Ordinal logistic model on large, classified windows data from Spence

Ruth Gómez Graciani

Prepare the data

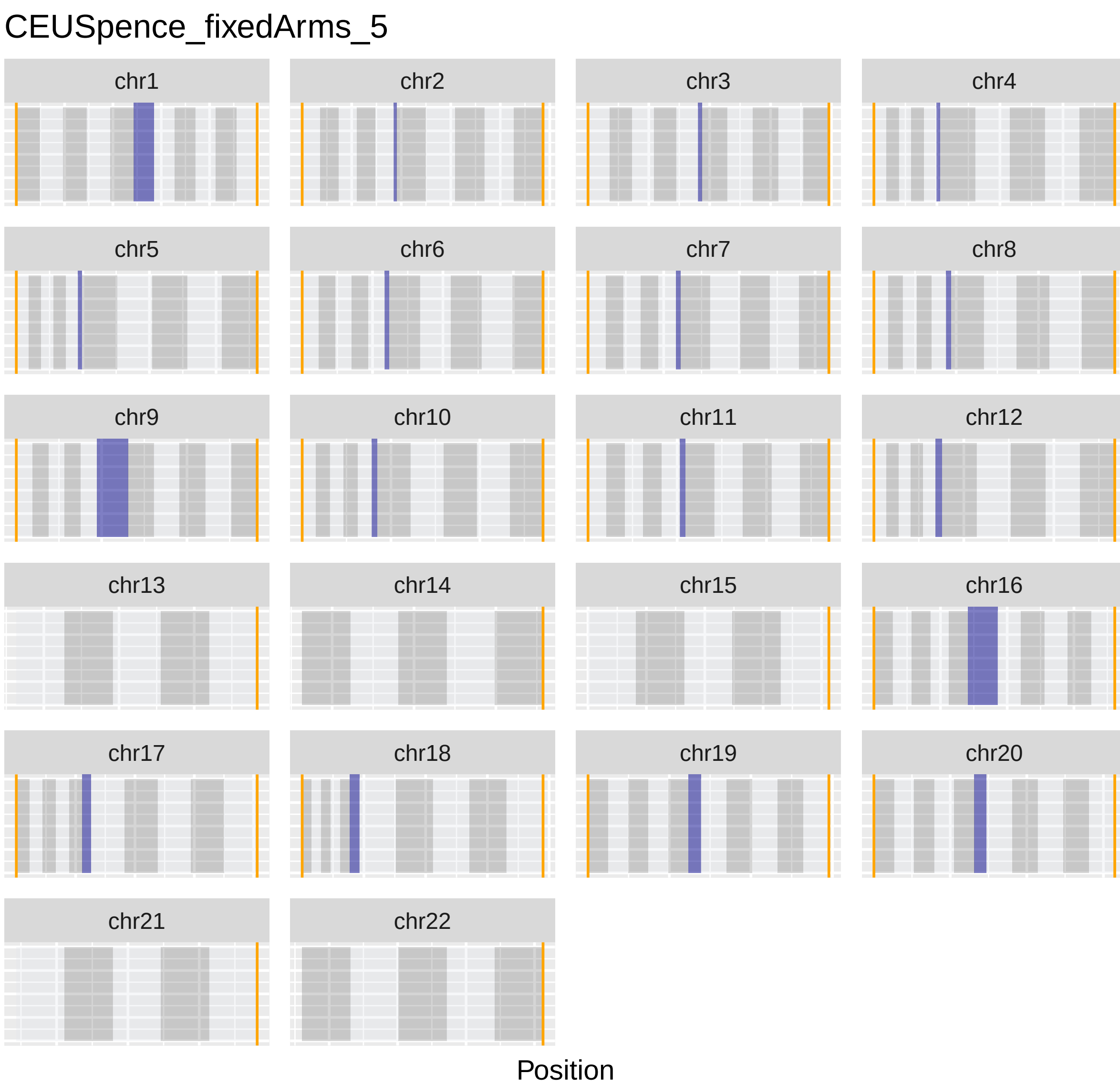


Figure 1: Crossover zones; centromeres in blue, workspace limits in orange.

Next, we define telomeric regions as the windows at the extremes of the chromosome. We will exclude centromeric regions because they have lower quality.

