Ordinal logistic model on large, classified windows data

Ruth Gómez Graciani

Prepare the data

First, we obtain the density distribution, and local minima and maxima for the recombination map.

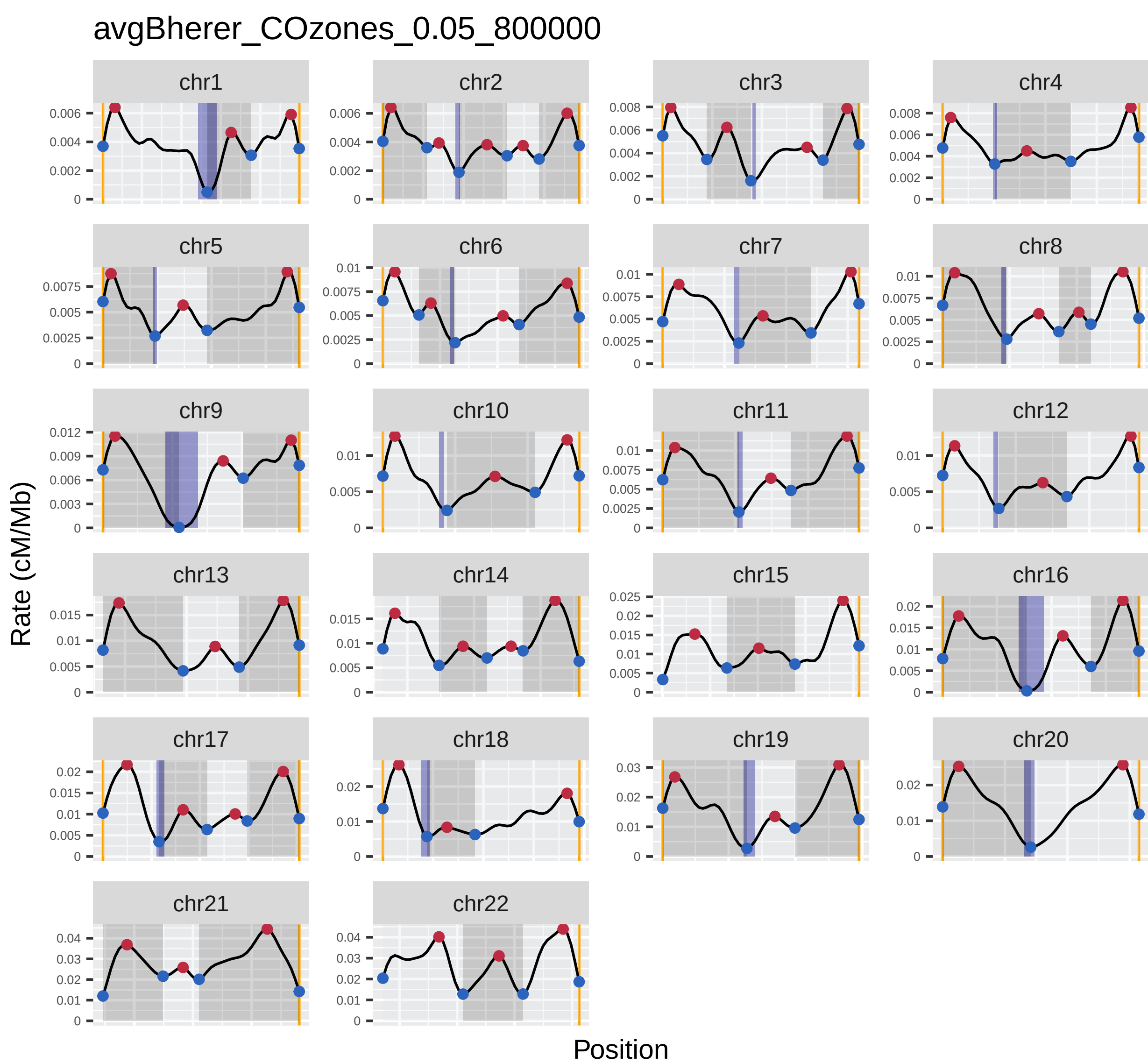


Figure 1: Crossover zones; centromeres in blue, workspace limits in orange.

Next, we define telomeric regions as the space between the chromosome start to the next local minimum, or between the chromosome end to the previous local minimum. We also define centromeric regions as the space between two local maxima that contains the centromere. When the local maximum delimiting a centromeric region is the same as the peak from the corresponding telomeric region (see chr1, chr5, chr7, chr8, etc.), the limit between the telomeric and centromeric regions is defined as the center point between the local maximum corresponding to the telomeric peak and the local minimum corresponding to the centromere valley. These categories will be represented as the "Color" variable in this analysis.

