Ordinal logistic model on large, classified windows data

Ruth Gómez Graciani

Prepare the data

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

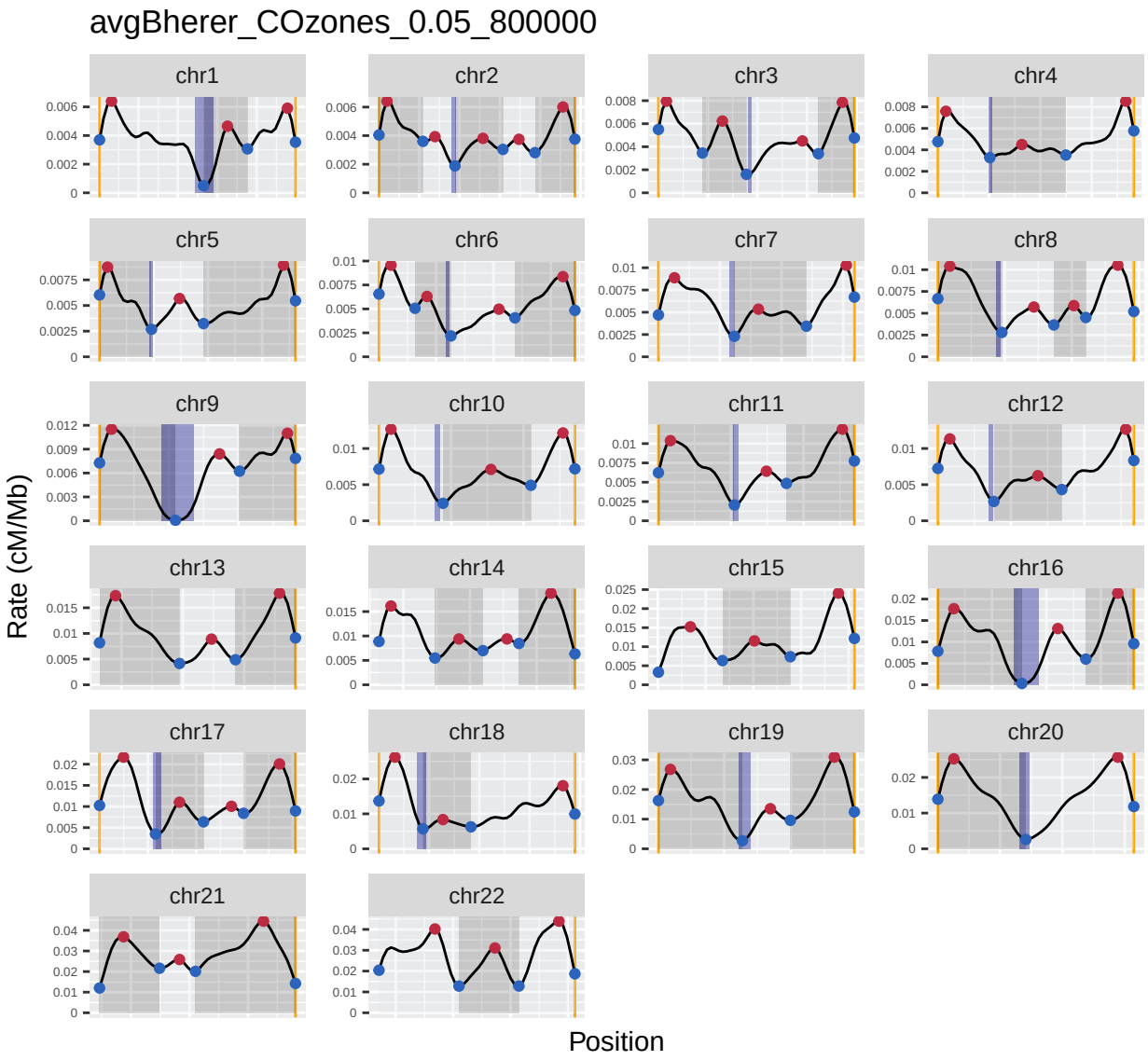


Figure 1: Crossover zones; centromeres in blue, workspace limits in orange.

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

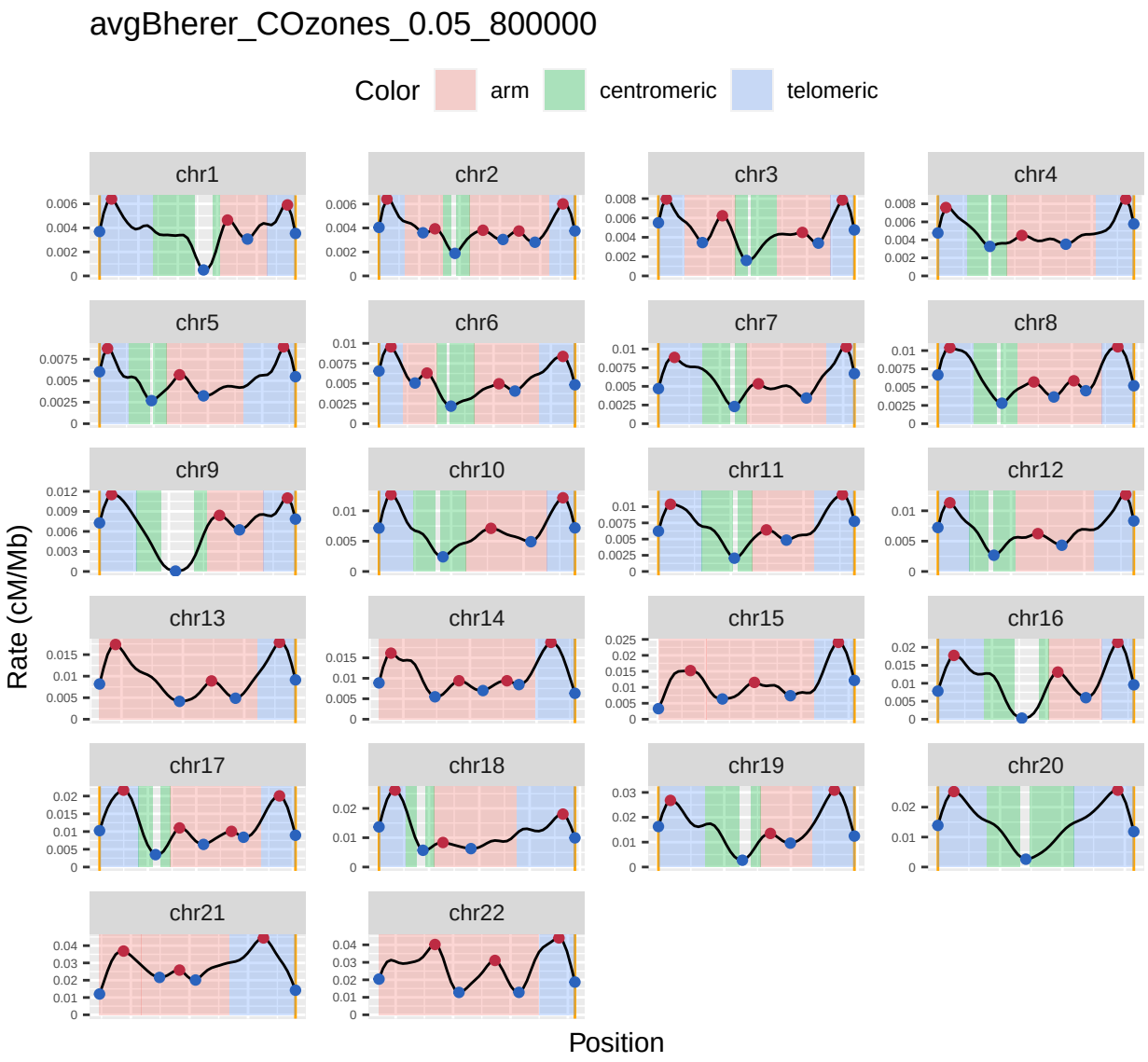


Figure 2: Color-coded windows for telomeric, centromeric and arm categories.

Numerical categories

Descriptive statistics

Raw data:

Chromosome	Start	End	Color	invCenters	NHCenters	NAHRCenters	Length.Mb	allRepCount	WAvgRate.perMb
chr10	158946	23770709	telomeric	3	2	1	23.61176	354	1.8516302
chr10	23770709	39097912	centromeric	1	0	1	15.32720	880	1.0408405
chr10	59958908	116142800	arm	2	2	0	56.18389	818	0.9967980
chr10	116142800	135473442	telomeric	1	1	0	19.33064	168	2.0570277
chr10	42436301	59958908	centromeric	2	2	0	17.52261	1678	0.5819161
chr11	241489	29941780	telomeric	2	1	1	29.70029	746	1.5167467

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

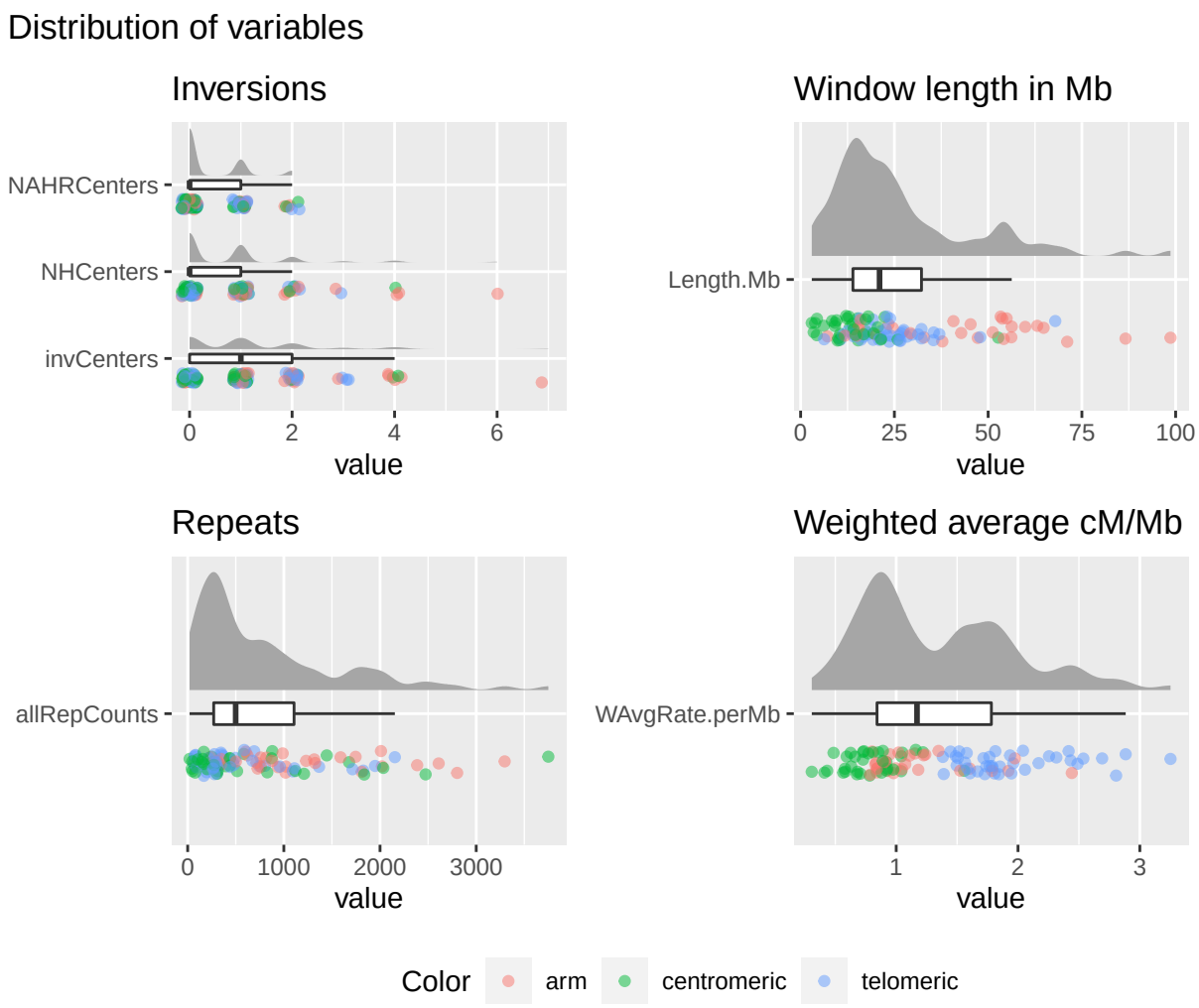


Figure 3: Distribution of variables.

