

Ordinal logistic model on large, classified windows data from Spence

Ruth Gómez Graciani

Prepare the data

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

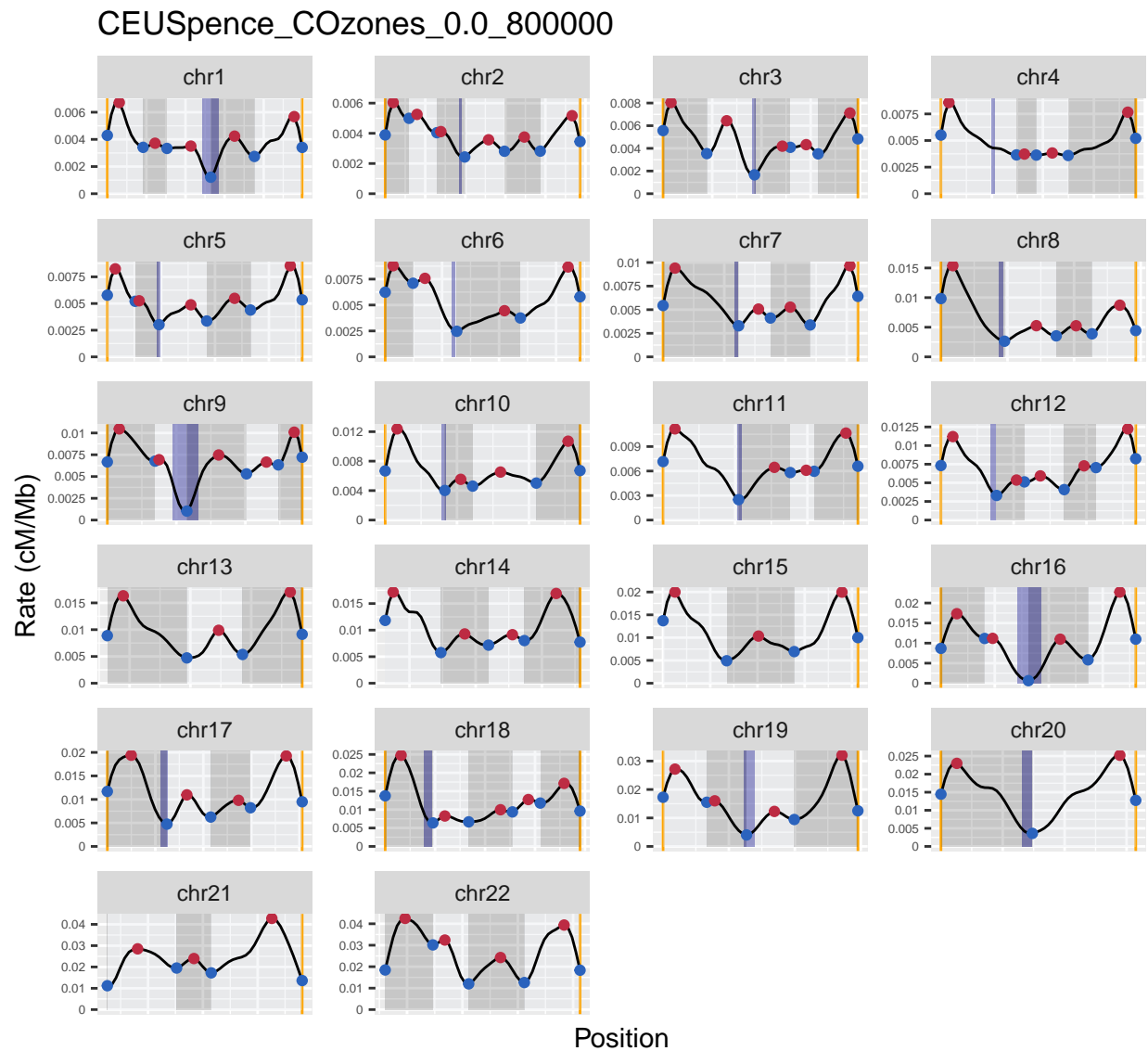


Figure 1: Crossover zones; centromeres in blue, workspace limits in orange.

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 the limits of the centromere by calculating the midpoint between the flanking maxima and the local minimum near the centromere (or the centromere itself if there is no local minimum). Actual centromeres will be excluded from centromeric regions to avoid biases, specially in the Spence data. These categories will be represented as the “Color” variable in the statistical analysis.