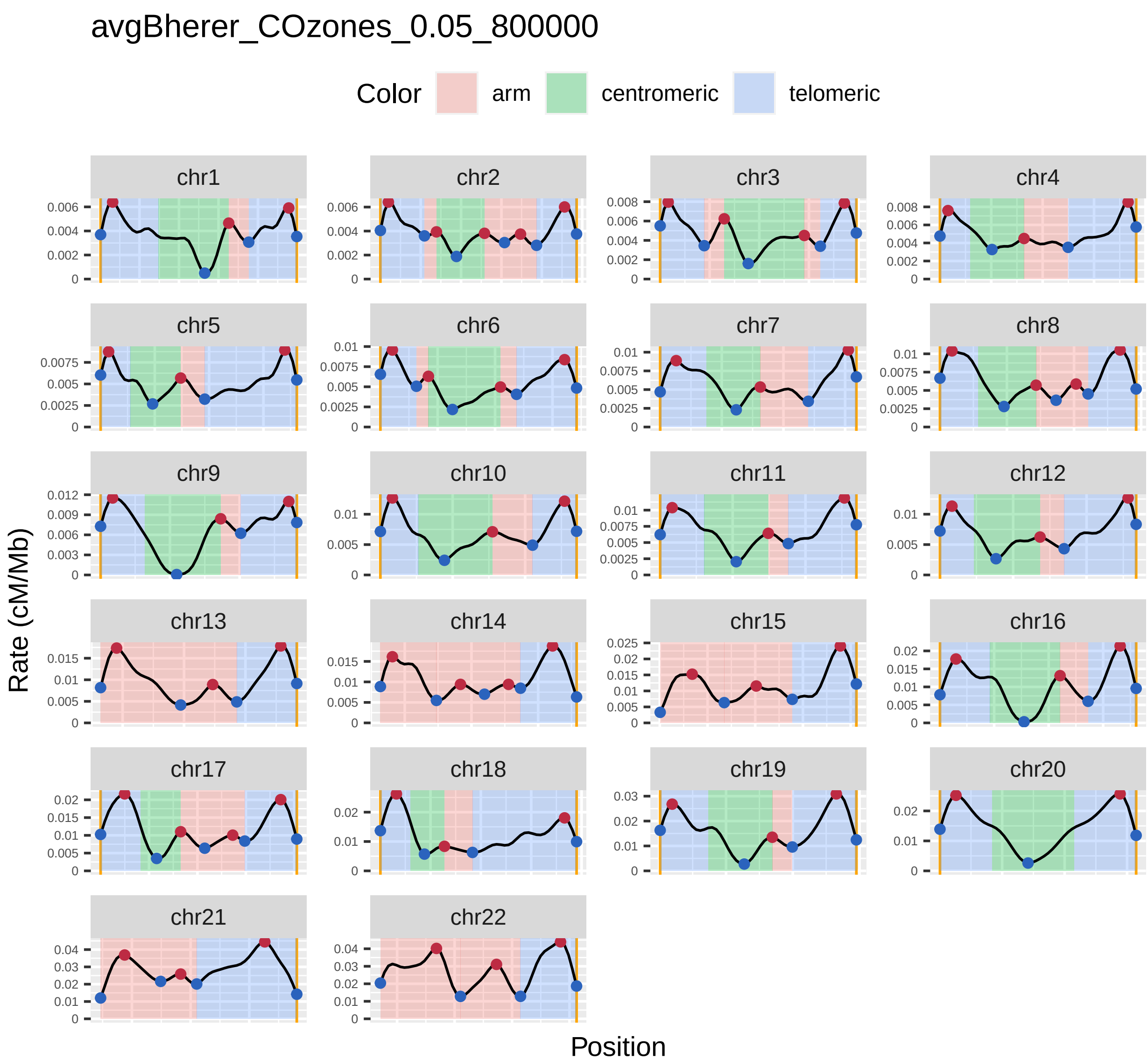


Figure 2: Color-coded windows for telomeric, centromeric and arm categories.

Descriptive statistics

Raw data:

```
##   Chromosome     Start       End       Color invCenters NHCenters NAHRCenters
## 1      chr10    158946  26393389   telomeric          3         2           1
## 2      chr10  26393389  77481515 centromeric          4         3           1
## 3      chr10  77481515 105096718         arm          1         1           0
## 4      chr10 105096718 135473442   telomeric          1         1           0
## 5      chr11    241489  30481001   telomeric          2         1           1
## 6      chr11  30481001  74465746 centromeric          3         2           1
##   Length.Mb allRepCounts WAvgRate.perMb
## 1  26.23444          360    0.013966282
## 2  51.08813         2730    0.007588011
## 3  27.61520          608    0.007725761
## 4  30.37672          200    0.011580532
## 5  30.23951          748    0.011815332
## 6  43.98474         2022    0.016501547
```

For each window, I calculated the number of total inversions, NH inversions, and NAHR inversions, the window length in Mb, number of repeats and the average recombination rate in cM/Mb.

I want to perform Ordinal Logistic Regressions on different subsets of the data. The assumptions of the Ordinal Logistic Regression are as follow:

1. The dependent variable is ordered.
2. One or more of the independent variables are either continuous, categorical or ordinal.
3. No multi-collinearity.
4. Proportional odds.

I show the data distributions in the figure below. The inversion counts have only a number of possible options, so they can be considered an ordinal variable. The independent variables are continuous and categorical, so assumptions 1 and 2 are satisfied

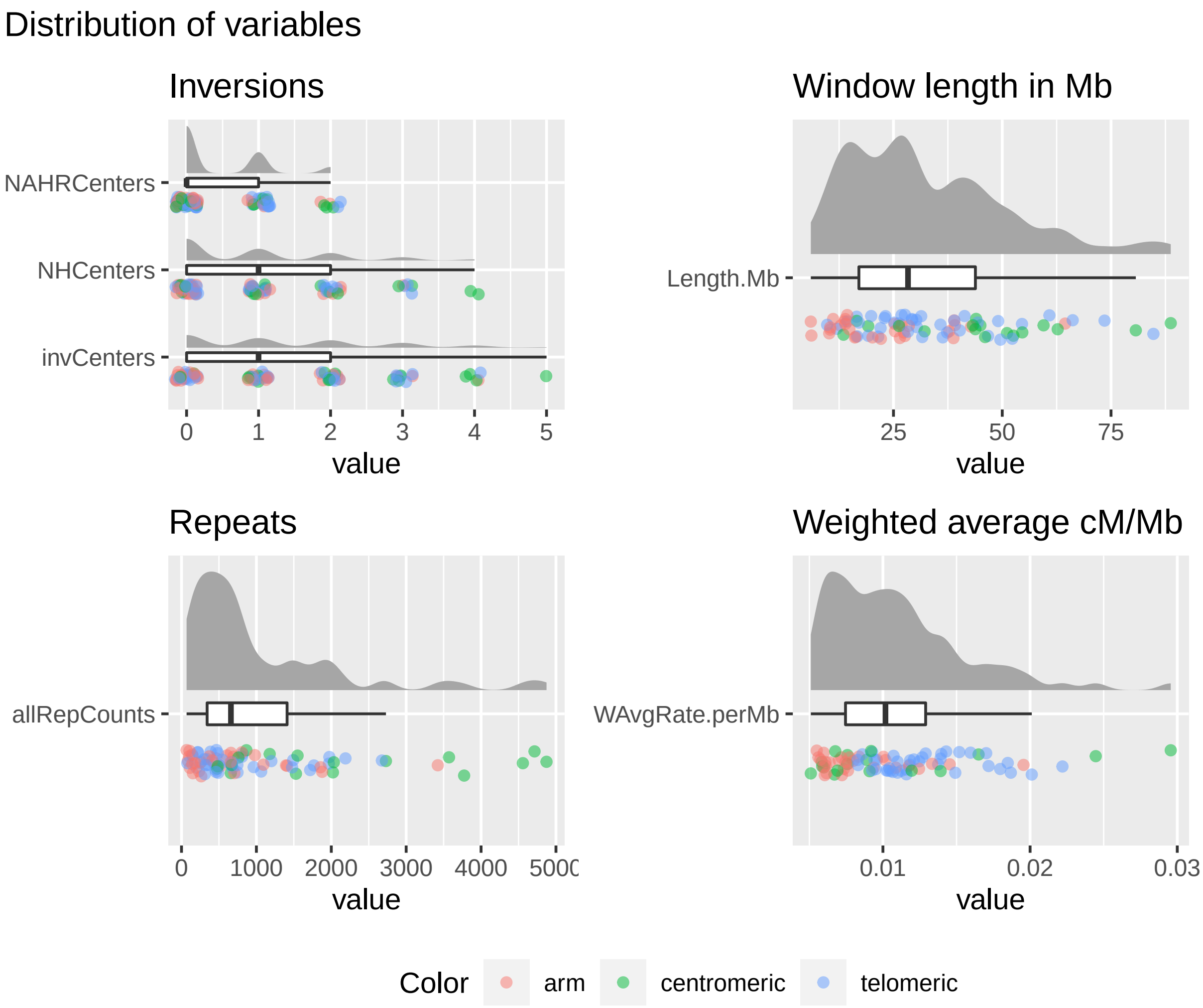


Figure 3: Distribution of variables.

We see that some categories have low number of cases, so I will make a "3 or more" category when relevant.

```
## [1] "Original counts"

##   CountGroups invCenters NHCenters NAHRCenters
## 1           0         29        41          54
## 2           1         22        22          24
## 3           2         17        14           7
## 4           3         11         6          NA
## 5           4          5         2          NA
## 6           5          1        NA          NA

## [1] "New counts"

##   CountGroups invCategory NHCategory NAHRCategory
## 1           0          29         41           54
## 2           1          22         22           24
## 3           2          17         14            7
## 4          3+          17          8           NA
```

With these groups, I visualize the relationships between dependent and independent variables.

