

# 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.

femBherer\_COzones\_0.05\_800000

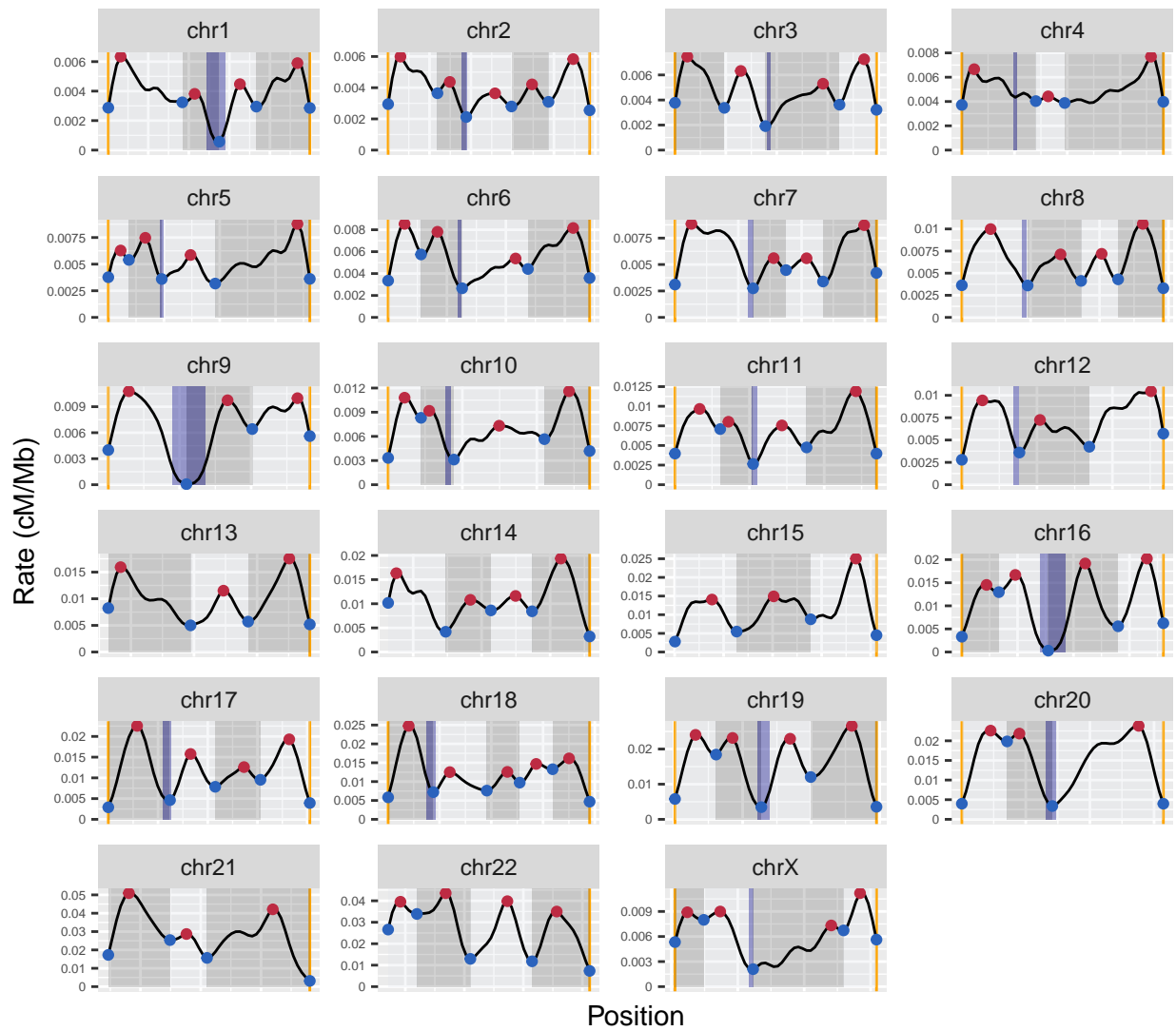


Figure 1: Crossover zones; centromeres in blue, workspace limits in orange.

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.

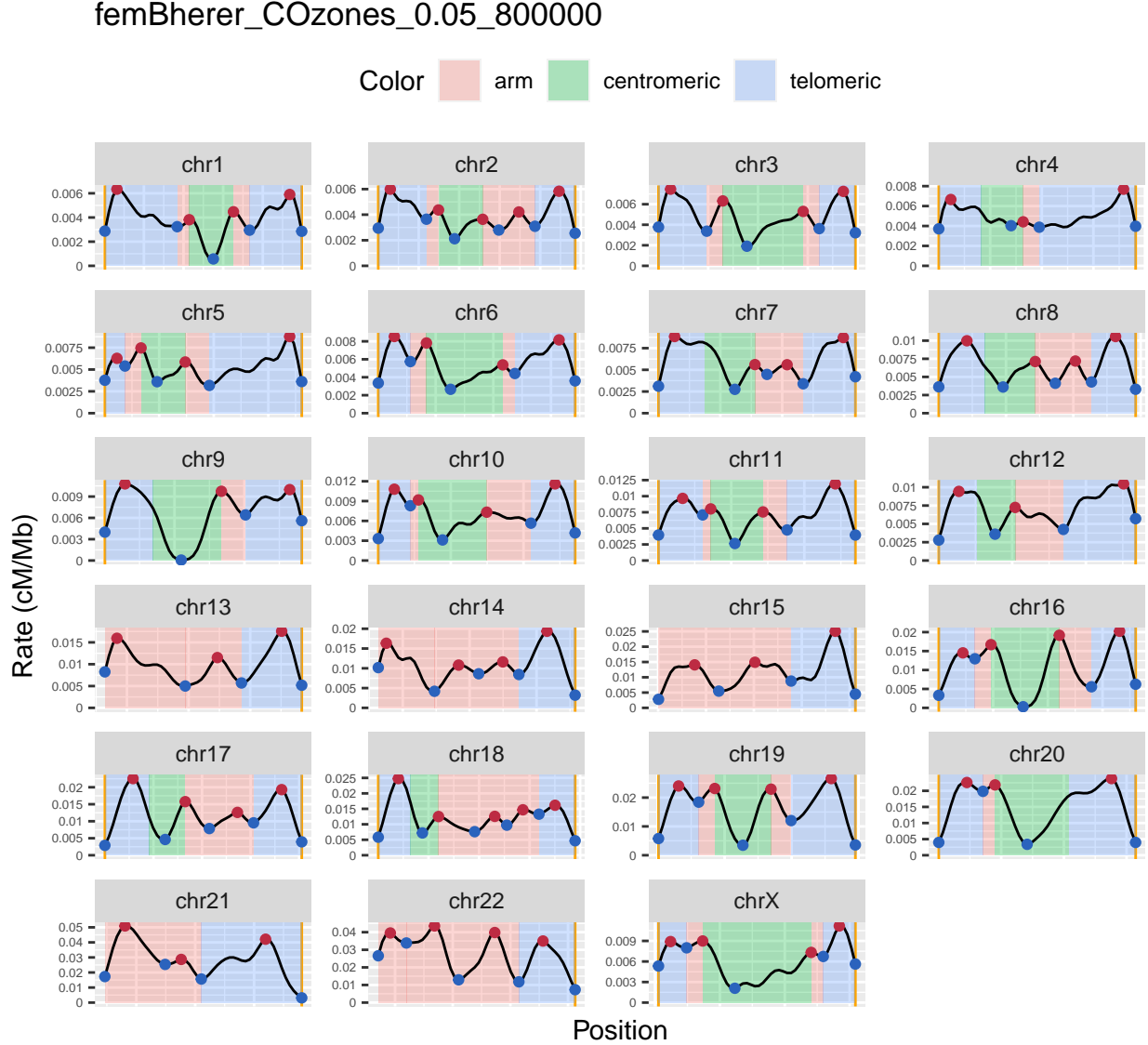


Figure 2: Color-coded windows for telomeric, centromeric and arm categories.