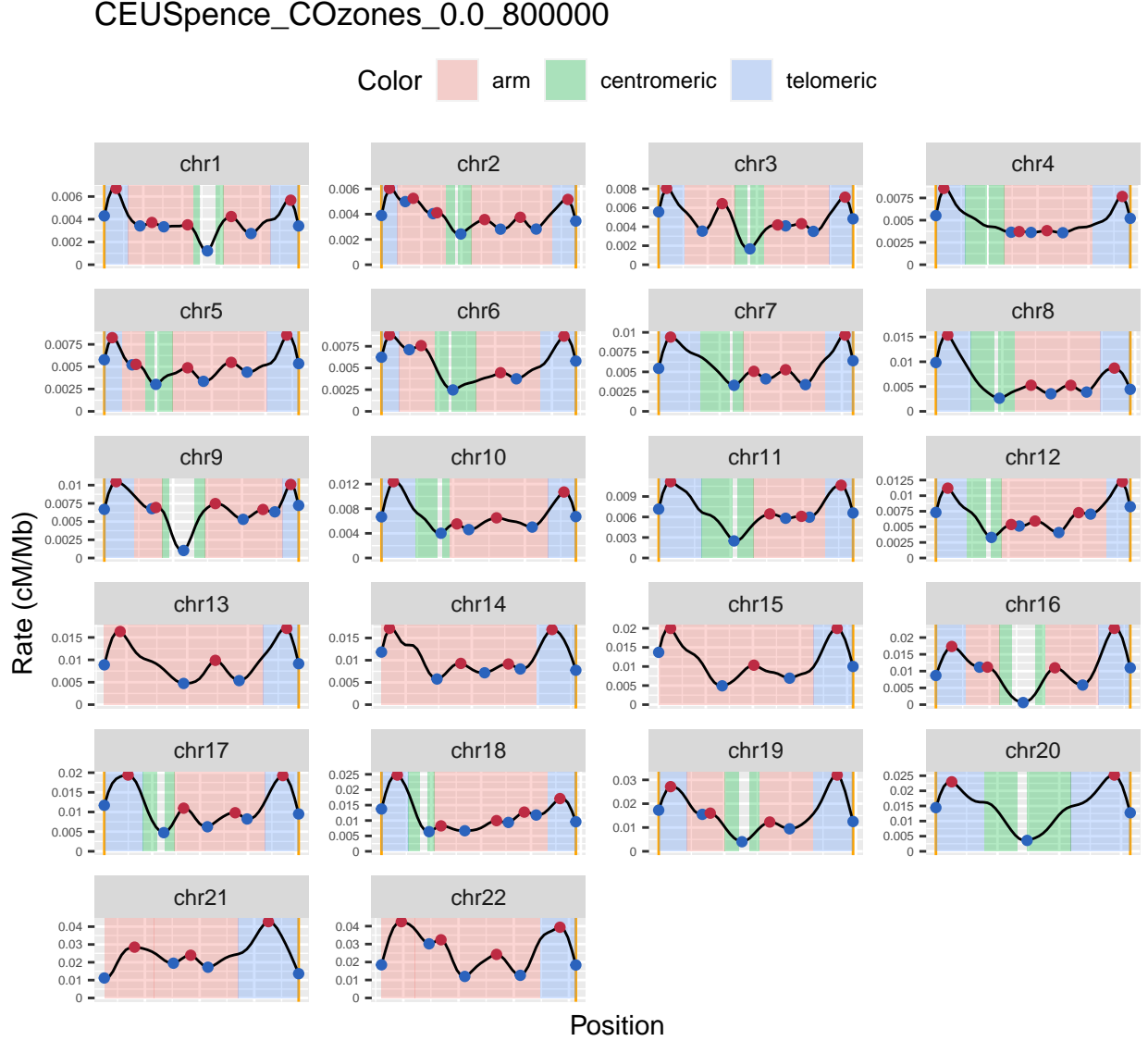


Figure 2: Color-coded windows for telomeric, centromeric and arm categories.

Descriptive statistics

Raw data:

##	Chromosome	Start	End	Color	invCenters	NHCenters	NAHRCenters
## 1	chr10	60683	23750219	telomeric	3	2	1
## 2	chr10	23750219	39146059	centromeric	1	0	1
## 3	chr10	47478464	116172416	arm	4	4	0
## 4	chr10	116172416	135524372	telomeric	1	1	0
## 5	chr10	42369508	47478464	centromeric	0	0	0
## 6	chr11	87267	29960849	telomeric	2	1	1

##	Length.Mb	allRepCounts	WAvgRate.perMb
## 1	23.689536	428	0.014480534
## 2	15.395840	880	0.007907887
## 3	68.693953	1542	0.007285414
## 4	19.351956	266	0.014043434
## 5	5.108956	954	0.006810256
## 6	29.873582	920	0.011957306

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

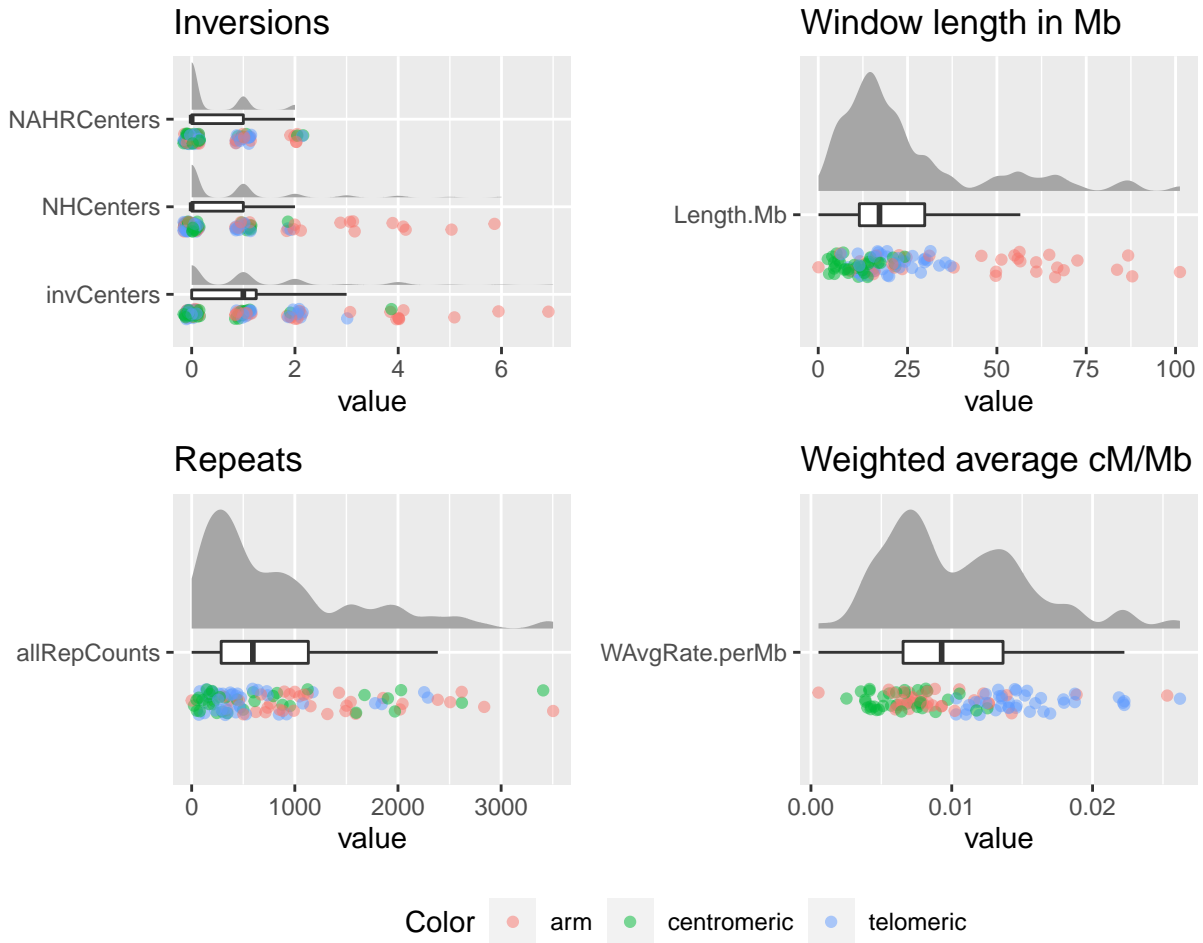


Figure 3: Distribution of variables.

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         49         65          78
## 2         1         32         27          22
## 3         2         15          7           8
## 4         3          2          4          NA
## 5         4          7          3          NA
## 6         5          1          1          NA
## 7         6          1          1          NA
## 8         7          1          NA          NA