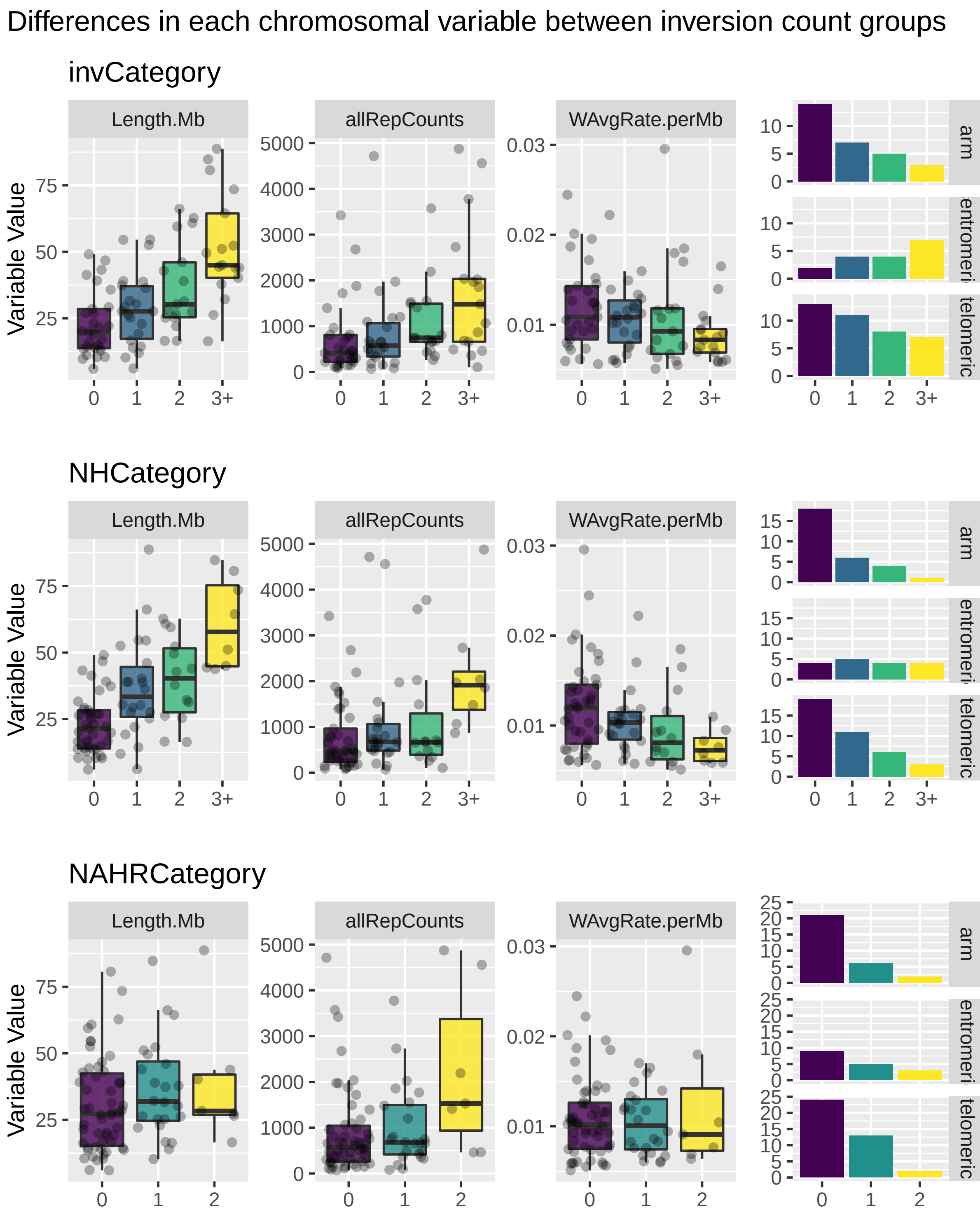


Figure 4: Potential effect of independent variables on the different types of invesions.

Finally, I will test assumption number 3, no multi-collinearity between independent variables.

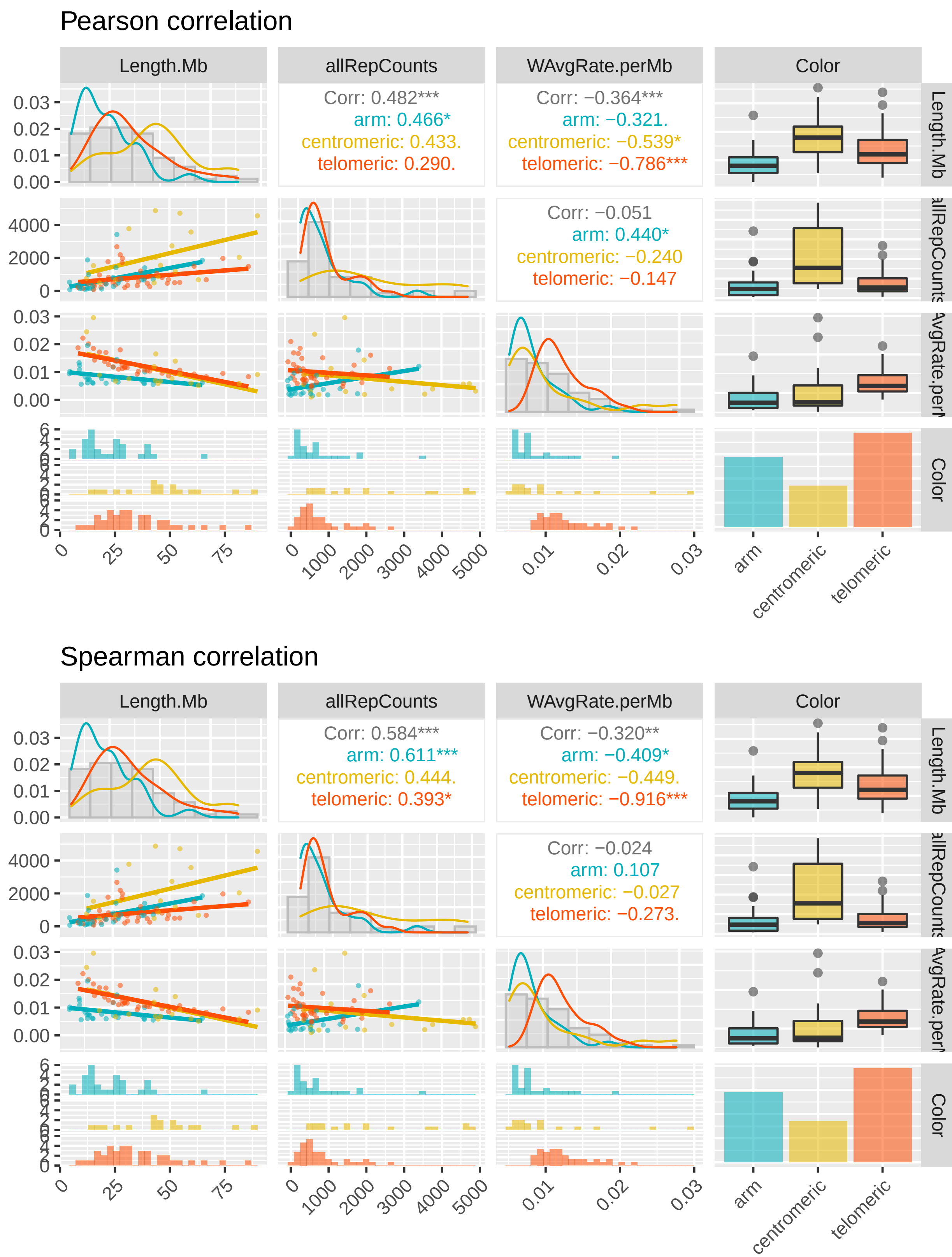


Figure 5: Correlations between variables.