## Descriptive statistics

Raw data:

##	Chromosome	Start	End	Color	invCenters	NHCenters	NAHRCenters
## 1	chr10	158946	22251109	telomeric	3	2	1
## 2	chr10	22251109	27774149	arm	0	0	0
## 3	chr10	27774149	74719995	centromeric	3	2	1
## 4	chr10	74719995	105096718	arm	2	2	0
## 5	chr10	105096718	135473442	telomeric	1	1	0
## 6	chr11	241489	30481001	telomeric	2	1	1

##	Length.Mb	allRepCounts	WAvgRate.perMb
## 1	22.092163	340	1.855788
## 2	5.523041	200	1.465803
## 3	46.945846	2488	1.262444
## 4	30.376724	670	1.282558
## 5	30.376724	200	1.864682
## 6	30.239512	748	1.591677

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

## Distribution of variables

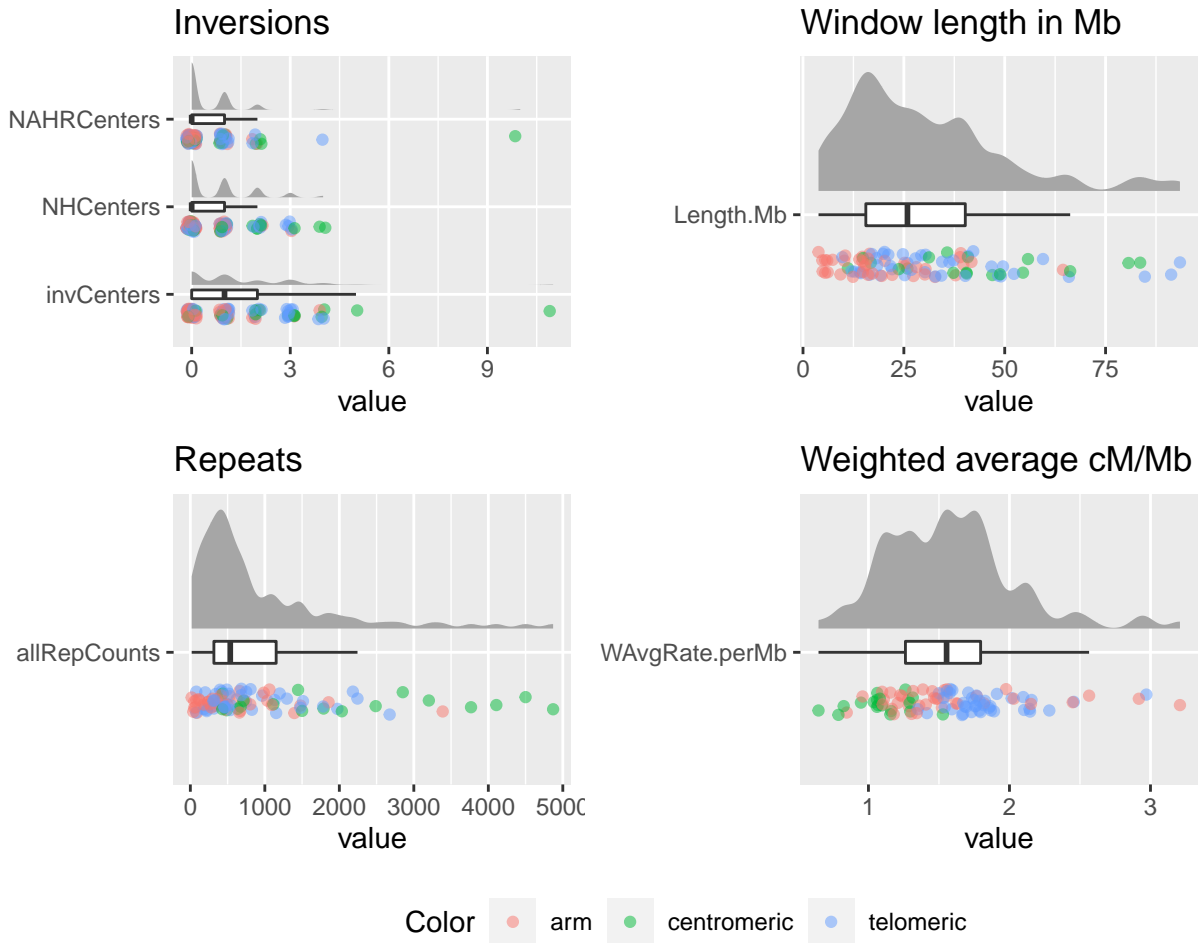


Figure 3: Distribution of variables.

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         36         50          64
## 2         1         28         26          24
## 3        10         NA         NA           1
## 4        11          1         NA          NA
## 5         2         12         13           7
## 6         3         14          6          NA
## 7         4          5          2           1
## 8         5          1         NA          NA