## [1] "New counts"

##   CountGroups invCategory NHCATEGORY NAHRCATEGORY
## 1         0         49         65          78
## 2         1         32         27          22
## 3         2         15          7           8
## 4         3+         12          9          NA
```

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

Differences in each chromosomal variable between inversion count groups

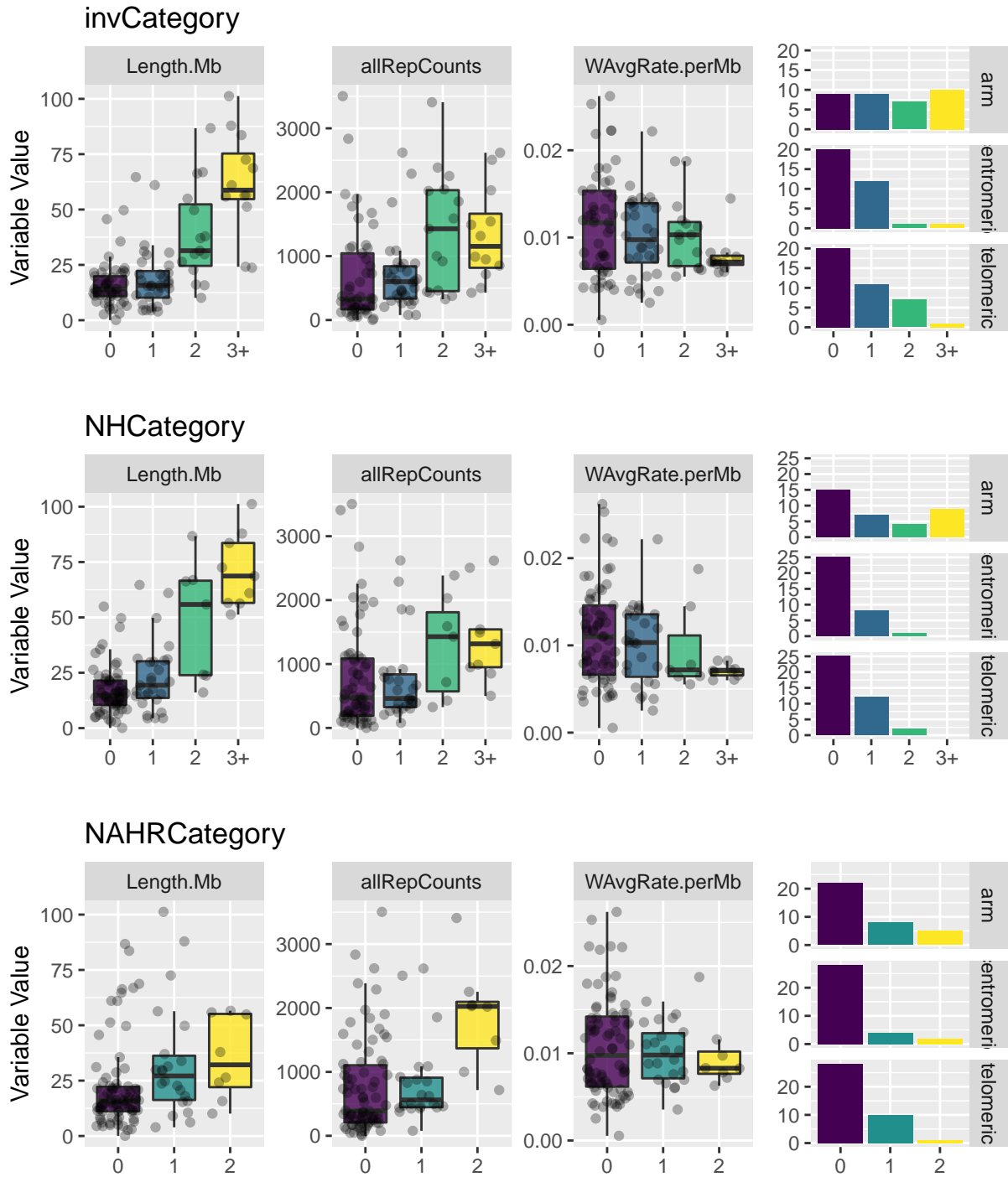
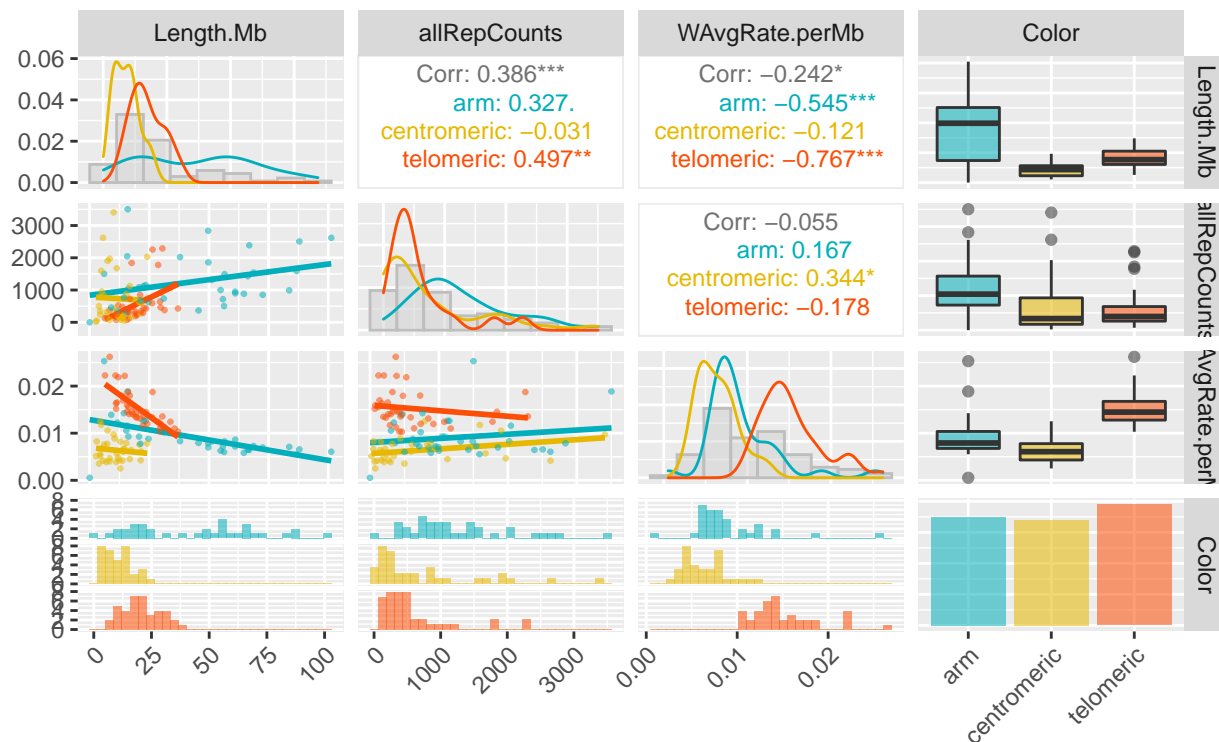


Figure 4: Potential effect of independent variables on the different types of inversions.

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

Pearson correlation



Spearman correlation

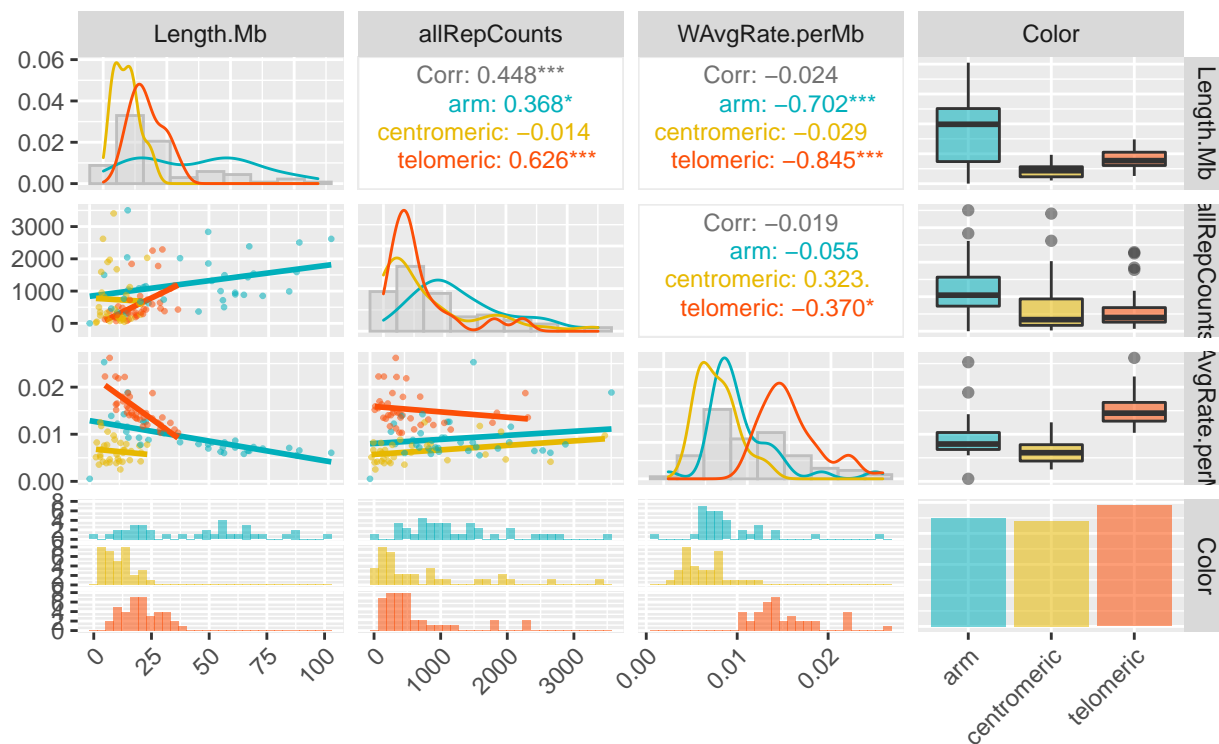


Figure 5: Correlations between variables.

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      2.554186  1      1.598182
## allRepCounts    1.329840  1      1.153187
## Color           5.084803  2      1.501649
## WAvgRate.perMb  3.040313  1      1.743649
```

