Ordinal logistic model on large, classified windows data

Ruth Gómez Graciani

Contents

Prepare the data

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

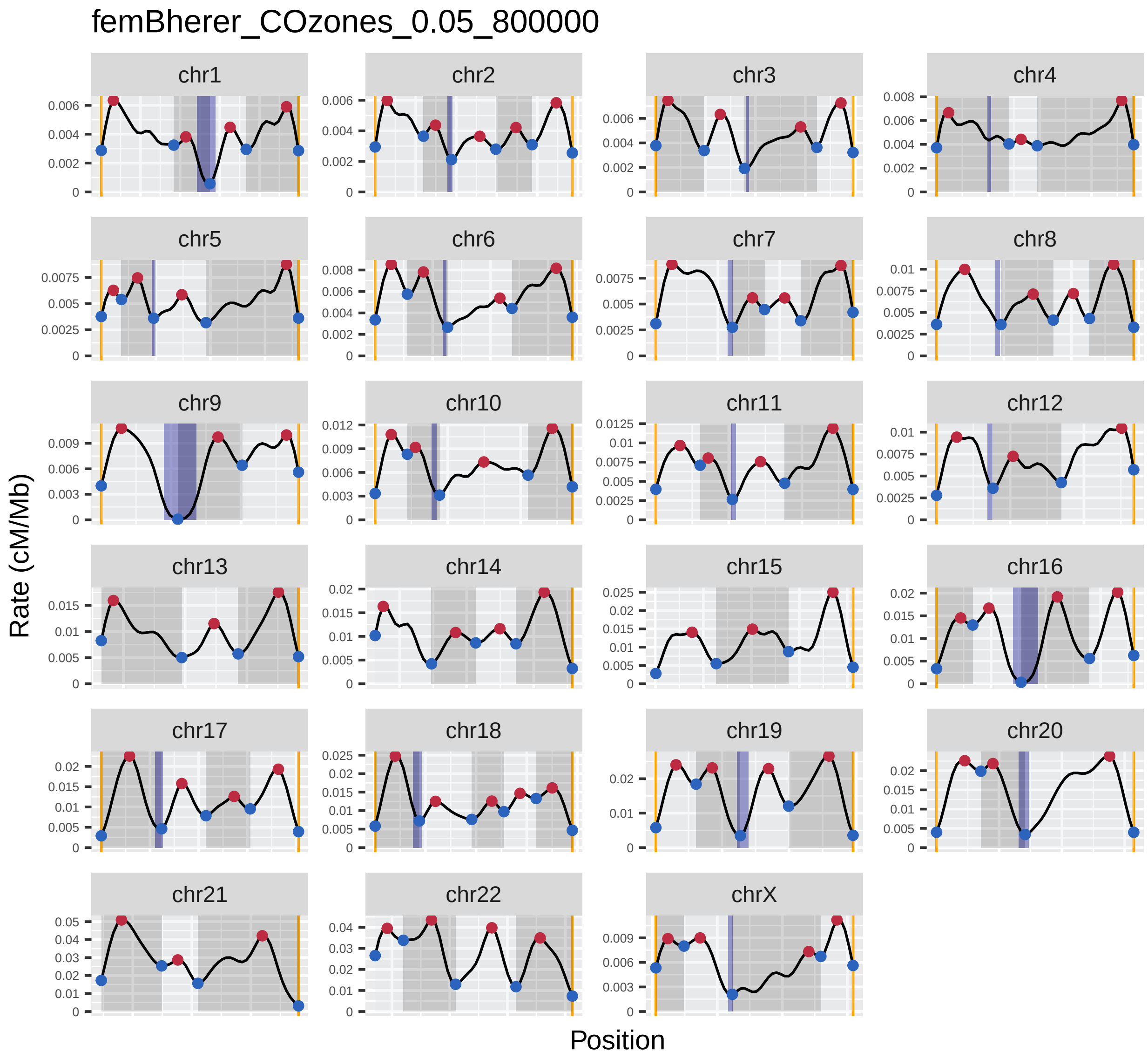


Figure 1: Crossover zones; centromeres in blue, workspace limits in orange.

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.

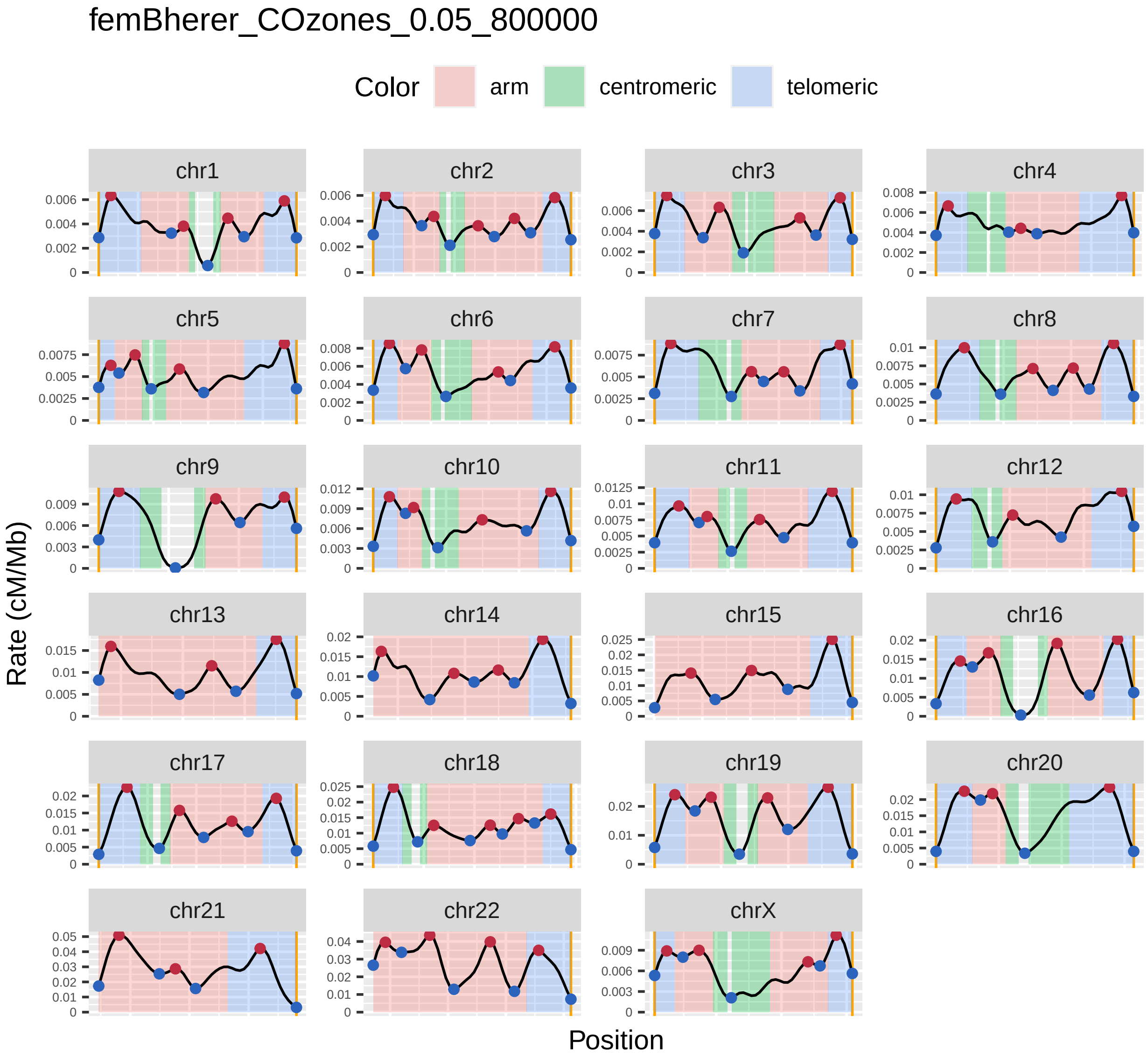
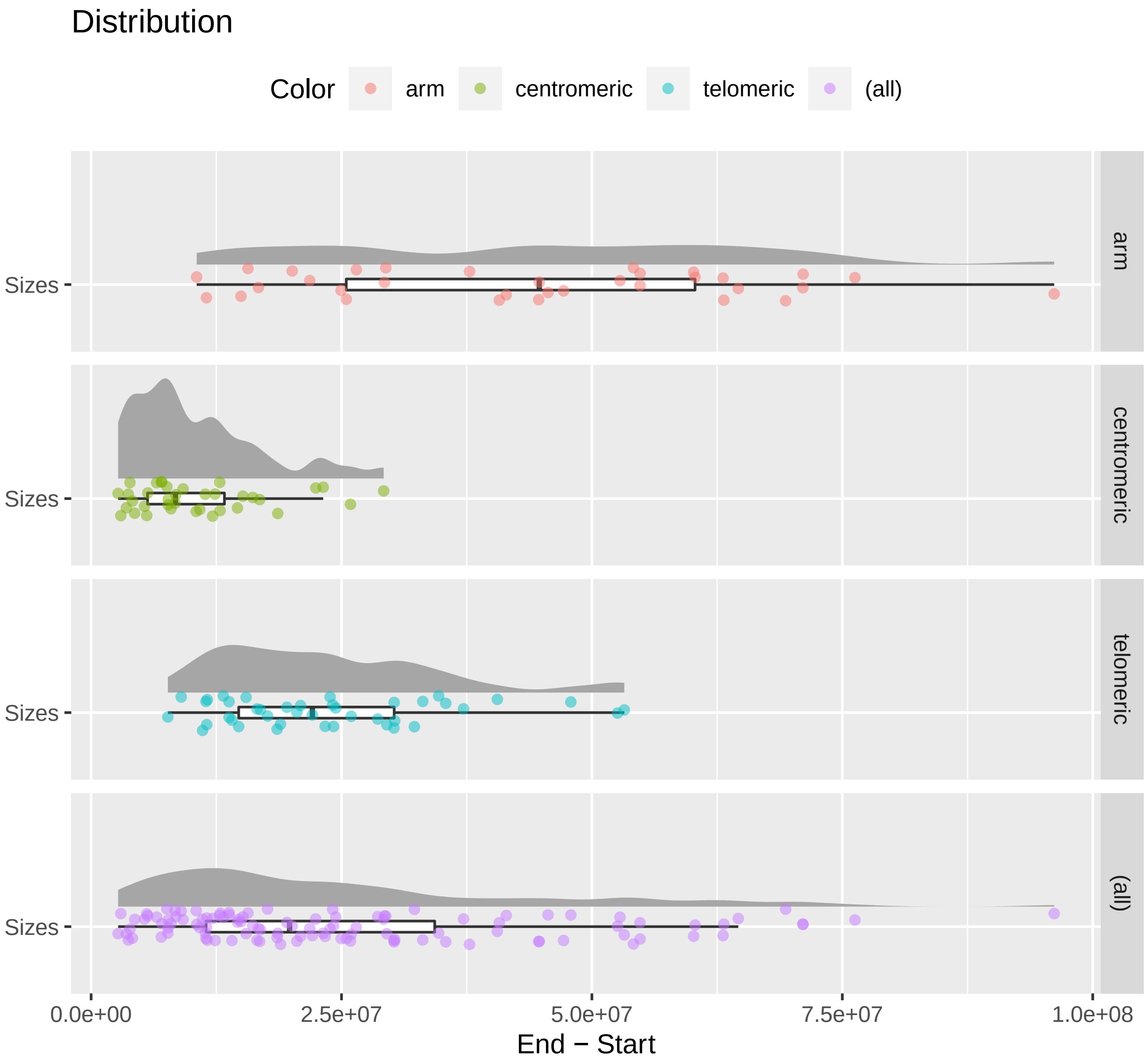
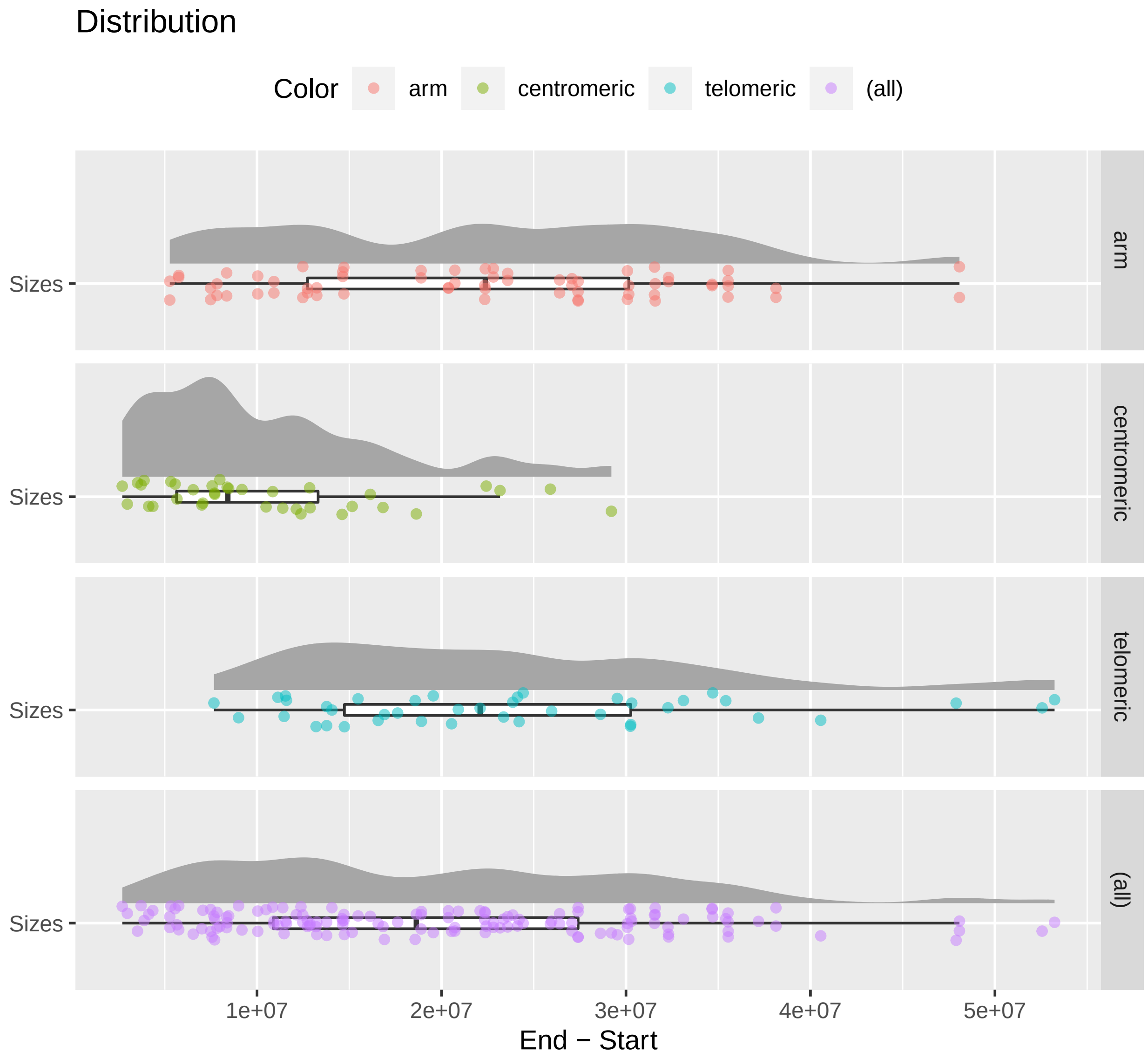


Figure 2: Color-coded windows for telomeric, centromeric and arm categories.



Distribution
Color
arm
centromeric
telomeric
(all)
Sizes
Sizes
Sizes
Sizes
arm
centromeric
telomeric
(all)
0.0e+00
2.5e+07
5.0e+07
7.5e+07
1.0e+08
End − Start


Distribution
Color
arm
centromeric
telomeric
(all)
Sizes
Sizes
Sizes
Sizes
arm
centromeric
telomeric
(all)
1e+07
2e+07
3e+07
4e+07
5e+07
End − Start

\title(Without the X)

Numerical categories

Descriptive statistics

Raw data:

Chromosome	Start	End	Color	invCenters	NHCenters	NAHRCenters	Length.Mb	allRepCounts	Log10RepCounts	WAvgRate.perMb	ChromType
chr10	158946	16728068	telomeric	3	2	1	16.569122	272	2.434569	2.0834355	A
chr10	33436031	39097912	centromeric	1	0	1	5.661881	556	2.745075	1.4181419	A
chr10	113381279	135473442	telomeric	1	1	0	22.092163	170	2.230449	2.1846155	A
chr10	42436301	58578148	centromeric	1	1	0	16.141847	1672	3.223236	0.9909238	A
chr11	241489	23608385	telomeric	1	0	1	23.366896	720	2.857333	1.7638010	A
chr11	43687013	51394932	centromeric	0	0	0	7.707919	494	2.693727	1.0575223	A

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

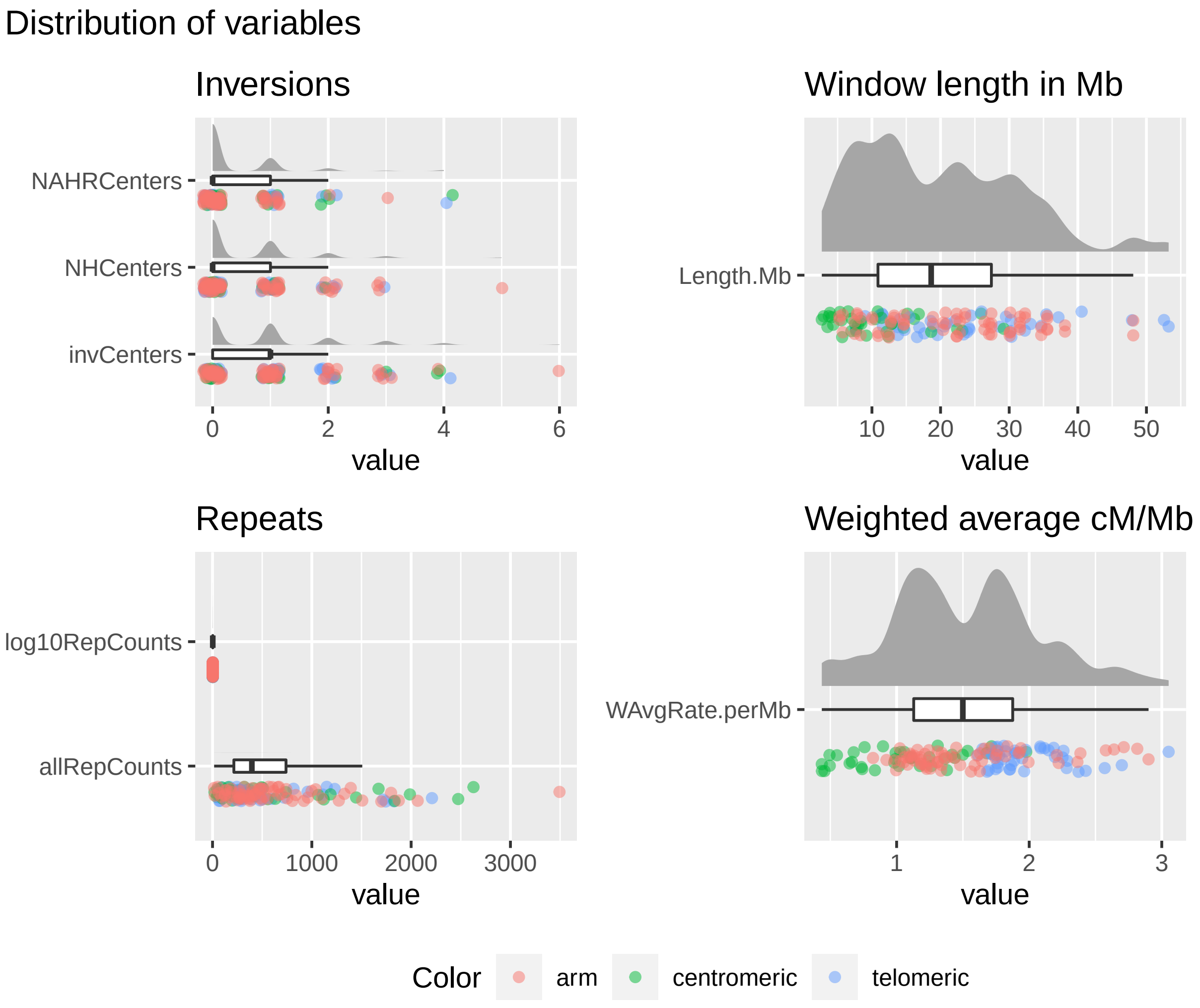


Figure 3: Distribution of variables.