## [1] "New counts"

##   CountGroups invCategory NHCategory NAHRCategory
## 1         0         36         50          64
## 2         1         28         26          24
## 3         2         12         13           7
## 4        3+         21          8           2
```

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

## Differences in each chromosomal variable between inversion count groups

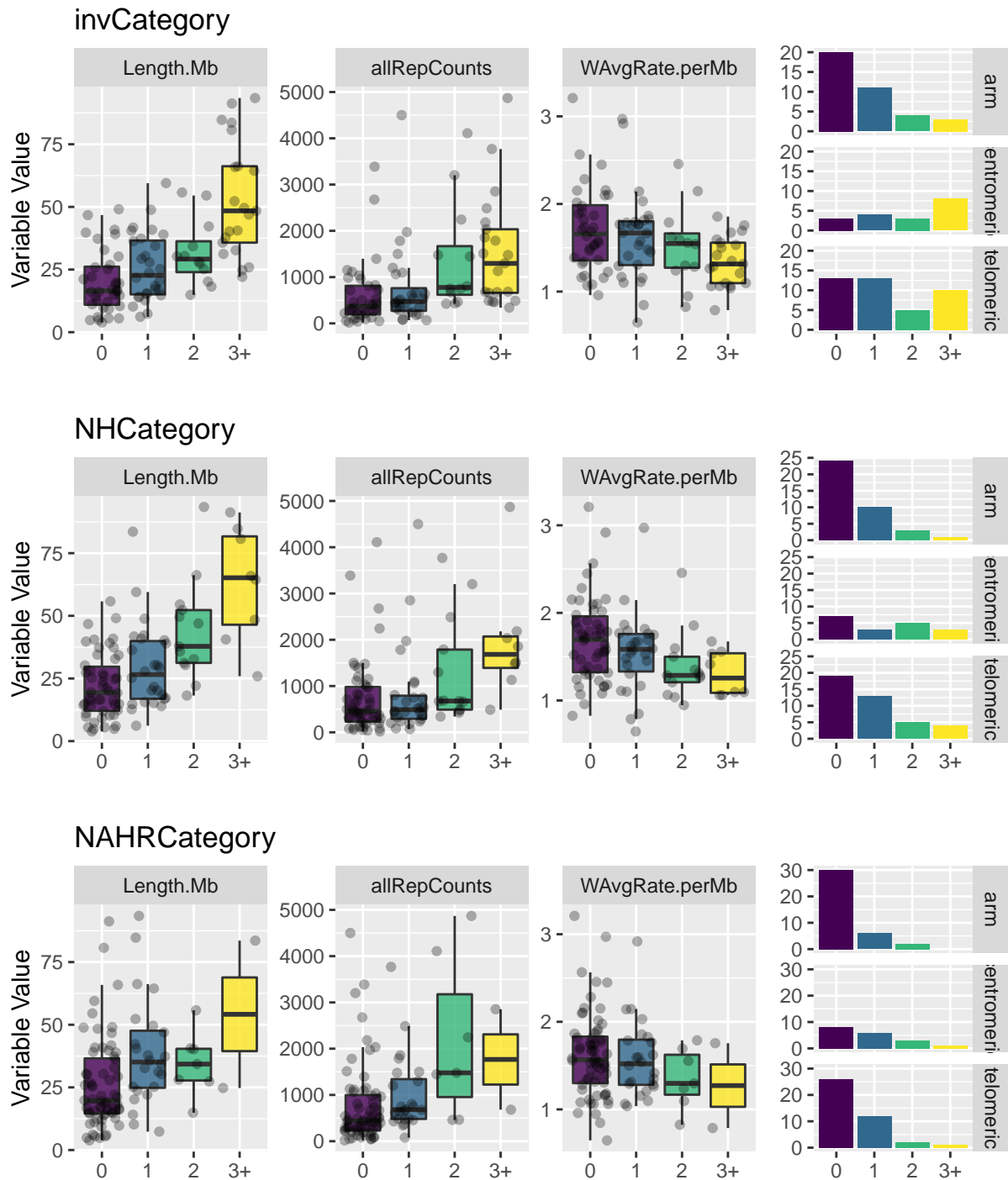
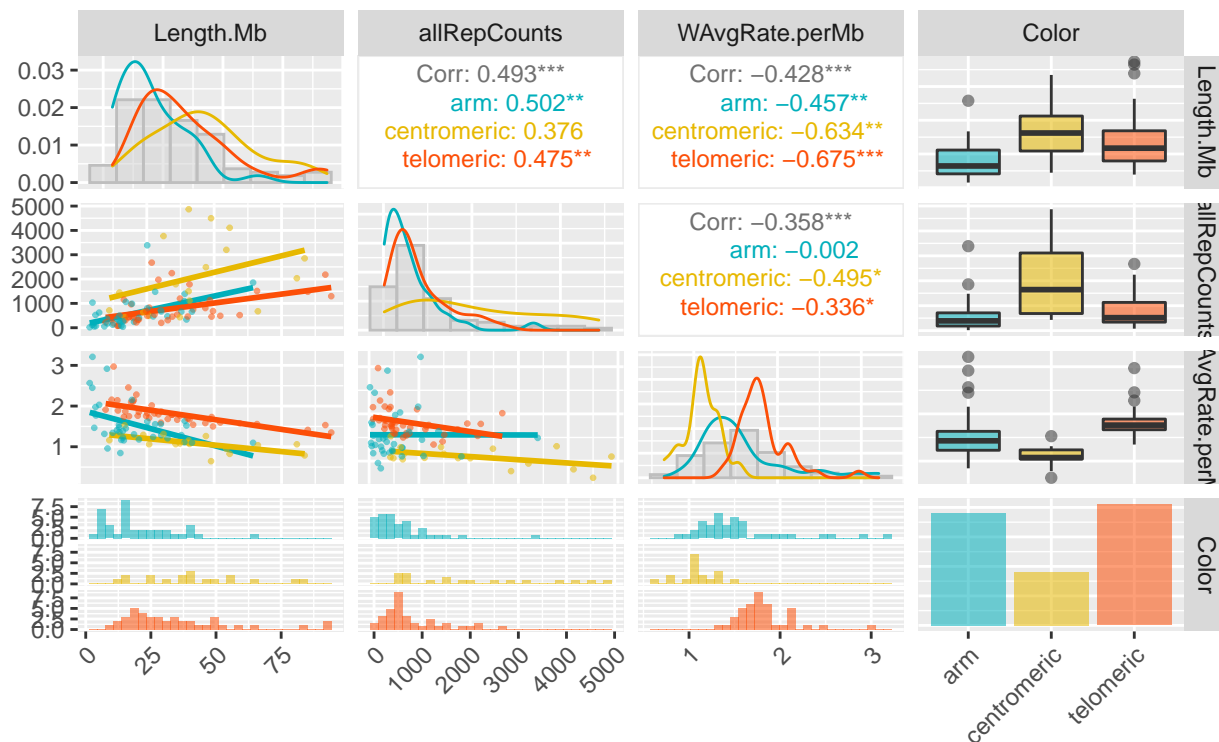


Figure 4: Potential effect of independent variables on the different types of inversions.

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

### Pearson correlation



### Spearman correlation

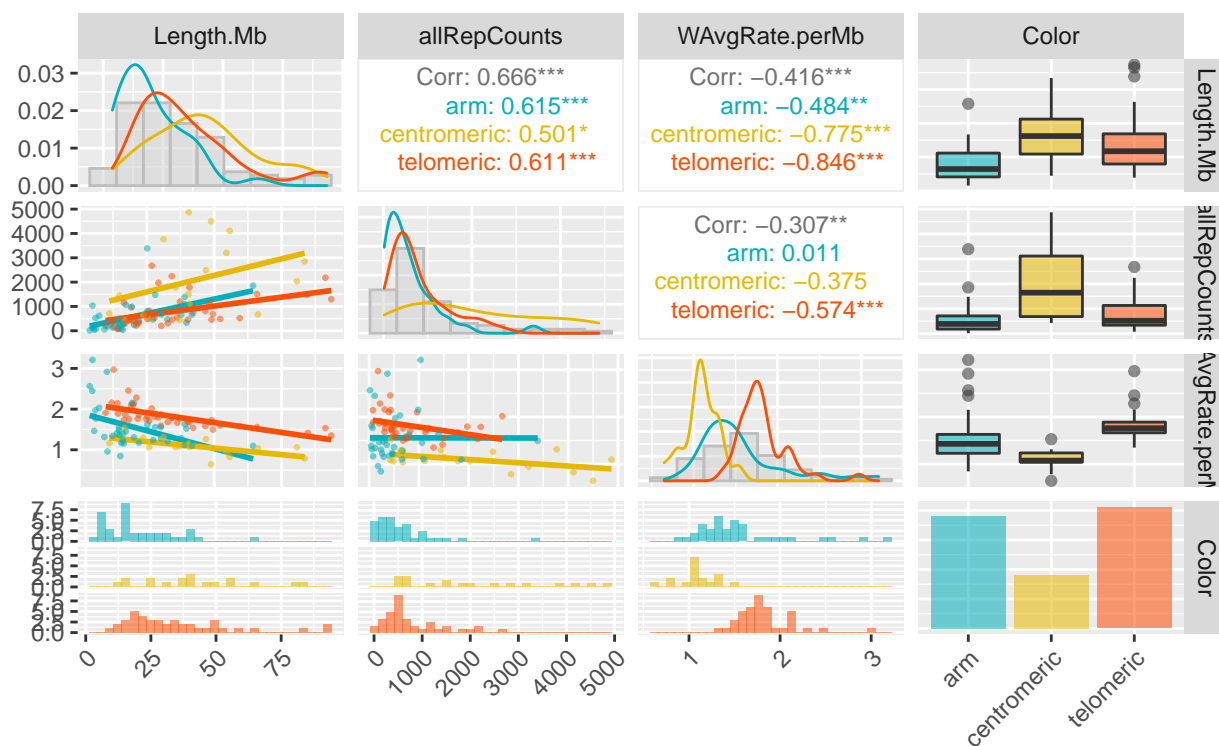


Figure 5: Correlations between variables.

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 corresponding linear model to further check the multi-collinearity.

