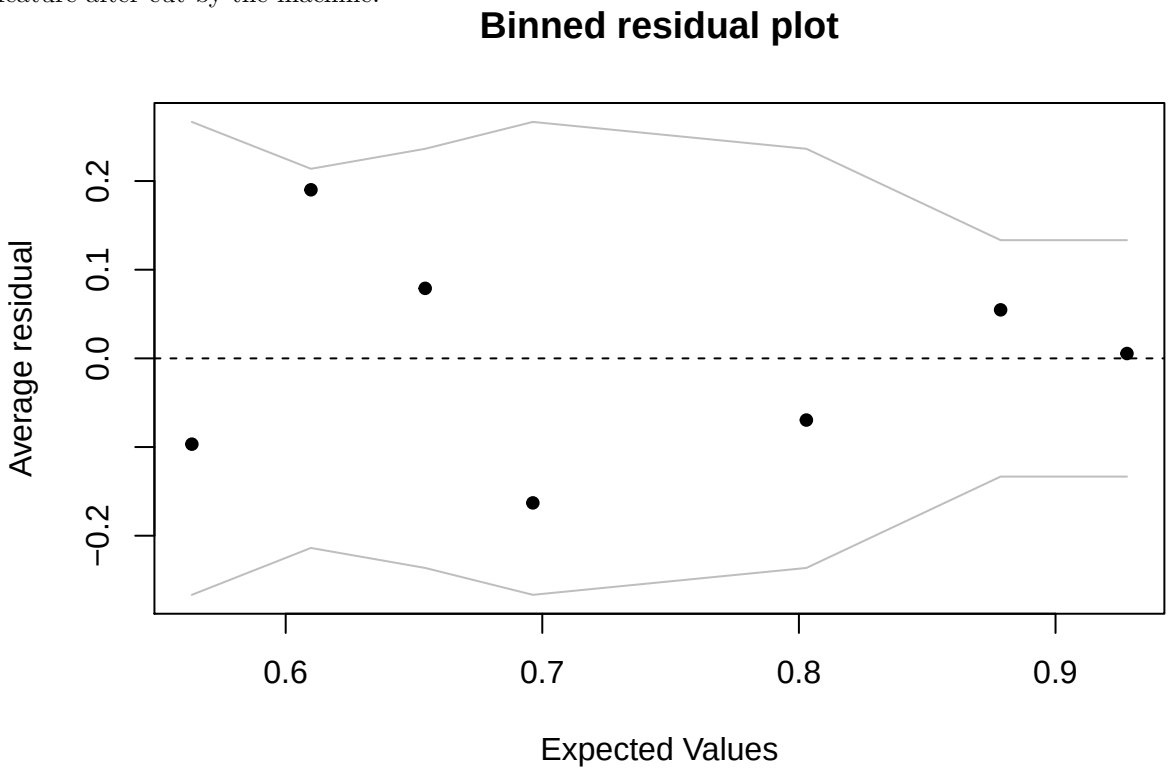
Fracture like part

We assume "Fracture-like" is a binary outcomes(0 or 1) and predictor is "days" as well, therefore, as idea with above part, we used logistic regression to figure out their relationship.
In addition, when "Fracture-like" is 0, it means the bone has no surface feature after cut by the standard fracturing machine, otherwise, it means fracturing machine made a mark on the bone.

```
## (Intercept)   bone$days
##      0.2551      0.0064
```

```
## (Intercept)   bone$days
##      0.4257      0.0055
```

From the output, the estimated intercept coefficent 0.2551 correspond to the log odds of made a mark on bone with average days, in other words, on average days, expected probability of made a mark on bone is expit(0.2551)=0.436=43.6%.
And one-unit increase in the predictor "days" increases the log odds of made a mark on bone(versus no surface feature on bone) by 0.0064.
Also, p-value of days is 0.0055 indicates that it is significant in determinging whether the bone has any surface feature after cut by the machine.



Same logic with above part, our model fit well in this case.