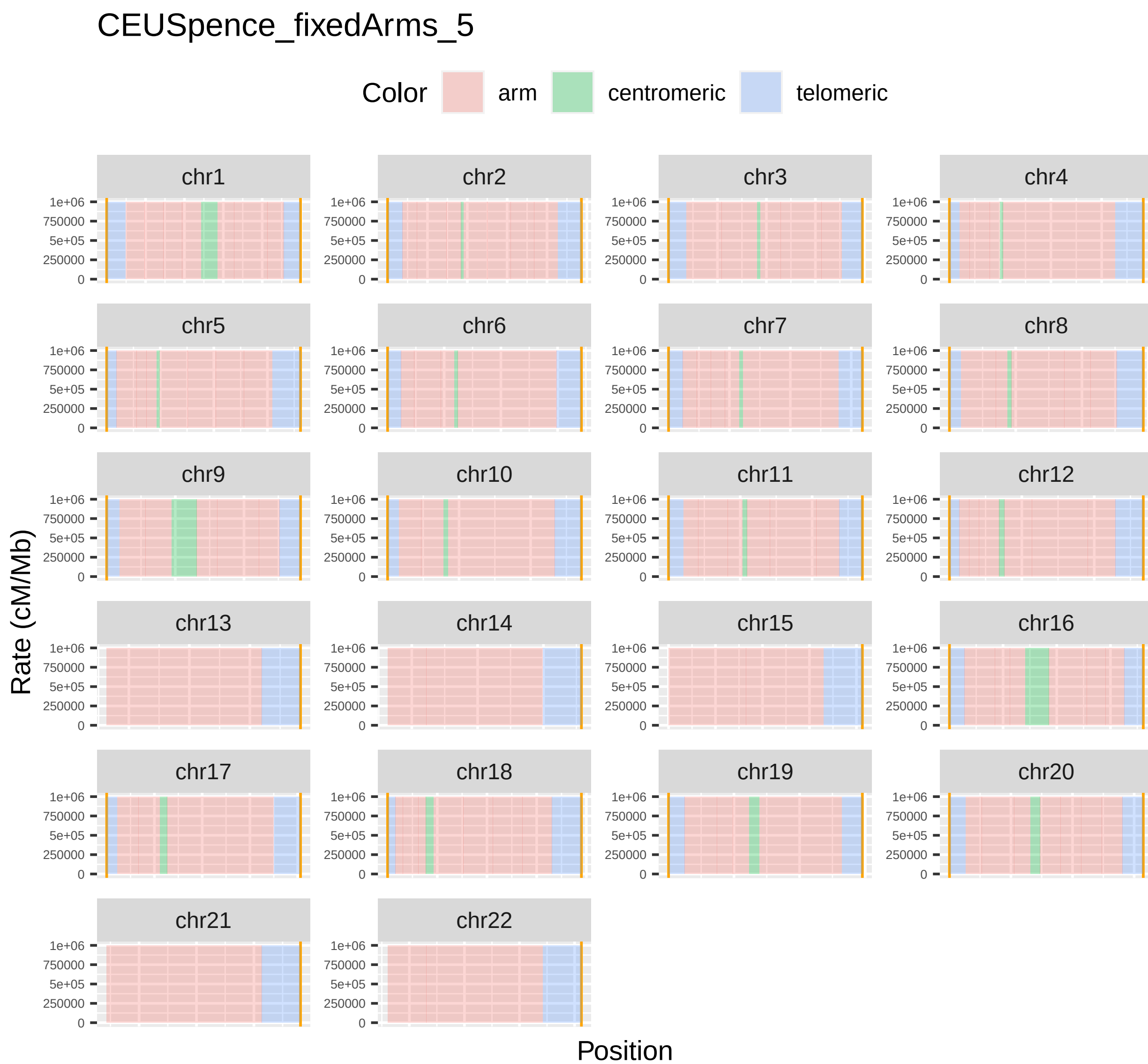


Figure 2: Color-coded windows for telomeric, centromeric and arm categories.

Descriptive statistics

Raw data:

```
##   Chromosome    Start      End     Color invCenters NHCenters NAHRCenters
## 1      chr10    60683  7877759 telomeric          1         0           1
## 2      chr10  7877759 15694835       arm          1         1           0
## 3      chr10 15694835 23511911       arm          1         1           0
## 4      chr10 23511911 31328987       arm          0         0           0
## 5      chr10 31328987 39146063       arm          1         0           1
## 6      chr10 42364408 60996401       arm          2         2           0
##   Length.Mb allRepCounts WAvgRate.perMb
## 1  7.817076          122    0.018522664
## 2  7.817076          224    0.015127506
## 3  7.817076           82    0.010026225
## 4  7.817076          292    0.010559735
## 5  7.817076          588    0.005206743
## 6 18.631993         1694    0.007113225
```

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

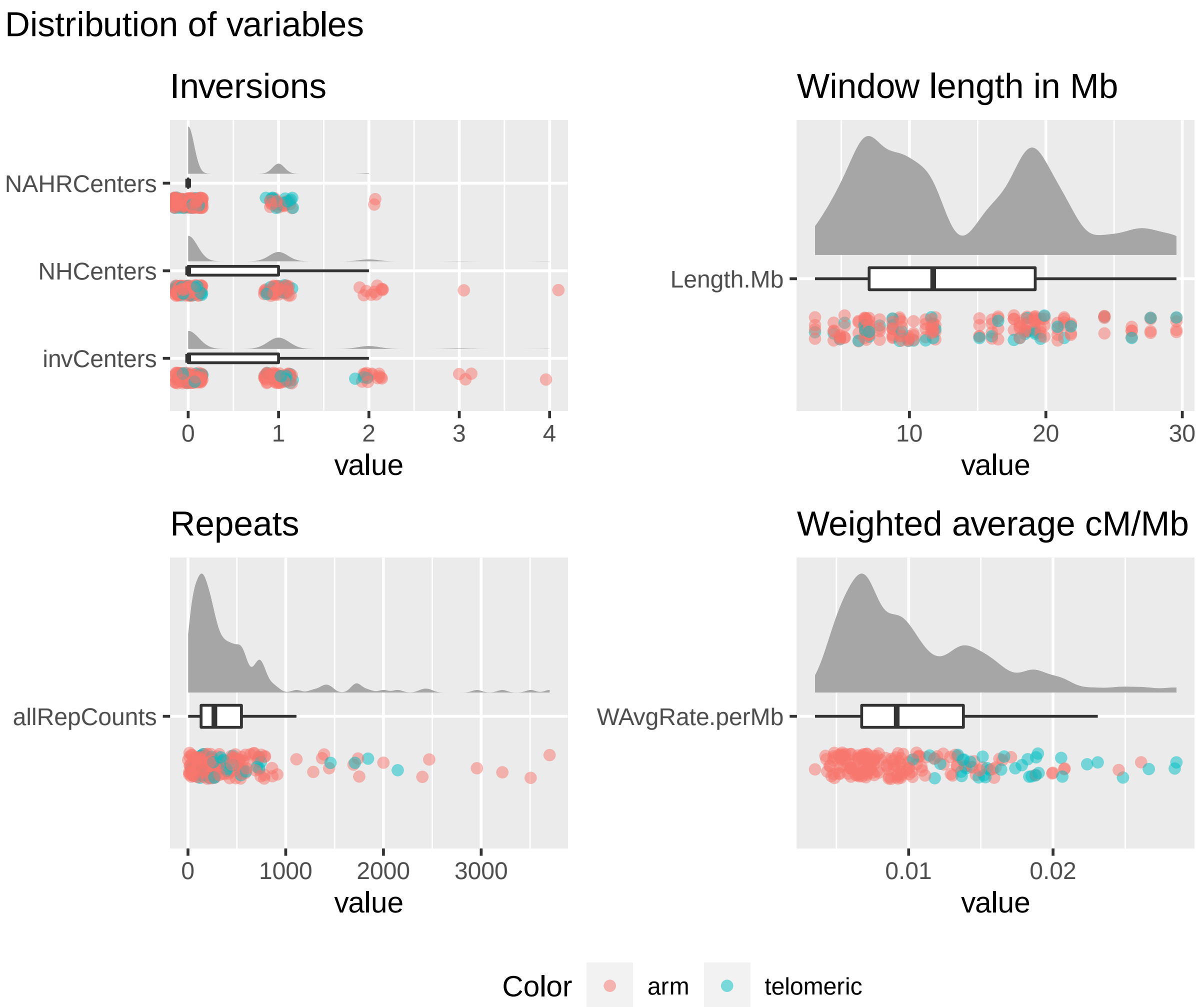


Figure 3: Distribution of variables.

We see that some categories have low number of cases, so I will make a "2 or more" category when relevant.

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

##   CountGroups invCenters NHCenters NAHRCenters
## 1           0        107       134         159
## 2           1         67        49          34
## 3           2         17        10           2
## 4           3          3         1          NA
## 5           4          1         1          NA