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

Table 2: Original counts

CountGroups	invCenters	NHCenters	NAHRCenters
0	64	88	105
1	49	39	29
2	16	11	6
3	9	4	1
4	4	NA	2
5	NA	1	NA
6	1	NA	NA

Table 3: New counts

CountGroups	invCategory	NHCategory	NAHRCategory
0	64	88	105

CountGroups	invCategory	NHCategory	NAHRCategory
1	49	39	29
2	16	11	6
3+	14	5	3

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

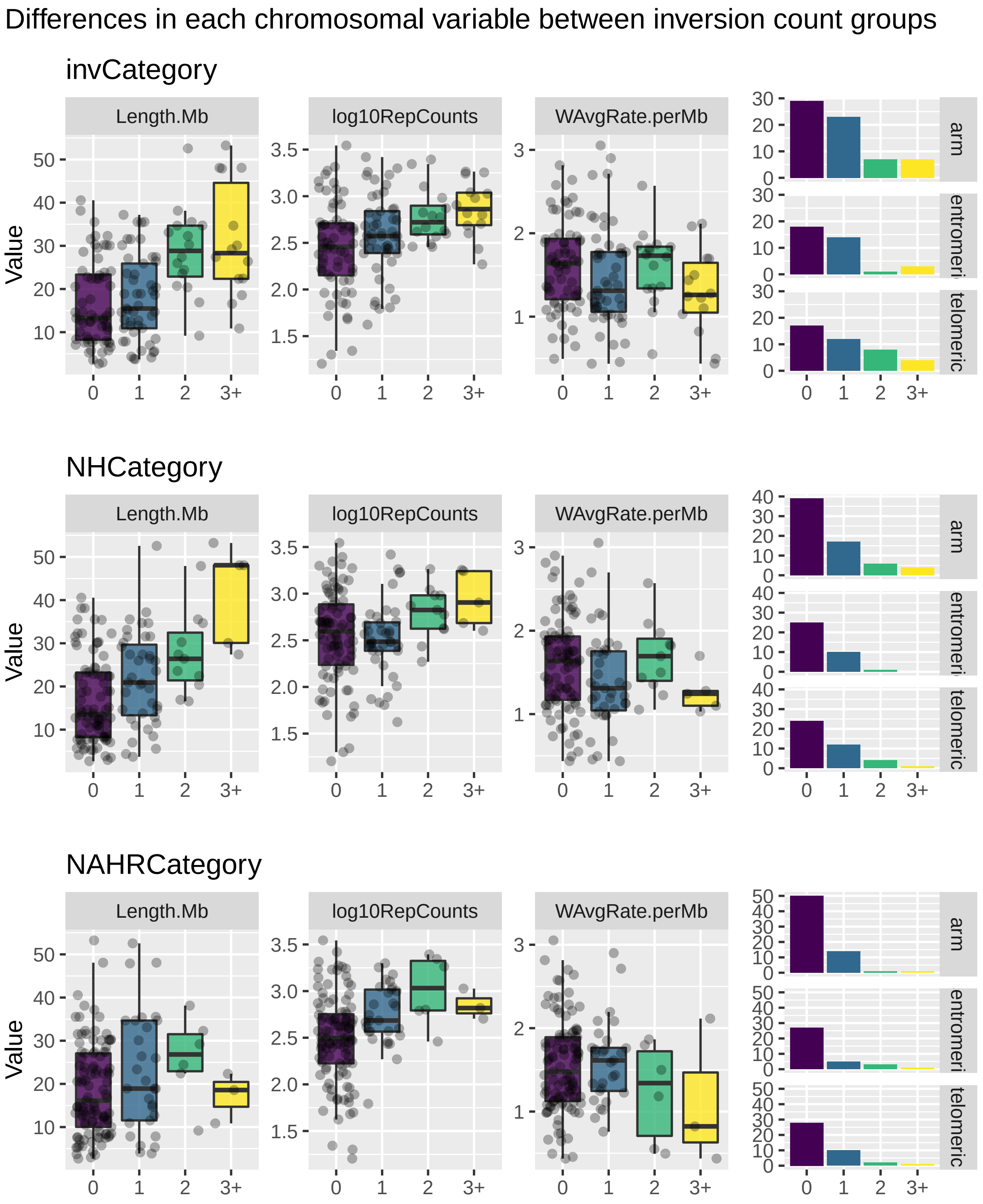


Figure 4: Potential effect of independent variables on the different types of invesions.

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

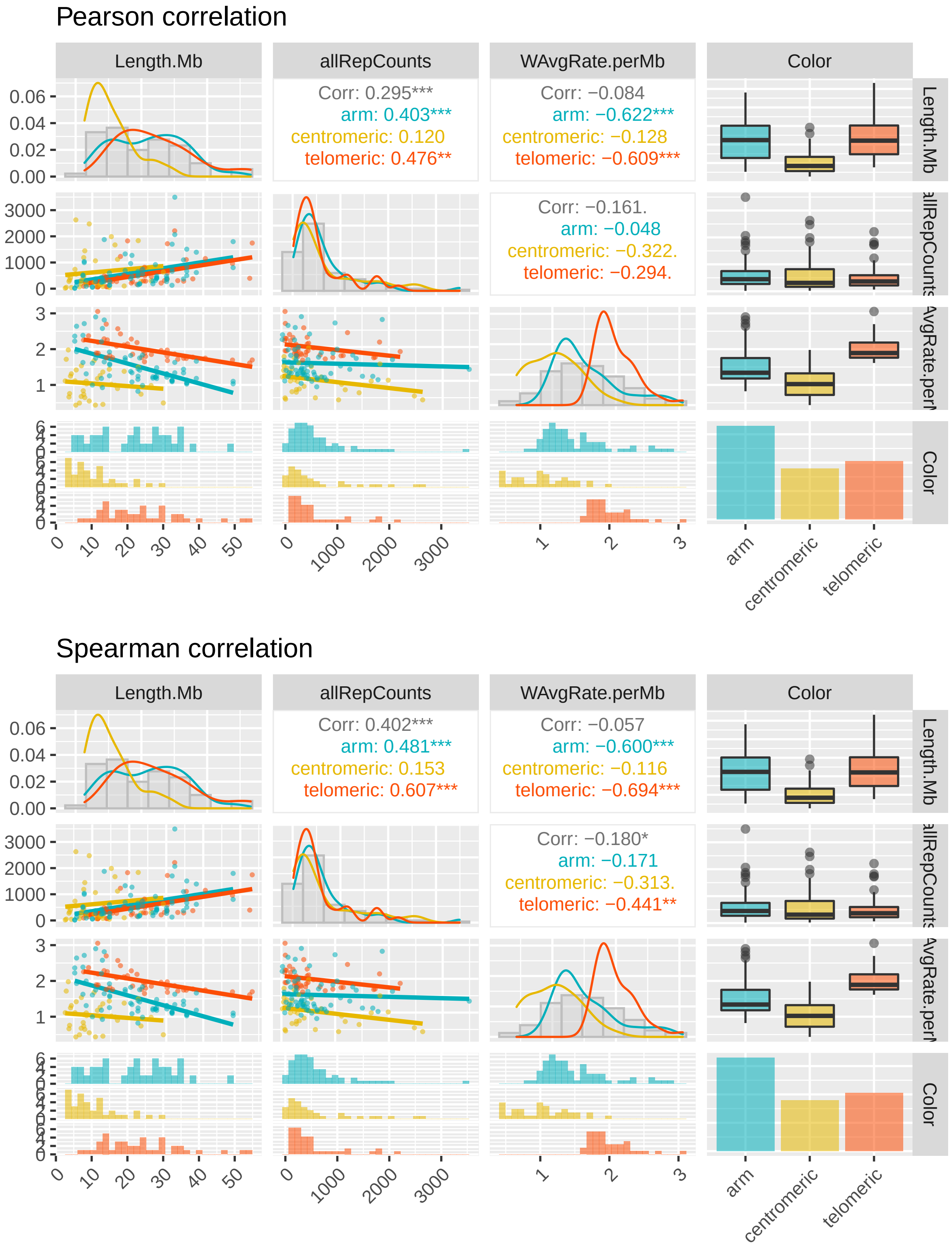


Figure 5: Correlations between variables.

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	1.954368	1	1.397987
allRepCounts	1.145729	1	1.070387
Color	3.035963	2	1.320001
WAvgRate.perMb	2.327808	1	1.525716

	GVIF	Df	GVIF^(1/(2*Df))
scale(Length.Mb)	1.954368	1	1.397987
scale(allRepCounts)	1.145729	1	1.070387
Color	3.035963	2	1.320001
scale(WAvgRate.perMb)	2.327808	1	1.525716

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	WAvgRate.perMb	WAvgRate.perMb.Scaled
Min.	2.694933	-1.4999406	16.0000	-0.9652404	0.4356883	-1.9908973
1st Qu.	10.882125	-0.7805224	215.0000	-0.6373992	1.1289521	-0.7351501
Median	18.633361	-0.0994121	396.0000	-0.3392120	1.4993333	-0.0642579
Mean	19.764700	0.0000000	601.9021	0.0000000	1.5348083	0.0000000
3rd Qu.	27.405822	0.6714345	740.0000	0.2275084	1.8756278	0.6173452
Max.	53.232426	2.9408488	3494.0000	4.7645668	3.0518090	2.7478277

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

Not scaled variables

Total inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error t value
## Length.Mb         0.0669211  0.0199802  3.3494
## allRepCounts      0.0003189  0.0003028  1.0532
## Colorcentromeric  0.2064001  0.5543936  0.3723
## Colortelomeric    0.2923966  0.4558337  0.6415
## WAvgRate.perMb   -0.3980183  0.4604821 -0.8644
##
## Intercepts:
##      Value   Std. Error t value
## 0|1   0.7183  1.0006     0.7179
## 1|2   2.4966  1.0247     2.4366
## 2|3+  3.5469  1.0669     3.3246
##
## Residual Deviance: 315.9039
## AIC: 331.9039
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0669211	0.0199802	3.3493779	0.0008099
allRepCounts	0.0003189	0.0003028	1.0532083	0.2922455
Colorcentromeric	0.2064001	0.5543936	0.3722988	0.7096704
Colortelomeric	0.2923966	0.4558337	0.6414546	0.5212274
WAvgRate.perMb	-0.3980183	0.4604821	-0.8643513	0.3873950
0|1	0.7183365	1.0006113	0.7178976	0.4728204
1|2	2.4966468	1.0246584	2.4365649	0.0148275
2|3+	3.5469431	1.0668854	3.3245774	0.0008855

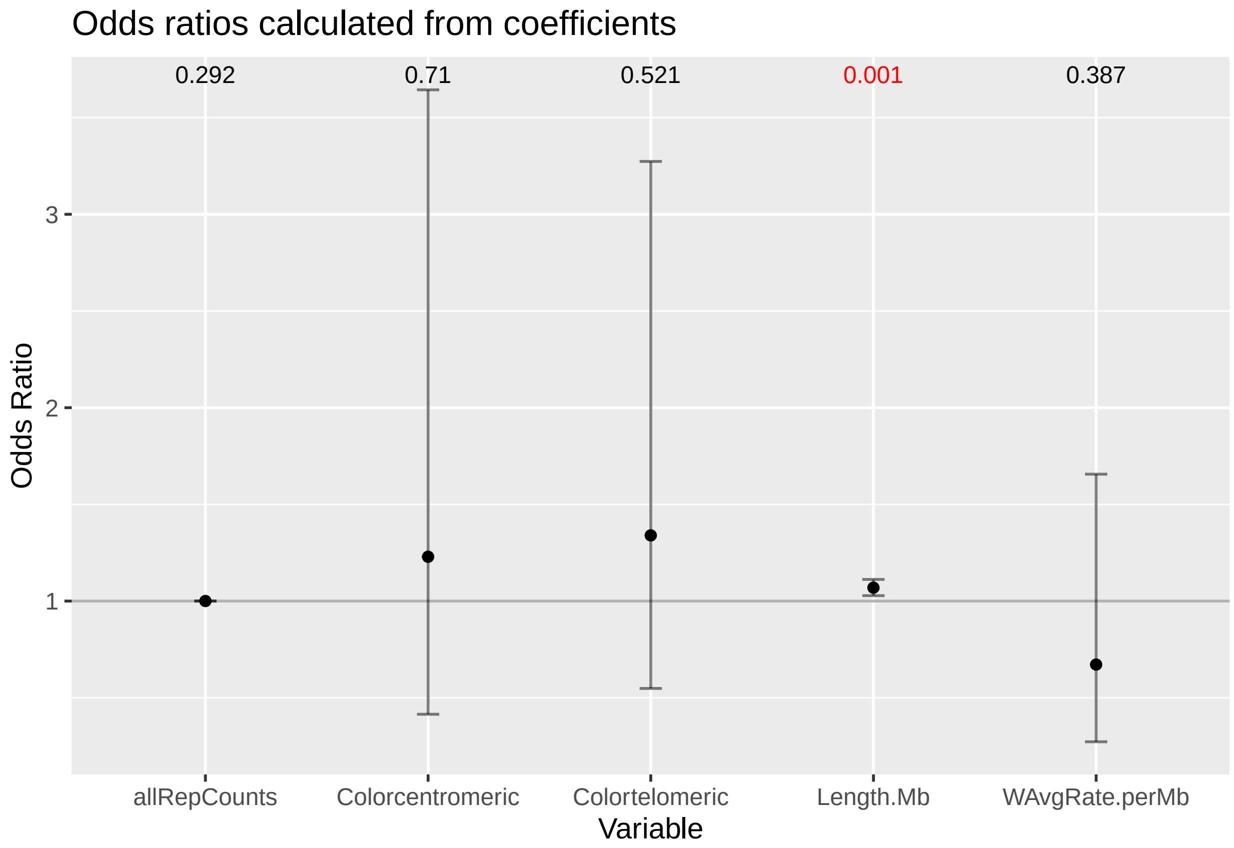
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0277607	0.1060815
allRepCounts	-0.0002745	0.0009122
Colorcentromeric	-0.8801914	1.2929915
Colortelomeric	-0.6010210	1.1858142
WAvgRate.perMb	-1.3005465	0.5045100

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	1.0692111	1.0281496	1.111913
allRepCounts	1.0003189	0.9997255	1.000913
Colorcentromeric	1.2292449	0.4147035	3.643670
Colortelomeric	1.3396342	0.5482516	3.273351
WAvgRate.perMb	0.6716498	0.2723829	1.656174

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0692111 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.

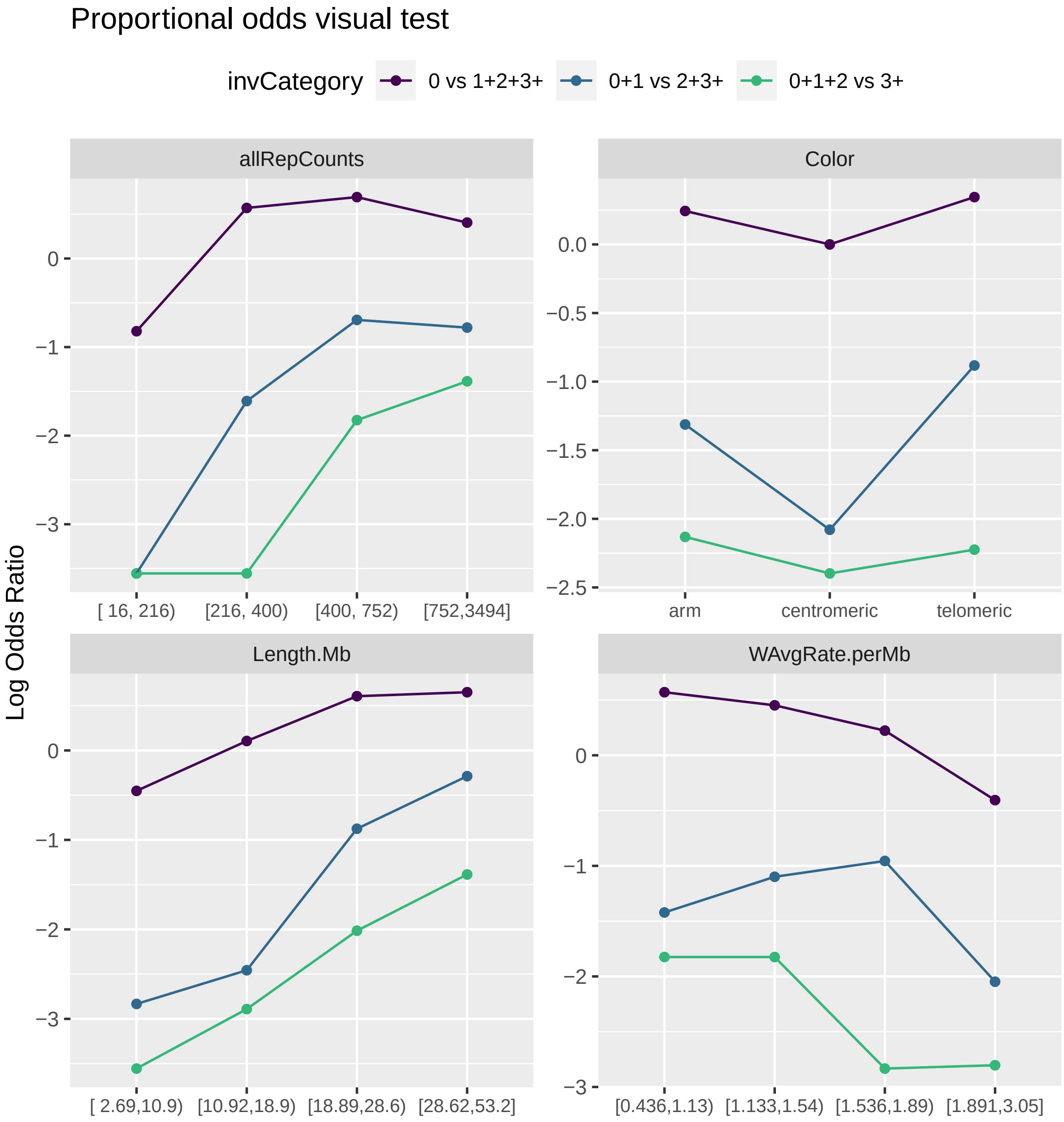
We test the parallel regression assumption with a Brant test:

```
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          14.42   10  0.15
## Length.Mb        6.35    2   0.04
## allRepCounts     0.02    2   0.99
## Colorcentromeric 0.6 2   0.74
## Colortelomeric       0.31    2   0.86
## WAvgRate.perMb       0.64    2   0.73
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	14.4229761	10	0.1545577
Length.Mb	6.3494544	2	0.0418055
allRepCounts	0.0176935	2	0.9911923
Colorcentromeric	0.6021317	2	0.7400291
Colortelomeric	0.3089363	2	0.8568708
WAvgRate.perMb	0.6426880	2	0.7251737

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.

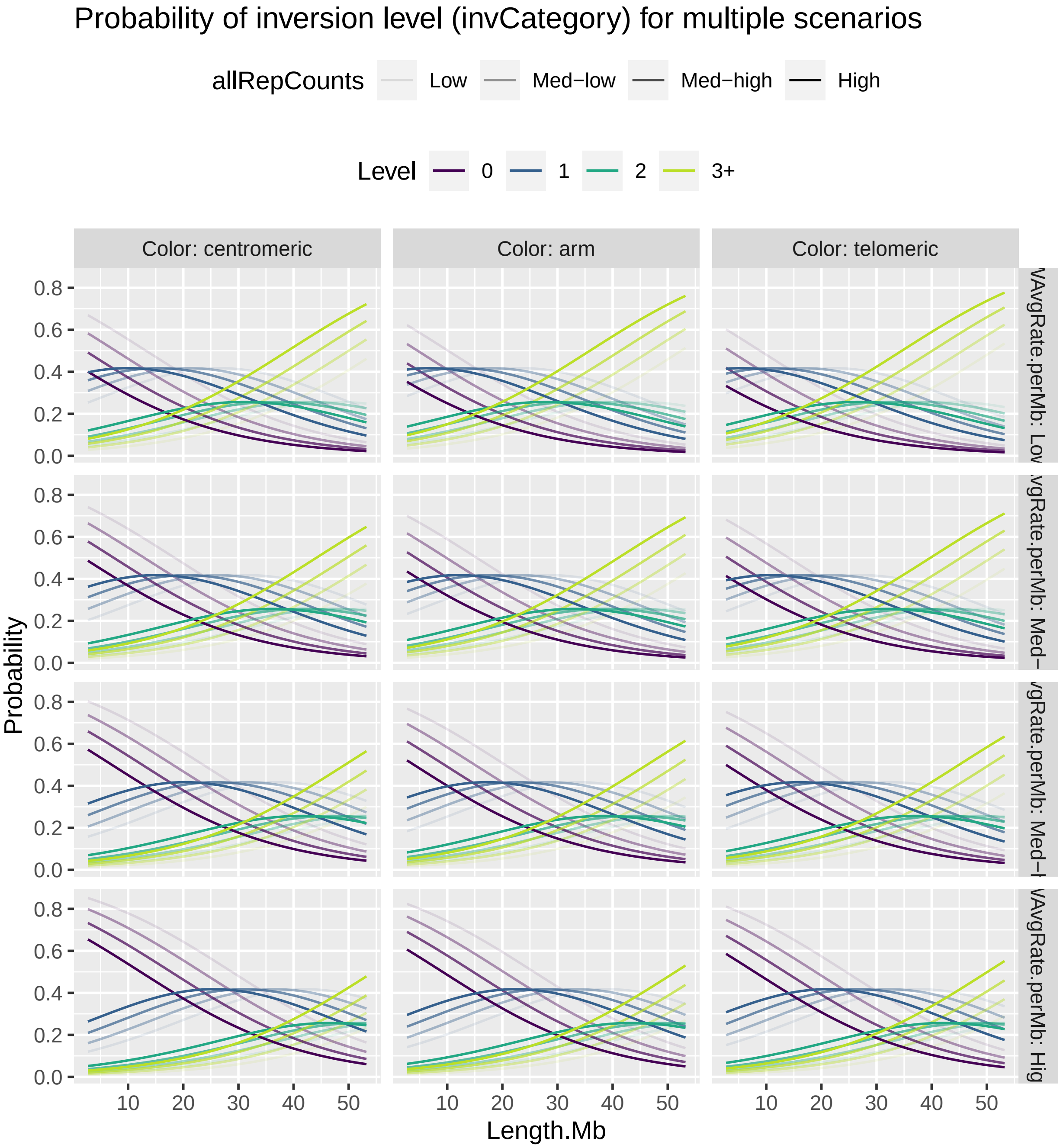


Figure 6: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error t value
## Length.Mb         0.0905463  0.0226136  4.0041
## allRepCounts     -0.0003138  0.0003485 -0.9004
## Colorcentromeric  0.2866621  0.6118505  0.4685
## Colortelomeric   -0.1733296  0.5213094 -0.3325
## WAvgRate.perMb   -0.2005601  0.5460425 -0.3673
##
## Intercepts:
##      Value   Std. Error t value
## 0|1   1.8229  1.1522     1.5821
## 1|2   3.6664  1.1974     3.0619
## 2|3+  5.0537  1.2851     3.9326
##
## Residual Deviance: 249.2127
## AIC: 265.2127
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0905463	0.0226136	4.0040606	0.0000623
allRepCounts	-0.0003138	0.0003485	-0.9003647	0.3679262
Colorcentromeric	0.2866621	0.6118505	0.4685166	0.6394152
Colortelomeric	-0.1733296	0.5213094	-0.3324889	0.7395201
WAvgRate.perMb	-0.2005601	0.5460425	-0.3672977	0.7133970
0|1	1.8229281	1.1522412	1.5820716	0.1136332
1|2	3.6664098	1.1974478	3.0618536	0.0021997
2|3+	5.0536669	1.2850747	3.9325862	0.0000840

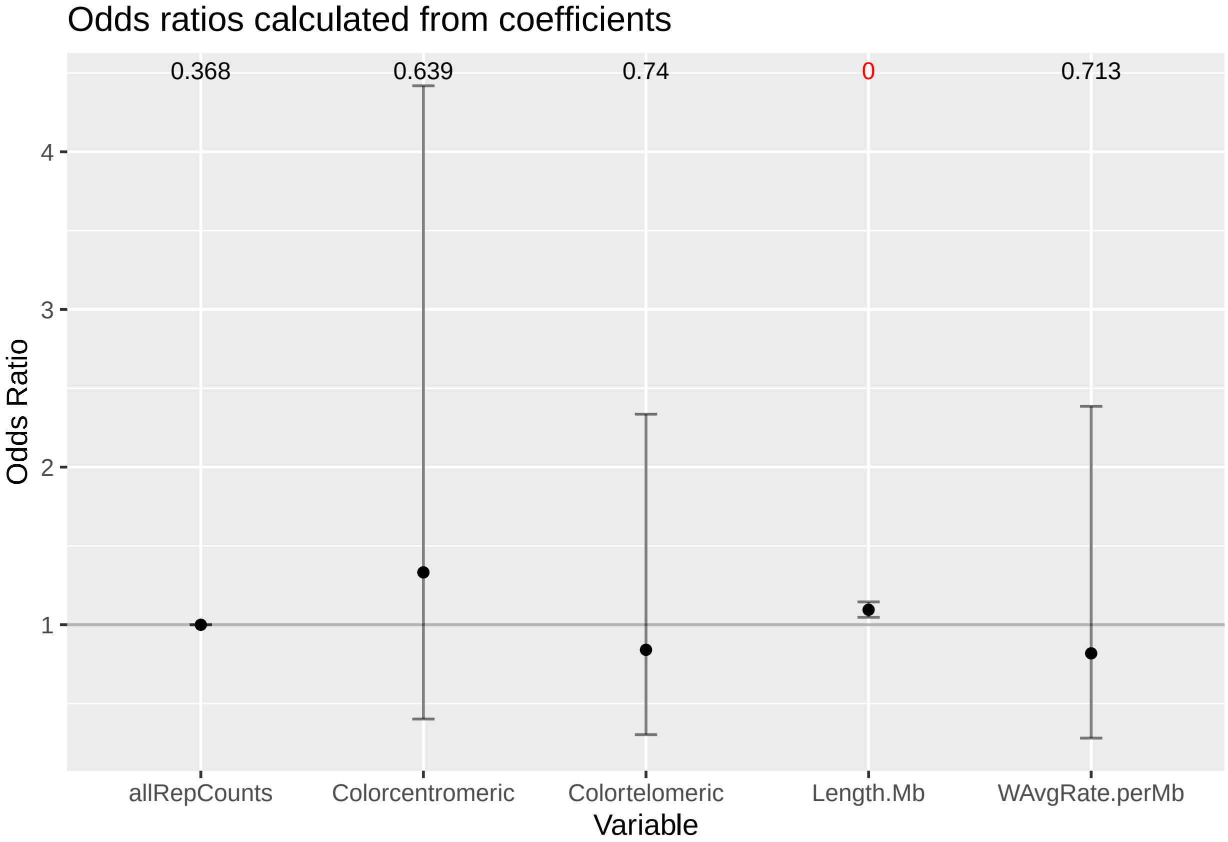
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0462244	0.1348682
allRepCounts	-0.0009968	0.0003693
Colorcentromeric	-0.9125429	1.4858671
Colortelomeric	-1.1950772	0.8484181
WAvgRate.perMb	-1.2707837	0.8696634

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	1.0947722	1.0473094	1.144386
allRepCounts	0.9996863	0.9990037	1.000369
Colorcentromeric	1.3319741	0.4015019	4.418795
Colortelomeric	0.8408604	0.3026806	2.335949
WAvgRate.perMb	0.8182723	0.2806116	2.386108

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0947722 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.

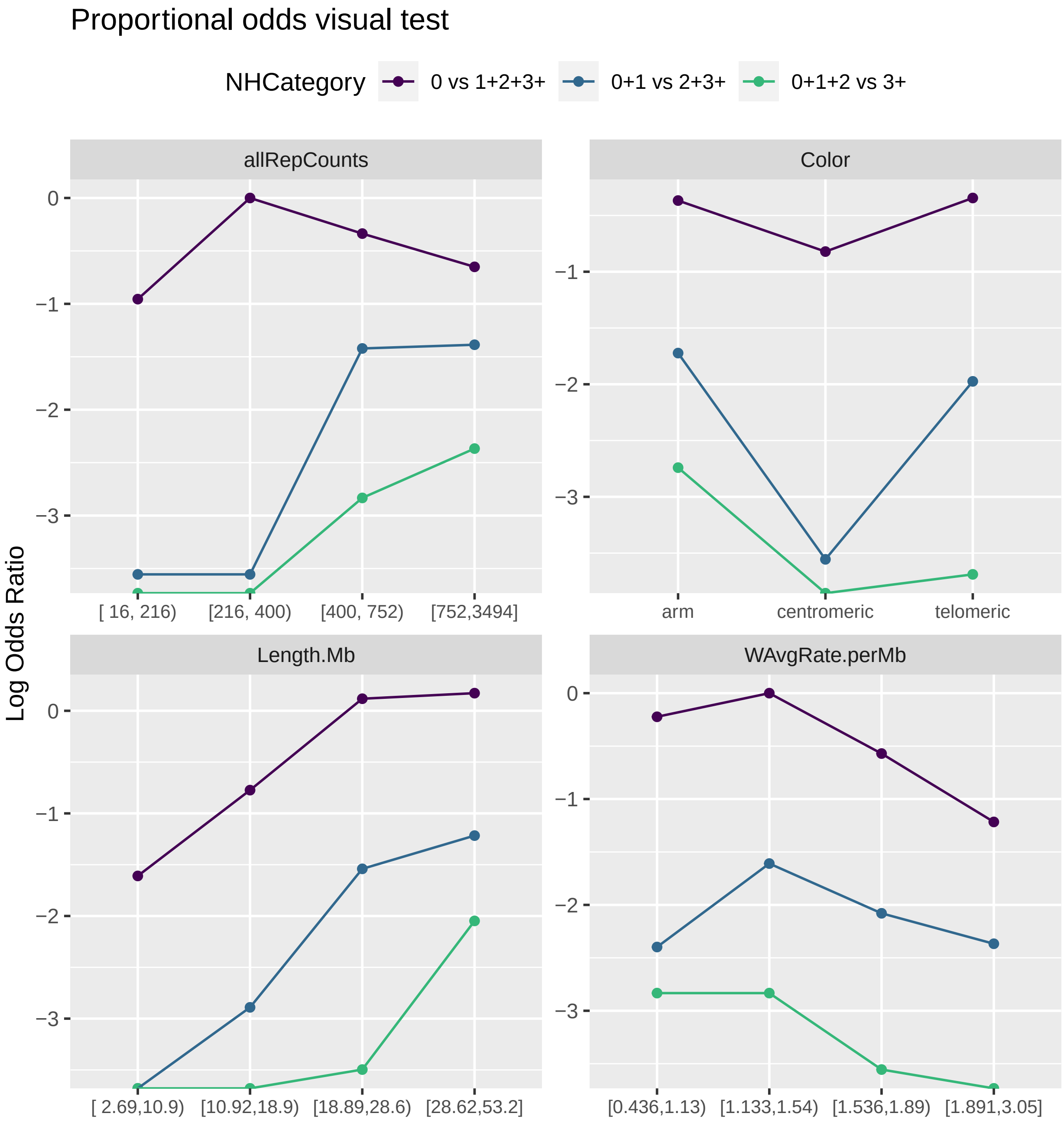
We test the parallel regression assumption with a Brant test:

```
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ----------------------------------------------------
## Test for      X2   df   probability
## ----------------------------------------------------
## Omnibus          19.8    10   0.03
## Length.Mb        3.78    2   0.15
## allRepCounts     1.04    2   0.59
## Colorcentromeric 0.09    2   0.96
## Colortelomeric       5   2   0.08
## WAvgRate.perMb       8.74    2   0.01
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	19.7974228	10	0.0312280
Length.Mb	3.7845530	2	0.1507283
allRepCounts	1.0407328	2	0.5943028
Colorcentromeric	0.0860027	2	0.9579101
Colortelomeric	5.0049955	2	0.0818802
WAvgRate.perMb	8.7382757	2	0.0126622

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.

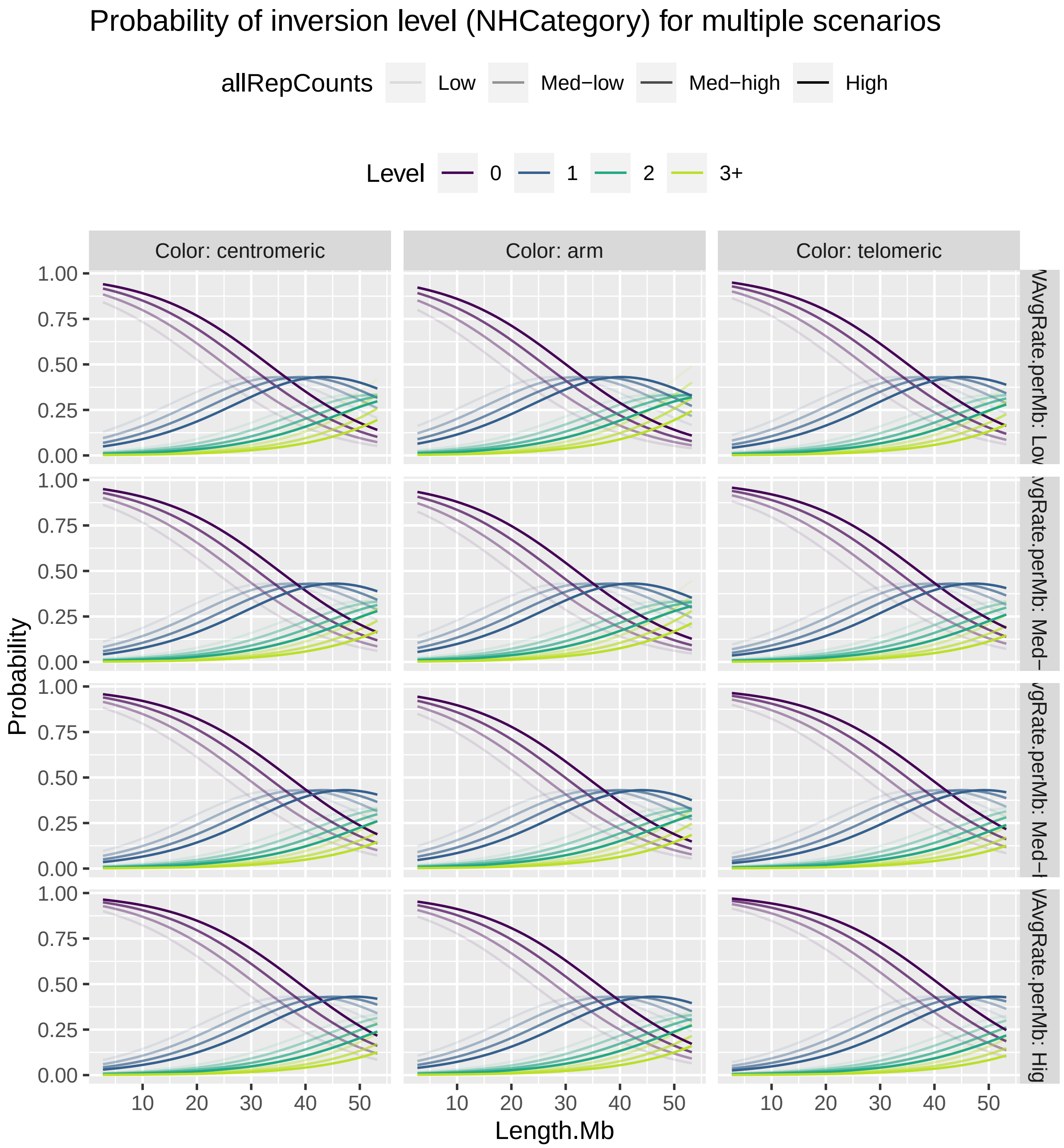


Figure 7: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error t value
## Length.Mb         0.0152104  0.0229120  0.6639
## allRepCounts      0.0007402  0.0003533  2.0948
## Colorcentromeric  0.2499320  0.7000256  0.3570
## Colortelomeric    0.5853430  0.5419634  1.0800
## WAvgRate.perMb   -0.2612893  0.5789971 -0.4513
##
## Intercepts:
##      Value   Std. Error t value
## 0|1   1.6442  1.2490     1.3164
## 1|2   3.4249  1.2954     2.6439
## 2|3+  4.6047  1.3796     3.3377
##
## Residual Deviance: 208.5344
## AIC: 224.5344
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0152104	0.0229120	0.6638604	0.5067796
allRepCounts	0.0007402	0.0003533	2.0948219	0.0361868
Colorcentromeric	0.2499320	0.7000256	0.3570327	0.7210674
Colortelomeric	0.5853430	0.5419634	1.0800416	0.2801236
WAvgRate.perMb	-0.2612893	0.5789971	-0.4512792	0.6517883
0|1	1.6442192	1.2489920	1.3164369	0.1880274
1|2	3.4248792	1.2954031	2.6438714	0.0081964
2|3+	4.6047002	1.3796089	3.3376852	0.0008448

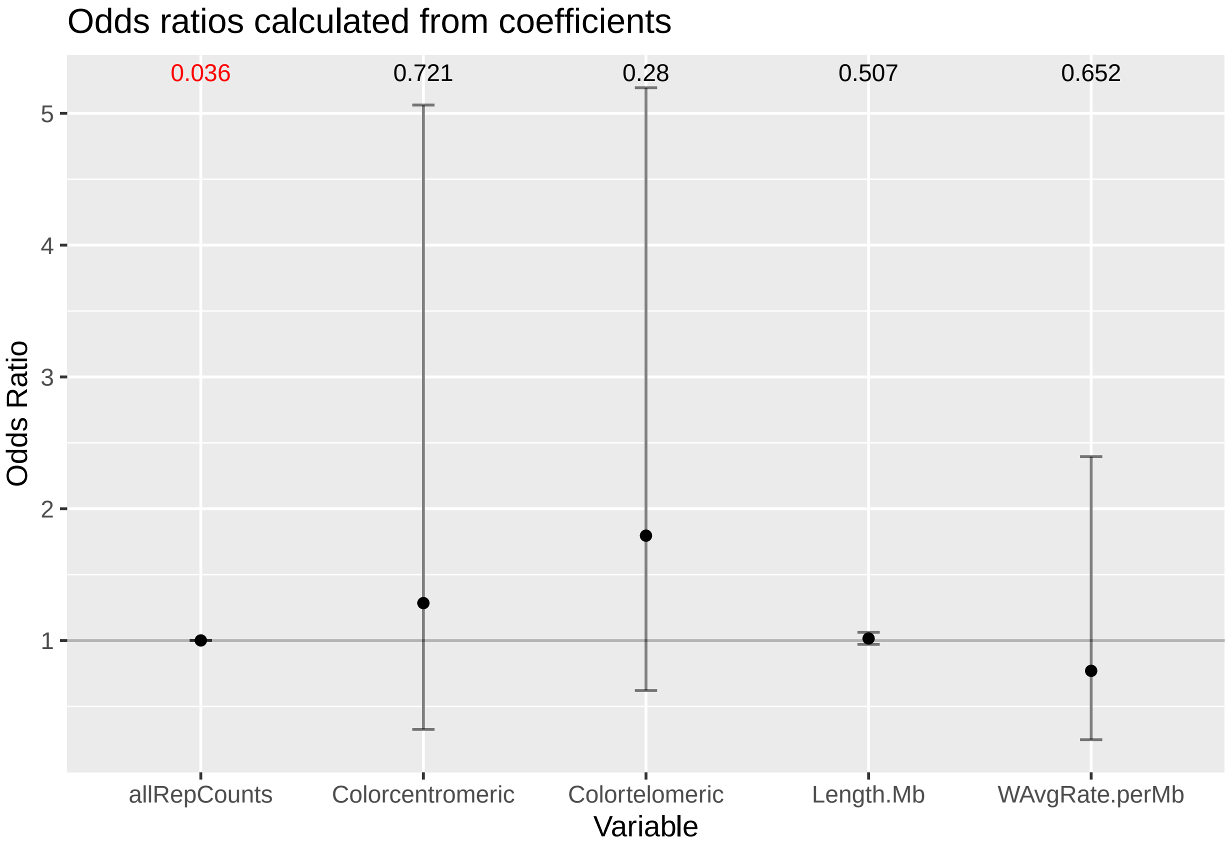
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	-0.0296963	0.0601171
allRepCounts	0.0000477	0.0014327
Colorcentromeric	-1.1220930	1.6219570
Colortelomeric	-0.4768857	1.6475717
WAvgRate.perMb	-1.3961028	0.8735241

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	1.0153266	0.9707403	1.061961
allRepCounts	1.0007405	1.0000477	1.001434
Colorcentromeric	1.2839381	0.3255976	5.062989
Colortelomeric	1.7956068	0.6207135	5.194351
WAvgRate.perMb	0.7700581	0.2475599	2.395338

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0153266 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.

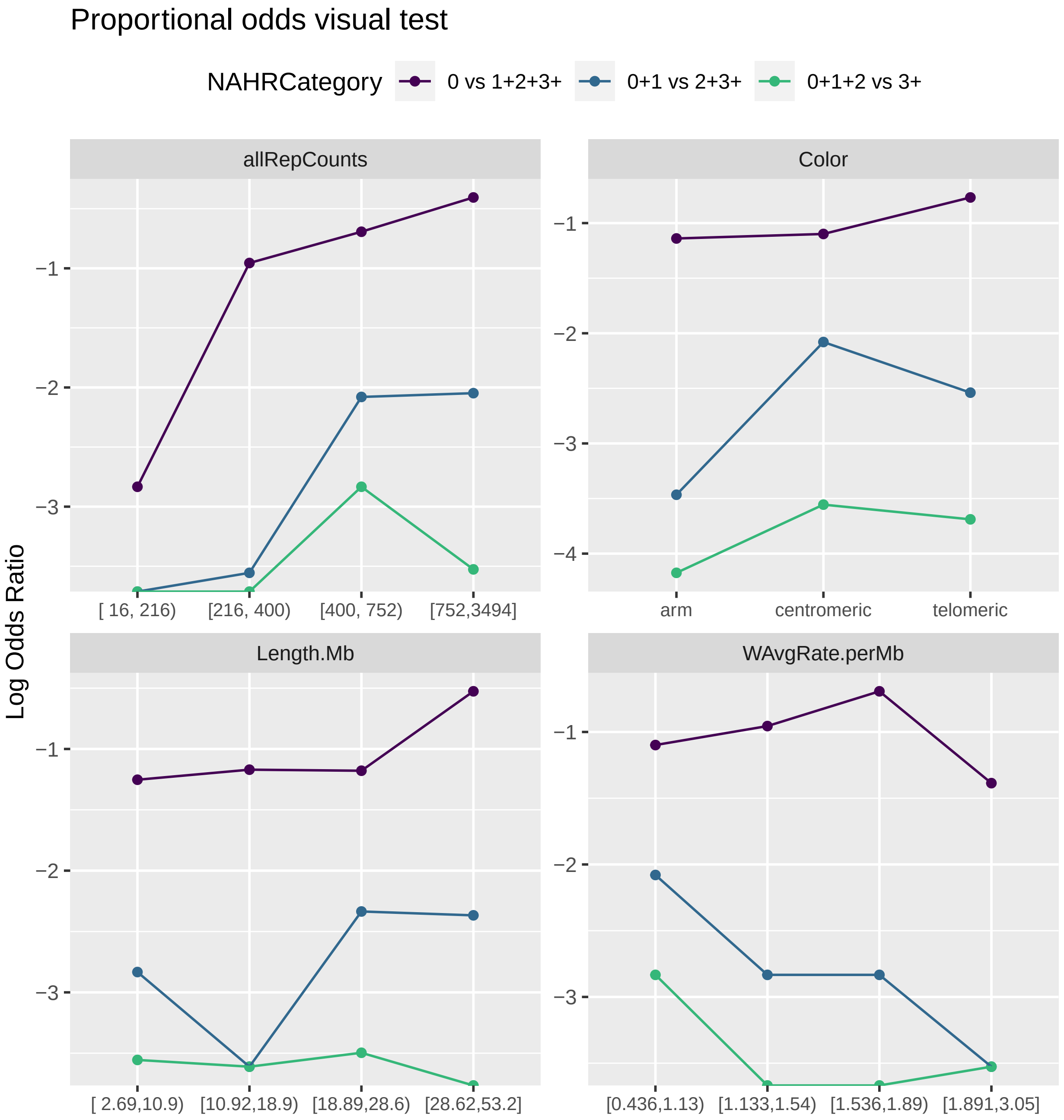
We test the parallel regression assumption with a Brant test:

```
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ----------------------------------------------------
## Test for      X2  df  probability
## ----------------------------------------------------
## Omnibus          15.26   10  0.12
## Length.Mb        2.33    2   0.31
## allRepCounts     0.87    2   0.65
## Colorcentromeric 4.24    2   0.12
## Colortelomeric       2.83    2   0.24
## WAvgRate.perMb       3.67    2   0.16
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	15.2555799	10	0.1230180
Length.Mb	2.3318105	2	0.3116404
allRepCounts	0.8707161	2	0.6470330
Colorcentromeric	4.2404296	2	0.1200058
Colortelomeric	2.8291963	2	0.2430233
WAvgRate.perMb	3.6735618	2	0.1593295

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.

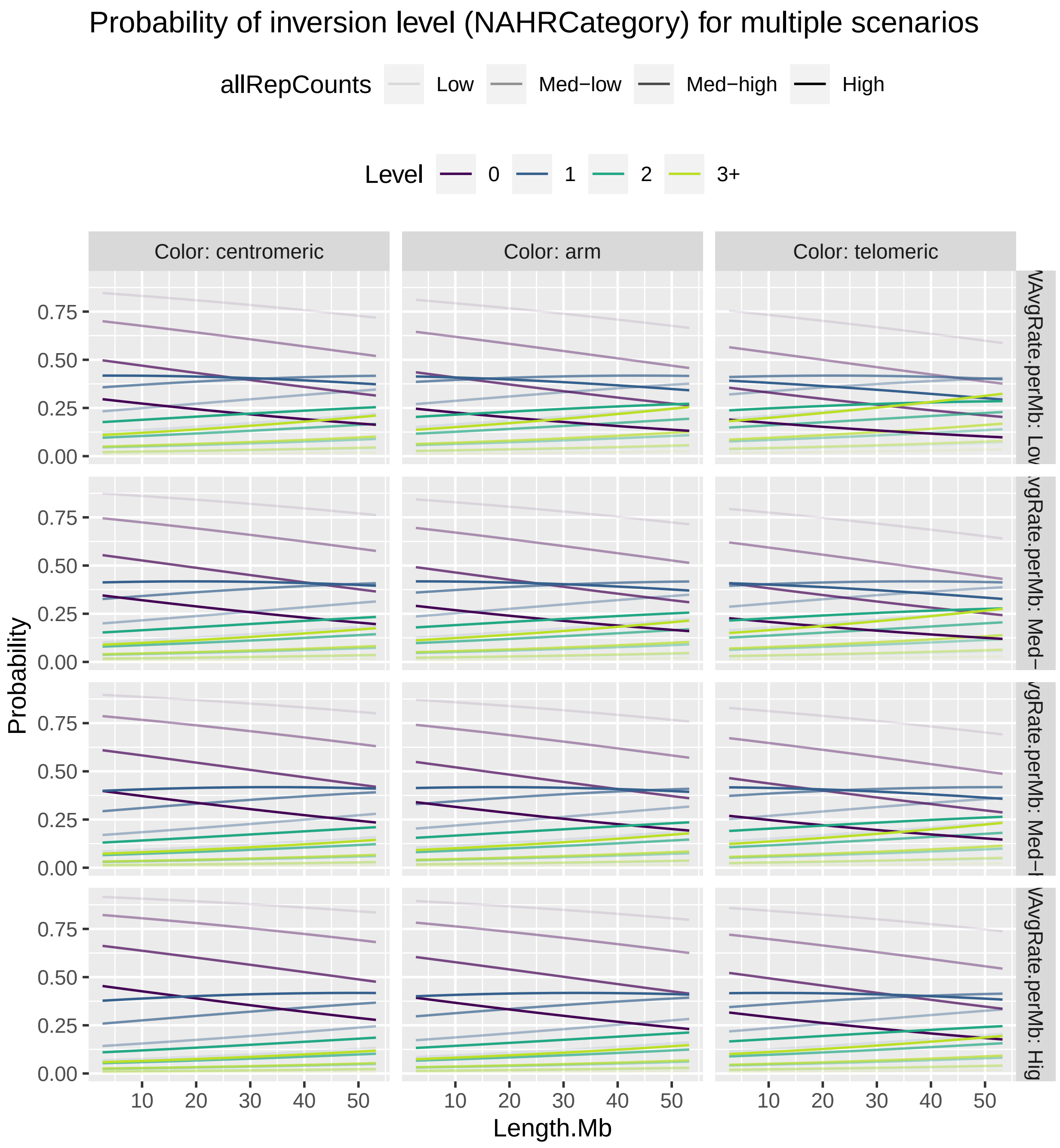


Figure 8: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

Scaled variables

Total inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                          Value Std. Error t value
## Length.Mb.Scaled        0.7616     0.2271  3.3540
## allRepCounts.Scaled     0.1936     0.1717  1.1270
## Colorcentromeric        0.2063     0.5546  0.3720
## Colortelomeric          0.2925     0.4556  0.6419
## WAvgRate.perMb.Scaled -0.2198     0.2542 -0.8644
##
## Intercepts:
##      Value   Std. Error t value
## 0|1  -0.1854  0.2545    -0.7284
## 1|2   1.5929  0.2897     5.4981
## 2|3+  2.6432  0.3638     7.2646
##
## Residual Deviance: 315.9039
## AIC: 331.9039
```