We see that our three variables are significantly correlated, but this does not confirm multi-collinearity. I perform a variance inflation factor test on the corresponging linear model to further check the multi-collinearity.

```
##                    GVIF Df GVIF^(1/(2*Df))
## Length.Mb      2.244535  1        1.498177
## allRepCounts   1.652985  1        1.285685
## Color          2.422486  2        1.247571
## WAvgRate.perMb 1.857273  1        1.362818
```

The general rule of thumbs for VIF test is that if the VIF value is greater than 10, then there is multi-collinearity, so we can say that the third assumption (no multi-collinearity) is satisfied.

The proportional odds assumption will be tested for each model that we fit in the following analyses.

Variable scalation (optional)

Standardized coefficients are useful in our case to compare effects of predictors reported in different units. The most straightforward way is using the Agresti method of standardization, applied with the `scale()` function.

```
##    Length.Mb     Length.Mb.Scaled   allRepCounts  allRepCounts.Scaled
##  Min.   : 6.00   Min.   :-1.4427   Min.   :  68   Min.   :-0.8920
##  1st Qu.:17.06   1st Qu.:-0.8468   1st Qu.: 342   1st Qu.:-0.6378
##  Median :28.34   Median :-0.2387   Median : 660   Median :-0.3428
##  Mean   :32.77   Mean   : 0.0000   Mean   :1030   Mean   : 0.0000
##  3rd Qu.:43.83   3rd Qu.: 0.5961   3rd Qu.:1410   3rd Qu.: 0.3529
##  Max.   :88.72   Max.   : 3.0158   Max.   :4874   Max.   : 3.5662
##  WAvgRate.perMb     WAvgRate.perMb.Scaled
##  Min.   :0.005099   Min.   :-1.2487
##  1st Qu.:0.007456   1st Qu.:-0.7431
##  Median :0.010175   Median :-0.1598
##  Mean   :0.010920   Mean   : 0.0000
##  3rd Qu.:0.012903   3rd Qu.: 0.4254
##  Max.   :0.029556   Max.   : 3.9983
```

Once the model is fitted, we can use the sd to transform scaled coefficients to natural coefficients and viceversa.

Total inversions (invCategory)

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error   t value
## Length.Mb         0.0611038  0.0153404   3.98319
## allRepCounts      0.0001466  0.0003108   0.47158
## Colorcentromeric  0.4188044  0.6787254   0.61705
## Colortelomeric   -0.0114797  0.4842683  -0.02371
## WAvgRate.perMb   -0.3213942  0.0063487 -50.62340
##
## Intercepts:
##      Value    Std. Error t value
## 0|1    1.2895   0.4848     2.6600
## 1|2    2.6478   0.5402     4.9013
## 2|3+   3.9947   0.6393     6.2485
##
## Residual Deviance: 198.9591
## AIC: 214.9591
```

We compare the t-value against the standard normal distribution to calculate the p-value.

```
##                          Value   Std. Error      t value    p value
## Length.Mb         0.0611037761 0.0153404205   3.98318783 0.00006800
## allRepCounts      0.0001465696 0.0003108054   0.47158002 0.63722659
## Colorcentromeric  0.4188043837 0.6787253867   0.61704541 0.53720477
## Colortelomeric   -0.0114797398 0.4842683421  -0.02370533 0.98108766
## WAvgRate.perMb   -0.3213942282 0.0063487286 -50.62339987 0.00000000
## 0|1               1.2895363757 0.4847873688   2.66000407 0.00781397
## 1|2               2.6478284930 0.5402323106   4.90127755 0.00000095
## 2|3+              3.9946895737 0.6392988143   6.24854839 0.00000000
```

We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI does not cross 0, the parameter estimate is statistically significant.

```
## [1] "Profiling likelihod"
```

```
##                         2.5 %       97.5 %
## Length.Mb         0.024944026 0.0974041078
## allRepCounts     -0.000380357 0.0006689941
## Colorcentromeric -0.942808000 1.9306061832
## Colortelomeric   -1.198121063 1.2140504509
## WAvgRate.perMb             NA           NA
```

```
## [1] "Assuming a normal distribtuion"
```

```
##                          2.5 %        97.5 %
## Length.Mb         0.0310371045  0.0911704477
## allRepCounts     -0.0004625978  0.0007557371
## Colorcentromeric -0.9114729296  1.7490816969
## Colortelomeric   -0.9606282491  0.9376687695
```

```
## WAvgRate.perMb   -0.3338375077 -0.3089509487
```

We convert the coefficients into odds ratios. To get the OR and confidence intervals, we just exponentiate the estimates and confidence intervals (here I used the likelihood confidence intervals).