Fracture Angle part

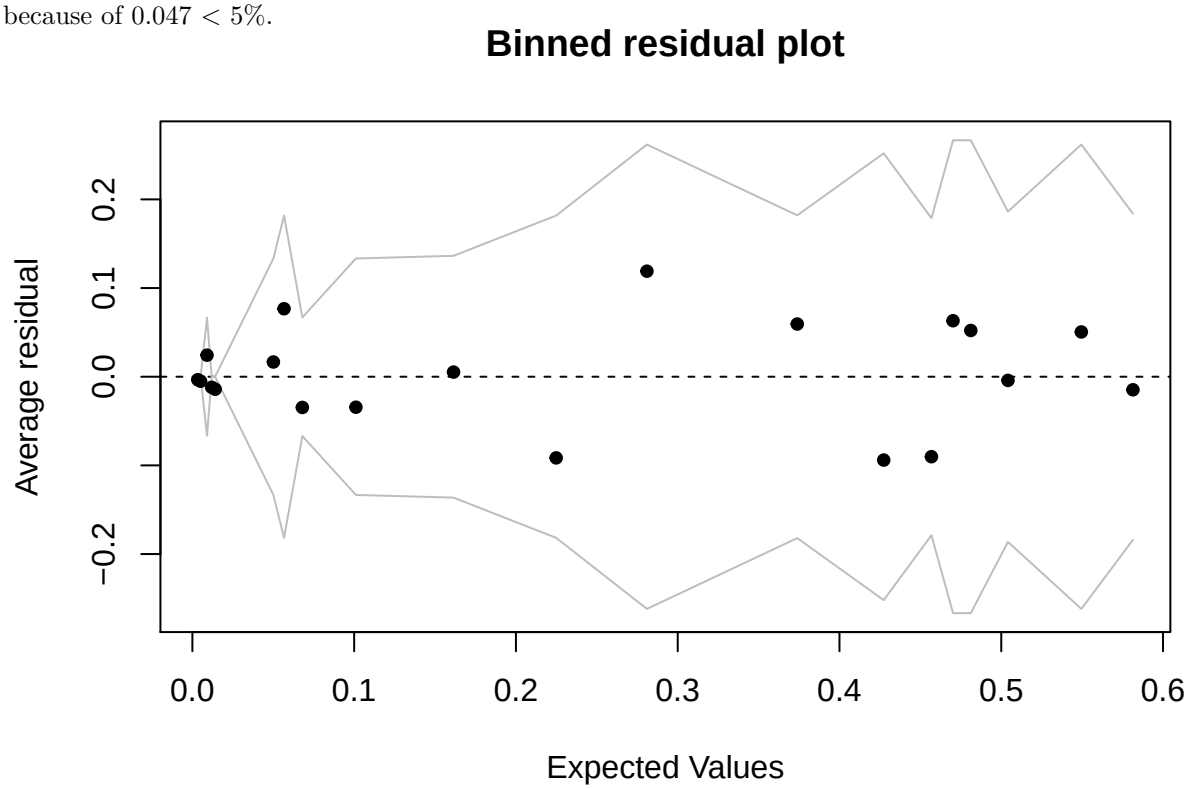
We assume "Fracture Angle" as no natural-order variable(nominal), since we cannot compare the order of angle of fresh bone and dry bone, and also N/A(non applicable) in this variable, so we used multinomial logit model in this case.

```
##                (Intercept) bone$days
## non applicable         2.2   -0.0057
## OBLIQUE                2.2   -0.0030
## RIGHT ANGLES          -1.2   -0.0077
```

From the output, the baseline is Fracture Angle = Intermediate.
A one-unit increase in the variable "days" is associated with the decrease in the log odds of being in Intermediate fracture angle vs. non applicable fracture angle in the amount of 0.0057.
A one-unit increase in the variable "days" is associated with the decrease in the log odds of being in Intermediate fracture angle vs. oblique fracture angle in the amount of 0.003.
In other words, if one-unit increase in "days", your chances of staying in the intermediate fracture angle category are higher compared to staying in non applicable and oblique.
Since summary of multinomial logit model does not provide p-values, we can get significance of the coefficients by calculating z score and input it into the formula:

```
##                (Intercept) bone$days
## non applicable     0.00037     0.047
## OBLIQUE            0.00034     0.266
## RIGHT ANGLES       0.41049     0.458
```

From output, it represents p-value of coeffients. Only the coefficient of non applicable is somewhat significant because of $0.047 < 5\%$.