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	38	54	71
1	35	32	24
2	18	10	7
3	4	2	NA
4	6	3	NA
6	NA	1	NA
7	1	NA	NA

Table 3: New counts

CountGroups	invCategory	NHCategory	NAHRCategory
0	38	54	71

CountGroups	invCategory	NHCategory	NAHRCategory
1	35	32	24
2	18	10	7
3+	11	6	NA

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

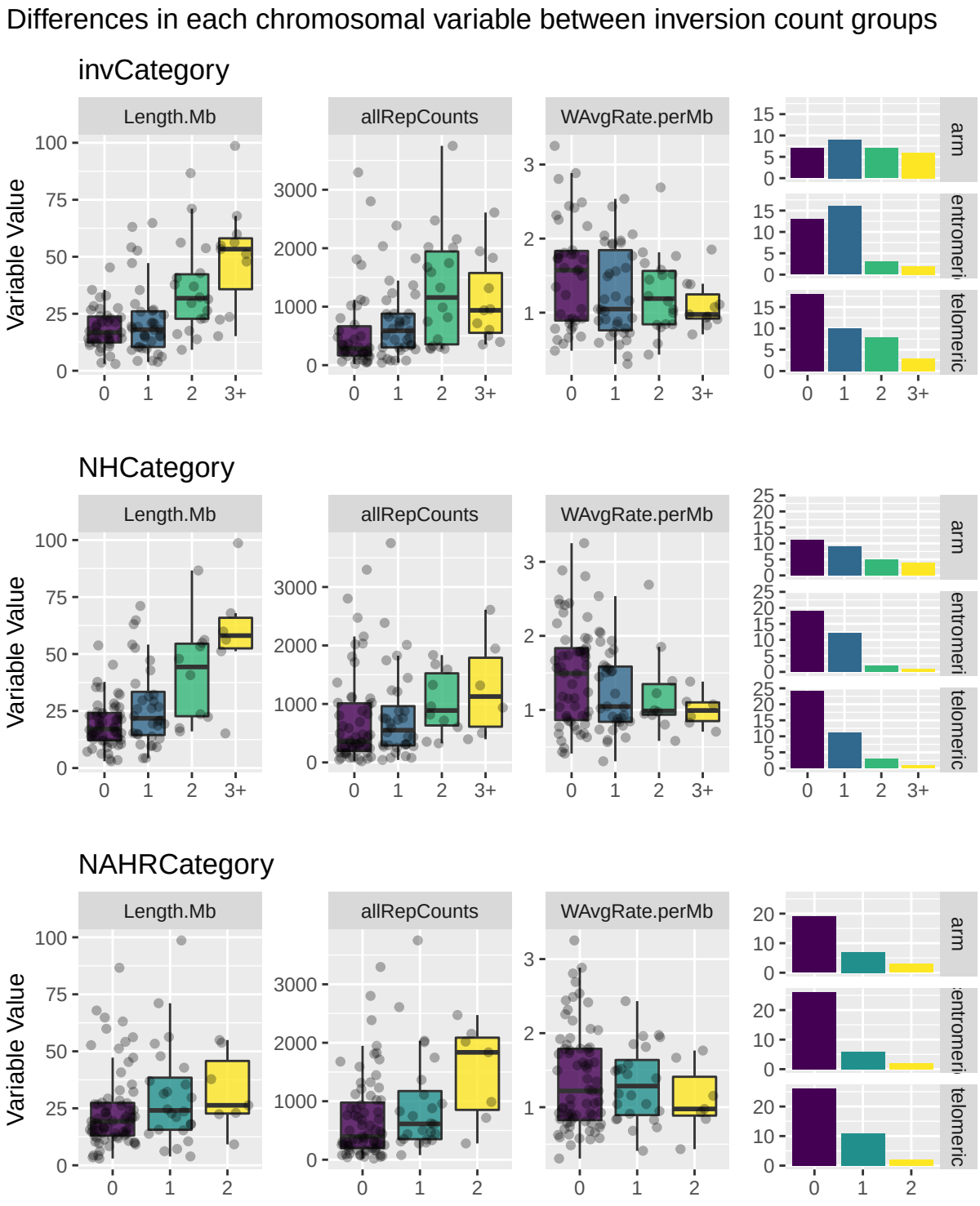


Figure 4: Potential effect of independent variables on the different types of invesions.

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

Figure 5: Correlations between variables.

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

	GVIF	Df	GVIF^(1/(2*Df))
Length.Mb	2.713019	1	1.647124
allRepCounts	1.248267	1	1.117259
Color	6.917414	2	1.621758
WAvgRate.perMb	4.344256	1	2.084288

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.	2.969812	-1.2438876	20.000	-1.0149666	0.3068634	-1.6361728
1st Qu.	13.949468	-0.6574034	270.500	-0.6908557	0.8424562	-0.7894017
Median	21.043519	-0.2784709	499.000	-0.3952096	1.1708581	-0.2701990
Mean	26.256815	0.0000000	804.451	0.0000000	1.3417622	0.0000000
3rd Qu.	32.224751	0.3187804	1106.000	0.3901610	1.7813523	0.6949911
Max.	98.630850	3.8658972	3750.000	3.8111162	3.2519378	3.0199836

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

Total inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error t value
## Length.Mb         0.0512268  0.0115267   4.444
## allRepCounts      0.0004253  0.0003051   1.394
## Colorcentromeric  0.3014468  0.3429438   0.879
## Colortelomeric    0.5325864  0.6171682   0.863
## WAvgRate.perMb   -0.5280284  0.3099066  -1.704
##
## Intercepts:
##      Value   Std. Error t value
## 0|1   0.5283  0.1393     3.7915
## 1|2   2.3491  0.3085     7.6145
## 2|3+  3.8876  0.4582     8.4849
##
## Residual Deviance: 229.2007
## AIC: 245.2007
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0512268	0.0115267	4.4441944	0.0000088
allRepCounts	0.0004253	0.0003051	1.3941441	0.1632741
Colorcentromeric	0.3014468	0.3429438	0.8789978	0.3794025
Colortelomeric	0.5325864	0.6171682	0.8629518	0.3881640
WAvgRate.perMb	-0.5280284	0.3099066	-1.7038306	0.0884127
0|1	0.5282537	0.1393242	3.7915427	0.0001497
1|2	2.3490725	0.3085006	7.6144827	0.0000000
2|3+	3.8876346	0.4581807	8.4849367	0.0000000

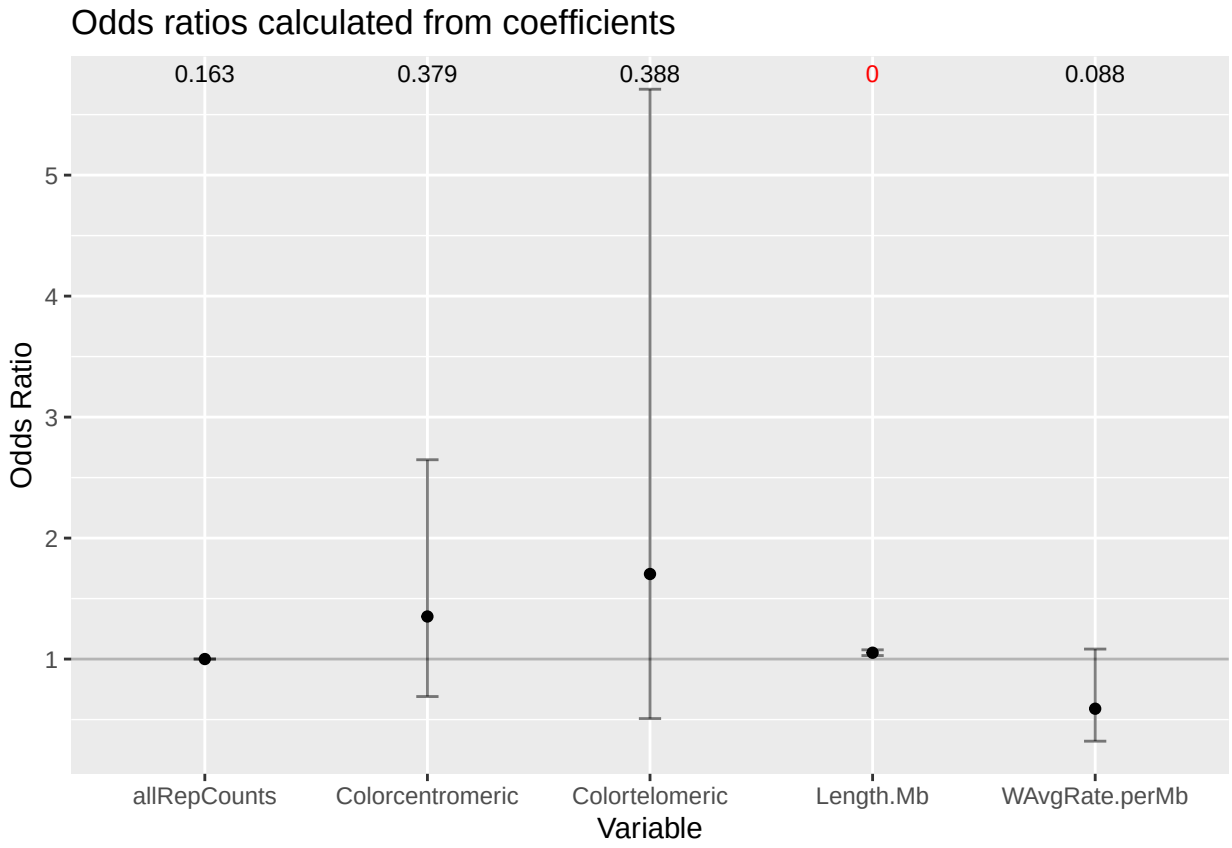
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0286349	0.0738187
allRepCounts	-0.0001726	0.0010233
Colorcentromeric	-0.3707107	0.9736044
Colortelomeric	-0.6770411	1.7422139
WAvgRate.perMb	-1.1354343	0.0793774

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	1.0525616	1.0290489	1.076612
allRepCounts	1.0004254	0.9998274	1.001024
Colorcentromeric	1.3518132	0.6902436	2.647470
Colortelomeric	1.7033322	0.5081183	5.709971
WAvgRate.perMb	0.5897666	0.3212826	1.082613

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0525616 times more likely to increase in inversion amount category."