```
##              GVIF Df GVIF^(1/(2*Df))
## Length.Mb      1.920456 1      1.385805
## allRepCounts   1.698738 1      1.303356
## Color          2.420771 2      1.247350
## WAvgRate.perMb 1.950086 1      1.396455
```

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.      : 3.847      Min.      :-1.3220      Min.      : 18.0      Min.      :-0.9148
## 1st Qu.:15.573      1st Qu.: -0.7355      1st Qu.: 316.0      1st Qu.: -0.6204
## Median :25.893      Median :-0.2193      Median : 536.0      Median :-0.4030
## Mean      :30.277      Mean      : 0.0000      Mean      : 943.8      Mean      : 0.0000
## 3rd Qu.:40.233      3rd Qu.: 0.4980      3rd Qu.:1152.0      3rd Qu.: 0.2057
## Max.      :93.438      Max.      : 3.1592      Max.      :4870.0      Max.      : 3.8798
## WAvgRate.perMb WAvgRate.perMb.Scaled
## Min.      :0.646      Min.      :-2.00474
## 1st Qu.:1.262      1st Qu.: -0.68458
## Median :1.554      Median :-0.05988
## Mean      :1.582      Mean      : 0.00000
## 3rd Qu.:1.796      3rd Qu.: 0.45718
## Max.      :3.209      Max.      : 3.48441
```

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      7.411e-02  0.0193467   3.83059
## allRepCounts   -3.669e-05  0.0003552  -0.10332
## Colorcentromeric 7.864e-01  0.7117957   1.10478
## Colortelomeric  2.359e-02  0.5393000   0.04374
## WAvgRate.perMb  2.417e-01  0.6036508   0.40037
##
## Intercepts:
##      Value  Std. Error t value
## 0|1  1.8540  1.1730    1.5806
## 1|2  3.4471  1.2028    2.8659
## 2|3+ 4.3822  1.2410    3.5313
##
## Residual Deviance: 211.3401
## AIC: 227.3401
```

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

```
##              Value  Std. Error  t value  p value
## Length.Mb      7.410938e-02  0.0193467035   3.83059473 0.00012783
## allRepCounts   -3.669477e-05  0.0003551589  -0.10331929 0.91770957
## Colorcentromeric 7.863778e-01  0.7117957065   1.10478014 0.26925487
## Colortelomeric  2.358897e-02  0.5392999748   0.04373998 0.96511167
## WAvgRate.perMb  2.416861e-01  0.6036507746   0.40037405 0.68888104
## 0|1            1.853987e+00  1.1729608003   1.58060468 0.11396845
## 1|2            3.447097e+00  1.2027848369   2.86592967 0.00415787
## 2|3+           4.382239e+00  1.2409775144   3.53128007 0.00041355
```

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 likelihood"
##              2.5 %      97.5 %
## Length.Mb      0.0404125820 0.1129061423
## allRepCounts   -0.0005668048 0.0004889206
## Colorcentromeric -0.5913307600 2.1808330363
## Colortelomeric  -1.0317611562 1.0940733802
## WAvgRate.perMb  -1.0049122656 1.4044052997
##
## [1] "Assuming a normal distribtuion"
##              2.5 %      97.5 %
## Length.Mb      0.0361905383 0.112028222
## allRepCounts   -0.0007327935 0.000659404
## Colorcentromeric -0.6087161888 2.181471709
## Colortelomeric  -1.0334195593 1.080597496
```



```
## WAvgRate.perMb -0.9414476745 1.424819880
```