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.   : 0.07717  Min.   : -1.1754  Min.   : 0.0  Min.   : -1.0881
## 1st Qu.: 11.49431  1st Qu.: -0.6367  1st Qu.: 285.5  1st Qu.: -0.7244
## Median : 17.14259  Median : -0.3701  Median : 593.0  Median : -0.3326
## Mean   : 24.98653  Mean   : 0.0000  Mean   : 854.1  Mean   : 0.0000
## 3rd Qu.: 29.80087  3rd Qu.: 0.2272  3rd Qu.: 1131.0  3rd Qu.: 0.3528
## Max.   :101.22393  Max.   : 3.5975  Max.   :3504.0  Max.   : 3.3761
## WAvgRate.perMb  WAvgRate.perMb.Scaled
## Min.   :0.0005371  Min.   : -1.9171
## 1st Qu.:0.0065417  1st Qu.: -0.7589
## Median :0.0092825  Median : -0.2302
## Mean   :0.0104759  Mean   : 0.0000
## 3rd Qu.:0.0136431  3rd Qu.: 0.6109
## Max.   :0.0262073  Max.   : 3.0344
```

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.Scaled    1.3686    0.3564  3.8402
## allRepCounts.Scaled  0.2787    0.2275  1.2249
## Colorcentromeric   -0.1699    0.7383 -0.2301
## Colortelomeric      0.1798    0.5987  0.3003
## WAvgRate.perMb.Scaled -0.2823    0.3588 -0.7867
##
## Intercepts:
##      Value Std. Error t value
## 0|1  -0.5144  0.4120   -1.2484
## 1|2   1.3368  0.4425    3.0208
## 2|3+  3.0134  0.5534    5.4453
##
## Residual Deviance: 214.4464
## AIC: 230.4464
```

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

```
##              Value Std. Error  t value  p value
## Length.Mb.Scaled    1.3686332  0.3563923  3.8402429 0.00012291
## allRepCounts.Scaled  0.2786941  0.2275226  1.2249075 0.22061006
## Colorcentromeric   -0.1699186  0.7383279 -0.2301398 0.81798316
## Colortelomeric      0.1798086  0.5986922  0.3003357 0.76392113
## WAvgRate.perMb.Scaled -0.2822953  0.3588499 -0.7866668 0.43147692
## 0|1                 -0.5144036  0.4120369 -1.2484406 0.21186976
## 1|2                  1.3367709  0.4425214  3.0208051 0.00252104
## 2|3+                 3.0134215  0.5534031  5.4452562 0.00000005
```

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.Scaled    0.7026388 2.1136993
## allRepCounts.Scaled -0.1760119 0.7232369
## Colorcentromeric   -1.6081776 1.3207391
## Colortelomeric     -0.9794620 1.3810275
## WAvgRate.perMb.Scaled -1.0125855 0.4121100
## [1] "Assuming a normal distribtuion"
##              2.5 %    97.5 %
## Length.Mb.Scaled    0.6701170 2.0671494
## allRepCounts.Scaled -0.1672419 0.7246301
## Colorcentromeric   -1.6170146 1.2771774
## Colortelomeric     -0.9936065 1.3532237
```



```
## WAvgRate.perMb.Scaled -0.9856282 0.4210376
```

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.Scaled      3.9299755 2.0190736 8.278811
## allRepCounts.Scaled    1.3214031 0.8386080 2.061094
## Colorcentromeric       0.8437335 0.2002522 3.746189
## Colortelomeric         1.1969883 0.3755131 3.978988
## WAvgRate.perMb.Scaled  0.7540510 0.3632785 1.510001
```

Example of interpretation: “For 1 unit increase in Length.Mb.Scaled, a window is 3.9299755 times more likely to increase in inversion amount category.”

