

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.

```
## Create Dummy Variable
```

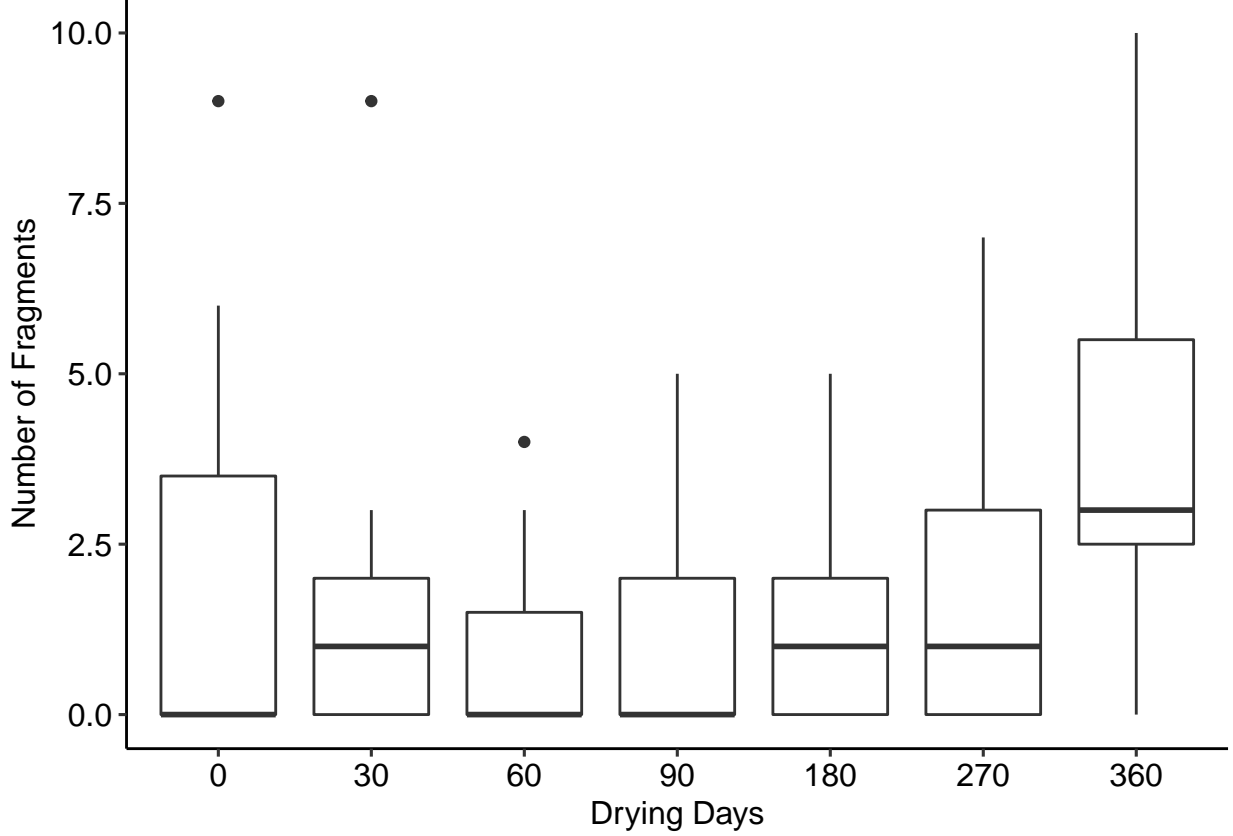
There are some observation have multiple feature in Fractographic Features variable. Then, we create 6 dummy variable to indicate which feature appaer for a given observation.

EDA (need revise, add explanation & change Fractographic features into 6 dummy varaibles.)

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



In the table, we give a bisecting percentage and fracture-like percentage for each groups of bones in different drying days. We can see a pattern that with the drying days increasing, it's more likely to have a bisecting or fracture-like happened during the standardized fracturing apparatus.

In the boxplot, we can see several outliers in the first three groups and large amount of zero value. It's hard to figure a clear pattern in the first four groups, but later on, then number of fragments are increasing as the drying days increasing.

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

We used logistic regression, poisson regression, and multinomial logit model based on different types of response variables in our dataset.

Fracture Type

###COMMINUTED relative to BUTTERFULY:

Intercept: This is the multinomial logit estimate for COMMINUTED relative to BUTTERFULY when

the predictor variables in the model are evaluated at zero. For day equals to zero, the logit for preferring COMMINUTED relative to BUTTERFULY is 0.8269632. Note that evaluating days equals to zero indicated that the day of starting the experiment Days: This is the multinomial logit estimate for a one unit increase in Days for COMMINUTED relative to BUTTERFULY. If a subject were to increase Days by one point, the multinomial log-odds of preferring COMMINUTED relative to BUTTERFULY would be expected to Increase by 10.00216.

T-statistics for days in $\text{Log}(\text{Prob}(\text{COMMINUTED})/\text{Prob}(\text{BUTTERFULY}))$ equals to 0.8158939, which is lower than 2. Then, the Days does not make any difference for log-odds of preferring COMMINUTED relative to BUTTERFULY

###non applicable relative to BUTTERFULY

Intercept: This is the multinomial logit estimate for non applicable relative to BUTTERFULY when the predictor variables in the model are evaluated at zero. For day equals to zero, the logit for preferring non applicable relative to BUTTERFULY is 1.5166589. Note that evaluating days equals to zero indicated that the day of starting the experiment Days: This is the multinomial logit estimate for a one unit increase in Days for non applicable relative to BUTTERFULY. If a subject were to increase Days by one point, the multinomial log-odds of preferring non applicable relative to BUTTERFULY would be expected to decrease by 0.0024548581.

T-statistics for days in $\text{Log}(\text{Prob}(\text{non applicable})/\text{Prob}(\text{BUTTERFULY}))$ equals to $|-0.0024548581/0.002647403| = 0.8948441$, which is lower than 2. Then, the Days does not make any difference for log-odds of preferring non applicable relative to BUTTERFULY

$\text{Log}(\text{Prob}(\text{OBLIQUE})/\text{Prob}(\text{BUTTERFULY})) = -0.1913727 - 0.0027752170 * \text{Days}$

OBLIQUE relative to BUTTERFULY:

Intercept: This is the multinomial logit estimate for OBLIQUE relative to BUTTERFULY when the predictor variables in the model are evaluated at zero. For day equals to zero, the logit for preferring OBLIQUE relative to BUTTERFULY is -0.1913727. Note that evaluating days equals to zero indicated that the day of starting the experiment Days: This is the multinomial logit estimate for a one unit increase in Days for OBLIQUE relative to BUTTERFULY. If a subject were to increase Days by one point, the multinomial log-odds of preferring OBLIQUE relative to BUTTERFULY would be expected to decrease by 0.0027752170.