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      1.0769246 1.0412403 1.119527
## allRepCounts    0.9999633 0.9994334 1.000489
## Colorcentromeric 2.1954296 0.5535901 8.853679
## Colortelomeric  1.0238694 0.3563788 2.986414
## WAvgRate.perMb  1.2733944 0.3660768 4.073104
```

Example of interpretation: “For 1 unit increase in Length.Mb, a window is 1.0769246 times more likely to increase in inversion amount category.”

### Proportional odds assessment

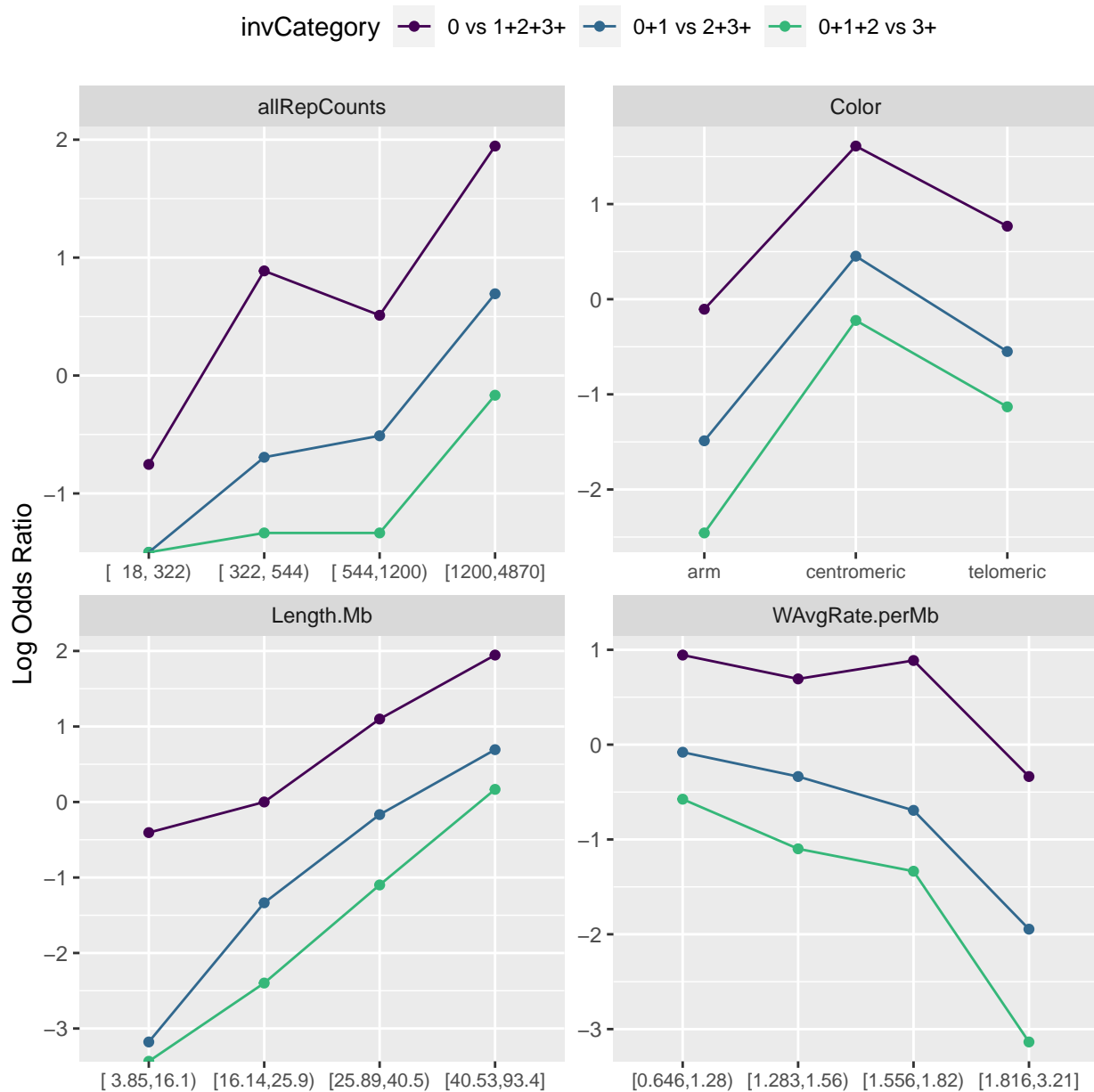
Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe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

We test the parallel regression assumption with a Brant test:

```
## -----
## Test for      X2  df  probability
## -----
## Omnibus           3.19   10  0.98
## Length.Mb         1.13    2  0.57
## allRepCounts      1.51    2  0.47
## Colorcentromeric  0.29    2  0.87
## Colortelomeric     0.12    2  0.94
## WAvgRate.perMb     0.03    2  0.98
## -----
##
## 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.

### Proportional odds visual test



## Predicted probabilities

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

### Probability of inversion level (invCategory) for multiple scenarios

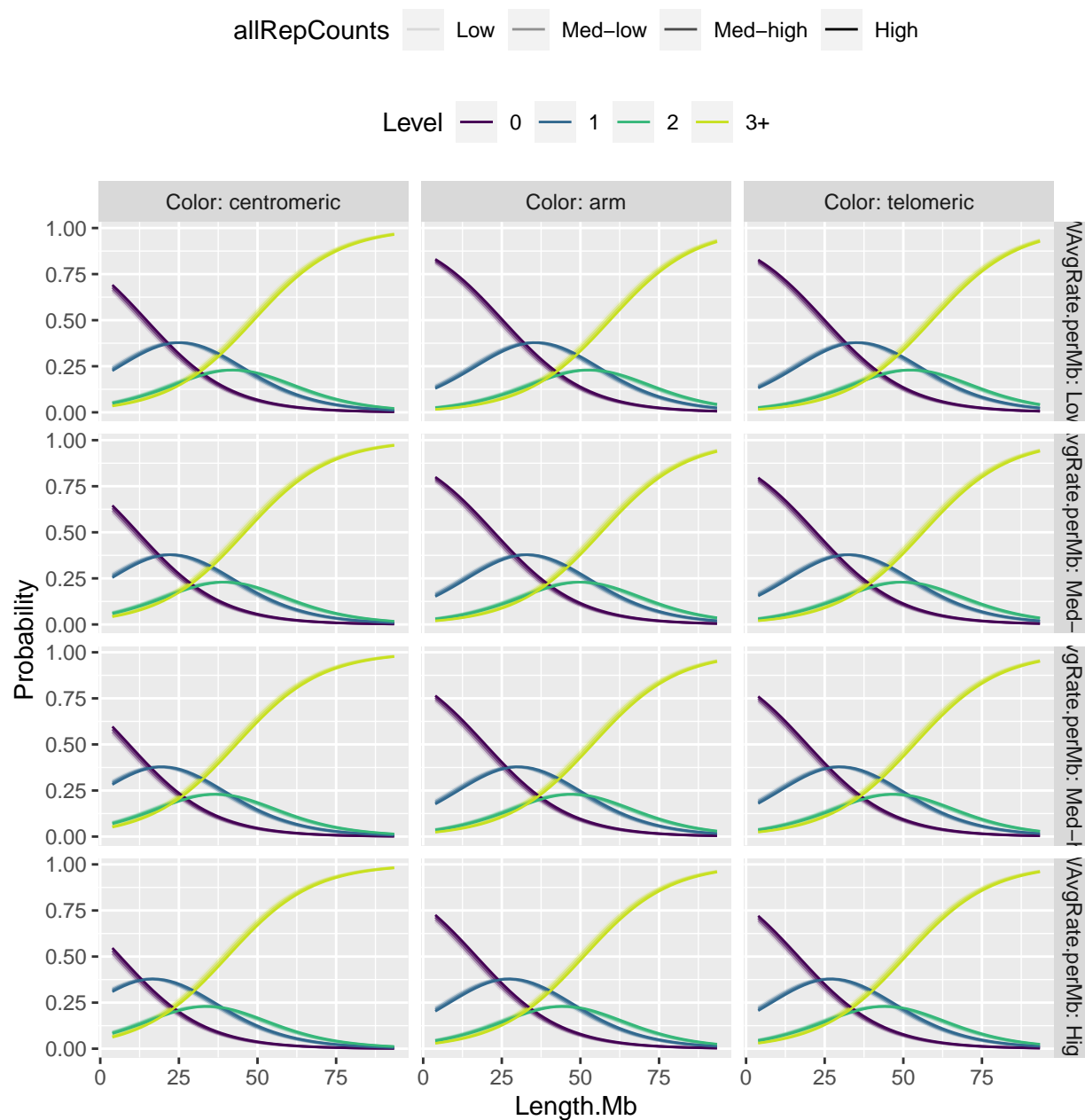


Figure 6: Probability of having 0 to >3 inversions depending on multiple independent variables

## Total inversions (NHCategory)

### Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##              Value Std. Error   t value
## Length.Mb      6.043e-02  0.0129009   4.6841477
## allRepCounts    2.104e-05  0.0003213   0.0654864
## Colorcentromeric 2.703e-04  0.7048769   0.0003834
## Colortelomeric   5.141e-02  0.5014151   0.1025277
## WAvgRate.perMb  -3.508e-01  0.1756986  -1.9967548
##
## Intercepts:
##      Value  Std. Error t value
## 0|1   1.2551   0.1641    7.6465
## 1|2   2.8774   0.3396    8.4721
## 2|3+  4.4314   0.5268    8.4125
##
## Residual Deviance: 191.6944
## AIC: 207.6944
```

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