Proportional odds assessment

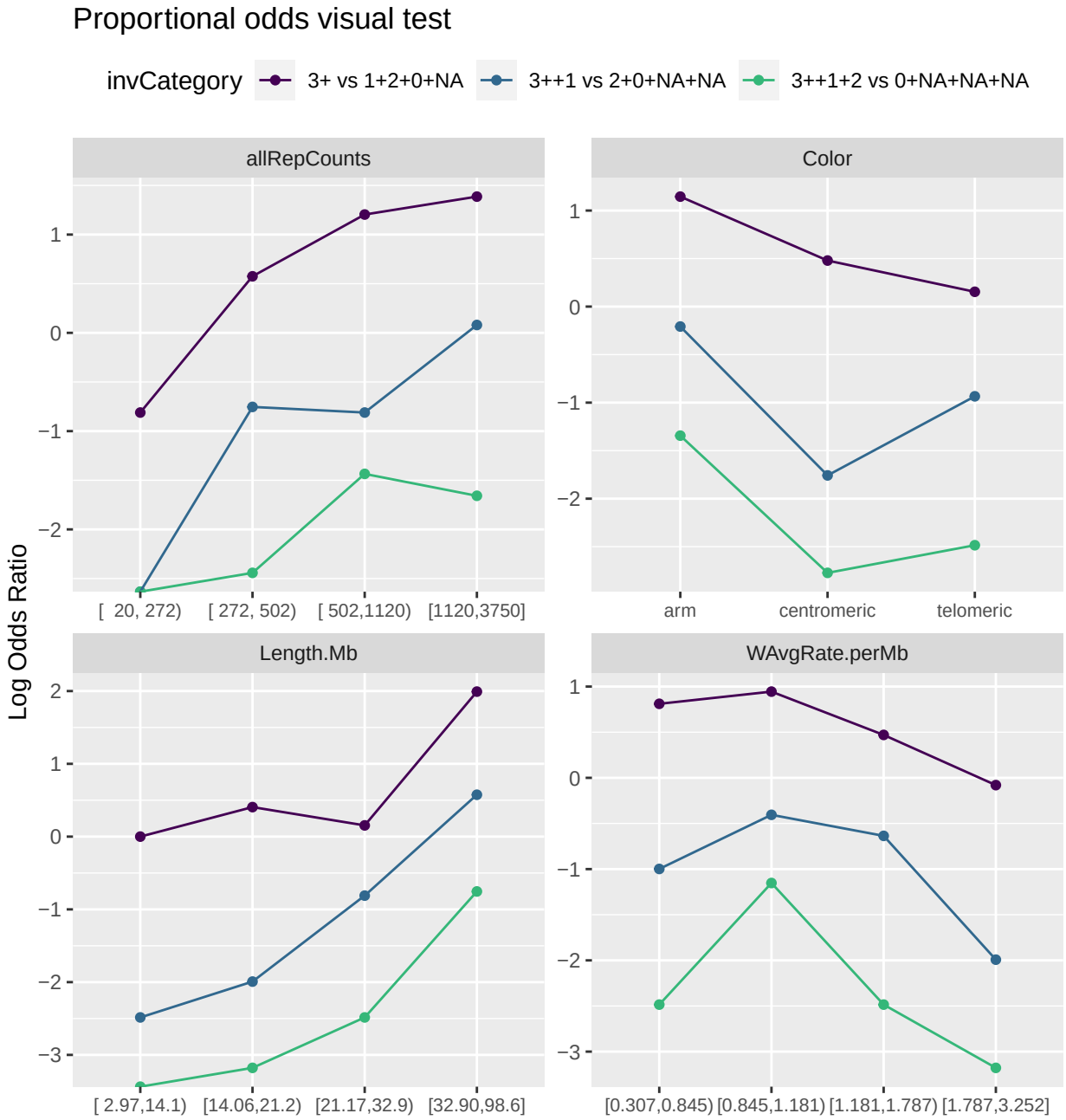
Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe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

We test the parallel regression assumption with a Brant test:

	X2	df	probability
Omnibus	10.1455607	10	0.4278170
Length.Mb	0.4219010	2	0.8098142
allRepCounts	3.7674007	2	0.1520265
Colorcentromeric	0.5246735	2	0.7692519
Colortelomeric	3.2684495	2	0.1951036

	X2	df	probability
WAvgRate.perMb	0.9840580	2	0.6113846

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.

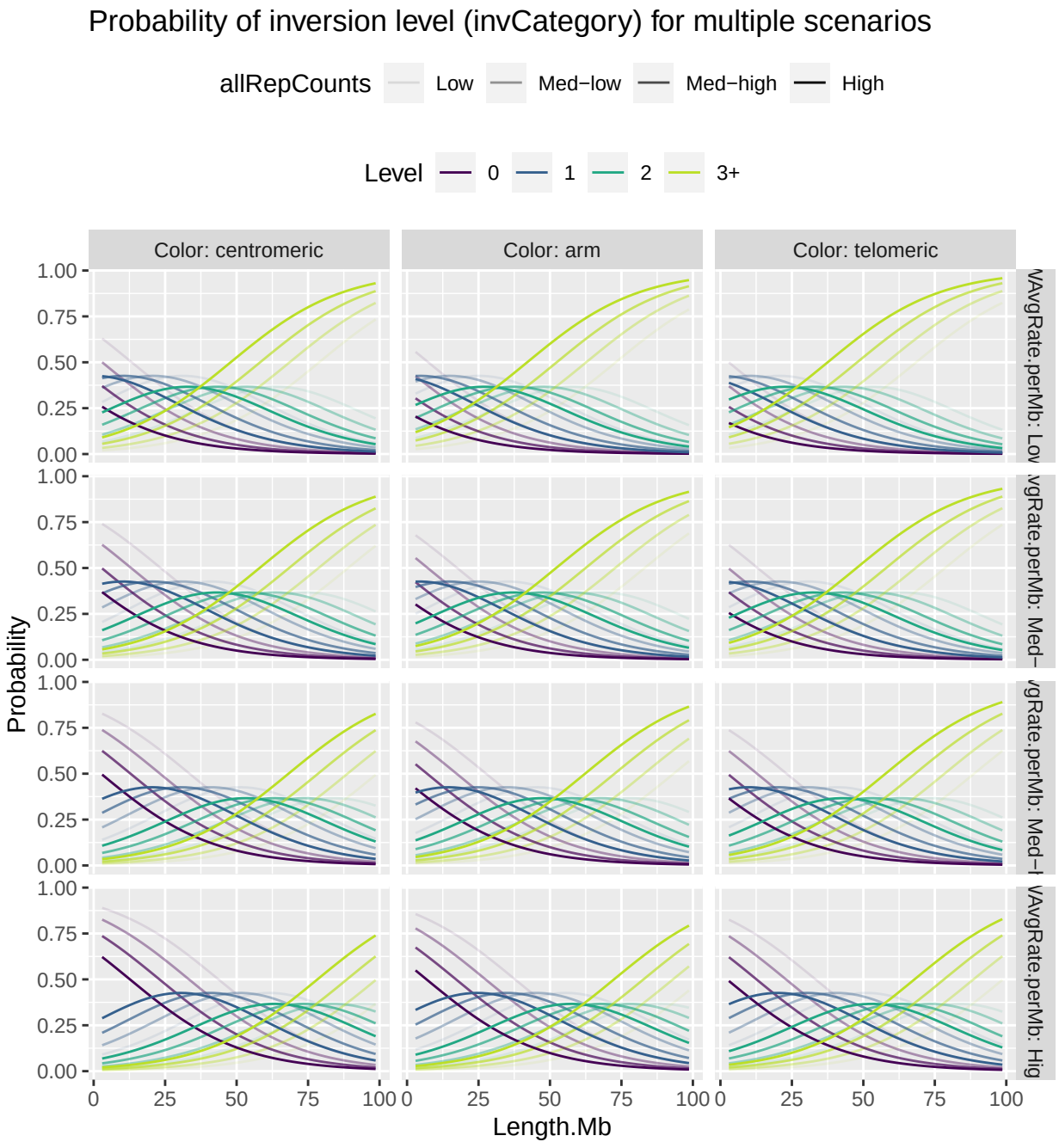


Figure 6: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error t value
## Length.Mb         0.0821705  0.0121600  6.7574
## allRepCounts     -0.0001384  0.0003289 -0.4206
## Colorcentromeric  1.3679474  0.3754413  3.6436
## Colortelomeric   -0.1655641  0.6652402 -0.2489
## WAvgRate.perMb    0.4879918  0.3486058  1.3998
##
## Intercepts:
##      Value   Std. Error t value
## 0|1   3.1020  0.1467    21.1466
## 1|2   5.1812  0.3780    13.7074
## 2|3+  6.6389  0.5709    11.6295
##
## Residual Deviance: 189.6291
## AIC: 205.6291
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0821705	0.0121600	6.7574483	0.0000000
allRepCounts	-0.0001384	0.0003289	-0.4206384	0.6740192
Colorcentromeric	1.3679474	0.3754413	3.6435716	0.0002689
Colortelomeric	-0.1655641	0.6652402	-0.2488787	0.8034546
WAvgRate.perMb	0.4879918	0.3486058	1.3998385	0.1615617
0|1	3.1019954	0.1466903	21.1465657	0.0000000
1|2	5.1812393	0.3779881	13.7074147	0.0000000
2|3+	6.6389198	0.5708704	11.6294691	0.0000000

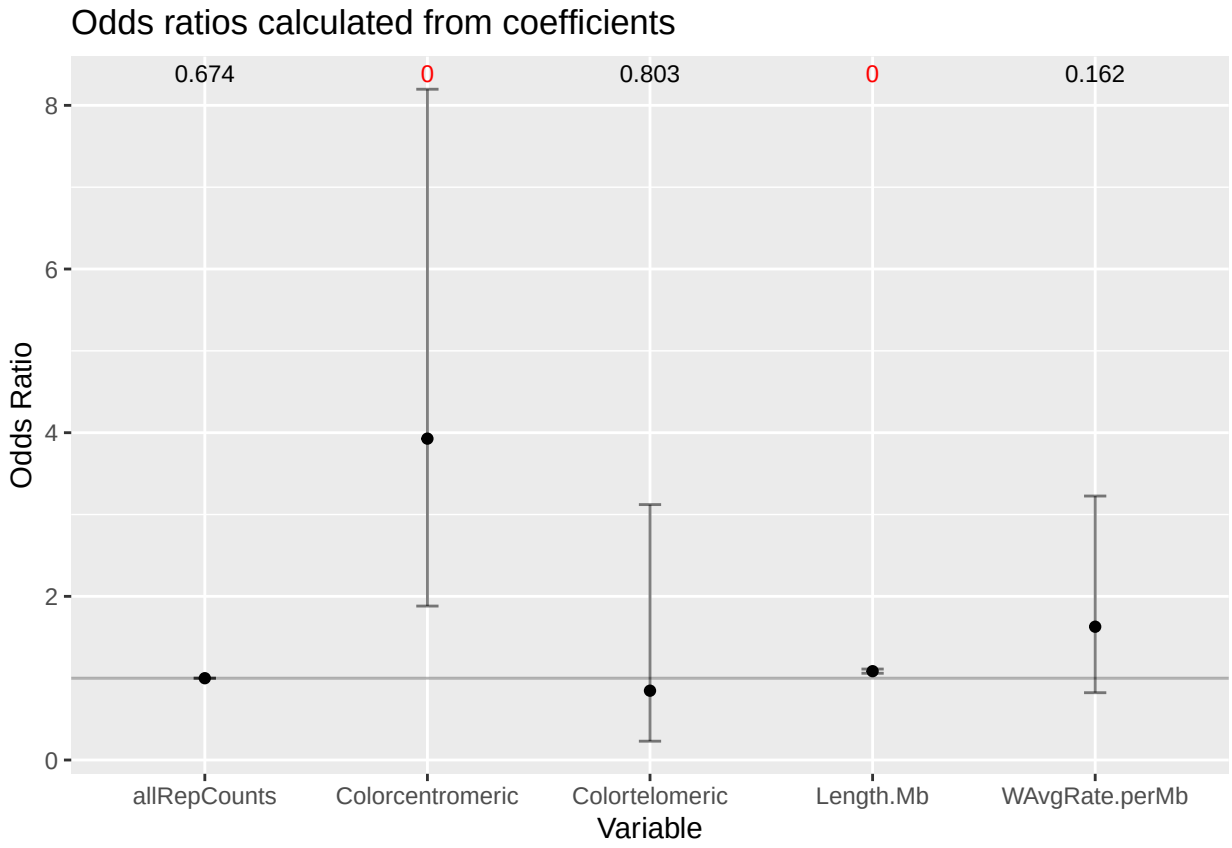
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0583374	0.1060037
allRepCounts	-0.0007831	0.0005064
Colorcentromeric	0.6320959	2.1037989
Colortelomeric	-1.4694110	1.1382827
WAvgRate.perMb	-0.1952630	1.1712467

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	1.0856409	1.0600726	1.111826
allRepCounts	0.9998616	0.9992172	1.000506
Colorcentromeric	3.9272811	1.8815499	8.197251
Colortelomeric	0.8474155	0.2300610	3.121403
WAvgRate.perMb	1.6290416	0.8226183	3.226012

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0856409 times more likely to increase in inversion amount category."