T-statistics for days in $\text{Log}(\text{Prob}(\text{OBLIQUE})/\text{Prob}(\text{BUTTERFULY}))$ equals to $|-0.0027752170 / 0.004192975| = 0.661873$, which is lower than 2. Then, the Days does not make any difference for log-odds of preferring OBLIQUE relative to BUTTERFULY

$\text{Log}(\text{Prob}(\text{SEGMENTAL})/\text{Prob}(\text{BUTTERFULY})) = -0.5994698 - 0.0006693986 * \text{Days}$

###SEGMENTAL relative to BUTTERFULY:

Intercept: This is the multinomial logit estimate for SEGMENTAL relative to BUTTERFULY when the predictor variables in the model are evaluated at zero. For day equals to zero, the logit for preferring SEGMENTAL relative to BUTTERFULY is -0.5994698. Note that evaluating days equals to zero indicated that the day of starting the experiment Days: This is the multinomial logit estimate for a one unit increase in Days for SEGMENTAL relative to BUTTERFULY. If a subject were to increase Days by one point, the multinomial log-odds of preferring SEGMENTAL relative to BUTTERFULY would be expected to decrease by -0.0006693986.

T-statistics for days in $\text{Log}(\text{Prob}(\text{SEGMENTAL})/\text{Prob}(\text{BUTTERFULY}))$ equals to $|-0.0006693986/0.004110398| = 0.1628549$, which is lower than 2. Then, the Days does not make any difference for log-odds of preferring SEGMENTAL relative to BUTTERFULY.

$\text{Log}(\text{Prob}(\text{TRANSVERSE})/\text{Prob}(\text{BUTTERFULY})) = -1.9804151 - 0.004371236 * \text{Days}$

###TRANSVERSE relative to BUTTERFULY:

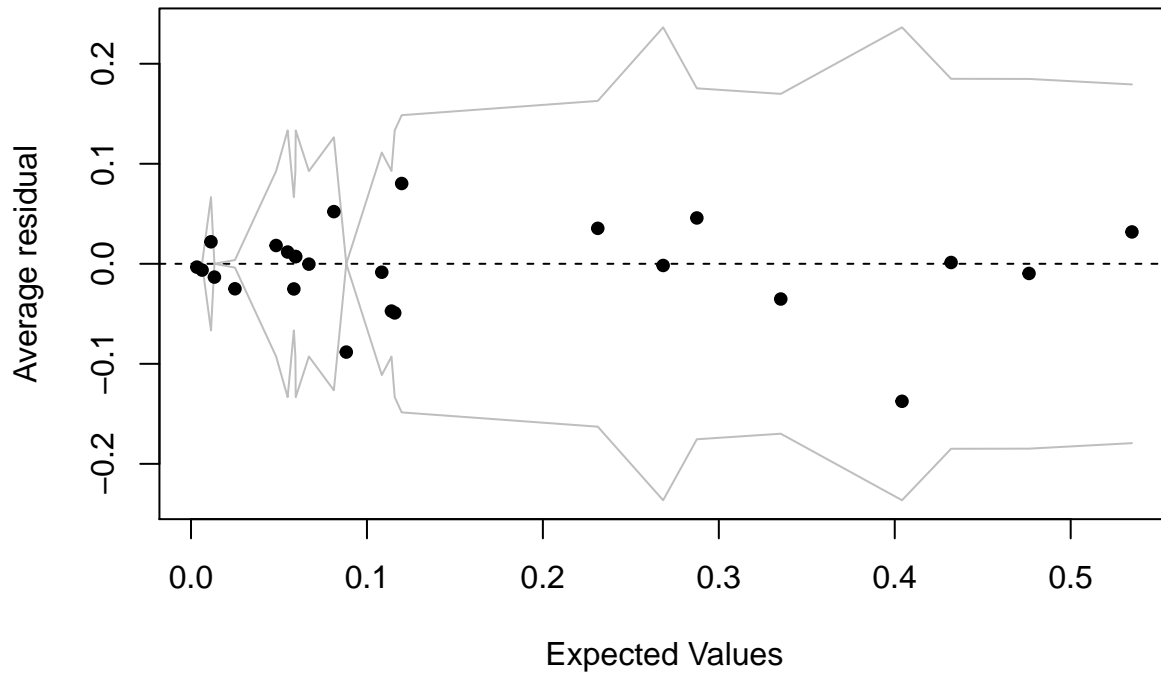
Intercept : This is the multinomial logit estimate for TRANSVERSE relative to BUTTERFULY when the predictor variables in the model are evaluated at zero. For day equals to zero, the logit for preferring TRANSVERSE relative to BUTTERFULY is -1.9804151. Note that evaluating days equals to zero indicated that the day of starting the experiment Days: This is the multinomial logit estimate for a one unit increase in Days for TRANSVERSE relative to BUTTERFULY. If a subject were to increase Days by one point, the multinomial log-odds of preferring TRANSVERSE relative to BUTTERFULY would be expected to decrease by 0.004371236.

T-statistics for days in $\text{Log}(\text{Prob}(\text{TRANSVERSE})/\text{Prob}(\text{BUTTERFULY}))$ equals to $|-0.0043712363/0.010436283| = 0.4188499$, which is lower than 2. Then, the Days does not make any difference for log-odds of preferring TRANSVERSE relative to BUTTERFULY

According to the Binned residual plot, there is not clearly pattern.

```
## Call:
## multinom(formula = bone$`Fracture Type` ~ bone$days, trace = FALSE)
##
## Coefficients:
##              (Intercept) bone$days
## COMMINUTED           0.83   0.00217
## non applicable        1.52  -0.00245
## OBLIQUE              -0.19  -0.00278
## SEGMENTAL            -0.60  -0.00067
## TRANSVERSE          -1.98  -0.00437
##
## Std. Errors:
##              (Intercept) bone$days
## COMMINUTED           0.52   0.0026
## non applicable        0.50   0.0027
## OBLIQUE              0.69   0.0042
## SEGMENTAL            0.76   0.0041
## TRANSVERSE           1.43   0.0104
##
## Residual Deviance: 281
## AIC: 301
```

Binned residual plot



##	(Intercept)	bone\$days
## COMMINUTED	1.58	0.82
## non applicable	3.04	-0.89
## OBLIQUE	-0.28	-0.66
## SEGMENTAL	-0.79	-0.16
## TRANSVERSE	-1.39	-0.42

Fracture Surface

###non applicable relative to INTERMEDIATE:

$\text{Log}(\text{Prob}(\text{non applicable})/\text{Prob}(\text{INTERMEDIATE})) = 2.185304 - 0.008706689 \cdot \text{Days}$

Intercept: This is the multinomial logit estimate for non applicable relative to INTERMEDIATE when the predictor variables in the model are evaluated at zero. For day equals to zero, the logit for preferring non applicable relative to INTERMEDIATE is 2.185304. Note that evaluating days equals to zero indicated that the day of starting the experiment Days: This is the multinomial logit estimate for a one unit increase in Days for non applicable relative to INTERMEDIATE. If a subject were to increase Days by one point, the multinomial log-odds of preferring non applicable relative to INTERMEDIATE would be expected to decrease by 0.008706689.