```
##                  Odds Ratio      2.5%    97.5%
## Length.Mb         1.0630092 1.0252577 1.102306
## allRepCounts      1.0001466 0.9996197 1.000669
## Colorcentromeric  1.5201430 0.3895325 6.893688
## Colortelomeric    0.9885859 0.3017607 3.367095
## WAvgRate.perMb    0.7251373        NA       NA
```

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0630092 times more likely to increase in inversion amount category."

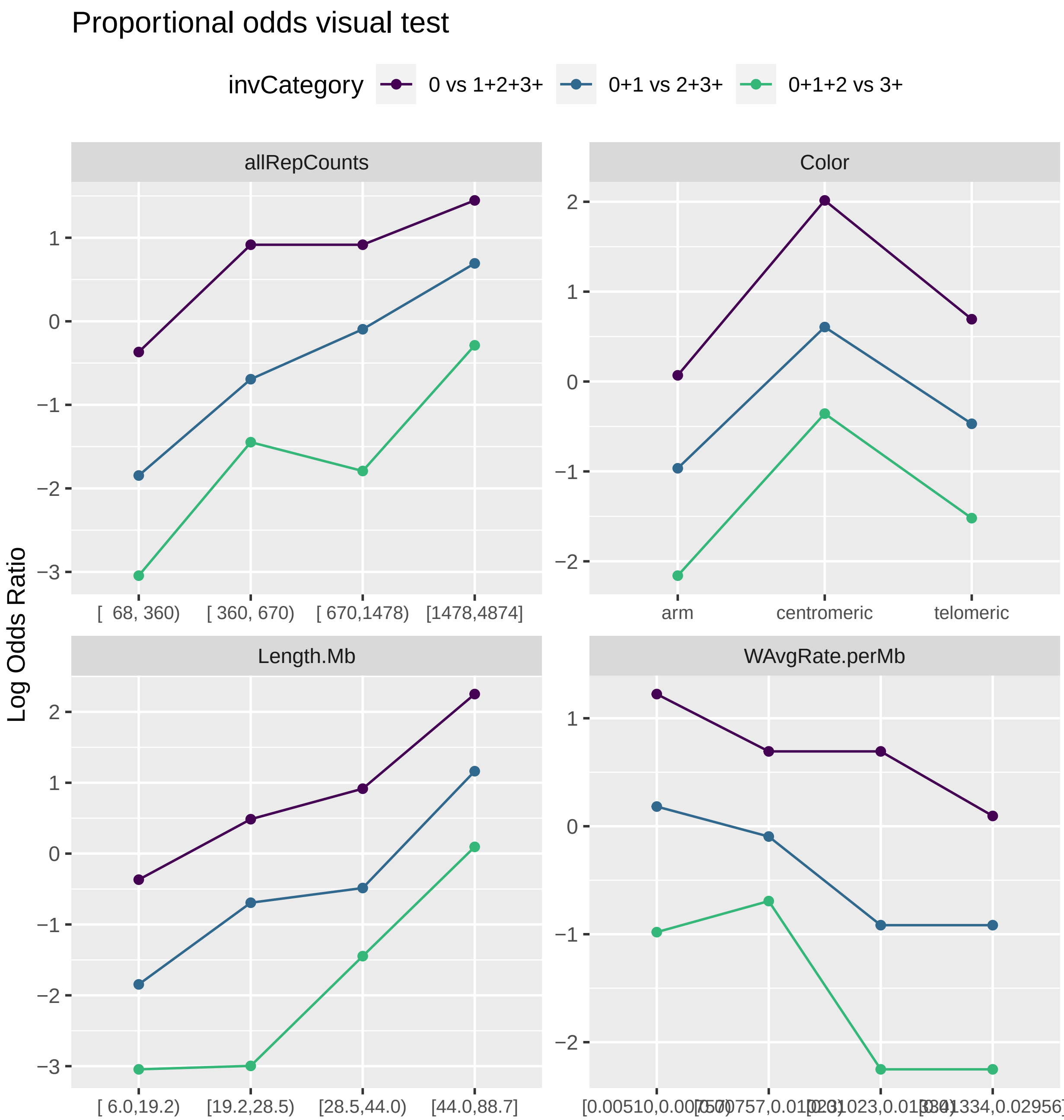
Proportional odds assessment

Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe the relationship between each pair of outcome groups.

We test the parallel regression assumption with a Brant test:

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          3.95    10  0.95
## Length.Mb        1.44    2   0.49
## allRepCounts     0.42    2   0.81
## Colorcentromeric 1.54    2   0.46
## Colortelomeric       1.83    2   0.4
## WAvgRate.perMb       2.62    2   0.27
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

We can also evaluate the parallel regression visually. We transform the ordinal dependent variable with k categories into a series of k-1 binary variables that indicate whether the dependent value is above or below a cutpoint (e.g. windows with at least 2 inversions vs windows with less than 2 inversions). We then calculate the observed Log Odds Ratio for each binary variable across multiple value ranges of the independent variables. The lines should be approximately parallel, that each independent variable affects the probability of increasing by 1 level the inversion count in the same way, for all transitions, and that we don't need a specific model for each level increase.



Predicted probabilites

Although our objective is to describe the dataset, predicted probabilities are usually easier to understand than either the coefficients or the Odds Ratios.

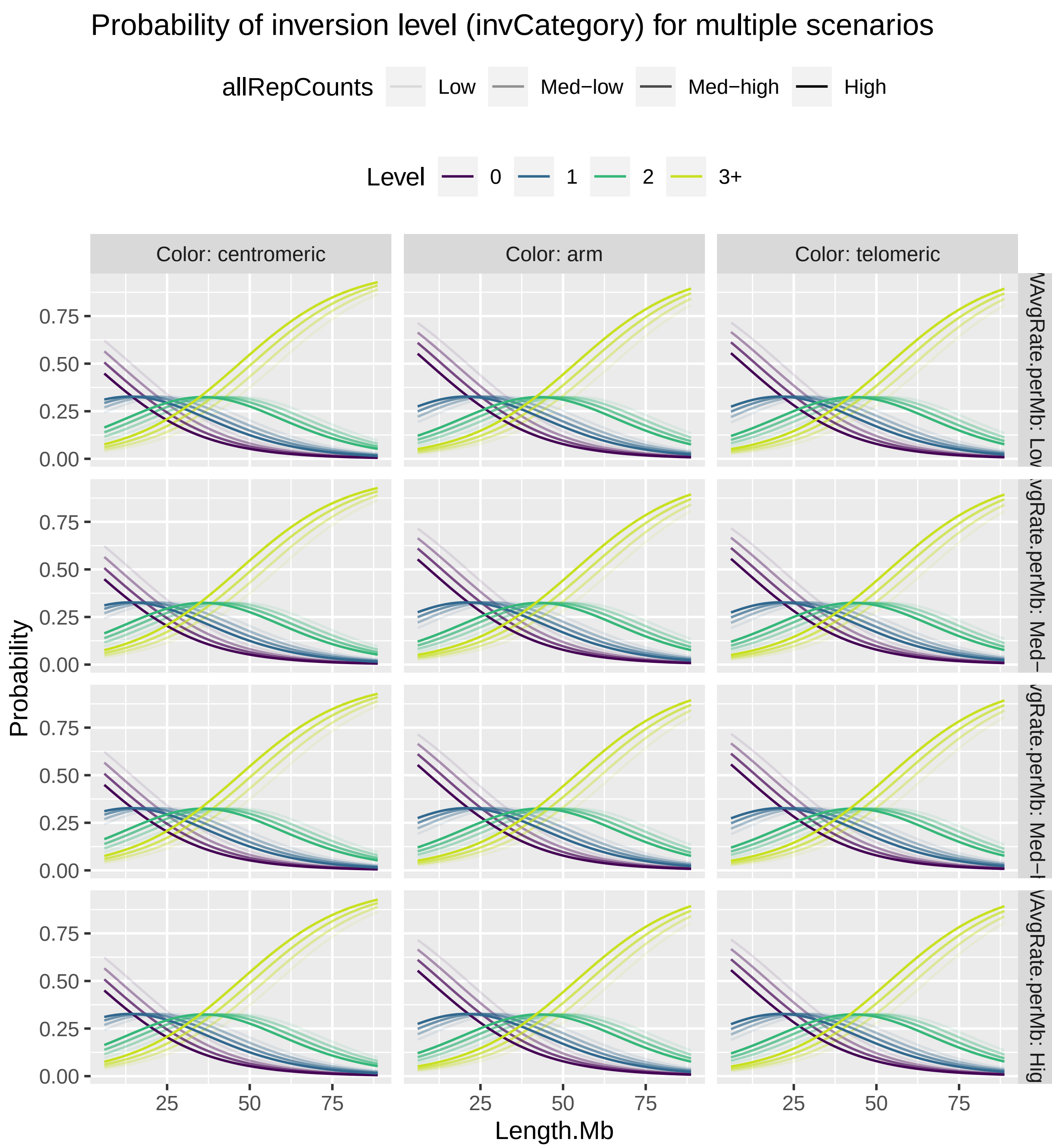


Figure 6: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions (NHCategory)

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error    t value
## Length.Mb         5.911e-02  0.0163860  3.607e+00
## allRepCounts      3.997e-05  0.0003112  1.285e-01
## Colorcentromeric  6.704e-01  0.7373105  9.093e-01
## Colortelomeric    2.951e-01  0.5295407  5.572e-01
## WAvgRate.perMb   -1.040e+02  0.0064252 -1.619e+04
##
## Intercepts:
##      Value       Std. Error  t value
## 0|1       0.9261      0.5418      1.7092
## 1|2       2.4915      0.6397      3.8950
## 2|3+      4.1239      0.7920      5.2067
##
## Residual Deviance: 170.2522
## AIC: 186.2522
```