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	0.7615620	0.2270616	3.3539881	0.0007966
allRepCounts.Scaled	0.1935553	0.1717393	1.1270297	0.2597299
Colorcentromeric	0.2063248	0.5546409	0.3719971	0.7098950
Colortelomeric	0.2924531	0.4556336	0.6418603	0.5209639
WAvgRate.perMb.Scaled	-0.2197687	0.2542300	-0.8644485	0.3873416
0|1	-0.1853984	0.2545331	-0.7283861	0.4663773
1|2	1.5929267	0.2897230	5.4981024	0.0000000
2|3+	2.6432067	0.3638460	7.2646311	0.0000000

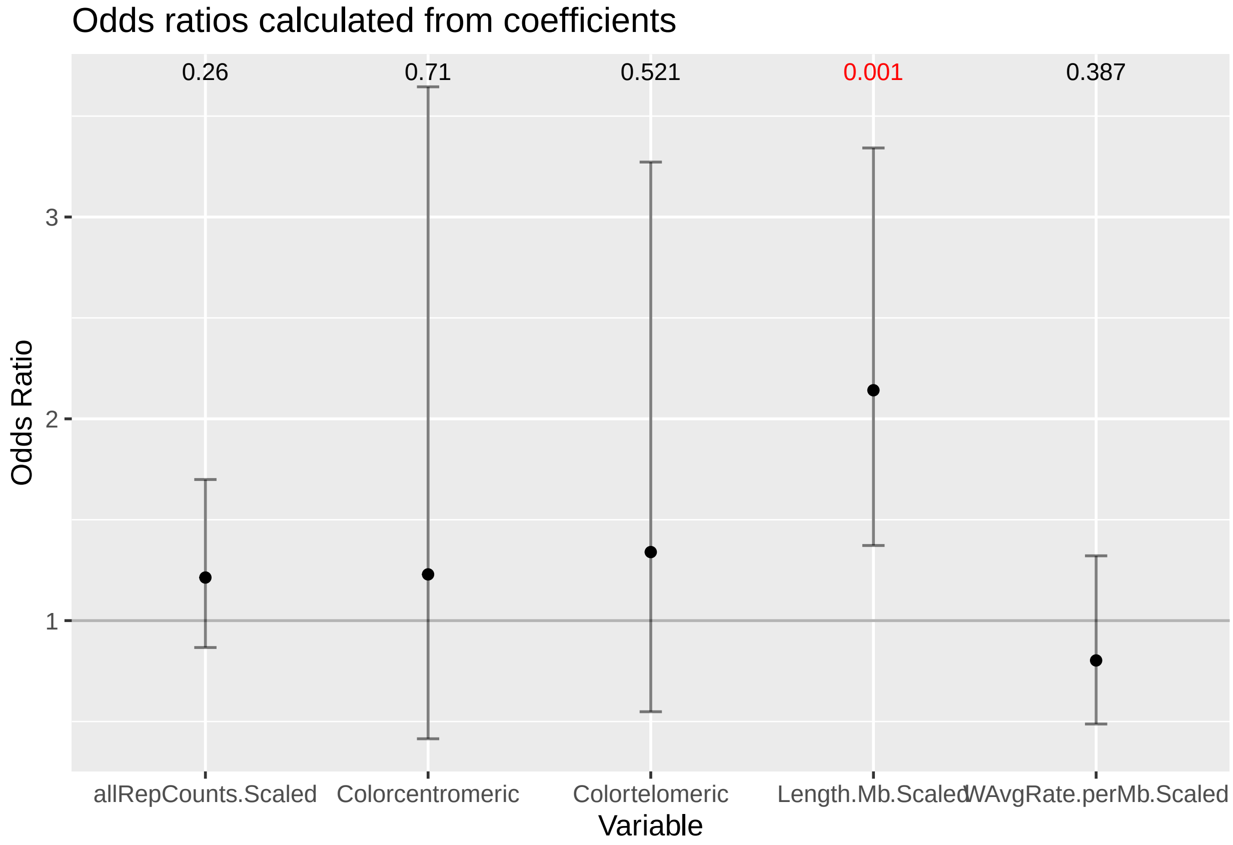
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	0.3165294	1.2065945
allRepCounts.Scaled	-0.1430476	0.5301581
Colorcentromeric	-0.8807514	1.2934010
Colortelomeric	-0.6005723	1.1854785
WAvgRate.perMb.Scaled	-0.7180504	0.2785129

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	2.1416187	1.3723565	3.342084
allRepCounts.Scaled	1.2135565	0.8667129	1.699201
Colorcentromeric	1.2291524	0.4144714	3.645163
Colortelomeric	1.3397099	0.5484977	3.272252
WAvgRate.perMb.Scaled	0.8027044	0.4877022	1.321164

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 2.1416187 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.

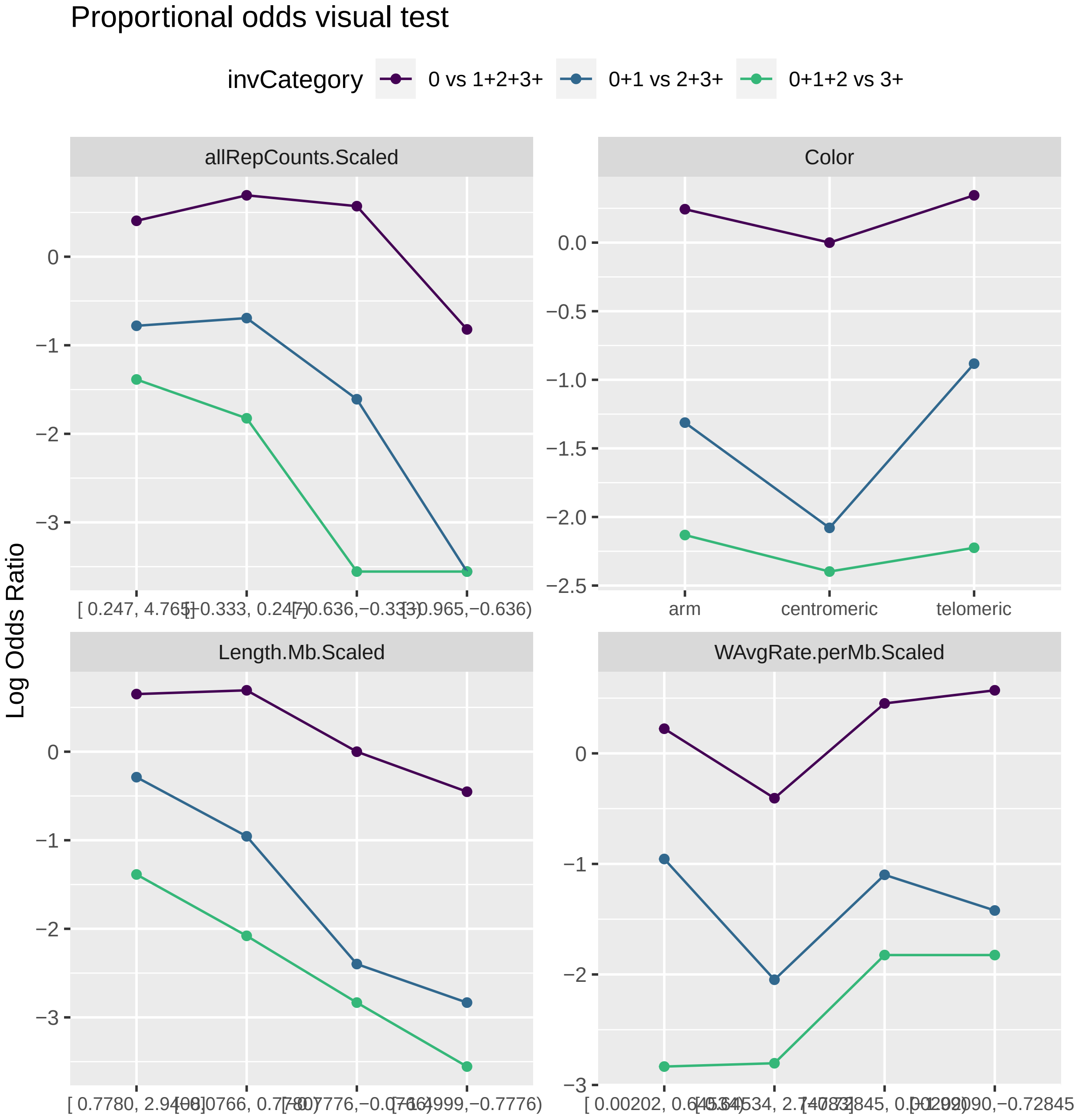
We test the parallel regression assumption with a Brant test:

```
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          14.42   10  0.15
## Length.Mb.Scaled 6.35    2   0.04
## allRepCounts.Scaled  0.02    2   0.99
## Colorcentromeric 0.6 2   0.74
## Colortelomeric       0.31    2   0.86
## WAvgRate.perMb.Scaled    0.64    2   0.73
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	14.4229761	10	0.1545577
Length.Mb.Scaled	6.3494544	2	0.0418055
allRepCounts.Scaled	0.0176935	2	0.9911923
Colorcentromeric	0.6021317	2	0.7400291
Colortelomeric	0.3089363	2	0.8568708
WAvgRate.perMb.Scaled	0.6426880	2	0.7251737

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.

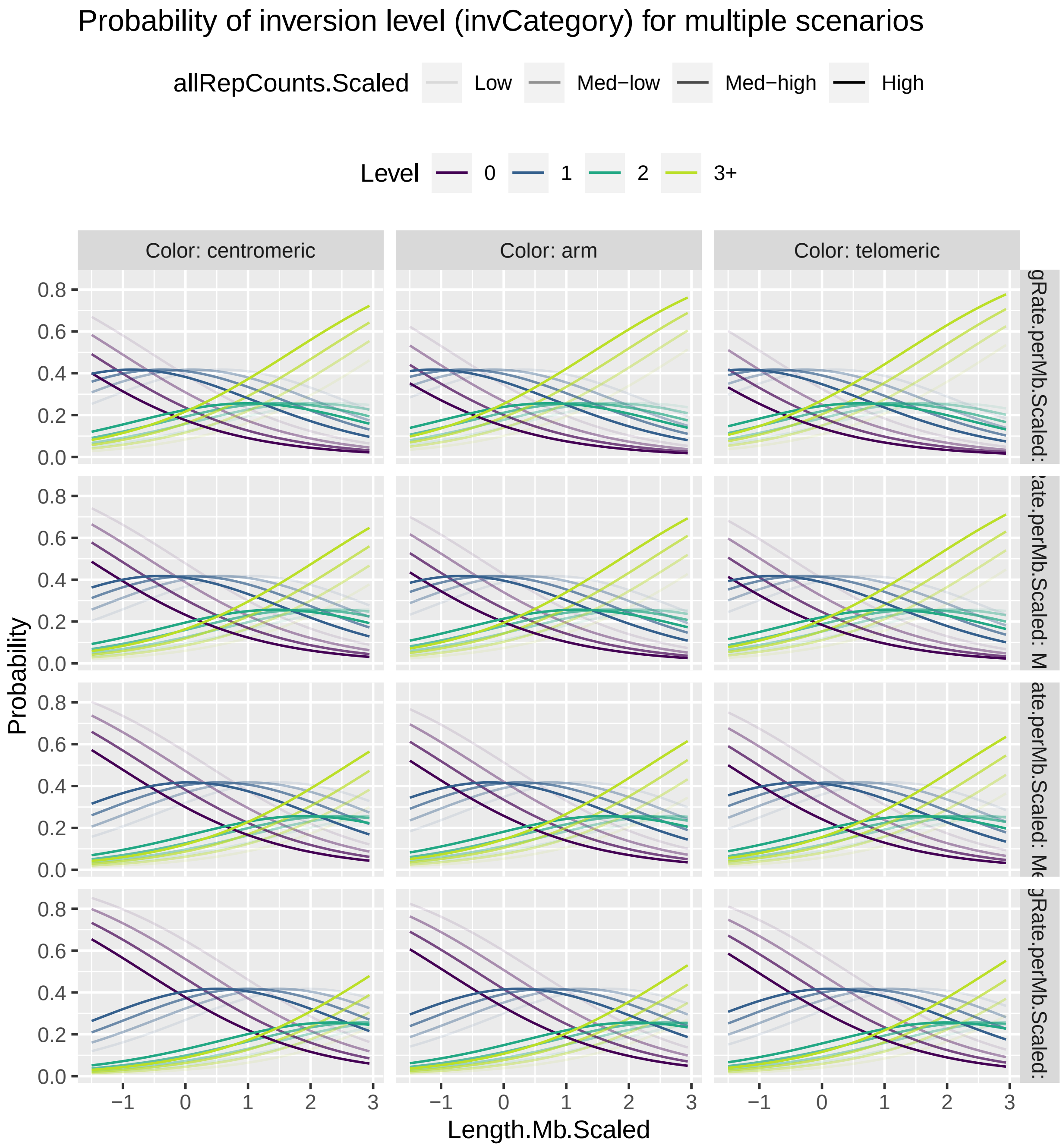


Figure 9: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                        Value Std. Error t value
## Length.Mb.Scaled       1.0304     0.2579  3.9951
## allRepCounts.Scaled   -0.1905     0.1950 -0.9770
## Colorcentromeric       0.2867     0.6114  0.4689
## Colortelomeric        -0.1733     0.5215 -0.3324
## WAvgRate.perMb.Scaled -0.1107     0.3015 -0.3673
##
## Intercepts:
##      Value   Std. Error t value
## 0|1   0.5300  0.2756     1.9232
## 1|2   2.3735  0.3564     6.6595
## 2|3+  3.7607  0.5268     7.1382
##
## Residual Deviance: 249.2127
## AIC: 265.2127
```

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	1.0304489	0.2579283	3.9950978	0.0000647
allRepCounts.Scaled	-0.1904647	0.1949515	-0.9769853	0.3285764
Colorcentromeric	0.2866821	0.6114006	0.4688940	0.6391454
Colortelomeric	-0.1733377	0.5215118	-0.3323755	0.7396058
WAvgRate.perMb.Scaled	-0.1107188	0.3014652	-0.3672691	0.7134183
0|1	0.5299767	0.2755762	1.9231586	0.0544601
1|2	2.3734664	0.3564052	6.6594611	0.0000000
2|3+	3.7607186	0.5268459	7.1381762	0.0000000

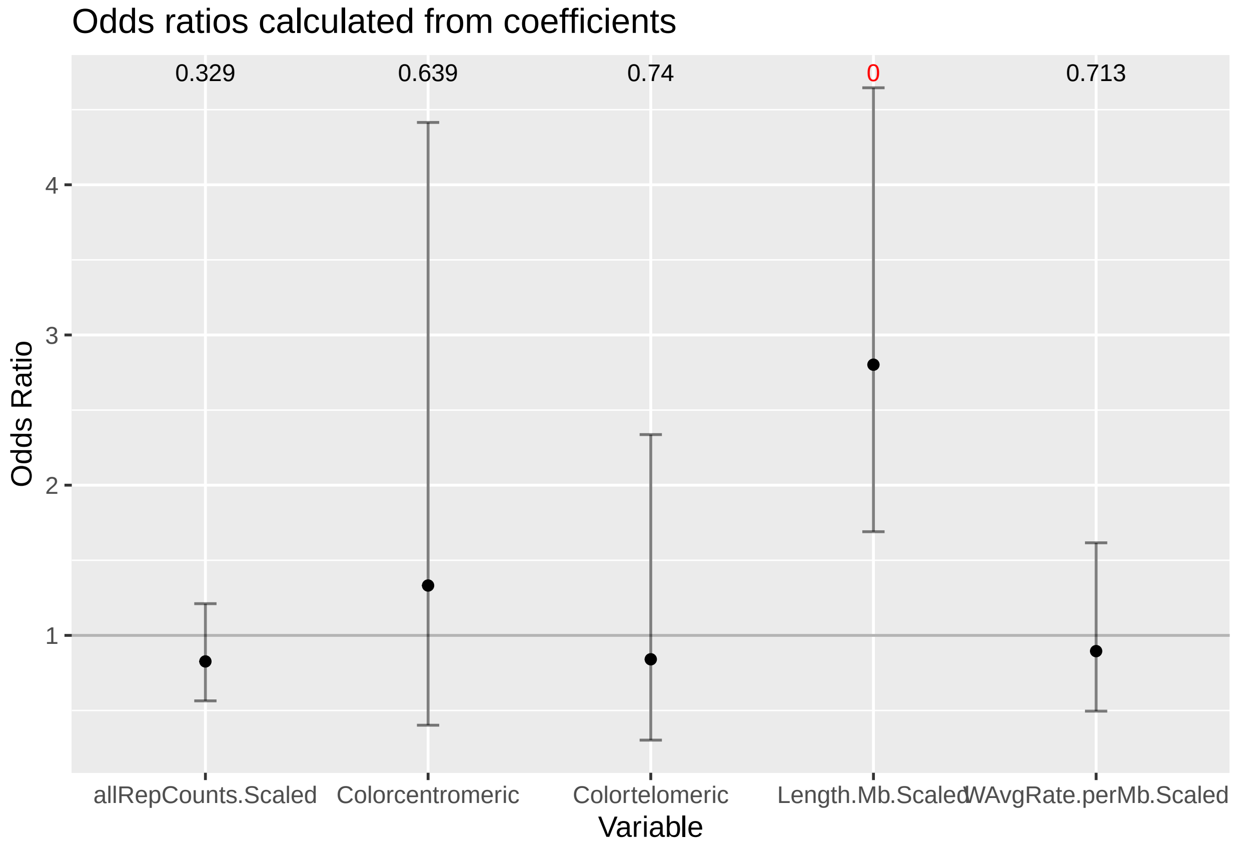
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	0.5249187	1.5359791
allRepCounts.Scaled	-0.5725626	0.1916332
Colorcentromeric	-0.9116412	1.4850053
Colortelomeric	-1.1954821	0.8488066
WAvgRate.perMb.Scaled	-0.7015797	0.4801420

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	2.8023234	1.6903213	4.645872
allRepCounts.Scaled	0.8265749	0.5640781	1.211226
Colorcentromeric	1.3320007	0.4018642	4.414989
Colortelomeric	0.8408536	0.3025581	2.336856
WAvgRate.perMb.Scaled	0.8951904	0.4958015	1.616304

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 2.8023234 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.