T-statistics for days in $\text{Log}(\text{Prob}(\text{non applicable})/\text{Prob}(\text{INTERMEDIATE}))$ equals to $|0.008706689/0.002536340| = 3.432777$, which is greater than 2. Then, the Days make a significant difference for log-odds of preferring non applicable relative to INTERMEDIATE

###ROUGH relative to INTERMEDIATE:

$\text{Log}(\text{Prob}(\text{ROUGH})/\text{Prob}(\text{INTERMEDIATE})) = -23.616556 + 0.060230466 \cdot \text{Days}$

Intercept: This is the multinomial logit estimate for ROUGH relative to INTERMEDIATE when the predictor variables in the model are evaluated at zero. For day equals to zero, the logit for preferring ROUGH relative to INTERMEDIATE is -23.616556. Note that evaluating days equals to zero indicated that the day of starting the experiment Days: This is the multinomial logit estimate for a one unit increase in Days for ROUGH relative to INTERMEDIATE. If a subject were to increase Days by one point, the multinomial log-odds of preferring ROUGH relative to INTERMEDIATE would be expected to increase by 0.060230466.

T-statistics for days in $\text{Log}(\text{Prob}(\text{ROUGH})/\text{Prob}(\text{INTERMEDIATE}))$ equals to $|0.060230466/0.002940204| = 20.48513$, which is greater than 2. Then, the Days make a significant difference for log-odds of preferring ROUGH relative to INTERMEDIATE

###SMOOTH relative to INTERMEDIATE:

$\text{Log}(\text{Prob}(\text{SMOOTH})/\text{Prob}(\text{INTERMEDIATE})) = 2.255591 - 0.007492372 * \text{Days}$

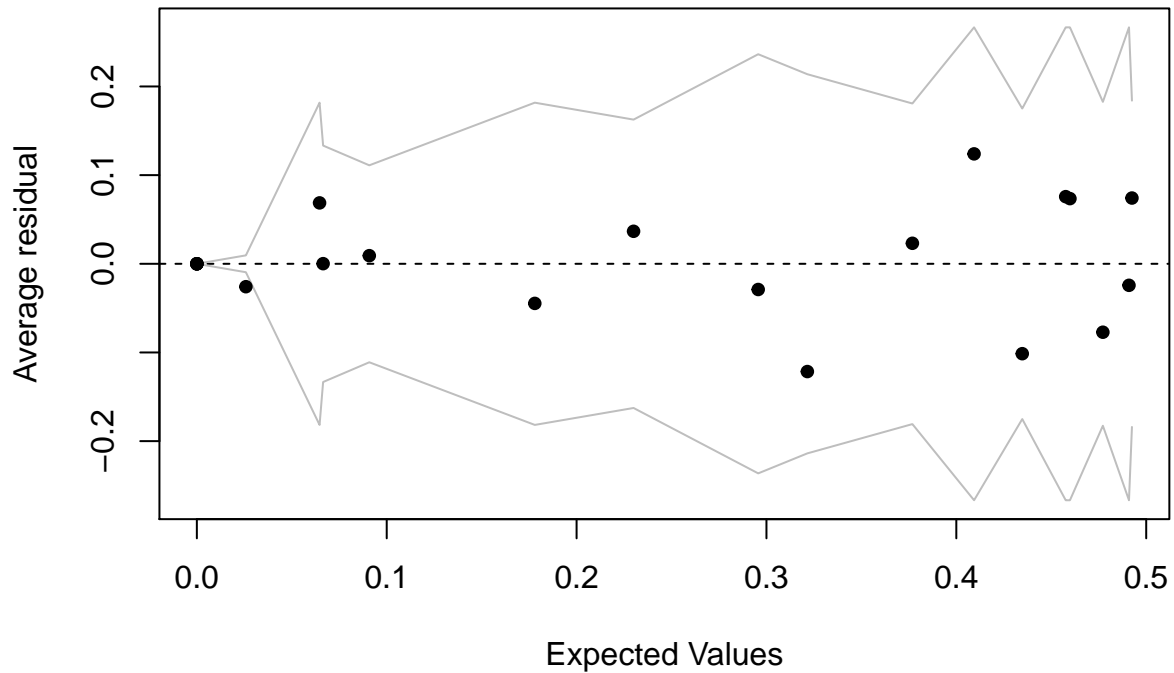
Intercept: This is the multinomial logit estimate for SMOOTH relative to INTERMEDIATE when the predictor variables in the model are evaluated at zero. For day equals to zero, the logit for preferring SMOOTH relative to INTERMEDIATE is 2.255591. Note that evaluating days equals to zero indicated that the day of starting the experiment Days: This is the multinomial logit estimate for a one unit increase in Days for SMOOTH relative to INTERMEDIATE. If a subject were to increase Days by one point, the multinomial log-odds of preferring SMOOTH relative to INTERMEDIATE would be expected to decrease by 0.007492372.

T-statistics for days in $\text{Log}(\text{Prob}(\text{SMOOTH})/\text{Prob}(\text{INTERMEDIATE}))$ equals to $|-0.007492372/0.002378579| = 3.149936$, which is greater than 2. Then, the Days make a significant difference for log-odds of preferring SMOOTH relative to INTERMEDIATE

According to the Binned residual plot, there is not clearly pattern.

```
## Call:
## multinom(formula = bone$`Fracture Surface` ~ bone$days, trace = FALSE)
##
## Coefficients:
##              (Intercept) bone$days
## non applicable          2.2   -0.0087
## ROUGH                 -23.6    0.0602
## SMOOTH                 2.3   -0.0075
##
## Std. Errors:
##              (Intercept) bone$days
## non applicable    5.7e-01    0.0025
## ROUGH             3.6e-05    0.0029
## SMOOTH            5.6e-01    0.0024
##
## Residual Deviance: 208
## AIC: 220
```

Binned residual plot



```
##               (Intercept) bone$days
## non applicable    3.8e+00    -3.4
## ROUGH            -6.6e+05    20.5
## SMOOTH           4.0e+00    -3.1
```

Bisect Fracture

Since response variable “bisect fracture” is a binary outcomes(0 or 1), 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.

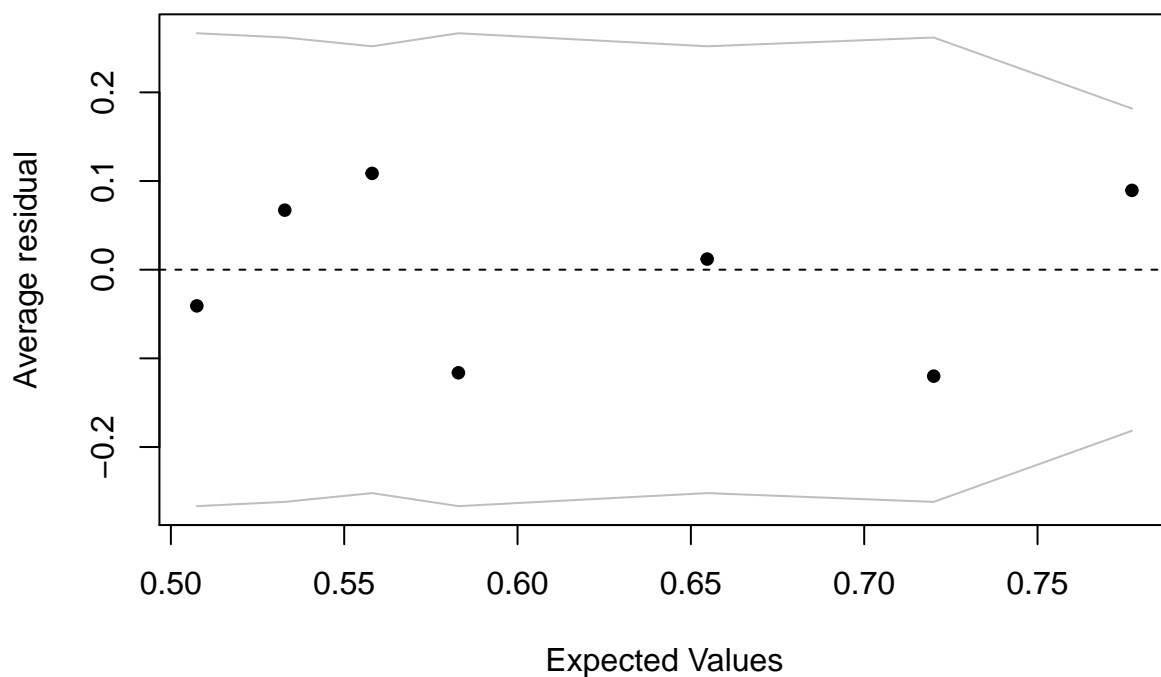
```
##
## Call:
## glm(formula = bone$`Bisecting Fracture` ~ bone$days, family = binomial())
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73   -1.23    0.71    1.04    1.16
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.02996   0.30127   0.10   0.921
## bone$days   0.00339   0.00174   1.94   0.052 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 139.55 on 104 degrees of freedom
## Residual deviance: 135.53 on 103 degrees of freedom
## AIC: 139.5
##
## Number of Fisher Scoring iterations: 4
```