## [1] "New counts"

##   CountGroups invCategory NHCategory NAHRCategory
## 1           0         107        134          159
## 2           1          67         49           34
## 3          2+          21         12            2
```

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

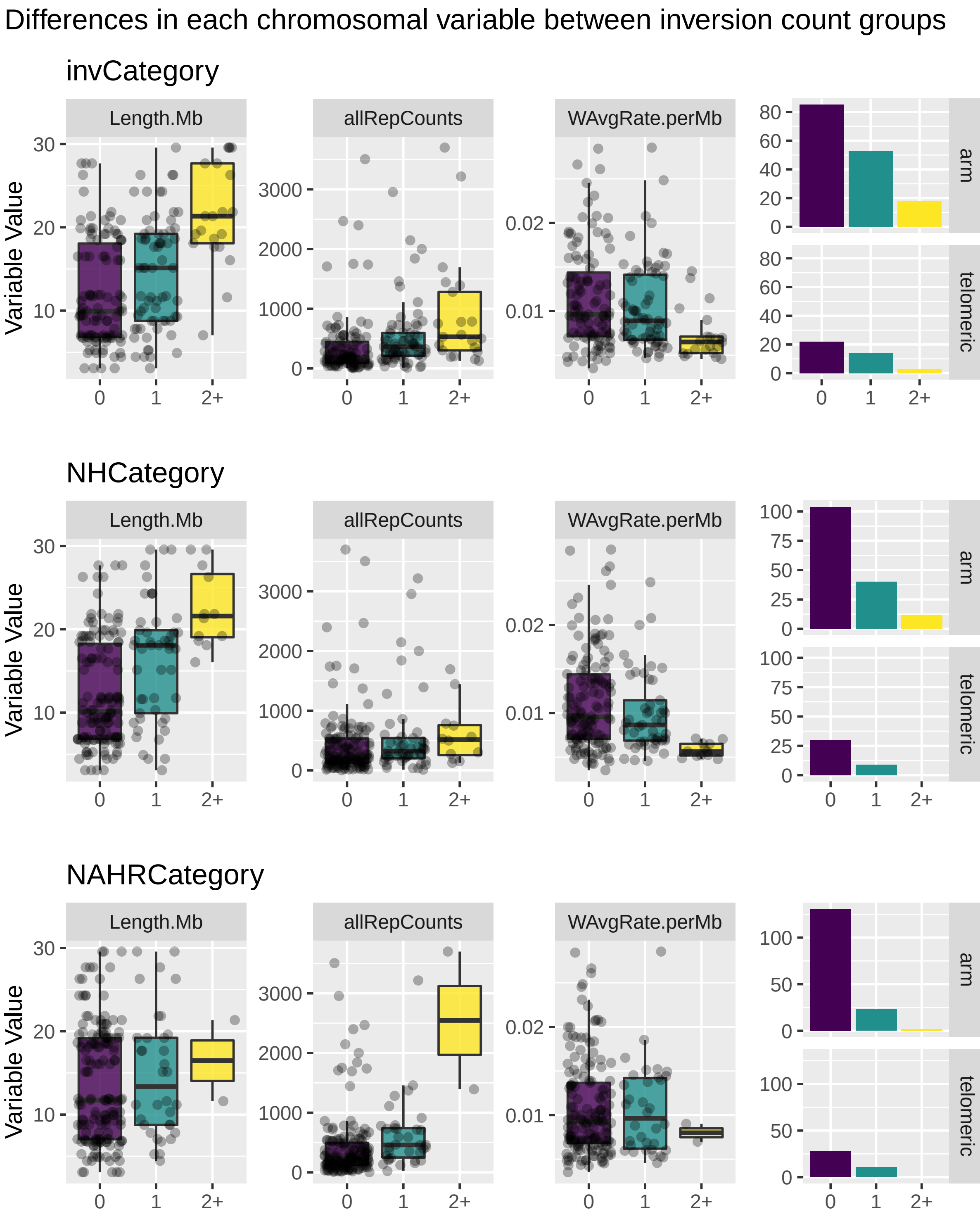


Figure 4: Potential effect of independent variables on the different types of invesions.

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

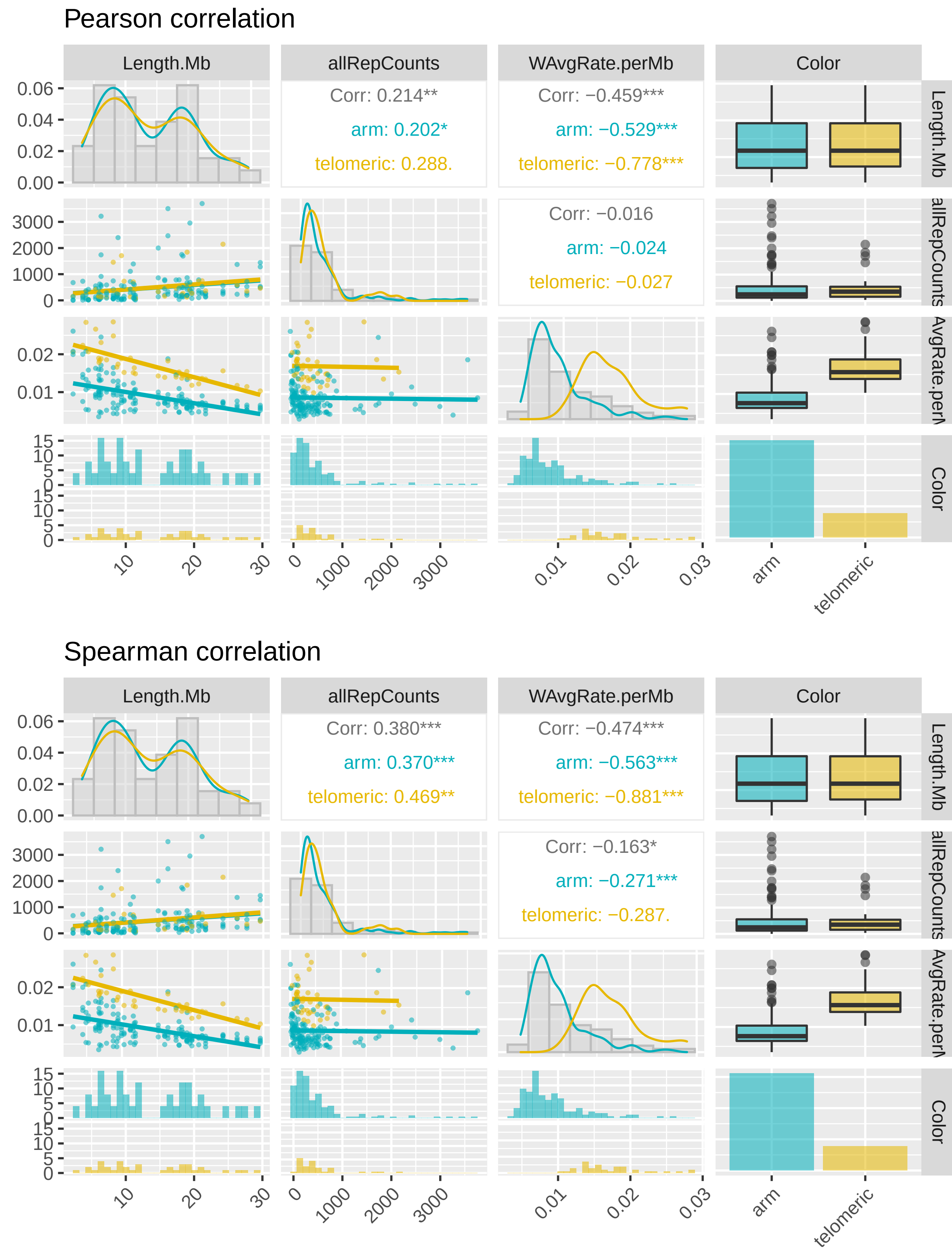


Figure 5: Correlations between variables.

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.