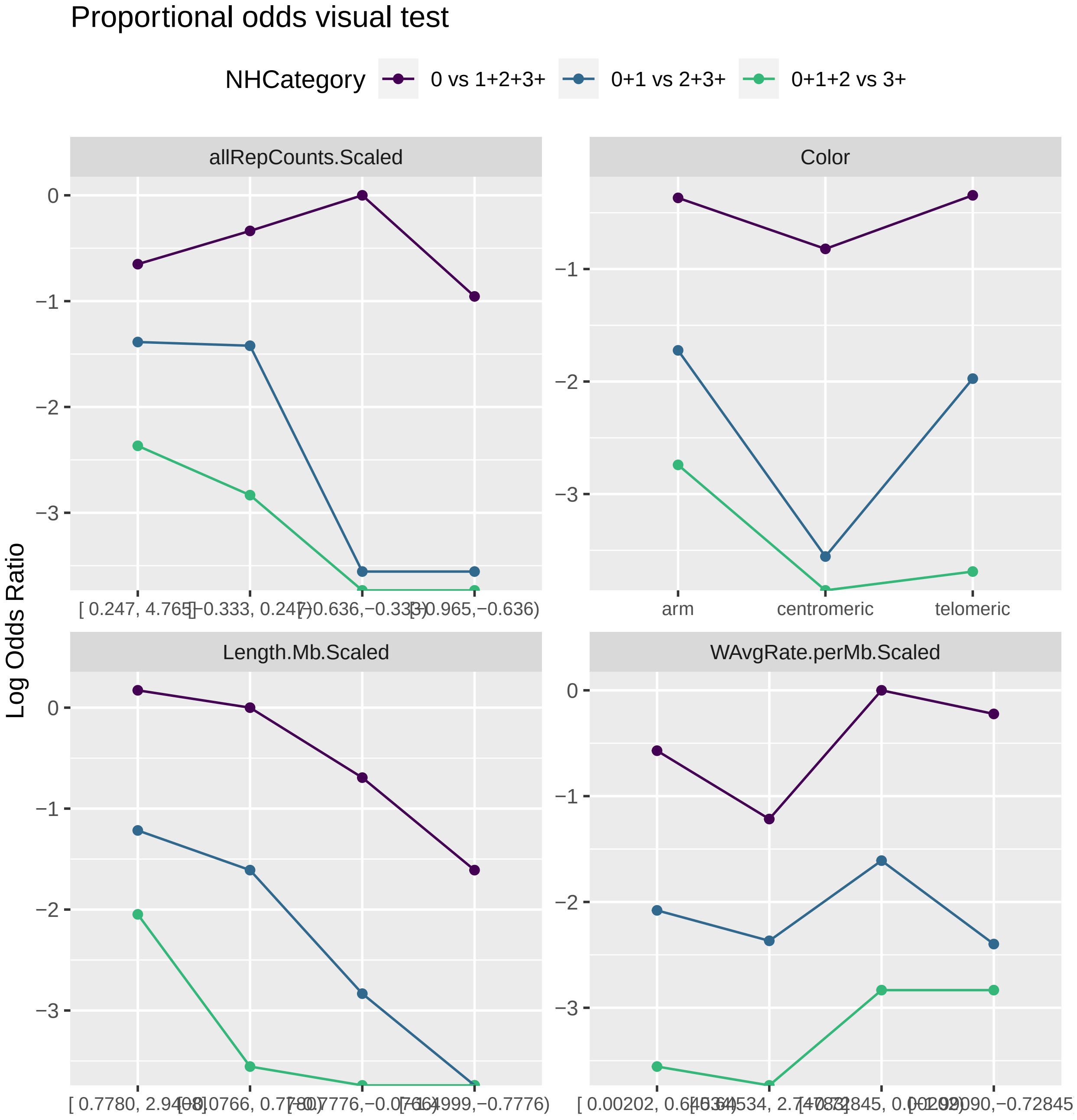
We test the parallel regression assumption with a Brant test:

```
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ----------------------------------------------------
## Test for      X2  df  probability
## ----------------------------------------------------
## Omnibus           19.8    10  0.03
## Length.Mb.Scaled 3.78    2   0.15
## allRepCounts.Scaled  1.04    2   0.59
## Colorcentromeric 0.09    2   0.96
## Colortelomeric       5   2   0.08
## WAvgRate.perMb.Scaled    8.74    2   0.01
## --------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	19.7974228	10	0.0312280
Length.Mb.Scaled	3.7845530	2	0.1507283
allRepCounts.Scaled	1.0407328	2	0.5943028
Colorcentromeric	0.0860027	2	0.9579101
Colortelomeric	5.0049955	2	0.0818802
WAvgRate.perMb.Scaled	8.7382757	2	0.0126622

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.

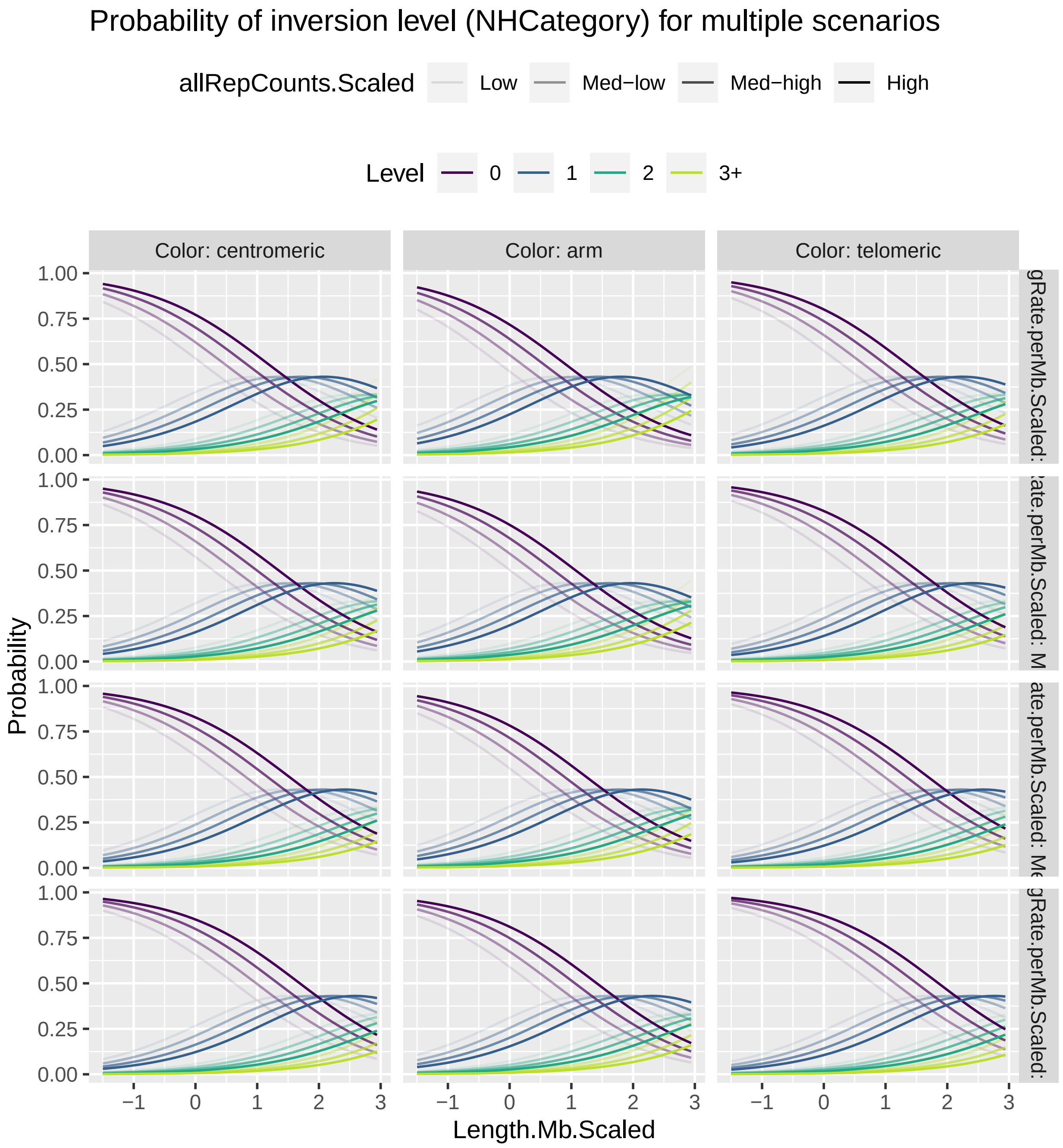


Figure 10: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                          Value Std. Error t value
## Length.Mb.Scaled        0.1730     0.2598  0.6660
## allRepCounts.Scaled     0.4493     0.1943  2.3127
## Colorcentromeric        0.2498     0.6993  0.3573
## Colortelomeric          0.5854     0.5415  1.0811
## WAvgRate.perMb.Scaled -0.1443     0.3196 -0.4515
##
## Intercepts:
##      Value   Std. Error t value
## 0|1   1.2991  0.3068     4.2341
## 1|2   3.0798  0.4361     7.0619
## 2|3+  4.2597  0.6477     6.5772
##
## Residual Deviance: 208.5344
## AIC: 224.5344
```

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	0.1730320	0.2597996	0.6660211	0.5053976
allRepCounts.Scaled	0.4492997	0.1942760	2.3126877	0.0207398
Colorcentromeric	0.2498344	0.6993049	0.3572610	0.7208964
Colortelomeric	0.5854294	0.5415076	1.0811102	0.2796481
WAvgRate.perMb.Scaled	-0.1442900	0.3195845	-0.4514924	0.6516347
0|1	1.2990912	0.3068163	4.2341006	0.0000230
1|2	3.0797966	0.4361166	7.0618657	0.0000000
2|3+	4.2597482	0.6476501	6.5772369	0.0000000

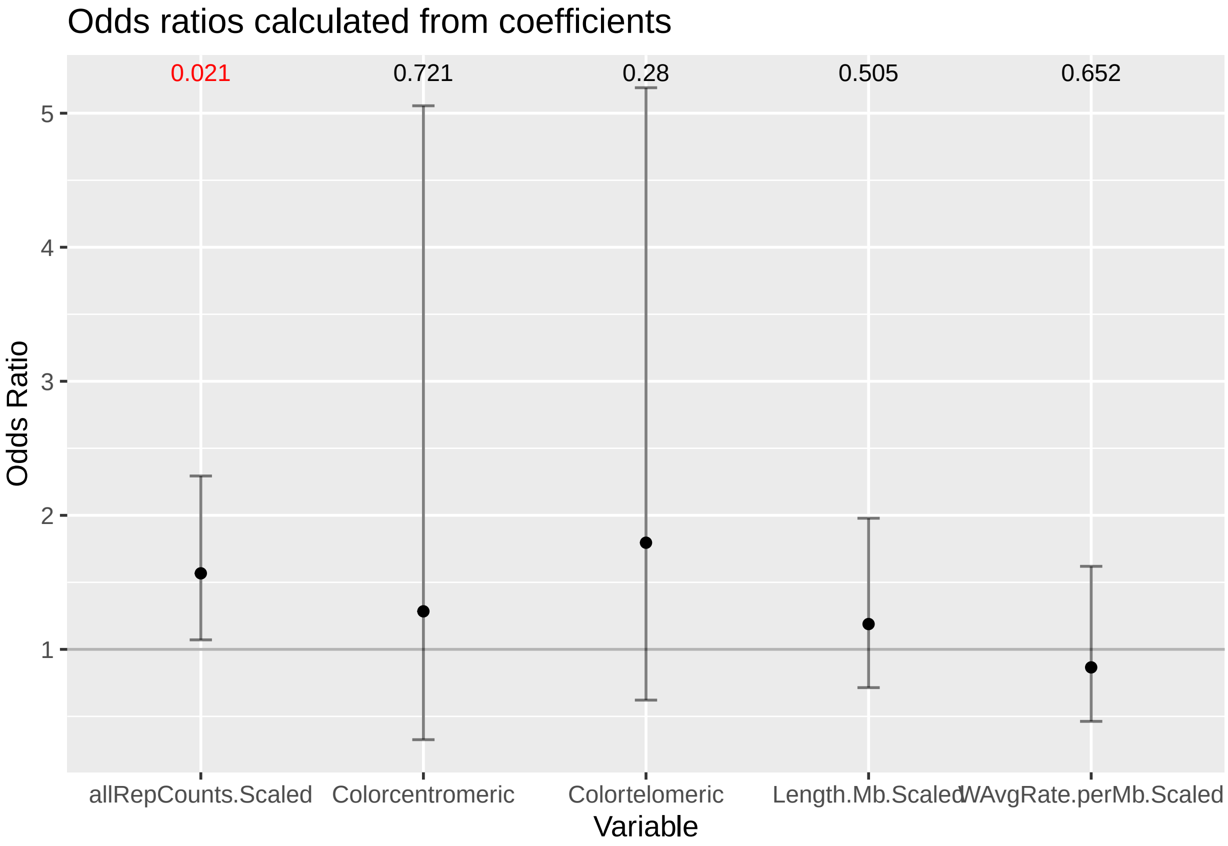
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	-0.3361658	0.6822297
allRepCounts.Scaled	0.0685257	0.8300736
Colorcentromeric	-1.1207780	1.6204468
Colortelomeric	-0.4759060	1.6467649
WAvgRate.perMb.Scaled	-0.7706642	0.4820842

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	1.1889041	0.7145046	1.978284
allRepCounts.Scaled	1.5672142	1.0709282	2.293487
Colorcentromeric	1.2838128	0.3260260	5.055349
Colortelomeric	1.7957620	0.6213219	5.190162
WAvgRate.perMb.Scaled	0.8656367	0.4627056	1.619446

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 1.1889041 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.