```
##              Value  Std. Error   t value   p value
## Length.Mb      6.042994e-02  0.0129009475   4.6841476552  0.00000281
## allRepCounts    2.103837e-05  0.0003212632   0.0654863854  0.94778675
## Colorcentromeric 2.702663e-04  0.7048768850   0.0003834234  0.99969407
## Colortelomeric   5.140894e-02  0.5014151342   0.1025277039  0.91833782
## WAvgRate.perMb  -3.508270e-01  0.1756985879  -1.9967547986  0.04585183
## 0|1             1.255055e+00  0.1641335807   7.6465434576  0.00000000
## 1|2             2.877420e+00  0.3396348904   8.4720985492  0.00000000
## 2|3+            4.431431e+00  0.5267700816   8.4124579018  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 likelihood"
##              2.5 %    97.5 %
## Length.Mb      0.0309808178  0.09217756
## allRepCounts   -0.0005146866  0.00054427
## Colorcentromeric -1.4174902293  1.38501111
## Colortelomeric  -1.0770779707  1.24817135
## WAvgRate.perMb           NA           NA
## [1] "Assuming a normal distribtuion"
##              2.5 %    97.5 %
## Length.Mb      0.035144551  0.0857153354
## allRepCounts   -0.000608626  0.0006507027
## Colorcentromeric -1.381263042  1.3818035744
## Colortelomeric  -0.931346662  1.0341645467
```

```
## WAvgRate.perMb -0.695189903 -0.0064640941
```

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      1.0622932 1.0314657 1.096560
## allRepCounts    1.0000210 0.9994854 1.000544
## Colorcentromeric 1.0002703 0.2423214 3.994870
## Colortelomeric  1.0527533 0.3405893 3.483966
## WAvgRate.perMb   0.7041056      NA      NA
```

Example of interpretation: “For 1 unit increase in Length.Mb, a window is 1.0622932 times more likely to increase in inversion amount category.”

### Proportional odds assessment

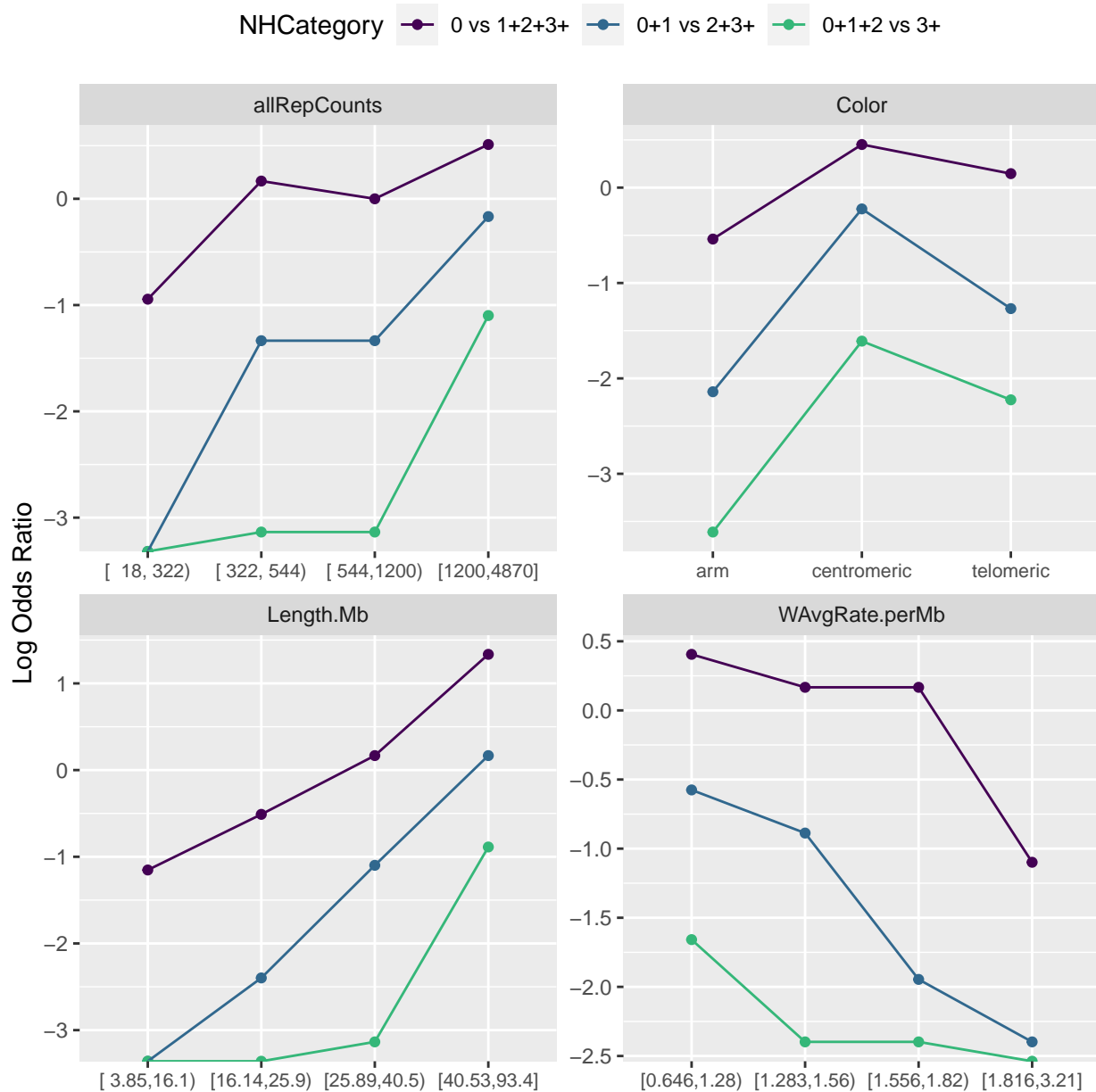
Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe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