```
##      Length.Mb   allRepCounts          Color WAvgRate.perMb
##       1.611037       1.065216       1.932037       2.471462
```

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.   : 3.079   Min.   :-1.5268   Min.   :   2   Min.   :-0.7671
##  1st Qu.: 7.044   1st Qu.:-0.9637   1st Qu.: 134   1st Qu.:-0.5521
##  Median :11.741   Median :-0.2967   Median : 270   Median :-0.3307
##  Mean   :13.830   Mean   : 0.0000   Mean   : 473   Mean   : 0.0000
##  3rd Qu.:19.218   3rd Qu.: 0.7651   3rd Qu.: 547   3rd Qu.: 0.1205
##  Max.   :29.571   Max.   : 2.2354   Max.   :3700   Max.   : 5.2553
##  WAvgRate.perMb     WAvgRate.perMb.Scaled
##  Min.   :0.003516   Min.   :-1.3495
##  1st Qu.:0.006748   1st Qu.:-0.7379
##  Median :0.009172   Median :-0.2792
##  Mean   :0.010647   Mean   : 0.0000
##  3rd Qu.:0.013791   3rd Qu.: 0.5948
##  Max.   :0.028546   Max.   : 3.3868
```

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       8.635e-02  0.0223652     3.8607
## allRepCounts    6.527e-04  0.0002818     2.3161
## Colortelomeric  1.311e-01  0.3674270     0.3567
## WAvgRate.perMb -4.401e+01  0.0049916 -8816.1289
##
## Intercepts:
##      Value      Std. Error t value
## 0|1      1.2284     0.3575     3.4363
## 1|2+     3.4314     0.4579     7.4939
##
## Residual Deviance: 330.3624
## AIC: 342.3624
```

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

```
##                        Value  Std. Error       t value    p value
## Length.Mb       8.634585e-02 0.022365204     3.8607228 0.00011305
## allRepCounts    6.527215e-04 0.000281815     2.3161345 0.02055093
## Colortelomeric  1.310581e-01 0.367426998     0.3566916 0.72132271
## WAvgRate.perMb -4.400674e+01 0.004991617 -8816.1288844 0.00000000
## 0|1             1.228351e+00 0.357459551     3.4363352 0.00058964
## 1|2+            3.431442e+00 0.457898713     7.4938884 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.

```
##                       2.5 %      97.5 %
## Length.Mb       0.034468773 0.139735975
## allRepCounts    0.000163073 0.001160119
## Colortelomeric -0.887645275 1.200688781
## WAvgRate.perMb           NA          NA
```

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

```
##                        2.5 %        97.5 %
## Length.Mb       4.251086e-02   0.130180849
## allRepCounts    1.003742e-04   0.001205069
## Colortelomeric -5.890856e-01   0.851201796
## WAvgRate.perMb -4.401652e+01 -43.996952058
```

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      1.090183e+00 1.043427e+00 1.139034e+00
## allRepCounts   1.000653e+00 1.000100e+00 1.001206e+00
## Colortelomeric 1.140034e+00 5.548344e-01 2.342460e+00
```