From the output, the estimated intercept coefficient 0.029 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 $\text{expit}(0.029)=0.4927=49\%$. 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 there is no strong evident but moderate evidence in support of the differences of the bone broke by the machine or not affected by drying time.

Binned residual plot



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

Fracture like

Since “Fracture-like” is a binary outcomes(0 or 1) 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.

```
##
## Call:
## glm(formula = bone$`Fracture-like` ~ bone$days, family = binomial())
```



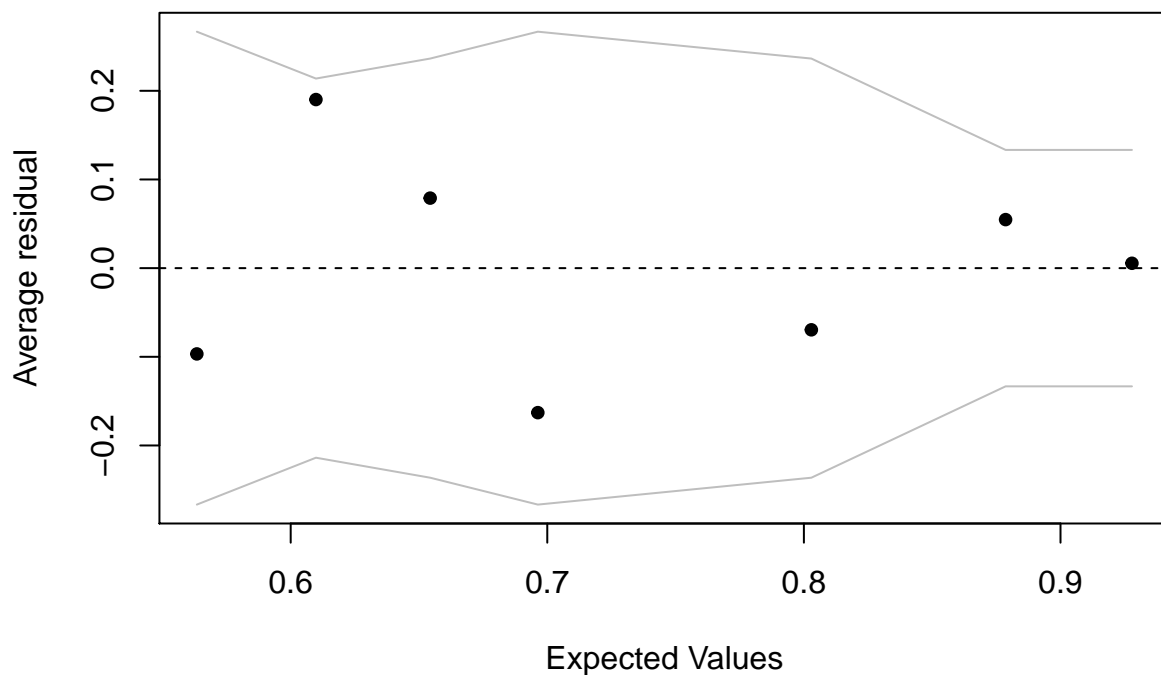
```
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.293  -1.288   0.509   0.921   1.071
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.25510    0.32022   0.80  0.4257
## bone$days    0.00639    0.00230   2.78  0.0055 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 121.78  on 104  degrees of freedom
## Residual deviance: 112.13  on 103  degrees of freedom
## AIC: 116.1
##
## Number of Fisher Scoring iterations: 4
```

From the output, the estimated intercept coefficient 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 $\text{expit}(0.2551) = 0.436 = 43.6\%$.

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 there is strong evidence in determining whether the bone has any surface feature after cut by the machine affected by drying time.

Binned residual plot



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

Fracture Angle

As “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, that the outcome variable in this model has more than 2 levels.

```
## Call:
## multinom(formula = bone$`Fracture Angle` ~ bone$days, trace = FALSE)
##
## Coefficients:
##              (Intercept) bone$days
## non applicable          2.2   -0.0057
## OBLIQUE                 2.2   -0.0030
## RIGHT ANGLES           -1.2   -0.0077
##
## Std. Errors:
##              (Intercept) bone$days
## non applicable          0.63    0.0029
## OBLIQUE                 0.62    0.0027
## RIGHT ANGLES           1.48    0.0104
##
## Residual Deviance: 200
## AIC: 212
```

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.