Looks like our model fit kind of well in this case, except there might not good enough when expected values really small.

Number of Fragments

Considering "Number of Fragments" as a numerical variable, we use the poisson regression at first.

```
## (Intercept)   bone$days
##      0.1964      0.0026
```

```
## [1] 1.5e-17
```

Here our null hypothesis is that our model is correctly specified. The low p-value shows that we can reject this hypothesis, thus the possion regression model fits poorly. Next, we consider to use a negative binomial model.

```
##
## Call:
## glm.nb(formula = bone$`Number of Fragments` ~ bone$days, init.theta = 0.833740805,
##     link = log)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.579  -1.301  -0.433   0.350   2.204
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.24931    0.20348    1.23    0.220
## bone$days    0.00225    0.00103    2.18    0.029 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.83) family taken to be 1)
##
##     Null deviance: 117.06  on 104  degrees of freedom
## Residual deviance: 111.39  on 103  degrees of freedom
## AIC: 386.7
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.834
##           Std. Err.:  0.212
##
##  2 x log-likelihood:  -380.656
```

```
## [1] 0.27
```

Using the negative binomial model, we have a better fit. The only variable "days" has a coefficient of 0.002247, which is statistically significant. This means that for each one-unit increase in "days", the expected log count of the number of fragments increases by 0.002247.

Fracture Outline

We assume "Fracture Outline" as no natural-order(nominal) variable, same reason as "Fracture Angle" part, thus we used multinomial logit model.

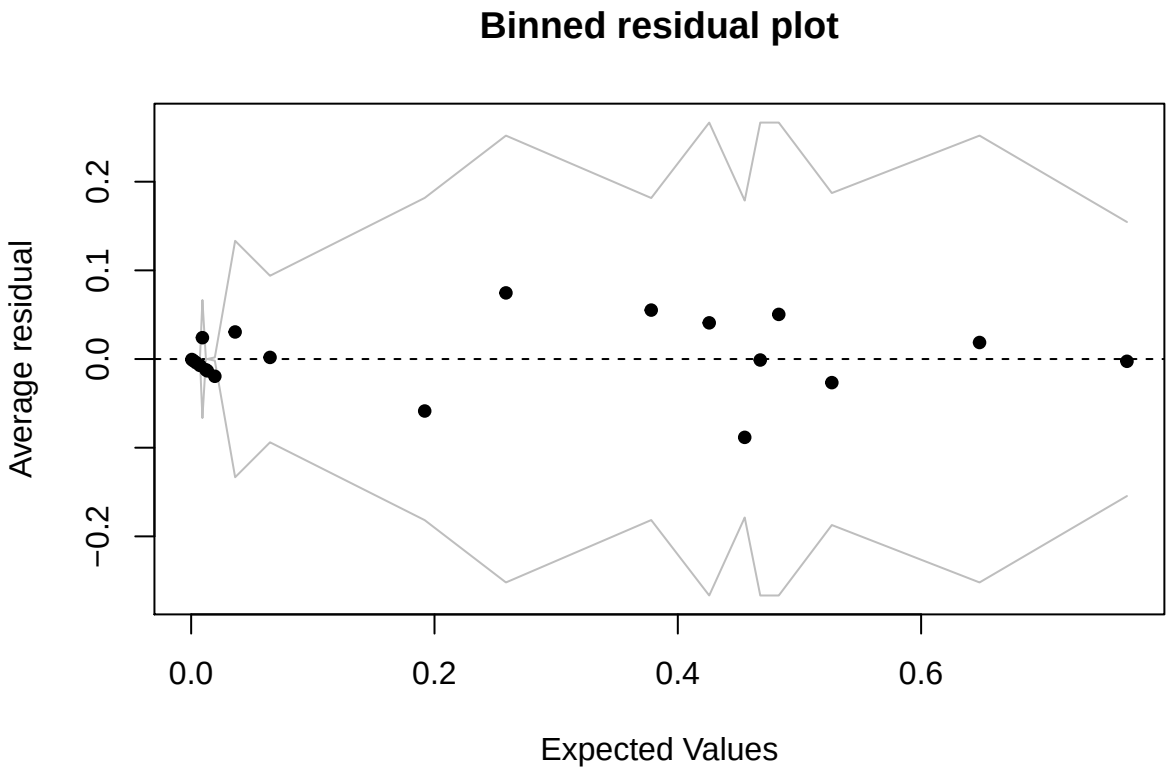
```
##                (Intercept) bone$days
## INTERMEDIATE         -1.71   -0.0155
## non applicable        0.13   -0.0043
## TRANSVERSE           -3.38   -0.0059
```