```
## WAvgRate.perMb 7.728899e-20 7.653653e-20 7.804885e-20
```

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0901833 times more likely to increase in inversion amount category."

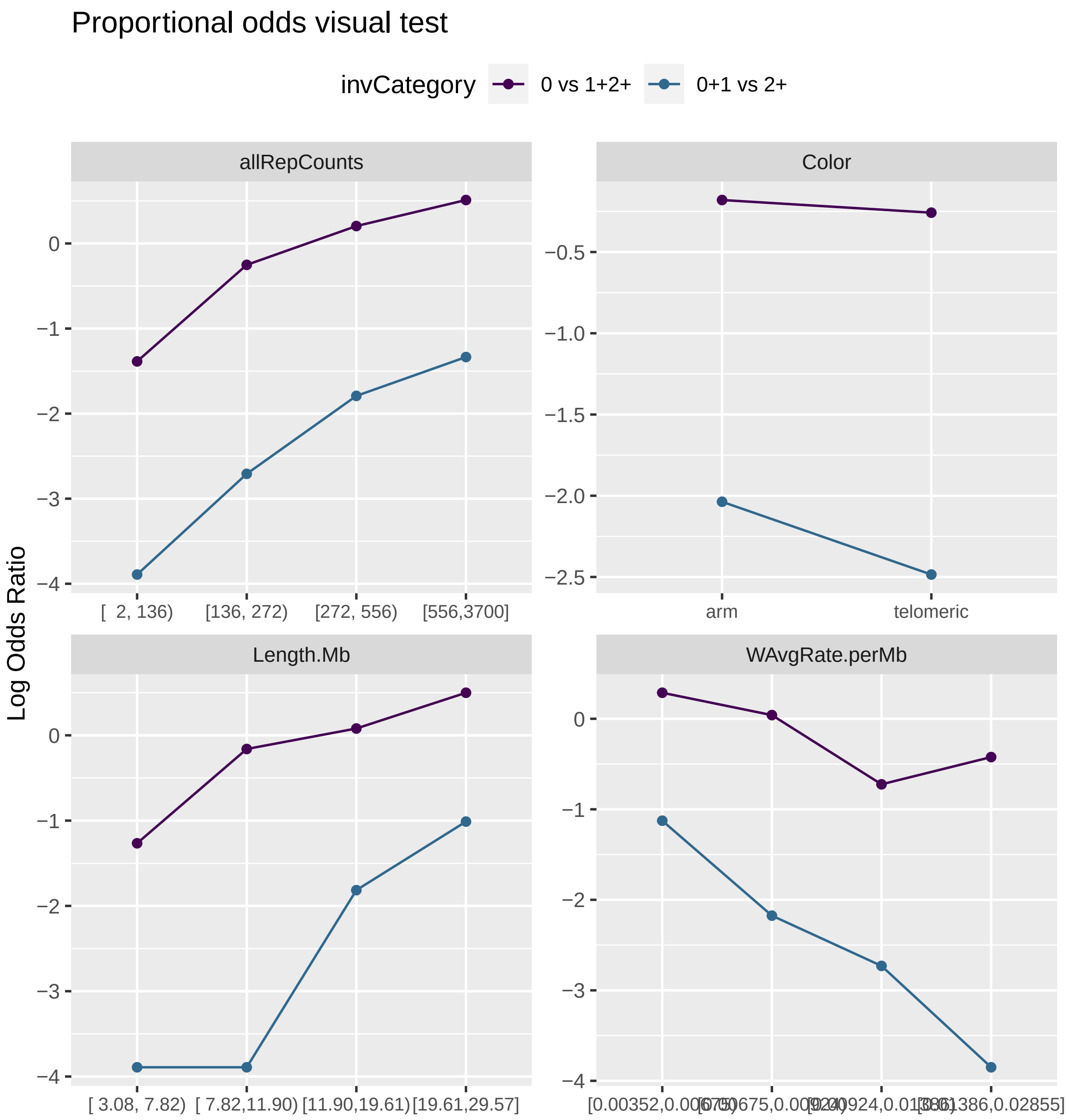
Proportional odds assessment

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:

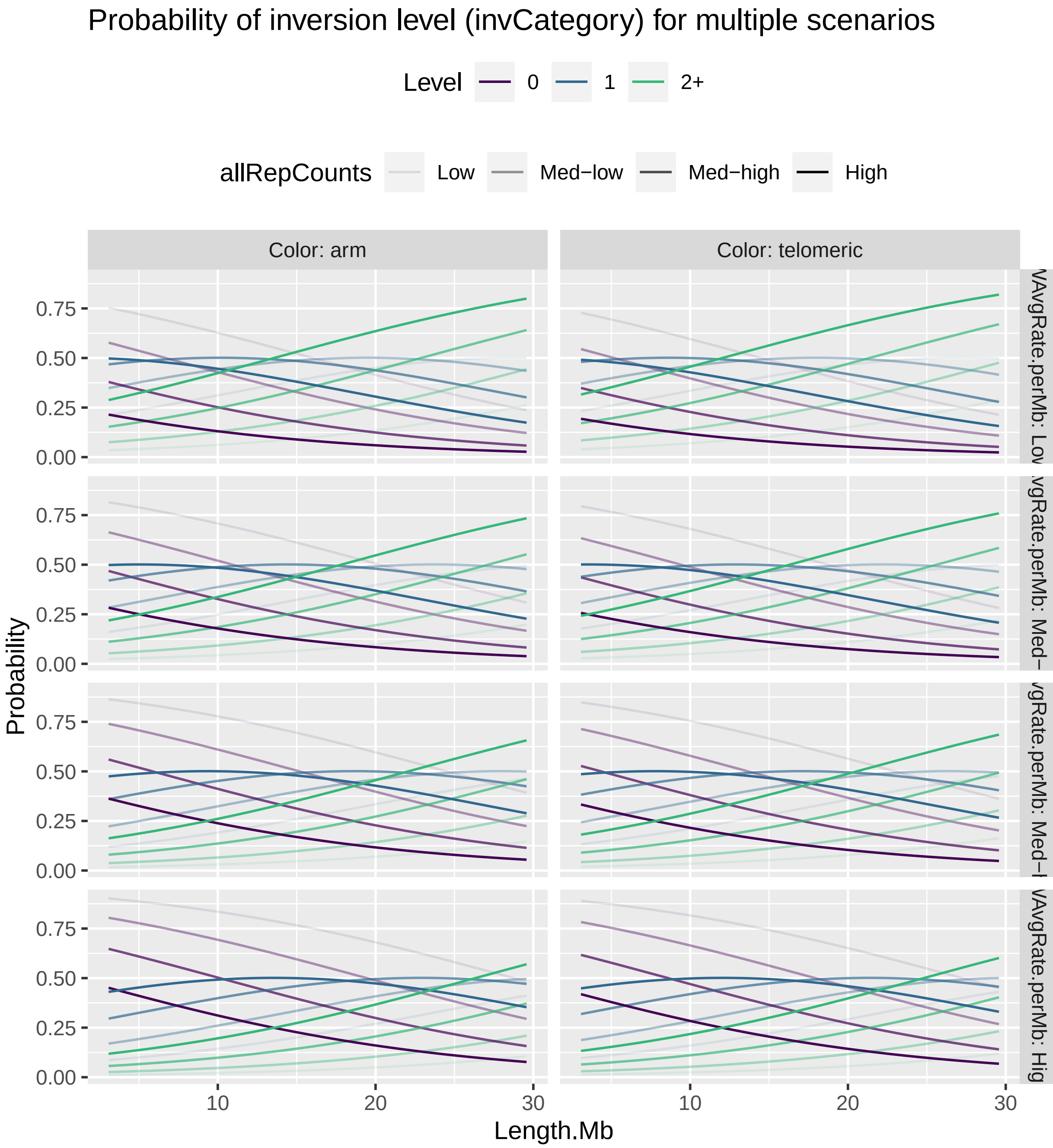
```
## --------------------------------------------
## Test for X2  df  probability
## --------------------------------------------
## Omnibus      8.26    4   0.08
## Length.Mb    2.38    1   0.12
## allRepCounts 1.49    1   0.22
## Colortelomeric   1.55    1   0.21
## WAvgRate.perMb   3.66    1   0.06
## --------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

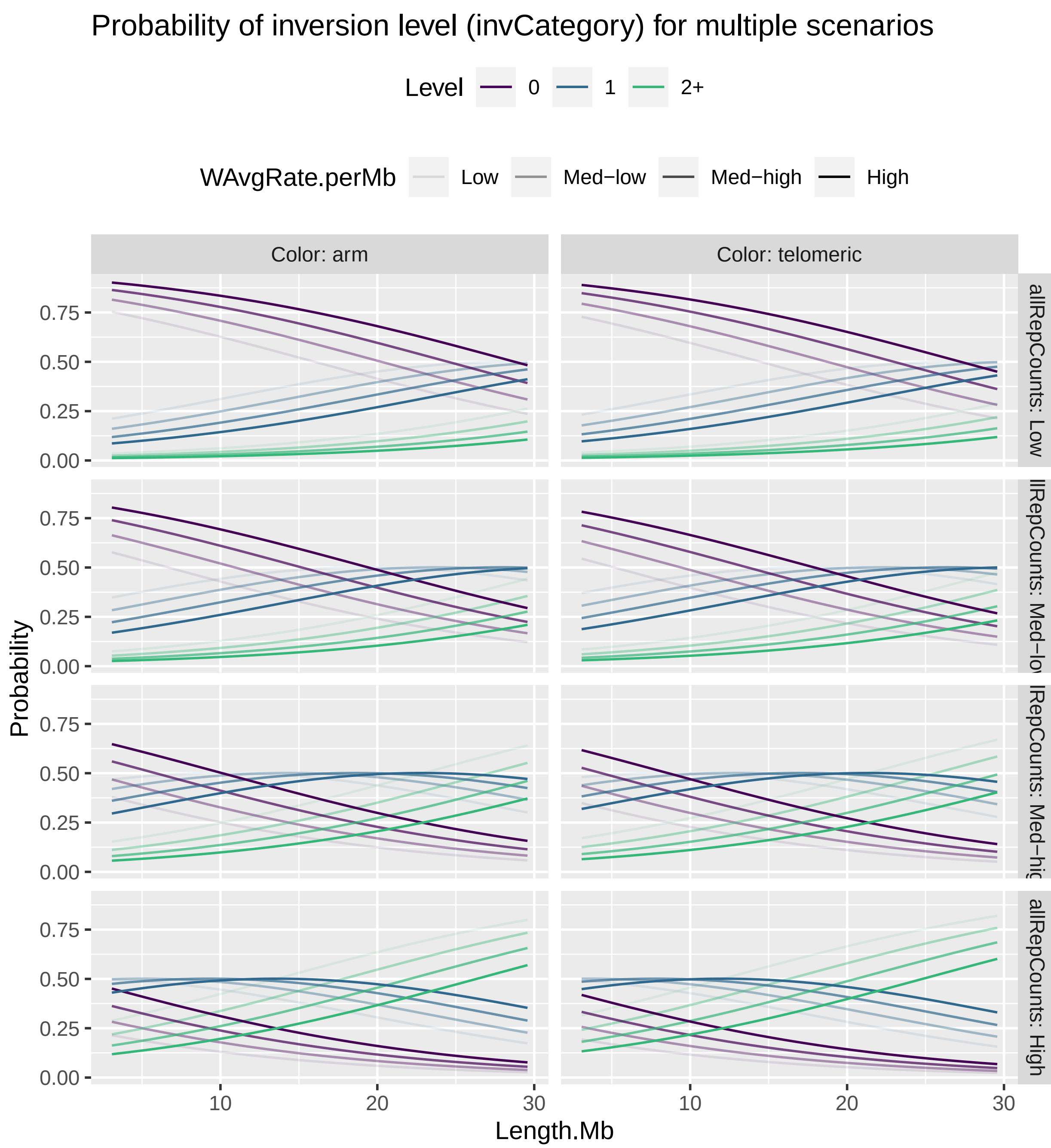
We can also evaluate the parallel regression visually. We transform the ordinal dependent variable with k categories into a series of k-1 binary variables that indicate whether the dependent value is above or below a cutpoint (e.g. windows with at least 2 inversions vs windows with less than 2 inversions). 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.