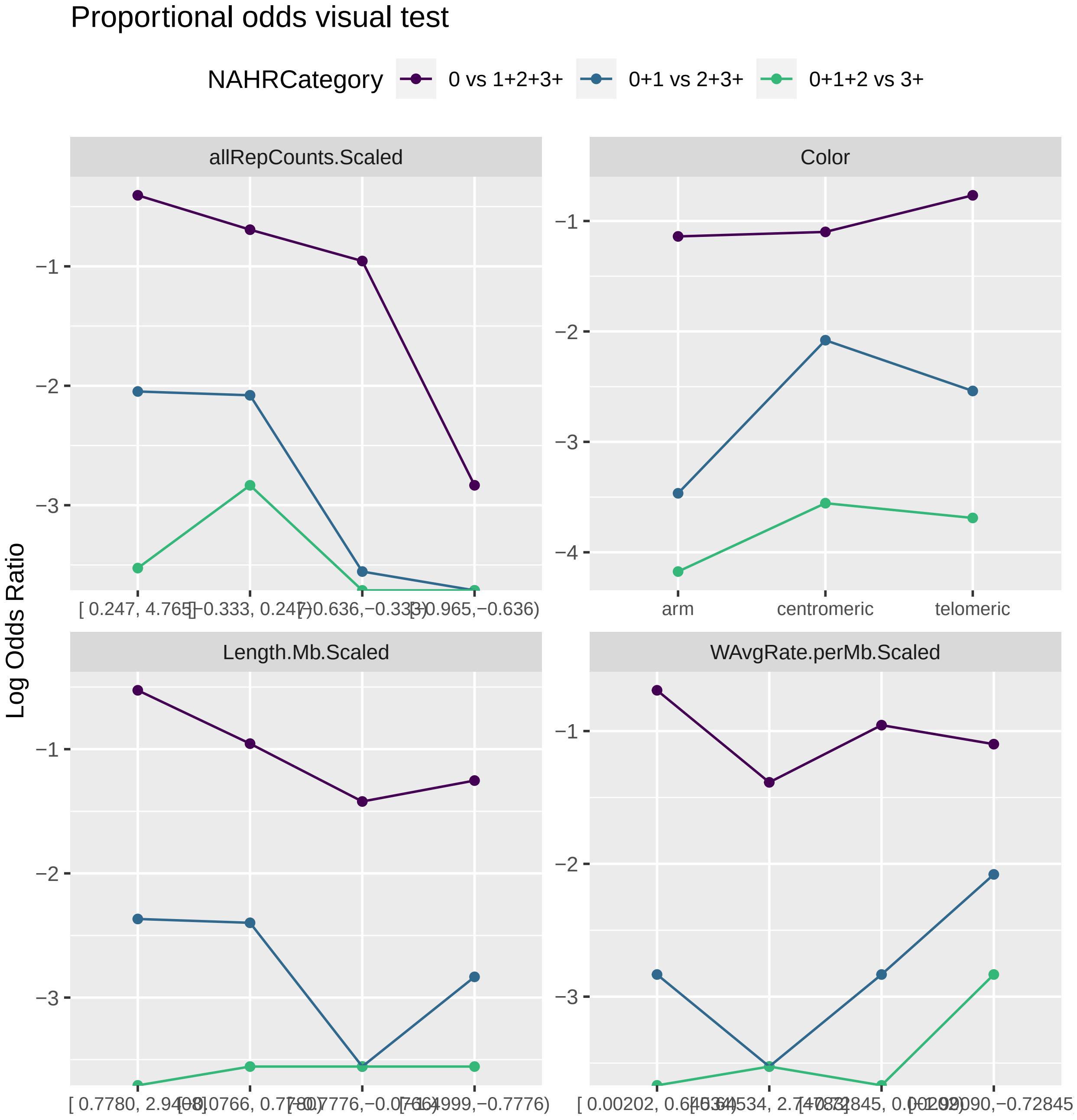
We test the parallel regression assumption with a Brant test:

```
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          15.26   10  0.12
## Length.Mb.Scaled 2.33    2   0.31
## allRepCounts.Scaled  0.87    2   0.65
## Colorcentromeric 4.24    2   0.12
## Colortelomeric       2.83    2   0.24
## WAvgRate.perMb.Scaled    3.67    2   0.16
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	15.2555799	10	0.1230180
Length.Mb.Scaled	2.3318105	2	0.3116404
allRepCounts.Scaled	0.8707161	2	0.6470330
Colorcentromeric	4.2404296	2	0.1200058
Colortelomeric	2.8291963	2	0.2430233
WAvgRate.perMb.Scaled	3.6735618	2	0.1593295

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.

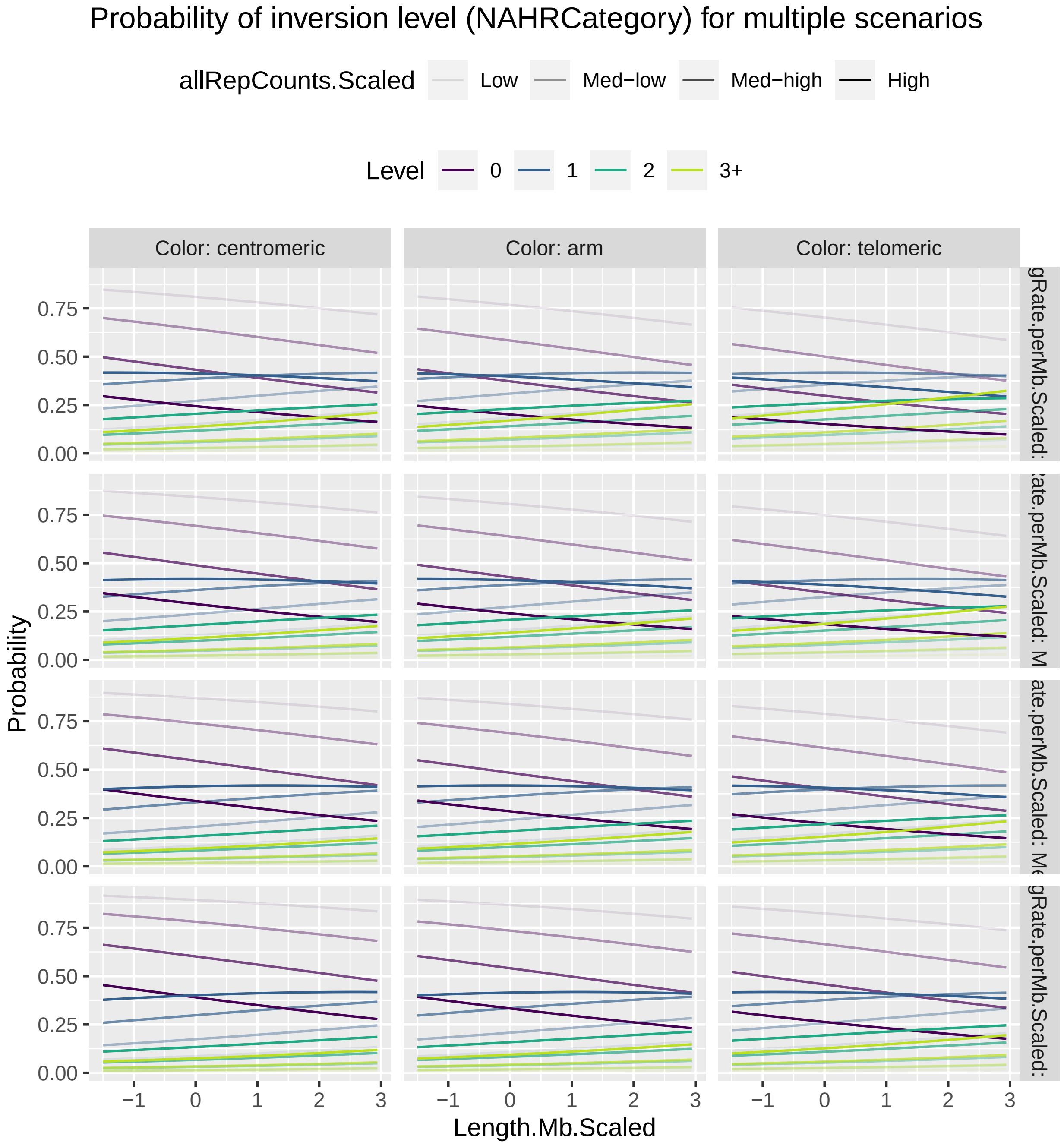


Figure 11: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

Descriptive categories

Descriptive statistics

Raw data:

Chromosome	Start	End	Color	invCenters	NHCenters	NAHRCenters	Length.Mb	allRepCounts	log10RepCounts	WAvgRate.perMb	ChromType
chr10	158946	16728068	telomeric	3	2	1	16.569122	272	2.434569	2.0834355	A
chr10	33436031	39097912	centromeric	1	0	1	5.661881	556	2.745075	1.4181419	A
chr10	113381279	135473442	telomeric	1	1	0	22.092163	170	2.230449	2.1846155	A
chr10	42436301	58578148	centromeric	1	1	0	16.141847	1672	3.223236	0.9909238	A
chr11	241489	23608385	telomeric	1	0	1	23.366896	720	2.857333	1.7638010	A
chr11	43687013	51394932	centromeric	0	0	0	7.707919	494	2.693727	1.0575223	A

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

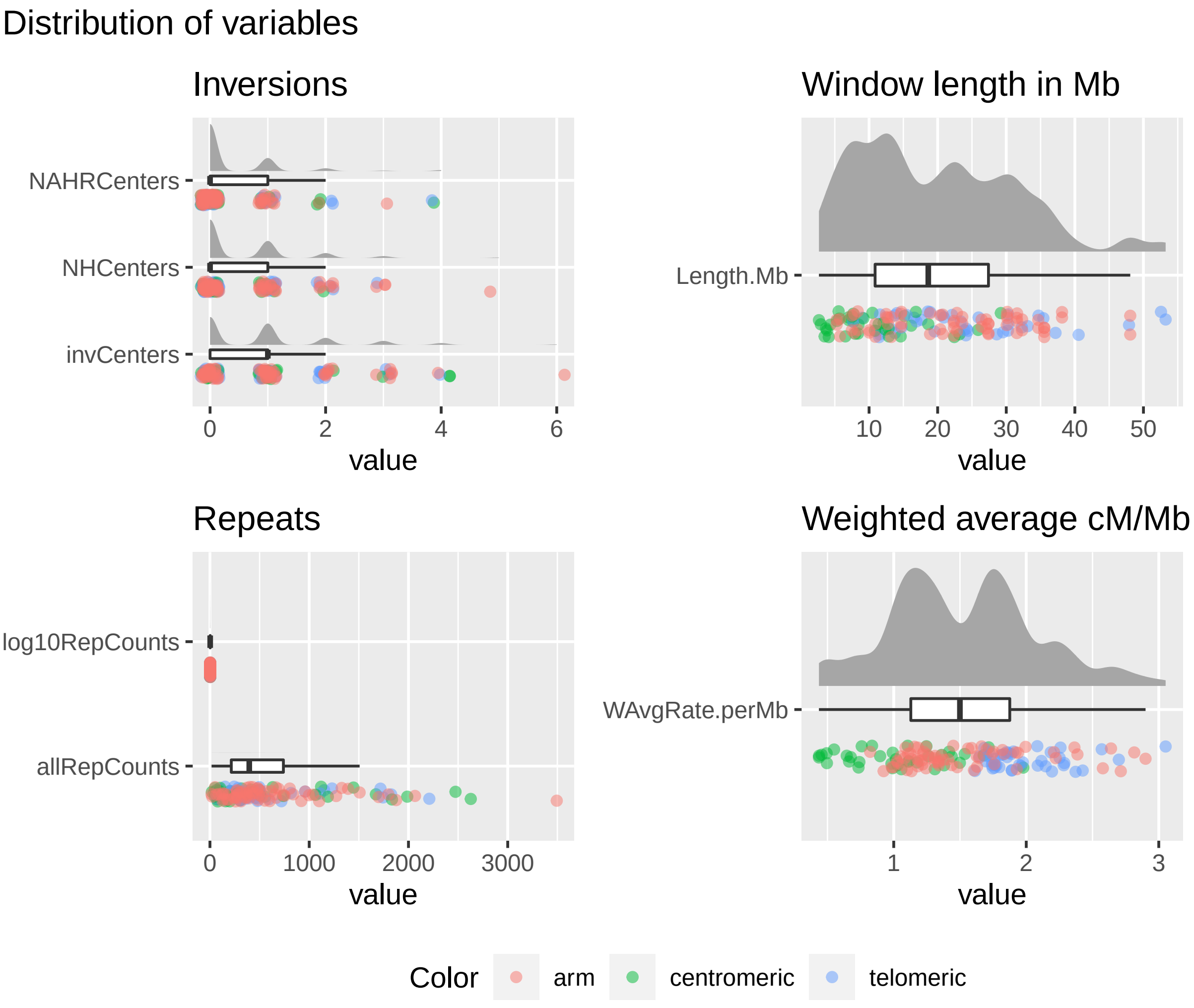


Figure 12: Distribution of variables.

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

Table 32: Original counts

CountGroups	invCenters	NHCenters	NAHRCenters
0	64	88	105
1	49	39	29
2	16	11	6
3	9	4	1
4	4	NA	2
5	NA	1	NA
6	1	NA	NA

Table 33: New counts

	CountGroups	invCategory	NHCategory	NAHRCategory
1	Absence	64	88	105

	CountGroups	invCategory	NHCategory	NAHRCategory
3	Presence	65	50	35
2	Abundance	14	5	3

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

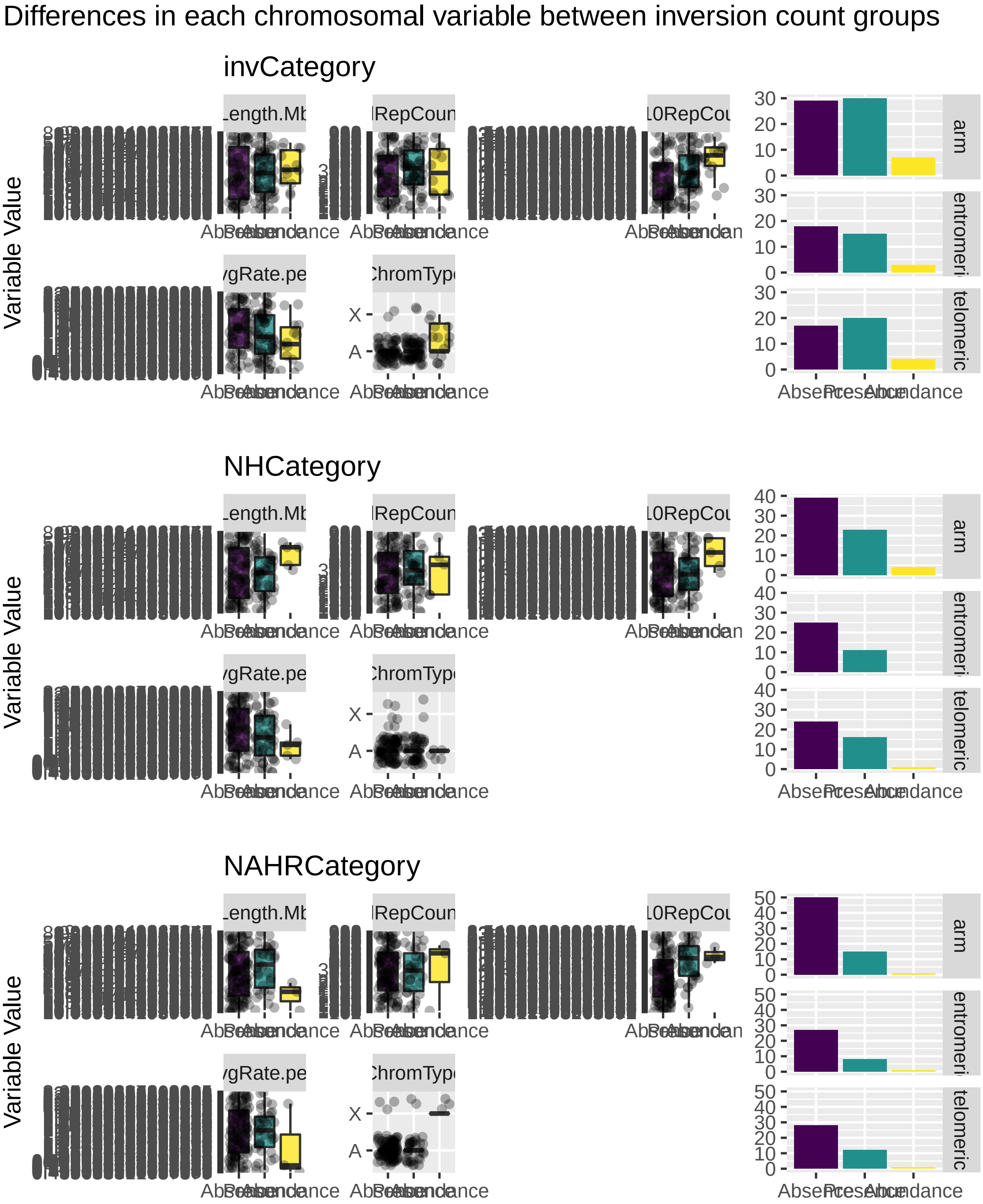


Figure 13: Potential effect of independent variables on the different types of invesions.

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

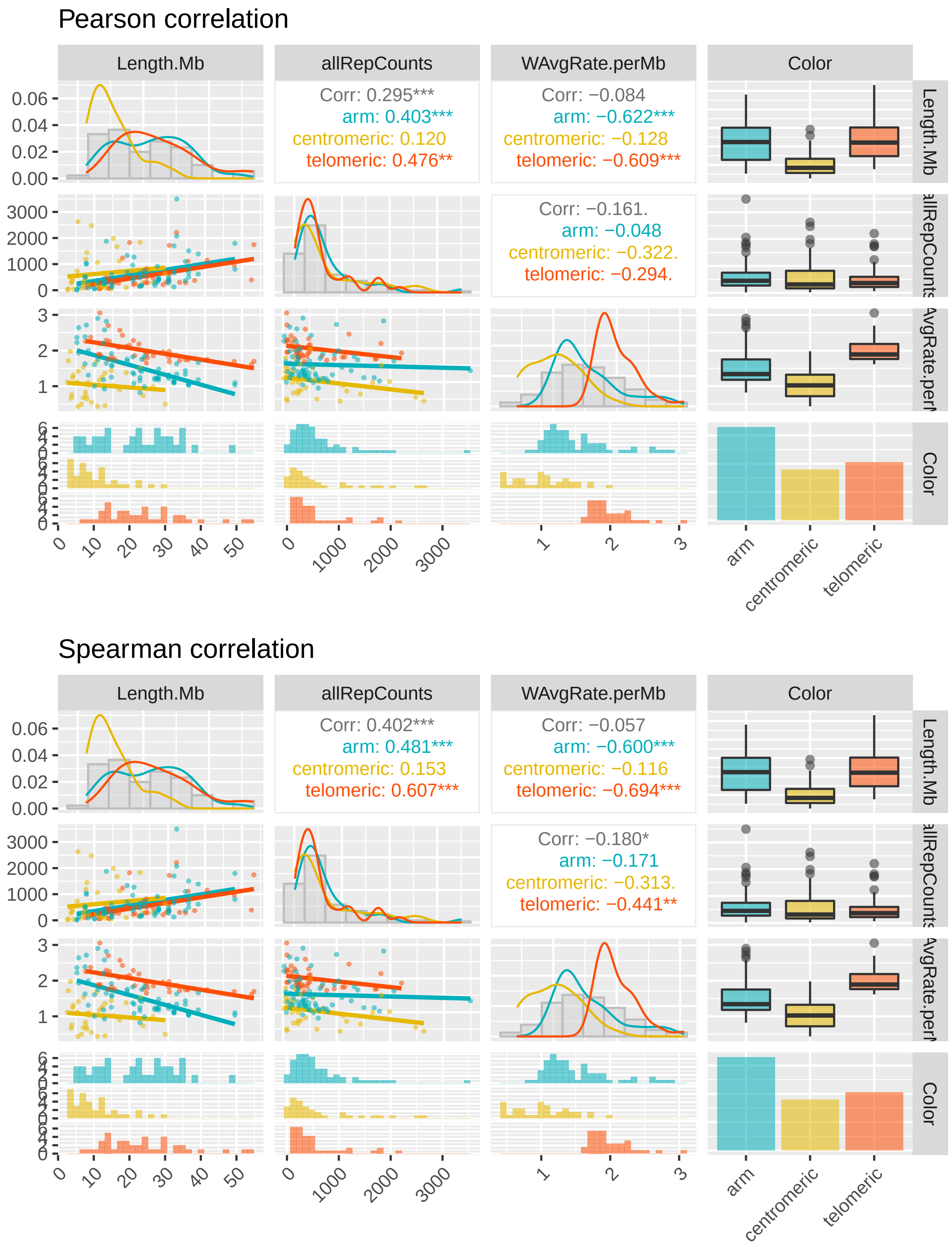


Figure 14: Correlations between variables.

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	1.954368	1	1.397987
allRepCounts	1.145729	1	1.070387
Color	3.035963	2	1.320001
WAvgRate.perMb	2.327808	1	1.525716

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	WAvgRate.perMb	WAvgRate.perMb.Scaled
Min.	2.694933	-1.4999406	16.0000	-0.9652404	0.4356883	-1.9908973
1st Qu.	10.882125	-0.7805224	215.0000	-0.6373992	1.1289521	-0.7351501
Median	18.633361	-0.0994121	396.0000	-0.3392120	1.4993333	-0.0642579
Mean	19.764700	0.0000000	601.9021	0.0000000	1.5348083	0.0000000
3rd Qu.	27.405822	0.6714345	740.0000	0.2275084	1.8756278	0.6173452
Max.	53.232426	2.9408488	3494.0000	4.7645668	3.0518090	2.7478277

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

Not scaled variables

Total inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                      Value Std. Error t value
## Length.Mb          0.0563006  0.0208051  2.7061
## allRepCounts       0.0002846  0.0003107  0.9158
## Colorcentromeric   0.1371802  0.5766406  0.2379
## Colortelomeric     0.2089953  0.4699373  0.4447
## WAvgRate.perMb    -0.5055305  0.4759624 -1.0621
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence     0.3270  1.0387     0.3149
## Presence|Abundance  3.0336  1.0908     2.7811
##
## Residual Deviance: 250.1644
## AIC: 264.1644
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0563006	0.0208051	2.7060957	0.0068079
allRepCounts	0.0002846	0.0003107	0.9158146	0.3597642
Colorcentromeric	0.1371802	0.5766406	0.2378955	0.8119622
Colortelomeric	0.2089953	0.4699373	0.4447301	0.6565148
WAvgRate.perMb	-0.5055305	0.4759624	-1.0621227	0.2881800
Absence|Presence	0.3270366	1.0386834	0.3148569	0.7528703
Presence|Abundance	3.0335951	1.0907886	2.7811026	0.0054175

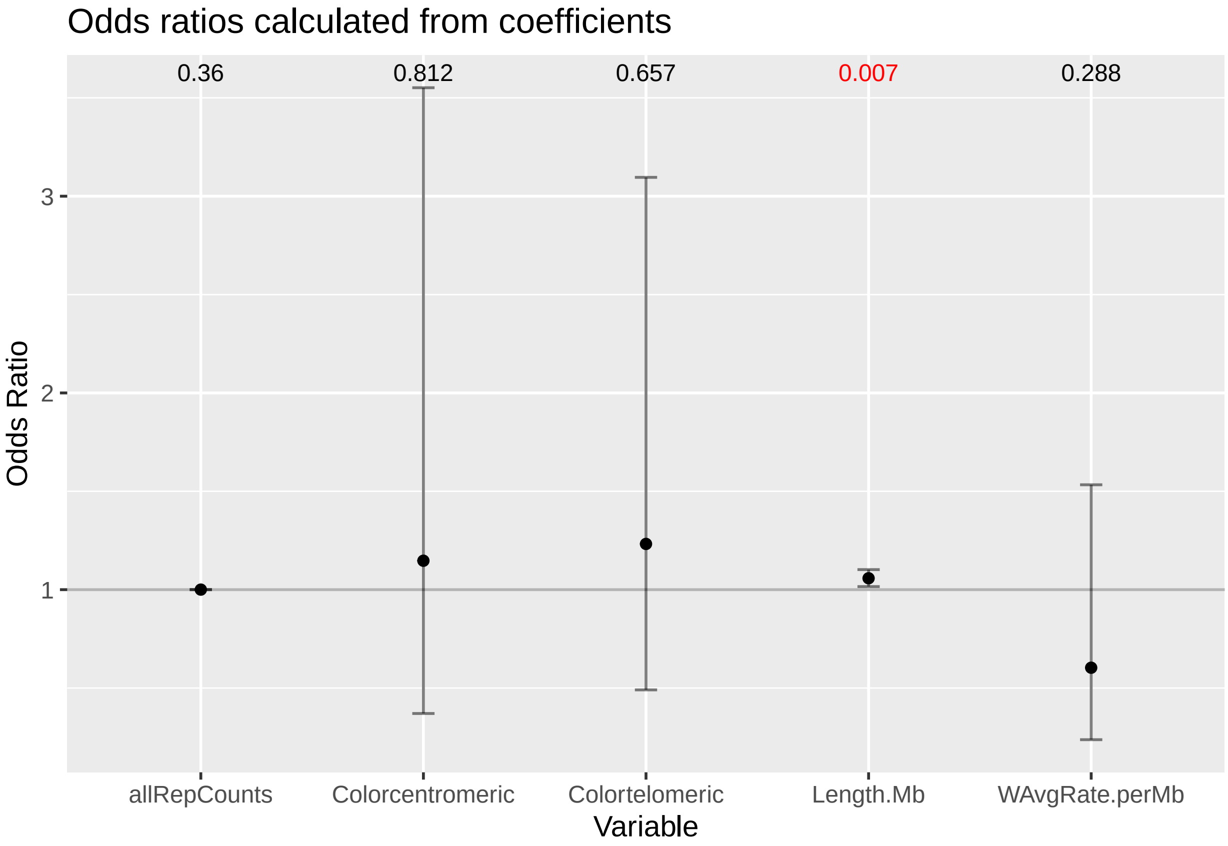
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0155234	0.0970779
allRepCounts	-0.0003244	0.0008936
Colorcentromeric	-0.9930146	1.2673750
Colortelomeric	-0.7120649	1.1300555
WAvgRate.perMb	-1.4383998	0.4273387

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	1.0579157	1.0156445	1.101946
allRepCounts	1.0002846	0.9996756	1.000894
Colorcentromeric	1.1470348	0.3704582	3.551518
Colortelomeric	1.2324392	0.4906300	3.095828
WAvgRate.perMb	0.6031855	0.2373072	1.533172

Example of interpretation: “For 1 unit increase in Length.Mb, a window is 1.0579157 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.

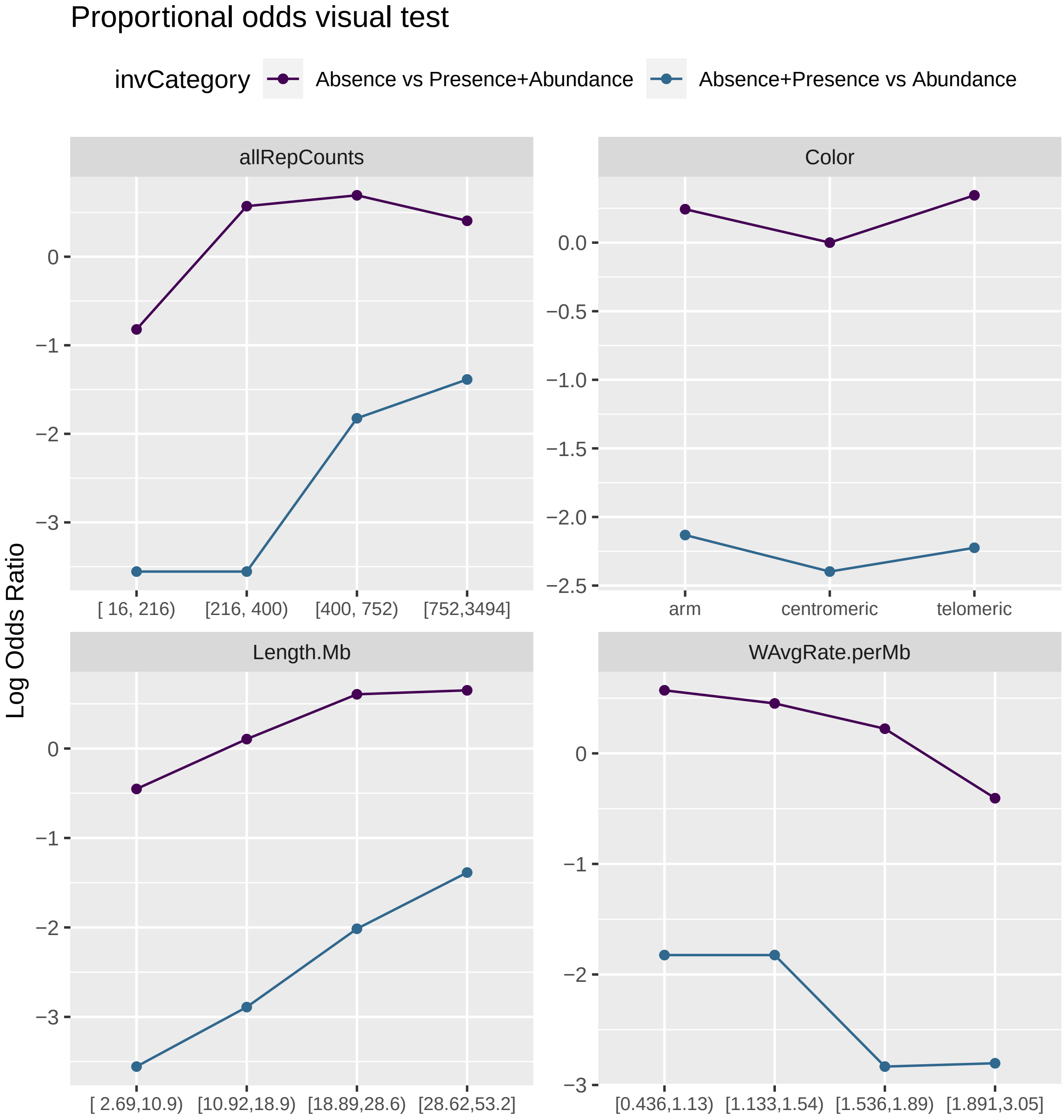
We test the parallel regression assumption with a Brant test:

```
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ----------------------------------------------------
## Test for      X2  df  probability
## ----------------------------------------------------
## Omnibus          4.46    5   0.49
## Length.Mb        3.08    1   0.08
## allRepCounts     0.01    1   0.91
## Colorcentromeric 0.56    1   0.45
## Colortelomeric       0.01    1   0.94
## WAvgRate.perMb       0.23    1   0.63
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	4.4608020	5	0.4851451
Length.Mb	3.0792596	1	0.0792966
allRepCounts	0.0133237	1	0.9081056
Colorcentromeric	0.5594005	1	0.4545019
Colortelomeric	0.0053902	1	0.9414733
WAvgRate.perMb	0.2333141	1	0.6290773

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.