We compare the t-value against the standard normal distribution to calculate the p-value.

```
##                          Value   Std. Error       t value    p value
## Length.Mb         5.910842e-02 0.0163859589  3.607260e+00 0.00030945
## allRepCounts      3.997291e-05 0.0003111618  1.284634e-01 0.89778224
## Colorcentromeric  6.704152e-01 0.7373105232  9.092711e-01 0.36320703
## Colortelomeric    2.950657e-01 0.5295407155  5.572106e-01 0.57738355
## WAvgRate.perMb   -1.040145e+02 0.0064252265 -1.618846e+04 0.00000000
## 0|1               9.260949e-01 0.5418415637  1.709162e+00 0.08742098
## 1|2               2.491493e+00 0.6396717166  3.894956e+00 0.00009822
## 2|3+              4.123887e+00 0.7920337286  5.206706e+00 0.00000019
```

We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI does not cross 0, the parameter estimate is statistically significant.

```
## [1] "Profiling likelihod"
```

```
##                         2.5 %       97.5 %
## Length.Mb         0.024457347 0.0977106249
## allRepCounts     -0.000482775 0.0005546513
## Colorcentromeric -0.880347037 2.1980994900
## Colortelomeric   -1.008671121 1.6408860361
## WAvgRate.perMb             NA           NA
```

```
## [1] "Assuming a normal distribtuion"
```

```
##                          2.5 %        97.5 %
## Length.Mb         2.699253e-02  9.122431e-02
## allRepCounts     -5.698930e-04  6.498388e-04
## Colorcentromeric -7.746869e-01  2.115517e+00
## Colortelomeric   -7.428150e-01  1.332946e+00
```

```
## WAvgRate.perMb   -1.040271e+02 -1.040019e+02
```

We convert the coefficients into odds ratios. To get the OR and confidence intervals, we just exponentiate the estimates and confidence intervals (here I used the likelihood confidence intervals).

```
##                     Odds Ratio      2.5%    97.5%
## Length.Mb         1.060890e+00 1.0247589 1.102644
## allRepCounts      1.000040e+00 0.9995173 1.000555
## Colorcentromeric 1.955049e+00 0.4146390 9.007878
## Colortelomeric    1.343215e+00 0.3647033 5.159739
## WAvgRate.perMb    6.715528e-46        NA       NA
```

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0608903 times more likely to increase in inversion amount category."

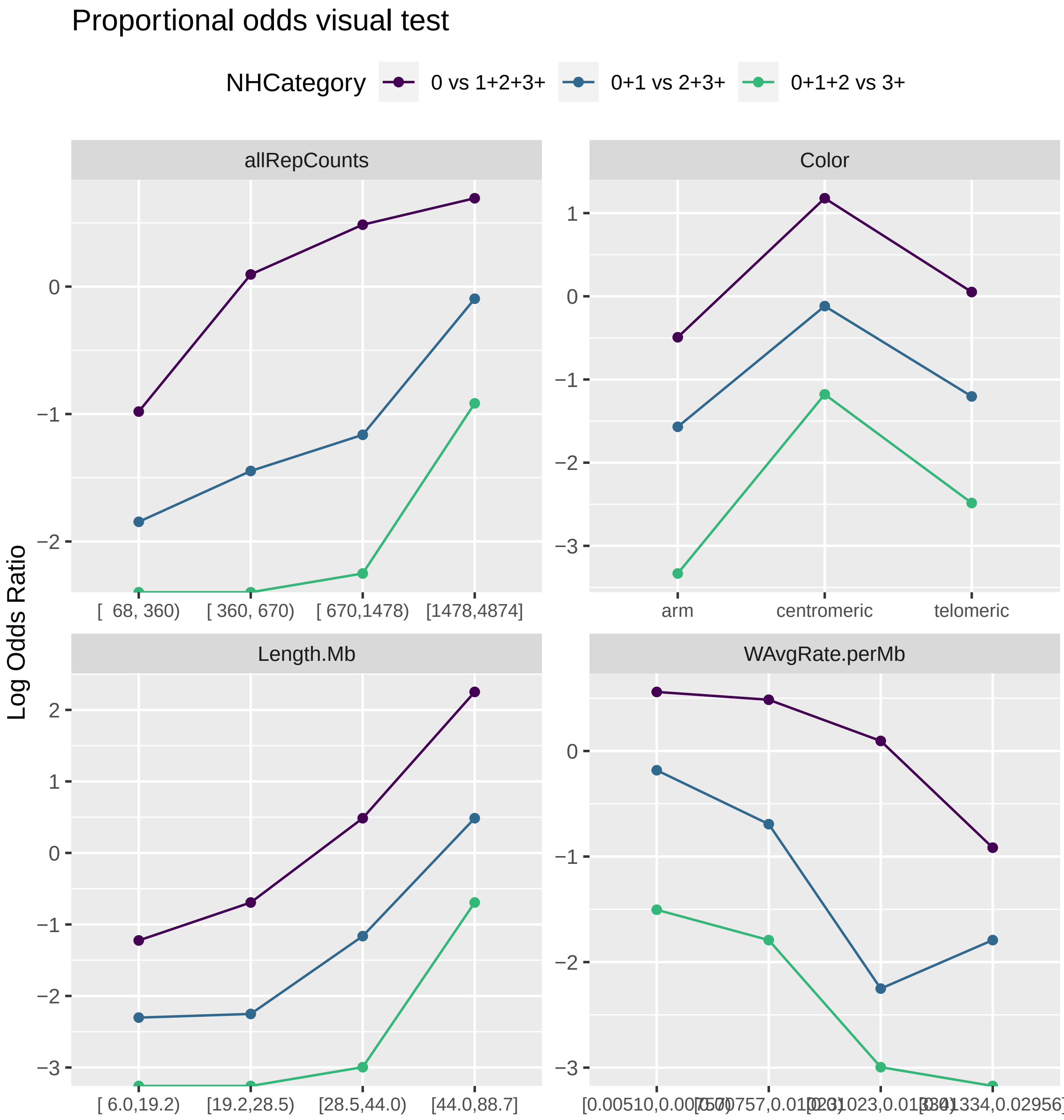
Proportional odds assessment

Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe the relationship between each pair of outcome groups.

We test the parallel regression assumption with a Brant test:

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          3.76    10  0.96
## Length.Mb        0.74    2   0.69
## allRepCounts     1.31    2   0.52
## Colorcentromeric 0.05    2   0.97
## Colortelomeric       0.92    2   0.63
## WAvgRate.perMb       1.81    2   0.4
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

We can also evaluate the parallel regression visually. We transform the ordinal dependent variable with k categories into a series of k-1 binary variables that indicate whether the dependent value is above or below a cutpoint (e.g. windows with at least 2 inversions vs windows with less than 2 inversions). We then calculate the observed Log Odds Ratio for each binary variable across multiple value ranges of the independent variables. The lines should be approximately parallel, that each independent variable affects the probability of increasing by 1 level the inversion count in the same way, for all transitions, and that we don't need a specific model for each level increase.



Predicted probabilites

Although our objective is to describe the dataset, predicted probabilities are usually easier to understand than either the coefficients or the Odds Ratios.