Probability of inversion level (invCategory) for multiple scenarios
Level 0 1 2+
WAvgRate.perMb Low Med−low Med−high High
Color: arm
Color: telomeric
allRepCounts: Low
allRepCounts: Med−lo
allRepCounts: Med−hig
allRepCounts: High
Probability
0.00
0.25
0.50
0.75
10
20
30
Length.Mb

NH inversions (NHCategory)

Model fitting

The comination telomere-2+ inversions did not occur, so I will not use the "Color" category

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                     Value Std. Error    t value
## Length.Mb       1.077e-01   0.024276     4.4379
## allRepCounts    1.719e-04   0.000283     0.6073
## WAvgRate.perMb -3.261e+01   0.005587 -5837.8549
##
## Intercepts:
##      Value      Std. Error t value
## 0|1      2.1051     0.4162     5.0575
## 1|2+     4.2735     0.5361     7.9721
##
## Residual Deviance: 272.938
## AIC: 282.938
```

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

```
##                          Value   Std. Error       t value    p value
## Length.Mb       1.077338e-01 0.0242757182     4.4379226 0.00000908
## allRepCounts    1.718616e-04 0.0002829885     0.6073097 0.54364541
## WAvgRate.perMb -3.261391e+01 0.0055866253 -5837.8549209 0.00000000
## 0|1             2.105068e+00 0.4162295943     5.0574674 0.00000042
## 1|2+            4.273489e+00 0.5360550097     7.9721093 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"
```

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

```
##                          2.5 %        97.5 %
## Length.Mb       6.015422e-02  1.553133e-01
## allRepCounts   -3.827856e-04  7.265088e-04
## WAvgRate.perMb -3.262486e+01 -3.260296e+01
```

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      1.113751e+00 1.043427e+00 1.139034e+00
## allRepCounts   1.000172e+00 1.000100e+00 1.001206e+00
## Colortelomeric 6.854245e-15 5.548344e-01 2.342460e+00
## WAvgRate.perMb 1.113751e+00 7.653653e-20 7.804885e-20
```

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.1137512 times more likely to increase in inversion amount category."

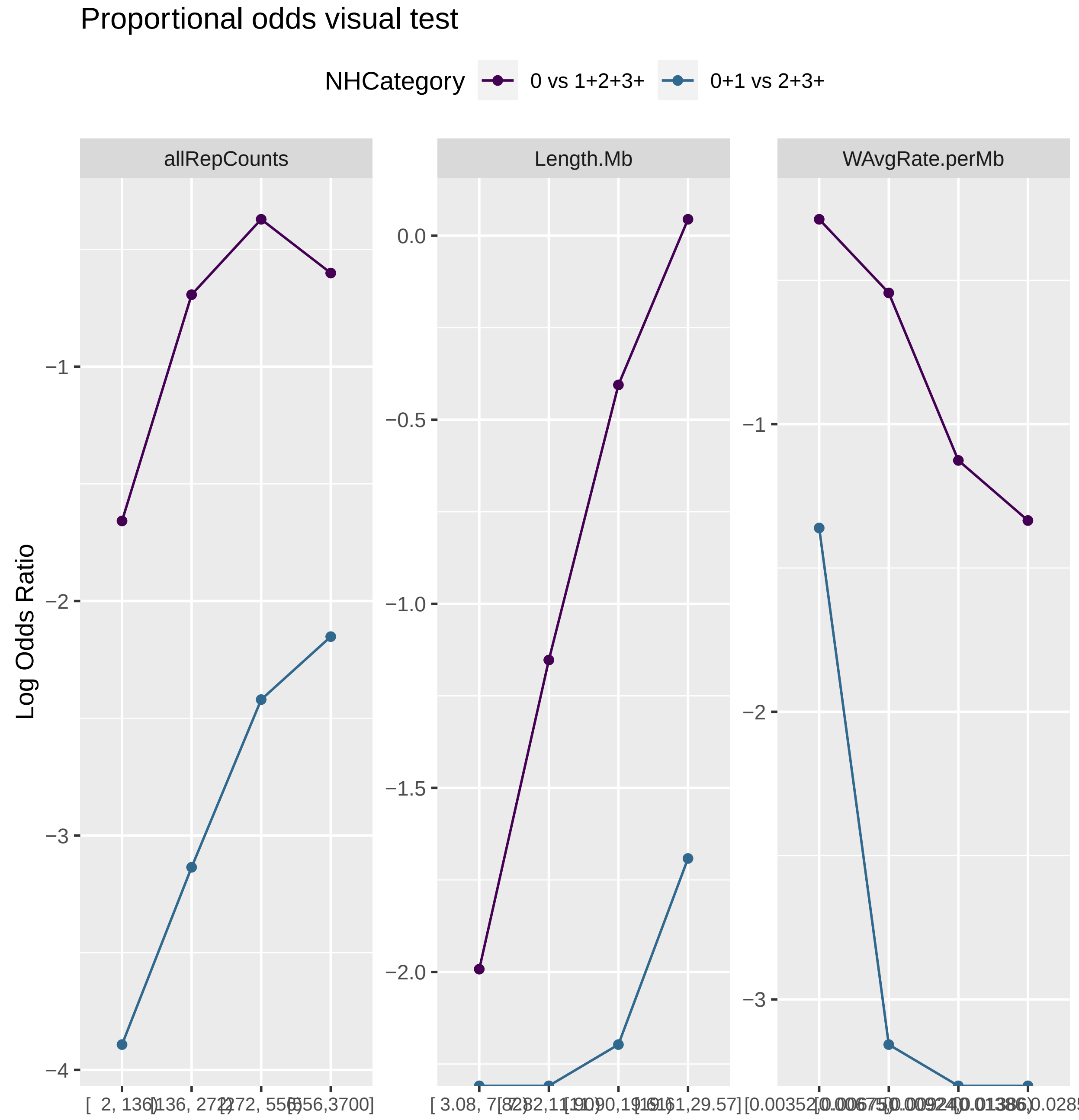
Proportional odds assessment

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      6.68    3   0.08
## Length.Mb    1.49    1   0.22
## allRepCounts 0.01    1   0.94
## WAvgRate.perMb   6.04    1   0.01
## --------------------------------------------
##
## 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.