From the output, the baseline is Fracture Angle = Cruved/Jagged.
A one-unit increase in the variable "days" is associated with the decrease in the log odds of being in curved fracture outline vs. intermediate fracture outline in the amount of 0.0155.
A one-unit increase in the variable "days" is associated with the decrease in the log odds of being in curved fracture outline vs. non applicable fracture outline in the amount of 0.0043.
In other words, if one-unit increase in "days", your chances of staying in the curved/jagged fracture outline category are higher compared to staying in non applicable, intermediate and transverse.
Get significance of the coefficients by calculating z score and input it into the formula:

```
##                (Intercept) bone$days
## INTERMEDIATE         0.027     0.184
```

```
## non applicable       0.686     0.017
## TRANSVERSE           0.014     0.565
```

From output, it represents p-value of coeffients. Only the coefficient of non applicable is significant because of $0.017 < 5\%$.



Looks like our model fit kind of well in this case, except there might not good enough when expected values really small.

Fractographic Features

Since "Fractographic Features" is an no natural-order(nominal) variable, we used multinomial logit model as before.

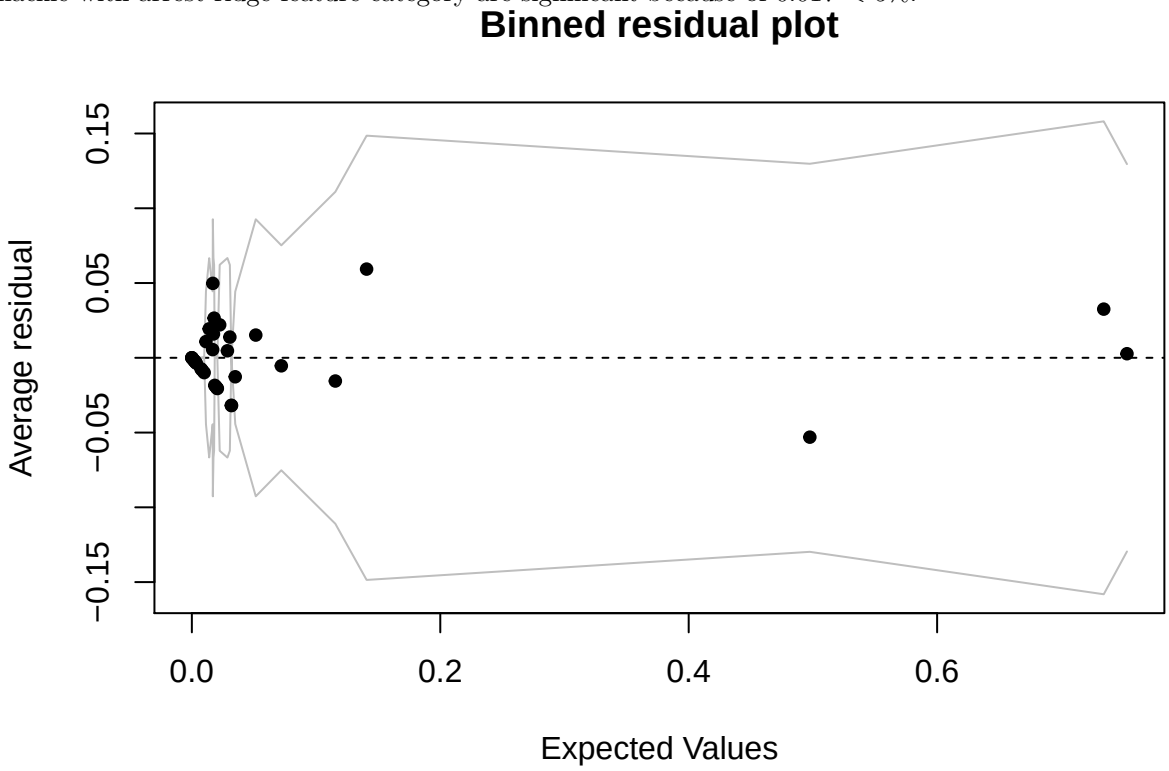
```
##                                     (Intercept) bone$days
## BONE HACKLE                           -4.5e+01   1.3e-01
## BONE HACKLE, WAKE HACKLE, ARREST RIDGE  1.7e+00  -4.2e-03
## BONE MIRROR                           2.4e+00  -7.9e-03
## COMPRESSION CURL                      1.8e+00  -2.9e-03
## COMPRESSIONAL CURL                    3.0e+00  -2.7e-02
## No Feature                            6.1e+00  -8.8e-03
## WAKE HACKLE                           4.6e+00  -1.1e-02
## WAKE HACKLE, ARREST RIDGE            -1.3e-04   4.5e-07
## WAKE HACKLE, BONE HACKLE              3.1e+00  -1.0e-02
## WAKE HACKLE, BONE MIRROR              1.9e-04  -8.1e-07
## WAKE HACKLE, WALLNER LINES            2.4e+00  -7.9e-03
```