Proportional odds assessment

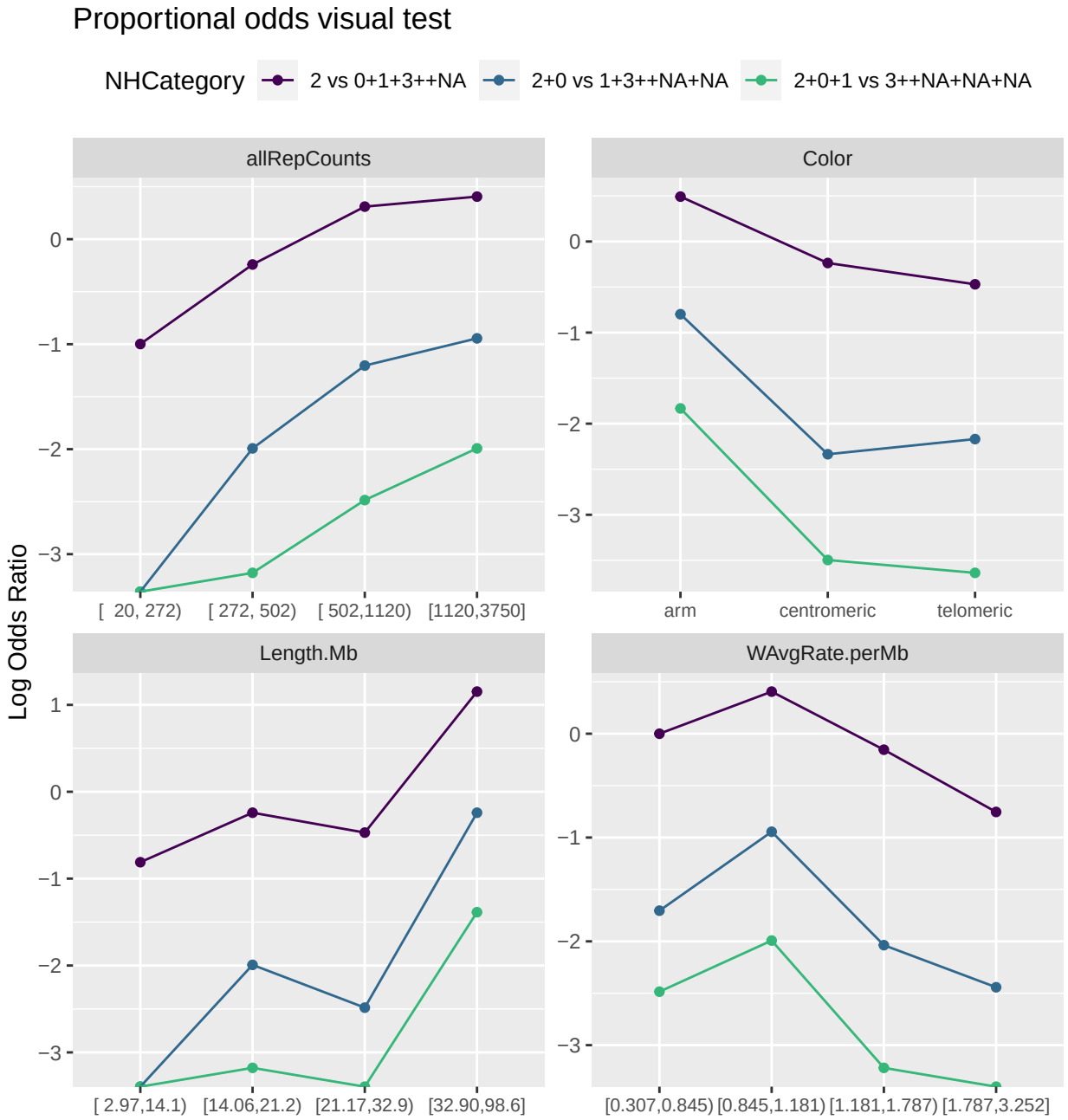
Now we should test the proportional odds or parallel regression assumption. If it is satisfied, the coefficients are valid for all the cases (i.e. the same coefficient is valid for increasing from 0 to 1 inversions, from 1 to 2, etc.). If this assumption is violated, different models are needed to describe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

We test the parallel regression assumption with a Brant test:

	X2	df	probability
Omnibus	1.3865641	10	0.9992473
Length.Mb	0.0923946	2	0.9548536
allRepCounts	0.1140866	2	0.9445532
Colorcentromeric	0.1116834	2	0.9456888
Colortelomeric	0.1790885	2	0.9143478

	X2	df	probability
WAvgRate.perMb	0.6416531	2	0.7255491

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.