Proportional odds assessment

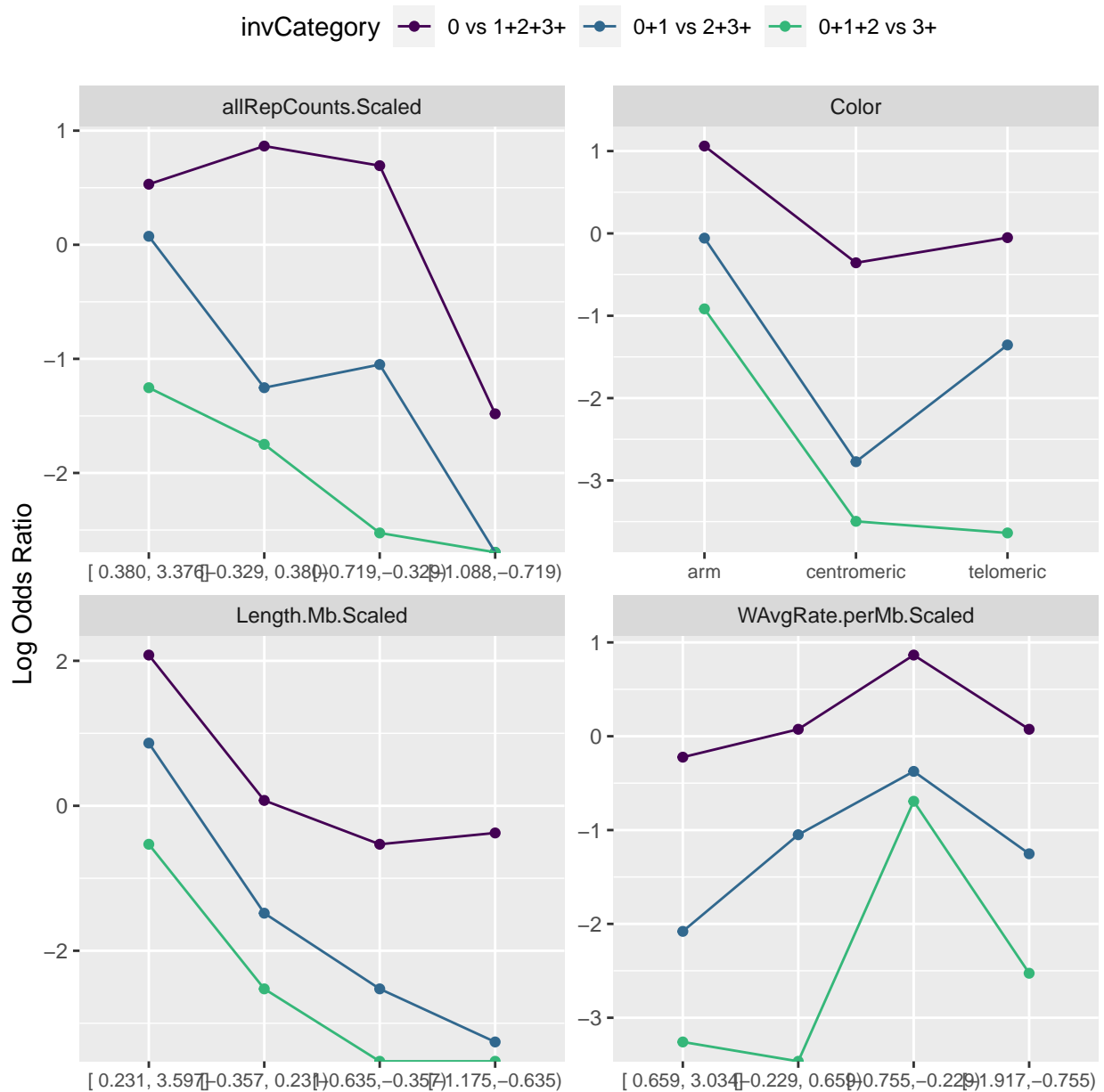
Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe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

We test the parallel regression assumption with a Brant test:

```
## -----
## Test for      X2  df  probability
## -----
## Omnibus           13.41   10   0.2
## Length.Mb.Scaled  3.59    2   0.17
## allRepCounts.Scaled 4.68    2   0.1
## Colorcentromeric  1.32    2   0.52
## Colortelomeric     3.18    2   0.2
## WAvgRate.perMb.Scaled 2.29    2   0.32
## -----
##
## 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

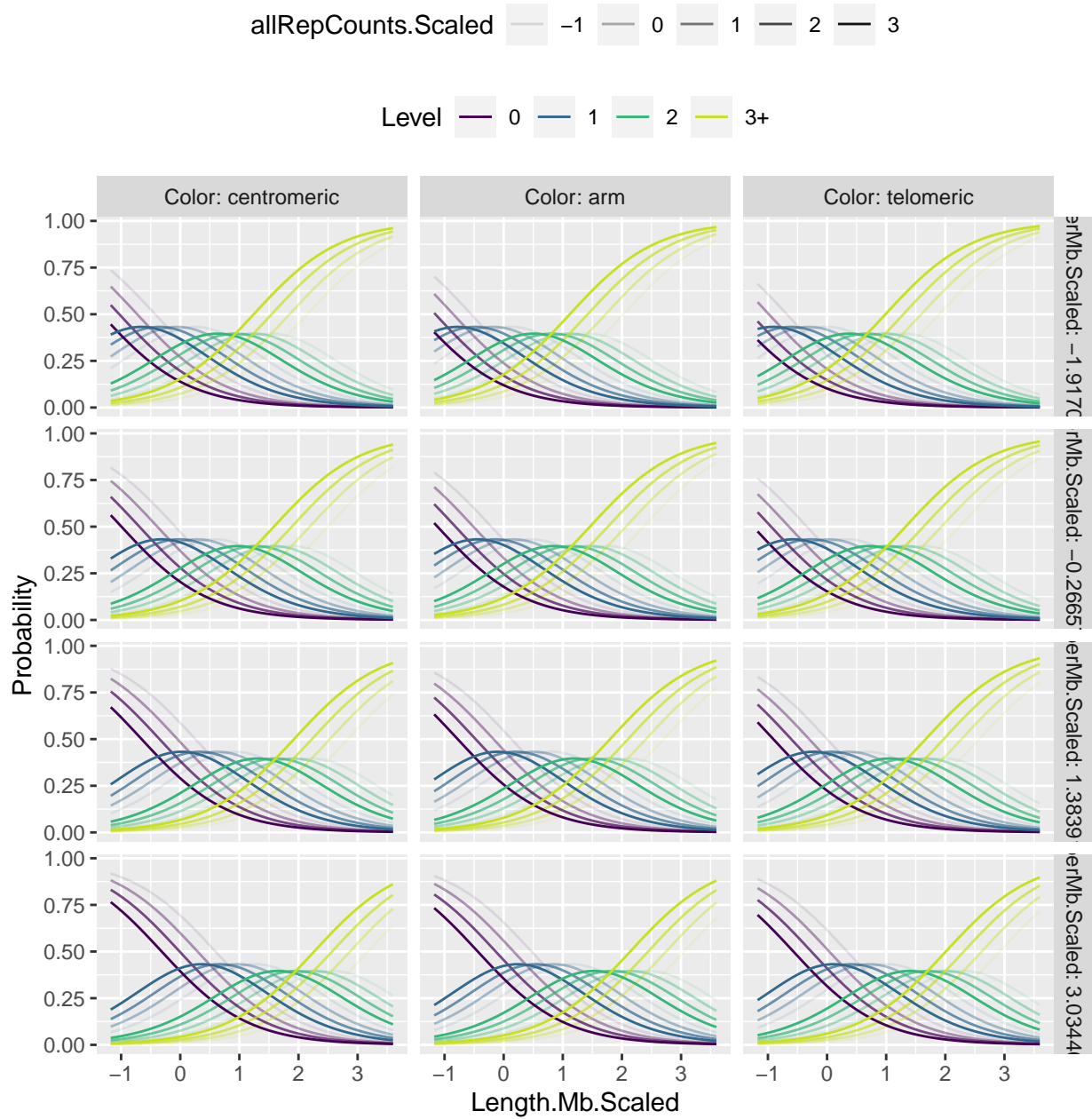


Figure 6: Probability of having 0 to >3 inversions depending on multiple independent variables

NH inversions (NHCategory)

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##              Value Std. Error  t value
## Length.Mb.Scaled    1.82483    0.4175  4.37095
## allRepCounts.Scaled -0.19475    0.2641 -0.73747
## Colorcentromeric     0.59203    0.8979  0.65932
## Colortelomeric       0.14965    0.6839  0.21882
## WAvgRate.perMb.Scaled -0.01151    0.4382 -0.02626
##
## Intercepts:
##      Value  Std. Error t value
## 0|1  0.5062  0.4906    1.0319
## 1|2  2.6307  0.5808    4.5290
## 2|3+ 3.8728  0.6967    5.5590
##
## Residual Deviance: 172.2763
## AIC: 188.2763
```