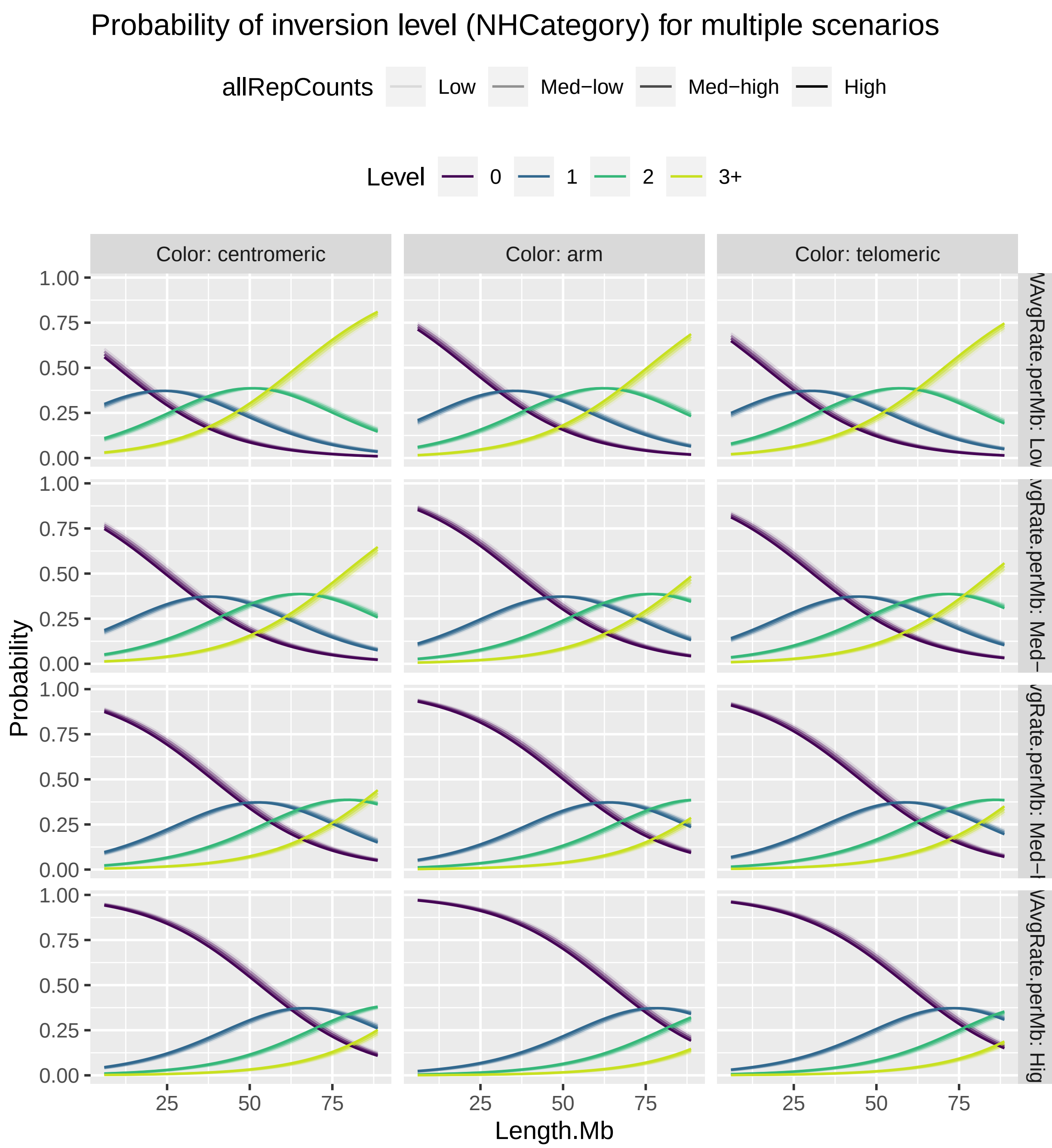


Figure 7: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions (NAHRCategory)

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error   t value
## Length.Mb         1.150e-02  0.0142774    0.8056
## allRepCounts      3.843e-04  0.0003726    1.0312
## Colorcentromeric 1.084e-01  0.7566870    0.1433
## Colortelomeric   1.113e-01  0.5511015    0.2019
## WAvgRate.perMb   3.472e+01  0.0065439 5306.0025
##
## Intercepts:
##     Value     Std. Error t value
## 0|1    1.7778    0.5422     3.2787
## 1|2    3.7569    0.6756     5.5605
##
## Residual Deviance: 138.6956
## AIC: 152.6956
```

We compare the t-value against the standard normal distribution to calculate the p-value.

```
##                          Value   Std. Error      t value    p value
## Length.Mb         1.150156e-02 0.0142773947    0.8055782 0.42048608
## allRepCounts      3.842658e-04 0.0003726262    1.0312368 0.30242978
## Colorcentromeric 1.084467e-01 0.7566869825    0.1433178 0.88603920
## Colortelomeric   1.112584e-01 0.5511014827    0.2018837 0.84000767
## WAvgRate.perMb   3.472217e+01 0.0065439422 5306.0024867 0.00000000
## 0|1               1.777790e+00 0.5422232208    3.2787054 0.00104284
## 1|2               3.756901e+00 0.6756455977    5.5604616 0.00000003
```

We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI does not cross 0, the parameter estimate is statistically significant.

```
## [1] "Profiling likelihod"
```

```
##                         2.5 %       97.5 %
## Length.Mb        -0.0221121466 0.0459839129
## allRepCounts     -0.0001506436 0.0009206725
## Colorcentromeric -1.4853806336 1.6483353354
## Colortelomeric   -1.1889853329 1.4435535816
## WAvgRate.perMb              NA           NA
```

```
## [1] "Assuming a normal distribtuion"
```

```
##                        2.5 %      97.5 %
## Length.Mb        -0.016481621  0.03948474
## allRepCounts     -0.000346068  0.00111460
## Colorcentromeric -1.374632525  1.59152594
## Colortelomeric   -0.968880668  1.19139745
## WAvgRate.perMb   34.709347854 34.73499964
```

We convert the coefficients into odds ratios. To get the OR and confidence intervals, we just exponentiate the estimates and confidence intervals (here I used the likelihood confidence intervals).