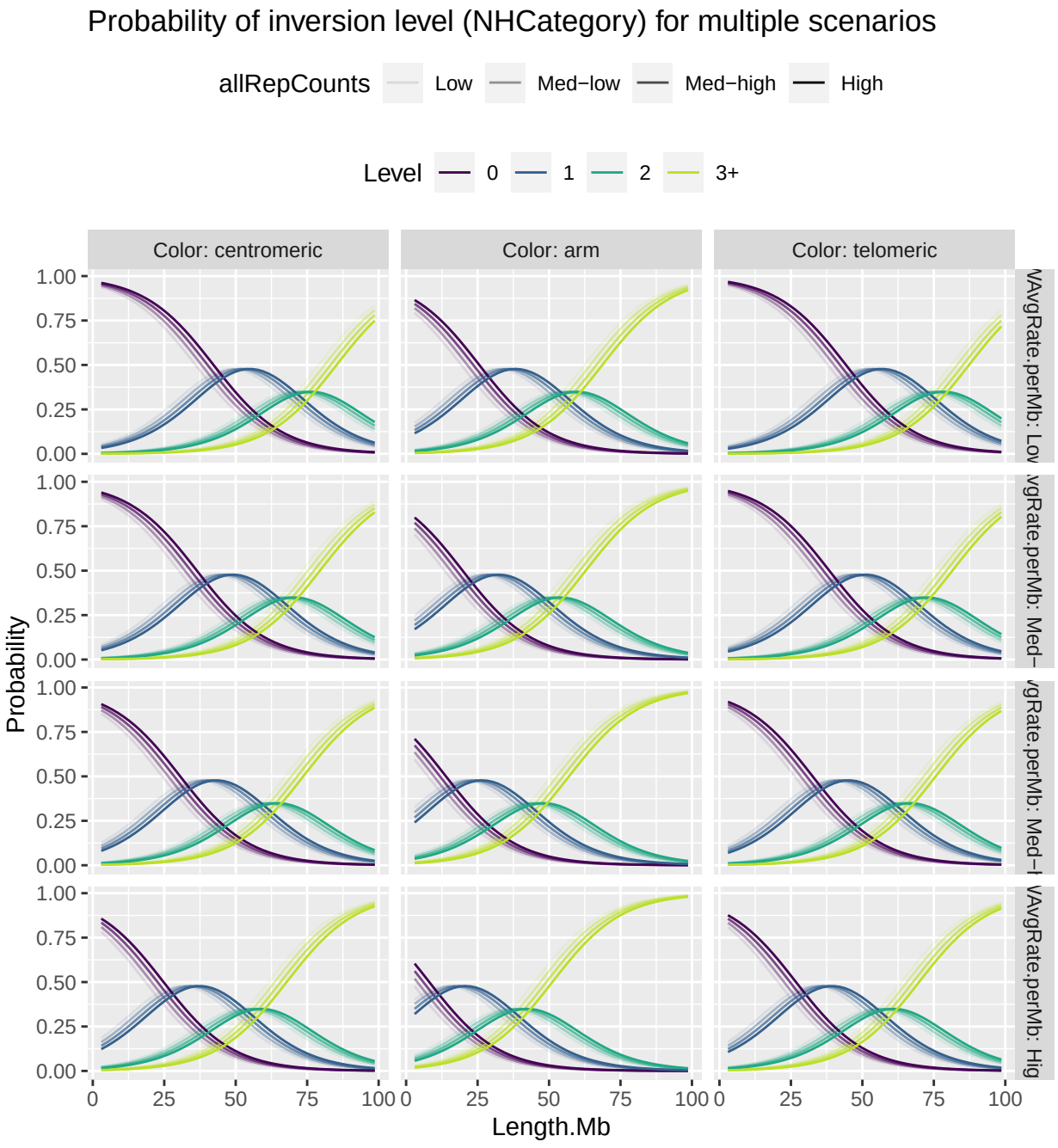


Figure 7: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                      Value Std. Error t value
## Length.Mb        -0.016494  0.0117780  -1.400
## allRepCounts      0.000742  0.0003632   2.043
## Colorcentromeric -1.387766  0.4536458  -3.059
## Colortelomeric    1.386489  0.7355248   1.885
## WAvgRate.perMb   -1.811670  0.3957835  -4.577
##
## Intercepts:
##     Value   Std. Error t value
## 0|1 -1.2545  0.2039    -6.1536
## 1|2  0.6749  0.4431     1.5233
##
## Residual Deviance: 145.4837
## AIC: 159.4837
```

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

	Value	Std. Error	t value	p value
Length.Mb	-0.0164936	0.0117780	-1.400373	0.1614017
allRepCounts	0.0007420	0.0003632	2.043255	0.0410272
Colorcentromeric	-1.3877657	0.4536458	-3.059139	0.0022197
Colortelomeric	1.3864887	0.7355248	1.885033	0.0594254
WAvgRate.perMb	-1.8116695	0.3957835	-4.577426	0.0000047
0|1	-1.2545096	0.2038664	-6.153586	0.0000000
1|2	0.6749200	0.4430593	1.523318	0.1276793

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.

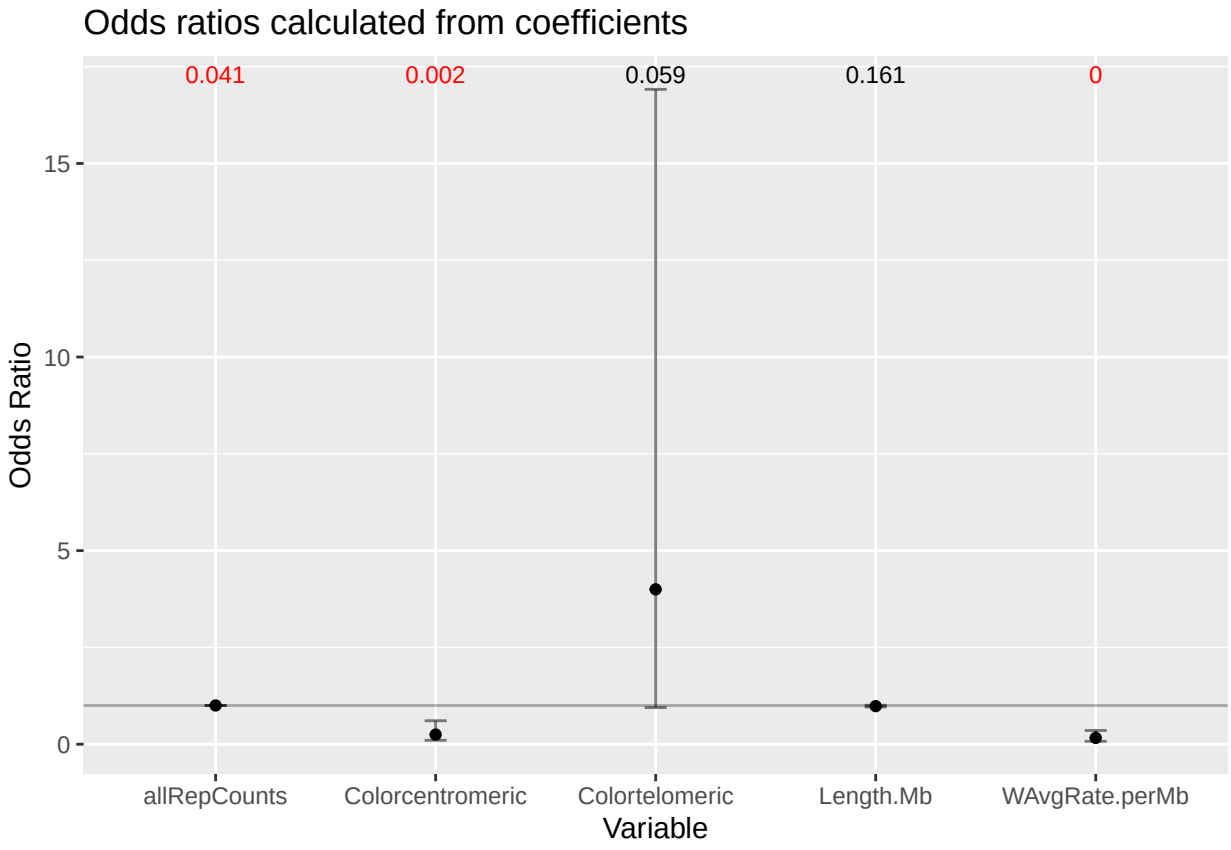
	2.5 %	97.5 %
Length.Mb	-0.0395780	0.0065909
allRepCounts	0.0000302	0.0014538
Colorcentromeric	-2.2768952	-0.4986362
Colortelomeric	-0.0551134	2.8280909
WAvgRate.perMb	-2.5873909	-1.0359482

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	0.9836417	0.9611950	1.0066126

	Odds Ratio	2.5%	97.5%
allRepCounts	1.0007423	1.0000302	1.0014548
Colorcentromeric	0.2496324	0.1026023	0.6073584
Colortelomeric	4.0007775	0.9463778	16.9131404
WAvgRate.perMb	0.1633811	0.0752160	0.3548897

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 0.9836417 times more likely to increase in inversion amount category."



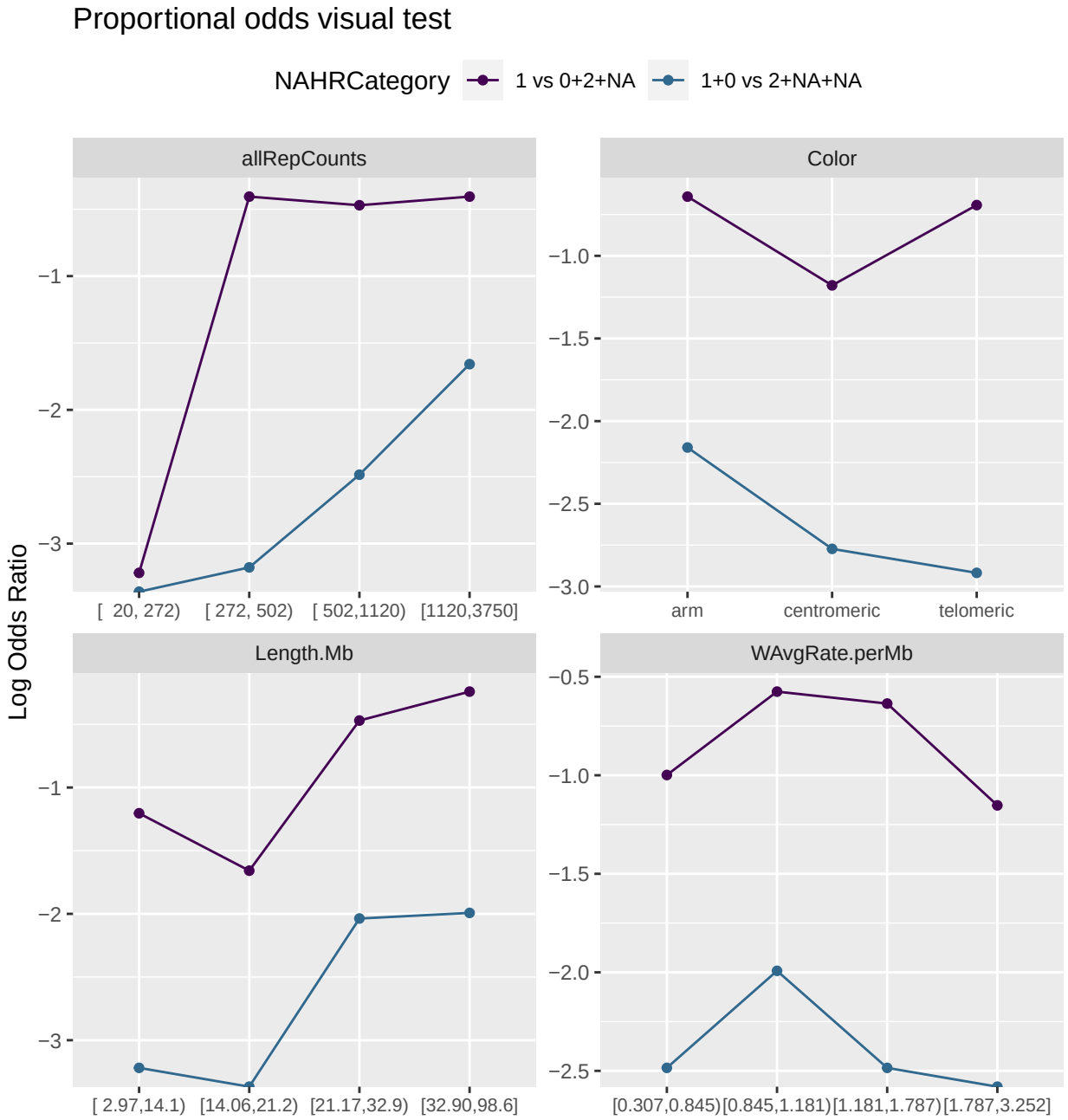
Proportional odds assessment

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:

	X2	df	probability
Omnibus	1.6632646	5	0.8934949
Length.Mb	0.4822376	1	0.4874105
allRepCounts	0.6150035	1	0.4329101
Colorcentromeric	0.4410626	1	0.5066101
Colortelomeric	0.0001382	1	0.9906216
WAvgRate.perMb	0.4516092	1	0.5015718

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.