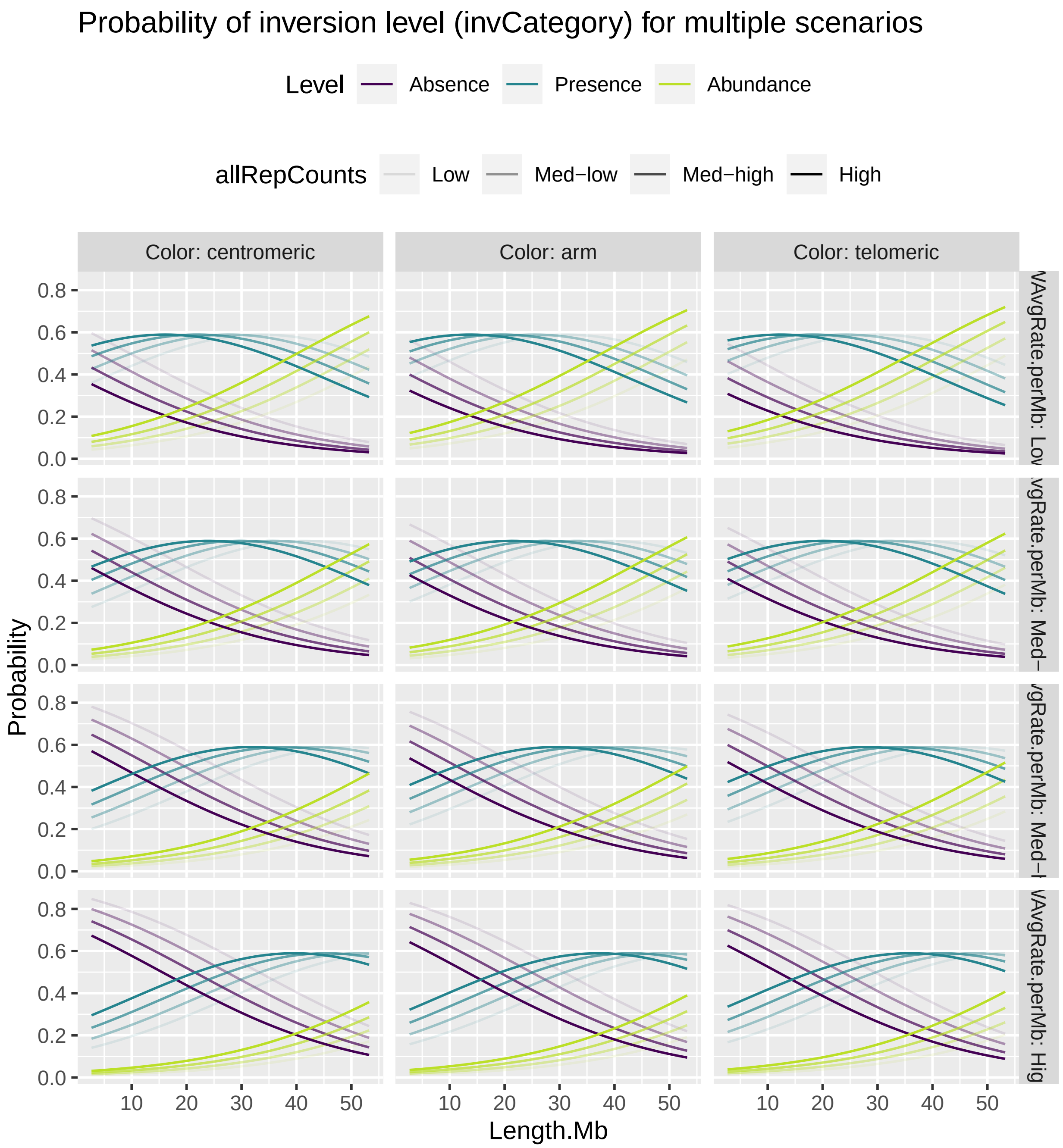


Figure 15: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                        Value Std. Error  t value
## Length.Mb          0.0906986  0.0237116  3.82508
## allRepCounts      -0.0004376  0.0003652 -1.19841
## Colorcentromeric  0.2609577  0.6299654  0.41424
## Colortelomeric    -0.0282214  0.5361529 -0.05264
## WAvgRate.perMb    -0.4382591  0.5649295 -0.77578
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence    1.4302  1.1774     1.2147
## Presence|Abundance  4.7000  1.3107     3.5860
##
## Residual Deviance: 196.6972
## AIC: 210.6972
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0906986	0.0237116	3.8250805	0.0001307
allRepCounts	-0.0004376	0.0003652	-1.1984137	0.2307560
Colorcentromeric	0.2609577	0.6299654	0.4142414	0.6786973
Colortelomeric	-0.0282214	0.5361529	-0.0526369	0.9580212
WAvgRate.perMb	-0.4382591	0.5649295	-0.7757766	0.4378809
Absence|Presence	1.4301787	1.1773897	1.2147029	0.2244794
Presence|Abundance	4.6999553	1.3106557	3.5859573	0.0003358

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.

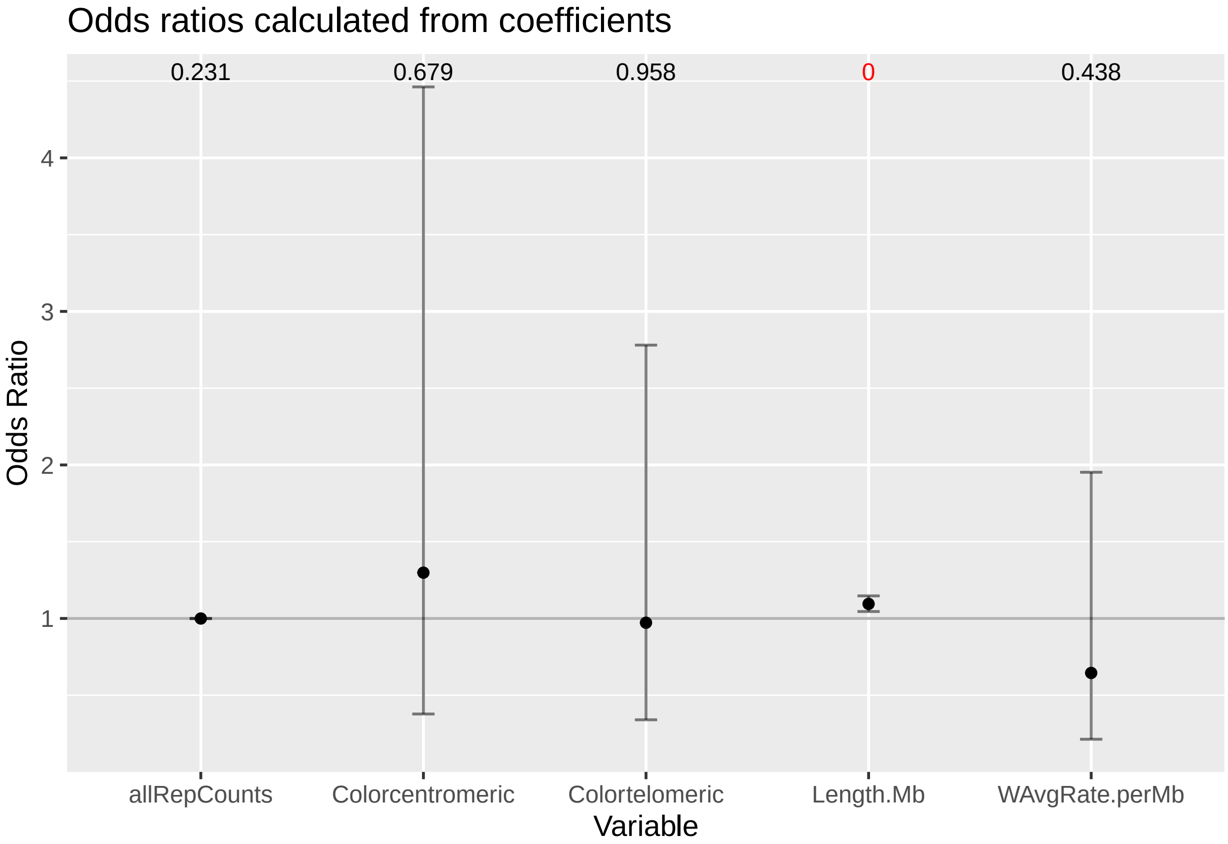
	2.5 %	97.5 %
Length.Mb	0.0442248	0.1371724
allRepCounts	-0.0011534	0.0002781
Colorcentromeric	-0.9737517	1.4956672
Colortelomeric	-1.0790618	1.0226189
WAvgRate.perMb	-1.5455006	0.6689824

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	1.0949390	1.0452173	1.147026

	Odds Ratio	2.5%	97.5%
allRepCounts	0.9995625	0.9988473	1.000278
Colorcentromeric	1.2981728	0.3776635	4.462313
Colortelomeric	0.9721731	0.3399143	2.780467
WAvgRate.perMb	0.6451586	0.2132051	1.952250

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.094939 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.

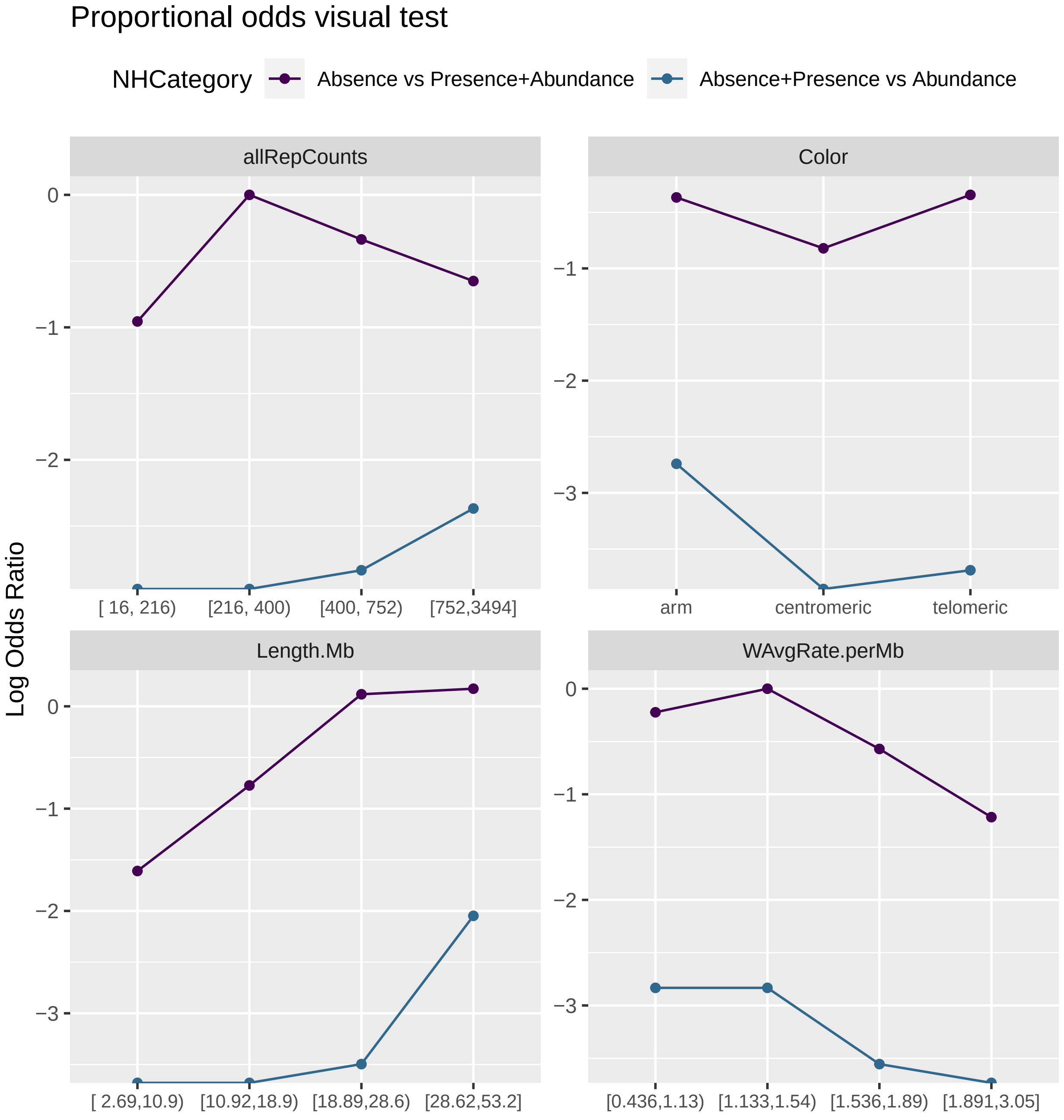
We test the parallel regression assumption with a Brant test:

```
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          4.41    5   0.49
## Length.Mb        2.45    1   0.12
## allRepCounts     0.35    1   0.56
## Colorcentromeric 0   1   1
## Colortelomeric       0.84    1   0.36
## WAvgRate.perMb       0.01    1   0.92
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	4.4143690	5	0.4914217
Length.Mb	2.4548196	1	0.1171646
allRepCounts	0.3467415	1	0.5559635
Colorcentromeric	0.0000360	1	0.9952119
Colortelomeric	0.8401731	1	0.3593473
WAvgRate.perMb	0.0090354	1	0.9242714

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.

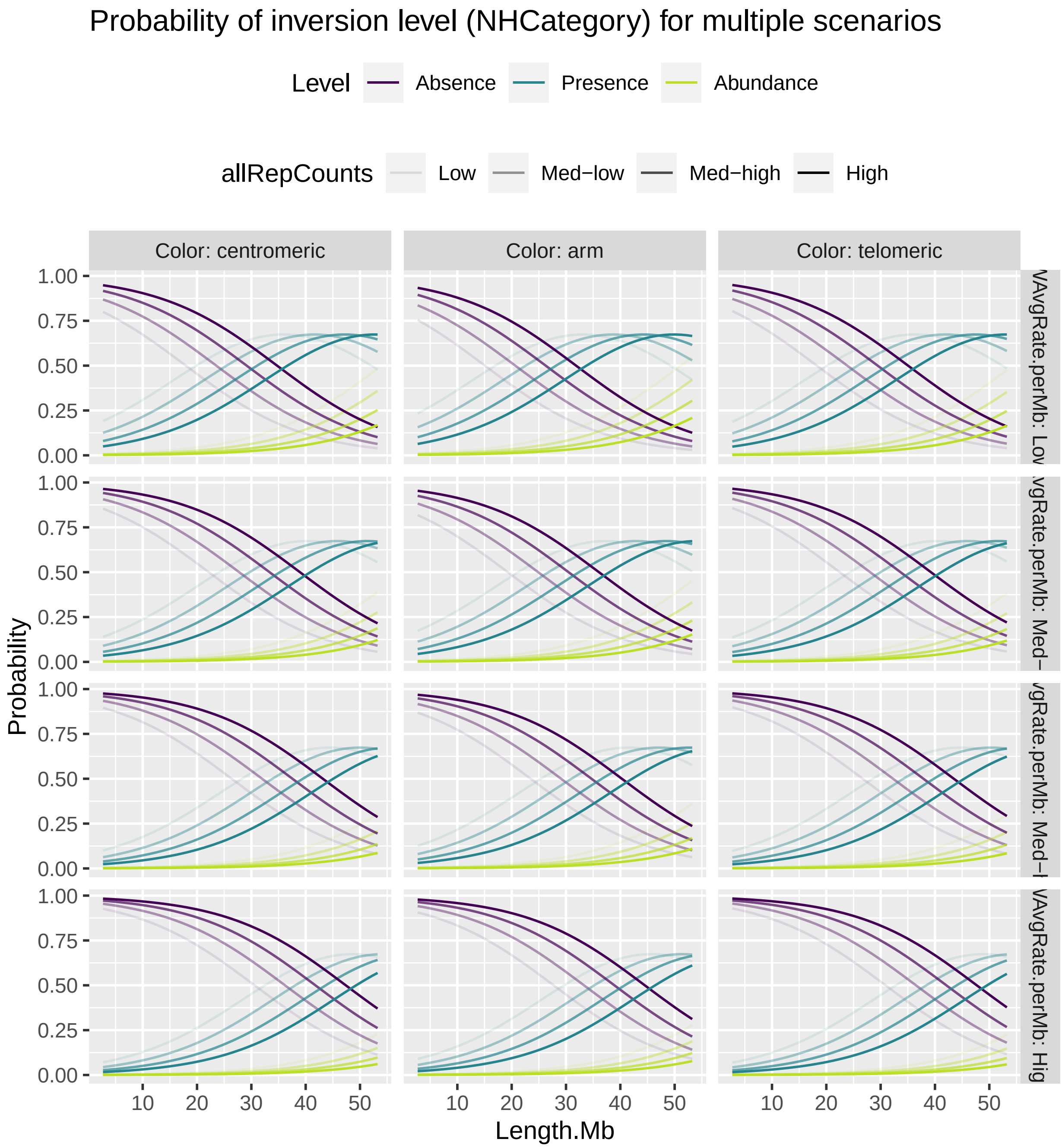


Figure 16: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

This cannot be done with ordinal logistic regression because we have only 2 categories, we would make a binomial logistic regression.

Scaled variables

Total inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                          Value Std. Error t value
## Length.Mb.Scaled        0.6407     0.2367  2.7072
## allRepCounts.Scaled     0.1727     0.1792  0.9640
## Colorcentromeric        0.1372     0.5768  0.2378
## Colortelomeric          0.2090     0.4699  0.4448
## WAvgRate.perMb.Scaled -0.2791     0.2628 -1.0621
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence   -0.1811  0.2582    -0.7016
## Presence|Abundance  2.5254  0.3588     7.0393
##
## Residual Deviance: 250.1644
## AIC: 264.1644
```

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	0.6407088	0.2366694	2.7071888	0.0067856
allRepCounts.Scaled	0.1727360	0.1791856	0.9640062	0.3350428
Colorcentromeric	0.1371708	0.5768325	0.2378000	0.8120362
Colortelomeric	0.2089934	0.4698517	0.4448072	0.6564591
WAvgRate.perMb.Scaled	-0.2790920	0.2627661	-1.0621309	0.2881763
Absence|Presence	-0.1811136	0.2581546	-0.7015700	0.4829474
Presence|Abundance	2.5254373	0.3587601	7.0393486	0.0000000

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.

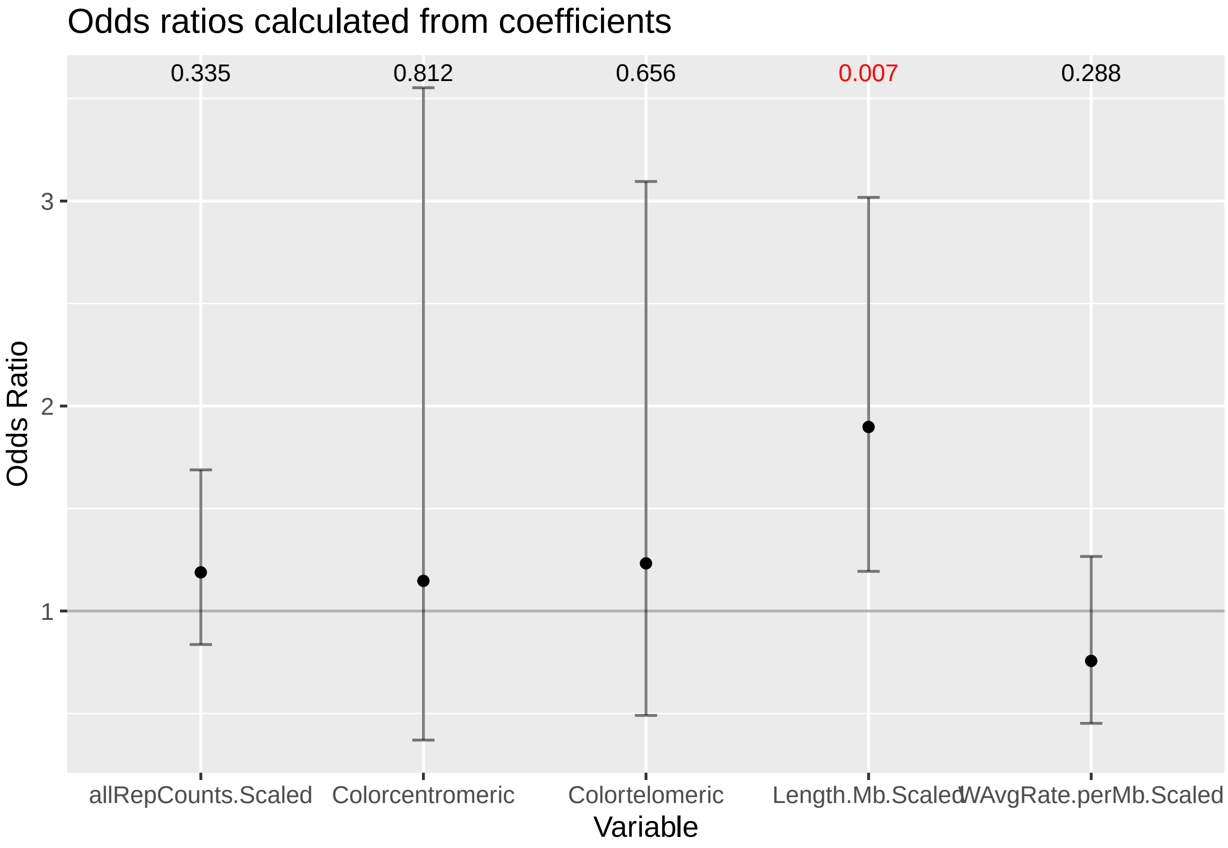
	2.5 %	97.5 %
Length.Mb.Scaled	0.1768453	1.1045723
allRepCounts.Scaled	-0.1784613	0.5239332
Colorcentromeric	-0.9934001	1.2677417
Colortelomeric	-0.7118991	1.1298859

	2.5 %	97.5 %
WAvgRate.perMb.Scaled	-0.7941042	0.2359201

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	1.8978255	1.1934464	3.017933
allRepCounts.Scaled	1.1885523	0.8365565	1.688657
Colorcentromeric	1.1470240	0.3703154	3.552820
Colortelomeric	1.2324369	0.4907114	3.095303
WAvgRate.perMb.Scaled	0.7564703	0.4519859	1.266073

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 1.8978255 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.