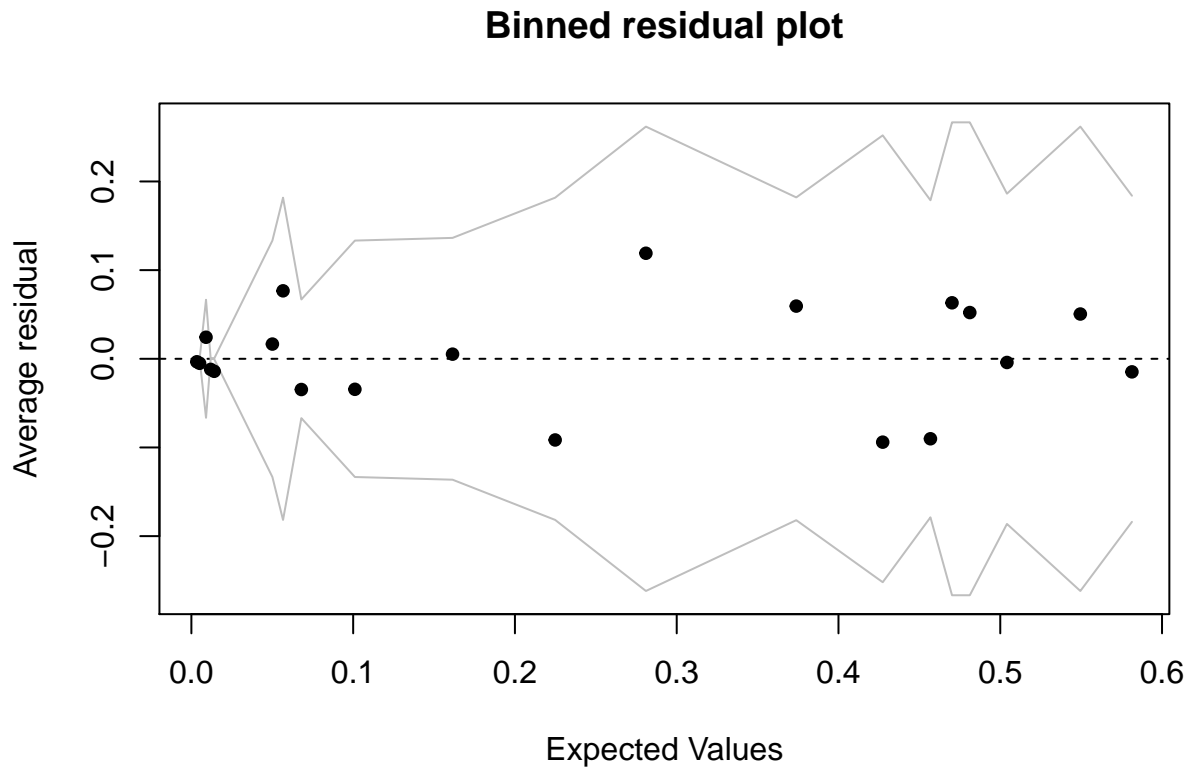
A one-unit increase in the variable “days” is associated with the decrease in the log odds of being in Intermediate fracture angle vs. oblique fracture angle in the amount of 0.003.

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: $z = (x_i - \bar{x})/s$ where coefficients of summary represents $x_i - \bar{x}$, and Std.Errors of summary represents s .

```
##              (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 coefficients. Only non applicable has strong evidence in determining whether affected by drying time because of its p-value is $0.047 < 5\%$.



It looks like our model fit kind of well in this case, but there might not good enough when expected values small.

Number of Fragments

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

```
fit.Number_of_Fragments1 <- glm(bone$`Number of Fragments` ~ bone$days, family="poisson")
summary(fit.Number_of_Fragments1) #estimated coefficients
```

```
##
## Call:
## glm(formula = bone$`Number of Fragments` ~ bone$days, family = "poisson")
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.480  -1.685  -0.741   0.649   4.522
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.196437   0.123824   1.59    0.11
## bone$days    0.002574   0.000549   4.69 2.8e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 296.20  on 104  degrees of freedom
```

```
## Residual deviance: 274.71 on 103 degrees of freedom
## AIC: 452
##
## Number of Fisher Scoring iterations: 6
```

```
pchisq(fit.Number_of_Fragments1$deviance,df=fit.Number_of_Fragments1$df.residual, lower.tail=FALSE)
```

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

```
regout <- glm(bone$`Number of Fragments`~bone$days, family=quasipoisson)
summary.glm(regout)$dispersion
```

```
## [1] 2.9
```

Here our null hypothesis is that our model is correctly specified. The low p-value shows that we can reject this hypothesis, thus the poisson regression model fits poorly. Next, we check the dispersion of the data. We can see the variance of random component is roughly 3 times the size of the mean. There may be a over-dispersion problem. So we consider to use a negative binomial model.

```
fit.Number_of_Fragments2 <- glm.nb(bone$`Number of Fragments` ~ bone$days)
summary(fit.Number_of_Fragments2)
```

```
##
## 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
```

```
pchisq(fit.Number_of_Fragments2$deviance,df=fit.Number_of_Fragments2$df.residual, lower.tail=FALSE)
```

```
## [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. It can be also interpreted as for each 100 unit increase in “days”, the expected count of the number of fragments will be increased by 1.25 times.

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.

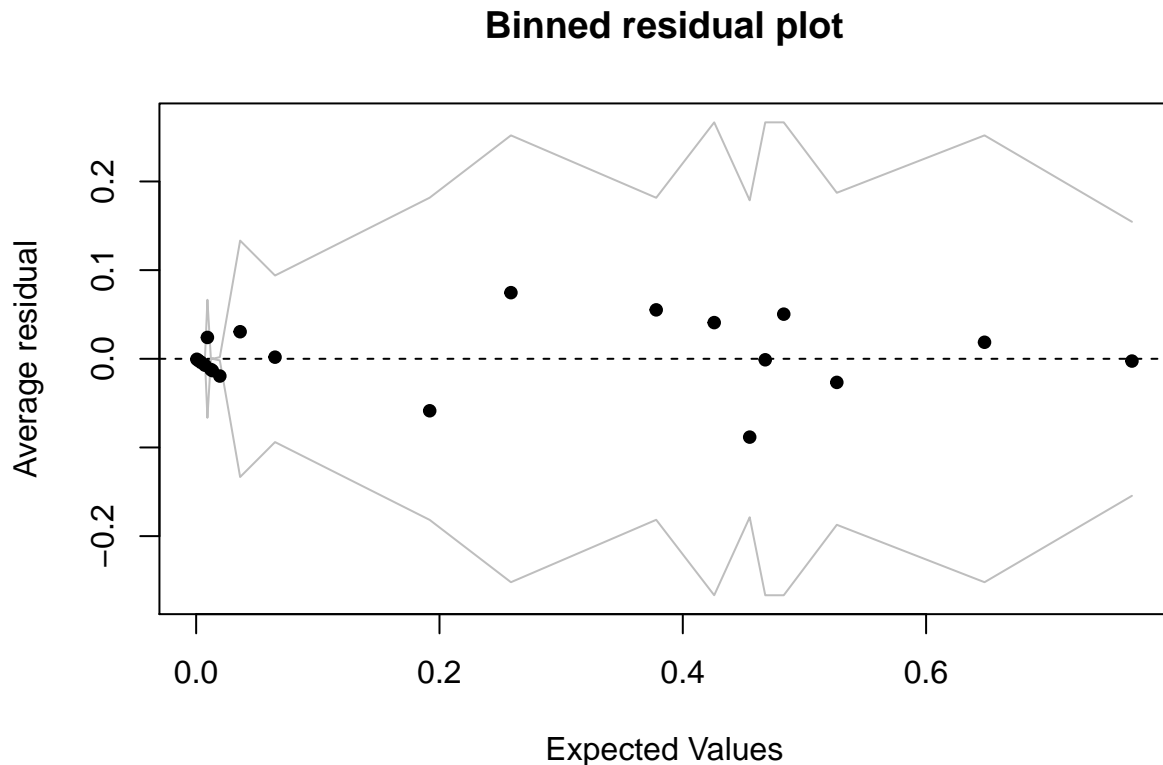
A one-unit increase in the variable “days” is associated with the decrease in the log odds of being in curved fracture outline vs. non applicable fracture outline in the amount of 0.0043.

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 coefficients. Only the non applicable has strong evidence to determining whether it affected by drying time because of its p value is $0.017 < 5\%$.