We test the parallel regression assumption with a Brant test:

```
## -----
## Test for      X2  df  probability
## -----
## Omnibus           6.29   10  0.79
## Length.Mb          1.22    2  0.54
## allRepCounts       1.53    2  0.46
## Colorcentromeric  1.54    2  0.46
## Colortelomeric      0.94    2  0.63
## WAvgRate.perMb      2.22    2  0.33
## -----
##
## 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.

### Proportional odds visual test



## Predicted probabilities

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

### Probability of inversion level (NHCategory) for multiple scenarios

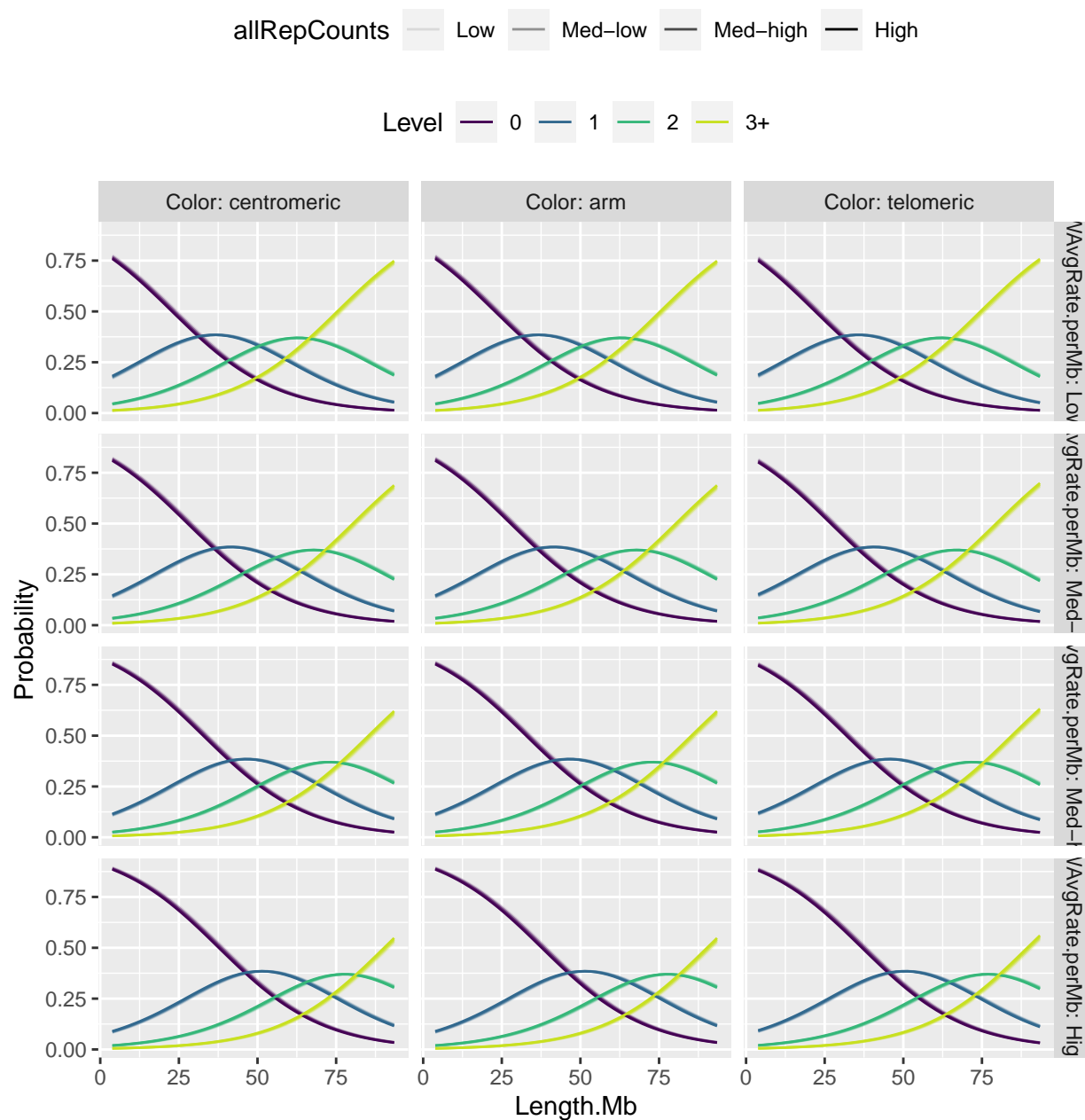


Figure 7: Probability of having 0 to >3 inversions depending on multiple independent variables

## Total inversions (NAHRCategory)

### Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##              Value Std. Error t value
## Length.Mb      0.0222537  0.011897  1.8705
## allRepCounts    0.0003363  0.000343  0.9805
## Colorcentromeric 0.9269353  0.676147  1.3709
## Colortelomeric  0.2712057  0.577790  0.4694
## WAvgRate.perMb  0.3550106  0.200470  1.7709
##
## Intercepts:
##      Value  Std. Error t value
## 0|1  2.5669  0.1532   16.7552
## 1|2  4.3722  0.3855   11.3409
## 2|3+ 6.0583  0.7650    7.9193
##
## Residual Deviance: 158.6448
## AIC: 174.6448
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0222536609	0.0118972461	1.8704884	0.06141602
allRepCounts	0.0003363444	0.0003430403	0.9804809	0.32684882
Colorcentromeric	0.9269352848	0.6761468838	1.3709082	0.17040359
Colortelomeric	0.2712057066	0.5777896788	0.4693848	0.63879459
WAvgRate.perMb	0.3550105530	0.2004696337	1.7708944	0.07657826
0 1	2.5669344400	0.1532024572	16.7551780	0.00000000
1 2	4.3722167836	0.3855248024	11.3409481	0.00000000
2 3+	6.0582756952	0.7650038797	7.9192745	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 likelihood"
##              2.5 %      97.5 %
## Length.Mb      -0.0059118574  0.0512096260
## allRepCounts    -0.0001812889  0.0008459939
## Colorcentromeric -0.5019966633  2.3687009731
## Colortelomeric  -0.9035429149  1.5079734433
## WAvgRate.perMb      NA              NA
## [1] "Assuming a normal distribtuion"
##              2.5 %      97.5 %
## Length.Mb      -0.0010645129  0.045571835
## allRepCounts    -0.0003360022  0.001008691
## Colorcentromeric -0.3982882557  2.252158825
## Colortelomeric  -0.8612412546  1.403652668
```



```
## WAvgRate.perMb -0.0379027091 0.747923815
```