```
##                    Odds Ratio      2.5%    97.5%
## Length.Mb        1.011568e+00 0.9781305 1.047058
## allRepCounts     1.000384e+00 0.9998494 1.000921
## Colorcentromeric 1.114546e+00 0.2264161 5.198319
## Coloretelomeric  1.117684e+00 0.3045301 4.235721
## WAvgRate.perMb   1.201292e+15        NA       NA
```

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.011568 times more likely to increase in inversion amount category."

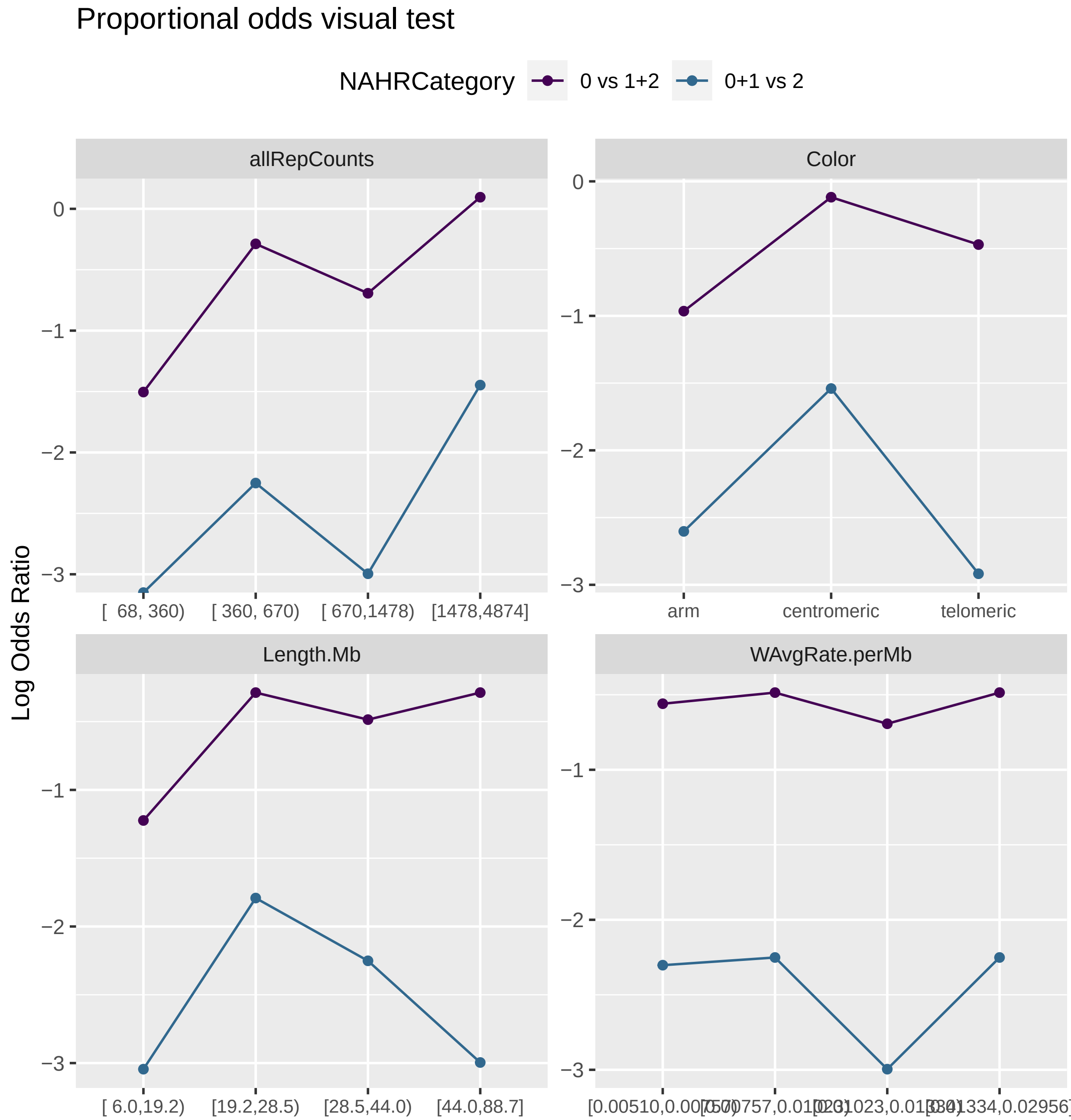
Proportional odds assessment

Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe the relationship between each pair of outcome groups.

We test the parallel regression assumption with a Brant test:

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          7.57    5   0.18
## Length.Mb        0.01    1   0.93
## allRepCounts     2.6 1   0.11
## Colorcentromeric 0.7 1   0.4
## Colortelomeric       1.24    1   0.27
## WAvgRate.perMb       2.55    1   0.11
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

We can also evaluate the parallel regression visually. We transform the ordinal dependent variable with k categories into a series of k-1 binary variables that indicate whether the dependent value is above or below a cutpoint (e.g. windows with at least 2 inversions vs windows with less than 2 inversions). We then calculate the observed Log Odds Ratio for each binary variable across multiple value ranges of the independent variables. The lines should be approximately parallel, that each independent variable affects the probability of increasing by 1 level the inversion count in the same way, for all transitions, and that we don't need a specific model for each level increase.



Predicted probabilites

Although our objective is to describe the dataset, predicted probabilities are usually easier to understand than either the coefficients or the Odds Ratios.

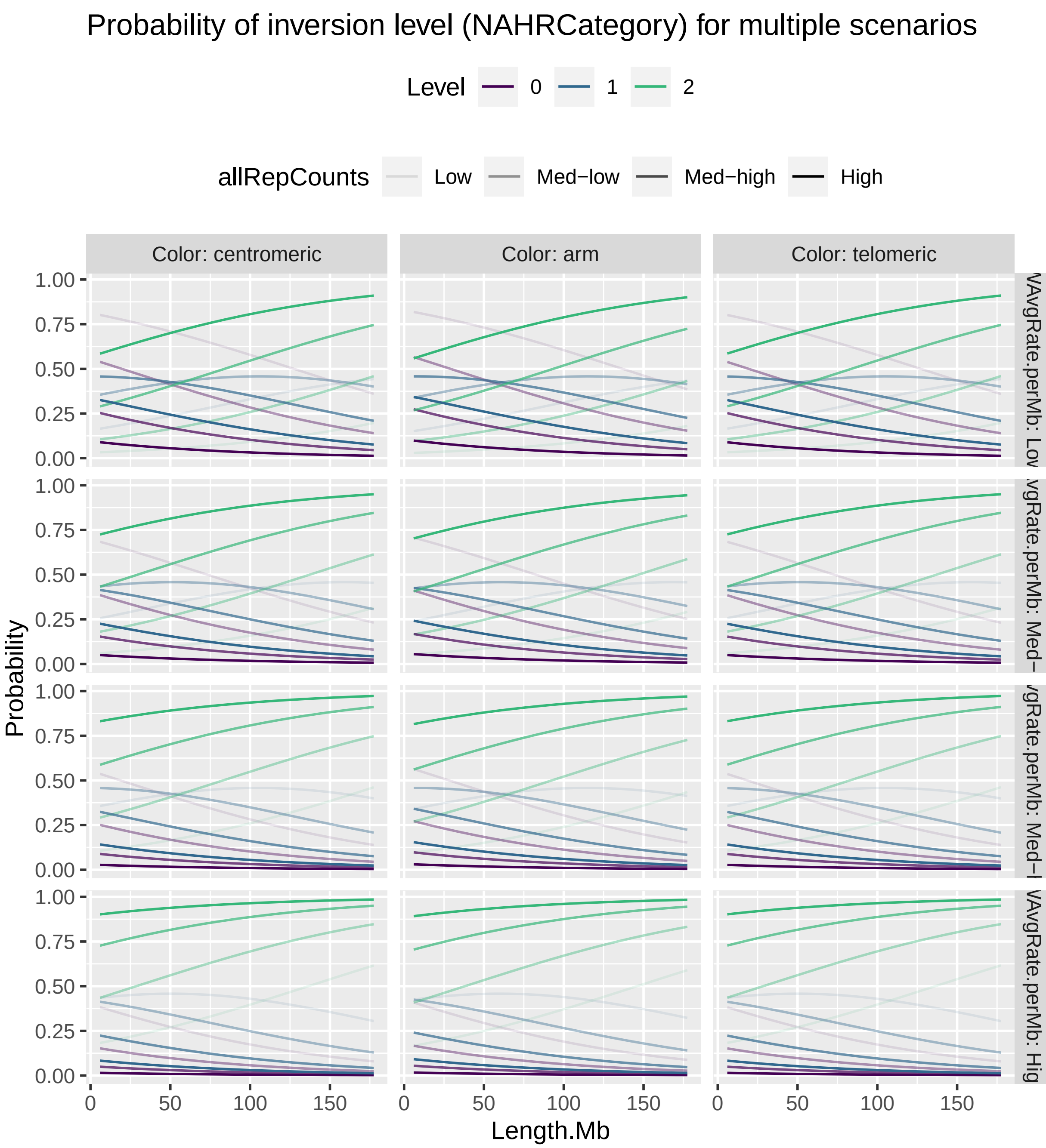


Figure 8: Probabiilty of having 0 to >3 inversions depending on multiple independent variables