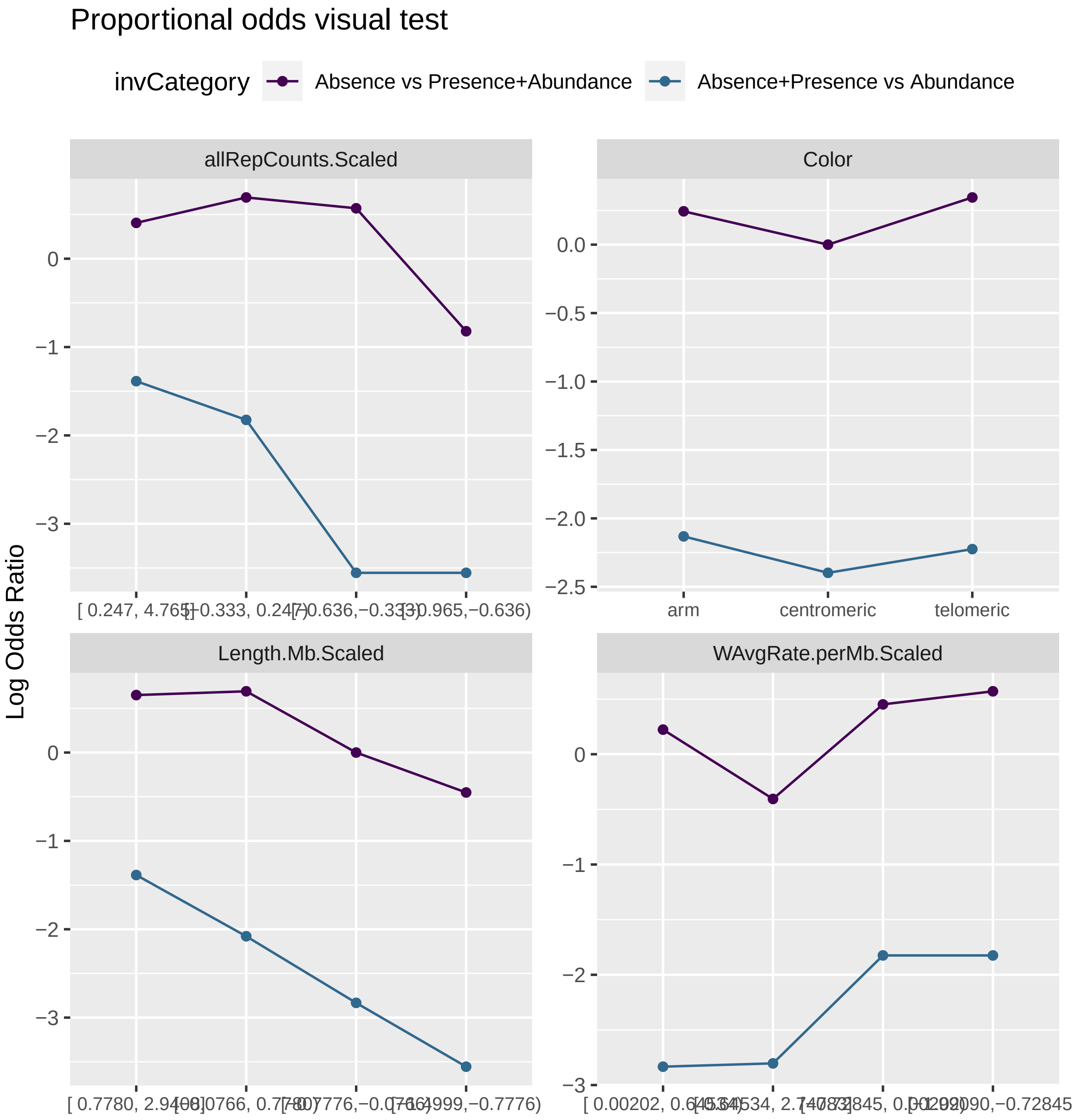
We test the parallel regression assumption with a Brant test:

```r
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          4.46    5   0.49
## Length.Mb.Scaled 3.08    1   0.08
## allRepCounts.Scaled  0.01    1   0.91
## Colorcentromeric 0.56    1   0.45
## Colortelomeric       0.01    1   0.94
## WAvgRate.perMb.Scaled    0.23    1   0.63
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	4.4608020	5	0.4851451
Length.Mb.Scaled	3.0792596	1	0.0792966
allRepCounts.Scaled	0.0133237	1	0.9081056
Colorcentromeric	0.5594005	1	0.4545019
Colortelomeric	0.0053902	1	0.9414733
WAvgRate.perMb.Scaled	0.2333141	1	0.6290773

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.

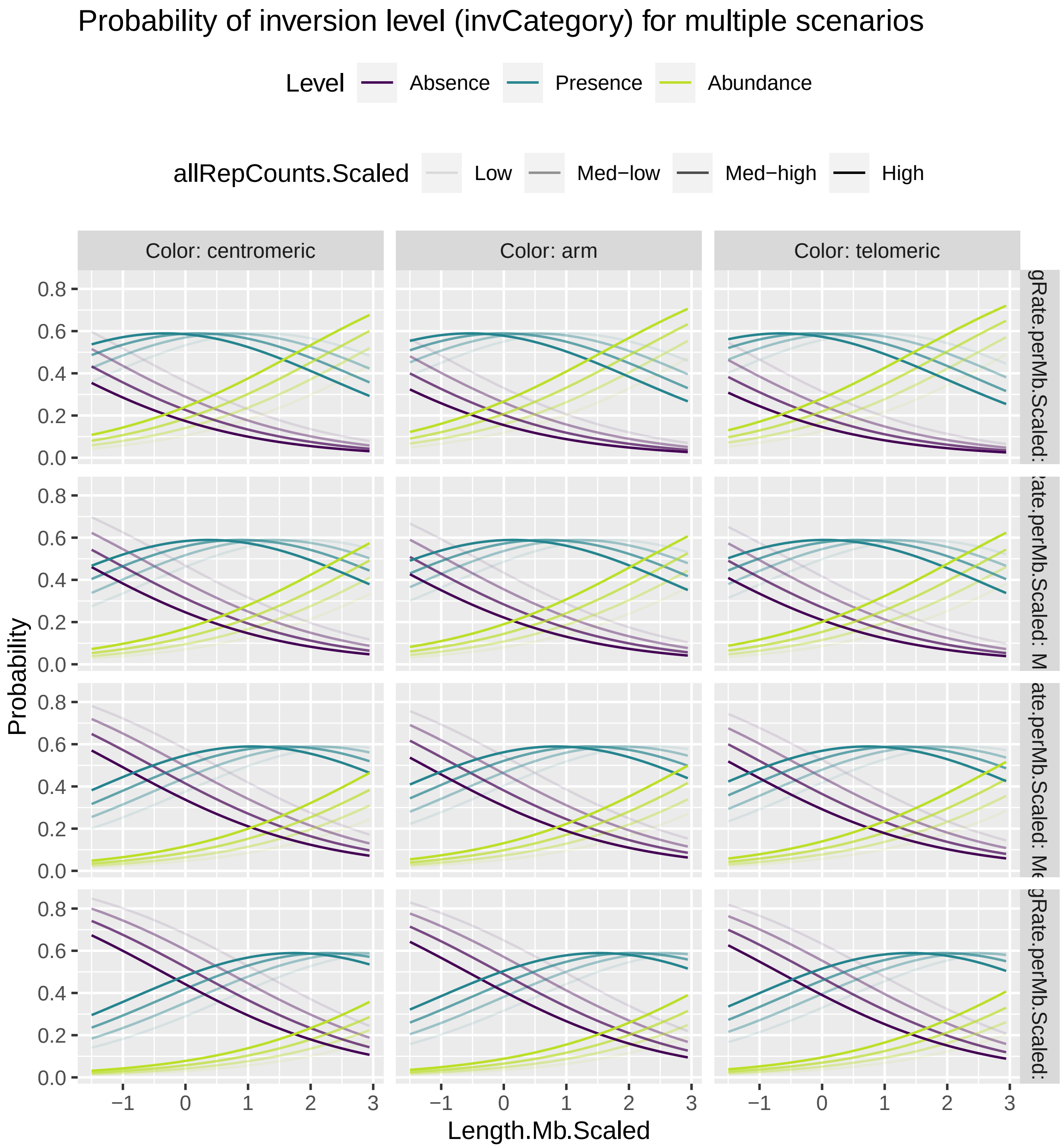


Figure 17: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                          Value Std. Error  t value
## Length.Mb.Scaled       1.03218     0.2704  3.81705
## allRepCounts.Scaled   -0.26565     0.2036 -1.30477
## Colorcentromeric       0.26095     0.6297  0.41441
## Colortelomeric        -0.02822     0.5364 -0.05261
## WAvgRate.perMb.Scaled -0.24195     0.3119 -0.77567
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence    0.5736  0.2837     2.0222
## Presence|Abundance  3.8434  0.5465     7.0329
##
## Residual Deviance: 196.6972
## AIC: 210.6972
```

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	1.0321776	0.2704127	3.8170452	0.0001351
allRepCounts.Scaled	-0.2656476	0.2035977	-1.3047668	0.1919723
Colorcentromeric	0.2609539	0.6297008	0.4144094	0.6785743
Colortelomeric	-0.0282170	0.5363699	-0.0526074	0.9580447
WAvgRate.perMb.Scaled	-0.2419525	0.3119271	-0.7756700	0.4379439
Absence|Presence	0.5736044	0.2836577	2.0221709	0.0431587
Presence|Abundance	3.8433807	0.5464861	7.0328970	0.0000000

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.

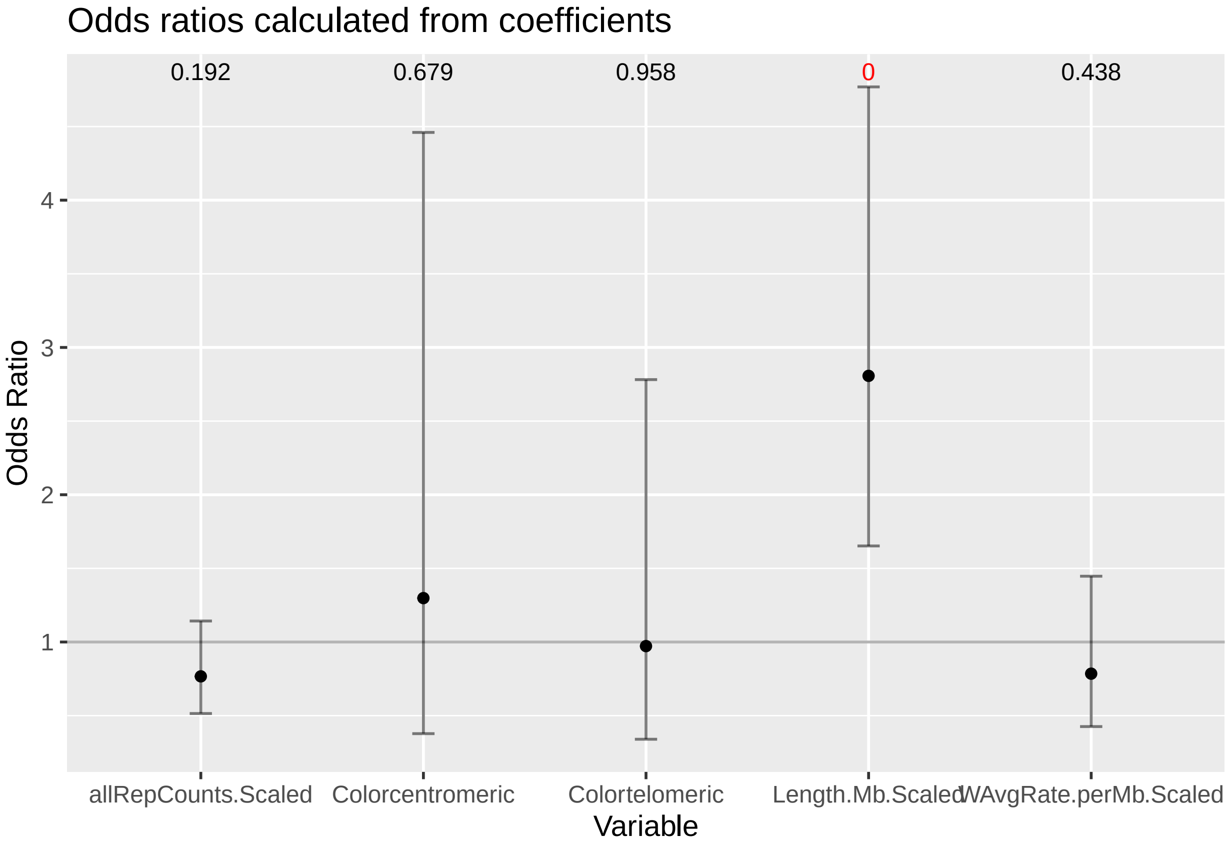
	2.5 %	97.5 %
Length.Mb.Scaled	0.5021784	1.5621768
allRepCounts.Scaled	-0.6646918	0.1333966
Colorcentromeric	-0.9732369	1.4951447
Colortelomeric	-1.0794827	1.0230486
WAvgRate.perMb.Scaled	-0.8533184	0.3694134

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	2.8071720	1.6523167	4.769191

	Odds Ratio	2.5%	97.5%
allRepCounts.Scaled	0.7667093	0.5144321	1.142703
Colorcentromeric	1.2981678	0.3778580	4.459982
Colortelomeric	0.9721774	0.3397712	2.781662
WAvgRate.perMb.Scaled	0.7850935	0.4259990	1.446886

Example of interpretation: “For 1 unit increase in Length.Mb.Scaled, a window is 2.807172 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.

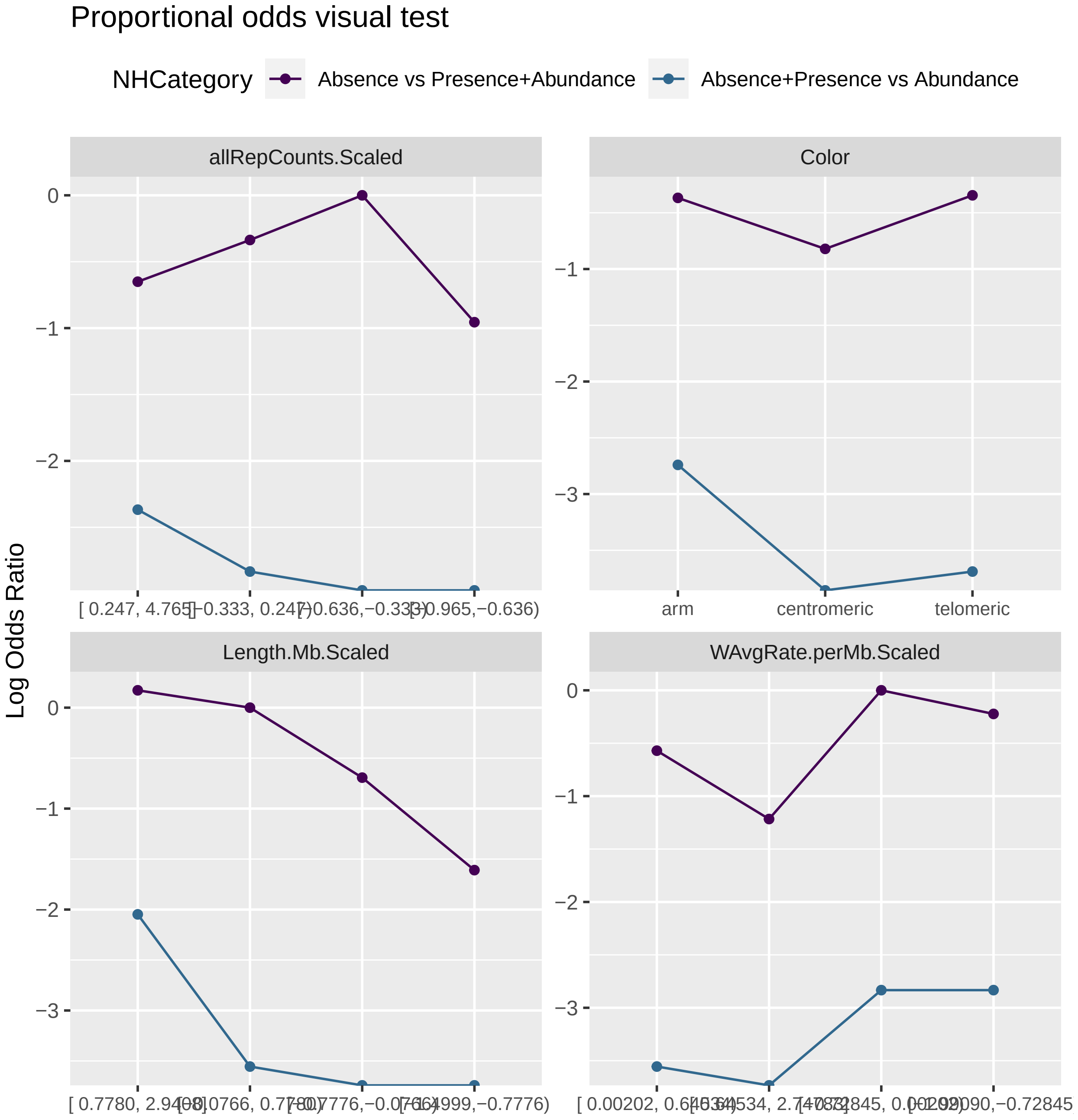
We test the parallel regression assumption with a Brant test:

```
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ----------------------------------------------------
## Test for      X2  df  probability
## ----------------------------------------------------
## Omnibus          4.41    5   0.49
## Length.Mb.Scaled 2.45    1   0.12
## allRepCounts.Scaled  0.35    1   0.56
## Colorcentromeric 0   1   1
## Colortelomeric       0.84    1   0.36
## WAvgRate.perMb.Scaled    0.01    1   0.92
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	4.4143690	5	0.4914217
Length.Mb.Scaled	2.4548196	1	0.1171646
allRepCounts.Scaled	0.3467415	1	0.5559635
Colorcentromeric	0.0000360	1	0.9952119
Colortelomeric	0.8401731	1	0.3593473
WAvgRate.perMb.Scaled	0.0090354	1	0.9242714

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.

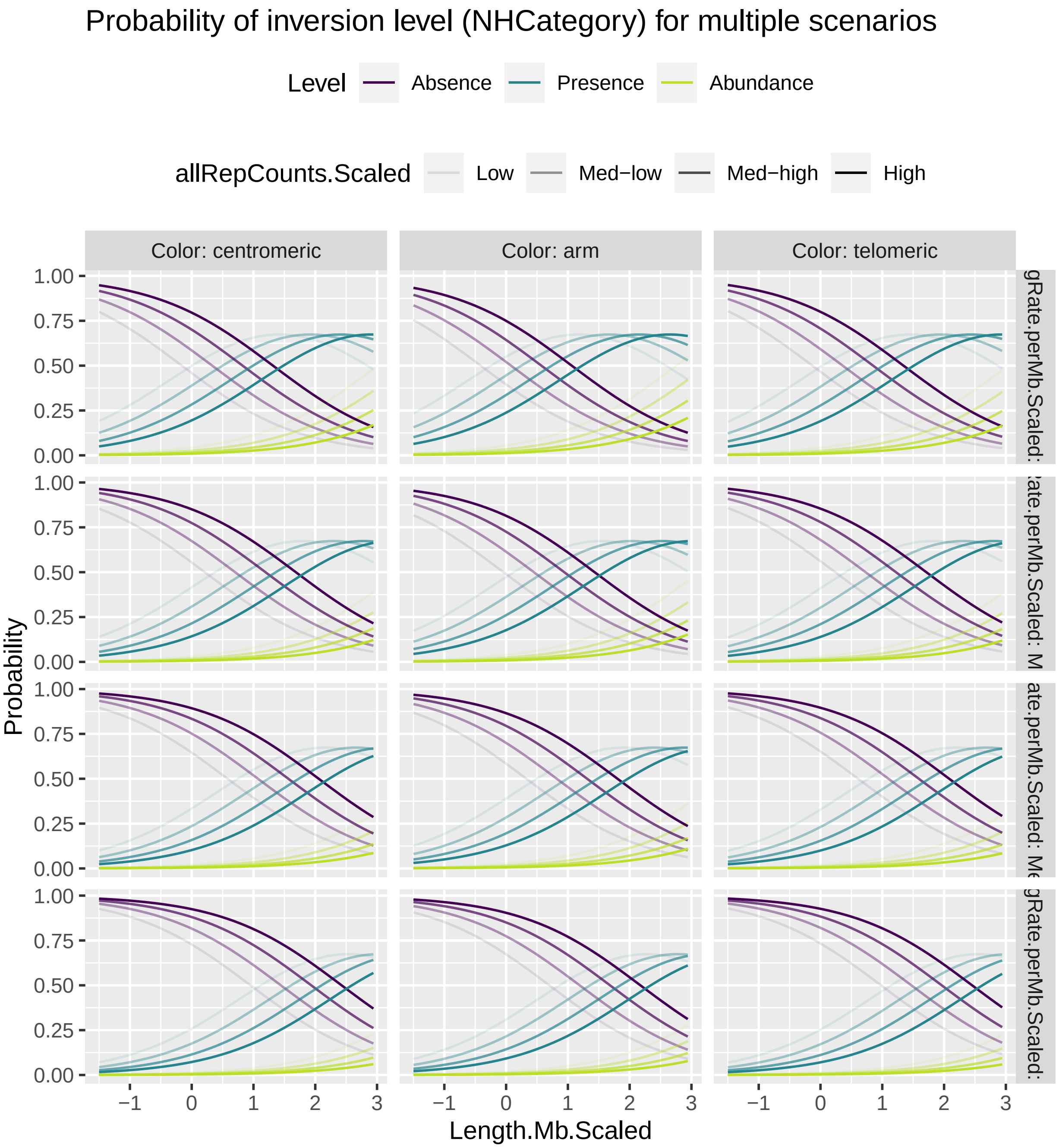


Figure 18: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

This cannot be done with ordinal logistic regression because we have only 2 categories, we would make a binomial logistic regression.