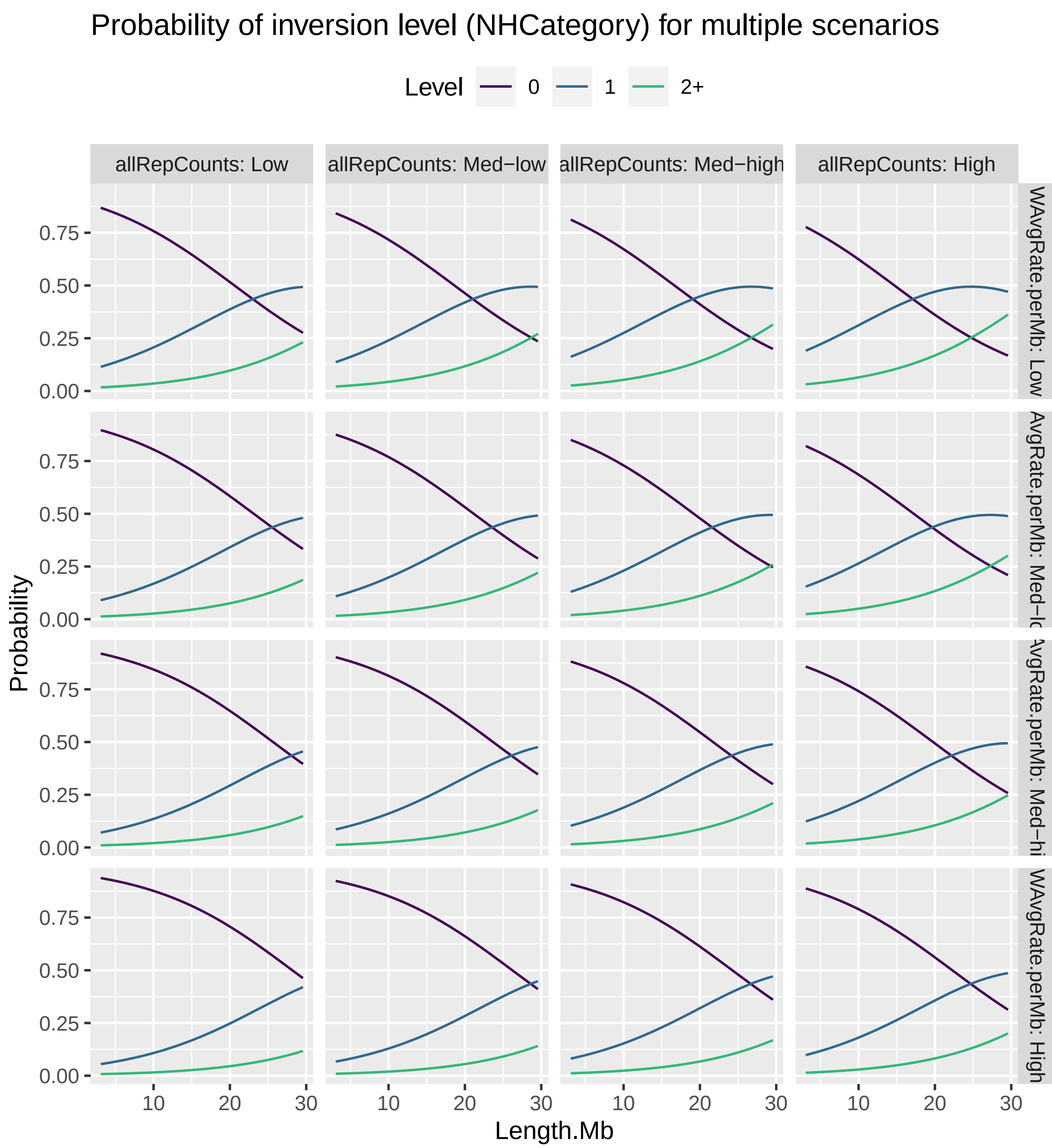


Figure 6: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions (NAHRCategory)

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                    Value Std. Error   t value
## Length.Mb       0.0123420  0.0268842    0.4591
## allRepCounts    0.0007666  0.0003133    2.4471
## WAvgRate.perMb -2.9669147  0.0066134 -448.6184
##
## Intercepts:
##      Value     Std. Error t value
## 0|1     2.0356    0.4430     4.5954
## 1|2+    5.2114    0.8389     6.2121
##
## Residual Deviance: 193.4798
## AIC: 203.4798
```

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

```
##                       Value   Std. Error      t value    p value
## Length.Mb       0.012341993 0.0268841765    0.4590802 0.64617658
## allRepCounts    0.000766619 0.0003132719    2.4471362 0.01439964
## WAvgRate.perMb -2.966914682 0.0066134487 -448.6183889 0.00000000
## 0|1             2.035550567 0.4429513406    4.5954270 0.00000432
## 1|2+            5.211352894 0.8389009889    6.2121191 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      -0.0471323870 0.070592993
## allRepCounts    0.0002264389 0.001314588
## WAvgRate.perMb            NA          NA
```

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

```
##                       2.5 %       97.5 %
## Length.Mb      -0.0403500249  0.065034010
## allRepCounts    0.0001526174  0.001380621
## WAvgRate.perMb -2.9798768033 -2.953952561
```

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      1.01241847 0.9539611 1.073144
## allRepCounts   1.00076691 1.0002265 1.001315
## WAvgRate.perMb 0.05146184        NA       NA
```

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0124185 times more likely to increase in inversion amount category."