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      1.022503 0.9941056  1.052544
## allRepCounts    1.000336 0.9998187  1.000846
## Colorcentromeric 2.526754 0.6053208 10.683505
## Colortelomeric  1.311545 0.4051318  4.517566
## WAvgRate.perMb   1.426196          NA          NA
```

Example of interpretation: “For 1 unit increase in Length.Mb, a window is 1.0225031 times more likely to increase in inversion amount category.”

### Proportional odds assessment

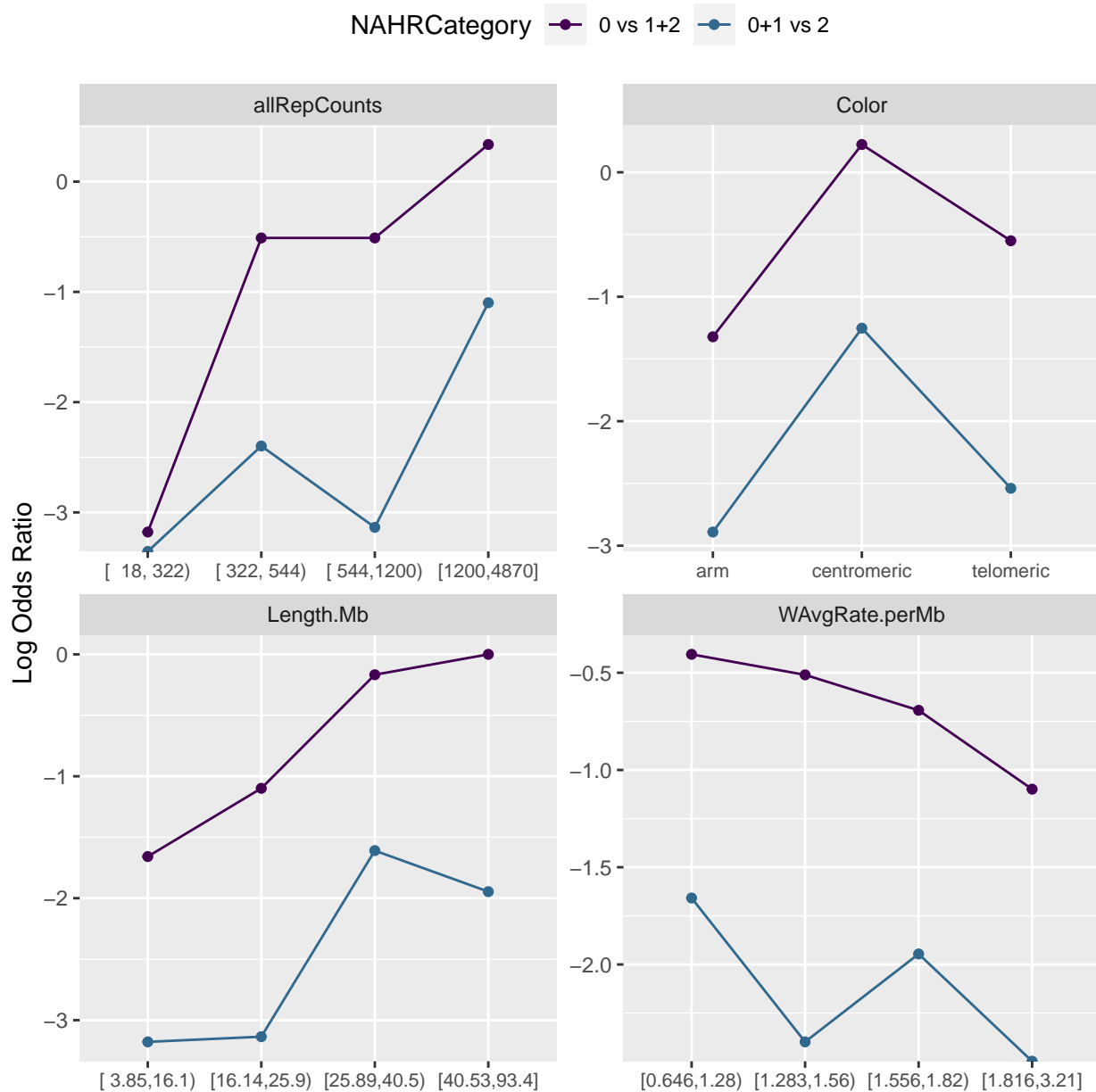
Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe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

We test the parallel regression assumption with a Brant test:

```
## -----
## Test for      X2  df  probability
## -----
## Omnibus           9.28   10  0.51
## Length.Mb         2.82    2  0.24
## allRepCounts      5.53    2  0.06
## Colorcentromeric 0.6 2    0.74
## Colortelomeric     0.16    2  0.92
## WAvgRate.perMb     1.22    2  0.54
## -----
##
## 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.

### Proportional odds visual test



## Predicted probabilities

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

### Probability of inversion level (NAHRCategory) for multiple scenarios

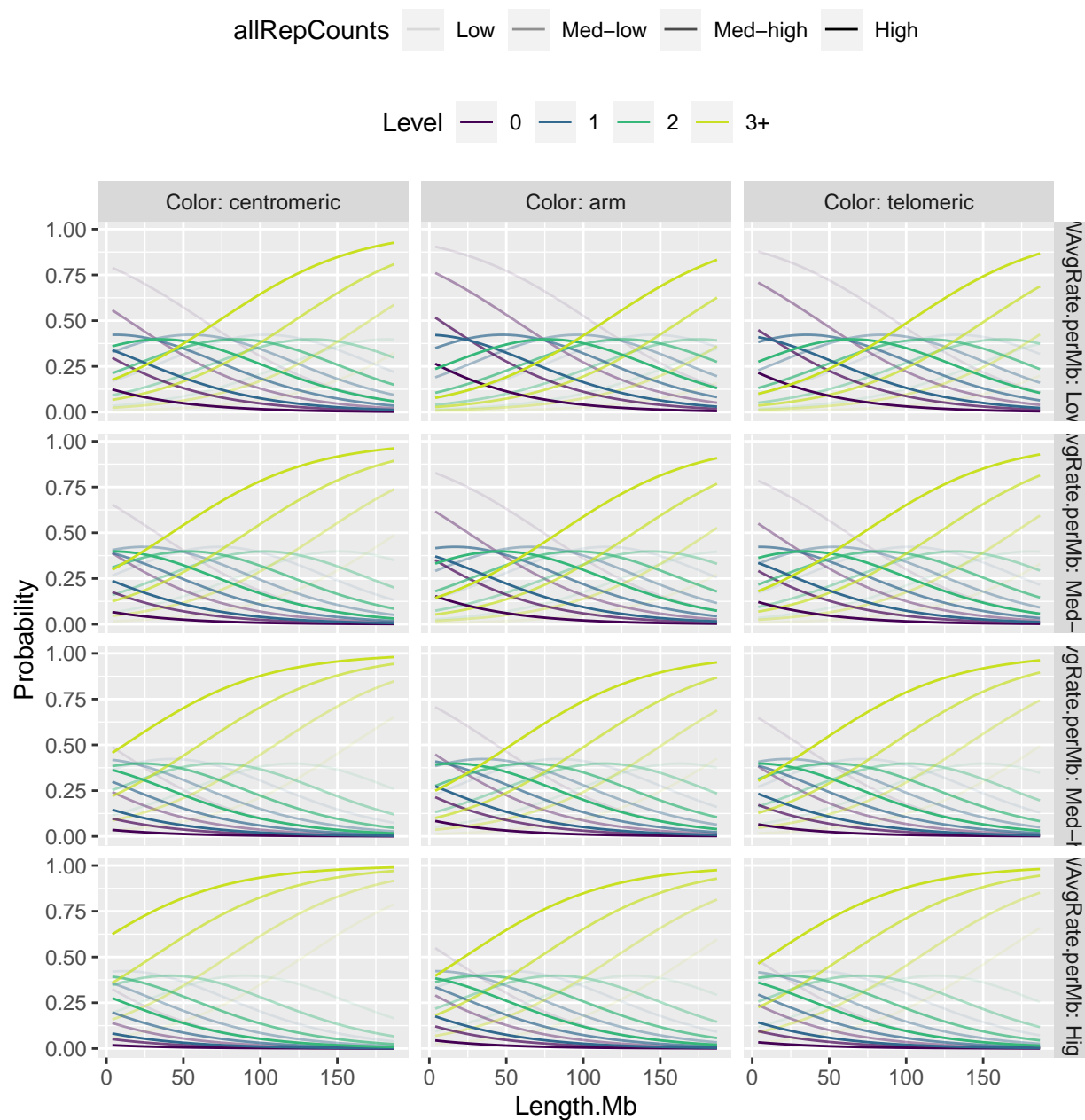


Figure 8: Probability of having 0 to >3 inversions depending on multiple independent variables