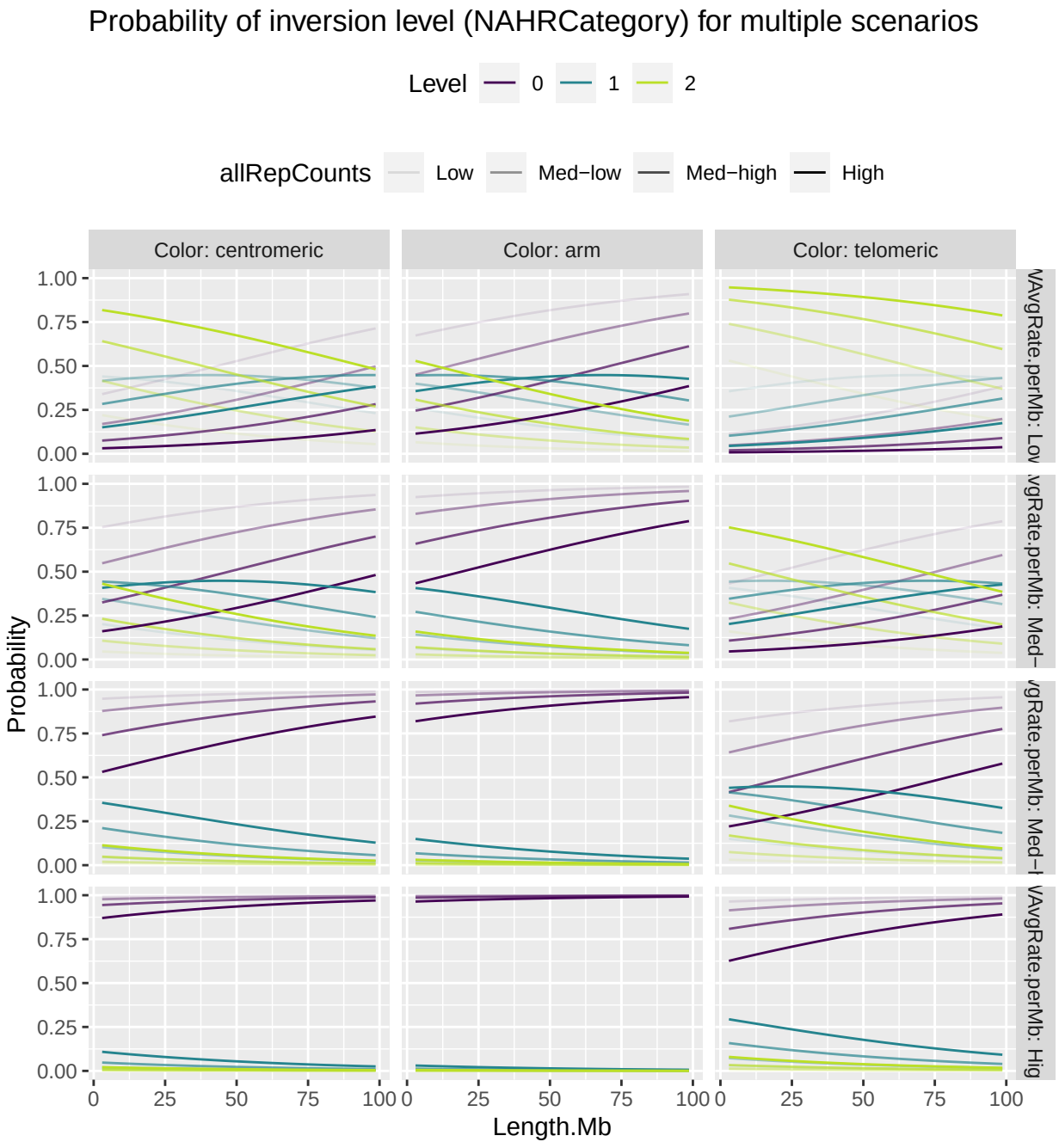


Figure 8: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

Descriptive categories

Descriptive statistics

Raw data:

Chromosome	Start	End	Color	invCenters	NHCenters	NAHRCenters	Length.Mb	allRepCount	WAvgRate.perMb
chr10	158946	23770709	telomeric	3	2	1	23.61176	354	1.8516302
chr10	23770709	39097912	centromeric	1	0	1	15.32720	880	1.0408405
chr10	59958908	116142800	arm	2	2	0	56.18389	818	0.9967980
chr10	116142800	135473442	telomeric	1	1	0	19.33064	168	2.0570277
chr10	42436301	59958908	centromeric	2	2	0	17.52261	1678	0.5819161
chr11	241489	29941780	telomeric	2	1	1	29.70029	746	1.5167467

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

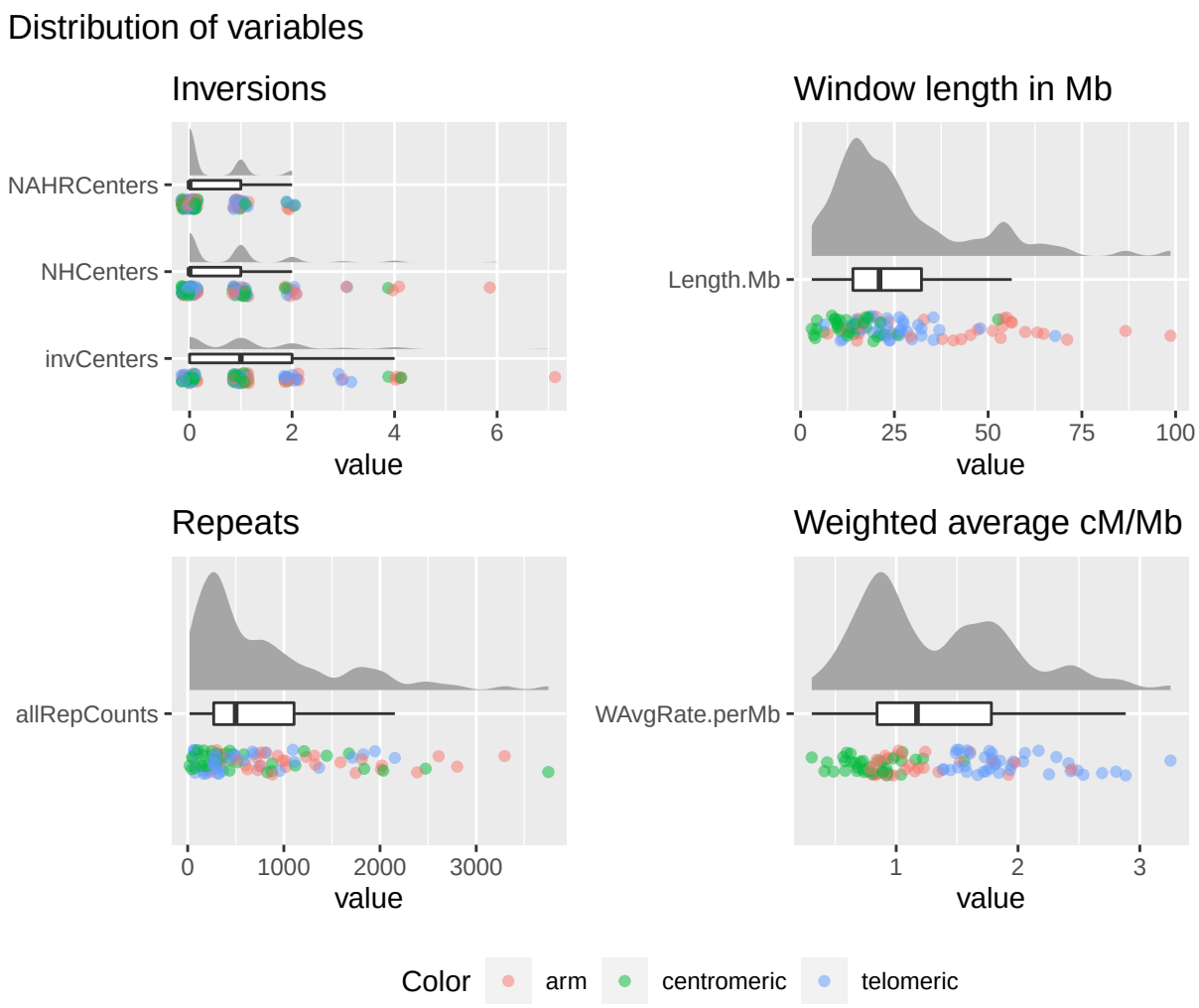


Figure 9: Distribution of variables.

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

Table 19: Original counts

CountGroups	invCenters	NHCenters	NAHRCenters
0	38	54	71
1	35	32	24
2	18	10	7
3	4	2	NA
4	6	3	NA
6	NA	1	NA
7	1	NA	NA

Table 20: New counts

	CountGroups	invCategory	NHCategory	NAHRCategory
1	Absence	38	54	71

	CountGroups	invCategory	NHCategory	NAHRCategory
3	Presence	53	42	31
2	Abundance	11	6	0

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

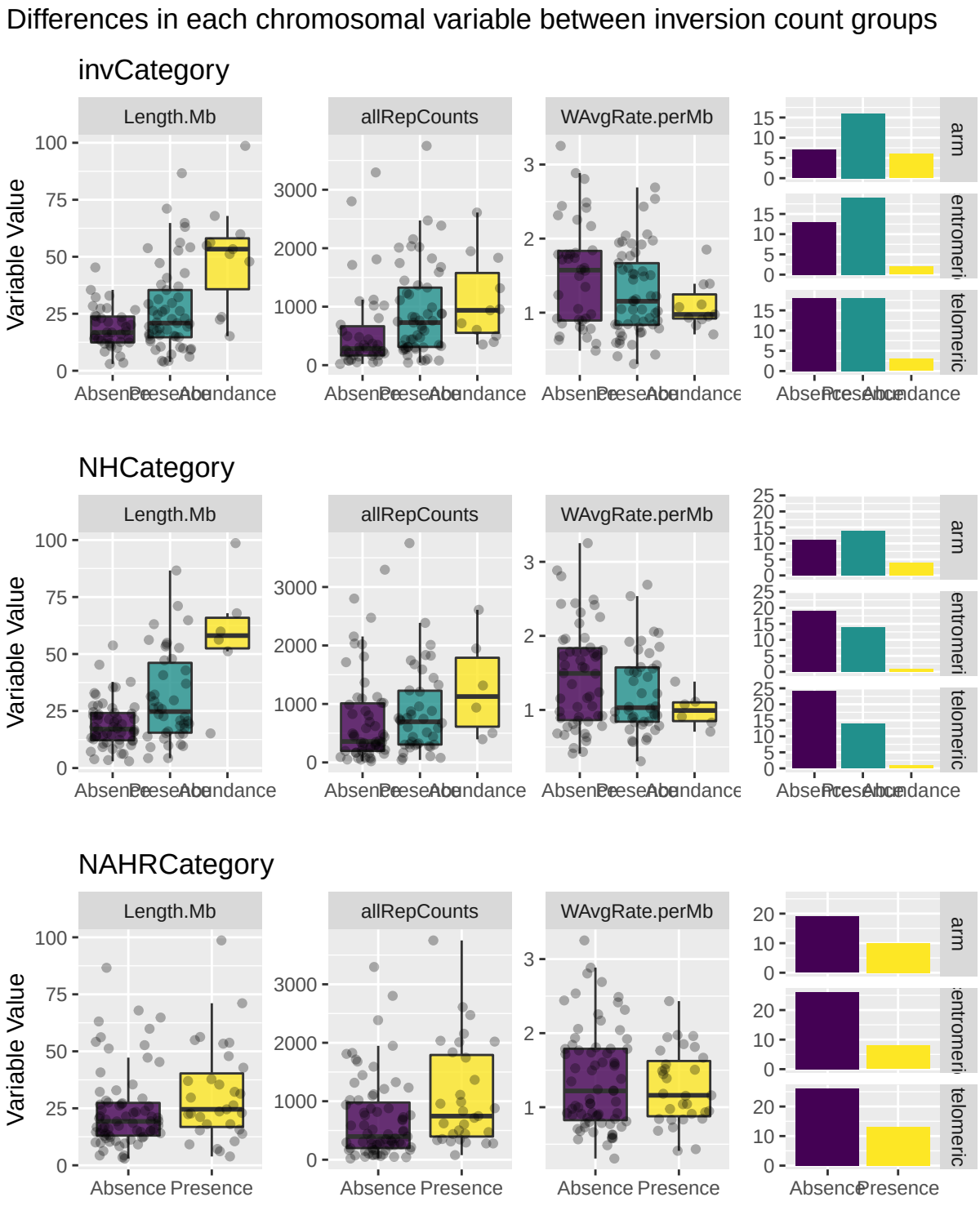


Figure 10: Potential effect of independent variables on the different types of invesions.

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

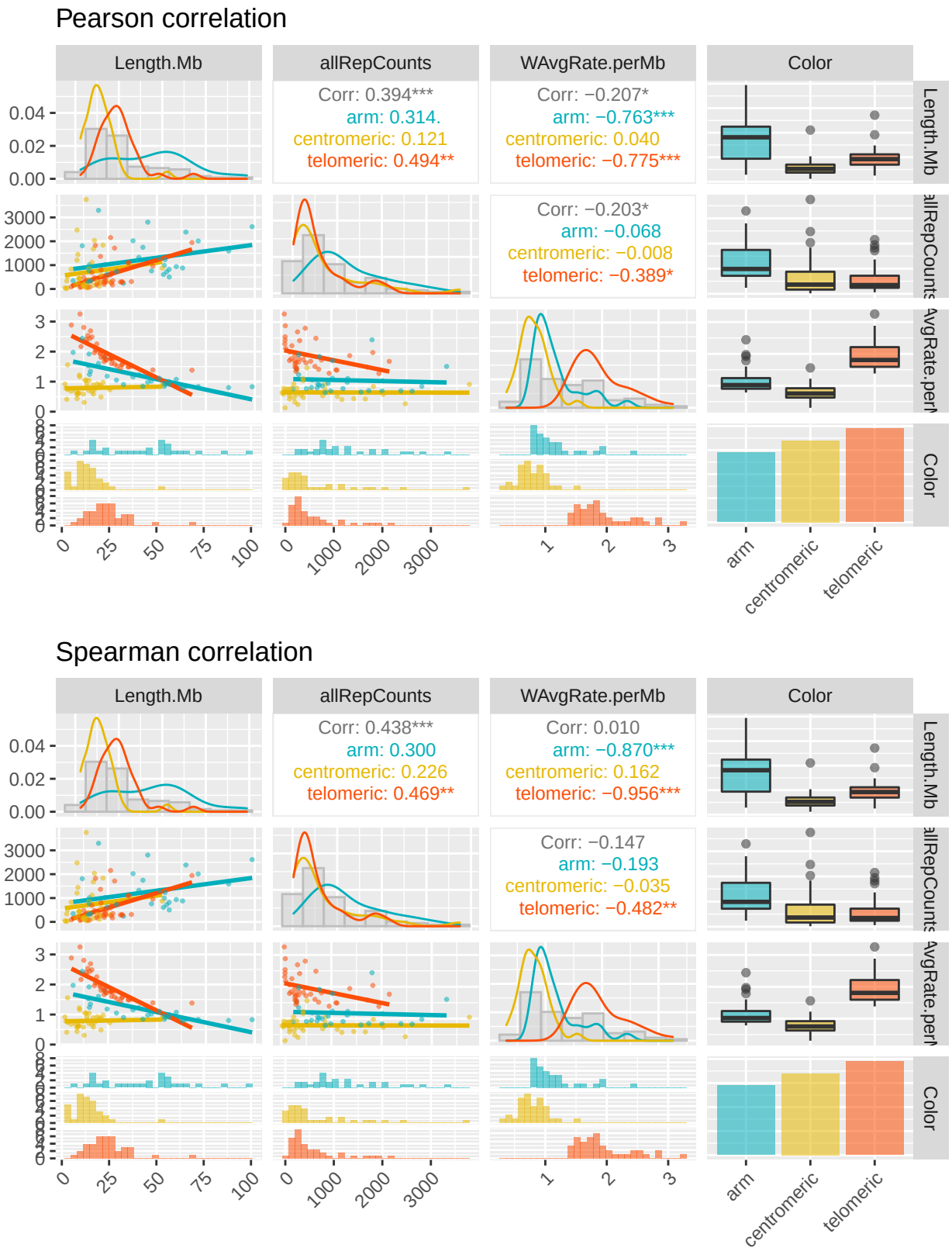


Figure 11: Correlations between variables.

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

	GVIF	Df	GVIF^(1/(2*Df))
Length.Mb	2.713019	1	1.647124
allRepCounts	1.248267	1	1.117259
Color	6.917414	2	1.621758
WAvgRate.perMb	4.344256	1	2.084288

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.	2.969812	-1.2438876	20.000	-1.0149666	0.3068634	-1.6361728
1st Qu.	13.949468	-0.6574034	270.500	-0.6908557	0.8424562	-0.7894017
Median	21.043519	-0.2784709	499.000	-0.3952096	1.1708581	-0.2701990
Mean	26.256815	0.0000000	804.451	0.0000000	1.3417622	0.0000000
3rd Qu.	32.224751	0.3187804	1106.000	0.3901610	1.7813523	0.6949911
Max.	98.630850	3.8658972	3750.000	3.8111162	3.2519378	3.0199836

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