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

```
##              Value Std. Error  t value  p value
## Length.Mb.Scaled    1.82482830  0.4174902  4.37094864 0.00001237
## allRepCounts.Scaled -0.19475006  0.2640801 -0.73746579 0.46083914
## Colorcentromeric     0.59202688  0.8979415  0.65931568 0.50969308
## Colortelomeric       0.14965463  0.6839101  0.21882209 0.82678864
## WAvgRate.perMb.Scaled -0.01150714  0.4381724 -0.02626168 0.97904862
## 0|1                  0.50620907  0.4905645  1.03189094 0.30212321
## 1|2                  2.63066232  0.5808437  4.52903687 0.00000593
## 2|3+                 3.87283588  0.6966735  5.55903968 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 likelihood"
##              2.5 %    97.5 %
## Length.Mb.Scaled    1.0634206 2.7085998
## allRepCounts.Scaled -0.7474781 0.2977162
## Colorcentromeric    -1.0988549 2.4425947
## Colortelomeric      -1.1555540 1.5476395
## WAvgRate.perMb.Scaled -0.9034309 0.8205340
## [1] "Assuming a normal distribtuion"
##              2.5 %    97.5 %
## Length.Mb.Scaled    1.0065625 2.6430941
## allRepCounts.Scaled -0.7123376 0.3228375
## Colorcentromeric    -1.1679060 2.3519598
## Colortelomeric      -1.1907845 1.4900938
```

```
## WAvgRate.perMb.Scaled -0.8703093 0.8472950
```

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.Scaled      6.2017301 2.8962611 15.008247
## allRepCounts.Scaled    0.8230403 0.4735593  1.346780
## Colorcentromeric      1.8076486 0.3332525 11.502848
## Colortelomeric        1.1614331 0.3148830  4.700362
## WAvgRate.perMb.Scaled  0.9885588 0.4051771  2.271713
```

Example of interpretation: “For 1 unit increase in Length.Mb.Scaled, a window is 6.2017301 times more likely to increase in inversion amount category.”

Proportional odds assessment

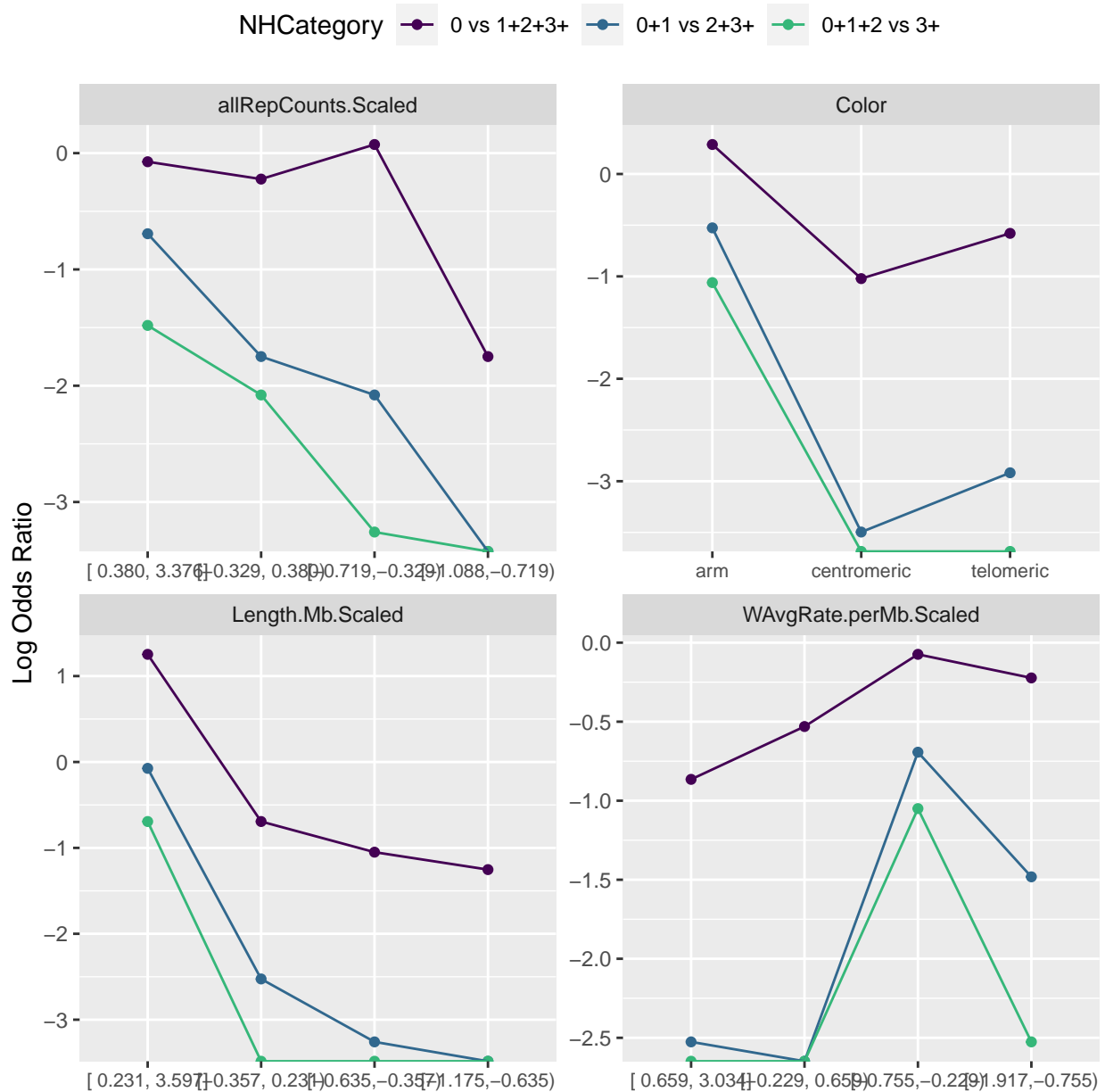
Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe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

We test the parallel regression assumption with a Brant test:

```
## -----
## Test for      X2  df  probability
## -----
## Omnibus           7.42   10  0.69
## Length.Mb.Scaled  5.91    2  0.05
## allRepCounts.Scaled 0.51    2  0.78
## Colorcentromeric   3.31    2  0.19
## Colortelomeric     0.02    2  0.99
## WAvgRate.perMb.Scaled 7.3  2  0.03
## -----
##
## 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