From the output, the baseline is Fractographic Features = Arrest Ridge.
A one-unit increase in the variable "days" is associated with the increase in the log odds of being in arrest ridge features vs. ibone hackle feature in the amount of 0.13.
A one-unit increase in the variable "days" is associated with the decrease in the log odds of being in arrest ridge features vs. bone hackle, wake hackgle, arrest ridge feature in the amount of 0.0042.

Same logic to interpret following coefficients.
In other words, if one-unit increase in "days", your chances of staying in the arrest ridge features category are higher compared to staying in bone hackle, wake hackle, arrest ridge feature category; bone mirror feature; compression curl feature category and so on, except bone hackle feature and wake hackle with arrest ridge feature.
Get significance of the coefficients by calculating z score and input it into the formula:

```
##                                       (Intercept) bone$days
## BONE HACKLE                                 0.000      0.00
## BONE HACKLE, WAKE HACKLE, ARREST RIDGE      0.570      0.70
## BONE MIRROR                                 0.414      0.47
## COMPRESSION CURL                            0.533      0.78
## COMPRESSIONAL CURL                          0.316      0.32
## No Feature                                  0.022      0.34
## WAKE HACKLE                                 0.091      0.24
## WAKE HACKLE, ARREST RIDGE                   1.000      1.00
## WAKE HACKLE, BONE HACKLE                    0.276      0.34
## WAKE HACKLE, BONE MIRROR                    1.000      1.00
## WAKE HACKLE, WALLNER LINES                  0.414      0.47
```

From output, it represents p-value of coeffients. Only coefficients of "days" for bone hackle feature and wake hackle with arrest ridge feature category are significant because of $0.017 < 5\%$.



From the plot, most of points fall left handside, and it may hard to see whether they fall inside the bounds or not. Therefore, the model might not good enough.

Conclusion