Total inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error t value
## Length.Mb         0.0589978  0.0126416  4.6669
## allRepCounts      0.0001951  0.0003129  0.6237
## Colorcentromeric  0.7040575  0.3586947  1.9628
## Colortelomeric    0.1081040  0.6461603  0.1673
## WAvgRate.perMb   -0.1728854  0.3236694 -0.5341
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence     1.0377  0.1474     7.0402
## Presence|Abundance   4.3360  0.4796     9.0413
##
## Residual Deviance: 168.0213
## AIC: 182.0213
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0589978	0.0126416	4.6669494	0.0000031
allRepCounts	0.0001951	0.0003129	0.6237340	0.5328023
Colorcentromeric	0.7040575	0.3586947	1.9628324	0.0496656
Colortelomeric	0.1081040	0.6461603	0.1673022	0.8671323
WAvgRate.perMb	-0.1728854	0.3236694	-0.5341421	0.5932432
Absence|Presence	1.0376613	0.1473915	7.0401716	0.0000000
Presence|Abundance	4.3359640	0.4795724	9.0413132	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.

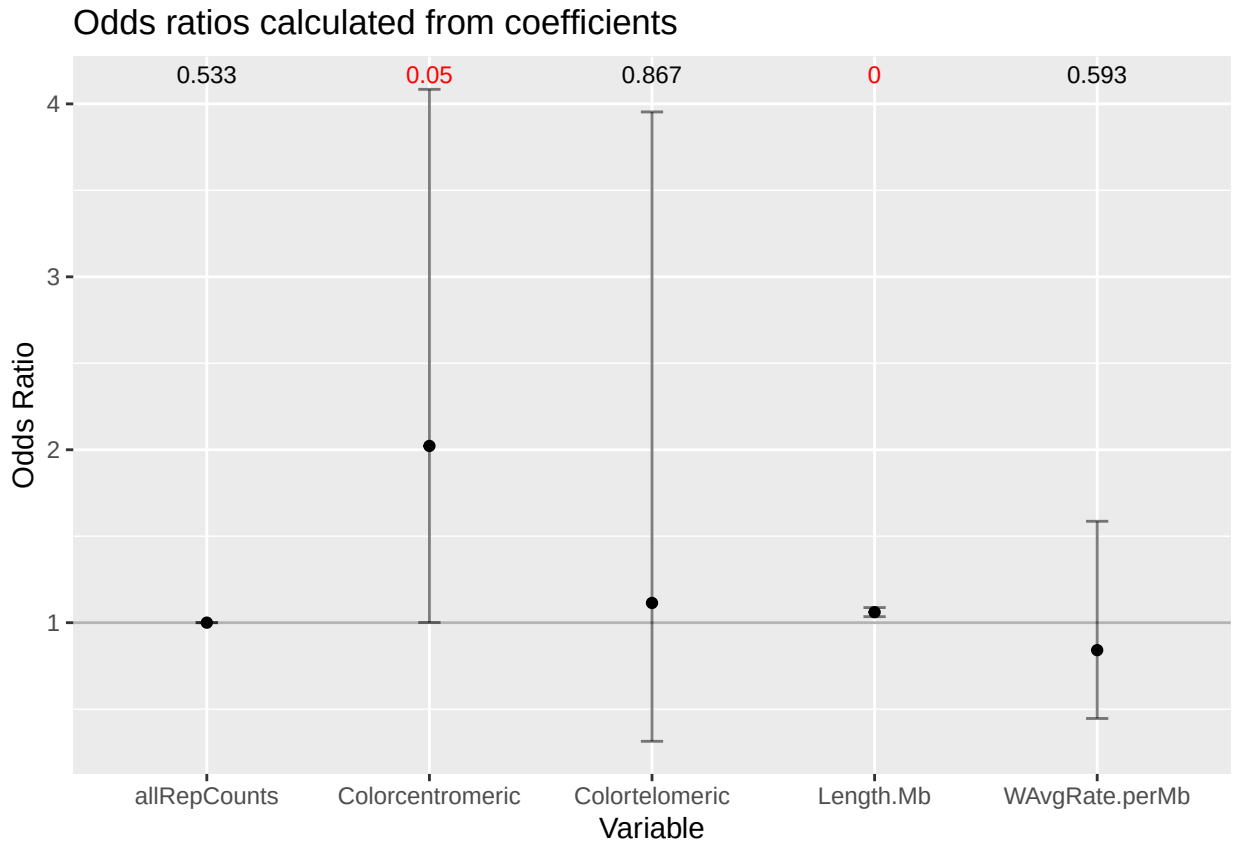
	2.5 %	97.5 %
Length.Mb	0.0342207	0.0837749
allRepCounts	-0.0004181	0.0008084
Colorcentromeric	0.0010289	1.4070861
Colortelomeric	-1.1583469	1.3745550
WAvgRate.perMb	-0.8072657	0.4614949

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	1.060773	1.0348129	1.087384

	Odds Ratio	2.5%	97.5%
allRepCounts	1.000195	0.9995820	1.000809
Colorcentromeric	2.021940	1.0010294	4.084038
Colortelomeric	1.114164	0.3140048	3.953317
WAvgRate.perMb	0.841234	0.4460761	1.586444

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0607729 times more likely to increase in inversion amount category."



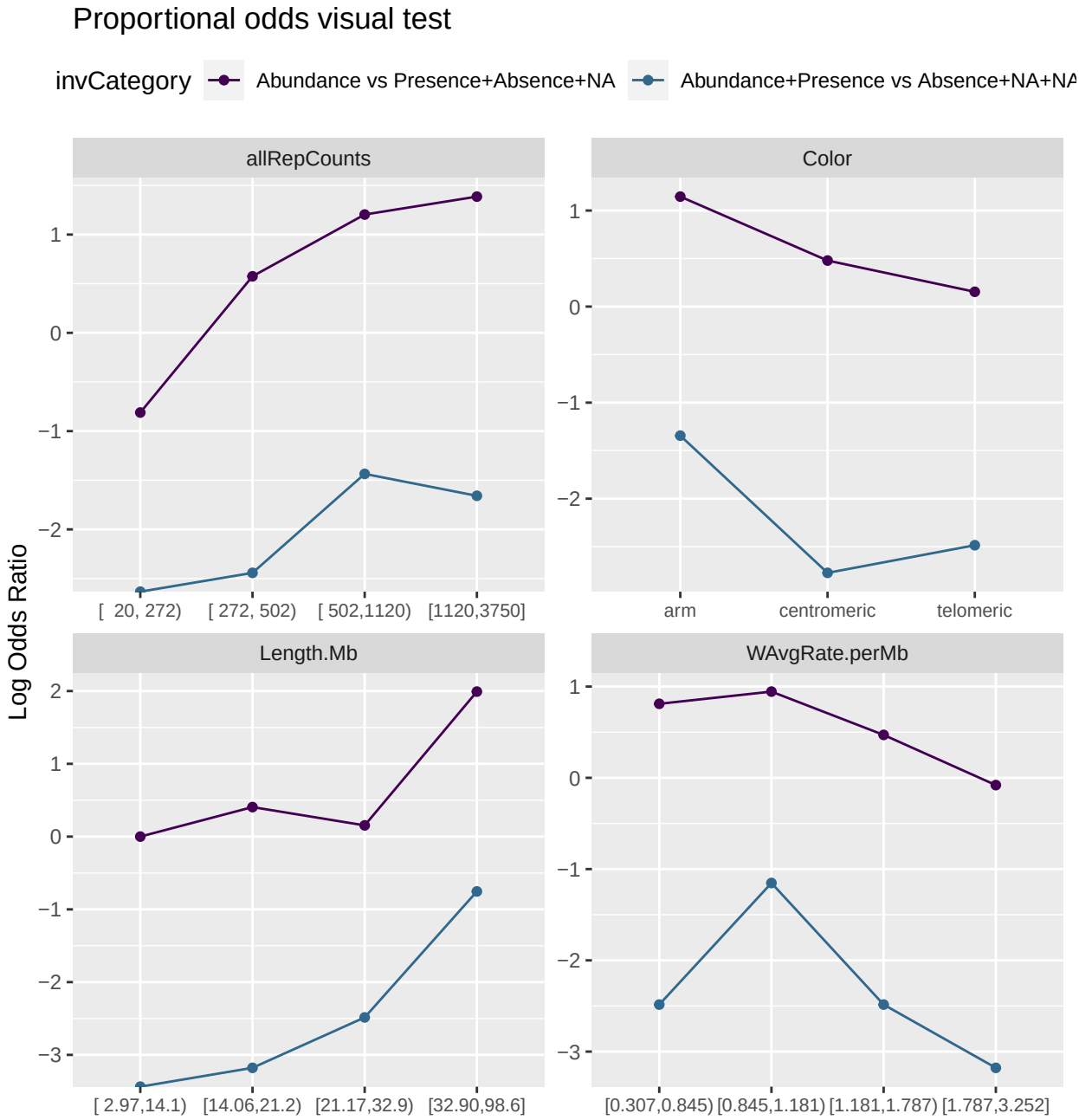
Proportional odds assessment

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:

	X2	df	probability
Omnibus	1.5918351	5	0.9022350
Length.Mb	0.3656788	1	0.5453692
allRepCounts	1.2042015	1	0.2724835
Colorcentromeric	0.0328914	1	0.8560850
Colortelomeric	0.1521839	1	0.6964570
WAvgRate.perMb	0.0000008	1	0.9992896

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.

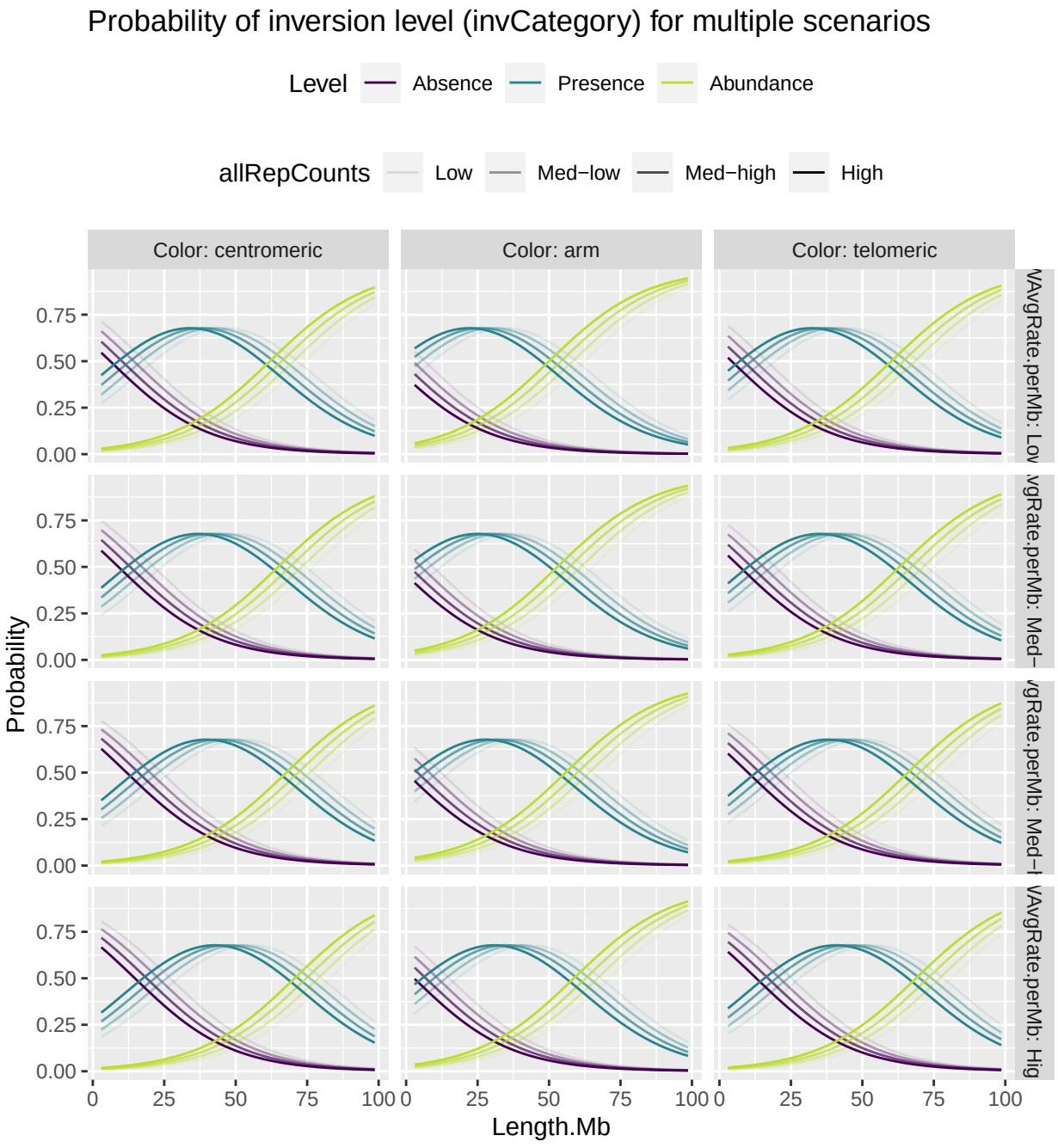


Figure 12: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error t value