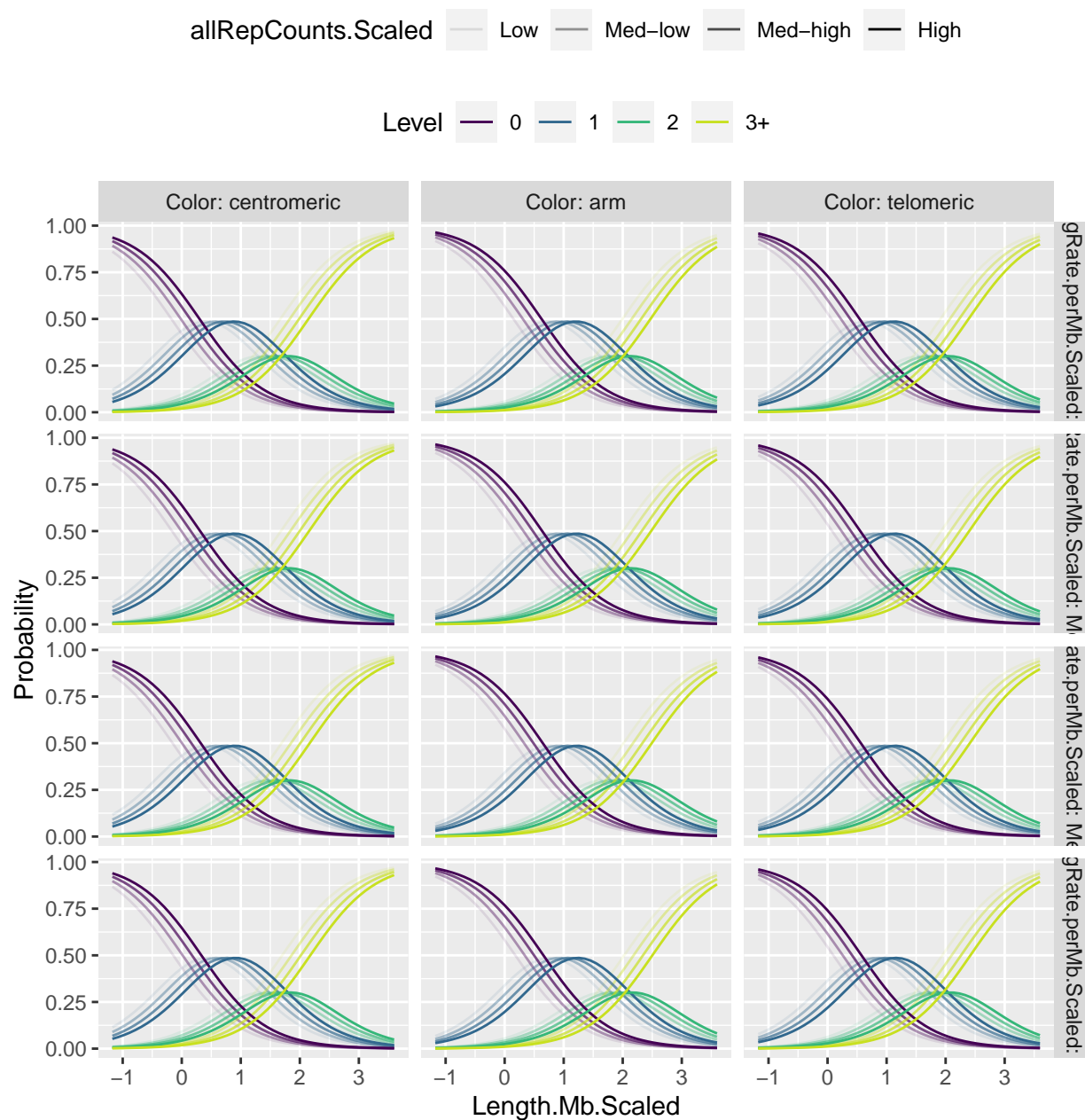


Figure 7: Probabilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions (NAHRCategory)

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##              Value Std. Error t value
## Length.Mb.Scaled    0.03331    0.3218  0.1035
## allRepCounts.Scaled  0.53884    0.2573  2.0943
## Colorcentromeric    -0.89716    0.8721 -1.0287
## Colortelomeric       0.60609    0.6917  0.8762
## WAvgRate.perMb.Scaled -0.55870    0.4290 -1.3024
##
## Intercepts:
##      Value  Std. Error t value
## 0|1  0.9808  0.4676    2.0976
## 1|2  2.6660  0.5615    4.7483
##
## Residual Deviance: 150.8831
## AIC: 164.8831
```

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

```
##              Value Std. Error   t value   p value
## Length.Mb.Scaled    0.03330661  0.3217541  0.1035158 0.91755365
## allRepCounts.Scaled  0.53884499  0.2572866  2.0943377 0.03622991
## Colorcentromeric    -0.89715538  0.8721037 -1.0287256 0.30360864
## Colortelomeric       0.60609197  0.6917230  0.8762062 0.38091796
## WAvgRate.perMb.Scaled -0.55870420  0.4289928 -1.3023626 0.19279247
## 0|1                   0.98078086  0.4675643  2.0976384 0.03593710
## 1|2                   2.66603101  0.5614660  4.7483390 0.00000205
```

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.Scaled   -0.59782693  0.6865963
## allRepCounts.Scaled  0.02708661  1.0492499
## Colorcentromeric    -2.63387408  0.8593236
## Colortelomeric      -0.71706974  2.0175760
## WAvgRate.perMb.Scaled -1.45946125  0.2484297
##
## [1] "Assuming a normal distribtuion"
##              2.5 %    97.5 %
## Length.Mb.Scaled   -0.59731975  0.6639330
## allRepCounts.Scaled  0.03457255  1.0431174
## Colorcentromeric    -2.60644719  0.8121364
## Colortelomeric      -0.74966020  1.9618441
## WAvgRate.perMb.Scaled -1.39951469  0.2821063
```