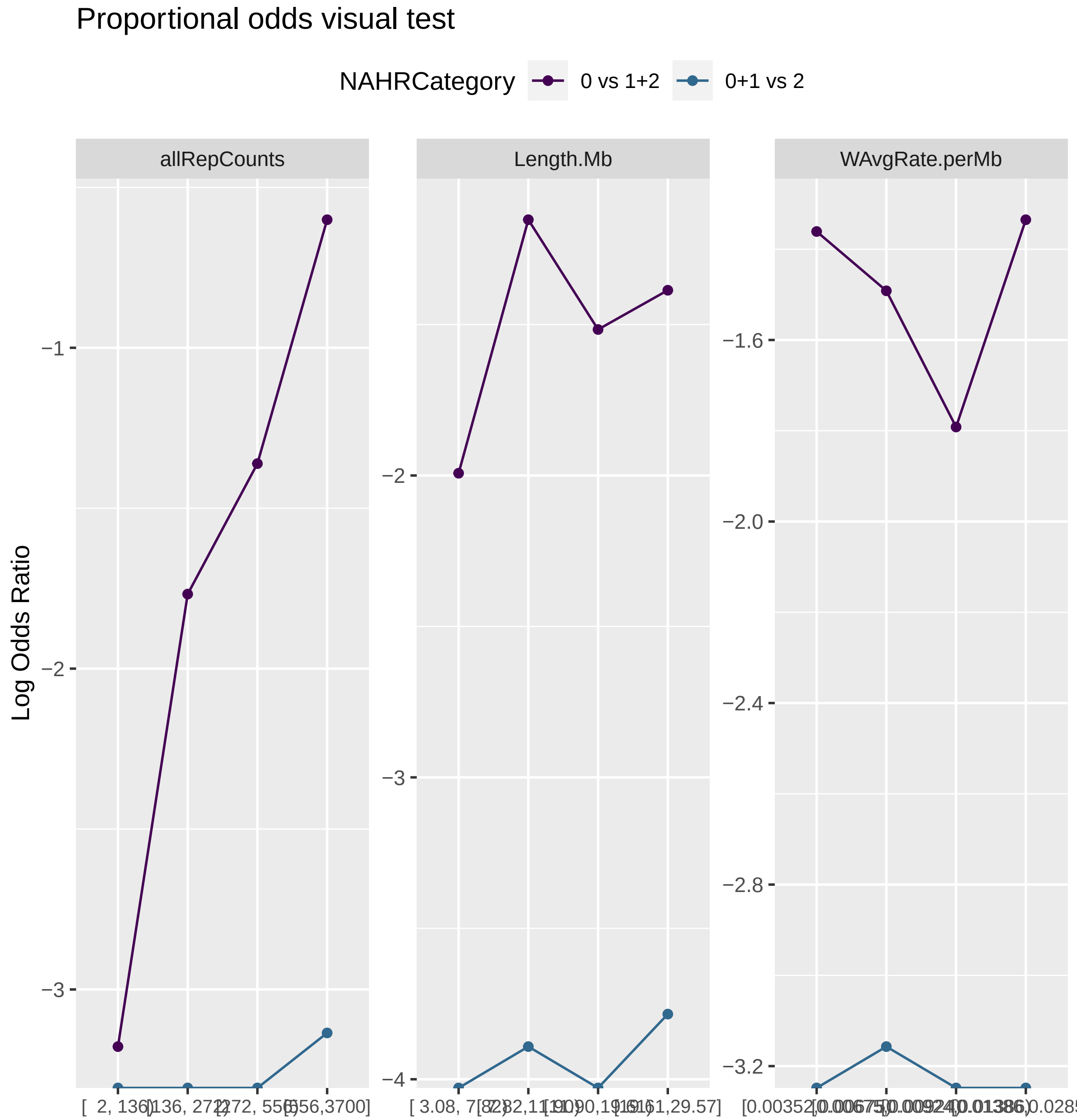
Proportional odds assessment

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      3.72    3   0.29
## Length.Mb    0   1   0.99
## allRepCounts 3.37    1   0.07
## WAvgRate.perMb   0.59    1   0.44
## --------------------------------------------
##
## 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.

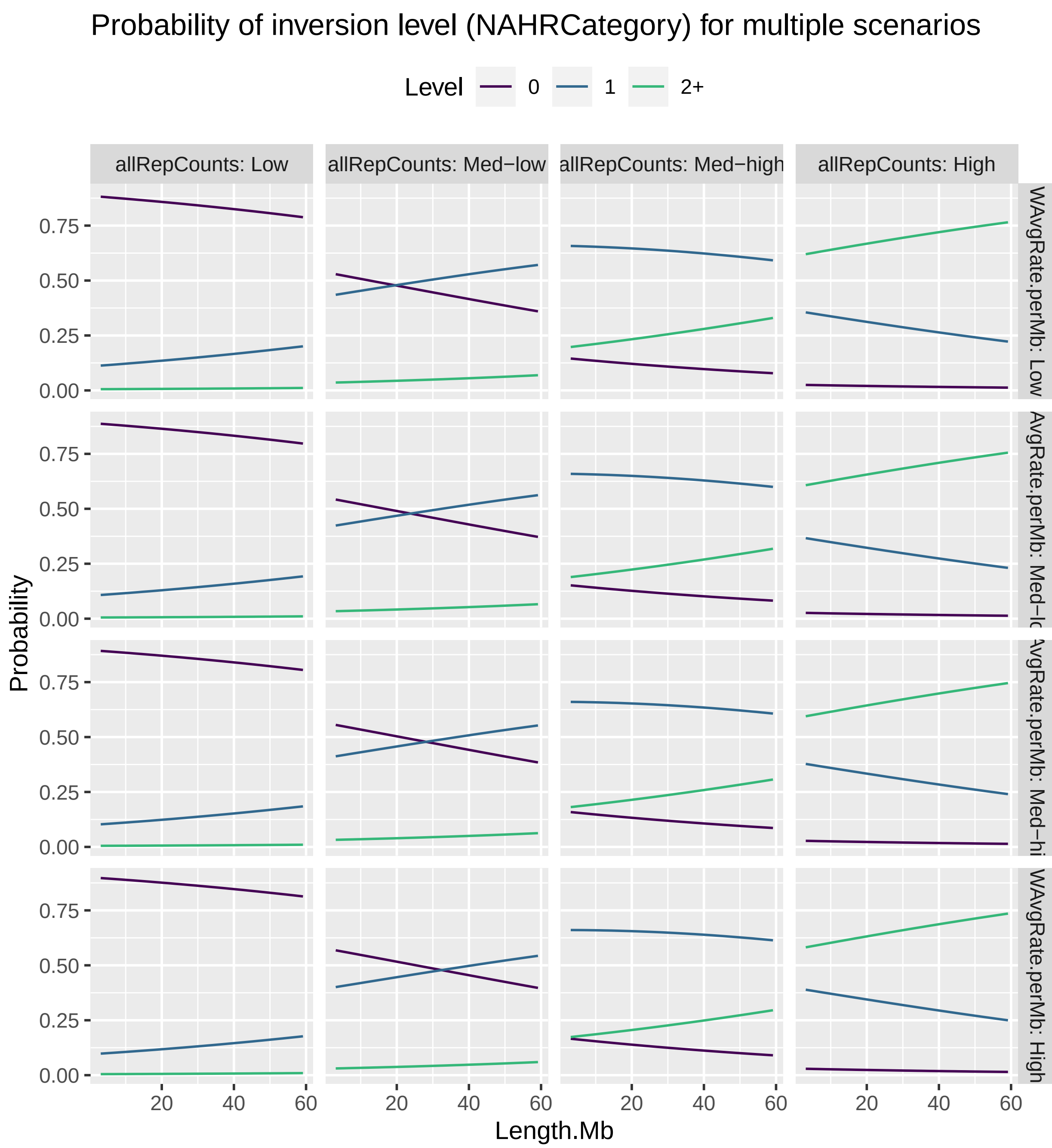


Figure 7: Probabiilty of having 0 to >3 inversions depending on multiple independent variables