## Length.Mb         0.0857681  0.0136976  6.2616
## allRepCounts     -0.0001485  0.0003459 -0.4292
## Colorcentromeric  1.4942540  0.3845879  3.8853
## Colortelomeric   -0.1429881  0.7004855 -0.2041
## WAvgRate.perMb    0.4766870  0.3614396  1.3189
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence     3.2222  0.1519    21.2100
## Presence|Abundance   6.8531  0.6347    10.7969
##
## Residual Deviance: 146.5986
## AIC: 160.5986
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0857681	0.0136976	6.2615571	0.0000000
allRepCounts	-0.0001485	0.0003459	-0.4292159	0.6677661
Colorcentromeric	1.4942540	0.3845879	3.8853383	0.0001022
Colortelomeric	-0.1429881	0.7004855	-0.2041272	0.8382541
WAvgRate.perMb	0.4766870	0.3614396	1.3188567	0.1872170
Absence|Presence	3.2222224	0.1519203	21.2099506	0.0000000
Presence|Abundance	6.8530757	0.6347271	10.7968853	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.

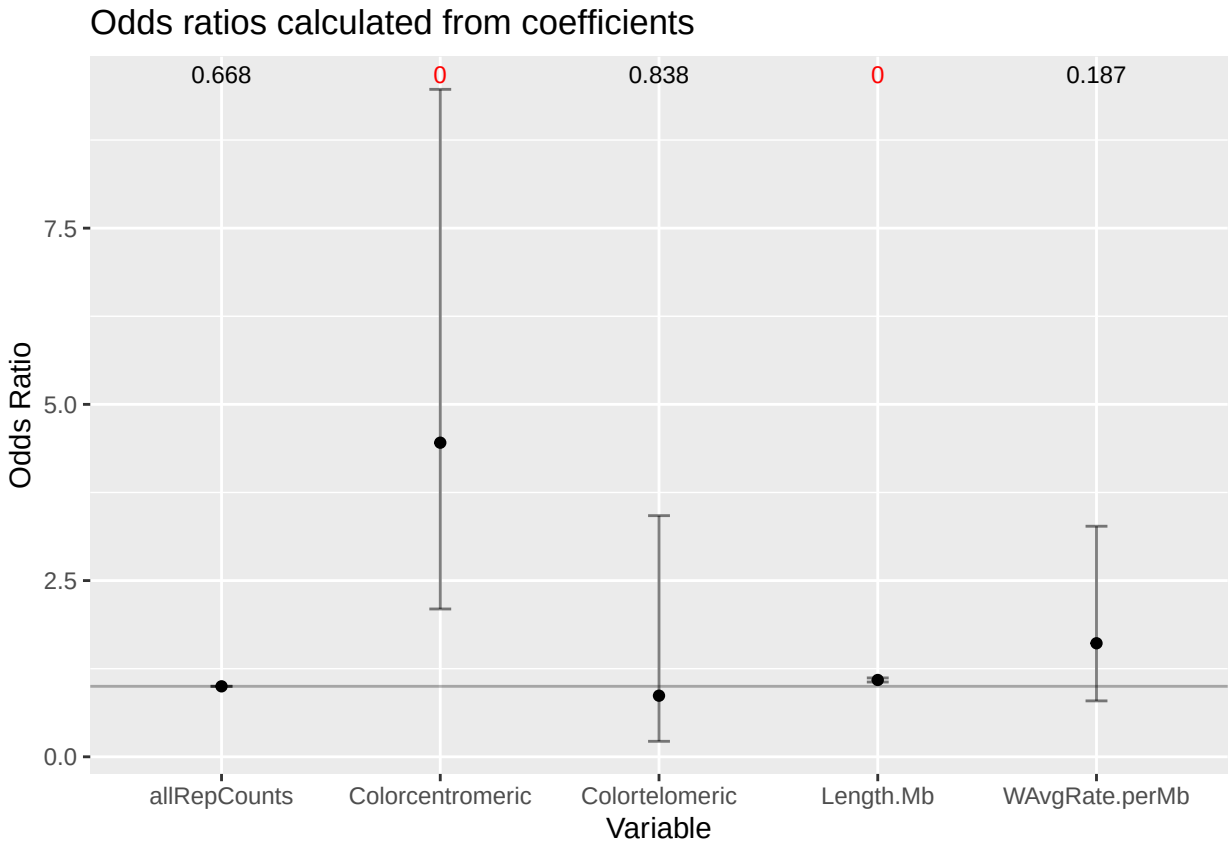
	2.5 %	97.5 %
Length.Mb	0.0589213	0.1126148
allRepCounts	-0.0008264	0.0005295
Colorcentromeric	0.7404756	2.2480324
Colortelomeric	-1.5159145	1.2299382
WAvgRate.perMb	-0.2317215	1.1850955

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	1.0895536	1.0606918	1.119201

	Odds Ratio	2.5%	97.5%
allRepCounts	0.9998515	0.9991739	1.000530
Colorcentromeric	4.4560110	2.0969326	9.469086
Colortelomeric	0.8667644	0.2196073	3.421018
WAvgRate.perMb	1.6107292	0.7931670	3.270999

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0895536 times more likely to increase in inversion amount category."



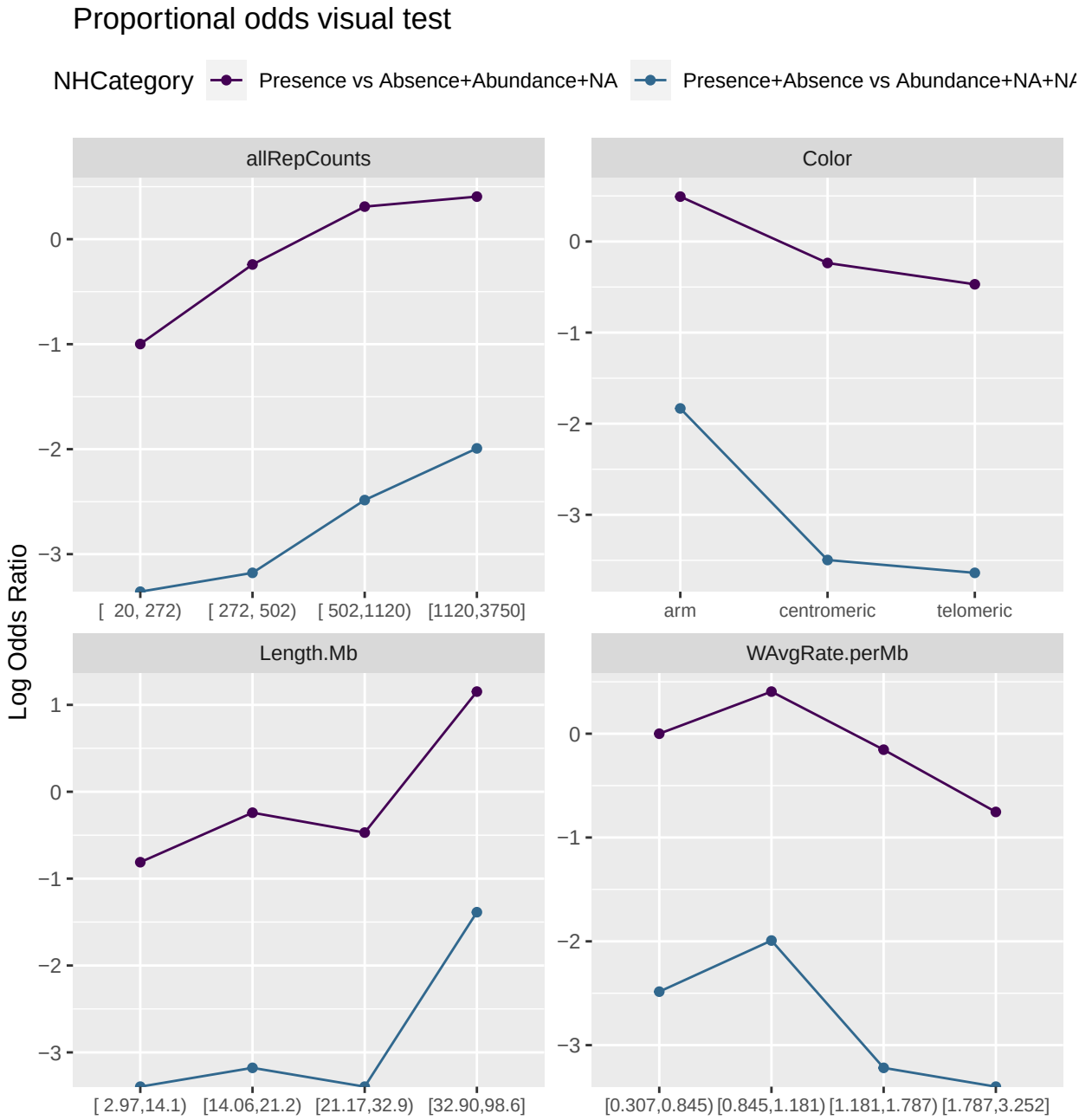
Proportional odds assessment

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:

	X2	df	probability
Omnibus	0.2296921	5	0.9987606
Length.Mb	0.0130211	1	0.9091504
allRepCounts	0.1128303	1	0.7369446
Colorcentromeric	0.0891067	1	0.7653158
Colortelomeric	0.0151338	1	0.9020916
WAvgRate.perMb	0.0665434	1	0.7964378

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.

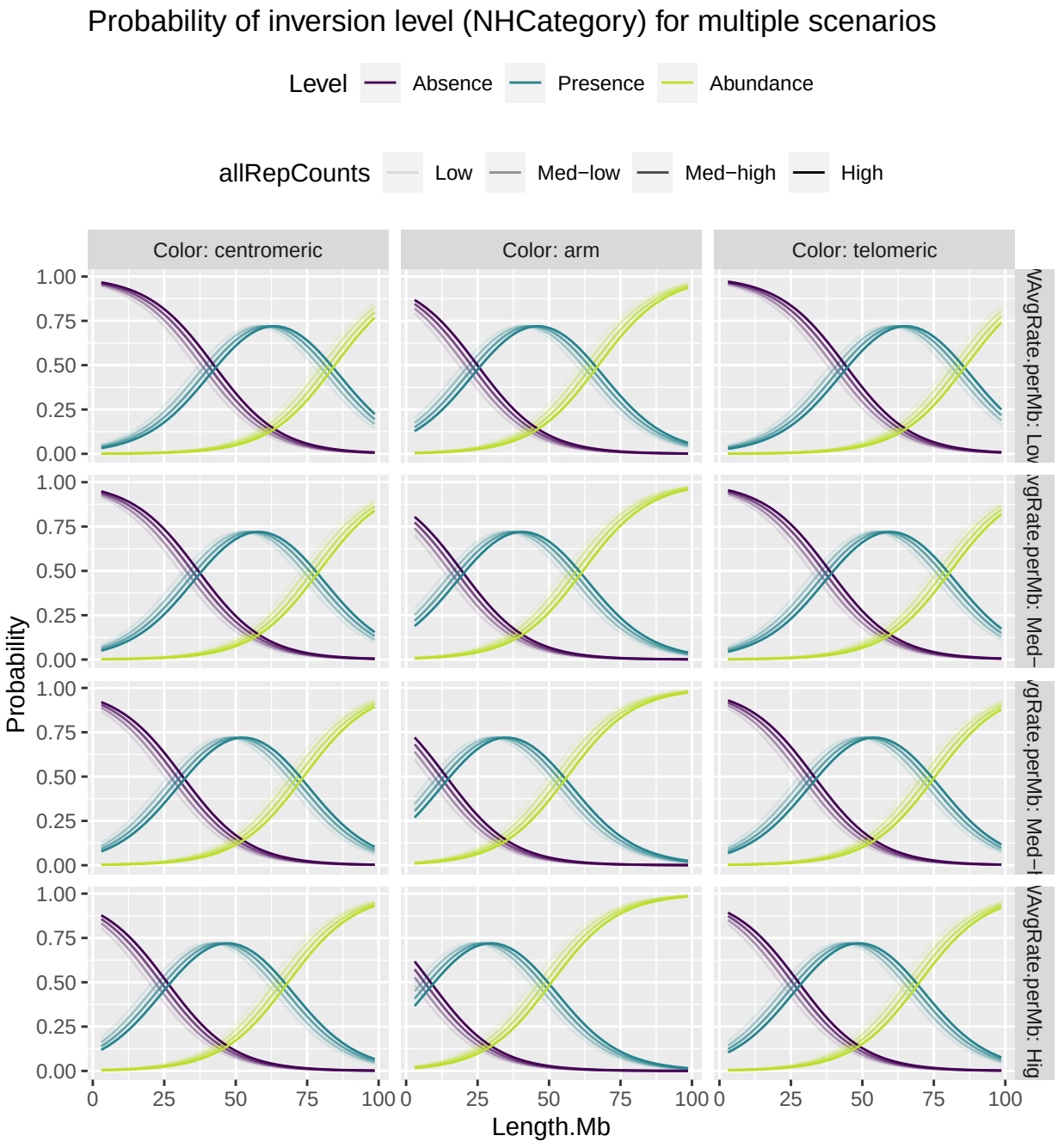


Figure 13: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

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