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.Scaled    1.0338675 0.55000554 1.986941
## allRepCounts.Scaled 1.7140260 1.02745679 2.855508
## Colorcentromeric    0.4077278 0.07179976 2.361563
## Colortelomeric      1.8332530 0.48818066 7.520074
## WAvgRate.perMb.Scaled 0.5719497 0.23236142 1.282011
```

Example of interpretation: “For 1 unit increase in Length.Mb.Scaled, a window is 1.0338675 times more likely to increase in inversion amount category.”

Proportional odds assessment

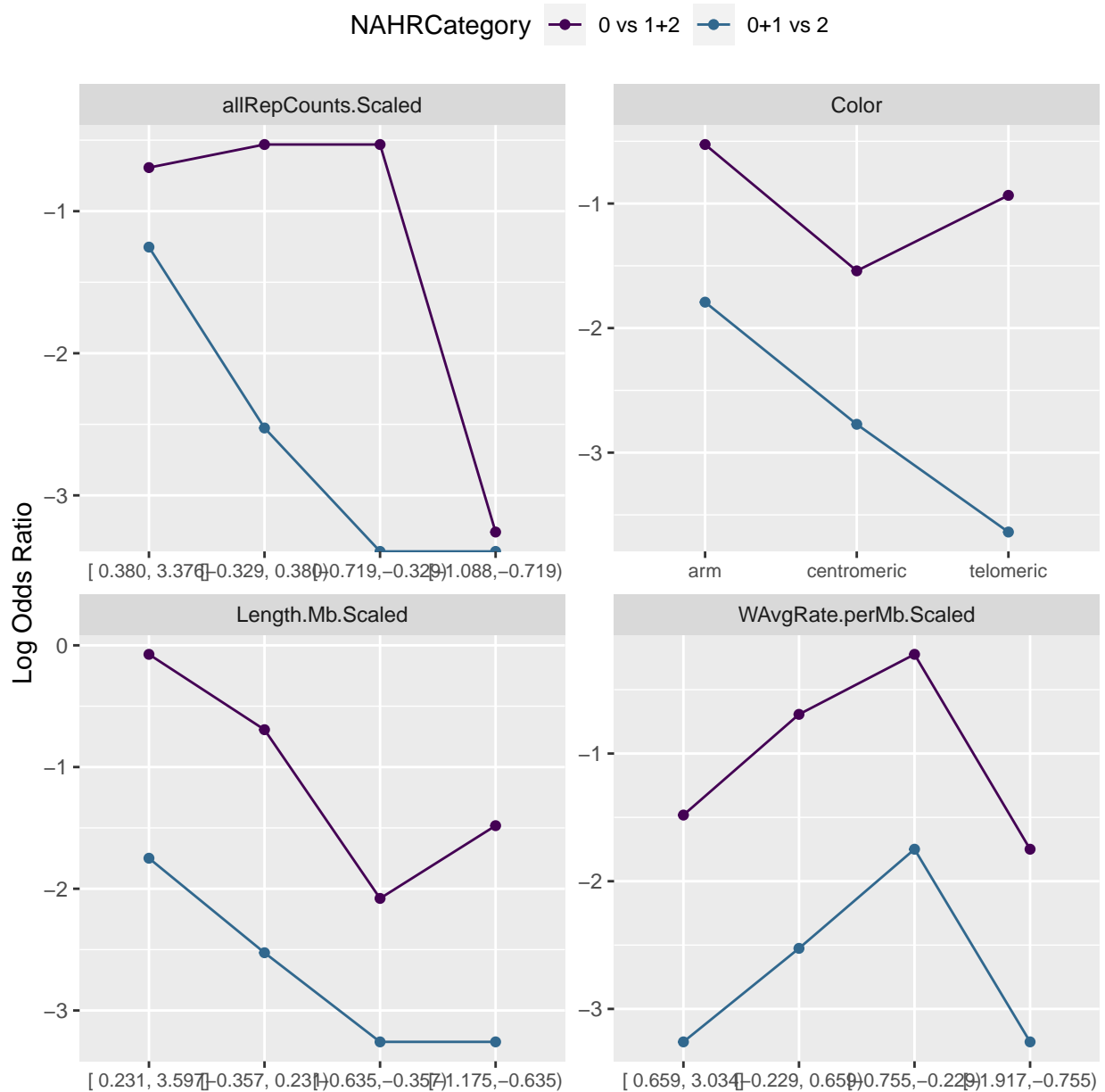
Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe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

We test the parallel regression assumption with a Brant test:

```
## -----
## Test for      X2  df  probability
## -----
## Omnibus           34.57   5    0
## Length.Mb.Scaled 0.49    1   0.49
## allRepCounts.Scaled 4.31    1   0.04
## Colorcentromeric 0.02    1   0.89
## Colortelomeric    2.35    1   0.13
## WAvgRate.perMb.Scaled 0.46    1   0.5
## -----
##
## 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

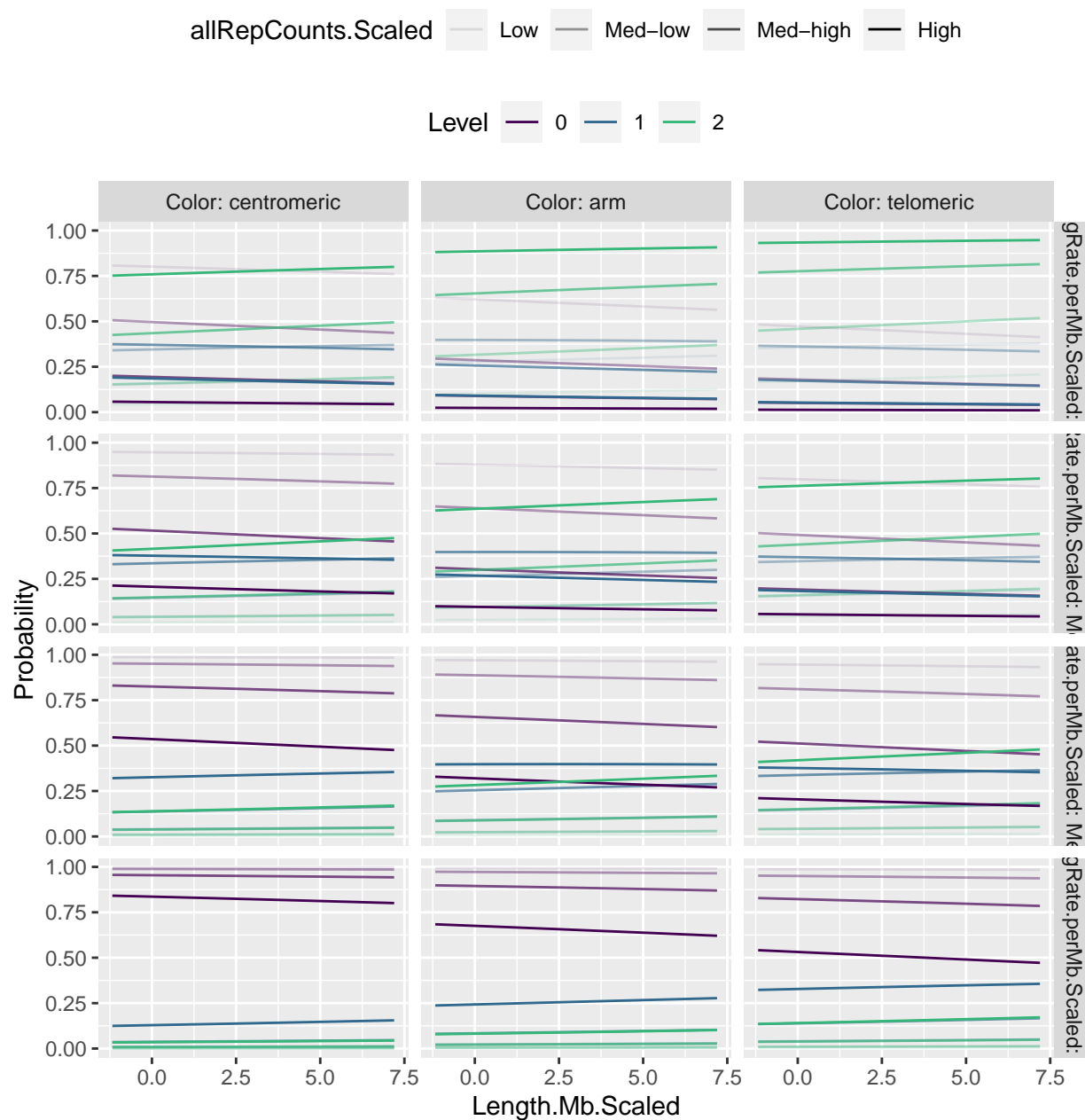


Figure 8: Probabililty of having 0 to >3 inversions depending on multiple independent variables