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

Fractographic Features(Revise)

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

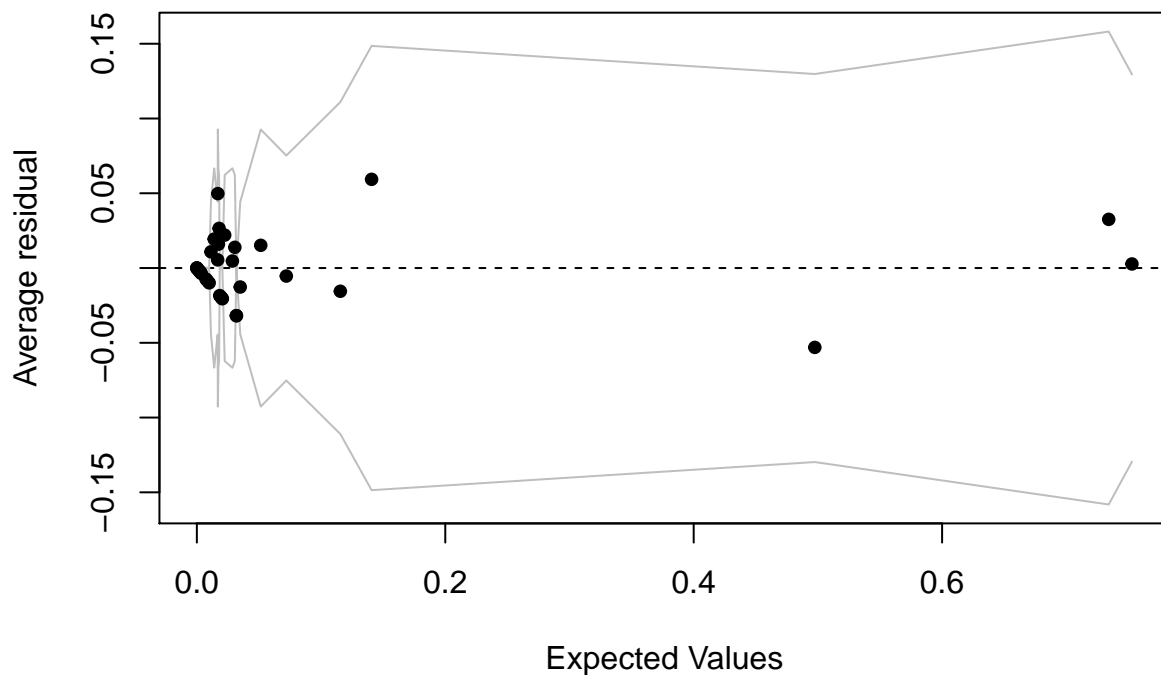
higher compared to staying in bone hackle, wake hackle, arrest ridge feature category; bone mirror feature; compression curl feature category and so on, except bone hackle feature and wake hackle with arrest ridge feature.

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 coefficients. Only coefficients of “days” for bone hackle feature and wake hackle with arrest ridge feature category are significant because of $0.017 < 5\%$.

Binned residual plot



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

```
##Chi-square test for Independence
```

```
## Warning in chisq.test(tb.1): Chi-squared approximation may be incorrect
```

```

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb.1
## X-squared = 12, df = 1, p-value = 5e-04
## Warning in chisq.test(tb.2): Chi-squared approximation may be incorrect
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb.2
## X-squared = 5e-30, df = 1, p-value = 1
##
## Chi-squared test for given probabilities
##
## data:  tb.3
## X-squared = 38, df = 1, p-value = 8e-10
## Warning in chisq.test(tb.4): Chi-squared approximation may be incorrect
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb.4
## X-squared = 5, df = 1, p-value = 0.03
## Warning in chisq.test(tb.5): Chi-squared approximation may be incorrect
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb.5
## X-squared = 4, df = 1, p-value = 0.05
## Warning in chisq.test(tb.6): Chi-squared approximation may be incorrect
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb.6
## X-squared = 7e-31, df = 1, p-value = 1
##
## Chi-squared test for given probabilities
##
## data:  tb.7
## X-squared = 82, df = 1, p-value <2e-16
## Warning in chisq.test(tb.8): Chi-squared approximation may be incorrect
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb.8
## X-squared = 8, df = 1, p-value = 0.005
## Warning in chisq.test(tb.9): Chi-squared approximation may be incorrect

```



```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb.9
## X-squared = 2e-30, df = 1, p-value = 1

##
## Chi-squared test for given probabilities
##
## data:  tb.10
## X-squared = 93, df = 1, p-value <2e-16
## Warning in chisq.test(tb.11): Chi-squared approximation may be incorrect

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb.11
## X-squared = 2e-30, df = 1, p-value = 1
## Warning in chisq.test(tb.12): Chi-squared approximation may be incorrect

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb.12
## X-squared = 1e-30, df = 1, p-value = 1

##
## Chi-squared test for given probabilities
##
## data:  tb.13
## X-squared = 90, df = 1, p-value <2e-16

##
## Chi-squared test for given probabilities
##
## data:  tb.14
## X-squared = 97, df = 1, p-value <2e-16
## Warning in chisq.test(tb.15): Chi-squared approximation may be incorrect

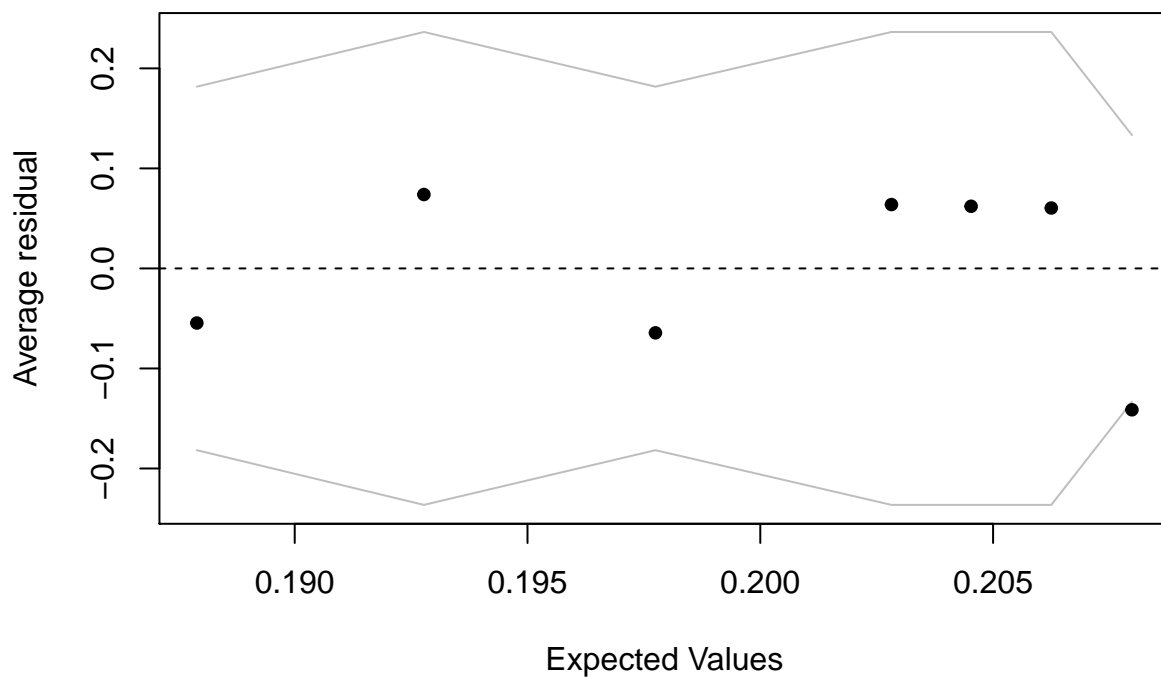
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb.15
## X-squared = 1e-30, df = 1, p-value = 1
```

Logistic Regression for six feature v.s. days

```
##
## Call:
## glm(formula = bone$`WAKE HACKLE` ~ bone$days, family = binomial())
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.683  -0.676  -0.664  -0.645   1.829
##
## Coefficients:
```

```
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.337128  0.367130  -3.64  0.00027 ***
## bone$days  -0.000352  0.001985  -0.18  0.85938
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 105.08  on 104  degrees of freedom
## Residual deviance: 105.05  on 103  degrees of freedom
## AIC: 109.1
##
## Number of Fisher Scoring iterations: 4
```

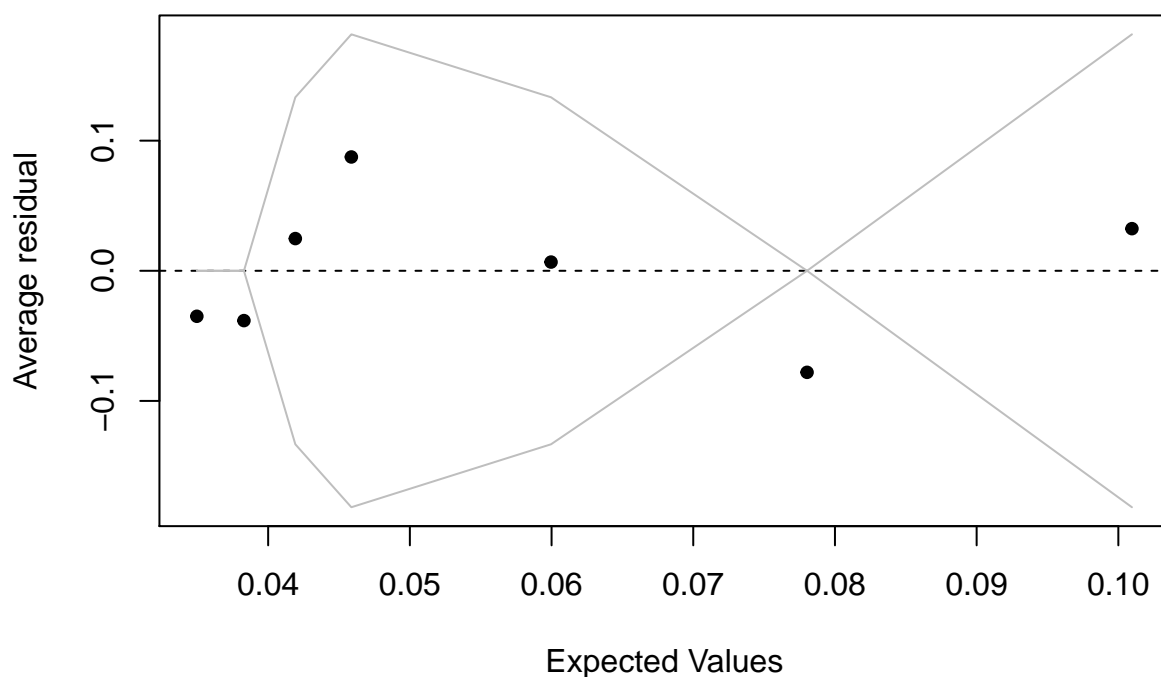
Binned residual plot



```
##
## Call:
## glm(formula = bone$`BONE HACKLE` ~ bone$days, family = binomial())
##
## Deviance Residuals:
##   Min       1Q   Median       3Q      Max
## -0.461 -0.403 -0.306 -0.279  2.519
##
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.31778   0.73705  -4.50  6.7e-06 ***
## bone$days   0.00314   0.00324   0.97   0.33
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 45.997  on 104  degrees of freedom
## Residual deviance: 45.067  on 103  degrees of freedom
## AIC: 49.07
##
## Number of Fisher Scoring iterations: 6
```

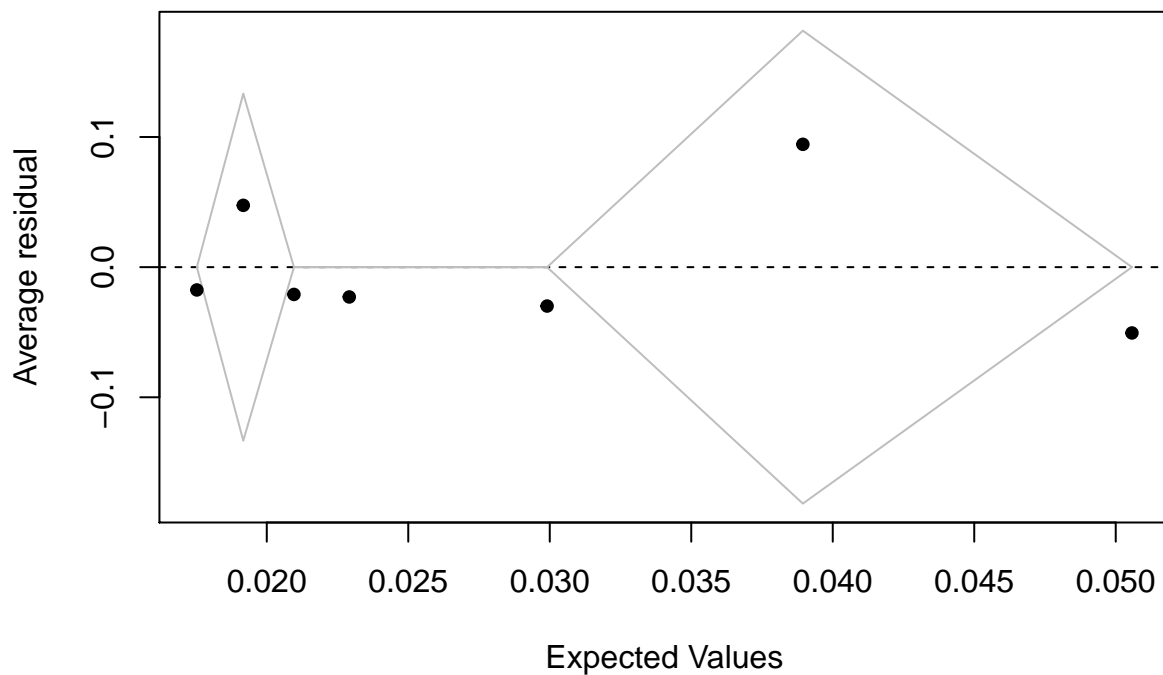
Binned residual plot



```
##
## Call:
## glm(formula = bone$`BONE MIRROR` ~ bone$days, family = binomial())
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.322  -0.282  -0.215  -0.197   2.812
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.02622    1.02982  -3.91  9.2e-05 ***
## bone$days    0.00304    0.00449   0.68    0.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

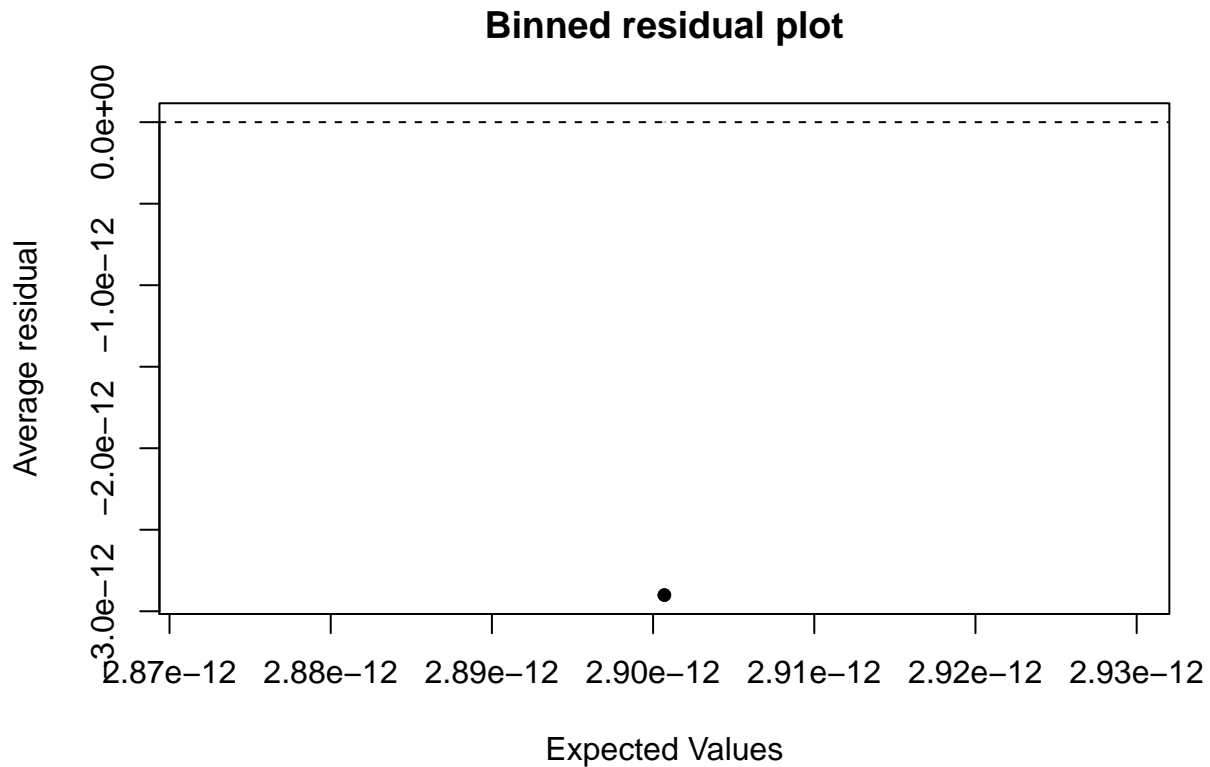
```
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 27.246 on 104 degrees of freedom
## Residual deviance: 26.795 on 103 degrees of freedom
## AIC: 30.8
##
## Number of Fisher Scoring iterations: 6
```

Binned residual plot



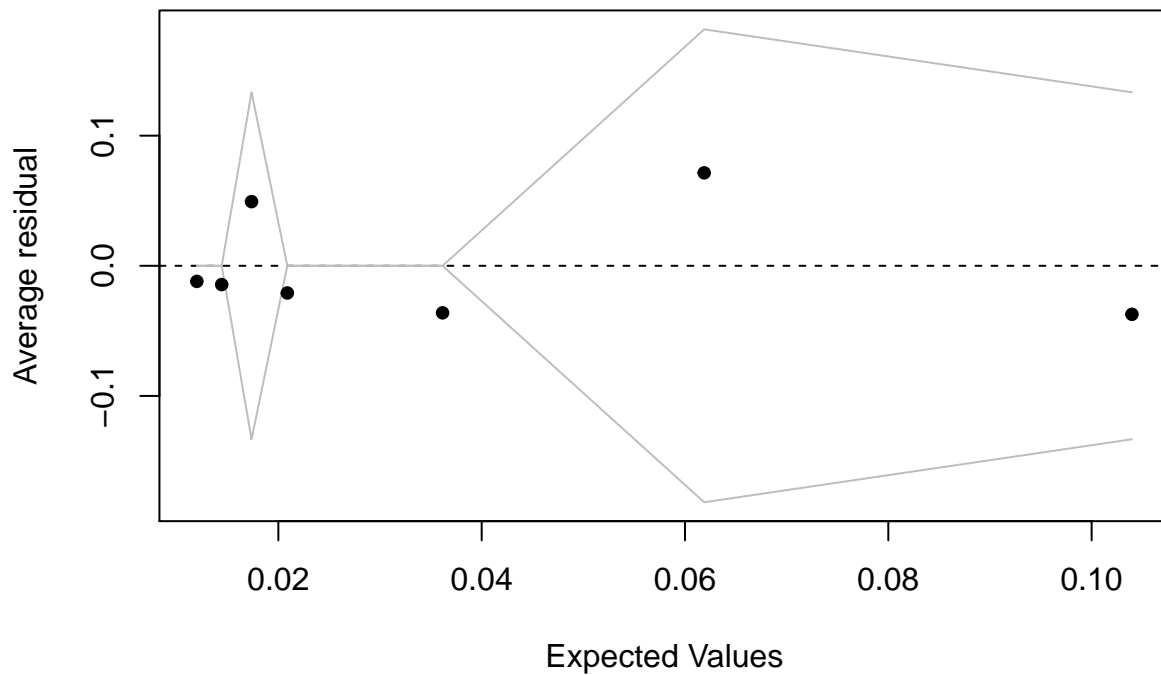
```
## Warning: glm.fit: algorithm did not converge
##
## Call:
## glm(formula = bone$`CANTILEVER CURL` ~ bone$days, family = binomial())
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.41e-06 -2.41e-06 -2.41e-06 -2.41e-06 -2.41e-06
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.66e+01  5.27e+04      0      1
## bone$days   -1.37e-17  2.80e+02      0      1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 0.0000e+00 on 104 degrees of freedom
## Residual deviance: 6.0917e-10 on 103 degrees of freedom
```

```
## AIC: 4
##
## Number of Fisher Scoring iterations: 25
```



```
##
## Call:
## glm(formula = bone$`ARREST RIDGE` ~ bone$days, family = binomial())
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.469  -0.357  -0.205  -0.171   2.847
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.41229    1.10916  -3.98  6.9e-05 ***
## bone$days    0.00627    0.00417   1.50    0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 33.987  on 104  degrees of freedom
## Residual deviance: 31.525  on 103  degrees of freedom
## AIC: 35.52
##
## Number of Fisher Scoring iterations: 6
```

Binned residual plot



```
##
## Call:
## glm(formula = bone$`WALLNER LINES` ~ bone$days, family = binomial())
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.208  -0.203  -0.193  -0.190   2.837
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.022938   1.108503  -3.63  0.00028 ***
## bone$days    0.000559   0.005660   0.10  0.92136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 19.805  on 104  degrees of freedom
## Residual deviance: 19.795  on 103  degrees of freedom
## AIC: 23.8
##
## Number of Fisher Scoring iterations: 6
```

Binned residual plot

